University of Pretoria

Master's Dissertation

Application of machine learning to retirement fund preservation

Identifying significant variables in retirement fund preservation decisions

Author:
Liezel Oberholzer
Student number: 18022431

Supervisors:
Prof Frederik Jakobus Conradie Beyers
Ms Marli Venter

A dissertation submitted in partial fulfilment of the requirements for the Master's Degree in Actuarial Science
in the Department of Actuarial Science
in the Faculty of Natural and Agricultural Sciences
at the University of Pretoria

2024

DATE OF SUBMISSION: 23 SEPTEMBER 2024

Abstract

This study aims to understand the retirement fund preservation field and determine which factors lead to low preservation of retirement funds. In addition, the study aims to build a machine learning model that classifies the retirement fund preservation data.

The study applied feature engineering to the preservation of retirement fund data from a large insurer in South Africa. The three feature-engineering methods applied were Ordinal Encoding, Dummy Encoding and Target Encoding. These methods were applied to build the three models: Logistic Regression, Random Forest and a Support Vector Machine (SVM).

All three models can accurately predict whether an individual will preserve or not. The random forest overall performed best but had the lowest precision. The SVM produces the highest precision of the three models.

The results from the logistic regression and the random forest showed that individuals who preserve part of the amount paid to them and take the other part in cash have better preservation rate than those who preserved their full amount or did not preserve at all. This is a strong indicator because it shows that if individuals can preserve more and still take a part of their funds in cash the overall preservation of their retirement funds is good.

This study could benefit the industry through identifying variables to focus on to improve the individual's preservation of their retirement funds.

***Keywords**: Retirement, Preservation Retirement Funds, Machine Learning, Logistic Regression, Random Forest, Support Vector Machine*

Declaration

I, Liezel Oberholzer, declare that this dissertation represents my own work for the fulfilment of the degree MSc Actuarial Science at the University of Pretoria, and has not been previously submitted by me for a degree at this or any other tertiary institution.

Date: 16 September 2024

Acknowledgements

My wish for this dissertation was to produce a piece of truth that will act as a form of guidance in the retirement preservation field. This led me to so much more than only retirement preservation.

The last three years of research has been a fluctuating path. It is a monumental opportunity to contribute a tiny piece of research to the numerous studies of retirement preservation and machine learning out there.

However, this dissertation is nothing without the village behind it. I would like to extend my gratitude to everyone who stood by me through this journey.

Thank you to each friend, tutor or family member for your words of encouragement, your prayers or just your interest in my research. It meant more to me than I can express.

Thank you to my supervisors, Prof FJC Beyers and Ms M Venter, for creating the opportunity for me to work with remarkable data. It made all the difference for the dissertation.

To the insurance company that provided the data, thank you for your trust. I am exceedingly grateful for this opportunity to work with real-world data.

I especially want to thank the two employees who assisted me through-out the study. Thank you for your patience with me while I tried to make sense of the data. The models would not have been nearly as good were it not for the two of you. I appreciate the time you took to answer all my never-ending questions. Thank you for your kindness. I will not forget what you have done for me.

I would like to extend my sincerest gratitude to my friend Ross Meyer. Thank you for lending a helping hand.

Thank you to Dr R Malan for editing my dissertation. I know this is a tough task. I appreciate your care and effort.

To my parents, thank you. I would not have any of these opportunities were it not for you.

To my aunt, I don’t know how to thank you. You went above and beyond to assist me. When the days got tough you helped keep me calm. Your guidance on structuring documents were very insightful. Thank you.

Lastly, but certainly not least, to my mother. You believe in me every day even when I lose hope. When nobody did, you fought for me. I know I would not be able to do any of this without you. Thank you for your constant support, encouragement and love. You have given me so much.

Abbreviations and Acronyms

LCH	Life-Cycle Hypothesis
BLCH	Behavioural Life-Cycle Hypothesis
DB	Defined Benefit
DC	Defined Contribution
SA	South Africa
US	United States
RR	Replacement Ratio
AI	Artificial Intelligence
OR	Odds Ratio
CART	Classification And Regression Trees
OOB	Out-of-Bag
SVM	Support Vector Machine
SV	Support Vector
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve

TABLE OF CONTENTS

List of Tables

List of Figures

1. INTRODUCTION

Insufficient retirement savings is a worldwide problem (Alemanni & Lucarelli, 2017). Several factors could contribute to insufficient retirement savings, including postponing the choice to save, failing to save enough, or withdrawing retirement savings when changing employers, i.e. not preserving retirement funds (Reyers et al., 2014). Not preserving the withdrawn retirement funds before retirement is an issue that has recently become much more recognised (Reyers et al., 2015). Subsequently, the number of studies focusing on this topic is growing (Ghilarducci et al., 2019; Hurd & Panis, 2006; Reyers et al., 2014, 2015).

Worldwide, a crucial investigation is being done into the improvement of the preservation of retirement funds (National Treasury, 2021a; Reyers et al., 2014). The lack of retirement fund preservation is an issue faced by many countries (Pratt, 2010; Reyers et al., 2014). Many South Africans use their retirement funds that are paid out if they are retrenched or resign from their current employers. The National Treasury (2021a) has identified that low preservation of retirement funds in South Africa is one of the main concerns regarding the retirement system.

Membership of retirement funds is either compulsory or voluntary. Compulsory funds are occupational funds to which employees must belong as a condition of their job. There are three main saving vehicles in the South African retirement system: pension funds, provident funds and retirement annuities. Pension and provident funds are funds to which regular contributions are made by the employer and/or the employee. The primary distinction between the two is how the members get their benefits upon retirement. A *pension fund* pays out up to a third of the pension benefits when a member retires or changes employers and the remaining balance should be used to buy an annuity or "income". Currently, a *provident fund* is similar to a pension fund, but before 1 March 2021, they differed. Provident funds used to pay out the entire provident fund benefit when a member retired or changed employer. *Retirement annuities* are funds that members voluntarily join and to which they make contributions; this is usually done independently from the employer. The most that can be paid out as a lump payment upon retirement, is one-third of the fund. (Momentum Corporate, 2021; Old Mutual, 2021; Pillay & Fedderke, 2022)

There is a lack of understanding of the factors that affect the preservation of retirement funds. Even though the topic is being recognised as significant, there are still minimal studies that explain what drives low preservation of retirement funds (National Treasury, 2021a; Reyers et al., 2014, 2015). If the factors that affect retirement fund preservation decisions are understood, optimal interventions can be established.

Many countries try to intervene in individuals' retirement savings decisions. Some countries have regulated their pension funds and other countries use tax to discourage certain decisions. Currently, South Africa has tax penalties in place to dissuade individuals from withdrawing retirement funds. The tax penalties assume that individuals will always behave rationally, but Reyers et al. (2015) show that individuals do not always behave rationally. Thus, the tax penalties aren't effective in encouraging individuals to preserve their retirement funds. (National Treasury, 2021a; Reyers et al., 2014, 2015)

Understanding what factors affect retirement fund preservation will enable companies to predict whether certain types of individuals will preserve or not preserve their retirement funds. A suitable model is required to classify an individual in this regard. Machine learning is an excellent prediction tool with the ability to classify data (Alzubi et al., 2018; Liu & Xie, 2019).

Machine learning is very effective today with the rapid growth of volumes of data. It is a tool that can make sense of large volumes of data. Machine learning is widely employed today and is the foundational idea for intelligent systems (Liu & Xie, 2019; Talwar & Kumar, 2013; E. Zheng et al., 2017).

Machine learning is a field in AI that gives computers the ability to reason for themselves. The idea behind machine learning is for computers to learn from their actions and improve their efficiency. The computer learns to adapt its actions to increase the accuracy of the predictions and/or classifications it makes (Alzubi et al., 2018). Concerning retirement fund preservation, machine learning can therefore be used by interested parties to classify individuals. Machine learning is expected to be able to model whether an individual will preserve the retirement funds, even taking into account all the differences between individuals.

Due to the lack of understanding of what drives preservation of retirement funds decisions, this study has set out to determine which factors affect retirement fund preservation significantly. Additionally, due to machine learning being an excellent classification tool, the study is apply machine learning models to retirement fund preservation data to determine if machine learning can classify this data.

1.1 Research Objectives

Since the preservation of retirement funds is not sufficiently recognised as a factor contributing to insufficient retirement savings, this dissertation will explore:

- *The factors that play a significant role in the preservation of retirement fund, highlighting behaviour as a major factor.*
- *Additionally, machine learning will be used to classify the preservation data and to determine which factors play a significant role in the preservation of retirement funds.*

The benefit the study could have is to improve the understanding of retirement fund preservation. Additionally, suitable intervention can be established to encourage retirement fund preservation. The study will also show whether machine learning is a technique that can classify retirement fund preservation data.

The research objectives are:

1. To understand the field of the preservation of retirement funds.
2. To determine which factors significantly drive the preservation of retirement funds.
    - The factors that will be studied are the variables present in the data provided.
3. Apply machine learning methods (Logistic Regression, Random Forests, and Support Vector Machine) to analyse the preservation of retirement funds .
4. To apply machine learning to retirement fund preservation data to determine whether machine learning can classify the data.
5. The study will also investigate whether the models can predict overall preservation with given factors, and whether the models can detect any patterns in the data.

1.2 Assumptions and Limitations of the Study

This study assumes that employed individuals are able and willing to save for retirement.

The study acknowledges that the lack of retirement fund preservation is not the only driver of low retirement savings.

Individuals have unique financial circumstances; thus many varied reasons may affect an individual's decision to withdraw retirement funds, and there is no one-size-fits-all assumption. In the data used to build the models, these different circumstances are not apparent. For example, many young employed individuals are known to have a lot of debt. If such individuals withdrew their retirement funds, it would not necessarily be the 'wrong' decision to take the cash and pay off their debt (Reyers et al., 2015).

In the study, a sample of South African retirement fund members is used to create the models. The results are therefore not necessarily relevant to members from other countries and/or other funds.

It should also be noted that the identified significant factors may not always stay the most significant.

1.3 Structure of the Study

Following the introductory research, the second and third chapters of the study present the background of retirement fund preservation and machine learning. Relevant literature on the topics is presented.

In chapter 4, the research methodology is discussed, the data used for the model is described as well as the way in which the models are built. Model validation is also discussed in this chapter. Chapter 6 presents the results delivered by the models. Chapter 7 discusses and analyses the results from chapter 6. In the final chapter, the study is concluded and recommendations for future studies are made.

2. RETIREMENT SAVINGS

The following chapter presents the background of retirement savings, the theoretical framework of saving behaviour, the importance of retirement planning, the switch from defined benefit retirement schemes to defined contribution retirement schemes and lasatly the preservation of retirement funds overall and in South Africa.

2.1 Background of Retirement Savings

Retirement is a significant step in an individual's life (Kelly & Swisher, 1998). Saving for retirement is probably one of the most important life decisions from a personal and financial perspective (Balasuriya et al., 2014; Kumar et al., 2019).

The lack of sufficient retirement savings is a growing problem in society. Insufficient retirement savings make individuals more vulnerable (Alemanni & Lucarelli, 2017; Kapoor & Prosad, 2017).

South Africa has a low rate of retirement savings. There aren't many South Africans who have saved enough money to support themselves in retirement when they reach the typical retirement age of 65 (Dhlembeu et al., 2022; National Treasury, 2012a). This is not only a problem in South Africa; it occurs worldwide (Alemanni & Lucarelli, 2017). For example, in 2014 more than 11 million people in the UK did not have enough savings for their retirement. Among a survey of 15 countries with insufficient retirement savings, the UK performed the worst. Their retirement savings covered only 37% of the required retirement provision. This could in part be due to the decreasing enrolment in pensions – both employer plans and private schemes (Balasuriya et al., 2014).

In America, it was found that employees moved from conventional pension plans to participant-directing 401(k) plans (Fisch et al., 2016), a retirement savings plan commonly offered by American employers. The two distinguishing features of 401(k) plans are voluntary participation and the fact that both employers and employees can make pre-tax contributions to the plan. By choosing whether to participate, how much to contribute, and how to invest the funds,

the employee has the important responsibility for providing for retirement. The decisions regarding participation and contributions are crucial because pension income frequently acts as a dividing line between individuals who are poor and those who have appropriate retirement income. When it comes to 401(k) programmes, the objective is to make sure that those who are eligible to participate do so and contribute as much as they can (Munnell et al., 2001). This has left American employees with the challenge of personally saving for retirement. Thus, individuals now have to make their own investment decisions. Since this shift took place, it has been evident that individuals make suboptimal investment decisions. This shift was even named "the greatest retirement crisis in history" in America. It is therefore important to understand why individuals are unable to make optimal investment decisions (Fisch et al., 2016).

Most employees are not actively involved in their retirement planning. They put little or no thought into their investment decisions. This is illustrated by the bankruptcy of the company Enron (Bailey et al., 2003). Enron, an energy-trading utility company based in Houston Texas, filed for bankruptcy in December 2001, after they had committed accounting fraud (Benston & Hartgraves, 2002). Most of the employees' investments were in Enron's company stock (Bailey et al., 2003). This led to many employees' retirement savings being drained and destroyed (Benston & Hartgraves, 2002). Thus, many individuals in America were left with no employment and no money for retirement (Bailey et al., 2003). Although this is a rather extreme example, it illustrates that individuals do not make the best investment decisions for themselves (Bailey et al., 2003). The Enron case is also an example of the effect of the change from defined benefit plans to defined contribution plans. The change from defined benefit plans to defined contribution plans are discussed later in the study.

South Africa was ranked as the worst of 30 countries for financial competency in a study by the Organisation for Economic Co-operation and Development (INFE, 2016). The Human Sciences Research Council recorded in its 2017 report that 48% of South Africans do not save at all, whereas 42% do not have any long-term savings. South Africans do not know how much to save for retirement, nor fully understand the role of inflation and interest rates on their contributions over time (Roberts et al., 2014). South Africans are showing that they are not taking advantage of opportunities to make provision for a life of financial security (Sanlam, 2023a).

2.2 A Theoretical Framework of Saving Behaviour

The theoretical foundations of various savings theories must first be examined to discover what elements may have an impact on retirement preservation decisions. Over the years there have been multiple savings theories. A common theme amongst most of them is the inclusion of psychological factors contributing to an individual's savings behaviour (Reyers et al., 2014).

This study is interested in the factors that affect retirement fund preservation decisions; however, it is still important to provide the framework of savings theories for these decisions.

2.3 Development of Savings Theories

2.3.1 Foundation for Modern Economic Saving Theories

The intertemporal choice was of great significance to economists in developing saving theories. Intertemporal decisions are choices between costs and benefits that occur at multiple points in time. Intertemporal decisions are significant and frequent. The foundation of the development of theories based on intertemporal choice was done by John Rae, William S. Jevons, Herbert S. Jevons, N.W. Senior, Eugen von Böhm-Bawerk and Irving Fisher (Frederick et al., 2002).

John Rae released "The Sociological Theory of Capital" in 1834. He was regarded as the father of intertemporal choice modelling. The subject of intertemporal choice originated from Rae. He was also the first to investigate and produce a detailed discussion of psychological factors influencing intertemporal choice (Frederick et al., 2002). According to Rae, people's "effective desire of accumulation" explains their time preferences. The two main factors affecting an individual's desire for accumulation are bequest motive and self-control. Bequest motive refers to accumulating wealth for future heirs to inherit. The limiting factors of his idea were the precariousness of human life and the concern for immediate consumption (Frederick et al., 2002; Loewe, 2006).

Building from Rae's work, William S. Jevons and his son, Herbert S. Jevons, argue that individuals only care about their immediate utility, and they use the utility from the prediction of future consumption to explain proactive behaviour (Frederick et al., 2002). The essential principle is that one should spread out consumption through time to maximise overall enjoyment rather than overconsume during one period at the expense of subsequent ones (Loewe, 2006). It is known as the pleasure of deferral.

In contrast to this opinion, N.W. Senior states that deferral causes pain. Senior is the most well-known proponent of the abstinence perspective. He claims that among human endeavours, depriving oneself of immediate pleasures to pursue long-term goals is one of the most painful. His theory shares the idea with Jevons that current/immediate emotions will drive decisions between costs and benefits at different points in time (Frederick et al., 2002).

Eugen von Böhm-Bawerk added to the theories of Rae, Jevons and Senior. He argues that individuals have a regular tendency to undervalue future desires, particularly distant ones (Frederick et al., 2002; Loewe, 2006).

Building on Böhm-Bawerk's work, Irving Fisher formalises time preference. He does it according to economic trade-offs between consumption in two distinct periods. According to Fisher, each person has a unique rate of "impatience" that is influenced by both subjective and objective elements. Objective considerations include the magnitude and risk of future income. Subjective considerations include foresight, willpower, habit, uncertainty, selfishness, and the influence of fashion. However, Fisher's theory attempts to describe the equilibrium in the financial market, and he believes that the market interest rate (through borrowing and lending) will homogenise all variable rates of impatience (Loewe, 2006).

These ideas saw intertemporal decisions as the result of numerous opposing psychological factors working together (Frederick et al., 2002). Fisher, Böhm-Bawerk and Jevons did not set out to achieve a specific mathematical function demonstrating a preference for a time in general (Loewe, 2006). This changed when Paul Samuelson introduced his theory, the discounted-utility (DU) model, in 1937 (Frederick et al., 2002; Loewe, 2006). Samuelson built a mathematical

structure that made it possible to illustrate intertemporal choice for multiple periods (Loewe, 2006). In his model, there was one parameter – the discount rate. This parameter combined all the psychological factors (Frederick et al., 2002). His model evolved into the established method for intertemporal choice in economics (Loewe, 2006).

Over time, however, irregularities were identified. The rival theory was hyperbolic discounting, which was primarily studied by Thaler. This theory concludes that individuals will choose a smaller immediate reward rather than wait for a better reward (Loewe, 2006; McKerchar & Renda, 2012).

2.3.2 Modern Economic Theories of Saving

The theories of intertemporal choice and the discounted-utility model formed the basis for modern economic theories of savings.

The three main modern theories that have been designed by economists to help explain how people spend and save their money are the absolute income hypothesis, the permanent income hypothesis, and the life-cycle hypothesis (Reyers et al., 2014).

Absolute Income Hypothesis

John Maynard Keynes developed a consumption theory in the early nineteen-hundreds. The absolute income hypothesis is the name usually used for this theory. The theory states that when income increases, consumption increases as well, but at a slower rate, since income growth outpaces consumption growth. This suggests that consumption and income are not inversely proportionate. Keynes explicitly states that individuals' subjective necessities, psychological tendencies and habits undeniably influence their consumption (Keynes, 1936).

Permanent Income Hypothesis

Milton Friedman believed that Keynes had presupposed that current consumption expenditure is a very reliable and stable function of current income (Friedman, 2016). Friedman proposed the permanent income hypothesis. He essentially says that both consumption and income

should be seen as consisting of permanent and temporary components and that while the permanent components of consumption and income are positively related to one another, there is no correlation between the temporary components or between the temporary component and the permanent component of the other variables (Eisner, 1958). The hypothesis simply argues that over the long run, an individual's consumption will remain consistent with the anticipated average income (Friedman, 2016).

According to Friedman, his theory is more comprehensive than that of Keynes, completely clarifies the wealth-income relationship, and makes sense of why the relative income hypothesis should hold true under specific conditions (Friedman, 2016).

Life-Cycle Hypothesis

Early savings theories still considered psychological factors. However, these were not considered in theories such as the permanent income hypothesis and the life-cycle hypothesis (LCH). The LCH theory only sets out how an individual *ought to* behave rather than how that individual *will* behave (Reyers et al., 2014).

Franco Modigliani and Richard Brumberg developed the LCH. Their hypothesis is closely linked to the permanent income hypothesis of Keynes. They move from Keynes' claim that as real income rises, a higher percentage of income is saved. Instead, they propose that the percentage of income saved is essentially not dependent on an individual's income (Modigliani & Brumberg, 1954).

The two premises for their hypothesis are (a) that the main goal of saving is to act as a safety net against the significant income fluctuations that frequently happen over the life cycle of a household, as well as against less predictable short-term changes in requirements and income, and (b) that the number of years over which these provisions can be made is largely independent of income levels, and that the provisions the household would wish to make and can afford to make for retirement and emergencies must be proportional, on average, to its basic earning capacity (Modigliani & Brumberg, 1954). According to the LCH, building up resources for future expenses

to maintain spending at the norm during retirement is the primary driver of saving (Jappelli & Modigliani, 1998).

The LCH is based on the assumption that individuals can solve difficult calculations to determine the sufficient savings amount and determine what they will consume over their lifetime. Thus, it is assumed that individuals will be able to decide how much of their present earnings they should set aside for retirement and how to manage a retirement savings plan that works for them. It is also assumed that they will not stray from their set-out plan. In such a scenario it is further assumed that individuals always make rational choices with optimal outcomes. However, if this were true, there would be no need for any type of intervention, such as tax penalties, to influence an individual's behaviour. (Reyers et al., 2014).

2.3.3 Behavioural Life-Cycle Hypothesis

Individuals' behaviour is subject to their intellect and emotions (Godoi et al., 2005). Emotions influence relationships with others, sleep habits, economic, political and policy decisions, and much more. Emotions drive decision-making and lead to negative and positive decisions. The potentially damaging consequences caused by emotions highlight the importance of understanding the way an individual behaves (Lerner et al., 2015). Thus, for the LCH theory to be accurate, it should include psychological factors affecting how an individual *will* behave (Reyers et al., 2014). Theories must not show how an individual *should* think but rather how an individual *thinks* (Godoi et al., 2005).

The theories did make a shift in the late 20th century and started to focus on the psychological drivers of individuals' behaviour affecting their savings. Richard Thaler and Hersh M. Shefrin introduced the behavioural life-cycle hypothesis (BLCH) (Canova et al., 2005; Reyers et al., 2014). The impact of an individual's behaviour and emotional factors on savings behaviour is the focal point of the theory (Reyers et al., 2014). It specifically looks at the psychological factors of self-control and mental accounting (Canova et al., 2005). The theory states that an individual's decisions will not always lead to ideal savings levels (Reyers et al., 2014), because people do not regard all of their wealth equally and instead make various purchases based on whether they view

their wealth as current income, current assets, or future assets (Canova et al., 2005). It emphasises the necessity of some type of intervention on individual behaviour (Reyers et al., 2014).

Learning from the history of saving theories it can be seen that individuals' emotions affect their behaviour and the choices they make (Bollen et al., 2011). It is important to understand and interpret the judgement and decision-making of individuals (Carminati, 2020; Madrian & Shea, 2001) and bear in mind the psychological factors that influence behaviour (Bollen et al., 2011; Kapoor & Prosad, 2017).

2.4 Behaviour in Savings Decisions

Individuals are expected to behave as experts in their decision-making about retirement savings without being experts (Bailey et al., 2003). Individuals are prone to deviate from the standard model regarding their preferences, beliefs and decision-making (DellaVinga, 2009).

As mentioned in the previous section, it is important to consider the behaviour of individuals, since it is a significant driver of their saving decisions and should not be overlooked (Bollen et al., 2011; Carminati, 2020; Kapoor & Prosad, 2017; Reyers et al., 2015).

Behavioural factors that generally affect an individual's decisions regarding retirement are related to either their bounded rationality or bounded willpower. Unfortunately, little is known about the rationality or decision-making process of an individual easing into retirement (Reyers et al., 2014).

Bounded rationality usually refers to the limits an individual has when making a decision. It is defined as "rational choice that takes into account the cognitive limitations of the decision-maker. These limitations can be both knowledge and computational capacity". Bounded rationality results from the difficulty that retirement decisions carry (Reyers et al., 2014, p. 419).

Bounded willpower refers to the way an individual behaves whilst knowing it is not the best long-term decision (Reyers et al., 2014). Situational and temporal inconsistencies will cause self-control and procrastination issues which are linked to bounded willpower (Reyers et al., 2015).

The behavioural factors that are prevalent in securing retirement savings are self-control and determination (Alemanni & Lucarelli, 2017; Carminati, 2020; Moffitt et al., 2011; O'Donoghue & Rabin, 1999), inconsistency of individuals' choices (Carminati, 2020; DellaVinga, 2009; O'Donoghue & Rabin, 1999), optimism or pessimism (Balasuriya et al., 2014), impulsiveness (Ainslie, 1975; Robayo-Pinzon et al., 2021), information avoidance (Duarte, 2021), procrastination (Duarte, 2021), affect heuristic (Duarte, 2021), and decision fatigue (Duarte, 2021). These factors are discussed in the following sections.

2.4.1 Self-control and Determination

It is frequently seen that people make choices that will ultimately or instantly hurt them even if the appropriate choice is known (Thaler & Benartzi, 2004).

Individuals tend to prefer the present more than the remote future. For example, many individuals fail to see the long-term effects of negative addictive behaviours such as smoking, obesity, and drug addiction. In addition to this, people are prone to impatience. This is an indication of their self-control problems. They like to experience their rewards now and put off costs for later (O'Donoghue & Rabin, 1999).

Self-control and determination are important characteristics related to individuals' retirement decisions (Alemanni & Lucarelli, 2017; Lerner et al., 2015). To be able to save adequately for retirement requires self-control (Thaler & Benartzi, 2004). A lack of self-control relates to choosing consumption now over saving for the future (Reyers et al., 2014).

The lack of self-control is usually very prominent in individuals who display situational inconsistencies. They intend to behave a certain way and then on impulse act differently due to tempting circumstances (Reyers et al., 2014). O'Donoghue & Rabin (1999) examined these self-control problems. They were modelled as time-inconsistent and present-bias preferences. Time-inconsistency refers to the fact that an individual's decision is determined at the time when the decision is made. Present-bias preferences refer to the decision an individual will put off to avoid immediate costs. The study distinguishes between immediate rewards or immediate costs as well as between sophisticated and naive individuals. Sophisticated individuals expect to have a lack of

self-control in the future and a naive individual does not. The results conclude that naive people procrastinate immediate-cost activities and do immediate-reward activities too soon. Thus, a naive individual is expected to have low savings for their future. They are mainly affected by the present-bias effect. Sophisticated individuals do not tend to procrastinate, but when it comes to immediate rewards, act too quickly. To wait is very unattractive to a sophisticated individual. Sophisticated people anticipate self-control problems in the future. Although some people are sophisticated, they are never completely sophisticated. Thus, naive individuals are more at risk with immediate costs, and sophisticated individuals are more affected with regard to immediate rewards (O'Donoghue & Rabin, 1999).

On the other hand, awareness of their self-control problems may often improve self-control in individuals (O'Donoghue & Rabin, 1999).

Similarly, it was found that self-control in children predicts their physical well-being, substance addictions, personal finances, and criminal records. Addressing self-control will affect an impressive range of societal costs, saving taxpayers money and improving wealth (Moffitt et al., 2011). Firms aim to analyse the self-control of clients to improve their wealth and health, as well as minimise crime. An intervention in self-control is beneficial to both the government and citizens (Carminati, 2020; Moffitt et al., 2011).

2.4.2 Procrastination

Research from Kelly and Swisher (1998) suggests that retirement planning should start sooner.

Individuals tend to procrastinate because they find it inconvenient and uncomfortable to complete the task in the present. Procrastination is delaying a task. It results from a lack of self-control and intertemporal inconsistencies (Duarte, 2021; Reyers et al., 2014), which refer to an individual who values current consumption more than future consumption (Reyers et al., 2014).

2.4.3 Inconsistency of Choices

Another relevant behavioural characteristic identified is the inconsistency of individuals' choices. As time goes on, an individual will not necessarily make the same choice even in a similar circumstance as before (Carminati, 2020). Economists have always assumed an individual's preferences are time-consistent, only affected by their own rewards, and independent of the decision-making. However, studies have found that individuals are time-inconsistent and do not only care for themselves. Individuals deviate from rational expectations, for example, by overestimating their skills and being overprotective (DellaVinga, 2009; O'Donoghue & Rabin, 1999). Individuals' preferences are time-inconsistent and present-biased. Present-biased preferences refer to a situation where an individual has to consider two future moments – they will delay the earlier of the two as it gets closer (O'Donoghue & Rabin, 1999).

2.4.4 Optimism or Pessimism

Behavioural biases, like optimism, have been a topic of interest to researchers dating back many years. Optimism influences different social domains in individuals' lives. Even though optimism holds the positive result of motivating individuals, it blinds them to the risks they face. Optimism can thus negatively affect an individual's outlook on the future. Optimism tends to lead an individual to hold investments with higher risks. They may expect that high-risk investments will lead to high returns; however, high-risk investments can also lead to low rewards/returns (Balasuriya et al., 2014). Balasuriya et al. (2014) found that optimism increases the probability of an individual not taking part in a pension scheme, both employer-run pension schemes and private pension schemes.

2.4.5 Impulsiveness

Impulsiveness affects the value of the choices individuals make, causing them not to gain effective rewards (Ainslie, 1975).

Impulsiveness is a specific behaviour characteristic that increases the probability of an individual not planning for the long term (Alemanni & Lucarelli, 2017). Impulsiveness refers to an individual freely deciding on the poorer reward between alternative rewards while being completely knowledgeable of the benefits of the alternatives (Ainslie, 1975).

Individuals with non-planning impulsiveness are affected by their emotions (Alemanni & Lucarelli, 2017). They are not considering what might happen in the future (Robayo-Pinzon et al., 2021).

Individuals who are less inclined to act upon their emotions and less impulsive are more willing to plan for retirement than emotive and impulsive individuals. They are more likely to preserve money for the future (Alemanni & Lucarelli, 2017).

2.4.6 Information Avoidance

Information is valuable and more often than not sought after. Individuals are sometimes willing to pay for information even when it is seen as useless. However, information is often only thought of as useful when it leads to better decision making. Generally, individuals tend to avoid information (Golman et al., 2017), especially if the information may have psychological costs (Duarte, 2021), such as affecting how the individual feels about the information. For example, investors sometimes avoid looking at how funds perform when the markets are down (Golman et al., 2017).

2.4.7 Affect Heuristic

Affect heuristic refers to the fact that people often make decisions based on their current emotions. *Affect* relates to the experience of feeling emotion, incorporating the specific quality of ‘goodness’ or ‘badness’ (Slovic et al., 2007). “To discover” is the meaning of the Greek word *heuristic*. Affect heuristic is thus a kind of mental shortcut to problem-solving that incorporates feelings and the individual’s past experiences. Heuristics offer methods for examining a constrained set of signals and/or potential options for making decisions. By diminishing the quantity of integrated information required to make the decision or pass judgement, heuristics

simplify the process of retrieving and storing information in memory, thus simplifying the decision-making process. Affect heuristics may accelerate our decision-making and problem-solving processes, but they can also add errors and biased judgements (Dale, 2015). Simply put, the affect heuristic is a stimulus that influences the process of making decisions and judgements (Duarte, 2021; Slovic et al., 2007). It may happen that an individual's emotions lead to polarised thinking, which in turn leads to all-or-nothing choices (Duarte, 2021). The emotions an individual may feel are important because decision-making happens quickly and according to the emotion. The affect heuristic proves that an individual's rationality is bounded, leading to decisions being satisfactory rather than optimal. Individuals rely on heuristic factors such as availability, representativeness, anchoring and adjustment to judgements, as well as the way in which they use simplified strategies to make their decisions (Slovic et al., 2007).

The affect heuristic is simultaneously extraordinary and terrifying. It is extraordinary in its speed, subtlety and sophistication. It is terrifying since it is dependent on the experiences and environment an individual is in. It can manipulate and mislead an individual without the individual knowing it. It allows individuals to be rational in some circumstances and irrational in others (Slovic et al., 2007).

Bounded rationality can cause issues for individuals who are not knowledgeable enough to carry out the necessary calculations. This is especially important with regard to retirement savings since they involve complicated decisions. When presented with a tough decision such as retirement savings, the mind of an individual uses mental shortcuts (heuristics) to solve a problem. This is not ideal if useful heuristics are not present. In the absence of useful heuristics, individuals may need to consult an expert for advice or they can just follow the behaviour of their peers, which is not necessarily good or bad (Reyers et al., 2015).

Peers can have an important influence on individuals' retirement savings decisions. Even though decisions may not directly correlate to peer influence, individuals are usually a part of social groups that are influenced by variables such as their environment, or they interact with individuals who have the same interests. The variables affect both group and individual behaviour (Duflo & Saez, 2002).

2.4.8 Decision Fatigue

A popular idea exists that the more choice an individual has the better. This is based on the belief that individuals want more choices and have the ability to manage more choices. Generally, it is believed that a large number of options is better, assuming an individual will be able to make rational decisions. However, a large number of choices may cause a lot of uncertainty for individuals and demotivate the individual to make a decision (Iyengar & Lepper, 2000; Iyengar et al., 2004). This is known as decision fatigue. It occurs when one decision leads an individual to make more decisions. The large number of decisions becomes overwhelming. Decision fatigue demonstrates an individual's bounded rationality (Duarte, 2021). Too much choice that leads to demotivating an individual is sometimes referred to as the 'tyranny of choice' (Iyengar & Lepper, 2000).

Limited choices will increase an individual's control and motivation (Iyengar et al., 2004). Iyengar and Lepper (2000) investigated whether individuals prefer many choices or limited choices. Overall the study showed that individuals prefer limited choices. Individuals with limited choices make more optimal decisions. Individuals with unlimited choices enjoyed the process of making decisions, but they felt more responsible for their choices, which resulted in frustration. Extensive choices cause joy, frustration and dissatisfaction at the same time (Iyengar & Lepper, 2000).

Today a wide range of different retirement savings plans are available. This could lead to sub-optimal decisions for retirement (Iyengar et al., 2004). Iyengar et al. (2004) found that the more 401(k) offered, the more demotivated an individual was to even participate in a 401(k) plan. This left many individuals without a retirement plan.

Another drawback of too much choice is that individuals tend to be overwhelmed when they are given multiple options to choose from. The multiple alternatives may make it difficult for them to decide on the best option (Carminati, 2020). Overreaction to information occurs almost as frequently as underreaction (Fama, 1998). Individuals tend to overreact in their decision-making, especially when probability is involved. An example of a decision with probability is: What is the

probability of the individual being able to retire at the age of 65 if he/she keeps paying their current contribution rate? (De Bondt & Thaler, 1987). In contrast to this, individuals also tend to stick to their current or previous decisions even when they are presented with alternatives. This is referred to as status quo bias. Status quo bias is of great significance for health and retirement products, considering that it is a leading cause of insufficient retirement rates (Samuelson & Zeckhauser, 1988).

2.4.9 Financial Literacy

Financial education is an indicator of investment outcomes. Many individuals, across different demographic categories, lack essential financial education. Individuals, regardless of whether they are male or female, young or old, make better investment decisions for their retirement plan when they have the relevant knowledge about the alternative options available to them (Fisch et al., 2016).

One of the barriers found to affect individuals' investment decisions is their low financial education. The low levels of necessary knowledge to invest in appropriate financial products for retirement savings are not only continually experienced by emerging economies but also by developed nations. For example, in the Australian market many people are not investing enough to build up sufficient money for a good retirement lifestyle. In developing economies, like India, where there is no adequate social security system, retirement preparation is particularly important (Kumar et al., 2019).

South Africa has a large financial education gap. Many South Africans do not understand basic financial concepts. Their financial illiteracy affects their ability to save and causes a lack of planning for retirement (Roberts et al., 2014). South Africans do not generally take the available opportunities to ensure a secure financial life. They miss these opportunities for a variety of reasons – from a lack of financial education, especially for young people, to the high expense of living, to failing to seek expert financial advice (Sanlam, 2023a). Addressing this gap with solutions that close the gap will also broaden the socio-economic inclusion in the country. In addition to individuals not being able to save for retirement, poor financial literacy also affects their mental

well-being. This is crucial, especially during a time like the Covid-19 pandemic that caused extra stress for people (Botha, 2021).

In 2021 research from Deloitte showed that about 70% of South Africans spend all their income (Botha, 2021). It is expected that the improvement of financial literacy will free up disposable income, making it possible for South Africans to save and invest for short-, medium- and long-term goals. Additionally, it will minimise individuals' debts, which will lead to an increase in disposable income. This is especially important in the light of the released statistics that South Africa's debt-to-income ratio for 2020 was 77%, an increase of 5% from the previous year. Financial literacy further protects individuals from falling for illegal financial dealings (Botha, 2021).

2.5 Importance of Retirement Planning

Planning is essential for a comfortable and enjoyable retirement. Individuals should make their money work for them. Retirement planning will enable them to achieve their goals later in life. The decision to plan for retirement affects one's quality of life and is a one-time strategic investment choice. Only 12% of the whole population has appropriately planned for their retirement despite the availability of numerous viable opportunities (Kumar et al., 2019).

Seeing that the population has longer life expectancies, the long-range planning attitude for retirement savings has become more crucial (M. Wang & Shultz, 2010). Individuals' retirement savings must be in line with longevity expectations (Van Solinge & Henkens, 2009).

The simplest approach to the challenging problem of insufficient retirement savings is retirement planning (Kumar et al., 2019). While proper planning can increase savings, individuals who have planned for their retirement also feel more confident about their financial future than those who haven't. Higher education levels are associated with more confidence in retirement planning (Kumar et al., 2019).

Individuals have choices. They need to make decisions about their future (Ainslie, 1975). To provide for retirement savings is a motivated choice made by an individual (M. Wang & Shultz,

2010). With the international trend of shifting from a defined benefit to a defined contribution approach in retirement savings, individuals are responsible for their retirement savings. This increases the importance of personal planning for retirement (Kumar et al., 2019; Reyers et al., 2015).

2.6 From Defined Benefit to Defined Contribution Retirement Schemes

The changes in the South African and global retirement systems highlight the importance of planning for retirement. The world of retirement planning and pensions shifted from a defined benefit to a defined contribution strategy (Beckker et al., 2019; Lusardi & Mitchell, 2014; M. Wang & Shultz, 2010).

In a conventional *defined benefit* (DB) pension plan, employees accumulate the promise of a consistent monthly payment from the time they retire until their death, or, in certain circumstances, until the death of their spouse (Broadbent et al., 2006). The benefits during retirement are usually known (Pillay & Fedderke, 2022).

In a *defined contribution* (DC) pension plan employees accumulate funds in an individual fund managed by the plan's sponsor. The sponsor is typically the employer. Employee contributions are often taken directly out of salaries, and frequently the company will match a part of these payments. DC assets grow at a relatively consistent pace over time. Contributions to DC plans are typically a fixed proportion of earnings (Broadbent et al., 2006). The benefits during retirement are usually not known before retirement.

Defined contribution funds dominate the SA retirement industry. DC causes a risk that is far greater for a household to withstand (Pillay & Fedderke, 2022). Individuals now have a greater obligation to *actively* plan for their retirement (Dhlembeu et al., 2022). Antolin (2008) found that individuals do not save sufficiently for retirement especially when they are left to themselves to make an investment choice.

A key difference between DB and DC plans is who bears the risk. It will either be the employer or the employee. Broadbent et al. (2006) highlight the risk bearers in DB and DC plans, as shown below:

Who bears the risk in a DB and DC Pension Plan		
Type of risk	**DB plan**	**DC plan**
Investment	Employer	Employee
Inflation	Employee / Employer	Employee
Longevity	Employer	Employee
Market timing (temporal)	Employer	Employee
Accrual (portability)	Employee	DC plans are portable
Vesting	Employee	Employee
Employer insolvency	Employee/taxpayers	DC plans are always fully funded
Salary replacement risk	Employer	Employee
Fiduciary/legal risk		Employer

Table 2.1: Risk Distribution in a DB and DC Pension Plan (Source: (Broadbent et al., 2006))

The change from defined benefit to defined contribution schemes for retirement savings places individuals in the position to make numerous decisions regarding their retirement (Beckker et al., 2019; Lusardi & Mitchell, 2014; M. Wang & Shultz, 2010). Individuals are now responsible for ensuring that (Sanlam, 2023b):

- Their contribution amount to their retirement fund is enough.
- Their decision of the risk-profiled portfolio suits them best over time.
- The decision to preserve their retirement savings is made if they change employment.

It is believed that this shift was largely the cause for the low levels of retirement fund preservation (Sanlam, 2023b).

2.7 Retirement Fund Preservation

Insufficient retirement savings are caused by individuals failing to prepare for the future, as well as the failure to preserve their savings when resigning from jobs or getting retrenched (National Treasury, 2021a). Quite a lot of research is done into the reasons for insufficient retirement savings, but little is known about the preservation of retirement savings (Reyers et al., 2014).

The preservation of retirement funds refers to the situation where an individual belonging to a retirement fund (such as pension or provident funds) changes jobs or gets retrenched, and then withdraws the retirement savings and places it into another retirement fund or a savings account to use for retirement (National Treasury, 2012a). Dissuading people from withdrawing their accumulated retirement funds when changing employment is an important area of study, as the issue of low levels of preservation is becoming widely acknowledged as a critical problem that may lead to insufficient savings at retirement (Reyers et al., 2015).

2.7.1 The Inadequacy of Retirement Fund Preservation

Employees frequently have several options for the payout of their pension funds. Typically, individuals take the payout without using that payment for another retirement savings vehicle. This causes impoverishment among the elderly during retirement (Hurd & Panis, 2006). It is found that more than half of individuals from the United States who change employment take the pre-retirement cash payout. South Africa has a similar issue where most individuals do not preserve either (Pratt, 2010; Reyers et al., 2014).

Given the rise in pension assets and payments on the one hand and proposed cuts to Social Security on the other in the US, the importance of pensions is certain to keep rising. However, pre-retirement withdrawals from pensions may diminish the security that private pensions offer in old age and jeopardise efforts to lower poverty among the elderly (Hurd & Panis, 2006).

Many countries, such as Australia, Germany, Switzerland, Canada and the United Kingdom, regulate the preservation of pension funds. Other countries, like South Africa and the United States of America, discourage cash payout before retirement with taxes and fines, but they do not regulate it (Reyers et al., 2014).

In South Africa, about 50% of employed individuals are members of a retirement fund (National Treasury, 2012a). Many retirement fund participants prefer to cash out every time they move jobs, not protecting their money (National Treasury, 2012b, 2021a).

2.7.2 Potential Reasons for Early Withdrawals

As mentioned, rational (bounded rationality) or irrational (bounded willpower) behaviour affects decisions of individuals regarding their savings. It also drives an individual's decision to withdraw retirement savings without preserving the withdrawn amount (Reyers et al., 2014).

Rational decisions relevant to the LCH and consumption smoothing behaviour may also be reasons for low preservation. The LCH theory forecasts that rational reasons for not preserving would include individuals that behave in such a way that they spend their money today rather than save it for the future, i.e. consumption smoothing behaviour. An example of these rational reasons may be adolescent adults who are in a life stage where they consume. Their focus is on their immediate needs or debt they have from their consumption phase. This type of behaviour is prominent among adolescent adults. Thus, based on this, a higher preservation of retirement funds is expected to be found among older individuals. However, individuals may be faced with situations where they need money, regardless of their age, and the rational individual will take the cash payout. In these types of situations, it is valuable to determine why the individual changed jobs. If an individual was laid off or retrenched, the cash payout may be a means to provide for the individual's consumption during unemployment. Another reason for individuals taking the cash can be that they need the money for their immediate survival needs (Reyers et al., 2014).

Deciding on the best amount to save for retirement is a difficult decision by nature. It is especially difficult for individuals who display bounded rationality, seeing that they will make decisions that are not ideal. When a decision has to be made whether funds should be preserved, a

person is required to be capable of comprehending and using the effects of compounding over a future period of time. Evidence suggests that many individuals do not have this ability. Numerous factors, including the duration of one's working life, inflation rates, potential investment returns, and retirement age, complicate this decision (Reyers et al., 2014).

If an individual displays bounded willpower it can affect their retirement savings. The absence of self-control and procrastination may be the cause of bounded willpower. Situational and temporal discrepancies lead to procrastination. These two inconsistencies are strongly related to time perspective and impulsiveness.

Time perspectives affect an individual's savings behaviour and financial planning. It comprises an individual's emotional perspective of the past, present and future. Impulsiveness describes a situation where individuals act faster than think about the future. This usually happens because individuals value temporary and immediate pleasures. An individual who does not act impulsively and has a time perspective that is focused on the future will have high self-control and low procrastination.

An individual's willpower develops with maturity. There is a relationship between brain maturation and an individual's futuristic attitude as well as their ability to implement control over their impulses. That means individuals should have more self-control as they age. However, age is an isolated factor in fund preservation because it will not give a comprehensible understanding of whether an individual behaved rationally or irrationally (with bounded willpower). It cannot be known whether the young adult did not preserve due to not being mature enough or whether they needed the cash payment to smooth consumption. Even though brain maturation shows differences between children, young adults and adults, people within these groups also show differences, which exist due to impulsiveness and time perspectives. Culture, religion, upbringing, education and specific social effects are drivers of time perspective. Impulsiveness and time perspective are strong character traits that affect the differences between individuals. Thus, individuals with bounded willpower are likely to have low retirement preservation rates due to them acting impulsively or not having a futuristic attitude (Reyers et al., 2014).

The factors that potentially drive low preservation of retirement savings are shown in Table 2.2.

Potential factors that could drive preservation levels			
Decision maker	**Potential causes**	**Potential factors**	**What would predict low levels of preservation**
Rational	Consumption smoothing	Age Liquidity constraints	Young Low levels of liquidity
Irrational	Bounded rationality	Education	Low levels of education/financial literacy
	Bounded willpower	Financial literacy Time perspective Level of impulsivity	Low level of future orientation High levels of impulsivity

Table 2.2: Potential Factors that could Drive Low Preservation Levels (Source: (Reyers et al., 2014))

2.7.3 Potential Solutions

It is crucial to manage the emotional component of choices for long-term planning decisions (Alemanni & Lucarelli, 2017). Studies have proven that strategic intervention supplying incentives or addressing behavioural barriers can encourage individuals to behave a certain way (Duarte, 2021). Intervention can lead to improved behaviour (P. Nguyen et al., 2022). Intervention does not need to be overly complicated; simple interventions have been found to have positive results on the behaviour of individuals (Duarte, 2021; Püschel et al., 2010).

Common interventions to improve the behaviour of saving retirement funds include financial education, constructive communication, and intermediaries.

Financial education has proven to be beneficial for individuals' savings (Duflo & Saez, 2002). When employees do not understand the provided information about a savings plan or take automatic enrolment as financial advice, the best solution will be financial education (Madrian & Shea, 2001).

Financial illiteracy, as mentioned, is one of the barriers faced by individuals to save sufficiently for their retirement. There is a positive relationship between the financial literacy of employees and higher enrolment rates or lower termination/withdrawal rates. Just the act of making individuals aware of the importance of preservation for retirement could lead to required outcomes, apart from improving financial literacy (Antolin, 2008). Using financial education and the provision of information separately and independently might be insufficient (Alemanni & Lucarelli, 2017).

The conclusion of an interview done with retired nurses or nurses close to retirement was that these individuals felt they had lacked financial knowledge when they made their retirement decisions. They agreed that the information provided before retirement was insufficient, which led to them making decisions that they now experience as inadequate. The need for communication is clear from these results (Kelly & Swisher, 1998).

Communication is found to be effective in managing the emotional component of taking long-term planning decisions. Constructive communication is likely to encourage individuals to try to prepare for their retirement (Alemanni & Lucarelli, 2017). The intervention through communication highlights the value of information provided to individuals (Karlan et al., 2016). Communication can be done through direct messages to individuals or through providing information via the internet (Duarte, 2021; Harrison et al., 2006; Karlan et al., 2016; Pop-Eleches et al., 2011). Antolin (2008) found that national awareness campaigns are successful in increasing individuals' knowledge of the retirement system and the importance of saving for retirement.

Karlan et al. (2016) found that reminding individuals with savings accounts through direct messages increased their commitment to their savings. The messages communicated in the study, contained information that stated the goal as well as the financial incentives of the savings account.

Not only did the reminders improve individuals' commitment to their savings account but they also increased the monetary value of the account.

Other methods used to improve retirement savings behaviour are automatic enrolment and default options (Alemanni & Lucarelli, 2017; Antolin, 2008; Duflo & Saez, 2002; Madrian & Shea, 2001). Default rules are rather significant for savings. They affect an individual's participation, contribution, asset allocation and participation in long-term future decisions as well as complement education and information (Alemanni & Lucarelli, 2017; Duflo & Saez, 2002).

Countries like South Africa and the United States of America discourage cash payouts of pension funds before retirement with taxes and fines. However, these measures are not always successful, seeing that individuals still take the cash (rather than preserving it) when they change employment. The South African government applies a high tax rate to retirement fund withdrawals when individuals change jobs, but it does not change individuals' behaviour about preserving their retirement savings (National Treasury, 2021b).

The National Treasury of South Africa has also acknowledged that the current tax system is insufficient to prevent or dissuade withdrawals of retirement funds. The government therefore tries to find ways to make preservation compulsory (Reyers et al., 2014). To this end, the government is in the process to introduce the "two-pot" system. It is a system that forces individuals to preserve two-thirds of their accumulated retirement funds when they change jobs. The remaining third remains accessible to individuals (National Treasury, 2021b). Interventions that are presently in place to affect retirement preservation rates only help an individual when that individual displays bounded willpower (Reyers et al., 2014).

Reyers et al. (2014) tested three types of intervention approaches on the different types of decision-makers to improve preservation rates. The three approaches tested were the libertarianism approach, the paternalism approach, and lastly the libertarian paternalism approach.

2.7.3.1 Libertarianism Approach

The libertarianism approach is driven by freedom of choice. This is an attractive approach to individuals who act rationally. In this approach, product providers will choose to do nothing to influence individuals. They assume individuals will know what the right decision is to make. This type of approach is only effective for individuals who behave rationally (Reyers et al., 2014).

2.7.3.2 Paternalism Approach

The paternalism approach uses regulation to make preservation mandatory, or uses some type of tool, such as taxes, to persuade individuals to preserve. This approach assumes that individuals behave irrationally. The regulation will therefore govern individuals' behaviour. The approach is successful for individuals who have low self-control and procrastinate (Reyers et al., 2014).

2.7.3.3 Libertarian Paternalism Approach

The libertarian paternalism approach combines the first two approaches to overcome the shortfalls of both of them. Libertarian paternalism entails that individuals are pointed towards a specific decision while retaining the freedom for the individual to opt out of the decision. This can be achieved through behavioural tools. The following behavioural tools were investigated: choice architecture and debiasing (Reyers et al., 2014).

Choice architecture is equal to "nudging". Individuals will be led to the optimal choice while retaining the freedom to choose another path. Default rules are the most used tool of choice architecture. This approach is successful for individuals who display procrastination problems (Reyers et al., 2014).

Debiasing focuses on education and training as well as using intelligent software of decision aids to influence an individual's behaviour positively. It is an approach used to overcome biases associated with irrational decision-makers. Debiasing has the potential to positively affect irrational decision-makers who display bounded rationality and those who display bounded

willpower. However, this approach has yet to be tested in relation to the preservation of funds (Reyers et al., 2014).

An approach is needed that penalises those who would not otherwise have preserved whilst simultaneously not penalises those who would have preserved (Reyers et al., 2014).

2.7.3.4 Intermediaries

Reyers et al. (2015) found that the probability of preservation of retirement funds increases if an individual follows professional advice.

There is a considerably large gap in financial knowledge between an ordinary investor and a financial advisor. A financial advisor is more financially literate and is capable of making better investment decisions among a range of investment options. A financial advisor is a potential solution for an individual to overcome the obstacle that is caused by the individual's financial illiteracy in retirement planning (Fisch et al., 2016). Financial advice and literacy will enable individuals to understand pension products, which are known to be quite complex (Meyll et al., 2020).

Individuals are known to try solving complex problems by seeking solutions by themselves. They are also affected by their temporary emotions while they make decisions (Della Vinga, 2009). Financial advisors offer assistance across a wide range of options (Fisch et al., 2016). They provide information, mitigate simple issues, provide clarification, and give aid on how to fix unexpected problems (Harrison et al., 2006). The advice from the professionals will help individuals to feel more comfortable with their investment decisions and maybe help encourage them to make more or better investments for their future (Fisch et al., 2016). The assistance offered by a financial advisor also simplifies the process of making investment decisions in contrast to following a set of complex rules. In addition, financial advisors are more equipped to manage risks (Bertram & Zvan, 2009).

Intermediaries can help individuals to control their emotions and can limit impulsive behaviour. Advisors also help individuals to reflect on their past investment mistakes. It can

therefore be deduced that consulting is beneficial for the wealth of an individual. Individuals furthermore benefit from increased confidence brought about by choosing a consultant (T. A. N. Nguyen et al., 2019).

Financial advisors could be beneficial to individuals during their retirement planning (Kelly & Swisher, 1998). Through their professional advice, financial advisors can guide individuals to prepare for retirement and enable them to retire with adequate funds (Stolper, 2018). Financial advisors are beneficial for group schemes as well (Thaler & Benartzi, 2004). Although advisors will be beneficial, the responsibility still lies with the investor to make the move to call in a professional advisor (Kelly & Swisher, 1998).

Intermediaries can have a positive influence on the long-term savings of individuals. They can encourage individuals to behave in a way that is beneficial to their future (Thaler & Benartzi, 2004). Financial advisors are often avoided (Fisch et al., 2016; Stolper, 2018) or advice given is not followed (Stolper, 2018). This is generally because of the costs of advisors and the moral hazard problem. The cost of advisors may deter individuals from consulting them. Clients, especially those with lower financial education, may also feel that the advice they receive is biased advice (Fisch et al., 2016; Lourenço et al., 2020; Stolper, 2018). However, it has been found that financial advisors complete the job they have been hired to do and seldom make biased choices for the advisee (Meyll et al., 2020).

2.8 Retirement Fund Preservation in South Africa

2.8.1 Current State of Retirement in South Africa

As mentioned, there are three main saving vehicles in the South African retirement system: pension funds, provident funds and retirement annuities.

Pension funds and insurance firms are the two biggest institutional investors in the world's financial markets. About 40% of the assets on the Johannesburg Stock Exchange in South Africa are owned by pension funds (Pillay & Fedderke, 2022).

Retirement planning is a worry for many people around the globe. This global trend has been dubbed the "retirement savings crisis" by academics. Previous predictions stated that 94% of South Africans would not be able to sustain their quality of living after retiring (Dhlembeu et al., 2022). With fewer than 10% of South Africans now thought to be financially independent when they retire, this number has stayed low. It shows that in South Africa, saving enough money for retirement has been difficult for a while. Retirement savings have taken a further hit as a result of the Covid-19 pandemic lockdowns and associated economic downturns (Dhlembeu et al., 2022). Old Mutual conducted a survey in South Africa with over 1 500 respondents that showed only 39% of South Africans have savings that are worth more than 3 months of income (Old Mutual, 2022). It is especially worrying for people whose companies' financial difficulties have led to job losses or income cutbacks, which has reduced their capacity to contribute to retirement savings schemes. According to a post-pandemic nationally representative poll, the percentage of South Africans who are certain that their retirement funds would support their living expenditures has dropped from 20% to 14% as a result of the pandemic (Dhlembeu et al., 2022).

South Africa has three main concerns regarding its retirement system (National Treasury, 2021a):

- Many individuals do not belong to a retirement scheme.
- Cost of retirement funds. The costs are considered to be uneconomical.
- Individuals do not preserve their savings when they change employers.

Alexander Forbes conducted a study on the members of the retirement funds that they administer. The sample was close to one million members. Based on the 2021 study, which is the most recent, the projected replacement ratio is 40.51% (Alexander Forbes, 2021). The Replacement Ratio (RR) is the ratio of retirement income to income during employment (Butrica et al., 2012). Most employers target a RR of 75%. Alexander Forbes (2021) found that only 6% will have an RR above 75%. The RR by age becomes worse as individuals become older. This can be seen in Figure 2.1

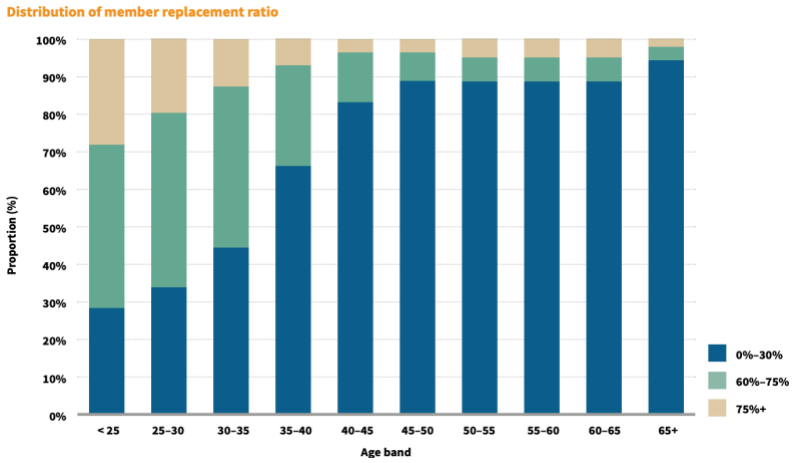


Figure 2.1: Alexander Forbes: Distribution of Member Replacement Ratio (Source: (Alexander Forbes, 2021))

Alexander Forbes (2021) suggests that the RR could have been better if individuals had preserved their retirement savings instead of focusing on their current life. The low RR at older ages is usually due to contributions not being sufficient. This illustrates that intervention should happen sooner during a member's participation in a scheme rather than later.

2.8.1.1 Participation in Retirement Savings

The private occupational retirement system is well-developed in South Africa. However, most of the South African population is not covered by it. Individuals in the lowest income bracket tend to rely on the state to provide pension grants for when they retire, even if the private system is tax-incentivised (Pillay & Fedderke, 2022).

2.8.1.2 Cost of Retirement Funds

The South African retirement funds are not operating at an efficient scale. In the industry, there are strong economies of scale. Administrative costs increase by 72% when the number of members increases by 50% (Pillay & Fedderke, 2022).

2.8.1.3 Preservation of Retirement Funds

South Africa has low preservation rates (National Treasury, 2012b). Preservation is essential for an individual to maintain their financial independence and to enjoy a satisfactory lifestyle during retirement while protecting their well-being (Sanlam, 2023b).

South Africa faces high unemployment levels. Yet, a trend that is recognised worldwide is experienced in South Africa. This trend is known as “the Great Resignation” (Old Mutual, 2022) and involves large numbers of people resigning from their jobs. This is concerning, seeing that South Africa already has low preservation rates.

2.8.2 The Inadequacy of Retirement Fund Preservation

South Africans are not saving enough and are not preserving their assets. Fewer than 10% of retirees can retain their quality of living after retirement. These individuals then become dependent on the state and their families for financial support. This state of affairs is largely due to individuals not preserving their retirement fund when they change employment. The lack of preserving retirement funds erodes financial security during retirement (Sanlam, 2023b). This is illustrated in Figure 2.1 above, which shows low replacement ratios due to poor preservation of retirement savings (Alexander Forbes, 2021).

Preservation rates apply to members who left their employment, were laid off, or were fired by their employers (Alexander Forbes, 2021). The number of resignations has increased over the years (Old Mutual, 2022). This can also be seen in the 2021 study by Alexander Forbes as illustrated in Figure 2.2:

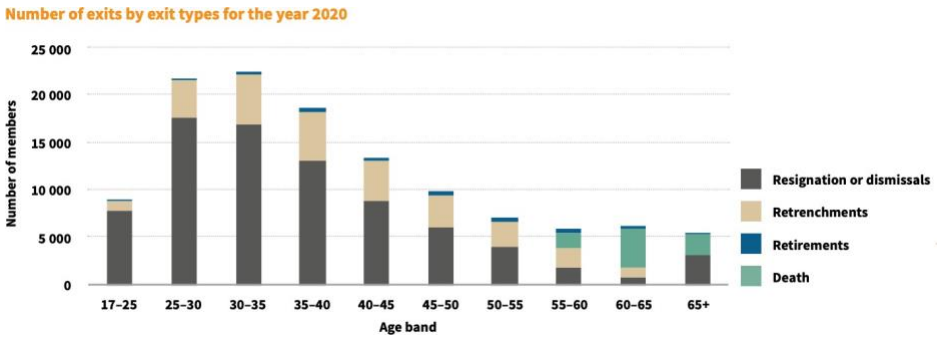


Figure 2.2: Alexander Forbes: Number of Exits by Exit Types (Source: Alexander Forbes, 2021)

The number of resignations is much higher than that of retrenchments, retirements and deaths.

Resignation is an individual's choice, although many factors play a role (Old Mutual, 2022). The number of resignations is concerning because of the country's already poor preservation rates (National Treasury, 2012b).

The preservation rates seem to be decreasing continuously in South Africa. Alexander Forbes (2021) recorded that the number of members preserving has decreased from 11.5% in 2012 to 9.6% in 2020. The 11.5% was not a high number to start with.

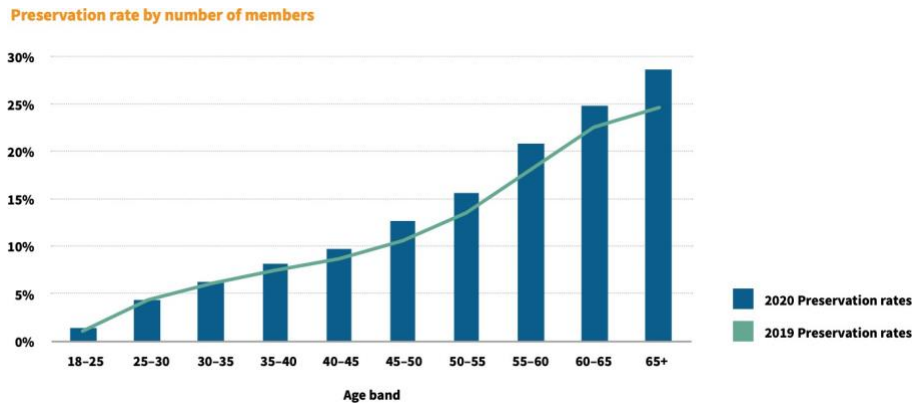


Figure 2.3: Alexander Forbes: Preservation Rates by Number of Members (Source: Alexander Forbes, 2021)

Even though preservation rates are low, Figure 2.3 shows that they do improve as people get older. The graph also shows that 5% of members between the ages of 25 and 30 preserve their savings. The low number is concerning, but it can be due to consumption smoothing.

These results show that most South Africans take the cash payout when they leave their employer rather than preserving their retirement funds. Thus, it has become of great importance to improve retirement fund preservation in South Africa (Reyers et al., 2014).

2.8.3 Legislation and Regulation of Retirement Funds in South Africa

South Africa uses taxes to dissuade individuals from withdrawing their retirement funds before retirement. South Africa's system is based on cumulative withdrawals. The original amount is tax exempt and a scale from 18% to 36% taxes is applied on withdrawals (Reyers et al., 2015). However, these taxes and fines that are put in place to discourage withdrawal are not very successful, seeing that individuals still take the cash (rather than preserving the funds) when they change employment (National Treasury, 2021b).

The National Treasury of South Africa has also acknowledged that the current tax system is insufficient regarding individuals' preservation. The government is therefore trying to find ways to make preservation compulsory (Reyers et al., 2014).

Since the early 1980s, there have been several suggestions to mandate preservation, but due to multiple factors such as the interference of unions and politicians, mandatory preservation failed. Default regulations, a suggestion made by the National Treasury, came closest to improving the nation's preservation rates. However, the rates stayed low even if they did improve (Sanlam, 2023b). The government is going to introduce a retirement reform, discussed below, to improve the preservation of retirement funds (Section 2.6.3.1). This proposal will make the intervention more regulated (Reyers et al., 2015).

2.8.3.1 Retirement Reform: Two-pot system

Due to the low preservation of retirement savings, the government is introducing the retirement reform referred to as the "two-pot" retirement system (National Treasury, 2023).

The two-pot system aims to (Sanlam, 2023b):

- Improve retirement outcomes through increasing preservation of retirement savings;
- Give members access to their retirement savings during times of need.

As mentioned, this system forces individuals to preserve two-thirds of their accumulated retirement funds when they change jobs. The remaining third is accessible to individuals when they change jobs (National Treasury, 2021b).

Preserving retirement funds is important, but having access to one's funds in case of emergencies is also important (Sanlam, 2023b).

The "two-pot" system will be effective from 1 March 2024 (National Treasury, 2023).

3. MACHINE LEARNING

Machine learning is a growing field. This study will apply machine learning models to retirement fund preservation data.

An overview of machine learning, the development of machine learning, the type of problems it can handle, the general machine learning model, the different machine learning paradigms and algorithms, and the limitations and application of machine learning are discussed in this chapter.

The different models used in the study will be discussed in this chapter and not in the methodology chapter. The background of the different models will be presented.

The study will apply three different models to the preservation of retirement funds data, namely, Logistic Regression, Random Forest, and Support Vector Machine. These three models were identified as good classification models (Alzubi et al., 2018; Cutler et al., 2012; Jakkula, 2006; Maalouf, 2011; Peng et al., 2002).

Classification models are used because they enable the study to predict preservation or non-preservation of retirement funds. Classification is performed using certain machine learning models to group data into certain classes. Training the model and testing the model through training and testing data, are two steps that have to be performed during classification. Training data is the data used by the model to learn, and testing data is the data used to test the performance of the model (Nivedha & Sairam, 2015). Fawcett (2006, p. 861) defines a classification model as a "mapping from instances to predicted classes".

3.1 Overview of Machine Learning

Machine learning is a concept introduced in the Artificial Intelligence (AI) field. Machine learning is widely employed today and is a foundational idea for intelligent systems, which paves the way for the introduction of cutting-edge technology and more sophisticated artificial intelligence concepts (Talwar & Kumar, 2013).

Machine learning entails that machines are enabled to recognise various patterns and adjust to changing conditions. Machine learning can be both an explanation and an experience based learning process (Talwar & Kumar, 2013).

The volume of data is growing at a rapid pace. This large volume of data is referred to as Big Data. Data can be structured or unstructured. It is difficult for an individual to interpret and make sense of unstructured data. Finding methods that can make predictions has been encouraged by literature on machine learning (Liu & Xie, 2019; E. Zheng et al., 2017). Lately, machine learning has become a tool that can interpret and make predictions regardless of the structure of the data (Liu & Xie, 2019; Nivedha & Sairam, 2015; Nusinovici et al., 2020; E. Zheng et al., 2017). Machine learning actually forms an excellent prediction tool (Liu & Xie, 2019).

Machine learning is a field in AI that gives computers the ability to reason for themselves. This reasoning is usually referred to as *learning*. The general process of learning involves gaining new or adjusting current behaviours, values, knowledge, skills, or preferences. Thus, the idea behind machine learning is for computers to learn from their actions and improve. The computer learns to adapt its actions to increase the accuracy of its predictions and/or classifications. Machine learning is a major tool in the technology field that can extract information from data that is already accessible. Machine learning solutions to challenging real-world issues make this a compelling area of study across sectors and nations (Alzubi et al., 2018).

Machine learning has a wide range of study fields supporting it. These fields include Psychology, Artificial Intelligence, Control Theory, Neuroscience, Information Theory, Philosophy, Bayesian Method and Computational Complexity Theory (Alzubi et al., 2018). Based on the wide applicability of machine learning to different fields, this study will apply it to retirement fund preservation data to determine how it will perform in this context.

3.2 Development of Machine Learning

Artificial intelligence and machine learning are not novel concepts. For more than 60 years, computer scientists, engineers, researchers, students, and members of the business community have studied, used and reinvented them. Algebra, statistics and probability are the three areas of mathematics that machine learning is built upon. The 1950s and 1960s saw the beginning of the significant growth of machine learning and artificial intelligence fields due to the work of researchers like Alan Turing, John McCarthy, Arthur Samuel, Alan Newell, and Frank Rosenblatt. The Optimising Checkers Programme was the subject of Samuel's initial functional machine learning model proposal (Alzubi et al., 2018).

Arthur Samuel first used the phrase Machine Learning in 1959. He defined it as an area of study that enables computers to learn without being explicitly programmed. Tom Mitchell has given a more modern definition. He defined machine learning as*: "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E"* (Alzubi et al., 2018, p. 1).

3.3 The Generic Machine Learning Model

Machine learning is a tool that is used to solve problems. A learning problem has the following three attributes (Alzubi et al., 2018):

1. Task classes. For example, in a board game like chess, the task would be to play the game. In this study, the task would be to preserve retirement funds.
2. Improvement made in performance measurement. In the game of chess, it is measuring the number of victories against the opposition.
3. The procedure of acquiring experience. In the board game it is practising by playing games against itself to improve performance.

According to Alzubi et al. (2018), the primary components of the generic machine learning model are:

1. Collection and preparation of data – there may be a lot of irrelevant or redundant data that should be filtered out for the model (Alzubi et al., 2018; Nivedha & Sairam, 2015).
2. Feature selection – not all the variables in a dataset are relevant to the learning problem; only the most important features should be used (Alzubi et al., 2018; Nivedha & Sairam, 2015).
3. Choice of algorithm – problems should first be categorised; not all algorithms are suitable for all problems (Alzubi et al., 2018).
4. Selection of model and parameters – most models require some type of intervention to set the parameters at the most appropriate (Alzubi et al., 2018).
5. Training – the model is created with a subset of the data; this subset is usually known as the training data (Alzubi et al., 2018).
6. Performance evaluation – the model should then be assessed with unseen data to determine the accuracy and precision of the model; this is known as the testing data (Alzubi et al., 2018).

3.4 Machine Learning Paradigms

Machine learning paradigms have ten subdivisions based on the methods used to train the algorithm as well as the availability of the output during the creation of the model with training data. These include instance-based learning, dimensionality reduction algorithms, reinforcement learning, evolutionary learning, ensemble learning, artificial neural networks, semi-supervised learning, supervised learning, and unsupervised learning (Alzubi et al., 2018). The following subsections provide explanations of the two main types of paradigms (Talwar & Kumar, 2013) and ensemble learning (Alzubi et al., 2018):

Supervised Learning: Supervised learning is an approach that allows for the perception of both inputs and outputs. Under supervised learning the model will receive training data with the

correct outputs. The algorithm then learns from this training data. The algorithm must generalise to be able to accurately respond to any input based on the training data. When given inputs that weren't seen during training, the algorithm is anticipated to generate the right results (Alzubi et al., 2018; Talwar & Kumar, 2013). In supervised learning, each example has a specific learning objective. When a trainer assigns the classification for each example, the classification is said to be supervised.

The supervised learning approach can be seen in Figure 3.1.

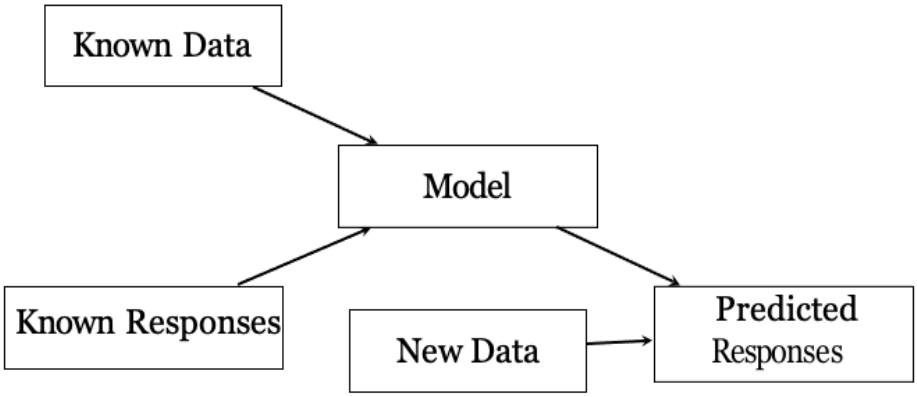


Figure 3.1: Supervised Learning Algorithm Explanation (Source: Broadbent et al., 2006)

The two main categories of supervised learning are (Talwar & Kumar, 2013):

1. The classification of answers that only have a few possible values, such as "true" or "false." The classification process only applies to nominal answer values, not to ordinal ones.
2. Regression for responses that are real numbers, such as a particular car's miles per gallon.

Examples of supervised learning algorithms are: Decision Trees, Neural Nets, Logistic Regression, Bagged Trees, Boosted Trees, Boosted Stamps, Support Vector Machines (SVM),

Naïve Bayes, Nearest Neighbor, and Discriminant Analysis (Talwar & Kumar, 2013). In this study, logistic regression and SVM are used.

Unsupervised Learning: In unsupervised learning, the computer just takes in the inputs, but it doesn't get any supervised goal outputs (Talwar & Kumar, 2013). This approach identifies distinct patterns in the data and then uses them to generate rules (Alzubi et al., 2018).

If the machine's objective is to create representations of the input that can be utilised for decision-making, predicting future inputs, effectively transmitting the inputs to another machine, etc., a formal framework for unsupervised learning may be created.

Examples of unsupervised learning algorithms are hierarchical clustering, dimensionality reduction and k-means clustering (Talwar & Kumar, 2013).

Ensemble Learning: In this approach, a very large number of individual models are taught to address a common issue. An attempt is made to develop a model by constructing a set of hypotheses from the data used for training the model and combining them to make a prediction model. This is done to reduce bias and variance or improve predictions. Other machine learning approaches use a single hypothesis (Alzubi et al., 2018).

Ensemble learning techniques produce several classifiers and integrate their findings. It is generally agreed that a collection of several weak classifiers performs better than a single classifier, given the same amount of training information (Oshiro et al., 2012).

Boosting, Bagging, and more recently Random Forests are well-known ensemble approaches (Oshiro et al., 2012). Random forest is used in this study.

3.5 Types of Problems

There are different algorithms in each paradigm. To apply the most suitable machine learning algorithm to a problem, the problem must first be appropriately classified before problem resolution can begin. The various types of problems are (Alzubi et al., 2018):

- *Classification* – the problem has a fixed output, like Yes or No. The problem can be dichotomous in nature or the problem can have numerous different outputs.
- *Anomaly Recognition* – this occurs when patterns are detected. These problems focus on identifying the outliers.
- *Regression* – these problems deal with outputs that are continuous and numeric.
- *Clustering* – this refers to algorithms trying to find similar characteristics in the data and group them, as in clustering.
- *Reinforcement* – past data is used by the algorithm to make decisions. This is a trial-and-error process done by the algorithm.

3.6 Machine Learning Algorithms

Different algorithms enable machines to learn in different ways and adapt to other virtual environmental factors. The machine makes a choice and does the specialised tasks based on these algorithms (Talwar & Kumar, 2013). Some popular machine learning algorithms are:

Decision Tree: A decision tree represents the learned function in a decision tree- based approach for approximating discrete-valued target functions. A decision tree sorts occurrences into different categories by arranging them according to the feature values from the root to a few leaf nodes. Every branch indicates a potential value for that feature, and each node represents a choice (test condition) on a characteristic of the instance. An instance is first classified at the decision node, which is the root node (Alzubi et al., 2018). This will be explained later in this chapter (Section 3.11).

K-Means Clustering: The objective of this popular unsupervised learning algorithm is to group the number of observations into k-clusters. Each observation with the nearest mean belongs to a certain cluster. The mean will serve as the template for the model. The cluster centre is determined by taking the mean of all the observations inside that specific cluster (Alzubi et al., 2018).

Support Vector Machines: SVMs are a viable solution for both classification and regression issues. It is an algorithm for supervised learning. It operates according to the idea of margin calculation. Each piece of data is plotted as a point in n-dimensional space using this procedure. In this case, n is the number of characteristics the dataset contains. Each feature's value is the matching coordinate's value. It divides the training datasets into classes by locating a line (hyperplane) that divides the data into those classes. It operates by increasing the spaces between the closest data point (in both classes) and the so-called margin hyperplane (Alzubi et al., 2018). This will be explained in a later section of the chapter (Section 3.12).

Regression Analysis: Regression analysis is a forecasting approach that examines the connection between an independent variable (predictor) and a dependent variable (target). In this procedure, it is attempted to minimise the variations in the spaces of the points from the curve or line by fitting the line or curve to the points. Regression analysis comes in a variety of forms, including logistic, polynomial, and linear (Alzubi et al., 2018). One of the popular regression analysis models, Logistic Regression, will be explained in a later section of the chapter (Section 3.10).

Random Forest: This algorithm is commonly used for classification and regression. It uses an approach that reduces variance on a random subset of the initial data. Multiple decision trees are created, which are then grouped to build the random forest.

After the random forest has been created, the model is then used to make forecasts (Alzubi et al., 2018). This will be explained later in this chapter (Section 3.11).

3.7 Applications and the Future of Machine Learning

Machine learning is the ideal solution to complex real-world issues. Seeing that machine learning has been in use for a long time, it has been applied and is still being applied to many real-world situations (Alzubi et al., 2018) such as detecting spam in email, detecting fraudulent activity on social media, online stock trading, face and shape detection, diagnosis of medical conditions, predicting what the traffic situation will be, detecting characters, product recommendation, and credit card fraud detection. Examples of companies using machine learning are Facebook for

friend suggestions, Amazon for recommendations of other products, Google's self-driving cars and Netflix's movie and series recommendations (Alzubi et al., 2018). In the financial sector, machine learning assists companies to minimise financial crime. To identify credit card fraud, algorithms like decision trees and neural networks have been employed (Bolton & Hand, 2001; Dal Pozzolo et al., 2014).

Artificial neural networks have been applied in studies in the insurance industry to assess the financial stability of insurance businesses and forecast solvency or insolvency (Ajibola et al., 2012).

Five categories of insurance have been classified using machine learning in prior experiments: life, annuity, health, accident, and investment-oriented insurance. The objective was to determine which insurance product would suit a customer best. This was created using feed-forward neural networks and the back-propagation algorithm (Lin et al., 2012).

Algorithms for machine learning have also been used to evaluate the accuracy of mortality predictions (Depre et al., 2017). Studies have been conducted in the pension sector to identify the explanatory variables influencing retirement choices. People's decisions to delay or improve their retirement are influenced by macroeconomic variables like the stock market and unemployment rate (Bosworth & Burtless, 2010; Coile & Levine, 2011). As previously mentioned, personality factors can predict a person's retirement date and path (Blekesaune & Skirbekk, 2012; Feldman & Beehr, 2011).

Machine learning is anticipated to be integrated into nearly all software programs. Machine learning uses natural language processing; that means it enables computers to comprehend the context and semantics of phrases (Alzubi et al., 2018).

3.8 Limitations of Machine Learning

Learning is a difficult process, seeing that many choices must be made. The process differs from model to model and from algorithm to algorithm in terms of how it comprehends a given situation and how it responds to it (Talwar & Kumar, 2013). These abundant choices make it

difficult for the machine to respond and react in some circumstances. The multiple decisions not only exacerbate existing issues, but also have an impact on how the model learns. The machine learning model depends on the input. The model has to take into account all the different outputs this input can produce and has to decide on the most appropriate and optimal output (Talwar & Kumar, 2013). Several of the typical problematic issues faced during the process of learning are:

- Volume of data: Machine learning algorithms demand large amounts of data that have high accuracy and efficiency. Such an amount of data is not always readily available to anyone. Front runners in technology such as Facebook and Google have been leaders in the Artificial Intelligence field, since they have had access to large amounts of data (Alzubi et al., 2018).

  Taigman et al. (2014) developed a software algorithm that can identify individuals' faces in photos. This algorithm also handles a human's ability to identify faces in photos. The authors state that the large number of photos available in recent years enables the machine to be more powerful.
- Bias: The preference of the machine for one hypothesis over another (Talwar & Kumar, 2013).

  Suppose there are two hypotheses present in a model and based on the data both hypotheses accurately predict all of the data. If the machine prefers one over the other, something external to the data would come into play (Talwar & Kumar, 2013).

  Bias is needed for a machine, because an inductive process to make predictions is needed on unseen data (Talwar & Kumar, 2013).

  The question that needs to be answered is what is considered a good bias.
- Noise: Noise refers to data that is not perfect. For example, missing entries, entry values that are assigned incorrectly, etc. An algorithm must be able to handle data with errors (Talwar & Kumar, 2013).

  Even though this is seen as one of machine learning's challenges, machine learning remains a tool that is very good at determining noise over a short period, and machine learning methods focus on dealing with heterogeneity in data (Liu & Xie, 2019).

- Pattern recognition: The main goal of pattern recognition algorithms is to conduct the "closest to" matching of the inputs, taking into account their statistical fluctuations and offering a fair response for all potential inputs. This stands in contrast to algorithms for pattern matching that match the precise values and dimensions. As with mathematical models, algorithms have well-defined values for forms like rectangles, squares, circles, etc. (Talwar & Kumar, 2013).
- Black box: Even professionals frequently struggle to comprehend the reasoning behind the choices made by the most effective machine learning systems. These systems remain complicated black boxes (Zhou et al., 2021).

  Black boxes refer to instances where machine learning systems aren't transparent. The lack of transparency may have serious repercussions or result in inefficient use of scarce, priceless resources (Zhou et al., 2021).

  Uncovering black boxes has become an interesting topic for researchers. Although many explanation methods have been investigated, evaluations are still necessary to quantify their quality, determine whether and to what extent the offered explainability achieves the defined objective, compare the available explanation methods, and recommend the best explanation for a given task based on the comparison (Zhou et al., 2021).

3.9 Potential Growth for Machine Learning

Machine learning algorithms are constantly being improved and will undoubtedly expand in the years to come (Alzubi et al., 2018). Alzubi et al. (2018) identified in their paper that one of the challenges machine learning faces is that machine learning is unsuccessful in identifying objects and images. However, ChatGPT, which was launched at the end of 2022, is successful with identifying objects in images (Hu et al., 2023; Taecharungroj, 2023). This proves that machine learning is continually developing.

Alzubi et al. (2018) also indicated that there are some areas that machine learning can still be applied to:

- deep learning, which may be used to create voice-controlled gadgets, diagnose diseases, build circuits, forecast stock market movements, and much more;
- big data analytics and data mining, for example, to forecast business market trends;
- hardware accelerators from companies like AMD and Intel for future AI architectures;
- healthcare for medical image processing, handling of clinical data, and deciphering large-scale population genetic data;
- evaluation and testing environments for self-driving automobiles and virtual reality.
- Human computer interaction should continue to advance with improved interfaces and usability amongst various devices with the growth of cloud computing and the IoT (Internet of Things).

This study will now expand on Logistic Regression, Random Forests and Support Vector Machines, seeing that those are the models that will be used to model the preservation of retirement fund data.

3.10 Logistic Regression

Logistic regression is a popular regression analysis method (Alzubi et al., 2018; Peng et al., 2002). Regression analysis is a popular statistical technique used for describing and quantifying the relationship between an outcome of interest and one or more other factors (Worster et al., 2007). Logistic regression allows one to examine the relationship between one or multiple independent variables and a dependent variable (Speelman, 2014).

Overview of Logistic Regression

Logistic regression models are used for classification and regression (Alzubi et al., 2018; Maalouf, 2011; Peng et al., 2002). The response variable is the variable of interest. Predictor variables are the variables used as part of the hypothesis that is believed to have an effect on the response (dependent) variable (Speelman, 2014). The variables used for logistic regression can be categorical or numerical/continuous (LaValley, 2008; Speelman, 2014). Usually, dummy variables must be created for categorical variables. The response variable will be dichotomous in nature, while the predictor variables do not have to be dichotomous (Speelman, 2014).

Dummy variables are created in such a way that the logistic regression model is feasible. The variables are created in such a manner that they are numeric variables with numbers assigned to the categories. Typically, only the numbers 0 and 1 are used. The value 1 is often referred to as the "*success outcome"*, but that is just terminology. An example of how a dummy variable is created, is that the independent variable gender, with possible values F and M, can be encoded as a single dummy variable genderM, with value 0 meaning "not male" and value 1 meaning "male" (this is an arbitrary choice; genderF could also have been encoded, in which 0 means "not female" and value 1 means "female"). In general, a set of k-1 dummy variables can be used to encode a category variable with k-levels, or k-categories (Speelman, 2014). However, many dummy variable methods can be applied. (This will be more extensively discussed in the Methodology chapter). The response variable with a dichotomous outcome will enable the logistic regression model to classify new data points as well as make forecasts (Peng et al., 2002).

Regression determines the relationship between the response variable (dependent variable) and the predictor variables (independent variables). Logistic regression will represent this relationship between the response and predictor variables as a probability, that is, logistic regression explains how the predictor variables affect the response variable. Logistic regression expresses this probability through the Odds Ratio (OR). The most significant elements are identified by comparing the ORs of the predictor variables (Worster et al., 2007). The OR is the probability of an outcome occurring, divided by the probability of the event not happening (LaValley, 2008).

The General Logistic Regression Model

The logistic regression model is based on the odds of the binary outcome happening. The natural logarithm of the odds is used by the logistic regression model as a regression function of the response variables. This has the formula of (LaValley, 2008):

$$ln[odds(Y = 1)] = \beta_0 + \beta_1 X \qquad (3.1)$$

where:

- ln is the natural logarithm.
- Y is the response variable:
  - If Y=1 then the event happens.
  - If Y=0 then the event does not happen.
- β_0 is the intercept term.
- β_1 represents the regression coefficiënt.
- X is the predictor.

A more formal equation would be (Hilbe, 2009):

$$logit(Y) = ln[\frac{\pi}{1-\pi}] = \alpha + \sum_{n=1}^{p} \beta_i \boldsymbol{X}_i \qquad (3.2)$$

Where:

- **X** is a vector with the number of instances.
- π is the probability that Y is equal to the outcome of interest, given the vector **x** (a specific value of **X**).
  - If Y =1 then the event happens.
  - If Y =0 then the event does not happen.
- α is the intercept term.
- β_i represents the regression coefficiënt for variable i ($i = 1, 2, \ldots, p$).

Limitations of Logistic Regression

Linear regression has an underlying linearity assumption. This assumption is often breached by using datasets that are not linearly separable. However, kernel approaches have made it possible to use non-linearly separable data in logistic regression (Peng et al., 2002). Kernels determine the relationships between variables in data (Canu & Smola, 2006).

For logistic regression, unlike linear regression, there is no formula for the estimators. Accurate estimators are obtained by testing different values of estimators until the best estimator for the most accurate prediction is obtained. While it is simple to do logistic regression on a computer and many statistical software packages can do it, this makes the method less comprehensible and more of a "black box" approach for many researchers (LaValley, 2008).

Although logistic regression is a strong and useful tool, using it is not a particularly simple process. Variable selection is one of the more challenging steps in the creation of this model. The process of selecting the independent variables from a set of possible ones for inclusion in the (final) model is known as variable selection (Speelman, 2014). The challenge faced with selecting the variables is finding a balance between misinterpreting or over-reducing the patterns in the data and overfitting the "noise" in the data (Speelman, 2014). The stepwise logistic regression approach is a useful technique for minimising the number of redundant and/or unnecessary variables and for

feature selection. The strategy, which includes or excludes variables based on the fitted model's deviation as a function of those variables, is comparatively simple to use (Peng et al., 2002).

The odds ratio is often mistaken as relative risk. Thus, the interpretation of the results from the model can be incorrect. Relative risk is the ratio of probabilities and not the ratio of odds (LaValley, 2008).

As mentioned, a limitation of machine learning is noise (Liu & Xie, 2019; Talwar & Kumar, 2013f). This is a common issue for logistic regression as well, but Maalouf (2011) found that logistic regression is effective in handling noise in data.

Benefits of Logistic Regression

One of the key benefits of logistic regression is that the latter may be extended to multi-class classification problems and automatically generates probabilities. The fact that most techniques employed in logistic regression model analysis adhere to the same principles as those in linear regression is another benefit (Maalouf, 2011).

Logistic regression is flexible enough to handle data-mining difficulties such as collinearity, missing data, duplicated characteristics, and non-linear separability, among others, making logistic regression a potent and robust data-mining technique (Peng et al., 2002).

It is a good method to use for unstructured data (Speelman, 2014). Binary logistic regression models serve as the building blocks for more complicated models (Maalouf, 2011).

Examples of Real-world Applications of Logistic Regression

The evaluation of gender as a response variable of operative mortality after coronary artery bypass grafting surgery, the assessment of the association between the TaqIB genotype and risk of heart diseases in a meta-analysis, and a study of the association between abnormalities in lipoprotein and the prevalence of diabetes are all examples where logistic regression was used (LaValley, 2008).

3.11 Random Forest

Random forests can be used for classification as well as regression, including multiclass classification. Random forests have been utilised effectively for a wide range of applications and are very popular across many fields (Cutler et al., 2012).

There are different types of random forests, each of which can be distinguished by the following characteristics (Boulesteix et al., 2012):

1. how each tree is built;
2. how modified data sets are produced;
3. how each tree's predictions are combined to create a singular consensus prediction.

Overview of Random Forest

The variables used in this model can either be categorical or numerical. This applies to the dependent and the independent variables. If the objective is classification, the dependent variable should be categorical and if the objective is regression, then the dependent variable should be numerical. The independent variables can be categorical or numerical regardless of using the model for classification or regression (Cutler et al., 2012).

In this study, categorical variables will be used to build the models. Thus, the study will only describe how the random forest model works when the model does classification. However, as mentioned, random forests are also good to use for regression.

Random forests have two parts to the model. The "*forests"* refer to the collection of decision trees and "*random"* refers to the bagging approach (Boulesteix et al., 2012; Oshiro et al., 2012).

Firstly, decision trees will be discussed, seeing that a random forest is an ensemble of decision trees.

A decision tree consists of a root node, leaf nodes and terminal nodes. The root node contains the condition to split the data. This condition reflects a choice between two options. These

options will then lead to a specific leaf node – depending on the answer of the initial condition. The leaf nodes will either have another condition or be terminal nodes, which indicates that an answer to the classification problem has been reached. If the leaf node has another condition, then the process will just be repeated until terminal nodes are reached (Cutler et al., 2012). This is shown in Figure 3.2.

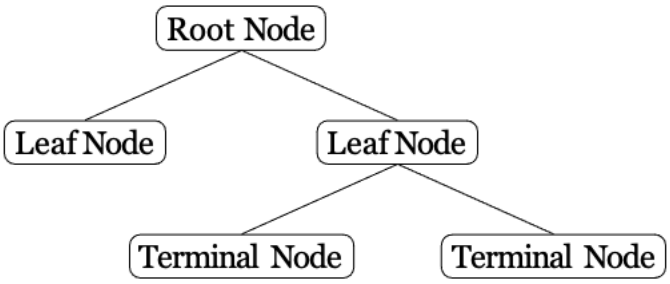


Figure 3.2: Decision Tree

The decision tree can be explained with an example. Suppose we have data on a population of individuals aged 0–19. The data includes the sex and age of the individuals. The objective of the decision tree is to determine whether an individual is a teenage girl (ages 10–19). Starting with all the data at the root node, the condition is whether the individual is male or female. All the males will then move to the leaf node, which is also a terminal node because no further classification is needed for this group of data. The individuals who are female will be presented with another condition in the other leaf node. The condition could ask whether the age is less than or equal to 9. If yes, then the data is moved to the terminal node which does not contain the objective, and if no, it is moved to the terminal node containing the objective.

This process involving the decision tree seems fairly simple. However, the model needs to determine the optimal split (this is what makes it machine learning). The split refers to the relevant condition. The optimal CART-split criterion is used to choose the appropriate cut at each node of each tree based on the so-called Gini impurity (for classification) or the prediction squared error (for regression) (Biau & Scornet, 2016). In simple words, the Gini impurity is the sum of the

probability that the data point is a certain class, multiplied by the probability of it not being that class (Nembrini et al., 2018). The probability of a certain class can be explained with an example. Suppose in the original data there are 60% males and 40% females. Then, if the data point is a male, the probability of 60% is used. For multiple data points, the Gini impurity will be the summation of each point's Gini impurity. This method helps reduce the impurity at the root nodes and leaf nodes with conditions (Nembrini et al., 2018).

The issue with decision trees is that they are very sensitive to the training data. Thus, the variance is high. The random forests mostly prevent this issue by the "random" component (Boulesteix et al., 2012; Cutler et al., 2012).

As for the random part, the randomness of a random forest is due to the "bagging" that is applied throughout the creation of the model (Biau & Scornet, 2016; Oshiro et al., 2012). As mentioned, random forests are ensembles of decision trees. Random samples of the original data are used to build the collection of independent decision trees. Random sampling with replacement is applied to the dataset. This is called "bootstrapping". Additionally, not all the features of the original dataset will be used in the decision tree. Different combinations of the features will be used to build the tree. These combinations are randomly chosen as well. The new data point will then pass through each of the different decision trees and the classification of each tree will be noted. Then, majority voting will be applied to the combined classifications of the decision trees to determine the final classification. This step is referred to as "aggregation". "Bagging" is bootstrap aggregation (Biau & Scornet, 2016; Breiman, 2001; Oshiro et al., 2012). This can be seen in Figure 3.3.

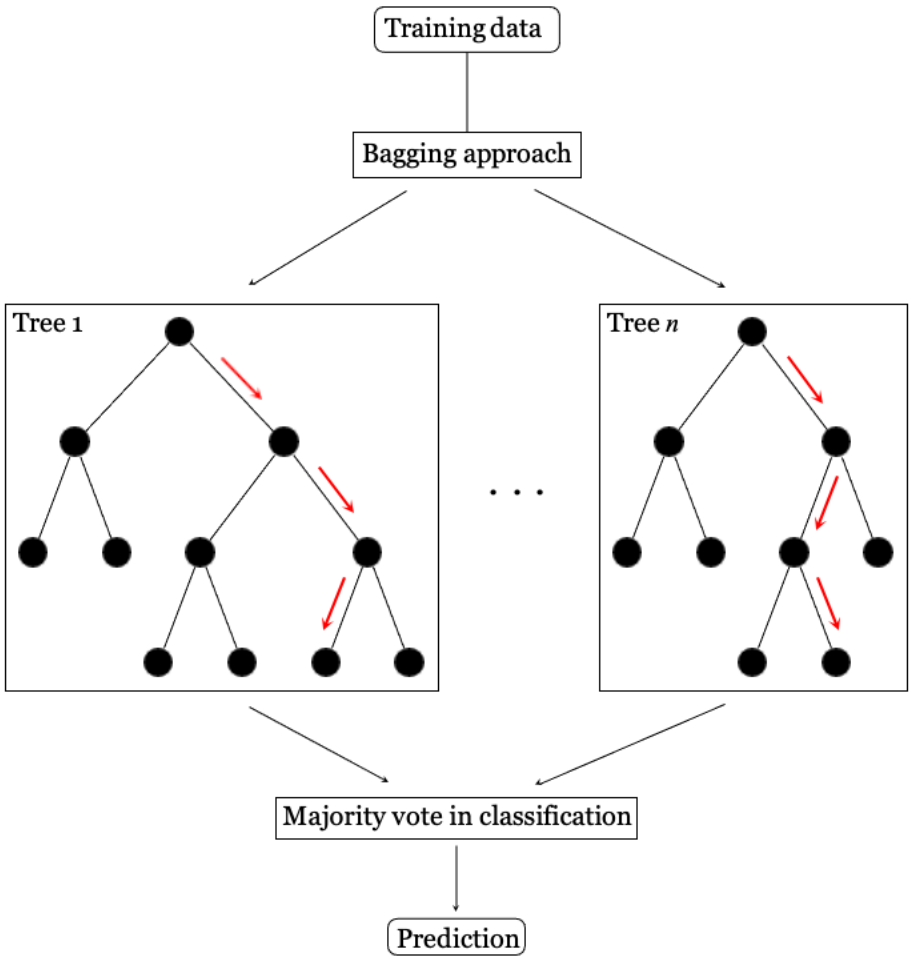


Figure 3.3: Random Forest

Bootstrapping refers to the situation where different decision trees do not use the same data to build the trees. This also reduces the random forest model to be less sensitive to the training data. The feature selection also helps to reduce the variance. This is because the feature selection reduces the correlation between the decision trees (Biau & Scornet, 2016; Oshiro et al., 2012).

Bagging has the side effect of not using all the data available in the original data set. These observations are referred to as the "out-of-bag" (OOB) observations and are used for determining which features (parameters) are the most optimal in the model and projecting the accuracy of the classification (Nembrini et al., 2018; Oshiro et al., 2012). A random forest model is more accurate

with categorisation when more decision trees were used to build the random forest model (Breiman, 2001). The original dataset is the training data and the new data is the test data.

The General Random Forest Model

Random forests have a less indisputable theoretical history, and despite their widespread application, little is understood about the mathematical underpinnings of the method (Biau & Scornet, 2016).

Numerous theoretical investigations have been carried out recently, analysing more complex models and getting closer to the real world (Biau & Scornet, 2016).

To build a *Decision Tree*, let $y_{k1},\ldots,y_{kn}$ represent the values from the dependent variable of the training data at node k, with k serving as the terminal node.

The following provides the dependent variable's predicted values for Classification:

$$\hat{h}(x) = argmax_y \sum_{i=1}^{n} I(y_{k_i} = y) \tag{3.3}$$

where

$$I(y_{k_i} = y) = \begin{cases} 1, & y_{k_i} = y \\ 0, & otherwise \end{cases}$$

Then, for the final prediction at a new point x of the *Random Forest* for classification (i.e. majority voting):

$$\hat{f}(x) = argmax_y \sum_{j=1}^{J} I(\hat{h}_j(x) = y) \tag{3.4}$$

Although bagging and the CART-splitting method are important components of the random forest process, formal mathematical analysis of both is challenging, which is why theoretical studies have up to this point focused on streamlined versions of the original procedure (Biau & Scornet, 2016).

Limitations of Random Forest

Random forests also have a black-box characteristic due to the approach having a blend of several elements, such as multiple decision trees, bootstrapping and feature selection. This makes it difficult to adequately analyse the model (Biau & Scornet, 2016).

Random forests still need to overcome several obstacles. In certain situations, they result in “odd unexpected results”, such as bias based on the kind of predictor (Boulesteix et al., 2012).

A benefit of random forests is the lack of a specific underlying stochastic model. This can, however, also be seen as a drawback because random forests do not fit into the statistical framework to which we are accustomed (which includes p-values, confidence intervals, etc.), and because it is challenging to understand what exactly happens in this dense jungle. Future analysis of the algorithm from a statistical perspective, maybe with the formulation of the method in terms of parameters and tests, may help to clarify both issues (Boulesteix et al., 2012).

Benefits of Random Forest

Random forests are flexible enough to be applied to supervised classifications and regression operations (Biau & Scornet, 2016).

The model is used to solve two basic challenges: to develop a prediction rule for a supervised learning issue and to rank variables according to their ability to predict the response. Random forests are thought to be particularly effective in identifying predictors involved in interactions (Boulesteix et al., 2012).

The random forest model has demonstrated exceptional performance in situations where there are many more variables than observations. It can handle intricate interaction patterns as well as

highly correlated variables and provide estimates of variable relevance (Biau & Scornet, 2016). It is independent of any specific stochastic model and is capable of capturing non-linear associations between predictors and responses (Boulesteix et al., 2012).

Multi-dimensional data can be handled by random forests, and the model can even be used in challenging situations with strongly correlated predictors (Boulesteix et al., 2012; Oshiro et al., 2012). Random forests have built-in variable significance measurements and are very versatile (Boulesteix et al., 2012; Cutler et al., 2012). Cutler et al. (2012) summarise why random forests are appealing from both a computational and statistical viewpoint. The appealing features from a computational standpoint include:

- Random forests naturally handle both regression and (multi-class) classification.
- Random forests are relatively quick to train and predict.
- They depend only on one or two tuning parameters.
- They have a built-in generalisation error estimate.
- They can be applied directly to high-dimensional problems.
- They are simple to implement in parallel.

The extra capabilities that random forests offer from a statistical viewpoint:

- measures of variable relevance,
- differential class weighting,
- missing value imputation,
- visualisation,
- outlier identification,
- unsupervised learning.

Random forests have attracted the attention of many, especially for classification tasks, due to their generalised performance, high accuracy and fast operation time (Du et al., 2015).

Examples of Real-world Applications of Random Forests

Random forests have been used in multiple fields in many recent studies and real-world applications (Oshiro et al., 2012). They have, for example, been utilised in the banking industry to distinguish between loyal and defaulting customers (Mu et al., 2018).

3.12 Support Vector Machine

A common tool for resolving multidimensional function estimate issues is the Support Vector approach (Vapnik et al., 1996). Support Vector Machines are a group of associated supervised learning techniques that are applied to regression and classification. They are a member of the generalised linear classifier family. Support Vector Machine (SVM) is a classification and regression prediction tool that automatically detects over-fitting to the data while maximising predicted accuracy using machine learning theory (Jakkula, 2006).

The foundation of SVM is statistical learning theory. It can be applied for future data prediction. By resolving a restricted quadratic optimisation issue, SVM is taught. SVM utilises a collection of non-linear basis functions to accomplish the mapping of inputs into a high dimensional space (Jakkula, 2006).

Overview of Support Vector Machine

SVM is a model that is very effective for classification. It is, in fact, one of the simplest and most elegant methods for classification (Chen et al., 2005; Jakkula, 2006). SVM transforms data from a low dimension to a higher dimension. It works in the following way concerning classification (Jakkula, 2006):

The objective is to classify a variable. This variable thus becomes the target variable. SVM will represent the target variable as a point in an n-dimensional space. The coordinate of this point is called features (Jakkula, 2006). For example, suppose you have data on a population of individuals aged 0–19. In the data, the variable present is the sex and age of the individuals. The objective is to classify whether an individual is a teenage girl (ages 10–19). Then, the coordinates of a data point will be: (age, sex).

SVM will classify the target variable by drawing a hyperplane. A hyperplane is a plane drawn in such a way that all the points of one category will be on one side of the plane and all the points of the other category will be on the other side of the plane. There are many ways to draw the hyperplane. SVM tries to find the hyperplane that best separates the two categories. The SVM will choose the hyperplane that maximizes the distance to points in both categories. This distance is the margin. The data points that fall on the margin are the supporting vectors (SV) (Cortes & Vapnik, 1995; Jakkula, 2006). This can be seen in Figure 3.4.

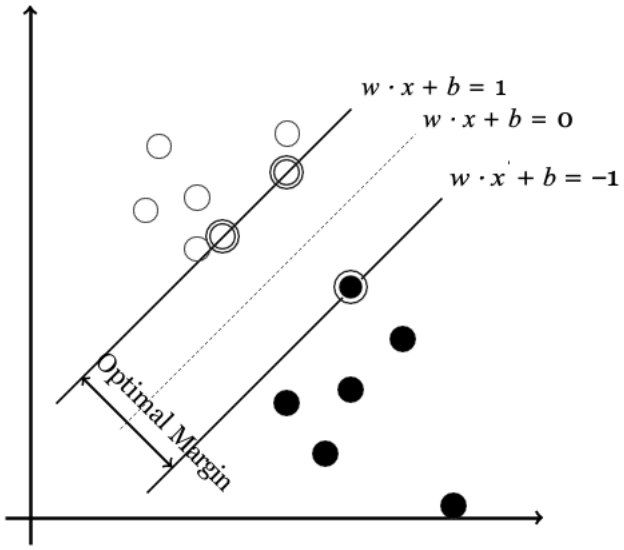


Figure 3.4: Representation of Hyperplane

The model is still left with the issue that data is not always linear, which means the points cannot be separated by a hyperplane. This is where the kernel trick comes into play (Chen et al., 2005; Jakkula, 2006; Talwar & Kumar, 2013).

The kernel trick allows the model to operate in the original n-dimensional space without separating the data points in a higher dimensional space. The kernel trick transforms the data by applying a relationship to the features. It can for example apply the observations to the power of 2. The kernel trick reduces computational time and avoids the complex math of working in a higher dimension (Chen et al., 2005).

Two common kernel functions are used: polynomial and radial basis function (Chen et al., 2005).

The General Support Vector Machine Model

Jakkula (2006) and Salcedo-Sanz et al. (2014) formally describe the SVM model to categorise data accurately.

Let training data be $\{x_i, y_i\}_{i=1}^{n}$ where $\boldsymbol{x_i} \epsilon \mathbb{R}^N$ and $y_i \epsilon \{-1, +1\}$ and a nonlinear mapping $\emptyset(.)$

In Figure 3.4 above, the circles with circles around them (on the solid lines) are the support vectors. The SVs define the margin of the largest separation between the two classes. x is a vector point and w is weight as well as a vector (Cortes & Vapnik, 1995).

By maximising x when x is on the hyperplane, it can be determined how far a point is from the origin of the hyperplane (Figure 3.4). The situation is the same for the opposing side points. The total of the two distances from the separating hyperplane to the nearest points is obtained by solving and subtracting the two distances (Jakkula, 2006; Salcedo-Sanz et al., 2014).

$$M = \frac{2}{\|w\|} \tag{3.5}$$

Maximising this margin (3.5) is equivalent to minimising it. Thus, the SVM method solves:

$$min_{w,\xi_i,b}\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \tag{3.6}$$

constrained to $\forall i = 1, \ldots, n$:

$$y_i(\langle \emptyset(x_i), w\rangle + b) \geq 1 - \xi_i \tag{3.7}$$

$$\xi_i \geq 0 \tag{3.8}$$

Variables ξ_i are called slack variables and these variables measure the error made at point $(\boldsymbol{x}_i, y_i)$. C is the regularisation parameter chosen by the user. This parameter is used to generalise the capability of the classifier (Salcedo-Sanz et al., 2014).

The next step is to find w and b in a quadratic optimisation function. The quadratic function under linear constraints needs to be optimised to resolve this issue. The answer entails creating a dual issue with an associated Lagrange's multiplier, α_i. To find w and b, $\|w\|^2$ needs to be minimised. The Lagrange multipliers should still correspond to the constraints (3.7) and (3.8). Thus, the decision function for any test vector x_* is given by:

$$f(x_*) = sgn\,(\sum_{i=1}^{n} y_i\,\alpha_i K(x_i\,, x_*) + b) \tag{3.9}$$

(Source for the formulation of SVM: Jakkula, 2006, and Salcedo-Sanz et al., 2014)

Limitations of Support Vector Machine

If there are only a few support vectors, SVM prediction performance and memory use are good, but if there are numerous support vectors, they may be inadequate. Although the default linear scheme is simple to understand, it might be challenging to understand how SVM classifies data when you use a kernel function (Talwar & Kumar, 2013). The requirement for a strong kernel function is one of the weaknesses of SVM (Jakkula, 2006).

Applications of Support Vector Machine

The three main benefits of SVM according to Deris et al. (2011):

1. Only two parameters, the upper bound and the kernel parameter, need to be selected.
2. The solution is singular, optimal and global for a linearly constrained quadratic problem.
3. The implementation of equation (3.6) – which minimises risk a – leads to an SVM model having good generalised performance.

These benefits have prompted academics to perform several studies on SVM theory and implementation (Deris et al., 2011).

The robust basis provided by both statistical learning theory and functional analysis is one of the most alluring aspects of kernel algorithms. Kernel approaches integrate statistics and geometry in a promising way, allowing us to interpret (and develop) learning algorithms geometrically in feature spaces unrelated to the input space (Chen et al., 2005).

SVM is also proven to not over-generalise during the classification of the model (Jakkula, 2006).

The training of an SVM model is quite simple, which is one of the main advantages (Jakkula, 2006). It is flexible to large dimensional data and allows for explicit control of the trade-off between classifier complexity and error (Jakkula, 2006).

Examples of Real-world Applications of Support Vector Machine

SVM became eminent when it was given pixel maps as input and performed handwriting recognition tasks with accuracy on par with complex neural networks that include intricate features (Jakkula, 2006).

SVM has successfully been applied in the wind energy field, estimation of non- invasive cardiac indices, antenna array processing, functional magnetic resonance imaging, image processing and remote sensing image classification (Salcedo-Sanz et al., 2014).

4. RESEARCH METHODOLOGY

The data used for this study was provided by a large South African insurance company. The data was given to help determine how to improve the preservation rates for retirement.

4.1 Data Preparation

The dataset consisted of members who belong to an umbrella fund. An umbrella fund is a fund to which employees of multiple employers belong (Pillay & Fedderke, 2022).

The dataset used in this study is the data consisting of disinvestments. Disinvestment means that a member of the fund either exited the fund or retired.

Originally this dataset with all the disinvestments had 684 298 observations. The data was filtered by only considering members with pension or provident funds and only members who changed employers, i.e. preservation members.

Pension and Provident Funds: The data was filtered to only consider members who have a pension or provident fund, because these are retirement funds. There were other fund plans in the dataset, but those plans were not disclosed by the company. Only the pension and provident plans were disclosed.

Preservation Members: This study is only interested in what individuals do with their retirement savings when they change employers. The company had a variable present in their data that indicated when an individual exited the fund on changing employers.

In the data, it did occur that individuals who changed employers preserved their payout and did not preserve the “other payout”. Thus, they had two preservation withdrawals, one preserved and the other not. A column was created to indicate when an individual did this. This was called the *Other Preservation Withdrawal*. This is considered a variable since it can be an indication that maybe an individual’s preservation of their retirement funds is better when they preserve some of their funds while still receiving an amount in cash.

A variable of interest is the one indicating when an individual preserves. This is indicated by the column of the company, as mentioned. When an individual had an “other preservation” withdrawal, their preservation status was determined by looking at how much they preserved. Thus, if the larger amount of the payout was preserved the member was marked as “Preservation” and if the larger amount was not preserved the member was marked as “No Preservation”.

After all the cleaning was applied, there were 262 202 unique members in the data. The full approach to cleaning the data is set out in the appendix.

4.2 Data Preprocessing

In this section, the dependent and independent variables are discussed. The preprocessing applied to these variables is also discussed.

Model Variables

The list of variables used for modelling is provided in Table 4.1.

Variable	Representation
Contract Number	Provident or pension fund
Gender	Male or female
Age band	Age of the member when first changing employer
Salary Band	Member's current salary
Commission Share	Proportion of commission payable to the broker responsible for the corresponding contract.
Industry	The industry in which the member is employed
Team Manager	Manager of the company's (company who provided the data) consulting team allocated to the member
Custom Broker Name	The broker of the member
Channel	Distribution channel through which the product was sold to members.
Other Preservation Withdrawals	When a member preserved part of the payout and did not preserve the other part of the payout.
Preservation	Indicates whether the member preserved the retirement funds or not

Table 4.1: Variables Used for Modelling

The literature review shows that in classification models dependent and independent variables are required. In this study, *Preservation* is the dependent variable in the models, and *Contract Number*, *Gender*, *Age band*, *Salary Band*, *Commission Share*, *Industry*, *Team Manager*, *Custom Broker Name*, *Channel* and *Other Preservation Withdrawals* are the independent variables.

To prepare the datasets for the models and to enhance the performance of the models several transformations were applied to the data.

Feature Engineering

It is an important step of machine learning to apply feature engineering. Preprocessing the data and applying feature engineering to the data improves the performance of models (A. Zheng & Casari, 2018).

"Features" refer to the input of a machine learning model, i.e. the variables. Features are the bridge between raw data and machine learning models.

The process of extracting features from unprocessed data and converting them into forms appropriate for machine learning models is known as feature engineering. Feature engineering can make modelling less challenging and improve the results obtained (A. Zheng & Casari, 2018).

A wide range of feature engineering steps can be applied. Some examples include: filtering, binning and scaling (A. Zheng & Casari, 2018). Seeing that feature engineering improves model performance, this study will apply it to the data.

Feature transformations are one of the steps of feature engineering that could be applied to data. Good features should not only accurately reflect important parts of the data but also adhere to the model's assumptions. Transformations are frequently required (A. Zheng & Casari, 2018).

This study only considered feature transformation for categorical variables, seeing that variables used for modelling are categorical variables. Feature transformation for categorical variables is usually referred to as "encoding" (Potdar et al., 2017; A. Zheng & Casari, 2018).

Categorical variable encoding is done to convert non-numerical variables to numeric variables. Common variable transformations are One-hot Encoding, Dummy Encoding and Ordinal Encoding. Usually, these are the first types of transformation are used (Potdar et al., 2017; A. Zheng & Casari, 2018). This study will use ordinal encoding and dummy encoding. One-hot encoding is not included because it is very similar to dummy encoding.

It has been found, however, that the two methods used for the study do not always treat datasets with a large number of categorical variables well (A. Zheng & Casari, 2018). Thus, another type of transformation, Target Encoding, which is a common method used for a large number of categorical variables (Nazyrova et al., 2022), is included.

Before the data was split into a training and test set, the encoding methods were applied to the data. Descriptions of each of these encoding methods are below:

Ordinal Encoding

Ordinal encoding allocates an integer to each category present in the variable. It does not add or remove any columns to or from the data (Potdar et al., 2017). An example of ordinal encoding is:

	e_1
Ordinal Encoding	1
Dummy Encoding	2
Target Encoding	3

Table 4.2: Ordinal Encoding of a Category of Encoding

Similarly, ordinal encoding was applied to each variable the model uses. How the integers are matched to each category of each variable can be found in the appendix.

The dependent variable, Preservation, has the integer 1 if individuals preserved their retirement funds and the value 0 if they did not preserve their retirement funds.

Dummy Encoding

Dummy encoding is very similar to one-hot encoding. The one-hot method is simple to apply (Potdar et al., 2017; Rodríguez et al., 2018). Every level of the category variable is compared to a predetermined reference level. A single variable with n observations and d different values is converted into d binary variables with n observations each using a one-hot encoding. Each observation indicates the dichotomous binary variable's presence (1) or absence (0) (Potdar et al., 2017).

However, one-hot encoding allows for d different values while only d-1 is needed. Dummy encoding removes this extra value. This feature is represented as a column of zeros (A. Zheng & Casari, 2018). Below is an example of dummy encoding:

	e_1	e_2
Ordinal Encoding	1	0
Dummy Encoding	0	1
Target Encoding	0	0

Table 4.3: Dummy Encoding of a Category of Encoding

Dummy encoding is preferred above one-hot encoding because it is easier to understand (A. Zheng & Casari, 2018).

Dummy encoding was applied to each variable the model uses. In the appendix a description of how the integers is matched to each category of each variable is included.

As with ordinal encoding the dependent variable, Preservation, has the integer 1 if individuals preserved their retirement funds and the value 0 if they did not preserve their retirement funds.

Target Encoding

Target encoding replaces each category with the average target value for samples that fall inside that category. The independent variable is the target value of the model (Nazyrova et al., 2022).

This method allows the preservation of the original raw data information, seeing that it does not reduce the raw data's dimensionality. The method is known to perform well with large datasets with categorical variables (Nazyrova et al., 2022). Target encoding is only applied to the independent variables. The dependent variable, Preservation, stayed the same as it was with ordinal and dummy encoding (the integer 1 if individuals preserved their retirement funds and the value 0 if they did not preserve their retirement funds).

After each encoding method had been applied to the data, the training dataset and a testing dataset were then created. Thus, there were three training datasets (ordinal, dummy and target) and three test datasets (ordinal, dummy and target) The training data was 70% of the initial dataset and the test data was 30%.

4.3 Modelling

Three different machine learning models are used: Logistic Regression, Random Forest and Support Vector Machine. These models will be used to classify the dataset and determine which factors affects the preservation of the retirement fund.

Furthermore, the different features engineering methods are applied.

The predictor variables used for the models are categorical variables that are used to predict a dichotomous outcome: Preservation (1) or No Preservation (0).

These models were all created in Python. The packages that were used for all the models are *pandas* and *sklearn*. Pandas is used to import datasets and sklearn has multiple functions that were used for the models. The specific functions used for specific models will be described below.

Additionally, the *statsmodels.api* package was used for the creation of the logistic regression model.

Logistic Regression

For the logistic regression model, a constant column (a column containing the number 1 for each member) was added. This was done because the specific method used in Python does not automatically do it. It was needed to determine the x-intercept (β_0), as in the formula.

The logistic regression model was built with the *sm* function from the statsmodels.api package. The confusion matrix was determined with the function *confusion_matrix*. The AUC curve was created with the *roc_curve* and *auc* functions. Similarly, the precision-recall curve was created with the *precision_recall_curve* function. The confusion_matrix, roc_curve, roc functions and precision_recall_curve are from the sklearn package. The AUC and PR curves were plotted with the *plt* function from the *matplotlib.pyplot* package.

Random Forest

The random forest model was built with the *RandomForestClassifier* function from the sklearn package. The confusion matrix was determined with the function *confusion_matrix*. The AUC curve was created with the *roc_curve* and *auc* functions. Similarly, the precision-recall curve was created with the *precision_recall_curve* function. The confusion_matrix, roc_curve, roc functions and precision_recall_curve are from the sklearn package.

The AUC and PR curves were plotted with the *plt* function from the *matplotlib.pyplot* package.

Support Vector Machine

The SVM model will be built as previously discussed.

One of the issues faced when building an SVM model is choosing parameters. Three parameters can be chosen, the parameter that defines the regularisation of the error (C), the kernel parameter and the gamma parameter (Jakkula, 2006).

Parameter C usually has a default of 1. If this parameter is small, it can lead to possible misclassifications. If the parameter is large, it can lead to fewer possible misclassifications. However, the higher the value parameter, the higher the probability of overfitting the model. Thus, the model will not generalise the test data very well (Jakkula, 2006).

Kernel parameter refers to the function that is used to transform the data. There are multiple options: linear, nonlinear, polynomial, radial basis function (rbf), and sigmoid. The default is usually rbf. This parameter determines the relationship between variables in the data (Jakkula, 2006).

The **gamma** parameter prevents overfitting the data. A low value of gamma will increase the variance, but lower the model bias, and inversely, a high value of gamma will increase the model bias, but decrease the variance (Jakkula, 2006).

For this study, a simple SVM model was created. The following parameters were used: *kernel=rbf, C=0.1, and gamma=0.5*. Thus, C and gamma are of neither too small nor too large value.

The SVM model was built with the *svm* and SVC functions from the sklearn package. The confusion matrix was determined with the function *confusion_matrix*. The AUC curve was created with the *roc_curve* and auc functions. Similarly, the precision-recall curve was created with the *precision_recall_curve* function. The confusion_matrix, roc_curve, roc functions and precision_recall_curve are from the sklearn package.

The AUC and PR curves were plotted with the *plt* function from the *matplotlib.pyplot* package.

4.4 Evaluation of the Models

Using the right assessment criteria is crucial. Accuracy, precision, inverse precision, recall, AUC and Precision-Recall curve will be used to evaluate this study's models. A description of what each metric is and why it is used is given in this section.

Confusion Matrix

Evaluation metrics such as precision, inverse precision, recall, accuracy and the ROC curve are best understood with the help of the confusion matrix (O'Reilly & Nielsen, 2013). The confusion matrix shows the following (Powers, 2020):

- True Positives (TP) – projected positive cases that were correctly classified.
- False Positives (FP) – projected positive cases that were incorrectly classified.
- True Negatives (TN) – projected negative cases that were correctly classified.
- False Negatives (FN) – projected negative cases that were incorrectly classified.

The confusion matrix usually takes on the following structure:

		Predicted	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

Table 4.4: Confusion Matrix

Accuracy: This metric presents the percentage of accurate instances – both positive and negative – that are found among all the results. Precision and inverse precision are weighted arithmetic means that make up accuracy (Dalianis, 2018).

Accuracy is calculated as follows (Nivedha & Sairam, 2015): $\frac{TP+TN}{TP+TN+FP+FN}$

Precision: By dividing the number of accurately predicted positive instances by the total number of positive instances predicted, precision is calculated (Dalianis, 2018).

Precision (P) is calculated as follows (Nivedha & Sairam, 2015): $P = \frac{TP}{TP+FP}$.

In the case of this study, the precision will be: out of all the individuals who changed employers (and exited the fund), the number that the model predicted to preserve their retirement funds were truly individuals who preserved their retirement funds.

Inverse Precision: By dividing the number of accurately predicted negative instances by the total number of negative instances retrieved, inverse precision is calculated (Powers, 2020).

Inverse Precision (IP) is calculated as follows (Nivedha & Sairam, 2015):$IP = \frac{TN}{TN+FN}$.

In the case of this study, the inverse precision will be: out of all the individuals who changed employers (and exited the fund), how many that the model predicted to not preserve their retirement funds were truly individuals who did not preserve their retirement funds.

Recall: Dividing the number of accurately predicted positive instances by the number of truly positive instances is called recall (Dalianis, 2018).

Recall (R) is measured as follows (Dalianis, 2018): $R = \frac{TP}{TP+FN}$

In the case of this study, the recall is from all the cases where an individual was predicted by the model to preserve the retirement funds and the individual did preserve correctly.

The Area Under the Curve

The Area Under the Curve (AUC) refers to the Receiver Operating Characteristics (ROC) curve. Thus, the ROC will be discussed first.

Receiver Operating Characteristics

A receiver operating characteristics graph shows how the correct positive classifications differ from the negative instances incorrectly classified (Davis & Goadrich, 2006). Using a ROC graph, classifiers can be visualised, grouped and chosen according to their performance. In signal detection theory, ROC graphs have long been used to show the trade-off between classifier hit rates and false alarm rates.

ROC became a popular metric due to accuracy not being a sufficient metric by itself (Fawcett, 2006). ROC shows how the correct positive classifications differ from the negative instances incorrectly classified (Davis & Goadrich, 2006).

The relative tradeoffs between false positives and true positives are shown on a ROC graph (Fawcett, 2006). Important points on the ROC curve are the lower left point (0,0) and the upper right point (1,1). The first point represents the classification never giving true positives or false positives and the latter point represents always giving positive classifications. The point (0,1) is the "perfect" point, representing a perfect classification. The goal of the curve is to be in the upper left triangle. If a classifier is close to the X-axis then the classifier can be seen as "conservative" and when a classifier is on the upper right side of the curve it can be seen as "liberal". Conservative classifiers make very few false positive errors and make positive classifications only. A liberal classifier makes a positive classification but does not have strong evidence. Thus, all positives are classified correctly but they have high false positive rates (Fawcett, 2006). The $y = x$ is the line representing the class of random guessing. If a classifier is on the right side of this line it is better off guessing its class than classifying it (Davis & Goadrich, 2006; Fawcett, 2006).

The ROC is a graph with the sensitivity (true positive rate) on the Y-axis and $1 - specificity$ (false positive rate) on the X-axis (Dalianis, 2018; Fawcett, 2006).

- sensitivity = $\frac{TP}{TP+FN}$
- specificity = $\frac{TN}{FP+TN}$

"Specificity" is the percentage of negatives that are accurately classified as negative and "sensitivity" is the proportion of negatives that are correctly detected (for example the proportion of healthy individuals who are accurately identified as not having the ailment), which is the same as recall (Dalianis, 2018). Sensitivity is also known as the true positive rate and the specificity can be given in terms of the false positive rate (specificity = 1-false positive rate). Sensitivity, specificity and accuracy are typically used together to measure the performance of a model. However, this may cause the performance to be optimistic rather than accurate, especially when there is a class imbalance (Fawcett, 2006).

Area Under the Curve

ROC and AUC have been used for pattern recognition (O'Reilly & Nielsen, 2013). It may be that the ROC's performance needs to be reduced to a single scalar number that represents expected performance to compare models. Calculating the area under the ROC curve (**AUC**) is a standard technique (Fawcett, 2006).

The AUC's value will always fall between 0 and 1 since it represents a fraction of the area of the unit square. No practical classifier should have an AUC below 50%. The AUC of a classifier is comparable to the likelihood that the classifier would score a randomly chosen positive instance higher than a randomly chosen negative instance, which is an essential statistical feature of the AUC (Fawcett, 2006).

Precision-Recall Curve

The ROC fails to give a complete picture of the model's performance when there is an imbalance in the data. Usually, the imbalance will be due to having more negative classes than positive classes (O'Reilly & Nielsen, 2013). Precision and recall are preferred metrics when working with a class imbalance in data. These metrics focus on positive events while ignoring the accurate predictions of negative classifications (Torgo & Ribeiro, 2009).

Precision-Recall (PR) curve gives a complete picture of how a model's algorithm performed (Davis & Goadrich, 2006). The precision-recall curve graphs the precision as a function of recall

(Cook & Ramadas, 2020). The objective of ROC is to be in the upper left-hand corner and the objective of PR is to be in the upper right-hand corner (Davis & Goadrich, 2006).

If the objective is to classify an instance that does not occur frequently, such as the preservation of retirement funds, then using a PR curve is informative in describing the performance of a model (Cook & Ramadas, 2020). This curve is especially beneficial when precision is the variable of interest (O'Reilly & Nielsen, 2013).

Precision and Recall are metrics used to evaluate the performance of a model. Sometimes high precision is preferred over high recall and vice versa (Dalianis, 2018). Generally, precision will increase if recall decreases and vice versa (Torgo & Ribeiro, 2009; Zhu, 2004). The decision to use either precision or recall usually depends on the goal of the model. If it is more important to have more true positives, then precision is more important, and if it is more important to have low false positives, then recall is more important (Cook & Ramadas, 2020).

Accuracy can also be high but precision low, indicating the system operates well but the results generated are significantly dispersed. Contrast this with striking the bullseye, which implies both high accuracy and high precision (Dalianis, 2018).

A misleading representation can be given when the model is created with unbalanced data (O'Reilly & Nielsen, 2013). It is expected in a dataset with a class imbalanced with more negative classes than positive classes to have a specificity that tends to 1. This is because TN will be a larger value and the FP a smaller value in comparison with TP. TP is a smaller value relative to the negative cases. If the FP is much larger than the TP a very low precision will be obtained. Thus, in datasets where the negative cases outweigh the positive ones the precision is much more important than the specificity. Sensitivity and accuracy also have lower use for datasets such as these (O'Reilly & Nielsen, 2013). AUC is insensitive to unbalanced data (Bradter et al., 2022). It can be higher with balanced data, but it would not be significant (Bradter et al., 2022; Xue & Hall, 2014).

ROC also gives an optimistic performance of the model when the data is unbalanced. Thus, the PR curve is preferred (Torgo & Ribeiro, 2009). However, it has been found that ROC should

not just be ignored. It is a tool that has been effective for many years. It should rather be used in conjunction with the precision-recall curve. The precision-recall curve will merely highlight precision more than the ROC does (O'Reilly & Nielsen, 2013). Only when a PR curve dominates, does a ROC curve (Davis & Goadrich, 2006) dominate.

PR curves are more volatile than ROC curves (Cook & Ramadas, 2020). PR curves are more suited to ensuring that all the instances are marked positive, whereas ROC is more suited to identify a high percentage of positives (Cook & Ramadas, 2020). If the objective of a model is to have all the positive instances, then the ROC is preferred. When there are many negatives it can be difficult to determine the number of false positives among the observations from the ROC (Cook & Ramadas, 2020).

The data in this study is unbalanced. However, one of the study's objectives is to understand the retirement fund preservation field. It is more important for the study to have higher precision. Thus, the unbalanced data is used which favours precision. This is supported by the study of O'Reilly and Nielsen (2013) that states in cases of having more negative classes than positive classes, precision is more important. Other metrics will still be given, but precision is considered as most important.

5. RESULTS

The objectives of this study are to determine what variables drive the preservation of retirement savings and to investigate whether machine learning models can classify retirement fund preservation data.

5.1 Descriptive Statistics of Data Used in Modelling

To identify significant drivers of retirement fund preservation decisions, the variables needs to be understood. Thus, before researching whether machine learning models can classify the preservation of retirement fund data, the study will first examine the initial statistics of the data used. This can also show why a variable might be significant.

5.1.1 Dependent Variable

As mentioned, the variable Preservation is the dependent variable that the models will try to classify. This variable indicates whether individuals preserved their retirement funds or did not preserve their retirement funds.

Only 17 626 of the members who changed employers preserved their retirement funds.

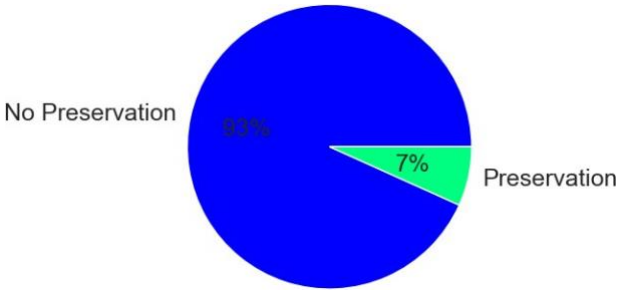


Figure 5.1: Number of Members that Preserved

In Figure 5.1 it can be seen that only 7% of the members preserved their retirement funds.

5.1.2 Independent Variables

Contract Number: As mentioned, only provident and pension funds were considered. About 27% of the members have a pension fund and 73% of the members have a provident fund.

Gender: 38% of the members in the dataset are females and 62% are males.

Age Band: The age band of the member who changed or exited an employer (i.e. preservation withdrawal).

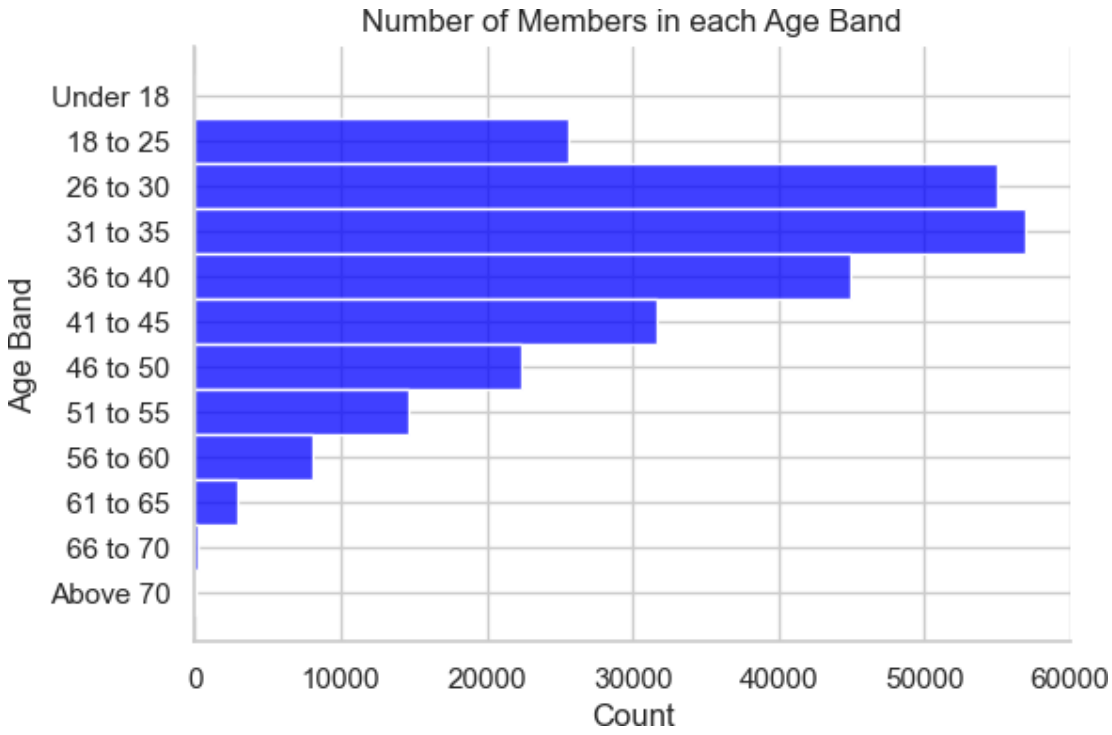


Figure 5.2: Number of Members in Age Band

Age Band was labelled as in Table 5.1 below for the models. The percentage of members in each band is added. (This information can also be seen in Figure 5.2 above):

Age Band	Approximate Percentage of Members
A - Under 18	0%
B - 18 to 25	10%
C - 26 to 30	21%
D - 31 to 35	22%
E - 36 to 40	17%
F - 41 to 45	12%
G - 46 to 50	8%
H - 51 to 55	6%
I - 56 to 60	3%
J - 61 to 65	1%
K - 66 to 70	0%
L - Above 70	0%

Table 5.1: Percentage of Members in Age Band

Based on the literature, the preservation rates are expected to be low due to the large percentage of young adult members.

Salary Band: Member's pensionable salary.

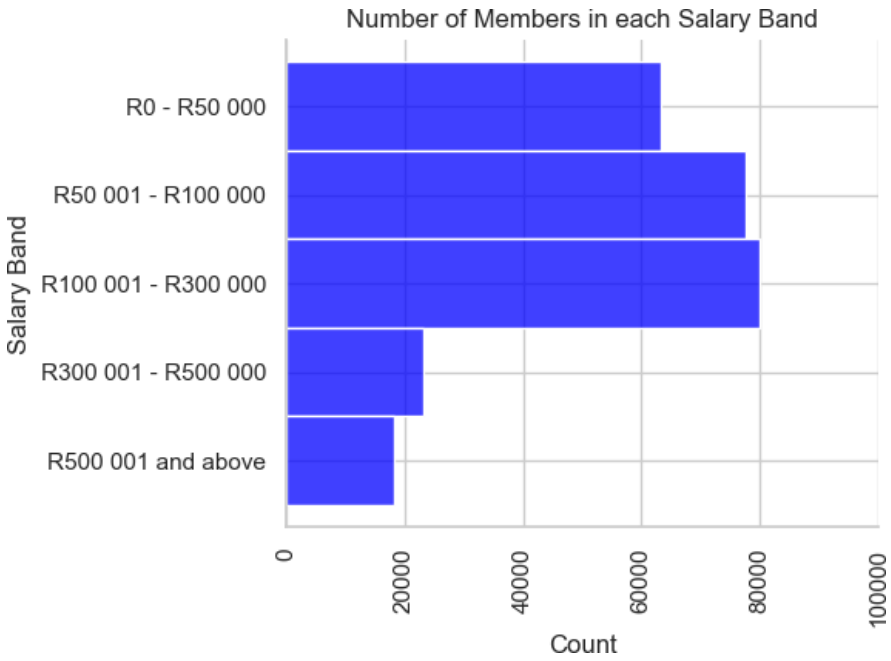


Figure 5.3: Number of Members in Salary Band

The Salary Band was labelled as in Table 5.2 below for the models. The percentage of members in each band is added. (This information can also be seen in Figure 5.3 above):

Salary Band	Approximate Percentage of Members
A – R0 - R50 000	24%
B – R50 001 - R100 000	30%
C – R100 001 - R300 000	30%
D – R300 001 - R500 000	7%
E – R500 001 and above	9%

Table 5.2: Percentage of Members in Salary Band

Usually, individuals in lower income brackets are less inclined to preserve their retirement funds. Most of the members in this case are in the lower salary bands, thus low preservation rates are expected.

Industry Code: This is the industry of the claiming member. The industries are distributed among the members as follows:

Industry	Percentage of Members
Agriculture, Forestry, Fishing and Co-ops	5.00%
Construction and Maintenance	8.22%
Education	2.93%
Entertainment and Hospitality	8.76%
Finance, Insurance and Real Estate	4.73%
Health and Welfare	2.94%
IT and Telecommunication	2.69%
Manufacturing	11.73%
Mining and Raw Materials	7.63%
Retail	12.29%
Security Services	2.15%
Transportation	5.31%
Unions and Labour Brokers	1.78%
Professional and Administrative Services	13.87%
Business Services	2.15%
Other	7.81%

Table 5.3: Number of Members in Each Industry

Members in the industry in the group Other, are members from industries that make up less than 2% of the data individually.

Custom Broker Name: The custom broker is the broker responsible for the employer group.

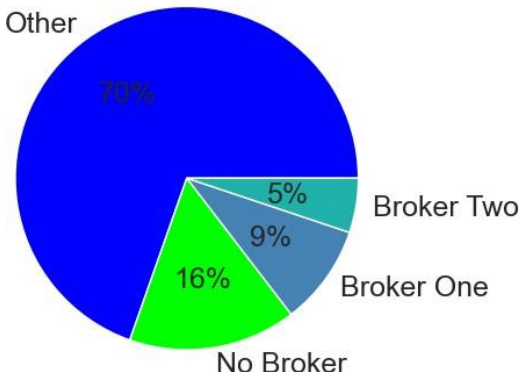


Figure 5.4: Custom Broker Name

As seen in Figure 5.4, four different groups of brokers were considered. The custom broker name was grouped as follows:

- Broker One – broker that belongs to the company that provided the data. 16% of the members belong to Broker One .
- Broker Two – most common broker. This broker does not belong to the company providing the data. 9% of the members belong to Broker Two .
- No Broker – members who do not have brokers associated with them. 5% of the members do not have a broker .
- Other – brokers who are neither Broker One nor Broker Two. 70% of the members belong to Other .

Commission Share: Proportion of commission payable to the broker responsible for the corresponding contract. This is the amount that the broker earns for handling the contract. It was included to see whether members with brokers who earn bigger commission percentages have better preservation behaviour.

Commission Share has the following bands:

Proportion of Commission Share	Percentage of Members
Under or equal to 40	5.17%
41 to 50	3.12%
51 to 60	0.05%
61 to 70	0.10%
71 to 80	0.10%
81 to 90	0.08%
91 to 100	91.38%

Table 5.4: Percentage of Members that pay the Proportion of Commission to Broker

As seen in Table 5.4, most members pay between 91% and 100% commission to their broker.

Team Manager: The team manager is the manager of the company's consulting team allocated to the client. The consulting team belonged to the company that provided the data. The team managers were grouped as follows:

- No Team Manager – no team manager was associated with the member. 2% of the members have no team manager.
- Team Manager One – most common team manager. 12% of the members have Team Manager One as their team manager.
- Team Manager Two – second most common team manager. 11% of the members have Team Manager Two as their team manager.
- Other – all other team managers. These members do have a team manager. 75% of the members have Other as their team manager.

Channel: The different distribution channels are:

- Independent – the company's distribution services, which usually sells group business. 102 871 members obtained their product through the independent distribution channel.
- Custom – 26 900 members obtained their product through the custom distribution channel.
- Direct Cluster – the company's internal team. 27 618 members bought their products through the direct cluster distribution channel.
- Channel not categorised – distribution channel used not known. 4 946 members' distribution channels are not indicated.
- Integrated – custom corporate brokers. 74 339 members obtained their products through the integrated distribution channel.
- Company's Business – A channel that only sells products from the company. 25 528 members obtained their products through the company's business distribution channel.

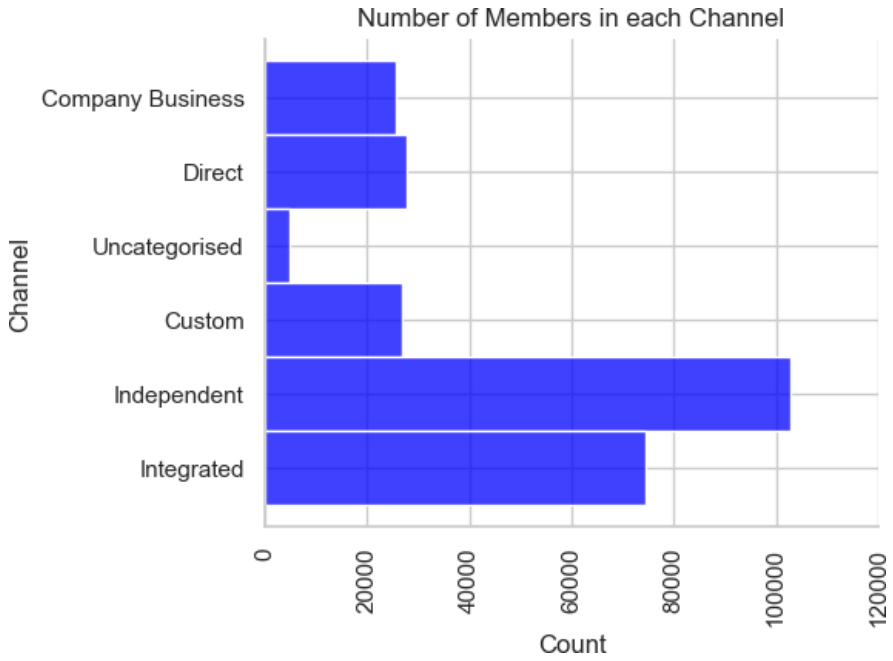


Figure 5.5: Number of Members by Different Distribution Channels

Other Preservation Withdrawals: This indicates members preserving part of their payout and not preserving the other part when changing employers. About 2% have an "other preservation" withdrawal. As mentioned, this is potentially an important variable because it can show that if the individual takes a part in cash and preserves the other part of their retirement funds, preservation of retirement funds can either improve or not improve.

5.1.3 Correlation of Variables

Before looking at the results of the machine learning models, the correlation of the variables is first examined. The correlation will help the study find whether any variables have a strong relationship with the dependent variable and if independent variables have strong mutual relationships. A strong correlation can be an indication of whether a variable is significant due to another variable's significance. Below is the correlation of all the variables that will be used in the model.

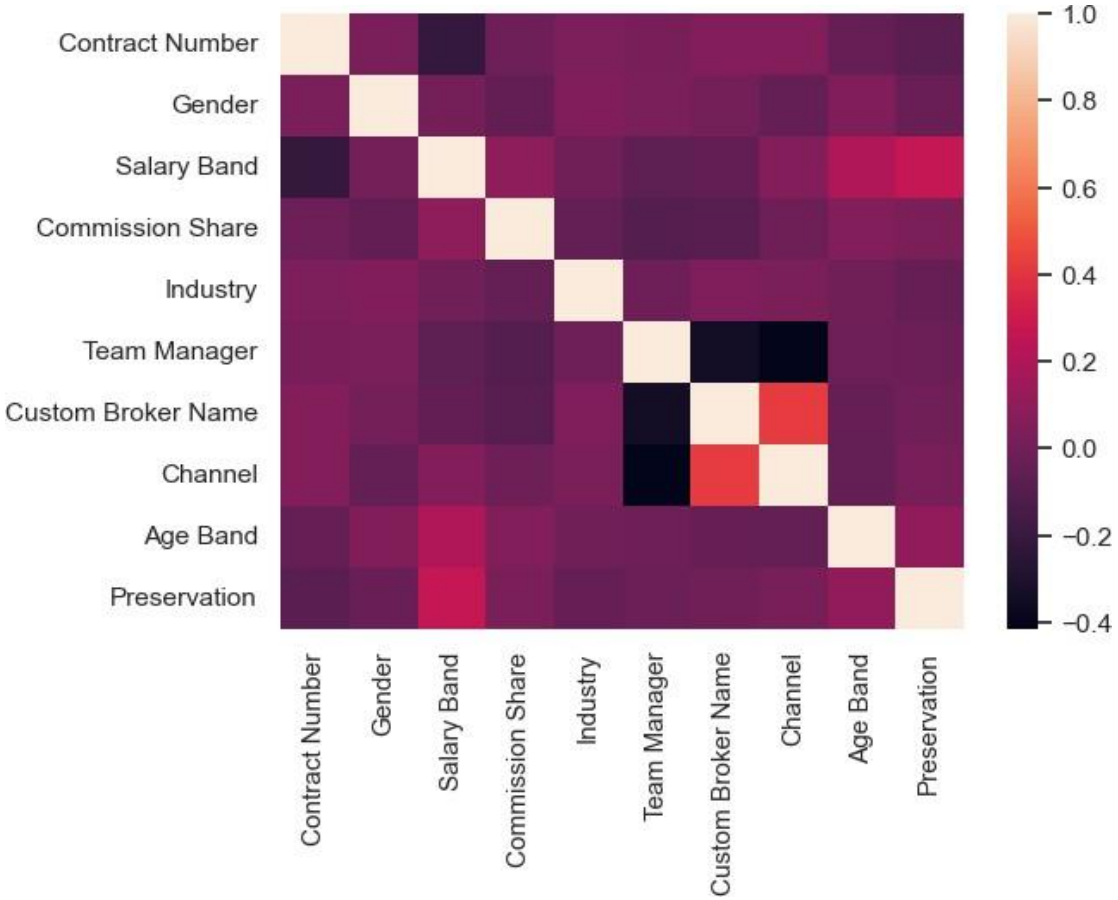


Figure 5.6: Correlation of Variables

From Figure 5.6 it seems that none of the variables are highly correlated (i.e. 0.6–1) to the dependent variable Preservation, except Other Preservation Withdrawal. Even though none of the other variables are highly correlated, none of the variables is negatively correlated with Preservation. The variables that are expected to have an impact on Preservation are Salary Band and Other Preservation Withdrawal. This is due to these variables having the highest correlation with Preservation.

The variables that have the highest correlation are:

- Custom Broker Name and Channel. This can be due to certain brokers only selling certain types of products.
- Age Band and Salary Band. This can be because younger individuals have not worked as long as older individuals and therefore do not earn as much.
- Salary Band and Preservation. This is expected because individuals who earn more are expected to be in a position to save more for their retirement.
- Other Preservation Withdrawal and Preservation. The high correlation between these two variables can be because individuals who did not preserve part of their payout know the consequences of not preserving. They can compare how the fund is performing with the money they preserved versus how the money they could have preserved is performing. It can also be that because they took part of the payout in cash they are content with the immediate gratification and therefore happy to preserve the rest of the money.

The following results were obtained for each model:

5.2 Logistic Regression

With ordinal encoding data, the following results were obtained from the logistic regression model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	859	257
	No Preservation	4 499	73 046

Table 5.5: Logistic Regression: Ordinal Encoding Confusion Matrix

The model correctly predicts 94% ($\frac{73\ 046}{73046+4\ 499}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts 77% ($\frac{859}{859+257}$) (Precision) of the individuals who preserve.

With dummy encoding data, the following results were obtained from the logistic regression model:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	940	306
	No Preservation	4 418	72 997

Table 5.6: Logistic Regression: Dummy Encoding Confusion Matrix

The model correctly predicts 94% ($\frac{72\ 997}{72\ 997+4\ 418}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts 75% ($\frac{940}{940+306}$) (Precision) of the individuals who preserve.

With target encoding data, the following results were obtained from the logistic regression model:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	947	297
	No Preservation	4 384	73 006

Table 5.7: Logistic Regression: Target Encoding Confusion Matrix

The model correctly predicts 94% ($\frac{73\,006}{73\,006+4\,384}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts 77% ($\frac{974}{974+297}$) (Precision) of the individuals who preserve.

The following metrics were obtained from the logistic regression model. How the model performed with the training data can be found in the appendix.

	Ordinal Encoding	**Dummy Encoding**	**Target Encoding**
Accuracy	94%	94%	94%
Precision	77%	75%	77%
Inverse Precision	94%	94%	94%
Recall	16%	18%	18%
AUC	82.12%	82.62%	82.06%

Table 5.8: Comparison of Logistic Regression Models with Different Encoding on Test Data

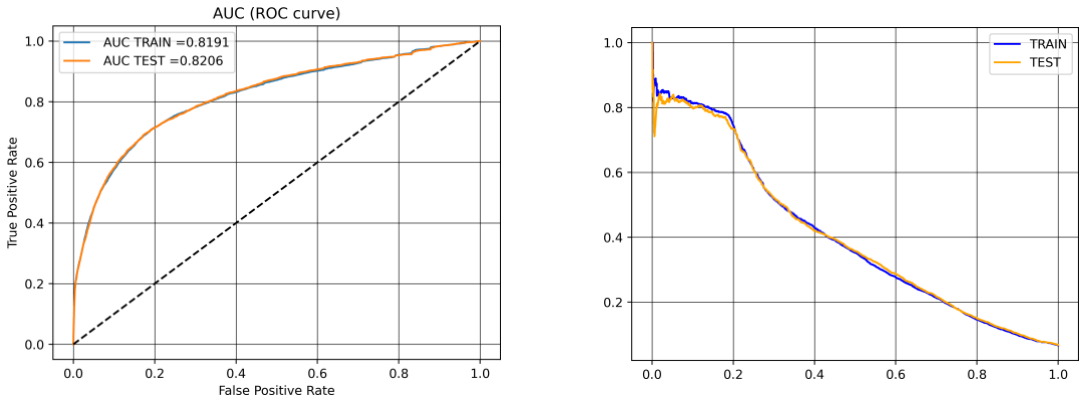
Based on the results presented in Table 5.8, the logistic regression model is accurate for all three encoding methods. All three methods produced an accuracy of 94%. However, accuracy is not a sufficient metric by itself.

The dummy encoding method has a higher AUC and lower precision than the other two methods. The ordinal encoding method has the lowest recall.

Based on this, the target encoding method performed the best of the three methods for the logistic regression model. However, the three methods performed very similarly.

High precision is preferred due to this data having more negatives (No Preservation) than positives (Preservation). The logistic regression model has an average precision of 76%. The model has higher precision than sensitivity and specificity.

Due to the fact that the target encoding method performed best, its AUC and PR-curve are shown below. Curves of the other methods can be found in the appendix.



a) ROC b) Precision-Recall Curve

Figure 5.7: Logistic Regression: Target Encoding

In Figure 5.7 the orange line represents the test data and the blue line the training data. Performance is based on the test data. Thus, we will only refer to the orange line when the performance is discussed.

The ideal for the ROC would be to be in the upper left corner of the graph and to have AUC larger than 50% to be an accurate classifier. The logistic regression model achieves this goal, with the classifier in the upper left corner and an AUC of 82%.

The PR curve shows that there is still room for improvement with this model. The ideal for the PR would be to be in the upper right corner. The PR curve is as expected. At points of low recall there is high precision and at points of high recall there is low precision. At a recall of 50%, precision of about 50% is still achieved. High precision is preferred, thus the performance of the model is good, based on the high precision.

5.3 Random Forest

With ordinal encoding data, the following results were obtained from the random forest model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	1 276	874
	No Preservation	4 082	72 429

Table 5.9: Random Forest: Ordinal Encoding Confusion Matrix

The model correctly predicts 95% ($\frac{72\,429}{72\,429+4\,082}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 59% ($\frac{1\,276}{1\,276+874}$) (Precision) of individuals who preserve.

With dummy encoding data, the following results were obtained from the random forest model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	1 280	836
	No Preservation	4 078	72 467

Table 5.10: Random Forest: Dummy Encoding Confusion Matrix

The model correctly predicts 95% ($\frac{72\,469}{72\,467+4\,078}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 60% ($\frac{1\,280}{1\,280+836}$) (Precision) of individuals who preserve.

With target encoding data, the following results were obtained from the random forest model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	1 275	865
	No Preservation	4 083	72 438

Table 5.11: Random Forest: Target Encoding Confusion Matrix

The model correctly predicts 95% ($\frac{72\,438}{72\,438+4\,083}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 60% ($\frac{1\,275}{1\,275+865}$) (Precision) of individuals who preserve.

	Ordinal Encoding	Dummy Encoding	Target Encoding
Accuracy	94%	94%	94%
Precision	59%	60%	60%
Inverse Precision	95%	95%	95%
Recall	24%	24%	24%
AUC	81.35%	81.44%	81.25%

Table 5.12: Comparison of Random Forest Models with Different Encoding on Test Data

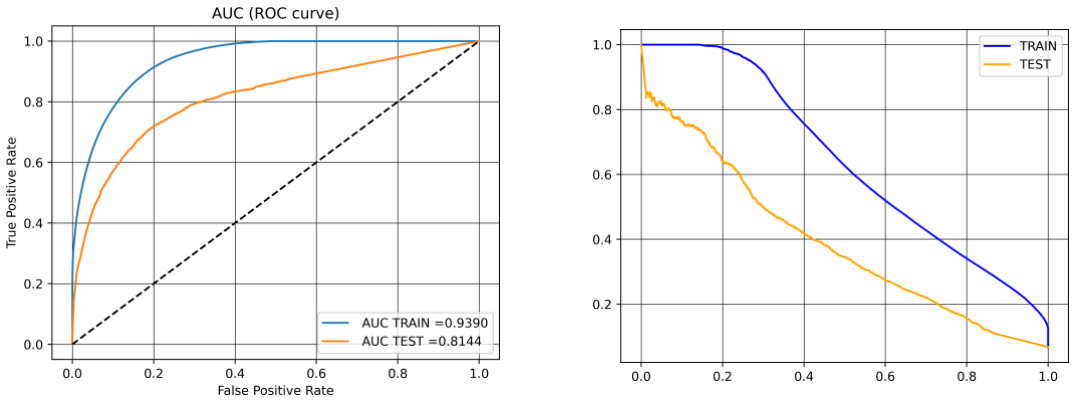
Based on the results presented in Table 5.12, the random forest model is accurate for all three encoding methods. All three methods produced an accuracy of 94%. However, accuracy is not a sufficient metric by itself.

The three models also have the same inverse precision and recall. The ordinal encoding method has the lowest precision. The dummy encoding method has the highest AUC.

As with logistic regression, the three encoding methods performed very similarly to one another. Based on small margins the dummy encoding method performed the best of the three.

The random forest model has an average precision of 60%. The model has higher precision than it has sensitivity and specificity. The random forest has about 16% lower precision than the logistic regression, but it has on average about 7% higher recall. High precision is preferred due to this data having more negatives (No Preservation) than positives (Preservation); thus based on precision, the logistic regression model performed better.

Due to the fact that the dummy encoding method performed the best, the AUC and PR-curve are shown below. The curves for the other methods can be found in the appendix.



a) ROC b) Precision-Recall Curve

Figure 5.8: Random Forest: Dummy Encoding

In Figure 5.8 the orange line represents the test data and the blue line the training data. Performance is based on the test data.

The ideal for the ROC would be to be in the upper left corner of the graph and to have AUC larger than 50% to be an accurate classifier. The random forest model achieves this goal, with the classifier in the upper left corner and an AUC of 81%. The PR curve shows that there is still room for improvement with this model. The ideal for the PR would be to be in the upper right corner. The PR curve is as expected. At points of low recall there is high precision and at points of high recall there is low precision. At a recall of 50%, precision of about 50% is still achieved. High precision is preferred, thus the performance of the model is good, based on the high precision.

5.4 Support Vector Machine

With ordinal encoding data, the following results were obtained from the SVM model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	262	65
	No Preservation	5 096	73 238

Table 5.13: Support Vector Machine: Ordinal Encoding Confusion Matrix

The model correctly predicts 94% ($\frac{73\ 238}{73\ 238+5\ 096}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 84% ($\frac{262}{262+65}$) (Precision) of individuals who preserve.

With dummy encoding data, the following results were obtained from the SVM model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	911	262
	No Preservation	4 447	73 041

Table 5.14: Support Vector Machine: Dummy Encoding Confusion Matrix

The model correctly predicts 94% ($\frac{73\ 041}{73\ 041+4\ 447}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 78% ($\frac{911}{911+262}$) (Precision) of individuals who preserve.

With target encoding data, the following results were obtained from the SVM model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	971	277
	No Preservation	4 387	73 026

Table 5.15: Support Vector Machine: Target Encoding Confusion Matrix

The model correctly predicts 94% ($\frac{73\,026}{73\,026+4\,387}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 78% ($\frac{971}{971+277}$) (Precision) of individuals who preserve.

	Ordinal Encoding	**Dummy Encoding**	**Target Encoding**
Accuracy	93%	94%	94%
Precision	80%	78%	78%
Inverse Precision	93%	94%	94%
Recall	5%	17%	18%
AUC	71.22%	72.60%	61.57%

Table 5.16: Comparison of SVM Models with Different Encodings on Test Data

Based on the results presented in Table 5.16, the SVM model is accurate for all three encoding methods. The Ordinal Encoding method has the lowest accuracy of 93%. However, accuracy is not a sufficient metric by itself.

The ordinal encoding method has the lowest inverse precision and low recall compared to the other two methods, but the ordinal encoding method has the highest precision. The ordinal encoding for the SVM has the highest precision of all the methods from all the models (Logistic

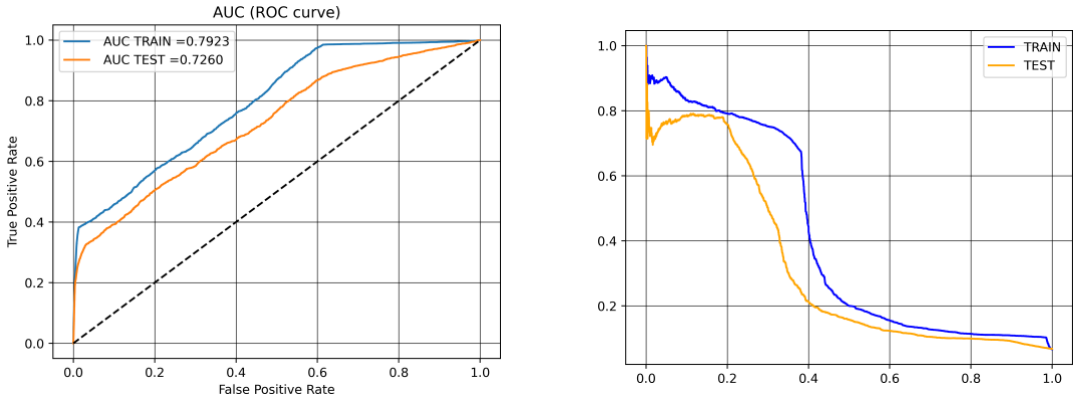
Regression, Random Forest and SVM). Even with the very high precision, the recall is so low that this method does not seem to have high performance.

The target encoding method has the highest recall and the lowest AUC.

As is the case with logistic regression and random forest, the three encoding methods perform very similarly to one another except for ordinal encoding's recall and target encoding's AUC.

In light of the ordinal encoding method's low recall and the target encoding method's low AUC, it was found the dummy encoding method performed the best.

Because the dummy encoding method was found to perform the best, the AUC and PR-curve are shown below. The curves for the other methods can be found in the appendix.



a) ROC b) Precision-Recall Curve

Figure 5.9: SVM: Dummy Encoding

In Figure 5.9 the orange line represents the test data and the blue line the training data. Performance is based on the test data.

Given that the AUC of the SVM is not as high as for the logistic regression and random forest models, it still achieves the goal of the ROC to be in the upper left corner of the graph and to have AUC larger than 50%, which is required to be an accurate classifier. The AUC is 73%. The PR-curve also shows worse precision than PR curves of other models. This PR curve reaches a 50%

precision at a recall lower than 50%, but it is still in the right vicinity. The PR curve shows that there is still room for improvement with this model. At points of low recall there is high precision and at points of high recall there is low precision. High precision is preferred, thus the performance of the model is good, based on the high precision.

5.5 Comparison of Models

	Logistic Regression	**Random Forest**	**SVM**
Accuracy	94%	94%	93%
Precision	77%	60%	80%
Inverse Precision	94%	95%	93%
Recall	16%	24%	5%
AUC	82.12%	81.35%	71.22%

Table 5.17: Comparison of Ordinal Encoding Models with Different Encoding on Test Data

For the ordinal encoding method, the SVM model has the lowest accuracy, inverse precision, recall and AUC. The random forest has the highest inverse precision, recall and AUC. Even though the random forest has the lowest precision, this model still has the overall best performance. The precision is still higher than 50%, which is required for it to be considered a good performance. However, based on precision only, the SVM outperforms the other models.

	Logistic Regression	Random Forest	SVM
Accuracy	94%	94%	94%
Precision	75%	60%	78%
Inverse Precision	94%	95%	94%
Recall	18%	24%	17%
AUC	82.62%	81.44%	72.60%

Table 5.18: Comparison of Dummy Encoding Models on Test Data

For the dummy encoding method, the SVM model has the highest precision and the lowest recall. The random forest has the highest inverse precision and recall. The logistic regression has the highest AUC. Even though the random forest has the lowest precision this model still has the overall best performance. The precision is still higher than 50%, which is required for it to be considered a good performance. However, based on precision only, the SVM outperforms the other models.

	Logistic Regression	Random Forest	SVM
Accuracy	94%	94%	94%
Precision	77%	60%	78%
Inverse Precision	94%	95%	94%
Recall	18%	24%	18%
AUC	82.06%	81.25%	61.57%

Table 5.19: Comparison of Target Encoding Models with Different Encoding on Test Data

For the target encoding method, the SVM model has the highest precision and the lowest AUC. For this method, the SVM had the same recall as the logistic regression. The random forest has the highest inverse precision and recall. The logistic regression has the highest AUC. Even though the random forest has the lowest precision, this model still has the overall best performance. The precision is still higher than 50%, which is required for it to be considered a good performance. However, based on precision only, the SVM outperforms the other models.

All three models can classify retirement fund preservation data with high accuracy and high precision overall. The SVM scores the highest for precision, which is a desired outcome based on the study of O'Reilly and Nielsen (2013). However, based on all the metrics, the random forest is the best model. The only downfall is that the lowest precision is produced by the random forest model, but the precision is not low enough for it to be a random guess by the model. The logistic regression also performed well for the different encoding methods, having metrics that were neither the highest nor the lowest of the models.

Additionally, based on the performance from the random forest and SVM models, the dummy encoding method seems to be a better encoding method for retirement fund preservation data for these machine learning models.

As mentioned, the data used to build the models was unbalanced. Balanced data was not created due to the high precision being obtained from the models with the unbalanced data. Balanced data would have led to low precision and high recall. This can be seen in the appendix.

5.6 Significant Preservation Variables

The logistic regression and random forest models were used to identify significant variables of retirement fund preservation.

From the above section, it can be seen that the models performed similarly for all three encoding methods. Thus, in this section, the study will only investigate the significant variables of the dummy encoding method. This method was chosen because the interpretation is easier to understand and it performed the best overall, as previously stated.

5.6.1 Logistic Regression

The following variables all had a p-value of 0. Usually, a p-value is chosen at 0.05. Most of the variables have a p-value less than this. Thus, the values that had a p-value of 0 were considered significant. The following tables show the significant variables (with a p-value of 0) as well as their Odds Ratio:

Variable	Odds Ratio
Gender (Male)	0.7763
Salary Band C	2.3441
Salary Band D	6.4384
Salary Band E	15.0971
Commission Share 81 to 90	9.5785
Industry - Education	1.5617
Industry - IT and Telecommunication	1.4895
Industry - Transportation	0.3525
Industry - Unions and Labour Brokers	0.3080
Other Preservation Withdrawal	55.7631

Table 5.20: Significant Variables from Logistic Regression

Table 5.20 shows that, given all other variables are kept constant, the odds of an individual having **Other Preservation Withdrawals** are 5476.31% ((55.7631-1) × 100) higher than an individual not having other preservation withdrawals. This is because every individual who made another preservation withdrawal preserves those funds.

Similarly, if an individual earns a salary of R500 001 and above (**Salary Band E**), given all the other variables are kept constant, this individual has the odds of higher preservation of 1509.71%. This is expected, seeing that individuals who earn more are able to save more.

Other variables in this table with higher odds of preservation are Salary Band C (R100 001–R300 000), Salary Band D (R300 001–R500 000), Commission Share 81 to 90, Education Industry, Transportation Industry, Unions and Labour Broker Industry and the IT and Telecommunication Industry.

The random forest model produces similar results. This can be seen in the following section.

5.6.2 Random Forest

It has been mentioned that random forest models have a built-in assessment that determines the important variables of the model.

The random forest concluded that the following variables were considered most significant: Other Preservation Withdrawal, Salary Band E (R500 001 and above), Male Gender, Provident Fund Contracts and Salary Band D (R300 001–R500 000). The 21 most significant variables of the random forest model are shown in Figure 5.10.

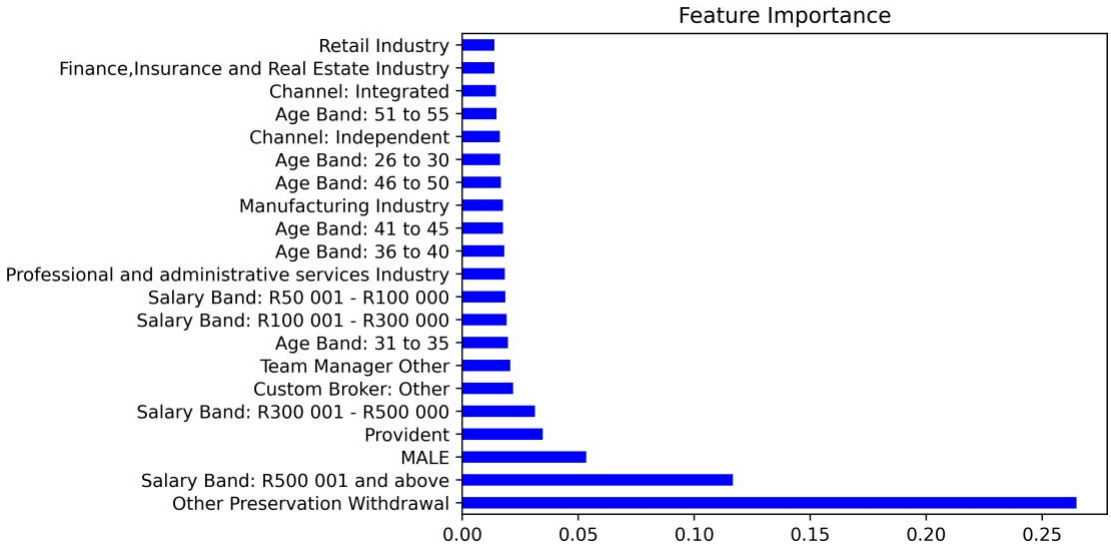


Figure 5.10: Significant Variables from Random Forest

From this, it can be seen that some of the Age Bands are considered significant as well. This was not as obvious with logistic regression.

All the variables that are significant for the logistic regression model are significant for the random forest as well, except for the industries. Different industries are considered significant in the random forest model.

The random forest model found that **Other Preservation Withdrawal** is the most useful when trying to predict whether an individual will preserve. Again, this is the case because every individual who had “other” preservation withdrawals preserved their funds. From the figure above, it can also be seen that individuals who earn a salary of R501 000 and above (**Salary Band E**) are more likely to be classified as individuals who preserve. Males are also more likely to be classified as individuals who will preserve.

5.7 Models with Significant Variables

Due to the logistic regression model and the random forest model indicating what variables are significant, a further analysis of these variables is done.

The following variables are considered in these models: Contract (Provident); Gender (Male); Salary Band R100 001–R300 000; Salary Band R300 001–R500 000; Salary Band R500 001 and above; Commission Share 71 to 80; Commission Share 81 to 90; Commission Share 91 to 100; Custom Broker Other; Industry – Retail; Industry – Finance, Insurance and Real Estate; Industry – Education; Industry – IT and Telecommunication; Industry – Transportation; Industry – Unions and Labour Broker; Industry – Construction and Maintenance; Industry – Manufacturing; Industry – Professional and Administrative Services; Team Manager Other; Independent Distribution Channel; Integrated Distribution Channel; Other Preservation Withdrawal, and the following Age Bands: 18–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60 and 61–65.

The Age Band decision is based on the random forest's variable importance and on the literature investigated. The literature states that age plays an important role in an individual's retirement decisions.

Additional commission share bands were also added due to a p-value very close to 0.

5.7.1 Logistic Regression

With the above-mentioned variables, a logistic regression model produces the following confusion matrix:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	929	308
	No Preservation	4 429	72 995

Table 5.21: Logistic Regression: Significant Variables Confusion Matrix

The model correctly predicts 94% ($\frac{72\ 995}{72\ 995+4\ 429}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 75% ($\frac{929}{929+308}$) (Precision) of individuals who preserve.

Table 5.22 shows the performance of the logistic regression model with the significant variables.

	Test Data
Accuracy	94%
Precision	75%
Inverse Precision	94%
Recall	17%
AUC	82.64%

Table 5.22: Evaluation of Logistic Regression Model with Significant Variables

Overall, this model performed well. Precision is high.

The following variables all had a p-value of 0. The following table shows the significant variables (with a p-value of 0) as well as their odds ratio:

Variable	Odds Ratio
Contract (Provident)	0.8579
Gender (Male)	0.7756
Salary Band C	2.1975
Salary Band D	6.0594
Salary Band E	14.2906
Commission Share 81 to 90	3.9512
Industry - IT and Telecommunication	1.7173
Industry - Transportation	0.4248
Industry - Education	1.7445
Industry - Unions and Labour Broker	0.3166
Industry - Finance, Insurance and Real Estate	1.5148
Industry - Manufacturing	1.2119
Age Band: 18 to 25	0.0214
Age Band: 26 to 30	0.0308
Age Band: 31 to 35	0.0279
Age Band: 36 to 40	0.0261
Age Band: 41 to 45	0.0269
Age Band: 46 to 50	0.0318
Age Band: 51 to 55	0.0420
Age Band: 56 to 60	0.0554
Age Band: 61 to 65	0.0679
Other Preservation Withdrawal	55.3280

Table 5.23: Significant Variables from Logistic Regression

Table 5.23 shows that, given all other variables are kept constant, the highest odds of an individual having higher preservation is better for an individual with **Other Preservation Withdrawals** or an individual who earns a salary of R500 001 and above (**Salary Band E**).

Other variables in this table with higher odds of preservation are Salary Band C (R100 001–R300 000), Salary Band D (R300 001–R500 000), Commission Share 81 to 90, Industry - Finance, Insurance and Real Estate, Industry - Education, Industry - Manufacturing and Industry - IT and Telecommunication.

The random forest model produces similar results. This can be seen in the following section.

5.7.2 Random Forest

With the above-mentioned variables random forest produces the following confusion matrix:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	1 145	685
	No Preservation	4 213	72 618

Table 5.24: Random Forest: Significant Variables Confusion Matrix

The model correctly predicts 95% ($\frac{72\ 618}{72\ 618+4\ 213}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 63% ($\frac{1\ 145}{1\ 145+685}$) (Precision) of individuals who preserve.

Table 5.25 shows the performance of the random forest model with the significant variables.

	Test Data
Accuracy	94%
Precision	63%
Inverse Precision	95%
Recall	21%
AUC	79.93%

Table 5.25: Evaluation of Random Forest Model with Significant Variables

The model has high precision, but not as high as the logistic regression. The random forest has higher recall than the logistic regression.

The random forest model again concluded that the following eight variables were most significant in this particular order: Other Preservation Withdrawal, Salary Band E (R500 001 and above), Custom Broker Other, Male Gender, Salary Band D R300 001–R500 000), Provident Fund Contracts, Team Manager Other and Independent Channel.

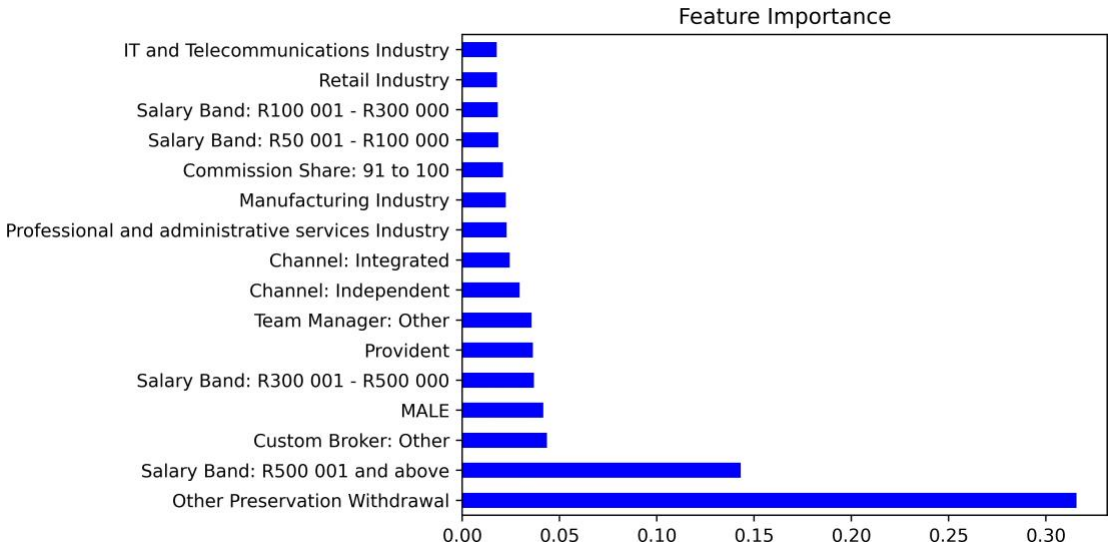


Figure 5.11: New Significant Variables from Random Forest

The random forest model found that Other Preservation Withdrawal is the most useful for trying to predict whether an individual will preserve.

6. DISCUSSION AND ANALYSIS OF FINDINGS

In this chapter, the results obtained will be discussed.

This study set out, as stated in the Research Objectives, to determine which factors play a significant role in driving low retirement fund preservation and to apply machine learning to the preservation data to determine whether machine learning can classify preservation of retirement fund data.

6.1 Factors that Drive Preservation Rates

Logistic regression and random forest models were used to determine the significant variables present in the data.

In the data used it is difficult to determine behavioural aspects that influence the preservation of funds. However, based on literature, good assumptions can be made of the variables proven to be significant.

6.1.1 Preservation of Variables

All the variables played a role in the classification of the data. Not all were equally significant; therefore, before discussing the most significant variables, a description of the preservation of other variables is given.

Distribution Channel

The random forest considered the Independent and Integrated Distribution Channels as two of the 21 most significant variables.

	Number of Members	
Channel	Did not Preserve	Preserved
Company Business	95%	5%
Direct	54%	46%
Uncategorised	95%	5%
Custom	92%	8%
Independent	95%	5%
Integrated	91%	9%

Table 6.1: Percentage of Members who Preserved within the Distribution Channel

Table 6.1 shows that the integrated channel has, out of all the members in that channel, the most individuals who preserve their retirement funds.

Team Manager

The random forest model took into account the team manager in the other group. Most individuals who preserve their retirement funds based on the team manager group are in the Team Manager Other group. This could be because most of the members belonged to this group. Thus, the probability of individuals in this group preserving their funds is higher.

Custom Broker

The random forest also considered the broker group Other as one of the significant variables. It was the third most significant in the reduced model. Again, most members belonged to this group of brokers. Even though this was a significant group of brokers, the most members who preserves are in the Broker Two group, which was not considered as significant by either the logistic regression or the random forest models.

Gender

It can be seen from the random forest and the logistic regression models that the gender variable is a strong indicator of whether the individual preserves.

	Number of Members	
Gender	Did not Preserve	Preserved
Male	92%	8%
Female	94%	6%

Table 6.2: Percentage of Members who Preserved based on Gender

Table 6.2 confirms that females have lower preservation numbers than males. Of all the females, only about 6% preserved their funds and of all the males about 8% preserved their retirement funds. This is in agreement with the models, which state that when a male makes a preservation withdrawal it can better predict whether the individual will preserve.

Contract Number

It can be seen from the random forest important variables, that the contract number variable is a strong indicator of whether the individual preserves. With the dummy encoding, it was shown that the provident fund contracts are strong indicators. The contract number is the fourth strongest indicator from the initial random forest model (Figure 5.10).

Commission Share

According to the logistic regression model, this variable is significant for predicting whether an individual will preserve or not. Commission Share 81 to 90 is the most significant group of the commission share variable.

This can mean that the members who pay higher commission rates to their brokers preserve more due to the brokers being better incentivised to assist them.

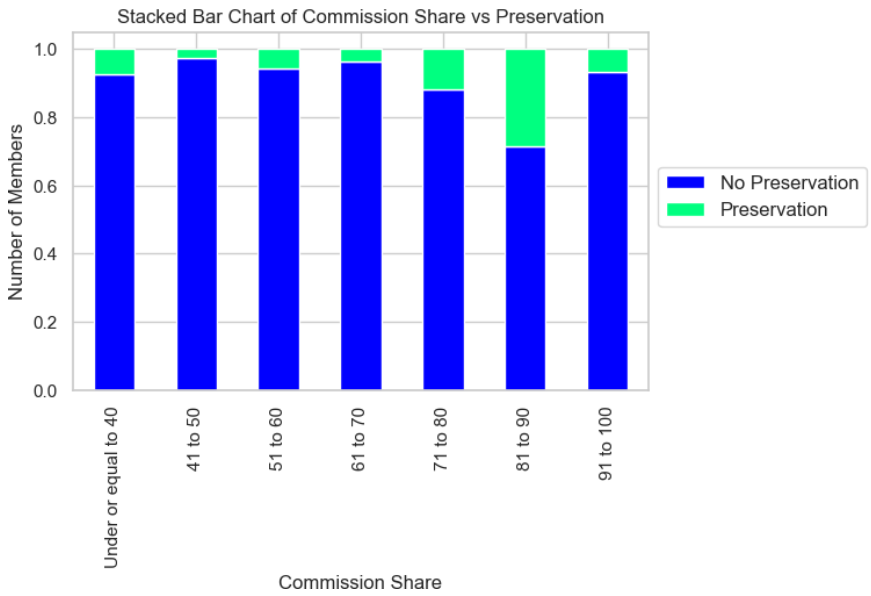


Figure 6.1: Preservation of Commission Share

From Figure 6.1 it can be seen that the members in the commission share band 81 to 90 preserve more, which agrees with the outcome of the logistic regression model that this specific commission band is the most significant commission share variable.

6.1.2 The Significant Variables

The logistic regression and random forest model considered almost all of the variables as important while classifying the preservation of retirement funds data. However, the most significant variables are Other Preservation Withdrawal, Industry, Salary Band, and Age Band.

Age Band

Based on past studies, it is expected that individuals in lower age brackets behave rationally and that they do not preserve (Table 2.2). However, it still cannot be known whether a young individual did not preserve due to consumption behaviour or just because they behaved irrationally.

In the logistic regression and random forest, age band was not one of the most significant variables. However, in the reduced logistic regression model it was found that age bands are significant. Even though this cannot be seen from the models, it is evident that some ages have better preservation than others.

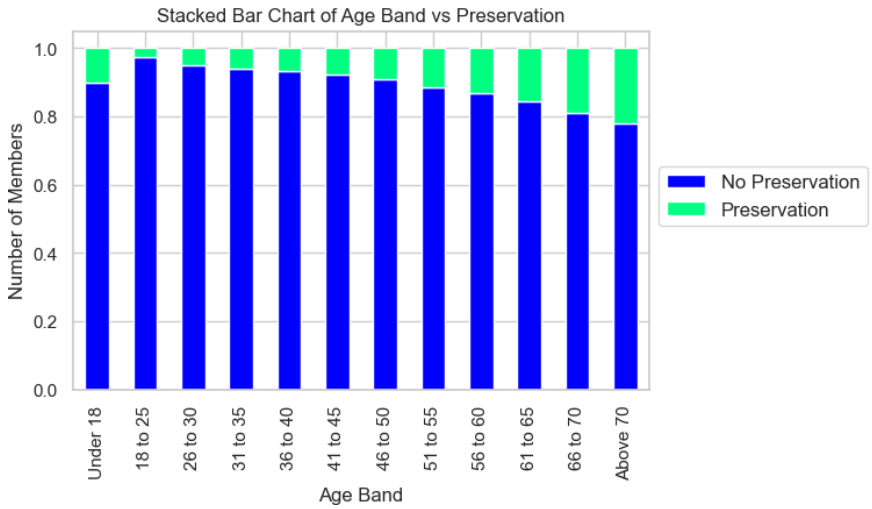


Figure 6.2: Preservation of Age Bands

Figure 6.2 confirms that higher ages preserve more.

Salary Band

The logistic regression models and random forest model consider salary as a significant driver of the preservation of retirement funds, especially a salary band of R500 001 and above. This is in agreement with the relevant literature. Literature states that individuals with lower incomes are less likely to preserve their retirement funds, usually due to rational behaviour (Table 2.2).

The second logistic regression model indicated the following salary bands were more significant:

- C - R100 001–R300 000
- D - R300 001–R500 000
- E - R500 001 and above

These bands are the brackets for the highest earning individuals.

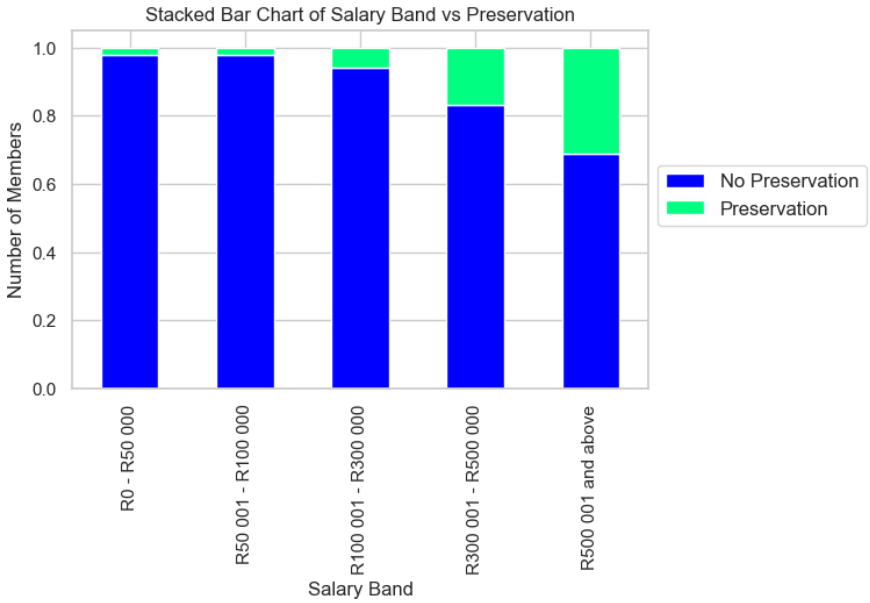


Figure 6.3: Preservation of Salary Bands

Figure 6.3 shows that these 3 salary bands preserve more than the other 2 bands.

Industry Code

The logistic regression model and random forest models consider the industry as a significant driver of the preservation of retirement funds.

In the logistic regression model, the following industries were considered more significant:

- IT and Telecommunication
- Education
- Transportation
- Unions and Labour Brokers

In the logistic regression model with the significant variables, the most significant industries are IT and Telecommunication, Education, Finance, Insurance and Real Estate, and the Manufacturing industry.

In the random forest model, the following industries are considered significant: Professional and Administrative Services, Manufacturing, Retail, Finance, Insurance and Real Estate.

In the random forest model with the significant variables, the most significant industries stayed the same while including the IT and telecommunication industry.

The two models did not show the same industries as significant. The models that were created with the significant variables agreed that the IT and Telecommunication, Education, Finance, Insurance and Real Estate, and the Manufacturing industries were significant.

For future studies regarding this variable, it may be interesting to compare salaries for these industries and or other variables such as age to determine why the members in these industries behave in a certain way.

Other Preservation Withdrawals

This variable was most significant every time for both the logistic regression and the random forest model.

	Number of Members	
Other Preservation Withdrawal	Did not Preserve	Preserved
No	243 716	14 390
Yes	860	3 236

Table 6.3: Number of Members who Preserved Based on Other Preservation Withdrawal

Table 6.3 shows that individuals who partially preserved their payout as preservation and did not preserve part of the payout had better preservation. Of the individuals who had Other preservation withdrawals, 79% had a status of preservation. Thus, more money was preserved than not preserved.

This is a significant variable because it indicates that maybe when an individual does take a small part of their payout in cash and preserve the rest, their preservation of retirement funds will be better. This can be because the small amount they take in cash can help them with any consumption needs they have while they still save most of their funds. This also indicates that South Africa's two-pot system may be a good encouragement for better preservation.

6.2 Machine Learning for Classifying Preservation

One of the research objectives of this study was to apply machine learning to the preservation data to determine whether machine learning can classify the data. The three different models used were Logistic Regression, Random Forest and Support Vector Machine. Three different encoding methods were also applied, with the expectation that one method would have better performance for the model. The models performed similarly in the confusion matrices and precision measures. Based on the performance of the three models, it seems that the dummy encoding method was effective for this data.

Overall the random forest model performed the best. The random forest is a good machine learning algorithm to employ due to its high accuracy with classification and because it can analyse the variables and derive important features. Even though this study did not find the recall rate as important as the precision rate, the random forest model had high precision and the highest recall rate (even if it was low). The only flaw would be that the random forest produced the lowest precision of the three models. The precision was still good, just not as high as that of the other models.

The logistic regression model is also a good model to use seeing that it can classify whether a member will preserve or not as well as generate a probability of whether a member will preserve or not. All three encoding methods performed similarly for the logistic regression model. Each of these methods has good performance and can predict whether an individual will preserve their retirement funds.

For the SVM model, the ordinal encoding method performed poorly on average. It had the lowest accuracy and recall, but it had the highest precision. Apart from finding that the SVM can

predict whether an individual will preserve, not much more can be derived from the SVM model. The model can be seen as a black box. This leaves a gap for further research to investigate what more this model can offer. Some studies are investigating the black box of machine learning, which could address issues such as these.

Due to the rapid growth of machine learning, its application to preservation data is a good step to take. These classification models can be used by insurance companies to determine whether their fund members will preserve or not. Machine learning also holds the advantage that the variables can be updated, i.e. new variables can be added to determine the value of the add to the preservation decision. Models that make predictions are a growing field. The retirement field should consider implementing more AI to improve results.

7. CONCLUSION AND RECOMMENDATIONS

Conclusion

Low retirement savings have been and still are an issue faced by many individuals worldwide.

In South Africa, and other countries, it has been identified that the lack of preserving one's retirement funds when an individual changes employer is one of the driving factors for low retirement savings.

This study had the following research objectives: to understand the field of the preservation of retirement funds, and to determine which factors significantly drive the preservation of retirement fund decisions. Seeing that machine learning is a growing field worldwide, this study also aims to understand machine learning and different machine learning methods and apply machine learning to the preservation data to determine whether machine learning can classify the data. The study specifically examines Logistic Regression, Random Forests and Support Vector Machines.

The study further found that many psychological factors play a role in an individual's retirement saving decisions. It is therefore important to also take these factors into account when trying to improve an individual's retirement savings. The following behavioural factors are found to be prevalent in retirement savings: self-control and determination, inconsistency of individuals' choices, optimism or pessimism, impulsiveness, information avoidance, procrastination, affect heuristics and decision fatigue.

In a similar way as behavioural factors affect retirement savings decisions, they also affect an individual's decision to preserve their retirement funds when they change employers. Potential reasons for poor preservation are usually either rational or irrational behaviour. Rational reasons are usually applicable to young individuals and individuals with liquidity constraints. These individuals tend to not preserve due to consumption smoothing. Irrational reasons are caused by an individual's bounded rationality or bounded willpower. Bounded rationality is usually due to

low levels of financial literacy. Bounded willpower is usually displayed in impulsive individuals or individuals who have low levels of future orientation.

Many countries attempt to influence poor retirement fund preservation with tax systems. A tax system is designed by making implicit assumptions that an individual will behave rationally. However, most countries assume individuals behave irrationally and introduce a tax system with rational behaviour assumptions. This can be due to not understanding what factors significantly impact retirement preservation.

Thus, in the preservation of the retirement fund field it is important to not neglect psychological factors when determining drivers of an individual's decisions. It should especially be considered when mitigation strategies are created.

Feature engineering was used with the expectation to improve machine learning models that are applied to preservation data. The following three methods of feature engineering were used: Ordinal Encoding, Dummy Encoding and Target Encoding.

The three models used were Logistic Regression, Random Forest and Support Vector Machine.

These three models were also very accurate in classifying the preservation of retirement funds, thus showing that machine learning models are good models to use for the classification of retirement data. Overall, all three models can accurately classify the preservation of retirement funds data. The random forest performs the best overall based on all the evaluation metrics used. It was also found that none of the encoding methods were superior to the others for all the models.

If mitigation of the non-preservation of retirement funds is investigated, machine learning models are good at classifying whether an individual will preserve or not. Machine learning models can also identify which factors drive certain preservation behaviours.

Additionally, the logistic regression and random forest models indicated that overall, the following variables are significant for the preservation models: whether an individual has other preservation withdrawals, the salary an individual earns and the industry they belong to. The Other preservation withdrawal indicated that preservation is better if an individual can take part of the

funds in cash and preserve the majority. Encouraging this behaviour thus may lead to effective mitigation. Industry is also a driver that can be investigated more. This can show what individuals are taught in specific industries to help them with their financial decisions.

Limitations

A limitation of the study is that the preservation of retirement funds is not the only factor that affects retirement savings. For example, in many cases it is the low contribution rates that affect poor retirement savings. Thus, the findings of this study should not be looked at as the main reason for poor retirement savings.

Individuals have unique financial circumstances and thus many varied reasons may affect an individual's decision to withdraw. There is no one-size-fits-all assumption. In the data used to build the models these different circumstances are not visible. For example, many young employed individuals are known to have a lot of debt. If these individuals withdrew retirement funds, it would not necessarily be the "wrong" decision to take the cash and pay off their debt.

A further limitation is that the model is built on a sample of South African retirement fund members. Thus, the results are not necessarily relevant to members from other countries.

Additionally, machine learning has the limitation of being a black box. This was particularly found with the SVM model. The model performed well, but more information is required. As mentioned, there are studies investigating this topic. It was not a part of the current study's objectives.

Lastly, the identified significant factors may not always stay the most significant.

Recommendations for further research

It is recommended that the findings of this study are applied to similar data to see whether the same results will be found.

As previously mentioned, a more specific study of how certain individuals behave in their respective industries can be undertaken. This can give better insight into what drives them to preserve their retirement funds.

Other machine learning models can be used to classify the data to see whether they are accurate at classifying the data.

Other variables from the data set can also be included to see if they have a significant impact on the preservation of retirement funds.

REFERENCES

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin, 82* (2), 463–496.

Ajibola, O. O. E., Ibiwoye, A., & Sogunro, A. (2012). Artificial neural network model for predicting insurance insolvency. *International Journal of Management and Business Research, 2*(1), 59–68.

Alemanni, B., & Lucarelli, C. (2017). Individual behaviour and long-range planning attitude. *The European Journal of Finance, 23* (5), 407–426.

Alexander Forbes. (2021). Alexander Forbes Member Insights 2021. Available from: https://connect.alexanderforbes.co.za/thought-article/146 (accessed: 03.08.2023).

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series, 1142*, 1–15.

Antolin, P. (2008). Coverage of funded pension plans. *OECD Working Papers on Insurance and Private Pensions, 19.*

Bailey, J. J., Nofsinger, J. R., & O'Neill, M. (2003). A review of major influences on employee retirement investment decisions. *Journal of Financial Services Research, 23* (2), 149–165.

Balasuriya, J., Gough, O., & Vasileva, K. (2014). Do optimists plan for retirement? A behavioural explanation for non-participation in pension schemes. *Economics Letters, 125* (3), 396–399.

Beckker, K. D., Witte, K. D., & Campenhout, G. V. (2019). Identifying financially illiterate groups: An international comparison. *International Journal of Consumer Studies, 43* (5), 490–501.

Benston, G. J., & Hartgraves, A. L. (2002). Enron: What happened and what we can learn from it. *Journal of Accounting and Public Policy, 21* (1), 105–127.

Bertram, R., & Zvan, B. (2009). Pension funds and incentive compensation: A story based on the Ontario teachers' experience. *Rotman International Journal of Pension Management, 2* (1), 30–35.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*, 197–227.

Blekesaune, M., & Skirbekk, V. (2012). Can personality predict retirement behaviour? A longitudinal analysis combining survey and register data from Norway. *European Journal of Ageing, 9*, 199–206.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2* (1), 1–8.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*, 235–255.

Bosworth, B. P., & Burtless, G. (2010). Recessions, wealth destruction, and the timing of retirement. *Boston College Retirement Research Center Working Paper, 2010–22*.

Botha, F. (2021). Closing the financial literacy gap. Available from: https://www.sanlam.co.za/blog/ articles/Pages/closing-the-financial-literacy-gap.aspx (accessed: 21.10.2023).

Boulesteix, A., Silke Janitza, J. K., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*, 493–507.

Bradter, U., Altringham, J. D., Kunin, W. E., Thom, T. J., O'Connell, J., & Benton, T. G. (2022). Variable ranking and selection with random forest for unbalanced data. *Environmental Data Science, 1*, e30.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Broadbent, J., Palumbo, M., & Woodman, E. (2006). The shift from defined benefit to defined contribution pension plans – implications for asset allocation and risk management. *Reserve Bank of Australia, Board of Governors of the Federal Reserve System and Bank of Canada, 1*, 54.

Butrica, B. A., Smith, K. E., & Iams, H. M. (2012). This is not your parents' retirement: Comparing retirement income across generations. *Social Security Bulletin, 72*, 37–58.

Canova, L., Rattazzi, A. M. M., & Webley, P. (2005). The hierarchical structure of saving motives. *Journal of Economic Psychology, 26*, 21–34.

Canu, S., & Smola, A. (2006). Kernel methods and the exponential family. *Neurocomputing, 69* (7-9), 714–720.

Carminati, L. (2020). Behavioural economics and human decision making: Instances from the health care system. *Health Policy, 124* (6), 659–664.

Chen, P., Lin, C., & Schölkopf, B. (2005). A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry, 21* (2), 111–136.

Coile, C., & Levine, P. B. (2011). The market crash and mass layoffs: How the current economic crisis may affect retirement. *The BE Journal of Economic Analysis & Policy, 11* (1).

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal, 20* (1), 131–148.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble machine learning: Methods and applications*, 157–175.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications, 41* (10), 4915–4928.

Dale, S. (2015). Heuristics and biases: The science of decision-making. *Business Information Review, 32* (2), 93–99.

Dalianis, H. (2018). Evaluation metrics and evaluation. *Clinical Text Mining*, 45–53.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on machine learning*, 233–240.

De Bondt, W. F. M., & Thaler, R. H. (1987). Further evidence on investor overreaction and stock market seasonality. *The Journal of Finance, 42* (3), 557–581.

DellaVinga, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature, 47* (2), 315–372.

Depre, P., Shevchenko, P. V., & Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal, 7,* 337–352.

Deris, A. M., Zain, A. M., & Sallehuddin, R. (2011). Overview of support vector machine in modeling machining performances. *Procedia Engineering, 24,* 308– 312.

Dhlembeu, N. T., Kekana, M. K., & Mvita, M. F. (2022). The influence of financial literacy on retirement planning in South Africa. *Southern African Business Review, 26,* 1–25.

Du, P., Samat, A., Waske, B., Liu, S., & Li, Z. (2015). Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS Journal of Photogrammetry and Remote Sensing, 105*, 38–53.

Duarte, F. (2021). Encouraging mammograms using behavioral economics: A randomized controlled trial in Chile. *Value Health, 24* (10), 1463–1469.

Duflo, E., & Saez, E. (2002). Participation and investment decisions in a retirement plan: The influence of colleagues' choices. *Journal of Public Economics, 85* (1), 121–148.

Eisner, R. (1958). The permanent income hypothesis: Comment. *American Economic Review,* 972–990.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics, 49* (3), 283–306.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters, 27* (8), 861–874.

Feldman, D. C., & Beehr, T. A. (2011). A three-phase model of retirement decision making. *American Psychologist, 66* (3), 193–203.

Fisch, J. E., Wilkinson-Ryan, T., & Firth, K. (2016). The knowledge gap in workplace retirement investing and the role of professional advisors. *Duke Law Journal, 66* (3), 633–672.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40* (2), 351–401.

Friedman, M. (2016). A theory of the consumption function. Golden Springs Publishing. eBook.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering, 2* (4), 42–47.

Ghilarducci, T., Radpour, S., & Webb, A. (2019). New evidence on the effect of economic shocks on retirement plan withdrawals. *SCEPA working paper series, 2018* (03).

Godoi, C. K., Marcon, R., & Da Silva, A. B. (2005). Loss aversion: A qualitative study in behavioural finance. *Managerial Finance, 31* (4), 46–56.

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature, 55* (1), 96–135.

Harrison, T., Waite, K., & Hunter, G. L. (2006). The internet, information and empowerment. *European Journal of Marketing, 40* (9/10), 972–993.

Hilbe, J. M. (2009). *Logistic regression models.* CRC press.

Hu, M., Pan, S., Li, Y., & Yang, X. (2023). Advancing medical imaging with language models: A journey from n-grams to ChatGPT. *ArXiv, abs/2304.04920.*

Hurd, M., & Panis, C. (2006). The choice to cash out pension rights at job change or retirement. *Journal of Public Economics, 90* (12), 2213–2227.

INFE, O. (2016). International survey of adult financial literacy competencies. Available from: https://www.oecd.org/financial/education/oecd-infe-survey-adult-financial-literacy-competencies.htm (accessed 03.06.2023)

Iyengar, S. S., Jiang, W., & Huberman, G. (2004). How much choice is too much? Contributions to 401(k) retirement plans. *Pension Design and Structure: New Lessons from Behavioral Finance, 83,* 84–87.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79* (6), 995–1006.

Jakkula, V. (2006). Tutorial on support vector machine (SVM). *School of EECS, Washington State University, 37* (2.5), 3.

Jappelli, T., & Modigliani, F. (1998). The age-saving profile and the life-cycle hypothesis. *Long-run Growth and Short-run Stabilization: Essays in Memory of Albert Ando,* 6.

Kapoor, S., & Prosad, J. M. (2017). Behavioural finance: A review. *Procedia Computer Science, 122,* 50–54.

Karlan, D., McConnell, M., Mullainathan, S., & Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. *Management Science, 62* (12), 3393–3411.

Kelly, N. R., & Swisher, L. (1998). The transitional process of retirement for nurses. *Journal of Professional Nursing, 14* (1), 53–61.

Keynes, J. M. (1936). *The general theory of employment, interest, and money.* London: Macmillan; Co.

Kumar, S., Shukla, G. P., & Sharma, R. (2019). Analysis of key barriers in retirement planning: An approach based on interpretive structural modeling. *Journal of Modelling in Management, 14* (4), 972–986.

LaValley, M. P. (2008). Logistic regression. *The Journal of Educational Research, 117,* 2395–2399.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology, 66,* 799–823.

Lin, Y., Huang, C., & Lin, C. (2012). Determination of insurance policy using neural networks and simplified models with factor analysis technique. *WSEAS Transaction on Information Science and Applications, 5* (10), 1405–1415.

Liu, Y., & Xie, T. (2019). Machine learning versus econometrics: Prediction of box office. *Applied Economics Letters, 26,* 124–130.

Loewe, G. (2006). The development of a theory of rational intertemporal choice. *Papers, 80,* 195–221.

Lourenço, C. J., Dellaert, B. G., & Donkers, B. (2020). Whose algorithm says so: The relationships between type of firm, perceptions of trust and expertise, and the acceptance of financial robo-advice. *Journal of Interactive Marketing, 49* (1), 107–124.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *American Economic Journal: Journal of Economic Literature, 52* (1), 5–44.

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies, 3*, 281–299.

Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly Journal of Economics, 89* (4), 1149–1187.

McKerchar, T. L., & Renda, C. R. (2012). Delay and probability discounting in humans: An overview. *The Psychological Record, 62*, 817–834.

Meyll, T., Pauls, T., & Walter, A. (2020). Why do households leave money on the table? The case of subsidized pension products. *Journal of Behavioural Finance, 21* (3), 266–283.

Modigliani, F., & Brumberg, R. (1954). Utility analysis and the consumption function: An interpretation of cross-section data. *Franco Modigliani, 1* (1), 388–436.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., Caspi, A., & Heckman, J. J. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, 108 (7), 2693–2698.

Momentum Corporate. (2021). Legal update: Demystifying the impact of the new annuitisation legislation on members of the FundsAtWork Umbrella Provident fund. https://eb.momentum.co.za/webDocumentLibrary/LegalUpdates/ 2021 / Legal Update 1 of 2021 Impact of new annuitisation legislation.pdf (accessed: 10.10.2023).

Mu, J., Xu, L., Duan, X., & Pu, H. (2018). Study on customer loyalty prediction based on RF algorithm. *Journal of Computers, 8* (8), 2134–2138.

Munnell, A. H., Sunden, A., & Taylor, C. (2001). What determines 401 (k) participation and contributions. *Social Security Bulletin, 64*, 64–75.

National Treasury. (2012a). Preservation, portability and governance for retirement funds. www.treasury.gov.za (accessed: 08.06.2022).

National Treasury. (2012b). Strengthening retirement savings. www.treasury.gov.za (accessed: 08.06.2022).

National Treasury. (2021a). Encouraging South African households to save more for retirement. www.treasury.gov.za (accessed: 08.06.2022).

National Treasury. (2021b). Release of two retirement reform discussion papers for public comment. www.treasury.gov.za (accessed: 08.06.2022).

National Treasury. (2023). Publication of the draft legislation for the “two pot” retirement system for public comment. www.treasury.gov.za (accessed: 11.07.2023).

Nazyrova, N., Chaussalet, T. J., & Chahed, S. (2022). Machine learning models for predicting 30-day readmission of elderly patients using custom target encoding approach. *International Conference on Computational Science*, 122–136.

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the gini importance? *Bioinformatics, 34* (21), 3711–3718.

Nguyen, P., Le, L. K., Ananthapavan, J., Gao, L., Dunstan, D. W., & Moodie, M. (2022). Economics of sedentary behaviour: A systematic review of cost of illness, cost-effectiveness, and return on investment studies. *Preventive Medicine, 156*, 106964.

Nguyen, T. A. N., Polách, J., & Vozňáková, I. (2019). The role of financial literacy in retirement investment choice. *Equilibrium. Quarterly Journal of Economics and Economic Policy, 14* (4), 569–589.

Nivedha, R., & Sairam, N. (2015). A machine learning based classification for social media messages. *Indian Journal of Science and Technology, 8*, 1–4.

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology, 122*, 56–69.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *The American Economic Review, 89* (1), 103–124.

Old Mutual. (2021). OMC annuitisation note Q&A. www.oldmutual.co.za (accessed: 21.10.2023).

Old Mutual. (2022). Old mutual savings & investment monitor 2022. www.oldmutual. co.za (accessed: 10.07.2023).

O'Reilly, C., & Nielsen, T. (2013). Revisiting the ROC curve for diagnostic applications with an unbalanced class distribution. *2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA)*, 413–20.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8*, 154–168.

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*, 3–14.

Pillay, N., & Fedderke, J. (2022). *Characteristics of the South African retirement fund industry* (tech. rep.).

Pop-Eleches, C., Thirumurthy, H., Habyarimana, J. P., Zivin, J. G., Goldsteing, M. P., De Walqueg, D., MacKeen, L., Haberer, J., Kimaiyo, S., Sidle, J., Ngare, D., & Bangsberg, D. R. (2011). Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: A randomized controlled trial of text message reminders. *AIDS, 25* (6), 825–834.

Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications, 175* (4), 7–9.

Powers, D. M. (2020). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Pratt, D. A. (2010). Focus on ... retirement plan leakage. *Journal of Pension Benefits, 17* (3), 3.

Püschel, K., Coronado, G., Soto, G., Gonzalez, K., Martinez, J., Holte, S., & Thompson, B. (2010). Strategies for increasing mammography screening in primary care in Chile: Results of a randomized clinical trial. *Cancer Epidemiology, Biomarkers & Prevention, 19* (9), 2254–2261.

Reyers, M., Schalkwyk, C. H. V., & Gouws, D. G. (2014). The rationality of retirement preservation decisions: A conceptual model. *Journal of Economics and Behavioral Studies, 6* (5), 418–431.

Reyers, M., van Schalkwyk, C. H., & Gouws, D. G. (2015). Rational and behavioural predictors of pre-retirement cash-outs. *Journal of Economic Psychology, 47*, 23–33.

Robayo-Pinzon, O., Foxall, G. R., Montoya-Restrepo, L. A., & Rojas-Berrio, S. (2021). Does excessive use of smartphones and apps make us more impulsive? An approach from behavioural economics. *Heliyon, 7 (*2), e06104.

Roberts, B., Struwig, J., Gordon, S., & Radebe, T. (2014). Financial literacy in South Africa: Results from the 2013 South African Social Attitudes Survey (SASAS) round. https://www.fscamymoney.co.za/Research%20Documents/Financial%20Literacy%20in%20South%20Africa%20Results%20from%20the%202013.pdf (accessed: 03.03.2023).

Rodríguez, P., Bautista, M. A., Gonzalez, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing, 75*, 21–31.

Salcedo-Sanz, S., Rojo-A´lvarez, J. L., Martínez-Ramón, M., & Camps-Valls, G. (2014). Support vector machines in engineering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4* (3), 234–267.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty, 1* (1), 7–59.

Sanlam. (2023a). Finances through the life stages: How is SA doing? A Sanlam Report. www.sanlam.co.za (accessed: 03.08.2023).

Sanlam. (2023b). Sanlam Benchmark 2023. www.sanlam.co.za (accessed: 03.08.2023).

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research, 177* (3), 1333–1352.

Speelman, D. (2014). Logistic regression. *Corpus methods for semantics: Quantitative studies in polysemy and synonymy, 43*, 487–533.

Stolper, O. (2018). It takes two to tango: Households' response to financial advice and the role of financial literacy. *Journal of Banking and Finance, 92*, 295– 310.

Taecharungroj, V. (2023). "What Can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on twitter. *Big Data and Cognitive Computing, 7* (1), 35.

Taigman, Y., Yang, M., Ranzanto, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA*, 1701–1708.

Talwar, A., & Kumar, Y. (2013). Machine learning: An artificial intelligence methodology. *International Journal of Engineering and Computer Science, 2* (12), 3400– 3404.

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy, 112* (S1), S164– S187.

Torgo, L., & Ribeiro, R. (2009). Precision and recall for regression. *Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12*, 332–346.

Van Solinge, H., & Henkens, K. (2009). Living longer, working longer? The impact of subjective life expectancy on retirement intentions and behaviour. *European Journal of Public Health, 20* (1), 47–51.

Vapnik, V., Golowich, S. E., & Smola, A. (1996). Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems, 9.*

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of classification methods on unbalanced data sets. *IEEE Access, 9,* 64606–64628.

Wang, M., & Shultz, K. S. (2010). Employee retirement: A review and recommendations for future investigation. *Journal of Management, 36* (1), 172–206.

Worster, A., Fan, J., & Ismaila, A. (2007). Understanding linear and logistic regression analyses. *Canadian Journal of Emergency Medicine, 9*, 111–113.

Xue, J.-H., & Hall, P. (2014). Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? *IEEE transactions on pattern analysis and machine intelligence, 37* (5), 1109–1112.

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: Principles and techniques for data scientists. O'Reilly Media, Inc.

Zheng, E., Tan, Y., Goes, P., Chellappa, R., Wu, D., Shaw, M., Sheng, O., & Gupta, A. (2017). When econometrics meets machine learning. *Data and Information Management, 1*, 75–83.

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics, 10* (5), 593.

Zhu, M. (2004). Recall, precision and average precision. Department of Statistics and Actuarial Science, *University of Waterloo, Waterloo, 2* (30), 6.

APPENDIX

A.1 Data

A.1.1 Variables

In the raw dataset, many variables took on different forms, but stayed essentially the same variable, for example, a column where the precise pensionable salary was given and a column where this salary's salary band was the variable used. There weren't many variables that were not used in the dataset. Additionally, the company advised that some variables should not be used. The following variables present in the raw dataset were not used:

Variable	Representation	Reason for not using
Paypoint Number	A number is allocated to each business unit in the employer that submits financial data to the company.	Does not affect the members' preservation.
Transaction Number	Unique row identifier. Distinct payment identifier.	The number is given after the preservation transaction and thus will not influence the preservation decision.
Trade Type	States whether the payment was income or a payment.	All the entries were payments.
Risk Benefits Accepted	Corresponding risk-based claim payable to member.	This is not a retirement policy.
Portal for claims	Employer uses the portal to load info each month.	Assume that the employer who uses the portal will not affect the member's preservation decision.
App use (3 similar columns)	Indicates whether employee used the app for payout	Assumed that the employer who uses the app will not affect the member's preservation decision, seeing that the app is used after that decision has been made.
Payee Name Mapping	Details of when preservation withdrawal happened and to which fund the payment went.	Used a summarised version.
Consultant	The internal consultant associated with the employer group contract if the broker is from the company.	Used another column similar to this.
Bank Account String	The last 4 digits of the bank account number.	Assumed this does not play a role in preservation decisions
Account Name	Lookup for the last 4 digits of the bank account number	Assumed this does not play a role in preservation decisions

Table A.1: Variables of Original Dataset not Used

A.1.2 Data Cleaning

In the dataset, there are 262 202 entries. This equals 262 202 unique members. Unique members were determined based on their assigned person number, group number (which corresponds to the employer they belong to), role number and contract number. Only members who changed employers or were retrenched were considered.

The following steps were taken to determine unique members in the data:

Firstly, unique members were determined with a unique combination based on the member's person number, the number of the employer group, the number assigned to them in that specific work environment, and their contract number. This was done at the suggestion of the data providers. If these members had multiple preservation withdrawals, similar ones were then grouped and the amount withdrawn was summed.

If there were still multiple members, the Other entry was considered as another preservation withdrawal. This happened when an individual took part of their payout and preserved it and did not preserve the other part. The two entries were combined, creating two columns: *first preservation withdrawal amount* and *second preservation withdrawal amount*. If an individual had these two withdrawals, they were marked as 'Yes' in the created column *Other preservation withdrawal*. Their preservation status was then determined by the larger amount in the columns' first preservation withdrawal amount and second preservation withdrawal amount.

The data contained entries that indicated whether members who withdrew their funds preserved or not. Based on this, the column Preservation was created to simplify what was already given in the data. Then the Overall Preservation (which is the independent variable) was determined based on the amounts withdrawn that were not preserved or preserved.

A.1.3 Ordinal Encoding

After ordinal encoding was applied to the data, the variables looked as follows:

Variable	Representation
Contract number	This variable takes the value 1 for a provident fund and 0 for a pension fund.
Gender	This variable takes the value 1 if the member is male and 0 if the member is female.
Age band	This variable takes the value 1 if the band is B, 2 if the band is C, 3 if the band is D, 4 if the band is E, 5 if the band is F, 6 if the band is G, 7 if the band is H, 8 if the band is I, 9 if the band is J, 10 if the band is K, 11 if the band is L and 0 if the band is A.
Salary band	This variable takes the value 1 if the band is B, 2 if the band is C, 3 if the band is D, 4 if the band is E and 0 if the band is A.
Industry	The following values were assigned for the different industries: Agriculture, Forestry, Fishing and Co-ops - 0, Business Services - 1, Construction and Maintenance - 2, Education - 3, Entertainment and Hospitality - 4, Finance Insurance and Real Estate - 5, Health and Welfare - 9, IT and Telecommunication - 10, Manufacturing - 8, Mining and Raw Materials - 9, Other - 10, Professional and administrative services - 11, Retail - 12, Security services - 13, Transportation - 14, Unions and Labour Brokers – 15.
Custom broker name	This variable takes the value 1 if the broker is Broker Two, 2 if the broker is No Broker, 3 if the broker is Other and 0 if the broker is Broker One.
Commission share	This variable takes the value 1 if the band is 51 to 60, 2 if the band is 61 to 70, 3 if the band is 71 to 80, 4 if the band is 81 to 90, 5 if the band is 91 to 100, 6 if the band is under or equal to 40 and 0 if the band is 41 to 50.
Team manager	This variable takes the value 1 if the team manager is Other, 2 if the team manager is Team Manager One, 3 if the team manager is Team Manager Two and 0 if the team manager is No Team Manager.
Channel	The following values were assigned for the different distribution channels: Channel not categorised - 0, Company's Business - 1, Custom - 2, Direct Cluster - 3, Independent - 4, Integrated - 5
Other Preservation Withdrawal	This variable takes the value 1 for another preservation withdrawal and 0 for no other preservation withdrawal

Table A.2: Predictor Variables Used in the Logistic Regression Model One

A.1.4 Dummy Encoding

In Python, the pandas library is used to create the dummy variables. The function used is the *get_dummies()*. The difference with this function is each variable is transformed into as many variables as there are different entries. Columns in the output are named after each entry. For example, for the variable Age Band, the following columns were created: AgeBand B, AgeBand C, AgeBand D, AgeBand E, AgeBand F, AgeBand G, AgeBand H, AgeBand I, AgeBand J, AgeBand K and AgeBand L.

A.1.5 Target Encoding

In Python, the *category_encoders* package and *ce* function were used.

For this method, the average for each variable was given as an example, determined as follows: First the gender variable was considered. Then for the target variable, which is Preservation, the average of the males who preserved was determined and similarly the average for females. Thus, the method determined the average males, given the male preserved, and did the same for females. For example, in the data, there are 163 757 males and 98 445 females of the 262 202 members. There are 10 047 males who preserved their funds and 7 579 females who preserved their funds. Thus, 6.135% ($\frac{10\ 047}{163\ 757}$) males preserved their funds and 7.699% ($\frac{7\ 579}{98\ 445}$) females preserved their funds. Thus, target encoding marked every male with 0.06135 and every female with 0.0799.

A.2 Logistic Regression: More Results

The equations that ordinal and target encoding methods produced for the logistic regression model are shown below. The dummy encoding produces a very large equation.

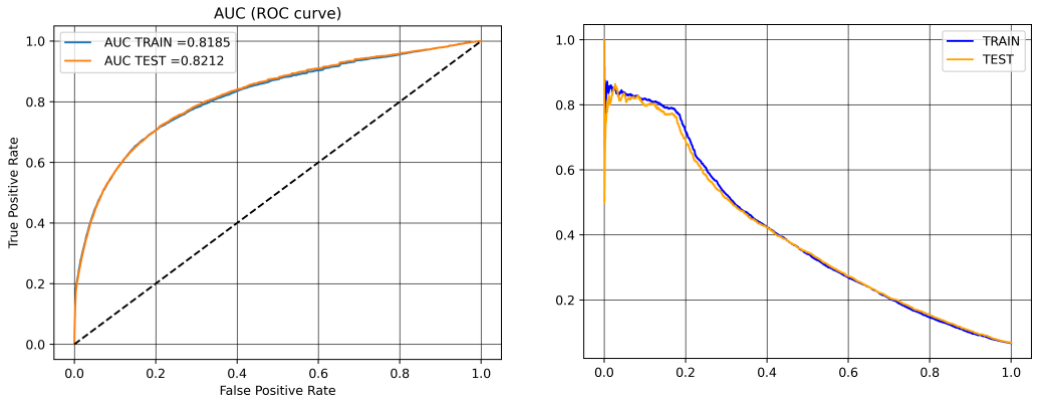
Ordinal Encoding

The logistic regression equation that resulted from ordinal encoding is as follows:

$$ln\left(\frac{p}{1-p}\right) = -4.7242 - 0.1188X_1 - 0.2762X_2 + 0.7976X_3 + 0.0933X_4 - 0.0362X_5 - 0.04892X_6 + 0.0112X_7 + 0.0214X_8 + 0.08869X_9 + 4.0285X_{10} \quad \text{(A.1)}$$

where the variables are marked as follows: Contract Number X_1, Gender X_2, Salary Band X_3, Commission Share X_4, Industry X_5, Team Manager X_6, Custom Broker Name X_7, Channel X_8, Age Band X_9 and Other Preservation Withdrawal X_{10}.

The AUC and PR-curve of the ordinal encoding method are shown below.



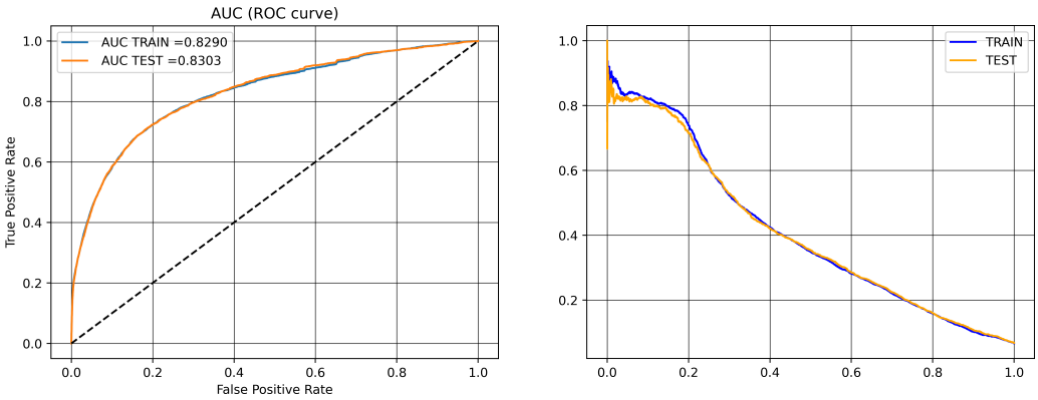
a) ROC b) Precision-Recall Curve

Figure A.1: Logistic Regression: Ordinal Encoding

In Figure A.1 the orange line represents the test data and the blue line the training data.

Dummy Encoding

The AUC and PR-curve of the dummy encoding method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.2: Logistic Regression: Dummy Encoding

In Figure A.2 the orange line represents the test data and the blue line the training data.

Target Encoding

The logistic regression equation that resulted from target encoding is as follows:

$$ln\left(\frac{p}{1-p}\right) = -7.4361 + 3.1097X_1 + 18.2874X_2 + 8.2356X_3 + 8.1521X_4 + 6.1404X_5 + 1.8525X_6 + 2.4568X_7 + 2.1609X_8 + 7.9571X_9 + 5.4596X_{10} \quad \text{(A.2)}$$

where the variables are marked as follows: Contract Number X_1, Gender X_2, Salary Band X_3, Commission Share X_4, Industry X_5, Team Manager X_6, Custom Broker Name X_7, Channel X_8, Age Band X_9 and Other Preservation Withdrawal X_{10}.

Model Performance with Training Data

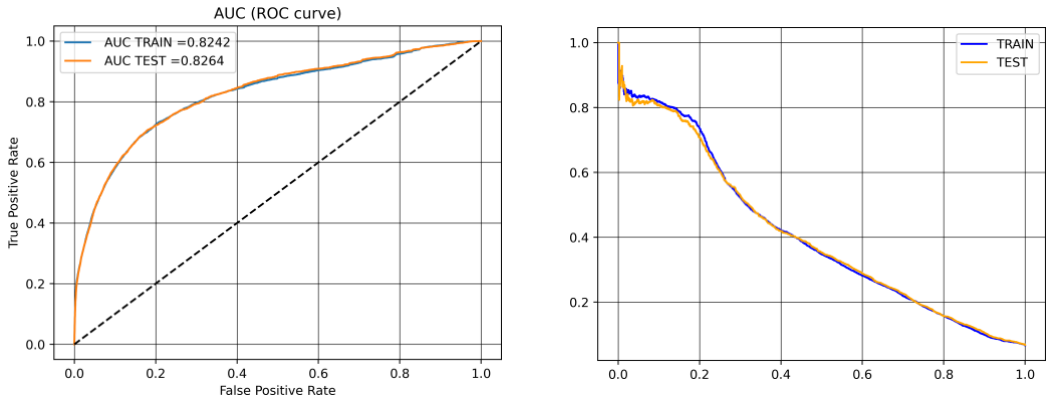
Table A.3 shows the performance of the logistic regression model with the three different encoding methods on the training data.

	Ordinal Encoding	Dummy Encoding	Target Encoding
Accuracy	94.00%	94.00%	94.00%
Precision	79.00%	78.00%	78.00%
Inverse Precision	94.00%	94.00%	94.00%
Recall	16.00%	17.00%	18.00%
AUC	81.85%	82.90%	81.91%

Table A.3: Comparison of Logistic Regression Models with Different Encoding on Training Data

Significant Variable Logistic Regression Model

The AUC and PR-curve of the logistic regression model with significant variables are shown below.



a) ROC b) Precision-Recall Curve

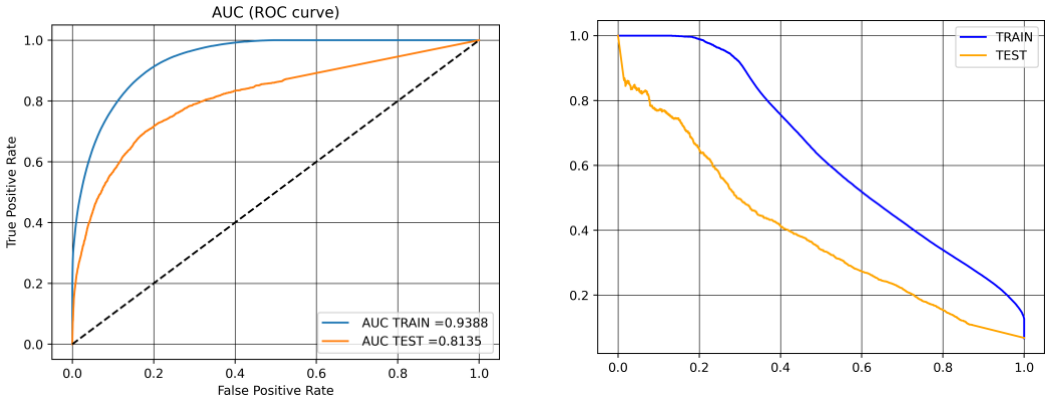
Figure A.3: Logistic Regression with Significant Variables

In Figure A.3 the orange line represents the test data and the blue line the training data.

A.3 Random Forest: More Results

Ordinal Encoding

The AUC and PR-curve of the ordinal encoding method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.4 Random Forest: Ordinal Encoding

In Figure A.4 the orange line represents the test data and the blue line the training data. The ordinal encoding method ranked the most significant variables as follows:

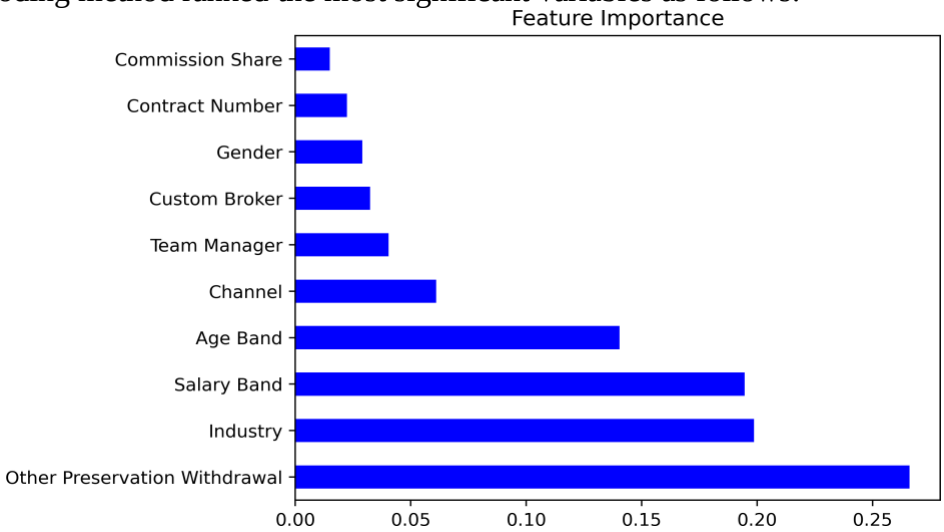
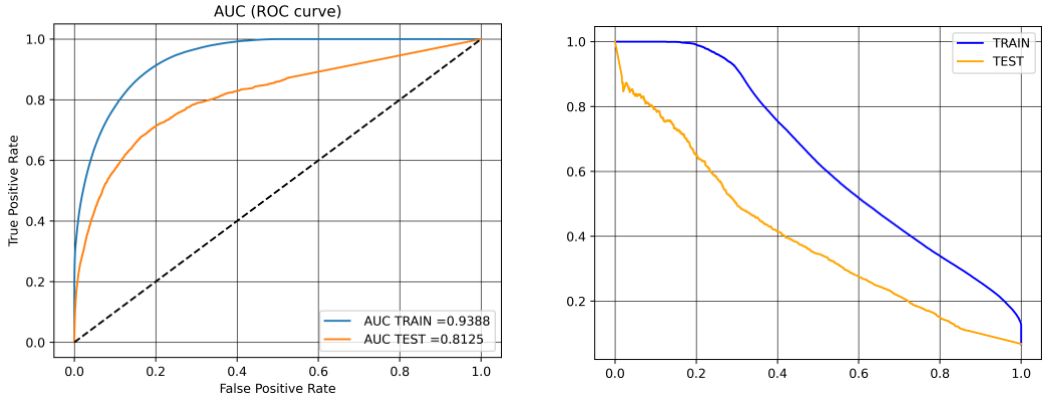


Figure A.5: Significant Variables from Random Forest with Ordinal Encoding

Target Encoding

The AUC and PR-curve of the target encoding method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.6: Random Forest: Target Encoding

In Figure A.6 the orange line represents the test data and the blue line the training data. The target encoding method ranked the most significant variables as follows:

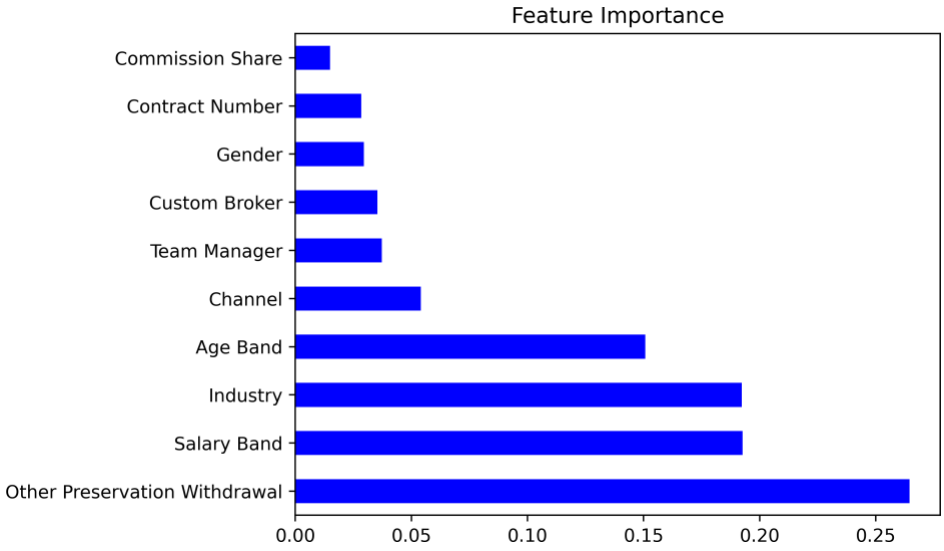


Figure A.7: Significant Variables from Random Forest with Target Encoding

Model Performance with Training Data

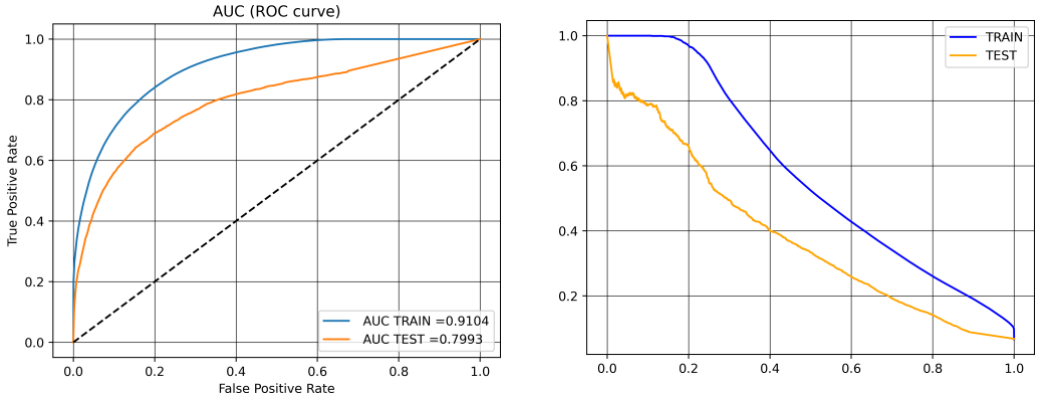
Table A.4 shows the performance of the random forest model with the three different encoding methods on the training data.

	Ordinal Encoding	Dummy Encoding	Target Encoding
Accuracy	95%	95%	95%
Precision	84%	84%	85%
Inverse Precision	95%	95%	95%
Recall	34%	34%	34%
AUC	93.88%	93.90%	93.88%

Table A.4: Comparison of Random Forest Models with Different Encoding on Training Data

Significant Variable Random Forest Model

The AUC and PR-curve of the random forest model with significant variables are shown below.



a) ROC b) Precision-Recall Curve

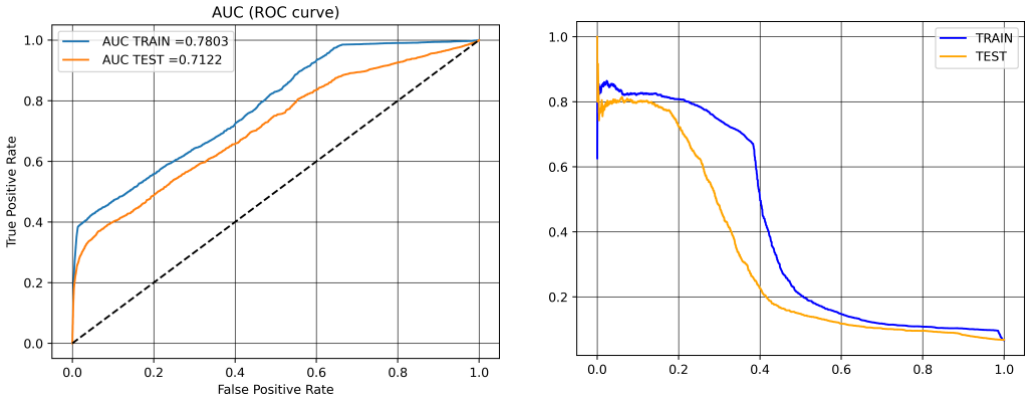
Figure A.8: Random Forest with Significant Variables

In Figure A.8 the orange line represents the test data and the blue line the training data.

A.4 Support Vector Machine: More Results

Ordinal Encoding

The AUC and PR-curve of the ordinal encoding method are shown below.



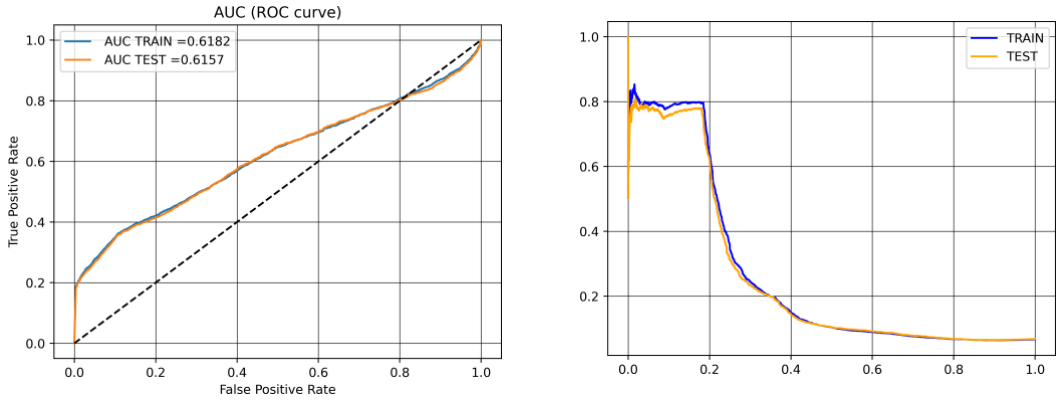
a) ROC b) Precision-Recall Curve

Figure A.9: SVM: Ordinal Encoding

In Figure A.9 the orange line represents the test data and the blue line the training data.

Target Encoding

The AUC and PR-curve of the target encoding method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.10: SVM: Target Encoding

In Figure A.10 the orange line represents the test data and the blue line the training data.

Model Performance with Training Data

Table A.5: Comparison of SVM Models with Different Encoding on Training Data shows the performance of the SVM model with the three different encoding methods on the training data.

	Ordinal Encoding	**Dummy Encoding**	**Target Encoding**
Accuracy	94%	94%	94%
Precision	84%	80%	80%
Inverse Precision	94%	94%	94%
Recall	5%	18%	18%
AUC	78.03%	79.23%	61.82%

Table A.5: Comparison of SVM Models with Different Encoding on Training Data

A.5 Balanced Data

There was an imbalance in the data. Many more individuals did not preserve versus the individuals who did preserve. This affected the recall rate of the models. However, if the recall rate is increased then the precision is decreased, which was not a desired outcome.

This section shows how the balanced data was sampled and some of the results of the models with the balanced data.

Data Resampling

An unbalanced dataset is a common issue in real-world applications. Unbalanced datasets can have a major impact on how well machine learning algorithms perform in terms of categorization. Over- and/or under-sampling and other artificial rebalancing techniques are frequently used to address the issue of imbalanced data (Ganganwar, 2012; L. Wang et al., 2021).

To address the imbalance in the data in this study, a hybrid, SMOTE-Tomek Links, resampling technique was used.

SMOTE-Tomek Links

SMOTE-Tomek Links is a hybrid of over- and under-sampling techniques.It works as follows: Using the SMOTE technique, which generates false instances based on k-nearest neighbours, the minority class is oversampled. Then, the Tomek Links technique is used to eliminate ambiguous entries from the dataset when they have two closest neighbour instances that belong to different classes. By using this strategy, the class separation close to the decision boundaries is improved (Nazyrova et al., 2022).

Even though over- and/or under-sampling methods are commonly used, it has been found that where the negative cases far exceed the positive instances, SVM’s performance is extremely constrained. Although undersampling the majority class does enhance SVM performance, there is a corresponding loss of important data. Thus, a method called modified support vector machines

also performs well on imbalanced data sets, but this resampling method is not expected to improve the SVM's performance (L. Wang et al., 2021).

A.5.1 Logistic Regression: Balanced Data

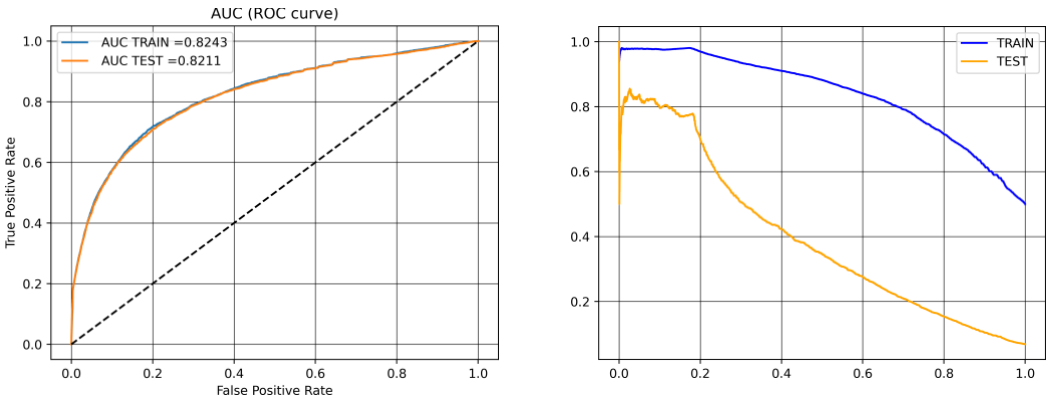
With ordinal encoding on the balanced data, the following results were obtained from the logistic regression model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	3 757	14 233
	No Preservation	1 601	59 070

Table A.6: Logistic Regression: Balanced Data Ordinal Encoding Confusion Matrix

The model correctly predicts 97% ($\frac{59\,070}{59\,070+1\,601}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 21% ($\frac{3\,757}{3\,757+14\,233}$) (Precision) of the individuals who do preserve.

The AUC and PR curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.11: Logistic Regression: Ordinal Encoding Balanced Data

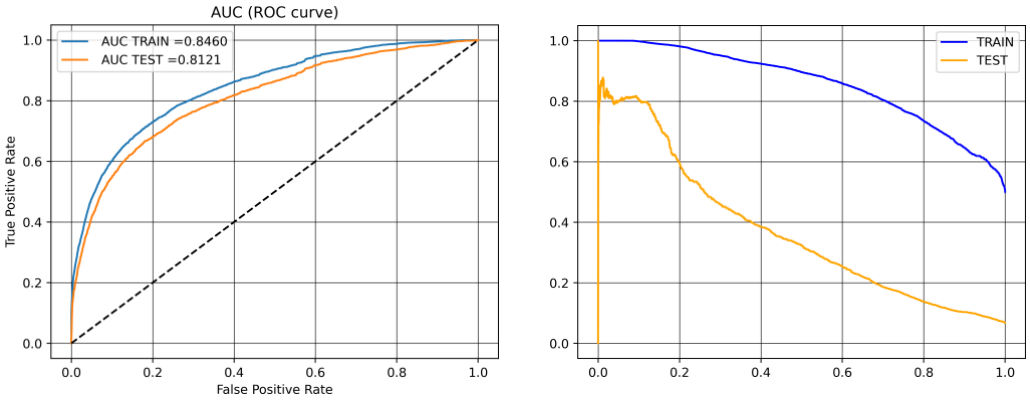
With dummy encoding on the balanced data, the following results were obtained from the logistic regression model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	3 613	14 128
	No Preservation	1 745	59 175

Table A.7: Logistic Regression: Balanced Data Dummy Encoding Confusion Matrix

The model correctly predicts 97% ($\frac{59\,175}{59\,175+1\,745}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 20% ($\frac{3\,613}{3\,163+14\,128}$) (Precision) of the individuals who do preserve.

The AUC and PR-curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.12: Logistic Regression: Dummy Encoding Balanced Data

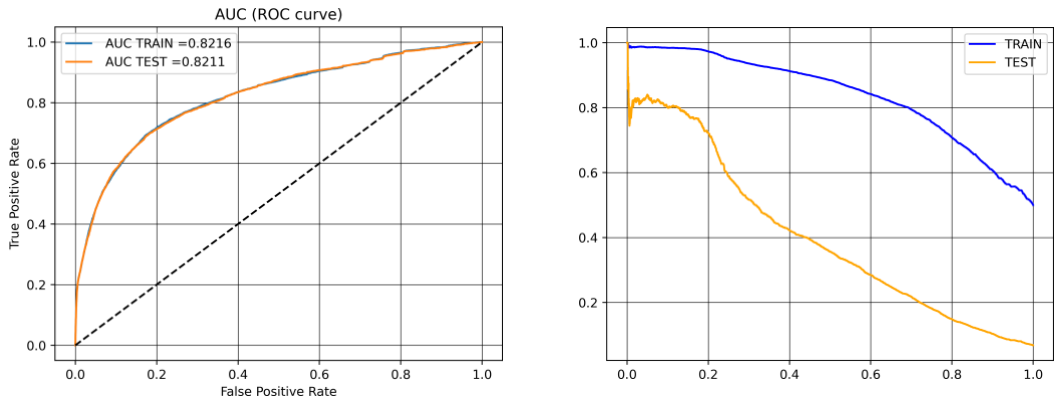
With target encoding on the balanced data, the following results were obtained from the logistic regression model:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	3 538	11 115
	No Preservation	1 820	62 188

Table A.8: Logistic Regression: Balanced Data Target Encoding Confusion Matrix

The model correctly predicts 97% ($\frac{62\ 188}{62\ 188+1\ 820}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 24% ($\frac{3\ 538}{3\ 538+11\ 115}$) (Precision) of the individuals who do preserve.

The AUC and PR-curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.13: Logistic Regression: Target Encoding Balanced Data

Table A.9: Comparison of Logistic Regression Models with Different Encoding on Balanced Training Data shows the performance of the logistic regression model with the three different encoding methods on the training data.

	Ordinal Encoding	**Dummy Encoding**	**Target Encoding**
Accuracy	76%	76%	75%
Precision	79%	79%	81%
Inverse Precision	74%	75%	72%
Recall	71%	72%	66%
AUC	82.43%	84.60%	82.16%

Table A.9: Comparison of Logistic Regression Models with Different Encoding on Balanced Training Data

Overall the models are accurate with the training data, with each of them having an accuracy of 76%. This is a lot less accurate than the imbalanced dataset.

The AUC of all three methods is also high, with the dummy encoding method being the highest. This was very similar to the imbalanced dataset.

	Ordinal Encoding	**Dummy Encoding**	**Target Encoding**
Accuracy	80%	80%	84%
Precision	21%	20%	24%
Inverse Precision	97%	97%	97%
Recall	70%	67%	66%
AUC	82.11%	81.21%	82.11%

Table A.10: Comparison of Logistic Regression Models with Different Encoding on Balanced Test Data

In Table A.10: Comparison of Logistic Regression Models with Different Encoding on Balanced Test Data, the precision is not as high for balanced test data as for the unbalanced data.

A.5.2 Random Forest: Balanced Data

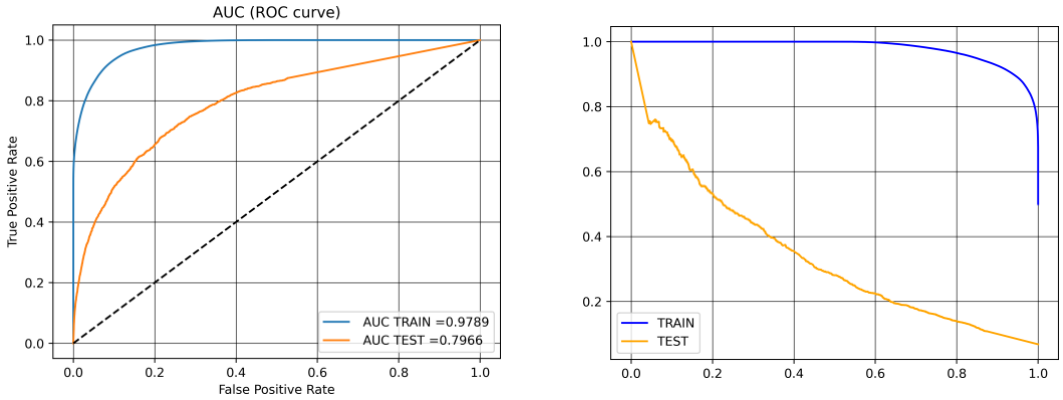
With ordinal encoding on the balanced data, the following results were obtained from the random forest model:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	2 818	7 808
	No Preservation	2 540	65 495

Table A.11: Random Forest: Balanced Data Ordinal Encoding Confusion Matrix

The model correctly predicts 96% ($\frac{65\ 495}{65\ 495+2\ 540}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 27% ($\frac{2\ 818}{2\ 818+7\ 808}$) (Precision) of the individuals who do preserve.

The AUC and PR-curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.14: Random Forest: Ordinal Encoding Balanced Data

The ordinal encoding method ranked the most significant variables as follows:

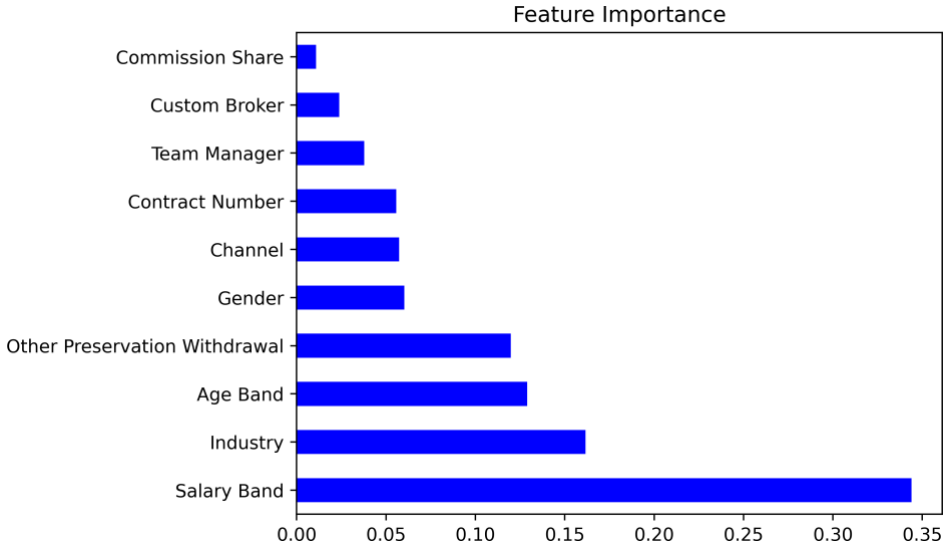


Figure A.15: Significant Variables from Random Forest: Balanced Data Ordinal Encoding

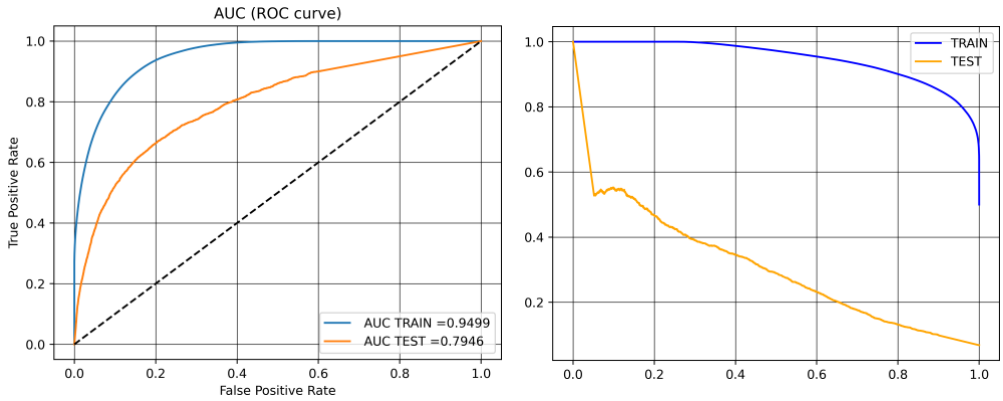
With dummy encoding on the balanced data, the following results were obtained from the random forest model:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	3 403	12 793
	No Preservation	1 955	60 510

Table A.12: Random Forest: Balanced Data Dummy Encoding Confusion Matrix

The model correctly predicts 97% ($\frac{60\ 510}{60\ 510+1\ 955}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 21% ($\frac{3\ 403}{3\ 403+12\ 793}$) (Precision) of the individuals who do preserve.

The AUC and PR curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.16: Random Forest: Dummy Encoding Balanced Data

The dummy encoding method ranked the most significant variables as follows:

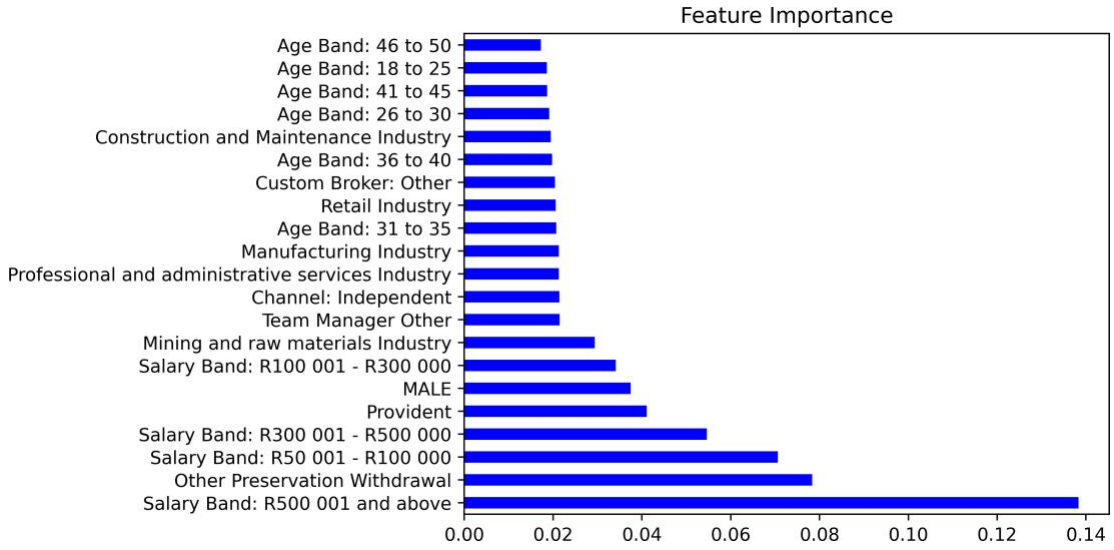


Figure A.17: Significant Variables from Random Forest: Balanced Data Dummy Encoding

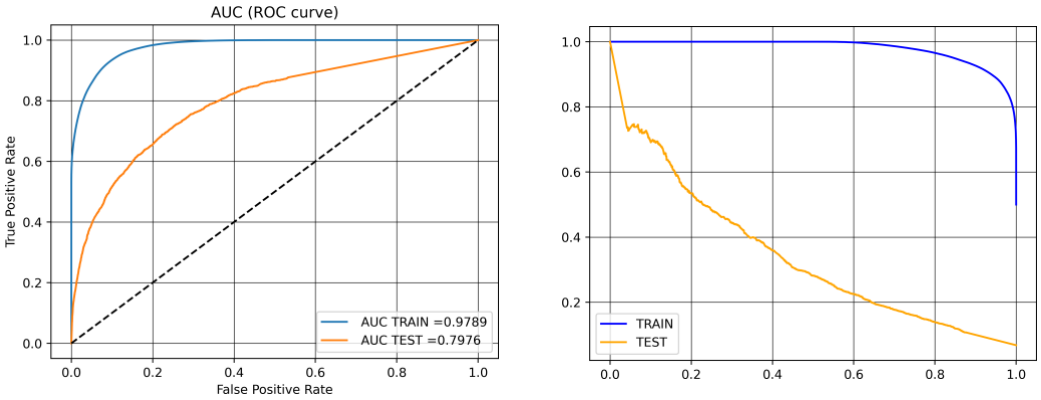
With target encoding on the balanced data, the following results were obtained from the random forest model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	2 822	7 806
	No Preservation	2 536	65 497

Table A.13: Random Forest: Balanced Data Target Encoding Confusion Matrix

The model correctly predicts 96% ($\frac{65\ 497}{65\ 497+2\ 536}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 27% ($\frac{2\ 822}{2\ 822+7\ 806}$) (Precision) of the individuals who do preserve.

The AUC and PR-curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.18: Random Forest: Target Encoding Balanced Data

The target encoding method ranked the most significant variables as follows:

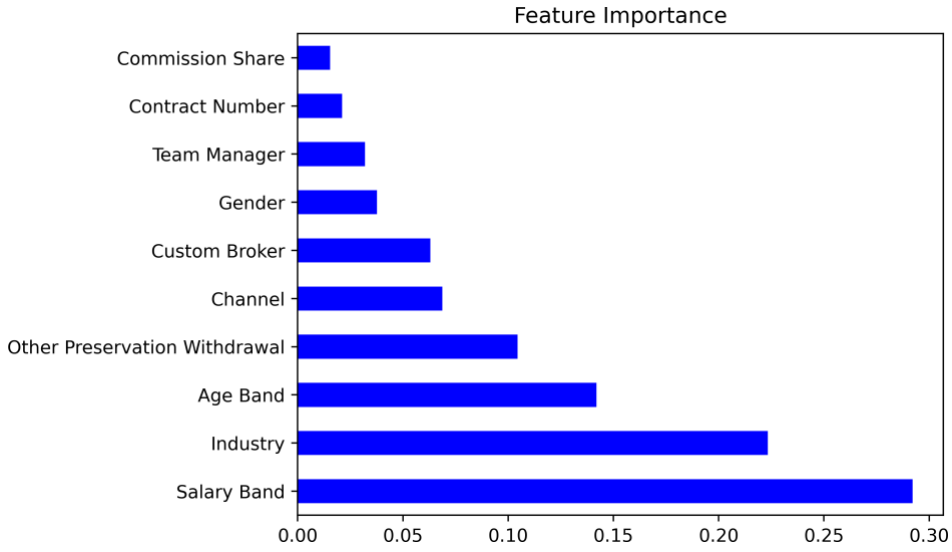


Figure A.19: Significant Variables from Random Forest: Balanced Data Target Encoding

Table A.14 below shows the performance of the random forest model with the three different encoding methods on the training data.

	Ordinal Encoding	Dummy Encoding	Target Encoding
Accuracy	92%	87%	92%
Precision	91%	85%	91%
Inverse Precision	93%	90%	93%
Recall	93%	90%	93%
AUC	97.89%	94.99%	97.89%

Table A.14: Comparison of Random Forest Models with Different Encoding on Balanced Training Data

Table A.15 shows the results on test data.

	Ordinal Encoding	Dummy Encoding	Target Encoding
Accuracy	87%	81%	87%
Precision	27%	21%	27%
Inverse Precision	96%	97%	96%
Recall	53%	64%	53%
AUC	79.66%	79.46%	79.76%

Table A.15: Comparison of Random Forest Models with Different Encoding on Balanced Test Data

A.5.3 Support Vector Machine: Balanced Data

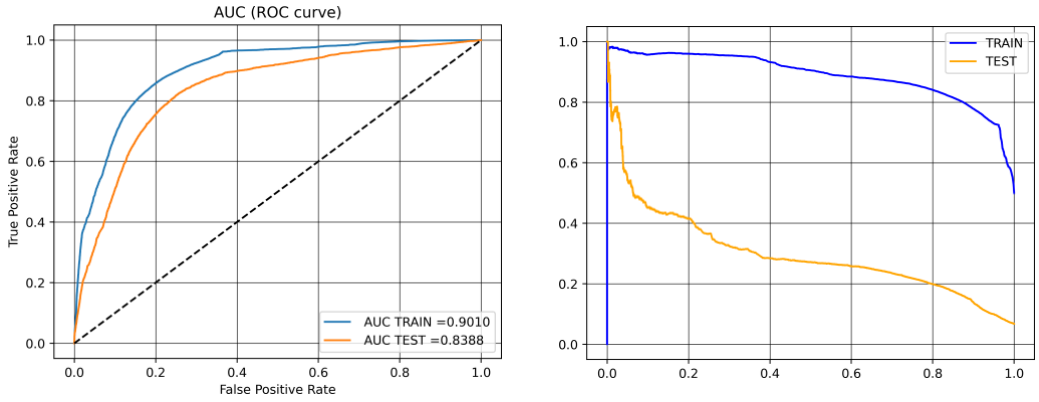
With ordinal encoding on the balanced data, the following results were obtained from the SVM model.

		Predicted	
		Preservation	No Preservation
Actual	Preservation	4 043	14 607
	No Preservation	1 315	58 696

Table A.16: Support Vector Machine: Balanced Data Ordinal Encoding Confusion Matrix

The model correctly predicts 98% ($\frac{58\ 696}{58\ 696+1\ 315}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 22% ($\frac{4\ 043}{4\ 043+14\ 607}$) (Precision) of the individuals who do preserve.

The AUC and PR-curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.20: SVM: Ordinal Encoding Balanced Data

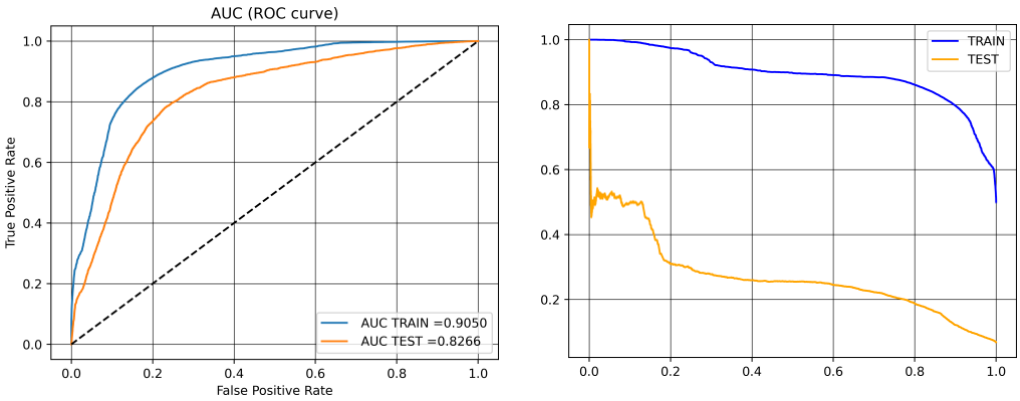
With dummy encoding on the balanced data, the following results were obtained from the SVM model:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	3 720	12 866
	No Preservation	1 638	60 437

Table A.17: Support Vector Machine: Balanced Data Dummy Encoding Confusion Matrix

The model correctly predicts 97% ($\frac{60\ 437}{60\ 437+1\ 638}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 22% ($\frac{3\ 720}{3\ 720+12\ 866}$) (Precision) of the individuals who do preserve.

The AUC and PR-curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.21: SVM: Dummy Encoding Balanced Data

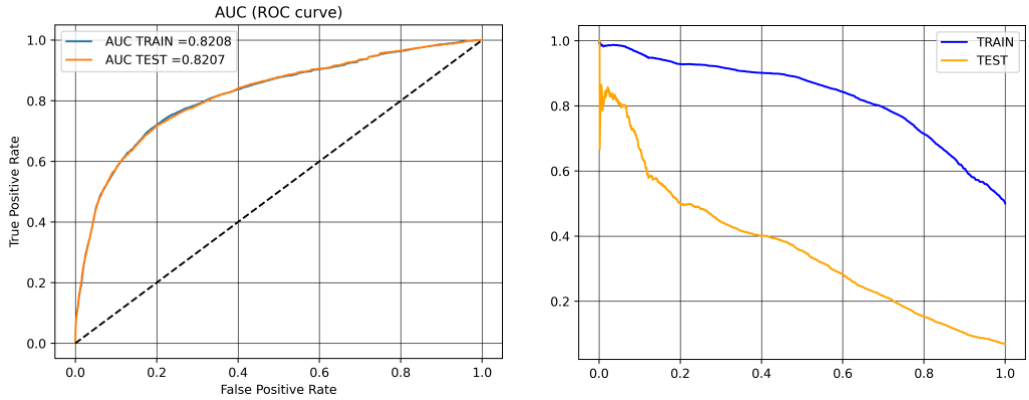
With target encoding on the balanced data, the following results were obtained from the SVM model:

		Predicted	
		Preservation	No Preservation
Actual	Preservation	3 385	9 868
	No Preservation	1 973	63 435

Table A.18: Support Vector Machine: Balanced Data Target Encoding Confusion Matrix

The model correctly predicts 97% ($\frac{63\ 435}{63\ 435+1\ 973}$) of the individuals who do not preserve (Inverse Precision) and correctly predicts the 26% ($\frac{3\ 385}{3\ 385+9\ 868}$) (Precision) of the individuals who do preserve.

The AUC and PR curve of this method are shown below.



a) ROC b) Precision-Recall Curve

Figure A.22: SVM: Target Encoding Balanced Data

Table A.19 shows the performance of the SVM model with the three different encoding methods on the training data.

	Ordinal Encoding	Dummy Encoding	Target Encoding
Accuracy	83%	84%	75%
Precision	81%	83%	82%
Inverse Precision	85%	85%	70%
Recall	85%	85%	64%
AUC	90.10%	90.50%	82.08%

Table A.19: Comparison of SVM Models with different encoding on Balanced Training Data

Table A.20 shows the results for balanced test data.

	Ordinal Encoding	Dummy Encoding	Target Encoding
Accuracy	80%	82%	85%
Precision	22%	22%	26%
Inverse Precision	98%	97%	97%
Recall	75%	69%	63%
AUC	83.88%	82.66%	82.07%

Table A.20: Comparison of SVM Models with Different Encoding on Balanced Test Data