Predicting Unexpected Retirement Benefits Withdrawals Using Machine Learning Algorithms

By

Simon Macharia

Adm. No. 151310

Master of Science in Data Science and Analytics

2024

Predicting Unexpected Retirement Benefits Withdrawals Using Machine Learning Algorithms

Simon Macharia

Adm. No. 151310

Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June 2024

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the research contains no material previously published or written by another person except where due reference is made in the research paper itself.

Student's Name: Simon Macharia.

Admission Number: 151310

Student Signature: _____ Date: 3/4/2024

The dissertation of Simon Macharia has been reviewed and approved by the following:

Dr. Collins Odhiambo.

Institute of Mathematical Sciences Strathmore University

Supervisor Signature: Collins Odhiambo Date: 3rd April, 2024

Abstract

Access to Retirement Benefits is normally through the attainment of the statutory retirement age, which we term as expected exits from a scheme. There are circumstances where the benefits may be accessed before attaining the retirement age i.e., upon separation from an employer in an occupational scheme where the employer is contributing on behalf of the member. The timing of these unexpected withdrawals is variable and therefore challenging to predict.

This research, therefore, sought to develop a machine-learning model that was able to accurately predict the unexpected withdrawals from the scheme and understand the factors that contributed to the withdrawal event.

The research applied secondary data from a Defined Contribution multi-employer retirement scheme in Kenya with over 63,000 participating members for a 1-year period between 2022 and 2023.

The research followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which is a widely used methodology for data science solutions. The data was cleaned and pre-processed in preparation for the model training stage. Further, the data was split into separate sets for training and testing to evaluate the performance of the model and check its generalisability. The Classification algorithms applied in this research were Logistic Regression, Support Vector Machines (SVM), Random Forests, and eXtreme Gradient boosting (XGB). The performance of the models was evaluated using accuracy, precision, recall, and F1-score.

The members were categorised into continuing members, expected exits, and unexpected exits, then the algorithms were trained and it was found that the Random Forests and XGB, consistently outperformed the Logistic Regression and SVM algorithms in classifying this data.

The logistic Regression and SVM algorithms' performance improved when the outliers were treated, then improved significantly when the Synthetic Minority Over-sampling Technique (SMOTE) technique was applied to treat the imbalanced data. The XGB and Random forests accuracy did not change but the precision improved slightly but recall marginally deteriorated on the SMOTE balanced data. The selected model was the XGB as it had superior performance across all the metrics with 97% accuracy, 87% precision, 92% recall, and 89% F1-score and it was noted to generalise well on unseen data.

The research further explored the explainability of the model to enhance transparency and conducted both global and local explainability. The global explainability was conducted through the XGBoost's feature importance and showed the top 3 features were Balance, Sponsor Code, and Age. The local explainability was explored using the Local Interpretable Model-agnostic Explanations (LIME) which showed the importance of the features in predicting specific instances

The prediction of unexpected exits is expected to help the retirement schemes in their planning for investments, make adequate budgetary provisions for these unpredictable exits, and contribute to interventions to reduce withdrawals.

KEY WORDS: *Retirement Benefits, Machine learning, algorithms, Big Data, Unexpected withdrawal Prediction, Classification Techniques.*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

AUC-ROC Area Under the Receiver Operating Characteristic Curve

AUC Area Under the Curve

ROC Receiver Operating Curve

CART Classification And Regression Tree

COVID-19 Coronavirus Disease 2019

CRISP-DM Cross-Industry Standard Process for Data Mining

DT Decision Tree

GDP Gross Domestic Product

LightGBM Light Gradient Boosting Machine

ML Machine Learning

RBF Radial Basis Function

RF Random Forests

SVM Support Vector Machines

XGB eXtreme Gradient boosting

SMOTE Synthetic Minority Over-sampling Technique

EDA Exploratory Data Analysis

EE Employee Contributions

ER Employer Contributions

AVC Additional Voluntary Contributions

LIME Local Interpretable Model-agnostic Explanations

IQR Interquartile Range

Acknowledgements

I would like to express my sincerest gratitude to my supervisor Dr. Collins Odhiambo for his patient guidance, encouragement, and advice. My sincere thanks also go to Dr. John Olukuru and the faculty and staff at Strathmore University. My heartfelt appreciation extends to my family and friends for their understanding and support during the period of the program without whom this success would not have been possible.

Chapter 1: Introduction

1.1 Background Information

Retirement benefits or Pensions are long-term savings that can be thought of as deferred earnings during working life to take care of your needs during the later stages i.e., post-retirement (Lagat, 2019).

Retirement is a situation where an individual leaves the labour force, usually in the later stages of their life (Salazar and Boado-Penas, 2019). If this individual was a member of a retirement scheme, this then triggers access to the accrued retirement benefits. There are a variety of events that allow access to the benefits in a retirement scheme which are determined by local regulations. These events include the easily predictable age-related retirement, which depends on the statutory retirement age or after a predefined period (Briere et al., 2022). There are also variable events that permit access in certain conditions (Briere et al., 2022) such as death, where the beneficiaries of the primary member of the scheme would access the accrued benefits. Secondly, there are pre-retirement withdrawals which may present in a variety of ways resignation, dismissal, ill-health, disability, transfers, and emigration (Briere et al., 2022).

The existing literature was noted to apply the terms "early retirement", "early withdrawal", "withdrawal" and "pre-retirement withdrawal" interchangeably as access to retirement benefits before the statutory retirement age (Salazar and Boado-Penas, 2019), (Briere et al., 2022), (Hamilton et al., 2023), (Panis, 2019) this may be a result of the different practices and regulations in different regions.

In financial planning, there are assumptions of the future cash flows. This is the case for Retirement Schemes, especially for their main liabilities and expenditures which are the payment of due benefits (Broeders et al., 2021). These payments generate liquidity requirements from the schemes (Broeders et al., 2021) as they are only payable in cash not in the various assets which the schemes hold. These payments include a predictable portion of the retirement age-related payments (Broeders et al., 2021) and the less predictable portion of the unexpected withdrawals.

Retirement schemes' financial stability like all other firms can be impacted by unforeseen circumstances and the traditional models of the retirement benefits liabilities may not produce accurate results, especially for the variable withdrawals and mortality categories, for example, the Coronavirus Disease 2019 (COVID-19) pandemic resulted in increased liquidity requirements

and challenges accessing funding markets (Bédard-Pagé et al., 2021) for schemes in Canada.

The payments associated with the occurrence of withdrawals and deaths introduce variability in the short-term liquidity requirements of schemes (Broeders et al., 2021) that may result in financial stress events necessitating premature liquidation of investments, (Bédard-Pagé et al., 2021), thereby foregoing some future income or in the extreme failure to pay the dues. To avoid this, schemes must improve the forecasting of unexpected withdrawals and ensure that the assets are able to settle the liabilities as and when they fall due.

Traditionally statistical methods have been applied based on past trends in historical data. These methods included linear regression models, Generalized Linear Models, and their variations (Baione et al., 2023). The traditional methods have limitations as the volume of available data grows (Salazar and Boado-Penas, 2019), some of the regression models include the assumption of the distribution of the outcome variable belonging to the exponential family (Baione et al., 2023) although this assumption may lead to simple and parsimonious models, it may not affect the accuracy of predictions.

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that allows for improvements in insights generated from data with minimal human intervention (Sunday et al., 2020). Machine learning techniques can be broadly categorised based on the intent as either descriptive or predictive and each category has various models. The choice of model can be influenced by the problem, available data characteristics, and resource availability among other considerations. The increase in data collected and availability has led to the promotion of (Sunday et al., 2020) machine learning for knowledge discovery from data.

1.2 Problem Statement

Retirement schemes allocate their assets to match their liabilities to ensure that they have adequate liquidity to meet their pension obligations at the required time. Their main liability is the accrued benefits to members (Broeders et al., 2021) and the payments of these benefits can be categorised into expected age-related retirement exits and unexpected withdrawals.

Exits due to retirement are based on a predefined regulatory retirement age,(Salazar and Boado-Penas, 2019), (Briere et al., 2022), (Hamilton et al., 2023), (Panis, 2019) that differs in different jurisdictions, and therefore can be accurately determined. On the other hand, unexpected withdrawals can arise from a variety of reasons including resignation, dismissal, emigration, and

death.

Previous studies noted that scheme participants opted to access their retirement benefits prematurely (Briere et al., 2022), up to the maximum possible amounts, including full withdrawals (Hamilton et al., 2023), i.e., prior to attaining the statutory retirement age. There were also notable events of employers withdrawing all their participating employees from multi-employer schemes (Panis, 2019). This therefore requires retirement schemes to plan for and pay withdrawals that may occur before a member attains the retirement age.

Due to the variability, Retirement schemes find it challenging to accurately predict these exits (Salazar and Boado-Penas, 2019) and they pose significant liquidity risks that can adversely affect the financial position of the pension schemes. Current methods for predicting unexpected withdrawals include actuarial, statistical, (Broeders et al., 2021),(Hamilton et al., 2023), and analytical methods (Briere et al., 2022) that focus on a few contributing factors that may result in sub-optimal accuracy in the changing environment and limit effectiveness in planning for them.

This research aimed to overcome these limitations by applying machine learning to explore the factors available in the data to gain a deeper understanding of unexpected exit patterns, develop accurate predictive models that predict them, and provide actionable insights for planning.

1.3 Research Objectives

The primary objectives of this dissertation are as follows:

1. To fit appropriate machine learning models for predicting unexpected exits from a retirement scheme.
2. To perform a comparative analysis of Machine Learning Algorithms for Predicting Unexpected Withdrawals
3. To investigate the features that contribute to unexpected withdrawals in a retirement scheme.

1.4 Research Questions

1. Which machine learning algorithms are most effective for predicting unexpected withdrawals from retirement benefits schemes?

2. How do different machine learning algorithms perform in the prediction of unexpected withdrawals from retirement schemes?
3. What are the characteristics that affect early withdrawals from retirement benefits schemes?

1.5 Scope and Limitations

The research focused on applying the Machine Learning models to data from a retirement scheme with members from all Counties in Kenya for 1 year from 2022 to 2023. The research will explore various machine learning algorithms used for classification and select the appropriate one based on the performance. The limitations include the timeline of the data available, data for a longer period would have allowed an analysis of trends and cycles along with comparison with external factors. Another limitation was that may not be able to capture all of the complexity in the features that influence withdrawals. The research focused on the data collected and therefore the accuracy of the machine learning models will be affected by the quality and quantity of data.

1.6 Justification

The outcomes of this research have the potential to significantly advance the administration of Retirement Benefits schemes and especially planning for withdrawals by harnessing the power of machine learning.

Accurate and interpretable withdrawal estimates can aid policymakers and insurance and pension organisations in allocating resources effectively, designing targeted interventions, and ultimately improving social protection programs.

Chapter 2: Literature Review

2.1 Introduction

There are different types of retirement benefits arrangements but they are all founded on a principle that is to provide a source of reliable income in the later years of one's life (Lagat, 2019) when there is reduced earnings potential supposedly due to physiological decline (Salazar and Boado-Penas, 2019). They aim to offer financial security when one is less productive and guard against old age poverty and over-dependency on family or the government.

This section discusses the concepts that underpin retirement benefits and informs the research on unexpected withdrawals. It helps establish a basis for understanding the contributing factors to unexpected withdrawals and the usage of machine learning to predict the withdrawals.

Unexpected withdrawals may cover a range of different exit events resulting in full or partial drawings from accrued retirement benefits that have limited access or stringent eligibility requirements. These vary, in the case of France where hardship conditions including home ownership, unemployment, death of the member or their next of kin, disability (Briere et al., 2022) all qualify a member of a scheme to access the long-term savings before the retirement age or qualification period.

Unexpected withdrawals from retirement benefits schemes pose a significant threat to both individual financial security and the overall stability of these crucial institutions (Salazar and Boado-Penas, 2019). Predicting these withdrawals becomes a vital undertaking for protecting present and future individual well-being, safeguarding the sustainability of schemes as a going concern and for regulatory purposes (Baione et al., 2023), and enabling proactive interventions including policy changes.

Unexpected withdrawals, while only allowed in certain conditions (Briere et al., 2022), affect the schemes by increasing short-term liquidity demands (Salazar and Boado-Penas, 2019), may force premature liquidation of assets or borrowing to meet the short-term liquidity requirements, and in the worst case scenario of mass withdrawals they may cause the bankruptcy of a scheme or even extending the consequences to the sponsoring employers firm (Panis, 2019)in the case of underfunded employer-sponsored Defined Benefit plans.

Withdrawals also affect the individual by increasing the risks of insufficiency of the accrued

benefits during old age (Hou et al., 2022) as the funds have competing priorities including the possibility of outliving the retirement savings, family responsibilities, losses, increased medical-related expenses

That is not to say that the withdrawal from retirement benefits should be completely outlawed, as the reasons for an allowable withdrawal are usually defined in the regulatory frameworks of a country, for example in France as noted by (Briere et al., 2022), or even caused by a change in the regulations such as a response to the COVID-19 pandemic, the Australian government allowed capped withdrawals (Hamilton et al., 2023).

Therefore, this research aims to examine the use of machine learning and develop a suitable model for predicting unexpected withdrawals from a retirement benefits scheme.

2.2 Comparison between insurance and pensions

There are similarities in life insurance contracts and pensions arrangements that allow crossovers in their analysis e.g. the surrenders in insurance contracts including whole life, endowment policies, are similar to withdrawals in retirement benefits arrangements (Baione et al., 2023). Life Insurance companies also provide retirement benefits products to willing participants (Salazar and Boado-Penas, 2019).

At inception, they have an initial expected period to be in force but have situations in which they can be prematurely terminated, wholly or partially, triggering a change in the expected cash flow thus affecting the organisations that are counter-parties to the contracts (Kiermayer, 2022), (Salazar and Boado-Penas, 2019).

Therefore, the research considered research conducted on insurance contract surrenders along with withdrawals from retirement schemes.

(Salazar and Boado-Penas, 2019).

2.3 Existing Approaches to Predicting Unexpected Withdrawals

Traditional approaches to prediction relied on analyzing various demographic and financial factors of members and included analytical and statistical methods like regressions, Traditional models were parsimonious and simple to understand, and interpret and allowed for comparison and bench-marking across portfolios, organisations, and industries

They had limitations including the limited amount of data they could analyse some assumptions such as linearity of the data or the underlying distribution belonging to the exponential family (Baione et al., 2023) which may not be the case in the real world data with complex relationships, their static nature that limited their adaptability (Salazar and Boado-Penas, 2019) and limited scope of factors considered (Bjerre, 2022).

The changing demographics, economic volatility, the growing complexity and availability of data, and outlier events such as the COVID-19 Pandemic (Hamilton et al., 2023) further exposed the limitations of traditional models.

While the traditional methods retain their value in providing a framework and historical context, the search for more sophisticated and adaptable methods, like machine learning algorithms, is becoming increasingly important for accurately predicting withdrawals and ensuring the long-term stability of retirement schemes.

However, it is worth noting that traditional statistical methods can be combined with machine learning algorithms to result in better overall predictions (Salazar and Boado-Penas, 2019) by combining the advantages and minimizing the disadvantages of each of the different approaches.

2.4 Case for application of Machine Learning to Predicting Unexpected Withdrawals

Machine Learning, a subset of AI, improves accuracy in decision-making by generating insights from the data (Salazar and Boado-Penas, 2019) and offers unique advantages that can improve the understanding and prediction of withdrawal patterns as they are uniquely able to handle data in large volumes and differentiated types (Salazar and Boado-Penas, 2019),

Machine learning algorithms can identify simple and complex relationships in data, and are inherently adaptive, continuously learning and updating their predictions based on new data and changing circumstances (Salazar and Boado-Penas, 2019). This ensures their forecasts remain relevant and accurate over time.

While Machine Learning is a powerful tool, it has its limitations and challenges including the models requiring substantial amounts of data for the analyses, data privacy and targeting concerns (Hamilton et al., 2023), model explainability, and potential biases. However, responsible ethical applications of the algorithms and data are possible.

2.5 Review of Existing Studies using Machine Learning for Withdrawal Prediction

The prediction of withdrawals applied supervised learning techniques including logistic regression, random forests, support vector machines, and neural networks. These methods require labelled data, which are data that have the target variable, denoting if an individual member withdrew or not.

(Salazar and Boado-Penas, 2019) applied machine learning techniques to determine scores for early withdrawals in claims from a Mexican insurance company, using data consisting of individual characteristics and macroeconomic variables. The study applied Logistic regression, RF, and SVM as the machine learning models. The Logistic regression model was selected as the model with the best performance as evaluated in terms of accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

This study applied the analysis on data consisting of claims, not on active participants and this may not be optimal in proactive planning as the claim event will already have occurred.

Machine learning has been successfully applied in the related insurance industry in the prediction of lapse and surrender events and the associated payment amounts. Various studies noted that modelling of surrenders in a supervised learning endeavour (Kiermayer, 2022) that may be presented as binary or multiple class problem (Baione et al., 2023). Therefore, it is a classification problem (Xong and Kang, 2019),(Baione et al., 2023) to determine the label of an outcome variable based on a new set of data points based on models developed prior (Salazar and Boado-Penas, 2019).

(Kiermayer, 2022) developed Machine Learning models to provide multi-period estimates of surrender probabilities in four different profiles based on the literature reviewed. The study found that mean surrender rates increased and decreased for the differentiated surrender profiles when reviewed as a time series i.e. they were non-stationary. The research applied Neural networks, logistic regression, tree-based classifiers (CART and Random forests), and the XGB which was selected due to superior modelling performance.

This study simulated data based on the literature, however, simulated data may not be as ideal as real-world data as it lacks complexity, may be prone to oversimplification, and produce misleading results thereby limiting their generalisability.

(Xong and Kang, 2019) in the related problem of the surrender of insurance contracts applied

the Neural Networks, Logistic Regression, k-Nearest Neighbours, SVM algorithms, and the SVM was preferred due to its performance in both in-sample and out-of-sample predictive performance. The study was conducted on a data set from a Malaysian insurance company that had 800 observations. This size data may not be sufficient to select the best features, model the complexities in the data and produce generalisable results.

(Baione et al., 2023) applied a two-fold methodology i.e., a multinomial Logistic regression to predict surrenders and withdrawals and secondly a beta regression analysis to measure the surrender value expected to be paid out in a life insurance portfolio. The study did not apply, contrast and compare different machine learning algorithms that may have improved on the results.

2.5.1 Studies that did not apply Machine learning algorithms

(Briere et al., 2022) studied the propensity of members of retirement benefits schemes in France to maintain their liquidity by accessing the restricted portion whenever they were able. The research noted that the inclusion of limited access in the scheme design affected the number of people who chose to enrol. The research also conducted hypothesis tests to determine if the early withdrawers were prone to accessing the restricted portion before accessing the unrestricted portion of the retirement benefits and found this to be the case.

This research was conducted on ample data 645,966 for a machine learning analysis but did not compare the modelling performance against any other algorithms, it also considered data in one calendar year which may not represent the trends or distribution of occurrences accurately as it may have been an outlier.

(Broeders et al., 2021) studied the asset allocations of pension funds in the Netherlands. The study noted that the main sources of liquidity risk in schemes were derivatives and short-term pension payments. The research stated that the pension payments were predictable and developed a mathematical function (inverse of the liability duration) to represent the pension payments and proceeded to focus on derivatives as the main driver of liquidity requirements. In their depiction of the retirement benefits function, it was noted that the fitted curve was least accurate on the shorter-term high benefits amounts. This then shows that the short-term liabilities from pension payments may require improved modelling for more accurate predictions.

(Panis, 2019) reviewed the risk of mass withdrawals from employer-sponsored pension schemes

and the possibility of a contagion of the withdrawals in the United States of America. The research applied data from an insurer of pension schemes and noted that there were cases of withdrawals where an employer pulled out of a scheme together with all their participating employees. The research applied a basic analysis of the events and developed a theoretical framework for the contagion risk. This study may have benefited from conducting an in-depth statistical and machine learning analysis to determine the levels of risk associated with mass withdrawals.

Therefore following a review of the studies, it is clear that the modelling of unexpected withdrawals from pension funds is possible and the research should cover gaps including applying multi-year real-world data, of an adequate volume, from the Kenyan market, developing and comparing the performance of various models and including data from active participants.

2.6 Factors Affecting Pension Fund Withdrawals

In studies on withdrawals and surrenders, authors have noted the importance of using the correct variables, and therefore the sources of data must be reputable, accurate, complete, and free from bias (Broeders et al., 2021). There were various sources applied in studies including internal administrative records from retirement schemes (Briere et al., 2022), insurance companies (Baione et al., 2023) (Xong and Kang, 2019), (Salazar and Boado-Penas, 2019), and aggregated data of participating organisations from industry supervisors (Broeders et al., 2021), insurers of pension plans in the case of (Panis, 2019), national databases and credit bureaus (Hamilton et al., 2023), survey records, (Hou et al., 2022) and simulated data (Kiermayer, 2022).

From the studies, it was noted that demographic variables including age, gender, and education level, were key predictors of withdrawals (Salazar and Boado-Penas, 2019). Their variations gave information that allowed the models to determine unique combinations that contributed to withdrawals.

Information on individuals' financial status was applicable as predictors of withdrawal(Hamilton et al., 2023) (Salazar and Boado-Penas, 2019). The individuals who chose to withdraw were found to have lower income from investments, lower salaries, and lower accrued retirement benefits, indicating either lower historical income or prior access to the retirement benefits.

Macroeconomic variables including unemployment returns from government bonds and the stock market were found to contribute to withdrawal rates (Salazar and Boado-Penas, 2019).

These variables may indicate the number of jobs available in the market as they may affect retrenchment and redundancy rates in the negative or periods of high return may result in voluntary withdrawals.

Occupational information such as the industry, job role and satisfaction levels affect the withdrawal rates. The studies found that individuals working in 'blue collar' jobs were more likely to withdraw than their 'white collar' counterparts (Panis, 2019), the same was noted where individuals working in roles with lower social class or job satisfaction were more likely to leave their jobs and withdraw their accrued benefits (Salazar and Boado-Penas, 2019)

The health status of the primary participant or their dependants was found to impact the decision to leave their job or withdraw from their retirement benefits (Salazar and Boado-Penas, 2019).

The calendar year was found to be a key component in the analysis of surrender rates as a time series as studied by (Kiermayer, 2022). The results of using time affected the predictive power of models and thus the accuracy of predictions made.

In the case where an employer withdraws from a scheme, the industry, state of employee and union relations, related connections among common employers, government policy such as bailouts, scheme financial health, and status of outstanding contributions to the scheme were found to contribute to the withdrawal event (Panis, 2019).

The COVID-19 pandemic resulted in regulatory or policy changes that affected the withdrawal rates from retirement schemes as noted in the study by (Hamilton et al., 2023) where following a change allowing capped access to previously inaccessible retirement benefits, 16% of the members withdrew amounts that added up to 2% of the Gross Domestic Product (GDP) of Australia within a short period.

The studies reviewed were noted to have applied the variables collected and conducted neither feature selection nor analysis on the importance of features applied in the models other than (Baione et al., 2023) who investigated the significance of the coefficients in the multinomial logistic regression model. Therefore this research aims to manage the issues of under/overfitting and to conduct feature importance to determine the contribution of each of the features to the models' predictions.

2.7 Machine Learning Algorithms for Prediction

Machine learning provides unique advantages to learning from complex relationships among diverse types of data such as in the prediction of unexpected withdrawals from retirement schemes. The models applied were as follows:

2.7.1 Logistic Regression

The logistic regression model is a simple and interpretable model suitable for analysing relationships between predictors and a binary outcome variable and it is advantageous as it can work with linear and nonlinear relationships within the data (Salazar and Boado-Penas, 2019). It was also noted to produce reasonable predictions even in cases of imbalanced data. Its ease of implementation allows for its popularity as seen by the number of implementations below.

The logistic regression model was applied in various studies to predict withdrawals from retirement benefits schemes (Salazar and Boado-Penas, 2019), or the related problem of surrenders in insurance contracts (Kiermayer, 2022), (Xong and Kang, 2019). (Baione et al., 2023) applied a multinomial extension of the logistic regression to cater for predictions of multiple classes.

In the reviewed studies, the logistic regression model performed well in the accuracy of predictions and was selected as the preferred model for scoring and predicting retirement scheme withdrawals in the study by (Salazar and Boado-Penas, 2019). Although the performance of the models was close, Logistic Regression slightly outperformed Support vector machines (SVM) and Random forests in the accuracy of predictions. The logistic regression model also directly generated the probability of early retirement in the study while the other two models would require additional computations (Salazar and Boado-Penas, 2019).

The model, in a bagged form, had a bias whereby it consistently underestimated surrender likelihood in contracts that had high surrender rates, and failed to capture a sharp change in the trend of mean surrender rates (Kiermayer, 2022) but the logistic regression models' performance improved on the naive baseline model, meaning it was better than random guessing, and even emerged as the best classifier for one of the surrender profiles studied.

2.7.2 Support Vector Machines (SVM)

The support vector machines separate data by applying a decision boundary known as a hyperplane even in higher dimensional spaces (Salazar and Boado-Penas, 2019) thereby being more

applicable in highly dimensional datasets. The SVM model was found to be the best classification model in predicting surrenders by (Xong and Kang, 2019). SVM fit the in-sample training data with the highest accuracy (Xong and Kang, 2019) but had slightly lower out-of-sample testing performance than the Neural networks and may have been overfitting. The model was applied and while the accuracy was commendable, it was comparatively the lowest performer and its results did not lend themselves to interpretation as easily as other models and may require additional computation (Salazar and Boado-Penas, 2019).

2.7.3 Random Forests (RF)

Random forests are machine learning algorithms that can be applied for classification tasks. They are a type of tree-based classifier that is trained on subsets of the data thereby creating branches or decision trees and then combined into a forest to improve their overall predictive performance. (Kiermayer, 2022)

The model, in a bagged form, had a bias whereby it consistently underestimated surrender likelihood in contracts that had high surrender rates (Kiermayer, 2022). The results from the RF models may require additional computation for interpretability (Salazar and Boado-Penas, 2019).

2.7.4 Neural Networks

Neural networks are algorithms that consist of nodes in layers, each with an input, output, and activation function (Kiermayer, 2022) as they aim to emulate the thought process in the brain. Multi-layer neural networks can outperform other models in classification problems as they are able to capture non-linear and complex relationships between predictors and outcome variables (Xong and Kang, 2019).

Neural networks can be prone to overfitting but this can be mitigated by implementation of early stopping (Kiermayer, 2022). Neural networks can also produce biased results such as in underestimating lapse likelihood of increased risk contracts (Kiermayer, 2022)

2.7.5 eXtreme Gradient boosting (XGBOOST)

XGB is a technique that combines the predictive power of multiple models to produce improved results. (Kiermayer, 2022) applied gradient-boosted decision trees to model surrender risk and the model emerged as the best performer. It was noted to be the only classifier among Logis-

tic Regression, Neural Networks, Classification And Regression Tree (CART), and Random forests to be unbiased among the profiles examined. It had the lowest prediction errors using Mean Absolute error and lowest variance and predicted the surrender rates reasonably well for all the years. However, as the researcher noted it was a slower model than other models e.g. LightGBM (Kiermayer, 2022)

It was also noted that the increasing model complexity may not improve predictions in the literature reviewed by (Xong and Kang, 2019)

2.8 Performance Metrics and Evaluation

As the classification subset of Machine learning aims to assign a label or category to an input based on some features or characteristics, Classifiers are normally evaluated on the accuracy of the predicted labels (Kiermayer, 2022). It is important to assess the quality and usefulness of machine learning models in classification problems like the prediction of unexpected withdrawals from retirement schemes. Therefore the models are evaluated on their accuracy, reliability, and generalizability.

2.8.1 Confusion Matrix and related measures

Classifiers are commonly evaluated using a confusion matrix and its related metrics of Accuracy, Specificity, Fβ-Score, and precision (Kiermayer, 2022). (Xong and Kang, 2019) and (Salazar and Boado-Penas, 2019) applied the confusion matrix to evaluate the performance of the classification models, and using accuracy selected the best-performing model.

In cases where there is a large disparity between the number of observations of the classes, known as imbalance, then the accuracy of the classifiers may be called into question (Kiermayer, 2022). The disparity may be due to the occurrence of the events being rare or having low probability rather than data quality issues therefore the predictions for all classes would have to follow a similar distribution. (Kiermayer, 2022) reviewed resampling techniques and noted that they can improve model performance but they also could introduce bias in the predictions and proposed the application of bootstrapping and ensemble techniques.

2.8.2 Receiver Operating Characteristics (ROC) Curve

The trade-off between the true positive rate, recall, and false positive rate, when plotted, gives the Receiver Operating Curve (ROC) curve (Kiermayer, 2022). The Area Under the Curve (AUC) is an accompanying metric that measures the area underneath the ROC curve whose outcome ranges from perfect modelling results at 1 and 0 where the model is inaccurate overall predictions and may be used for validation (Baione et al., 2023). Performance above the diagonal is interpreted to mean that the classifier performs better than random guessing (Kiermayer, 2022) where the probability of an accurate result is 50%.

(Xong and Kang, 2019) and (Salazar and Boado-Penas, 2019) applied the ROC curve to evaluate the performance of the classification models, and select the best-performing model.

Although the AUC-ROC method is primarily used for binary classification, it can be extended for multi-class classification predictions by implementing micro-averaging while also treating imbalances in the data (Baione et al., 2023).

(Kiermayer, 2022) noted that the main drawbacks of the ROC curve were that its constituent metrics focused evaluation on a single class label's predictions and that it wasn't particularly sensitive to inequalities in the classes. The research presented an argument for the precision-recall curve as it improves on this drawback by considering both binary class labels and improving the sensitivity to data balance.

2.8.3 Time Sensitive Confidence Bands

(Kiermayer, 2022) proposed and applied time-sensitive confidence bands in a time-series prediction of surrender rates as a probabilistic evaluation in contrast with the frequentist approaches of confusion matrix-based metrics. The results showed that the confidence bands were more interpretable and applicable in multi-year prediction models, but less so for single-year predictions.

Through the review of existing literature, it was clear that robust evaluation methodologies were key in any Machine learning research, and for insightful decision-making. The evaluation methods are essential tools for ensuring integrity, effectiveness, and progress and this research applied them for model evaluation.

2.9 Conclusion

This literature review provided an overview of the current state of machine learning applications to predict withdrawals from retirement schemes. It has discussed the problem area of unexpected withdrawals in retirement benefits schemes and surrender in insurance companies and reviewed traditional approaches to predicting withdrawals. The review also explored the progress made in previous studies applying machine learning and the factors considered in those studies and found that machine learning algorithms, especially classifiers can be successfully applied to predict withdrawals. Their success is predicated on the quality of data, including its size, and the different categories of variables considered including demographic, economic, regulatory, and financial data. Then the developed algorithms should be evaluated using robust measures to build trust, gain valuable insights, and make informed decisions.

The literature review has also identified gaps such as inadequate machine learning analyses of retirement scheme data in Kenya, data quality, model selection, explainability and evaluation, data privacy and targeting concerns, and potential biases. The research aimed to fill these gaps by developing machine learning models that apply real-world data from a retirement scheme in Kenya to predict unexpected withdrawals, selecting the best-performing model using robust evaluation criteria, and determining the contribution of each feature in the model.

Chapter 3: Methodology

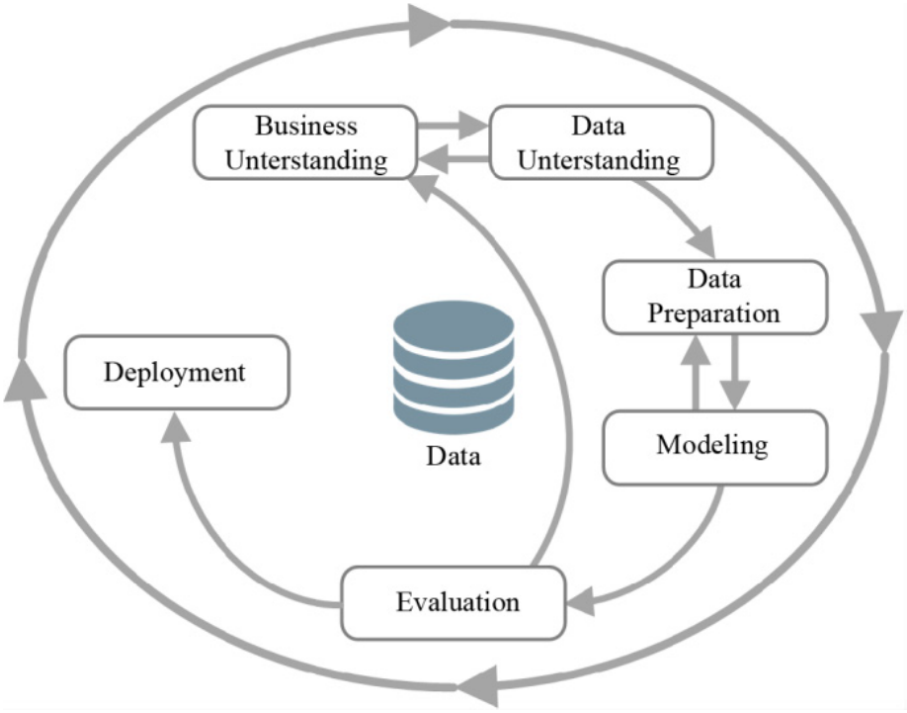
3.1 Introduction

3.1.1 Research Design

This research applied machine learning predictive modelling on a cross-sectional cohort of pension scheme participants for a one-year period and used the CRISP-DM methodology, which is a widely used framework for conducting data mining projects (Schröer et al., 2021).

This framework has the following steps:

Figure 3.1: Steps in the CRISP-DM Method



Source: (Alcober et al., 2020)

3.2 Business Understanding

This was the primary planning phase of the research and involved brainstorming to decide the general area of research, the initial broad objectives, and then refining the research objectives and the research questions (Schröer et al., 2021). The preliminary business understanding was conducted and detailed in the initial chapters namely the Introduction and the Literature review.

This step considered the problem of predicting unexpected withdrawals in retirement schemes and how Machine learning can be applied to solve it. The potential sources for data applied for

machine learning purposes were primarily secondary sources as primary data required recording over a long period (Kamiri and Mariga, 2021). Further review of the literature noted that providers of retirement benefits data included insurance companies (Salazar and Boado-Penas, 2019), pension schemes (Briere et al., 2022), regulatory authorities, and national Bureaus of statistics (Hamilton et al., 2023) who were sources of data that was previously applied for machine learning.

Further, this step explored data mining goals to solve the problem of predicting withdrawals from retirement schemes as a classification data mining problem as in the case of (Salazar and Boado-Penas, 2019).

The final stage of this step was to develop a plan to conduct this research which culminated in a research proposal outlining the refined research objectives and questions, the past literature, and a methodology.

3.3 Data Understanding

The data understanding step followed the exploration of the potential data sources based on the research objectives defined for retrieval of the data (Schröer et al., 2021). The potential data sources explored were the main providers of retirement benefits i.e., insurance and pension organisations, but the researchers settled on secondary data from a multi-employer Defined Contribution retirement scheme in Kenya for a 1-year period that experienced sufficient membership and had withdrawal rates during the year.

3.3.1 Data description

The analysis was conducted using Python, a programming language widely used for machine learning (Kamiri and Mariga, 2021). The data was loaded and it was noted that it had various variables for the 63,852 scheme participants for the period 2022-2023 as shown in table 1.

Table 1: Variables in the data and a summary description of their contents

Variable Name	Description
No.	The index / unique identifier
Gender	The gender of the member (Male or Female)
Date of Birth	The date the member was born

Sponsor Code	The code relating to the specific employer of the member
Join Scheme Date	The date the member enrolled into the scheme
Member Status	The status of the member (Pending, Active, Inactive, or Deferred)
Marital Status	The reported marital status (Single, Married, or Divorced)
DORet	The expected date of retirement (when the member attains 60 years)
Last Contribution Period	The most recent contribution by the member
Last Contribution Date	The date of the most recent contribution
EE	Amount of the most recent Employees Statutory Contribution
ER	Amount of the most recent Employers Statutory Contribution on behalf of the member
EEAVC	Amount of the most recent Employees Additional Voluntary Contribution
ERAVC	Amount of the most recent Employers Additional Contribution on behalf of the member
Balance	Total accrued benefits for the member
Has Next of Kin	Indicates if the member has a next of kin (True or False)
Reason For Exit	The Reason for exit recorded The values are End of Contract, Termination, Resignation, Normal Retirement, Withdrawal, Dismissal, Early Retirement, Death In Service, Transfers, Ill Health, Emigration, Death In Retirement, Restructuring, and Null values for currently active members.

3.3.2 Exploratory Data Analysis

The Data exploration stage involved acquiring a cursory understanding of the data that included a review of the variables in the data, initial statistical analysis, exploratory data analysis, and preliminary visualization of the variables in the data (Schröer et al., 2021). The Exploratory Data Analysis (EDA) included univariate, bivariate and multivariate analyses (Bundi, 2023).

3.4 Data Preparation

The data preparation stage was primarily about making the data ready for the analytical steps (Schröer et al., 2021). The step involved data cleaning, outlier detection, feature transformation and data balancing.

3.4.1 Data Cleaning

The data cleaning task began with reviewing an initial statistical analysis of the variables including the count, mean, standard deviation, percentiles, and for the categorical variables the number of unique occurrences, and the most frequent occurrence (Bundi, 2023).

Then the next step was to identify missing values in the data. The nature of the values was investigated as detailed by (Bundi, 2023) and based on the nature of the missing values the values were imputed or the records were dropped (Kamiri and Mariga, 2021).

3.4.2 Outlier Detection and handling

The next step involved the detection and handling of outliers, as outliers may represent noise (Kamiri and Mariga, 2021) that may affect the pattern learning of the models. The outliers were detected using statistical measures i.e., any value that was outside the range identified by the Interquartile range method (Bundi, 2023) and box plots.

The Interquartile Range (IQR) is a statistical measure that quantifies the dispersion of data within the middle 50%. It is calculated on sorted data as the difference between the 25th percentile (Q1) and the 75th percentile (Q3). The IQR and the limits set to determine potential outliers were (Bundi, 2023):

$$\text{IQR} = Q3 - Q1 \text{Lower limit} = Q1 - 1.5 \times \text{IQR} \text{Upper limit} = Q3 + 1.5 \times \text{IQR} \quad (1)$$

The outliers were treated by applying the winsorization technique applied in the research by (Bundi, 2023) where outliers are replaced by values at the percentile limits.

3.4.3 Data Pre-processing

Data Pre-processing is a key step for machine learning as the quality of outputs from machine learning models is highly dependent on the quality and type of data (Kamiri and Mariga, 2021).

The data preprocessing stage involved the transformation of variables, feature scaling and feature engineering.

The categorical variables including 'Gender', 'Sponsor Code', MemberStatus, and 'Has Next of Kin' were transformed into a form more applicable for machine learning. This was done using label encoding as applied by (Sunday et al., 2020).

For a categorical column with n unique values which can be represented in the following manner:

$$X = x_1, x_2, \ldots, x_n$$

An integer label is assigned to each category, beginning at 0 and increasing by 1. For each category x_i: The label assigned to x_i is (i). The label-encoded value for a category x_i is denoted as

$$L(x_i) : [L(x_i) = i] \tag{2}$$

Feature scaling was done through the MinMax scaling technique to reduce the range of the features in the data to a consistent scale, as implemented by (Kiermayer, 2022) because some machine learning algorithms are sensitive to the magnitude of the values.

The MinMax scaling formula for a value y is given by:

$$y_{\text{scaled}} = \frac{y - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \tag{3}$$

where y_{min} and y_{max} are the minimum and maximum observations in the variable. To scale to a different range, say $[a, b]$, the formula is modified to:

$$y_{\text{scaled}} = a + \frac{(y - y_{\text{min}})(b - a)}{y_{\text{max}} - y_{\text{min}}} \tag{4}$$

Through the use of this transformation, variance in the ranges of values was preserved while bringing all characteristics to a single scale.

Feature engineering involved extracting new features from the available data. The aim is to extract more information for the model to learn from. New features were extracted from the

Date-Time features in the form of elapsed time as noted in the research by (Garriga et al., 2022) including

Table 2: Initial variables and the variables derived from them by Feature engineering

Initial variable name	New variable
Date of Birth	Age
Date of Joining Scheme	Years of membership
Last Contribution Period	Months since last contribution
Reason for Exit	Status

For the target variable, we noted that the 'Reason for Exit' variable had multiple causes for exits from the scheme including End of Contract, Termination, Normal Retirement, Death In Service, Withdrawal, Resignation, Early Retirement, Dismissal, Transfers, Death In Retirement, Emigration, Ill Health, Restructuring. The area of interest for this research was predicting an unexpected withdrawal, i.e., an exit event that occurred before the member attained retirement age, therefore the one vs all method (Abramovich et al., 2021) was applied to reduce the multiple reasons for the exit, identified through the age of the event to a single class of unexpected withdrawal.

3.4.4 Data imbalance

The research investigated the balance of the classes in the target variable. Following the identification of imbalances, the Synthetic Minority Over-sampling Technique (SMOTE) applied by (Zhang et al., 2022) was applied to the data. It randomly generates synthetic data points, fitting the distribution of the minority class to reduce the imbalance (Wang et al., 2021).

Mathematically, a synthetic sample y_{new} is created as follows:

$$y_{new} = y_i + (y_{zi} - y_i) \times \delta \tag{5}$$

Where: - y_i is a vector representing the minority class sample. - y_{zi} is a vector representing the randomly chosen neighbour from the k-nearest neighbours. - δ is a random number between 0 and 1.

SMOTE algorithm can improve the classifiers' performance on imbalanced data by randomly generating new minority sample points thereby creating a more balanced data set for the analysis (Wang et al., 2021).

It was also noted that there were evaluation metrics that were not biased towards the majority class such as the F1-score, precision, recall curves, and the area under the ROC curve (Alshammari, 2023) that can be applied alongside the accuracy of the model.

The data preparation stage output data that was of high quality to enable optimal performance during data modelling. The final step of the Pre-processing was splitting the data to enable in-sample training and in-sample and out-of-sample evaluation of the model. The larger part of the data 70% was applied to train the models and the remainder 30% was to evaluate the performance on unseen data as implemented by (Salazar and Boado-Penas, 2019).

3.5 Data Modelling

The data modelling step used data mining techniques to find patterns and relationships in the prepared data. Supervised learning was used to train a classifier to predict the target class for unseen data instances (Salazar and Boado-Penas, 2019). Supervised learning is a form of artificial intelligence, where the model is trained on a labelled data set. The labelled data set contained examples of the input data and the corresponding target class. The model learned from this data to predict the target class for new, unseen data instances.

The performance of the classifier was then evaluated and validated by a test set. The models explored include:

3.5.1 Logistic Regression

The logistic regression model is a type of statistical analysis applied in the examination of the relationship between a binary dependent variable, and one or more independent variables. The outcome variable sometimes represented numerically as 0 or 1, represents two mutually exclusive categories such as the case of withdrawals. (Salazar and Boado-Penas, 2019) .

(Salazar and Boado-Penas, 2019) found that the Logistic regression model had higher prediction accuracy in the case of withdrawal analysis. The predictor variables can be categorical and or continuous. The logistic regression model estimates the probability of the outcome variable being 1 for a given set of independent values, using a logistic function defined as:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}} \tag{6}$$

where $p(x)$ is the probability of the outcome being 1, e is the exponent function in natural logarithm, β_0 is the intercept, and $\beta_1, ..., \beta_k$ are the coefficients for the predictor variables $x_1, ..., x_k$. The coefficients represent the change in the natural logarithm of the odds of the outcome being 1 for a one-unit increase in the corresponding predictor variable, holding all other predictors constant. The odds are the ratio of the probability of the outcome being 1 to the probability of the outcome being 0.

The values of the coefficients, which increase the likelihood that data will be observed, are therefore used to construct a Logistic Regression Model.

3.5.2 Support Vector Machines (SVM)

This is a Machine learning model that is generally for non-linear relationships between the predictor and the outcome variables.(Sunday et al., 2020)

It aims to find an optimal hyperplane that separates the data points of different classes with the maximum distance. (Xong and Kang, 2019) The hyperplane is defined by a linear function of the form:

$$f(x) = w^T x + b \tag{7}$$

where w is the weight vector and b is the bias term. The optimal hyperplane is the one that minimizes the following objective function:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{8}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n \tag{9}$$

where C is a regularization parameter that controls the trade-off between the margin and the training error, y_i is the class label of the i-th data point, and ξ_i is the variable that measures the extent of classification errors. The data points that lie on or within the margin are called support vectors, and they determine the hyperplane.

A key feature of the SVM model is mapping data to higher dimensional feature spaces in cases where the data is linearly inseparable to separate the features using a kernel function. A kernel function is a function that computes the inner product of two data points in the feature space without explicitly performing the mapping. A common kernel function is the radial basis function (RBF) kernel, which is defined as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{10}$$

where γ is the kernel width control parameter. Using a kernel function, the SVM model becomes:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{11}$$

where α_i are Lagrange multipliers that can be obtained by solving a dual optimization problem.

SVM was noted to be a powerful and flexible machine-learning technique that could handle nonlinear and high-dimensional data with good generalization performance and has been widely used for various applications, including surrender analysis and prediction. (Xong and Kang, 2019)

3.5.3 Random Forests (RF)

Random forests are machine learning algorithms that can be applied to classification tasks. They are a type of tree-based classifier and apply an ensemble learning method that operates by constructing multiple decision trees (Kiermayer, 2022) by creating random samples of the training dataset with replacement, i.e., bootstrapping, and training a decision tree on each sample. It then selects the classification prediction by voting the majority class from individual trees. (Salazar and Boado-Penas, 2019)

The mathematical setting of the random forest is as follows:

Let X be the feature space and Y be the output space.

For a training dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i \in X$ and $y_i \in Y$, the Random Forest algorithm can be formalized as follows:

Randomly select B bootstrap samples $\{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), ..., (x_{in}, y_{in})\}$ from the training dataset.

For each bootstrap sample, grow a decision tree T_b by recursively partitioning the data based on a random subset of features at each node.

Combine the predictions of all trees by majority voting for classification or averaging for regression.

In the classification case:

$$\hat{y} = \text{argmax}_y \sum_{b=1}^{B} I(\hat{y}_b = y) \tag{12}$$

3.5.4 XGBoost

Extreme Gradient Boosting (XGBoost), an ensemble machine learning algorithm used widely for data science problem solving (Chen et al., 2020). It's an implementation of gradient-boosted decision trees (Kiermayer, 2022) on classification and regression (Chen et al., 2020). It has been noted to sometimes perform well on imbalanced data while sometimes it gives mixed results (Chen et al., 2020).

The core principle behind XGB is the idea of creating new models that correct the errors of existing models. Mathematically, it involves the minimisation of a loss function and the addition of regularisation terms to prevent over-fitting.

The methodology of XGB can be described in three main steps:

1. XGB initialises a model with a single tree or a constant prediction. This is represented by the formula

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} l(y_i, \gamma) \tag{13}$$

where l is the loss function, y_i are the true values, and γ is the initial model prediction.

2. Subsequently, new trees are added to the model. Each new tree, $h_t(x)$, is fitted on the negative gradient of the loss function. The update equation is

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x) \tag{14}$$

where η is the learning rate that scales the contribution of each new tree.

3. XGB seeks to improve generalisability by adding regularisation terms $\Omega(h_t)$ into the objective function, which penalises the complexity of the model. The overall objective at each step is to minimise

$$L(F_t) = \sum_{i=1}^{n} l(y_i, F_{t-1}(x_i) + h_t(x_i)) + \Omega(h_t) \tag{15}$$

These procedures are continued until a predetermined benchmark is attained, such as a maximum number of trees or no further improvement in the loss function. As a result, a better model is produced that enhances predictive performance by reducing over-fitting and leveraging the advantages of several decision trees (Zhang et al., 2022).

3.6 Evaluation

The performance of the models was evaluated based on the ability to predict the target variable in an out-of-sample testing set split from the main data set. To test for over-fitting or under-fitting, the model should perform well in seen and unseen data (López et al., 2022), therefore the metrics were evaluated for both splits of the data.

3.6.1 Assessment and Ranking

The models were assessed and ranked based on the following performance metrics: Performance criteria Calculation (Moulaei et al., 2022)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FP} \tag{16}$$

where T stands for True, F for False, P for Positive, and N is negative.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{17}$$

$$Sensitivity/Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{18}$$

$$F1score = 2 * \frac{Precision * Recall}{Precison + Recall} \tag{19}$$

Contained in the Confusion Matrix (Moulaei et al., 2022) as shown below:

Table 3: Design of the Confusion Matrix

	Predicted Values	
Actual Values	True Positive False Positives	False Negatives True Negatives

3.6.2 Model Explainability

Model explainability is a crucial part of ethics in machine learning. It promotes transparency through the understanding and validation of the decisions and builds trust in the models (Belle and Papantonis, 2021).

3.6.2.1 Global Explainability

Global explainability aims to provide insight into the predictions of the model (Belle and Papantonis, 2021). This research applied the feature importance approach which is a model-specific explanation of the relevance of each feature in the model (Belle and Papantonis, 2021).

3.6.2.2 Local Explainability

Local Explainability selects an instance(s) and seeks to clarify the decision of the model based on the available features in the data. The research applied the LIME, one of the more popular techniques (Belle and Papantonis, 2021). The granularity allowed interpretation of how the features influenced the outcome of the prediction of the instance.

3.7 Deployment

The final step involved the deployment of the results of the analysis. The model exhibiting superior performance was selected and then saved into a .pkl file using the pickle module in Python. The Streamlit module in Python was then used to deploy the model. The deployment

consisted of an authentication page and then a page to enter the different variables and have the model predict and display the results as implemented in the study by (Saxena et al., 2021).

3.8 Ethical Considerations

The ethical considerations of applying individualised data include privacy, confidentiality, and informed consent. The data applied did not have any personally identifiable information thereby ensuring privacy and confidentiality considering that the identities of individuals were not required for this research.

Authorisation was sought from and granted by the organisation, that preferred to remain anonymous and the study only applied authorised data for the express purpose of the research. The proposal was submitted for review by Strathmore University's Ethics Review Board and approved.

Chapter 4: Discussion of Results

4.1 Introduction

This section details the results of the machine learning analysis and follows the CRISP-DM methodology (Schröer et al., 2021). The objective was to predict unexpected withdrawals, a critical challenge in the retirement benefits sector that impacts liquidity and revenue forecasting. The model's performance was evaluated using a robust set of metrics.

4.2 Data Understanding

This section delves into the initial data collection, and proceeds to describe the data in detail, providing insights into the various attributes and their respective roles in the predictive model. It also outlines the preliminary steps taken to assess the quality of the data, ensuring that the subsequent analysis is built on a solid foundation of reliable and relevant information. This foundational understanding is essential for developing a robust machine-learning model capable of accurately predicting unexpected withdrawal events.

4.2.1 Data Description

A description of the variables available in the data was provided in the methodology section in table 1.

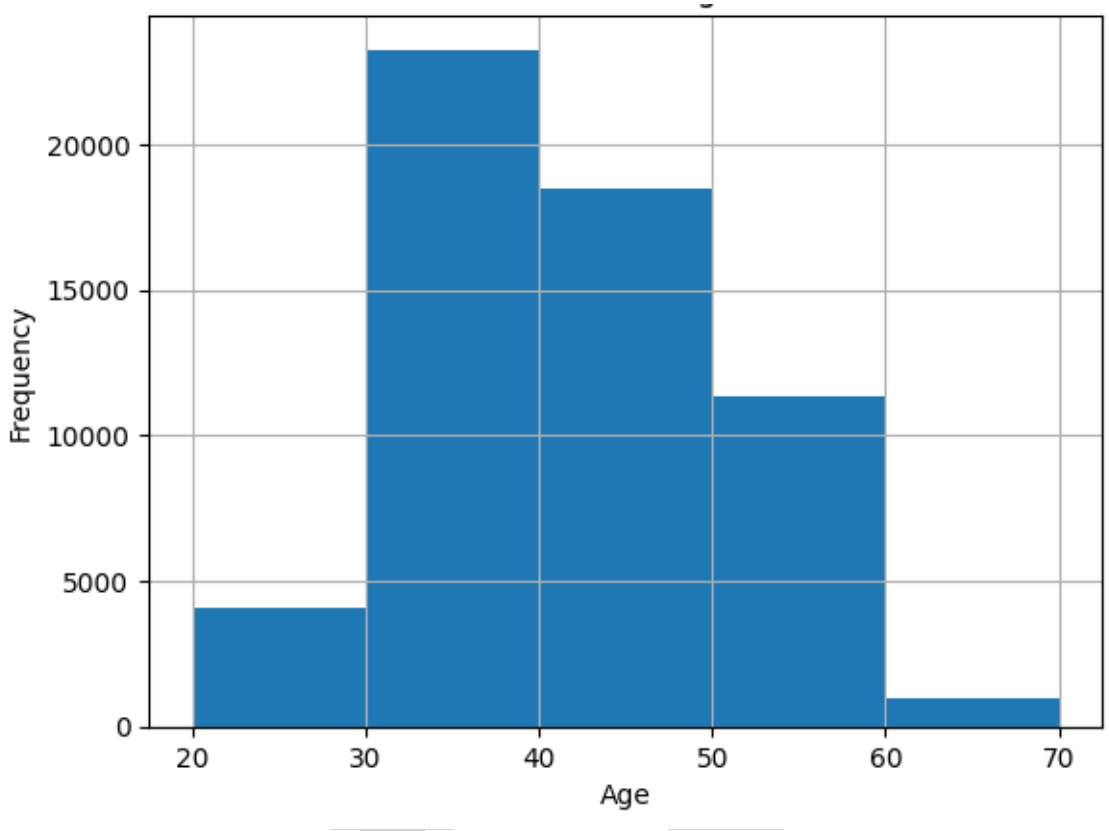
4.2.2 Exploratory Data Analysis

The research explored the data to gain an understanding of the variables, their distribution, and relationships.

1. Univariate Analysis

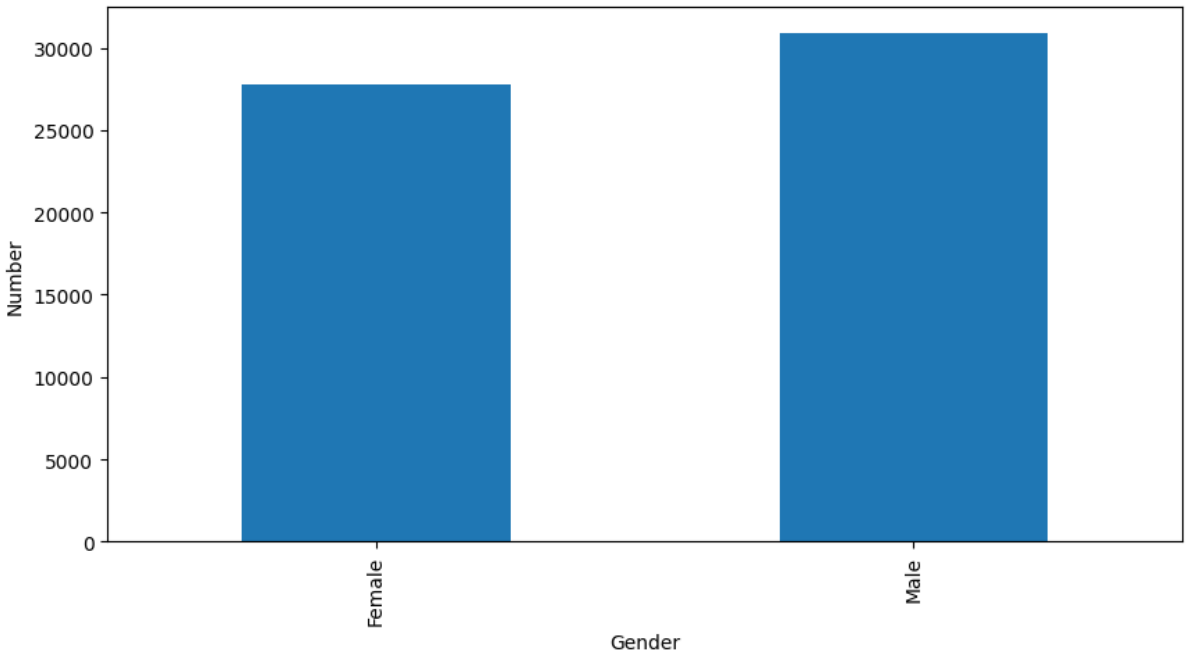
The age distribution of the members in the data was visualised using a histogram. The distribution was noted to be skewed to the right indicating that that there are more people in the scheme who are younger than the average age (41 years) than there are people who are older than the average age.

Figure 4.1: Histogram showing the Distribution of the values in the Age Variable



The membership in the data was also noted to be balanced in the representation of gender with 52% male and 48% female members of the scheme in figure 4.2.

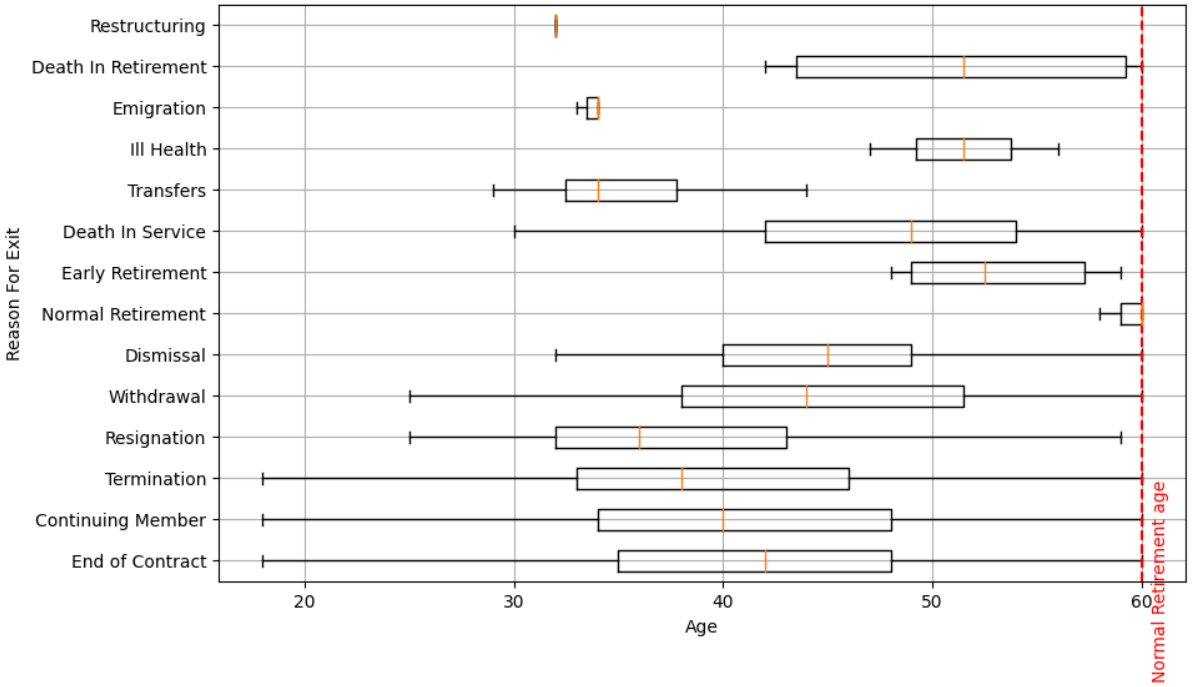
Figure 4.2: Bar chart showing the distribution of the categories Gender Variable



2. Bivariate Analysis

The bi-variate analysis was applied to review the interaction of the age and Reason for exit variables in figure 4.3. it was noted that the 'Normal Retirement' reason for exit had most of its members above the retirement age, with some expected to turn 60 during the following year, and was, therefore, the predictable reason for exiting the scheme. The age of the members in the other reasons for exit was noted to vary significantly up to the age of retirement.

Figure 4.3: Box-plots showing the variation of Age by each Reason for Exit

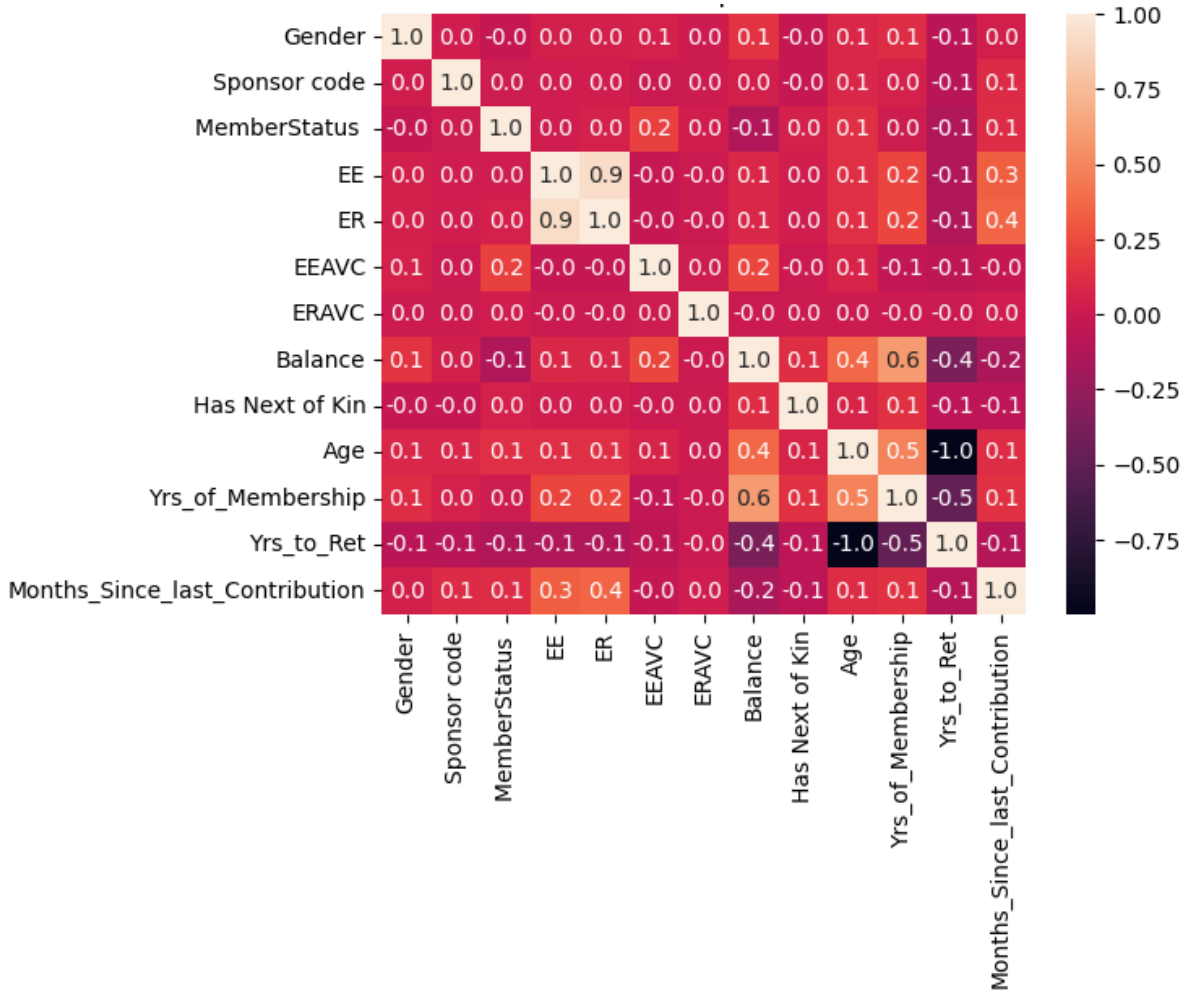


3. Multivariate Analysis

The research reviewed and summarised the multivariate analysis in the form of a correlation matrix of the variables.

There were few high correlations such as a strong negative correlation between age and the years to retirement, and strong positive correlations between the Employee Contributions (EE) and Employer Contributions (ER) as these are based on a percentage of the monthly salary. The negative correlation shows that the older an individual was the fewer the number of years to retirement which is true in practice. The was a moderate positive correlation between the balance and the years of membership as this shows that the longer a member has participated in the scheme the more likely they were to have more benefits accrued.

Figure 4.4: Correlation Heatmap of the variables



4.3 Data Preparation

The results of the data preparation are presented below:

4.3.1 Data Cleaning

The initial statistical analysis revealed that there were missing values in some variables which had a lower count, there were also some data quality issues such as negative values for contributions (EE, ER Additional Voluntary Contributions (AVC)) and balances.

There were 179 different employers in the data as denoted by the sponsor code, and 13 different reasons for exit from the scheme. Most of the members were active as denoted by the MemberStatus variable.

Following the above, missing values were identified using the MissingNo module in python as per figure 4.7.

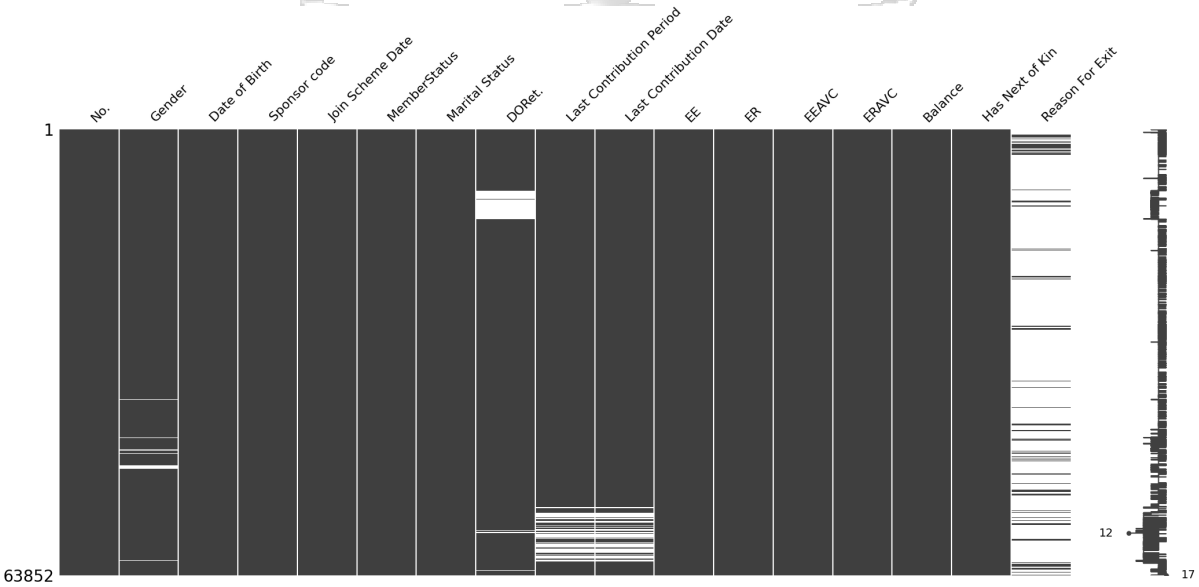
Figure 4.5: Summary Statistics of the Numerical Variables

	EE	ER	EEAVC	ERAVC	Balance
count	58,679.00	58,679.00	58,679.00	58,679.00	58,679.00
mean	14,637.43	19,818.11	68,768.58	164.51	794,094.21
std	84,742.64	109,979.17	443,117.10	18,778.99	1,176,047.89
min	-31,382.07	-14,050.86	-13,387.20	-100,000.00	-8,033,735.72
25%	0.00	0.00	0.00	0.00	90,810.31
50%	3,724.00	4,705.00	0.00	0.00	310,792.22
75%	5,992.00	7,581.53	0.00	0.00	1,014,741.05
max	3,436,983.41	3,971,181.46	52,699,997.03	3,152,908.13	52,699,997.03

Figure 4.6: Summary Statistics of the Nonnumerical Variables

	No.	Gender	Date of Birth	Sponsor code	MemberStatus	Marital Status	DORet.	Reason For Exit
count	63852	63027	63852	63830	63852	63852	59568	7942
unique	63852	2	11390	179	4	3	11221	13
top	AM00576	Male	1980-01-01 00:00:00	SP - 003A	Active	Single	2040-01-01 00:00:00	End of Contract
freq	1	32917	1593	3584	55227	62024	1475	5322

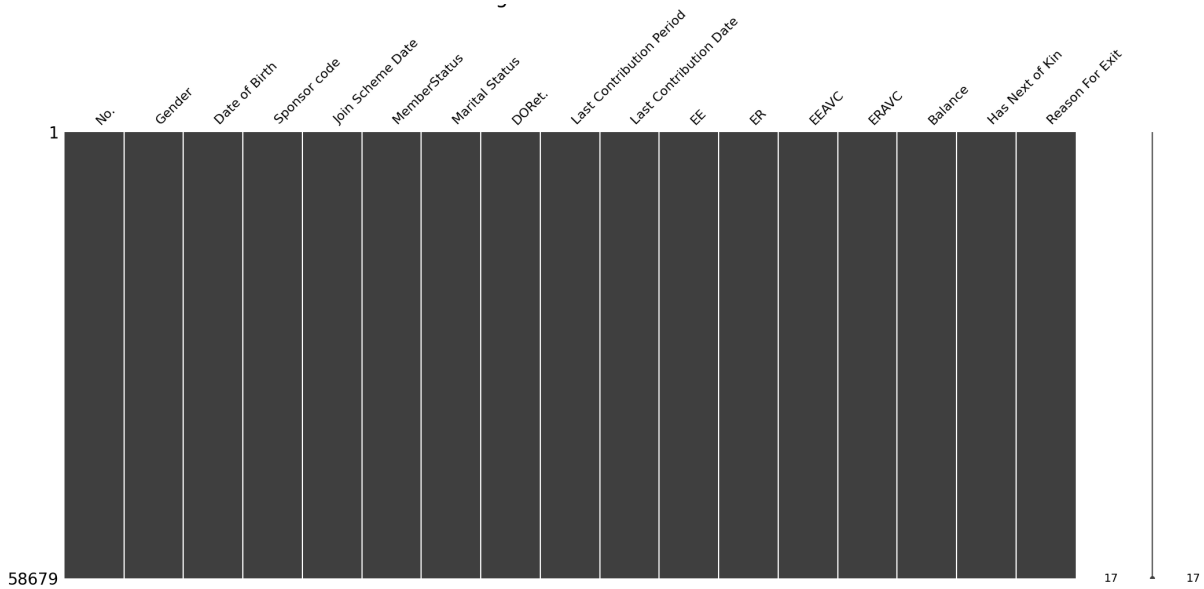
Figure 4.7: Matrix showing the Initial Missing Values per variable



The missing values in 'Reason for Exit' were noted to be members who had not exited the scheme and were therefore imputed as 'Continuing Members', missing values from 'DORet',

which is the expected date of retirement, were imputed based on the corresponding Date of Birth. The variables that had values that were missing completely at random but were a relatively small number were dropped from the analysis (Bundi, 2023). Following this the cleaned values were as per figure 4.8

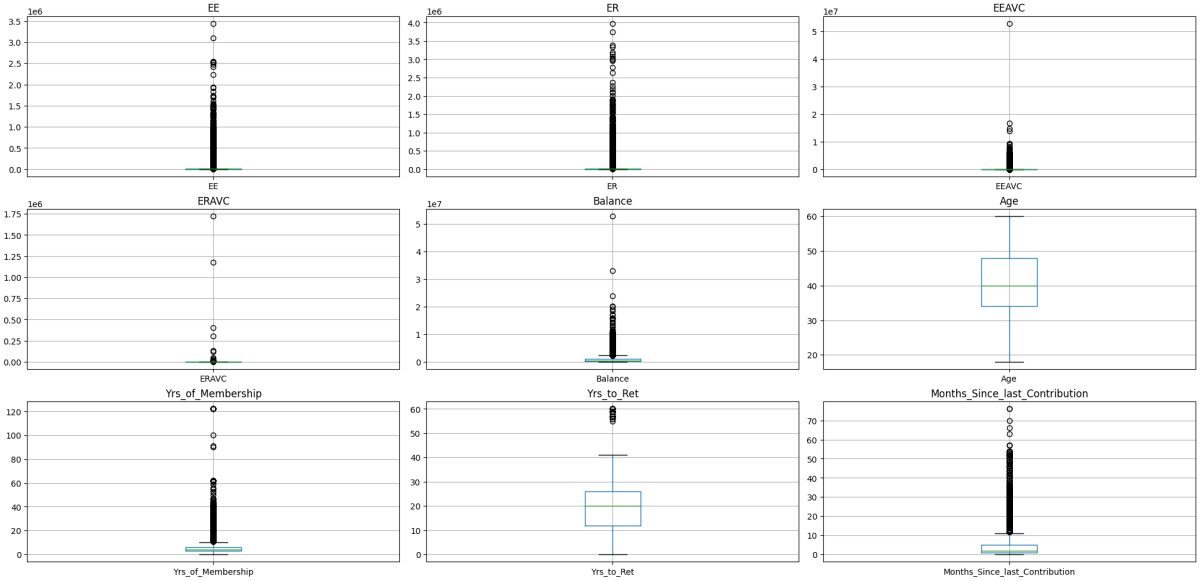
Figure 4.8: Matrix showing Cleaned data without Missing Values



4.3.2 Outlier Detection and Handling

The outliers were detected using the descriptions in figure 4.5 and box-plots and then treated. The negative monetary variables were set to 0 then the Interquartile range method was applied to reduce them (Bundi, 2023).

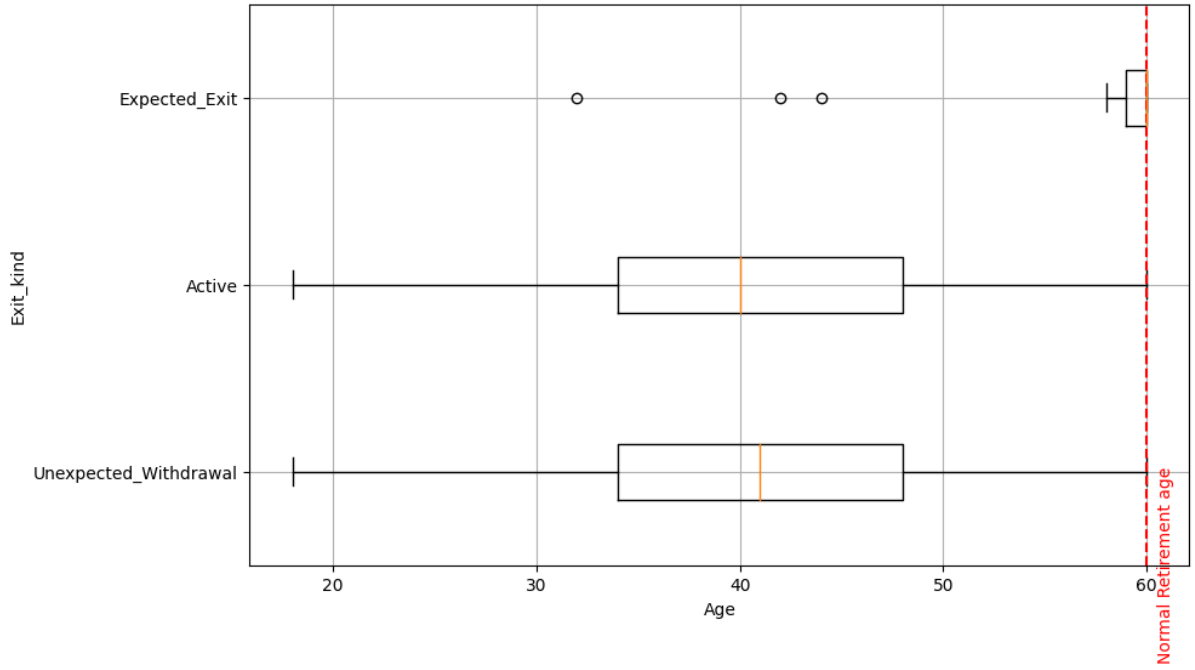
Figure 4.9: Boxplots of the Numerical Variables for Outlier Detection



4.3.3 Data Preprocessing

A variable 'Status' was derived from the 'Reason for Exit' variable whereby the reasons were grouped into Active members, Unexpected Withdrawals, and Expected exits for use in the modelling as shown in figure 4.10. The Expected exits were based on Normal retirements that occur at age 60 (The Parliament of Kenya, 2020) and are therefore deterministic and were later dropped from the predictive analysis.

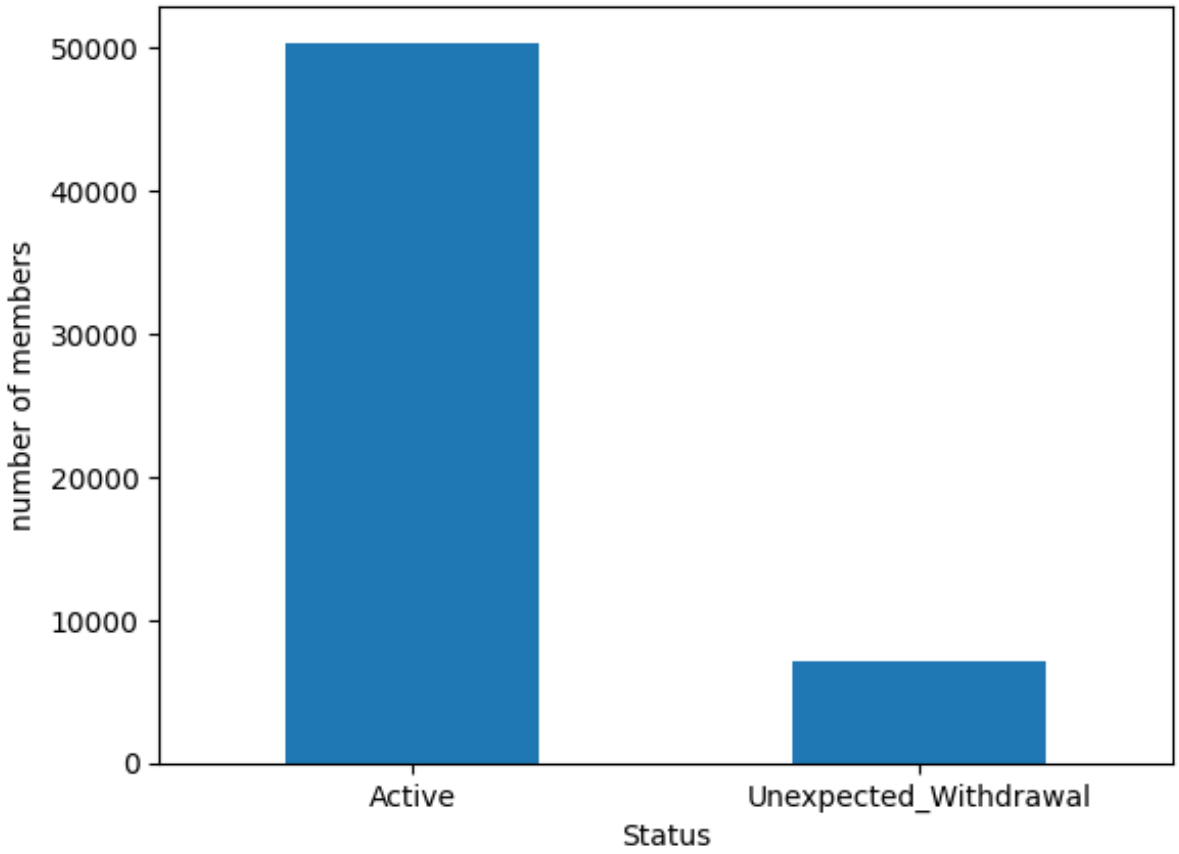
Figure 4.10: Box plots showing the distribution of Age by the different categories in the Status variable



4.3.4 Data Imbalance

The balance of the status variable was investigated using the plot of the number of occurrences of the different classes. It was noted that the distribution was imbalanced with the minority class being the Unexpected withdrawals accounting for 12% of the values as shown in figure 4.11

Figure 4.11: Bar chart showing the imbalance in the categorical Status variable



The data was split into a training and testing set and the imbalance was treated using the SMOTE technique for training the model.

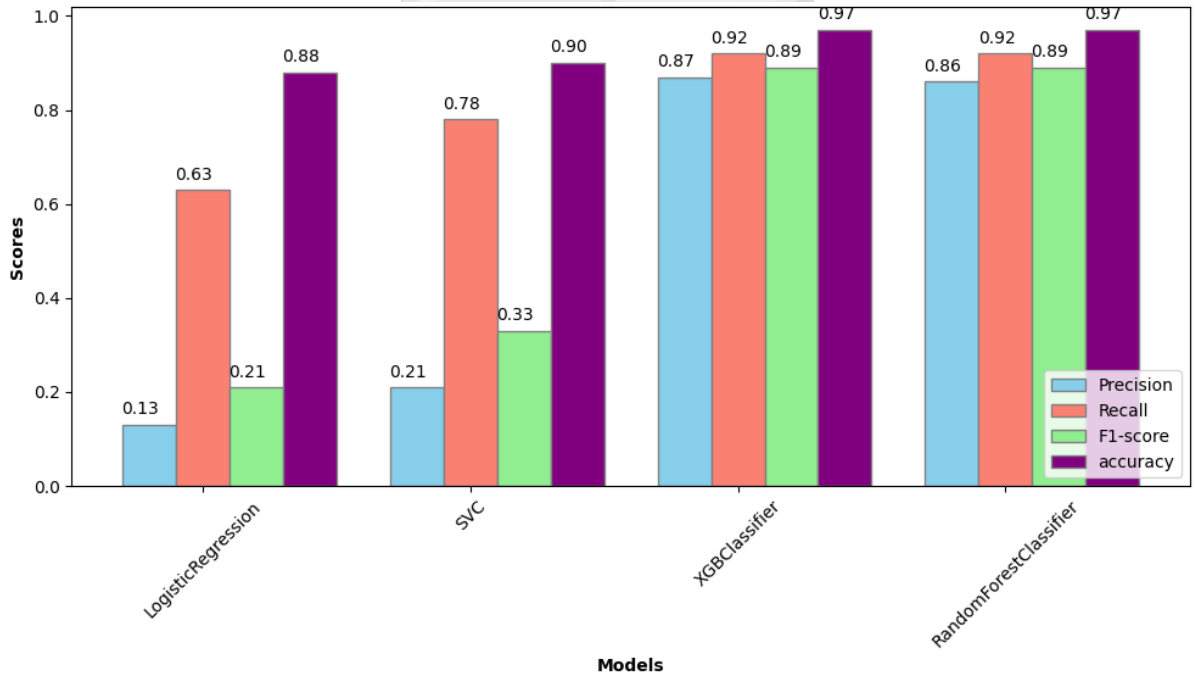
4.4 Data Modelling and Evaluation

This section presents the performance of the models selected for the analysis namely Logistic Regression, SVM, XGB, and RF.

1. Initial Data

   The models were fit on the data before the application of the Interquartile technique to handle outliers and application of the SMOTE technique. The models had high accuracy, above 88% across all classes but the precision, recall, and F1-score were low for the Logistic Regression and the SVM classifier as shown in figure 4.12. The XGB and RF classifiers had excellent performance, with the former having slightly better performance in recall. Therefore overall the XGB classifier had better performance across the metrics.

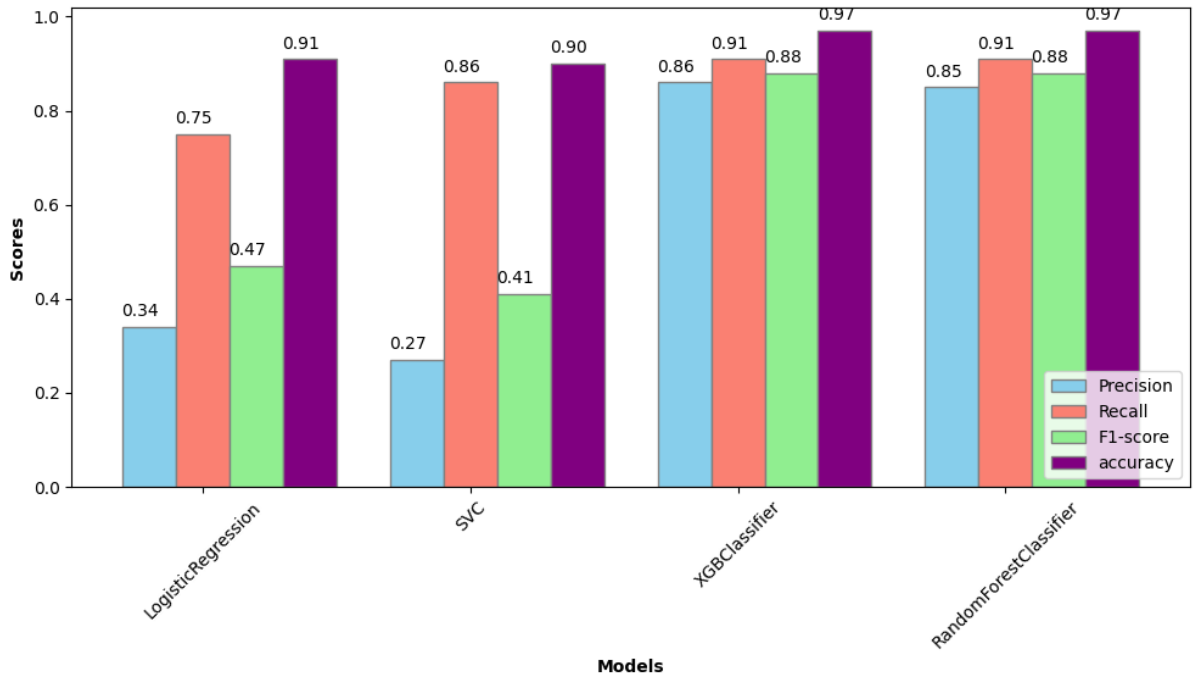
Figure 4.12: Bar chart showing the Performance Metrics of the various models on the Initial Data



2. Dataset without Outliers

   The models were fit on the data without outliers. This improved the performance of the Logistic Regression and SVM models, especially on the precision and recall. The XGB and RF classifiers outperformed the Logistic Regression and SVM. Therefore overall the XGB classifier had better performance across the metrics on this data manipulation.

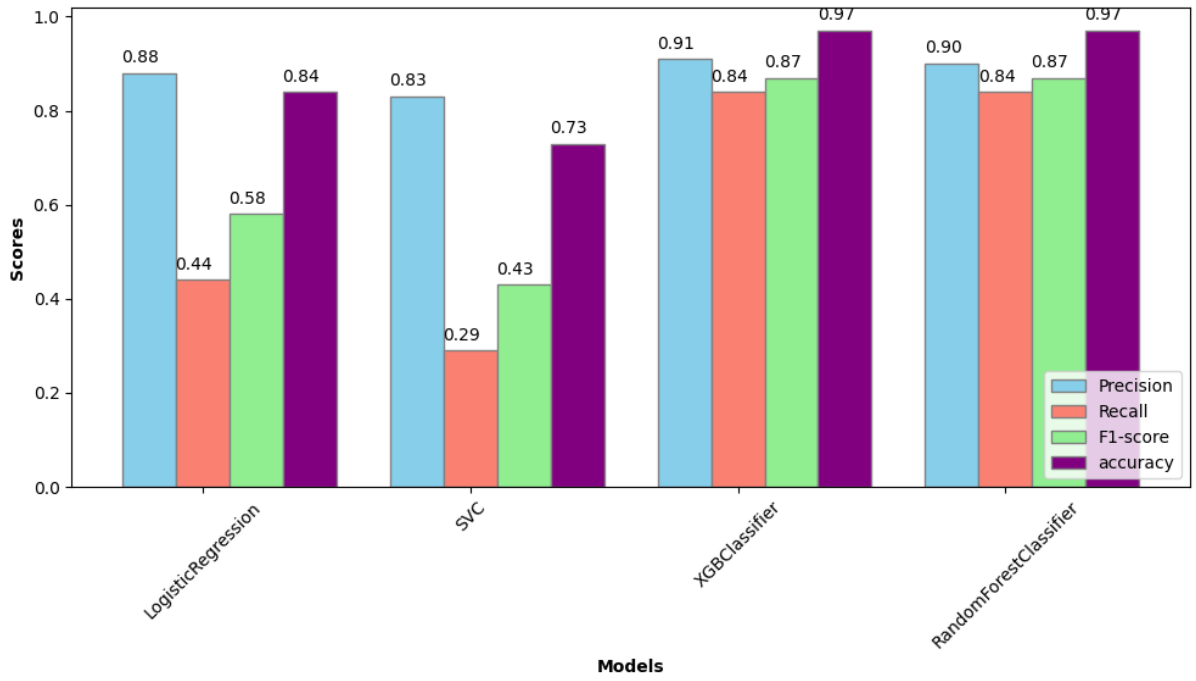
Figure 4.13: Bar chart showing the Performance Metrics of the various models on Data with treated Outliers



3. SMOTE Balanced Data

The models were fit on the data balanced using the SMOTE technique. This improved the precision of the Logistic Regression and SVM models but affected the recall of all the models. The XGB and RF classifiers had very similar performance in this research. Therefore overall the XGB classifier had slightly better performance across the metrics on this data manipulation as shown in figure 4.14.

Figure 4.14: Bar chart showing the Performance Metrics of the various models on the SMOTE Balanced Data

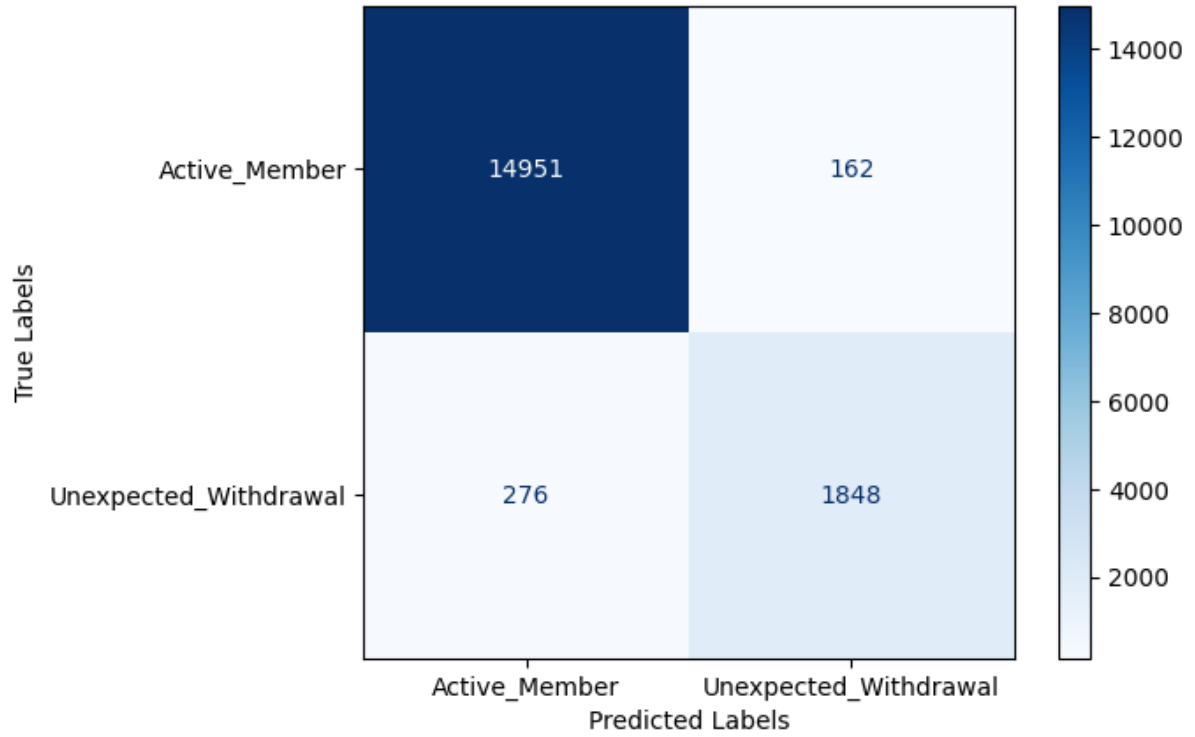


The overall performance metrics of the trained models are shown in table 4. From this, we note that the XGB and RF models consistently perform well on this data but the superior model across the metrics is the XGB trained on the initial data where the outliers and balancing were not implemented. The XGB model was also noted by (Kiermayer, 2022) to give superior performance. The confusion matrix showing the predictive classification of the selected model on the testing set is shown in figure 4.15.

Table 4: Performance metrics for the models trained on the different data manipulations

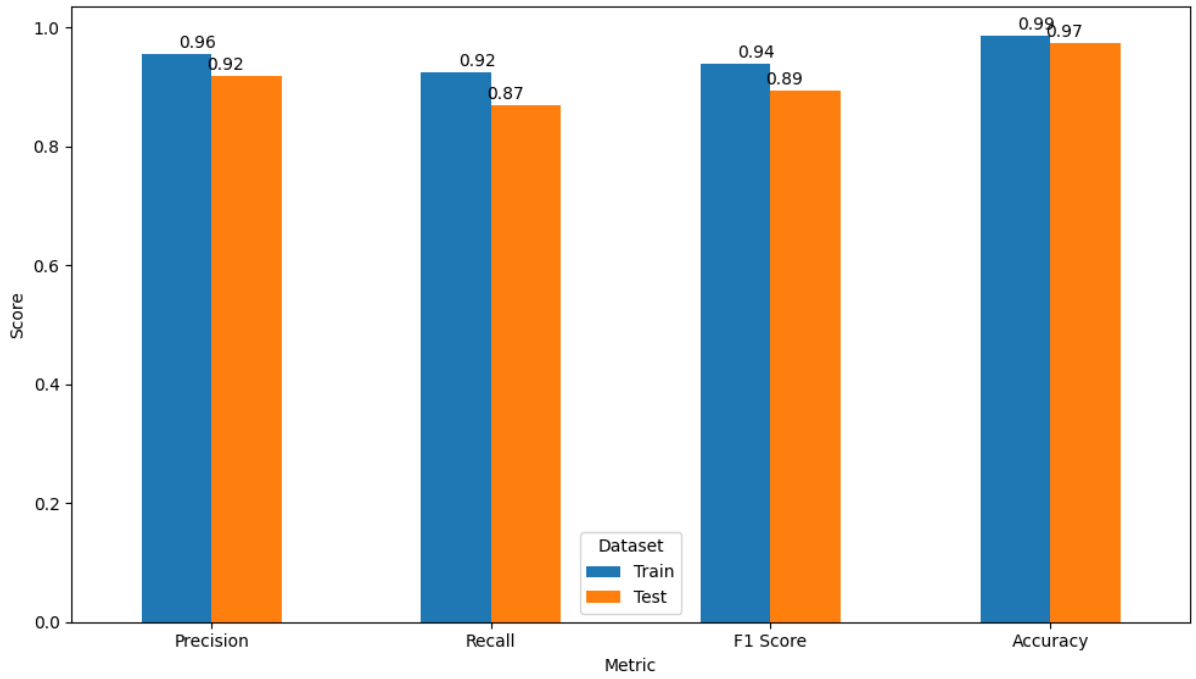
Model	Initial Data				Data w/o Outliers				SMOTE balanced Data			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
LR	88%	13%	63%	21%	91%	34%	75%	47%	84%	88%	44%	58%
SVM	90%	21%	78%	33%	90%	27%	86%	41%	73%	83%	29%	43%
XGB	97%	87%	92%	89%	97%	86%	91%	88%	97%	91%	84%	87%
RF	97%	86%	92%	89%	97%	85%	91%	88%	97%	90%	84%	87%

Figure 4.15: Confusion Matrix showing the number of Actual and Predicted values by the selected XGB model



To check for overfitting or underfitting the model was evaluated for both the training and testing sets. The model generalised well to unseen data as the difference in the evaluation metrics of the training and testing data was very small as seen in the figure 4.16.

Figure 4.16: In-sample and Out-of-sample evaluation Metrics for selected XGB model showing that the model is neither over-fitting nor under-fitting

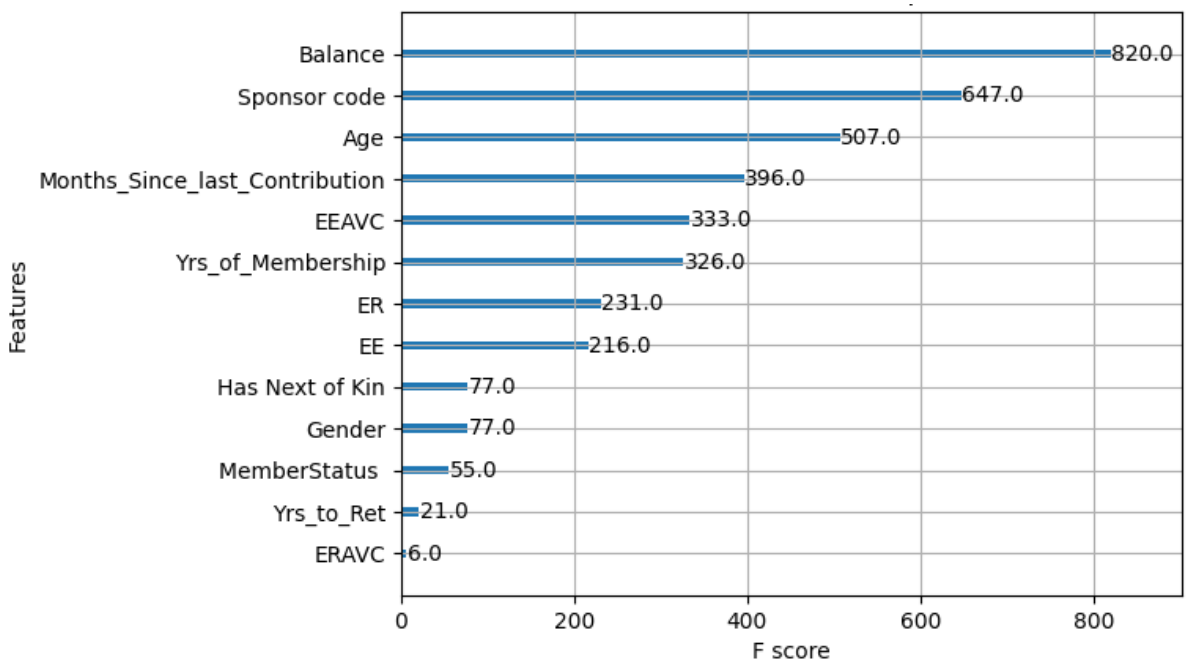


4.4.1 Model Explainability

Following the selection of the model with superior performance, the research sought to investigate the decisions the model made based on the input features as was described in the first research objective.

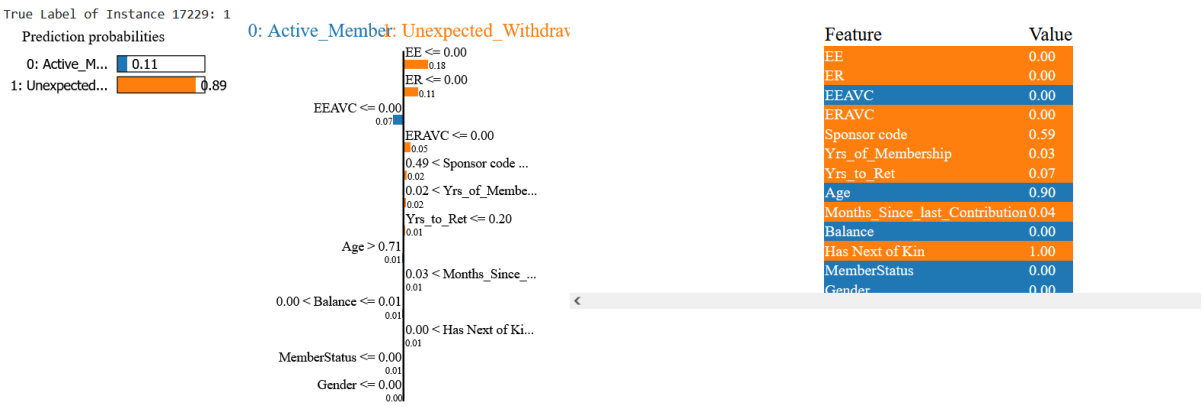
The global explainability was conducted based on the model-specific feature importance which showed the top 5 most important features were Balance, Sponsor Code, Age, Months since Last Contribution, and Employee Additional voluntary contributions. This shows that the features above had the most influence on the unexpected withdrawal event predicted by the model.

Figure 4.17: Global model explainability using a horizontal bar chart of ranked Feature Importance scores for the selected XGB model



The research further investigated the local explainability by reviewing the predictions of singular instances using LIME. The instance was selected from the unseen data and had a true label of 1 which means that the member had withdrawn. The model predicted the withdrawal with a probability of 89%. This also showed that for the model, the values of the Employee Contributions (EE), Employer contributions (ER), Employer Additional Voluntary Contributions (EEAVC), Sponsor Code, and Years of membership contributed to the probability of predicting Class 1 - Unexpected withdrawal while Employees Additional Voluntary Contributions (EEAVC), Age and balance contributed to the probability of predicting class 0.

Figure 4.18: Local explainability of the model predicting an instance showing the importance of each variable in the prediction, using LIME



4.4.2 Findings Coherent with Previous Works

The findings of this research were coherent with previous literature, where (Briere et al., 2022), in their research, noted that there was a propensity to seek access to retirement benefits before attaining retirement age. Studies conducted by (Salazar and Boado-Penas, 2019) and (Baione et al., 2023) noted that the demographic and non-demographic variables affected the withdrawal probabilities of the participants and that machine learning models were able to learn the patterns in the data to provide accurate predictions. (Kiermayer, 2022) found that the XGB algorithm produced superior modelling performance, and also noted that although resampling may improve the performance of the algorithms on some metrics it was possible that it would introduce bias in the predictions.

4.4.3 Findings Contrasting with Previous Works

There were also noted differences with existing literature where (Salazar and Boado-Penas, 2019) found that the Logistic Regression algorithm outperformed the SVM, and RFs contrary to the findings of this research where the RFs outperformed the previous two algorithms. (Xong and Kang, 2019) found that the SVM algorithm had superior performance in their research, but this study found that the XGB and RF consistently outperformed the SVM.

In conclusion, the comparative analysis of machine learning models within this section underscores the robustness and efficacy of the XGB algorithm in fitting the dataset presented. While Logistic Regression, SVM, and RF each displayed particular strengths, it was the XGB model that consistently outperformed the others in terms of accuracy, precision, recall and f1-score. This may be attributed to its ability to handle a large number of features and its use of gradient boosting, which optimises predictive performance.

Chapter 5: Conclusions, Recommendations, and Future Work

5.1 Conclusion

The study showed that the application of the CRISP-DM method developed a machine learning model that was able to accurately predict unexpected withdrawals from a Retirement Benefits scheme. The classification models applied were the Logistic Regression, SVM, RF, and finally the XGB. The XGB model exhibited superior performance as evaluated by the robust metrics of accuracy, precision, recall, and f1-score. The study also identified the importance of each variable in the data that influenced its contribution to the model on both the global scale and local. The results affirmed earlier findings in the literature that showed machine learning models were powerful and able to understand the complex relationships and patterns in the data to identify potential unexpected withdrawals.

5.2 Recommendations

The accurate identification of these withdrawals has the potential to help retirement schemes. We recommend that schemes apply the eXtreme Gradient boosting (XGB) machine learning algorithm to identify potential withdrawals, help plan their resources, and develop targeted outreach programs to incentivise long-term savings. Schemes will be able to better understand and manage their liquidity risk by ensuring that there are adequate cash reserves available to refund the unexpected withdrawals accrued benefits as and when they fall due.

5.3 Future Work

To enhance our research scope, future endeavours should broaden the application of our models to encompass multi-class classification, where each exit reason is predicted individually. This approach will offer a more nuanced understanding of withdrawal patterns. Additionally, extending our research over longer periods, such as rolling 3-year or 5-year spans, will facilitate medium-term planning by enabling predictive insights into withdrawal trends. By analysing withdrawals over time, we can glean valuable insights into behavioural patterns, identify trends, detect cycles, and observe the evolution of data variables.

Furthermore, integrating economic variables into our analysis will provide insights into the impact of external environmental factors on withdrawal behaviour. This holistic approach will enrich our understanding of the complex interplay between individual decisions and broader

economic dynamics, thereby enhancing the predictive accuracy and applicability of our models.

References

Abramovich, F., Grinshtein, V., and Levy, T. (2021). Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory*, 67(7):4637–4646. DOI:10.1109/TIT.2021.3075137.

Alcober, G. M. I., Lagman, A. C., and Revano, T. F. (2020). Predicting the mortality of female patients suffering from myocardial infarction using data mining methods: A comparison. In *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–6. DOI: 10.1109/HNICEM51456.2020.9400093.

Alshammari, T. (2023). Applying machine learning algorithms for the classification of sleep disorders. *IEEE Access*.

Baione, F., Biancalana, D., and De Angelis, P. (2023). A two-part beta regression approach for modeling surrenders and withdrawals in a life insurance portfolio. *North American Actuarial Journal*, 27(2):380–395.

Bédard-Pagé, G., Bolduc, D., Demers, A., Dion, J.-P., Pandey, M., Berger-Soucy, L., and Walton, A. (2021). Covid-19 crisis: Liquidity management at canada's largest public pension funds.

Belle, V. and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*. DOI: 10.3389/fdata.2021.688969.

Bjerre, D. (2022). Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin*, 52:1–23. DOI: 10.1017/asb.2022.11.

Briere, M., Poterba, J., and Szafarz, A. (2022). Precautionary liquidity and retirement saving. In *AEA Papers and Proceedings*, volume 112, pages 147–50.

Broeders, D. W., Jansen, K. A., and Werker, B. J. (2021). Pension fund's illiquid assets allocation under liquidity and capital requirements. *Journal of pension economics & finance*, 20(1):102–124.

Bundi, L. E. (2023). Application of machine learning in customer analytics.

Chen, W., Chengyuan, D., and Suzhen, W. (2020). Imbalance-xgboost: leveraging weighted and focal losses for binary label-imbalanced classification with xgboost. *Pattern Recognition Letters*, 136. https://DOI.org/10.1016/j.patrec.2020.05.035.

Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., and Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*. https://DOI.org/10.1038/s41591-022-01811-5.

Hamilton, S., Liu, G., and Sainsbury, T. (2023). Early pension withdrawal as stimulus. *Available at SSRN 4389699*.

Hou, W. et al. (2022). How well do retirees assess the risks they face in retirement? Technical report.

Kamiri, J. and Mariga, G. (2021). Research methods in machine learning: A content analysis. *International Journal of Computer and Information Technology*, 10.

Kiermayer, M. (2022). Modeling surrender risk in life insurance: theoretical and experimental insight. *Scandinavian Actuarial Journal*, 2022(7):627–658.

Lagat, C. A. (2019). An inclusive pension model for kenya's informal sector with late entries and early exit rates.

López, Osval Antonio, M., López, A. M., and Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction*.

Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., and Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting covid-19 mortality. https://DOI.org/10.1186/s12911-021-01742-0.

Panis, C. W. (2019). Contagion of mass withdrawals from multiemployer pension plans.

Salazar, J. d. J. R. and Boado-Penas, M. d. C. (2019). Scoring and prediction of early retirement using machine learning techniques: application to private pension plans. In *Anales del Instituto de Actuarios Españoles*, pages 119–145.

Saxena, A., Dhadwal, M., and Kowsigan, M. (2021). Indian crop production: prediction and model deployment using ml and streamlit. *Turkish Journal of Physiotherapy and Rehabilitation*, 32:3.

Schröer, C., Kruse, F., and Marx Gómez, J. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534. DOI: 10.1016/j.procs.2021.01.199.

Sunday, A., Roseline, I., John, O., and Ademola, A. (2020). Accuracy of machine learning models for mortality rate prediction in a crime dataset. pages 150–160.

The Parliament of Kenya (2020). Public Service Commission Regulations, 2020 -Sect 70. Kenya Gazette. Available at: https://www.publicservice.go.ke/index.php/publications/acts-legislation?download=282:the-public-service-commission-regulations-2020.

Wang, S., Dai, Y., and Shen, J. e. a. (2021). Research on expansion and classification of imbalanced data based on smote algorithm. *Scientific Reports*. https://DOI.org/10.1038/s41598-021-03430-5.

Xong, L. J. and Kang, H. M. (2019). A comparison of classification models for life insurance lapse risk. *International Journal of Recent Technology and Engineering*, 7(5):245–250.

Zhang, P., Jia, Y., and Shang, Y. (2022). Research and application of xgboost in imbalanced data. *International Journal of Distributed Sensor Networks*. DOI:10.1177/15501329221106935.

Appendices

Appendix I: Ethical Approval

Strathmore UNIVERSITY

15th March 2024

Mr Macharia Simon,
simon.macharia@strathmore.edu

Dear Mr Macharia,

RE: Predicting Unexpected Retirement Benefit Withdrawals using Machine Learning Algorithms

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2068/24.** The approval period is from **15th March 2024 to 14th March 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
15-Mar-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix II: Similarity Report