

BEYOND INTUITION: REDEFINING INTERNAL AUDIT
RISK ASSESSMENTS THROUGH MACHINE LEARNING
PARADIGMS

VIVEK SHIVRAM

Professional Doctorate - Data Science

Institute of Computing Science and Mathematics

University of Stirling

February 2025

ABSTRACT

Established literature has found that Internal Audit Functions (IAF) increasingly view AI-powered innovation as being indispensable to continue staying relevant in the light of increased organisational complexity and regulatory expectations.

This thesis explores opportunities to deploy Machine Learning (ML) capabilities to assess how IAFs can benefit from increased precision, objectivity and efficiency for delivering internal audit risk assessments. Specifically, this thesis explores how the risk assessment process, a fundamental component of the audit lifecycle, can be redefined through ML to generate measurable value for the IAF. In doing so, this thesis makes several researched assumptions and acknowledges the impact of relevant limitations.

Three research questions were framed for evaluation throughout the thesis, to explore the relevance and findings of prevailing literature, identify suitable ML techniques for practical implementation, and measure value from any proposed implementations. In doing so, the thesis supplements the findings of prior academic research, whilst laying the groundwork for further practical improvement by other IAFs.

The thesis applied multiple ML techniques to answer these research questions, combining academic ML principles with real-world datasets and implementation techniques, offering a meaningful starting point for IAFs looking to improve their approach to internal audit risk assessments. Three rounds of experiments were performed to reframe risk assessments as specific problem statements for ML models, supported by (1) an examination of the extant risk assessment methodology used by the Senior Leadership Team (SLT) at the subject organisation, a world-leading financial services firm, (2) an analysis of organisational data relevant to the risk assessment, anonymised for the purposes of this thesis, and (3) the construction and application of several ML hypotheses on this dataset. The conclusions of these experiments were consolidated to answer the research questions, including an assessment of how the value arising from such efforts could be measured for continued sponsorship and improvement.

Overall, the thesis noted that there is no single best ML approach for practical implementation in all cases, and that the choice of ML models would vary based on the quality of underlying data, consistent audit methodologies, as well as the size of the dataset including the final set of input features. The thesis found that a supervised machine learning approach was most suitable for internal audit risk assessments, given the high class-specific

precision ratio of the final model (c.88%) on the test dataset. However, the thesis noted significant dependencies on high-quality datasets and a requirement to carefully curate relevant features to ensure meaningful model outputs. Although the thesis evaluated unsupervised machine learning techniques, these were discounted from practical implementation due to comparatively lower accuracy, inhibiting practical implementation for the given dataset. However, the thesis also noted that such techniques may generate better results in alternative practical settings, where significantly larger datasets are used for risk assessment cycles.

ETHICAL APPROVAL

I have obtained formal approval for the ethical approaches of my project [Ethics Reference Number EC 2024 18600 13483] prior to proceeding with the research work undertaken.

Stirling, February 2025

Vivek Shivram

DECLARATION: USAGE OF GENERATIVE ARTIFICIAL
INTELLIGENCE

I confirm that generative artificial intelligence software has not been used in the generation of data and /or writing the final text submitted as part of this thesis.

Stirling, February 2025

Vivek Shivram

DECLARATION: PLAGIARISM

I understand the nature of plagiarism, and I am aware of the university's policy on this. I certify that this thesis represents my original work, undertaken as part of my University's doctoral programme. No additional collaborative work has been included in this thesis. References have been cited per the University's academic policy requirements. I consent to the use of plagiarism detection software prior to the formal submission of my thesis.

Stirling, February 2025

Vivek Shivram

ACKNOWLEDGMENTS

This academic journey has been a thrilling adventure and I am immensely grateful for the opportunity afforded by the University of Stirling for me to explore the fascinating and ever evolving world of Machine Learning in the context of Internal Audit, my profession of 18 years.

More than my individual contribution for a University programme, this thesis has inspired me to explore a new chapter in my professional life. As I look forward to my career beyond the professional doctorate, I hope to formally pursue opportunities to share the practical knowledge I've gathered through the doctoral programme with my colleagues in the Internal Audit profession, as they seek to harness the capabilities of Data Science in their roles. I truly believe that Data Science has the potential to transform Internal Audit from being a compliance function to a genuine organisational change agent, and I'm grateful to have had a wholesome opportunity to learn about and apply the various machine learning pathways for enabling this transformation.

This work would not have been possible without my supervisor David Cairns' unwavering support, guidance, and seemingly endless patience with my draft! Thank you Dr. Cairns, for taking your time with this random industry professional as he navigated the quirks of academia. Your counsel and encouragement has made this thesis possible.

I am also eternally grateful to my wife Anju, and my children, Siddharth and Samriddhi, for all their support and encouragement over the last two years, without which none of this would have been possible. All the late evenings and weekends I've spent working on my doctorate were rightfully theirs to enjoy with me, and I look forward to making up for lost time.

LIST OF PUBLICATIONS

1. **Position Paper** - Shivram, V. (2024). AUDITING WITH AI: A THEORETICAL FRAMEWORK FOR APPLYING MACHINE LEARNING ACROSS THE INTERNAL AUDIT LIFECYCLE. EDPACS, 69(1), 22–40. <https://doi.org/10.1080/07366981.2024.2312025>

CONTENTS

i INTRODUCTION	1
1 INTRODUCTION	2
ii CURRENT STATE & RESEARCH OVERVIEW	11
2 LITERATURE REVIEW	12
2.1 Understanding Risk Assessments	12
2.2 Systematic literature review	15
2.3 Comparative Studies of Machine Learning in Other Settings	26
3 OPPORTUNITIES FOR INNOVATION	30
3.1 Preliminary takeaways from literature review	30
3.2 Research methodology	33
iii GETTING STARTED: WHAT DATA DO WE NEED?	35
4 DATA COLLECTION	36
4.1 Risk assessment methodology	36
4.2 Datasets required	36
4.3 Base data collection	37
5 DATA PROFILING	43
5.1 Data distribution	43
5.2 Normality of distributions: Conclusions	56
5.3 Preliminary insights	60
5.4 Independence of variables	69
iv EXPERIMENTS WITH MACHINE LEARNING	75
6 SUPERVISED MACHINE LEARNING	76
6.1 Framing the model - where does supervised ML come in?	76
6.2 Independent and dependent Variables	77
6.2.1 Feature Analysis for Inherent Risk	77
6.2.2 Feature Analysis for Control Effectiveness	82
6.2.3 Hypothesis evaluation	85
6.3 Model Development and Evaluation	86
6.3.1 Data pre-processing and hyperparameters	86
6.3.2 Inherent Risk: Hyperparameters	87
6.3.3 Inherent Risk: Modelling results	99

6.3.4	Inherent Risk: Model performance on test dataset	103
6.3.5	Control Effectiveness: Hyperparameters	107
6.3.6	Control Effectiveness: Modelling results	108
6.3.7	Control Effectiveness: Model performance on test dataset	111
6.4	Noisy data: A brief analysis of excluded features on model results	114
6.5	Feature importance assessment	116
6.5.1	Feature importance using SHAP - Inherent Risk	118
6.5.2	Feature importance using SHAP - Control Effectiveness	119
6.6	Interim observations	120
7	A DEEP LEARNING APPROACH WITH NEURAL NETWORKS	124
7.1	Why neural networks?	124
7.2	ANN Methodology	125
7.3	ANN - Preliminary Iteration	127
7.4	ANN - Tuned Hyperparameters	127
7.5	Interim observations	131
8	UNSUPERVISED MACHINE LEARNING FOR RISK ASSESSMENTS	132
8.1	Why use unsupervised ML?	132
8.2	Proposed approach for leveraging unsupervised ML	134
8.3	Clustering techniques	135
8.4	An anomaly-based approach to risk assessments	140
8.5	Interim observations	144
V	EMERGING CONTRIBUTIONS	146
9	BRINGING IT ALL TOGETHER	147
9.1	Experiment outcomes	147
9.2	Key limitations and assumptions	149
9.3	Contributions to practice & suggested techniques for implementation	152
9.4	Contributions to literature	154
9.5	Measuring value	157
10	CONCLUDING REMARKS	160
10.1	Research questions revisited	160
10.2	Further research avenues	161
vi	BIBLIOGRAPHY	163
A	APPENDIX A: CODE USED TO DEVELOP MODELS USED IN EXPERIMENTS	1

LIST OF FIGURES

Figure 1.1	The IIA's Three Lines Model (IIA, 2020)	3
Figure 1.2	The Internal Audit Lifecycle (Deloitte, n.d.)	6
Figure 1.3	An aspirational evolutionary cycle of an IAF (PwC, n.d.)	8
Figure 2.1	Annual Internal Audit Development Plan (IIA, n.d.b)	13
Figure 2.2	Basic Internal Audit Lifecycle (UNC, n.d.)	14
Figure 2.3	Audit Firm DA Adoption Process (Krieger et al., 2021)	23
Figure 4.1	Data snippet for four BUs	42
Figure 5.1	Key Applications	44
Figure 5.2	Production Changes	44
Figure 5.3	Risk Events	45
Figure 5.4	Risk Event - Action Plans	45
Figure 5.5	Audit Findings	46
Figure 5.6	Audit Findings	47
Figure 5.7	Customer Complaint Volumes	48
Figure 5.8	People Attrition Ratio	48
Figure 5.9	Unauthorised Changes and Resilience Breaches	49
Figure 5.10	Change Programmes	50
Figure 5.11	Key Applications	51
Figure 5.12	Production Changes	51
Figure 5.13	Risk Events	52
Figure 5.14	Risk Event - Action Plans	52
Figure 5.15	Audit Findings	53
Figure 5.16	DR and ARP Non-Compliance	53
Figure 5.17	Customer Complaint Volumes	54
Figure 5.18	People Attrition Ratio	54
Figure 5.19	Unauthorised Changes and Resilience Breaches	55
Figure 5.20	Change Programmes	55
Figure 5.21	Key Statistics - Numeric Variables	56
Figure 5.22	JB Test Statistics	58
Figure 5.23	JB Test p-values	59
Figure 5.24	JB Test Statistic: High RE Actions	60
Figure 5.25	Correlation between numeric features collected	61
Figure 5.26	Numeric features and control effectiveness	63

Figure 5.27	Numeric features and inherent risk	64
Figure 5.28	Non-numeric features and inherent risk - 1 of 2	67
Figure 5.29	Non-numeric features and inherent risk - 2 of 2	68
Figure 5.30	Spearman's R Correlation with ρ and α	71
Figure 6.1	Chi Square p-value for Inherent Risk	80
Figure 6.2	Chi-square p-values for Control Effectiveness	84
Figure 6.3	Random Forest Classifier: Confusion Matrix (IR Validation)	101
Figure 6.4	IR Confusion Matrix - Log. Regression	105
Figure 6.5	IR Confusion Matrix - KNN	105
Figure 6.6	IR Confusion Matrix - Decision Tree	105
Figure 6.7	IR Confusion Matrix - Random Forest	105
Figure 6.8	IR Confusion Matrix - Support Vector Machines	106
Figure 6.9	IR Confusion Matrix - XGBoost	106
Figure 6.10	IR Confusion Matrix - Gradient Boosting Machines	106
Figure 6.11	IR Confusion Matrix - LightGBM	106
Figure 6.12	Random Forest Classifier: Confusion Matrix (CE Validation)	109
Figure 6.13	CE Confusion Matrix - Log. Regression	112
Figure 6.14	CE Confusion Matrix - KNN	112
Figure 6.15	CE Confusion Matrix - Decision Tree	112
Figure 6.16	CE Confusion Matrix - Random Forest	112
Figure 6.17	CE Confusion Matrix - Support Vector Machines	113
Figure 6.18	CE Confusion Matrix - XGBoost	113
Figure 6.19	CE Confusion Matrix - Gradient Boosting Machines	113
Figure 6.20	CE Confusion Matrix - LightGBM	113
Figure 6.21	IR Confusion Matrix for Raw Dataset	115
Figure 6.22	CE Confusion Matrix for Raw Dataset	116
Figure 6.23	SHAP feature importance scores for inherent risk	119
Figure 6.24	SHAP feature importance scores for control effectiveness	120
Figure 7.1	Confusion matrix for Artificial Neural Network - Tuned Hyperparameter Values	130
Figure 8.1	DBSCAN Clustering with $\epsilon = 6$, $n = 506$	137
Figure 8.2	Data structure following conversion of multi-class labels to binary class	140

LIST OF TABLES

Table 3.1	Research methodology	34
Table 4.1	Non-Numeric Variables	41
Table 5.1	Volume of missing observations across dataset	65
Table 5.2	Statistically Significant Correlations - Independence Hypothesis Rejected	73
Table 6.1	List of IR features	79
Table 6.2	List of CE features	83
Table 6.3	Imbalanced data for IR	87
Table 6.4	Training dataset labels post imbalance treatment	88
Table 6.5	Hyperparameter Tuning Approach	99
Table 6.6	Classifier Accuracy for Inherent Risk	100
Table 6.7	Classifier Precision for Inherent Risk	102
Table 6.8	Classifier Recall for Inherent Risk	102
Table 6.9	Classifier Accuracy for Inherent Risk: Test Dataset	103
Table 6.10	Imbalanced data for CE	107
Table 6.11	Training dataset labels post imbalance treatment	107
Table 6.12	Classifier Accuracy for Control Effectiveness	108
Table 6.13	Classifier Precision for Control Effectiveness	110
Table 6.14	Classifier Recall for Control Effectiveness	110
Table 6.15	Classifier Accuracy for Control Effectiveness: Test Dataset	114
Table 7.1	Artificial neural network performance	127
Table 8.1	DBSCAN Results	136
Table 8.2	DBSCAN Results - All features	138
Table 8.3	Anomaly Detection Accuracy	143
Table 8.4	Anomaly Detection Recall	143
Table 8.5	Anomaly Detection Results: Test Dataset	144

LIST OF ACRONYMS

AAA	American Accounting Association
ADAM	Adaptive Moment Estimation
ANN	Artificial Neural Networks
AQ	Audit Quality
ARP	Application Recovery Plans
AS	Auditing Standard
BU	Business Units
CA	Continuous Auditing
CAE	Chief Audit Executive
CE	Control Effectiveness
DA	Data Analytics
DBSCAN	Density-based spatial clustering of applications with noise
DR	Disaster Recovery
FCA	Financial Conduct Authority
HDBSCAN	High density-based spatial clustering of applications with noise
IAF	Internal Audit Functions
ICAS	Institute of Chartered Accountants of Scotland
IIA	Institute of Internal Auditors
IR	Inherent Risk
IRS	Internal Revenue Service
JB	Jarque-Bera
KNN	k-Nearest Neighbours
LIME	Local Interpretable Model-agnostic Explanations
LOF	Local Outlier Factor
LLM	Large Language Models
ML	Machine Learning
NFM	Non-financial measures
PCA	Principal Component Analysis
PCAOB	Public Company Accounting Oversight Board
PDF	Probability Density Function
PII	Personally Identifiable Information
RBIA	Risk based Internal Auditing
RCSA	Risk and Control Self Assessment

RELU	Rectified Linear Unit
RR	Residual Risk
SEC	Securities and Exchange Commission
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive Explanations
SLA	Service Level Agreements
SLR	Systematic Literature Review
SLT	Senior Leadership Team
SMOTE	Synthetic Minority Oversampling Technique
SQL	Structured Query Language
SVM	Support Vector Machine
TBAT	Technology-based Audit Techniques

Part I

INTRODUCTION

INTRODUCTION

Organisations around the world are on an ongoing journey of innovation-driven improvement, which is increasingly marked by the complexity brought on by an ever-evolving technology and data landscape (Boyles, 2022). These complexities naturally increase the level of risks an organisation must appropriately mitigate in order to ensure the smooth running of daily business operations and compliance with regulatory expectations. An organisation's Internal Audit department provides independent assurance that their risk management, governance, and internal control processes are operating effectively (IIA, 2020). An evolving risk landscape also translates into a requirement for a robust internal controls system, resulting in an organisationally significant role for the IAF (Martino et al., 2021).

Although management teams are responsible for identifying areas of key organisational risk and implementing an appropriate suite of internal controls in response, the IAF plays an important role in ensuring the continued currency and value of these controls through periodic audit testing (IIA, 2023a). This segregation of roles is well established in practice by the "three lines of defence" model, originally introduced by the Institute of Internal Auditors (IIA) and reinforced in their 2020 publication (IIA, 2020) (see figure 1.1).

The IIA's Three Lines Model

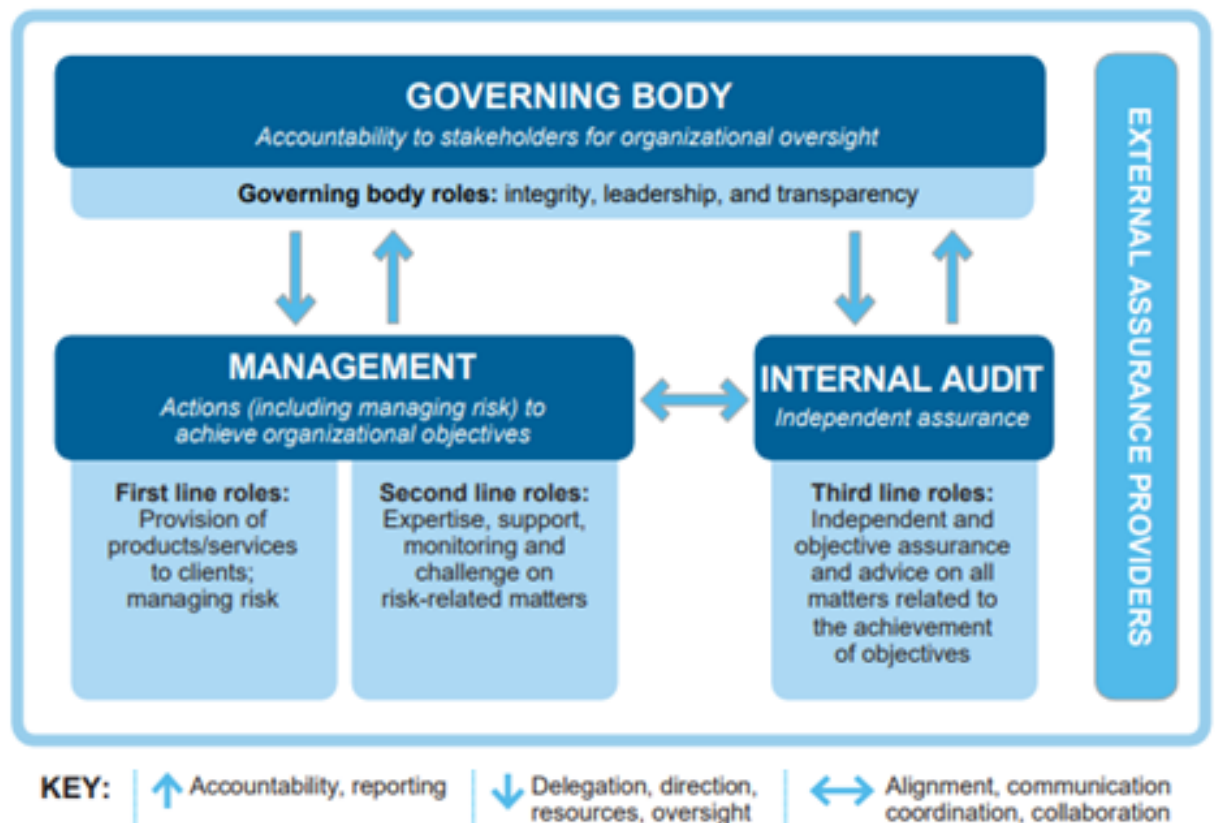


Figure 1.1: The IIA's Three Lines Model (IIA, 2020)

Positioned as the third line of defence, Internal Audit are well placed to deliver meaningful change and value to an organisation. Given the strong and intrinsic link to risk, the IIA (2023b) state that the need to manage risks has become recognised as an essential part of good corporate governance practice, and that one of the key roles of internal audit is to provide assurance that those risks have been effectively managed. In their article, the IIA (2023b), define Risk based Internal Auditing (RBIA) as a methodology that links internal auditing to the organisation's overall risk management framework. A risk-based audit plan enables an organisation to focus on the most important areas of the audit universe, thereby adding value to stakeholders and regulators (Anon et al., 2020; Coetzee, 2010).

Established accounting bodies, such as the Institute of Chartered Accountants of Scotland (ICAS), have recognised the seminal role of risk assessments within the internal audit process. As an example, in his 2018 article, Bruce (2018) asserts that the risk assessment process is foundational to the internal audit lifecycle (see figure 1.2 (Deloitte, n.d.) below), as a robust assessment would help inform when an audit should be performed and when, including the most appropriate audit produces to use and required skillsets (Bruce, 2018).

As such, a core competency of IAFs is the ability to systematically identify key risks, and do so regularly in order to maintain focus on the highest risk areas of the organisation. This systematic evaluation, known as risk assessment, forms the basis for directing audit resources toward areas of greatest importance. Indeed, research has shown the presence of a strong relationship between the outcomes of risk assessments, and the auditors' risk perception towards the organisation as a whole, which ultimately drives the mechanics of audit delivery (Porcuna-Enguix et al., 2021).

During the risk assessment process, the internal audit function would both identify and assess the likelihood and potential impact of various risks to the organization (Bruce, 2018). This clearly articulates the pivotal role of internal audit risk assessment as a foundational step in the internal auditing process, as the identification of high-risk areas would naturally be the starting point of an IAF's focus in reviewing the prevalent risks across an organisation.

The RBIA methodology has found favour in academic research as well. An example of this is Kurniawan and Erwin (2023)'s research, which undertook a quantitative study to examine the impact of the RBIA model on the effectiveness of the IAF. The (Kurniawan and Erwin, 2023) analysis highlighted that "Professional Development, Top Management Commitment, and the Role of Internal Audit had a significant positive effect on the successful implementation of risk management with the RBIA method" (Kurniawan and Erwin, 2023).

It is therefore pertinent to note that the starting point of a robust internal audit methodology is the assessment of risk, which contributes to the audit strategy and periodic audit planning. This activity, referred to as "internal audit risk assessments" throughout this dissertation, takes the form of a formula, as represented below, based on the Public Company Accounting Oversight Board (PCAOB)'s Auditing Standard (AS) 1101: Audit Risk (PCAOB 2010):

$$RR = IR - CE$$

Where:

- Inherent Risk (IR) refers to the susceptibility of an assertion to a misstatement, due to error or fraud, that could be material, individually or in combination with other misstatements, before consideration of any related controls.
- Control Effectiveness (CE) is the risk that a misstatement due to error or fraud that could occur in an assertion and that could be material, individually or in combination with other misstatements, will not be prevented or detected on a timely basis by the company's internal controls framework. Control risk is therefore a function of the effectiveness of the design and operation of internal controls.

Residual Risk (RR) is therefore the risk level remaining after the controls have been implemented.

As shown in the flow diagram in figure 1.2 (Deloitte, n.d.), the risk assessment process is fundamental to the auditing process, as it requires the audit function to subjectively assess the residual risk across the organisation as a whole, one audit entity at a time. This enables the creation of an audit plan, targeting individual departments of the organisation for review, referred to as the “Internal Audit Lifecycle” for the rest of this dissertation.

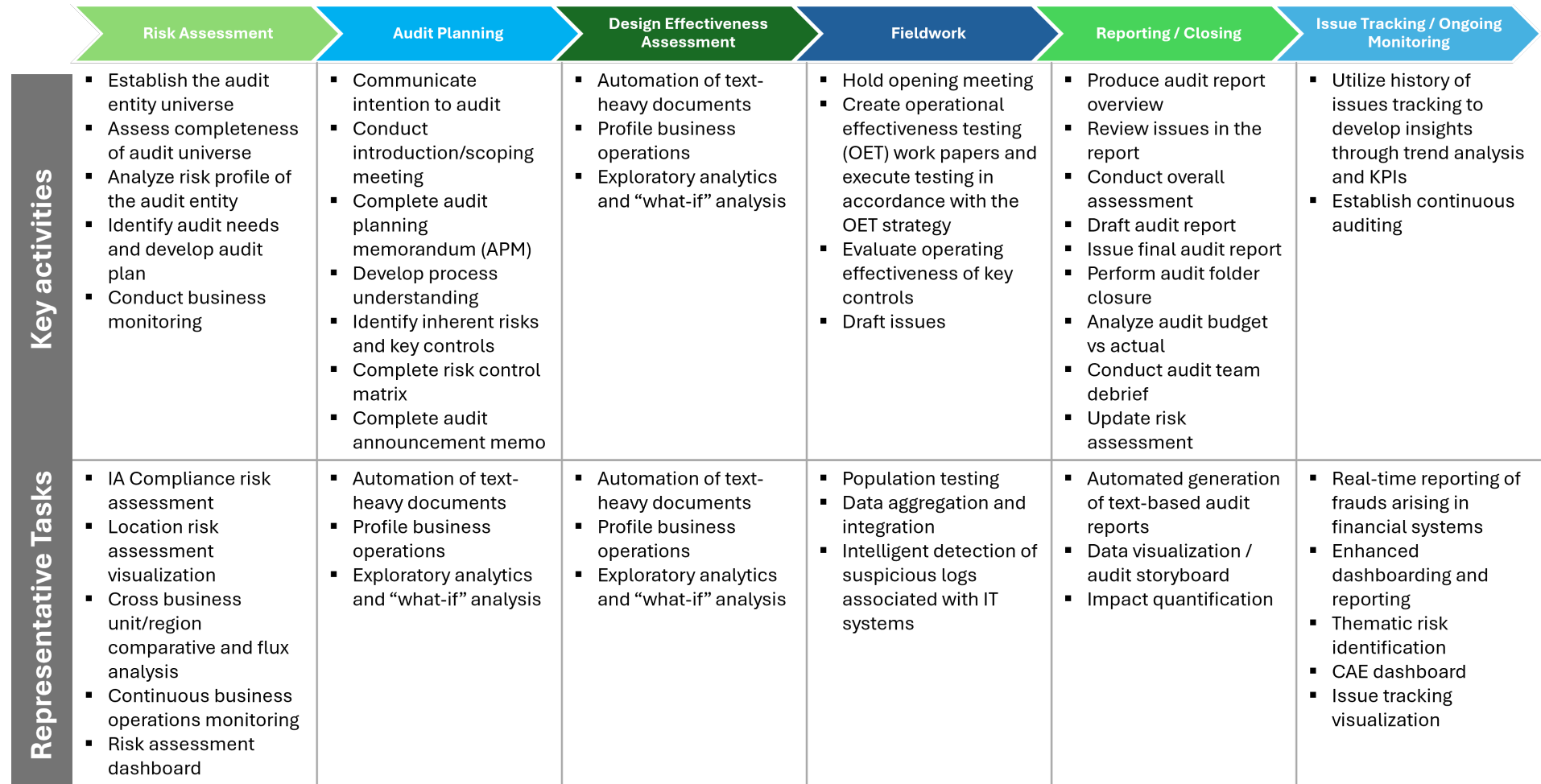


Figure 1.2: The Internal Audit Lifecycle (Deloitte, n.d.)

Unsurprisingly, audit risk assessment activities may also be closely linked to fraud risk assessment activities (Hammersley, 2011), elevating the importance of this stage for the organisation as a whole. Prior research by Imoniana and Gartner (2008) has also established that internal audit risk assessments are key to establishing a confidence level for an audit engagement, which can directly contribute to an increase in the probability of detecting potential misstatements in financial statements by auditors.

Additionally, the internal audit risk landscape is a complex and ever-evolving environment, and prior research has discussed the influence of digitisation and innovation on the internal audit profession (Betti and Sarens, 2021; Bibler and Gold, 2023; Eilifsen et al., 2020), with studies also increasingly offering empirical analyses of auditing topics, such as the role of the IAF in preventing and detecting fraud (Bonrath and Eulerich, 2023). Indeed, risks can originate from various sources, both internal and external to the organization.

- Internal risks stem from weaknesses in internal controls, human error, process inefficiencies, or technological vulnerabilities. Examples include fraudulent activities, errors in financial reporting, or IT security breaches.
- External risks arise from factors beyond the organization's direct control, such as economic downturns, regulatory changes, industry disruptions, or natural disasters. These external factors can significantly impact the organization's financial performance, operational efficiency, and overall sustainability.

As such, the pivotal positioning of risk assessments, especially in the context of increased organisational complexity and an ever-evolving risk landscape, raises several interesting questions:

- If organisations evolve, why not their internal audit functions?
- Is innovation within an internal audit function “essential” or “desirable”?
- Can an innovative risk assessment drive greater value from the IAF?
- What does an innovative risk assessment for internal audit look like?

A study by Ditkaew and Suttipun (2023) highlights the importance of innovation by IAFs and specifically mentions Data Analytics (DA) as the main vehicle for ushering in this change, given the results of their quantitative study highlighting the positive impact of DA on Audit Quality (AQ). Additionally, industry leaders in the field of internal audit, such as PwC (n.d.), focus on helping organisations increase their overall investment in data analytics, in order to generate value from internal audit and benefit the organisation holistically. As seen from figure 1.3 (PwC, n.d.) below, the effectiveness in monitoring key risks increases as an internal audit function moves away from “manual” activities to more

advanced, “predictive” capabilities, which, along with being innovative, has the effect of adding holistic value to an organisation.

Manual IA function

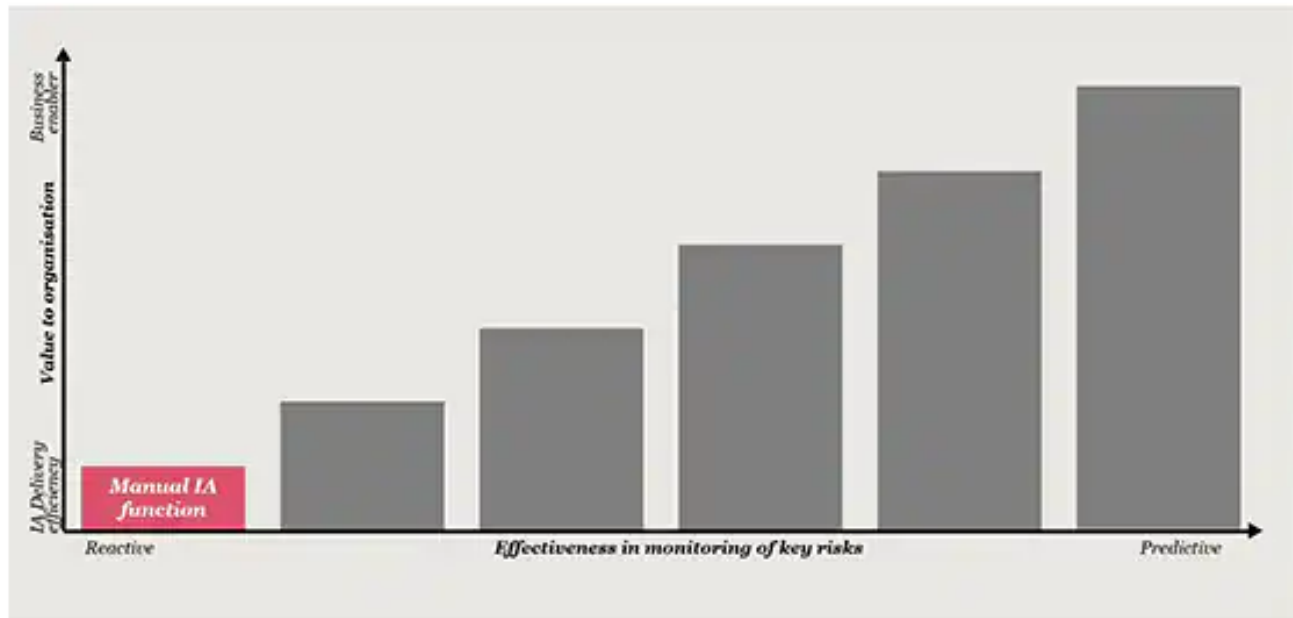


Figure 1.3: An aspirational evolutionary cycle of an IAF (PwC, n.d.)

In a world defined by big data, DA within Internal Audit can take several forms, from an automated analysis of historic datasets to advanced machine learning capabilities. Whilst the possibilities and value-add for the audit profession are immense, the challenges are manifold. And finally, whilst there **may be an established case for pursuing innovation in the sphere of internal audit risk assessments**, it is not clear as to how such innovation might be generated, to result in a **practical and scalable** methodology.

As such, the core research questions that have been evaluated as part of this dissertation are as follows:

- RQ1.** Has existing academic literature explored the impact of ML and DA capabilities on the audit lifecycle?
- RQ2.** How can machine learning capabilities be effectively leveraged to deliver risk assessments for internal audit functions?
- RQ3.** How can practitioners measure value from the deployment of ML techniques on the audit lifecycle?

This research therefore began with a detailed study of extant academic literature, which directly addressed RQ1. Through the literature study, seminal trends and research patterns in the field of internal audit and specifically, risk assessments were examined.

In chapter 2 of this thesis, RQ1 is addressed by exploring previous efforts to profile and implement ML techniques across academia and practice. A consensus was found on the positive influence and future potential of ML capabilities on the assurance lifecycle, which has been empirically researched on specific, isolated audit use cases to assess overall value-add. Combined with a desire to improve AQ and the overall efficacy of the IAF, the push to continue innovating across the audit lifecycle is only set to increase. However, the Systematic Literature Review (SLR) found a research gap specifically in the area of internal audit risk assessments, which are performed differently across individual IAFs and vary across industries and geographies. Despite being a foundational step in the audit lifecycle, the lack of practical case studies exploring the application of ML techniques for risk assessments on the audit lifecycle represented the principal focus for this research.

The SLR findings in chapter 2, were used to identify appropriate ML techniques for risk assessments, which are summarised in chapter 3. This enabled the identification of specific ML techniques, framed as individual experiments, to help progress the research in more detail. To support these experiments, a range of foundational steps in chapters 4 and 5 were performed. These involved the extraction, transformation and loading of base datasets, as well as feature selection and preprocessing.

This formed the basis for the experiments with ML across chapters 6, 7 and 8. As part of these experiments, an analysis was performed into the use of specific ML techniques for internal audit risk assessments to help answer RQ2. The collected data set was examined for risk assessments at the subject organisation, including underlying features to assess their suitability for the experiments, splitting out a test data set prior to feature selection and engineering. As part of this stage, key hyper-parameters for in-scope ML techniques were examined and fine-tuned, with the approach drawing on prior research as appropriate.

Finally, a comparative evaluation of the various techniques was performed and approaches were considered to measure the value-add from the use of ML paradigms for risk assessments, with a view to answering RQ3. In doing so, the pros and cons of each ML technique were highlighted and overlaid on the business considerations impacting successful practical implementation.

Chapters 9 and 10 detail the conclusions and proposals, as well as key limitations identified from the experiment phase, which are intended to support practitioners looking to leverage this work for practical implementation. This assessment of the theoretical techniques underpinning various ML techniques in light of practical implementation challenges represents

a further contribution to this emerging field of study, which would help further additional research by future academics and practitioners.

Part II

CURRENT STATE & RESEARCH OVERVIEW

LITERATURE REVIEW

The research begins with a considered study of existing literature, in support of RQ1. This chapter profiles the existing academic literature exploring potential machine learning techniques relevant to internal audit, focusing on the risk assessment phase. This chapter also considers literature exploring risk assessments in fields other than internal audit, to examine potential read-across that could be used in support of both RQ1 and RQ2. The outputs of this chapter are used to inform the research methodology that has been applied in subsequent chapters.

2.1 UNDERSTANDING RISK ASSESSMENTS

In recent years, expectations of IAFs to innovate their approach to delivering assurance have grown manifold. These expectations are triggered by a range of factors, from multi-million dollar accounting frauds to increased stakeholder and regulatory expectations. Therefore, the Internal Audit function's mandate of providing independent and objective assurance must be sustainable and scalable.

The IIA in their 2023 review paper on how audit should be conducted (IIA, 2023a) , state that the “audit planning process should reflect upon the extent to which risks to the achievement of key objectives and major changes and emerging risks are reflected in the **overall risk assessment** of the organisation and, linked to this, the extent to which internal audit is involved in the major changes and developments taking place across the organisation.” Therefore, when considering the end-to-end flow of the internal audit process, risk assessments can be considered the all-important starting point. The approach typically taken by IAFs to design the approach for delivering their work is shown in figure 2.1 (IIA, n.d.b).

As highlighted below, an important aspect of a robust internal audit lifecycle is the need to keep the annual plan up to date through periodic reviews. This proposes the position that a “dynamic” or “real-time” update to the internal audit plan, in response to an evolving risk landscape, is crucial to enable currency and value from the internal audit function.



Figure 2.1: Annual Internal Audit Development Plan (IIA, n.d.b)

Once selected through risk assessment, an audit's lifecycle moves across three stages, as shown in figure 2.2 below.

The Internal Audit Process

The process employed by the Office of Internal Audit in performing audits follows three general phases comprising planning, fieldwork, and reporting.

The following illustration documents the process typically employed in the performance of an internal audit.

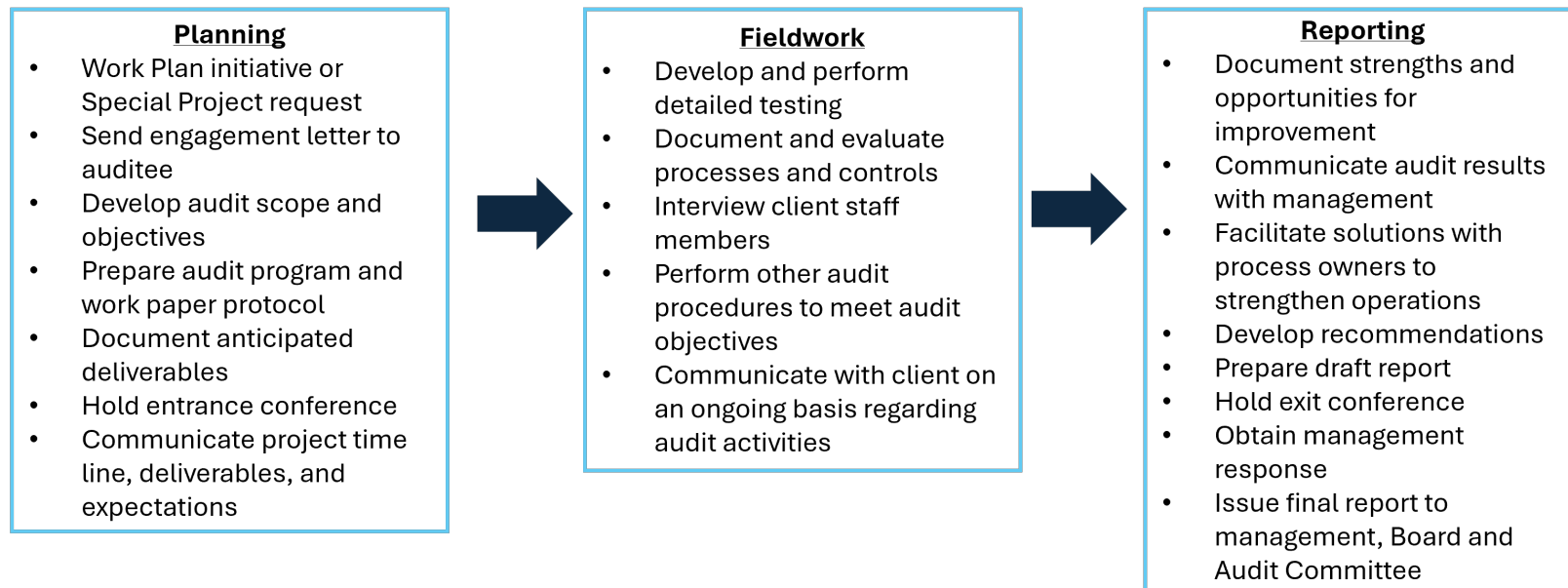


Figure 2.2: Basic Internal Audit Lifecycle (UNC, n.d.)

A 2022 PwC survey (PwC, 2022), based on 3,584 business and risk, audit and compliance executives, highlighted the need for audit functions to take what PwC describes as a “panoramic view of risk”, which, by definition, would require a more pronounced and continuous application of advanced techniques in the above audit lifecycle. This view is supplemented by Irwansyah and Zega (2023) who suggest a negative correlation between the role of effective Internal Audit and the tendency to commit fraud, which further accentuates the importance and value-add from a robust internal audit function.

2.2 SYSTEMATIC LITERATURE REVIEW

An SLR of the existing literature on Internal Audit innovation was undertaken to evaluate RQ1. This SLR was designed to profile previous attempts to introduce such innovation within academia and across practice. This review collated a preliminary set of research papers for detailed analysis, which have been explored in detail across four primary themes as set out below.

EMPIRICAL ANALYSIS

Firstly, an assessment of internal audit literature shows that multiple attempts have been made to introduce innovation through the creation and deployment of advanced data analytics and machine learning capabilities, with mixed success and limited evidence of scalable effectiveness (Alles and Gray, 2016). Equally, these studies point to the positive attitude by the internal audit community towards the use of advanced data analytics (Zhang et al., 2015; Yoon et al., 2015). There exists evidence of machine learning for risk assessments in fields other than audit, governance, risk and compliance, such as safety risk assessments. An example of this is research carried out by Paltrinieri et al. (2019), whose work showed that the use of deep neural networks for safety risk assessments, when properly constructed showed reasonable accuracy and suitability for the industry. Notwithstanding the above, the literature review showed that the use of empirical data to analyse the effectiveness of innovation techniques to improve internal audit processes, and specifically, internal audit risk assessments, is quite limited, and offers plenty of room for additional research. The uptake of innovation on the internal audit lifecycle using DA appears to be limited by a) awareness and b) lack of training opportunities outside the “big four accounting firms” (Islam and Stafford, 2022).

Secondly, research shows that a range of factors impact the effectiveness of DA in auditing (Jacky and Sulaiman, 2022). Indeed, research has shown that the Big Four accounting firms lead the way in implementing data analytics as part of the auditing process (Oyewo and Ajape, 2021) and (Alles and Gray, 2016), with press reports suggesting collective investment in the range of \$9billion by three of the big four firms (Kapoor, 2020). Such significant sums

of investment indicate quality and differentiation as drivers for adopting DA or AI for the auditing process. Additionally, Dagilienė and Klovienė (2019) cite compliance with regulatory requirements as a key motivator for focusing on DA for auditing.

Specifically, Gepp and Smith (2018) assert that auditing is “lagging behind” the other research streams in the use of valuable big data techniques. Similar conclusions were reached by De Santis and D’Onza (2021) who reaffirm the positive influence of DA on audit quality whilst calling for more empirical studies on the utilisation of DA within auditing. The SLR identified a limited set of empirical studies that serve as suitable examples illustrating the impact of effective and considered innovation on the audit lifecycle. Five of these are presented below.

1. Big data analytics usage by the IRS:

A concerted effort by a government regulator, the Internal Revenue Service (IRS) (responsible for tax administration and collections in the United States), indicates what is possible when it comes to the identification, deployment and value-add from advanced data analytics. Specifically, Houser and Sanders (2018) show that the IRS has invested in both predictive and descriptive analytics to support a preliminary review of tax returns, with a view to detecting potential non-compliant behaviour and profile taxpayers for potential audits, where needed. The IRS’ significant investment in a dedicated analytics programme, namely, the Automated Substitute for Returns data analytics program, staffed with just 65 full-time equivalent employees finalised c.400,000 cases resulting in \$542.8 million in additional assessments, in fiscal year 2016 alone. This value-add works out to a cost of 35c. for every \$100 collected.

2. Correlation between advanced analytics and audit scepticism:

Barr-Pulliam et al. (2020) conducted empirical research to study the influence and impact of data analytics on how auditors apply professional scepticism, a key quality attribute expected of the internal audit community. The Barr-Pulliam analysis performed an experiment to observe how auditors assess visualisations from a data analytic tool that had analysed financial and Non-financial measures (NFM) for a hypothetical audit engagement. The study demonstrated that an increase in the rate of false positives from data analytics had a negative correlation with the auditor’s level of professional scepticism, on the basis that auditors were “more comfortable” dismissing any red flags highlighted by data analytics. Indirectly, this experiment highlights the growing influence and perception of data analytics on audit effectiveness.

3. Impact of process mining techniques on Internal Audit effectiveness:

In their submission to the International Journal of Accounting Information Systems, Bäßler and Eulerich (2023) offer a rare empirical insight into how the deployment of

a machine learning techniques on a classic internal audit test of payment timeliness enabled the auditors on the team to “provide assurance, reduce risk and prevent undesired outcomes”. The study showed that the use of machine learning techniques (in this case, supervised machine learning using classification) could enable the prediction of control failures in the payments process prior to the formal completion of the audit, whilst retaining the objectivity and independence of the audit. The empirical study considered the impact and usefulness of:

- Predicting control failures too close to the audit deadline (which showed the highest accuracy; however, with limited additional value, given the inability of the first- and second-line Management teams to prevent the control failure in good time).
- Predicting control failures too early, which could result in a potential confusion of roles and responsibilities between the IAF and the first and second lines of defence.
- Identifying the ‘optimal cutoff’ – in this experiment, defined as seven business days prior to the audit deadline. This Bäßler study identified this threshold as the optimal point that enabled the highest level of value-add from advanced analytics, whilst retaining auditor objectivity and independence.

4. The application of advanced technology, such as Blockchains, on Internal Audits:

Vishnia and Peters (2019) performed experimental research leveraging Blockchain technology to build a model, sequential and tamper-proof audit trail for a trading platform, referred to in their research as “AuditChain”. Given the core internal audit requirement to assure integrity and demonstrability of an audit trail, the hypothesis was that the Blockchain constructed could be used to verify the integrity of the trading data executed on the platform using a smart contract. The experiment was able to successfully demonstrate the creation of a Blockchain-based audit trail, within an acceptable timeframe. Their work concluded that the usage of blockchain made it a “perfect candidate for auditing usage”, which is a justified conclusion given the immutable nature of blockchain and its direct application on auditing requirements (integrity and traceability of an audit trail).

5. Predictive modelling to generate predictions of audit opinions:

Stanisic et al. (2019) show that a combined application of statistical techniques and machine learning capabilities could be used to develop predictive models to generate predictions of internal audit opinions, with high accuracy and importantly, interpretability. The empirical analysis in the Stanisic study was based on a hand-collected dataset of 13,561 pairs of audit reports, which makes the findings of this study impact-

ful. Additionally, the Stanistic study was stratified across audit reports where prior client information was available and cases where such information was missing, supporting scalable adoption of the modelling approach across organisations.

Interestingly, whilst the above literature assesses the impact of specific analytics techniques and technology-based innovation capabilities (such as Blockchain) on the Internal Audit process, the above examples do not illustrate sufficient usage of advanced data analytics for internal audit risk assessments. However, there is sufficient literature to enable us to conclude that the deployment of advanced analytics is a positive step, widely viewed by academics and audit practitioners as a step in the right direction, with indeed many (including the IRS) viewing the adoption of analytics as a key component of the future of internal auditing.

QUALITATIVE IMPACT

As part of their exploratory study into the usage of audit data analytics in practice, Eilifsen et al. (2020) interviewed the Heads of Professional Practice at five international public accounting firms in Norway and analysed the “perceptions towards audit data analytics” by 216 senior audit practitioners in the country. The Eilifsen analysis concluded that a) firms differ in strategies on how to implement advanced data analytics, b) the use of advanced data analytics is “rare” and c) there is a general lack of consensus on the response of supervisory inspection authorities with respect to the use of analytics. The paper, published in the respected American Accounting Association (AAA) journal, has been cited repeatedly since publication and serves as a seminal study into the use of advanced data analytics by auditors.

The Eilifsen study and conclusions are representative of the broader attempts made to assess the impact of data-driven audit capabilities and innovation in the field of audit. More recent research, by Kahyaoglu and Aksoy (2021), has assessed the impact of AI capabilities on internal audit risk assessments in particular, noting the potential of AI to boost assurance quality through the early identification of potential organisational risks.

Similar literature examining the qualitative impact of audit data analytics appears to take the approach of focusing on individual aspects of the internal audit lifecycle, as set out in figure 2 above, and is largely motivated by two drivers – firstly, the use of analytics to enable greater coverage of transactions subject to auditing (typically referred to as “full population testing”), as opposed to basing audit conclusions on sample-based test work (“audit sampling”) (Chiu and Jans, 2017; Huang et al., 2022; Nonnenmacher and Gomez, 2021) and secondly, the expectation that technology-based innovation, such as audit data analytics, could allay pressure on internal auditors (Cardinaels et al., 2021; Eulerich et al., 2020, 2022).

Neither driver is particularly surprising. From the use of advanced descriptive analytics capabilities leveraging tools such as advanced Excel and Structured Query Language (SQL), to predictive analytics capabilities such as machine learning techniques, such as classification, regression, and clustering techniques, enable auditors to test the full population of an audit scope area, which automatically obviates the need to base an audit opinion on a haphazard or statistical sample, which was previously the case. A direct consequence of this approach is the reduction of audit risk (which refers to the risk of an inaccurate audit opinion) (Huang et al., 2022; Barr-Pulliam et al., 2022). Indeed, Huang et al. (2022) assert that audit sampling is beginning to lose its relevance in the era of big data and offer a qualitative framework to apply audit data analytics to full population testing on an audit. This literature advocates the positive, qualitative impact of data analytics on the execution of individual internal audit assignments and is supported by multiple supplementary studies.

Given that academic literature discussing the qualitative impact of data analytics for audit tends to focus on individual audit subject areas (as broadly set out in figure 2 above), a brief overview of the approach and findings of prior literature, supported by examples, is given below. The resultant qualitative impact, if any, as identified by the research example, is highlighted in bold.

1. An empirical exploratory study carried out by Hampton and Stratopoulos (2016), found a positive relationship between audit opinion confidence and decreased litigation risk, significant for both medium and large-size audit firms, following a survey conducted of 394 Canadian audit professionals across small, medium, and large-sized audit firms.
2. In their paper, Eulerich et al. (2022) surveyed 268 auditors to assess the qualitative impact and effectiveness of Technology-based Audit Techniques (TBAT), including data analytics, and found that the use of TBATs is positively associated with audit effectiveness, and consistent with auditors' and Chief Audit Executive (CAE)' perceptions of TBATs. The Eulerich analysis also showed that the use of TBATs positively impacts audit efficiency and calculated that one standard deviation increase in TBAT usage is associated with a 13.5% decrease in the number of days necessary to complete an audit.
3. Gao et al. (2020) specifically examined the impact of data analytics on audit quality and found a higher qualitative impact of the use of audit data analytics on firms with complex business operations, given the ability of analytics to draw out underlying risks more effectively. Additionally, Gao et al. (2020) suggest an increased investment in human capital investment in data analytics to contribute to increased audit quality.

4. Existing literature also suggests that the use of data analytics for internal auditing could reduce audit pressure (Cardinaels et al., 2021), which serves as a powerful case study advocating the qualitative impact of data analytics. Cardinaels et al. (2021) found that auditors were less pressured in drawing their final conclusions when their test work was supported by data analytics, and that the use of analytics improves the overall internal audit quality of internal auditors as hypothesised.
5. Research conducted by Butkė and Dagilienė (2022) on a selection of companies across agricultural, retail and manufacturing sectors, where a risk-based approach to audit was implemented, showed that the use of advanced data analytics yielded much higher productivity and completeness of testing. This finding was consistent across all analytics techniques deployed as part of the research, which included Classification, Clustering, Regression, Association Rules, Text Analysis and Visualisation. Additionally, this study found that the results yielded by most data analytics tools deploying the above techniques were easily interpretable, without requiring additional skills or expertise.
6. One of the most relevant examples in prior literature examining the impact of advanced analytics in the area of internal audit risk assessments is research performed by Eulerich and Eulerich (2020), who specifically examined the use of Continuous Auditing (CA) as an audit methodology to support risk assessment. The Eulerich study included analysis of responses provided by 264 Chief Audit Executives and concluded that CA plays a growing role in risk-based audit planning, there is consensus in the audit practitioner community as to the increasing importance of data analytics and the improvement of risk assessment is one of the main benefits of adopting forensic data analytics techniques, such as CA. However, this particular study does not offer an insight into how such an improvement might be achieved.
7. A niche assessment of the qualitative impact of data analytics using advanced technology, such as Blockchain, was performed by Lombardi et al. (2021) in their systematic literature review of 40 articles analysing the impact of blockchain in the field of internal auditing. Lombardi et al. (2021) not only find that the use of Blockchain technology is disrupting the auditing profession, but that the majority of these articles (36, representing 90% of the research population) are qualitative analyses of the use of blockchain in auditing. For instance, Lombardi et al. (2021) cite numerous examples of how prior literature exploring the use of blockchain for auditing highlights the possibility to have continuous audit procedures for auditors' clients without adding software integrations, representing a notable efficiency gain in the internal audit life-cycle.

8. Ruhnke (2022)'s empirical research exploring the qualitative implications of the use of audit data analytics highlights the positive impact of the use of data analytics on the auditor-client relationship. Ruhnke (2022) also considers whether the use of audit data analytics can serve simultaneously as a risk assessment procedure, substantive procedure, or as a test of controls; however, offers no conclusion that could be used to support answering RQ2.
9. Similar to the Lombardi et al. (2021) analysis, Cao et al. (2019) argue that the use of decentralised ledger technology as an advanced analytics capability to support internal auditing opens up the possibility of improved verification efficiency for transactions recorded on audit databases, but boosts audit quality more holistically, by enabling cross-auditor verifications by preserving data privacy, a cornerstone of internal audit processes.
10. Chiu and Jans (2017) focused on the usage of process mining techniques to explore the impact of auditing the full population of subject matter areas, as opposed to adopting a sample-based approach to auditing. Chiu and Jans (2017) were able to conduct analysis over the event logs of a large European bank and solely use this dataset to detect potential risks, ineffective internal controls and processes in a comprehensive manner that sample-based auditing cannot support. Similar outcomes were found by Jans et al. (2014) in a prior study of process mining techniques on event logs, with the results asserting potential "transformation" of the auditing process as a result.

In conclusion, prior literature supports the assertion that the use of advanced data analytics has a significant, positive qualitative impact, across multiple dimensions of the audit lifecycle.

However, a strong theme of prior literature exploring the qualitative improvement in internal auditing is the application of such techniques on the actual execution of audit assignments, also known as audit fieldwork. Indeed, a noticeable omission of prior literature examining the qualitative impact of analytics is in the area of internal audit risk assessments and / or audit planning, an assertion backed up in Awuah et al. (2021)'s research paper. Awuah et al. (2021), analysed data from 75 private and state-owned enterprises and noted that the actual usage of data analytics in the areas of risk assessment, fraud detection and substantive testing are still low. Similarly, whilst the Cao et al. (2019) analysis offers an encouraging analysis of the application of blockchain technology for internal auditing, alternative literature, such as Cagle (2020) and Fuller and Markelevich (2019), asserts that there is a lack of consensus on how such technology would be realistically applied for auditing purposes.

As such, whilst there is a strong case showing the positive qualitative impact of the use of advanced analytics for internal audit, there exists an important research gap showing how analytics techniques could be leveraged for internal audit risk assessments and planning in particular.

SCALABLE INNOVATION DELIVERY

Existing academic literature has also highlighted that there are multiple elements feeding into the process of innovation in the audit domain (Krieger et al., 2021), however, research also suggests that the prior audit literature has not significantly contributed to a holistic appreciation of the key factors that contribute to scalable innovation within audit functions (Kotb and Halabi, 2020; Rawashdeh, 2022). In their seminal work, Krieger et al. (2021) propose that the complexity of deriving innovation, “comprising of several activities across different organisational units”, coupled with the requirements to have increasingly complex technology, are pre-requisites for scalable adoption. The Krieger analysis produced a model of their process theory explaining the possible reasons for non-adoption of advanced data analytics at scale by several organisations, which is shown below.

Figure 2.3 (Krieger et al., 2021) shows that there is a multi-step journey from ideation to deployment that organisations have to go through before unlocking value from innovation or advanced data analytics.

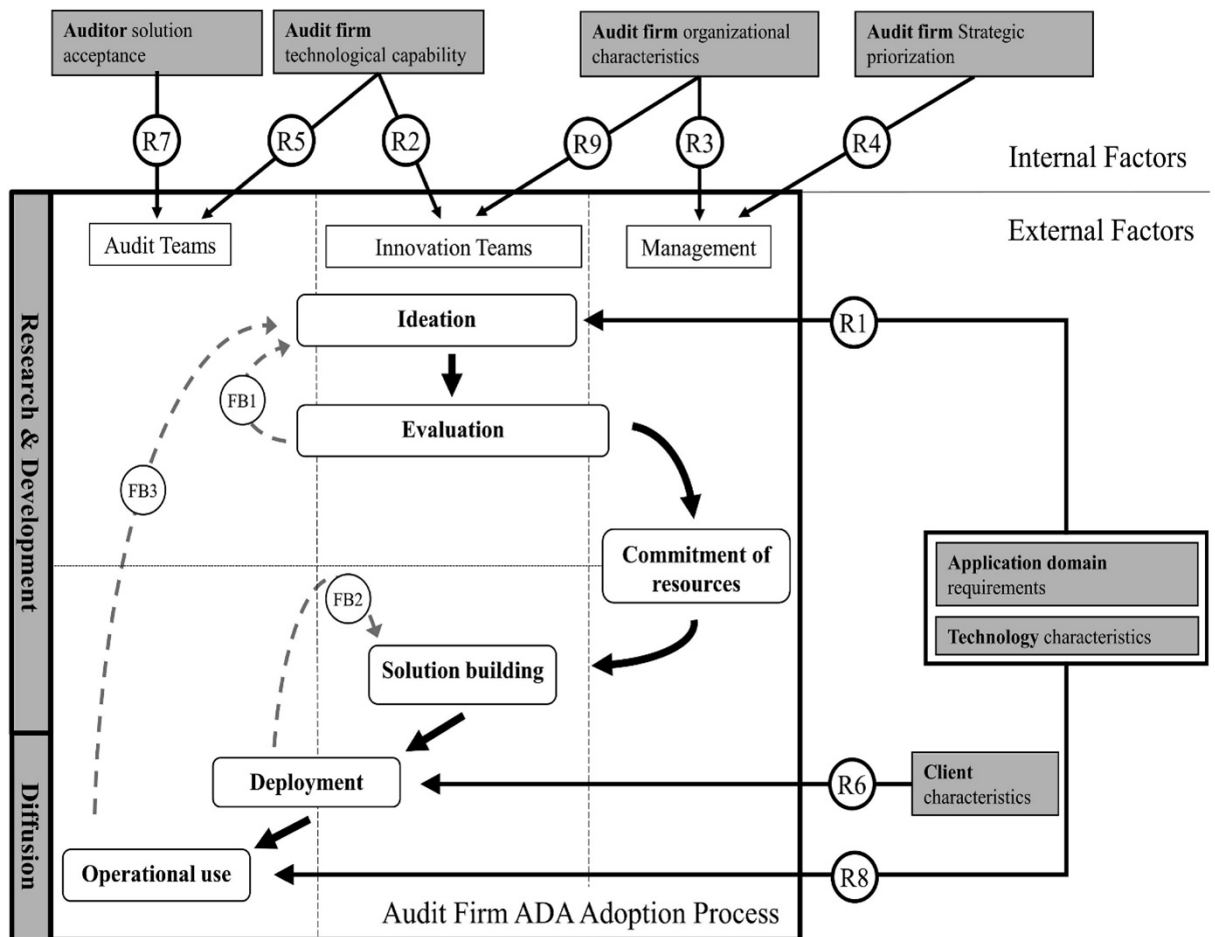


Figure 2.3: Audit Firm DA Adoption Process (Krieger et al., 2021)

This proposed model highlights a number of useful characteristics. For starters, there are a number of dependencies to adoption of advanced data analytics techniques. For example, whilst a focused commitment of resources between the evaluation and solution stage could build a solution, the deployment and uptake would require support of the customer. It is therefore no surprise that the Kreiger analysis concluded that the scalable adoption of advanced data analytics techniques, despite the obvious advantages for audit quality (as discussed previously in this section), is rare and largely limited to the major/ Big four accounting firms.

In their paper examining the determinants of audit procedures, Blokdijs et al. (2003) also noted that the use of advanced audit techniques such as data analytics is applied differently across the major accounting firms versus other audit functions. Whilst the Blokdijs research was based on the state of the art in audit practices about 20 years ago, the requirement for the internal audit process, and internal audit planning in particular, remains a constant ongoing requirement. In their research, Betti and Sarens (2021) highlight that the increas-

ing digitisation of the business environment directly impacts the IAF and specifically, the “agility” of the internal audit planning.

Existing literature has also examined the correlation between the adoption of technology-based audit techniques and scalable analytics adoption, often synonymising the two concepts. For example, Sifa et al. (2019) cite “audit efficiency” as a main qualitative benefit of their research into the adoption of an audit software (Automated List Inspection), which utilises machine learning techniques (focused on text analysis and natural language processing) to support automated auditing. The emerging use of Large Language Models (LLM), such as ChatGPT, is also linked to improvements to audit automation and efficiency (Emett et al., 2023). Specifically, Emett et al. (2023) propose from preliminary research that efficiency gains of between 50 and 80% for various processes are estimated from the adoption of emerging technologies such as ChatGPT. Whilst this is a clear example of how value can be measured from the use of such technology, there are limited examples of value add for the use of ML techniques in the area of internal audit risk assessments, from an analysis of extant literature. Assertions of scalable analytics adoption have also been sporadically tested empirically (especially in the past decade and prior to the emergence of generative AI), with a case in point being Brennan and Teeter (2010) demonstrating how the creation of automated audit rules in place of a pre-existing audit work plan resulted in scalable usage of audit tests by the subject organisation. Therefore, the requirement and desire for scalable and holistic adoption of analytics and / or automation techniques in audit functions is not new. The common link across these research studies is automation, which would enable feasible and scalable adoption of analytics capabilities to drive innovation in an IAF.

REGULATORY EXPECTATIONS

Prior literature has recognised the important role played by regulators in the adoption of advanced analytics techniques and emerging technologies (Warren and Smith, 2006; Chatzigiannis and Konstantinos, 2021; Wellmeyer, 2022). This aspect appears to be pronounced in multiple ways.

Given the regulatory focus on IAFs by bodies such as the IIA and Financial Conduct Authority (FCA), a major concern with the design of auditing techniques is the ability to deliver “explainable” innovation in the light of regulatory reviews and challenge. This view finds favour in recent research, with Wellmeyer (2022) finding that despite increasing investments in data analytics technologies, auditors continue to rely heavily on audit sampling techniques, and citing “fear of inspection findings and lack of clarity from regulators / firms on testing full populations, cross-border data usage and independence issues” being key drivers for this approach.

This observation appears to be true not only for audit data analytics, but also emerging technologies such as blockchain, where the enforcement of policy regulations and availability of auditing mechanisms are considered crucial to scalable adoption (Chatzigiannis and Konstantinos, 2021). Therefore, as a corollary, one could also argue that a positive spell of regulatory interaction could also hasten the scalable adoption of innovation arising from emerging technologies, as supported by Auer (2022). Indeed, the fear of regulatory inspection or censure has not prevented the application of advanced technologies for innovating in audit, with research by Dai and Vasarhelyi (2017) and Munevar and Patatoukas (2020) exploring the application of blockchain technology on auditing. However, despite arguing the merits and possibilities of blockchain technology for the field of auditing, the Munevar analysis highlighted regulatory uncertainty as the “top barrier” to technology adoption.

At the same time, regulators recognise the need for innovation, a sentiment taken almost as a given in recent academic literature in the field (Bibler and Gold, 2023). This is also by no means a recent phenomenon. Warren and Smith (2006) pushed their working paper arguing for the need for continuous auditing as part of the internal audit toolkit to address fraud and offer a modern definition of continuous auditing in the light of disruptive technology. In doing so, they assert that “Once-a-year reviews”, typical of IAFs, are often insufficient. This view is supported by Omidirouzi and Wang (2020), who assert that a “silo approach” to auditing fails in the light of rising AI capabilities.

Further research, such as Perols et al. (2015) has explored the possibility of developing analytics models to detect financial statement fraud, asserting that one of the possible benefits of this approach would be improving the ability of the regulator (the Securities and Exchange Commission (SEC) in this case) to detect fraudulent filings and support IAFs with audit planning. Similarly, Locke and Bird (2020) argue that technology may lend “real benefit” to governance practice and encourage legislative amendments where required, presumably to accommodate the impact of growing technology.

Therefore, literature does suggest an inextricable linkage between the adoption of innovation (via either emerging technologies or advanced data analytics), and regulatory interactions. A safe conclusion from this literature study is the need for auditors to keep innovating to make auditing a meaningful and valuable endeavour, a necessity in the light of disruptive technology. However, prior literature also highlights the importance of being cognizant of the pace of regulatory change, and enabling the “explainability” of innovation whilst simultaneously ensuring compliance with regulatory expectations. As such, research question RQ2 represents a valuable starting point for further study, which will contribute to a scant but steadily growing body of scholarship supporting academics and practitioners in the field of Internal Audit.

As discussed in section 2.2, there exists a clear research gap in prevailing literature assessing the impact of ML on Internal Audit Risk Assessments. This is perhaps also attributable to the fact that internal audits are inherently a subjective activity; indeed, with varying perceptions of auditors and auditees having a direct correlation on perceived efficacy and value (L'elis and Pinheiro, 2012). This also means that different IAFs could end up adopting different approaches to evaluate and prioritise organisational risks for audit's attention. However, as discussed by Glette-Iversen et al. (2023), there is a pressing need for reforming risk assessment techniques more generally, given the growing recognition of the importance of risk-based decision making - which in audit would be translated as RBIA.

The scope of the SLR was therefore expanded to consider prior attempts to leverage AI and ML techniques on risk assessments in other fields, to study whether research question RQ2 has been explored in the context of a non-auditing field.

Firstly, there is a large extent of literature exploring the value of risk assessments in other fields, given the foundational nature of this process within any risk management framework (Lyon and Hollcroft, 2012). Three unconnected examples are reproduced below, where the relevant research has explored the approach, impact and methods to be used for incorporating ML techniques for supporting judgemental risk assessments:

1. Berk et al. (2019) serving as an example of how ML techniques could be used for risk assessments in criminal justice settings. The Berk et al. (2019) study explored a range of supervised ML techniques, starting with linear regression, to support decision making in support of crime prevention. The study framed the business objective (crime prevention) as an ML problem statement, namely, forecasting the different kinds of crime in a particular area, and exploring in detail various supervised ML techniques, such as regression and classification to work through the problem. In doing so, a range of key assumptions and limitations (such as missing data and the need for practitioners to possess the requisite skills to deploy ML techniques) were considered by the authors as they developed their models and arrived at their conclusions. The study highlights "transparency, accuracy and fairness" as a bespoke evaluation model for assessing the efficacy of the techniques deployed, and the need for continuous improvement of data quality and stakeholder interaction - principles also found in the realm of internal auditing.
2. Similar conclusions arrive from an examination of research by Hegde and Rokseth (2020), who deployed a range of ML methods to perform engineering risk assessments. The study, which was a considered literature review, showed that the automotive industry is leading the adoption of ML algorithms for risk assessments, with

artificial neural networks being the most commonly applied technique for gathering insights through ML capabilities, such as predicting car crash severity or evaluating the probability of harm in the safety of machinery. Whilst it remains to be seen whether this ML technique would be directly relevant to research question RQ2 for this dissertation, the findings of this study demonstrate the relevance of ML for the principle of risk identification and management.

3. Zennaro et al. (2021)'s research offers a detailed analysis exploring the potential of ML for climate change risk assessment. The Zennaro et al. (2021) analysis notes an increasing interest in ML within the climate change risk assessment literature since 2015, with an assessment of the types of data used to inform the models. Similar to the Hegde and Rokseth (2020) and Berk et al. (2019) analyses, the Zennaro et al. (2021) study looks to structure the core business issue, or underlying research questions, as ML problems, identifies the most suitable methods for assessment, and catalogues limitations, improvement opportunities and suggestions for value-add. Specifically, the study notes ensemble and hybrid models as being "extensively applied" to improve performance in risk assessment.

These three case studies, though unconnected to internal audit and each other, exhibit several useful traits that are transferable to an evaluation of RQ2:

- Firstly, each problem considered by these three examples focuses on a core topical subject of wider social, organisational or international importance. None of the three topics in the previous paragraph are directly correlated to statistics or mathematical sciences alone, which reaffirms the potential of machine learning capabilities to be suitably deployed at scale to solve specific business issues.
- Secondly, each of these topics revolves around a subjective activity. From predicting the volume of crimes in a neighbourhood (Berk et al., 2019) to flood possibilities (Zennaro et al., 2021), the subject matter being solved for with ML techniques is judgement-based, and requires subject matter expertise. In each case, the subject matter of research was reframed as a machine learning problem, or several such problems, with each ML-based problem statement evaluated using a range of methods. There are parallels in related research, such as by (Alamri, 2022; Koivu et al., 2018), among others.
- Thirdly, a range of ML techniques have been deployed to solve for risk assessment outcomes in prior research. In addition to the research cited above, similar research by Kourehpaz and Hutt (2022) exploring ML techniques for seismic risk assessments also adopts a comparative evaluation-based approach, testing a range of methodologies from logistic regression, k-nearest neighbours, decision trees, random forest,

ADABOOST and gradient boosting, among others. The choice of methodologies does not appear influenced by a single attribute of study, and is instead driven by a pre-determined value measurement technique relevant to the specific study. However, these studies not only evaluate multiple models, but highlight a final, successful model for practical implementation, which serves as a useful pathway for answering research question RQ₃ on the subject of internal audit risk assessments. Identical approaches have been taken in similar studies such as (Lin et al., 2021; Dixit, 2021; Ossai, 2019).

- Fourthly, all studies exploring the impact of ML for risk assessments highlight the importance of ensuring adequate data quality prior to commencing any modelling efforts. This is a key requirement and constraint in equal measure, which has also been highlighted in related studies such as Bello (2023), who leveraged ML techniques for credit risk assessments, adopting an identical approach to the other studies referred to in this section. Bello (2023) identified the potential of ML techniques to "revolutionize credit risk assessment practices in financial institutions", whilst also noting the importance of data availability and quality to enable successful implementation of ML-based techniques for similar risk assessment initiatives.
- Finally, there is no single value measurement technique to assess the best performing ML model for risk assessments, and the choice of value measurement, which directly influences practical implementation, is driven by the attributes of the subject matter being evaluated. For example, the Kourehpaz and Hutt (2022) study considered "prediction success" to be the F1 score of the final chosen model, given the subject matter of analysis as being the identification of earthquake-induced damage and collapse risk. However, analysis by Koivu et al. (2018) which evaluated the feasibility machine learning algorithms for risk assessing Down's syndrome reported the efficacy of models based on the false positive rate, given the importance of the study and impact on healthcare. Indeed, research in the field of healthcare has also focused on identifying new metrics to support risk assessments, such as Casanova et al. (2013), who put forth a bespoke metric based on their research to risk assess Alzheimer's disease. As such, a considered view of multiple evaluation options should be taken prior to the building of suitable modelling techniques, with a clear justification for the choice of ML metrics for assessing value-add.

An exploration of risk assessment-specific literature across myriad fields has yielded interesting insights to help answer RQ₁. The overarching importance of ML for the financial sector is also underscored in extant and emerging research, such as Bello (2023) and Bauguess (2017). Specifically, Bauguess (2017) note the growing importance of ML and AI in the financial sector, examining the growth of similar initiatives in associated compliance

and regulatory activities. The Bauguess (2017) study notes the growing importance, potential and limitations of ML and AI in market regulation, supplementing the findings of the SLR in section 2.2, where regulatory impact was noted as a key driver influencing the uptake or otherwise of ML in the field of auditing. This limitation is not specifically highlighted in the supplementary studies reviewed in this section. This is perhaps attributable to one of the key findings of RQ1, i.e., the fact that regulatory interest in internal auditing is constant and ever-evolving, and recognises the role that innovation has to play in the audit lifecycle. In any case, the promising potential of ML for risk assessments in other fields, combined with the research gap in exploring ML methods for internal audit risk assessments and measuring the value add from doing so, render RQ2 a meaningful starting point for detailed analysis. The SLR findings and broader comparative literature evaluation will be examined as part of RQ2 in the subsequent chapters. Pursuant to an examination of RQ2, RQ3 will consider how the resultant value can be measured by IAFs as they seek to adopt the findings of this work in practical settings.

OPPORTUNITIES FOR INNOVATION

This chapter collates the findings of the literature review performed in chapter 2, and frames a research methodology that is subsequently applied across chapters 6, 7 and 8.

3.1 PRELIMINARY TAKEAWAYS FROM LITERATURE REVIEW

The SLR found examples of empirical research exploring the method, value, and application of advanced data analytics, including emerging technologies, in the field of internal auditing. However, despite the examples and qualitative benefit thereof, there is a gap in research literature exploring the possibilities of applying advanced analytics or machine learning specifically in the field of internal audit risk assessments and audit planning, which could boost the effectiveness of an IAF. Equally, whilst literature highlighted potential roadblocks to scalable adoption of analytics, the SLR also found that the research questions RQ2 and RQ3, framed in chapter 1 have not been adequately answered in extant academic literature.

The limitations highlighted in previously completed academic research pertaining to risk assessments, as noted in sections 2.2 and 2.3 are not showstoppers to experimental empirical research on the use of advanced data analytics or machine learning to support innovation in the area of internal audit risk assessments, given the potential of ML methods for risk assessments in the field, as highlighted by Bauguess (2017) and Kahyaoglu and Aksoy (2021). Drawing on the SLR findings, RQ1 is evaluated via a detailed assessment of the literature across each research theme, as below:

1. **Empirical Analysis:** Researchers have attempted to test the application of various data analytics and machine learning techniques across the audit lifecycle, taking anonymised datasets and well-considered assumptions to base their empirical research. Furthermore, there are limited examples of how innovation and data analytics could benefit the area of internal audit risk assessments. Given the pivotal nature of risk assessments as highlighted previously, experimental research in this area could lay the foundation for scalable industry-wide adoption of a new technology and support further research.

2. **Qualitative Impact:** Almost all studies unanimously concluded the qualitative impact of innovation in the field of internal auditing. A range of approaches have been used to arrive at this outcome, from full population testing / process mining to classification algorithms to predict test outcomes. The qualitative assessments are based on individual case studies that are primarily the preserve of audit fieldwork. It is unclear whether the use of similar innovation capabilities could support the starting position of the audit lifecycle, namely risk assessments. The relatively high success rate of innovation in areas of internal auditing other than internal audit risk assessments offer an encouraging opportunity to apply a similar approach to testing the value-add in the area of internal audit risk assessments. In particular, the Bäßler and Eulerich (2023) analysis offers a useful example of how classification techniques were successfully applied to identify control failures prior to formal audit completion. Given that a similar requirement exists in the area of internal audit risk assessments, the approach taken by the Bäßler and Eulerich (2023) study could be adapted to new research.
3. **Scalability:** The topic of scalable adoption of advanced data analytics has not been thoroughly explored by prior literature. However, the examples that do touch on scalable application of new and innovative solutions invariably make reference to automation techniques to enable such scalability. Research around the usage of automation capabilities to streamline the development and delivery of real-time or near real-time assurance techniques appears to have been limited to the fieldwork stage of the audit lifecycle, with little evidence of research exploring the application and / or scalability of innovation to internal audit risk assessments. Given the foundational nature of risk assessments, the feasibility of widespread adoption of innovation in the area of internal audit risk assessments would be a key component of future experimental research.
4. **Regulatory Expectations:** Regulatory expectations in the field of auditing are ever-present and appear to serve as catalysts and roadblocks in equal measure. However, what is clear from prior literature is the need to keep regulatory expectations on-board, whilst designing and developing a new innovation proposal. There are no examples in prior literature assessing regulatory expectations specifically in relation to the use of innovation for internal audit risk assessments. The IIA (2023a) state that the annual plan should be informed through a risk-based approach and need not necessarily be annual in nature. The effects of data or technology-based innovation in the field of internal audit risk assessments can therefore be studied in this context as part of experimental research, which would assure alignment with regulatory expectations.

This research work aims to leverage the opportunities summarised above to develop a framework for a machine learning-powered approach to risk assessments, which is data-

driven and sensitive to changes in data. The research aims to test a range of machine learning approaches applicable to the risk assessment process and closely examine the pros and cons of each approach. In addition to the opportunities arising from the gaps in existing research building up to this position (outlined in sections 2.2 and 2.3 above), there are several additional benefits to the proposed research subject area:

1. **Gaps in existing research:** Three out of four research themes explored as part of the literature review highlighted the lack of adequate comparable examples or attempts at empirical analysis of innovation in the field of internal audit risk assessments. This is despite the otherwise strong academic interest in the field of internal auditing and the focus on fieldwork-oriented research, where a range of prior examples are noted.
2. **Foundational nature of internal audit risk assessments:** Internal audit risk assessments serve as the basis of the audit plan at IAFs (Fukukawa and Mock, n.d.). The foundational nature of this step therefore serves as a strong motivation to identify and deploy new techniques to support judgements and decision making. In particular, the impact of innovation-based risk assessment delivery would be felt across the internal audit lifecycle, as opposed to individual business areas or specific audit steps.
3. **Regulatory alignment:** The requirement for an audit function to adopt a “risk-based approach” can be best fulfilled when the view of risk is dynamic and adaptable to real-time changes in the organisation. Therefore, the achievement of the proposed research objectives would also simultaneously support regulatory compliance objectives.

Whilst the research methodology would be flexible and adaptive to ongoing industrial and academic developments, the ultimate objectives, or success criteria of research are proposed as under:

1. To supplement a manual, judgement-based, subjective approach to internal audit risk assessments, with a **data-driven** approach.
2. To develop the solution with **automation capabilities**, so as to highlight updated risk assessment outcomes in near-real time, or on demand, to obtain early visibility of an evolving risk landscape.
3. To demonstrate the **executive and regulatory acceptance** of the solution, through qualitative feedback from stakeholders, and developing a value measurement framework to drive continuous confidence in the developed capability.

3.2 RESEARCH METHODOLOGY

Based on the findings of the literature review, this dissertation is structured as a five-step process as shown in table 3.1. This approach will be used to support ML model design decisions, with a view to initially answering RQ2 using a series of methods which have been covered in chapters 6 - 8 of this dissertation.

Ref	Stage	Description
1	Defining the framework	The first step in this project would be the translation of the business problem statement to selected machine learning frameworks.
2	Data Collection	This stage would include the gathering of all relevant variables to support the development of various ML models. This stage would also consider data pre-processing and / or cleaning as required to render it suitable for modelling. Key attributes to be considered for data collection include, but are not limited to the following: • Identification of relevant features supporting internal audit risk assessments. • Accessibility of data for modelling. • Security considerations, including the anonymisation of data. • Internal vs. external features for model consumption. The selection of features to develop these models will be based on the prevailing approach adopted at the sponsoring organisation, given the research gap in extant literature and diversity in risk assessment methodologies across IAFs as shown in the different approaches adopted by Kahyaoglu and Aksoy (2021) and Arzhenovskiy S.V. (2019) in their recent empirical analyses of the subject.

3	Model Development	Given the exploratory nature of this research and consistent with the findings of the SLR as noted in section 2.2, this project will consider not one but multiple machine learning approaches, with a view to identifying the most suitable strategy to support internal audit functions with a machine learning based, data-driven risk assessment. This stage will involve iterative experimentation and the documentation of final models chosen for conclusion purposes will include a record of the hyperparameters used for model design, including rationale for the selected values and fine-tuning, if any.
4	Results	Following the development of multiple research models, the work will be holistically reviewed to enable conclusions by a wide range of academics and practitioners, as the objective of the research is to identify multiple possibilities for machine learning implementation, rather than suggest a one-size-fits-all solution.
5	Future Research	The final stage in the research will include a discussion on potential future research in this field, leveraging the outcomes of this project.

Table 3.1: Research methodology

Across chapters 6 - 8, the risk assessment exercise is approached as a standalone machine learning problem, drawing out the relative effectiveness and value-add from switching the risk assessment process from a manually immersive one to a data-driven, automated exercise. This analysis is based on the risk assessment process actively used at the IAF of the sponsoring organisation, leveraging actual data for c. 150 Business Units (BU), across multiple variables used for IR and CE assessments. These have been anonymised to preserve confidentiality. Regardless of the specifics of the final modelling approach identified from research, the expectation is that the use of machine learning capabilities such as classification would enable the IAF to generate value for this foundational step, thereby enabling the IAF to move beyond auditor intuition alone for risk assessing the enterprise.

Part III

GETTING STARTED: WHAT DATA DO WE NEED?

DATA COLLECTION

This chapter introduces and examines the features used to perform machine learning experiments across chapters 6 - 8. Given the bespoke nature of internal audit risk assessments performed across organisations, the chapter also links generic risk assessment features used in practice with the specific features used within the sponsoring organisation, which serves as the baseline input dataset considered for machine learning models used in the research.

4.1 RISK ASSESSMENT METHODOLOGY

In order to understand the nature, extent and volume of variables required to support a machine learning exercise involving the risk assessment process, an analysis of how risk assessment is carried out in practice is required.

Extant research shows that risk assessment techniques range from the use of questionnaires (Demirović and Proho, 2021) to the use of big data (Kahyaoglu and Aksoy, 2021), however, the fundamental principles of risk assessment techniques include the following steps:

1. Divide the organisation being risk assessed into **individual units** for risk assessment, commonly referred to as "audit universe entities" or BU.
2. For individual BUs, collect a range of quantitative and qualitative **variables** to score the inherent risk (IR) and control effectiveness (CE) factors (Demirović and Proho, 2021) driving the auditor's overall appreciation of residual risk for the BU.
3. **Holistically assess the residual risk profile** of the organisation, to identify seminal trends and patterns, to arrive at an audit plan for the target period.

4.2 DATASETS REQUIRED

There is insufficient evidence in literature (Awuah et al., 2021; Shivram, 2024) to outline a comprehensive and consistent list of the data that may drive the evaluation of inherent risk and control effectiveness scores, especially given how organisations differ across geographies, industries and cultures. However, the following list of variables serves as a general, illustrative list of "input groups" for subjective assessments of IR and CE, which would serve

as the target variables in the event these features are used to design a machine learning model for predicting risk assessment outcomes.

1. **A - Prior Audit Findings:** A numeric variable showing the number of open audit observations raised by the IAF from prior audit work, typically for a more recent timeframe (e.g. six months).
2. **B - Known Control Failings:** A numeric variable identifying the number of actual control failures self-identified by Management.
3. **C - Technology Landscape, such as applications and production change volumes:** The number of applications used to support the business unit, along with multiple technology-based variables, such as the volume of change programmes or production change tickets deployed in these applications, with higher figures across these attributes typically representative of a higher inherent risk for the BU as a whole.
4. **D - Regulatory Exposure:** A variable highlighting whether or not the BU is subject to periodic regulatory inspection.
5. **E - People Attrition:** An numeric variable highlighting people attrition for each business unit, used to identify potential volatility in the BU, or detect key person dependencies, if any.

The above list is illustrative and additional variables can be added or removed over time to reflect the most important attributes driving IAF perceptions of the risk landscape. However, given existing working practices for risk assessment at the sponsoring organisation, base data collection has been performed with a view to incorporating the above variables as a minimum, over and above additional features used to inform risk assessment cycles. The above features are referenced where appropriate in section 4.3 below.

4.3 BASE DATA COLLECTION

For the purposes of this research, input variables used by auditors to perform risk assessments across c. 150 BUs at one organisation were compiled into a single dataframe. The dataset included the following variables used to support the IAF's assessment of Inherent Risk or Control Effectiveness for each BU, which serve as two separate target variables, as these are the outcomes that the machine learning experiments would be designed to predict.

A mixture of numeric and non-numeric variables were used for risk assessments, and research began by assessing these for normality and spread in section 5.1. The IAF does not proactively assign higher weighting to individual variables, although there exists the possib-

ility of unconscious bias by auditors to base their subjective risk assessments on a subset of variables rather than the entirety of data at their disposal.

A profile of these variables is provided below for reference:

Numeric Variables:

1. **Key Applications:** The number of business critical or sensitive applications used by the business unit to perform regular operational activities.
2. **DR Non-Compliance Rate:** Expressed on a scale of 0 to 100%, this variable denotes the proportion of applications used by the business unit that are not compliant with Disaster Recovery (DR) test requirements as set out by the organisation's technology policies. For the purposes of risk assessment, the root cause of the non-compliance itself is ignored; and the rationale of non-compliance could range from documentation failures to the absence of any robust DR strategies or testing.
3. **ARP Non-Completion Rate:** Each BU is required to maintain Application Recovery Plans (ARP) to demonstrate how such applications may be recovered in the event of an extended business disruption. The non-completion rate of ARPs reflects the proportion of applications used by the BU that do not have such a plan, which could adversely impact their ability to operate effectively.
4. **Number of DR changes (last 6 months):** This variable represents the volume of change tickets raised by the BU to test the disaster recovery adequacy of applications used.
5. **Number of Production Changes Deployed (last 6 months):** A high volume of changes deployed to production environments within the applications is typically associated with BUs that are inherently high risk or business critical and is a common variable used by the IAF to support risk assessment activities.
6. **Number of unauthorised changes (last 6 months):** Where change tickets are not deployed through a controlled, authorised process flow, these are referred to as "unauthorised", and typically represent a control failure on the part of the BU. As such, a review of the volume of such unauthorised changes is important as part of risk assessment activities.
7. **Number of Resilience SLA breaches (last 6 months):** IAFs, and increasingly, regulators such as the Financial Conduct Authority (FCA)'s, are required to assess the effectiveness of their organisation's operational resilience arrangements (Murdock, 2021). As such the number of breaches in the Service Level Agreements (SLA) for adequate operational resilience serves as a key variable for the organisation's risk assessment activities. This variable also forms part of input group C.

8. **Risk Events:** A risk event represents a materialised control failure that has impacted the organisation, and may have varying levels of severity. Risk events are typically self-labelled by Management teams as being either "High", "Medium" or "Low" severity and data collection has been stratified across "High" and "Medium" level risk events for the purposes of this study.
9. **Risk Event Action Plans:** This dataset represents the open action plans raised to mitigate risk events raised in the six months prior to this thesis. A high volume of open action plans for risk events indicates that remediation activities are yet to be implemented by Management to enable satisfactory closure of the relevant risk event(s), and therefore this variable would serve as a useful input to risk assessment activities. Similar to risk events themselves, these action plans are stratified across High and Medium labels.
10. **Open Audit Findings:** Similar to Risk Events, a key variable used by IAFs to perform risk assessments is the volume of observations raised by the IAFs against BUs from previously completed audit work, and specifically, the volume of observations that are yet to be satisfactorily mitigated by Management to sufficiently reduce the open risk to an acceptable level. As the IAF at the sponsoring organisation only considers audit findings rated "High" or "Medium", and does not consider "Low"-rated findings for risk assessment, only "High" or "Medium" rated findings have been collected.
11. **Open Compliance Monitoring Findings:** Audit observations may also be reported by assurance functions other than internal audit, such as those based in the organisation's second line of defence. These are referred to as "Compliance Monitoring" findings for the purposes of this thesis.
12. **Open Change Projects:** The number of change programmes within a BU signifies volatility in operations and as such has an impact on the BU's risk profile. This variable is therefore used to support an assessment of inherent risk. Only the volume of approved, funded change programmes has been identified for each BU for the purposes of this study.
13. **Customer Impacting Risk Events:** In addition to collecting the volume of risk events that a BU has been "responsible" for, i.e., where the risk event occurred due to a failure in underlying controls maintained or owned by the relevant BU, the number of risk events with a customer impact are also identified, regardless of whether the risk event has been as the result of a failure of the BU's controls framework or not. This variable has been split across retail and non-retail customers, given the organisation's customer profile.

14. **Customer Complaint Volumes:** This shows a snapshot of customer complaint volumes (considering only upheld complaints), as these are indicative of a sub-optimal controls framework (particularly for customer-facing BUs).
15. **People Attrition Ratios:** This variable represents the average attrition ratios for each BU, as the IAF associates high people turnover with a higher risk level for the relevant BU.

Reviewing the above list, the specific variables used to perform risk assessments at the sponsoring organisation are linked to the general set of variables compiled in section 4.2 as follows:

1. Variables 1 to 7 inclusive are examples of input group **C - Technology Landscape**.
2. Variables 8 and 9 are examples of input group **B - Known Control Failings**, given that risk events represent known control gaps self-identified by Management.
3. Variables 10 and 11 are examples of input group **A - Prior Audit Findings**, and bring together inputs from both Internal Audit and Compliance Monitoring functions.
4. Input group **E - People Attrition** is reflected in variable 15.
5. The remaining variables are bespoke to the sponsoring organisation, and part of internal audit methodology.

The above numeric variables comprise of a mixture of Continuous and Discrete variable types, which are defined as follows:

1. **Continuous Variables:** These are numeric variables that can take on any value within a given range. Examples include DR non-compliance ratios, people attrition ratios and ARP non-completion ratios. This is because such variables can theoretically assume any value within a range of 1 - 100%.
2. **Discrete Variables:** These represent countable, specific values. The majority of variables presented in the list above are discrete, such as the volume of production changes, risk events, open audit findings and the number of change programmes.

For consistency, this research refers to both continuous and discrete variables above as "numeric" variables.

In addition to the above variables, the risk assessments also leverage the auditor's subjective assessment of factors that also serve to inform the ultimate risk level of a particular BU, even if not quantifiable as the variables listed above. This includes a consideration of input group **D - Regulatory Exposure**, which is included as item 3 in the list below.

Non-Numeric Variables:

Non-Numeric Variable	Risk Rationale	Scoring or Encoding Approach (E.g.: 1 = Low, 5 = High)
Does the BU have a formal project management team to support change programmes?	Lack of formal project governance may lead to uncontrolled risks during transformations.	No = 0, Yes = 1
Is the BU customer-facing or not?	Customer-facing units have higher reputational and operational risk.	No = 0, Yes = 1
Is the BU subject to regular regulatory scrutiny or reviews?	Regulatory oversight implies increased compliance risk and audit attention.	No = 0, Yes = 1
Are the operational activities or processes owned / performed by the BU manually intensive?	Manual processes increase error, fraud, and inefficiency risks.	No = 0, Yes = 1
Does the BU support a strategic business offering as identified by Senior Management?	Strategic BUs are critical to organisational success and reputation.	No = 0, Yes = 1
Does the BU handle or own confidential, or highly confidential datasets, such as Personally Identifiable Information (PII) relating to customers?	Data privacy and cybersecurity risks increase when sensitive data is handled.	No = 0, Yes = 1
Does the BU reflect a complex organisational structure?	Complexity can mask control failures and increase operational risk.	No = 0, Yes = 1
Is the BU subject to regular penetration testing?	BUs with externally facing applications are naturally subject to a heightened risk of cyberattacks by a malign actor, and are therefore required to undergo periodic penetration testing.	Annual = 1, Every Two Years = 2, Every Three Years = 3

Table 4.1: Non-Numeric Variables

Each variable in the table above is marked as either Yes or No by the auditor, based on their business expertise and knowledge of the relevant BU's processes and operational activit-

ies. These variables are manually marked by the auditor on each round of risk assessment, and the final entry for these variables from the preceding four rounds of risk assessment is used for modelling purposes. The above variables were compiled from multiple applications used by the sample organisation, which were identified by the IAF as reliable or authoritative sources of such data. These variables are categorical in nature, as they represent qualitative aspects of the organisation. The majority of these variables are nominal in nature, as they do not represent any intrinsic order of magnitude and are typically binary assessments by the IAF from previously completed risk assessment cycles. The exception to this is the penetration testing frequency, which is an ordinal categorical variable, given the meaningful order of potential values (annual, biennial or every three years). However, categorical variables collected in table 4.1 above are consistently referred to as "non-numeric" variables throughout this thesis.

Once collected, the data was suitably mapped to each BU on a best efforts basis, where assumptions were required (as the variables spawn different taxonomies). There are no known biases with any variable identified in the lists above.

Finally, the observations for each of these variables were refreshed for four cycles of risk assessment and individually labelled by the IAF as supporting either Inherent Risk or Control Effectiveness, enabling us to build two dataframes for analysis, with either IR or CE being the target variable for each. This would result in 158×4 , or 632 records for analysis, which would also incorporate temporal variation in risk profiles arising from changes to variables over time. An overview of the data collected is provided in figure 4.1.

Org Unit	Key Applications	DR not Compliance Rate (%)	ARP not Completion Rate (%)	Number of DR changes (last 6 months)	Number of Production Changes Deployed (last 6 months)	Number of unauthorised changes in ServiceNow (last 6 months)	Number of Resilience SLA breaches (last 6 months)	Eligibility for penetration testing coverage	...	CUST_FACING_DIV_YN	CUST_IMPACT_AU_YN
A1	49	0.00	0.00	0	0	0	0	Every Two Years	...	No	No
A2	0	10.00	0.00	0	32	0	0	Not Applicable	...	No	No
A3	2	50.00	0.00	1	84	0	0	Not Applicable	...	No	No
A4	22	0.00	0.00	5	400	0	0	Not Applicable	...	No	Yes
A5	9	23.81	14.29	1	59	0	0	Every Three Years	...	No	NaN

Figure 4.1: Data snippet for four BUs

DATA PROFILING

Following the completion of base data collection in chapter 4, this chapter outlines a detailed profiling and examination of each variable collected, with a view to incorporating these as features for machine learning experiments. Additionally, whilst these are the same features used by the sponsoring organisation to perform manual risk assessments, they required appropriate pre-processing and feature engineering to ensure their suitability as inputs for machine learning models. As such, this chapter also includes an overview of key statistics prepared for the numeric data, the approach to handling missing data, preliminary insights individual feature relationships with the target variables (IR and CE), and an assessment of feature independence. The insights from this chapter serve as the baseline for choosing key statistical tests used to examine hypotheses prepared in chapters 6 - 8.

5.1 DATA DISTRIBUTION

Having collected the data in the previous chapter, the numeric variables are now examined for normality in their distribution. The key statistics are given in figure 5.21 and are also represented here utilising both scatter plots and box plots as appropriate. These plots are shown in figure 5.1 to figure 5.10 for all numeric variables listed in section 4.3.

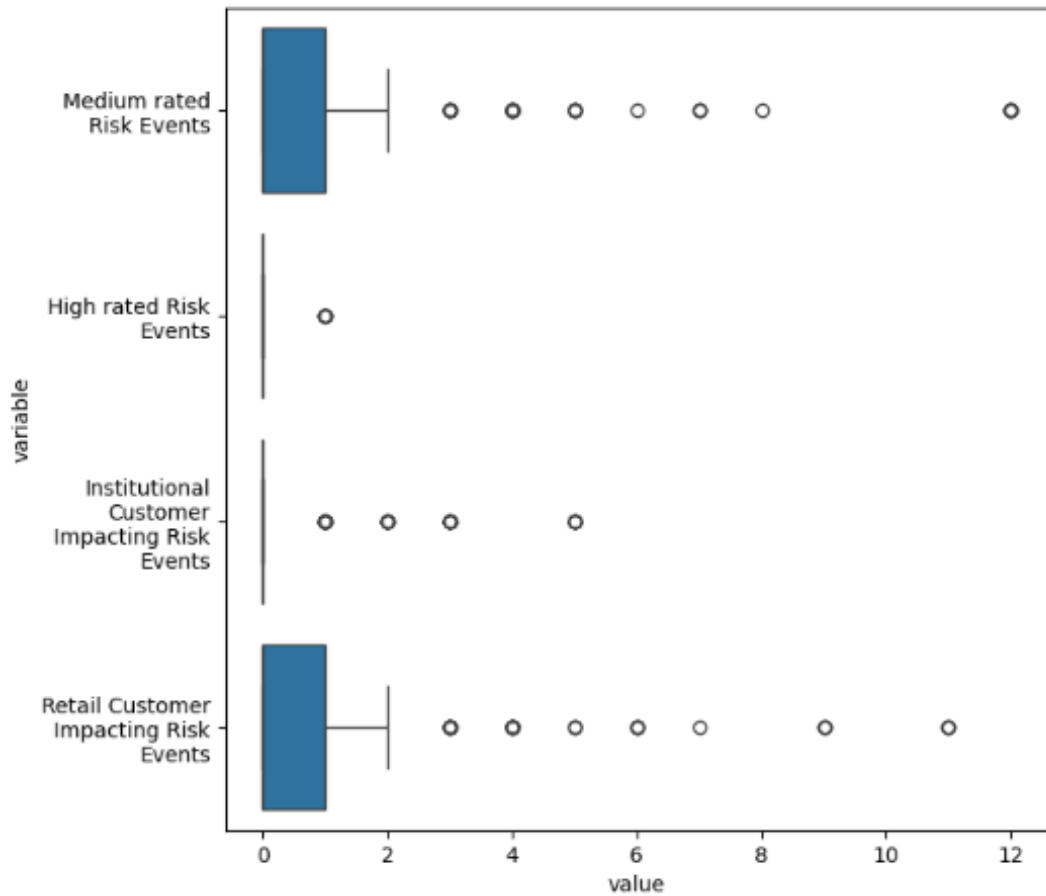


Figure 5.3: Risk Events

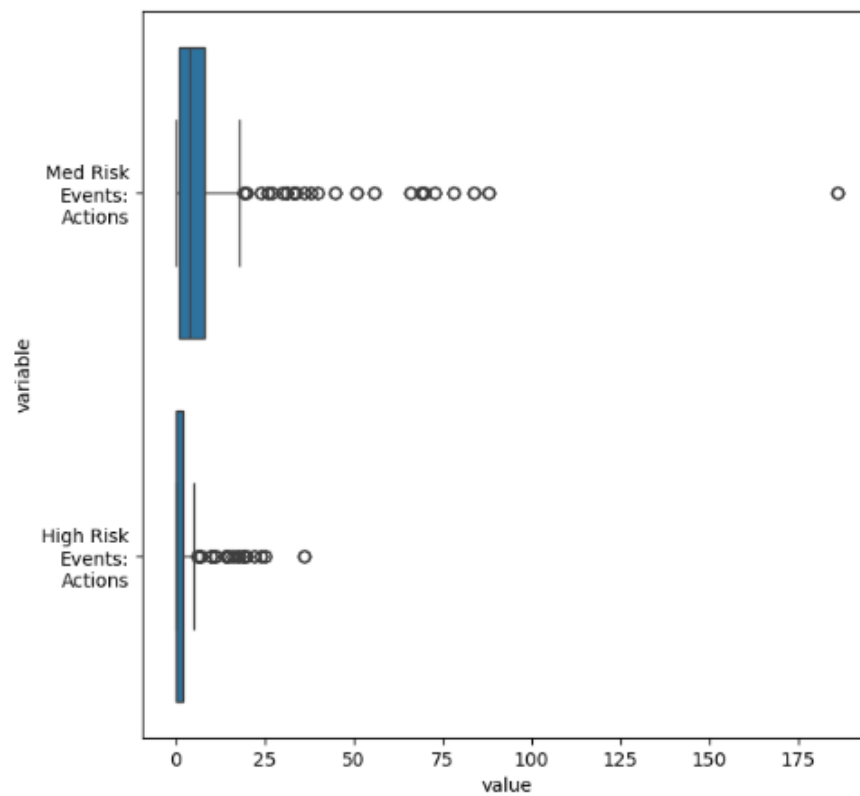


Figure 5.4: Risk Event - Action Plans

As risk events are expected to be a rare occurrence, it is not surprising that an observation of two high-risk events, or more than two institutional customer-impacting risk events, are marked as outliers. Medium-rated risk events, as well as those impacting retail customers are comparatively more frequent given the higher absolute number of retail customers, which is seen in the spread of observations. The boxplots for risk events (figure 5.3) and corresponding actions in figure 5.4 highlight expected volumes for such occurrences within the dataset, which corroborates and aligns with the IAF’s own conclusions and practical expectations.

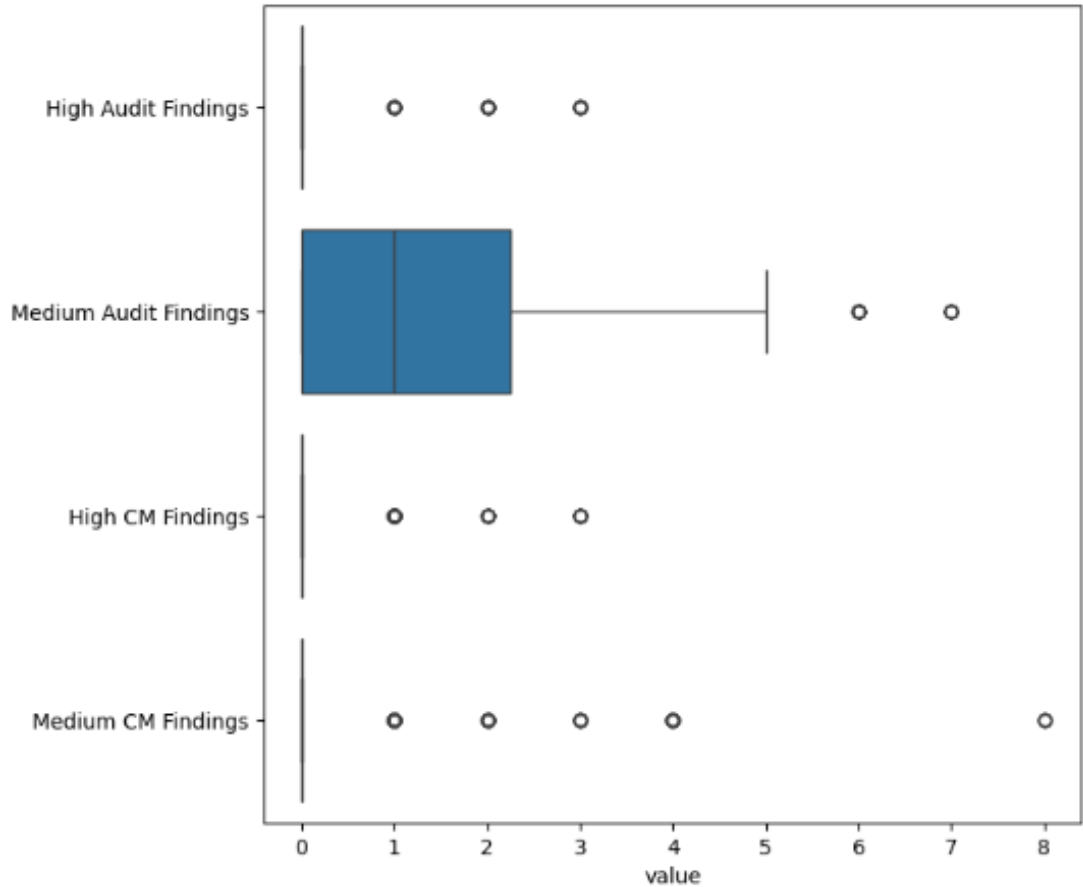


Figure 5.5: Audit Findings

Similar to risk events, high risk audit findings are not as frequently reported compared to medium risk findings, given the relative stability of the sponsoring organisation. As such, the boxplot justifiably marks all instances of high risk audit findings as outliers. The same is true of findings raised by compliance monitoring, a supplementary assurance function.

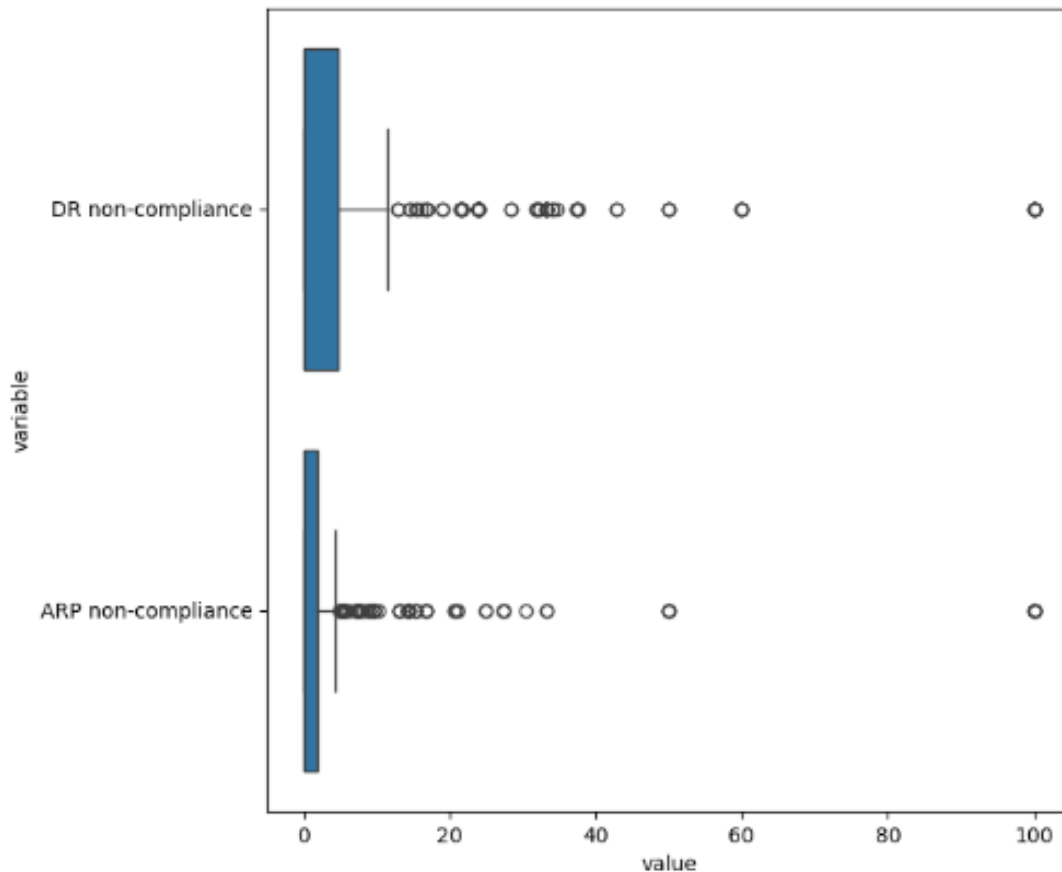


Figure 5.6: Audit Findings

The spread of applications non-compliant with DR requirements are consistent with IAF expectations. The boxplot in figure 5.6 highlights that the median non-compliance ratio with DR requirements, including the lack of ARP, is approximately 5%, which indicates overall compliance with technology policy requirements, with some concentrated areas of the organisation being non-compliant. Whilst such non-compliance could be attributable to a range of factors, from human error to organisational restructuring, significant non-compliance, such as lack of DR testing for more than 10% of the applications tagged to a BU, is justifiably marked as an outlier observation that requires IAF focus.

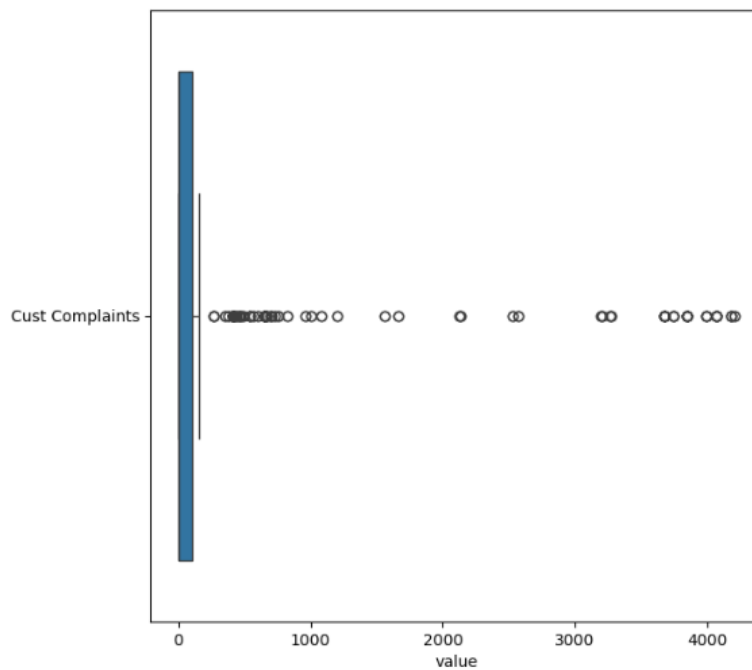


Figure 5.7: Customer Complaint Volumes

Customer complaint volumes are typically limited to those BUs that liaise with the organisation's customers, and a concentration of complaint levels within a restricted group of BUs results in outliers being called out in the box plots. This observation is aligned with the IAF's understanding of the organisation.

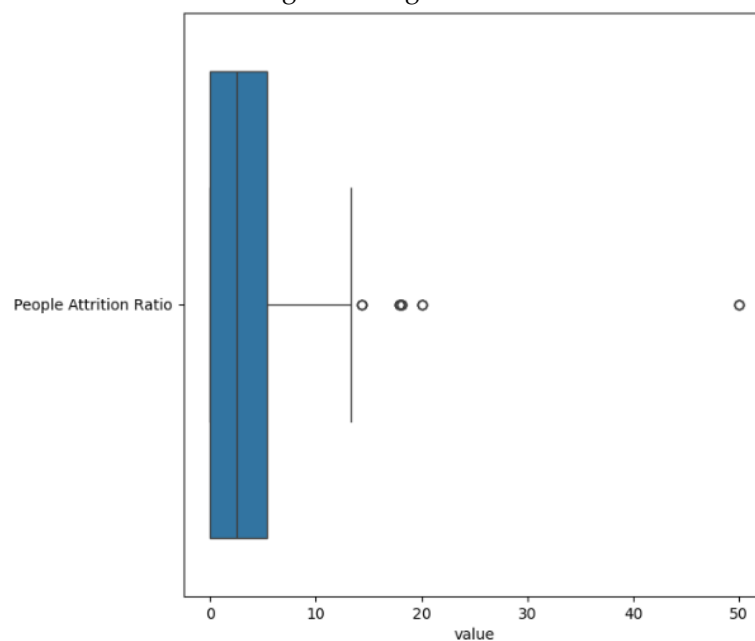


Figure 5.8: People Attrition Ratio

The median people attrition ratio is shown to be approximately 5%, which is standard for a large financial services organisation. The boxplot highlights attrition ratios greater than 15%

as anomalous, which is justified as such large levels of people turnover is not expected in an otherwise stable organisational environment.

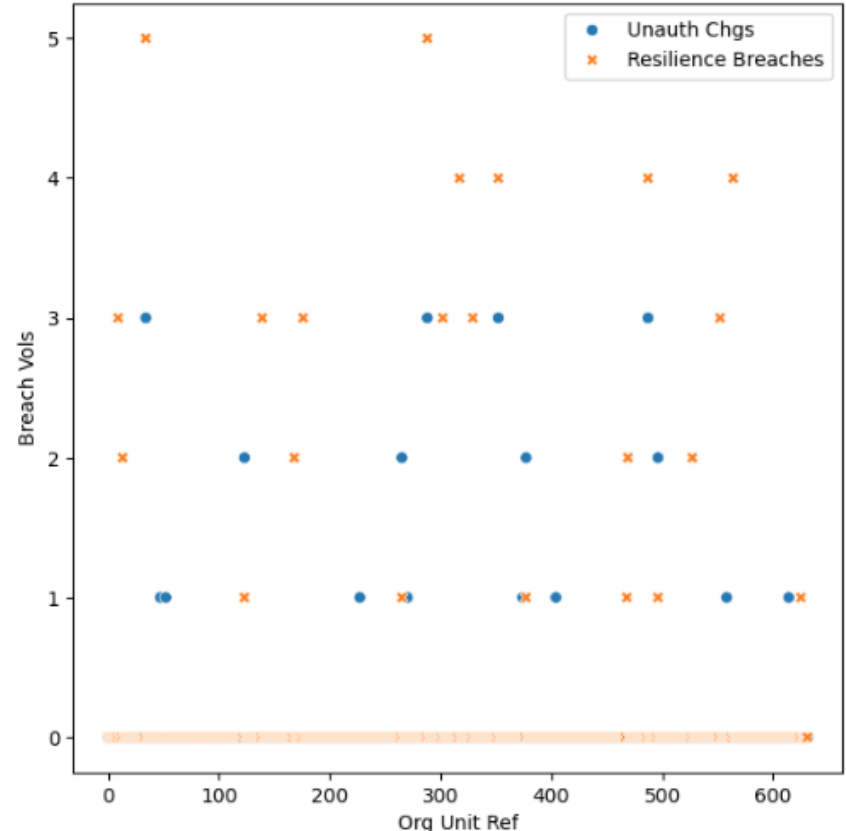


Figure 5.9: Unauthorised Changes and Resilience Breaches

Similar to observations around control gaps - borne out by both audit findings (figure 5.5) and risk events (figure 5.3) - unauthorised changes released to application production environments represent a failure of controls in place to mitigate the risk of erroneous code deployments. Barring a few exceptions, the concentration of organisational units towards the bottom half of the scatter plot highlights overall organisational compliance with change management principles, consistent with IAF expectations. The same logic applies to breaches of pre-defined resilience requirements.

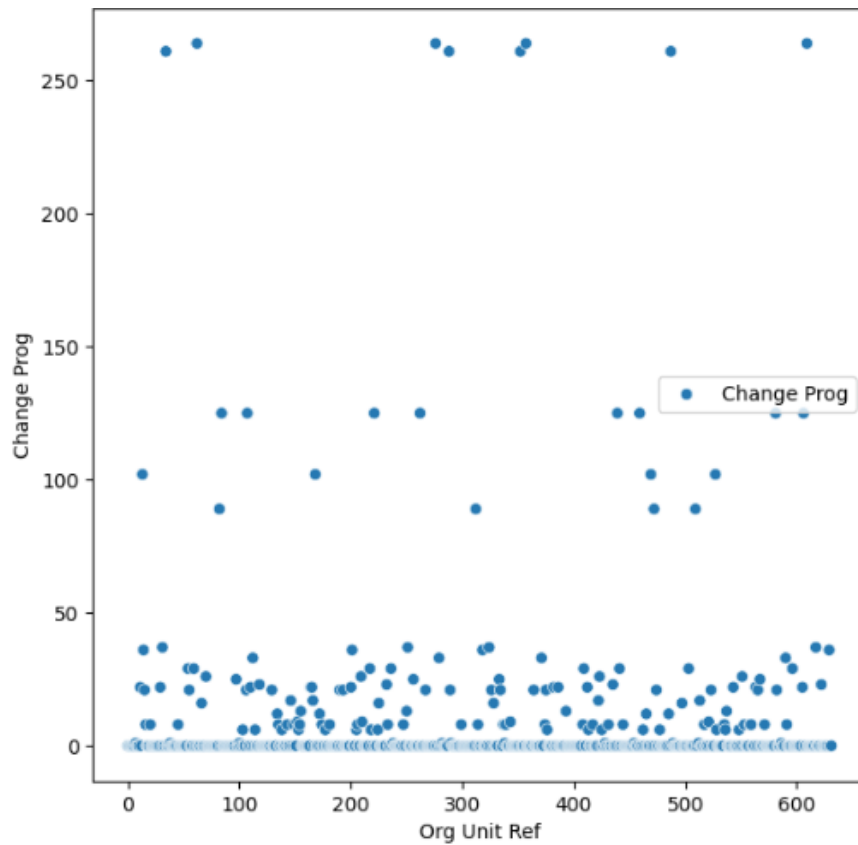


Figure 5.10: Change Programmes

At the time of this thesis, the sponsoring organisation is undergoing a significant organisational restructuring, which is reflected in a large volume of change programmes. Some of these change programmes impact certain parts of the organisation more than others, which is reflected in the scatter plot above.

Histograms for each of these features have been reproduced below to supplement the analysis of data distribution ahead of modelling. These complement the observations from the boxplots, highlighting the abnormal data distributions for all features collected.

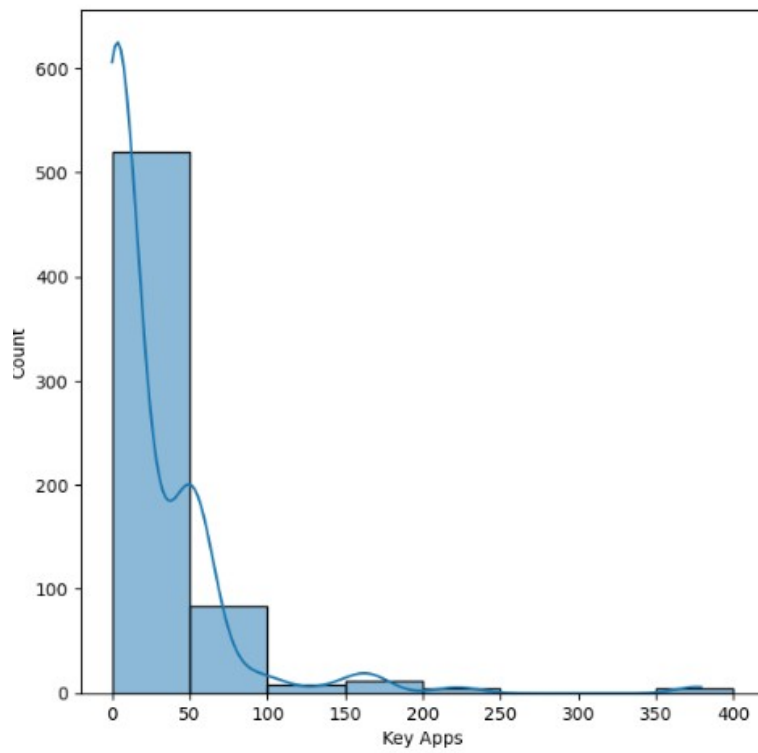


Figure 5.11: Key Applications

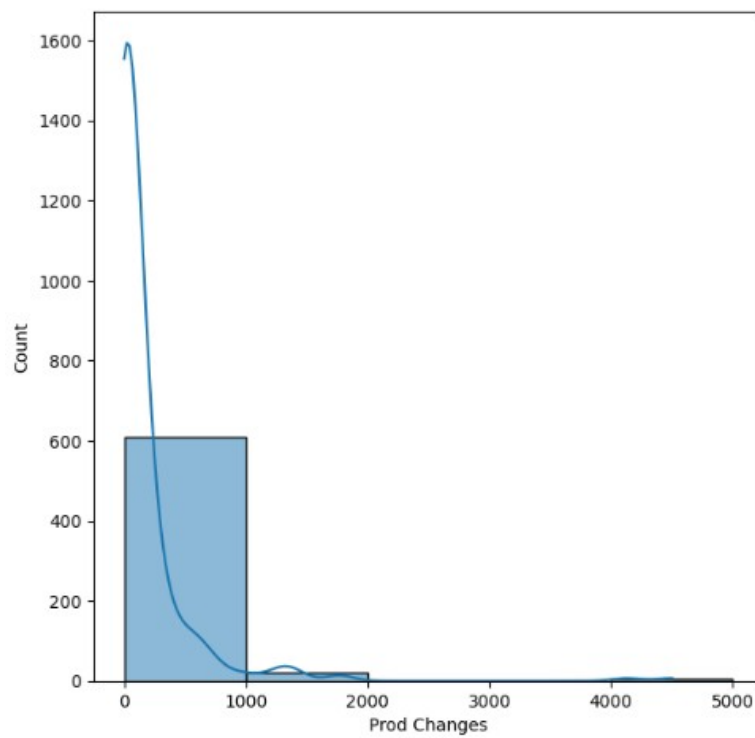


Figure 5.12: Production Changes

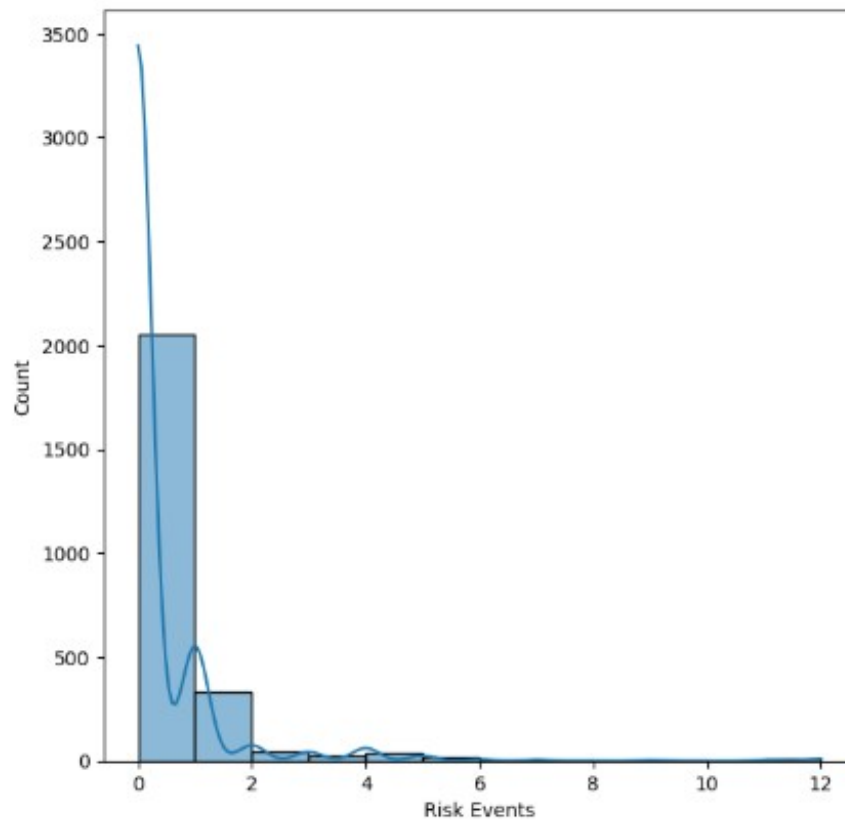


Figure 5.13: Risk Events

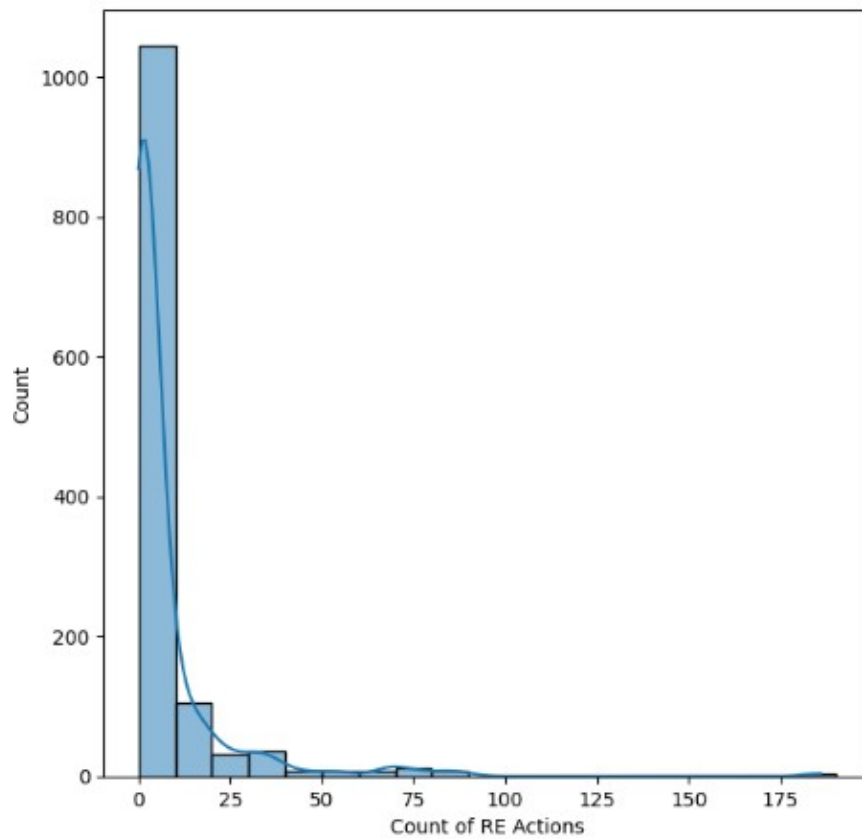


Figure 5.14: Risk Event - Action Plans

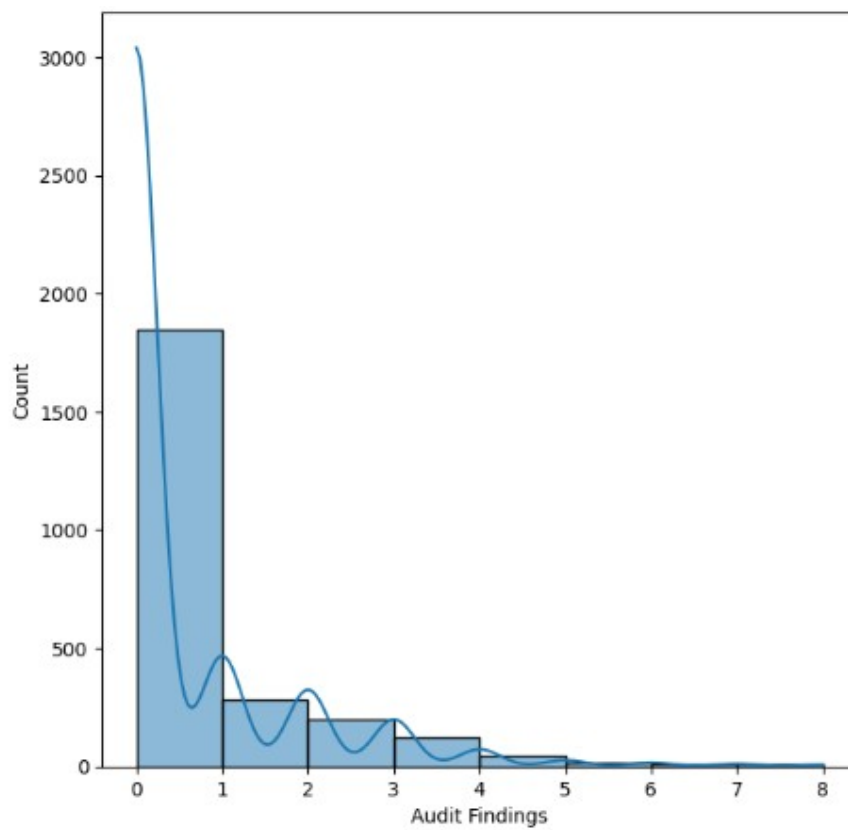


Figure 5.15: Audit Findings

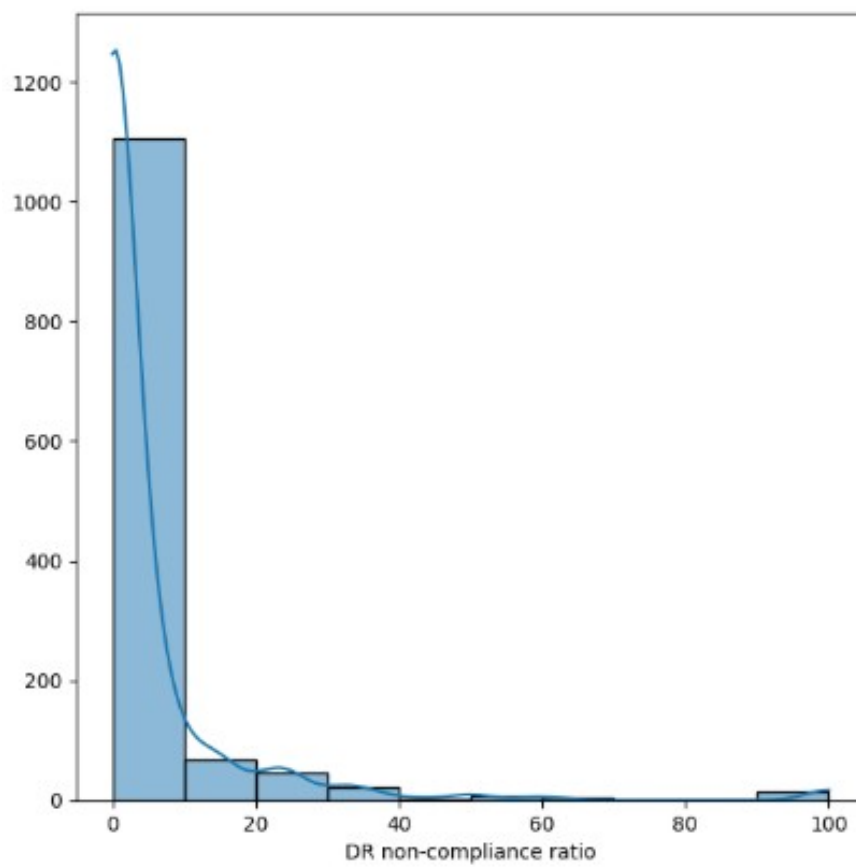


Figure 5.16: DR and ARP Non-Compliance

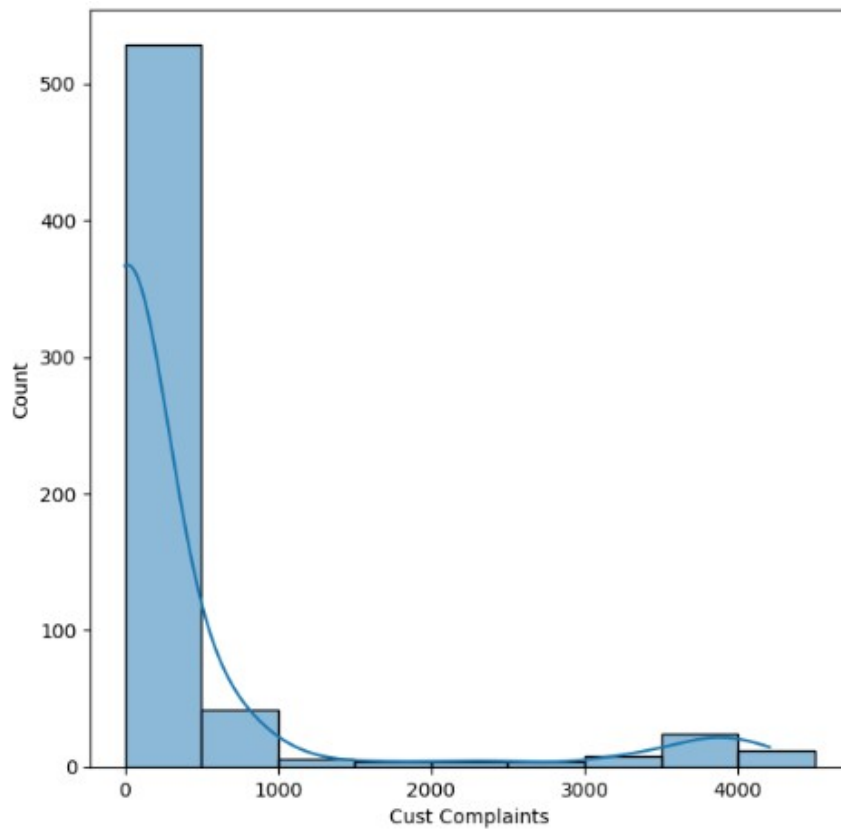


Figure 5.17: Customer Complaint Volumes

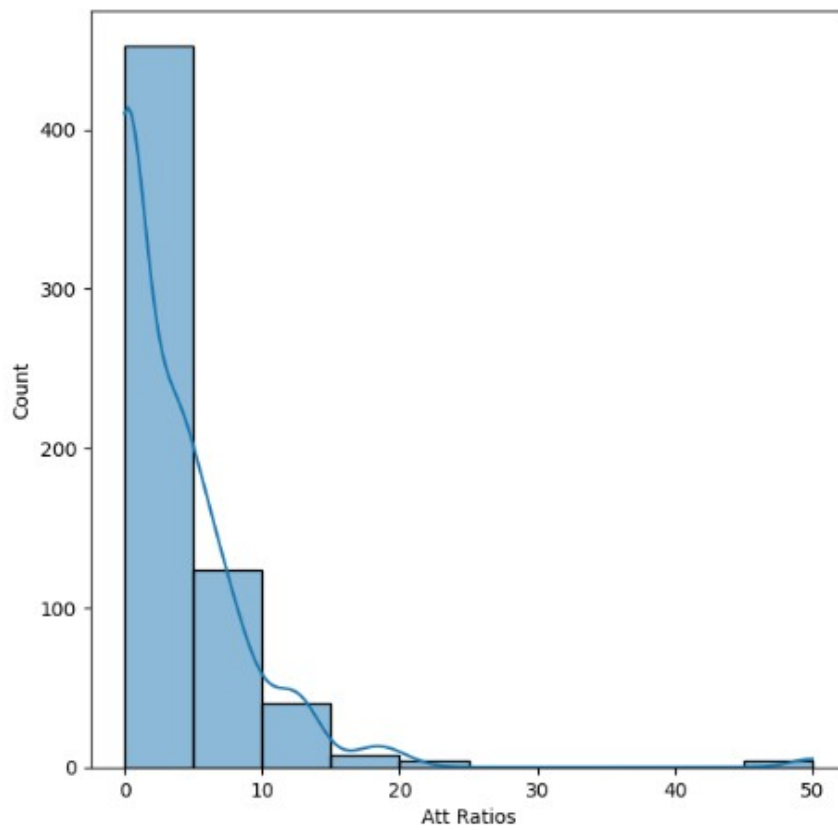


Figure 5.18: People Attrition Ratio

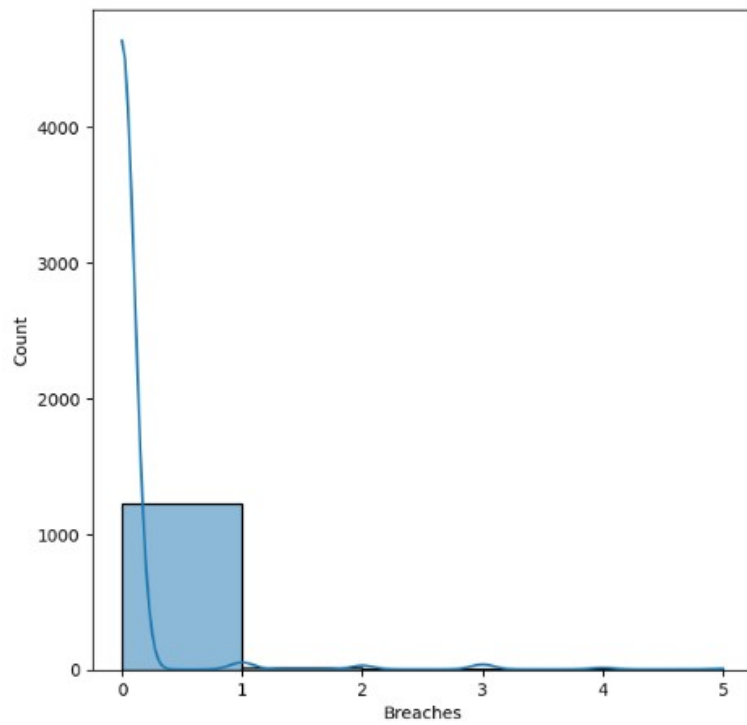


Figure 5.19: Unauthorised Changes and Resilience Breaches

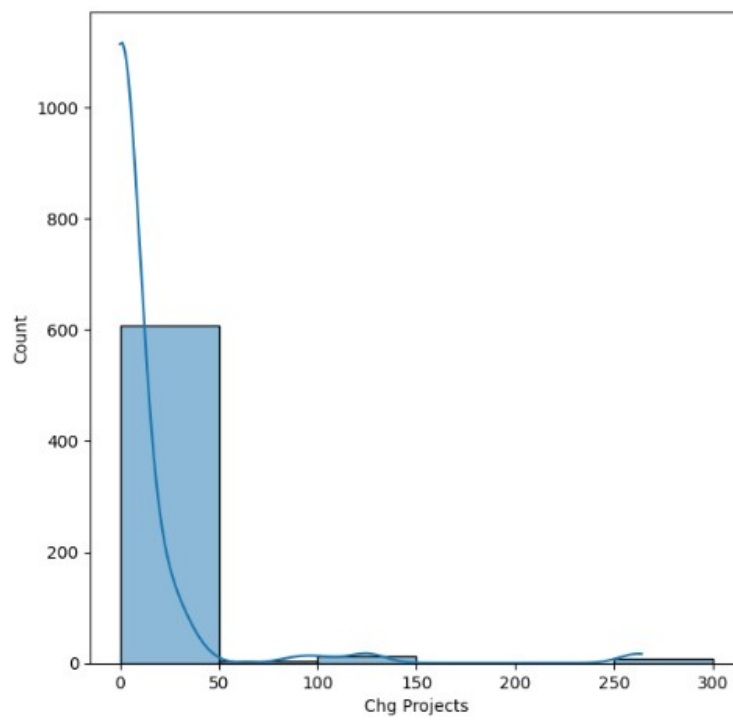


Figure 5.20: Change Programmes

	Key Applications	DR not Compliance Rate (%)	ARP not Completion Rate (%)	Number of DR changes (last 6 months)	Number of Production Changes Deployed (last 6 months)	Number of unauthorised changes in ServiceNow (last 6 months)	Number of Resilience SLA breaches (last 6 months)	Medium rated Risk Events	High rated Risk Events	Medium rated Risk Events: Open Action Plans	High rated Risk Events: Open Action Plans	High Rated Open Audit Findings
count	632.00	632.00	632.00	632.00	632.00	632.00	632.00	632.00	632.00	632.00	632.00	632.00
mean	25.13	5.52	3.66	3.38	163.79	0.04	0.09	0.72	0.01	11.16	2.84	0.11
std	45.84	14.07	11.40	9.25	433.43	0.31	0.54	1.61	0.11	22.51	5.84	0.46
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
50%	8.00	0.00	0.00	0.00	33.00	0.00	0.00	0.00	0.00	4.00	0.00	0.00
75%	43.00	4.65	1.89	3.00	171.00	0.00	0.00	1.00	0.00	8.00	2.00	0.00
max	379.00	100.00	100.00	79.00	4500.00	3.00	5.00	12.00	1.00	186.00	36.00	3.00
skew	4.08	4.28	6.38	5.72	6.41	7.84	6.57	4.29	8.74	4.19	2.90	4.75
median	8.00	0.00	0.00	0.00	33.00	0.00	0.00	0.00	0.00	4.00	0.00	0.00
kurt	23.36	22.53	48.09	38.63	53.74	65.34	45.18	23.55	74.61	23.82	9.15	23.28

	High Rated Open Compliance Monitoring Findings	Medium Rated Open Compliance Monitoring Findings	Number of Clarity Change Projects	Institutional Customer Impacting Risk Events	Retail Customer Impacting Risk Events	Customer Complaints - By Business Channel	People Attrition Ratio
count	632.00	632.00	632.00	632.00	632.00	632.00	632.00
mean	0.18	0.40	9.52	0.13	0.58	385.28	3.64
std	0.50	1.03	34.25	0.64	1.36	998.42	5.63
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00	0.00	0.00	2.50
75%	0.00	0.00	1.00	0.00	1.00	101.50	5.40
max	3.00	8.00	264.00	5.00	11.00	4207.00	50.00
skew	3.34	4.03	5.81	5.76	4.38	2.91	4.10
median	0.00	0.00	0.00	0.00	0.00	0.00	2.50
kurt	12.79	20.99	37.27	35.96	24.27	7.17	27.96

Figure 5.21: Key Statistics - Numeric Variables

As seen from the key statistics compiled, the collected data has a total of 31 features, split across 20 numeric and 11 non-numeric features. It can be seen from the positions of the quartiles that there is a lack of normality in the distribution of the data. The data is not symmetric around the median and is skewed in all cases. These features will play a role in determining the choice of statistical tests to be applied as part of model selection and design.

5.2 NORMALITY OF DISTRIBUTIONS: CONCLUSIONS

The statistics compiled in 5.21 highlight the lack of normality within the distribution of numeric variables. This is an important point of assessment that drives the choice of key statistical tests and models developed throughout this thesis, given that the validity of certain statistical tests, such as Pearson's correlation analysis, t-tests and linear regressions, relies on the presence of normally distributed data. Whilst the statistics and figures compiled in section 5.1 strongly indicate the lack of normality, as reinforced by the high skewness and kurtosis for each of the numeric variables compiled in figure 5.21, it is key to

confirm the normality or otherwise of each numerical distribution at this stage, given the impact on the choices of statistical techniques and machine learning models.

To this end, the Jarque-Bera (JB) test has been employed to review the normality of distributions within the collected base dataset. This test will augment the visualisations in section 5.1, and descriptive statistics in figure 5.21, by calculating a composite statistic to quantify deviations from normality for all numeric variables within the dataset. The hypothesis for assessing normality of distributions with the JB test statistic is framed as under:

Hypothesis 0 (Null Hypothesis) *The dataset is normally distributed.*

Hypothesis 1 (Alternate Hypothesis) *The dataset is not normally distributed.*

Given the dataset has 632 data points in total, the threshold for rejecting normality would be a high JB value, e.g. 1,000 or higher, which would be strong evidence of severe non-normality, given preliminary indications of skewness and kurtosis as seen in figure 5.21. Additionally, the p-values for the JB test statistic will also be calculated for each numeric variable, to verify the significance of the calculated JB test statistics. A significant JB test statistic (i.e. $p < 0.05$) would confirm systemic non-normality in the dataset.

Given the preliminary indications of outliers in the dataset, the dataset was firstly scaled to standardise the range of values, to ensure that comparisons across features are meaningful, and limit the influence of measurement units that may differ from one numeric variable to another.

The first set of calculations (figure 5.22) indicates high JB test statistics for each numeric variable, with the overall statistics within the range of 2,000 to 140,000, indicating severe non-normality of distributions. The statistical significance of these calculations is further accentuated by the accompanying visualisation of p-values for each JB test statistic (figure 5.23), which shows the p-value close to 0, below the defined α (significance) level of 0.05, reinforcing the statistical significance of this test.

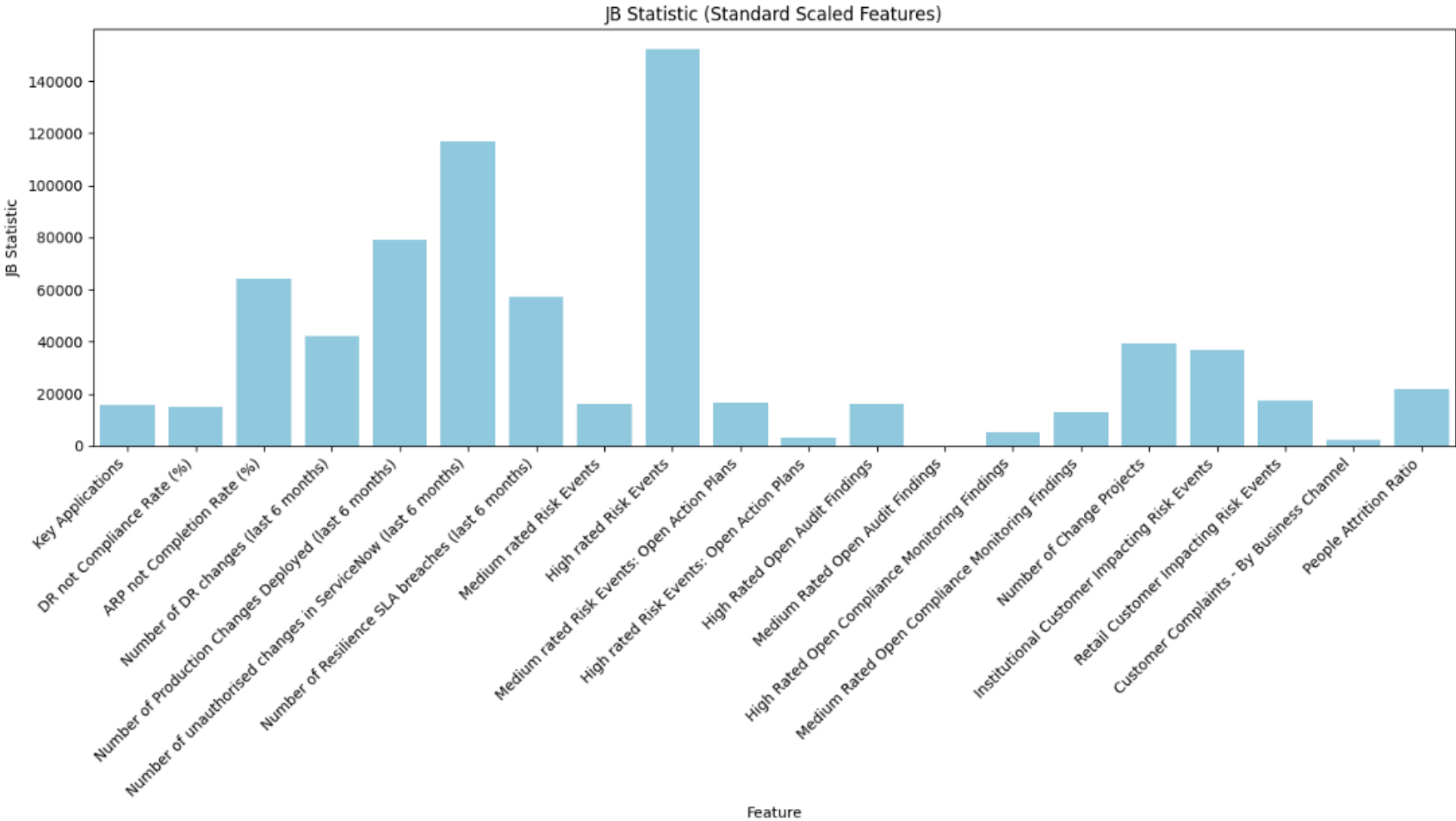


Figure 5.22: JB Test Statistics

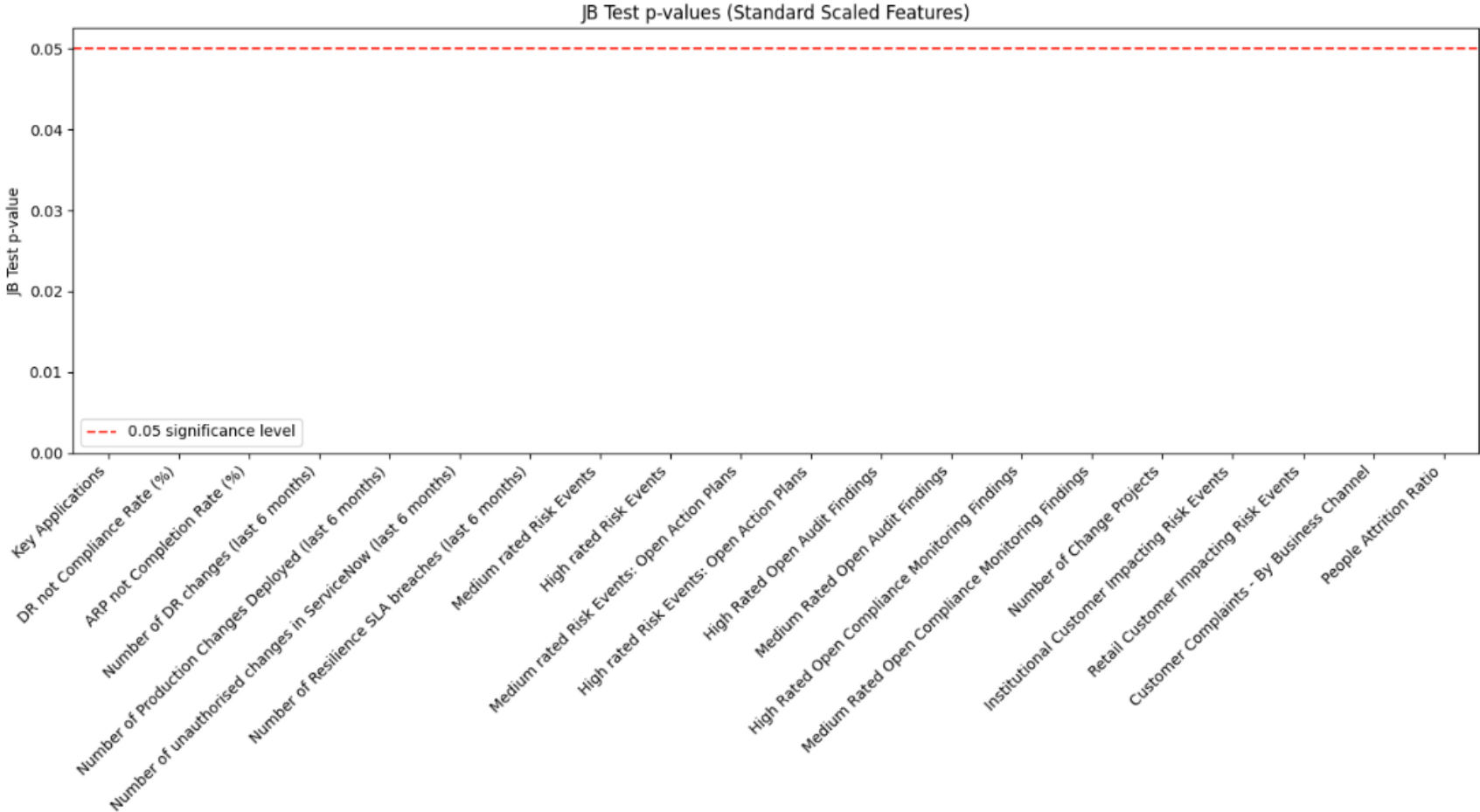


Figure 5.23: JB Test p-values

A visual inspection of the above graphs shows that the lowest JB test statistics are seen for two numeric variables: High Risk Events: Open Action Plans and Customer Complaint Volumes - By Business Channel. A spot inspection of the High Risk Events: Open Action Plans in the following figure reconfirms the high JB test statistic against a Probability Density Function (PDF), which represents the probability distribution of this variable under the assumption of normality. A visual inspection shows that the actual distribution is inconsistent with the PDF, which confirms the lack of normality in the distribution of this variable.

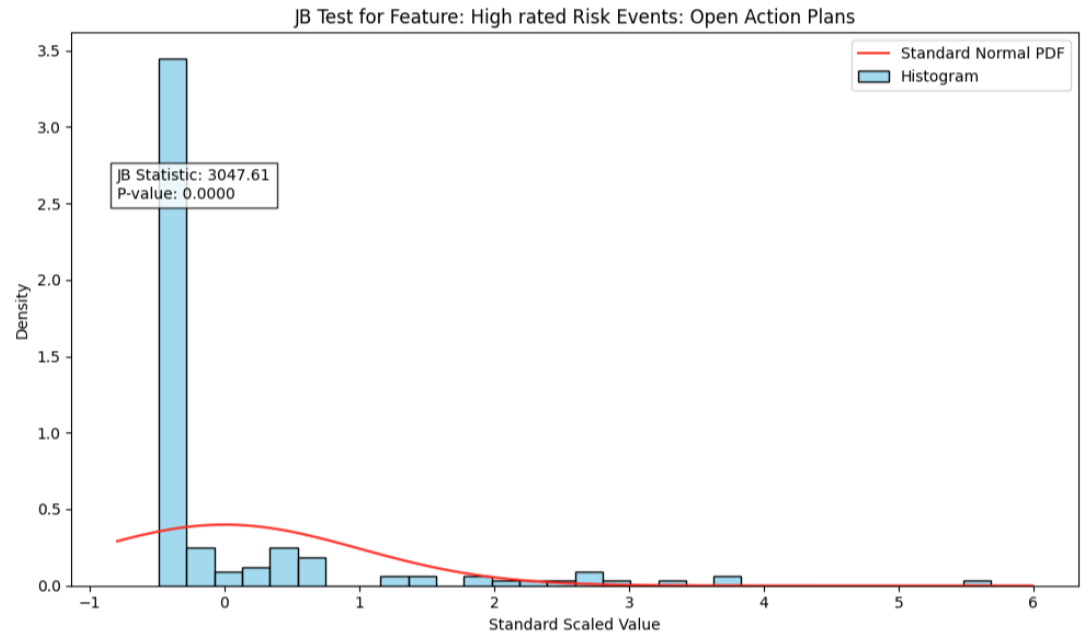


Figure 5.24: JB Test Statistic: High RE Actions

Based on the above results, the null hypothesis is rejected, with all numeric distributions confirmed as non-normal. This conclusion will be the basis of key statistical test choices as well as machine learning models that are produced as part of this thesis.

5.3 PRELIMINARY INSIGHTS

A targeted analysis of the dataset profile was now performed with a view to generating insights to be leveraged for the design of machine learning models. In doing so, observed relationships identified in the dataset were evaluated to assess whether they correspond to extant notions of risk assessment, such as an expected linear relationship between audit findings (denoting control failures) and a weak control environment (see Doyle et al. (2007), a seminal study which found a link between material control weaknesses, and complex

operations or restructuring programmes (a feature included in the data compiled in section 4.3).

The first step is an analysis of correlation amongst the numeric features collected in the data, performed with a view to identifying and assessing the extent of any observed multicollinearity. The resulting correlation heat map shown in figure 5.25 corroborated pre-existing business expectations of internal relationships within the dataset.

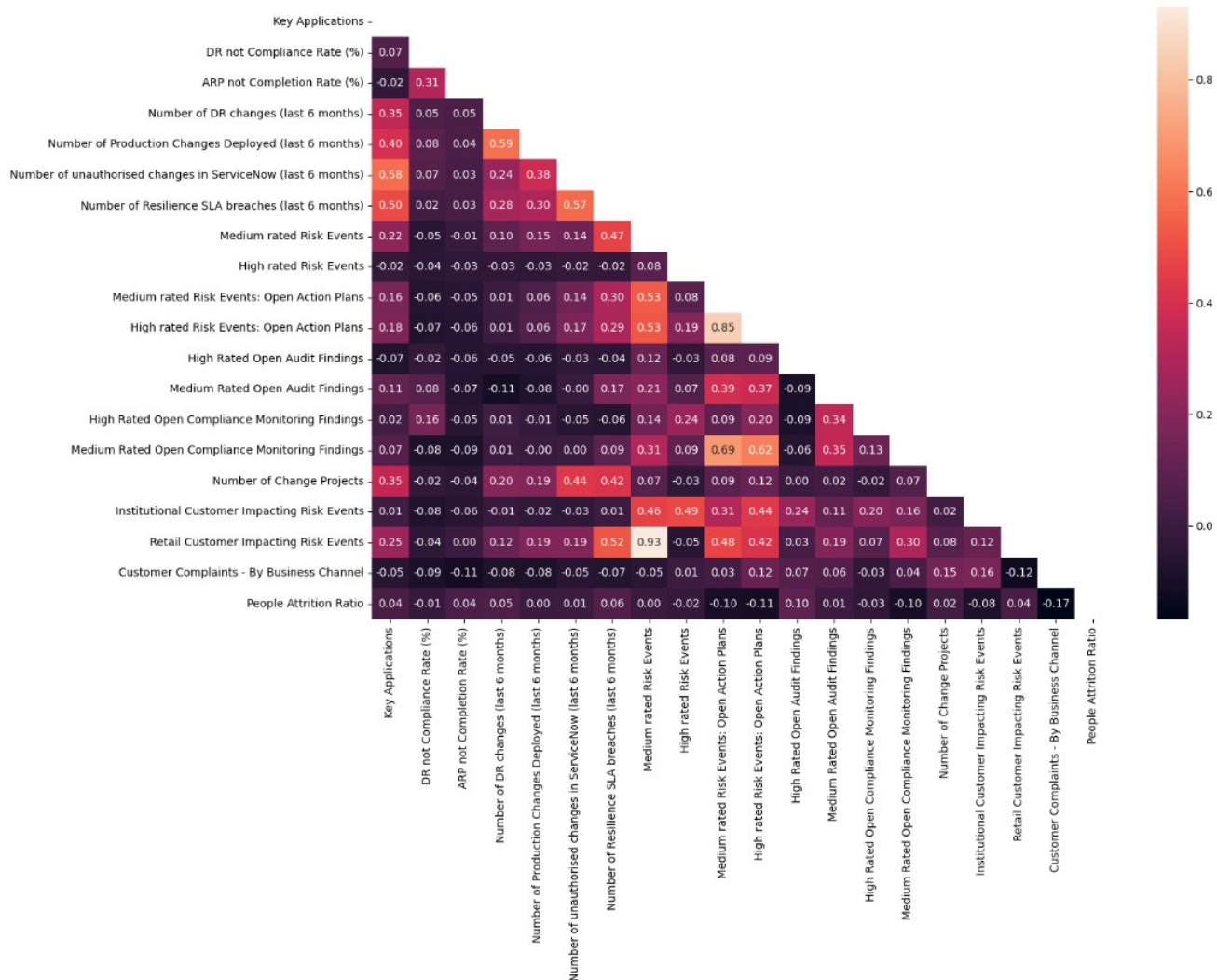


Figure 5.25: Correlation between numeric features collected

As an example, there is strong positive correlation (a score of 0.93) between the "Retail Customer Impacting Risk Events" and the overall population of risk events collected in the data, captured in the "Medium" or "High rated Risk Events" variables. Customer-impacting risk events are expected to be a subset of the overall population of risk events and as such, the removal of "customer-impacting risk events" will not result in information loss, given that all risk events are collected, regardless of whether or not they impact customers. Additionally, the subject nature of internal audit risk assessments means that individual

auditors and IAFs view the impact and relevance of individual risk events differently, given that customer-impacting risk events would only impact those parts of the organisation responsible for serving and supporting customers. Therefore, the appropriate treatment for the observed multicollinearity between the "Retail Customer Impacting Risk Events" and "Medium" or "High rated risk events" variables would be to remove the "customer-impacting risk events" variable from the dataset prior to model development. Whilst this instance of multicollinearity can be suitably treated through disposing of one feature without incurring information loss, other observations of multicollinearity may not be as clear cut. Feature independence has been assessed in this chapter; see section 5.4. Prior to removing any multicollinear features where a clear subset relationship is not established, the default position for model development purposes is to build two comparative models; one with all features included, and one with multicollinearity addressed. In doing so, the impact of any potential information loss from the removal of the multicollinear features on the model's performance would be identified. Detailed work around model development is discussed in section 6.3.

Individual relationships, if any, between the numeric features and the relevant target variables, are now examined to gain any exploratory insights from the datasets collected. Extending the findings of Doyle et al. (2007), correlation between the volume of audit findings and the weakness level of the overall control environment for a particular BU would be expected. Therefore, a more effective control environment would have relatively a lower average volume of audit findings, with a more adverse control environment reflecting a higher average volume of audit findings. This assertion was now evaluated, alongside other related expectations as noted below in figures 5.26 and 5.27, with targeted visualisations of the findings.

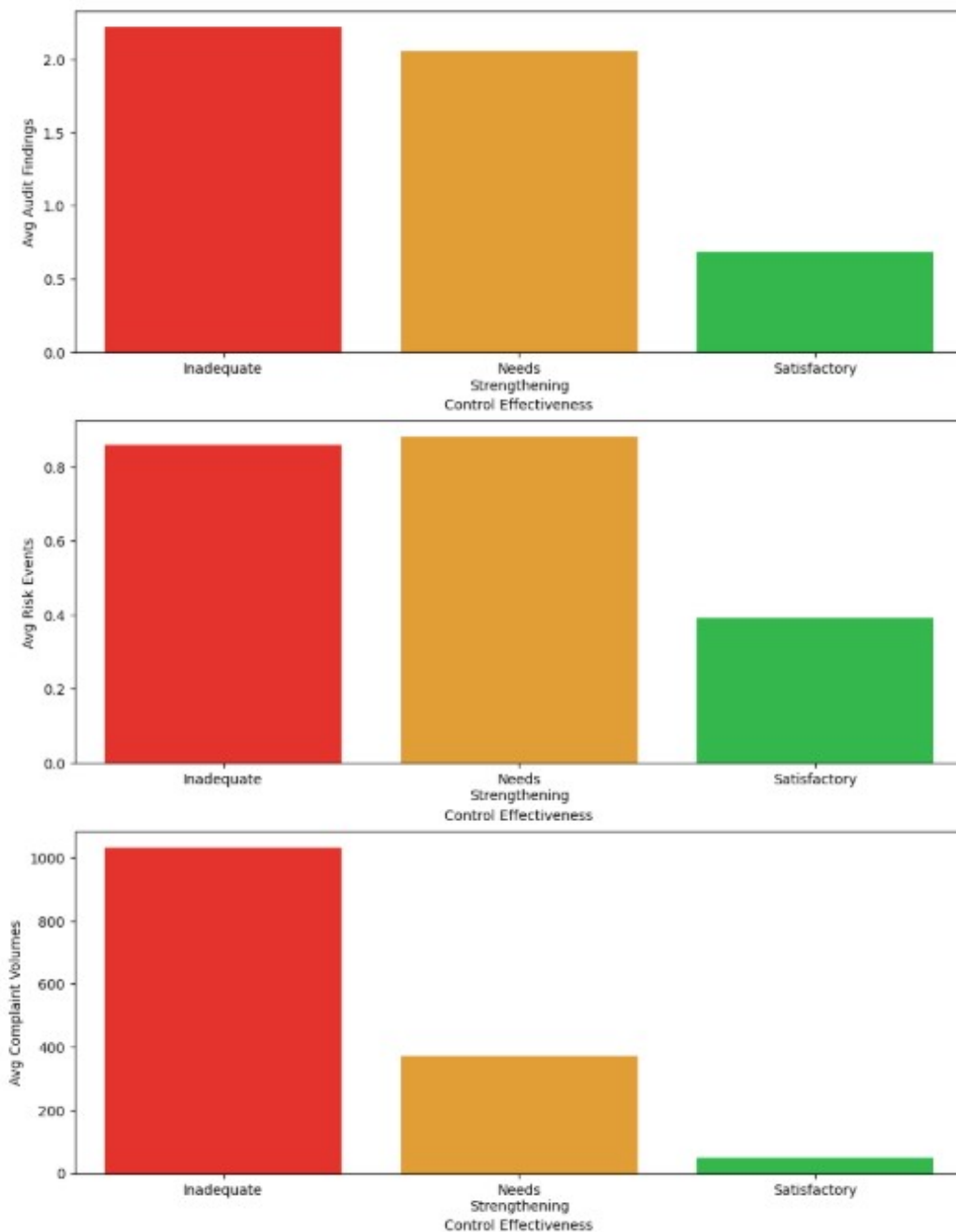


Figure 5.26: Numeric features and control effectiveness

The above figures bring out key observations as below:

1. The highest average volume of audit findings is associated with the most adverse control environment (i.e., an "Inadequate" assessment).
2. The average volumes of risk events (known control gaps) are highest for BUs with weaker control effectiveness ratings, i.e. Inadequate or Needs Strengthening labels.

3. The lowest customer complaint volumes are seen in BUs that have the strongest relative control environments.

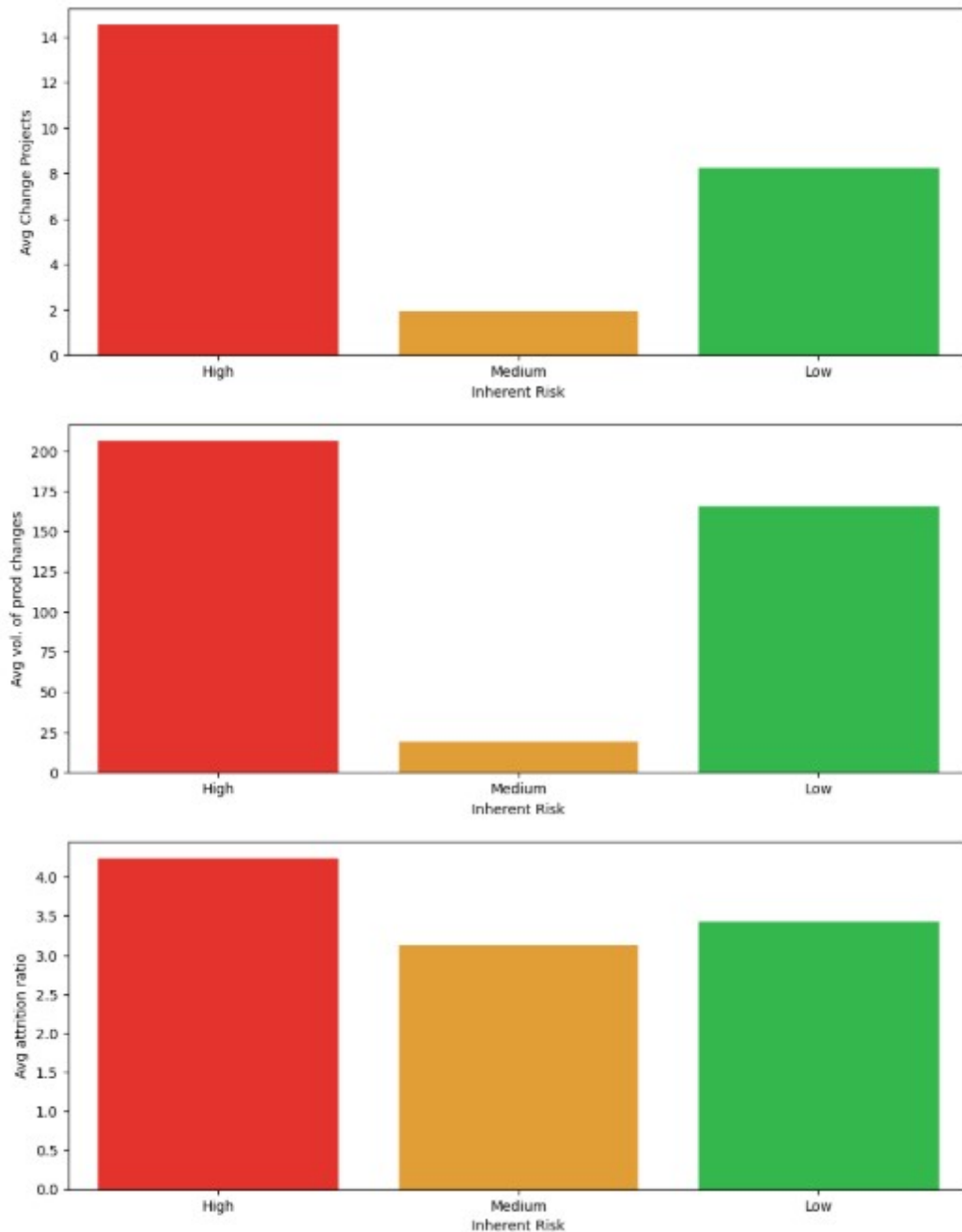


Figure 5.27: Numeric features and inherent risk

The following observations emerge from a review of figure 5.27:

1. High levels of organisational volatility (denoted by the number of change projects as well as production changes deployed to applications) are associated with a weaker control environment assessment, directly supporting the findings previously reported in Doyle et al. (2007).

2. Similarly, there is high people turnover in business areas that have weaker control environments, and lower attritions in areas with comparatively stable control environments.

Collectively, figures 5.26 and 5.27 demonstrate expected linear relationships between individual features and the target variables, and correspond to well-established approaches taken by industry practitioners for risk assessments. These insights are encouraging as they point to expected relationships between the features and target variables.

Similar insights using non-numeric variables are now evaluated. Firstly, the non-numeric variables collected in the dataset relate to one target variable, namely, Inherent Risk.

Targeted analysis of the data quality, performed as part of data collection, showed the following profile of missing observations for a few variables.

Volume of missing observations across dataset:		
Feature	Missing Values	Missing values as a % of population
RCSA - Inherent Risk Assessment	450	71.20
RCSA - Control Environment	452	71.52
Customer-facing division - Yes / No	8	1.27
Customer-impacting division - Yes / No	248	39.24
Organisational complexity observed in division?	248	39.24
Process delivery manually intensive	212	33.54
Formal project management team in place?	8	1.27
Important business service supported?	16	2.53
Regulatory impact on BU?	8	1.27
Does BU store or process confidential data?	68	10.76

Table 5.1: Volume of missing observations across dataset

No missing observations were noted for any of the other variables. Missing observations for the relevant variables were dealt with by adopting the following approach:

1. Five variables, namely, Risk and Control Self Assessment (RCSA) by Management (for both Inherent Risk and Control Effectiveness), the associated customer impact, organisational complexity of a particular business unit and the identification of manually intensive processes, were removed due to the lack of sufficient data points and the high proportion of missing data. This is because the significant lack of

observations, for the majority of the audit universe, is unlikely to be treated adequately or add meaningful insight to the research.

2. Divisions that are not populated as customer-facing divisions were assumed to be non customer-facing.
3. The "Project Management Team in place - Yes/ No" attribute is intended to track whether a project management team is in place to support complex change transformation programmes. No such teams were assumed to be in place for the eight missing records.
4. A similar approach was adopted for two other features, namely, "Important Business Service supported?" and "Regulatory impact on BU?" - in the absence of affirmative labels for these features, cases with missing observations for these features were assigned "No".
5. The majority of business areas within any organisation are expected to store and / or process data in some shape or form, so the BUs with missing values were also assumed to process / store confidential data.

Having addressed the missing data, similar visualisations were built to assess the relationships between the remaining non-numeric features and the Inherent Risk (target) variable. This approach is identical to the one used for examining the relationships between the various numeric features and the target variables and the results can be seen in figures 5.28 and 5.29.

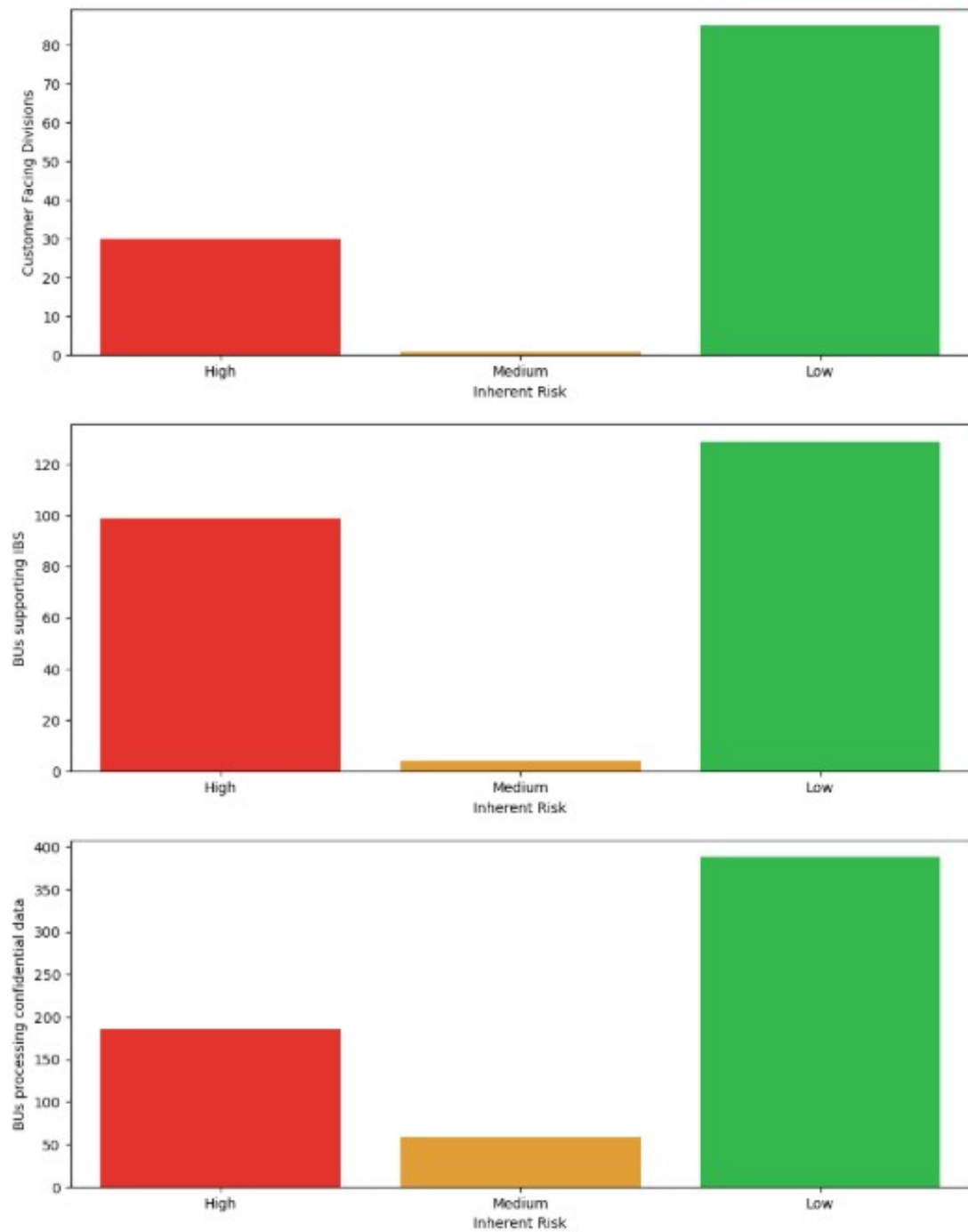


Figure 5.28: Non-numeric features and inherent risk - 1 of 2

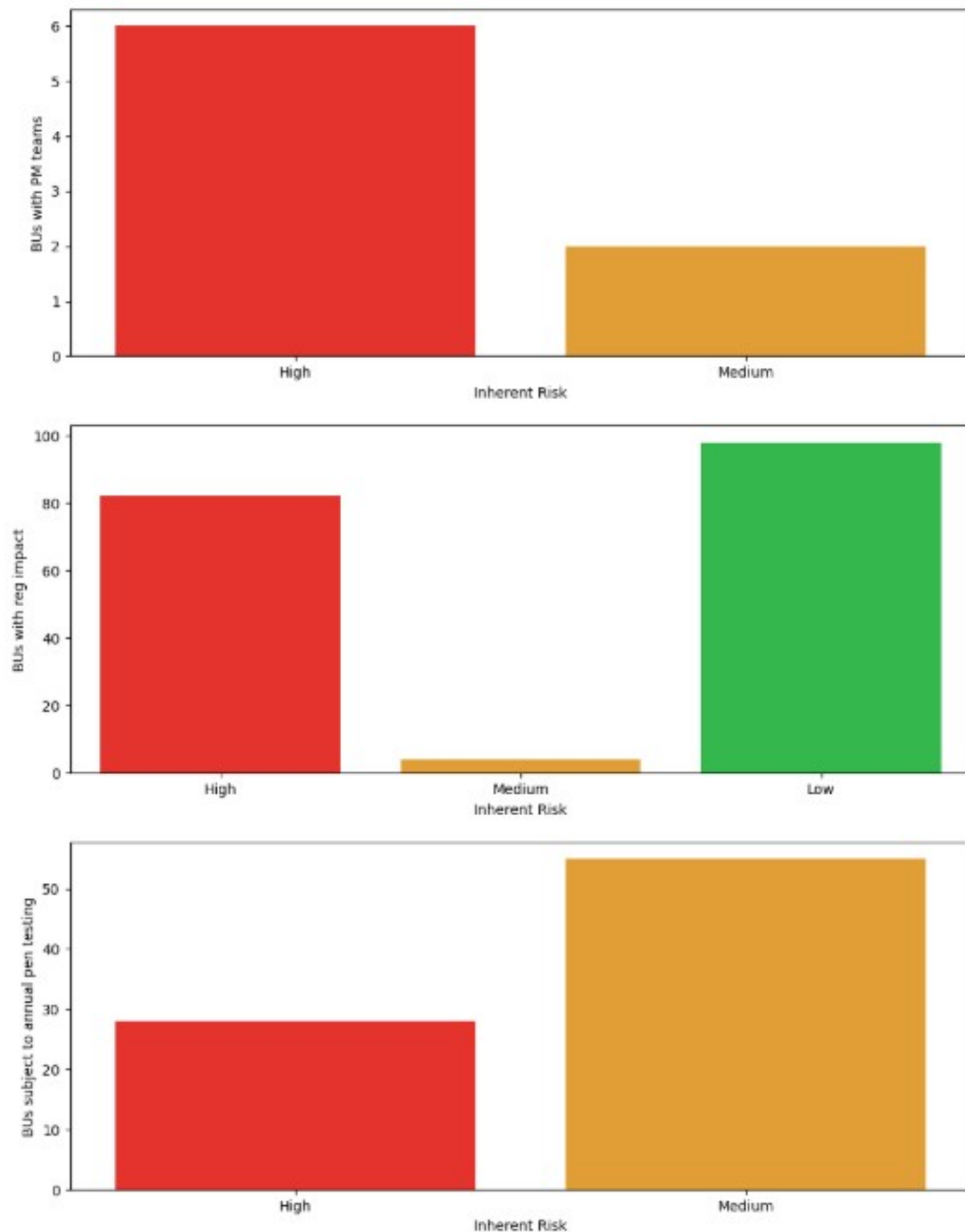


Figure 5.29: Non-numeric features and inherent risk - 2 of 2

Unlike the observations with numeric variables, there are surprising relationships, or the lack thereof, between the non-numeric features and the target variable. For example, the "customer-facing" aspect of a BU does not always correspond to the division being inherently high risk, despite the additional regulatory and operational pressures that are likely to affect such a BU. Similarly, BUs subject to annual penetration test exercises would be expected to be inherently high risk, given the reputational impact of a nefarious actor

succeeding in infecting the organisation's businesses. However, exploratory data analysis shows that this assumption is not always valid, as the majority of divisions with annual penetration test exercises are those which the IAF has historically assessed as being "Medium" rather than "High". The most surprising of all though, is the spread of BUs with a regulatory impact, where the majority of BUs facing off to regulatory inspections are inherently rated as "Low" rather than "High".

The above insights point to a general lack of consistency in the risk assessment exercise, and is reflective of the historically manual nature of the risk assessment process. Indeed, prior research by Demirović and Proho (2021) shows that risk assessment exercises have previously been conducted manually and rely on subjective expertise and auditor perceptions (Norman et al., 2010), rather than through a machine learning approach. Perhaps this is intentional, with such observations serving to lend "weight" or "biases" to an auditor's assessment of the prevalent risk profile of a particular business unit or audit entity. However, research by Hellström et al. (2020) has shown that the inclusion of bias in machine learning exercises is not always negative and that the use of a certain level of bias within the dataset might serve to support model training. There is also a generic requirement to drive the greater uptake of machine learning within the audit profession, with recent research by Law and Shen (2020) supporting the assertion that the use of advanced capabilities such as ML or AI are only set to grow in the future. On this basis, the non-numeric features will be retained within the dataset, and included for machine learning and modelling purposes, with a view to testing their relevance in machine learning outcomes.

5.4 INDEPENDENCE OF VARIABLES

The correlation matrix produced in section 5.3 highlights various relationships between the features and the target variables, namely, Inherent Risk and Control Effectiveness. This base analysis offers a starting point for framing the risk assessment exercise as a supervised machine learning approach, where the features are seen as the independent variables driving either IR or CE (the dependent variables in question). Whilst this is of course one of several possibilities being explored as part of the research, the pre-requisite to being able to setup a supervised machine learning model is the independence of variables collected. As such, a hypothesis is framed to test and establish the independence of the numeric variables collected in section 4.3 above, using Spearman's R correlation. The choice of statistical test is influenced by the following factors:

1. **Non-parametric distribution:** As evidenced in section 5.1, the datasets contain outliers and do not demonstrate normal distributions, as expected in a real-world dataset.
2. **Monotonic relationships:** Given the insights collected from section 5.3, it is clear no strong linear patterns exist amongst the variables, even if traces of multicollinearity are observed.
3. **Feature engineering:** Spearman's rank correlation can be used to identify highly correlated features and therefore corroborate the correlation matrix designed in section 5.3. In turn, this can support next steps in feature selection and engineering, leading to more robust models.

The following hypothesis is framed to test the independence of the numeric variables collected in section 4.3, using Spearman's R correlation.

Hypothesis 0 (Null Hypothesis) *There is no statistically significant correlation between any two variable-pairs within the dataset, suggesting that the variables are independent and any observed association in the sample is due to random variation.*

Hypothesis 1 (Alternate Hypothesis) *There is a statistically significant correlation between any two variable-pairs within the dataset, suggesting that the variables are not independent and the observation is not due to chance alone.*

For the purposes of hypothesis testing, the α , or significance level, is set to 5%. Spearman's R coefficient was calculated and compiled for each of the variable pairs, and visualised using the following heat map shown in figure 5.30.

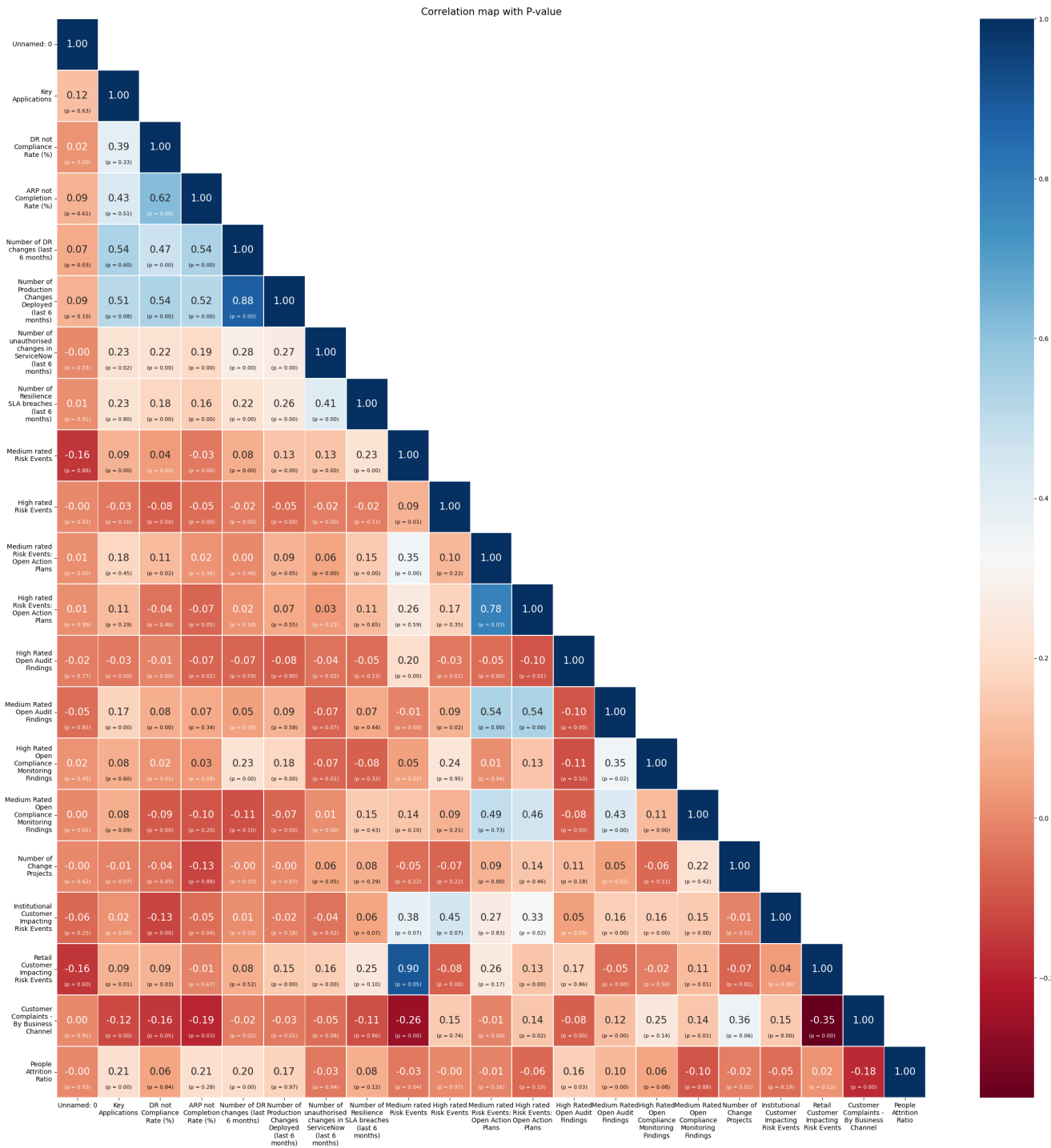


Figure 5.30: Spearman's R Correlation with ρ and α

The assessment of statistical significance is performed as a two-step process. The p-value for each variable-pair was examined across the dataset, ignoring the ρ value for correlation strength. This is because a p-value greater than the preset significance threshold of 0.05 would result in the null hypothesis being upheld, establishing the independence of the variables in question. Where the p-value was less than 0.05, the strength of the correlation

was evaluated with reference to the p value. For this second test, a threshold of $|0.49|$ was defined, with weaker correlation strengths (i.e., below this threshold) reasserting the independence of the two variables.

1. The pre-defined α threshold of 0.05 was met or breached for 88 out of 190 variable-pairs tested. In these cases, the null hypothesis is upheld, and confirms the independence of the relevant variable pairs. Examples of such cases include the correlation between the Number of Production Changes deployed across applications used by the BU in the last six months, and the DR non-compliance ratio, as well as the number of change programmes in-flight within the BU and the volume of Medium-rated risk events raised by the BU. These variables are also expected to remain independent from a business / IAF perspective.
2. A borderline case of correlation was observed with one relationship, namely, that of retail customer impacting risk events with Medium Rated risk events, where a high correlation strength (0.9) is observed, however, alongside a p -value score of 0.05. These variables are retained given the pre-defined test threshold, but nevertheless exhibit a strong relationship, given that risk events with an impact on customers are expected to be tracked through to resolution, justifying the predominant "medium"-level risk rating.
3. 96 variable-pairs showed a statistically significant ($p\text{-value} < 0.05$) result, but did not meet the expected threshold for correlation strength (set at $|0.49|$). As such, these variable combinations are also concluded as being independent, as the established correlation is not sufficiently strong to distort using them for classification modelling purposes.
4. The remaining six combinations showed a statistically significant relationship, and a strong correlation score in excess of $|0.49|$. These are reproduced below in table 5.2.

Statistically Significant Correlations - Independence Hypothesis Rejected		
Variable 1	Variable 2	Spearman's (ρ)
Number of Production Changes Deployed	Number of DR Changes	0.88
Medium-rated audit findings	Medium-rated Risk Event Action Plans	0.54
Medium-rated audit findings	High-rated Risk Event Action Plans	0.54
Number of Production Changes Deployed	ARP Non-Completion Ratio	0.52
Number of Production Changes Deployed	Key Applications	0.51
Number of DR Changes	ARP Non-Completion Ratio	0.54

Table 5.2: Statistically Significant Correlations - Independence Hypothesis Rejected

Upon closer inspection of these six combinations, the resulting multicollinearity is unsurprising. IAFs are required to raise audit findings to report the results of audit work, which would invariably require action plans to be produced in response, to enable appropriate levels of mitigation. Whilst action plans framed in response to risk events are not the same as those framed in response to audit observations, it is not unsurprising to see the collinearity between audit findings and risk-event action plans, especially when considering the audit findings in question to be the ones that are rated "medium". Equally, the strong relationship observed between the volumes of production changes deployed to applications used by BUs and the corresponding DR changes is expected, as DR changes would be a subset of production changes deployed on the applications landscape and indeed a higher number of applications used by a particular BU would require a correspondingly higher number of DR changes to test the robustness and availability of these applications.

Based on the above analysis, the following conclusions emerge:

1. The null hypothesis is upheld for all but six combinations of variable-pairs as listed above - highlighting that the variables are sufficiently independent so as to warrant inclusion for classification modelling purposes. The "Key Applications" and "Number of DR Changes (Last six months)" features were removed from the set of features used to model Inherent Risk, and "Medium rated Risk Events: Open Action Plans" and "High rated Risk Events: Open Action Plans" from the set of features used to model

Control Effectiveness, given their multicollinearity with other features as noted above in table 5.2.

2. As discussed in section 5.3, the impact of multicollinearity was assessed for the remaining six combinations by building a first round of ML models that exclude one of the multicollinear features, and a separate round that considers all features. This exercise is performed in section 6.3. In doing so, any information loss from the removal of the multicollinear features is considered and explained as part of the experiment.

Having tested the independence of the variables, model development is now performed in the next section, whereby these independent variables are used to generate risk assessment outcomes with a view to evaluating RQ2.

Part IV

EXPERIMENTS WITH MACHINE LEARNING

SUPERVISED MACHINE LEARNING

This chapter details how internal audit risk assessments can be performed using a supervised machine learning approach as an initial experiment to answer RQ2. The chapter begins by framing the risk assessment process as a supervised machine learning problem, with IR and CE framed as the target variables to be predicted using the collective set of features collected and explored in the preceding chapters. The chapter tests the influence of the features on the target variables and sets out key assumptions and limitations of working with the collected set of features. Thereafter, the chapter designs and evaluates a series of supervised machine learning models, where only influential features are used to generate prediction labels for the target variables. The chapter also demonstrates the impact of ineffective or noisy features on the overall efficacy of the model. Finally, the chapter concludes with an assessment of how the model's results are communicated to professionals in a practical setting, which serves as an input to answering RQ3.

6.1 FRAMING THE MODEL - WHERE DOES SUPERVISED ML COME IN?

The above research has led to the identification of a set of features that can be used to structurally approach the risk assessment exercise across IAFs. Given that IAFs are required to assess the residual risk of each organisational unit based on these features, the first round of experiment considers risk assessment as an exercise where these features serve as individual independent variables, with the residual risk being the desired target variable. However, given that the residual risk itself is an amalgam of two constituent analyses, namely, inherent risk and control effectiveness, it leads to the evaluation of not one, but two supervised machine learning attempts, with IR and CE acting as the target variables for a subset of the collected features.

Therefore, the first attempt to leverage machine learning for risk assessments is constructed as a supervised machine learning task with the following hypothesis, which supports the examination of RQ2:

Hypothesis: The risk assessment exercise is viewed as a **subjective assessment** (of the IR and CE scores), with the supporting features serving as objective variables that are hypothesized to influence the outcomes of periodic risk assessments.

Hypothesis 0 (Null Hypothesis) *Features used for risk assessments do not influence the risk assessment outcomes for IR and CE.*

Hypothesis 1 (Alternate Hypothesis) *Features used for risk assessments influence the risk assessment outcomes for IR and CE.*

Given that the hypothesis postulates two separate target variables, IR and CE, the hypothesis is tested twice, once for each target variable. Each independent variable is assigned to either IR or CE, as each variable is only relevant to one or the other outcome from the perspective of the auditor’s risk assessment exercise. The final risk assessment outcome of RR, will be calculated on the constituent underlying predictions of IR and CE generated by the classification model. A similar approach to evaluating fraud risk outcomes has been performed in recent research by Özbaltan (2024).

6.2 INDEPENDENT AND DEPENDENT VARIABLES

6.2.1 Feature Analysis for Inherent Risk

Relevant features were firstly mapped to the first target variable of interest, namely, IR. Prior research by Monroe et al. (2008) has attempted to compile an inventory of features relevant to IR given the impact on audit perceptions; however, there is limited evidence in literature of such an attempt being made in the context of an internal audit risk assessment exercise. Therefore, the practical approach adopted by the IAF is adopted for the purposes of research, and reproduced below in table 6.1.

Features used for Inherent Risk Prediction		
Feature Name	Feature Type	Rationale
Key Applications	Numeric	Where a BU utilises a high number of key applications, it would be typically reflective of the inherent risk level of that area rather than an assessment of underlying controls.
Number of Production Changes Deployed (last 6 months)	Numeric	Production change volumes for applications maintained or owned by a BU reflect the nature of operations carried out by these units and as such are reflective of risk levels.
Number of DR changes (last 6 months)	Numeric	Similar rationale to Prod Changes; the volume of change tickets for disaster recovery is a subset of the Prod Changes population.

Process Delivery Deemed Manual?	Non-numeric	This is an auditor's subjective assessment based on their understanding of the BU.
Number of Change Projects	Numeric	The change projects carried out within a BU would introduce volatility and potential uncertainty within the organisation.
Management's self-assessment of IR	Non-numeric	Management regularly self-assess their Inherent Risk environment and the output of this self-asst may be considered an input for the IAF's own risk assessment.
People Attrition Ratio	Numeric	The staff attrition rate is a reflection of the inherent risk of that BU, as a high staff turnover may adversely impact the organisation's ability to deliver operations at scale.
Customer Facing Area?	Non-numeric	This is an auditor's subjective assessment based on their understanding of the BU. A BU that faces off to or engages directly with customers would be considered higher risk to the organisation than one which does not.
Customer Impact Noted?	Non-numeric	This is an auditor's subjective assessment based on their understanding of the BU. A BU whose operations impact a customer's experience would be considered higher risk to the organisation than one which does not.
Complex Org Structure?	Non-numeric	This is an auditor's subjective assessment based on their understanding of the BU. Where an auditor considers the organisation structure of a BU to be inherently complex, this would be likely to introduce additional risk to the organisation.
Reg Impact Noted?	Non-numeric	Where a BU's operations are subject to regulatory scrutiny, the BU would be considered inherently higher risk than one which is not.
Confidential Data Processed by BU?	Non-numeric	The IAF reported that the handling of confidential datasets by a BU would result in it being considered higher risk than one where such data is not handled.

BU supports an Important Business Service?	Non-numeric	The IAF reported that the BU's support for an important business service (as defined by regulation) increases the inherent risk level of that area.
--	-------------	---

Table 6.1: List of IR features

Having mapped specific features against the target variable, a Chi-Square test is used to determine the influence of these features on the target variable (Inherent Risk), with the hypothesis framed as under:

Hypothesis 0 (Null Hypothesis) *There is no association between the set of independent variables and the dependent variable. In other words, the independent variables do not influence the Inherent Risk level assessed by the IAF.*

Hypothesis 1 (Alternate Hypothesis) *There is a statistically significant association between the set of independent variables and the Inherent Risk level.*

The decision to use a Chi-square test is justified by the nature of the dataset which constitutes a mixture of both numeric and non-numeric variables. Given that the Chi-square test, a non-parametric method, does not assume a specific statistical distribution, it is considered particularly advantageous in this context and suitable for dealing with both types of variables. The greater the discrepancy between the observed and expected frequencies, the stronger the evidence against the null hypothesis. These observations were overlaid with the correlation tests performed in section 5.4.

The significance level for the Chi-square test (α) was set as 0.05, and where the p-value from the test is less than this level, the null hypothesis of no association can be rejected, suggesting a significant association between the variables. In turn, this would render the Chi-square test a valuable tool for identifying whether the stated independent variables of various types are meaningful in predicting the category of the dependent variable. Features that do not display a statistically significant association with the dependent variable would be excluded at this step, enabling effective feature selection to support the robustness of the classification model.

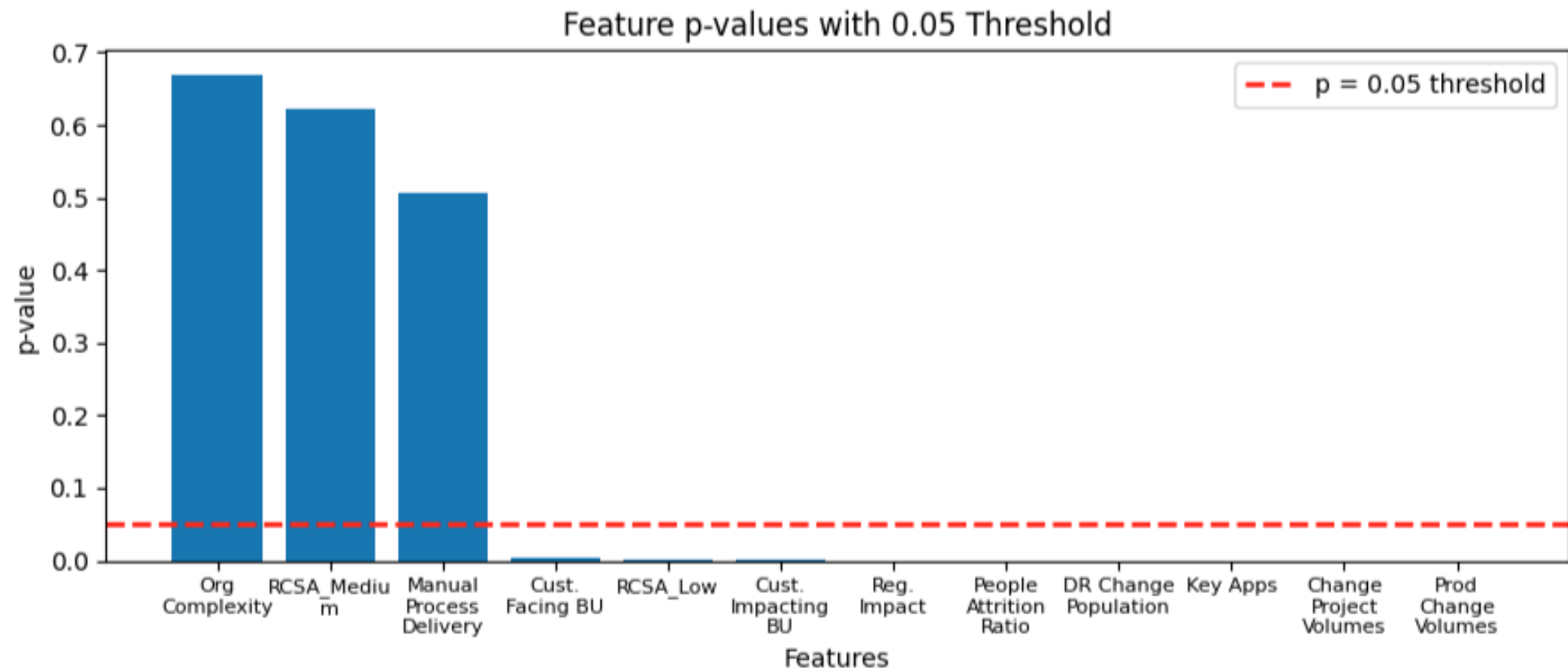


Figure 6.1: Chi Square p-value for Inherent Risk

The p-values for the Chi-square test for Inherent Risk variables are shown in figure 6.1. From the analysis, three features are shown to not have a statistically significant relationship with Inherent Risk, given $p\text{-value} > 0.05$. These variables are:

1. **Complex Org Structure?:** This variable is highly subjective and would vary from one auditor to the other within the IAF, as the definition of a "complex" organisation structure cannot be objectively defined. It is therefore not surprising that this variable does not exhibit a statistically significant association with the Inherent Risk assessment. Another explanation would be the lack of accurate labels assigned to various BUs by the IAF for this study, which would also negate the usefulness of this variable for machine learning purposes. This variable is therefore excluded from the model.
2. **Management's self-assessment of IR:** The lack of a statistically significant association between two separate risk assessments, one carried out by Management and the other by the IAF, is neither surprising nor unexpected, as these are meant to serve different objectives. The findings show that the IAF largely ignores alternate assessments of risk levels across the organisation, which is a fair approach to take, as it preserves the objectivity of the IAF's own risk assessment. As such, this variable is also excluded from the model for training purposes, as it does not add additional value to an independent risk assessment.
3. **Process Delivery Deemed Manual?:** The evaluation of this variable is very similar to the one for "Complex Org Structure" above, as a "manually immersive process" is hard to define objectively. As such, this variable is also excluded from the model.

Additionally, these three features were considered for removal in table 5.1, due to the lack of a sufficient volume of observations.

The remaining variables show a statistically significant association with the Inherent Risk variable, since the p-value for the Chi-square test is below the stated significance threshold of 0.05. The null hypothesis is therefore rejected in these cases and these remaining variables (with the exception of already excluded IR features at the end of section 5.4) are concluded as influencing the Inherent Risk, making them meaningful features for selection in machine learning. This approach was validated through consultation with the sponsoring organisation to ensure that there is no significant information loss as a result of using Chi-square test results to feature selection. Additionally, given the limited number of features for predicting IR labels, the model will be re-run incorporating all features to verify their impact on the model's overall efficacy and therefore, there is no adverse impact noted as a result of removing the statistically insignificant features at this stage.

6.2.2 Feature Analysis for Control Effectiveness

This exercise is now repeated for Control Effectiveness levels, the other target variable in the study. The relevant variables have been mapped to CE in table 6.2.

Variables used for Control Effectiveness Prediction		
Variable Name	Variable Type	Rationale
Project Management Team?	Non-numeric	The presence of a project management team for executing change programmes within the BU is reflective of the control environment's maturity, and is therefore chosen as a potential feature for analysis.
Risk Events	Numeric	The volume of risk events signifies actual control failures materialising, and is relevant to an auditor's assessment of control effectiveness levels.
Eligibility for Penetration Testing	Non-numeric	Penetration testing is carried out by specialist information security functions (separate to the IAF), and target BUs where internet-facing applications are used to support daily operations. The penetration test coverage outcomes are therefore linked to control effectiveness assessments.
Audit Findings (including observations raised by other assurance functions)	Numeric	Similar to risk events, these represent instances of actual control failures or improvement opportunities, influencing an auditor's assessment of the controls maturity within the BU in question.
Resilience Breaches	Numeric	The number of breaches in operational resilience service levels signifies control failures, as the BU is expected to be operationally robust, especially where it supports customers.
Management's self-assessment of CE	Non-numeric	Management regularly self-assess their Control Effectiveness environment and the output of this self-asst may be considered an input for the IAF's own risk assessment.

ARP non-completion ratio	Numeric	The lack of compliance with application recovery plans (ARP) would suggest a weaker control environment overall and therefore this variable is used to support control effectiveness assessments.
DR non-completion ratio	Numeric	Similar to resilience breaches and ARP non-completion, the lack of adequate disaster recovery test arrangements would suggest a weaker control environment overall and therefore this variable is used to support control effectiveness assessments.
Customer Complaint Volumes	Numeric	Whilst relevant only to customer-facing BUs, the volume of customer complaints would be generally relevant to control effectiveness outcomes, given the linkage of poor customer experience with the BU's inability to prevent or detect such instances.

Table 6.2: List of CE features

A Chi-Square test is used to test the influence of these variables on the target variable (Control Effectiveness), with the hypothesis framed as below. The rationale for using a Chi-square test is identical to the one stated above, i.e., for Inherent Risk levels, given that the dataset of independent variables is a mixture of variable types, with no assumption of normality in the distribution of numeric variables:

Hypothesis 0 (Null Hypothesis) *There is no association between the set of independent variables and the dependent variable. In other words, the independent variables do not influence the Control Effectiveness level assessed by the IAF.*

Hypothesis 1 (Alternate Hypothesis) *There is a statistically significant association between the set of independent variables and the Control Effectiveness level.*

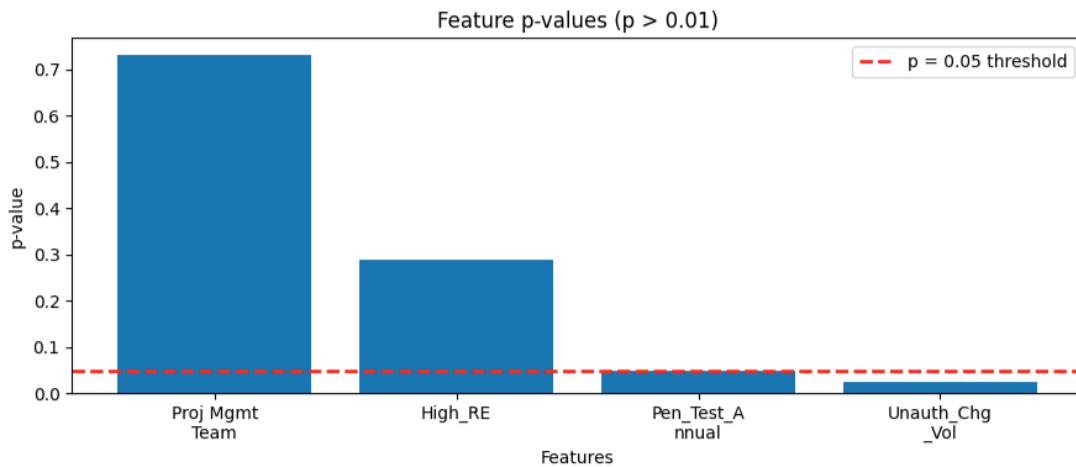


Figure 6.2: Chi-square p-values for Control Effectiveness

Figure 6.2 identifies the features hypothesized to have a statistically significant relationship with the target variable (control effectiveness), but with a Chi-square test p-value score greater than 0.01. Only a filtered selection of features are shown in figure 6.2 given the high number of features examined. Additionally, as the remaining figures have a p-value below the defined significance threshold of 0.05, they are already concluded to have a statistically significant relationship with the target variable.

From the analysis, two features are noted as not having a statistically significant relationship with the Control Effectiveness, given $p\text{-value} > 0.05$. These are "Project Management Team?" and "High-rated Risk Events". Given that the presence or indeed identification of a project management team is not always clear for individual BUs (as many parts of the organisation rely on centralised project management teams, rather than individual / bespoke ones), there is no significant value in continuing to use this feature for machine learning purposes. Similarly, the boxplots in section 5.3) show that the High-risk events are rare events, which is expected in an organisational context. Given the low information loss from the exclusion of these two variables, the final set of features for modelling the control effectiveness scores will exclude these two features. These two features will be removed in addition to the open action plans for Medium- and High-risk events, which were not found to be independent based on the work in section 5.4.

Whilst the Chi-square test has enabled an objective assessment of the independent variables for the purposes of feature selection, the above conclusions may not always align with IAF decisions on features that may require mandatory inclusion in a risk assessment model. For example, the established methodology for performing risk assessments within an IAF may require auditors to select customer complaint volumes for evaluating control effectiveness for customer-facing business units, regardless of whether the feature has any statistically significant relationship with the Control Effectiveness target variable. As such, whilst

excluded for the purposes of this experiment, a practical implementation of ML for risk assessments should always include an assessment of business considerations prior to excluding or including specific features, in order to align with audit methodology requirements. The Chi-square may be used as a supplementary aid to evaluate the robustness of included features, but should not be seen as an approach to exclude features that may appear statistically insignificant but are otherwise relevant for audit considerations. Additionally, a Chi-square based approach to feature selection assumes feature independence, an assertion tested using a Spearman's R test in section 5.4. Neither does a Chi-square test effectively capture non-linear relationships between the features and the target variables in this experiment, which is expected given the inherent subjectivity of the risk assessment exercise. However, these limitations aside, the Chi-square test does offer a meaningful starting point to identify noisy features within the dataset, and the results observed in this section also correspond with the IAF's pre-established expectations in a practical audit setting.

6.2.3 *Hypothesis evaluation*

The analysis in the preceding sections has established a statistically significant association between the mapped independent variables and the target variables. On this basis, the null hypothesis is rejected and it is concluded that the selected features do have a statistically significant influence on the risk assessment outcomes, and that the historic selection of features by auditors is not random selection.

The findings of the preceding sub-sections underscore the importance of careful selection and validation of features for the risk assessment exercise. However, it is acknowledged that the list of features is subject to constant change and updates, given an organisation's evolution over time. As a practical implementation guidance, it is proposed that organisations consider the application of similar statistical tests to assess the relevance and influence of proposed new features on the relevant target variables in future.

The final (tuned) models were also rerun with the removed features incorporated, to assess the efficacy of these statistical tests and examine the resulting impact on model accuracy and precision.

6.3 MODEL DEVELOPMENT AND EVALUATION

6.3.1 *Data pre-processing and hyperparameters*

Having established the connection between the independent and dependent variables, classification models were developed to assess the possibility of leveraging machine learning capabilities to perform risk assessment, and assess value-add from doing so.

Given that the models would be developed on the collected datasets, it is also essential to demonstrate the efficacy of any developed models on an identical dataset and evaluate the reliability of model predictions. To this end, it is essential to set aside a "test" dataset prior to initiating any modelling. This would enable pre-processing, scaling and model training on the remaining portion of the dataset and avoid a scenario where the model adapts to the patterns identified within the dataset, but performs poorly when tested against a real-world dataset, commonly referred to as "overfitting". Prior research by Joseph (2022), Joseph and Vakayil (2022), Xu and Goodacre (2018) and Gholamy et al. (2018) has not only emphasised the importance of removing a test dataset before initiating model development but research such as Gholamy et al. (2018) and Joseph (2022) has also evaluated the optimal split ratio between training and test datasets, with a split of 80:20 in favour of training data seen as an appropriate approach to segregating a real-world dataset for testing model performance. Additionally, research by Joseph and Vakayil (2022) also shows the importance of further segmentation of the training dataset into training and "validation" datasets, in order to enable random subsampling of individual data points within the training dataset and consequently, strengthen model robustness.

Based on the above, 20% of the dataset was removed as a "test" cohort to assess the efficacy of the final models identified upon the completion of the experiment. The remaining 80% of the dataset (resulting in $n = 506$) has been subject to preprocessing, and further segmentation into training and validation sets, to support the modelling exercise.

Research shows the extensive use of open source or low-code approaches to model design, such as the python library - PyCaret - which has been successfully used in similar empirical research work involving model design and development (Rye et al., 2022; Kobayashi and Yoshida, 2022; Ladds and G Harcourt, 2017). As such, this approach was adopted modelling, ensuring the selection and tuning of hyperparameters is performed through ten-fold cross validation to maximise the model's performance. The analysis of specific hyperparameters used for the final, tuned models (each of which were executed for 10 epochs) is documented in this chapter.

The experiment was executed in two phases - firstly, for Inherent Risk and secondly, for Control Effectiveness - given the difference in features used to assess these two target variables.

6.3.2 Inherent Risk: Hyperparameters

The initial view of the target class population (in this case for IR) showed an imbalance in the dataset as shown below in table 6.3.

Spread of BUs by Inherent Risk		
Ref.	Inherent Risk Label	Count of BUs
1	High	140
2	Medium	313
3	Low	53

Table 6.3: Imbalanced data for IR

Furthermore, upon splitting the full population of 506 entities across training and validation datasets, the training dataset is reduced by 20% to 404 entities, accentuating the adverse impact of the data imbalance. Prior research by Thabtah et al. (2020) and Luque et al. (2019) has shown that imbalance can adversely impact the accuracy and reliability of a classification model. Additionally, multiple techniques could be used to treat data imbalance, including:

1. **Oversampling:** Increasing the number of instances from the minority class.
2. **Undersampling:** Reducing the number of instances from the majority class.
3. **Hybrid Approaches:** Combining oversampling and undersampling techniques.
4. **Cost-Sensitive Learning:** Assigning different costs to mis-classifications of different classes.

Given the relatively small volume of training data ($n = 404$) within the dataset, an oversampling approach was adopted, leveraging the synthetic minority oversampling technique (SMOTE), to produce an additional set of data using the majority class as a reference point for one (resulting in 250 observations across all classes, or 750 observations in total). These two datasets were used to assess the impact of data balancing on tuned classifier performance. This approach draws on learnings from prior research work, such as Thabtah et al. (2020) and Jeong et al. (2018), who have experimented with multiple techniques to assess the impact of data imbalance on classification models. As such, there

are two datasets following imbalance treatment, with populations tabled as below in table 6.4:

Dataset size pre and post-imbalance			
Ref.	Inherent Risk Label	train (original)	train sm (majority oversampling)
1	High	112	250
2	Medium	250	250
3	Low	42	250

Table 6.4: Training dataset labels post imbalance treatment

Given the data imbalance observed, the final models developed as part of this experiment round will be assessed on not only the accuracy of the classifiers, but also class-specific precision. This is because the true test of an ML capability expected to be used within a risk assessment exercise at an IAF is the ability of the model to successfully highlight "High risk" entities that require timely IAF intervention, and to ensure numeric recalibration of the model to reflect changes in features and organisational volatility. As seen in table 6.4 above, the Medium risk class is the dominant category of inherent risk within the dataset and as such accuracy will not be a genuine indicator of model efficacy. Whilst the correction of data imbalance carried out above would lessen the impact of imbalance on accuracy ratios, the usage of precision to assess the accuracy of class-specific model predictions would offer a more nuanced approach to model evaluation. To this end, this experiment will provide accuracy, precision and recall metrics to enable holistic model evaluation, and the choice of final model will be influenced not only by the overall accuracy, but also by the ability of the model to specifically highlight the "High risk" entities within the dataset, which will be identified through an assessment of relative precision ratios for each classification model.

Having addressed data imbalance and adopted a starting position for final model evaluation, a three-fold approach was adopted for model development: firstly, data was normalised for all features, as prior research by Cabello-Solorzano et al. (2023) and Singh and Singh (2020) has shown that data normalisation is a key pre-processing stage for machine learning models used for data classification. Multicollinear features identified in section 5.4 were then removed, for both IR and CE target variables. Secondly, multiple classification models were developed, employing cross validation techniques to prevent overfitting as much as possible. Thirdly, 10 epochs for each developed model were executed to examine the variance in classifier performance and obtain comfort of stable model

accuracy for reporting. This iterative process was repeated twice, for each dataset as captured in table 6.4.

The rationale for producing multiple models is to study the impact of complex interactions between different attributes, and model non-linear relationships that are often present in audit risk assessments. Furthermore, any final machine learning models should not require assumptions about the distribution of variables as in other statistical techniques and should be robust to outliers and skewed data.

The following assumptions were made as part of modelling:

1. The training-validation split was 80% and 20% respectively, given the limited size of the dataset, with the populations for each cohort tabled above in table 6.4.
2. A fairly consistent approach to risk assessment has been undertaken by the IAF in the previous iterations, acknowledging the potential impact of bias.
3. The primary metric for model robustness is model accuracy, defined as the ability of the model to arrive at the same "label" for a BU, for Inherent Risk and Control Effectiveness predictions.

Using the above approach, eight models were constructed below, in order to draw out as many patterns among the independent and dependent variables as possible. Model selection was not biased towards a single approach, and instead used a combination of bagging, boosting, and ensemble methods to identify the best-performing classifier. For each model, a combined grid-search of all relevant hyperparameters was performed to arrive at the tuned values eventually used for model training, validation and testing, drawing on prior literature where appropriate and relevant. The individual tuned hyperparameters and grid search ranges used for each are discussed in table 6.5 below:

Hyperparameters and Tuning Approach				
Ref.	Model	Hyperparameters	Description	Tuning Approach
1	Logistic Regression	Regularisation Strength (C)	Controls the inverse of regularisation strength. Smaller values specify stronger regularisation, but drive higher bias. Conversely, a higher value of C would reduce the regularisation impact, increasing variance but lowering bias at the same time.	Cross-validation (10 folds) was used to assess the impact of the algorithm's performance on multiple values of C, on a parameter grid range between 1 and 51, and found C=1 as the optimal hyperparameter value in driving the highest model accuracy across training and validation datasets, considering computing efficiency and model performance.
		Solver	The solver hyperparameter specifies the algorithm used to optimise the logistic regression objective function. Different solvers have different strengths and are suitable for different types of datasets, with some suitable for binary classification and others for multiclass problems such as ours.	Given that the target variable has multiple classes, the choice was narrowed to either 'lbfgs' or 'newton-cg' as these are both quasi-Newton methods considered suitable for datasets with a large number of features. Prior research by Bhatia et al. (2021) has shown that the 'newton-cg' hyperparameter yields a marginally better model accuracy ratio, with minimal additional compute time. Both hyperparameters were used to evaluate model accuracy and noted a slightly higher validation dataset accuracy ratio with lbfgs, which was the final choice for this hyperparameter.

		Penalty	Norm used in the penalisation ('l1', 'l2', 'elasticnet'), which controls the type of regularisation applied to the model. L1 refers to Lasso and L2 to Ridge regularisation, where the squared value of the coefficients is added as a penalty term to the loss function.	At this stage of the experiment, all features used in the dataset are considered relevant to analysis, except the ones discounted following the statistical analysis performed in subsections 6.2.1 and 6.2.2. For this reason, L2 is the chosen penalty term for the model.
2	k-Nearest Neighbours	Neighbours (k)	The primary hyperparameter for the kNN algorithm, k refers to the number of neighbours to consider for classification. A smaller value of k may result in overfitting and a higher value may result in underfitted models due to the algorithm's inability to effectively generalise the data by considering a large number of neighbours	A grid-search underpinned by cross-validation was used to select the optimal value of k, by specifying a range of 1 to 51 - this was influenced by the need to assess a range of values, whilst being mindful of computing time. This is because a very high value of k, such as 500, with 10-fold cross validation, resulted in inefficient code execution. The execution of this grid search showed that the optimal value of k for the experiment is 5, which was re-verified through multiple (10) epochs of code execution.

		Distance metric	This hyperparameter is used to determine how the distance between data points is calculated. Possible values are 'Euclidean', which is a straight-line distance between the two datapoints, 'Manhattan', the sum of absolute differences between the coordinates of the points or 'Minkowski', a generalisation of both Euclidean and Manhattan distances).	Similar to the above approach used to select k, 10-fold cross-validation was used to arrive at the optimal value of this hyperparameter. Cross-validation testing showed that the 'Minkowski' distance metric was most suitable for the dataset, which is also supported by the fact that a mixture of non-numeric and numeric variables is compiled for risk assessments.
		Weights	The weights hyperparameter determines how the influence of each neighbour is weighted when making a prediction. Neighbours could be accorded equal weight ('Uniform') or weighted by the inverse of their distance, resulting in closer neighbours having a greater impact on the prediction ('Distance').	The experiment showed that the 'Uniform' approach is the most suitable value of the Weights hyperparameter, which is also supported by the fact that audit universe entities of a similar type (such as High-risk entities) will naturally vary in characteristics compared to a different type.

3	Support Vector Machines	Kernel type	The kernel type enables the transformation of data into a higher dimensional space, where it can be easier to separate into classes using a hyperplane. This hyperparameter has various values depending on whether the data is linear, non-linear or a combination of the two.	Prior research by Drewnik and Pasternak-Winiarski (2017) has shown that the radial basis function (RBF) is the most common choice of kernel type in a support vector machine, particularly where non-linear datasets, such as ours, are used for classification. This was corroborated using a grid-search to identify the most optimal kernel type between RBF, polynomial, linear and sigmoid kernels, and 10 epochs of grid-search (10-fold cross validation) showed that the RBF was most suited for the dataset.
		Regularization parameter (C)	Controls the trade-off between achieving a low error on the training data and minimizing the norm of the weights, similar to Logistic Regression.	An identical approach was used to the one used for logistic regression, but examined a manual range of values for C - 0.001, 0.01, 0.1, 1, 10 and 100 - with cross-validation, to arrive at the most suitable value of this hyperparameter. The choice of values was influenced by the desire to strike an appropriate balance between high and low values of C, whilst maintaining reasonable computing efficiency for multiple epochs of code execution. This experiment showed that the optimal value of C was 100.

		Gamma (for RBF kernel)	This hyperparameter defines how far the influence of a single training example reaches, with a low Gamma value offering a smoother decision boundary and a higher value driving a more complex boundary that captures potential intricacies in underlying features but introducing potential overfitting.	Grid-search using the same parameter grid as shown above for C identified an optimal value of 1 for this hyperparameter, which was used for the model.
4	Decision Trees	Maximum depth	This hyperparameter controls how deep a tree can grow, which influences the model's complexity. A shallow tree (e.g. a value of 3 for this hyperparameter) would typically have fewer splits in the decision tree, resulting in a simpler model and promote better explainability of the decision-making approach.	Consistent with the approach to the other algorithms, cross-validation was used to select the optimal values of the decision tree, controlling the range as a distribution of values from 1 to 50. This approach identified that the default value of "None" for this hyperparameter was the most suitable, as this would enable the model to expand the nodes until all leaves contained less than the minimum samples split feature.
		Minimum samples split	Minimum number of samples that a node must have, before being split further. For example, if the value of hyperparameter is set to 10, the nodes must have at least 10 samples to be split. For this reason, a higher value for this hyperparameter can reduce overfitting.	A grid search between the values of 2 and 20 identified an optimal value of 2 for the optimal model, which was leveraged for the experiment.

		Criterion	This hyperparameter is a function used to measure the quality of a split. Default values used for this hyperparameter include 'gini', which measures the impurity of a node, and 'entropy', which refers to the information gain from a split.	Research by Aning and Przybyła-Kasperek (2022) has cross-examined the 'entropy' and 'twoing' criteria, which represents a less commonly used technique for splitting nodes. The 'twoing' criterion aims to find the best split between groups by considering how well the split separates the dataset into two groups. Cross validation using a grid-search, tested across four possible values of this hyperparameter, namely, gini, entropy, log_loss and twoingm, identified 'gini' as the most suitable criterion for modelling purposes, which was used for the experiment.
5	Random Forest	Number of trees	Number of trees in the random forest classifier that are built and averaged to produce the final prediction.	This hyperparameter can support model performance by reducing variance and improving generalisation. Using 10-fold cross validation for a parameter grid search range of 1 - 101, an optimal value of 100 was identified for this hyperparameter.
		Maximum depth	Max depth of individual trees, same hyperparameter as in decision tree classifiers.	An identical approach was used for selecting the maximum depth for this algorithm as discussed above for decision trees.
		Minimum samples split	Minimum number of samples required to split an internal node, same hyperparameter as in decision tree classifiers.	An identical approach was used for selecting the minimum samples split for this algorithm as discussed above for decision trees, which identified an optimal value of 2 for this hyperparameter.

6	Gradient Boosting Classifier	Number of boosting stages (n_estimators)	This hyperparameter determines how many weak learners are to be sequentially added to the model, with each stage aiming to correct the errors of the previous stage.	A range of values - 50, 100, 200, 500 - were tested to find the optimal number of stages, using a cross validation approach. This was performed to identify the optimal value of n_estimators, where a higher number of stages could increase the risk of overfitting, and a lower number could lead to underfitting. Hyperparameter tuning identified an optimal value of n_estimators as 100.
		Learning rate	This hyperparameter controls the contribution of each weak learner to the final model. A lower value for this hyperparameter can help the model develop a better understanding and generalisation of the dataset, reducing overfitting.	The learning rate was calculated by assessing a range of lower-end learning rates, from 0.001 to 0.1 - as the experiment's objective is to identify the best performing model, regardless of practical efficiency at this stage. This approach identified a hyperparameter value of 0.1.
		Maximum depth	Max depth of individual trees, same hyperparameter as in decision tree classifiers.	An identical approach was used to select the maximum depth for this algorithm as discussed above for decision trees, which identified an optimal value of 3 for this hyperparameter.
		Subsample	The fraction of samples used for fitting the individual base learners. This hyperparameter introduces randomness into the model, preventing overfitting.	Cross validation based grid searching across a range of 0.5, 0.7, 0.9 and 1.0 identified an optimal value of 1.0 for the subsample, effectively resulting in all training samples being used for the model.

		Minimum samples split	Minimum number of samples required to split an internal node, same hyperparameter as in decision tree classifiers.	An identical approach was used to select the minimum samples split for this algorithm as discussed above for decision trees, which identified an optimal value of 2 for this hyperparameter.
7	XGBoost	Number of boosting rounds	Number of boosting rounds to be added to each model, similar to the gradient boosting classifier.	An independent grid search approach with cross validation, with a specified parameter search range of 50 - 500 identified an optimal value of 200 for this hyperparameter.
		Learning rate	Shrinks the contribution of each tree by this factor, controlling the contribution of each tree to the final model.	As with gradient boosting, a range of low value learning rates (0.001, 0.01, 0.1, 0.2) were evaluated to arrive at the best value for this hyperparameter, with the grid search exercise identifying an optimal value of 0.2.
		Maximum depth	Max depth of individual trees, same hyperparameter as in decision tree classifiers.	An identical approach was used to select the maximum depth for this algorithm as discussed above for decision trees, resulting in an optimal value of 10 for this hyperparameter.
		Subsample	The fraction of samples used for fitting the individual base learners. This hyperparameter introduces randomness into the model, preventing overfitting.	Cross validation based grid searching for a range of 0.5 to 1.0 identified an optimal value of 0.7 for the subsample.

		Colsample_bytree	The fraction of features used for fitting the individual base learners. This hyperparameter introduces randomness into the model, preventing overfitting.	As a lower value of this hyperparameter can increase bias (whilst reducing overfitting), and vice-versa, a cross validation based approach for a range of 0.5 to 1.0 to identify the optimal hyperparameter value showed this to be 0.9.
8	LightGBM	Number of boosting rounds	Number of boosting rounds to be added to each model, similar to the gradient boosting classifier.	A 10-fold cross validation on a parameter grid range of 50 - 300 estimators identified an optimal value of 100 for this hyperparameter.
		Learning rate	Shrinks the contribution of each tree by this factor, controlling the contribution of each tree to the final model.	As with gradient boosting, a range of low value learning rates were examined to identify the best value for this hyperparameter, with grid search identifying an optimal value of 0.1.
		Maximum depth	Max depth of individual trees, same hyperparameter as in decision tree classifiers.	An identical approach (10-fold cross validation) was used to select the maximum depth for this algorithm as discussed above for decision trees (with a range of values between 1 and 10), which showed that the default value of -1 was the most suitable for this hyperparameter.
		Num_leaves	This hyperparameter controls the maximum number of leaves in one tree, determining the tree's complexity.	Whilst more leaves may capture complex patterns within the data, a higher value for this hyperparameter could result in overfitting. Therefore, a cross validation grid search range between 20 and 100 was defined for this hyperparameter, which identified the optimal value to be 31.

		Subsample	The fraction of samples used for fitting the individual base learners. This hyperparameter introduces randomness into the model, preventing overfitting.	Cross validation based grid search on a range of values between 0.5 to 1.0 identified an optimal value of 1.0 for the subsample.
		Colsample_bytree	The fraction of features used for fitting the individual base learners. This hyperparameter introduces randomness into the model, preventing overfitting.	As a lower value of this hyperparameter can increase bias (whilst reducing overfitting), and vice-versa, a grid search on a range of values between 0.5 to 1.0 identified the optimal hyperparameter value as 1.0, effectively retaining all features for model training.

Table 6.5: Hyperparameter Tuning Approach

The hyperparameters were tuned based on the balanced dataset, with the final tuned models applied on the validation dataset. The train and validation scores for these tuned models are reported below in columns one and three respectively.

As a separate exercise, the models were also tuned on the original, imbalanced dataset to assess the impact of data imbalance on model development, with the results of the tuned models on the imbalanced train dataset reported in column two of table 6.6.

6.3.3 Inherent Risk: Modelling results

The approach using the standard classification approach produced the following results. The results shown in 6.6 have an average variance of $\pm 1\%$, following ten iterations of code execution. Algorithms with comparable best performance within this variance are shown in bold.

Classifier Accuracy				
Ref.	Classifier	X train (balanced)	X train (imbalanced)	Validation dataset (balanced)
1	Logistic Regression	43.07%	63.61%	42.16%
2	K-nearest neighbours	76.24%	63.86%	77.45 %

3	Support vector machines	35.89%	63.61%	35.29%
4	Decision tree classifier	75.99%	62.87%	72.55%
5	Random forest classifier	76.24%	62.87%	71.57%
6	XGBoost classifier	75%	62.87%	75.49%
7	Gradient boosting classifier	71.29%	71.29%	69.61%
8	LightGBM classifier	72.52%	66.09%	72.55%

Table 6.6: Classifier Accuracy for Inherent Risk

The results shown in 6.6 demonstrate an interesting spread of observations across the three datasets. Firstly, the adverse impact of data imbalance on the performance of various classifiers is evident, with all eight models improving their accuracy on the validation dataset as it moved from imbalanced to balanced data.

High classifier model performance with k-Nearest Neighbours (KNN) is primarily attributable to the small dataset used for modelling, as well as the impact of data normalisation on the small dataset as part of the preprocessing stage.

High training accuracy is also seen with ensemble methods such as random forest classifier, which was the best performing model of all when considering the computational efficiency of all models. As such, a confusion matrix for random forest was produced to further evaluate the model's precision and recall across the prediction classes on the validation dataset. The model was shown to be most effective in returning "Medium" rated business units, which is unsurprising, given that this risk label represents the majority class of the dataset. Encouragingly, the model also did well with "High" rated business units, correctly identifying 20 High-rated business units as originally risk assessed by the auditor.

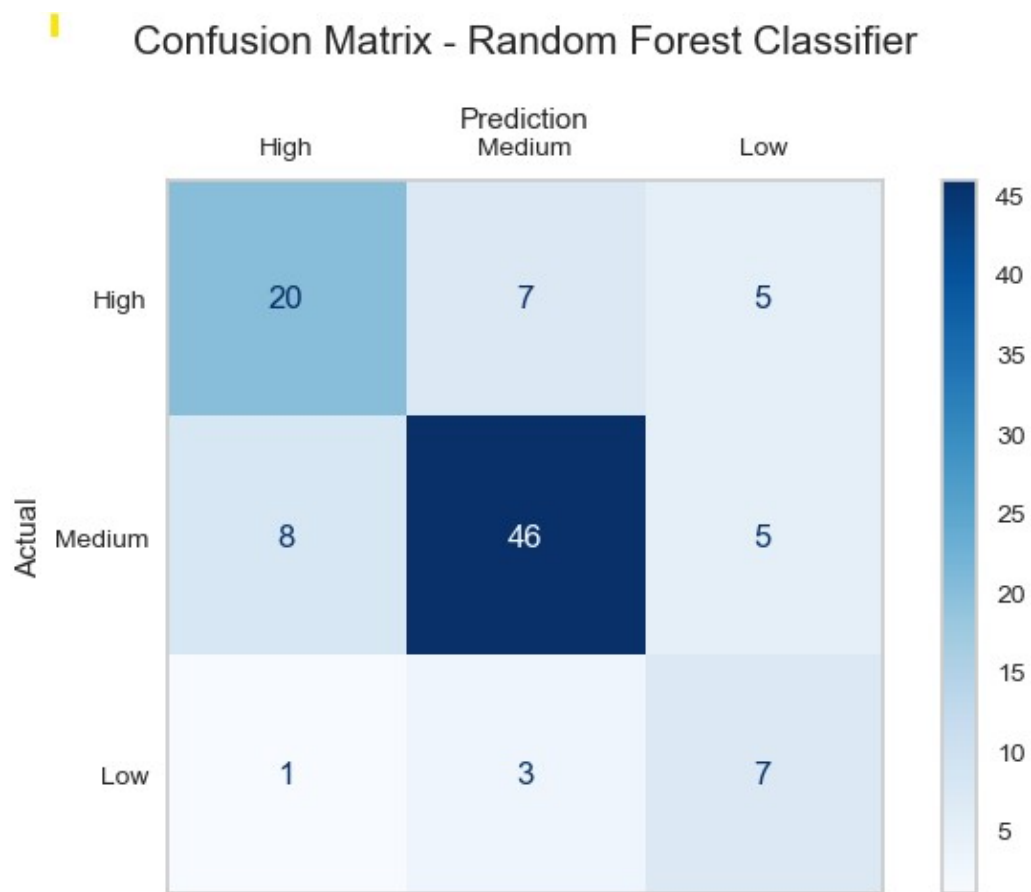


Figure 6.3: Random Forest Classifier: Confusion Matrix (IR Validation)

The individual precision and recall scores (reported as a weighted average across all three classes) for each of the models produced as part of this stage of the experiment are reported below. Similar to the work on measuring accuracy, these results have an average variance of $\pm 1\%$, following 10 epochs of code execution for each dataset. These scores are reported as before, for the tuned models on the training and validation dataset in columns one and three, with the training performance of the tuned models on the imbalanced dataset reported in column two. Algorithms with comparable best performance within this variance are shown in bold.

Classifier Precision				
Ref.	Classifier	X train (balanced)	X train (imbalanced)	Validation dataset (balanced)
1	Logistic Regression	59.23%	55.08%	54.56%
2	K-nearest neighbours	78.69%	62.24%	78.36%

3	Support vector machines	60.92%	66.56%	49.24%
4	Decision tree classifier	80.36%	39.53%	75.6%
5	Random forest classifier	79.51%	39.53%	73.59%
6	XGBoost classifier	78.05%	39.53%	77.79%
7	Gradient boosting classifier	73.87%	72.58%	72.43%
8	LightGBM classifier	75.64%	64.96%	74.43%

Table 6.7: Classifier Precision for Inherent Risk

Classifier Recall				
Ref.	Classifier	X train (balanced)	X train (imbalanced)	Validation dataset (balanced)
1	Logistic Regression	43.07%	63.61%	42.16%
2	K-nearest neighbours	76.24%	63.86%	77.45%
3	Support vector machines	35.89%	63.61%	35.29%
4	Decision tree classifier	75.99%	62.87%	72.55%
5	Random forest classifier	76.24%	62.87%	71.57%
6	XGBoost classifier	75%	62.87%	75.49%
7	Gradient boosting classifier	71.29%	71.29%	69.61%
8	LightGBM classifier	72.52%	66.09%	72.55%

Table 6.8: Classifier Recall for Inherent Risk

These results show the importance of curating the features for model design, and removing noise from the dataset prior to model development. Furthermore, the use of bagging and ensemble techniques has also helped allay fears of auditor bias impacting the assessment results, as the use of techniques like Gradient Boosting or Random Forest would involve a sample of individual features being combined to produce the final set of predictions. Thirdly, the adverse impact of data imbalance on model performance is evident, with each of the key metrics - accuracy, precision and recall - falling significantly when the models are developed on imbalanced training datasets. Overall, the reported accuracy scores are encouraging as they offer robust challenge to subjective risk assessment outcomes

previously reported by the IAF, with the best performing models having reasonably generalised the training dataset based on the relationships between the features and target variable.

6.3.4 Inherent Risk: Model performance on test dataset

Having completed model design, each of the final tuned models were tested using the final set of hyperparameters on the test dataset that was set aside at the start of this experiment. The results have been summarised in individual confusion matrices for each model, and the results are consistent with those observed at training / validation stage. Specifically, the application of tuned models on the test dataset shows that the best performing model is the random forest classifier (see figure 6.7), which outperforms others not only because of high accuracy (71%), but has a high precision ratio of 88% for the "High" risk entities identified by the IAF. These entities are of particular interest to IAFs, given that such entities are subject to audit as part of the IAF's audit plan ((Wang et al., 2023; Eulerich and Eulerich, 2020)). The kNN classifier continues to be the best performing model from the standpoint of classifier accuracy, but has a relatively lower precision of 74% for the "High" risk entities within the dataset, rendering the random forest classifier a better choice. Table 6.9 below summarises the accuracy of each classifier on the test dataset in column two.

Classifier Accuracy: Test Dataset				
Ref.	Classifier	X train (balanced)	Test Dataset	Validation dataset (balanced)
1	Logistic Regression	43.07%	41.27%	42.16%
2	K-nearest neighbours	76.24%	72.22%	77.45 %
3	Support vector machines	35.89%	35.71%	35.29%
4	Decision tree classifier	75.99%	72.22%	72.55%
5	Random forest classifier	76.24%	71.43%	71.57%
6	XGBoost classifier	75%	69.04%	75.49%
7	Gradient boosting classifier	71.29%	65.87%	69.61%
8	LightGBM classifier	72.52%	69.04%	72.55%

Table 6.9: Classifier Accuracy for Inherent Risk: Test Dataset

In the "weaker" models, a relatively high accuracy ratio is seen in the identification of "Medium" risk entities. The confusion matrices produced for Logistic Regression and Support Vector Machines - bring forth this pattern clearly, with the former accurately identifying only 16 "Medium"-risk audit universe entities, and the latter similarly identifying only 10 "Medium"-risk audit universe entities as labelled by the auditor. This outcome renders these two models being rejected for the experiment, as the Medium risk class is the majority class of the dataset, and even weaker models are expected to generalise the underlying patterns within the dataset and identify the correct label for this class. Therefore, a model's ability to correctly understand the complexities of either "extreme" class, i.e. High or Low-risk entities, would increase its robustness by supporting the auditor in making efficient, data driven risk assessment decisions. This is most pronounced with the random forest classifier, which as shown in figure 6.7, which not only has the highest overall accuracy, but has also accurately predicted the highest number of both "High" and "Medium" risk entities, relative to all other models.

The subjectivity of risk assessments (Redmill, 2002; Aven, 2016) means that it is possible for auditors to be flexible with their final assessments, or even override them. As such, the better performing models in the experiment - decision trees, random forest and the three boosting algorithms - all show useful results, with the misclassified entities typically only one level up or down, with extreme errors, such as the four "High" risk entities identified by the IAF being misclassified as "Low" inherent risk, not being likely to undermine auditor faith in a machine learning based risk assessment exercise. Extreme differences (where the difference between predicted and actual labels are two levels apart) are the lowest with the kNN classifier, which increase this model's overall reliability.

From the following confusion matrices, ensemble models are seen to perform better than others. This supports the assertion that the datasets have complex underlying non-linear relationships and patterns, and also reinforces the need for auditor consistency in completing future risk assessments. These results also underscore the importance of hyperparameter tuning, which in the case of the best performing model - random forest - has effectively handled the bias variance trade off to minimise overfitting as seen from the model's ability to largely identify the two audit universe categories of interest (High and Medium risk entities).

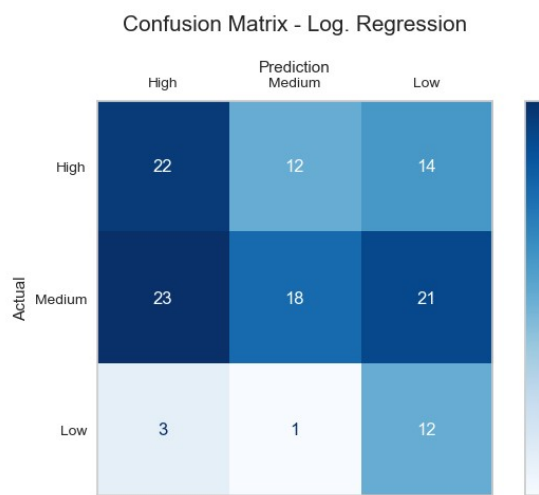


Figure 6.4: IR Confusion Matrix - Log. Regression

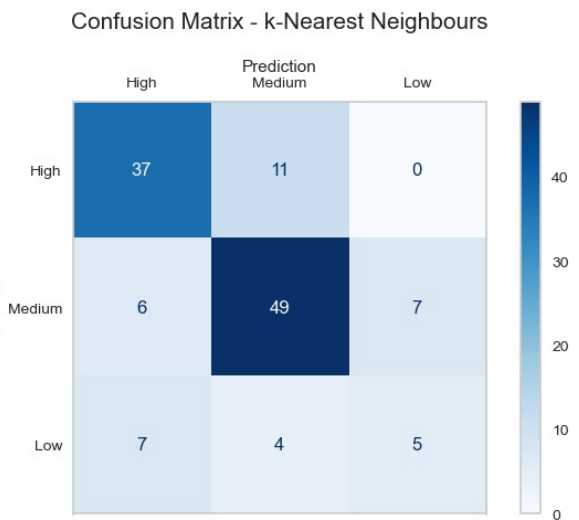


Figure 6.5: IR Confusion Matrix - KNN
Confusion Matrix - Random Forest Classifier

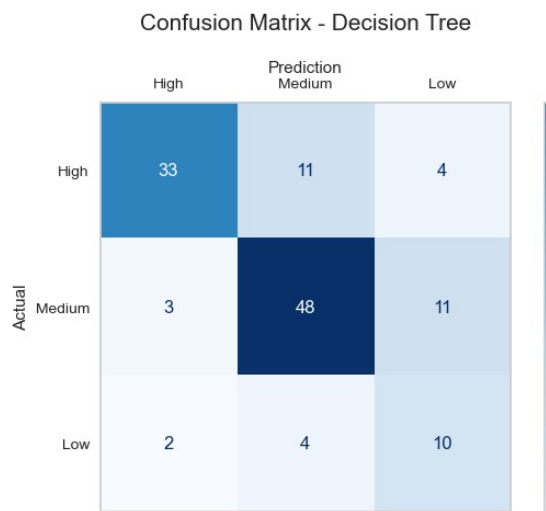


Figure 6.6: IR Confusion Matrix - Decision Tree

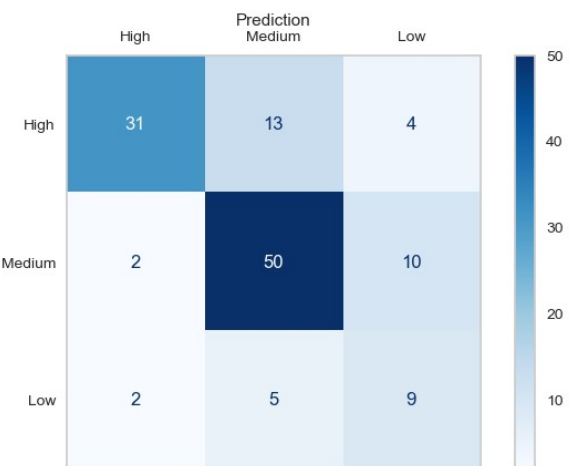


Figure 6.7: IR Confusion Matrix - Random Forest

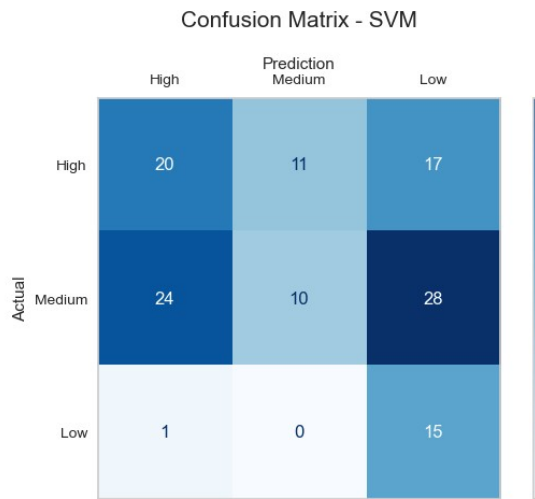


Figure 6.8: IR Confusion Matrix - Support Vector Machines

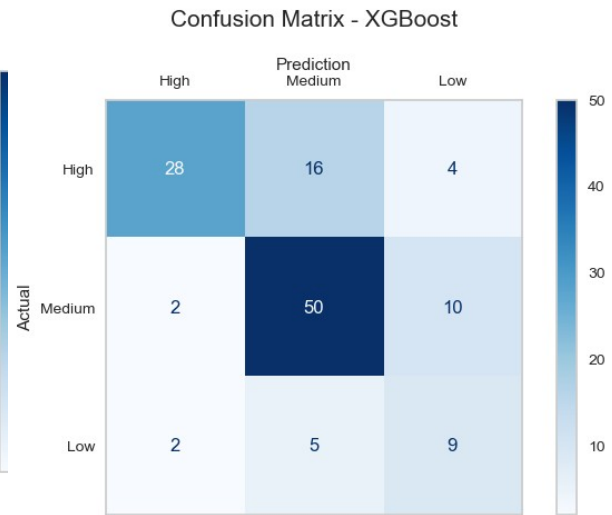


Figure 6.9: IR Confusion Matrix - XGBoost

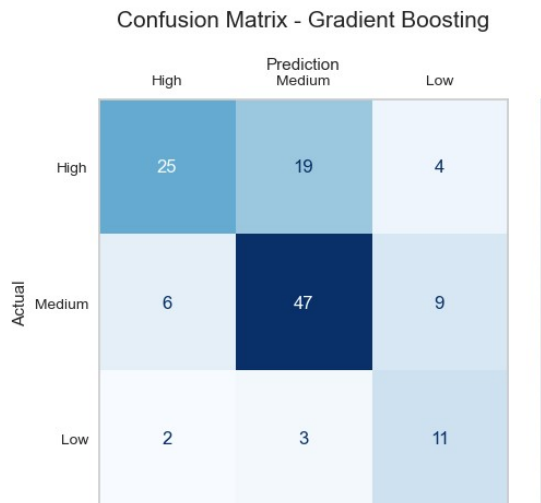


Figure 6.10: IR Confusion Matrix - Gradient Boosting Machines

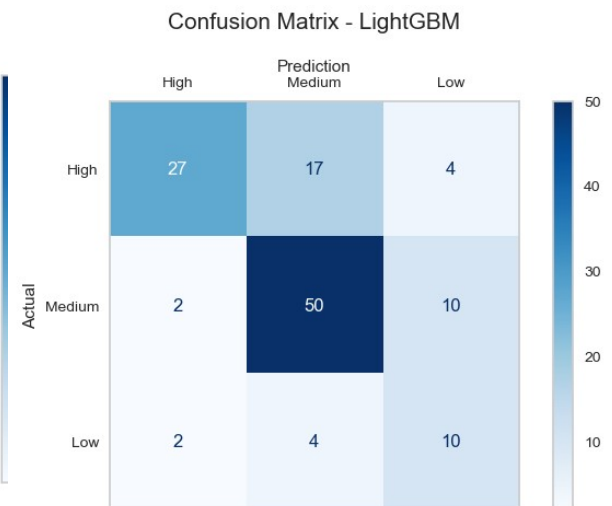


Figure 6.11: IR Confusion Matrix - LightGBM

Despite the added advantage of ensemble methods, the use of boosting techniques (XGBoost, Gradient Boosting and LightGBM being the three boosting algorithms tested in the experiment) has identified comparatively fewer High risk entities compared to the random forest or kNN classifiers. Further, the computational complexity generally seen with boosting algorithms (Mayr et al., 2014) also means that the real-world implementation of such algorithms on an internal audit dataset might be practically challenging, especially when the requirement to tune hyperparameters across a large training dataset is taken into context.

Based on the reported results and confusion matrices, the random forest classifier is concluded as the best performing algorithm, given the high accuracy and class-specific precision scores on the test dataset. This experiment is now reperformed for the other target variable of interest - Control Effectiveness - in the next section.

6.3.5 Control Effectiveness: Hyperparameters

The initial view of the target class population (in this case for CE) showed an imbalance in the dataset as shown below in table 6.10.

Spread of BUs by Control Effectiveness		
Ref.	Ctrl Effectiveness Label	Count of BUs
1	Satisfactory	155
2	Needs Strengthening	283
3	Inadequate	68

Table 6.10: Imbalanced data for CE

The approach to dealing with imbalanced data is repeated as with the previous target variable, and the overall approach to ML model design for CE is identical to IR. As such, there are two datasets following imbalance treatment, with populations tabled as below in table 6.11:

Dataset size pre and post-imbalance			
Ref.	Ctrl Effectiveness Label	train (original)	train sm (majority oversampling)
1	High	124	226
2	Medium	226	226
3	Low	54	226

Table 6.11: Training dataset labels post imbalance treatment

Using the same approach as with IR, the same eight models were constructed as before, in order to draw out as many patterns among the independent and dependent variables as possible. Model selection was not biased towards a single approach, and instead used a combination of bagging, boosting, and ensemble methods to identify the best-performing classifier. The approach to tuning the optimal hyperparameters for each algorithm is discussed in table 6.5, with each of the eight models tuned with the most optimal

hyperparameters, developed on the balanced, training dataset to arrive at the reported results. Similar to IR, the final tuned models were eventually applied on the validation dataset. The train and validation scores for these tuned models are reported below in table 6.12 columns one and three respectively, with a separate training result on imbalanced data reported in column two.

6.3.6 Control Effectiveness: Modelling results

The use of the standard classification approach produced the following results. The results shown below have an average variance of $\pm 1\%$, following ten iterations of code execution:

Classifier Accuracy				
Ref.	Classifier	X train (balanced)	X train (imbalanced)	Validation dataset (balanced)
1	Logistic Regression	66.34%	68.56%	58.82%
2	K-nearest neighbours	78.71%	68.07%	78.43%
3	Support vector machines	66.58%	65.1%	62.75%
4	Decision tree classifier	76.49%	72.77%	75.49%
5	Random forest classifier	76.49%	68.32%	75.49%
6	XGBoost classifier	72.52%	78.71%	65.69%
7	Gradient boosting classifier	74.75%	78.96%	66.67%
8	LightGBM classifier	74.5%	76.98%	62.75%

Table 6.12: Classifier Accuracy for Control Effectiveness

The above results show an interesting departure from IR. Firstly, whilst data imbalance continues to adversely impact the performance of the classifiers on the training dataset, the magnitude of impact is not as pronounced as with IR, given that some classifiers such as logistic regression and XGBoost seem to have better results with imbalanced data than with balanced data. Secondly, simpler classifiers, such as the logistic regression model, performed better for control effectiveness than they did for Inherent Risk, where the ensemble models were required to return the most reliable outcomes. This is possibly attributable to the nature of features used for CE modelling, which comprise exclusively of numeric features unlike IR where a mixture of numeric and non-numeric features are used

to perform risk assessments. Finally, the kNN and random forest classifiers continue to perform strongly on the validation dataset.

As before, a confusion matrix was produced for the random forest classifier to further evaluate the model's precision and recall across the prediction classes on the validation dataset. The model was most effective in returning "Needs Strengthening" rated business units, which is unsurprising, given that this risk label represents the majority class of the dataset. Encouragingly, the model also did well with both outer classes, correctly identifying 24 "Satisfactory"-rated business units as originally assessed by the auditor.

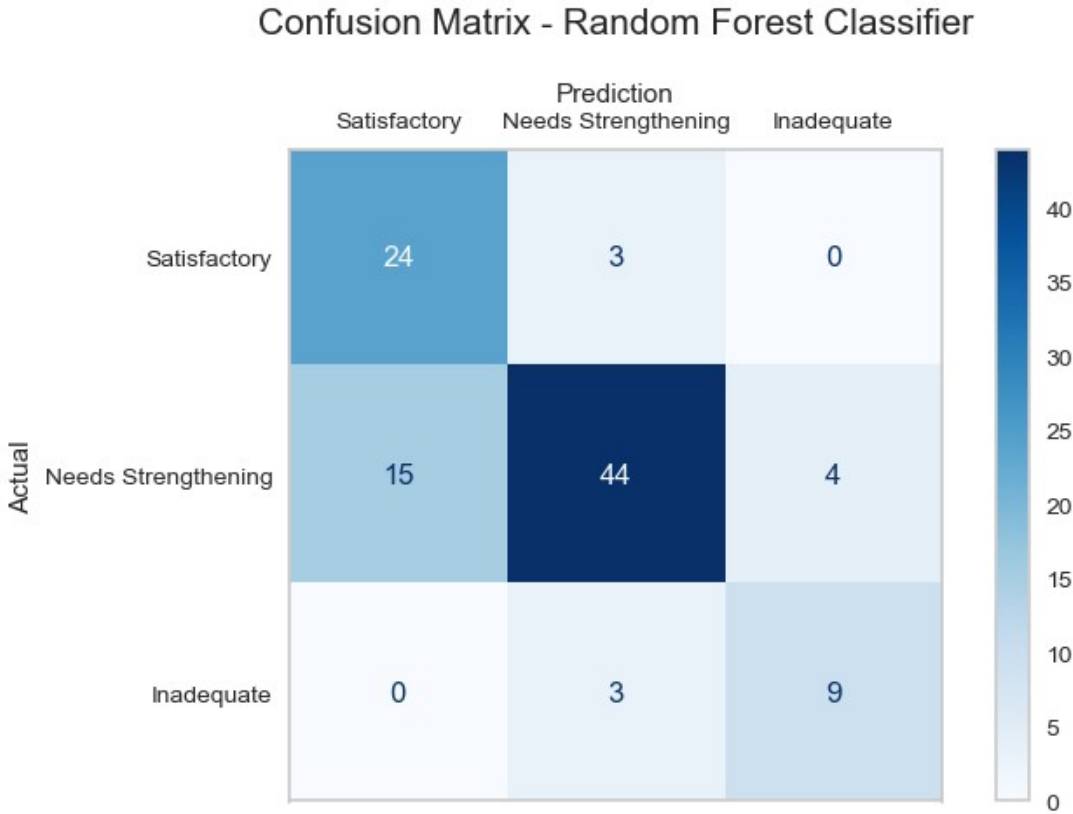


Figure 6.12: Random Forest Classifier: Confusion Matrix (CE Validation)

The individual precision and recall scores (reported as a weighted average across all three classes) for each of the models produced as part of this stage of the experiment are reported below. Similar to the work on measuring accuracy, these results have an average variance of $\pm 1\%$, following 10 epochs of code execution for each dataset. These scores are again reported as before, for the tuned models on the training and validation dataset in columns one and three, with the training performance of the tuned models on the imbalanced dataset reported in column two.

Classifier Precision				
Ref.	Classifier	X train (balanced)	X train (imbalanced)	Validation dataset (balanced)
1	Logistic Regression	69.83%	68.76%	66.05%
2	K-nearest neighbours	79.66%	67.84%	79.1%
3	Support vector machines	69.44%	68.44%	67.5%
4	Decision tree classifier	78.33%	73.17%	78.05%
5	Random forest classifier	78.87%	68.46%	78.79%
6	XGBoost classifier	73.33%	78.8%	70.27%
7	Gradient boosting classifier	75.74%	79.29%	70.07%
8	LightGBM classifier	74.55%	76.97%	66.1%

Table 6.13: Classifier Precision for Control Effectiveness

Classifier Recall				
Ref.	Classifier	X train (balanced)	X train (imbalanced)	Validation dataset (balanced)
1	Logistic Regression	66.34%	68.56%	58.82%
2	K-nearest neighbours	78.71%	68.07%	78.43%
3	Support vector machines	66.58%	65.1%	62.75%
4	Decision tree classifier	75.99%	72.77%	75.49%
5	Random forest classifier	76.49%	68.32%	75.49%
6	XGBoost classifier	72.52%	78.71%	65.69%
7	Gradient boosting classifier	74.75%	78.96%	66.67%
8	LightGBM classifier	74.5%	76.98 %	62.75%

Table 6.14: Classifier Recall for Control Effectiveness

These above results continue to underscore the importance of careful feature selection by the IAF, and offer an encouraging start to adopt ML techniques for assessing control

effectiveness. It is unclear at this point whether additional features, such as external or non-numeric features, will be added over time to support an analysis of control effectiveness. If added in future, the above results should be recalculated to assess the most appropriate algorithm for an ML-based decision making approach to assessing control effectiveness.

As a final step, confusion matrices are produced below for the test dataset, similar to the work performed for IR.

6.3.7 *Control Effectiveness: Model performance on test dataset*

Having completed model design, each of the final, tuned models was tested in turn using the final set of hyperparameters, on the test dataset that was set aside at the start of this experiment. The results have been summarised in individual confusion matrices for each model, and the observed results are consistent with those observed at training / validation stage.

As with IR, the best performing model continues to be the random forest classifier (see figure 6.16), which outperforms others not only because of high accuracy, but has a high precision ratio of 65% for both outlier labels, namely, the "Satisfactory" and "Inadequate" control effectiveness labels. These entities pose a higher risk overall to risk assessments given the heightened regulatory and organisational attention to areas that are designated as such by the IAF ((Wang et al., 2023; Eulerich and Eulerich, 2020)). Conversely, the kNN classifier performed better than the random forest classifier in identifying "Inadequate" entities, with a precision ratio of 75% for this class, but had a similar precision ratio for "Satisfactory" entities.

The confusion matrices produced for Logistic Regression show impressive results, with a simple algorithm such as logistic regression identifying a relatively high number of entities correctly across all three classes. Combined with the nature of the dataset for CE modelling - a combination of just numeric features unlike IR, and the normalisation of the dataset as part of model setup - it is not surprising to see this model performing much better for CE unlike IR. However, the inaccuracy of LR in identifying the outlier labels of "Inadequate" and "Satisfactory" highlight two traits within the dataset. Firstly, the assessment of control effectiveness is not a linear exercise, with non-linear relationships clearly present within the dataset. Secondly, whilst logistic regression may be a useful theoretical exercise, these results may not drive the confidence levels required in ML-based techniques. As such, ensemble models such as the random forest classifier continue to be preferable for risk assessing control effectiveness.

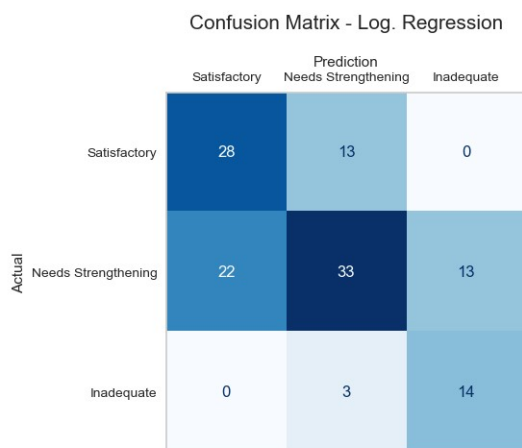


Figure 6.13: CE Confusion Matrix - Log. Regression

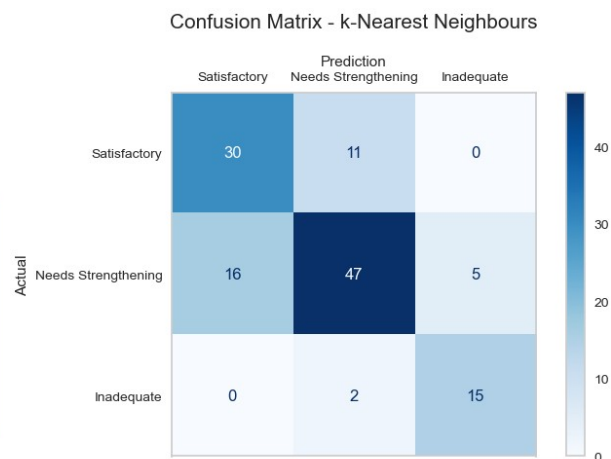


Figure 6.14: CE Confusion Matrix - KNN

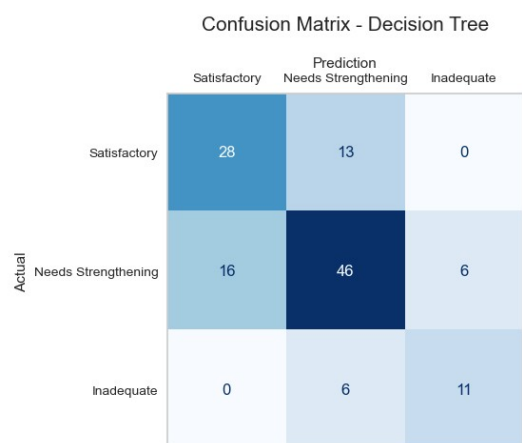


Figure 6.15: CE Confusion Matrix - Decision Tree

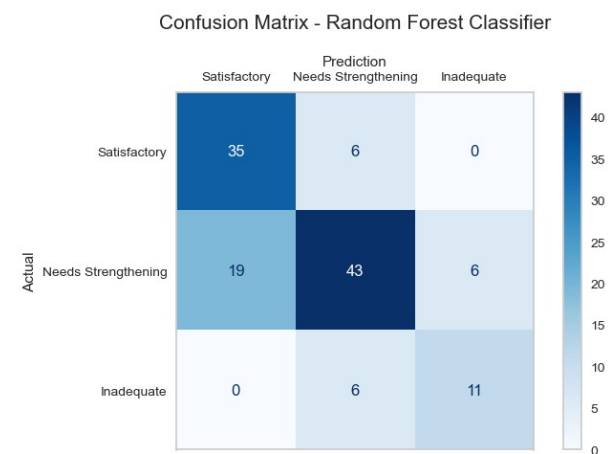


Figure 6.16: CE Confusion Matrix - Random Forest

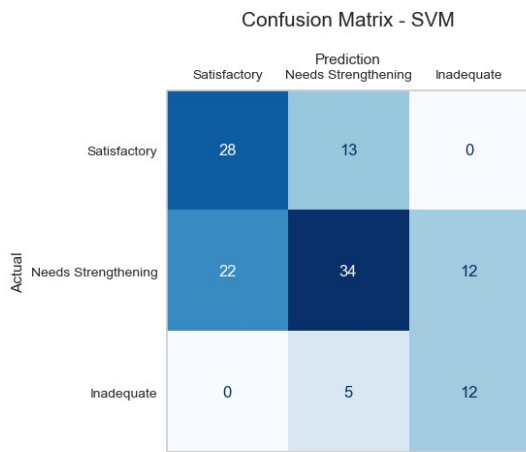


Figure 6.17: CE Confusion Matrix - Support Vector Machines

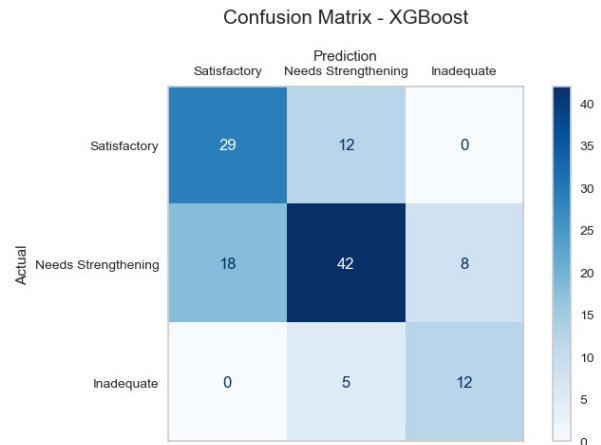


Figure 6.18: CE Confusion Matrix - XGBoost

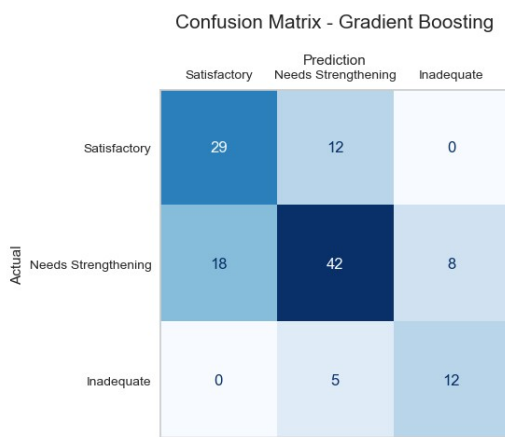


Figure 6.19: CE Confusion Matrix - Gradient Boosting Machines

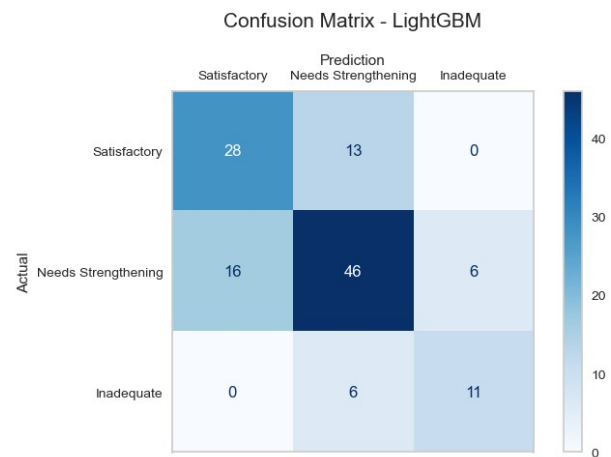


Figure 6.20: CE Confusion Matrix - LightGBM

Based on the reported results and confusion matrices, the random forest classifier continues to be the best performing algorithm, given the high accuracy and class-specific precision scores on the test dataset. The model-specific accuracy ratios as observed on the test dataset are reproduced below in column two.

Classifier Accuracy: Test Dataset				
Ref.	Classifier	X train (balanced)	Test Dataset	Validation dataset (balanced)
1	Logistic Regression	66.34%	59.52%	58.82%
2	K-nearest neighbours	78.71%	73.01%	78.43%

3	Support vector machines	66.58%	58.73%	62.75%
4	Decision tree classifier	76.49%	67.46%	75.49%
5	Random forest classifier	76.49%	70.63%	75.49%
6	XGBoost classifier	72.52%	65.87%	65.69%
7	Gradient boosting classifier	74.75%	65.87%	66.67%
8	LightGBM classifier	74.5%	67.46%	62.75%

Table 6.15: Classifier Accuracy for Control Effectiveness: Test Dataset

A final, short experiment on the dataset is now performed by considering all features, given the findings from statistical tests completed in sections 5.4, 6.2.1 and 6.2.2.

6.4 NOISY DATA: A BRIEF ANALYSIS OF EXCLUDED FEATURES ON MODEL RESULTS

As a final step, the original dataset, from which a number of features were removed following feature independence analysis, is used for modelling purposes. Prior research by Gupta and Gupta (2019), Hasan and Chu (2022) and Saseendran et al. (2019) shows that the occurrence of noise within a dataset can significantly impact the results, and by extension, the confidence in any ML models offered to supplement decision making. To assess the impact of noise on the dataset, and to corroborate the decision to select the features used to build the final, tuned models, the random forest classifier was used on the original dataset, i.e., incorporating the excluded features for IR and CE. This follows from the interim conclusion in section 6.2.3 above. No preprocessing steps, such as normalising the dataset, data cleaning, or multicollinearity assessment, were performed. Following 10-fold cross validation to identify optimal hyperparameters for the classifier, the following model scores were observed:

1. Accuracy - 70.3%
2. Precision - 65.21%
3. Recall - 70.3%

A confusion matrix for the above results shows that the model's training performance is only slightly better than random chance, with most of the High-risk entities incorrectly classed as shown in figure 6.21.

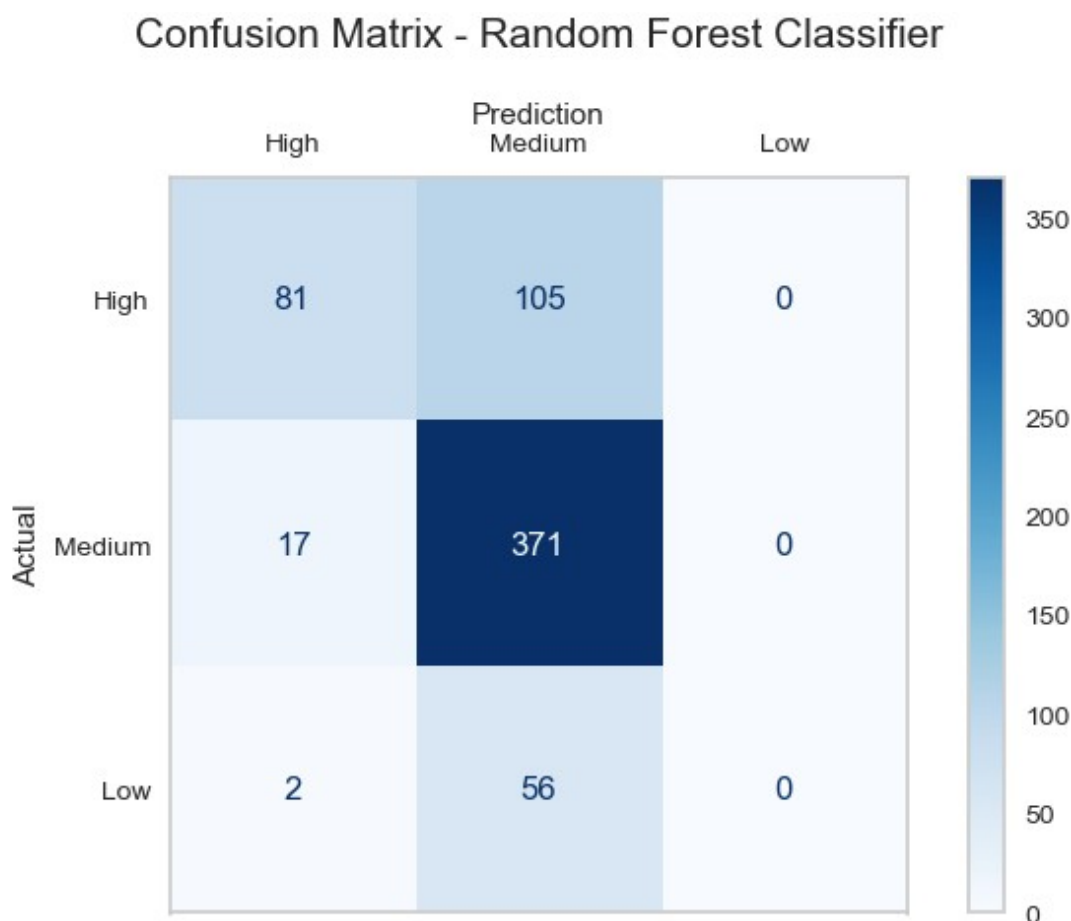


Figure 6.21: IR Confusion Matrix for Raw Dataset

Clearly, the above scores show a significant degradation in model performance when the model was trained on all features, without preprocessing or conscious feature selection. This exercise is now repeated for control effectiveness, noting the following training scores:

1. Accuracy - 71.3%
2. Precision - 74.15%
3. Recall - 71.3%

The lack of data preprocessing has not had a pronounced impact on the CE modelling, perhaps due to the limited number of features originally excluded from the dataset at the start of the experiment - only two features, Medium Risk Event Action Plans and High Risk Event Action Plans were removed for the original round of testing. These features have a strong correlation with the number of Medium and High Risk Events reported by the organisation as seen in section 5.4, but are not expected to materially alter the performance of the final model. Indeed with the final confusion matrix in figure 6.22 for this work, the Random Forest Classifier continues to perform optimally in identifying "Satisfactory"

entities, but has nevertheless misclassified the majority of "Inadequate" entities that have been labelled as such by the IAF.

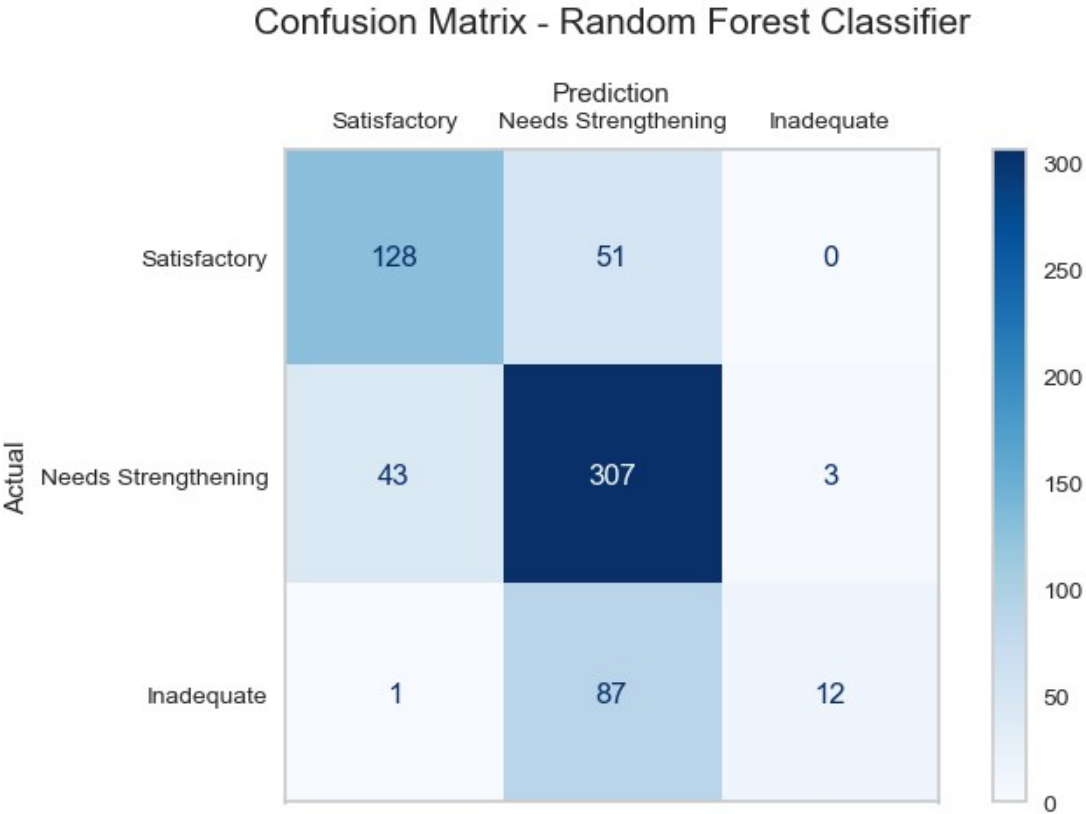


Figure 6.22: CE Confusion Matrix for Raw Dataset

Despite the lack of material impact on the model’s overall accuracy, the degradation in the model’s precision in identifying "Inadequate" Control Effectiveness serves as yet another example of careful and considered feature selection prior to model design. This aspect is explored further in the next section of research.

6.5 FEATURE IMPORTANCE ASSESSMENT

The next step is to investigate the key features driving the model’s predictive power for both variables. Despite widespread adoption, ML models remain mostly black boxes (Ribeiro et al., 2016). Research on ML model interpretability by Carvalho et al. (2019), Rodríguez-Pérez and Bajorath (2020), Honegger (2018) and Linardatos et al. (2021) has shown that the key requirement for a model to be interpretable is for it to be simple enough for a non-technical end user to interpret in full. With multiple features used in the dataset, interpretability of the ML model’s predictions might appear complex at first instance, but when considering the collective expertise of the IAF and the end user being well versed on

the subject of risks and internal controls implemented within the organisation, model interpretability is not a challenge as long as the features driving the model's predictions are clearly identifiable. Whilst there are multiple methods to interpret machine learning models, extant research by Carvalho et al. (2019), Ribeiro et al. (2016), Salih et al. (n.d.) and Rodríguez-Pérez and Bajorath (2020) specifically calls out the use of two primary techniques, namely, Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), as being model agnostic methods that are favoured by Data Scientists to assess the predictive capabilities of ML models. Salih et al. (n.d.) states that SHAP and LIME are two of the most popular methods to use for interpreting the machine learning black box, especially when considering the dataset as being tabular data, such as in this case. These two methods were used to evaluate feature importance, and the underlying methodology adopted for each is discussed below.

SHAP values are based on the concept of cooperative game theory, which provides a fair allocation of contributions to each feature depending on how much it contributes to the prediction. This is particularly relevant in the context of auditing where understanding the contribution of each feature to risk assessment is crucial, and especially considering auditors might accord greater weight to some features compared to others in light of organisational priorities or audit reporting lines (see Norman et al. (2010) where an empirical analysis found that internal auditors perceive higher personal threats when they report high levels of risk directly to the audit committee instead of management). SHAP works by calculating the contribution of each feature by considering all combinations of features and computing the average marginal contribution of a feature across all possible subsets of features, ensuring a fair distribution of importance. SHAP values are "additive", which means the sum of all SHAP values for all features equals the difference between the model's prediction and the average prediction, promoting interpretability and consistency.

Moreover, unlike traditional feature importance metrics which only provide an aggregate measure of each feature's importance, SHAP values are able to provide detailed insights into how each feature contributes to each individual prediction, an approach that has been tested successfully in prior research (Rodríguez-Pérez and Bajorath, 2020). This allows for a more nuanced understanding of the model, making it easier to identify and understand complex interactions between features.

SHAP also provides a global interpretability measure, giving an overview of the impact of all features across the entire dataset, as well as local interpretability that explains individual predictions. This dual perspective can help in validating the model and in explaining specific audit risk assessments, aiding decision making.

Conversely, LIME focuses on providing local explanations for individual predictions by approximating the model locally with simpler, interpretable models like linear regressions or decision trees.

LIME's model agnostic applicability may be useful here and is discussed in detail by Ribeiro et al. (2016) who advocates LIME as a novel explanation technique for any classifier. Unlike SHAP, LIME works by generating a new dataset by perturbing the input data around the instance being explained. For each perturbed instance, LIME records the model's predictions and fits a simpler, interpretable model to these new data points, which is intended to locally approximate the complex model's behaviour in the local region around the instance. Closer data points are given higher weighting to ensure that the local approximation is more accurate around the point of interest. Finally, the coefficients of the simpler model provide insights on the importance and direction of impact of each feature on the specific prediction, indicating how a change in feature could locally impact the model's output.

Based on the methodology used by SHAP and LIME, SHAP is a clearer choice for the experiment, given the existence of non-linear relationships within the dataset as shown in sections 5.1 and 5.3. The use of a simpler technique to fit local permutations of linear variables, using LIME, may not particularly yield useful insights to an auditor given these non-linear relationships. Finally, Linardatos et al. (2021)'s seminal research cited by over 2,500 researchers, advocates SHAP as a desirable approach to assessing feature importance not only on the basis of differentiating between output classes, but also "in terms of better aligning with human intuition" when compared to other existing methods. This is a particularly important trait of machine learning interpretability that specifically applies to internal audit risk assessments, which are heavily subjective (Redmill, 2002; Aven, 2016).

On this basis, SHAP visualisations are now built for the two target variables of interest in the following subsections.

6.5.1 *Feature importance using SHAP - Inherent Risk*

The SHAP feature importance scores for the random forest classifier model used to generate inherent risk predictions are shown in figure 6.23.

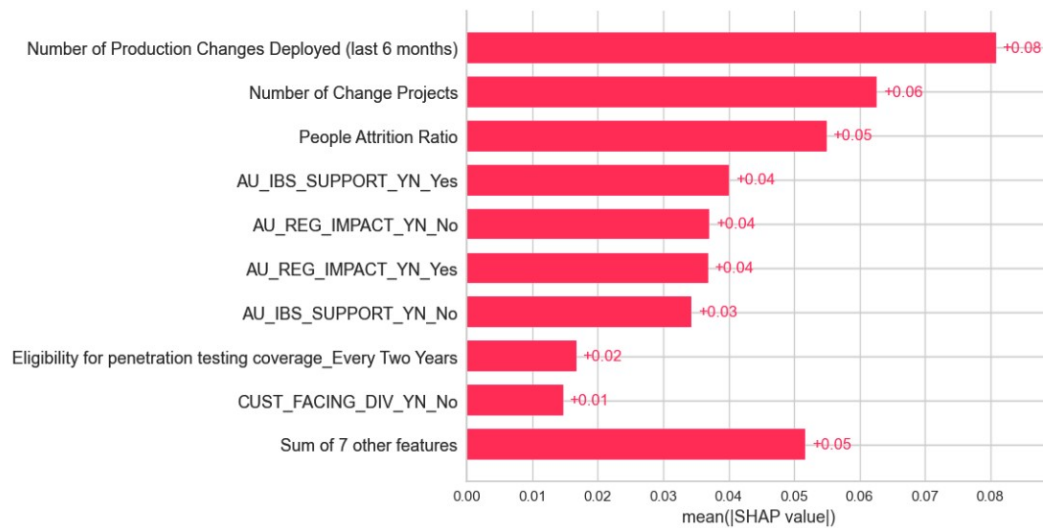


Figure 6.23: SHAP feature importance scores for inherent risk

As can be seen above, the use of SHAP has clearly outlined the key IR features that are significantly driving the model's prediction power. These features corroborate pre-existing subjective risk assessment outcomes and notions established within the sponsoring organisation, and also give auditors confidence in the model's predictions, as change projects, production change volumes and regulatory impact, key features as highlighted by SHAP, are also deemed significant for manual risk assessments. However, given the already high accuracy, corroborated with desirable precision as evident from the classifier's performance on test data, no further changes to the IR model are proposed and the model's ability to generate ongoing value to the IAF is considered optimal. Whilst no changes have been made to reconfigure the final, tuned random forest classifier, the use of SHAP may be used in future executions of the model development to assess whether further feature selection is required to improve model accuracy or precision.

6.5.2 Feature importance using SHAP - Control Effectiveness

The above approach is now repeated with the control effectiveness features. The best performing model was again a random forest classifier, for which the feature importance scores are visualised below in figure 6.24.

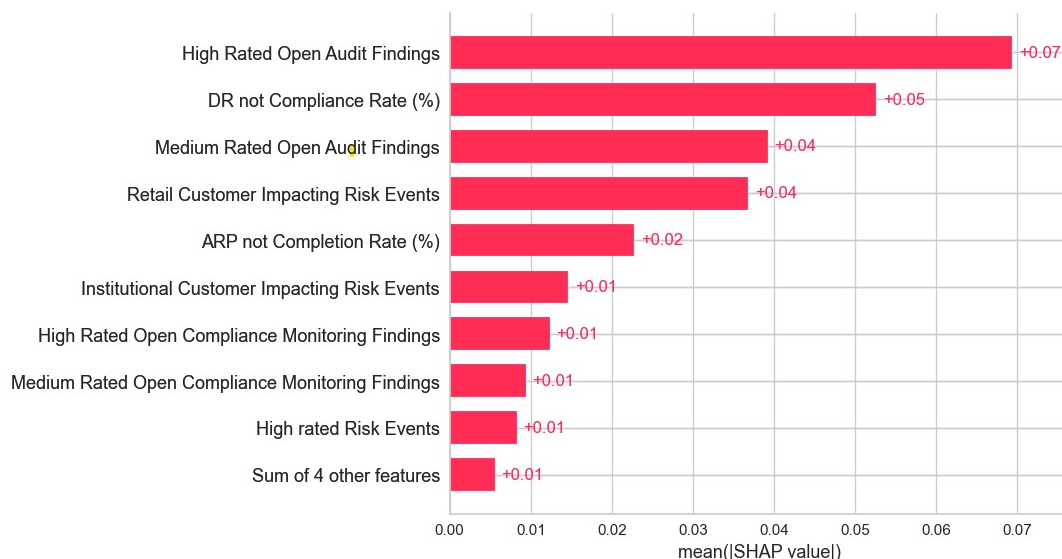


Figure 6.24: SHAP feature importance scores for control effectiveness

Similar to the position with IR, feature selections were not modified, given the already high accuracy ratio of the final CE model.

6.6 INTERIM OBSERVATIONS

This experiment has identified statistically significant relationships between independent variables typically used for risk assessments by IAFs, and offers a valuable alternative to manual risk assessments generally adopted by IAFs. The performance and reliability of the classification approach was validated with the IAF at the sponsoring organisation by engaging with auditors responsible for performing quarterly risk assessments. The IAF's internal review of the BUs highlighted interesting observations as this first part of the experiment involving the use of machine learning for risk assessments concluded, and identified significant benefits to the use of an ML-based approach to risk assessments. Given the observed benefits, the supervised ML-based approach has been incorporated into regular usage at the sponsoring organisation to supplement manual internal audit risk assessments.

Firstly, the final accuracy and precision ratios of the developed classification models, grounded in the application of ensemble decision tree models, and SHAP for deciphering the models, has highlighted challenges to the established positions taken by the IAF as part of previously concluded risk assessment cycles. In particular, the final model's ability to draw out High risk entities was used to support the recalibration of the audit plan at the subsequent round of risk assessment finalisation, highlighting a practical business benefit of an ML-based approach to delivering risk assessments. As such, the use of supervised

machine learning was welcomed by the IAF not only for its innovation, but for the ability to challenge human bias through a scalable, data-driven, machine learning-based approach to a cyclical and subjective process.

Additionally, the IAF acknowledged the model's ability to accommodate complex interactions and non-linear relationships between the various features used for IR and CE, coupled with its robustness against outliers and skewed data, making it a valuable alternative against manual risk assessments.

Whilst the model's highly complex and technical nature posed a challenge for interpretability, the IAF benefitted greatly from the transparency offered by SHAP. SHAP's ability to provide detailed insights into how each feature contributes to individual predictions, as well as an overview of the impact of all features across the entire dataset, was instrumental in building trust with the IAF. Auditors appreciated the contribution of each feature to the risk assessment modelling, which also facilitated a more nuanced understanding of the model.

However, the model's outputs required significant manual override at this stage, which is one of the main practical constraints inhibiting widespread adoption. This is partly attributable to foundational data quality issues, which highlight opportunities to improve foundational data practices at the sponsoring organisation with a view to strengthening the accuracy and robustness of future training datasets. Despite the model's inaccuracy rate for a minority of BUs, and the IAF's override of model-generated scores, the data-driven objectivity introduced by the IAF was acknowledged as a key benefit of the model. The subjective aspect of risk assessments also offered a window of opportunity for experienced auditors to apply their individual expertise for arriving at risk assessment conclusions, rather than being entirely reliant on the outputs of a machine. The key takeaways from the internal validation of the classifier performance were as follows:

1. **Constant retraining** of the final model is required to ensure that the classifier adapts to changes in the audit function's risk assessment methodology, organisational structures and emerging regulations and industry practices, such as the IIA's new code of conduct impacting entities from January 2025 (IIA, n.d.a).
2. The IAF recognised that the model is also adversely impacted by **inconsistent auditor assessments** that have been manually completed in previous quarterly cycles, which by itself was seen as a significant benefit of introducing a machine learning based approach to internal audit risk assessments. Given that previous auditor-assigned risk assessment labels have been used as the training dataset, the lack of consistent manual judgements by auditors directly prevents the model's ability to pick up nuances in the underlying dataset. Therefore, the IAF concluded the need for further

training to be given to auditors in the function, to drive more consistent conclusions in risk assessment cycles, which should reflect in stronger precision across all three classes in future risk assessment cycles.

3. Given the intrinsic subjectivity of the risk assessment process, the IAF concluded that an ML-based approach to risk assessment would be **supplementary to manual auditor review, rather than an outright replacement** for manual / subjective risk assessments. This effectively affords a "human in the loop" component within the process, which strengthens the reliability of final IAF decisions which lead to the formation of the audit plan.
4. The IAF acknowledged the impact of **changing the features** used for machine learning, noting that these may evolve as the audit function's methodologies change over time. This means that the risk assessment process would be subject to constant review and change, requiring the development of new models to deliver data-driven risk assessments. Additionally, three consecutive quarters of risk assessment data are used to develop the final model, which would presumably be required to develop a future model with new features.
5. Finally, the development of ML models requires the considered application of technical skills to tune hyperparameters, select models, manage the input data and support visualisations which are ultimately served to IAF leadership. The typical auditor skill profile should evolve to include data science capabilities, a position endorsed by Huang et al. (2022) and Chiang et al. (2021). This would enable the continued usage of ML within the IAF, driving innovation and objectivity in risk assessments.

Additionally, the use of machine learning also highlighted significant efficiencies for the risk assessment process, which was previously almost entirely manual in nature. For example, the IAF reported that an average of 400 hours time are expended each quarter to perform risk assessments, which would be largely reduced in the event of model adoption. This introduces a third dimension of model value - efficiency - that further supports the assertion of value-add from machine learning, the subject of RQ3.

The primary outcome of the research is that the independent variables used in the analysis, which comprise a mixture of non-numeric and numeric features (representative of those used by IAFs), substantially influence the dependent variables, namely the Inherent Risk and Control Effectiveness scores of business units. Therefore, the extended premise of this conclusion is that alterations within any of these independent variables would engender perceptible shifts in the risk assessment results, thereby affirming their crucial role in shaping audit risk evaluations.

Secondly, the research also evaluated the independence of the variables used for hypothesizing the association with the risk assessment outcomes, which was found to be the case. However, the study did not extend to testing the weight or sensitivity of these features to assess their quantitative influence on risk assessment evaluations. These could be avenues for future research or industrial adoption. Moreover, alternative approaches could also be considered in future research efforts. For instance, the application of non-linear models or the inclusion of additional variables could enhance the robustness of the research and yield more comprehensive insights into the dynamics of risk assessments.

Finally, the experiment's principal value metrics were the model's accuracy and precision, which are two of several dimensions that an IAF would consider as part of practical risk assessment efforts. For instance, the IAF may want to specifically focus on identifying the highest risk areas of the organisation, rather than labelling every business unit through a machine learning model as part of risk assessments. This is further nuanced when considering that a "gold standard" for machine learning training purposes is not available, given the innate subjectivity of risk assessments. Indeed, IAFs comprise of individuals with a range of expertise and skillsets, which could impact the consistency of training labels that would ultimately impact the reliability and robustness of the final model. Therefore, a further extension of this research could be in the area of value measurements, where a more reliable metric for machine learning (or similar data-driven or innovative approaches) could be developed, to engender trust in these approaches and enable adoption at scale by IAFs.

The potential for further research in this area is significant. The insights gleaned from such research could prove instrumental in assisting other IAFs with similar objectives, providing them with a robust empirical basis to guide their risk assessment strategies. In this regard, this research not only contributes to the existing body of knowledge on internal audit risk assessments but also paves the way for future research initiatives aimed at refining and expanding upon these findings.

In the subsequent phases of this research, different machine learning techniques were evaluated to assess their resulting value-add. The next round of experiments with machine learning for risk assessments leverages artificial neural networks, discussed in chapter 7.

A DEEP LEARNING APPROACH WITH NEURAL NETWORKS

This chapter extends the use of supervised machine learning, with the internal audit risk assessment problem continued to be framed as a classification exercise. This chapter examines the value from using Artificial Neural Networks (ANN) to identify predicted inherent risk and control effectiveness labels, with a value to incrementally increasing the complexity of the underlying model to assess the model's ability to capture non-linear relationships seen in the input dataset. This chapter continues to use the same training and validation dataset approach as outlined in chapter 6, with the same set of features as before to ensure comparability of results. The chapter sets out the approach to tuning ANN hyperparameters, which includes an impact assessment of introducing deep learning techniques, to holistically examine value-add from ANN, before concluding on experiment outcomes.

7.1 WHY NEURAL NETWORKS?

The dataset for risk assessments comprises multiple features, categorical and continuous for two separate target variables. The use of multiple processing techniques to reduce overfitting, as seen in section 6.3.1 has shown that these only serve to partially improve accuracy.

Detailed literature reviews by Liao and Wen (2007) has shown that ANN have evolved over time, with data being presented to the network with a view to extracting "maximal separation" between the various classes of examples. The use of ANNs also affords multiple layers of training, rather than recursive epochs of the same dataset which may or may not result in a classifier uncovering new patterns or relationships between the inputs (features) and outputs (the target variables). The Liao and Wen (2007) analysis also highlights the ability of ANNs to be applied to all types of machine learning problems, supervised and unsupervised.

The risk assessment exercise is therefore reframed as ANN, leveraging a popular machine learning library in Python (TensorFlow), as the next step in evaluating RQ2. Consistent with the approach in chapter 6, k-fold cross validation was used to inform hyperparameter

selection, combined with a parametrised grid search to tune the optimal set of artificial neural network parameters.

7.2 ANN METHODOLOGY

The same dataset as the one used for section 6.3 was used for this round of experiments. Similar to the approach described in section 6.3.1, 20% of the dataset was segregated as a "test" dataset, prior to any preprocessing, scaling or modelling steps. Given the stochastic nature of ANNs, multiple models were developed with 10 epochs of execution, across varying hyperparameters. The following settings were used to vary the settings of the models developed over these iterations:

1. **Loss function:** The choice of the correct loss function, used to quantify the "error" or "difference" between the predicted and actual class is essential to being able to correctly draw out predictions using an ANN. Research by Demirkaya et al. (2020) shows that the use of cross-entropy loss functions greatly supports ANN model performance, as it simplifies the training process for a multi-class classification problem such as an IAF's risk assessment. Similar findings were noted by Semenov et al. (2019), who state that the use of the appropriate cross-entropy function contributes to the increase in classification accuracy using ANNs. The loss function was set as a sparse categorical cross entropy, given the multiclass aspect of risk assessments, the shape of the dataframe, prevalent data imbalance, and the sparse distribution of the data as shown in section 5.1.
2. **Layers:** To draw out the best performing neural network, multiple values of layers in the ANN were used, ranging from 3 to 10. The value of output dimension was set to three, the same as the number of classes in the classification problem.
3. **Activation function:** Prior work in this research has highlighted the presence of non-linear relationships among the features. An activation function boosts model performance by introducing non-linearity into neural networks, by enabling them to learn complex patterns. Extant research (Sharma et al., 2020; Wang and Cao, 2023; Dubey et al., 2022) demonstrates efforts to identify the optimal activation function for various types of ANNs; however, as shown in Ramachandran et al. (2017), where the Swish function was first proposed, the Rectified Linear Unit (RELU) is considered most suitable for classification problems. Further research (see Nwankpa et al. (2018) and Wang and Cao (2023) compares two suitable activation functions for multiclass classification problems, namely, ReLU and softmax, with softmax being the most directly suitable activation function for a multiclass model. On this basis, the experiment uses softmax across all epochs.

4. **Optimizer:** The default optimizer used for the ANN within most common machine learning libraries is the Adaptive Moment Estimation (ADAM) optimizer, which adapts the learning rate of the neural network with each layer, enabling more nuanced learning by the model. Given ADAM's ability to handle noisy data and suitability for multiclass machine learning problems (Kingma and Ba, 2017) for a widely cited paper providing a holistic overview of ADAM, advocating the use of the ADAM optimizer for stochastic models and highlights its computational efficiency.
5. **Learning rate and Momentum:** Given that the learning rate hyper-parameter controls the step size taken during the optimization process and drives the timeliness of the updates to the model's weights and biases in response to the gradient of the loss function, the optimal value of the learning rate is essential for computational efficiency. Additionally, the momentum controls the acceleration of the convergence of the gradient descent to enable appropriate learning. These hyper-parameters are less relevant for each iteration given the use of an adaptive optimizer like ADAM. However, various learning rates were tested to assess their impact on the model's performance.
6. **Epochs:** Finally, a range of low to high values of epochs, varying from 5 (very low) to 500 (very high), were used to identify potential loss of training information, as well as arrive at the optimum value of epochs to best train the ANN.

With a view to assessing the optimal ranges of hyperparameters for grid-search evaluation, a two-step approach was adopted for this experiment. In the preliminary iteration, the ANN was developed with randomly chosen input values, with complexity gradually increasing from model 1 to 9 for individual hyperparameters. This enabled a targeted assessment of changes to individual hyperparameters, including the impact of high complexity introduced for the number of layers for capturing non-linear relationships between the features and target variables. As before, each set of nine artificial neural networks was repeated ten times to identify potential variance in the ANN's outputs. This process was repeated for the balanced dataset (defined in section 6.3.3 as $y_{train\ sm}$), with Synthetic Minority Oversampling Technique (SMOTE) again being used to balance the observations across all classes. Once completed, the results of the preliminary iteration were used to inform the appropriate range or choice of inputs for each hyperparameter, which were fed into a combined grid search for a final, tuned set of hyperparameters. This approach helped achieve two objectives - firstly, the impact of increased model complexity using deep learning techniques was assessed in a controlled manner, to identify the point at which increased complexity does not yield improved results. Secondly, the results of the preliminary iteration helped identify an informed choice of the grid search evaluation range to arrive at a robust ANN.

7.3 ANN - PRELIMINARY ITERATION

Based on the ANN methodology described in section 7.2, the results of the preliminary ANN are reported below. The work found that the results stayed within a $\pm 1\%$ variance following ten iterations:

ANN performance				
Ref.	Epochs	Accuracy - original data	Accuracy - balanced data	Notes
Model 1	5	26.47%	39.37%	
Model 2	50	56.86%	53.54%	
Model 3	100	57.84%	62.73%	Number of layers increased to two.
Model 4	125	57.84%	59.84%	Number of layers increased to two.
Model 5	200	57.84%	60.03%	Number of layers increased to three.
Model 6	125	55.88%	56.69%	ADAM replaced with a custom learning rate of 0.05, and a momentum of 0.5 (arbitrarily chosen for experimentation)
Model 7	200	54.91%	55.12%	ADAM replaced with a custom learning rate of 0.05, and a momentum of 0.5 (arbitrarily chosen for experimentation), and increased the number of layers to four.
Model 8	300	57.84%	39.37%	Number of layers increased to ten.
Model 9	500	57.84%	39.37%	Number of layers increased to ten.

Table 7.1: Artificial neural network performance

7.4 ANN - TUNED HYPERPARAMETERS

The ANN experiment produced a number of interesting observations as summarised below:

1. **Accuracy:** As expected, the neural network's performance gradually increased with an increase in the number of epochs across the nine models. However, beyond 125 epochs, there is no increase in the network's accuracy. Indeed, an increase in the

number of epochs coupled with an increase in complexity (brought about from the increase in the number of network layers) adversely impacted the network's accuracy, with the accuracy ratio declining to 58.27%.

2. **Optimization:** Tuning the learning rate and momentum did not produce any meaningful improvement in accuracy, as seen in the delta in accuracy from model 5 to 7. This shows that the use of a static learning rate, which is not adaptable to nuances in the underlying relationships among the inputs, is not suitable for a practical implementation of machine learning capabilities, and especially for an important activity such as risk assessments. This finding corroborates the work of Igiri et al. (2015), who found improvement in model accuracy with an increased learning rate, but advocate the application of an appropriate optimization strategy in machine learning to obtain the best possible result.
3. **Interpretability:** Whilst the accuracy undoubtedly increases with tuning of hyperparameters, there is no way of identifying the drivers for the increases or decreases in neural network accuracy based on a study of the underlying features. For an organisationally sensitive activity like risk assessments, this presents an important limitation in the approach, given that the defence of an audit strategy or plan would require explanations, including to regulatory bodies, from time to time.
4. **Comparison with ensemble classification:** The highest network accuracy ratio is 57.84% with the most optimal set of hyperparameters used for the artificial neural network. This is significantly lower than the classification approach adopted in section 6.3.4 using an ensemble model (random forest classifier) which resulted in a higher training accuracy ratio of 89% for the Inherent Risk target variable.

In order to identify the optimal set of hyperparameters, a five-fold cross validation grid search was used to search for the best combination of:

- **Batch_size**, from a distribution of 10, 20 and 40, drawing on the observations from the initial round of ANN development.
- **Epochs**, from a distribution of 10, 50 and 125, considering the accuracy starts flatlining after an increase in epochs beyond 125.
- **Optimizer**, to select between Stochastic Gradient Descent (SGD) and ADAM, drawing on previous research by Igiri et al. (2015) and Zhou et al. (2020), who performed a comparative study between the two optimizers under different data attributes.
- **Initialiser or init**, to choose between either 'uniform' or 'normal' - with the former initialising the weights of the neural network through a uniform distribution, and the latter using a normal distribution to select the initial weights for the network.

- **Learning Rate**, to specify the optimal rate at which the network understands the underlying relationships and patterns within the dataset. Research by Jepkoech et al. (2021) and Opěla et al. (2022) has shown that a lower learning rate contributes to a better model performance, albeit impacts model speed and efficiency. This is not considered significant at this stage, given that the purpose of this stage is to experiment with the optimal hyperparameter values for the ANN.
- **Neurons**, the basic units in a neural network to process information, from a distribution of 8, 12 or 16, chosen randomly for hyperparameter tuning purposes.

Hyperparameter tuning using grid search identified the following optimal values for an artificial neural network:

- **Batch_size** - 10
- **Epochs** - 100
- **Optimizer** - Adam
- **init** - uniform
- **learning_rate** - 0.001
- **neurons** - 16.

Plugging the above hyperparameters into a further ANN, a validation dataset accuracy of 53.92% was observed, significantly lower compared to the results noted in section 6.3.3, where the best performing supervised machine learning model, the random forest classifier, produced a much higher validation accuracy ratio of 71.57%.

As a final step in this exercise, the final model was applied on the test dataset, which was segregated at the start of supervised machine learning exercise in chapter 6. The ANN returned an overall accuracy of only approximately 57.93%, with the following distribution of predictions as seen in the confusion matrix:

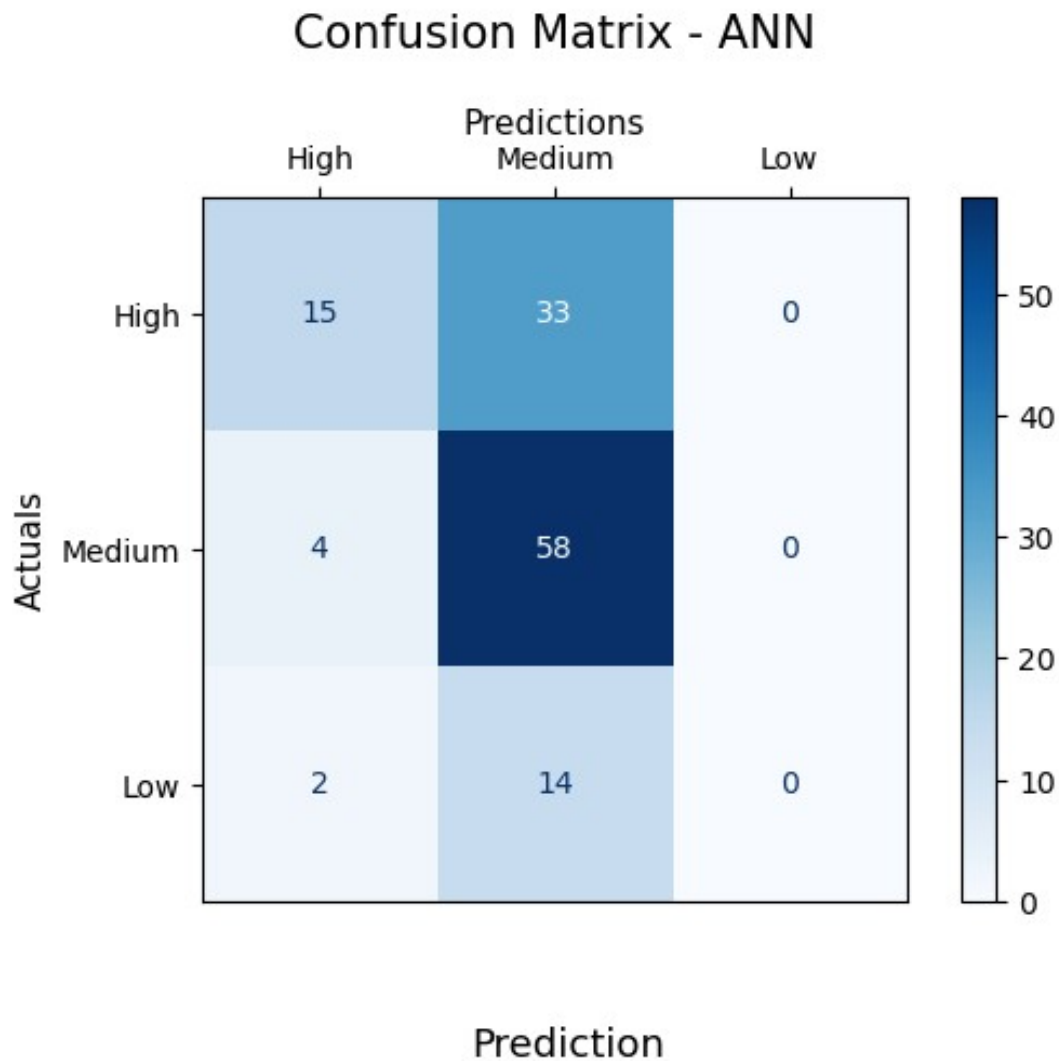


Figure 7.1: Confusion matrix for Artificial Neural Network - Tuned Hyperparameter Values

Reviewing the above confusion matrix, the ANN shows high precision in identifying the "Medium"-risk audit entities, consistent with some of the models in section 6.3.4. This is not surprising given that the Medium class dominates the dataset as the majority class. However, the neural network has not satisfactorily picked up complex non-linear relationships that characterise High or Low entities, with only 15 High audit universe entities being correctly labelled as such by the ANN. The performance of this ANN is very similar to that of a Gradient Boosting Classifier, which was developed as part of section 6.3.4, and the results of which are available in figure 6.10.

7.5 INTERIM OBSERVATIONS

On this basis, there is insufficient basis for advocating the use of an ANN over the supervised machine learning techniques discussed in section 6.3.4. For this reason, this experiment does not proceed to construct a similar ANN for CE. Whilst more powerful in analysing non-linear relationships and nuances in underlying input features, research by Dreiseitl and Ohno-Machado (2002) identified neural networks to be more susceptible to overfitting due to the flexibility offered by the deep learning model. A very similar pattern is seen from the above results as well. Similarly, research by Otchere et al. (2021) showed that a Support Vector Machine (SVM) model fared better than an artificial neural network for a small dataset due to improved parameter selection offered by a conventional machine learning model like SVM. Therefore, whilst the use of ANN cannot be ruled out entirely, an IAF would need to consider the volume of features, organisational universe size, and perform additional fine tuning of hyperparameters prior to leveraging a similar approach taken in this thesis for practical implementation of a neural network for risk assessments. An ANN-based approach to modelling risk assessments may find favour with large enterprises with a more complex approach to internal audit risk assessments, such as vast amounts of data points and / or additional features, including externally sourced data. A large amount of data may be required by an ANN to capture and reflect the impact of non-linear patterns within the dataset, an observation confirmed by exploratory data analysis performed in sections 5 and 5.1.

UNSUPERVISED MACHINE LEARNING FOR RISK ASSESSMENTS

Having experimented with supervised machine learning and ANNs in the preceding sections, the use of unsupervised learning techniques serves as the final segment of the evaluation of RQ2. Contrary to supervised machine learning, this chapter seeks to examine whether the audit universe itself can be grouped into various clusters, or groups, based on underlying relationships between the variables. An extension of this technique - anomaly detection - has also been used by reframing the risk assessment exercise as a binary set of audit universe entities, effectively segregating High-risk entities as the anomaly set. The chapter concludes with an evaluation of these techniques, including potential limitations to practical implementation.

8.1 WHY USE UNSUPERVISED ML?

As shown through literature studies (see section 2.2), ensemble methods involving supervised and unsupervised techniques have been trialled in the past for related areas such as credit risk assessment, as shown in Bao et al. (2019) for example.

For the purposes of this experiment with unsupervised machine learning, two broad categories of techniques are considered as follows:

1. **Clustering:** Given the ability of clustering algorithms to group similar data points, clustering techniques were leveraged to discover unexpected patterns or anomalies within the dataset, which could potentially be used to assess high risk areas of the organisation. This approach relies on prior risk assessments being completed consistently by the audit function, and clustering could well be impacted by non-linear relationships within the dataset as these may have been dealt with subjectively by different members of the IAF in prior risk assessment exercises. Clustering audit universe entities to produce a risk profile support an IAF's risk assessment exercise does not appear to have been subject to empirical research in past as there is little evidence of such attempts in existing literature as shown in section 2.2.
2. **Anomaly Detection:** Separately, the experiment also considered the use of algorithms such as isolation forest, to detect outlier BUs within the audit universe that may be

indicative of high risks. However, given the lack of labels within the data, the clustering of BUs may not result in a risk-based profiling of the audit universe and draw out an unknown pattern within the data. Anomaly detection in areas relevant to financial statement risk, such as fraudulent journal postings or tax fraud has been attempted in prior research as shown in Bakumenko and Elragal (2022) and de Roux et al. (2018). As shown in Bakumenko and Elragal (2022), unsupervised machine learning (specifically, the isolation forest technique) could be used to detect anomalies in general ledger entries (which would be a subset of the area of interest for a risk assessment exercise).

As shown in Bakumenko and Elragal (2022), Huang et al. (2022) and de Roux et al. (2018), supported by the literature review in section 2.2, there is little evidence of prior research being carried out to support machine learning for areas other than forensic auditing or fraud detection. However, there are potential advantages in adopting an unsupervised machine learning approach to support internal audit risk assessment cycles, as shown below:

1. **Identification of Hidden Risks:** Unsupervised learning can uncover patterns and anomalies that may not be apparent through traditional risk assessment methods, where human bias may impact decision making as discussed in Fukukawa and Mock (n.d.) and Cassell et al. (2022).
2. **Prioritization of Risks:** The detection of clusters or even anomalous business units within an audit universe could enable the detection of high risk entities within an organisation, enabling effective prioritisation of an IAF's limited resources. Prior research by Xiao et al. (2020) has highlighted an attenuation of audit quality in larger audit firms, so it is crucial to allocate sparse manpower within an IAF to the most critical or high-risk areas of the organisation.
3. **Early Warning Systems:** Both supervised and unsupervised techniques can lead to the development of early warning systems within the organisation, enabling the flagging of high risk areas of the organisation to enable timely audit or compliance intervention.

Prior research (Bao et al., 2019; Bakumenko and Elragal, 2022) has highlighted a range of critical assumptions prior to using unsupervised machine learning. These are summarised below, in the context of risk assessments:

1. **Data Quality:** As with supervised machine learning, the quality of the data used for unsupervised learning is crucial. As part of section 4.2, the quality of input data used by the IAF for prior risk assessment exercises has been tested and is a reasonably representative dataset of the organisation. The same dataset used for supervised

machine learning, following cleaning and pre-processing in section 6.3, has been used for this round of experimentation.

2. **Choice of technique:** Whilst multiple unsupervised machine learning techniques could be used for clustering data or identifying anomalous business units, the choice of technique must be informed by the nature of the data, as shown in section 5.3. For example, the use of k-means clustering for grouping similar observations relies on these being spherically distributed (Amayri and Bouguila, 2013), whilst density-based spatial clustering is able to deal with non-spherical data distributions, which are typically characteristic of organisational data and as demonstrated in section 5.3.
3. **Interpretation:** Interpreting the results of unsupervised learning can be challenging, as it often requires domain expertise to understand the patterns and anomalies identified, despite the use of visualisation techniques as seen in prior research by Venderley et al. (2022). This is not expected to be a challenge given that the user base of a practically implemented solution leveraging unsupervised machine learning would be a skilled IAF responsible for performing risk assessments for the organisation.

8.2 PROPOSED APPROACH FOR LEVERAGING UNSUPERVISED ML

Based on the considerations outlined in the previous section, two separate experiments were developed to cluster the datasets using density-based clustering libraries and to separately detect anomalous business entities using multiple anomaly detection libraries.

Two popular density-based clustering algorithms, density-based spatial clustering of applications with noise (DBSCAN) and High density-based spatial clustering of applications with noise (HDBSCAN) were developed for this experiment.

This choice was influenced by the fact that the dataset is not spherical in nature, which means that an alternative clustering technique such as k-means clustering would not be an appropriate choice. Neither was the exact number of clusters specified to mirror the number of risk classes (three), as the expectation with unsupervised machine learning is that the machine learning model detects unknown patterns within the data to highlight clusters independently.

Separately, three anomaly detection models were developed to identify anomalous business units based on the dataset provided, with "anomalous" defined as an entity with "High" inherent risk. This is because an organisation is primarily interested in the ability to timely detect "High" risk entities to drive proactive audit planning efforts, with accurate labelling

of medium or low risk entities less of a business concern. To strengthen confidence on the final models developed for anomaly detection, 20% of the original dataset was segregated, prior to detailed preprocessing or model development.

Both experiments were executed for 10 iterations to assess variance in observed results and table the results of the work below.

8.3 CLUSTERING TECHNIQUES

The unsupervised machine learning experiment commenced with the use of DBSCAN, a powerful clustering algorithm that has been effectively applied in the past to generate meaningful clusters of data (Deng, 2020; Kumar and Reddy, 2016). DBSCAN's ability to automatically determine the number of clusters based on the data density (Kumar and Reddy, 2016) makes it a valuable starting point for unsupervised machine learning research, given that a data profile of the audit universe is desired, based purely on the quality of input features assigned to each organisational unit. DBSCAN's density-based approach also makes it less sensitive to outliers, which are common in organizational data. Additionally, as shown in section 5.1, the dataframe is not normally distributed across the set of organisational units and DBSCAN's ability to identify clusters of arbitrary shapes and handle noise renders it well-suited for the complex and often non-spherical nature of organizational data. By applying DBSCAN to organizational units, natural groupings of organisational units are expected to emerge, possibly based on shared characteristics such as size, function, location, or risk exposure.

DBSCAN requires us to specify two hyperparameters:

1. **Epsilon ϵ :** DBSCAN works by grouping data points within a pre-defined radius based on their density, or the concentration of data points within a specific region (also referred to as a "neighbourhood"). The algorithm works by identifying core points within a neighbourhood cluster, expanding clusters from these core points, and labelling non-clustered units as noise. Therefore, the first hyperparameter required to be specified is the radius of the neighbourhood to consider, which is known as ϵ . Prior research such as Ramadan et al. (2022) and Esmaelnejad et al. (2010) has attempted to define structured techniques to identify the value of ϵ for a dataset. For example, the Ramadan et al. (2022) analysis proposes the repeated application of DBSCAN for as many times as required, evaluating the results against a pre-defined criteria of clusters until the optimal ϵ has been identified. The same technique has been adopted for this experiment.

2. **Minimum data points:** Whilst DBSCAN works out the optimal number of clusters based on the data, the user is required to specify the minimum number of data points for a neighbourhood to be formed for clustering purposes. A risk assessment dataset is not balanced by definition, as the number of high risk entities within an organisational setup would be far outweighed by non-high risk entities. Therefore, a range of minimum data points for a cluster were specified for the experiment to enable a meaningful formation of densities. This value was initially set as 5 for the purposes of the experiment, with the epsilon value varied across multiple iterations to assess impact and arrive at an optimal value for this hyperparameter.

The experiment was further segmented to initially consider only numeric features in a first run of unsupervised machine learning, to assess whether the lack of categorical markers (reflecting auditor bias and judgement, such as an organisation's customer facing aspect or regulatory exposure) have any impact on the formation of clusters.

Having segmented the dataset to consider only numeric values, the dataset was scaled prior to running DBSCAN, and varying values of ϵ applied, with 10 iterations for each run. The results are as follows, all within a $\pm 1\%$ variance following ten iterations:

DBSCAN Results			
Ref.	Epsilon	Clusters identified	BUs classed as Noise
Model 1	1	6	291
Model 2	2	6	136
Model 3	3	1	96
Model 4	4	2	59
Model 5	5	2	41
Model 6	6	3	21
Model 7	7	2	17
Model 8	8	2	4
Model 8	9+	1	4

Table 8.1: DBSCAN Results

The experiment showed that an increase of the epsilon value beyond 9 rendered the exercise meaningless, as there was no further clustering of data points to highlight changes in density. Whilst this round of the experiment only focused on numeric variables, interesting shifts in the clustering approach were observed with the changes to epsilon

values. Firstly, as expected, the observations classed as noise gradually reduced as the epsilon value was increased. Secondly, the value of $\epsilon = 6$ resulted in the most optimal clustering of data points, as DBSCAN segmented the dataset into three clusters, plus a separate set of 16 observations classed as noise. This result is shown in figure 8.1. There was no single pattern affecting all observations in a DBSCAN cluster.

To visualise the DBSCAN outputs, the full set of 22 numeric features was standardised to a common scale, and compressed into two dimensions, using Principal Component Analysis (PCA). PCA achieves dimensionality reduction by identifying the directions that capture the greatest variance in the data, effectively compressing the essential information contained within the entirety of the feature set while minimising information loss. The resulting two-dimensional representation shown in figure 8.1 offers a starting point to effectively visualise any natural groupings within the dataset, and identify useful insights for supporting risk assessment outcomes.

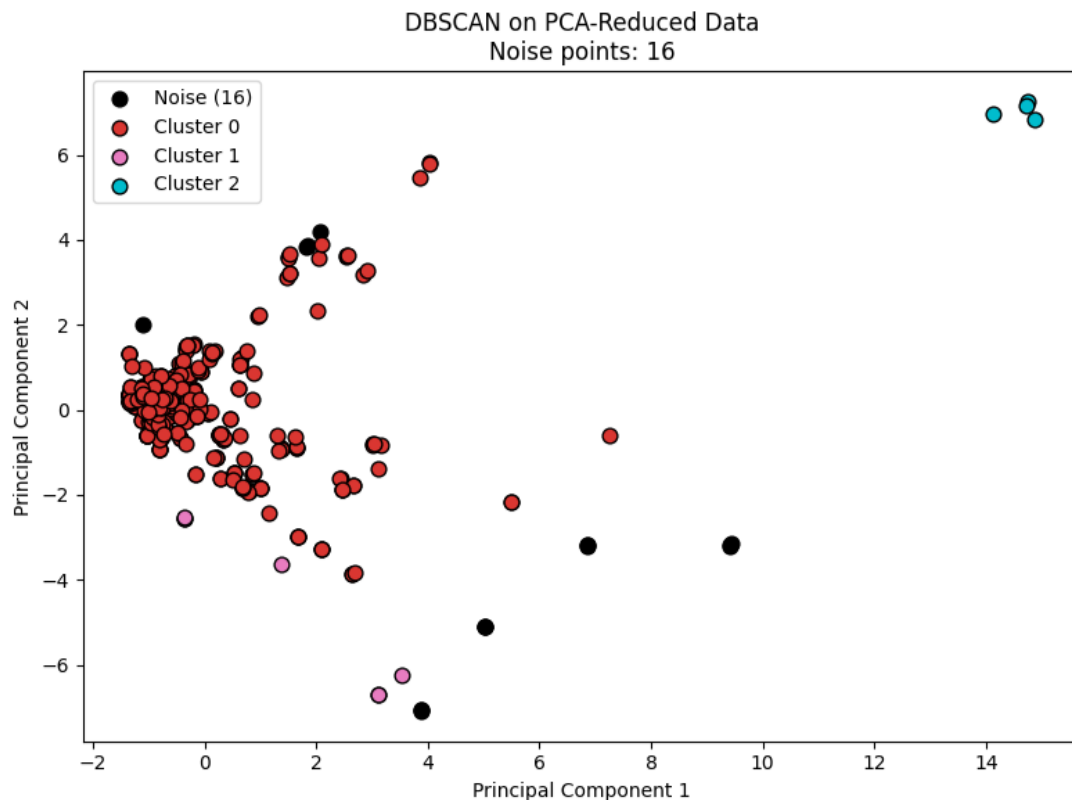


Figure 8.1: DBSCAN Clustering with $\epsilon = 6$, $n = 506$

Given that the number of clusters was the same as the number of risk assessment outcomes, the results of this iteration were explored further:

1. Group 0 had the highest number of organisational units (150), with three BUs grouped in Group 1 and one in Group 2. Seven BUs represent the 21 noisy

observations (a BU would be repeated across the results given that observations have been collected for the same BU across multiple quarters).

2. A gradual increase in the inherent risk profile of clustered organisational units was seen within BUs across clusters 0, 1 and 2. For example, the people attrition ratio was in single digits for BUs clustered in group 2, but at least 2.7% for each BU grouped in cluster 2. Whilst this offered a promising start for the review, the lack of an organisational alignment of BUs across or even within a single cluster did not allow a deeper analysis.
3. The results of prior risk assessments could not be overlaid on the clusters identified from this exercise. Neither did the data points captured in any of these groups show a clear pattern that would have enabled an experienced or competent auditor to meaningfully identify these as belonging to a specific group of audit entities requiring further analysis.

Despite the potential insight offered by model 6, this exercise does not reflect important features required by auditors as part of a risk assessment exercise, which are only represented in categorical attributes that have been separately captured. To this end, the exercise completed above is limited in usage and adoption and the clustering of data into three groups is not reflective of the all overall risk profile of the relevant BUs within each cluster.

This experiment was therefore repeated with all features included, with results reported below:

DBSCAN Results			
Ref.	Epsilon	Clusters identified	BUs classed as Noise
Model 1	1	19	28
Model 2	2	19	28
Model 3	3	12	14
Model 4	4	2	2
Model 5	5	2	0
Model 6	6	2	0
Model 6	7+	2	0

Table 8.2: DBSCAN Results - All features

The experiment showed that an increase of the epsilon value beyond 5 rendered the exercise meaningless, as there was no further granular clustering of data points to highlight changes in density. Similar to the first experiment above, there was a gradual drop in observations classed as noise as the epsilon value increased. Additionally, the number of clusters dropped sharply as the epsilon value increased from 3 to 4, but this has led to considerable information loss, as the model was no longer able to discern meaningfully between various groups of organisational units from this point onwards.

The nature of organisational units grouped as noise across each model were spot checked at this stage. As expected, there was no consistent profile of noise points in model 1, given the low epsilon value of 1. Indeed, the 28 observations classed as noise by DBSCAN were found to be across all organisational units, and exhibited a range of attributes across all numeric features, with the number of key applications held by an organisational unit in this grouping ranging from 0 to 163. However, the individual clusters themselves showed a meaningful segmentation of the data. For example, cluster 1 comprised only of those organisational units without any key applications, or requirement for penetration testing, due to the lack of customer or external-facing applications. These are characteristic of low risk entities, although all entities in cluster 1 did not correspond to a single organisational area. Similarly, cluster 2 showed entities with a slight increase in key applications, and interestingly, a deterioration of the control environment reflected in a higher DR non-compliance rate for these applications. There was also an increase in regulatory impacted BUs from this cluster onwards, alongside an increase in BUs that are customer-facing in nature. From cluster 4 and above, organisational units had an average of 25 key applications, with penetration testing required at least every two years. This shows that the DBSCAN clustering algorithm has labelled organisational units on an increased risk profile, ranging from cluster 1 to 19 in this experiment, with inconsistently assessed risk assessment units being classed as noise.

Despite the potential insight offered by models 1 to 3, the number of clusters remains quite high at 12, which is not useful for a risk assessment exercise, requiring the profiling of organisational units across three categories of High, Medium and Low categories to enable effective audit planning.

The clustering experiment concluded with a rerun of the same exercise using HDBSCAN, which is similar to DBSCAN in approach, but does not require the user to specify the epsilon value. Existing literature (Gao et al., 2023) highlights the limitation of HDBSCAN unnecessarily classifying outliers as noise as a potential limitation. This observation was corroborated with the results of the experiment. Using the same parameter of five as the minimum cluster size, HDBSCAN returned 13 potential clusters of the data, with a concentration of organisational units in group 10 of the clusters. This was found to be true

for 10 iterations of the code, with very low variance (less than 1%) in the observed results. Similarly, the observations and limitations observed with the DBSCAN experiment were identical, and therefore, the results of the HDBSCAN experiment cannot be meaningfully extrapolated for risk assessment purposes.

8.4 AN ANOMALY-BASED APPROACH TO RISK ASSESSMENTS

The next step is to draw out specific anomalies from the dataset, with the expectation of identifying observations that indicate high risk areas of the organisation.

In order to achieve this objective, the number of classes were reduced from three to two, to simplify the problem statement. For the purposes of this experiment, an "anomalous" business unit is defined as a "High" risk business unit. The target variable is reframed as a binary class of 0 and 1, with 1 representing such anomalies. The anomaly detection approach was designed to train the dataset on the non-anomalous business units, in order to successfully profile outliers or anomalies, for which the population of "High" risk business units would serve as a test dataset.

The initial demarcation of the data shows the following split of business units in readiness for the deployment of anomaly detection techniques. Of the 506 entities that are subject to analysis, there are 140 High risk entities and the remaining 366 entities are labelled as either "Medium" or "Low", which are collectively referred to as "Non High" in figure 8.2 below. In other words, the number of "High" risk entities is 38% of the population of "Non High" entities.

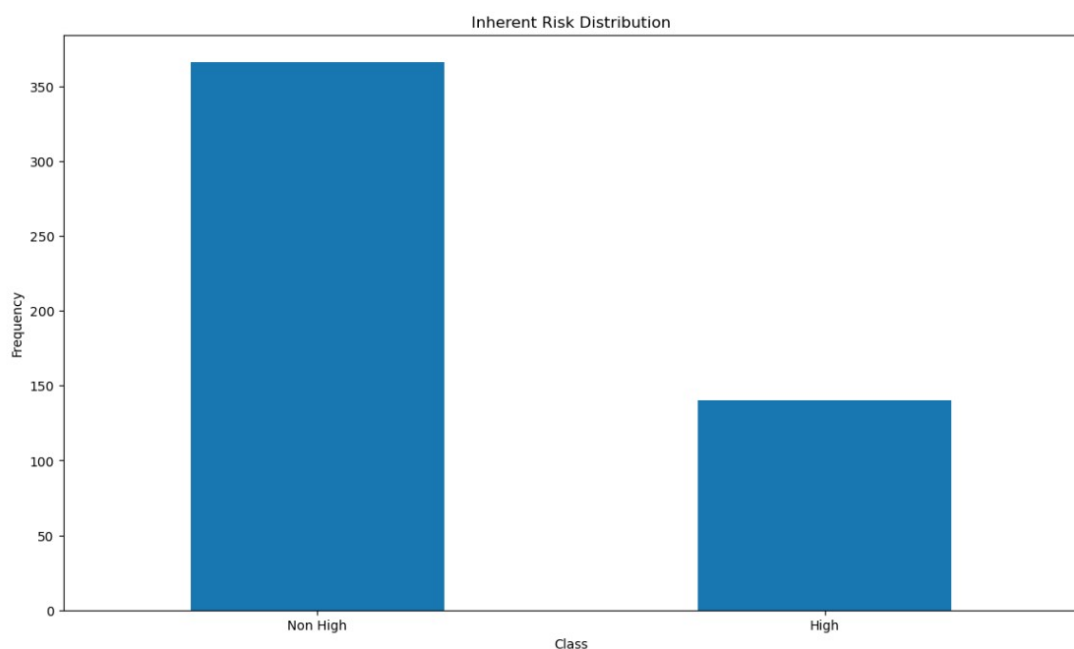


Figure 8.2: Data structure following conversion of multi-class labels to binary class

A preliminary view of the dataset suggests that whilst it is imbalanced in favour of the negative class (non high business units), the imbalance is not so severe as to render the experiment insightful. Indeed, prior research (Kong et al., 2020; Kim et al., 2020) shows that the treatment of the imbalanced classes is not always the preferred approach in anomaly detection, with Kong et al. (2020) using an alternative approach, of introducing additional information to the dataset with synthetic data.

To this end, three datasets are prepared for anomaly detection purposes - the original dataset used for all experiments, showing the true number of observations (506) spread across all three classes (labelled alpha), a second dataset balanced at 366 observations (the majority class) for both classes (labelled beta), and a final dataset that leveraged oversampling techniques in favour of the positive (Non High) class (labelled gamma), with a view to training the anomaly detection models further on a wider range of synthetic data modelled on the original dataset. This resulted in populations of 100,000 for the Non High class and 10,000 for the High class.

Similar to all experiments completed in this research, the anomaly detection classifiers were executed for a total of 10 separate iterations and the accuracy ratios being within a tolerance range of <1%, supporting the reliability of these findings.

The key variation in this experiment was to also track the recall of the model's performance, in addition to the accuracy metric. This is because the model's ability to draw out the High risk business units is paramount, given the heightened importance assigned to these as part of the problem statement. Accuracy is also used to verify the model's performance and ability to predict each class as expected.

For anomaly detection, three models are built as explained below:

1. **Isolation forest:** Used in prior research involving large datasets for anomaly detection purposes (Al Farizi et al., 2021; Xu et al., 2023; Lesouple et al., 2021; Ding and Fei, 2013), isolation forest is a popular technique that relies on outliers or anomalies being the minority class, and isolates them by randomly selecting features from the dataset and splitting them to isolation. The premise of an isolation forest technique is that anomalies have a shorter path to isolation and the depth of these anomalous units would help identify these as anomalies for further investigation. The lower the calculated isolation score (or depth of isolating a particular data point after splitting all datasets based on randomly selected features), the greater the probability of a data point being marked as an anomaly.
2. **Local outlier factor:** An alternative anomaly detection technique, Local Outlier Factor (LOF) works by measuring the local density of each data point, similar to the DBSCAN clustering technique. The LOF of each data point is the ratio of the average

density of its neighbours to its own density and therefore, higher the LOF score, higher the probability that the data point is an outlier. Prior research by Cheng et al. (2019), Fadul (2023) and John and Naaz (2019) has undertaken comparative studies of isolation forest and LOF, with both algorithms successfully deployed for detecting anomalous observations across high dimensional and high volume datasets.

3. **One class SVM:** A variant of the supervised machine learning technique (support vector machines), this anomaly detection technique works by drawing a boundary around a set of "normal" data points, with points falling outside the boundary being considered outliers. The boundary itself is controlled by the hyperparameter "nu" within the algorithm, and a higher value of "nu" indicates lower sensitivity to outliers. The distance of each observation from the hyperplane boundary is used to determine its classification as an outlier. As shown in Hejazi and Singh (2013), this technique can be very useful for large datasets and successfully leveraged to detect outliers in practical settings such as credit card fraud.

The methodology involved four steps as enumerated below:

1. The negative class observations were firstly isolated into a dedicated dataset to clearly profile the business units and observations that do not represent high risk entities. Devoid of any contamination, the above models returned an expected accuracy score of 100% when fed the non-high observations in isolation.
2. Having setup the anomaly detection models with exposure to the "Non High" population alone, the same models were tested against the original (i.e., imbalanced) and evenly balanced datasets comprising of negative (medium or low rated entities) and positive class (high risk entities) observations. This approach began with the balanced dataset, where the contamination ratio hyperparameter was set to 0.5, given that both isolation forest and LOF require this value to be within a range of 0 (no contamination) and 0.5 (a dataset split evenly across non-anomalous and anomalous observations. For the purposes of this exercise, "contamination" is defined as the proportion of the dataset that is anomalous. Given step 1 above, a dataset with 10% high risk entities would be said to have a contamination ratio of 0.1. A contamination ratio of 0.5 was achieved through oversampling the dataset and creating a "beta" dataset where the positive and negative classes have 366 observations each.
3. A separate set of anomaly detection tests were executed on the original dataset, which had a contamination ratio of 38% as noted above. This is the crucial dataset for the purposes of the experiment, as it would be the "real life" dataset where the anomaly detection techniques are expected to draw out high risk entities as desired for risk assessments.

4. Finally, the anomaly detection models were applied on the gamma dataset, where a synthetically large dataset of "Non High" observations was generated to better support the anomaly detection models in detecting potential outliers within the dataset.

The results for each iteration are reported separately for accuracy and recall for each of the three sets of data based on the performance of the various anomaly detection algorithms as at the end of step four above.

Anomaly Detection Accuracy			
Dataset	Isolation forest	LOF	One class SVM
Alpha	57.51%	54.94%	57.11%
Beta	57.11%	42.62%	48.36%
Gamma	62.66%	62.55%	79.44%

Table 8.3: Anomaly Detection Accuracy

The accuracy ratio of all three models is below average, demonstrating the inability of the models to correctly draw out the high risk entities based on the data provided. There is some improvement in accuracy following synthetic data augmentation, based on the results of the gamma dataset. The recall ratios are reported below for each anomaly detection model for the positive class (High risk entities) only:

Anomaly Detection Recall (positive class only)			
Dataset	Isolation forest	LOF	One class SVM
Alpha	42%	37%	35%
Beta	45%	31%	33%
Gamma	55%	54%	23%

Table 8.4: Anomaly Detection Recall

The recall of the predicted positive class is extremely poor, highlighting the complex relationships within the underlying dataset which have rendered it very difficult for any of the three models to pick out the anomalous entities. There is a high recall for the negative class (averaging 93% for all three datasets), and it is therefore clear that the accuracy is skewed as a result of the strong performance of the anomaly detection algorithms to profile the negative class.

Therefore, as a final step in the anomaly detection experiment, the final models were applied on the originally segregated "test" dataset, set aside prior to initiating data

preprocessing for unsupervised machine learning. This dataset comprised of 126 observations in total, split as 78 and 48 across "Non Highs" and "High" risk entities respectively. 10 epochs of the anomaly detection were executed on the test dataset with an accuracy ratio of 55.55%, with recall of 42% for the "High" risk entities in the dataset. Similar to the earlier reported metrics, the recall remains low and is not indicative of a useful solution that lends itself to practical implementation within the IAF. The full set of results across all three anomaly detection models on the test datasets are reproduced below:

Anomaly Detection Results: Test Dataset			
Metric	Isolation forest	LOF	One class SVM
Accuracy	55.55%	60.32%	55.55%
Recall (Positive class)	42%	48%	62%

Table 8.5: Anomaly Detection Results: Test Dataset

8.5 INTERIM OBSERVATIONS

Unsupervised machine learning does not appear to have been sufficiently studied in extant literature in the context of internal auditing and specifically, risk assessments. The experiments in this chapter have therefore explore a range of possible techniques to support IAFs with unsupervised techniques for risk assessments, but highlight a number of precautions that are required to be adopted prior to concluding on the efficacy or otherwise of such an approach.

Firstly, the IAF must place significant emphasis on the underlying quality of data, as poor quality data will significantly impact the reliability of any unsupervised technique and will not enable the IAF to draw out useful patterns or insights from either clustering or anomaly detection techniques. Density-based clustering shows relatively high promise for the dataset as it was able to successfully detect some patterns among audit universe entities, but the high number of features required for risk assessment render clustering a potentially unsuitable technique for direct application. The experiment with anomaly detection has shown relatively the worst performance of all approaches, with the low recall primarily attributable to the high number of features within the data, and the potentially inconsistent labelling of past data by various auditors within the same IAF.

Going back to the the early 2000s, Redmill (2004)) and including more recent research, such as Radbruch and Schiprowski (2024), studies involving risk assessments across relevant fields outside of auditing have highlighted the importance and place of subjective expertise

by subject matter experts, which hinder the ability of machine learning to build a truly objective, "gold standard" profiling of data that may be used for aiding risk assessments.

Secondly, the use of unsupervised machine learning has highlighted benefits that are not seen in the previous two rounds of experiments. For example, the use of clustering identified organisational units exhibiting similar attrition ratios within the organisation, as well as holding similar inventories of key applications, but spread across a range of business areas. These are not intuitively the starting point for a manual risk assessment, which tends to focus efforts by specific audit entity or business area.

Experiments complete, the next chapter summarises the results of these experiments, including next steps for the adoption of this work by practitioners and IAFs to support their risk assessment cycles. This primarily includes measuring value from an ML based approach to risk assessments, which forms the premise of RQ3.

Part V

EMERGING CONTRIBUTIONS

BRINGING IT ALL TOGETHER

This chapter summarises the results of experiments conducted in chapters 6 - 8, with a view to answering RQ2. The experiment outcomes are explored further in this chapter to assess value-add, the subject of RQ3.

9.1 EXPERIMENT OUTCOMES

This thesis has explored five separate approaches to the use of data and machine learning capabilities for analysing an IAF's risk assessment cycles. Research has shown not only promising results, but has put a spotlight on the requirement to carefully consider the quality of input parameters and data and the significant impact of poor quality data on the outputs and reliability of machine learning results. In this section, the findings of the experiments are consolidated to make a novel contribution to emerging literature, identify practical opportunities for innovation by IAFs, and discuss avenues for future research.

RQ2 has been evaluated through the three sets of experiments conducted in chapters 6, 7 and 8, with the results of each experiment summarised below.

1. **Supervised Machine Learning:** The experiments began by framing the risk assessment exercise as the interplay of independent and dependent variables, with risk assessment features identified by auditors acting as the independent variables and the risk assessment outcomes serving as the target or dependent variables. The problem statement was tested across multiple hypotheses, with a range of machine learning techniques employed to drive risk assessment predictions. The results were evaluated on **model accuracy**, given the objective of supplementing IAF expertise with machine learning and the requirement for the model to predict each class correctly. Whilst classifier accuracy on the validation and test datasets was not impressive when reviewed as an overall metric across all classes, the precision rate for the random forest classifier was an impressive 75% for "High risk" entities within the test dataset, which supports the case for wider practical implementation of the best performing model, as this enables the IAF to draw confidence in the model's identification of potentially high risk entities within the organisation. The supervised

ML technique was built on all features used by the IAF to perform manual risk assessments and is therefore all-inclusive in nature.

2. **Deep learning - artificial neural networks:** Following the encouraging results of supervised machine learning, the complexity of the underlying modelling approach was increased, by leveraging a deep learning technique on the same problem statement as before, treating the risk assessment outcomes as target variables. The neural network hyper-parameters were adjusted across a total of nine model setups, to assess the increase in accuracy and / or loss of information, if any. The use of artificial neural networks reduced complexity in implementation, but did not generate the same level of accurate predictions as observed with supervised machine learning techniques. The accuracy plateaued after a certain level of epochs. The experiment highlighted the importance of defining a consistent framework for risk assessments by IAFs, as the complexity and variation in subjective evaluation by IAFs impaired the neural network's accuracy to develop a holistic understanding of the dataset.
3. **Clustering:** For the final two experiments, the problem statement was reframed to consider the audit universe as a whole without labelled risk assessment results. The pre-set expectation was to identify meaningful groups of risk assessment units based on the input features used. The choice of density based spatial clustering models was influenced by the lack of spherical data, number of features and lack of normality in data distribution. Interesting and meaningful insights were seen from the clustering results, with groups of audit universe entities being formed for specific epochs displaying consistent features. However, these features could not be extrapolated specifically for the purposes of risk assessing the audit universe or formulating an audit strategy, which would be the eventual goal of the risk assessment exercise. The results of the clustering algorithms may be a useful supplement to any final practical implementation of machine learning for risk assessment, once suitable pre-processing has been completed on the input features and the data quality is assured. This aspect is discussed further over the next two chapters.
4. **Anomaly detection:** The final round of experiment involved reframing the dataset into a binary model, attempting to build three anomaly detection models to detect anomalous business units, defined for the purposes of this experiment as "high risk" units. Similar to the previous approaches, multiple epochs of this experiment were run across three datasets after considering the impact of data imbalance on the model. This was the least accurate approach of all five experiments, as the model was unable to precisely identify the positive class as expected. Similar to remarks against the third experiment (involving artificial neural networks), the experiment may

potentially be rerun to consider better quality data, or changes to imbalance treatment approaches to potentially increase model precision.

The above results highlight the use of supervised machine learning as the most suitable way forward, on the basis that the model's precision (for High risk entities) and interpretability (using SHAP) are significantly higher, and by extension, more reliable than any of the other techniques explored in this experiment. Emerging research, such as Nur Muslihatun et al. (2021), also corroborates the findings of this work. The Nur Muslihatun et al. (2021) study in particular comes closest to the findings of original research performed as part of this thesis, and supports the assertion that risk assessments in auditing would greatly benefit from the use of artificial intelligence capabilities, and in particular the use of classification techniques.

Indeed, whilst the use of artificial neural networks might allow for more powerful interpretations in the event of larger datasets; however, the "black box effect" (Dayhoff, 2001) limits the power of decision makers to make truly informed decisions; a limitation all the more acute when considering the requirement to explain audit approaches to third-party entities, such as a regulator. Similarly, unsupervised machine learning, by virtue of the unlabelled data approach, would not neatly fit into risk assessment-driven decision making; although spatial techniques may help with the identification of unknown patterns, or anomalies, within audit universe entities that may miss human supervision given the structured approach typically adopted by audit functions to perform risk assessments. This experiment for example identified a group of Investment Management audit entities as a single cluster; however this insight by itself is not relevant to a subject matter expert looking to opine on the highest risk areas of an organisation for future audit work.

9.2 KEY LIMITATIONS AND ASSUMPTIONS

The experiment outcomes are limited by a range of assumptions, which not only influence the chosen approach to model training and approach selection, but also directly impact the conclusions and recommendations for practical implementation. These are discussed in detail below.

1. **"One size fits all" dataset:** Firstly, a common dataset of independent variables without variation is required by all machine learning libraries, discussed in detail in practical implementations as well as prior research (Rashidi HH, 2023). Practically speaking, arriving at a consistent set of features truly representative of every aspect of a complex (or large) organisation is difficult to construct, and auditors might place greater weight on factors specific to parts of the organisation that may not read across to all business units. As an example, portfolio management areas within an

organisation involved in investment management activities would be risk assessed with reference to specific business metrics relevant to portfolio management, which would not read across to a back-office unit of the organisation, such as payroll or accounts payable. This analysis considers such metrics in some cases, such as customer complaint volumes, which does not apply to parts of the organisation that are not customer facing. The impact of such "bespoke" features on the machine learning models cannot be quantified reliably, and may pose a limitation to the effective interpretation of results. Numerous studies (Rashidi HH, 2023; Zainal Abidin, 2017; Arzhenovskiy S.V., 2019) have highlighted the impact of careful feature selection within risk assessments for the purposes of an effective audit risk assessment model, with the Nur Muslihatun et al. (2021) study showing that careful feature selection increased their model accuracy to 88.37% in one instance. Whilst this thesis has taken the findings of studies like Nur Muslihatun et al. (2021) forward, it is nevertheless limited to a single set of features across all audit universe entities, to enable a comparable risk assessment exercise.

2. **Feature taxonomy:** Secondly, a consistent taxonomy of features at the business unit level is assumed, which is not always available in practice and requires one-off efforts as part of the data pre-processing stage. As an example, whilst audit findings may be mapped directly to business units, production changes are mapped to applications, which are required to be mapped separately to business units to enable a holistic assessment of data for the entirety of the audit universe. At a practical level, it is quite difficult for practitioners to reliably map features on a single scale or taxonomy, and doing so would invariably require assumptions that may dilute the efficacy or accuracy of the assigned feature score to a single unit of measurement, such as an audit entity. Additionally, the maintenance of such a taxonomy, often exclusively for the purposes of risk assessments, is not a trivial task, with dedicated teams required to constantly update the taxonomy with changes arising from organisational restructures, new applications, removal of audit universe entities and new features requiring consideration, among other dynamics. For the purposes of the experiment, complete accuracy is assumed in the quality of these underlying feature mappings, the analysis of which is beyond the scope of this thesis. The quality of these mappings is subject to constant change given organisational restructures and requirements, and inaccurate data mappings will adversely impact the quality of model training, and by extension, the reliability of any machine learning outcomes.
3. **Audit Methodology:** Thirdly, the framing of the experiment as a supervised machine learning endeavour was premised on the original methodology used by the IAF. However, audit methodologies are not consistent across IAFs, with recent research

(Ege et al., 2020) finding that the global network audit firms, such as the big four professional services firms do not always enforce consistent audit methodologies across their network. Indeed, the Ege et al. (2020) study found that the role of global network methodologies in global financial reporting comparability is more pronounced across stronger investor protection jurisdictions and across jurisdictions that have adopted International Standards on Auditing. To this end, the findings of the experiment might be rendered invalid, or unusable, in an environment where the methodology adopted for risk assessments significantly differs from the one used in the example IAF. The same limitation would also be triggered in the event of a change in audit methodology within an organisation, which, albeit rare in practice, cannot be ruled out in the event of changes in team composition and leadership.

4. **Auditor skillset:** Fourthly, the experiment was carried out through careful application of statistical techniques and machine learning models that were underpinned by subject matter experience in internal audit. These are not typical skill sets currently seen within IAFs, although emerging practices and research (Nur Muslihatun et al., 2021; Li and Wang, 2024; Kahyaoglu and Aksoy, 2021) indicate that these skills are increasingly being emphasised. The uptake of machine learning for data-driven risk assessments within internal audit would be limited by the ability of auditors to understand, code and explain such approaches to senior audit leadership, as well as regulators, to enable decision making and adoption.
5. **Tooling:** Finally, the availability of appropriate audit tooling is a crucial limiting factor for many IAFs. As shown in research by Van der Nest et al. (2017), the evolution of IAF tooling and adoption of approaches involving data analytics and machine learning is overdue and traditional audit methodologies have to be refreshed. This is not only a general call to action across the board for IAFs, but also a specific limiting factor for adopting the findings of this work. IAFs seeking to implement the approaches suggested in this experiment would not only require the knowledge and skillsets for applying and interpreting the findings of this work, but also need appropriate tooling, such as machine learning libraries, on-premise databases (to download and host datasets required for supporting data collection, synthesis and exploration) and visualisation solutions for relaying the findings of machine learning models to the intended audience.

The above limitations notwithstanding, the experiment results are encouraging and offer a useful starting point for practical implementation of such capabilities, even if as a pilot initiative for large organisations. A feedback loop could help in constant tracking and processing of requisite changes to modelling approach, which could improve the quality

and reliability of machine learning results over time, helping IAFs unlock significant efficiency gains and value from machine learning.

9.3 CONTRIBUTIONS TO PRACTICE & SUGGESTED TECHNIQUES FOR IMPLEMENTATION

Whilst the experiment results above highlight supervised machine learning as the most suitable option, these results are reflective only of the complexity of the dataset used for modelling, and may not be extrapolated across all IAFs and audit universe populations as a standard approach to machine learning for risk assessments. Indeed any of the other approaches described in this thesis may be potentially suitable avenues for risk assessment and a similar comparative evaluation should be performed to identify the most suitable approach across each IAF. To this end, specific contributions from this thesis are discussed below, focusing on the perspective of practical implementation of the techniques used in the research. In this section, workarounds for each of the limitations stated in the previous section are also offered to enable practical implementation by a range of IAFs within the industry.

1. **Framing the approach:** To begin with, any machine learning exercise has to commence with the IAF carefully considering the overall exercise objectives, in light of extant audit methodology and availability of data, whilst also assessing the overall department culture and appetite for innovation. The overall approach must be clearly defined and approved by the audit leadership prior to any practical implementation. Consistency of audit methodology, and specifically for risk assessments, is a cornerstone of effective machine learning based delivery, as the above experiment results have highlighted. Where an IAF clearly identifies a risk assessment methodology, the possibility of demarcating the collected dataset into features and target variables, similar to this experiment, and approaches recently undertaken (Nur Muslihatun et al., 2021; Özbaltan, 2024) would enable a classification-based approach to risk assessment.

Consistency in applying risk assessment methodology, and objectivity in evaluating features, might also identify opportunities for adopting the artificial neural network approach, as shown in Nur Muslihatun et al. (2021), whose research identified a 90%+ accuracy rate with ANNs, which introduces significant confidence in machine learning driven approaches. Additionally, research by Nonnenmacher and Gomez (2021) and Bakumenko and Elragal (2022) has discussed the use of anomaly detection techniques for auditing, a practice that can gain widespread adoption among IAFs who have basic but evolving risk assessment methodologies. Indeed, the lack of a structured or pre-defined approach to risk assessment can lend itself suitably to

profiling techniques, where an unsupervised machine learning model similar to those tested in this experiment could highlight unknown anomalies and patterns in the data, enabling IAFs to take a bespoke approach to auditing for each audit cycle. However, a key risk with such an implementation approach would be the comparability of the organisation's overall risk profile, which is measurable with other approaches.

2. **Curating the dataset:** Secondly, a very wide range of continuous and categorical features were consumed in this thesis for building various supervised and unsupervised models. An IAF should consider adopting a workshop-led approach to identifying, and curating, the requisite datasets for informing a machine learning based approach to risk assessments. As mentioned above, a key limitation with practical implementation is the "one size fits all" approach taken by machine learning. To address this problem, especially where an IAF deems it inappropriate or impractical to identify a common set of variables to develop a machine learning methodology, the IAF could consider breaking up the overall audit universe into smaller segments, such as by business area, and treat each segment as a standalone audit universe, with bespoke features for each segment driving the overall assessment of residual risk. In the previous example of Investment Management, a subset of specific features for Investment Management, such as credit or liquidity risk variables, may not apply to the risk assessment question for central / group functions, such as accounts payable or payroll. This would enable more targeted approaches to assessing residual risk requiring audit attention; however, the use of different feature sets would hinder overall comparability. This can be resolved using a bespoke / judgement based methodology to obtain an overarching view of residual risk for the organisation. As part of dataset curation, care should also be taken to ensure that the various features collected (for individual organisational areas, or for the entirety of the audit universe) are mapped to a single taxonomy to facilitate the use of all features for model training purposes.
3. **Skillsets:** Thirdly, the IAFs should invest substantially in upskilling the function, bringing in much needed expertise in data analytics, engineering and machine learning, whilst comprehensively engaging all auditors within the function to drive a targeted, machine learning driven approach to risk assessments. This implementation attribute will not be straightforward as it will require considered efforts, budgets and time to implement effectively. However, as shown in (Kahyaoglu and Aksoy, 2021; Nur Muslihatun et al., 2021; Shivram, 2024; Van der Nest et al., 2017), the use of advanced analytics and machine learning capabilities is not a nice to have but a necessity for IAFs, given the growing importance of technology and innovation for

assurance. IAFs could consider including the achievement of one or more machine learning skills as part of individual performance goals to encourage uptake.

4. **Tooling:** Additionally, the availability of appropriate audit tooling is critical to implement machine learning capabilities. To this end, IAFs should consider a staggered approach - beginning in year one with data collection and curation, and gradually moving to more advanced levels with data science goals over a two to three year horizon. Successful implementation of data science techniques would be contingent on not only the right tools but effective implementation and investment, which requires senior management sponsorship. Combined with data engineering, investment in tooling is a key implementation attribute that should be considered early on to drive innovation for auditing.
5. **Model Drift:** Finally, the any practical implementation of machine learning, especially supervised machine learning, should consider the possibility of model drift to ensure timely correction and / or refresh of underlying training data and modelling algorithms. Over time, the spectrum of features and their associated risk levels are likely to evolve, resulting in originally designed models being no longer fit for purpose. Therefore, the model implementation is not a "one and done" exercise, but rather a continuous improvement endeavour that brings forth the most relevant risks to an organisation for each risk assessment cycle, whilst generating triggers in the event of a significant shift in the volume of any key features so as to enable timely identification of new risk factors requiring audit's attention.

9.4 CONTRIBUTIONS TO LITERATURE

This doctoral thesis has made a number of contributions to extant literature, firstly by setting out an empirical, evidence-based approach to implementing machine learning capabilities for risk assessments across the internal audit lifecycle. When reviewed alongside recent research by Özbaltan (2024) and Nur Muslihatun et al. (2021), this thesis brings out a range of techniques, applied on real-world data, to demonstrate the power of machine learning capabilities on the foundations of internal auditing, along with suggestions for practical implementation. The key themes highlighted by the literature review originally performed are now revisited, to highlight the emerging contributions of this work:

EMPIRICAL ANALYSIS:

Firstly, the literature review shows that the uptake of innovation on the internal audit lifecycle using DA is generally limited by a) awareness and b) lack of training opportunities outside the “big four accounting firms” (Islam and Stafford, 2022). Secondly, research shows that a range of factors impact the effectiveness of DA in auditing (Jacky and Sulaiman, 2022), such as availability of budgets and compliance with regulatory requirements. In the experiment, the usage of open source machine learning libraries, underpinned with subject matter knowledge in auditing, has enabled the empirical analysis of datasets for risk assessments, which appears quite limited to a select few authors such as Özbaltan (2024) and Nur Muslihatun et al. (2021), whose work appears in recent research. As the literature review showed that the use of empirical data to analyse the effectiveness of innovation techniques to improve internal audit processes, and specifically, internal audit risk assessments, is quite limited, this thesis makes a valuable contribution to the field by suggesting methods and practical techniques for implementation. To this end, experiments such as this thesis help address concerns expressed by authors such as Gepp and Smith (2018) and De Santis and D’Onza (2021), who reaffirmed the positive influence of DA on audit quality whilst calling for more empirical studies on the utilisation of DA within auditing.

QUALITATIVE IMPACT:

Prior literature examining the qualitative impact of audit data analytics appears to take the approach of focusing on individual aspects of the internal audit lifecycle, as set out in figure 2 above, and is largely motivated by two drivers – firstly, the use of analytics to enable greater coverage of transactions subject to auditing (typically referred to as “full population testing”), as opposed to basing audit conclusions on sample-based test work (“audit sampling”) (Chiu and Jans, 2017; Huang et al., 2022; Nonnenmacher and Gomez, 2021) and secondly, the expectation that technology-based innovation, such as audit data analytics, could allay pressure on internal auditors (Cardinaels et al., 2021; Eulerich et al., 2020, 2022).

By leveraging machine learning capabilities for risk assessments, rather than full population testing or efficiency gains (the common quality drivers for audit innovation), this thesis offers a fresh dimension of quality for auditors, and in doing so, a new avenue of research literature that has not been made available before. The qualitative impact of machine learning for foundational elements of the audit lifecycle could directly influence the overall efficacy of the IAF, an aspect that should be considered for future research avenues by academics, as well as practical implementation by practitioners.

SCALABLE INNOVATION DELIVERY:

Existing academic literature has also highlighted that there are multiple elements feeding into the process of innovation in the audit domain (Krieger et al., 2021), however, research also suggests that the prior audit literature has not significantly contributed to an overall appreciation of the key factors that contribute to scalable innovation within audit functions (Kotb and Halabi, 2020; Rawashdeh, 2022). The research leveraged a range of techniques and features to generate value through an objective, data-driven risk assessment exercise that facilitates scalability. The use of supervised machine learning, in particular, would require supervision of model drift as explained in the prior section, an example of how a "once trained" model can be suitably scaled and kept fit for purpose in a changing environment. This enables the innovation delivery to be future proof, adding to an IAF's arsenal of skills and techniques for delivering value through the audit function.

REGULATORY EXPECTATIONS:

Prior literature has recognised the important role played by regulators in the adoption of advanced analytics techniques and emerging technologies (Warren and Smith, 2006; Chatzigiannis and Konstantinos, 2021; Wellmeyer, 2022). In particular this focus has revolved around the interpretability of innovation, as shown in research by Wellmeyer (2022), Ribeiro et al. (2016) and Salih et al. (n.d.). At the same time, regulators recognise the need for innovation, a sentiment taken almost as a given in recent academic literature in the field (Bibler and Gold, 2023).

The use of explainability techniques such as SHAP, offers a useful solution to the important consideration of explainable innovation, and therefore a fresh literary dimension that shows how regulatory expectations can be met by an audit function that invests significantly in innovation.

The use of innovation within the risk assessment cycle brings another further benefit from the perspective of satisfying regulatory expectations, one which could benefit from the work presented in the research and taken forward by academics and practitioners alike. By identifying key and emerging risk factors through a data driven approach such as the techniques discussed in this work, the audit function can demonstrate significant value to regulatory stakeholders. This is because unlike manually performed risk assessments, machine learning based risk assessments would identify triggers and outcomes a lot more efficiently, highlighting areas of focus and enable timely audit intervention. This could be a valuable support for an audit function that bases its risk assessment techniques on machine learning.

As shown in the literature study, Warren and Smith (2006) pushed their working paper arguing for the need for continuous auditing as part of the internal audit toolkit to address fraud and offer a modern definition of continuous auditing in the light of disruptive technology. In doing so, they assert that “Once-a-year reviews”, typical of internal audit functions, are often insufficient. This view is supported by Omid Firoouzi and Wang (2020), who assert that a “silo approach” to auditing fails in the light of rising artificial intelligence (AI) capabilities. This thesis directly addresses these challenges by presenting a practical solution for implementation, which should be studied and taken forward by the research community as well as IAFs to generate further interest in this all-important subject area.

9.5 MEASURING VALUE

The final research question framed at the start of this work, RQ₃ is now evaluated. Prior research has established that an IAF has a critical role in identifying and assessing the risks of any organisation (Yousif and Mohamed, 2022). Given the primary purpose of risk assessments is to drive the suitability of the audit plan (IIA, 2023a,b) and identify the most critical / high-risk areas of the organisation, two key drivers emerge for measuring the value-add of any risk assessment cycle, based on practical business benefits observed from practical implementation of machine learning techniques explored in this thesis:

1. **Objectivity:** Risk assessments by IAFs are an inherently subjective exercise, as shown by prior research (Redmill, 2004; Porcuna-Enguix et al., 2021; Demirović and Proho, 2021). Therefore, higher objectivity in the identification of high risk areas of the organisation warranting audit focus is a key value measurement metric for any IAF. Machine learning introduces objectivity by greatly reducing any human bias in the identification of high risk labels across the organisation, but relies on a robust set of features and consistent evaluation of base data by the entirety of the audit function, both of which are hard to achieve in practice (Kahyaoglu and Aksoy, 2021). However, over time, the consistency of manual risk assessment outcomes can be sufficiently improved so as to render it a meaningful training base for a machine learning model, especially as articulated from the findings of the first experiment involving supervised machine learning. The IAF at the sponsoring organisation acknowledged the impact of meaningful data-driven challenge to pre-existing manual risk assessment labels based on the final classification model, which led to practical recalibration of the IAF’s audit plan. This value attribute continues to be observed at the sponsoring organisation, despite the limitation that the training dataset requires continuous improvement to ensure continued value-add and reliability.

2. **Efficiency:** There is very limited research in extant literature that may enable meaningful benchmarking of the time expended to perform a cycle of risk assessments by auditors, however, research by Eulerich and Eulerich (2020) has attempted to define, measure and communicate the value of internal audit as a whole. The Eulerich and Eulerich (2020) analysis highlights timeliness of audit finding disclosures, efficiency of the audit function overall and identification of key insights as three drivers to be considered by audit functions as key performance indicators. These could be read-across the risk assessment landscape as well, with efficiency being understood as timely identification of potentially new high risk entities within an organisation and less reliance on manual evaluation of input features. A suitable machine learning approach helps meet this expectation by being an on-demand repository of insights and analysis. From the initial implementation of a supervised machine learning technique ultimately implemented at the sponsoring organisation, a significant reduction in manual hours expended (approximately 400 hours each quarter) on risk assessments was observed, which continues to be a key driver for promoting the continued usage of such techniques for periodic risk assessments.

Based on the above and the results of practical implementation observed, future implementations in alternative professional settings could be tested against the above two objectives to help evaluate the efficacy and results of machine learning based approaches against manual risk assessments. Whilst this research has found supervised machine learning to be the most suitable implementation approach at the sponsoring organisation, other audit functions with varying team sizes, features or dataset populations might arrive at different outcomes and results should be based on a holistic evaluation of all proposed techniques and the above two-point framework prior to implementation. Value measurement, especially for efficiency, should consider the impact of one-off efforts required to perform the following steps:

1. **Collect underlying data:** This research has demonstrated the use of 31 separate variables collected from across a medium-sized organisation to experiment with on multiple machine learning techniques, with data collection requiring over 200 hours of effort in total. This figure may be considerably higher for other organisations depending on the complexity of the underlying audit universe, and any efficiency gains should make an allowance for data collection, including pre-processing, as part of result or value evaluation.
2. **Data quality assessment:** Once collected, one-off efforts are required to pre-process and clean the underlying data before applying suitable modelling techniques to generate insights.

3. **Model design:** Machine learning development and implementation requires specific skillsets which are not conventionally found in IAFs, although recent developments and research (Huang et al., 2022; Shivram, 2024; Wassie and Lakatos, 2024) highlight the importance of IAFs to upskill in data science and machine learning skills. Where unavailable within the function, the engagement of third-party experts or suppliers is likely to impact the time and cost required to build a supplementary machine learning-based approach to risk assessments.

Ultimately, any value measurement exercise will be specific to each organisation, but leveraging the findings of this research will likely support a holistic and considered approach to practical implementation of a suitable machine learning technique, elevating the value-add from the internal audit function through risk assessments.

CONCLUDING REMARKS

10.1 RESEARCH QUESTIONS REVISITED

Having completed experiments across a range of ML paradigms, the research questions set out at the end of Chapter 1 can now be evaluated in-turn.

1. RQ1 has been answered through the literature review performed in section 2.2. The literature review shows that a range of analyses have been undertaken by the academic community in the area of internal auditing, as well as the role played by machine learning to support data driven risk assessments in other fields. However, there are little references that show the interplay between machine learning and internal audit, and specifically, risk assessments within the audit lifecycle. The subjective nature of risk assessments lends itself to a machine learning exercise; however, the work performed as part of the SLR has shown that existing academic literature has yet to fully explore the impact, ideal approach, and considerations for the use of ML and DA capabilities on the audit lifecycle.
2. The work performed in section 2.3, following on from the SLR, was used as the basis for evaluating RQ2, which has been answered in stages, across chapters 6, 7 and 8. The work showed that machine learning capabilities can indeed take different forms for practical implementation in an internal audit setting, with differing considerations dictating the most appropriate approach for a particular IAF. For example, the use of an artificial neural network, in chapter 7 was intended to strengthen machine learning prediction quality based on the ANN's ability to draw out non-linear relationships within the underlying dataset. However, this was not found to be the case, as the ANN performed worse than simpler supervised machine learning models that were built in chapter 6. This is perhaps attributable to the limited dataset at the sponsoring organisation; however, a much larger or more complex IAF could effectively gather the large volumes of data for completing a risk assessment exercise, which could drive a superior ANN accuracy. In effect, there is no single best type of machine learning approach that an IAF could take for solving a risk assessment exercise, and the choice of approach would be driven by the IAF's ability to continuously collect,

and curate, the relevant features driving risk assessments, as demonstrated in chapters 4 and 5.

3. Combined with the observations of the SLR as detailed in section 2.2, there is very little comparable literature that can be used to directly answer RQ3. Based on the results of the experiments performed in chapters 6, 7 and 8, a proposed value measurement framework was constructed in section 9.5, with a view to evaluating RQ3. In summary, any value measurement framework for a machine learning algorithm that is deployed to model risk assessments should be based on efficiency gains as well as objectivity, which can offer robust and meaningful challenge to the IAF to support risk assessment outcomes. Given that risk assessment methodologies can vary from one organisation to the other, and given the inherently subjective nature of risk assessment methodologies in play across IAFs, a prescriptive value measurement model for evaluating machine learning effectiveness cannot be constructed. Similar to the answer above for RQ2, the choice of the appropriate value measurement framework would be driven by the relevant IAF methodology for risk assessment.

10.2 FURTHER RESEARCH AVENUES

This research concludes by suggesting the following avenues for further academic and practical exploration.

1. **Comparative Analysis:** Future research efforts could consider a comprehensive comparative analysis of the various machine learning models developed in this research work. Following a more comprehensive examination of the strengths, weaknesses, and performance metrics of each model, researchers and practitioners may be able to identify the most suitable approaches for specific risk assessment scenarios that could be applicable specifically for individual organisations or industries.
2. **Auditor skills and competencies:** As internal audit functions increasingly adopt machine learning, it is important to develop associated skills and competencies within audit teams. This includes a strong understanding of the principles underpinning machine learning models, including data science, statistical analysis, and the approach adopted by various types of supervised and unsupervised techniques. Of perhaps greater significance is the need to interpret and communicate the results of such models to stakeholders effectively. However, as highlighted in section 2, there is limited literature exploring the development of auditor skills and competencies in the realm of machine learning and artificial intelligence.

3. **Model Maintenance:** The implementation of advanced machine learning models within internal audit requires not only the requisite skills and competencies, but also ongoing maintenance and updates following implementation. Additionally, as the organization's data evolves and new risks emerge, model retraining becomes a pre-requisite to assure ongoing effectiveness and reliability. This necessitates a robust data management infrastructure and a dedicated team responsible for model maintenance and development. The nascent nature of machine learning-driven innovation in IAFs has resulted in limited research work on the patterns of value arising from novel approaches in the fields of governance, risk and compliance. This is a particularly important field for future research, as the lack of a defined value framework specific to machine learning-based approaches and artificial intelligence can hinder uptake by IAFs, as discussed in existing research by Winter et al. (2021).

By exploring the application of various machine learning models to enhance internal audit risk assessments, this research has demonstrated the potential of innovative capabilities to identify hidden patterns, anomalies, and emerging risks within organisational data. In doing so, this research contributes to the growing body of academic literature on the topic of artificial intelligence and innovation in the fields of internal audit and compliance and drive continuous innovation for increased efficiency and value-add for IAFs.

Part VI

BIBLIOGRAPHY

BIBLIOGRAPHY

- Al Farizi, W. S., Hidayah, I. and Rizal, M. N. (2021), Isolation forest based anomaly detection: A systematic literature review, in '2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)', pp. 118–122.
- Alamri, A. H. (2022), 'Application of machine learning to stress corrosion cracking risk assessment', *Egyptian Journal of Petroleum* **31**(4), 11–21.
URL: <https://www.sciencedirect.com/science/article/pii/S1110062122000642>
- Alles, M. and Gray, G. L. (2016), 'Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors', *International Journal of Accounting Information Systems* **22**, 44–59.
- Amayri, O. and Bouguila, N. (2013), 'On online high-dimensional spherical data clustering and feature selection', *Engineering Applications of Artificial Intelligence* **26**(4), 1386–1398.
URL: <https://www.sciencedirect.com/science/article/pii/S0952197612002709>
- Aning, S. and Przybyła-Kasperek, M. (2022), 'Comparative study of twoing and entropy criterion for decision tree classification of dispersed data', *Procedia Computer Science* **207**, 2434–2443. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022.
URL: <https://www.sciencedirect.com/science/article/pii/S1877050922011905>
- Anon, E., Nasution, A., Yahya, I. and Atmanegara, A. (2020), 'The role of risk based internal audit in improving audit quality', *INTERNATIONAL JOURNAL OF MANAGEMENT* **11**.
- Arzhenovskiy S.V., Bakhteev A.V., S. T. H. N. (2019), 'Audit risk assessment model', *International Journal of Economics and Business Administration* **7**, 74–85.
- Auer, R. (2022), 'Embedded supervision: How to build regulation into decentralised finance', *CESifo Working Paper No. 9771* .
- Aven, T. (2016), 'Risk assessment and risk management: Review of recent advances on their foundation', *European Journal of Operational Research* **253**(1), 1–13.
URL: <https://www.sciencedirect.com/science/article/pii/S0377221715011479>
- Awuah, B., Onumah, J. M. and Duho, K. C. T. (2021), 'Information technology adoption within internal auditing in ghana: Empirical analysis', *Dataking Working Paper Series No. WP2021-04-04* .

- Bakumenko, A. and Elragal, A. (2022), 'Detecting anomalies in financial data using machine learning algorithms', *Systems* **10**(5).
URL: <https://www.mdpi.com/2079-8954/10/5/130>
- Bao, W., Lianju, N. and Yue, K. (2019), 'Integration of unsupervised and supervised machine learning algorithms for credit risk assessment', *Expert Systems with Applications* **128**, 301–315.
URL: <https://www.sciencedirect.com/science/article/pii/S0957417419301472>
- Barr-Pulliam, D., Brazel, J. F., McCallen, J. and Walker, K. (2020), 'Data analytics and skeptical actions: The countervailing effects of false positives and consistent rewards for skepticism'.
- Barr-Pulliam, D., Brown-Liburd, H. L. and Sanderson, K. A. (2022), 'The Effects of the Internal Control Opinion and Use of Audit Data Analytics on Perceptions of Audit Quality, Assurance, and Auditor Negligence', *Auditing: A Journal of Practice Theory* **41**, 25–48.
- Bauguess, S. W. (2017), 'The role of big data, machine learning, and ai in assessing risks: A regulatory perspective', *Machine Learning, and AI in Assessing Risks: A Regulatory Perspective* (June 21, 2017). SEC Keynote Address: OpRisk North America .
- Bello, O. A. (2023), 'Machine learning algorithms for credit risk assessment: an economic and financial analysis', *International Journal of Management* **10**(1), 109–133.
- Berk, R., Berk, D. and Drougas, D. (2019), *Machine learning risk assessments in criminal justice settings*, Springer.
- Betti, N. and Sarens, G. (2021), 'Understanding the internal audit function in a digitalised business environment.', *Journal of Accounting Organizational Change* **17**, 197–216.
- Bhatia, B., Michalski, J. and Shebin, J. (2021), 'Comparison of the ease of use and accuracy of two machine learning algorithms – forestry case study', *Journal of Emerging Investigators* .
- Bibler, Sara, C. T. C. M. and Gold, A. (2023), 'Thinking Outside of the Box: Engaging Auditors' Innovation Mindset to Improve Auditors' Fraud Actions in a Data-Analytic Environment.', *Foundation for Auditing Research Working Paper 2023/03-13* pp. 1–47.
URL: <https://foundationforauditingresearch.org/files/papers/2023-03-13-far-wp-gold2019b02-full.pdf>
- Blokdijk, H., Driehuisen, F., Simunic, D. A. and Stein, M. T. (2003), 'Determinants of the mix of audit procedures: Key factors that cause auditors to change what they do.', *Sauder School of Business, The University of British Columbia, Forthcoming* .

- Bonrath, A. and Eulerich, M. (2023), 'Internal auditing's role in preventing and detecting fraud: An empirical analysis', *International Journal of Auditing (Forthcoming)* .
- Boyles, M. (2022), 'Innovation in business: What it is why it's so important.', *Harvard Business School Online Business Insights Blog*. .
URL: <https://online.hbs.edu/blog/post/importance-of-innovation-in-business>
- Brennan, G. and Teeter, R. A. (2010), 'Aiding the audit: Using the it audit as a springboard for continuous controls monitoring'.
- Bruce, S. (2018), 'Internal audit: Understanding the risk assessment of the internal audit universe', *The Institute of Chartered Accountants of Scotland* .
URL: <https://www.icas.com/professional-resources/audit-and-assurance/internal-audit/internal-audit-understanding-the-risk-assessment-of-the-internal-audit-universe>
- Butkè, K. and Dagilienè, L. (2022), 'The Interplay between Traditional and Big Data Analytic Tools in Financial Audit Procedures', *SSRN Electronic Journal* pp. 5, 17–19.
- Bäßler, T. and Eulerich, M. (2023), 'Predictive process monitoring for internal audit: Forecasting payment punctuality from the perspective of the three lines model', pp. 6–7, 14–19.
- Cabello-Solorzano, K., Araujo, I., Peña Cubillos, M., Correia, L. and Tallón-Ballesteros, A. J. (2023), *The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis*, pp. 344–353.
- Cagle, M. (2020), 'A Mapping Analysis of Blockchain Applications Within the Field of Auditing.', *The World of Accounting Science* **22**, 695–724.
URL: <https://ssrn.com/abstract=3878577>
- Cao, S. S., Cong, L. and Yang, B. (2019), 'Financial reporting and blockchains: Audit pricing, misstatements, and regulation'.
- Cardinaels, E., Eulerich, M. and Salimi Sofla, A. (2021), 'Data Analytics, Pressure, and Self-Determination: Experimental Evidence from Internal Auditors', *SSRN Electronic Journal* pp. 3–7, 22–24.
- Carvalho, D. V., Pereira, E. M. and Cardoso, J. S. (2019), 'Machine learning interpretability: A survey on methods and metrics', *Electronics* **8**(8).
URL: <https://www.mdpi.com/2079-9292/8/8/832>
- Casanova, R., Hsu, F.-C., Sink, K. M., Rapp, S. R., Williamson, J. D., Resnick, S. M., Espeland, M. A. and Initiative, A. D. N. (2013), 'Alzheimer's disease risk assessment using large-scale machine learning methods', *PloS one* **8**(11).

- Cassell, C. A., Dearden, S. M., Rosser, D. M. and Shipman, J. E. (2022), 'Confirmation Bias and Auditor Risk Assessments: Archival Evidence', *AUDITING: A Journal of Practice Theory* **41**(3), 67–93.
URL: <https://doi.org/10.2308/AJPT-2020-035>
- Chatzigiannis, Panagiotis, B. F. and Konstantinos, C. (2021), 'Applied cryptography and network security', *19th International Conference, ACNS 2021, Kamakura, Japan, June 21–24, Proceedings, Part II* pp. 311–227.
- Cheng, Z., Zou, C. and Dong, J. (2019), Outlier detection using isolation forest and local outlier factor, in 'Proceedings of the Conference on Research in Adaptive and Convergent Systems', RACS '19, Association for Computing Machinery, New York, NY, USA, p. 161–168.
URL: <https://doi.org/10.1145/3338840.3355641>
- Chiang, C. C., Agnew, K. S. and Korol, K. (2021), 'Knowledge and skills essential for auditors in the age of big data—the early evidence from a survey.', *International Journal of Organizational Innovation* **13**(4).
- Chiu, T. and Jans, M. J. (2017), 'Process mining of event logs: A case study evaluating internal control effectiveness', pp. 16–22.
- Coetzee, G. P. P. (2010), 'A risk-based audit model for internal audit engagements'.
- Dagilienė, L. and Klovienė, L. (2019), 'Motivation to use big data and big data analytics in external auditing', *Managerial Auditing Journal* **34**, 750–782.
- Dai, J. and Vasarhelyi, M. A. (2017), 'Toward Blockchain-Based Accounting and Assurance', *Journal of Information Systems* **31**(3), 5–21.
URL: <https://doi.org/10.2308/isys-51804>
- Dayhoff, J. E. (2001), 'Artificial neural networks: opening the black box.', *National Library of Medicine: National Center for Biotechnology Information* .
- de Roux, D., Perez, B., Moreno, A., Villamil, M. d. P. and Figueroa, C. (2018), Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach, in 'Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', KDD '18, Association for Computing Machinery, New York, NY, USA, p. 215–222.
URL: <https://doi.org/10.1145/3219819.3219878>
- De Santis, F. and D'Onza, G. (2021), 'Big data and data analytics in auditing: in search of legitimacy', *Meditari Accountancy Research* **29**, 1088–1112.

- Deloitte (n.d.), 'Perspectives - adopting robotic process automation in internal audit'.
URL: <https://www2.deloitte.com/us/en/pages/risk/articles/internal-audit-robotic-process-automation-adoption.html>
- Demirkaya, A., Chen, J. and Oymak, S. (2020), Exploring the role of loss functions in multiclass classification, in '2020 54th Annual Conference on Information Sciences and Systems (CISS)', pp. 1–5.
- Demirović, Lejla, I.-K. and Proho, M. (2021), 'Internal audit risk assessment in the function of fraud detection', *Journal of Forensic Accounting Profession* **1**(1), 35–49.
URL: <https://intapi.sciendo.com/pdf/10.2478/jfap-2021-0003>
- Deng, D. (2020), Dbscan clustering algorithm based on density, in '2020 7th International Forum on Electrical Engineering and Automation (IFEAA)', pp. 949–953.
- Ding, Z. and Fei, M. (2013), 'An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window', *IFAC Proceedings Volumes* **46**(20), 12–17. 3rd IFAC Conference on Intelligent Control and Automation Science ICONS 2013.
URL: <https://www.sciencedirect.com/science/article/pii/S1474667016314999>
- Ditkaew, K. and Suttipun, M. (2023), 'The impact of audit data analytics on audit quality and audit review continuity in thailand', *Asian Journal of Accounting Research* pp. 269–278.
- Dixit, R. R. (2021), 'Risk assessment for hospital readmissions: Insights from machine learning algorithms', *Sage Science Review of Applied Machine Learning* **4**(2), 1–15.
- Doyle, J., Ge, W. and McVay, S. (2007), 'Determinants of weaknesses in internal control over financial reporting', *Journal of Accounting and Economics* **44**(1), 193–223. Conference Issue on Corporate Governance: Financial Reporting, Internal Control, and Auditing.
URL: <https://www.sciencedirect.com/science/article/pii/S0165410106000905>
- Dreiseitl, S. and Ohno-Machado, L. (2002), 'Logistic regression and artificial neural network classification models: a methodology review', *Journal of Biomedical Informatics* **35**(5), 352–359.
URL: <https://www.sciencedirect.com/science/article/pii/S1532046403000340>
- Drewnik, M. and Pasternak-Winiarski, Z. (2017), Svm kernel configuration and optimization for the handwritten digit recognition, in 'Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16', Springer, pp. 87–98.
- Dubey, S. R., Singh, S. K. and Chaudhuri, B. B. (2022), 'Activation functions in deep learning: A comprehensive survey and benchmark'.
URL: <https://arxiv.org/abs/2109.14545>

- Ege, M. S., Kim, Y. H. and Wang, D. (2020), 'Do Global Audit Firm Networks Apply Consistent Audit Methodologies across Jurisdictions? Evidence from Financial Reporting Comparability', *The Accounting Review* **95**(6), 151–179.
URL: <https://doi.org/10.2308/tar-2018-0294>
- Eilifsen, A., Kinserdal, F., Messier Jr, W. F. and McKee, T. (2020), 'An Exploratory Study into the Use of Audit Data Analytics on Audit Engagements', *American Accounting Association* pp. 1–57.
- Emett, S. A., Eulerich, M., Lipinski, E., Prien, N. and Wood, D. A. (2023), 'Leveraging chatgpt for enhancing the internal audit process – a real-world example from a large multinational company', pp. 1–27.
- Esmaelnejad, J., Habibi, J. and Yeganeh, S. H. (2010), A novel method to find appropriate ϵ for dbscan, in 'Intelligent Information and Database Systems: Second International Conference, ACIIDS, Hue City, Vietnam, March 24-26, 2010. Proceedings, Part I 2', Springer, pp. 93–102.
- Eulerich, A. and Eulerich, M. (2020), 'What Is the Value of Internal Auditing? – A Literature Review on Qualitative and Quantitative Perspectives.', *Maandblad Voor Accountancy en Bedrijfseconomie* **94**, 83–92.
- Eulerich, M., Georgi, C. and Schmidt, A. (2020), 'Continuous Auditing and Risk-Based Audit Planning', *SSRN Electronic Journal* pp. 4, 7, 10–12.
- Eulerich, M., Masli, A., Pickerd, J. S. and Wood, D. A. (2022), 'The Impact of Audit Technology on Audit Task Outcomes: Evidence for Technology-Based Audit Techniques', *SSRN Electronic Journal* pp. 3–7, 16–20.
- Fadul, A. M. A. (2023), Anomaly Detection based on Isolation Forest and Local Outlier Factor, PhD thesis, Africa University.
- Fukukawa, H. and Mock, T. J. (n.d.), 'Audit risk assessments using belief versus probability'.
- Fuller, S. and Markelevich, A. J. (2019), 'Should Accountants Care About Blockchain?', *SSRN Electronic Journal* pp. 1–34.
- Gao, C. X., Dwyer, D., Zhu, Y., Smith, C. L., Du, L., Filia, K. M., Bayer, J., Menssink, J. M., Wang, T., Bergmeir, C., Wood, S. and Cotton, S. M. (2023), 'An overview of clustering methods with guidelines for application in mental health research', *Psychiatry Research* **327**, 115265.
URL: <https://www.sciencedirect.com/science/article/pii/S0165178123002159>

- Gao, R., Huang, S. and Wang, R. (2020), 'Data analytics and audit quality', *Singapore Management University School of Accountancy Research Paper No. 2022-151* pp. 2–3, 14–15.
- Gepp, A., L. M. O. T. and Smith, T. (2018), 'Big data techniques in auditing research and practice: Current trends and future opportunities', *Journal of Accounting Literature* **40**, 102–115.
- Gholamy, A., Kreinovich, V. and Kosheleva, O. (2018), 'Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation', *Int. J. Intell. Technol. Appl. Stat* **11**(2), 105–111.
- Glette-Iversen, I., Flage, R. and Aven, T. (2023), 'Extending and improving current frameworks for risk management and decision-making: A new approach for incorporating dynamic aspects of risk and uncertainty', *Safety Science* **168**, 106317.
URL: <https://www.sciencedirect.com/science/article/pii/S092575352300259X>
- Gupta, S. and Gupta, A. (2019), 'Dealing with noise problem in machine learning data-sets: A systematic review', *Procedia Computer Science* **161**, 466–474. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.
URL: <https://www.sciencedirect.com/science/article/pii/S1877050919318575>
- Hammersley, J. S. (2011), 'A review and model of auditor judgments in fraud-related planning tasks', *A Journal of Practice and Theory* .
- Hampton, C. and Stratopoulos, T. C. (2016), 'Audit Data Analytics Use: An Exploratory Analysis', *SSRN Electronic Journal* pp. 3, 16–17.
- Hasan, R. and Chu, C. (2022), 'Noise in datasets: What are the impacts on classification performance? [noise in datasets: What are the impacts on classification performance?]', *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods* .
URL: <https://par.nsf.gov/biblio/10319530>
- Hegde, J. and Rokseth, B. (2020), 'Applications of machine learning methods for engineering risk assessment—a review', *Safety science* **122**, 104492.
- Hejazi, M. and Singh, Y. P. (2013), 'One-class support vector machines approach to anomaly detection', *Applied Artificial Intelligence* **27**(5), 351–366.
- Hellström, T., Dignum, V. and Bensch, S. (2020), Bias in machine learning - what is it good for?, in 'NeHuAI@ECAI'.
URL: <https://api.semanticscholar.org/CorpusID:221716464>

- Honegger, M. (2018), 'Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions', *arXiv preprint arXiv:1808.05054* .
- Houser, K. and Sanders, D. (2018), 'The use of big data analytics by the irs: What tax practitioners need to know', *Journal of Taxation* **128**, 4–5.
URL: <https://ssrn.com/abstract=3120741>
- Huang, F., No, W. G., Vasarhelyi, M. and Yan, Z. (2022), 'Audit Data Analytics, Machine Learning, and Full Population Testing', *KeAi: The Journal of Finance Data Science* pp. 1–16.
- Igiri, C. P., Anyama, O. U. and Silas, A. I. (2015), 'Effect of learning rate on artificial neural network in machine learning'.
- IIA (2020), 'The iia's three lines model: An update of the three lines of defence', pp. 3–6.
URL: <https://www.theiia.org/globalassets/documents/resources/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense-july-2020/three-lines-model-updated-english.pdf>
- IIA (2023a), 'Annual internal audit coverage plans', pp. 2–4.
URL: <https://www.iaa.org.uk/resources/managing-internal-audit/how-to-prepare-annual-internal-audit-coverage-plans/>
- IIA (2023b), 'Risk based internal auditing', pp. 1–3.
URL: <https://www.iaa.org.uk/resources/risk-management/risk-based-internal-auditing/doing-a-risk-based-audit/>
- IIA (n.d.a), 'Position paper: Uk corporate governance code 2024 - what does it mean for internal audit?'
- IIA (n.d.b), 'Supplemental guidance practice guide: Developing a risk-based internal audit plan'.
- Imoniana, J. O. and Gartner, I. R. (2008), 'Towards a multi-criteria approach to corporate auditing risk assessment in brazilian context'.
- Irwansyah, I. and Zega, W. (2023), 'The influence of good corporate governance, the role of internal audit, the effectiveness of internal controls and the appropriate of compensation on fraud trends', *International Business and Accounting Research Journal* **7**.
- Islam, S. and Stafford, T. (2022), 'Factors associated with the adoption of data analytics by internal audit function', *Managerial Auditing Journal* **37**, 193–223.
- Jacky, Y. and Sulaiman, N. (2022), 'The use of data analytics in external auditing: a content analysis approach', *Asian Review of Accounting* **30**, 31–58.

- Jans, M. J., Alles, M. G. and Vasarhelyi, M. A. (2014), 'Process Mining of Event Logs in Auditing: Opportunities and Challenges.', *SSRN Electronic Journal* pp. 1–32.
- Jeong, H., Jang, Y., Bowman, P. J. and Masoud, N. (2018), 'Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data', *Accident Analysis Prevention* **120**, 250–261.
URL: <https://www.sciencedirect.com/science/article/pii/S0001457518305232>
- Jepkoech, J., Mugo, D. M., Kenduiywo, B. K. and Too, E. C. (2021), 'The effect of adaptive learning rate on the accuracy of neural networks'.
- John, H. and Naaz, S. (2019), 'Credit card fraud detection using local outlier factor and isolation forest', *Int. J. Comput. Sci. Eng* **7**(4), 1060–1064.
- Joseph, V. R. (2022), 'Optimal ratio for data splitting', *Statistical Analysis and Data Mining: An ASA Data Science Journal* **15**(4), 531–538.
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11583>
- Joseph, V. R. and Vakayil, A. (2022), 'Split: An optimal method for data splitting', *Technometrics* **64**(2), 166–176.
- Kahyaoglu, S. B. and Aksoy, T. (2021), *Artificial Intelligence in Internal Audit and Risk Assessment*, Springer International Publishing, Cham, pp. 179–192.
URL: https://doi.org/10.1007/978-3-030-72624-9_8
- Kapoor, M. (2020), 'Big four invest billions in tech, reshaping their identities. bloomberg tax'.
- URL:** <https://news.bloombergtax.com/financial-accounting/big-four-invest-billions-in-tech-reshaping-their-identities>
- Kim, J., Jeong, K., Choi, H. and Seo, K. (2020), Gan-based anomaly detection in imbalance problems, in A. Bartoli and A. Fusiello, eds, 'Computer Vision – ECCV 2020 Workshops', Springer International Publishing, Cham, pp. 128–145.
- Kingma, D. P. and Ba, J. (2017), 'Adam: A method for stochastic optimization'.
- URL:** <https://arxiv.org/abs/1412.6980>
- Kobayashi, Y. and Yoshida, K. (2022), 'Automated retention time prediction of new psychoactive substances in gas chromatography', *Procedia Computer Science* **207**, 654–663. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022.
URL: <https://www.sciencedirect.com/science/article/pii/S1877050922010018>

- Koivu, A., Korpimäki, T., Kivelä, P., Pahikkala, T. and Sairanen, M. (2018), 'Evaluation of machine learning algorithms for improved risk assessment for down's syndrome', *Computers in Biology and Medicine* **98**, 1–7.
URL: <https://www.sciencedirect.com/science/article/pii/S0010482518301112>
- Kong, J., Kowalczyk, W., Menzel, S. and Bäck, T. (2020), Improving imbalanced classification by anomaly detection, in T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich and H. Trautmann, eds, 'Parallel Problem Solving from Nature – PPSN XVI', Springer International Publishing, Cham, pp. 512–523.
- Kotb, A., E. H. and Halabi, H. (2020), 'Mapping of internal audit research: a post-enron structured literature review', *Accounting, Auditing Accountability Journal* **33**, 1969–1996.
- Kourehpaz, P. and Hutt, C. M. (2022), 'Machine learning for enhanced regional seismic risk assessments', *Journal of Structural Engineering* **148**(9).
- Krieger, F., Drews, P. and Velte, P. (2021), 'Explaining the (non-) adoption of advanced data analytics in auditing: A process theory', *International Journal of Accounting Information Systems* **41**.
- Kumar, M. and Reddy, A. R. M. (2016), 'A fast dbscan clustering algorithm by accelerating neighbor searching using groups method', *Pattern Recognition* **58**, 39–48.
URL: <https://www.sciencedirect.com/science/article/pii/S0031320316001035>
- Kurniawan, Taufik, B. R. B. and Erwin, K. (2023), 'The factors influencing the risk based internal audit in improving the effectiveness of internal audit', *International Journal of Social Science and Business* **7**, 1030–1041.
- Ladds, Monique A., T. A. P. K. J.-P. J. S. D. P. H. D. and G Harcourt, R. (2017), 'Super machine learning: improving accuracy and reducing variance of behaviour classification from accelerometry', *Animal Biotelemetry* **5**.
- Law, K. and Shen, M. (2020), 'How does artificial intelligence shape audit firms?', *Management Science* .
- L'elis, D. L. M. and Pinheiro, L. E. T. (2012), 'Auditor and auditee perceptions of internal auditing practices in a company in the energy sector', *Revista Contabilidade Financas* **23**, 212–222.
- Lesouple, J., Baudoin, C., Spigai, M. and Tournet, J.-Y. (2021), 'Generalized isolation forest for anomaly detection', *Pattern Recognition Letters* **149**, 109–119.
- Li, Y. and Wang, X. (2024), 'A hybrid decision support framework for planning a risk-based audit engagement', *Journal of the Operational Research Society* pp. 1–16.

- Liao, S.-H. and Wen, C.-H. (2007), 'Artificial neural networks classification and clustering of methodologies and applications – literature analysis from 1995 to 2005', *Expert Systems with Applications* **32**(1), 1–11.
URL: <https://www.sciencedirect.com/science/article/pii/S0957417405003027>
- Lin, S.-S., Shen, S.-L., Zhou, A. and Xu, Y.-S. (2021), 'Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods', *Automation in Construction* **122**, 103490.
- Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. (2021), 'Explainable ai: A review of machine learning interpretability methods', *Entropy* **23**(1).
URL: <https://www.mdpi.com/1099-4300/23/1/18>
- Locke, N. and Bird, H. L. (2020), 'Perspectives on the current and imagined role of artificial intelligence and technology in corporate governance practice and regulation', *Australian Journal of Corporate Law* .
URL: <https://ssrn.com/abstract=3534898>
- Lombardi, R., de Villiers, C., Moscariello, N. and Pizzo, M. (2021), 'The disruption of blockchain in auditing - a systematic literature review and an agenda for future research', *Accounting, Auditing Accountability Journal* .
- Luque, A., Carrasco, A., Martín, A. and de las Heras, A. (2019), 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition* **91**, 216–231.
URL: <https://www.sciencedirect.com/science/article/pii/S0031320319300950>
- Lyon, B. K. and Hollcroft, B. (2012), 'Risk assessments', *Professional Safety* **57**(12), 28–34.
- Martino, P., D'Onza, G. and Melville, R. (2021), 'The relationship between cae leadership and the iaf's involvement in corporate governance', *Journal of Accounting, Auditing & Finance* **36**(2), 430–448.
- Mayr, A., Binder, H., Gefeller, O. and Schmid, M. (2014), 'The evolution of boosting algorithms', *Methods of information in medicine* **53**(06), 419–427.
- Monroe, G., Ng, J. and Woodliff, D. (2008), 'The importance of inherent risk factors: Auditors' perceptions', *Australian Accounting Review* **3**, 34 – 46.
- Munevar, K. and Patatoukas, P. N. (2020), 'The blockchain evolution and revolution of accounting. information for efficient decision making: Big data, blockchain and relevance', pp. 1, 5, 19.
- Murdock, H. (2021), *Operational auditing: Principles and techniques for a changing world*, CRC Press.

- Nonnenmacher, J. and Gomez, J. M. (2021), 'Unsupervised anomaly detection for internal auditing: Literature review and research agenda', *The International Journal of Digital Accounting Research* pp. 1–22.
- Norman, C. S., Rose, A. M. and Rose, J. M. (2010), 'Internal audit reporting lines, fraud risk decomposition, and assessments of fraud risk', *Accounting, Organizations and Society* **35**(5), 546–557.
URL: <https://www.sciencedirect.com/science/article/pii/S0361368209001056>
- Nur Muslihatun, F. A., Sunarfri Hantono, B. and Fauziati, S. (2021), Using artificial intelligence technology for decision support system in audit risk assessment: A review paper, in '2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)', pp. 326–331.
- Nwankpa, C., Ijomah, W., Gachagan, A. and Marshall, S. (2018), 'Activation functions: Comparison of trends in practice and research for deep learning'.
URL: <https://arxiv.org/abs/1811.03378>
- Omidi Firouzi, H. and Wang, S. (2020), 'A Fresh Look at Internal Audit Framework at the Age of Artificial Intelligence (AI).', *SSRN Electronic Journal* pp. 1–11.
- Opěla, P., Schindler, I., Kawulok, P., Kawulok, R., Rusz, S. and Sauer, M. (2022), 'Shallow and deep learning of an artificial neural network model describing a hot flow stress evolution: A comparative study', *Materials Design* **220**, 110880.
URL: <https://www.sciencedirect.com/science/article/pii/S0264127522005020>
- Ossai, C. I. (2019), 'A data-driven machine learning approach for corrosion risk assessment—a comparative study', *Big Data and Cognitive Computing* **3**(2), 28.
- Otchere, D. A., Arbi Ganat, T. O., Gholami, R. and Ridha, S. (2021), 'Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ann and svm models', *Journal of Petroleum Science and Engineering* **200**, 108182.
URL: <https://www.sciencedirect.com/science/article/pii/S0920410520312365>
- Oyewo, B., A. O. and Ajape, M. (2021), 'Characteristics of consulting firms associated with the diffusion of big data analytics', *Journal of Asian Business and Economic Studies* **28**(4), 281–302.
- Paltrinieri, N., Comfort, L. and Reniers, G. (2019), 'Learning about risk: Machine learning for risk assessment', *Safety Science* **118**, 475–486.
URL: <https://www.sciencedirect.com/science/article/pii/S0925753518311184>

- Perols, J., Bowen, R. M., Zimmermann, C. and Samba, B. (2015), 'Finding needles in a haystack: Using data analytics to improve fraud prediction', pp. 1, 10–20.
- Porcuna-Enguix, L., Bustos-Contell, E., Serrano-Madrid, J. and Labatut-Serer, G. (2021), 'Constructing the audit risk assessment by the audit team leader when planning: Using fuzzy theory', *Mathematics* **9**(23).
URL: <https://www.mdpi.com/2227-7390/9/23/3065>
- PwC (2022), 'Global risk survey: Internal audit's response to the new risk multiverse'.
URL: <https://www.pwc.com/gx/en/services/audit-assurance/assets/pwc-global-risk-survey-internal-audits-response-to-new-risk-multiverse.pdf>
- PwC (n.d.), 'Data analytics - transforming your internal audit function'.
URL: <https://www.pwc.co.uk/services/risk/insights/data-analytics-transforming-your-internal-audit-function.html>
- Radbruch, J. and Schiprowski, A. (2024), 'Interview sequences and the formation of subjective assessments', *Review of Economic Studies* p. rdae039.
- Ramachandran, P., Zoph, B. and Le, Q. V. (2017), 'Searching for activation functions'.
URL: <https://arxiv.org/abs/1710.05941>
- Ramadan, H. S., Maghawry, H. A., El-Eleamy, M. and El-Bahnasy, K. (2022), 'A heuristic novel approach for determination of optimal epsilon for dbscan clustering algorithm', *J Theor Appl Inf Technol* **15**(7).
- Rashidi HH, Albahra S, R. S. T. N.-H. B. (2023), 'Common statistical concepts in the supervised machine learning arena', *Front Oncol* .
- Rawashdeh, A. (2022), 'Adoption of robotic process automation (rpa) and its effect on business value: An internal auditors perspective', *Journal of Positive School Psychology* **6**.
URL: <https://journalppw.com/index.php/jpsp/article/view/9497/6180>
- Redmill, F. (2002), 'Risk analysis - a subjective process', *Engineering Management Journal* **12**, 91 – 96.
- Redmill, F. (2004), 'Exploring risk-based testing and its implications', *Software testing, verification and reliability* **14**(1), 3–15.
- Ribeiro, M. T., Singh, S. and Guestrin, C. (2016), "why should i trust you?": Explaining the predictions of any classifier, in 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, Association for Computing Machinery, New York, NY, USA, p. 1135–1144.
URL: <https://doi.org/10.1145/2939672.2939778>

- Rodríguez-Pérez, R. and Bajorath, J. (2020), 'Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions', *Journal of Computer-Aided Molecular Design* **34**.
- Ruhnke, K. (2022), 'Empirical research frameworks in a changing world: The case of audit data analytics', *Journal of International Accounting, Auditing and Taxation* .
- Rye, I., Vik, A., Kocinski, M., Lundervold, A. S. and Lundervold, A. J. (2022), 'Predicting conversion to alzheimer's disease in individuals with mild cognitive impairment using clinically transferable features.', *Scientific reports* **12**.
- Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K. and Menegaz, G. (n.d.), 'A perspective on explainable artificial intelligence methods: Shap and lime', *Advanced Intelligent Systems* **n/a(n/a)**, 2400304.
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202400304>
- Saseendran, A., Setia, L., Chhabria, V., Chakraborty, D. and Barman Roy, A. (2019), 'Impact of noise in dataset on machine learning algorithms'.
- Semenov, A., Boginski, V. and Pasiliao, E. L. (2019), Neural networks with multidimensional cross-entropy loss functions, in A. Tagarelli and H. Tong, eds, 'Computational Data and Social Networks', Springer International Publishing, Cham, pp. 57–62.
- Sharma, S., Sharma, S. and Athaiya, A. (2020), 'Activation functions in neural networks', *International Journal of Engineering Applied Sciences and Technology* **4**, 310–316.
- Shivram, V. (2024), 'Auditing with ai: A theoretical framework for applying machine learning across the internal audit lifecycle', *EDPACS* **69**(1), 22–40.
URL: <https://doi.org/10.1080/07366981.2024.2312025>
- Sifa, R., Ladi, A., Pielka, M., Ramamurthy, R., Hillebrand, L., Kirsch, B., Biesner, D., Stenzel, R., Bell, T., Lübbering, M., Nütten, U., Bauckhage, C., Warning, U., Fürst, B., Khameneh, T. D., Thom, D., Huseynov, I., Kahlert, R., Schlums, J., Ismail, H., Kliem, B. and Loitz, R. (2019), Towards automated auditing with machine learning, in 'Proceedings of the ACM Symposium on Document Engineering 2019', DocEng '19, Association for Computing Machinery, New York, NY, USA.
URL: <https://doi.org/10.1145/3342558.3345421>
- Singh, D. and Singh, B. (2020), 'Investigating the impact of data normalization on classification performance', *Applied Soft Computing* **97**, 105524.
URL: <https://www.sciencedirect.com/science/article/pii/S1568494619302947>
- Stanisic, N., Radojevic, T. and Stanic, N. (2019), 'Predicting the Type of Auditor Opinion: Statistics, Machine Learning, or a Combination of the Two?', *The European Journal of Applied Economics* **16**(2), 1–58.

- URL:** <https://ssrn.com/abstract=3471395>
- Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A. (2020), 'Data imbalance in classification: Experimental evaluation', *Information Sciences* **513**, 429–441.
URL: <https://www.sciencedirect.com/science/article/pii/S0020025519310497>
- UNC (n.d.), 'University of north carolina - the internal audit process'.
URL: <https://internalaudit.unc.edu/audit-processes/the-internal-audit-process/>
- Van der Nest, D., Smidt, L. and Lubbe, D. (2017), 'The use of generalised audit software by internal audit functions in a developing country: The purpose of the use of generalised audit software as a data analytics tool', *Risk Governance and Control: Financial Markets Institutions* **7**, 100–112.
- Venderley, J., Mallayya, K., Matty, M., Krogstad, M., Ruff, J., Pleiss, G., Kishore, V., Mandrus, D., Phelan, D., Poudel, L., Wilson, A. G., Weinberger, K., Upreti, P., Norman, M., Rosenkranz, S., Osborn, R. and Kim, E.-A. (2022), 'Harnessing interpretable and unsupervised machine learning to address big data from modern x-ray diffraction', *Proceedings of the National Academy of Sciences* **119**(24), e2109665119.
URL: <https://www.pnas.org/doi/abs/10.1073/pnas.2109665119>
- Vishnia, G. and Peters, G. (2019), 'Auditchain: A trading audit platform over blockchain', pp. 15, 17–18, 22–23.
- Wang, S. and Cao, G. (2023), 'Multiclass classification for multidimensional functional data through deep neural networks'.
URL: <https://arxiv.org/abs/2305.13349>
- Wang, X., Yan, P. and Ferreira, F. A. F. (2023), 'A multi-objective optimization approach for integrated risk-based internal audit planning.', *Annals of operations research* pp. 1–29.
- Warren, J. D. and Smith, M. (2006), 'Continuous auditing: An effective tool for internal auditors', *Internal Auditing* **21**, 27–35.
URL: <https://ssrn.com/abstract=905144>
- Wassie, F. A. and Lakatos, L. P. (2024), 'Artificial intelligence and the future of the internal audit function', *Humanities and Social Sciences Communications* **11**.
- Wellmeyer, P. (2022), 'In the era of audit data analytics, what's happened to audit sampling?'.
- Winter, P. M., Eder, S., Weissenböck, J., Schwald, C., Doms, T., Vogt, T., Hochreiter, S. and Nessler, B. (2021), 'Trusted artificial intelligence: Towards certification of machine learning applications'.
URL: <https://arxiv.org/abs/2103.16910>

- Xiao, T., Geng, C. and Yuan, C. (2020), 'How audit effort affects audit quality: An audit process and audit output perspective', *China Journal of Accounting Research* **13**(1), 109–127.
URL: <https://www.sciencedirect.com/science/article/pii/S1755309120300034>
- Xu, H., Pang, G., Wang, Y. and Wang, Y. (2023), 'Deep isolation forest for anomaly detection', *IEEE Transactions on Knowledge and Data Engineering* **35**(12), 12591–12604.
- Xu, Y. and Goodacre, R. (2018), 'On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning.', *Journal of analysis and testing* **2**, 249–262.
- Yoon, K., Hoogduin, L. and Zhang, L. (2015), 'Big Data as Complementary Audit Evidence', *Accounting Horizons* **29**(2), 431–438.
URL: <https://doi.org/10.2308/acch-51076>
- Yousif, N. and Mohamed, S. (2022), 'The role of internal audit in assessing the risks of management decisions regarding strategic operations)acquisition)', *Journal of Economics and Administrative Sciences* **28**, 172–186.
- Zainal Abidin, N. H. (2017), 'Factors influencing the implementation of risk-based auditing', *Asian Review of Accounting* **25**.
- Zennaro, F., Furlan, E., Simeoni, C., Torresan, S., Aslan, S., Critto, A. and Marcomini, A. (2021), 'Exploring machine learning potential for climate change risk assessment', *Earth-Science Reviews* **220**, 103752.
- Zhang, J., Yang, X. and Appelbaum, D. (2015), 'Toward Effective Big Data Analysis in Continuous Auditing', *Accounting Horizons* **29**(2), 469–476.
URL: <https://doi.org/10.2308/acch-51070>
- Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H. et al. (2020), 'Towards theoretically understanding why sgd generalizes better than adam in deep learning', *Advances in Neural Information Processing Systems* **33**, 21285–21296.
- Özbaltan, N. (2024), 'Applying machine learning to audit data: Enhancing fraud detection, risk assessment and audit efficiency', *EDPACS* **69**(9), 70–86.

APPENDIX A: CODE USED TO DEVELOP MODELS USED IN EXPERIMENTS

The code used to build the various models developed in this thesis are available at the following github link: <https://github.com/vivekshivram/stirling-profd-notebooks>

The code has been written across five individual notebooks, to facilitate a detailed review of the code and / or reproduce any results. The base dataset collected from the sponsoring organisation has been anonymised and saved to a csv file (base_dataset.csv), which is referenced across each of the notebooks below.

1. **VS_Classification_Models_Visuals_Stats:** Contains statistical tests, visualisations and exploratory data analyses performed in the initial stages of the thesis, prior to model development.
2. **VS_Classification_Models_Final_IR:** Contains the code used to develop the final set of algorithms used to model inherent risk assessments.
3. **VS_Classification_Models_Final_CE:** Contains the code used to develop the final set of algorithms used to model control effectiveness assessments.
4. **Artificial_Neural_Network:** Contains the code used to build an artificial neural network, developed as part of chapter 7.
5. **Unsupervised_ML_DBSCAN_kMeans:** Contains the code used to develop unsupervised machine learning models over the dataset.