Health Insurance Cross-Selling Predictions with Machine Learning for South African Consumers

Submission in completion of the requirements for the Degree of

Master of Information and Communications Technology

in the Faculty of Accounting and Informatics

at the Durban University of Technology

By

Khulekani Mavundla

(20907985)

Date Submission: 2024-02-25

Supervisor: Prof. Surendra Thakur *DTech(ICT)* **Date:** 2024-02-25

Declaration

I, Khulekani Mavundla, hereby certify that the dissertation's content is entirely original. I have included detailed references for every source that I have used in the text. No earlier version of this dissertation in any format has been submitted to the Durban University of Technology or to any other organisation for review or for any other reason.

Student: [redacted] **Date:** 2024-07-02

Approved for final submission

Supervisor: [redacted] **Date:** 2024-08-13

Dedication

To God, the unerring source of wisdom and guidance, I dedicate this dissertation. Your unwavering love and divine grace have been my constant companions on this academic journey.

To my family, for their boundless support, understanding, and sacrifice, I owe an immeasurable debt of gratitude. You have been my rock, my motivation, and my unwavering source of strength.

To my friends, who have shared in both my triumphs and tribulations, I am forever thankful. Your encouragement, laughter, and unwavering faith in me have made this endeavour all the more rewarding.

May this work stand as a testament to the values and principles instilled in me by my faith, my family, and my friends. With all my heart, I dedicate this dissertation to you, with love and appreciation.

Acknowledgements

I would like to express my heartfelt gratitude to several individuals and groups who have played pivotal roles in the completion of this dissertation.

First and foremost, I am deeply indebted to my dissertation supervisor, Prof. Surendra Thakur, whose unwavering guidance, expertise, and patience have been invaluable throughout this research journey. Your mentorship, supervision, and insightful feedback have been instrumental in shaping the quality and direction of this work.

This dissertation also benefited from the academic support provided by the research committee at Durban University of Technology. Their expertise and resources have been vital in the successful completion of this research.

I extend my appreciation to my colleagues, both current and former, who have offered their support, collaboration, and camaraderie. Their contributions to the dataset, during discussions, and in various brainstorming sessions have enriched this research and my overall experience.

Abstract

Cross-selling is the practice of selling additional products or services to an existing customer to increase business revenue. Cross-selling health insurance is challenging for companies, as they spend significant time meeting with prospective clients without knowing the likelihood of a sale. A health insurance provider often markets additional insurance products to its clients through different channels. This study aims to develop a robust ML model to help health insurance companies identify potential customers likely to engage in cross-selling. Objectives include extracting and preparing customer data from a large South African insurance company using suitable ML techniques. The study also seeks to determine effective algorithms for predicting health insurance cross-selling and to identify influential features for algorithm selection. This study adopted a quantitative research approach focused on extracting health insurance customer data. To achieve this, the study applied ML techniques by using the Python language using a dataset obtained from a large South African insurance company which is a rich repository that contains demographics, health conditions, and policy information. The study applied various ML algorithms, including Random Forest, K-Nearest Neighbors, XGBoost classifier, and Logistic Regression, feature engineering techniques were employed to enhance predictive accuracy. Analyzing 1,000,000 customer records with 17 features, Random Forest emerged as the top model with an accuracy of 0.91 and an F1 score of 1.00. The study found that customers aged 25-70, with prior insurance and longer service history, are more likely to purchase additional health insurance. This study will assist insurance providers in developing a strategy for reaching out to those clients in order to enhance their business operations and revenue.

Keywords

Cross-selling, Machine learning algorithms, Health insurance, Prediction, Feature engineering, Model training, Model evaluation metrics, Supervised machine learning, Unsupervised machine learning.

Table of Contents

List of Figures

List of Tables

Output

Journal papers arising from this study (Submitted)

Mavundla, K. and Thakur, S. 2024. Predicting the likelihood of cross-selling health insurance products to existing policyholders using machine learning techniques.

Conferences arising from this study

Mavundla, K. and Thakur, S. 2023. Analysing Health Insurance Customer Dataset to Determine Cross-Selling Potential. In: 2023 International Conference on Artificial Intelligence and Its Applications (ICARTI 2023). Preskil Island Resort, Mahebourg, Mauritius, 9-10 November 2023. 219-226.

Mavundla, K. and Thakur, S. 2024. Employing Machine Learning Techniques to Analyse Customer Records for Cross-Selling Probability. In: NEMISA Digital Skills Summit and Colloquium 2024. East London International Convention Centre, East London, 21-23 February 2024.

Presentations arising from this study

Mavundla, K. 2023. Artificial Intelligence and its applications in learning, teaching, and assessment. *Centre for Excellence in Learning and Teaching (CELT).* Online workshop. Durban University of Technology, 25 July 2023.

Mavundla, K. 2023. Lecture for Engineering Honors Degree Students: Exploring the World of 4IR/5IR Technology. *Exploring 4IR/5IR and Its applications.* Lecture. Durban University of Technology, 14 August 2023.

Mavundla, K. 2023. Machine Learning and Artificial Intelligence for Student Researchers. *Artificial Intelligence and Its applications.* Lecture. Eduvos Private College, Durban Campus, 25 August 2023.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Health insurance is a means of financial protection from health risks where an insurance company provides insurance to its customers which details the conditions and circumstances under which the insurance company will compensate the customer (Hung *et al.* 2020). With the rapid increase in the human population and diseases, and the impact of the COVID-19 pandemic, the health insurance industry has had to quickly adapt their operations and the precision of their cross-selling of health insurance to customers. Thus, the global pandemic impact has challenged the insurance companies to adopt new approaches of selling of insurance. Models using machine learning (ML) predictions can assess the likelihood that recent clients from the previous year may be interested in purchasing more health insurance policies (Kirti and Shin 2020).

Cross-selling is a key to increase revenue from existing customers by selling additional health insurance products to increase customer retention and company profits. Customer retention is one of the crucial priorities of every insurance company because customers are a direct link with the company's future revenue streams (Hammah 2020).

The primary goal of this study was to obtain useful knowledge and information from the existing research by specialists regarding data analytics and predictions tools applicable for health insurance cross-selling prediction (Wang 2020). Machine learning and predictive algorithms can be used to examine historical data and extract useful information to support companies to identify and target potential markets by revealing customer behaviour patterns to gain business insights (Larzelere 2021). In the context of South African health insurance, these techniques play a crucial role in optimizing cross-selling strategies tailored to local demographics and market dynamics (Kumar, Amgoth and Annavarapu 2018). By leveraging machine learning models, insurers in South Africa can enhance their understanding of customer needs,

improve targeting of additional insurance products, and ultimately drive business growth while improving customer satisfaction and retention.

1.2 Research problem

South African service providers have the good fortune of having a pool of customers who generally are reluctant to leave the company services they subscribed to, regardless of the quality of the service or products, and are almost indifferent to any challenges they face (Nkolele and Wang 2021).

Health insurance has been found to improve healthcare operations and reduce devastating health expenditures globally (Kumar *et al.* 2019). Extracting useful knowledge and information in massive customer data, and obtaining more customer resources has become a major competitive challenge among insurance companies. Contacting potential clients at the appropriate moment to offer the appropriate range of products that meet their needs has grown to be a significant issue for all insurance companies (Kumar *et al.* 2019).

The health insurance industry in South Africa is perceived to be reluctant to join the digital revolution space when compared with other industries. By enhancing cross-selling to current clients and cutting expenses, digital transformation presents a chance to improve health insurance performance and boost income for businesses (Kumar, Amgoth and Annavarapu 2018).

This study focuses on health insurance product use by South African consumers to predict cross-selling. Health insurance product sales have an uneven bias towards high earners and their families; therefore, there is a huge gap in other income bands. The Covid-19 pandemic experience has demonstrated the need for health insurance for every South African citizen (McKinney, Swartz and McKinney 2020). The National Health Insurance (NHI) initiative is a transformative megaproject. This study, however, is not about health insurance as such, rather it is about increasing health insurance quality and the range of cross-selling products at all levels (Ayanore *et al.* 2019).

Cross-selling health insurance products by leveraging ML and data prediction is a way to gain competitive advantage, customer retention, and create business insights (Wang 2020). Based on customer behaviour or purchasing trends, an insurance company would be interested to know if the existing insured customers who are on

lower insurance monthly premiums and customers with many years of using company services would be interested in purchasing additional health insurance products. Such analysis can also reveal the purchasing trends according to customer demographics such as age bands (Bielawska and Lyskawa 2021).

1.3 Research aim and objectives

The aim of this research was to determine how ML can be applied to increase the cross-selling revenue of companies in the South African health insurance industry.

In order to accomplish this research aim, the following research objectives were set:

- To use suitable data mining techniques to extract and preprocess health insurance customer data from a large insurance company in South Africa.
- To determine which machine learning algorithms are most useful for predicting health insurance cross-selling.
- To determine which features influence the machine learning algorithms chosen for cross-selling health insurance.
- To evaluate the performance of the machine learning algorithms using standard performance metrics.

1.4 Research questions

The following research questions were constructed to achieve each of the objectives.

1. What is the importance of cross-selling health insurance products using data analytics?
2. What data mining techniques can be used for extracting health insurance customer data from a large insurance company in South Africa?
3. What existing machine learning techniques can be applied to predict cross-selling of health insurance?

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The chapter provides the reader with an explanation of data analytics and the theoretical framework that will be used. It explains what cross-selling is in the insurance industry, and gives the reader background information about health insurance (the health industry, products, stakeholders, sellers). It also explains what open-source and private-source databases exist where customer data can be extracted and explains what type of customer data is stored in the database.

2.2 Cross-selling in the health insurance industry

In the health insurance sector, cross-selling refers to the practice of advertising additional insurance plans that complement the ones that current clients already use (Sekeroglu 2021)

Cross-selling techniques can be effectively managed by revealing a customer's purchasing behaviour pattern from the database where it has been stored (Sekeroglu 2021).

2.2.1 Cross-selling

Cross-selling is the sale of additional products or services that are considered related to the primary products that an existing customer has already purchased (Rajesh and Vijayakumar 2021). It is far simpler and less expensive to cross-sell products to an existing customer than to find new ones. This is considered one of the best and easiest methods of generating additional revenues and strengthening customer relationships with the business.

In the health insurance sector customers tend to purchase more goods from the same supplier because switching to a rival results in higher switching costs (Sumartha and Samopa 2017). Cross-selling is therefore seen as a powerful tool for reducing client attrition and raising lifetime value.

Cross-selling assumes that the demand for or sale of products will lead to an additional demand for its associated products (Zhang *et al.* 2018). Cross-selling is a tactic used by sellers to give their current clientele the chance to buy more items from them. This is done by offering clients things that will go well in some way with their first purchase/s (Rajesh and Vijayakumar 2021). The main purpose of cross-selling is to either enhance the revenue from the client or to save the interrelation.

Cross-selling is a well-known method used to boost an organisation's earnings by combining the sale of multiple related products into one transaction. Cross-selling is essentially the practice of offering more services or goods to customers while they are still in the process of making a purchase. For example, if a new customer opens a bank account, the seller will also cross-sell the life insurance in addition to the original product (Sekeroglu 2019).

All organisations seek to increase profit by boosting the quantity of products sold through cross-selling methods to generate consistent recommendations for products. A variety of cutting-edge techniques, such as collaborative filtering and natural language processing (Sekeroglu, 2019), can be employed to generate personalized suggestions for products. These techniques analyze customer behavior and preferences based on historical data, enabling companies to tailor their cross-selling strategies effectively.

In summary, the purpose of cross-selling is to support companies to increase their profitability, enhance customer loyalty, generate more leads, streamline the customer's purchase processes, and enhance convenience.

2.2.2 The health insurance industry

Health insurance commonly offers protection against the potential costs of individual medical treatments or for time lost caused due to injury or illness (Dawson *et al.* 2018).

In the realm of health insurance, policyholders maintain enduring connections with their insurance service provider by relying on information about their attributes, financial interactions, demographic details, and conduct that are securely stored within the company's database. This customer data can be leveraged to identify preferred clients and through cross-selling offer them complementary products they may not currently possess (Markovic 2020).

The number of insurance industry personnel, fixed asset investments, premium income, and compensation costs are some of the factors that have a substantial impact on the growth of the health insurance sector (Darrat and Flaherty 2019).

The insurance company's premium structure and market share increase in direct proportion to the level of fixed asset investment. In the health insurance sector, premium income serves as a direct indicator of the companies' gross productivity (Darrat and Flaherty 2019).

In the insurance sector, compensation costs are one of the key indicators of an organisation's capacity for accepting risk (Wang *et al.* 2021).

2.2.2.1 Global health insurance industry

The health insurance industry in India is highly competitive and significant within the service sector. This is due to its dual role of not only offering risk protection for individuals but also fostering a culture of saving and investment (Ramamoorthy and Kumar 2018; Nayak, Bhattacharyya and Krishnamoorthy 2019). This sector has grown crucial in offering preventive and wellness measures due to the rising prevalence of lifestyle-related ailments in emerging markets like India. Health insurance requirements and risks vary based on an individual's age, gender, lifestyle, occupation, and economic standing. Therefore, the uptake of health insurance could be higher with the creation of tailored products for specific needs and risk categories, as opposed to generic offerings for everyone (Prakash, Pandey and Pareek 2018).

A health insurance firm can obtain a competitive edge from a variety of sources. Firstly, insurance firms must understand the hospital sector, hospitalisation costs, and healthcare funding. Only then can they concentrate on identifying these sources. Secondly, firms must research the major companies in the health insurance market. Thirdly, they can then develop a suitable business model for health insurance (Kumar and Duggirala 2021).

Health insurance markets in the United States are categorised by the presence of imperfect information as well as complex products and costs. Health insurance customers can choose from an extensive range of insurance products that vary in many dimensions such as deductibles, coverage of services, and medications, and

out-of-pocket expenses for different health services(Karaca-Mandic, Feldman and Graven 2018).

All people in the Netherlands are required to purchase health insurance directly from a private health insurance market (Warner *et al.* 2020). The Health Insurance Act of 2006 established a managed competition model in which the contents of the basic health insurance package are subject to stringent government regulation.

The health insurance system in Germany has a strong foundation in solidarity among the insured population, a principle that is evident in both the financial contributions and the provision of services within statutory health insurance. In this system, all insured individuals, regardless of their health risk, make financial contributions based on their income. These contributions grant them access to benefits that are tailored to their medical expenses and healthcare needs (Busse *et al.* 2017).

2.2.2.2 Health insurance in Africa

The expense of delivering health insurance services is becoming a significant financial obstacle and a growing concern for many developing nations. The cost of health insurance services constitutes a substantial portion of household incomes, leading to an inability to afford the expenses of health insurance services (Alesane and Anang 2018). Health insurance in Ghana covers citizens by law, with exemption entitlement to some segments of the population (Fenny, Yates and Thompson 2017).

In Ghana, the National Health Insurance Scheme (NHIS) was instituted to enhance access to fair and quality healthcare for all residents, regardless of their socio-economic status (Solanki *et al.* 2020).

Tanzania and Kenya have similar health insurance programmes, although their target groups differ. They have separated insurance for formal and informal sectors which allows them to have no specific exemption schemes for the poor, but some waivers are given to patients in Tanzania who are assessed to be poor so that they can pay their bills (Fenny, Yates and Thompson 2017).

2.2.2.3 Health insurance in South Africa

In South Africa, disparities and inequalities in the availability, acceptability, and affordability of health insurance between the public and private sectors have been on

the rise, marked by rapid changes and an increasing need for accessible social services like healthcare. As a result, the idea of universal health coverage (UHC) has become increasingly important in numerous developing and underdeveloped nations. UHC aims to guarantee fair access to high-quality healthcare services. In South Africa, the government has highlighted UHC as an essential element of their sustainable development objectives, seeking to shield vulnerable communities from the economic challenges linked to medical costs and encourage increased involvement in healthcare insurance registration (Akokuwebe and Idemudia 2022).

The initial step in providing Universal Health Coverage (UHC) for its citizens is acknowledging the significance of advancing population health nationwide and the government's responsibility to ensure accessible health insurance for all. In 2011, the South African government decided to institute a health insurance scheme that would provide subsidisation of medical costs and reduce out-of-pocket payments with a single fund to cover all individuals, no matter their level of earnings (Akokuwebe and Idemudia 2022).

Provision of affordable health insurance to the population of the low and middle class is a persistent development issue in South Afirca (Akokuwebe and Idemudia 2022). The World Health Organisation (WHO) estimates that many people are pushed into poverty and other people suffer financial devastation because of high and out-of-pocket expenditure on health services, thus South Africa is implementing the NHI scheme to improve the accessibility of quality health insurance services for all South African citizens (Chimusoro *et al.* 2018).

South Africa has a sophisticated financial services industry in terms of products, services, and distribution infrastructure (Mohr *et al.* 2018). The highly developed health insurance industry in South Africa makes a significant contribution to these services compared to other countries; the assets of health insurance companies and pension funds correspond to well over 80% of GDP, a ratio that is higher than in the United Kingdom and Canada, though lower than in Switzerland, the Netherlands and the United Kingdom. More significantly, South Africa has the highest level of health insurance premiums in relation to GDP in the world (White, McAllister and Munro 2017).

2.2.3 Health insurance products

Insurance products are any health insurance coverage provided by an insurer or service contract provider in its insurance.

Health insurance products can include a variety of products such as hospital cash plans, gap cover, medical travel insurance, and primary healthcare.

For an individual or a group of people, a health insurance product is one of the most important risk management strategies by transferring risk from the insured to the insurer (Zhang and Nie 2021).

Customers may have varying preferences when it comes to health insurance products, especially concerning the types of coverage offered such as sum insured, hospitalisation benefits, inclusions, and exclusions (Kumar and Duggirala 2021). Customer preferences result in whether the customer is willing to pay the health insurance premium for a particular private insurance company's product compared to obtaining that service in the public sector (Kumar and Duggirala 2021).

The insurance sector holds significant importance in the business landscape, as numerous companies and individuals avail themselves of its services. Insurance firms provide a range of insurance offerings including life, property, unemployment, accident, as well as casualty and health coverage (Wen *et al*. 2021). Life insurance providers predominantly concentrate on personalized coverage and interest compensation. Health insurance shields against medical costs. Incidents involving vehicles, goods, and residences are addressed by accident insurance (Wen *et al*. 2021).

2.2.4 Stakeholders (customers and sellers)

The customer in health insurance refers to the person who buys the health insurance products and that person is called "insured" because he/she is the owner of the health insurance policy or the person with the health insurance coverage for medical and health expenses (Shilpa *et al.* 2021).

A health insurance seller refers to the person who presents one or more health insurance companies and their health insurance products to customers and the

company that provides health insurance medical services or contracts to the insured is called an "insurer" (Li *et al*. 2021).

Health insurance rates by customer's age

The age of the consumer is the main factor that affects health insurance premiums; young individuals often pay the lowest rates and older people the highest.

Health insurance customers consider age to understand how that will affect their premium when they purchase insurance, whereas insurance sellers consider customer age to offer necessary health insurance products. There are three insurance rates considered when customers are shopping for health insurance, namely, young adult health insurance, middle-age health insurance, and health insurance for seniors (Van Deventer 2022).

Young adult health insurance

Young folks (18 to 35 years old) typically have good health and may only require a minimal quantity of health insurance, which could lead to cheaper premiums of health insurance purchase. Young adults often opt for term health insurance, which provides financial flexibility to accommodate their spending plans and address urgent financial needs like debt repayment or burial costs.

Health insurance for middle-aged people

If the premiums are paid to the health insurance seller, individuals in the age range of 40 to 60 may be eligible for a permanent health insurance policy, in which the insurance provider provides lifetime protection. In the event that a middle-aged policyholder's spouse passes away, their health insurance may be designed to assist the surviving spouse in paying off the mortgage and all outstanding bills.

Health insurance for seniors

Each life stage comes with different needs for health insurance coverage, in most cases, senior customers between 70 to 80 years old may want to purchase health insurance to cover funeral costs and other final expenses to meet estate planning goals (Hagiwara 2022).

Public vs private insurance industry

Choosing a good health insurance plan that meets your needs is very difficult because you must understand the difference between public versus private health insurance. This study discusses the public and private health insurance sectors, and focuses on how they operate in terms of offering health insurance services to their customers (Duckett and Nemet 2019).

Public insurance industry

Public health insurance is insurance that is subsidised or paid by the public (government) with an agreement that covers the whole or part of the medical expenses incurred by the insured on the basis of the collective sharing of the healthcare risks of a large number of persons and the government (Kiri and Ojule 2020).

The government offers public health insurance programs to low-income individuals and families, the elderly, and other people who meet certain eligibility requirements. These policies are more affordable than private insurance (Berrone *et al.* 2019).

At the international level, the conceptions of universality in health insurance have become a challenge and are polarised based on a universal health system and ideas of universal health coverage. The term universality in developed countries is used to refer to the public coverage of national systems which means basic health services coverage (Giovanella *et al.* 2018).

Comprehensive health coverage ensures that everyone has access to the quality health services they require without financial problems. Public health insurance is highly subsidised by the government with normal payments required from policyholders (Joarder, Chaudhury and Mannan 2019).

In South Africa, the government has implemented the NHI system with the goal of establishing a healthcare financing mechanism. The NHI is designed to aggregate funds and offer access to high-quality and affordable health services for all South African individuals and families, aligning with their healthcare requirements, without causing undue financial burden (Katurura and Cilliers 2018).

The quality of public healthcare in South Africa is not of a good standard because there are too few doctors and outdated facilities and long waiting times for health

services. This has resulted in many people taking private health insurance in order to receive the standard of care they need (Katuu 2018).

In South Africa, everyone can access the public healthcare system, irrespective of nationality or immigration status. The public system is designed to support those that are low-income and cannot afford to secure private health insurance.

Private insurance industry

Private health insurance refers to plans provided by private insurance companies that are paid for by the individuals being covered. This insurance is often provided by an employer or other organisation with which the policyholder is affiliated. Private healthcare can provide more extensive care choices and extra services that might not always be offered by public health insurance (Filc, Rasooly and Davidovitch 2020).

Private health insurance pays for most patient services, and physicians are able to decide what services and treatments are in the patient's interest (Michel *et al*. 2020). Many individuals opt for private healthcare due to the increased flexibility it offers in healthcare decisions, such as selecting preferred hospitals and specialists. Additionally, some consumers acquire private health insurance services in response to governmental incentives and tax regulations (Coughlan 2021). For certain segments of the populace excluded from public health care programs, private health insurance plans may represent their only source of coverage. For this reason, private health insurance plans serve as a substitute for public policies.

Consumers purchase complementary private health insurance programmes for many reasons such as wanting to get services quicker and get better or more comfortable services than in the public system. Consumers are likely to switch from public to private health insurance due to the level of service delivery. Thus, both the public and private health insurance industries need to adopt cross-selling prediction models to know the possibility of customer retention.

2.2.5 Open-source and private-source databases

Open-source databases are any database that is free to view, download, modify, distribute, and reuse. However, there is a lack of publicly available health insurance open-source datasets because much of the data in the health insurance industry is

confidential and proprietary (Hsieh et al. 2020). This is a challenge from the point of view of cross-selling prediction.

With increased interest in applying machine learning and predictive modelling techniques across all health insurance industries, access to health insurance data is becoming more essential. Having access to realistic health insurance data from insurance companies can allow researchers to solve more real-world problems and validate the developed methodologies for health insurance cross-selling prediction (Hsieh 2020). If health insurance data is more publicly accessible/available this can allow researchers and health insurance companies to open source their methodologies and that will encourage other researchers to build models and extend research based on existing research. Most insurance industries are willing to share anonymised datasets, however, the effort required from them to create dummy data that is similar to the real data structure can prevent them from sharing anonymised data.

Health insurance datasets are available from various open-sources such as kaggle.com, which is a free repository that contains limited health insurance data including demographics, health conditions, and policy information that cannot be used to analyse and build a health insurance cross-selling predictive model (Babuna *et al.* 2020). The researcher opted not to use an open-source dataset to build a predictive model due to limited data availability and instead utilised a dataset from a large insurance company in South Africa for the study and the data was anonymised for confidentiality purposes.

2.3 Customer churn rates internationally and in South Africa

This section focuses on comparing health insurance customer churn rates internationally and in South Africa.

2.3.1 Customer churn rates internationally

Due to the hypercompetitive industry, service providers offer various ranges of services or products to their customers. However, companies are facing major challenges with revenue losses due to increased market competition and a high rate of customer loss. Customer can easily switch from one company to another if they are

not satisfied with the service or product offered. This switching is called "customer churn" (Dulhare and Ghori 2018).

Customer churn rate is described in business as the percentage of customers that stopped using a company's product or service during a certain time frame and this is indicated as one of the most critical metrics across industries because it is much less expensive to retain existing customers than it is to acquire new customers (Napitu and Putri 2018). In commerce, customer attrition refers to the departure of clients, wherein loyal or existing customers terminate their association with one company to transition their purchases to another. The rate of customer attrition is the likelihood that a customer will cease engagement with a company and is frequently assessed in sectors such as banking, telecommunications, and insurance. The churn rate is used to estimate company growth and is considered as an essential metric for business revenue and financial profit (Agrawal *et al.* 2018).

Napitu and Putri (2018) explored the factors that cause the customer churn and found that utilizing online social media platforms like X enables customers to readily and transparently voice their feedback regarding their brand or service encounters, with the quantity of feedback being linked to the level of customer attrition. This mode of communication has paved the way for a novel avenue of 21st-century research and challenged many companies across the world to change their customer service delivery and to understand the importance and need for customer churn to be evaluated and calculated by using data mining techniques to forecast the likelihood of a customer opting to switch (Lappeman *et al*. 2022).

Internationally, studies have indicated that the customer churn rate is high due to high competition in countries such as Taiwan, Turkey, Pakistan, and India resulting in companies not focusing on predicting customer churning factors but rather focusing on what measures should be considered for customer retention (Saghir *et al.* 2019). Customer loyalty is a fundamental element which is viewed as central to customer relationship management (CRM). Meanwhile, customer attrition is a concept linked to customer loyalty (Iranmanesh *et al*. 2019).

Churn forecasting techniques can assist in identifying customers who may likely leave in the upcoming period, and companies that can pinpoint these potential churners can offer them incentives to reduce the overall churn rate within the organisation. These

customers can be identified using their behaviour and demographic information (Johny and Mathai 2017).

Customer churn is divided into two categories, namely, voluntary and involuntary churn. Voluntary churn occurs when the existing customer leaves the company due to poor customer service and joins a competing company, while involuntary customer attrition takes place when a company requests a customer to halt or terminate their service due to issues like non-payment for the services rendered (Johny and Mathai 2017). Voluntary churn includes deliberate and incidental churn. Deliberate churn occurs through a desire to change service providers or package price rates. Incidental churn occurs when customers have not planned to leave the service provider but do so for reasons such as a change of location or a change in financial position.

The primary reasons that prompt the customer to cease using the company's services or products are: no understanding of the service plan, high costs, unstable customer support, and no satisfaction with the customer service. These main causes incurred significant costs for the business because attracting new clients is more expensive than keeping hold of current ones.

A client churn predictive model supports many companies internationally by detecting customers likely to switch to another company, enabling proactive retention strategies to reduce churn and maintain a stable customer base (Yousefli, Nasiri and Moselhi 2017). Such a model should be able to predict the right customers and estimate their time of churn in order to incorporate their requirements for avoiding churning.

The high rate of customer churn in the telecommunication and health insurance industries is due to the fact that customers are able to choose a service provider from a range of companies and have the right to switch from one service provider to another if they perceive that they will receive better service and support from that other provider. Telecommunication is the most highly affected industry—the challenge of revenue losses in this industry is mainly related to customer churn (Ahmed and Linen 2017).

2.3.2 Customer churn rates in South Africa

Digitalisation and an increased number of firms have created highly competitive markets in most industries worldwide. In South Africa the retail banking sector experiences intense competition between the five largest South African banks as well

as two new entrant banks which are Discovery and TymeBank. This competition increases the likelihood of customers deciding to switch their purchasing, which has made customer retention a primary concern in the South African banking sector thus ensuring that customer service satisfaction is prioritised to keep customers happy with all the services or products offered (Lappeman *et al*. 2022).

In South Africa, most companies consider that their profits are directly proportional to their customer base, and measure their profit by business growth meaning that customer growth must always be higher than the churn rate (Agrawal *et al*. 2018). This business model lowers the customer churn rate in while at the same time increasing the cross-selling opportunities to existing customers within the organisation.

Customer churn is faced by almost every industry nowadays especially the service industries, but South African service providers have the good fortune of having a pool of customers who generally are reluctant to leave the company services they subscribe to, regardless of the quality of the service or products, and, secondly, they are almost indifferent to any challenges they face. This is the major reason why customer churn rate is low (Nkolele and Wang 2021). Detecting churners in advance supports companies to identify which customers are going to leave and enables them to cross-sell additional products to the existing customers.

Customer churn prediction or detecting customers who are likely going to cease or cancel their subscription is an ever-growing field of study international and in South Africa. Marketing and actively managing customer churn is now more critical and costly than ever due to the high competition. The Covid-19 pandemic influenced many customers to switch from one company to another. Churn prediction can support many industries to segregates churners from non-churners. Companies will then be able to reprioritise their business strategies and customer service support to target only those customers instead of all customers by giving some incentives to retain them, and focusing on cross-selling additional products to increase business revenue (Dulhare and Ghori 2018).

2.4 Data analysis

2.4.1 Importance of using data analytics for cross-selling health insurance products

Data analytics does not support companies to cross-sell more effectively to the potential customer, however, it does support companies to change how they cross-sell additional products to existing customers (Koutsomitropoulos and Kalou 2017). By analysing data from existing health insurance customers, a company can identify patterns and find opportunities within the current customer base to cross-sell. Data analytics can provide opportunities to stop and reconnect with customers and view insights into the price of health insurance products and services, and these insights can lead to increased cross-selling and business revenue (Sekeroglu 2021).

Data analytics can enable health insurance companies to extract valuable insights regarding customer behaviour to build a health insurance cross-selling predictive model to support the marketing department in targeting the right customers. Analysing the existing customer data can also offer insights into improving health insurance customer satisfaction as satisfied customers are more likely to opt for policy renewals (Subramani 2022).

Data analytics has proven to be a game changer in numerous ways for equally competitive industries. Thus data analytics for health insurance cross-selling can support insurers in identifying profitable customer segments in order to design strategies to target the right customers with the right health insurance products (Das *et al.* 2022).

2.4.2 Data mining techniques for extracting health insurance customer data

In this study, the researcher will compare data mining techniques and choose one which is most appropriate.

Data mining is the data-driven extraction of information from a large dataset for predictive and descriptive analysis. This process aims to identify customer behavior and purchasing patterns, enabling accurate predictions for cross-selling opportunities (Schneeweiss *et al.* 2021).

There are several primary data mining techniques for extracting health insurance customer data including supervised learning and unsupervised learning.

Data mining techniques are used to carry out tasks such as classification, clustering, regression, and associated rules which are most commonly used for the study of prediction (Figure 2.1).

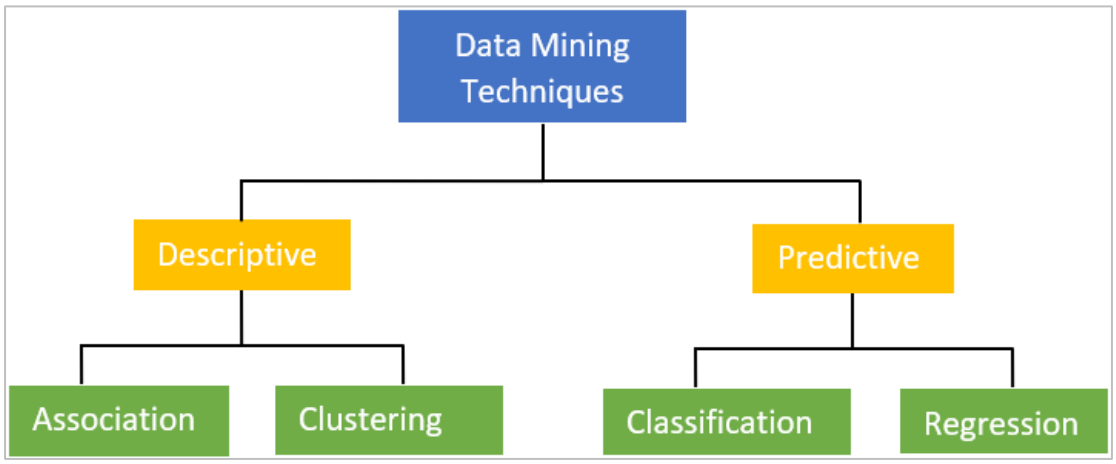


Figure 2.1: Data mining techniques

The study examines descriptive and predictive data mining methods related to supervised and unsupervised learning techniques that will support insurance companies to price their products profitably and promote new offers by cross-selling to their new or existing health insurance customers (Taylor *et al.* 2022).

Descriptive data mining describes the extracted health insurance customer data, summarises and converts it into a human-readable form in order to identify interesting patterns in the data so that the questions like what happened and what is happening can be easily addressed and answered. The two main descriptive mining tasks are clustering and association (Amani and Fadlalla 2017).

The **association** data mining method is a form of unsupervised learning technique that employs various rules to discover relationships among variables within a provided dataset.

The **clustering** data mining method is used to group unlabelled customer data based on their similarities or differences.

Predictive data mining is analysis used to predict what will happen in the future for insurance companies to make better decisions. In this study, predictive analytics were used to extracted health insurance data to predict which existing customers would be

most likely to be interested to purchase additional health insurance products (Fernandes and Ferreira 2023). The study focuses on the two primary predictive data mining tasks including classification analysis and regression analysis which are most appropriate for health insurance cross-selling prediction.

The **classification** data mining method is used to retrieve important and relevant information about customer data and classify it into different categories. This approach helps produce a general hypothesis for forecasting future customer data instances.

The **regression** data mining method is used to identify and analyze the relationship between variables, such as independent and dependent variables, within customer data.

2.4.3 Machine learning algorithms

An ML algorithm is a method for automatically creating models from data. It is generally used to turn datasets into a model for predicting output values from given input data (Ambika 2020).

Machine learning algorithms are generally categorised into three learning approaches are three types of learning including reinforcement, unsupervised, and supervised. Supervised and unsupervised learning methods are commonly used for ML prediction. Whereas supervised learning is employed to solve classification and regression problems, unsupervised learning is used to solve clustering problems (Banerjee 2021) (Figure 2.2).

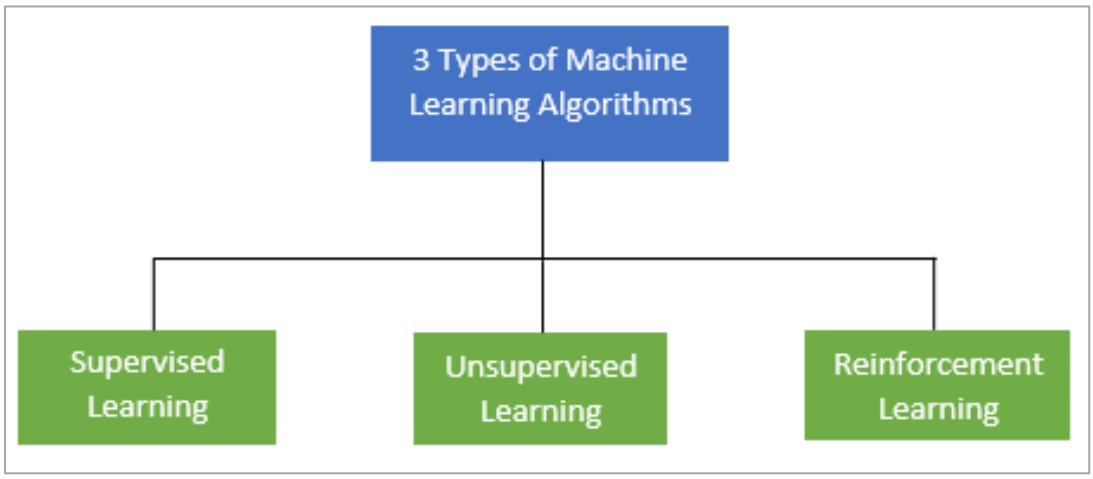


Figure 2.2: Machine learning algorithms

Supervised machine learning algorithms

Supervised ML is a technique that can be used according to what has been previously learned to get new data using labelled data and to predict potential outcomes or labels (Ambika 2020). Supervised ML techniques entail learning a mapping between a set of input variables ***X*** and an output variable ***Y*** for data analysis. This algorithm is considered to be the most essential methodology and is common in the classification of problems for ML models, and has been applied in this study of health insurance cross-selling prediction for mapping input and output variables to predict the outputs for unseen health insurance data for cross-selling (Nasteski 2017).

Regarding the supervised ML technique, the input dataset is divided into a train and a test dataset. The training dataset has an output variable that can predict and classify customer data to teach the cross-selling prediction model to yield the desired output. For data mining, supervised learning is separated into two types of methods including classification and regression (Mahesh 2018).

Several machine learning algorithms and computational methods are employed in the supervised learning process, such as Neural Networks, Naïve Bayes, Linear Regression, Logistic Regression, Support Vector Machine, K-Nearest Neighbor (KNN), and Random Forest.

The most primarily used supervised ML algorithms are:

Neural Networks algorithm

Commonly employed in deep learning algorithms, this method manages training data by mirroring the interconnected structure of the human brain using layers of nodes.

Naïve Bayes algorithm

A classification technique that utilizes the concept of class conditional independence based on Bayes theorem. The main advantage of Bayesian classification is its ability to address predictive challenges, and it is frequently applied in text categorization, spam detection, and recommendation systems.

Linear Regression algorithm

The objective of Linear Regression is to determine the relationships and associations between variables to forecast future events or outcomes when there is a single independent variable and a single dependent variable. Linear Regression falls under the supervised machine learning algorithms, where the model is trained on a set of labeled data (training data) to enable the prediction of labels for unlabeled data (testing data) (McEvoy *et al.* 2019).

Logistic Regression algorithm

Linear Regression is used when the dependent variables are continuous, Logistic Regression is a regression method that forecasts the probability of an event's occurrence by fitting data to a logistic function. It utilizes multiple predictor variables, which can be numerical or categorical. Binary outcomes, like true or false, or yes or no, are the representations used (McEvoy *et al.* 2019).

Support Vector Machine algorithm

Vladimir Vapnik created the well-known supervised learning model known as the Support Vector Machine. This model is used for both data classification and regression. Having said that, it is often used to solve classification problems by creating a hyperplane where the distance between two classes of data points is at its greatest.

K-Nearest Neighbour algorithm

The KNN algorithm combines data points based on their proximity to and correlation with other available pieces of information using a non-parametric method. Assumptions made by this method include the discovery of linked data points nearby. Following an attempt to calculate the distance between data points, usually using the Euclidean distance, it allocates a category based on the most common category or average.

Random Forest Algorithm

Regression and classification both use this. To reduce variance and generate more accurate data predictions, a group of uncorrelated decision trees known as the "forest" are joined.

The most commonly used unsupervised ML algorithms:

Unsupervised ML algorithms analyse and cluster unlabelled datasets in order to discover hidden patterns or data groupings without the need for human intervention. They are not used for any target or outcome variable to predict or estimate. This algorithm is good when working with a large dataset to identify online shoppers/customers that often purchase groups of products at the same time and segment them into different groups for specific interventions (Palacio-Niño and Berzal 2019).

In this study, unsupervised learning was applied for the process of inferring underlying hidden customer behaviour and purchase trends/patterns from health insurance customer historical data (Bojanowski and Joulin 2017). Unsupervised learning techniques include the apriori algorithm and K-means clustering.

Reinforcement Learning technique

During this training approach, the algorithm gets placed in an environment where the system self-educates and adjusts (not using labels) to make specific decisions continually, using trial and error. This machine-learning algorithm learns from previous experience and attempts to capture the best possible knowledge to make accurate business decisions (Mahesh 2018).

2.4.3.1 Features influencing machine learning algorithm selection for health insurance cross-celling

Feature selection is the process of reducing the input variables to the model by using only relevant data and getting rid of noise in the data.

The following are the benefits of feature selection for ML algorithms:

- **Diminished overfitting-** decreased redundant information leading to fewer chances of decisions being influenced by irrelevant data.

- **Enhanced precision-** reduced misleading information leading to improved modeling accuracy.
- **Decreased training duration-** fewer data points leading to simplified algorithm and expedited algorithm training.

Flowchart in Figure 2.3 shows the various methods used for feature selection.

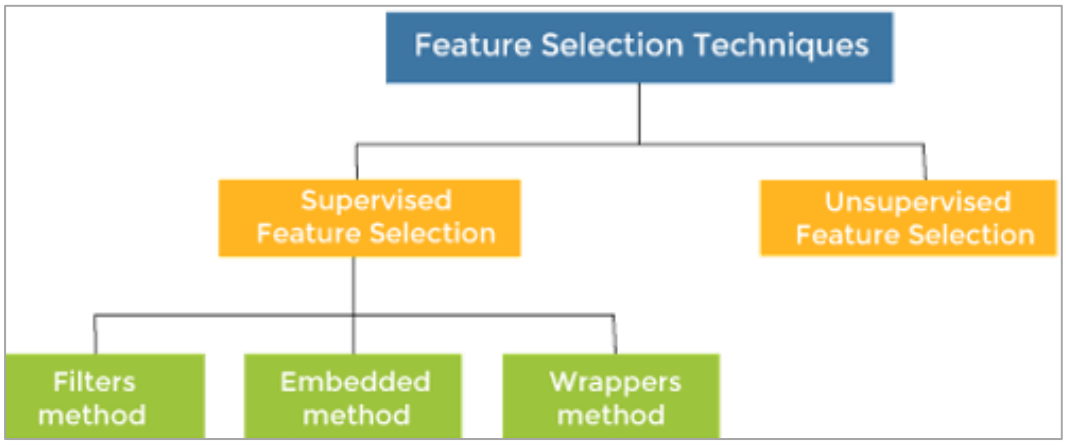


Figure 2.3: Feature selection techniques

The health insurance industry relies heavily on data, housing vast amounts of customer data which often contains redundant and irrelevant attributes. These attributes can have a detrimental impact on the accuracy of health insurance cross-selling predictions. Therefore, it is imperative to employ feature selection techniques before constructing ML models to mitigate the influence of low-impact features on health insurance industry data (Taha, Cosgrave and Mckeever 2022). In training datasets for supervised learning, the selection of features can affect the performance of the ML model. Before training any model huge quantities of data need to be collected in order to facilitate improved ML. However, a considerable portion of the collected data will contain redundant, irrelevant, and noisy information that may not make a substantial contribution to the performance of the health insurance predictive model.

Machine learning techniques are increasingly utilised to improve the effectiveness of insurance datasets which often face challenges related to poor data quality caused by the inclusion of noisy features. Choosing the appropriate features is a crucial pre-processing step in constructing an effective predictive ML model for health insurance cross-selling (Taha, Cosgrave and Mckeever 2022). Applying feature selection for ML

algorithms is important prior to using ML techniques so as to enable the algorithm to concentrate on important features.

There are two categories of feature selection techniques:

Supervised technique – the term "supervised feature selection" refers to a technique used with labelled data to find the pertinent features and boost the performance of supervised models like regression and classification.

Unsupervised technique- unsupervised feature selection is a technique that uses unlabelled data and does not require an output label class in order to choose features.

Three categories of feature selection methods based on the selection strategy exist:

Filter method- techniques for feature selection can be applied regardless of the ML algorithm being employed, relying instead on general properties like variance, consistency, correlation, and information.

Embedded method- approaches that combine the best aspects of filtering and wrapping. Similar to wrapper methods, they often combine feature selection with a predetermined learning process, however unlike the original wrapper methods, they do not iterate the learning algorithm.

Wrapper method- use supervised or unsupervised learning algorithms to identify feature subsets enhancing classification accuracy or clustering quality. Wrapper techniques yield high performance but require significant computation time and may change selected features with different learning algorithms.

In this study, a supervised feature selection technique was applied for feature selection, taking into account the target variable and utilising labelled datasets related to health insurance customers.

2.4.4 The Python programming language

Python has risen to become one of the world's most popular programming languages in recent years. It is open-source and freely available, and is widely employed for tasks such as data analysis and the efficient and rapid construction of ML models (Udawant and Srinath 2019). This study used the Python programming language to build the

health insurance cross-selling prediction model. Python programming language was used for data preparation, building and training models, and deployment and prediction.

2.5 Review of related works

Following a review of the literature on health insurance cross-selling prediction, all the employed methods and selected procedures are conceptually detailed below.

2.5.1 Review of the literature on health insurance cross-selling prediction modelling

The literature revealed that the cross-selling of health insurance products is a major task for the insurance industry to be able to target potential customers who will be interested in purchasing additional health insurance products to increase business revenue and gain customer retention. It is widely acknowledged that retaining existing customers is more cost-effective than acquiring new ones (Bellani 2019).

In order to do health insurance cross-selling predictions using ML algorithms, building a prediction model is essential.

A survey of previous studies on the methods and processes for health insurance cross-selling predictions shows that the most commonly used technique is logistic regression (Wang 2023). Focusing on the insurance sector, previous research has shown that a thorough theoretical foundation of ML models and algorithms is necessary to comprehend the process of health insurance cross-selling modelling.

2.5.2 Machine learning model for cross-selling prediction

Machine learning, at its core, represents the practical manifestation of artificial intelligence, encompassing a wide array of techniques and processes aimed at empowering machines to autonomously acquire the skills required to tackle specific challenges.

Supervised and unsupervised techniques are the two main categories for algorithms in machine learning. Supervised learning focuses on known variables, while unsupervised learning does not rely on such predefined information (Ozdemir and Bayrakli, 2022).

In ML one of the most common tasks for supervised learning is classification, where the goal is to determine the category to which an object belongs. Predictive models for health insurance fall under the category of supervised classification models. They utilise historical health insurance data and ML techniques to leverage customer information to predict whether existing customers are likely to be interested in purchasing additional health insurance products. In this case, health insurance customer data has an unbalanced target.

To ensure the effectiveness of a ML model designed for cross-selling predictions, it is imperative that the model possesses strong generalisation capabilities when it comes to classification tasks. Evaluating the model's generalisation ability and performance commonly involves employing established techniques such as data splitting and K-fold cross-validation.

Data splitting involves partitioning the available dataset into two distinct sets: the training set and the test set. The training set serves the purpose of constructing the model, while the test set is employed to assess the model's generalisation ability.

In K cross-validation the performance of the cross-selling predictive ML model will be evaluated using the dataset for building a more generalised model.

2.5.3 Features' pre-processing and selection

A ML predictive model leverages available data to categorise objects, with this data representing a collection of object features that are pertinent and valuable for the classification task.

Feature selection in ML algorithms is influenced by various factors, encompassing accuracy, completeness, consistency, timeliness, and interpretability. Consequently, it becomes imperative to regard data pre-processing as an integral component of the feature selection process.

The core steps involved in data pre-processing are as follows:

- Data cleaning – involves the procedure of identifying and addressing missing values while rectifying inconsistencies in the dataset.

- Data transformation – encompasses the process of extracting new features from the existing data or modifying the data in a way that enhances its usefulness.
- Data reduction – pertains to the process of identifying and eliminating redundancy within the dataset and conducting relevance analysis to determine which features are most significant for the task at hand.

The primary objectives of feature selection are to enhance model performance, increase computational speed, reduce the cost of predictors, and offer improved insights into the underlying processes responsible for generating health insurance customer data.

2.5.4 Supervised predictive models

Logistic Regression, Random Forest, and Decision Tree models have frequently been employed in prior research for cross-selling predictions. This is due to their common utilisation as binary classification algorithms. These models have been extensively investigated, and this study provides a comprehensive theoretical review of their application in cross-selling prediction.

- **Logistic Regression**

  This statistical model is frequently employed for classification and predictive analytics, as it is designed to estimate the probability of an event taking place. In this case Logistic Regression will predict if the existing health insurance customer will be interested to purchase additional insurance products. In the context of a provided health insurance dataset with independent variables, this model is used to predict the probability of an event occurring, which results in the dependent variable being constrained within the range of 0 to 1.
- **Random Forest**

  Random Forest is an ensemble method comprising multiple tree predictors, where each tree provides a classification. The forest then makes its decision by considering all these individual votes. This characteristic grants Random Forests the capability to rectify the tendency of decision trees to overfit.

- **Decision Tree**

  Decision Trees have gained significant popularity for tackling classification tasks due to their user-friendliness and interpretability. This model can be depicted as a collection of if-then rules, and the classification of instances is carried out by traversing the tree, starting from the root and progressing to the leaf nodes. The leaves signify the class labels, while the branches symbolise combinations of features that guide the assignment of those class labels.

2.5.5 Health insurance cross-selling predictive data analysis process

Adapting cross-selling prediction with machine learning models offers a business the easiest way to increase revenue by leveraging its present portfolio of health insurance customers. At the lowest cost of customer acquisition, cross-selling tactics can boost profitability and increase customer retention. Delivering the appropriate cross-sell product to the appropriate policyholder at the appropriate time and pricing presents a difficulty for the insurer.

Figure 2.4 shows an overview of the health insurance cross-selling data analysis process.

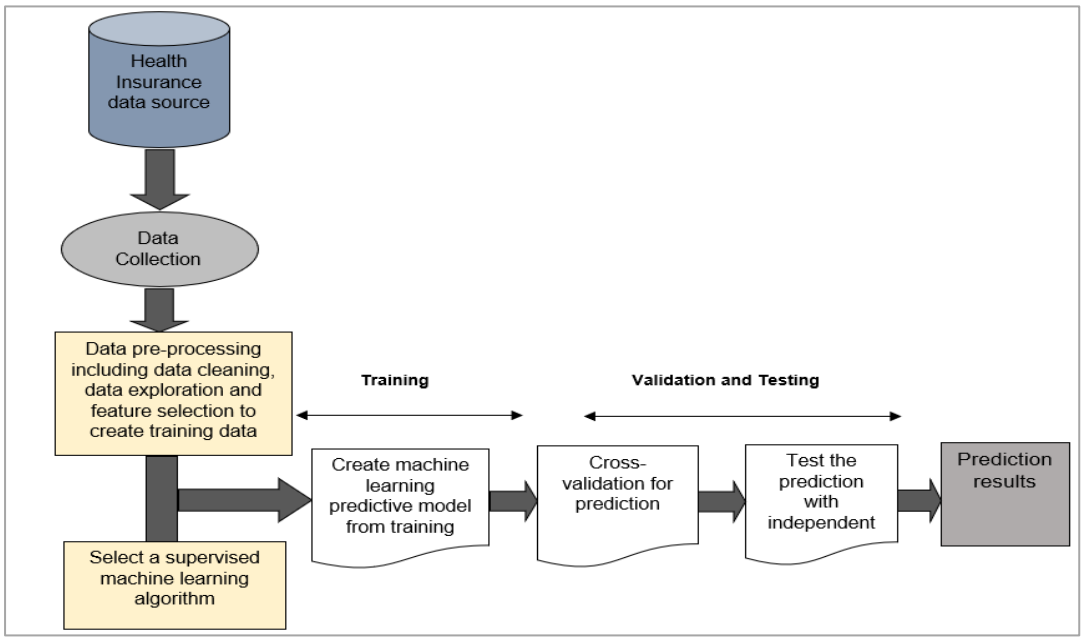


Figure 2.4: Health insurance cross-selling data analysis

Health insurance data source

Previous studies indicate that there is no common ground in financial data disclosure across financial institutions because many of them do not disclose financial data of insurance companies. Most previous researchers indicated that data has been generated from surveys, interviews, and experiments, specially designed for understanding and solving the research problem. These data collection methods have been regarded as insurance data sources.

In this study, the health insurance data source will be the existing health insurance customer's data from a large insurance company data source where all health insurance customer data is restored.

In the realm of ML, the quality of the data utilised for training predictive models holds as much importance as its quantity. It is crucial that datasets are both representative and balanced to yield improved predictive outcomes. This is important to train predictive models. Generally, insurance companies struggle to provide relevant data for training ML models.

Data collection

Data collection is the process of gathering raw data collected at the source and analysing specific information on variables for the purpose of utilising that data to provide a solution to relevant questions and evaluate results. Irrespective of the field of study or definition of data qualitative or quantitative, precise data collection is vital for preserving the integrity of research.

Data collection for insurance companies enables the creation of a single customer view, comprising a comprehensive data profile of each customer, along with their interactions with the company based on the main existing purchased insurance product. This can support insurance companies to strategise the process of identifying the potential customers who might be keen to purchase additional health insurance products.

Collected data can be used for hypothesis testing to eliminate the researcher's assumptions for predictions about future probabilities and trends of health insurance cross-selling.

In this study, the data was collected from health insurance data sources that were used for training and testing the health insurance ML predictive model. This data was de-identified/anonymised due to the risk of data leaks and security breaches. The researcher complied with company data ethics which is aligned with POPIA (Protection of Personal Information Act).

Data pre-processing, as described previously in the context of feature selection, pre-processing aims to enhance the model's performance, expedite computational speed, and enhance the cost-effectiveness of predictor variables.

Data pre-processing is a fundamental technique for preparing data, involving cleaning and organising, to transform raw data into a refined dataset suitable for training ML predictive models. This process plays a pivotal role in ML because real-world data is often incomplete, inconsistent, and inaccurate, containing noise and missing values.

There are seven key steps in the data pre-processing process when constructing ML predictive models, which should be diligently followed:

- **Data acquisition**
  Acquiring the dataset is the initial step in data pre-processing when developing a health insurance ML model. To build and train ML models, the first task involves obtaining the relevant dataset. This dataset typically comprises data collected from various disparate sources, which are then amalgamated into a consistent format to create a comprehensive dataset.

- **Library importation**
  Given that the Python programming language is commonly used for constructing ML models, it is crucial to import the appropriate Python libraries for data pre-processing. The most frequently utilised Python libraries for this purpose include:

  NumPy: NumPy serves as the fundamental package for scientific calculations, enabling the execution of various mathematical operations.

  Pandas: This open-source Python library specialises in data manipulation and analysis, facilitating dataset importation and management.

Matplotlib: Matplotlib, a Python 2D plotting library, is instrumental in generating various types of data visualisations.

- **Dataset importation**

  Importing the dataset is a pivotal step in data pre-processing for ML. In this step dependent and independent variables are extracted, differentiating between the independent variables (referred to as the matrix of features) and the dependent variables within the dataset.

- **Handling missing values**

  In the data pre-processing phase, it is critical to identify and appropriately address any missing values, as they can result in inaccurate predictions and flawed conclusions.

- **Categorical data transformation to numerical format**

  Categorical data consists of information categorised into specific classes within the dataset. Since ML models generally rely on mathematical equations, it is necessary to convert categorical data into numerical formats. This ensures that the ML model operates effectively, as it requires numerical data for computation.

- **Dataset splitting**

  Every dataset used for constructing a ML model during the data pre-processing stage must be divided into two distinct sets: training and testing sets. The training set serves as the subset used to train the ML model, while the test set is employed to evaluate the model's performance.

- **Feature selection and scaling**

  Feature selection is a technique employed in data pre-processing to standardise the independent variables of a dataset, typically within a specific range, ensuring that they are conducive to the ML process.

Feature selection in machine learning algorithms is also explained previously as this is very important to apply before starting the actual building of the ML model under the supervised feature selection method and unsupervised feature selection method. This

will be applied prior to building a health insurance cross-selling predictive model to ensure that the performance of the model will be accurate.

This process of feature selection will be used to identify and remove unneeded, irrelevant, and redundant attributes from the health insurance dataset, removing unneeded variables will assist in creating an accurate health insurance cross-selling predictive model. Feature selection methods will be used in data pre-processing to achieve efficient data reduction. Feature selection is focused on optimising the construction of a ML predictive model by maximising the importance of features while minimising any redundant ones.

The following are the feature selection techniques:

- Enhanced generalisation is achieved by reducing overfitting, which can be formally described as a reduction in variance.
- Simplifying models is aimed at making them more accessible and interpretable for researchers.
- Accelerated training times.
- Mitigating the challenges posed by high dimensionality.

The ML predictive model was built and trained based on the data collected, prepared data, features selected, and algorithm selected.

The health insurance cross-selling prediction results are shown as a binary where the interested customer will be marked as 1 and not interested as 0.

Building this model resulted in the proposal of a new framework for health insurance cross-selling prediction with machine learning and in South Africa will be called ***Cross-Selling Health Insurance Framework*** (CHIF). This framework will be defined along with the research process.

2.6 Summary

This study focuses on cross-selling in the health insurance industry which has been identified as a very effective strategy with a positive impact on revenues and profitability for insurance companies and involves cross-selling additional health products or services to existing customers; this has a direct effect on both eliminating customer churn rates and increasing business profits (Sidorowicz, Peres and Li 2022).

The objective of this study was to forecast health insurance policyholders who are more likely to show interest in acquiring additional health insurance products. This information can aid the company's marketing department in bolstering its cross-selling strategies.

In the health insurance industry, the fundamental basis of cross-selling is to offer more products while the initial purchase transaction is not completed, and this is a technique being used to increase a company's profit and at the same time decrease customer churn by selling more complementary products in a single transaction.

Within the health insurance sector, predictive analytics encompasses a range of statistical and analytical methods employed to create models that forecast future events or anticipate customer behaviour s and purchasing trends.

Machine learning models are essential in health insurance cross-selling prediction to be able to get meaningful insights after data analysis to increase sales through customer purchasing patterns and reduce the rate at which customers want to switch to another company's insurance products.”

CHAPTER 3: MACHINE LEARNING FOR INSURANCE CROSS-SELLING PREDICTION

3.1 Introduction

The prediction of cross-selling in health insurance employs ML techniques method to predict whether an existing health insurance customer will purchase an additional insurance policy (Sidorowicz, Peres and Li 2022). This is crucial for insurance firms because it helps them spot chances for prospective revenue growth and customer expansion (Haag *et al.* 2022).

3.2 Machine learning techniques

Machine learning falls under the category of artificial intelligence, utilizing algorithms and statistical methods to allow computer systems to glean insights from data, make forecasts or determinations, and enhance their proficiency in a particular task using accumulated datasets (Kumar 2019). In other words, ML involves training computer systems to learn from data, rather than relying on explicit instructions or rules (Sekeroglu 2021).

ML can be applied to the prediction of health insurance cross-selling to create models that analyse historical data on customer demographics, past purchase behaviour, and other factors to predict whether the existing customer will be interested to purchase an additional health insurance product (Joshi and Nair 2018). For example, ML can be used in the processing and analysing of human languages such as language translation, sentiment analysis, voice recognition, and chatbots using the natural language processing (NLP) technique which is a branch of ML (Brunila *et al.* 2021).

Another example where ML is being used is fraud detection by analysing transaction data and detecting fraudulent activities (Thennakoon, Miharanga and Kuruwitaarachchi 2019).

There are three types of ML namely supervised learning, unsupervised learning, and reinforcement learning (Gottlieb *et al.* 2022). These are described below

3.2.1 Supervised learning

In supervised learning, the model is trained on a labelled dataset where each input is associated with a corresponding output (Zhou and Li 2017). The objective is to understand the correlation between input and output, allowing the model to offer precise forecasts for unfamiliar data (Nasteski 2017). An example of supervised learning is medical diagnosis where the extracted patient dataset will be labelled with patients' corresponding diagnoses to build a supervised learning model to predict the diagnosis of new, unseen patients based on their symptoms and medical history (Richens, Lee and Johri 2020).

3.2.2 Unsupervised learning

In unsupervised learning, the model is trained using an unlabelled dataset, where no associated output is provided for each input. The goal is to learn patterns or structures in the data, through techniques such as clustering or dimensionality reduction (Carcillo *et al*. 2021).

An example of unsupervised learning is clustering where similar customer data points are grouped together based on their characteristics using a customer dataset with purchase history to identify other groups of customers with similar purchasing patterns by using K-means algorithms (Anitha and Patil, 2019). K-mean algorithm is described in section 3.3.4.2.

3.2.3 Reinforcement learning

In reinforcement learning, the model adapts its decision-making based on environmental feedback. The objective is to optimize a reward signal, like a score or profit, by selecting actions that result in favorable results (Francois-Lavet *et al.* 2018).

An example of reinforcement learning is autonomous vehicles whereby the model is trained to make a decision in self-driving cars to detect when to change lanes, accelerate or brake, and navigate intersections (Zade and Mansouri 2021).

3.3 Machine learning in health insurance cross-selling predictions

Machine learning can be used to develop a predictive model to analyse health insurance historical data on customer demographics, past purchase behaviour, and

other factors to predict whether an existing customer is likely to purchase an additional health insurance product (Perera 2022).

For example, a ML model can be used in customer segmentation based on characteristics, such as age, gender, income, location, and previous purchase behaviour to identify customers who are most likely to be interested in cross-selling (Jansen 2018).

In the context of this study, the literature informs that ML model can be used as a recommender system which can recommend additional health insurance products to customers based on their past purchase trends and preferences (Portugal, Alencar and Cowan 2018).

Supervised learning is applied in health insurance cross-selling predictions by analysing the historical data on customer demographics, past purchase behaviour, and other factors as input to an ML algorithm (Amani and Fadlalla 2017). The health insurance historical customer data will be labelled with information on whether the existing customer will be interested to purchase an additional insurance product, providing a target variable for the algorithm to predict (Taylor *et al.* 2022). By analysing historical data with ML algorithms, the insurance company can identify patterns in customer behaviour and make data-driven decisions on how to target cross-selling efforts for increasing business revenue and improving customer retention (Mahesh 2018).

In the current study a supervised learning model mapped between a set of input variables ***X*** and an output variable ***Y*** for health insurance data that was used to train a model to accurately predict and provide an indication of whether a customer would be likely to buy an additional insurance product. To achieve this the health insurance dataset was divided into a training set and test (validation) set (Ambika 2020). The training set was used to train the model, while the test set was used to evaluate the performance of the model.

An example of how supervised learning can be used to predict a continuous value, is the price of a house based on its features and its location, size, number of bedrooms, and bathrooms. In this example a Random Forest model can be used to train data and

predict new unseen house prices using input variables to output predicted house values and the true sale prices (Ihre and Engstrom 2019).

The process of applying supervised learning to the health insurance dataset involved the following steps:

1. **Data collection** - collecting historical data on existing health insurance customers, including demographics, medical history, and previous insurance policies.
2. **Pre-processing the data** - cleaning it up, dealing with null and missing values, encoding categorical variables, and scaling numerical variables.
3. **Feature selection** - selecting the most relevant features from the existing health insurance data for the prediction task to reduce noise and improve model performance.
4. **Model selection** - choosing a suitable ML algorithm for the health insurance cross-selling prediction, such as Random Forest, KNN, XGBoost, or Logistic Regression.
5. **Model training** - using the training set of labelled data to train the chosen model.
6. **Model evaluation** - measuring the model's performance on the test set of labelled data using metrics such as accuracy, precision, recall, and F1 score.
7. **Model deployment** -predicting outcomes on fresh data using the trained model such as current customer data, to identify cross-selling opportunities.

3.3.1 Factors to consider when selecting a machine learning model

1. Task type - the type of task the model will perform, such as classification or regression.
2. Dataset size - the size of the dataset can impact which models are most effective. Some models may require more data to achieve good performance.
3. Dataset complexity - the complexity of the dataset, including the quantity of noise and characteristics present in the data, can impact model performance.
4. Model complexity - Performance can be impacted by a model's complexity, which includes its type and amount of parameters.

5. Hyperparameters - the hyperparameters of the model, such as learning rate, regularisation strength, and many hidden layers, can be tuned to optimize model performance.

3.3.2 Common machine learning model selection techniques

1. Cross-validation - process entails dividing the insurance data into testing and training sets, then assessing the model's output for each iteration. This method guarantees the model's generalisability and lessens overfitting.
2. Grid search - this involves systematically testing different hyperparameter combination to find the best set of hyperparameters for the given task.
3. Random search - this involves randomly selecting hyperparameters to test, rather than systematically testing all possible combinations.

Cross-validation was applied in this study so the insurance dataset was separated into training and testing sets to gauge a model's generalisation performance on new data. This method was selected for its ability to mitigate overfitting, a condition that arises when a model is overly complex and closely fits the training data, leading to suboptimal generalization performance. Cross-validation was utilized to obtain a more precise estimate of the model's performance (Raschka, 2018). As an illustration, envision a situation where cross-validation is utilised to construct a predictive model for a binary classification problem. This challenge involves employing neural networks to anticipate the probability of a credit card application getting approved, taking into account diverse attributes like age, income, and credit score (Markova et al., 2021).

3.3.3 Machine-learning algorithms

For the purpose of health insurance cross-selling prediction the following ML algorithms were selected, trained, and their performance compared to choose the best and most accurate algorithm.

3.3.3.1 Random Forest algorithm

This algorithm works by randomly selecting a subset of the features (variables) from the insurance dataset for each decision tree and splitting the data based on the selected features. The result is a collection of decision trees that make independent

predictions, which are combined to form the final prediction of whether a customer will accept a cross-selling offer or not.

For example, the random forest can be used to predict credit card application approval based on the customer's existing dataset using features such as income, age, credit score, and employment status (Choubey *et al.* 2020). Random forest can also be used for credit card fraud detection by analysing the training dataset to detect fraudulent transactions (Jonnalagadda, Gupta and Sen 2019).

3.3.3.2 KNN algorithm

The KNN model was trained on the training set by assigning each data point to the class of its k-nearest neighbours (Ahmed. 2019). The value of k can be selected through cross-validation to predict the possibility of cross-selling additional insurance products to the existing customer.

For example, suppose there is a dataset of customer purchase history in a retail store, along with their corresponding features such as gender, age, and purchase history, KNN can be used to predict whether the customer will be interested to make a purchase based on their features by analysing the historical dataset (Lico and Enesi 2021).

3.3.3.3 XGBoost algorithm

This model is built iteratively, with each new tree attempting to correct the mistakes of the previous tree. The hyperparameters of the model, Common machine learning approaches like grid search and random search can be used to adjust parameters like learning rate, maximum tree depth, and number of trees. XGBoost is widely used for classification problems and to handle missing values without imputation pre-processing for health insurance cross-selling using insurance customers' historical data (Rusdah and Murfi 2020).

For example, the XGBoost is commonly used for building ML models for recommendation systems by training insurance datasets to identify customer purchasing patterns and recommend them for insurance cross-selling additional products (Shahbazi and Byun 2019).

3.3.3.4 Logistic Regression algorithm

Logistic Regression, using the dependent variables age, gender, previous insured, monthly income, and vintage, can predict the health insurance customer likelihood of accepting the cross-selling offer. The logistic regression model can calculate the probability of the customer accepting the offer using a sigmoid function where the result is presented as 0 and 1, zero means the customer is not interested and 1 means interested.

After evaluating the performance of the models, the best and most accurate model was selected and applied to health insurance cross-selling prediction. The most used prediction model was Logistics Regression and, in this study, was considered depending on its performance.

The Logistics Regression model was trained on historical data that included features such as customer demographics, previous purchase history, and other relevant factors that may impact a customer's decision to purchase an additional health insurance product.

The algorithm calculated the probability of the existing customer purchasing the health insurance product based on the input features. This probability was then used as a binary classification algorithm to predict whether an existing customer would be likely to purchase an additional health insurance product or not.

For example, if a university wants to predict whether a student will be admitted to their programme based on their scores in two entrance exams (Exam 1 and Exam 2). The university has historical data on previous students who applied and whether they were admitted or not. The university can use a logistic regression model to identify the probability of students being admitted based on their exam scores (Helal *et al.* 2018).

3.3.4 Model evaluation for health insurance cross-selling prediction using machine learning

After training, the model's performance was evaluated using the testing set. Metrics such as **accuracy**, **precision**, **recall**, and **F1-score** were used to assess the model's performance. If a model performs well it can be deployed in a real-world setting to

forecast whether a customer is inclined to accept a cross-selling offer using their demographic and insurance plan information.

1. Accuracy- the proportion of correctly classified samples to the total number of samples in the dataset. This is a good metric for balanced datasets.
2. Precision- the ratio of true positives to the total predicted positives. This assesses the model's capability to accurately recognize positive cases.
3. Recall- the ratio of true positives to the overall number of real positives. This assesses the model's capability to detect all positive cases.
4. F1-Score- the balanced average of precision and recall. This provides a balanced measure between precision and recall.

Examples of accuracy, precision, and recall can be used if a company wants to classify spam emails using a dataset of 100 emails, where 60 are legitimate emails and 40 are spam emails. Will train a binary classification model such as logistic regression or a decision tree to classify new emails as either legitimate or spam based on their features extracted from the email content, sender information, and metadata. (Ahmed, Hameed and Bawany 2022).

After training, the model will evaluate its performance on a test set of 20 emails, where 10 are legitimate and 10 are spam.

The confusion matrix for the model's predictions for example shown in Table 3.1.

Table 3.1: Confusion matrix

	Predicted Legitimate Emails	Predicted Spam Emails
Legitimate Emails	8	2
Spam Emails	1	9

From the confusion matrix, we can calculate the following metrics:

1. Accuracy: The percentage of emails that were correctly classified out of all emails.

   Accuracy = (true positives + true negatives) / total emails = (8 + 9) / 20 = 0.85
2. Precision: Out of all emails anticipated to be spam, the percentage of emails that were actually classed as spam.

   Precision = true positives / (true positives + false positives) = 9 / (9 + 2) = 0.82
3. Recall (also known as sensitivity): The percentage of spam emails that were accurately categorised out of all spam emails.

   Recall = true positives / (true positives + false negatives) = 9 / (9 + 1) = 0.9

An example of the F1-score gives a balanced measurement of both metrics and the precision and recall harmonic mean is found.

The formula for the F1-score is as follows:

$$F1 - score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.1)$$

Since the precision in this instance is 0.82 and the recall is 0.9, the F1 score is:

$$F1 - score = \frac{2 \times 0.82 \times 0.9}{0.82 + 0.9} = 0.86$$

3.3.5 Evaluation metric for machine learning that assesses a model's performance

The predictive effectiveness of a model on a dataset is represented by a confusion matrix. A confusion matrix for a binary class dataset with, let's say, "positive" and "negative" classes includes four key elements:

1. True Positives (TP): The quantity of samples that were accurately classified as positive
2. False Positives (FP) are samples that were incorrectly classified as positive.
3. True Negatives (TN): The number of samples that were accurately classified as negative.
4. False Negatives (FN) are samples that were incorrectly classified as negative.

TP and TN

TP and TN are critical measures for evaluating the effectiveness of a classification model, particularly in binary classification problems. Along with other essential measurements like false positives and false negatives, these metrics are calculated (Chicco and Jurman 2020).

True positives

In a confusion matrix, a situation known as a TP occurs when a sample is correctly classified as belonging to a particular class by the model (positive). In other words, the model properly identified a positive because its prediction and the ground truth label coincide (Kulkarni, Chon and Batarseh 2020).

A TP, for instance, in a medical diagnosis scenario means that a model has correctly identified a patient with a disease, and that patient is indeed positive for that disease (Kennedy *et al.* 2019).

In binary classification tasks, TPs can be used as a performance statistic to assess how well a model can recognise positive cases. In other words, a true positive occurs when both the ground truth label and the model accurately anticipate the positive class (Grandini, Bagli and Visani 2020).

True negatives

In binary classification problems, TN is a performance statistic that measures how well a model can recognise negative cases. A TN is an instance where the model predicts the negative class with accuracy (Diamandis, Prassas, and Diamandis 2020).

For example, in the case of classifying spam emails, a TN will indicate that a model successfully classified a non-spam email and indeed it is true that the email is not spam.

True negative is particularly useful when the cost of a FP is high, whereby a negative case is incorrectly labelled as positive. The use of the TN classification model will ensure that the model correctly identifies all negative cases (Grandini, Bagli and Visani 2020).

False positives and false negatives

In binary classification tasks, FP and FN are significant performance indicators that complement TP and TN (Rottmann, Maag and Chan 2019).

False positives

FP refers to the cases where the model predicts a positive class, but the actual ground truth label is negative. In other words, the model incorrectly identifies a negative case as a positive. For example, in health insurance cross-selling FP are situations in which a model predicts that a consumer is likely to buy health insurance but, in fact, the customer is not interested in the product or does not qualify for it. This could cause the business to invest resources in marketing or sales efforts for a client that won't convert, resulting in lost time and money (Hanif 2019).

False negatives

FN refers to the cases where the model predicts a negative class, but the actual truth label is positive. In other words, the model incorrectly identifies a positive case as a negative.

For example, the scenario where a credit card provider developed a ML model to identify fraudulent transactions based on trends in past data. However, some fraudulent transactions that have fresh patterns or aren't visible in the previous data may go undetected by the model, producing false negatives (Syed *et al.* 2020).

3.3.6 Formulas used for machine learning evaluation

F1-Score

The F1 score is defined based on the precision and recall scores, which are mathematically defined as follows:

$$Precision = \frac{TP}{TP+FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3.3}$$

As illustrated below, the F1 score is obtained by calculating the harmonic mean of the precision and recall scores. A higher F1 score signifies a higher-quality classifier, and it is typically expressed as a percentage ranging from 0% to 100%.

$$\text{F1 } Score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \tag{3.4}$$

$$= \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall}$$

The F1 score, which is equal to the precision and recall scores in the equation above, can alternatively be expressed as follows:

$$F1\ Score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \tag{3.5}$$

F1 score macro-averaged

The F1 scores for individual classes are averaged to determine the model's macro-F1 score. In a dataset comprising "n" classes, this approach is employed, it is mathematically stated as follows:

$$Macro\ F1\ Score = \frac{\sum_{i=1}^{n} F1\ Score_i}{n} \quad (3.6)$$

The macro-averaged F1 score is effective only when the dataset has an equal number of data points in each class. Nonetheless, many real-world datasets exhibit class imbalance, with varying amounts of data in different categories. In such scenarios, a simple average can be a misleading performance metric.

3.4 Summary

This study explores the application of machine learning techniques to predict health insurance cross-selling opportunities. It emphasizes the critical role of these predictions in driving revenue growth and expanding customer bases for insurance companies. Machine learning, a subset of artificial intelligence, enables computers to learn from historical data, make predictions, and improve decision-making without explicit programming. The study categorizes machine learning into supervised, unsupervised, and reinforcement learning. Supervised learning is particularly highlighted for its application in predicting customer behavior based on labeled data, such as demographics and past purchase history. Model evaluation techniques, such as accuracy, precision, recall, and F1-score, are employed to assess the performance of predictive models in cross-selling prediction. Formulas used for machine learning evaluation are utilized to quantify the effectiveness of these models. By leveraging these insights, insurers can segment customers effectively and tailor cross-selling strategies to individual preferences, thereby enhancing customer satisfaction and retention.

CHAPTER 4: RESEARCH METHODOLOGY

4.1 Introduction

In the previous chapter, an in-depth review of the literature on health insurance cross-selling, data-mining techniques, and ML algorithms was presented. In this chapter, the research design, data collection, data analysis, and the validity and reliability of the collected data are discussed.

4.2 Research design

Research design is a method of setting the procedures, methods, and techniques to be applied for collecting and analysing the required data to answer the research questions (Asenahbi 2019; Sileyew 2019).

4.2.1 Research paradigm (philosophy)

The research paradigm is a conceptual and practical tool that is used as a guideline for developing research methodology to solve specific research problems (Kaushik and Walsh 2019). The research paradigm comprises different types of paradigms including positivism, post-positivism, critical theory, interpretivism, and pragmatists (Rapport and Braithwaite 2018).

Postpositivist

The post-positivism approach aims to find objective solutions by identifying and addressing them. This approach challenges the positivist notion that a researcher can remain an impartial observer of the social world by presenting its arguments (Gu *et al.* 2021). Postpositivism aligns with the qualitative research method which uses descriptive qualitative research whereby the research describes the situation as observed in the field. Postpositivism describes the event so that the data to be collected is descript in nature to identify and analyse the effectiveness of flexible work arrangements to improve health insurance cross-selling in the health insurance companies (Rusilowati and Pratiwi 2022).

Critical theory paradigm

Critical theory challenges predictable knowledge bases and methodologies either qualitative or quantitative. Critical theory is concerned with the critical meanings of experiences as they correlate to gender, race, class, and other kinds of social oppression that apply to health insurance customers. This paradigm attempts to discover the social-historical specificity of knowledge and shed light on how particular pieces of knowledge reproduce structural relations of imbalance and oppression (Paynton and Hahn 2021).

Interpretivist paradigm

The interpretivist paradigm is a method to address a problem predicated on the notion that reality and human behaviour are marked by continuous shifts, adaptations, and simultaneous transformations which provide the research with greater scope to address issues of influence and impact and to ask questions such as 'why' and 'how' particular technological trajectories are created. The interpretive approach aims to grasp the world based on individuals' subjective experiences as it genuinely exists. This approach does not predetermine dependent and independent variables but centers on the complete intricacy of human sense-making as the situation unfolds (van der Walt 2020). Interpretivism uses both qualitative and quantitative research methods, however, the perspective is that there is no one right path to knowledge (Nickerson 2022).

Pragmatist paradigm

The pragmatist paradigm refuses to get involved in contentious metaphysical concepts such as truth and reality. Instead, it accepts that there can be single or multiple realities that are open to empirical inquiry (Kaushik and Walsh 2019). Pragmatism is a paradigm that claims to bridge the gap between the scientific method and structuralist orientation of older approaches and the naturalistic methods and freewheeling orientation of newer approaches.

Positivist paradigm

This study adopted the positivism paradigm that is most used for quantitative methodology (Antwi and Hamza 2015).

The positivist paradigm focuses on investigating social reality under the belief that the most effective understanding of human behaviour comes from observable phenomena that can be quantified. This allows for statistical analysis to determine the relationships between variables. This approach prioritizes experimentation, observation, control, measurement, reliability, and validity in the research process for predicting cross-selling in health insurance (Park, Konge and Artino 2020).

Positivism is used to explain and predict what will happen in the future (Bonache and Festing 2020). This research analysed and predicted health insurance customer behaviour to identify whether the customer would be interested in buying a health insurance product.

Positivism aligns with quantitative research. Quantitative research is frequently applied in the health insurance industry and social care research that emphasises quantification in the collection and analysis of data (Jain *et al.* 2018). Quantitative research uses objective measurements with statistical methods, mathematics, and computational modelling to examine trends and patterns which is associated with the positivism paradigm (Bonache and Festing 2020; Goertzen 2017).

The quantitative research method focuses on investigating the answers to the questions starting with how many, how much, and to what extent (Bonache and Festing 2020). For this study, health insurance data was extracted from a large insurance company database, focusing on customer behaviour which could be quantified and patterned to interpret their meanings to create data insights.

4.2.2 Research approach

Inductive approach

The inductive approach is a qualitative research methodology that involves crafting general theories from specific observations. This methodology is defined by the iterative process of observing a particular phenomenon multiple times and inferring that it will manifest similarly in the future (Shemshurenko *et al.* 2018).

The inductive approach consists of three steps: observation, pattern recognition, and theory formulation. The inductive method is employed in analysing quantitative data for model evaluation to ensure a direct correlation between the study's objectives and the concise conclusions derived from data analysis. This method is also applied to

create a framework for the process's fundamental structure, which is visible in the unprocessed data (Acimovic *et al.* 2019). For any studies using the quantitative approach, the inductive approach is regarded as less complicated than other approaches. While the inductive approach may not be as robust as other analytical strategies for model development, it does offer a straightforward method for generating results in the context of evaluation questions.

Deductive approach

The deductive approach commences with a theory, from which hypotheses are formulated followed by the collection and analysis of data to assess the validity of these hypotheses. The deductive research approach comprises four stages, beginning with a current theory, formulating a falsifiable suggestion derived from that perception, gathering information, analysing the data, and then making a determination about whether to reject the null hypothesis (Nisbet, Miner and Kale 2018).

This study adopted a deductive research approach because this is typically suitable for research under positivism. This study primarily involves a large quantity of measurable health insurance data (Moreau, Pichault and Mertens de Wilmars 2014). In this research, an objective numerical analysis of the data was conducted, followed by validation and generalization processes. According to Park, Bahrudin, and Han (2020), the deductive approach is frequently employed in quantitative studies.

Using the deductive approach, the following occurred:

i) The existing health insurance data was extracted from a large insurance company database to be prepared and cleaned.
ii) Based on existing health insurance data a machine learning predictive model (MLPM) was built.
iii) The MLPM was verified iteratively with many quantifiable health insurance data points

4.2.3 Research strategy

There are various types of research strategies including case study, experimental, survey, and action research (Johannesson *et al.* 2014; Walia *et al*. 2021). This study

adopted an experimental research strategy which is a scientific research design with variables that seek to determine a relationship between the dependent variable and independent variable (Harland *et al.* 2019). This strategy was used for health insurance cross-selling prediction focusing on the variables that could be measured, calculated, and compared and linked to quantitative research.

4.3 Data collection

Data collection is the process of gathering relevant data from various sources (Singh and Dhillon 2022). In order to build a model for health insurance cross-selling prediction this study used ML tools for data preparation and modelling.

The health insurance customer's data was collected from a large insurance company in South Africa and the company name was redacted for confidentiality purposes and the data was pre-processed and converted into specific models (see Appendix A). As this data is anonymised, customer privacy will not be compromised.

Data science techniques for data collection

The data collection in this study involved primary data collection and secondary data collection.

Primary data collection methods refer to data collected from first-hand experience directly from the main source, being data that has never been used in the past. This data is considered the best kind of data in research; however, this method is time-consuming. The methods of collecting primary data can be divided into quantitative data collection and qualitative data collection methods. For this method, the researcher asks questions of a large sample of people, either by direct interviews, questionnaires, or communication such as telephone or email.

Secondary data collection methods refer to data that has already been collected and saved. This data collection method is much more inexpensive and less time-consuming compared to the primary data collection methods. In the secondary data collection method, there are no specific collection methods that need to be considered, because the information has already been collected. For this study the researcher extracted data from a large insurance company in South Africa. Data ML techniques were used to extract and clean the health insurance customer data.

There are two major machine learning techniques, namely, descriptive and predictive.

Descriptive techniques

The descriptive technique identifies what happened in the past by analysing health insurance customer data. This method is frequently used to extract new, important information from a data set and to uncover data indicating patterns.

Predictive techniques

Predictive technique describes what can happen in the future with the help of past data analysis. This study adopted a predictive technique to use health insurance customer historical data to make predictions about future outcomes. The predictive technique carries out and inductive process regarding the current and past data so that the predictions can be made.

4.4 Data analysis and building predictive model

In this study, ML was applied to build a prediction model to increase the cross-selling revenue of companies in the health insurance industry.

The following steps were undertaken:

i) **Identify existing ML algorithms by comparing and selecting suitable ML algorithms from the literature and by appropriate identifying criteria.**
   - Research and review existing ML algorithms suitable for tasks like health insurance cross-selling prediction, considering criteria such as performance metrics (accuracy, precision, recall), scalability, interpretability, and dataset suitability.

ii) **Extract health insurance customer data from a large insurance company in South African.**
   - Obtain the dataset containing relevant information about health insurance customers. This data may include demographic information, medical history, previous interactions with the insurance company, and any other relevant variables.

iii) Data cleaning and pre-processing.

- Clean the extracted data to handle missing values, outliers, and inconsistencies. This involves techniques such as imputation, normalization, scaling, and encoding categorical variables.
- Pre-process the data to make it suitable for machine learning algorithms. This includes feature selection, feature engineering, and splitting the data into training and testing sets.

iv) ML model selection.

- Based on the identified algorithms from step (i) and after pre-processing the data, select the most suitable machine learning model(s) for the task.
- Consideration factors may include the nature of the problem (classification, regression), size and complexity of the dataset, computational resources available, and the interpretability of the model.

v) Use Python programming language to train the health insurance customer MLPM for health insurance cross-selling prediction.

- Implement the chosen machine learning model using Python libraries.
- Train the model using the pre-processed data from step (iii). This involves feeding the training data into the model, optimizing model parameters, and validating the model's performance using appropriate evaluation metrics.

vi) Analyse the effectiveness of the ML model for predicting cross-selling in the health insurance industry.

- Evaluate the trained model's performance on the test dataset to assess its effectiveness in predicting cross-selling opportunities.
- Analyse metrics such as accuracy, precision, recall, F1-score to understand how well the model performs compared to expectations and benchmarks.

vii) Build a health insurance customer model for health insurance cross-selling prediction.

- After confirming the model's effectiveness, implement it to forecast cross-selling chances for both new and existing health insurance customers. Continuously monitor and update the model to uphold its accuracy and relevance over time.

The model selection for health insurance cross-selling prediction using ML involved choosing the best ML algorithm and hyperparameters and comparing the model performance as the performance of the model can vary significantly depending on the algorithm, hyperparameters, and environment used such as cloud (Kaggle.com) or on a local machine (Ozdemir and Bayrakli 2021).

4.5 Validity and reliability

Validity pertains to the extent to which research findings accurately and reliably depict the events studied. This differs from the objective numeric and statistical assessments employed to validate and verify quantitative data (FitzPatrick 2019). In the realm of quantitative research, the concept of validity can assume various connotations, encompassing attributes like rigour, trustworthiness, suitability, and even overall quality. The description and characterisation of validity can take on a multitude of forms (Hayashi, Abib and Hoppen 2019). Validity is the extent to which a measurement provides an accurate and correct response (McDonald, Schoenebeck and Forte 2019).

Reliability refers to the degree to which results can be consistently reproduced, such as between different tests or when evaluated by different observers. In the context of quantitative research, reliability implies a sense of constancy and repeatability over time, rather than focusing on accuracy (Bhattacherjee 2012; Dong *et al.* 2023). Reliability is the extent to which a measurement consistently yields the same or consistent results when repeated multiple times (McDonald, Schoenebeck and Forte 2019).

In this research study data validity and reliability were maintained throughout by employing methodological approaches such as data validation and verification. These measures included implementing comprehensive checks and validations throughout the phases of data collection and analysis. These processes were instrumental in identifying potential errors and discrepancies, thereby ensuring the accuracy and integrity of the study's findings.

4.6 Summary

This study adopted the positivism paradigm, deductive approach, and experimental research strategy.

For the data collection, data science techniques were used to extract South African health insurance customer data from a large insurance company. The data were then pre-processed (cleaned).

For the data analysis, ML was applied to build a prediction model to increase the cross-selling revenue of companies in the health insurance industry.

Data validity and reliability were maintained throughout the study.

The next chapter presents the health insurance cross-selling predictive model and the empirical findings of the analysed data.

CHAPTER 5: FINDINGS AND RESULTS

5.1 Introduction

This chapter presents the findings which are the answers to the study's questions, and is guided by the data analysis steps outlined in Chapter 4 section 4.4. This chapter presents the ML algorithm identification process, the model building and training process, the deployment process, and the results and discussion of the deployment process.

5.2 Identification of ML algorithms for a health insurance prediction model

In order to identify ML algorithms, the business problem was understood which is the first stage in ML Lifecycle. The ML algorithms were identified by comparing and selecting the most suitable algorithms from the literature and by appropriately identifying the criteria based on the research questions and objectives. The results and findings are presented in Section 5.9 outlining the performance of each ML algorithm after training and testing. This involved understanding the problem and the types of tasks the study wanted to address, the dataset size, structure, and any unique properties such as missing values, and duplicates to consider whether the data required pre-processing steps before applying ML algorithms (Dhieb *et al*. 2019).

5.3 Health insurance dataset extraction

In this section, the process of data collection was guided by ML technique **explained** in section **4.3** which is the descriptive technique. The deductive research approach was used to process the extraction of current and historical health insurance data from the database. The health insurance data were collected using Structure Query Language (SQL) by selecting the required fields for this study. The extracted dataset was anonymised thereby ensuring that customer privacy would not be compromised and maintaining confidentiality in order to comply with data ethics and the POPI Act. The extracted dataset contained information about demographics and policy details for existing health insurance customers.

The target variable, which is also referred to as the dependent variable or response variable, is the variable in a ML model that is being predicted based on the input features (Ray 2019).

5.3.1 Health insurance dataset overview and description

The collected health insurance dataset from the insurance database contained information related to insurance policies, members, claims, and other relevant variables which was in line with the recommendations of Rawat *et al.* (2021).

The extracted health insurance dataset was a comprehensive collection of anonymised and aggregated information related to health insurance claims and policyholders. It encompassed a wide range of demographics and policy details (Table 5.1). The selected dataset contained one million records, with 17 features, reflecting diverse individuals and their respective insurance policy information.

This dataset was cleaned and split into two parts the validation set and the training set, the 'response' variable was excluded from the split, as it served as the target variable.

Table 5.1: Health insurance dataset table

No#	Variable	Definition	Data Type
1	ID	Unique ID for consumer e.g., 1, 2,3, 4, anonymous data	Integer
2	Gender	Gender of the consumer 0= Other, 1=Male, 2=Female	Integer
3	Age	Age of the consumer	Integer
4	RegionCode	Unique code for the region of the customer	Float
5	RaceCode	Different race (Unknown, White, Black, Indian, Coloured)1=Unknown, 2=White, 3=African, 4=Coloured, 5=Asian	Integer
6	PreviouslyInsured	1: Customer already has Health Insurance, 0: Customer does not have Health Insurance	Integer
7	InitialSumAssured	An original fixed amount that will be paid to the nominee	Float
8	CurrentSumAssured	A current fixed amount that will be paid to the nominee	Float
9	MonthlyIncome	Current consumer monthly income	Float
10	MonthlyPremium	Amount consumer needs to pay every month for Health Insurance	Float
11	AnnualPremium	Amount consumer needs to pay as a premium in the year for Health Insurance	Float
12	Vintage	Number of days, the consumer has been associated with the company	Integer
13	InsurerType	1= Consumer uses Internal Health Insurance product, 0=Customer uses external Health Insurance product	Integer
14	PolicyStatusType	1= Active, 0=Inactive	
15	ProductTypeType	1=Consumer has comprehensive health insurance cover, 2= consumer has accident only cover, 3= consumer has standard cover	Integer
16	InsuranceCondition	1= Health Insurance condition is Compulsory, 0=health insurance condition is optional	Integer
17	Response	1: Consumer is perceived to be interested, 0: Customer is perceived not to be interested	Integer

Table 5.2 displays the variables employed for model training and testing, omitting the features utilized for the submission dataset after the completion of model training and evaluation. These variables are indicative of the probability of customers acquiring supplementary health insurance products.

Table 5.2: Health insurance submission dataset table

Variable	Definition	Data Type
ID	Unique ID for the customer	Integer
Response	1: Customer is perceived to be interested 0: Customer is perceived not to be interested	Integer

The extracted health insurance dataset was split into training and validation sets. The dataset shape for both training and validation or test sets are shown in Table 5.3. After the data extraction process was completed it was imported using Python import libraries called pandas and during the process the data description and the number of rows and columns were identified using Python code, and the matching of the extracted dataset and data shape was confirmed.

Table 5.3: Training and validation dataset shape

	Training Dataset Shape	Validation Dataset Shape
Number of Rows	1 000 000	1 000 000
Number of Columns	17	16

5.4 Health insurance data preparation (cleaning and pre-processing)

Misra and Yadav (2019) explain that data cleaning and pre-processing are crucial steps in preparing data for ML. These steps involve converting raw data into a refined, uniform, and suitable format that can be effectively employed by ML algorithms. Data preprocessing process was applied to the health insurance extracted dataset for data cleaning and pre-processing purposes.

The dataset consisted of the following demographic information and policy information: Demographic fields (*Gender, Age, RegionCode, RaceCode*), and Policy fields: *PreviouslyInsured, InitialSumAssured, CurrentSumAssured, MonthlyPremium, MonthlyIncome, AnnualPremium, InsurerType, Vintage, ProductType, PolicyStatusKey, InsuranceCondition* (Table 5.4). These features were all selected.

A selection of the top 10 rows of the health insurance training dataset along with the associated 13 headings as well as the dataset data type is provided in Table 5.4.

Table 5.4: Training dataset

ID	Gender	Age	Region Code	Race Code	Previously Insured	Initial SumAssured	Current SumAssured	Monthly Income	Monthly Premium	Annual Premium	Vintage	Resp
int64	int64	int64	int64	int64	int64	float64	float64	float64	float64	float64	int64	int64
1	2	62	4093	2	1	195332.4	138986.52	0	367.23	4244.143	8	0
2	2	54	2162	3	1	575058.65	563762.120	23994	0	0	10	0
3	1	44	2302	2	1	542714.65	729075.35	51878.45	992.17	11357.52	3	0
4	1	41	3610	3	1	2600187.1	2544079.21	0	479.25	5485.024	9	0
5	2	60	4037	5	1	252602.74	136684.35	0	430.22	5011.394	4	0
6	1	53	1813	2	1	1457869.95	1000000	0	1271.11	14712.9	1	0
7	1	43	163	2	1	1109078.81	759778.59	24500	707.04	8189.997	10	0
8	1	45	182	3	1	445814.88	445814.88	0	557.3	6439.693	15	0
9	2	54	7490	4	1	361963.09	290403.15	0	376.14	4326.319	8	0
10	2	56	2190	3	1	301754.35	252398.73	18840.25	618.49	7106.655	9	0

5.4.1 Data cleaning

The data cleaning process was conducted using Python programming language and libraries for pre-processing. This involved removing null values, duplicate values, and missing values and handling noise data to improve the dataset's quality and reliability by ensuring the data was consistent, complete, and appropriate for training the model.

5.4.2 Data pre-processing

Features from the chosen dataset were selected as inputs, and inconsistent data was eliminated to make the data ready for analysis and ML algorithms. An initial exploration of the dataset in the data pre-processing step was conducted to gain insights into the distribution, correlations, and data visualisation. A feature engineering process was performed to ensure that all the feature data types were correct and ready for training a model.

Unique values were checked from the health insurance dataset, and the sample size selected.

After performing the aforementioned processes, the selected health insurance dataset records were eliminated and dropped, meaning that the dataset shape had changed as illustrated in Table 5.6.

Table 5.5: Dataset shape after cleaning and pre-processing

Dataset Shape after Cleaning and Preprocessing	
Number of Rows	713538
Number of Columns	17

Post-processing the dataset containing 713538, it was split into 80% (570830 entries) for training and 20% (142708 entries) for testing or validation sets.

5.4.2.1 Data exploration

The data exploration process was followed within the data pre-processing process. Wilke (2019) states that data exploration is a crucial process in data analysis which involves thoroughly examining and understanding the extracted dataset at hand. It goes beyond simply looking at raw data by applying various techniques to gain insights, identify patterns, and detect relationships within the data. During data exploration, the researcher delved into the structure of the extracted health insurance dataset, evaluating factors such as the number of rows, columns, and data types of variables present, their descriptive, distributions, correlations, and data visualisation.

5.4.2.1.1 Data descriptive and distribution

Data distribution refers to the distribution of the input data that is used to train and evaluate the ML models (Doupe, Faghmous and Basu 2019). This step plays a crucial role in determining the performance and generalisation capabilities of the models.

Summary statistics are a concise way to describe the key characteristics of a dataset.

Table 5.6: Summary statistics

No#	Feature names	count	unique	top	freq
1	ID	1 000 000	1 000 000	1	1
2	Gender	1 000 000	3	1	522 163
3	Age	1 000 000	63	40	43 199
4	RegionCode	1 000 000	1 798	0	25 804
5	RaceCode	1 000 000	6	3	501 164
6	PreviouslyInsured	1 000 000	2	1	998 348
7	InitialSumAssured	1 000 000	74 490	424 963	230
8	CurrentSumAssured	1 000 000	72 887	1 500 000	25 342
9	MonthlyIncome	1 000 000	77 501	0	180 973
10	MonthlyPremium	1 000 000	47 754	0	123 371
11	AnnualPremium	1 000 000	65 308	0	105 109
12	Vintage	1 000 000	30	8	67 481
13	Response	1 000 000	2	0	831 591

Count: refers to the total number of non-missing or non-null values in a variable.
Unique: represents the number of distinct or unique values in a variable
Top: represents the most frequent value or mode in a variable.
Freq: indicates the frequency or count of occurrences of the "top" value.

Table 5.7 displays the descriptive statistics and the distribution of the health insurance dataset.

Table 5.7: Descriptive statistics and distribution

No#		count	mean	std	min	25%	50%	75%	max
1	ID	1 000 000	500 001	288 675	1	250 001	500 001	750 000	1 000 000
2	Gender	1 000 000	1	1	0	1	1	2	2
3	Age	1 000 000	44	9	0	37	43	50	82
4	RegionCode	1 000 000	3 515	2 858	0	1 459	2 190	6 211	9 992
5	RaceCode	1 000 000	3	1	0	2	3	3	5
6	PreviouslyInsured	1 000 000	1	0	0	1	1	1	1
7	InitialSumAssured	1 000 000	631 557	379 414	-9 508	389 197	554 474	757 006	5 045 003
8	CurrentSumAssured	1 000 000	566 008	342 136	-120 849	336 604	514 533	725 175	3 500 000
9	MonthlyIncome	1 000 000	20 729	26 276	0	6 762	16 862	27 562	2 768 806
10	MonthlyPremium	1 000 000	551	496	0	233	439	747	7 533
11	AnnualPremium	1 000 000	6 425	5 657	0	2 801	5 114	8 613	86 617
12	Vintage	1 000 000	9	6	1	4	8	14	30
13	Response	1 000 000	0	0	0	0	0	0	1

Count: refers to the total number of non-missing or non-null values in a variable.
Std: refers to the standard deviation, which measures the dispersion or spread of the values around the mean.
Min: the minimum value in a variable, indicating the smallest observed value in the dataset.
25%: "25th percentile" represents data point below which 25% of the observations fall.
50%: "50th percentile" represents middle value in a sorted dataset.
75%: "75th percentile" represents data point below which 75% of the observations fall.
Max: represents the maximum value in a variable, indicating the largest observed value in the dataset.

These descriptive statistics provide valuable insights into the characteristics of the dataset and help with making informed decisions, identifying patterns, detecting outliers, and performing further analysis.

In the process of data exploration, the collected health insurance dataset, variables were separated into dependent variables as well as independent factors.

The variable being assessed and evaluated is referred to as the dependent variable. An independent variable is a factor that is manipulated or controlled to examine its influence on the dependent variable.

Table 5.8 Shows the top 10 selected records from the dataset after data cleaning and pre-processing process took place.

Table 5.8: The top 10 selected records

ID	Gender	Age	Region Code	Race Code	Previously Insured	Sum Assured	Monthly Income	Monthly Premium	Annual Premium	Vintage	Resp
3	1	44	2302	2	1	729 075.35	51 878.45	992.17	11 357.52	3	0
7	1	43	163	2	1	759 778.59	24 500.00	707.04	8 190.00	10	0
10	2	56	2190	3	1	252 398.73	18 840.25	618.49	7 106.65	9	0
11	2	59	9786	3	1	285 339.33	15 748.00	471.27	5 465.18	2	0
15	1	31	157	4	1	823 986.79	29 750.00	526.57	6 055.17	12	0
17	1	46	1818	3	1	521 995.54	36 904.16	755.98	8 589.81	20	0
18	2	52	7100	3	1	188 678.85	7 820.55	243.68	2 790.08	20	0
20	1	40	5201	3	1	759 578.71	40 789.75	767.42	8 761.35	19	0
23	2	46	4066	3	1	653 353.80	37 303.87	700.46	7 933.77	2	0
24	2	52	1618	3	1	672 240.25	29 313.59	2 347.55	26 980.11	18	0

5.4.2.1.2 Correlation

Mikalef *et al*. (2020) state that correlation refers to the statistical relationship between two or more variables. It measures the strength and direction of the linear association between variables, indicating how changes in one variable are related to changes in another variable. In a correlation process, statistical measurements are made to describe how much two variables are linearly related to one another and to determine whether two variables are correlated when both move in the same direction.

Figure 5.1 shows the Pearson correlation of 17 features (source: researcher), illustrating how the data were correlated.

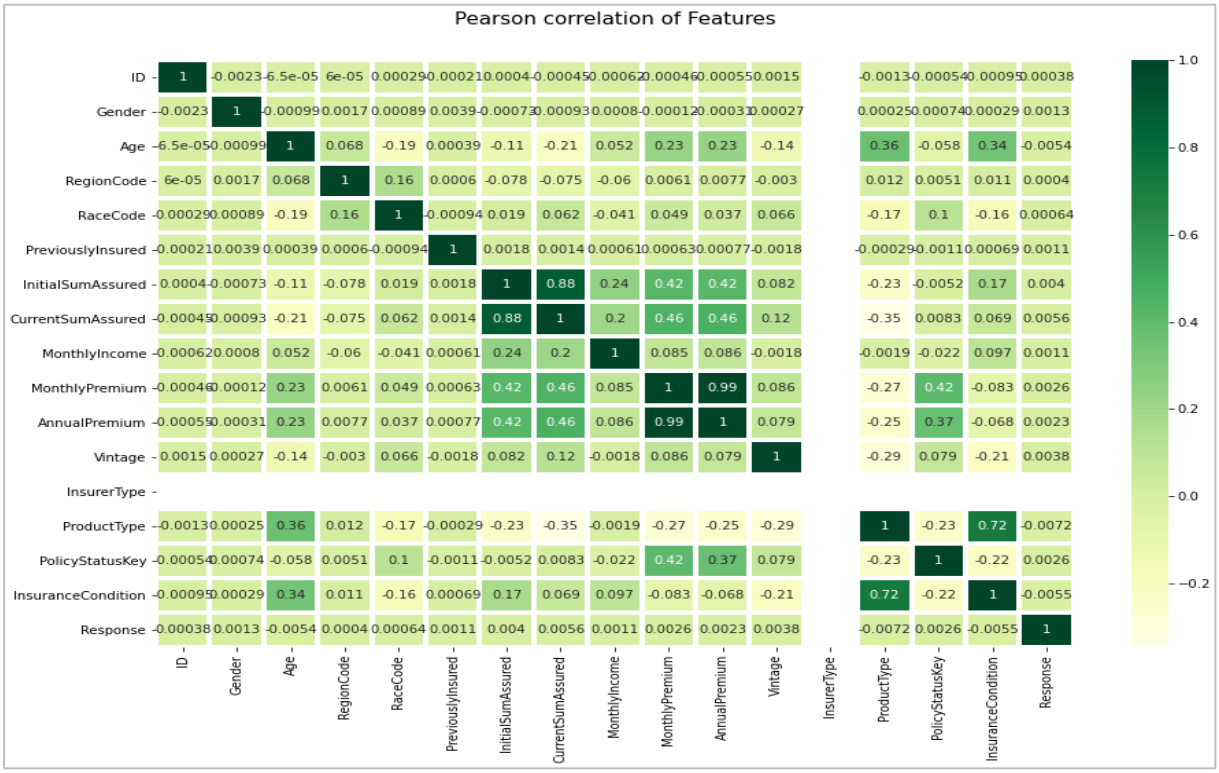


Figure 5.1: Pearson's correlation

Examining the correlation analysis of features for predicting cross-selling in health insurance, we focused on several key correlations, namely age, monthly premium, and gender.

Age (-0.054): As age increases, the likelihood of purchasing the cross-selling insurance product decreases, indicating younger individuals may be more inclined to buy additional insurance compared to older individuals.

Monthly Premium (+0.0026): As the monthly premium increases, the likelihood of purchasing the cross-selling insurance product rises, indicating that customers paying higher premiums may be more inclined to buy additional products.

Gender (+0.0013): While gender has a slight positive correlation with purchasing the cross-selling insurance product, its small coefficient suggests it may not be a major influencing factor compared to other features.

5.4.2.1.3 Data visualisation

Data visualisation is the graphical form representation of data and information using visual elements such as charts, graphs, maps, and infographics (Wang *et al.* 2020). The data has been presented in different visual elements to ensure that extracted health insurance dataset is distributed correctly. This process helps the researcher to explore the data, uncover patterns, identify outliers, and communicate the study's findings effectively.

Data is presented in graphical using features that are more affecting the target variable.

Figure 5.2 illustrates the count of Other, Male, Female in a bar graph and pie chart as percentages.

Gender attribute has been defined in section 5.3.1 whereby 0 stands for other, 1 stands for male, and 2 stands for female.

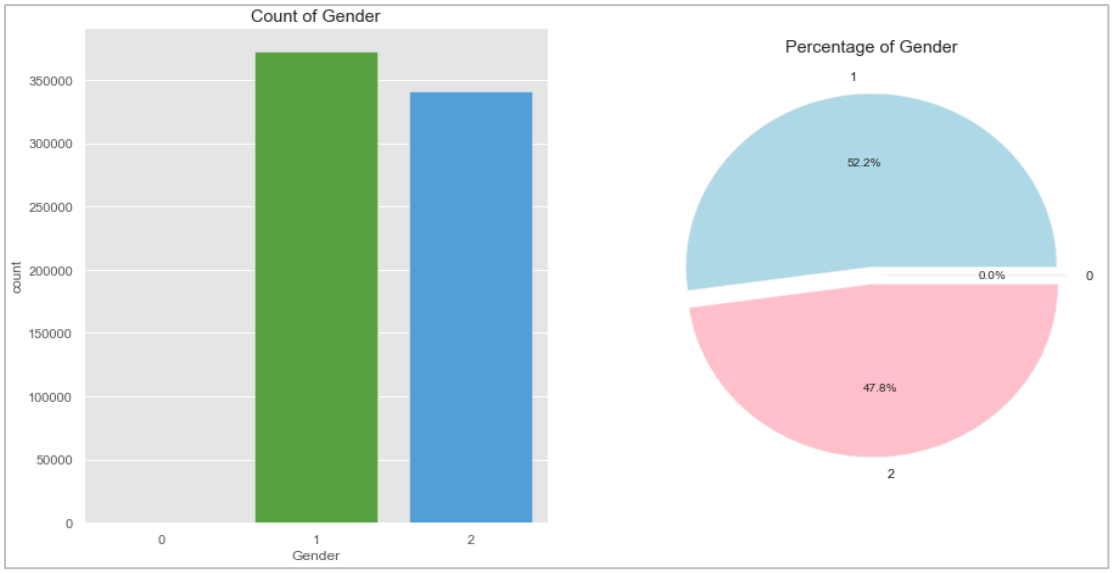


Figure 5.2: Gender

Figure 5.2 presented that the dataset contained more male compared to female records. The positive response of both males and females was slightly different, where response 1 means interested customers and 0 is not interested in purchasing additional health insurance products.

Figure 5.3 compares two curves blue and red showing Age Vs Response based on customer age analysis. 1 indicates interest in purchasing additional health insurance, and 0 indicates no interest.

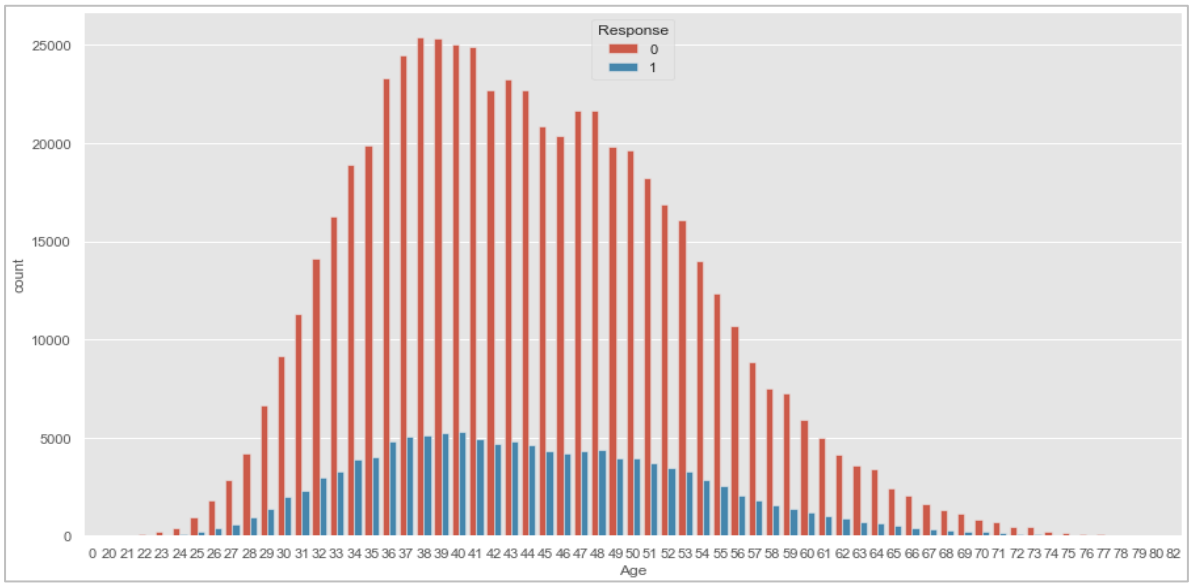


Figure 5.3: Age Vs response

Figure 5.3 illustrates the age versus Frequency of Likelihood curve, after analyzing existing health insurance consumer data based on consumers who are previously insured. Data analyses indicated that consumers aged 25 to 70 exhibit a higher likelihood of expressing interest in purchasing additional health insurance products compared to customers younger than 25 years old.

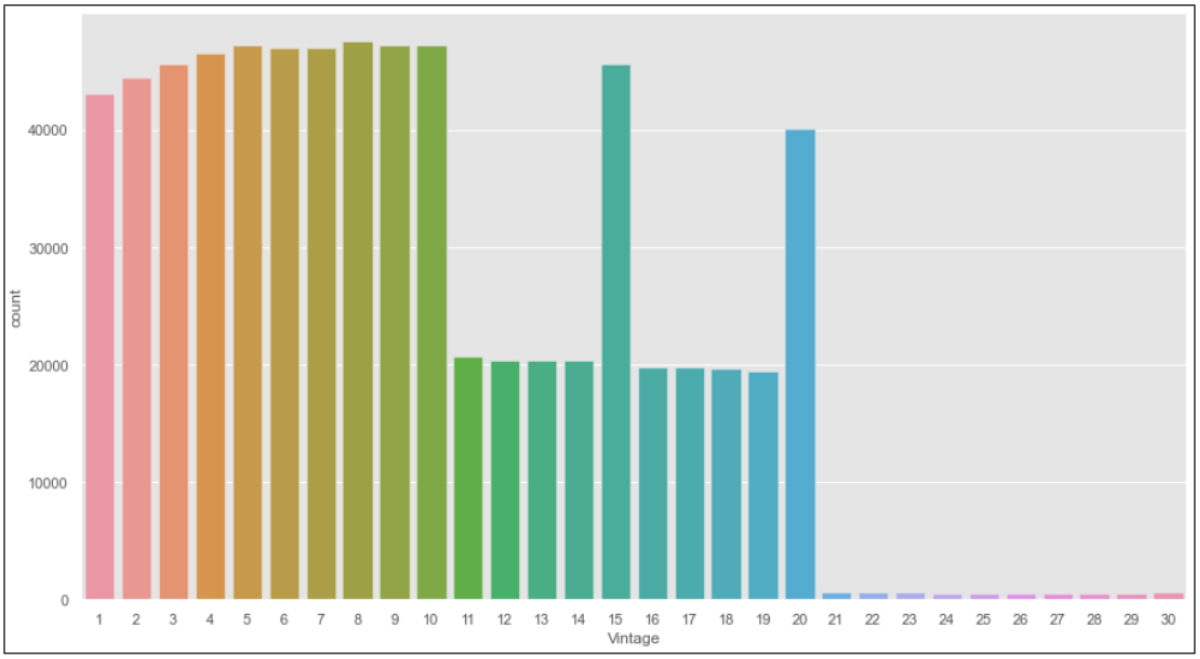


Figure 5.4: Vintage Response

Figure 5.4 depicts the vintage response, indicating the duration of customer association with the company in terms of years. The vintage represents the number of years consumers use company services from 1 to 30 years versus the number of consumers in the count. This graph shows that previously insured customers are more likely to respond to cross-selling of an additional health insurance product compared to customers who are not previously insured.

Before model selection, the step of data exploration was carried over by performing a feature selection process. Feature selection is also known as variable selection.

5.5 ML model selection for health insurance cross-selling predictions

Vandrangi (2022) explains that ML model selection involves choosing the most suitable ML algorithm or model for a given problem or task. The selection process,

therefore, depends on several factors, including the nature of the data, and the problem type such as classification, regression, or clustering.

In order to choose the ML algorithms referred to in Chapter 2, the literature review proposed that the following ML algorithms were commonly used due to their effectiveness in managing complex datasets, robustness against overfitting, and efficient handling of large datasets. These algorithms provide insights into feature importance and are interpretable, aligning well with the predictive modeling objectives of the study. The researcher consequently used these ML algorithms for the predictive model.

The four identified models are:

- Random Forest
- KNN
- XGBoost
- Logistic Regression

5.6 Health insurance cross-selling model training

ML model training refers to the process of teaching an ML algorithm to recognise patterns and make predictions based on input data. Training a model involves providing it with a set of labelled historical data, known as the training dataset, and allowing it to learn from that data to improve its performance (Ozdemir and Bayrakli, 2022). The four models identified in section 5.5 were used to train the model for this study of health insurance cross-selling prediction with ML for South African consumers.

Before the model training commenced, data exploration was carried out by performing a feature selection process as explained in Section 2.4.4.1 in Chapter 2 and Section 3.3.1 in Chapter 3. Feature selection, also known as variable selection attribute selection, is the process of selecting a subset of relevant features (variables) from a larger set of available features in a dataset. Taha, Cosgrave and Mckeever (2022) state that the objective of feature selection is to optimize the performance of an ML model by decreasing the data's dimensionality, discarding irrelevant or redundant features, and emphasizing the most relevant ones.

In this study the features were selected based on the nature of the collected health insurance dataset, the number of features available, and the chosen ML algorithms in order to train a model.

The extracted dataset from the large insurance company database was split into training and testing sets using Python programming language. A training dataset is a portion of the data that the model learns from to understand patterns, relationships, and underlying data structures used to train a machine-learning model.

The training dataset consisted of input data (features) and corresponding target variables (labels) for the health insurance extracted dataset. The model was trained using the training dataset to discover the patterns and relationships for predictions.

The testing dataset is a separate portion of the data that is used to evaluate the performance of a trained machine-learning model. The test set also consists of input features and corresponding target variables. The target variables in the testing dataset are used as ground truth to compare the model's predictions. The test set was applied after the model was trained on the training dataset using Python programming language to make health insurance cross-selling predictions.

5.7 Health insurance ML predictive model evaluation (performance)

Model evaluation is the process of assessing the performance of an ML model on a given task. It is an important step in the ML pipeline, as it helps to identify potential problems with the model and guides improvements in the modelling process (Chekani *et al.* 2023).

Using the selected training dataset, the performance of the four models was assessed by comparing their predictions to the actual values in the testing dataset. The assessment criteria, including accuracy, precision, recall, and F1 score, were employed to determine the effectiveness of the model in predicting health insurance cross-selling.

5.7.1 Definition of assessment metrics

Model performance of an ML model is evaluated using evaluation metrics which measure how well the generalises to unseen data. The choice of evaluation metrics depends on the specific task and the nature of the problem being solved (Vujovic

2021). Here are some commonly used evaluation metrics for various types of ML models, explained with examples in Chapter 3, section 3.3.5.

Confusion matrix

A matrix of confusion is a tool used to evaluate the performance of a classification model. It is particularly useful for binary classification problems, divided into two groups: positive and negative predictions.

Table 5.9: Confusion matrix

Actual	Negative	Positive
Negative	(True Negative(TN))	(False Positive (FP))
Positive	(False Negative (FN))	(True Positive (TP))

The following evaluation metrics were used to measure the performance of each model.

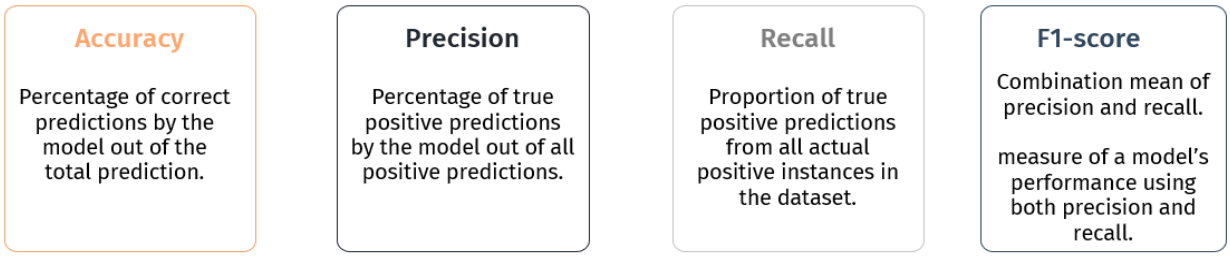


Figure 5.5: Evaluation metrics

5.8 Build of the ML prediction model for health insurance cross-selling

To build a ML prediction model for health insurance cross-selling, the researcher was guided by the crucial steps performed in section 5.4 regarding data preparation including data cleaning and pre-processing. After these steps were completed the dataset was ready for training. The extracted health insurance dataset was trained, and a predictive model was built. The quality of the data, feature engineering, model selection, and appropriate evaluation were all important factors in the health insurance cross-selling prediction model's success.

5.9 Health insurance cross-selling predictive model results

In this section, the ML algorithms named in section 5.5 were applied to predict cross-selling potential in the health insurance domain. The results demonstrate that all the ML algorithms achieved relatively high accuracy and performed well in identifying potential cross-selling additional health insurance products to existing customers.

The results obtained from the training and testing of each ML model using 80% of the training dataset (570830 records) and 20% of the validation/testing dataset (142708 records)

5.9.1 Random Forest

Table 5.10: Random forest classifier training results

Random Forest Accuracy Score: 0.99877				
Confusion matrix: [[474506, 95] [607, 95622]]				
Classification report:	precision	recall	F1-score	support
0	1.00	1.00	1.00	474601
1	1.00	1.00	0.99	96229
accuracy			1.00	570830
macro avg	1.00	1.00	1.00	570830
weighted avg	1.00	1.00	1.00	570830

Table 5.11: Random forest classifier testing results

Random Forest Accuracy Score: 0.79605				
Confusion matrix: [[112264, 6189], [22916, 1339]]				
Classification report:	precision	recall	F1-score	support
0	0.83	0.95	0.89	118453
1	0.18	0.06	0.08	24255
accuracy			0.80	142708
macro avg.	0.50	0.50	0.48	142708
weighted avg.	0.72	0.80	0.75	142708

5.9.2 KNN

Table 5.12: KNN training results

K-Neighbors Classifier Accuracy Score: 0.86007				
Confusion matrix: [[460904, 13697], [66175, 30054]]				
Classification report:	precision	recall	F1-score	support
0	0.87	0.97	0.92	474601
1	0.69	0.31	0.43	96229
accuracy			0.86	570830
macro avg.	0.78	0.64	0.67	570830
weighted avg.	0.84	0.86	0.84	570830

Table 5. 13: KNN testing results

K-Neighbors Classifier Accuracy Score: 0.78016				
Confusion matrix: [[109395, 9058], [22314, 1941]]				
Classification report:	precision	recall	F1-score	support
0	0.83	0.92	0.87	118453
1	0.18	0.08	0.11	24255
accuracy			0.78	142708
macro avg.	0.50	0.50	0.49	142708
weighted avg.	0.72	0.78	0.74	142708

5.9.3 XGBoost

Table 5.14: XGBoost training results

XGBoost Accuracy Score: 0.83157				
Confusion matrix: [[474587, 14], [96128, 101]]				
Classification report:	precision	recall	F1-score	support
0	0.83	1.00	0.91	474601
1	0.88	0.00	0.00	96229
accuracy			0.83	570830
macro avg.	0.85	0.50	0.46	570830
weighted avg.	0.84	0.83	0.76	570830

Table 5.15: XGBoost testing results

XGBoost Accuracy Score: 0.82995				
Confusion matrix: [[118436, 17], [24250, 5]]				
Classification report:	precision	recall	F1-score	support
0	0.83	1.00	0.91	118453
1	0.23	0.00	0.00	24255
accuracy			0.83	142708
macro avg.	0.53	0.50	0.45	142708
weighted avg.	0.73	0.83	0.75	142708

5.9.4 Logistic Regression

Table 5.16: Logistic regression training results

Logistic Regression Accuracy Score: 0.83142				
Confusion matrix: [[474601, 0], [96229, 0]]				
Classification report:	precision	recall	F1-score	support
0	0.83	1.00	0.91	474601
1	0.00	0.00	0.00	96229
accuracy			0.83	570830
macro avg.	0.42	0.50	0.45	570830
weighted avg.	0.69	0.83	0.75	570830

Table 5.17: Logistic regression testing results

Logistic Regression Accuracy Score: 0.83003				
Confusion matrix: [[118453, 0], [24255, 0]]				
Classification report:	precision	recall	F1-score	support
0	0.83	1.00	0.91	118453
1	0.00	0.00	0.00	24255
accuracy			0.83	142708
macro avg.	0.42	0.50	0.45	142708
weighted avg.	0.69	0.83	0.75	142708

Among the algorithms, the Random Forest demonstrated superior overall performance, boasting an accuracy of 1.00 and an F1 score of 1.00. Consequently, the Random Forest model is selected as the optimal choice for building a health insurance cross-selling prediction model with the historical health insurance customer dataset that was utilized in this study. The exceptional accuracy and F1 score of the Random Forest model signify its robustness and effectiveness in accurately predicting potential customers for health insurance cross-selling. This model's high performance makes it a reliable tool for guiding marketing strategies and maximizing cross-selling opportunities within the insurance industry.

Table 5.18 shows comprehensive single view for training and testing results

Table 5.18: Evaluation metrics for ML algorithms

ML Algorithms Model	Accuracy Score		F1-Score	Computation (min & sec)
	Training	Testing		
Random Forest	0.9987	0.7960	1.00	6m 34.6s
K-Nearest Neighbors	0.8600	0.7801	0.92	30m 12.4s
XGboost	0.8315	0.8299	0.91	5m 5.5s
Logistic Regression	0.8314	0.8300	0.91	4m 5.2s

Interpretation of the F1-scores for the four models are shown in Table 5.18. The Random Forest model stands out with the highest F1-Score of 1.00, indicating its superior balance between precision and recall. The other models all exhibited an F1-Score of 0.92 and 0.91 which is also a good score in terms of accurately classifying positive instances and maintaining a balance between precision and recall, but the random forest model offers a slightly better overall classification performance.

The results indicated that previously insured customers are more likely to respond to the health insurance additional product compared to a customer who was not previously insured.

Figure 5.6 shows interested vs not Interested customers illustrated in the form of a bar graph and pie chart.

The visual representation in the below figure provides a comparative analysis between consumers with a history of prior insurance coverage, who exhibit a higher likelihood of responding to the health insurance supplementary product, and consumers without previous insurance. This visual comparison sheds light on the distinctive response patterns between these two customer segments, contributing valuable insights into the factors influencing the propensity to engage with additional health insurance products based on their insurance history.

The following results indicating 83.1% of consumers responded negatively to cross-selling and 16.9% responded positively were obtained by analyzing a dataset containing consumers' policy information and their previous responses to health insurance cross-selling offers within. The dataset sample size was used to calculate the percentage of consumers interested and not interested in the supplementary product. Additionally, a comparative analysis was conducted between consumers with and without previous insurance coverage to determine their respective likelihood of responding to the cross-selling offer.

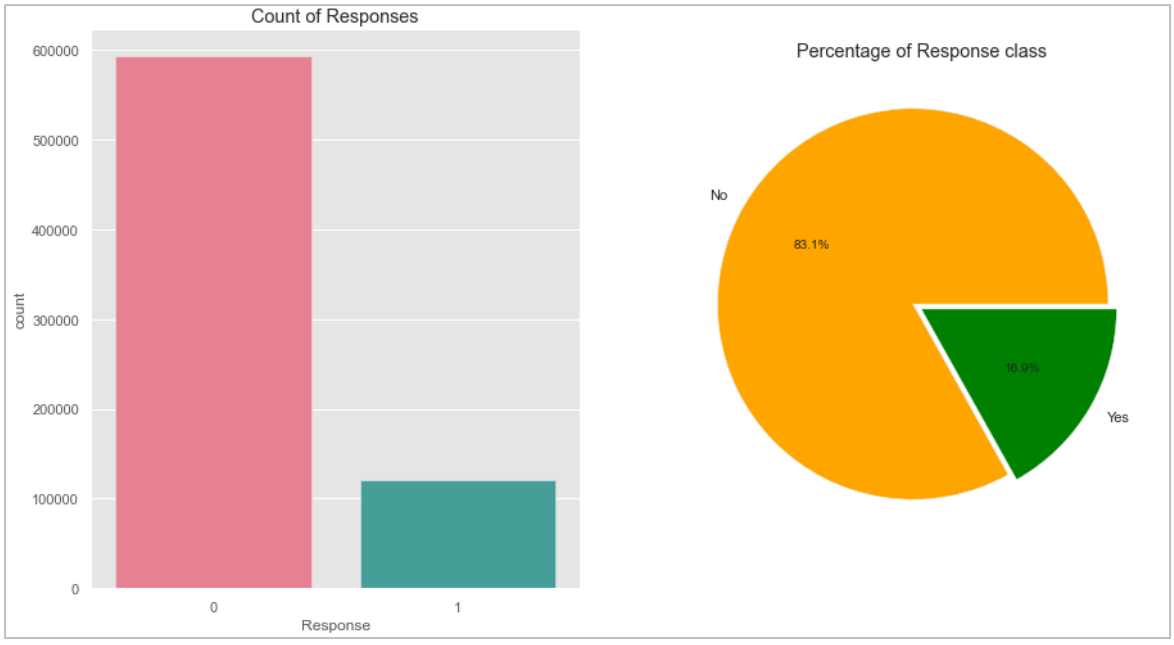


Figure 5.6: Interested vs not interested customers

In this study the process of building the ML prediction model was carried over, with the researcher randomly selecting 5 customers after each ML model's output results were produced to test the model accuracy. Based on the researcher's prior experience, 3 out of 5 customers would be interested in purchasing additional health insurance products.

The researcher trained and tested the following dataset with 5 records knowing that 3 of them would be interested. The selected dataset features were *Age, Gender, PreviouslyInsured, InitialSumAssured, CurrentSumAssured, MonthlyIncome, MonthlyPremium, AnnualPremium,* and *Vintage* (Table 5.19).

Table 5.19: Health insurance dataset for 5 randomly selected customers

ID	Gender	Age	Previously-Insured	Initial-SumAssured	Current-SumAssured	Monthly-Income	Monthly-Premium	Annual-Premium	Vintage
5	2	34	1	226 479,08	167 230,65	11 942,88	111,04	1 279,32	15
2	2	47	1	439 242,89	416 440,67	20 304,12	958,81	10 860,02	17
6	2	36	1	551 320,84	533 374,24	19 663,75	327,76	3 768,21	12
4	2	50	1	753 597,95	553 894,29	30 574,25	589,37	6 656,12	20
10	1	31	1	622 840,90	541 964,98	11 810,00	494,92	5 638,01	19

Table 5.20 shows the submission of results indicating the customers who were interested and not interested to purchase health insurance additional products. After completing the model testing the results produced for the 3 manually predicted customers with IDs 2, 4, and 10 from the selected dataset based on the researcher's experience showed that all 3 were positively interested in purchasing the additional health insurance products, with the response field marked as 1 for interested customers, and marked as 0 for not interested customers. Thus the model was accurate and its performance is good.

Table 5.20: Submission of results

ID	Response
5	0
2	1
6	0
4	1
10	1

The selected features for the above output results after the model was tested indicated the following:

- Customers of age between 25 to 70 are more likely to purchase additional health insurance products.
- Customers with a high vintage (number of days customers are associated with the company) have a higher chance of purchasing additional health insurance.
- Customers who have previously been insured are likely to purchase additional health insurance products.
- Features such as *Gender, Age, PreviouslyInsured, MonthlyIncome, MonthlyPremium,* and *AnnualPremium* have more of an effect on the target variable compared to *InitialSumAssured, CurrentSumAssured,* and *Vintage*.

5.10 Conclusion

These findings illustrated the significant impact of utilising ML algorithms in enhancing cross-selling strategies within the health insurance industry. With vast amounts of customer data available in the health insurance sector, ML models can analyse this data to identify patterns, behaviours, and preferences, thereby helping identify potential customers who may be interested in purchasing additional insurance products. Furthermore, ML algorithms can accurately predict the likelihood of a customer's interest in purchasing additional products, enabling insurance companies to personalise their cross-selling efforts.

This personalisation can lead to increased customer engagement and satisfaction while optimising revenue generation for the business. Moreover, ML algorithms continuously learn and adapt based on historical datasets extracted from health insurance customers. By harnessing the power of ML, the health insurance industry can enhance the effectiveness of cross-selling, drive business growth, and address the evolving needs of its customers more effectively. Future research expands and builds on these findings. The findings are discussed in detail in the next chapter.

CHAPTER 6: DISCUSSION OF FINDINGS

6.1 Introduction

All of the study's findings were presented in Chapter 5. A discussion of these results is presented in this chapter.

6.2 Identified ML algorithms for health insurance prediction model

In Chapter 5, section 5.9, the results for identified commonly used ML algorithms, including Random Forest, K-Nearest Neighbors (KNN), XGBoost, and Logistic Regression, were presented for the health insurance prediction model. The findings regarding these algorithms are discussed in detail here in section 6.2.1.

6.2.1 ML algorithms – findings and discussion

A suitable ML algorithms were identified considering the health insurance dataset and the type of research question to be solved. The dataset consists of health insurance customer information extracted from a large insurance company. The algorithms were applied to train and test the extracted dataset from the health insurance data (section 6.3)

6.3 Findings of extracted health insurance dataset

The extracted dataset had enough entries and features with customer demographic details and policy details to train the model. This dataset was a rich repository with historical and current customer information which supported a deductive research approach. The dataset needed to undergo the process of dataset overview and descriptions to identify the number of records and features extracted (section 6.4).

The extracted dataset consisted of one million entries (rows) and 17 features (columns). The features were selected for model training. The dataset data type was checked for every feature and the findings showed that all the data types were correct. The data was prepared through cleaning and pre-processing (see section 6.5).

Gender representation

The dataset contained a relatively balanced representation of gender, with slightly more males (52.2%) than females (47.8%). This suggests that the dataset included a diverse set of individuals, allowing for potential gender-based analysis and modelling.

Missing values: The finding of 0% for the unknown gender category suggests that the dataset had been well-preprocessed or does not contain any missing values or invalid entries for gender. This is important as missing or unknown values can affect the accuracy and reliability of subsequent analyses and modelling.

Gender-based analysis

The gender distribution in the dataset enables the possibility of performing gender-based analysis. Researchers or analysts can explore potential variations or differences in health insurance preferences, needs, or behaviour s between males and females. This analysis can provide valuable insights for insurance companies to better understand their customer base and tailor their products and services accordingly.

Gender-specific marketing strategies

The gender distribution finding can also guide marketing strategies for health insurance companies. By understanding the proportion of males and females in the dataset, companies can develop targeted marketing campaigns that address the unique needs, preferences, and communication styles of each gender segment. This personalised approach can enhance customer engagement and improve marketing effectiveness.

Equal representation

Although the gender distribution appears relatively balanced in this dataset, it is important to ensure equal representation and avoid biases in data collection and analysis. Unbalanced gender representation can introduce biases in modelling and analysis results, leading to skewed or inaccurate insights.

6.4 Findings of data preparation

The data preparation means that data was cleaned and preprocessed. The findings in this section indicated that the dataset contained duplicate, null values, and missing

values. After cleaning and pre-processing the dataset, the resulting shape was 713 538 rows and 17 columns. The cleaning and pre-processing steps were conducted correctly because the duplicate, null, and missing values were removed from the extracted health insurance dataset from the large insurance company database.

Findings also showed that the dataset was ready for the model selection process which is discussed in section 6.6.

6.5 Findings of model selection

The results presented in section 5.9 for the selected machine learning algorithms, including Random Forest, K-Nearest Neighbors (KNN), XGBoost, and Logistic Regression, revealed distinct performance outcomes in building a model for predicting cross-selling opportunities in health insurance. Each algorithm demonstrated varying degrees of effectiveness based on metrics like accuracy, precision, recall, and F1 score, providing insights into their suitability for this specific application. Detailed discussions on the outcomes of each model are provided in section 6.6.

6.6 Findings of the trained ML model

After preprocessing, the health insurance dataset was trimmed down to 713,538 records. This dataset was subsequently divided, with 80% (570,830 records) allocated for training all machine learning models, and the remaining 20% (142,708 records) reserved for testing purposes. The findings from the trained machine learning models provided valuable insights into predicting health insurance outcomes. Detailed results for the trained model are discussed below:

Upon analyzing the F1-Scores for the four models presented across four tables, it becomes evident that the random forest model stands out with 6 minutes and 35.6 seconds computation, boasting the highest F1-Score of 1.00 and an accuracy score of 0.99. This underscores its exceptional equilibrium between precision and recall, indicating proficiency in accurately identifying positive instances while minimizing both false positives and false negatives.

Conversely, the K-Nearest Neighbours (KNN) demonstrated an F1-Score of 0.92, XGBoost Classifier, and Logistic Regression demonstrated an F1-Score of 0.91,

showcasing comparable performances in accurately classifying positive instances and maintaining a balance between precision and recall. Despite the effectiveness of these models, the Random Forest model appears to offer a superior overall prediction performance.

The differences in computation times among the training algorithms for health insurance cross-selling predictions suggest variations in their computational complexity and efficiency. Random Forest with 6m 34.6s and Logistic Regression 4m 5.2s demonstrated relatively shorter computation times compared to K-Nearest Neighbours took 30m 12.4s, implying faster processing and model training. XGBoost classifier fell between these extremes in terms of computation time of 5m 5.5s. However, while computation time provides insights into efficiency, it alone doesn't determine model performance; other performance metrics like accuracy, precision, and recall should also be considered for a comprehensive assessment of each model's effectiveness in making accurate predictions for health insurance cross-selling.

6.7 Findings of the model evaluation

Referring to Chapter 5, where all results were detailed for model performance evaluation using a testing dataset comprising 20% (142,708 records), this section elaborates on the findings of the evaluated models for constructing a health insurance cross-selling prediction model based on the testing dataset.

6.7.1 Random Forest algorithm

The performance evaluation of the Random Forest model resulted in an accuracy score of 0.80, using 20% of the testing dataset, which amounts to 142708 records. It showed higher precision and recall for the first class (0) compared to the second class (1), with an F1-Score of 0.89 for class 0. However, the macro-average F1-Score is relatively low at 0.48, and the weighted average F1-Score is 0.75, indicating varying performance across both classes.

6.7.2 Findings of the KNN algorithm

The evaluation of the K-Nearest Neighbors (KNN) Classifier resulted in an accuracy score of 0.78, using 20% of the testing dataset, equivalent to 142708 records. In the KNN results, it was evident that the classifier displayed higher precision and recall for

the first class (0) compared to the second class (1), resulting in an F1-Score of 0.87 for class 0. The macro-average F1-Score, reflecting a balanced evaluation across both classes, was observed at 0.49, while the weighted average F1-Score stood at 0.74.

6.7.3 Findings of the XGBoost algorithm

The performance evaluation of the XGBoost model achieved a commendable accuracy score of 0.83, utilizing a testing dataset comprising 20%, which totals 142708 records, indicating its strong overall performance. Particularly noteworthy is the model's precision of 0.83 for class 0 and recall of 1.00 for the first class (0), highlighting its capability to precisely identify instances of this class. This accuracy score of 0.83 underscores the model's effectiveness in making accurate predictions, taking into account both true positives and true negatives.

6.7.4 Findings of the Logistic Regression algorithm

The performance evaluation of the logistic regression model revealed a significant accuracy score of 0.83, employing a testing dataset consisting of 20%, totaling 142708 records, highlighting its strong overall performance. The logistic regression model displays a commendable precision of 0.83 and a recall of 1.00 for the first class (0), indicating its capability to accurately identify instances of this class. With an accuracy of 0.83, the model demonstrates its adeptness in making precise predictions, covering both true positives and true negatives.

6.8 Findings of the ML predictive model

The significant difference in the percentages suggests a class imbalance issue in the dataset. Class imbalance occurs when one class (in this case "not interested") is dominant compared to the other class ("interested"). Class imbalance can potentially affect the model's performance, as the algorithm may be biased toward predicting the majority class.

Only 16.9% of customers were predicted to show an interest in additional health insurance, highlighting the importance of targeted marketing efforts. Insurance companies can focus their marketing strategies on reaching out to the specific segment of customers who are interested in additional health insurance, tailoring their messaging and offerings to attract and engage this particular group.

Although there is a class imbalance, the presence of customers interested in health insurance still represents a market opportunity. Companies can concentrate on understanding the characteristics, preferences, and needs of these interested customers to design targeted insurance products and services that cater to their requirements. By doing so, they can capitalise on the potential demand and attract more customers.

The findings are that a significant majority of customers, specifically 83.1%, are not likely to be interested in purchasing additional health insurance products. The large percentage of customers who express little interest in purchasing additional health insurance policies raises the possibility that there may not be much of a market for cross-selling additional health insurance. Insurance companies should take this into account when developing their marketing strategies and setting realistic expectations regarding the potential customer base.

6.9 Conclusion

Finding out that the majority of customers are likely to not be interested in additional health insurance products can still provide insights into their preferences and priorities.

It would be beneficial to conduct further analysis of the interested customers to identify any common patterns or characteristics. Exploring variables such as age, gender, income, or previous insurance status could provide insights into the specific customer segments that are more likely to be interested in health insurance. This information can guide marketing strategies and assist in developing personalised approaches to engage these potential customers effectively.

Analysis of the health insurance dataset revealed that customers between the ages of 25 and 70 showed a higher likelihood of purchasing additional health insurance products. This age range corresponds to the prime working and earning years for many individuals, where they may have a greater need for comprehensive health coverage to protect themselves and their families.

Understanding this age-related pattern is crucial for developing targeted marketing strategies and tailoring product offerings to cater to the needs and preferences of customers within this age group. Insurance companies can focus their efforts on

designing plans that align with the specific life stages, financial goals, and healthcare requirements of individuals in this demographic.

The findings highlight the importance of segmenting the customer base by age and customising marketing messages and channels accordingly. By crafting age-appropriate communication strategies, insurance companies can effectively convey the value and benefits of additional health insurance products to customers in the targeted age range.

CHAPTER 7: FINAL REMARKS AND RECOMMENDATIONS

7.1 Introduction

This chapter presents a comprehensive overview of the entire study, encompassing a summary of the undertaken work, an exploration of areas for further research, and an elucidation of the distinctive contributions made by this research. Additionally, it revisits the objectives initially set forth in Chapter 1, assessing their accomplishment. Furthermore, this section presents the research's conclusions along with recommendations, while also delving into potential future possibilities of study that have been illuminated by the findings.

The study of health insurance cross-selling prediction using ML for South African consumers has been conducted successfully. Various ML algorithms were employed to assess the likelihood of customers acquiring additional health insurance products. Chapter 1 provided an introduction to the study and outlined its aims and objectives. Chapter 2 surveyed the existing literature on health insurance cross-selling, while Chapter 3 provided an explanation of the ML algorithms used. Chapter 4 presented the methodology and data collection and analysis methods. Chapter 5 unveiled the study's findings and results, while Chapter 6 engaged in a discussion of these findings.

a) Research Objective 1

To use suitable machine learning techniques to extract health insurance customer data from a large insurance company in South Africa.

The first objective of this research was to employ appropriate ML techniques to extract customer data related to health insurance from a large insurance company. Subsequently, the extracted health insurance data was subjected to a rigorous process of cleaning and preparation to ensure its quality and suitability for further analysis. This was achieved.

b) Research Objective 2

To determine which machine learning strategies are most useful for anticipating health insurance cross-selling.

The second objective of this research was to identify the most effective ML algorithms from among Random Forest, KNN, XGBoost, and Logistic Regression in terms of their ability to predict health insurance cross-selling opportunities. The objective was to determine the algorithms that demonstrate the highest level of accuracy, reliability, and applicability in forecasting customer behaviour related to health insurance cross-selling probabilities. This was achieved and logistic regression was established to be the most effective ML algorithm.

c) Research Objective 3

To determine which features influence the machine learning algorithms chosen for health insurance customer cross-selling.

Third objective of this study was to identify the characteristics that are crucial in selecting machine learning algorithms for predicting health insurance cross-selling.

7.2 Recommendations and future work

7.2.1 Research Objective 1

1. Researchers are encouraged to explore the ML techniques employed in this study to extract and analyze health insurance customer data from larger datasets, thereby enhancing the accuracy of their models.
2. Recommend other researchers to employ the ML techniques utilised in this study for extracting and analysing customer datasets from various sectors, based on the nature of the problem and datasets available.

7.2.2 Research Objective 2

1. Given that the random forest algorithm proved useful for this study, it is recommended that this algorithm be deployed to other health insurance products.
2. Use the random forest algorithm for analysis and customer cross-selling prediction in other sectors.

7.2.3 Research Objective 3

1. The identification of the features deployed for this study proved to be non-trivial because it required insurance domain-specific exposure and experience. It is

recommended that unsupervised ML techniques be employed for further exploration and refinement of feature selection processes.

7.3 Conclusion

The research conducted on the health insurance dataset aimed to gain insights into customer behaviour and preferences regarding likelihood of customers purchasing additional health insurance products as a result of cross-selling. The findings revealed that approximately 16.9% of the 713 538 health insurance customers in the dataset could be interested in purchasing additional health insurance products. This indicates a potential market segment that insurance companies can target for cross-selling initiatives.

The gender distribution showed that 52.2% of customers were male, while 47.8% were female, highlighting the importance of considering gender-specific marketing strategies. Furthermore, customers between the ages of 30 and 60 exhibited a higher likelihood of purchasing additional health insurance products, suggesting that individuals in their prime working and earning years may have a greater need for comprehensive coverage.

Another significant finding is that customers with a high number of days associated with the insurance company, known as "vintage", would be more likely to purchase additional health insurance. This suggests that customer loyalty and long-term relationships play a role in decision-making processes when it comes to additional coverage. Insurance companies can leverage this insight by implementing customer retention strategies and offering tailored cross-selling opportunities to customers with a high vintage.

Furthermore, the research indicated that customers who have previously been insured are more likely to purchase additional health insurance products. This finding highlights the potential value of targeting customers who already have some level of insurance coverage, as they may be more receptive to expanding their coverage.

The main research question of this study was to explore how ML can be applied to increase cross-selling revenue for companies in the South African health insurance industry. The research's conclusions show that it is possible to employ predictive

models and ML algorithms to identify clients who are more likely to buy additional health insurance coverage.

The analysis revealed that certain customer characteristics, such as *Gender, Age, Previously Insured, Monthly Income, Monthly Premium, and Annual Premium* are significant factors. These variables play a role in predicting customer interest in additional health insurance products. Insurance companies can leverage this information to develop targeted marketing campaigns and personalised product offerings that align with the preferences and needs of their customer base.

The research highlighted the importance of using ML algorithms to identify customer behaviour patterns and preferences for effective cross-selling. It also explored various ML techniques suitable for extracting health insurance customer data from a large insurance company in South Africa.

The findings emphasised the significance of cross-selling health insurance products using data analytics, as it enables insurance companies to make informed decisions, tailor marketing strategies, and improve customer satisfaction. ML techniques, such as association rule mining, clustering, and classification, were identified as effective approaches for extracting meaningful insights from customer data. Additionally, a variety of ML algorithms, such as Random Forest, KNN, Logistic Regression, and XGBoost were recognised as appropriate choices for forecasting customer behaviour and enhancing cross-selling initiatives.

Based on the findings and discussion, several recommendations for future research can be made. Firstly, conducting in-depth customer surveys with a representative sample can provide deeper insights into the motivations and barriers influencing customers' decisions to purchase additional health insurance. Secondly, further analysis using advanced segmentation techniques can help uncover distinct customer segments with unique preferences and behaviours, enabling more targeted marketing strategies. Thirdly, comparing the performance of different ML algorithms and models can guide the selection of the most effective approach for health insurance cross-selling prediction.

Fourthly, investigating the customer lifetime value of those who purchase additional health insurance products compared to those who do not can provide insights into

long-term profitability and retention rates. Fifthly, integrating external data sources, such as socio-economic and lifestyle data, with the existing dataset can enhance the understanding of customer profiles and potential predictors of interest in additional coverage. Finally, conducting a comparative industry analysis can provide benchmarks and best practices for insurance companies, facilitating the identification of successful cross-selling strategies in other markets.

By addressing these research directions, insurance companies can refine their marketing strategies, better understand customer needs, and drive business growth in the competitive health insurance market. Thus, this research adds to the existing body of knowledge within the field and offers valuable insights for decision-making and the advancement of future improvements in health insurance cross-selling practices.

References

Acimovic, J., Erize, F., Hu, K., Thomas, D.J. and Mieghem, J.A.V. 2019. Product life cycle data set: raw and cleaned data of weekly orders for personal computers. *Manufacturing & Service Operations Management*, 21(1): 171-176.

Agrawal, S., Das, A., Gaikwad, A. and Dhage, S. 2018. Customer churn prediction modelling based on behavioural patterns analysis using deep learning. In *2018 International conference on smart computing and electronic enterprise (ICSCEE)* (pp. 1-6). IEEE.

Ahmed, A. and Linen, D.M. 2017. A review and analysis of churn prediction methods for customer retention in telecom industries. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1-7). IEEE.

Ahmed, H.A., Hameed, A. and Bawany, N.Z. 2022. Network intrusion detection using oversampling technique and machine learning algorithms. *PeerJ Computer Science*, 8: e820.

Ahmed, S. 2019. Density based clustering algorithm for distributed datasets using mutual k-nearest neighbors. *International Journal of Advanced Computer Science and Applications*, 10(3): 320-360.

Akokuwebe, M.E. and Idemudia, E.S., 2022. A comparative cross-sectional study of the prevalence and determinants of health insurance coverage in Nigeria and South Africa: a multi-country analysis of demographic health surveys. *International Journal of Environmental Research and Public Health*, 19(3): 1766.

Alesane, A. and Anang, B.T. 2018. Uptake of health insurance by the rural poor in Ghana: determinants and implications for policy. *Pan African Medical Journal*, 31(1).

Amani, F.A. and Fadlalla, A.M., 2017. Data mining applications in accounting: a review of the literature and organizing framework. *International Journal of Accounting Information Systems*, 24: 32-58.

Ambika, P. 2020. Machine learning and deep learning algorithms on the industrial internet of things (IIoT). *Advances in Computers*, 117(1): 321-338.

Anitha, P. and Patil, M.M. 2019. RFM Model for customer purchase behaviour using K-Means Algorithm. Journal of King Saud University - Computer and Information Sciences, 34(5): 1785-1792. Available: https://www.sciencedirect.com/science/article/pii/S1319157819309802 (Accessed 03 April 2023).

Antwi, S.K. and Hamza, K., 2015. Qualitative and quantitative research paradigms in business research: a philosophical reflection. *European Journal of Business and Management*, 7(3): 217-225.

Asenahabi, B.M., 2019. Basics of research design: A guide to selecting appropriate research design. *International Journal of Contemporary Applied Researches*, *6*(5): 76-89.

Ayanore, M.A., Pavlova, M., Kugbey, N., Fusheini, A., Tetteh, J., Ayanore, A.A., Akazili, J., Adongo, P.B. and Groot, W. 2019. Health insurance coverage, type of payment for health insurance, and reasons for not being insured under the National Health Insurance Scheme in Ghana. *Health Economics Review*, 9: 1-15.

Babuna, P. Yang, X. Gyilbag, A. Awudi, D.A. Ngmenbelle, D. and Bian, D. 2020. The impact of Covid-19 on the insurance industry: *International Journal of Environmental Research and Public Health*, 17(16): 5766. Available: https://www.mdpi.com/1660-4601/17/16/5766 (Accessed 29 July 2021).

Banerjee, A., Chen, S., Fatemifar, G., Zeina, M., Lumbers, R.T., Mielke, J., Gill, S., Kotecha, D., Freitag, D.F., Denaxas, S. and Hemingway, H. 2021. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. *BMC Medicine*, 19: 1-14.

Bellani, C. 2019. Predictive churn models in vehicle insurance. Doctoral thesis, NOVA Information Management School, Lisbon, Portugal. Available at: https://run.unl.pt/bitstream/10362/90767/1/TAA0042.pdf (Accessed 24 August 2023).

Berrone, P., Ricart, J.E., Duch, A.I., Bernardo, V., Salvador, J., Piedra Peña, J. and Rodríguez Planas, M., 2019. EASIER: An evaluation model for public–private partnerships contributing to the sustainable development goals. *Sustainability*, *11*(8): 2339.

Bhattacherjee, A. 2012. *Social science research: Principles, methods, and practices*. Tampa, FL: University of South Florida.

Bielawska, K. and Lyskawa, K. 2021. Retirees' healthcare needs and satisfaction with their coverage. *European Research Studies Journal*, 24(2): 1007-1018.

Bojanowski, P. and Joulin, A. 2017. Unsupervised learning by predicting noise. In *International Conference on Machine Learning* (pp. 517-526). PMLR.

Bonache, J. and Festing, M. 2020. Research paradigms in international human resource management: an epistemological systematisation of the field. *German Journal of Human Resource Management*, 34(2): 99-123.

Brunila, M., Zhao, R., Mircea, A., Lumley, S. and Sieber, R. 2021. Bridging the gap between supervised classification and unsupervised topic modelling for social-media assisted crisis management. *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pp. 33-49. Available: https://aclanthology.org/2021.adaptnlp-1.5.pdf (Accessed 12 April 2023).

Busse, R., Blümel, M., Knieps, F. and Bärnighausen, T. 2017. Statutory health insurance in Germany: a health system shaped by 135 years of solidarity, self-governance, and competition. *The Lancet*, 390(10097): 882-897.

Carcillo, F., Le Borgne, Y.A., Caelen, O., Kessaci, Y., Oblé, F. and Bontempi, G., 2021. Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, 557: 317-331.

Chekani, F., Zhu, Z., Khandker, R.K., Ai, J., Meng, W., Holler, E., Dexter, P., Boustani, M. and Ben Miled, Z. 2023. Modeling acute care utilization: practical implications for insomnia patients. *Scientific Reports*, 13(1): 2185.

Chicco, D. and Jurman, G.2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC*

Genomics, 21: 6. Available: https://link.springer.com/article/10.1186/s12864-019-6413-7 (Accessed 15 April 2023).

Chimusoro, A., Maphosa, S., Manangazira, P., Phiri, I., Nhende, T., Danda, S., Tapfumanei, O., Munyaradzi Midzi, S. and Nabyonga-Orem, J. 2018. Responding to cholera outbreaks in Zimbabwe: building resilience over time. *Current Issues in Global Health*, pp.45-64. Available: https://www.intechopen.com/profiles/249286 (Accessed 17 April 2023).

Choubey, R. and Gautam, P. 2020. Combined technique of supervised classifier for the credit card fraud detection. *Shodah Sarita*, 7: 27-32.

Coughlan, J. 2021. Accessible health: an evidenced based approach to improve user experience and clinical sustainability within rural healthcare. Master's dissertation, University of Nebraska-Lincoln, Lincoln, USA.

Darrat, M. and Flaherty, G.T. (2019). Retrospective analysis of older travellers attending a specialist travel health clinic. *Tropical Diseases, Travel Medicine and Vaccines,* 5(1): 1-8.

Das, H., Naik, B., Behera, H.S., Jaiswal, S., Mahato, P. and Rout, M. 2022. Biomedical data analysis using neuro-fuzzy model with post-feature reduction. *Journal of King Saud University-Computer and Information Sciences*, 34(6): 2540-2550.

Dawson, B.K., Fereshtehnejad, S.M., Anang, J.B., Nomura, T., Rios-Romenets, S., Nakashima, K., Gagnon, J.F. and Postuma, R.B. 2018. Office-based screening for dementia in Parkinson disease: the Montreal Parkinson Risk of Dementia Scale in 4 longitudinal cohorts. *JAMA Neurology*, 75(6): 704-710.

Dhieb, N., Ghazzai, H., Besbes, H. and Massoud, Y. 2019. Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE international conference on vehicular electronics and safety*: 1-5. Available: https://www.researchgate.net/profile/Najmeddine-Dhieb/publication/337508754_Extreme_Gradient_Boosting_Machine_Learning_Algorithm_For_Safe_Auto_Insurance_Operations/links/5e9de20d299bf13079ad7e6d/Extreme-Gradient-Boosting-Machine-Learning-Algorithm-For-Safe-Auto-Insurance-Operations.pdf (Accessed 20 May 2023).

Dhillon, A., Singh, A., Vohra, H., Ellis, C., Varghese, B. and Gill, S.S., 2022. IoTPulse: machine learning-based enterprise health information system to predict alcohol addiction in Punjab (India) using IoT and fog computing. Enterprise Information Systems, 16(7): 1-34.

Diamandis, P., Prassas, I. and Diamandis, E.P. 2020. Antibody tests for COVID-19: drawing attention to the importance of analytical specificity. *Clinical Chemistry and Laboratory Medicine*, 58(7). Available: https://www.degruyter.com/document/doi/10.1515/cclm-2020-0554/html (Accessed 15 April 2023).

Dong, H., Wang, X., Qiu, Y., Lou, C., Ye, Y., Feng, H., Ye, X. and Chen, D., 2023. Establishment and visualization of a model based on high-resolution CT qualitative and quantitative features for prediction of micropapillary or solid components in invasive lung adenocarcinoma. *Journal of Cancer Research and Clinical Oncology*, 1-12.

Doupe, P., Faghmous, J. and Basu, S. 2019. Machine learning for health services researchers. *Value in Health*, 22(7): 808-815. Available: https://www.sciencedirect.com/science/article/pii/S1098301519301469 (Accessed 22 May 2023).

Duckett, S. and Nemet, K. 2019. *The history and purposes of private health insurance.* Grattan Institute. Available: https://grattan.edu.au/wp-content/uploads/2019/07/918-The-history-and-purposes-of-private-health-insurance.pdf (Accessed 03 April 2022).

Dulhare, U.N. and Ghori, I. 2018. An efficient hybrid clustering to predict the risk of customer churn. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)* (pp. 673-677). IEEE.

Fenny, A.P., Yates, R. and Thompson, R. 2017. Strategies for financing social health insurance schemes for providing universal health care: a comparative analysis of five countries. *Global Health Action*, 14(1): 1868054.

Fernandes, H.E. and Ferreira, F.A., 2023. Health insurance risk assessment using cognitive mapping and multiple-criteria decision analysis. *International Transactions in Operational Research*, *30*(5): 2158-2188.

Filc, D., Rasooly, A. and Davidovitch, N. 2020. From public vs. private to public/private mix in healthcare: lessons from the Israeli and the Spanish cases. *Israel Journal of Health Policy Research*, 9(1): 1-14.

FitzPatrick, B. 2019. Validity in qualitative health education research. *Currents in Pharmacy Teaching and Learning*, 11(2): 211-217.

Francois-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G. and Pineau, J., 2018. An introduction to deep reinforcement learning. *Foundations and Trends in Machine Learning*, *11*(3-4): 219-354.

Galetsi, P. Katsaliaki, K. and Kumar, S. 2020. Big data analytics in the health sector: Theoretical framework, techniques, and prospects*: International Journal of Information Management*, 50(I): 202-216. Available: https://www.sciencedirect.com/science/article/pii/S0268401219302890 (Accessed 02 August 2021).

Giovanella, L., Mendoza-Ruiz, A., Pilar, A.D.C.A., Rosa, M.C.D., Martins, G.B., Santos, I.S., Silva, D.B., Vieira, J.M.D.L., Castro, V.C.G.D., Silva, P.O.D. and Machado, C.V. 2018. Universal health system and universal health coverage: assumptions and strategies. *Ciencia & Saude Coletiva*, 23: 1763-1776.

Goertzen, M.J. 2017. Introduction to quantitative research and data. *Library Technology Reports*, 53(4). Available: https://journals.ala.org/index.php/ltr/article/view/6325 (Accessed 19 October 2022).

Gottlieb, E.R., Samuel, M., Bonventre, J.V., Celi, L.A. and Mattie, H. 2022. Machine learning for acute kidney injury prediction in the intensive care unit. *Advances in Chronic Kidney Disease*, 29(5): 431-438.

Grandini, M., Bagli, E. and Visani, G. 2020. Metrics for multi-class classification. Available: https://arxiv.org/pdf/2008.05756.pdf (Accessed 15 April 2023).

Gu, Y., Zalkikar, A., Liu, M., Kelly, L., Hall, A., Daly, K. and Ward, T. 2021. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Scientific Reports*, 11(1): 18961.

Haag, F., Hopf, K., Vasconcelos, P.M. and Staake, T. 2022. Augmented cross-selling through explainable AI--a case from energy retailing. *arXiv preprint arXiv:2208.11404*. Available: https://arxiv.org/abs/2208.11404 (Accessed 27 July 2021).

Hagiwara, Y., Harada, K., Nealon, J., Okumura, Y., Kimura, T. and Chaves, S.S. 2022. Seasonal influenza, its complications and related healthcare resource utilization among people and older: A descriptive retrospective study in Japan. *Plos one*, 17(10): p.e0272795.

Hammah, C.A. 2020. A customer retention strategy for phoenix insurance company. Ashesi University. Available: https://air.ashesi.edu.gh/server/api/core/bitstreams/3977478e-954a-4d10-927b-1af0e3a16bc0/content (Accessed 27 July 2021).

Hanif, E. 2019. Applications of data mining techniques for churn prediction and cross-selling in the telecommunications industry. Master's dissertation, Dublin Business School, Dublin, Ireland. Available: https://esource.dbs.ie/handle/10788/3709 (Accessed 01 June 2022).

Harland, C., Telgen, J., Callender, G., Grimm, R. and Patrucco, A. 2019. Implementing government policy in supply chains: an international coproduction study of public procurement. *Journal of Supply Chain Management*, 55(2): 6-25.

Hayashi, P., Abib, G. and Hoppen, N. 2019. Validity in qualitative research: a processual approach. *The Qualitative Report*, 24(1): 98-112.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D.J. and Long, Q. 2018. Predicting academic performance by considering student heterogeneity. *Knowledge Based Systems,* 161: 134-146. Available: http://4llab.net/publication/1-s2.0-S0950705118303939-main.pdf (Accessed 04 April 2023).

Hsieh, W.T., Chien, T.W., Kuo, S.C. and Lin, H.J. 2020. Whether productive authors using the national health insurance database also achieve higher individual research metrics: a bibliometric study. *Medicine*, 99(2).

Hung, M., Lauren, E., Hon, E. S., Birmingham, W. C., Xu, J., Su, S., Hon, S. D., Park, J., Dang, P. and Lipsky, M. S. (2020). Social network analysis of COVID-19

sentiments: application of artificial intelligence. *Journal of Medical Internet Research*, 22(8): 3-32.

Ihre, A. and Engstrom, I. 2019. *Predicting house prices with machine learning methods*. Available at: https://www.diva-portal.org/smash/get/diva2:1354741/FULLTEXT01.pdf (Accessed September 10, 2023).

Iranmanesh, S.H., Hamid, M., Bastan, M., Hamed Shakouri, G. and Nasiri, M.M. 2019. Customer churn prediction using artificial neural network: an analytical CRM application. In *Proceedings of the International Conference on Industrial Engineering and Operations Management, Pilsen, Czech Republic* (pp. 23-26).

Jain, N.K., Kamboj, S., Kumar, V. and Rahman, Z., 2018. Examining consumer-brand relationships on social media platforms. *Marketing Intelligence & Planning*, *36*(1): pp.63-78.

Jansen, S. 2018. *Hands-on machine learning for algorithmic trading: design and implement investment strategies based on smart algorithms that learn from data using Python*. Birmingham: Packt Publishing.

Joarder, T. Chaudhury, T. and Mannan. 2019. Universal health coverage in Bangladesh: activities, challenges, and suggestions. *Advances in Public Health.* Available: https://www.hindawi.com/journals/aph/2019/4954095/ (Accessed 07 April 2022).

Johannesson, P. and Perjons, E. 2014. *An introduction to design science*. Cham: Springer International, pp. 39-73.

Johny, C.P. and Mathai, P.P. 2017. Customer churn prediction: a survey. *International Journal of Advanced Research in Computer Science*, 8(5): 2178-2181.

Jonnalagadda, V., Gupta, P. and Sen, E. 2019. Credit card fraud detection using random forest algorithm. *International Journal of Advance Research, Ideas and Innovations in Technology*, 5(2): 1-5.

Joshi, S. and Nair, M.K., 2018. Survey of Classification Based Prediction Techniques in Healthcare. *Indian journal of science and Technology*, 11(15), 1-19.

Karaca-Mandic, P., Feldman, R. and Graven, P. 2018. The role of agents and brokers in the market for health insurance. *Journal of Risk and Insurance*, 85(1): 7-34.

Kathirgamanathan, B. and Cunningham, P., 2020. A feature selection method for multi-dimension time-series data. In *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6* (pp. 220-231). Springer International Publishing.

Katurura, M.C. and Cilliers, L. 2018. Electronic health record system in the public health care sector of South Africa: a systematic literature review. African *Journal of Primary Health & Family Medicine,* 10(1): 1746. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6295973/ (Accessed 01 April 2022).

Katuu, S. 2018. Healthcare systems: typologies, framework models, and South Africa's health sector. *International Journal of Health Governance*, 23(2): 134-148.

Kaushik, V. and Walsh, C.A. 2019. Pragmatism as a research paradigm and its implications for social work research. *Social Sciences*, 8(9): 255.

Kennedy, N., Brodbelt, D.C., Church, D.B. and O'Neill, D.G. 2019. Detecting false-positive disease references in veterinary clinical notes without manual annotations. *NPJ Digital Medicine*, 2(1): 33.

Kiri, V.A. and Ojule, A.C. 2020. Electronic medical record systems: a pathway to sustainable public health insurance schemes in sub-Saharan Africa. *Nigerian Postgraduate Medical Journal*, 27(1): 1-7.

Kirti, D. and Shin, M.Y. 2020. *Impact of COVID-19 on insurers*. International Monetary Fund Special Series on Covid-19. Washington DC: IMF.

Koutsomitropoulos, D.A. and Kalou, A.K. 2017. A standards-based ontology and support for big data analytics in the insurance industry. *ICT Express*, 3(2): 57-61. Available: https://www.sciencedirect.com/science/article/pii/S2405959517300875 (Accessed 01 July 2022).

Kulkarni A., Chon, D. and Batarseh, F. A. 2020. Foundations of data imbalance and solutions for a data democracy In: Batarseh, F. A. and Yang, R. eds. *Data democracy.*

Cambridge, MA: Academic Press. Available: https://doi.org/10.1016/B978-0-12-818366-3.00005-8 (Accessed 15 April 2023).

Kumar, D.P., Amgoth, T. and Annavarapu, C.S.R. 2019. Machine learning algorithms for wireless sensor networks: a survey. *Information Fusion*, 49: 1-25.

Kumar, N., Srivastava, J.D. and Bisht, H. 2019. Artificial intelligence in insurance sector. *Journal of The Gujarat Research Society,* 21.

Kumar, R. and Duggirala, A. 2021. Health insurance as a healthcare financing mechanism in India: key strategic insights and a business model perspective. *Vikalpa*, 46(2): 112-128.

Lappeman, J., Franco, M., Warner, V. and Sierra-Rubia, L. 2022. What social media sentiment tells us about why customers churn. *Journal of Consumer Marketing*, 39(5): 385-403.

Larzelere, S.P., 2021. The Cognitive and Emotional Reactions of Commercial Casualty Insurance Underwriters to the Use of Predictive Analytics. Available at: https://repository.upenn.edu/entities/publication/91dcd0ba-e2c4-4de8-8fa8-e7d091728965 (Accessed: 18 September 2023).

Li, J., Varnfield, M., Jayasena, R. and Celler, B., 2021. Home telemonitoring for chronic disease management: Perceptions of users and factors influencing adoption. *Health Informatics Journal*, 27(1): 1460458221997893.

Lico, L. and Enesi, I. 2021. Performance analysis of and neural KNN networks for predicting customer purchases in a real retail department store. Available: https://ieeexplore.ieee.org/abstract/document/9461316 (Accessed 05 April 2023).

Mahesh, B. 2018. Machine learning algorithms. Available:https://www.researchgate.net/profile/Batta-Mahesh/publication/344717762_Machine_Learning_Algorithms_A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf?eid=5082902844932096 (Accessed 05 August 2022).

Markova, E., Bajtos, T., Sokol, P., Mezesova, T. and Pekarčík, P. 2021. Malicious emails classification based on machine learning. In *Proceedings of the Computational Methods in Systems and Software* (pp. 797-810). Cham: Springer International.

Markovic, D.I. 2020. The relationship between civilian and military health insurance: The condition for more efficient healthcare. *Vojno delo*, 72(1): 71-88.

McDonald, N., Schoenebeck, S. and Forte, A. 2019. Reliability and inter-rater reliability in qualitative research: norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1-23.

McEvoy, C.R., Xu, H., Smith, K., Etemadmoghadam, D., San Leong, H., Choong, D.Y., Byrne, D.J., Iravani, A., Beck, S., Mileshkin, L. and Tothill, R.W., Nasteski, V. 2019. Profound MEK inhibitor response in a cutaneous melanoma harboring a GOLGA4-RAF1 fusion. *The Journal of Clinical Investigation*, 129(5): 1940-1945.

McKinney, V., Swartz, L. and McKinney, E.L.M. 2020. COVID-19, disability and the context of healthcare triage in South Africa: Notes in a time of pandemic. *African Journal of Disability*, 9(1): 1-9.

Michel, J., Obrist, B., Bärnighausen, T., Tediosi, F., McIntyre, D., Evans, D. and Tanner, M., 2020. What we need is health system transformation and not health system strengthening for universal health coverage to work: Perspectives from a National Health Insurance pilot site in South Africa. *South African Family Practice*, *62*(3).

Mikalef, P., Krogstie, J., Pappas, I.O. and Pavlou, P., 2020. Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities. *Information & Management*, *57*(2): p.103169.

Misra, P. and Yadav, A. 2019. Impact of preprocessing methods on healthcare predictions. Available: https://www.researchgate.net/publication/332436103_Impact_of_Preprocessing_Methods_on_Healthcare_Predictions (Accessed 21 May 2023).

Mohr, E., Snyman, L., Mbakaz, Z., Caldwell, J., DeAzevedo, V., Kock, Y., Trivino Duran, L. and Venables, E. 2018. “Life continues”: patient, health care and community

care workers perspectives on self-administered treatment for rifampicin-resistant tuberculosis in Khayelitsha, South Africa. *PLoS One*, 13(9): e0203888.

Moreau, C., Pichault, F. and Mertens de Wilmars, S. 2014. Beyond the limits of a deductive approach based on ideal types and configurations. In *30th EGOS Colloquium*.

Napitu, J.J. and Putri, Y.R. 2018. Strategi Komunikasi Pemasaran Pt. Dirgantara Indonesia Dalam Memasarkan Pesawat Terbang N219. *eProceedings of Management*, 5(3).

Nasteski, V.2017. An overview of the supervised machine learning methods. *Horizons B*, 4: 51-62. Available: https://www.researchgate.net/profile/Vladimir Nasteski/publication/328146111_An_overview_of_the_supervised_machine_learning_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf (Accessed 01 August 2022).

Nayak, B. Bhattacharyya, S.S. and Krishnamoorthy, B. 2019. Integrating wearable technology products and big data analytics in business strategy. a study of health insurance firms. Available: https://www.emerald.com/insight/content/doi/10.1108/JSIT-08-2018-0109/full/html (Accessed 01 May 2022).

Nickerson, C. 2022. Interpretivism paradigm and research philosophy. *Journal of Systems and Information Technology*, 21(2): 255-275. Available: https://simplysociology.com/interpretivism-paradigm.html (Accessed 22 October 2022).

Nisbet, R., Miner, G. and Yale, K. 2018. The data mining and predictive analytic process. *Handbook of statistical analysis and data mining applications*. Cambridge, MA.: Academic Press, pp. 39-54. Available: https://www.sciencedirect.com/topics/mathematics/deductive-approach (Accessed 23 October 2022).

Nkolele, R. and Wang, H. 2021. Explainable machine learning: a manuscript on the customer churn in the telecommunications industry. *In 2021 Ethics and Explainability for Responsible Data Science* (EE-RDS) (pp. 1-7). IEEE.

Ozdemir, Y.E and Bayrakli, S. 2022. A case study on building a cross-selling model through machine learning in the insurance industry. *European Journal of Science and Technology*, 35: 364-372. Available: https://dergipark.org.tr/en/pub/ejosat/article/895069 (Accessed 20 May 2023).

Palacio-Niño, J.O. and Berzal, F. 2019. Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*.

Park, D., Bahrudin, F. and Han, J. 2020. Circular reasoning for the evolution of research through a strategic construction of research methodologies. *International Journal of Quantitative and Qualitative Research Methods,* 8(3): 1-23.

Park, Y.S, Konge, L. and Artino, A.R. 2020. The positivism paradigm of research. *Academic medicine*, 95(5): 690-694. Available: https://hsrc.himmelfarb.gwu.edu/cgi/viewcontent.cgi?article=1075&context=smhs_hs_facpubs (Accessed 20 October 2022).

Paynton, S. T. and Hahn, L. K. 2021. *Critical theories paradigm*. Available: https://socialsci.libretexts.org/Bookshelves/Communication/Introduction_to_Communication/Introduction_to_Communication_(Paynton_and_Hahn)/05%3A_Communication_Theory/5.09%3A_Critical_Theories_Paradigm (Accessed 23 October 2022).

Perera, W. 2022. Prediction of interest for motor insurance products based on machine learning approaches using information on existing health insurance customers support cross selling opportunities. Master's dissertation, Robert Gordon University, Aberdeen, Scotland. Available: http://dlib.iit.ac.lk/xmlui/handle/123456789/1401 (Accessed 18 October 2022).

Portugal, I., Alencar, P. and Cowan, D. 2018. The use of machine learning algorithms in recommender systems: a systematic review. *Expert Systems with Applications*, 97: 205-227.

Prakash, A.C., Pandey, M.K. and Pareek, M. 2018. Wearables technology: awareness, adoption and applications in Indian health insurance industry. *Turkish Journal of Physiotherapy and Rehabilitation*, 32(3).

Rajesh, P.G. and Dr. K. Vijayakumar, 2021. A study on customer's insight towards cross selling practices of ICICI Bank in Thrissur district, Kerala. *Journal of Xi'an Shiyou University, Natural Science Edition*. 17(09), 193-203.

Ramamoorthy, R. and Kumar, S.A. 2018. A study on innovative strategies to growth of Indian health insurance sector. *SAARJ Journal on Banking & Insurance Research*, 7(4): 19-25.

Rapport, F. and Braithwaite, J. 2018. Are we on the cusp of a fourth research paradigm? Predicting the future for a new approach to methods-use in medical and health services research. *BMC Medical Research Methodology*, 18: 1-7.

Raschka, S. 2018. MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, 3(24): 638.

Rawat, S., Rawat, A., Kumar, D. and Saitha, A.S. 2021. Application of machine learning and data visualization techniques for decision support in the insurance sector. Available: https://www.sciencedirect.com/science/article/pii/S2667096821000057 (Accessed 21 May 2023).

Ray, S. 2019. A quick review of machine learning algorithms. Available: https://translateyar.ir/wp-content/uploads/2021/12/A-Quick-Review-of-Machine.pdf (Accessed 21 May 2023).

Richens, J.G., Lee, C.M. and Johri, S. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications,* 11. Available: https://www.nature.com/articles/s41467-020-17419-7 (Accessed 29 March 2023).

Rottmann, M., Maag, K. and Chan, R. 2019. Detection of false positive and false negative samples in semantic segmentation. Available: https://arxiv.org/pdf/1912.03673.pdf (Accessed 15 April 2023).

Rusdah, D.A. and Murfi, H. 2020. XGBoost in handling missing values for Life Insurance risk prediction. *SN Applied Sciences,* 2. Available: https://link.springer.com/article/10.1007/s42452-020-3128-y (Accessed 03 April 2023).

Rusilowati, U. and Pratiwi, A. 2022. Lecturers' real Contributions to a resilient Indonesia in the era of society 5.0. *Scientific Journal Of Reflection: Economic, Accounting, Management and Business*, 5(4): 877-890.

Saghir, M., Bibi, Z., Bashir, S. and Khan, F.H. 2019. Churn prediction using neural network based individual and ensemble models. In *16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (pp. 634-639). IEEE.

Sarkar, S.S., Sheikh, K.H., Mahanty, A., Mali, K., Ghosh, A. and Sarkar, R., 2021. A harmony search-based wrapper-filter feature selection approach for microstructural image classification. *Integrating Materials and Manufacturing Innovation*, 10: 1-19.

Schneeweiss, S., Rassen, J.A., Brown, J.S., Rothman, K.J., Happe, L., Arlett, P., Dal Pan, G., Goettsch, W., Murk, W. and Wang, S.V. 2021. Graphical depiction of longitudinal study designs in health care databases. *Annals of Internal Medicine*, 170(6): 398-406.

Sekoroglu, A.G. 2021. Impacts of feature selection techniques in machine learning algorithms for cross selling: a comprehensive study for insurance industry. Available: https://www.researchgate.net/publication/353072980_Impacts_of_Feature_Selection_Techniques_in_Machine_Learning_Algorithms_for_Cross_Selling_A_Comprehensive_Study_for_Insurance_Industry (Accessed 21 May 2023).

Shahbazi, Z. and Byun, Y. 2020. Product recommendation based on content-based filtering using XGBoost classifier. Available: https://www.researchgate.net/publication/342864588_Product_Recommendation_Based_on_Content-based_Filtering_Using_XGBoost_Classifier (Accessed 04 April 2023).

Shaikh, R., Rafi, M., Mahoto, N.A., Sulaiman, A. and Shaikh, A., 2023. A filter-based feature selection approach in multilabel classification. *Machine Learning: Science and Technology*, 4(4): 045018.

Shemshurenko, O.V., Nizamieva, L.R., Nazarova, G.I. and Broussois, G. 2018. Inductive approach when forming informative and practical autonomy in Fl Training. *Turkish Online Journal of Design, Art & Communication*, 8.

Shilpa, M., Shivakumar, M.R., Hamritha, S., Ajay Kumar, V.G. and Shreyansh, S. 2021. Analysis of lean six sigma implementation indicators in health care sector—a customer perspective. In *Advances in Industrial and Production Engineering: Select Proceedings of FLAME 2020* (pp. 227-235). Singapore: Springer.

Sidorowicz, T., Peres, P. and Li, Y. 2022. A novel approach for cross-selling insurance products using positive unlabelled learning. Available: https://ieeexplore.ieee.org/abstract/document/9892762 (Accessed 20 May 2023).

Sileyew, K.J. 2019. Research design and methodology. In: Abu-Taieh, E., Mouatasim, A.E and Al Hadid, I.H. eds. *Cyberspace*. IntechOpen, pp.1-12.

Solanki, G.C., Daviaud, E., Fawcus, S. and Cornell, J.E. 2020. Caesarean section rates in South Africa: a case study of the health systems challenges for the proposed National Health Insurance. *South African Medical Journal*, 110(8): 747-750.

Subramani, S.S, Shakeel, M.P, Bin Mohd Aboobaider, B. and Salahuddin, L.B. 2022. Classification learning assisted biosensor data analysis for preemptive plant disease detection. *ACM Transactions on Sensor Networks*. https://doi.org/10.1145/3572775

Sumartha, P. S. and Samopa, F. 2017. Cross-selling product bundling based on customer satisfaction: study case meat & food supplier X. *International Journal of Educational Research*, 5(1): 241-252.

Syed, N.F., Baig, Z., Ibrahim, A. and Valli, C. 2020. Denial of service attack detection through machine learning for the IoT. *Journal of Information and Telecommunication*, 4(4): 482-503.

Taha, A. Cosgrave, B. and Mckeever, S. 2022. Using feature selection with machine learning for generation of insurance insights. *Applied Sciences*, 12(6): 3209. Available: https://www.mdpi.com/2076-3417/12/6/3209/htm (Accessed 22 June 2022).

Taylor, H.A., Simmons, K.J., Clavane, E.M., Trevelyan, C.J., Brown, J.M., Przemyłska, L., Watt, N.T., Matthews, L.C. and Meakin, P.J. 2022. PTPRD and DCC are novel BACE1 substrates differentially expressed in Alzheimer's disease: a data mining and bioinformatics study. *International Journal of Molecular Sciences*, 23(9): 4568.

Thennakoon, A., Miharanga, S. and Kuruwitaarachchi, N. 2019. Real-time credit card fraud detection using machine learning. Available: https://www.researchgate.net/publication/334761474_Real-time_Credit_Card_Fraud_Detection_Using_Machine_Learning (Accessed 02 April 2023).

Udawant, P. and Srinath, P. 2019. Diseased portion classification & recognition of cotton plants using convolution neural networks. *International Journal of Engineering and Advanced Technology*, 8(6): 3492-3496.

Van der Walt, J.L. 2020. Interpretivism-constructivism as a research method in the humanities and social. *International Journal of Philosophy and Theology*, 8(1): 59-68. Available: http://ijptnet.com/journals/ijpt/Vol_8_No_1_June_2020/5.pdf (Accessed 22 October 2022).

Van Deventer, M. 2022. Differences in South African Generation Y banking consumers' bank identification and selected brand personality dimensions. Conference: *Social Sciences International Research Conference (SSIRC)* 2022, Mauritius.

Vandrangi, S.K. 2022. Predicting the insurance claim by each user using machine learning algorithms. *Journal of Emerging Strategies in New Economics*, 1(1): 1-11.

Vujovic, Z. 2021. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6): 599-606.

Walia, B., Shridhar, A., Arasu, P. and Singh, G.K., 2021. US physicians' perspective on the sudden shift to telehealth: survey study. *JMIR human factors*, *8*(3): e26336.

Wang, H.D. 2020. Research on the features of car insurance data based on machine learning. *Procedia Computer Science,* 166: 582-587. Available: https://www.sciencedirect.com/science/article/pii/S1877050920301381 (Accessed 05 September 2020).

Wang, Q., Chen, Z., Wang, Y. and Qu, H. 2020. Applying machine learning advances to data visualization. Available: https://www.researchgate.net/profile/Yong-Wang-149/publication/346555391_Applying_Machine_Learning_Advances_to_Data_Visualization_A_Survey_on_ML4VIS/links/603cd29e92851c4ed5a5590d/Applying-

Machine-Learning-Advances-to-Data-Visualization-A-Survey-on-ML4VIS.pdf (Accessed 22 May 2023).

Wang, S., 2023. Prediction of Health Insurance Cross-Selling Through Binary Logical Regression and Decision Tree Model. *Highlights in Science, Engineering and Technology*, 78: 172-184.

Warner, J.J., Benjamin, I.J., Churchwell, K., Firestone, G., Gardner, T.J., Johnson, J.C., Ng-Osorio, J., Rodriguez, C.J., Todman, L., Yaffe, K. and Yancy, C.W. 2020. Advancing healthcare reform: the American Heart Association's 2020 statement of principles for adequate, accessible, and affordable health care: a presidential advisory from the American Heart Association. *Circulation*, 141(10): e601-e614.

Wen, Chen, Ke Gao, Yuanzhi Xiao and Nandapala, K. 2021. Data-driven market segmentation in insurance industry and other related sectors. *Journal of Finance and Accounting,* 9(6): 268-272.

White, S., McAllister, I. and Munro, N. 2017. Economic inequality and political stability in Russia and China. *Europe-Asia Studies*, 69(1): 1-7.

Wilke, C.O. 2019. *Fundamentals of data visualization: a primer on making informative and compelling figures*. Sebastopol, CA: O'Reilly Media.

Yousefli, Z., Nasiri, F. and Moselhi, O. 2017. Healthcare facilities maintenance management: a literature review. *Journal of Facilities Management*, 15(4): 352-375.

Zade, B.M.H and Mansouri, N. 2021. PPO: a new nature-inspired metaheuristic algorithm based on predation for optimization. *Soft Computing*, 26(3): 1331-1402.

Zhang, R. Q., Yi, M., Wang, Q. Q. and Xiang, C. 2018. Polynomial algorithm of inventory model with complete backordering and correlated demand caused by cross-selling. *International Journal of Production Economics*, 199: 193-198.

Zhang, X. and Nie, H. 2021. Public health insurance and pharmaceutical innovation: Evidence from China. *Journal of Development Economics*, 148: 102578.

Zhou, Q. and Li, Z. 2017. A comparative study of various supervised learning approaches to selective omission in a road network. *The Cartographic Journal*, 54(3): 254-264.

Appendix A

Company logo is anonymous

Date: 19 September 2022

Company email address is anonymous

Email: KhulekaniM@[redacted]

Dear: Khulekani Mavundla

Company name is anonymous

RE: REQUEST TO UTILIZE [redacted] DATA FOR MASTER'S DEGREE RESEARCH

We refer to your request dated 15 September 2022 pertaining to use of information belonging to [redacted] for purposes of gathering information for your master's degree thesis.

Company name is anonymous

In my capacity as Information Officer, this letter serves to inform you that your request has been considered and granted for deidentified data to be provided to you for purposes of your master's degree research for your thesis.

All information provided to you is to be strictly treated in accordance with the Protection of Personal Information Act, 2013 and any other applicable legislation.

In the event you have any questions please feel free to contact the writer on [redacted].

Company email address is anonymous

Kind Regards,

[redacted] Company Compliance Manager name is anonymous

Senior: Compliance Manager

Company address and directors information are anonymous

Appendix B

2. Importing Libraries

[2]:
```
# The step of importinf Libraries wich are the collections of code that provide specific functionality.
# Such as Pandas which is a popular open-source data manipulation and analysis library for the Python

# Basic
import numpy as np
import pandas as pd

# Plotation
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Evaluation Metrics
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import log_loss

# Hyper Parameter Tuning
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.experimental import enable_halving_search_cv
from sklearn.model_selection import HalvingRandomSearchCV
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Miscellaneous
import time
from sklearn.preprocessing import MinMaxScaler
from sklearn.feature_selection import mutual_info_classif
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings('ignore')
```

2.0.1. Importing DataSets

[3]:
```
train = pd.read_csv("C:\KhulekaniM\Health Insurance Cross-Selling Dataset\HealthInsurance_Cross_Selling_Dataset_Train.csv")
```

3. Summarize dataset

[5]:
```
#Training Dataset column names
train.columns
```

[5]:
```
Index(['ID', 'Gender', 'Age', 'RegionCode', 'RaceCode', 'PreviouslyInsured',
       'InitialSumAssured', 'CurrentSumAssured', 'MonthlyIncome',
       'MonthlyPremium', 'AnnualPremium', 'Vintage', 'InsurerType',
       'ProductType', 'PolicyStatusKey', 'InsuranceCondition', 'Response'],
      dtype='object')
```

[6]:
```
# Testing Dataset column names
test.columns
```

[6]:
```
Index(['ID', 'Gender', 'Age', 'RegionCode', 'RaceCode', 'PreviouslyInsured',
       'InitialSumAssured', 'CurrentSumAssured', 'MonthlyIncome',
       'MonthlyPremium', 'AnnualPremium', 'Vintage', 'InsurerType',
       'ProductType', 'PolicyStatusKey', 'InsuranceCondition'],
      dtype='object')
```

Appendix C

3.0.1. Column Descriptions

[7]:
```
from prettytable import PrettyTable
from tabulate import tabulate


#Create data
data= [["ID", "Unique ID for Customer e.g 1, 2,3, 4, anonymousing data", "Integer"],
      ["Gender", "Gender of the customer 0= Other, 1= Male, 2= Female", "Integer"],
      ["Age", "Age of the customer", "Integer"],
      ["RegionCode", "Unique code for the region of the customer", "Integer"],
      ["RaceCode", "1= Unknown, 2= White, 3= African, 4= Coloured, 5= Asian", "Integer"],
      ["PreviouslyInsured", "1= Insured previously, 0= Not Insured previously", "Integer"],
      ["InitialSumAssured", "Original fixed amount that will be paid to the nominee", "Float"],
      ["CurrentSumAssured", "Current fixed amount that will be paied to the nominee", "Float"],
      ["MonthlyIncome", "Current customer monthly income", "Float"],
      ["MonthlyPremium", "Amount customer needs to pay every month for Health Insurance", "Float"],
      ["AnnualPremium", "Amount customer needs to pay as a premium in the year for Health Insurance", "Float"],
      ["Vintage", "Number of days, customer has been associated with the company", "Integer"],
      ["InsurerType", "1= Internal Health Insurance product, 0= External Health Insurance product", "Integer"],
      ["PolicyStatus", "1= Active, 0= Inactive", "Integer"],
      ["ProductType", "1= Comprehensive cover, 2= Accident Only cover, 3= Standard cover", "Integer"],
      ["InsuranceCondition", "1= Compulsory, 0= Optional ", "Integer"],
      ["Response", "1= Interested to buy additional health insurance product, 0= Not interested", "Integer"]]

#Define header names
col_names=(["FeatureName", "Description", "DataType"])

#Display table
print(tabulate(data, headers=col_names, tablefmt="fancy_grid", showindex="always"))
```

	FeatureName	Description	DataType
0	ID	Unique ID for Customer e.g 1, 2,3, 4, anonymousing data	Integer
1	Gender	Gender of the customer 0= Other, 1= Male, 2= Female	Integer
2	Age	Age of the customer	Integer
3	RegionCode	Unique code for the region of the customer	Integer
4	RaceCode	1= Unknown, 2= White, 3= African, 4= Coloured, 5= Asian	Integer
5	PreviouslyInsured	1= Insured previously, 0= Not Insured previously	Integer
6	InitialSumAssured	Original fixed amount that will be paid to the nominee	Float
7	CurrentSumAssured	Current fixed amount that will be paied to the nominee	Float
8	MonthlyIncome	Current customer monthly income	Float
9	MonthlyPremium	Amount customer needs to pay every month for Health Insurance	Float
10	AnnualPremium	Amount customer needs to pay as a premium in the year for Health Insurance	Float
11	Vintage	Number of days, customer has been associated with the company	Integer
12	InsurerType	1= Internal Health Insurance product, 0= External Health Insurance product	Integer
13	PolicyStatus	1= Active, 0= Inactive	Integer
14	ProductType	1= Comprehensive cover, 2= Accident Only cover, 3= Standard cover	Integer
15	InsuranceCondition	1= Compulsory, 0= Optional	Integer
16	Response	1= Interested to buy additional health insurance product, 0= Not interested	Integer

Appendix D

3.0.2. Data Dimensions

[8]:
```
# showing number of rows and columns for train dataset
print('Number of Rows: {}'.format(train.shape[0]))
print('Number of Columns: {}'.format(train.shape[1]))
```

```
Number of Rows: 1000000
Number of Columns: 17
```

•••

[10]:
```
#showing first train 10 rows
#rain.head(10)
filtered_train = train[train['Response'] == 0]
# Display the first 10 rows
filtered_train.head(10)
```

[10]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey
1	2	2	54	2162	3	1	575058.650	5.637621e+05	23994.00	0.00	0.000000	10	1	3	0
2	3	1	44	2302	2	1	542714.650	7.290753e+05	51878.45	992.17	11357.517510	3	1	1	1
3	4	1	41	3610	3	1	2600187.100	2.544079e+06	0.00	479.25	5485.024304	9	1	2	1
4	5	2	60	4037	5	1	252602.740	1.366844e+05	0.00	430.22	5011.393522	4	1	3	1
5	6	1	53	1813	2	1	1457869.950	1.000000e+06	0.00	1271.11	14712.900000	1	1	3	1
6	7	1	43	163	2	1	1109078.810	7.597786e+05	24500.00	707.04	8189.997171	10	1	3	1
7	8	1	45	182	3	1	445814.880	4.458149e+05	0.00	557.30	6439.692878	15	1	1	1
8	9	2	54	7490	4	1	361963.090	2.904032e+05	0.00	376.14	4326.319064	8	1	3	1
9	10	2	56	2190	3	1	301754.350	2.523987e+05	18840.25	618.49	7106.654937	9	1	1	1
10	11	2	59	9786	3	1	361337.939	2.853393e+05	15748.00	471.27	5465.177648	2	1	3	1

[12]:
```
#Descriptive Statistics

# Count: refers to the total number of non-missing or non-null values in a variable.
# Std: refers to the standard deviation, which measures the dispersion or spread of the values around the mean.
# Min: represents the minimum value in a variable, indicating the smallest observed value in the dataset.
# 25%: The value at the "25th percentile" represents the data point below which 25% of the observations fall.
# 50%: The value at the "50th percentile" represents the middle value in a sorted dataset.
# 75%: The value at the "75th percentile" represents the data point below which 75% of the observations fall.
# Max: represents the maximum value in a variable, indicating the largest observed value in the dataset.

train.describe()
```

[12]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage
count	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1.000000e+06	1.000000e+06	1.000000e+06	1000000.000000	1000000.000000	1000000.000000
mean	500000.500000	1.477821	44.003477	3515.384350	2.981510	0.998348	6.315567e+05	5.660082e+05	2.072933e+04	551.073758	6424.854520	9.137682
std	288675.278933	0.499524	9.032324	2857.825374	0.927688	0.040611	3.794142e+05	3.421361e+05	2.627569e+04	496.273474	5657.079899	5.772057
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-9.507890e+03	-1.208491e+05	0.000000e+00	0.000000	0.000000	1.000000
25%	250000.750000	1.000000	37.000000	1459.000000	2.000000	1.000000	3.891966e+05	3.366040e+05	6.761920e+03	232.747500	2801.132873	4.000000
50%	500000.500000	1.000000	43.000000	2190.000000	3.000000	1.000000	5.544740e+05	5.145328e+05	1.686200e+04	438.510000	5114.180000	8.000000
75%	750000.250000	2.000000	50.000000	6211.000000	3.000000	1.000000	7.570060e+05	7.251750e+05	2.756188e+04	746.500000	8612.784641	14.000000
max	1000000.000000	2.000000	82.000000	9992.000000	5.000000	1.000000	5.045003e+06	3.500000e+06	2.768806e+06	7532.950000	86617.165300	30.000000

Appendix E

[13]:
```
#Prints information for Train and Test DataFrame, i.e
#DataFrame is a data structure provided by the Pandas library in Python

print(train.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 17 columns):
 #   Column               Non-Null Count    Dtype
---  ------               --------------    -----
 0   ID                   1000000 non-null  int64
 1   Gender               1000000 non-null  int64
 2   Age                  1000000 non-null  int64
 3   RegionCode           1000000 non-null  int64
 4   RaceCode             1000000 non-null  int64
 5   PreviouslyInsured    1000000 non-null  int64
 6   InitialSumAssured    1000000 non-null  float64
 7   CurrentSumAssured    1000000 non-null  float64
 8   MonthlyIncome        1000000 non-null  float64
 9   MonthlyPremium       1000000 non-null  float64
 10  AnnualPremium        1000000 non-null  float64
 11  Vintage              1000000 non-null  int64
 12  InsurerType          1000000 non-null  int64
 13  ProductType          1000000 non-null  int64
 14  PolicyStatusKey      1000000 non-null  int64
 15  InsuranceCondition   1000000 non-null  int64
 16  Response             1000000 non-null  int64
dtypes: float64(5), int64(12)
memory usage: 129.7 MB
None
```

[14]:
```
#Checking Data Type
train.dtypes
```

[14]:
```
ID                      int64
Gender                  int64
Age                     int64
RegionCode              int64
RaceCode                int64
PreviouslyInsured       int64
InitialSumAssured     float64
CurrentSumAssured     float64
MonthlyIncome         float64
MonthlyPremium        float64
AnnualPremium         float64
Vintage                 int64
InsurerType             int64
ProductType             int64
PolicyStatusKey         int64
InsuranceCondition      int64
Response                int64
dtype: object
```

Appendix F

[19]:
```
# Drop all rows with any NaN and NaT values
df = pd.DataFrame(train)
df.dropna()
print(df)
```

```
             ID  Gender  Age  RegionCode  RaceCode  PreviouslyInsured  \
0             1       2   62        4093         2                  1
1             2       2   54        2162         3                  1
2             3       1   44        2302         2                  1
3             4       1   41        3610         3                  1
4             5       2   60        4037         5                  1
...         ...     ...  ...         ...       ...                ...
999995   999996       2   45        1501         3                  1
999996   999997       2   37        2192         3                  1
999997   999998       1   53        4091         4                  1
999998   999999       2   63        2091         3                  1
999999  1000000       2   29        7784         3                  1

        InitialSumAssured  CurrentSumAssured  MonthlyIncome  MonthlyPremium  \
0               195332.40       1.389865e+05           0.00          367.23
1               575058.65       5.637621e+05       23994.00            0.00
2               542714.65       7.290753e+05       51878.45          992.17
3              2600187.10       2.544079e+06           0.00          479.25
4               252602.74       1.366844e+05           0.00          430.22
...                   ...                ...            ...             ...
999995          514593.99       4.048864e+05       14753.00          423.04
999996          402149.15       3.426325e+05           0.00          163.98
999997          436675.34       4.686650e+05       15000.00          980.92
999998          665757.32       3.560788e+05        8164.00            0.00
999999          742090.02       7.270843e+05           0.00          586.85
```

4. Data Cleaning and Pre-processing

[15]:
```
from sklearn.model_selection import train_test_split
from  sklearn.preprocessing import PowerTransformer
```

[16]:
```
#Checking for Duplicate values from Training dataset:
train[train.duplicated()]
print("No duplicate records found")
```

```
No duplicate records found
```

[17]:
```
#Checking for Null Values from training Dataset:
train.isna().sum()
print("There are no NULL values found")
```

```
There are no NULL values found
```

[18]:
```
# Checking the number of unique values in each column of the Training DataFrane
train.nunique()
```

[18]:
```
ID                    1000000
Gender                      3
Age                        63
RegionCode               1798
RaceCode                    6
PreviouslyInsured           2
InitialSumAssured       74490
CurrentSumAssured       72887
MonthlyIncome           77501
MonthlyPremium          47754
AnnualPremium           65308
Vintage                    30
InsurerType                 1
ProductType                 3
PolicyStatusKey             2
InsuranceCondition          2
Response                    2
dtype: int64
```

Appendix G

```
        AnnualPremium  Vintage  InsurerType  ProductType  PolicyStatusKey  \
0         4244.142582        8            1            1                1
1            0.000000       10            1            3                0
2        11357.517510        3            1            1                1
3         5485.024304        9            1            2                1
4         5011.393522        4            1            3                1
...               ...      ...          ...          ...              ...
999995    4800.955423        4            1            1                1
999996    1899.476192       10            1            3                1
999997   11318.550770        9            1            3                1
999998       0.000000       15            1            3                0
999999    6783.042178        4            1            1                1

        InsuranceCondition  Response
0                        1         1
1                        1         0
2                        1         0
3                        1         0
4                        1         0
...                    ...       ...
999995                   0         0
999996                   1         0
999997                   1         0
999998                   1         0
999999                   0         0

[1000000 rows x 17 columns]
```

[20]:
```
# Remove/drop any rows with Null values
train.dropna()
```

[20]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	Ins
0	1	2	62	4093	2	1	195332.40	1.389865e+05	0.00	367.23	4244.142582	8	1	1	1	
1	2	2	54	2162	3	1	575058.65	5.637621e+05	23994.00	0.00	0.000000	10	1	3	0	
2	3	1	44	2302	2	1	542714.65	7.290753e+05	51878.45	992.17	11357.517510	3	1	1	1	
3	4	1	41	3610	3	1	2600187.10	2.544079e+06	0.00	479.25	5485.024304	9	1	2	1	
4	5	2	60	4037	5	1	252602.74	1.366844e+05	0.00	430.22	5011.393522	4	1	3	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
999995	999996	2	45	1501	3	1	514593.99	4.048864e+05	14753.00	423.04	4800.955423	4	1	1	1	
999996	999997	2	37	2192	3	1	402149.15	3.426325e+05	0.00	163.98	1899.476192	10	1	3	1	
999997	999998	1	53	4091	4	1	436675.34	4.686650e+05	15000.00	980.92	11318.550770	9	1	3	1	
999998	999999	2	63	2091	3	1	665757.32	3.560788e+05	8164.00	0.00	0.000000	15	1	3	0	
999999	1000000	2	29	7784	3	1	742090.02	7.270843e+05	0.00	586.85	6783.042178	4	1	1	1	

1000000 rows × 17 columns

[21]:
```
#Count the total number of missing values
pd.isnull(train).sum().sum()
print("No missing values found")
```

```
No missing values found
```

Appendix J

[22]:
```
# Detect a list of missing values with df.isin() from Training Dataset, True means there's a missing value, False means no missing value
missing_vals=["NA", "", None, np.NaN]
missing=train.isin(missing_vals)
missing.head(10)
```

[22]:

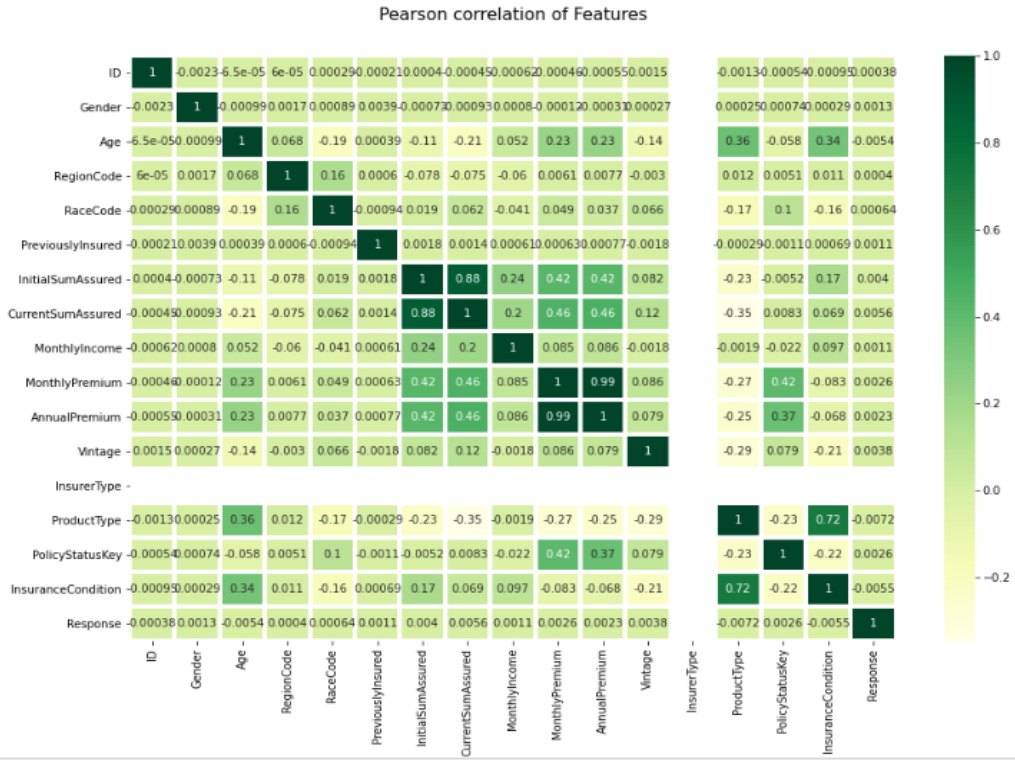
	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	InsuranceCo
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
5	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
6	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
7	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
8	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
9	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	

5. Exploratory Data Analysis and Visualization

5.0.1. Correlation Matrix for every feature with target to see which variables have relatively strong correlation

[23]:
```
# A Statistical measure that expresses the extent to which two variables are linearly related.
# A relationship between two variables in which both variables move in the same direction.
corr = train.corr()
f, ax = plt.subplots(figsize=(15,10))
sns.heatmap(corr, ax=ax, annot=True,linewidths=3,cmap='YlGn')
plt.title("Pearson correlation of Features", y=1.05, size=15)
```

[23]:
```
Text(0.5, 1.05, 'Pearson correlation of Features')
```



Appendix I

[24]:
```
#Drop all records where MonthlyIncome is zero
train = df[ df['MonthlyIncome'] == 0.00 ].index
df.drop(train, inplace = True)
df
```

[24]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	InsuranceCondition	Response
1	2	2	54	2162	3	1	575058.650	563762.1208	23994.00	0.00	0.000000	10	1	3	0	1	0
2	3	1	44	2302	2	1	542714.650	729075.3500	51878.45	992.17	11357.517510	3	1	1	1	1	0
6	7	1	43	163	2	1	1109078.810	759778.5900	24500.00	707.04	8189.997171	10	1	3	1	1	0
9	10	2	56	2190	3	1	301754.350	252398.7300	18840.25	618.49	7106.654937	9	1	1	1	0	0
10	11	2	59	9786	3	1	361337.939	285339.3300	15748.00	471.27	5465.177648	2	1	3	1	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
999993	999994	1	48	4275	4	1	975688.080	789607.6900	29945.94	1749.66	20191.540000	11	1	1	1	0	0
999994	999995	2	37	2095	3	1	439459.220	392996.6900	8920.00	190.83	2172.626473	10	1	3	1	1	0
999995	999996	2	45	1501	3	1	514593.990	404886.3900	14753.00	423.04	4800.955423	4	1	1	1	0	0
999997	999998	1	53	4091	4	1	436675.340	468665.0500	15000.00	980.92	11318.550770	9	1	3	1	1	0
999998	999999	2	63	2091	3	1	665757.320	356078.8300	8164.00	0.00	0.000000	15	1	3	0	1	0

19027 rows × 17 columns

[25]:
```
#Drop all records where MonthlyPremium is zero
train = df[ df['MonthlyPremium'] == 0.00 ].index
df.drop(train, inplace = True)
df
```

[25]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	InsuranceCondition	Response
2	3	1	44	2302	2	1	542714.650	729075.35	51878.45	992.17	11357.517510	3	1	1	1	1	0
6	7	1	43	163	2	1	1109078.810	759778.59	24500.00	707.04	8189.997171	10	1	3	1	1	0
9	10	2	56	2190	3	1	301754.350	252398.73	18840.25	618.49	7106.654937	9	1	1	1	0	0
10	11	2	59	9786	3	1	361337.939	285339.33	15748.00	471.27	5465.177648	2	1	3	1	1	0
14	15	1	31	157	4	1	823986.790	823986.79	29750.00	526.57	6055.174949	12	1	1	1	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
999992	999993	2	49	6229	3	1	957792.520	957792.52	21947.33	1663.19	19142.328380	9	1	1	1	0	0
999993	999994	1	48	4275	4	1	975688.080	789607.69	29945.94	1749.66	20191.540000	11	1	1	1	0	0
999994	999995	2	37	2095	3	1	439459.220	392996.69	8920.00	190.83	2172.626473	10	1	3	1	1	0
999995	999996	2	45	1501	3	1	514593.990	404886.39	14753.00	423.04	4800.955423	4	1	1	1	0	0
999997	999998	1	53	4091	4	1	436675.340	468665.05	15000.00	980.92	11318.550770	9	1	3	1	1	0

713546 rows × 17 columns

[26]:
```
#Drop all records where AnnualPremium is zero
train = df[ df['AnnualPremium'] == 0.00 ].index
df.drop(train, inplace = True)
df
```

[26]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	InsuranceCondition	Response
2	3	1	44	2302	2	1	542714.650	729075.35	51878.45	992.17	11357.517510	3	1	1	1	1	0
6	7	1	43	163	2	1	1109078.810	759778.59	24500.00	707.04	8189.997171	10	1	3	1	1	0
9	10	2	56	2190	3	1	301754.350	252398.73	18840.25	618.49	7106.654937	9	1	1	1	0	0
10	11	2	59	9786	3	1	361337.939	285339.33	15748.00	471.27	5465.177648	2	1	3	1	1	0
14	15	1	31	157	4	1	823986.790	823986.79	29750.00	526.57	6055.174949	12	1	1	1	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
999992	999993	2	49	6229	3	1	957792.520	957792.52	21947.33	1663.19	19142.328380	9	1	1	1	0	0
999993	999994	1	48	4275	4	1	975688.080	789607.69	29945.94	1749.66	20191.540000	11	1	1	1	0	0
999994	999995	2	37	2095	3	1	439459.220	392996.69	8920.00	190.83	2172.626473	10	1	3	1	1	0
999995	999996	2	45	1501	3	1	514593.990	404886.39	14753.00	423.04	4800.955423	4	1	1	1	0	0
999997	999998	1	53	4091	4	1	436675.340	468665.05	15000.00	980.92	11318.550770	9	1	3	1	1	0

713538 rows × 17 columns

Appendix K

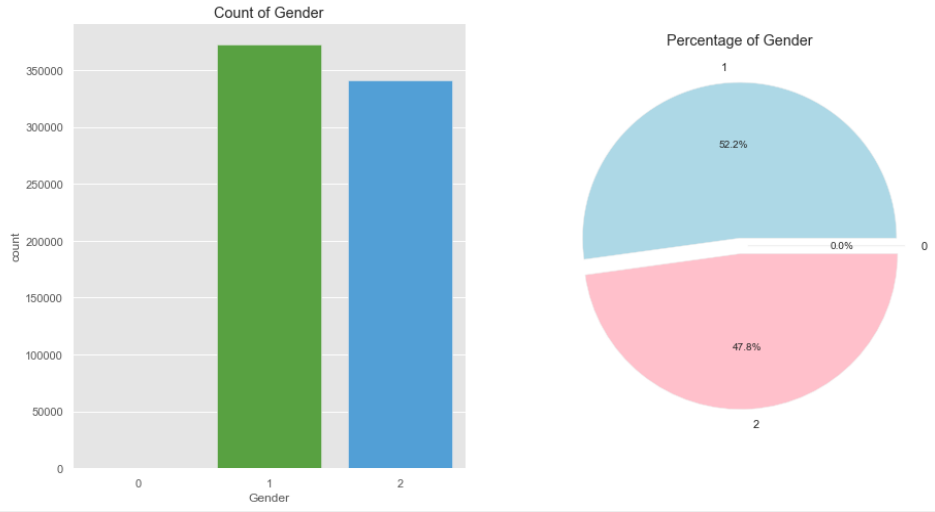
[34]:
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(15, 8))

plt.subplot(1, 2, 1)
plt.title('Count of Gender')
sns.countplot(x='Gender', data=df, palette='husl')

plt.subplot(1, 2, 2)
gender_counts = df['Gender'].value_counts()
plt.pie(gender_counts, explode=[0.05] * len(gender_counts), colors=['lightblue', 'pink'], autopct='%.1f%%', labels=gender_counts.index, labeldistance=1.1)
plt.title('Percentage of Gender')

plt.show()
```



[27]:
```
from tabulate import tabulate
# Showing number of rows and columns for train dataset after data cleaning

#create data
train = [[df.shape[0], df.shape[1]]]

#define header names
col_names = ["Number of Rows", "Number of Columns"]

#display table
print(tabulate(train, headers=col_names, tablefmt="fancy_grid"))
```

Number of Rows	Number of Columns
713538	17

Appendix L

5.0.4. Age Vs Response

[32]:
```
#Age VS Response- do a curve comparison between the blue curve and red curve
plt.figure(figsize=(15,8))
sns.countplot(x='Age',hue='Response',data=df)
```

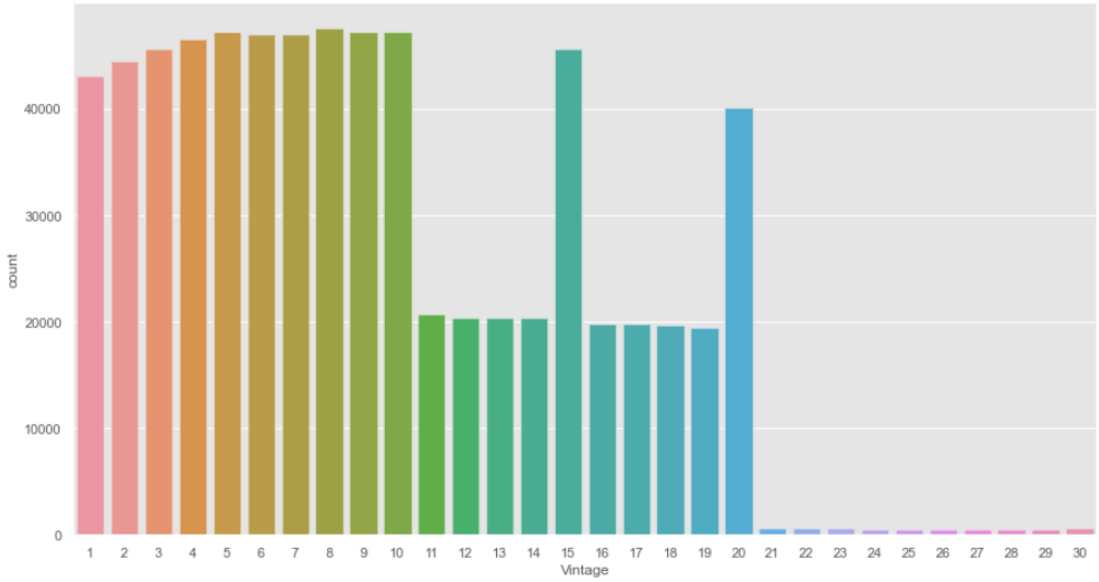
[32]:
```
<AxesSubplot:xlabel='Age', ylabel='count'>
```

5.0.5. Vintage Response

#Number of days customer has been associated with the company train['Vintage'].value_counts()

[33]:
```
#Number of years customer has been associated with the company
plt.figure(figsize=(15,8))
sns.countplot(x='Vintage',data=df)
```

[33]:
```
<AxesSubplot:xlabel='Vintage', ylabel='count'>
```



Appendix M

5.0.6. Previously Insured Vs Response

[34]:
```
#Count number of previous insured customers vs not previous insured ie. (1 previous Insured , 0 not previous insured)
df['PreviouslyInsured'].value_counts()
```

[34]:
```
PreviouslyInsured
1    712372
0      1166
Name: count, dtype: int64
```

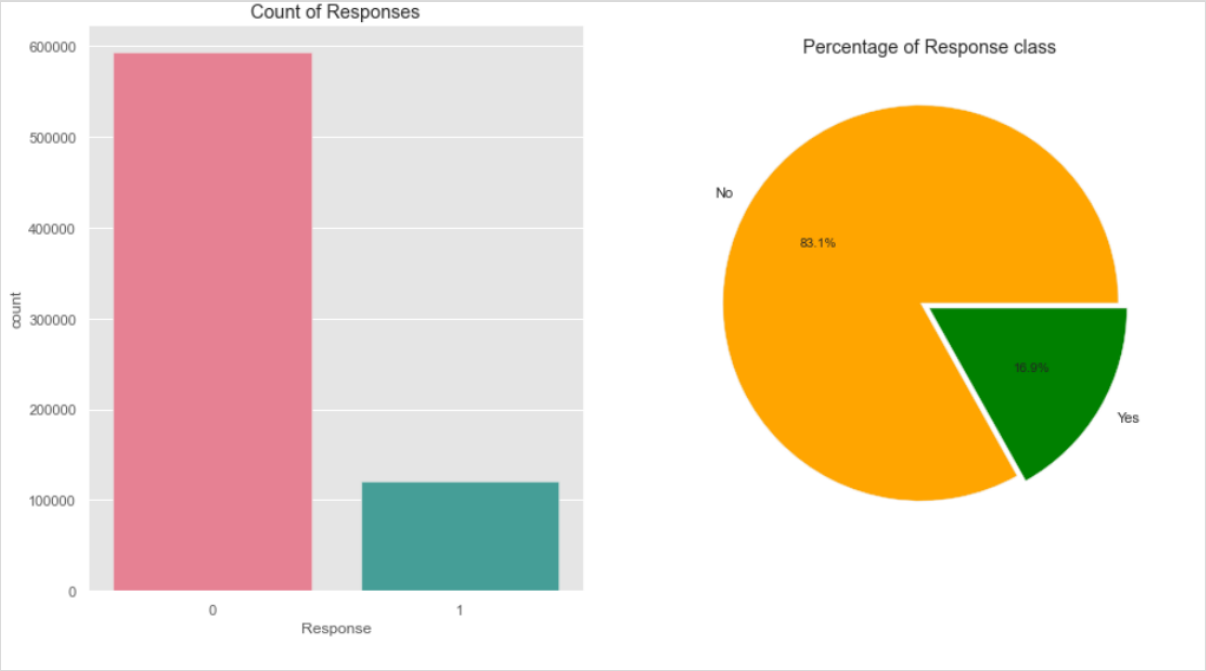
[35]:
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(15, 8))

plt.subplot(1, 2, 1)
plt.title('Count of Responses')
sns.countplot('Response', data=df, palette='husl')

plt.subplot(1, 2, 2)
plt.pie(df['Response'].value_counts(), explode=[0.05, 0], colors=['orange', 'green'], autopct='%.1f%%', labels=['No', 'Yes'], labeldistance=1.1)
plt.title('Percentage of Response class')

plt.show()
```



5.0.7. Seperating dependent and independent variables

[36]:
```
# Dependent and Independent variables

# Dependent variable - A dependent variable is the variable being tested and measured.
# Independent variable - An independent variable is a variable that is changed or controlled in order to test the effects it will have on the dependent variable.

X=df.drop(['Response'],axis=1)      #independent variable take all 17 columns and remove last comlumn "Response"

y=df['Response']                    #dependent variable take all columns and start counting from index 1 to 18 last column "Response" index 1 is on count zero
X
```

[36]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	InsuranceCondition
2	3	1	44	2302	2	1	542714.650	729075.35	51878.45	992.17	11357.517510	3	1	1	1	1
6	7	1	43	163	2	1	1109078.810	759778.59	24500.00	707.04	8189.997171	10	1	3	1	1
9	10	2	56	2190	3	1	301754.350	252398.73	18840.25	618.49	7106.654937	9	1	1	1	0
10	11	2	59	9786	3	1	361337.939	285339.33	15748.00	471.27	5465.177648	2	1	3	1	1
14	15	1	31	157	4	1	823986.790	823986.79	29750.00	526.57	6055.174949	12	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
999992	999993	2	49	6229	3	1	957792.520	957792.52	21947.33	1663.19	19142.328380	9	1	1	1	0
999993	999994	1	48	4275	4	1	975688.080	789607.69	29945.94	1749.66	20191.540000	11	1	1	1	0
999994	999995	2	37	2095	3	1	439459.220	392996.69	8920.00	190.83	2172.626473	10	1	3	1	1
999995	999996	2	45	1501	3	1	514593.990	404886.39	14753.00	423.04	4800.955423	4	1	1	1	0
999997	999998	1	53	4091	4	1	436675.340	468665.05	15000.00	980.92	11318.550770	9	1	3	1	1

713538 rows × 16 columns

Appendix N

5.0.8. Feature Selection

```
[38]: X = df.drop(['Response'], axis=1)
      y = df['Response']
      X
```

[38]:

	ID	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	InsuranceCondition
2	3	1	44	2302	2	1	542714.650	729075.35	51878.45	992.17	11357.517510	3	1	1	1	1
6	7	1	43	163	2	1	1109078.810	759778.59	24500.00	707.04	8189.997171	10	1	3	1	1
9	10	2	56	2190	3	1	301754.350	252398.73	18840.25	618.49	7106.654937	9	1	1	1	0
10	11	2	59	9786	3	1	361337.939	285339.33	15748.00	471.27	5465.177648	2	1	3	1	1
14	15	1	31	157	4	1	823986.790	823986.79	29750.00	526.57	6055.174949	12	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
999992	999993	2	49	6229	3	1	957792.520	957792.52	21947.33	1663.19	19142.328380	9	1	1	1	0
999993	999994	1	48	4275	4	1	975688.080	789607.69	29945.94	1749.66	20191.540000	11	1	1	1	0
999994	999995	2	37	2095	3	1	439459.220	392996.69	8920.00	190.83	2172.626473	10	1	3	1	1
999995	999996	2	45	1501	3	1	514593.990	404886.39	14753.00	423.04	4800.955423	4	1	1	1	0
999997	999998	1	53	4091	4	1	436675.340	468665.05	15000.00	980.92	11318.550770	9	1	3	1	1

713538 rows × 16 columns

```
[39]: # Determining x and y
      X=df[["Gender", "Age", "RegionCode", "RaceCode","PreviouslyInsured", "InitialSumAssured", "CurrentSumAssured", "MonthlyIncome", "MonthlyPremium", "AnnualPremium","Vintage", "InsurerType","ProductType","Policy
      X
```

[39]:

	Gender	Age	RegionCode	RaceCode	PreviouslyInsured	InitialSumAssured	CurrentSumAssured	MonthlyIncome	MonthlyPremium	AnnualPremium	Vintage	InsurerType	ProductType	PolicyStatusKey	InsuranceCondition
2	1	44	2302	2	1	542714.650	729075.35	51878.45	992.17	11357.517510	3	1	1	1	1
6	1	43	163	2	1	1109078.810	759778.59	24500.00	707.04	8189.997171	10	1	3	1	1
9	2	56	2190	3	1	301754.350	252398.73	18840.25	618.49	7106.654937	9	1	1	1	0
10	2	59	9786	3	1	361337.939	285339.33	15748.00	471.27	5465.177648	2	1	3	1	1
14	1	31	157	4	1	823986.790	823986.79	29750.00	526.57	6055.174949	12	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
999992	2	49	6229	3	1	957792.520	957792.52	21947.33	1663.19	19142.328380	9	1	1	1	0
999993	1	48	4275	4	1	975688.080	789607.69	29945.94	1749.66	20191.540000	11	1	1	1	0
999994	2	37	2095	3	1	439459.220	392996.69	8920.00	190.83	2172.626473	10	1	3	1	1
999995	2	45	1501	3	1	514593.990	404886.39	14753.00	423.04	4800.955423	4	1	1	1	0
999997	1	53	4091	4	1	436675.340	468665.05	15000.00	980.92	11318.550770	9	1	3	1	1

713538 rows × 15 columns

5.0.8.1. Spliting of Data into train and test Data

```
[46]: from sklearn.model_selection import train_test_split

      # Assuming X contains the features and y contains the labels

      # Split the data into training 80% and testing 20% sets
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

      # Output the shapes of the resulting datasets
      print("X_train shape:", X_train.shape)
      print("y_train shape:", y_train.shape)
      print("X_test shape:", X_test.shape)
      print("y_test shape:", y_test.shape)
```

```
X_train shape: (570830, 15)
y_train shape: (570830,)
X_test shape: (142708, 15)
y_test shape: (142708,)
```

Appendix O

6. Machine Learning Modeling Selection

- Idea is to start selection of models as:

6.1. Random Forest

6.1.0.0.1. Is a supervised machine learning algorithm which is extremely popular and is used for Classification and Regression problems in Machine Learning.

6.2. KNeighborsClassifier

6.2.0.0.1. Is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure.

6.3. XGBClassifier

6.3.0.0.1. Is a popular supervised-learning algorithm used for regression and classification on large datasets.

6.4. Logistic Regression

6.4.0.0.1. Is a Machine Learning classification algorithm that is used to predict the probability of certain classes based on some dependent variables.

```
2]: # Importing the necessary libraries for Machine Learning Models
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from xgboost import XGBClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn import model_selection

from sklearn.metrics import f1_score, recall_score, accuracy_score, roc_auc_score, precision_score, auc, roc_curve,classification_report,confusion_matrix
```

6.4.1. RandomForestClassifier Model

•••

```
[47]: # Create a Random Forest classifier model and fit it to the training data
rfc = RandomForestClassifier(n_estimators=100, random_state=42)
rfc.fit(X_train, y_train)

# Use the model to make predictions on the training set of 80%
y_pred = rfc.predict(X_train)
rfc_accuracy = rfc.score(X_train, y_train)

# Evaluate the performance of the model
print("RandomForest Accuracy score: ", accuracy_score(y_train, y_pred))
print("\nConfusion matrix:\n", confusion_matrix(y_train, y_pred))
print("\nClassification report:\n", classification_report(y_train, y_pred))
```

```
RandomForest Accuracy score:  0.9987702117968572

Confusion matrix:
 [[474506     95]
 [   607  95622]]

Classification report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    474601
           1       1.00      0.99      1.00     96229

    accuracy                           1.00    570830
   macro avg       1.00      1.00      1.00    570830
weighted avg       1.00      1.00      1.00    570830
```

Appendix P

[48]:
```
# Create a Random Forest classifier model and fit it to the training data
rfc = RandomForestClassifier(n_estimators=100, random_state=42)
rfc.fit(X_train, y_train)

# Use the model to make predictions on the testing set of 20%
y_pred = rfc.predict(X_test)
rfc_accuracy = rfc.score(X_test, y_test)

# Evaluate the performance of the model
print("RandomForest Accuracy score: ", accuracy_score(y_test, y_pred))
print("\nConfusion matrix:\n", confusion_matrix(y_test, y_pred))
print("\nClassification report:\n", classification_report(y_test, y_pred))
```

```
RandomForest Accuracy score:  0.7960520783698181

Confusion matrix:
 [[112264   6189]
 [ 22916   1339]]

Classification report:
              precision    recall  f1-score   support

           0       0.83      0.95      0.89    118453
           1       0.18      0.06      0.08     24255

    accuracy                           0.80    142708
   macro avg       0.50      0.50      0.48    142708
weighted avg       0.72      0.80      0.75    142708
```

6.4.2. KNeighborsClassifier Model

[49]:
```
# Create a KNN classifier model and fit it to the training data
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)

# Use the model to make predictions on the testing set of 80%
y_pred = knn.predict(X_train)
KNeighborsClassifier = knn.score(X_train, y_train)

# Evaluate the performance of the model
print("Accuracy score: ", accuracy_score(y_train, y_pred))
print("\nConfusion matrix:\n", confusion_matrix(y_train, y_pred))
print("\nClassification report:\n", classification_report(y_train, y_pred))
```

```
Accuracy score:  0.8600774311090867

Confusion matrix:
 [[460904  13697]
 [ 66175  30054]]

Classification report:
              precision    recall  f1-score   support

           0       0.87      0.97      0.92    474601
           1       0.69      0.31      0.43     96229

    accuracy                           0.86    570830
   macro avg       0.78      0.64      0.67    570830
weighted avg       0.84      0.86      0.84    570830
```

Appendix Q

[53]:
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# Create a KNN classifier model and fit it to the training data
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)

# Use the model to make predictions on the testing set of 20%
y_pred = knn.predict(X_test)

# Evaluate the performance of the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)

print("Accuracy score: ", accuracy)
print("\nConfusion matrix:\n", conf_matrix)
print("\nClassification report:\n", class_report)
```

```
Accuracy score:  0.7801664938195476

Confusion matrix:
 [[109395   9058]
 [ 22314   1941]]

Classification report:
              precision    recall  f1-score   support

           0       0.83      0.92      0.87    118453
           1       0.18      0.08      0.11     24255

    accuracy                           0.78    142708
   macro avg       0.50      0.50      0.49    142708
weighted avg       0.72      0.78      0.74    142708
```

6.4.3. XGBClassifier Model

...

[54]:
```
# Create an XGBoost classifier model and fit it to the training data
xgb = XGBClassifier(learning_rate=0.1, n_estimators=1000, max_depth=5, subsample=0.8, colsample_bytree=0.8, objective='binary:logistic', seed=42)
xgb.fit(X_train, y_train)

# Use the model to make predictions on the training set pf 80%
y_pred = xgb.predict(X_train)
XGBClassifier = xgb.score(X_train, y_train)

# Evaluate the performance of the model
print("Xgboost Accuracy score: ", accuracy_score(y_train, y_pred))
print("\nConfusion matrix:\n", confusion_matrix(y_train, y_pred))
print("\nClassification report:\n", classification_report(y_train, y_pred))
```

```
Xgboost Accuracy score:  0.8315750748909483

Confusion matrix:
 [[474587     14]
 [ 96128    101]]

Classification report:
              precision    recall  f1-score   support

           0       0.83      1.00      0.91    474601
           1       0.88      0.00      0.00     96229

    accuracy                           0.83    570830
   macro avg       0.85      0.50      0.46    570830
weighted avg       0.84      0.83      0.76    570830
```

Appendix R

[56]:
```
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# Assuming you have already defined X_train, X_test, y_train, y_test

# Create an XGBoost classifier model and fit it to the training data
xgb = XGBClassifier(learning_rate=0.1, n_estimators=1000, max_depth=5, subsample=0.8, colsample_bytree=0.8, objective='binary:logistic', seed=42)
xgb.fit(X_train, y_train)

# Use the model to make predictions on the testing set of 20%
y_pred = xgb.predict(X_test)

# Evaluate the performance of the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)

print("Xgboost Accuracy score: ", accuracy)
print("\nConfusion matrix:\n", conf_matrix)
print("\nClassification report:\n", class_report)
```

```
Xgboost Accuracy score:  0.8299534714241669

Confusion matrix:
 [[118436     17]
 [ 24250      5]]

Classification report:
              precision    recall  f1-score   support

           0       0.83      1.00      0.91    118453
           1       0.23      0.00      0.00     24255

    accuracy                           0.83    142708
   macro avg       0.53      0.50      0.45    142708
weighted avg       0.73      0.83      0.75    142708
```

6.4.4. Logistic Regression Model

•••

59]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# Assuming you have already defined X_train, X_test, y_train, y_test

# Create a Logistic Regression model and fit it to the training data
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

# Use the model to make predictions on the training set of 80%
y_pred_train = logreg.predict(X_train)

# Evaluate the performance of the model on the training set
train_accuracy = accuracy_score(y_train, y_pred_train)
train_conf_matrix = confusion_matrix(y_train, y_pred_train)
train_class_report = classification_report(y_train, y_pred_train)

print("Logistic Regression Accuracy score on training set: ", train_accuracy)
print("\nConfusion matrix on training set:\n", train_conf_matrix)
print("\nClassification report on training set:\n", train_class_report)
```

```
Logistic Regression Accuracy score on training set:  0.8314226652418408

Confusion matrix on training set:
 [[474601      0]
 [ 96229      0]]

Classification report on training set:
              precision    recall  f1-score   support

           0       0.83      1.00      0.91    474601
           1       0.00      0.00      0.00     96229

    accuracy                           0.83    570830
   macro avg       0.42      0.50      0.45    570830
weighted avg       0.69      0.83      0.75    570830
```

Appendix S

[60]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# Assuming you have already defined X_train, X_test, y_train, y_test

# Create a Logistic Regression model and fit it to the testing data
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

# Use the model to make predictions on the testing set of 20%
y_pred_train = logreg.predict(X_test)

# Evaluate the performance of the model on the testing set
train_accuracy = accuracy_score(y_test, y_pred_train)
train_conf_matrix = confusion_matrix(y_test, y_pred_train)
train_class_report = classification_report(y_test, y_pred_train)

print("Logistic Regression Accuracy score on training set: ", train_accuracy)
print("\nConfusion matrix on training set:\n", train_conf_matrix)
print("\nClassification report on training set:\n", train_class_report)
```

```
Logistic Regression Accuracy score on training set:  0.8300375592118171

Confusion matrix on training set:
 [[118453      0]
 [ 24255      0]]

Classification report on training set:
               precision    recall  f1-score   support

           0       0.83      1.00      0.91    118453
           1       0.00      0.00      0.00     24255

    accuracy                           0.83    142708
   macro avg       0.42      0.50      0.45    142708
weighted avg       0.69      0.83      0.75    142708
```