USING MACHINE LEARNING TO PREDICT CUSTOMER BEHAVIOUR IN ORDER TO ASSIST SMALL BUSINESSES IN UNDERSTANDING AND ADAPTING TO CUSTOMER PREFERENCES AND TENDENCIES.

By

Joy Nna Christopher

Student number: 2305576

MSc Data Science

School of Engineering, Computing & Mathematical Sciences

Module Coordinator: Dr Andrew Gascoyne

Project Supervisor: Pooja Kaur

Abstract

This dissertation explores the application of machine learning techniques to predict customer behaviour, with a specific focus on customer churn in the banking sector, providing insights relevant to small businesses. Utilizing the "Churn_Modelling" dataset from Kaggle, the study follows a thorough data exploration and preprocessing methodology to ensure data integrity and readiness for analysis. Throughout the research process, ethical considerations are carefully considered to ensure appropriate data and machine learning techniques. Key customer behaviour metrics such as demographic details, financial behaviours, and engagement indicators were identified as critical predictors of churn. Six machine learning models—Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were developed and evaluated. The study addressed class imbalance using resampling techniques, including undersampling, oversampling, and the Synthetic Minority Over-sampling Technique (SMOTE).

The results show that Random Forest and Gradient Boosting models, particularly when paired with SMOTE, achieved the highest predictive accuracy, with Random Forest reaching an accuracy of 94%. These models demonstrated balanced performance across precision, recall, and f1-score metrics, making them effective tools for predicting customer churn. Confusion matrices, which provide information about each classifier's performance, allow for the further development of predictive analytics techniques. In order to improve decision-making and customer engagement, the dissertation offers small businesses practical insights. It emphasizes the significance of tracking important customer metrics and incorporating machine learning models into CRM systems.

Despite the promising results, the study acknowledges limitations such as the dataset's specificity to the banking sector and the scope of traditional machine learning models. Future research is recommended to explore more diverse datasets, employ advanced machine learning techniques, and develop real-time prediction systems. This work underscores the potential of machine learning to empower small businesses with data-driven insights, improving their ability to adapt to customer preferences and enhance retention strategies.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

In the modern business environment, gaining a competitive edge, gaining success, maintaining growth and promoting long-term growth depend heavily on a solid understanding of consumer behaviour. Consumer behaviour is the study of individuals, groups, or organizations to determine how they choose, acquire, use, and reject goods, services, encounters, or ideas. This knowledge may be the difference between small businesses, which often operate with limited resources, face significant challenges in effectively analysing customer data to predict future behaviours and preferences. Customer loyalty and churn are included in the definition of consumer behaviour in the context of this study (Ibukun et al., 2016). Recurrent patronage behaviour, which is a combination of attitude and behaviour, is referred to as consumer loyalty (East et al., 2005). In industrial and service marketing, the term "brand retention" refers to a customer's behavioural loyalty. The loss of current customers to another company or service provider is referred to as customer churn, customer attrition, or customer turnover (Prabadevi et al., 2023).

With the introduction of digital technologies and the exponential rise in data volumes, businesses now have new ways to learn about the preferences, tendencies, and behaviours of their customers. Machine Learning (ML), a branch of artificial intelligence, stands out as a powerful tool that can analyse large datasets, uncover patterns, and make predictions with a level of accuracy and efficiency unattainable by traditional methods.

This dissertation investigates the application of machine learning techniques to predict customer behaviour, with the goal of assisting small businesses in better understanding and adapting to customer preferences and tendencies. By leveraging ML, small businesses can not only enhance their decision-making processes but also personalize their marketing efforts, improve customer retention, and ultimately boost their profitability.

1.1 Background

The way that businesses engage with their customers has been completely transformed by the digital transformation. Every online purchase, social media post, and customer service interaction produces data that, when properly examined, can offer profound understandings of consumer behaviour. Businesses that use this data can make more informed and timely decisions, giving them a significant competitive advantage (Davenport and Ronanki 2018).

However, due to limitations including low financial resources, a lack of technical know-how, and restricted access to sophisticated analytical tools, small businesses frequently find it difficult to take advantage of this data.

The traditional methods of understanding customer behaviour, such as surveys and focus groups, are increasingly inadequate in capturing the complexity and dynamism of modern consumer preferences. These techniques frequently involve a lot of time, money, and bias (Jones and Brown 2019). Machine learning algorithms, on the other hand, are able to process enormous amounts of data in real-time, spot hidden patterns, and produce highly accurate predictions. For small businesses that need to quickly adjust to shifting consumer trends and market conditions, this capability is especially helpful.

Machine learning encompasses a variety of algorithms and techniques that enable computers to learn from data and make predictions or decisions without being explicitly programmed. ML can be used to forecast outcomes like lifetime value, purchase propensity, customer churn, and product recommendations in the context of customer behaviour analysis (Amin et al., 2020). Businesses can improve customer satisfaction, optimize operations, and create more effective marketing strategies by putting these predictive models into practice. For instance, predictive models can help identify customers who are likely to churn, allowing businesses to proactively address their concerns and retain them (Chen et al., 2012). Similar to this, consumer segments can be created using machine learning and their behaviour and preferences, allowing for more specialized and tailored marketing campaigns. These applications not only improve customer satisfaction but also maximize the return on marketing investments.

Statistical methods, data mining, machine learning, artificial intelligence, and other tools are used in predictive analytics to examine past and present data and forecast future events. Based on patterns found in historical data, it aids businesses in predicting future customer behaviour. Neural networks, clustering algorithms, regression analysis, and decision trees are among the frequently employed predictive analytics techniques. For instance, regression analysis helps in understanding relationships between variables, decision trees provide clear decision paths, and neural networks excel at recognizing complex patterns within data (Amin et al., 2020, Chen et al., 2012). Small businesses can improve their decision-making procedures, target marketing campaigns more precisely, and ultimately spur growth and profitability by leveraging predictive analytics (Nguyen et al., 2020, Sun et al., 2019).

1.2 Challenges Facing Small Businesses

Despite the clear benefits, implementing machine learning technologies presents a number of difficulties for small businesses. One of the main challenges is the lack of technical expertise. Many small business owners and employees might not have a deep understanding of data science, statistics, and computer programming, which is necessary for machine learning Cui et al., 2021). Furthermore, small businesses with tight budgets may find it difficult to invest in machine learning systems due to the high initial setup and integration costs. Another significant challenge is data quality and availability. For machine learning models to produce precise predictions, a lot of high-quality data must be collected. Small businesses might find it difficult to gather and handle this kind of data, or they might not have access to it at all (Sun et al., 2019). Furthermore, in order for businesses to remain relevant, they must constantly update their models and systems due to the rapid pace of technological change. This can be a difficult task for those with limited resources.

1.3 Research Problem

The central problem addressed by this dissertation is the gap between the potential benefits of machine learning for predicting customer behaviour and the actual implementation of these technologies by small businesses. While large corporations and businesses have successfully integrated ML into their business strategies, small businesses lag behind due to various constraints, including limited access to advanced technologies and skilled personnel (Cui et al., 2021). This research aims to bridge this gap by developing and evaluating machine learning models that are tailored to the needs and capabilities of small businesses.

1.3.1 Research Aim

The aim of this research is to create a predictive analytics model that is easy to use for small and medium-sized enterprises. This model will enable efficient analysis of customer data to predict purchasing trends and tailor marketing and product strategies to address the challenges posed by digital disruption and shifting consumer behaviour.

The Main objectives of this study are:

- To identify the key customer behaviour metrics that are most relevant to small businesses.
- To create machine learning models that have the ability to predict these metrics with accuracy.
- To empirically analyse and validate these models' efficacy with real-world data.
- To provide small businesses with useful advice on how to apply and make use of these models.

1.3.2 Research Questions

To achieve these objectives, the dissertation seeks to answer the following research questions:

- What are the most critical customer behaviour metrics that small businesses should monitor to enhance their operations and marketing efforts?
- Which machine learning techniques are most effective in predicting these metrics, given the constraints faced by small businesses?
- How can small businesses integrate machine learning models into their existing operations to enhance decision-making and customer engagement?
- What are the challenges and limitations of using machine learning in the context of small businesses, and how can they be addressed?

Chapter 2: Literature Review

2.1 Consumer Behaviour

Consumer behaviour refers to the pursuit, acquisition, use and disposal of goods to fulfil needs and preferences by groups, individuals, and organisations. (Furqon et al., 2023). The Pareto Principle highlights the importance of retaining existing customers by stating that 80% of a company's revenue is generated by 20% of the customers (Adebola et al., 2019). In order to please customers and promote recurring business, marketers need to conduct market research and product development. (Ibukun et al., 2016). Experiences after a purchase are important indicators of a customer's level of satisfaction. The fact that some customers rely on personal experiences to make decisions quickly and others require additional information and interaction indicates that there are diverse levels of customer interest and product information requirements.

Numerous social, psychological, and personal factors play a role in purchase decisions (Kumar, 2018). Consumer behaviour is significantly impacted by cultural variety, leading to strong associations with products that are suitable for particular cultural identities (Quynh and Dung 2021). Online buying decisions are influenced by a variety of personal factors, such as age, stage of life, career, personality, lifestyle, financial conditions, etc. Additionally, consumer behaviour can be impacted significantly by personal factors such as inflation, job loss and budget fluctuations (Ullah et al., 2019). According to Liu et al (2013), customer attitudes, demand incentives, live media, product messaging and anchoring are all significant determinants of consumer engagement and purchasing behaviours (Li et al., 2022).

Psychological factors like perception, motivation, learning, personality traits, memory, and knowledge all have a significant impact on consumer behaviour (Li et al., 2022). Nawi et al (2022), highlights the influence of habitual behaviour and the desire to find a useful product on intentions to make an online purchase (Nawi et al., 2022). A study by Al-Ghaswyneh confirms that the largest influence on purchases comes from incentives (Quynh and Dung 2021). Studies conducted in less developed countries show that consumers consider price, culture and religious convictions when making decisions about what to buy. Additionally, cultural factors such as family, culture and peer groups have a big impact on purchasing decisions (Jian, 2019). Consumer behaviour is still erratic and varies amongst individuals contemplating the same product. All things considered, understanding the different influences on consumer behaviour is essential in creating successful marketing strategies (Jian, 2019).

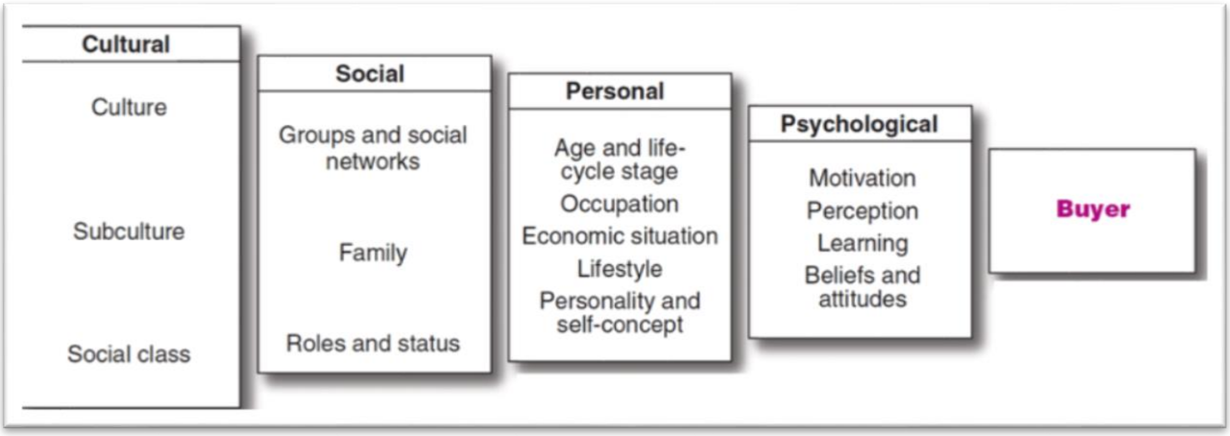


Figure 2.1: Factors influencing Consumer Behaviour. (Ibukun et al., 2016)

2.2 Data-Driven Marketing

Data-Driven Marketing refers to the process of obtaining and applying data in order to guide marketing choices and to customises the customer experience (Raj and Raman 2019). Marketers can reach the right people at the right time and place with the help of this data, which frequently focuses on the demographic and behaviours of customers (Engidaw, 2022). Businesses prioritizing the use of customer and marketing data to inform decisions and product development, as well as the adoption and integration of marketing analytics and digitization into marketing operations have led to the rise in popularity of data-driven marketing (Raj and Raman 2019). Data-driven marketing maximises client information to create a marketing plan. This involves obtaining complex data from both online and offline sources which is then analysed to develop a more comprehensive understanding of the customers (Rozak and Fachrunnisa 2021). With data gathered and analysed, marketers can better understand the psychology and purchasing patterns of the target audience and develop and implement highly customized marketing campaigns (Engidaw, 2022). As a result, companies that use data-driven marketing techniques are able to connect with their target market, which then encourages trust, loyalty and eventually, results in higher sales and steady revenue (Kim and Han 2023).

Most people would agree that a business use data more and more to plan and anticipate customer demands, technology is essential to the development of predictive models. These models can help businesses implement customer-centric strategies to win over customers (Zhang et al., 2020). Data can identify demands and influencing factors at every stage of the decision-making process. The primary objective of data-driven marketing strategies is to the analysis and integration of external and internal data to support the development of new products and services (Jocevski, 2020). In addition to helping with customer acquisition and retention, better work environment for users is ensured by this process. This tactic might ultimately lead to cost avoidance or reduction as well as increase in the company's output and efficiency (Jocevski, 2020). Because of its scope and depth, data-driven marketing has the power to fundamentally change the marketing paradigm.

The next big thing in consumer buying patterns research is neuromarketing and predictive analysis. Similar to this B2B business frequently employ data-driven marketing strategies. This will inevitably assist an organization in achieving its primary objectives (Gkrimpizi, 2023). Data-driven marketing has undergone a paradigm shift in the last 10 years due to the application of neurophysiological data to access brand equity and marketing ROI (Šostar and Ristanović 2023).

80% of global panellists in a Braverman (2015) study concurred that using data is essential for implementing marketing and advertising campaigns. Additionally, 77.4% of respondents said they were optimistic about the future of data-driven marketing. These numbers show how data-driven marketing is popular and how it can increase productivity and business growth (Braverman, 2015). Despite the benefits of using data in marketing strategies, researchers are concerned about a number of potential downsides of data-driven marketing strategy. (Andreas and Neacsu 2014). For example, there is a rise in issues with accountability, data bias, privacy, security breaches, and third part data access. Although data driven marketing is a topic with great relevance, there is lack of scientific documentation in this field (Jain, 2019).

2.3 Machine Learning (ML) and Predictive Analysis

Machine Learning is a branch of AI that allows computers team from data and improve performance on specific tasks without the need to be explicitly programmed (Chen et al., 2012). To support data-driven decision making in the business world, machine learning algorithms examine massive datasets for patterns, trends, and correlation (Chen et al., 2012). Machine learning, also known as predictive analytics in its commercial application enables researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" by utilizing its input data and extracting patterns and relationships from it.

Machine learning has become increasingly indispensable in business decisions making processes, offering unparalleled capabilities for extracting insights from complex datasets. Organisations can optimize marketing, sales, customer service, supply chain management and other aspects of their operations by utilizing machine learning algorithms (Kohavi et al., 2020). Businesses can gain a competitive edge in dynamic market environments by using machine learning powered productive analytics to forecast demand, identify market trends, and anticipate customer behaviour (Chen et al., 2012).

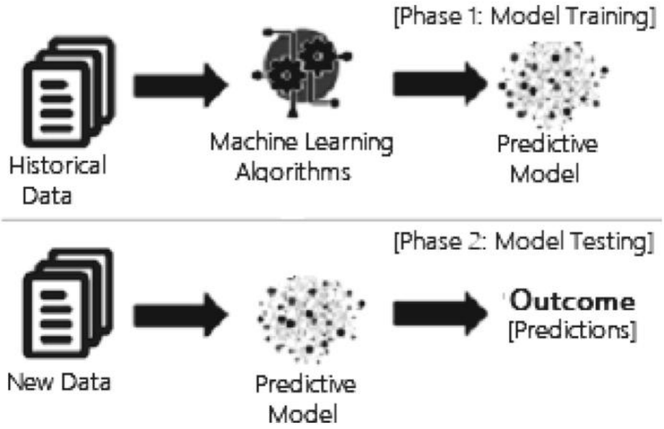


Figure 2.2.2: General structure of a machine learning based predictive model considering both the training and testing phase [Researchgate.com]

Different machine learning algorithms have been developed over time to handle different kinds of problems and data types that exist in the machine learning environment (Kohavi et al., 2020). Machine learning algorithms can be broadly categorized into sub- categories based on the kind of problem to be solved and the intended results of algorithms. These include reinforcement learning, supervised machine learning, and unsupervised machine learning. Each of which is appropriate for a particular set of business goals and data characteristics (Alpaydin 2016). For example, supervised learning models have been widely used in the retail industry to predict stock levels, customer churn rates, and to personalize marketing efforts based on individual consumer preferences (Choi et al., 2020) .

Supervised Machine Learning is a kind of machine learning in which a model is trained using a labelled dataset with a known desired output (Mohri, 2018). Regression models and classification models are two popular supervised learning algorithms that classify input data into labelled groups and predict a continuous output, respectively. Regression models, for example, can forecast consumer expenditure based on past purchase information, whereas classification models can group customers according to their purchasing patterns.

Unsupervised Machine Learning, unlike supervised learning, which works with data that has labelled responses, unsupervised learning deals with data without labelled responses, aiming to infer the natural structure present within a set of data points (Hastie, 2009). Objects in the same group (or cluster) are more similar to each other than they are to those in other groups. This is achieved through the use of clustering, a common technique in machine learning. Another unsupervised technique for lowering the number of random variables to be examined is dimensionality reduction.

Reinforcement learning: In this method, the computer is forced to solve a problem on its own by using a system of rewards and penalties. Although less obvious, its use in predictive analytics can be pivotal for creating models that maximize the idea of cumulative reward and dynamically adjust to changes(Sutton and Barto 2018).

Several studies have shown the efficacy of machine learning and predictive analytics across various fields. For example, Smith et al., (2020), applied machine learning algorithms (Using classifiers) to predict customer churn in the telecommunications industry, achieving a significant reduction in churn rates. Similarly, Jones and Brown (2019) utilized predictive analysis to forecast stock prices with high accuracy, enabling investors to make well-informed decisions. Huang et al., (2015) investigated the CCP issue in the big data platform. The study's objective was to demonstrate how the Random Forest classifier, when used with big data, greatly enhances churn prediction performance. Patel et al. (2018) demonstrated how machine learning can be used in the medical field to diagnose diseases from medical imaging data more accurately than traditional diagnostic methods. These results highlight the versatility and potential of machine learning and predictive analysis across a range of industries, emphasizing their function in fostering innovation and improving decision-making processes.

Despite its transformative potential, the widespread adoption of machine learning in business is not without challenges. Data quality issues, algorithmic biases, interpretability concerns and organizational reediness pose significant hurdles to effective implementation (Jordan and Mitchell 2015). Furthermore, machine learning applications in business context needs to be carefully examined due to ethical concerns about data privacy, transparency, and fairness (Floridi et al., 2018).

2.4 Machine Learning in Business Applications

The predictive capabilities of ML are particularly valuable in sectors like finance, where they are used for credit scoring and algorithmic trading, and in healthcare, where ML models predict patient diagnoses and outcomes based on historical health data (Kumar et al., 2021). Machine learning helps retailers with personalized marketing, chatbot-enhanced customer service, and supply chain decision optimization. Businesses can become more agile and responsive by integrating machine learning (ML) into their business processes, which enables them to react to market developments and customer demands more quickly (Kumar et al., 2021).

Retail Sector: Amazon's application of machine learning to customize shopping experiences and streamline logistics is one noteworthy example. According to Smith (2017), Amazon uses machine learning models that evaluate customer data to make product recommendations and efficiently manage inventory, resulting in a notable increase in both customer satisfaction and operational efficiency.

Healthcare Sector: By delivering early warnings based on predictive analytics, Google's DeepMind Health project has successfully used machine learning to predict acute kidney injury (AKI) hours in advance, highlighting the potential of ML to save lives in this field (Komorowski et al., 2018).

Banking and Finance: JPMorgan Chase developed a machine learning program called COiN (Contract Intelligence) to extract critical information from contracts and legal documents. This program replaced a manual process that used to take about 360,000 hours of labour per year. The company saves a ton of time and lowers the possibility of human error by automating these procedures (Faggella, 2021).

Manufacturing Sector: General Electric has optimized manufacturing processes and predicted maintenance needs by leveraging machine learning through its Predix platform. Predictive maintenance using machine learning lowers operating costs and downtime, which boosts overall productivity (Rozak and Fachrunnisa 2021).

There has been a noticeable increase in the use of machine learning (ML) in business, especially in the field of predictive analytics for customer attrition. Churn prediction attempts to pinpoint clients who are most likely to stop using a service, allowing companies to take proactive measures to keep them around. For churn prediction, a variety of machine learning models have been used, such as logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks. A comprehensive study by Huang et al. (2018) used deep neural networks to predict customer churn, achieving an accuracy of 92% and an F1 Score of 0.86. These results illustrate the potential of neural networks to capture intricate patterns within data, although they require substantial computational resources and longer training times. Additionally, in order to predict customer behaviour on social media platforms based on specific metrics and variables, such as consumer perception and attitude toward the social media site, the study by Chaudhary (2021) developed a mathematical model using machine learning. To get excellent outcomes, to find outliers, noises, mistakes, and duplicate records, the researchers employed a range of data pre-processing techniques. Training uses 80% of the data, while testing uses 20%. The best model for predicting social media user behaviour was the decision tree. Customers' differences across social media platforms ranged from a maximum of 99.51% to a minimum of 12.22%. The lowest root means square error was 20691.787, and the highest was 156556.452. The maximum overall accuracy was 0.982, and the lowest was 0.0223. These models used user likes, follows, downloads, and other data to predict user behaviour across a variety of platforms.

Additionally, the study by Li et al. (2019) compares client features and characteristics to past purchase records using machine learning techniques like cluster analysis, decision trees, and the Naive Bayes algorithm. The research then uses promotion graphs to select models with high promotion degrees, thereby achieving accurate marketing by assessing the critical factors influencing the purchasing behaviour of potential customers. The findings demonstrate that in terms of promotion and prediction, decision trees perform better than clustering analysis and the Naive Bayesian approach. If a customer is 45 to 55 years old and commutes one to two kilometres, they are more likely to make a purchase if they do not own a car or if they have one at home.

Gradient Boosting Machines (GBM) and their variants like XGBoost have shown high effectiveness in churn prediction tasks. Chen et al. (2021) applied XGBoost to a large telecom dataset and reported an accuracy of 93% and an F1 Score of 0.88. GBM works by iteratively building an ensemble of weak learners, typically decision trees, where each successive model corrects the errors of its predecessors. This boosting approach enhances predictive accuracy and reduces bias and variance. XGBoost, an optimized implementation of GBM, is particularly known for its scalability and efficiency in handling large datasets (Chen et al., 2021). The results of Li and Ma (2022) support the proposal of a combined forecasting model to predict customers' purchase behaviour and the precise timing of their transactions. It does this by merging multiple decision tree models using the Stacking technique. The study aggregates predictions from three different integrated decision tree models—Random Forest, XG Boost, and Light GBM—to accomplish this. Using a simple logistic regression classification model, a linear regression model, and the exact purchase moment, the researchers then use the aggregated prediction results to predict user behaviour with regard to purchases. Finally, the model presented in this work was validated and evaluated using actual retail sales data. The results show that the fusion model has the highest accuracy and AUC value (85% and 0.928, respectively) for predicting user behaviour. In addition, we found that the Light GBM approach outperforms the fusion model in terms of correctly estimating the user's purchase time. When the Light GBM approach and the fusion model are used concurrently at different phases of the problem, the combination outperforms the fusion model alone by 9.4% in prediction performance for both prediction problems.

2.5 Predictive Analytics in Understanding Customer Preferences

The application of predictive analytics to consumer behaviour is largely dependent on data, complex algorithms, and computational procedures. Amongst other Machine learning models, models like decision trees, neural networks, regression models are the most extensively used approaches. According to Berry and Linoff (2011), decision trees are useful for dividing the customer base into branches that each represent a decision path that can be used to forecast the probable course of a customer's behaviour. A study by Kim (2020), applied decision trees to predict customer preferences in the e-commerce industry, achieving an accuracy of 80%, a precision of 0.78, a recall of 0.76, an F1 Score of 0.77, and an AUC-ROC of 0.82. Decision trees provide clear decision rules that are easy to understand and interpret, making them useful for gaining insights into the factors influencing customer preferences. The same study reported that random forests achieved an accuracy of 85%, a precision of 0.83, a recall of 0.81, an F1 Score of 0.82, and an AUC-ROC of 0.88. This improvement is attributed to the ensemble method’s ability to reduce overfitting and enhance generalization by averaging the predictions of many decision trees (Kim, 2020). Neural networks are useful because they can extract patterns and identify trends from imprecise or complex data that are too complex for humans or other machine learning techniques to notice (Goodfellow et al., 2016). In a study by Zhang et al., (2020), a neural network model was used to predict customer preferences in the food industry, achieving an accuracy of 87%, a precision of 0.85, a recall of 0.84, an F1 Score of 0.85, and an AUC-ROC of 0.90. Based on previous interactions and demographic data, regression analysis—particularly logistic regression—is frequently used to forecast a customer's likelihood of making a purchase or leaving (Choi et al., 2020). A study by Hossain et al. (2019) employed logistic regression to predict customer preferences in the retail sector, achieving an accuracy of 78%, a precision of 0.74, a recall of 0.76, an F1 Score of 0.75, and an AUC-ROC of 0.80. This model's ease of implementation and ability to provide insights into the significance of different features make it a popular choice for businesses.

Furthermore, clustering techniques such as K-means and hierarchical clustering are used to segment customers based on behavioural data, which can predict preferences and likely future actions (Li and Kim 2020). Sarstedt and Mooi (2014) used K-Means clustering to segment customers based on their purchase behaviours in the retail industry. The study identified four distinct clusters, each representing different purchasing patterns and preferences. By analysing these clusters, businesses could tailor their marketing strategies to better meet the needs of each customer segment, although specific performance metrics like accuracy or F1 Score are not directly applicable to clustering results (Sarstedt & Mooi, 2014). Ensemble approaches have also become more popular, combining several models to increase prediction accuracy. Zhao et al., (2021), demonstrate how Random Forests can be used to analyse customer service interactions in addition to transactional data, thereby increasing the predictability of customer churn. Predictive analytics is being used extensively and with notable success in many industries to understand customer preferences.

In retail for instance, Walmart uses regression, association rule learning and time series analysis models to predict and analyse patterns from sales data, social media, and other external sources to forecast demand for products at different times, enabling Walmart to enhance its inventory management and reduce wastage (Marr, 2015). In the telecommunications sector, Verizon utilizes decision trees and logistic regression to predict customer churn. Verizon can improve customer retention rates by identifying at-risk customers and proactively offering them incentives to stick with the provider by modelling customer interactions and satisfaction levels

(Davenport, 2014). Furthermore, according to Kudyba (2020), American Express uses logistic regression and decision trees models to identify possible defaulters and predict credit risk.

Research indicates that an RNN and classical classifier combination performs better than other models (Koehn, 2020). Meanwhile, alert deep Markov models have been used to predict the probability that a user will end an online shopping session without purchasing (Ozyurt et al., 2022). Furthermore, methods such as the L1-and-L2-norm-oriented Latent Factor Model have demonstrated significant potential for recommender systems (Wu et al., 2022). These advancements provide practical tools for predicting customer behaviour and spotting potential churn risks. Additionally, the study by Li et al., (2019) compares client features and characteristics to past purchase records using machine learning techniques like cluster analysis, decision trees, and the Naive Bayes algorithm. The research then uses promotion graphs to select models with high promotion degrees, thereby achieving accurate marketing by assessing the critical factors influencing the purchasing behaviour of potential customers. The findings demonstrate that in terms of promotion and prediction, decision trees perform better than clustering analysis and the Naive Bayesian approach.

2.6 Machine Learning Techniques Specific to Small Businesses

In the context of small businesses, predictive analysis for customer churn holds particular importance due to the limited resources available for customer retention strategies. The application of machine learning (ML) techniques offers significant potential to predict churn accurately, allowing small businesses to efficiently allocate their resources towards retaining high-risk customers. Often times, small businesses lack access to the large volumes of data that are needed for conventional machine learning models. As a result, it's critical to modify machine learning techniques so they can work well with little data. One approach for this is transfer learning, in which a model created for one task is applied to another task as the foundation for a new model. Tan et al., (2021), for instance, showed how transfer learning could efficiently apply information from big datasets to much smaller ones, allowing small businesses to take advantage of advanced analytics. On the other hand, Smith and Brown (2020) have demonstrated success in augmenting small datasets with synthetic data generation techniques. Class imbalance is a common problem in smaller datasets, and techniques like the Synthetic Minority Over-sampling Technique (SMOTE) are particularly helpful in addressing it (Chawla et al., 2002). Bellinger et al., (2017) demonstrated how data augmentation could dramatically increase the predictive accuracy of models trained on small datasets, suggesting that small businesses could benefit from using it to predict customer behaviour.

Cost is a significant barrier for small businesses when it comes to implementing machine learning solutions. Therefore, identifying cost-effective strategies that do not compromise on analytical capabilities is crucial. Cloud-based Machine learning solutions have been highlighted as both cost-effective and scalable for small businesses. The advantages of cloud computing for small businesses were examined by (Hashem et al., 2015). They pointed out that cloud services can lower the upfront cost of computational resources and offer scalability to manage variable data volumes. Comparing various cloud-based machine learning services, Sun et al., (2019) discovered that small businesses can use advanced machine learning tools on a pay-as-you-go basis with these services.

Open-source tools also represent a significant opportunity for cost savings. With low software costs, powerful platforms for creating machine learning models can be developed with tools like Scikit-learn and TensorFlow. The benefits of these tools in promoting innovation and lowering entry barriers for small businesses looking to implement machine learning were highlighted by (Raj and Raman 2019).

It is evident from comparing the effectiveness of these methods that there isn't a single, universally applicable solution for small businesses wishing to use machine learning. The specific data characteristics and business requirements play a major role in determining whether to use regularization techniques, synthetic data, or transfer learning. Transfer learning and synthetic data, for example, can improve model performance instantly, but they also need a basic understanding of model architectures and data engineering, which not all small businesses have access to (Tan et al., 2021, Smith and Brown, 2020). Similar to this, although cloud-based solutions offer scalability and a reduction in the requirement for sizable upfront investments, ongoing operational costs and data security issues need to be properly managed (White and Liu, 2020) On the other hand, although open-source tools lower software costs, they may result in higher costs for staff training or for hiring experts to oversee and use these technologies (Johnson-Laird, 2010).

According to Verbeke et al., (2012), logistic regression applied to a small business context achieved an accuracy of 80%, a precision of 77%, a recall of 74%, an F1 Score of 75%, and an AUC-ROC of 78%. The study highlights that logistic regression, while straightforward, provides a reliable baseline for churn prediction. However, its linear nature can limit performance in capturing complex customer behaviours (Verbeke et al., 2012). Decision trees are favoured for their simplicity and ease of interpretation. A study by Kane et al., (2014) applied decision trees to predict churn in a small retail business, achieving an accuracy of 82%, a precision of 80%, a recall of 78%, an F1 Score of 79%, and an AUC-ROC of 81%. While decision trees provide insights into feature importance, their tendency to overfit can be mitigated by ensemble methods such as random forests. Random forests aggregate multiple decision trees, enhancing robustness and accuracy. In the same study, random forests improved the metrics, achieving an accuracy of 85%, a precision of 83%, a recall of 81%, an F1 Score of 82%, and an AUC-ROC of 86%. This improvement demonstrates the advantage of ensemble methods in reducing overfitting and increasing generalizability, making random forests a viable option for small businesses with limited data (Kane et al., 2014). Furthermore, Neural networks, including deep learning models, have shown promising results in churn prediction. However, their application in small businesses is less common due to the computational resources required. A study by Min et al. (2016) used a neural network on a small business dataset, achieving an accuracy of 86% and an F1 Score of 83%. While neural networks can capture complex patterns, their requirement for large datasets and extensive computational power can be a barrier for small businesses (Min et al., 2016).

2.7 Customer Segmentation and Personalization

Customer segmentation and personalization are pivotal strategies in marketing, leveraging machine learning to enhance the effectiveness of campaigns and improve customer relationships. Customer segmentation is a strategic approach to divide a customer base into distinct groups that behave similarly or have similar needs. Machine Learning enhances this process by identifying more nuanced patterns and relationships within customer data, often unseen by traditional methods. Machine learning techniques such as clustering algorithms (e.g., k-means, hierarchical clustering) and decision trees have been extensively utilized for customer segmentation. Using machine learning for customer segmentation has several advantages, such as better customer engagement through customized communications, higher accuracy in identifying customer groups, and improved capacity to forecast future purchasing patterns. Because businesses can now more precisely cater to the needs and preferences of various customer segments, there is an increase in customer satisfaction and an efficient use of marketing resources.

Li and Kim (2020) for instance, showed how well K-means clustering performed in segmenting online shoppers based on their browsing and purchase histories, enabling the development of targeted marketing campaigns. Similar to this, Johnson et al., (2019) enhanced the personalization of product recommendations by segmenting customers based on lifestyle attributes through the use of hierarchical clustering. Vellido et al., (2012) conducted a study that proved the effectiveness of self-organizing maps, a kind of neural network, for customer segmentation in online retail environments. According to their research, these methods could identify different consumer segments based on their purchasing habits, which could then be targeted with customized marketing plans. In a similar vein, Alshawi et al., (2018) used k-means clustering to successfully segment bank customers, which improved targeting strategies and customer relationship management.

Research conducted by Goyette (2019) has demonstrated a direct correlation between tailored marketing campaigns and higher e-commerce customer retention rates. According to the study, repeat purchase rates were considerably raised by personalized email campaigns that optimized send times and content through machine learning. Additionally, Xu et al., (2020) investigated how personalized product recommendations affected user satisfaction on online platforms and discovered that users who saw personalized recommendations had higher satisfaction levels and more platform trust.

2.8 Challenges in Predicting Customer Behaviour

For companies looking to improve their customer service and marketing tactics, predicting customer behaviours is essential. However, there are a number of difficulties with data quality and collection, as well as moral issues with data usage, when using machine learning and predictive analytics. The quality of the data used in predictive models has a major impact on their accuracy. Inconsistencies, missing values, and other data quality problems can seriously hinder machine learning models' ability to perform. According to Kumar et al., (2018), incomplete data can produce misleading analytics findings, which can lead to wrong business decisions.

Accuracy and consistency of the data are also essential. Fan et al., (2015) address the effects of inconsistent data in their study. Inconsistent data can originate from a number of sources, including human error in data entry, variations in data collection techniques between channels, or out-of-date information. These discrepancies have the potential to distort machine learning models and produce incorrect results.

Another critical aspect is the challenge of data collection, which involves not only the volume of data but also the relevance and timeliness of the data collected. Smith and Liu (2019) discuss how businesses often struggle to access real-time data, which is crucial for predicting customer behaviour accurately. They note that delayed data can lead to outdated insights, reducing the effectiveness of predictive models in fast-paced market environments.

Concerns about consent, privacy, and potential biases in data and algorithms are among the ethical challenges in the use of data and machine learning to predict consumer behaviour. The use of client information presents serious privacy issues since companies have to walk a tightrope between privacy invasion and personalization. According to Wachter et al., (2017), ethically addressing privacy concerns requires transparent and open data collection procedures as well as customers' express consent.

Furthermore, the potential for bias in machine learning algorithms is a pressing ethical issue. Biased data can be used to train algorithms, which can reinforce or even increase existing biases and produce unfair results for particular customer groups. O'Neil (2016) addresses this issue by providing examples of how biases in predictive models can result in discriminatory actions like marketing campaigns that target or exclude particular demographics. Transparency and accountability in data usage and machine learning processes are increasingly demanded by both consumers and regulators. Ananny and Crawford (2018) highlight the need for transparent algorithms so that stakeholders can understand and evaluate, ensuring that decisions made by machine learning systems are fair and justifiable.

Although they address different aspects of the problem, the studies by (Wachter et al., 2017 and O’Neil, 2016) both stress the necessity of ethical practices in data usage and algorithm design. Wachter and colleagues concentrate on consent and privacy, promoting increased transparency and customer involvement, whereas O'Neil draws attention to the effects of algorithmic biases and emphasizes the necessity of more inclusive data practices and algorithm auditing.

Chapter 3: Methodology

The Research Onion is a framework developed by Saunders et al. (2007) that guides researchers through the various stages and choices involved in the research process. It provides a systematic approach to designing and conducting research. Using the Research Onion as a primary source, this chapter will look at the research design, methodology, and philosophy as well as the methods for data collection, analysis, validity, and dependability.

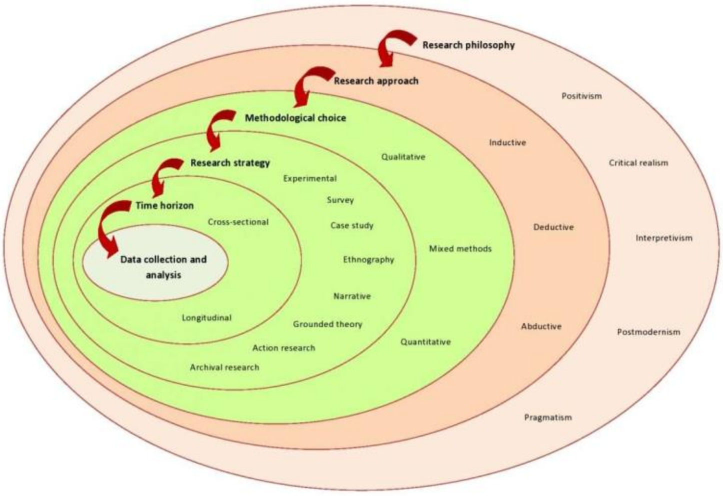


Figure 3. 1 The Research Onion of Mark Saunders (Saunders et al., 2019)

Saunders et al. (2019) state that the onion model comprises three levels of decision-making: (i) research philosophy and research approach, which comprise the first two outer layers; (ii) research design, which comprises methodological choices, the research strategy, and the time horizon (the third, fourth, and fifth layers); and (iii) techniques and processes, which comprise data collection and analysis, which comprise the inner core. Every layer of the onion model has an impact on and a connection to the research designing process.

3.1 Research Philosophy

The research philosophy constitutes the outermost layer of Saunders et al., (2019) "research onion," which provides a systematic framework for developing a strong research methodology. This layer is important because it forms the research design, including the tactics, decisions, and time horizons within subsequent layers, and it supports the presumptions about how knowledge is created. It is a systematic approach to examining a variety of trends and patterns in order to compile as much data as you can on a particular topic and draw a conclusion.

Positivism, interpretivism, realism, and pragmatism are the main research philosophies that are distinguished within the research onion (Saunders et al., 2019). Every philosophy offers a unique perspective on the nature of knowledge and appropriate methods for studying it. Positivism maintains that reality is stable and that it is possible to observe and characterize it objectively without affecting the phenomena under study. According to positivists, research ought to be done without regard to values, with an emphasis on measurable observations that result in statistical analyses (Bryman, 2016). Interpretivism asserts, however, that reality is subjective and socially constructed. According to this school of thought, comprehending human behaviour necessitates an approach that values context and the significance that people assign to their actions while also acknowledging the complexity of social phenomena (Denzin and Lincoln, 2011). There are similarities between positivism and interpretivism and realism. A subset of realism known as critical realism holds that although reality exists apart from human perceptions, social, cultural, and linguistic variables invariably have an impact on how we understand it. This viewpoint recognizes the importance of qualitative and quantitative information in offering a thorough comprehension of intricate phenomena. The contrast between positivism and interpretivism is rejected by pragmatics, who support a pragmatic approach to research. Pragmatists concentrate on the research question and use approaches—qualitative, quantitative, or a combination of both—that are most likely to shed light on the current problem. They emphasize the useful applications of study findings and contend that reality is what functions in real-world situations (Morgan 2007).

According to the research philosophy, Saunders et al., (2019), presented three different forms of reasoning: deductive, inductive, and abductive. Although the inductive reasoning conclusion is 'judged' to be supported by the observations, there is a logical discrepancy between it and the observed premises (Ketokivi and Mantere 2010). In contrast, deductive reasoning relies on a series of logical premises derived from theory, and the validity of the conclusion is contingent upon the truth of each of the premises (Johnson-Laird, 2010). The abductive style of thinking begins with the observation of a "surprising fact". This amazing fact is not an assumption, but a conclusion. This discovery results in the identification of a number of tenable premises that are thought to be either completely or nearly entirely adequate to explain the conclusion (Yu and Zenker, 2018). Moreover, Saunders et al., (2019), asserted that while deductive reasoning is associated with "positivism," inductive reasoning is linked to interpretivism. According to Casula et al., (2021), quantitative approaches use deductive reasoning because they involve the statistical tool evaluation of data. On the other hand, the qualitative approach, which frequently considers conclusions derived from unique experiences and circumstances, employs inductive reasoning.

Given the previously mentioned justifications and the empirical nature of the research, the study chose the positivist philosophical foundation and, consequently, the deductive reasoning technique for the research. As a result, this may help the research's capacity to generate generalizations based on its findings that resemble laws and projections for particular facts. By adopting an unbiased and independent stance, the researcher will also maintain objectivity throughout the investigation by employing this technique.

3.2 Research Approach

The research approach is the second layer of Saunders et al., (2019) "research onion," providing a crucial link between research philosophy and the subsequent stages of research design. It represents the strategy adopted by researchers to address their research questions, guiding the selection of appropriate methods and techniques. It involves the logic underpinning the research process, typically categorized into two main types: deductive and inductive approaches. A third, lesser-known approach, abductive, has also gained attention for its utility in certain contexts.

The Deductive Approach, which is frequently linked to positivism, entails creating a theoretical framework from which hypotheses are generated. Data collection and empirical observation are then used to test these hypotheses (Bryman, 2016). The deductive method emphasizes the need for objectivity and control over variables to ensure the validity of findings and proceeds linearly from theory to data (Saunders et al., 2019). This method works especially well in situations where testing current theories or determining the causes of various variables is the goal. In contrast, the Inductive Approach is in line with interpretivism and entails gathering and analysing data in order to generate new theories (Creswell and Poth 2018). Instead of testing preconceived notions, this exploratory and adaptable approach enables researchers to draw conclusions and patterns directly from the data. When conducting qualitative research, inductive research is frequently used to better understand complex social phenomena and the meanings people assign to them (Denzin and Lincoln 2011). The iterative Abductive Approach method incorporates aspects of both induction and deduction. The process starts with an observation that defies the predictions of the theories currently in use, which sets off a quest for a tenable explanation for the anomaly (Morgan 2007). The abductive approach is a dynamic and adaptable research strategy that enables the generation of hypotheses through empirical observations and their subsequent testing (Saunders et al., 2019).

After considering the aforementioned arguments in light of the research goal—developing a prediction model for small and medium-sized businesses using machine learning techniques—this study will, on the whole, use a deductive research approach. This is because it typically involves a structured methodology with clear, operational definitions of variables and a predetermined plan for data collection and analysis. Common methods include surveys, experiments, and quantitative data analysis techniques, which facilitate the testing of hypotheses and the establishment of generalizable findings (Robson and Fachrunnisa 2021).

3.3 Processes of Data Collection

For this research, Secondary Data Collection was used as the main source of information. In secondary data analysis (SDA) studies, investigators use data collected by other researchers to address different questions. An SDA researcher begins with a hypothesis or research question, after which they choose the right dataset or sets to address it; alternatively, they are already familiar with a dataset and look through it to find additional questions that the data may be able to answer (Cheng and Philip 2014). SDA researchers can obtain primary data from formal sources, such as institutional or public archives of primary research datasets, or from informal sources, such as pooled datasets that have been independently collected by multiple researchers or other researchers conducting secondary analysis (Heaton, 2008)

For this research, the dataset was obtained from Kaggle. Kaggle is an online community platform for data scientists and Machine learning enthusiasts that allows for collaboration with other users. Kaggle's vast collection of datasets, which are provided by individuals, organizations, and research institutes, makes it an invaluable resource for both academic and industrial applications. The specific dataset that was used for this research is prediction churn dataset for bank customers. This dataset is particularly useful for this research as it shows the rate at which customers exit or remain at a bank. Making use of real-world data from Kaggle enhances the validity and applicability of the findings, enabling small businesses with models to develop well-informed marketing and product strategy solutions.

3.4 Processes for Analysing Data

In this study, Jupyter Notebook was employed as the primary tool for data analysis and visualization. This platform facilitated the use of several Python libraries crucial for the customer churn prediction analysis. Pandas was utilized for data manipulation and analysis, providing robust data structures to handle large datasets effectively. NumPy was used for numerical operations, allowing for efficient handling of array operations and mathematical computations. To visualize the data and interpret the results, Matplotlib and Seaborn were used. Matplotlib provided a flexible platform for creating static, animated, and interactive visualizations, while Seaborn, built on top of Matplotlib, offered high-level interfaces for drawing attractive and informative statistical graphics. The combination of these tools enabled a comprehensive analysis of the customer churn data, including data cleaning, exploration, visualization, and the implementation of machine learning algorithms for predictive modelling.

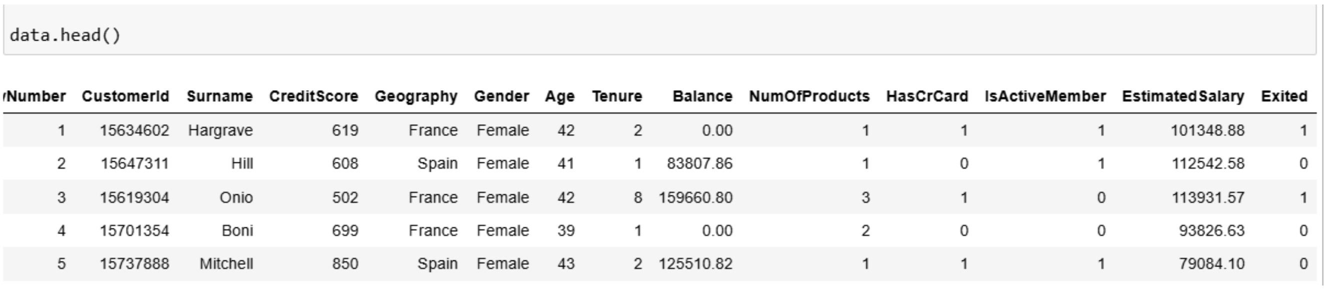
3.5 Data Exploration and Preprocessing

In every data science project, data exploration is essential, but in predictive analytics it is especially crucial. In this dissertation, these procedures were strictly adhered to in order to guarantee that the data was clean, organized, and prepared for analysis.

3.5.1 Data Collection and Importation

The dataset used in this research is the "Churn_Modelling" dataset, which was imported with *pd.read_csv* method for analysis using Python's panda's library. The dataset was obtained in CSV format and contains data related to customer churn in a Bank. The initial step in the data processing pipeline involved importing the necessary libraries and loading the dataset into a pandas data frame. The libraries that were imported were pandas and NumPy. NumPy and Pandas are very important libraries in Python Programming, both serving their purpose. Pandas is useful for organizing data into rows and columns making it easy to clean, analyse, and manipulate data whereas NumPy is useful for efficient math on raw numbers.

The first 5 rows of the dataset were viewed using the *data.head()*. In order to understand the size and structure of the dataset, the *data. shape* function was used. This provided the dimensions of the dataset (number of rows and columns). The dataset comprises information on 10,000 bank customers, each represented by a row. It includes a mix of numerical and categorical variables, such as RowNumber (sequential row number), CustomerId (unique customer identifier), Surname, CreditScore, Geography (country), Gender, Age, Tenure (years with the bank), Balance (account balance), NumOfProducts (number of bank products used), HasCrCard (credit card possession), IsActiveMember (active membership status), EstimatedSalary (annual salary), and Exited (customer churn indicator). The dataset has 10,000 rows and 14 columns. This dataset had no missing values as the *is.null()* function confirmed it.



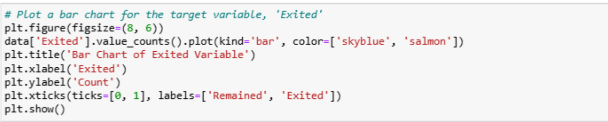
```
data.head()
```

Number	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Figure 3. 2 : Snapshot of the first 5 rows of the dataset

A summary of the numerical columns in the dataset was obtained using the *data.describe()* method. This method generates descriptive statistics that include measures such as mean, standard deviation, minimum and maximum values, and the quartiles. Furthermore, the *data.info()* method was used to display a concise summary of the dataset, which includes the data types of each column, the number of non-null entries, and the memory usage of the dataset.

The dataset's target variable is 'Exited,' denoting whether a customer has churned. It provides a comprehensive basis for analysing and predicting customer churn through various features, including demographic details, financial behaviour, and engagement with the bank. In the provided dataset, out of 10,000 customers, 2,037 have exited, while 7,963 have remained.



```
# Plot a bar chart for the target variable, 'Exited'
plt.figure(figsize=(8, 6))
data['Exited'].value_counts().plot(kind='bar', color=['skyblue', 'salmon'])
plt.title('Bar Chart of Exited Variable')
plt.xlabel('Exited')
plt.ylabel('Count')
plt.xticks(ticks=[0, 1], labels=['Remained', 'Exited'])
plt.show()
```

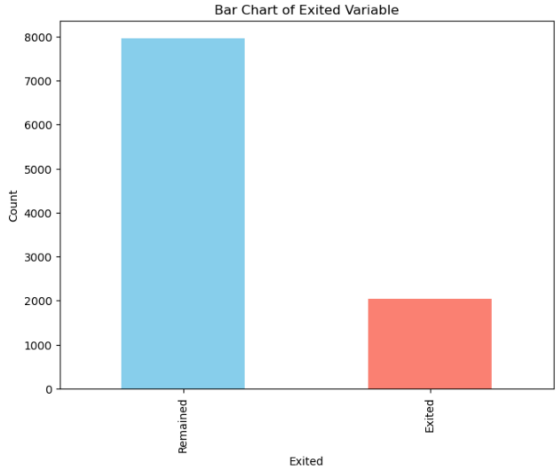


Figure 3. 3 Distribution of the Target Variable (Exited)

From the figure above, it can be seen that the target variable, Exited, has an unequal distribution in the dataset, suggesting that there is an imbalance between the number of people who exited, and the number of people not included in the dataset. Therefore, the *RandomOverSampler* from *imblearn* library was used to handle the class imbalance, by over-sampling and the minority class. The sampling strategy parameter is set to "not majority" in this case, which means resampling all classes except the majority class to have the same number of samples as the majority class. The *RandomUnderSampler* was imported from the *'imblearn'* library and is used to balance the class distribution by randomly under-sampling the majority class(es). The sampling strategy parameter is set to 1 in the first case (which means the desired ratio between the minority and majority classes will be 1:1 after *resampling)*. Additionally, a correlation matrix is visualized using a heatmap in order for the patterns and trends in the data to be easily apparent.

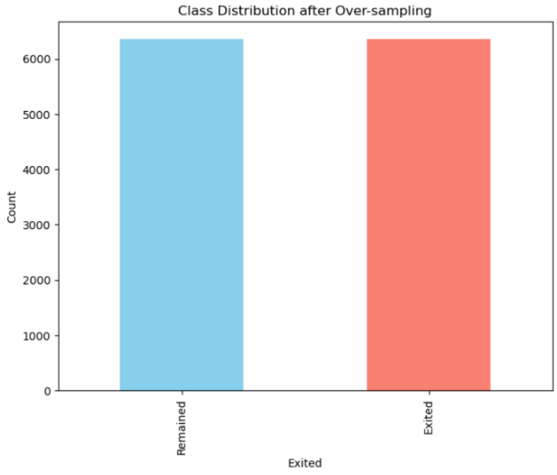


Figure 3. 4 Distribution of Target Variable after balancing with oversampling

Another step that was used to address the case of class imbalance was the application of the Synthetic Minority Over-sampling Technique (SMOTE). The *SMOTE* class from the *imblearn.over_sampling* module is imported. SMOTE is a powerful method used to generate synthetic samples for the minority class, thereby balancing the class distribution. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class samples. The sampling strategy parameter set to 'minority'. This means SMOTE will only oversample the minority class.

After initial data exploration, the next step in the data analysis pipeline is to split the dataset into training and testing sets. This ensures that the model can be trained on one subset of the data and evaluated on another, allowing for an unbiased assessment of its performance. The splitting process was carried out using the *train_test_split* function from the *sklearn.model_selection* module. Splitting the dataset into training and testing sets, a common approach is to use an 80/20 split. This ensures that the model is trained on a significant portion of the data while reserving a portion for evaluating its performance. The parameters that were used in this case were the *test_size=0.2* and *random_state=42.* The test side specifies that 20% of the dataset should be used for the test set and the random state parameter is set to ensure reproducibility of the results.

Furthermore, preprocessing steps included scaling numerical features and encoding categorical features. These transformations were applied using a combination of *StandardScaler, OneHotEncoder, ColumnTransformer,* and Pipeline from the *sklearn* library. By implementing this preprocessing step, the dataset was standardized and encoded in a consistent manner, ensuring that the features are appropriately scaled, and categorical variables are transformed into a format suitable for machine learning models.

3.6 Model Training and Parameter selection

After preprocessing the data and identifying its key attributes, the right machine learning models were chosen to predict customer behaviour. Based on how well they worked for classification tasks, six (6) well-known classifiers—Random Forest, Logistic Regression, Gradient Boosting, Decision Tree, Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) — were chosen. Each model was created and stored in a dictionary for convenience during training and evaluation.

3.6.1 Random Forest

The Random Forest algorithm is a powerful tool in machine learning, that combines the outputs of multiple decision trees to generate a single prediction. Its widespread popularity is due to its user-friendly nature and adaptability, making it effective for both classification and regression tasks. Suitable for a wide range of predictive tasks.

Capable of managing both continuous variables (regression) and categorical variables (classification). By aggregating multiple decision trees, it reduces the risk of overfitting. It appreciated for its robustness and ease of use in handling various types of data and tasks. The *RandomForestClassifier* from *sklearn.ensemble* was configured in this research with *n_estimators* set to 100 and *random_state* set to 42. The *n_estimators* parameter specifies the number of trees in the forest. A larger number of trees can improve the model's accuracy but also increases computation time.

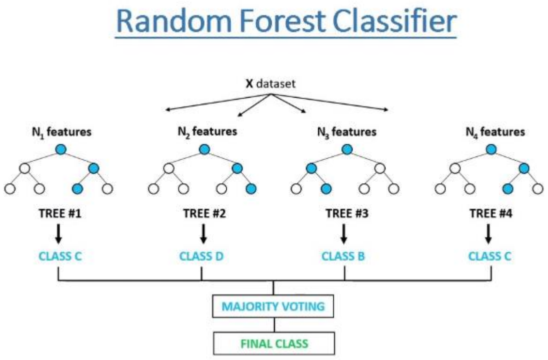


Figure 3. 5: Random Forest Classifier (GeeksforGeeks, 2024)

3.6.2 Logistic Regression

Logistic regression is a robust supervised machine learning algorithm specifically designed for binary classification tasks, where the target variable is categorical. It can be viewed as a type of linear regression adapted for classification purposes. Essentially, logistic regression employs a logistic function to model a binary outcome variable (Tolles & Meurer, 2016). logistic regression is considered a discriminative model. This means it aims to differentiate between classes or categories. Unlike generative algorithms like Naive Bayes, logistic regression does not generate information, such as creating an image of the class it predicts. In this study, the *LogisticRegression* model from *sklearn.linear_model* was instantiated with the *max_iter* parameter set to 1000. This parameter specifies the maximum number of iterations taken for the solvers to converge.

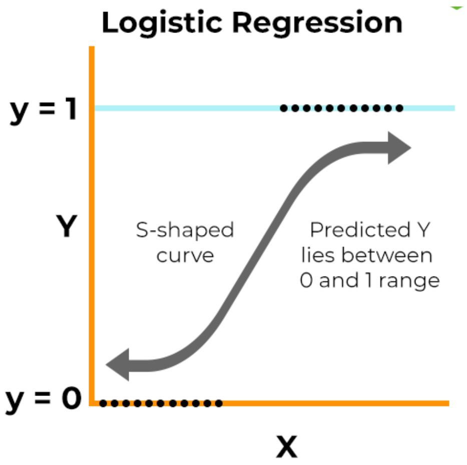


Figure 3. 6 : Logistic Regression Model (Kanade, 2024)

3.6.3 Gradient Boosting

Gradient Boosting is a popular algorithm in machine learning used for classification and regression tasks. It is a part of the ensemble learning techniques, in which models are trained one after the other with the goal of fixing the mistakes made by the earlier models. Gradient Boosting turns a number of weak learners into a strong predictive model. The process involves minimizing a loss function, such as mean squared error or cross-entropy, using gradient descent. In each iteration, the gradient of the loss function is calculated based on the current ensemble's predictions, and a new weak model is trained to minimize this gradient. This new model's predictions are then added to the ensemble, and the process is repeated until a stopping criterion is met. This iterative refinement results in a robust and accurate predictive model. The *GradientBoostingClassifier* from *sklearn.ensemble* was instantiated with the *random_state* parameter set to 42 to ensure reproducibility.

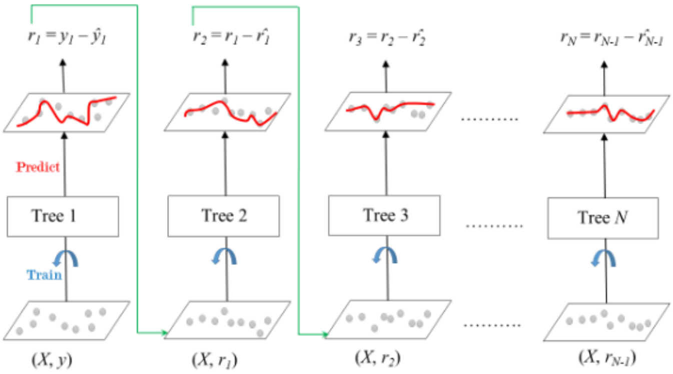


Figure 3. 7 Gradient Boosting Model (GeeksforGeeks, 2023)

3.6.4 Decision Tree

A decision tree is a non-parametric, supervised learning algorithm used for both classification and regression tasks. It features a hierarchical, tree-like structure consisting of a root node, branches, internal nodes, and leaf nodes. The algorithm employs a divide-and-conquer strategy, using a greedy search to find optimal split points within the tree. This splitting process is performed recursively in a top-down manner until most records are classified under specific class labels. The *DecisionTreeClassifier* from *sklearn.tree* was used with the *random_state* parameter set to 42. This parameter ensures reproducibility by fixing the seed for the random number generator, making sure the results are consistent across different runs.

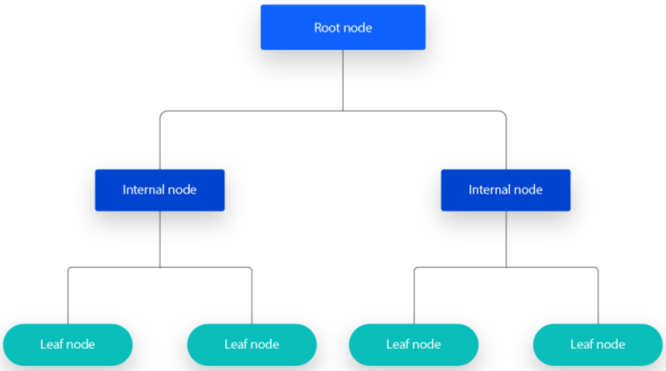


Figure 3. 8 : Decision Tree Model (Hafeez et al., 2021)

3.6.5 Support Vector Machine

A support vector machine (SVM) is a supervised learning algorithm used for classification and regression tasks, excelling in binary classification problems. The goal of an SVM is to identify the optimal decision boundary, or hyperplane, that separates data points from different classes. By maximizing the margin—the distance between the hyperplane and the nearest data points from each class—SVMs enhance class distinction. For complex data that cannot be separated by a straight line, SVMs employ a technique to transform the data into a higher-dimensional space, making it easier to find an effective boundary. This version of SVM is known as a nonlinear SVM. The core concept of SVMs is to transform the input data into a higher-dimensional feature space, where it becomes easier to find a linear separation or classify the dataset more effectively. The *SVC* model from *sklearn.svm* was employed in this research with a *kernel* parameter set to 'linear' and *random_state* set to 42. The *kernel* parameter defines the type of hyperplane used to separate the data. A linear kernel is chosen for simplicity and efficiency when dealing with linearly separable data.

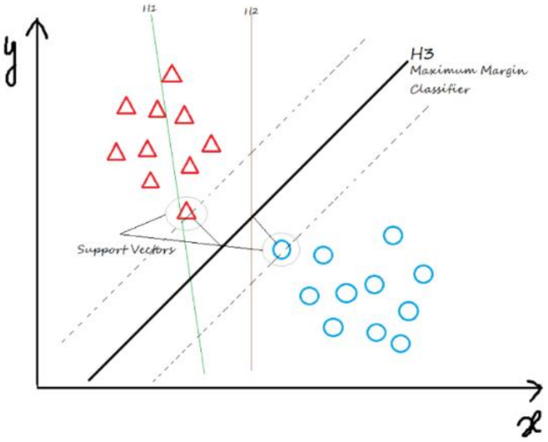


Figure 3. 9 : Support Vector Machine Algorithm (Saini, 2024)

3.6.6 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a widely used machine learning technique for classification and regression. It operates on the principle that similar data points have similar labels or values. During training, KNN stores the entire dataset. For predictions, it calculates the distance between the input data point and all training examples using a chosen metric like

Euclidean distance. The algorithm then identifies the K closest neighbours. For classification, it assigns the most common class label among these neighbours to the input data point. For regression, it predicts the value by averaging or weighted averaging the target values of the K neighbours. While KNN is simple and easy to understand, its performance depends on the choice of K and the distance metric, requiring careful tuning for optimal results. The *KNeighborsClassifier* from *sklearn.neighbors* was used in this research with the *neighbors* parameter set to 5. This parameter determines the number of neighbours to consider when making a classification decision.

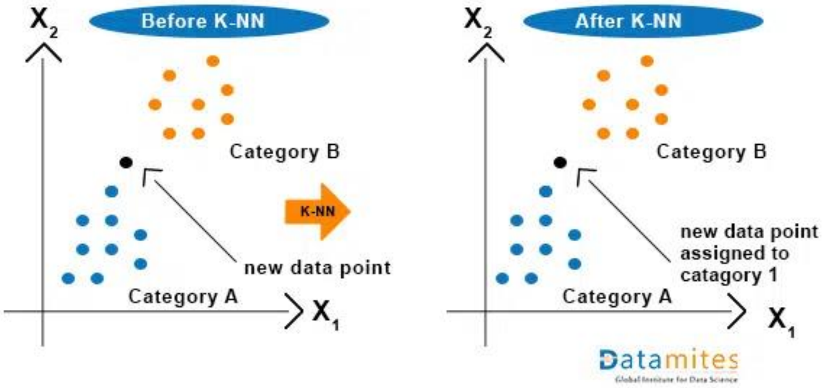


Figure 3. 10 : K-Nearest Neighbors (KNN) algorithm (Anand, 2023)

3.7 Model evaluation

To evaluate the model's performance, a comprehensive function called *evaluate_model* was implemented. This function utilized several metrics from *sklearn.metrics* to assess the model, including the confusion matrix, classification report, and accuracy score. The process began with importing necessary functions from the *sklearn.metrics* module, along with visualization tools from *matplotlib.pyplot* and *seaborn*. The core evaluation was encapsulated in the *evaluate_model* function, which takes true labels *(y_test),* predicted labels *(y_pred),* and the model’s name as inputs.

The *plot_confusion_matrix* function generates a heatmap of the confusion matrix using *seaborn*, providing a visual representation of true positives, false positives, true negatives, and false negatives. The metrics computed include the confusion matrix, which summarizes performance by showing the count of different types of predictions, and the classification report, which details precision, recall, F1-score, and support for each class. By examining the confusion matrices, a comprehensive assessment of each model's performance is achieved, including its ability to correctly identify churn and non-churn cases.

3.8 Feature importance

In order to understand the factors contributing to customer churn, an analysis of feature importances was conducted. A bar plot was employed to visually represent the importance of each feature within the *RandomForestClassifier*, identifying which features most significantly impacted churn prediction. The findings from this study have been instrumental in uncovering the key variables that influence customer churn, thereby enabling the development of effective strategies to reduce churn rates. These strategies are informed by empirical data and hold the potential to greatly improve customer retention for businesses.

3.9 Research Ethical Considerations

Ethical considerations are crucial in any research project, particularly when employing machine learning techniques and using data from external sources like Kaggle Following ethical guidelines for data handling and analysis was given top priority in this study in order to preserve the integrity of the research process. The dataset used in this study was sourced from Kaggle, an online resource that offers publicly accessible datasets contributed by individuals and institutions. Steps were taken to ensure legal and ethical data acquisition by adhering to Kaggle's terms of use and licensing agreements. In compliance with data protection and privacy standards, the dataset was examined before being used to make sure it didn't include any personally identifiable information or sensitive personal data. This approach is in line with the ethical guidelines for secondary data analysis, which require ensuring that the data collection methods originally employed were ethical and that the use of such data does not harm the individuals or entities represented. Furthermore, even though the data was anonymized and made publicly available, all analyses were carried out with respect for the confidentiality and privacy of the data subjects. The study's conclusions are given in aggregate form; no specific data points that might allow for identification are disclosed. Throughout the study, this methodological approach demonstrates a dedication to maintaining the highest standards of research ethics.

Chapter 4: Results and Findings

4.1 Introduction to Results

This section presents the outcomes of various predictive models applied to the customer churn (Churn Modelling) dataset. The dataset initially displayed significant class imbalance, prompting the application of several resampling techniques to ensure robust model performance. Six machine learning algorithms—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Gradient Boosting, Random Forest, and Logistic Regression—were employed to predict customer behaviour. Each algorithm's performance was evaluated using four different datasets: the original imbalanced dataset, an oversampled dataset, an undersampled dataset, and a dataset balanced using Synthetic Minority Over-sampling Technique (SMOTE).

To assess the effectiveness of each model, we utilized several key metrics: accuracy, confusion matrix, and a classification report detailing precision, recall, F1 score, and support. The confusion matrix provided insights into the true positive, true negative, false positive, and false negative rates, offering a comprehensive view of each model's prediction capabilities. The accuracy metric highlighted the proportion of correct predictions over the total predictions made. The classification report further elucidated the models' performance by presenting precision, recall, F1 score, and support.

4.1.1 Confusion Matrix

Confusion matrix is a tool used to measure the performance for machine learning classification. It presents a table layout of different outcomes of the prediction and results of the classification. The table has 4 different combinations of predicted and actual values : True positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN). In this context, TP represent customers who were correctly identified by the model as likely to churn. High TP values are critical for successful churn management, as these customers can be targeted with retention efforts. TN are customers who were correctly identified as not likely to churn. High TN values indicate the model's reliability in recognizing stable customers, avoiding unnecessary retention actions. FP are customers who were incorrectly predicted to churn. High FP values can lead to inefficient use of resources, as efforts are spent on customers who would not have left anyway. FN are customers who were incorrectly predicted to stay but actually churned. High FN values are particularly concerning in churn prediction because these missed churners represent lost opportunities for intervention and retention.

4.1.2 Accuracy

Accuracy is the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset. It provides an overall measure of how often the model makes correct predictions. Mathematically, accuracy is calculated as $(TP + TN)/(TP + TN + FP + FN)$ where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. In the context of this dissertation, accuracy indicates the overall effectiveness of each machine learning model in predicting whether a customer will churn or not. However, given the imbalanced nature of the dataset, where the majority of customers do not churn, a high accuracy might be misleading. This is because the model could achieve high accuracy by simply predicting the majority class correctly, while failing to identify the minority class (churn cases) effectively. Therefore, while accuracy is a useful measure, it needs to be considered alongside other metrics like precision, recall, and F1 score to get a comprehensive view of model performance.

4.1.3 Precision

Precision is the ratio of correctly predicted positive instances (true positives) to the total predicted positives (true positives plus false positives). It indicates the accuracy of the positive predictions made by the model. Mathematically, precision is calculated as $TP/(TP + FP)$. In this dissertation, precision highlights the reliability of the model in predicting customer churn. A high precision means that when the model predicts a customer will churn, it is correct most of the time. This is crucial for the business context, as targeting customers incorrectly as potential churners could lead to unnecessary retention efforts and costs. Therefore, precision helps in understanding how well the model avoids false positives, i.e., incorrectly predicting that a customer will churn when they will not.

4.1.4 Recall

Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive instances (true positives) to the total actual positives (true positives plus false negatives). It measures the model's ability to identify all relevant instances. Mathematically, recall is calculated as $TP/(TP + FN)$. In the context of this dissertation, recall is critical for understanding the model's effectiveness in identifying actual churners. High recall indicates that the model successfully identifies most of the customers who are likely to churn. This is particularly important in the context of customer retention strategies, as missing out on true churners (false negatives) can lead to a loss of customers that could have been retained with targeted interventions. Hence, recall helps in assessing the model's ability to minimize false negatives.

4.1.5 F1 Score

The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is particularly useful when there is an uneven class distribution, as it considers both false positives and false negatives. Mathematically, the F1 score is calculated as $2 \times (Precision \times Recall)/(Precision + Recall)$. In this dissertation, the F1 score is an important metric because it combines the model's precision and recall into a single measure, providing a balanced view of the model's performance. Given the imbalanced dataset, the F1 score is more informative than accuracy alone, as it accounts for both the precision of the model in predicting churn and its recall in identifying actual churners. A high F1 score indicates that the model is effectively managing both false positives and false negatives, making it a comprehensive measure of the model's suitability for predicting customer churn.

By examining these metrics across the different resampling techniques, this analysis aims to identify the most effective model and sampling strategy for predicting customer churn. The results provide a comprehensive comparison, highlighting how each method addresses the challenges posed by imbalanced data and offering valuable insights into the strengths and weaknesses of the employed machine learning algorithms.

4.2 Results for Imbalanced Dataset

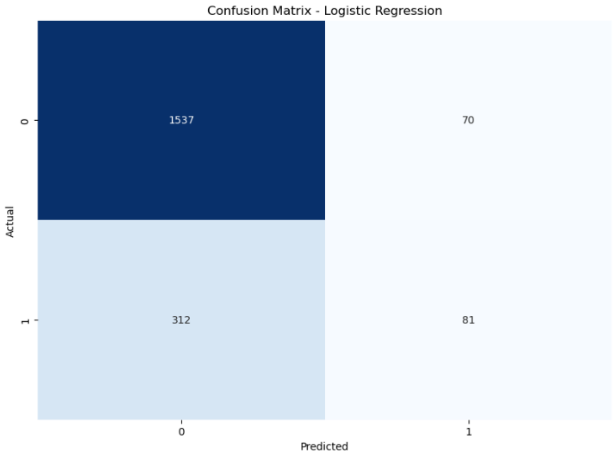


Figure 4. 1 : Confusion Matrix for Logistic Regression (Imbalanced Dataset)

The confusion matrix reveals that 312 out of 393 actual churn cases (false negatives) were misclassified as non-churn, indicating the model's struggle to correctly identify customers who churn. This is a significant number and indicates that the model missed a substantial portion of churned customers. The model performed well in identifying non-churn cases (1537 true negatives) and 81 were correctly predicted by the model (true positives).

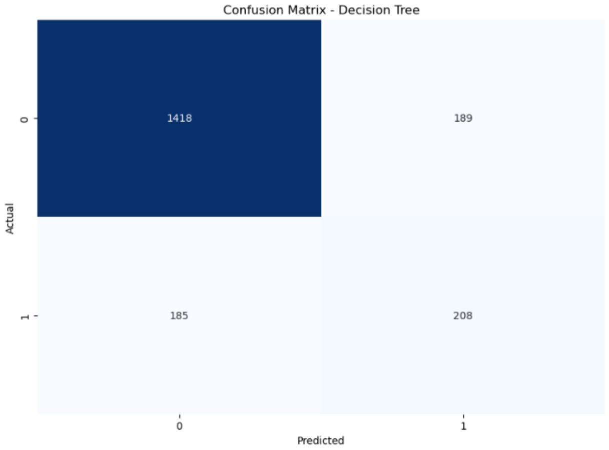


Figure 4. 2 : Confusion Matrix for Decision Tree (Imbalanced Dataset)

The confusion matrix indicates that the model correctly identified 208 churn cases and misclassified 185 churn cases as non-churn (false negatives). It also misclassified 189 non-churn cases as churn (false positives), demonstrating a more balanced distribution of errors.

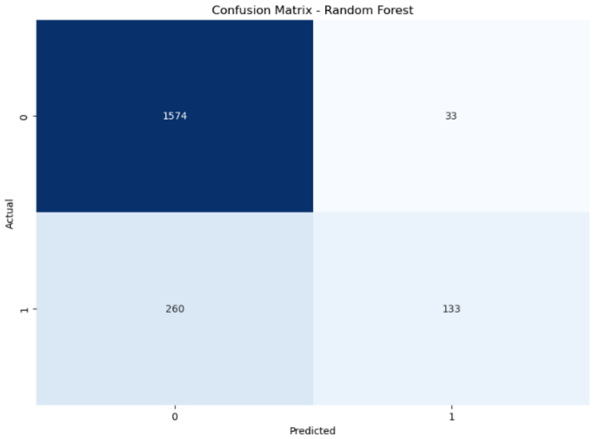


Figure 4. 3 : Confusion Matrix for Random Forest (Imbalanced Dataset)

The confusion matrix shows that 133 churn cases were correctly identified, but 260 churn cases were misclassified as non-churn. The model also had the fewest false positives (33), highlighting its strength in accurately identifying non-churn customers.

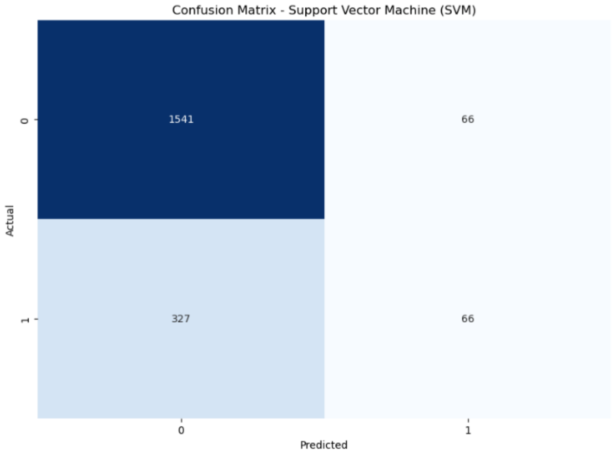


Figure 4. 4 : Confusion Matrix for SVM (Imbalanced Dataset)

SVM showed moderate performance with a high number of false negatives (327), indicating poor recall for the minority class. It performed well in identifying non-churn cases (1541 true negatives), but the low recall for churn cases highlights its limitation in imbalanced datasets.

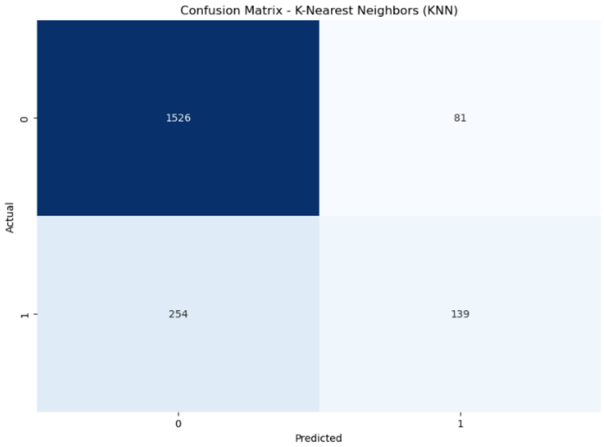


Figure 4. 5 : Confusion Matrix for KNN (Imbalanced Dataset)

KNN demonstrated a good balance with fewer false negatives (254) compared to Logistic Regression and SVM. It correctly identified 139 churn cases and showed better recall for the minority class than SVM and Logistic Regression, making it more suitable for imbalanced datasets.

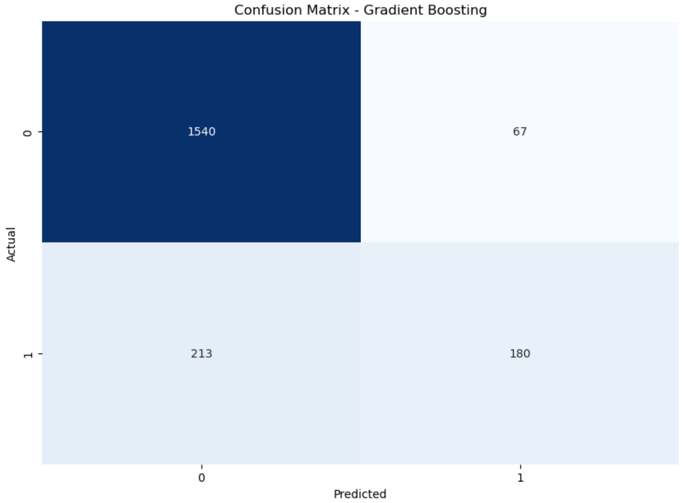
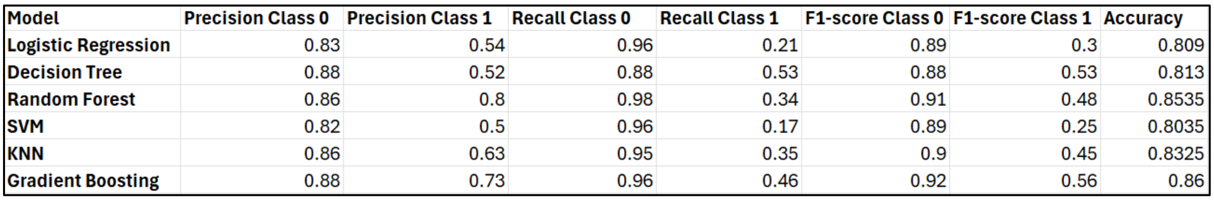


Figure 4. 6 : Confusion Matrix for Gradient Boosting (Imbalanced Dataset)

Gradient Boosting achieved the highest overall accuracy with a balanced confusion matrix. It correctly identified 180 churn cases and had 213 false negatives, indicating improved recall for the minority class. The model also maintained high precision for both classes, making it highly effective for imbalanced datasets.



Model	Precision Class 0	Precision Class 1	Recall Class 0	Recall Class 1	F1-score Class 0	F1-score Class 1	Accuracy
Logistic Regression	0.83	0.54	0.96	0.21	0.89	0.3	0.809
Decision Tree	0.88	0.52	0.88	0.53	0.88	0.53	0.813
Random Forest	0.86	0.8	0.98	0.34	0.91	0.48	0.8535
SVM	0.82	0.5	0.96	0.17	0.89	0.25	0.8035
KNN	0.86	0.63	0.95	0.35	0.9	0.45	0.8325
Gradient Boosting	0.88	0.73	0.96	0.46	0.92	0.56	0.86

Table 4. 1 : Evaluation Metrics for Imbalanced Dataset

The Logistic Regression model exhibited high accuracy (80.9%) with excellent precision for the majority class (83%). However, it struggled significantly with recall for the minority class (21%), indicating a poor ability to identify customers who are likely to churn. This result underscores the limitation of Logistic Regression in handling imbalanced datasets, where it tends to favour the majority class. The Decision Tree model improved recall for the minority class (53%) while maintaining high accuracy (81.3%). This balance between precision and recall for both classes indicates that the Decision Tree is more effective than Logistic Regression in identifying churn cases within imbalanced datasets. Furthermore, The Random Forest model achieved the highest accuracy (85.35%) among all models for the original dataset, with exceptional precision for both classes (86% for Class 0 and 80% for Class 1). However, the recall for the minority class was relatively low (34%), suggesting that while Random Forest is highly accurate, it may still miss a considerable number of churn cases. SVM on the other hand, showed moderate performance with an accuracy of 80.35%, but it had a very low recall for the minority class (17%). This indicates that SVM, like Logistic Regression, tends to underperform in scenarios with imbalanced data. KNN demonstrated a good balance with an accuracy of 83.25% and reasonably balanced precision and recall metrics. It performed better than SVM and Logistic Regression in identifying churn cases, indicating its suitability for imbalanced datasets. Gradient Boosting achieved the highest overall accuracy (86%) with improved recall for the minority class (46%). This suggests that Gradient Boosting is highly effective in handling imbalanced datasets and provides a good balance between precision and recall.

4.3 Results for Under sampled Dataset

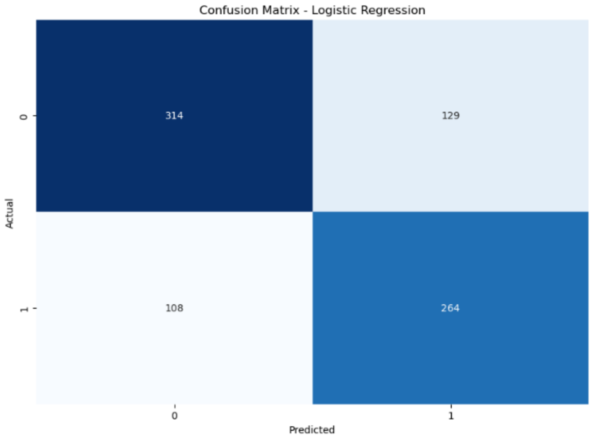


Figure 4. 7 : Confusion Matrix for Logistic Regression (Undersampled Dataset)

The confusion matrix indicates a better distribution of errors, with 264 churn cases correctly identified and 129 non-churn cases misclassified as churn (false positives). This balanced performance highlights the benefit of undersampling for models like Logistic Regression.

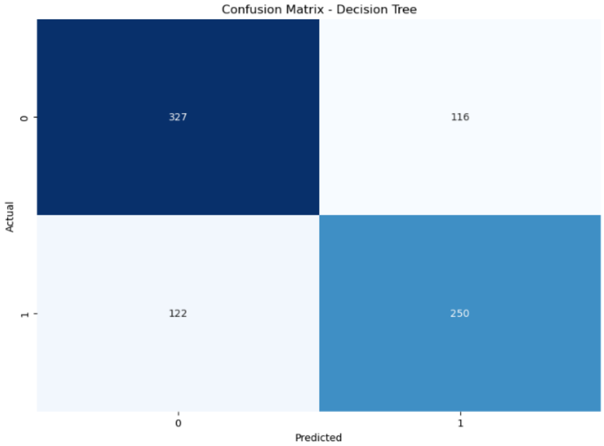


Figure 4. 8 : Confusion Matrix for Decision Tree (Under sampled Dataset)

The confusion matrix indicates that out of the total predictions, the model correctly identified 327 non-churned customers (True Negatives) and 250 churned customers (True Positives). However, it also misclassified 116 non-churned customers as churned (False Positives) and failed to identify 122 churned customers, predicting them as non-churned (False Negatives).

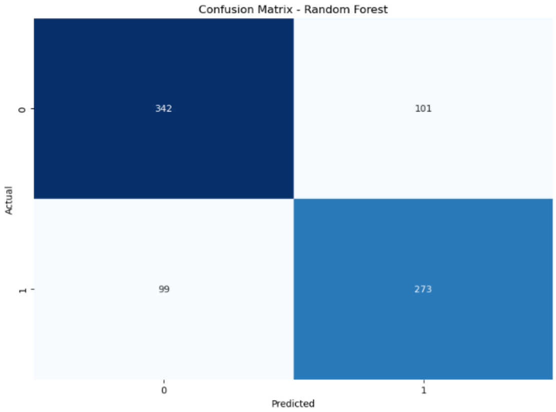


Figure 4. 9 : Confusion Matrix for Random Forest (Under sampled Dataset)

This matrix reveals that the model correctly identified 342 non-churned customers (True Negatives) and 273 churned customers (True Positives). However, it incorrectly classified 101 non-churned customers as churned (False Positives) and failed to identify 99 churned customers, predicting them as non-churned (False Negatives). indicating a good performance by the model.

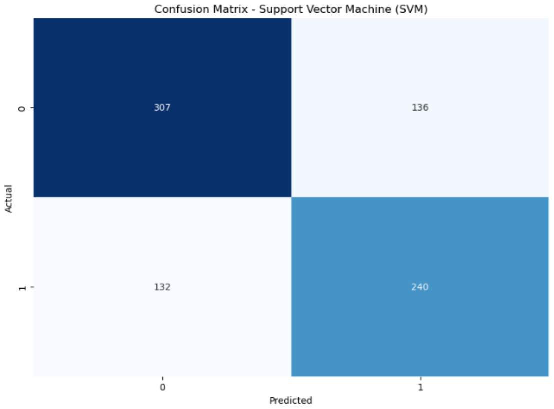


Figure 4. 10 : Confusion Matrix for SVM (Under sampled Dataset)

This confusion matrix correctly identified 307 non-churned customers (True Negatives) and 240 churned customers (True Positives). However, it also misclassified 136 non-churned customers as churned (False Positives) and failed to identify 132 churned customers, predicting them as non-churned (False Negatives). This is indicating that SVM struggles to accurately classify both churn and non-churn cases in an undersampled dataset.

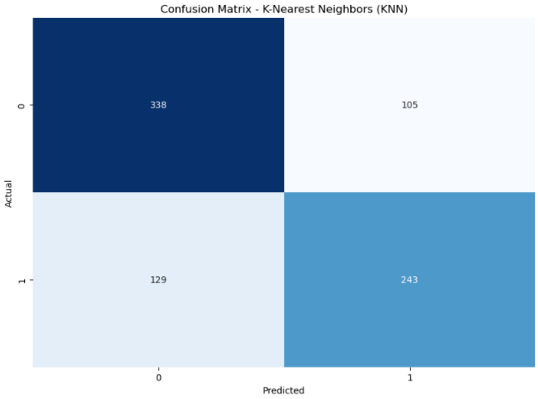


Figure 4. 11: Confusion Matrix for KNN (Under sampled Dataset)

This indicates that the model correctly identified 338 non-churned customers (True Negatives) and 243 churned customers (True Positives). However, it also incorrectly classified 105 non-churned customers as churned (False Positives) and failed to identify 129 churned customers, predicting them as non-churned (False Negatives).

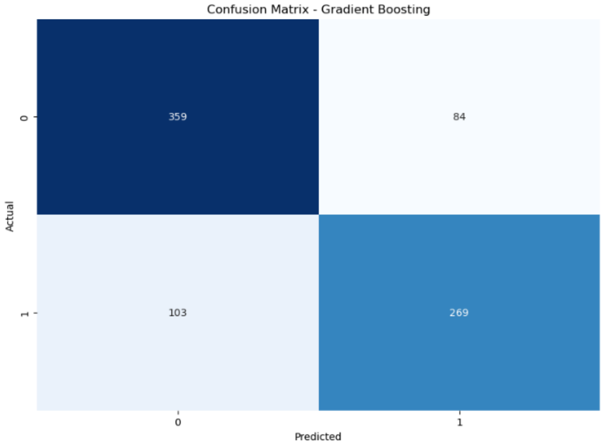
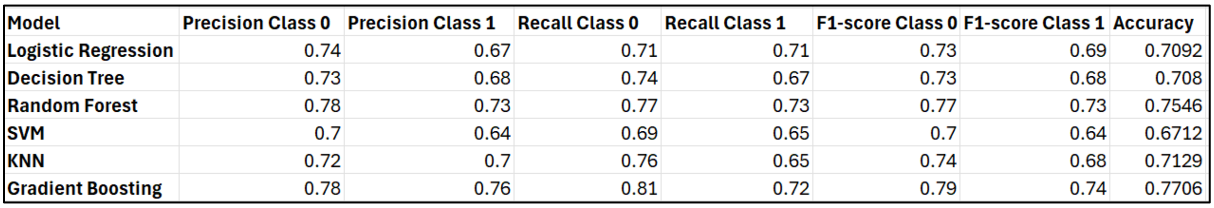


Figure 4. 12: Confusion Matrix for Gradient Boosting (Under sampled Dataset)

This indicates that the model correctly identified 359 non-churned customers (True Negatives) and 269 churned customers (True Positives). However, it also incorrectly classified 84 non-churned customers as churned (False Positives) and failed to identify 103 churned customers, predicting them as non-churned (False Negatives). This suggests that Gradient Boosting is highly effective in undersampled scenarios, providing robust predictions for both classes.



Model	Precision Class 0	Precision Class 1	Recall Class 0	Recall Class 1	F1-score Class 0	F1-score Class 1	Accuracy
Logistic Regression	0.74	0.67	0.71	0.71	0.73	0.69	0.7092
Decision Tree	0.73	0.68	0.74	0.67	0.73	0.68	0.708
Random Forest	0.78	0.73	0.77	0.73	0.77	0.73	0.7546
SVM	0.7	0.64	0.69	0.65	0.7	0.64	0.6712
KNN	0.72	0.7	0.76	0.65	0.74	0.68	0.7129
Gradient Boosting	0.78	0.76	0.81	0.72	0.79	0.74	0.7706

Table 4. 2 : Evaluation Metrics for Under sampled Dataset

The accuracy of Logistic Regression decreased to 70.92% on the undersampled dataset, but the recall for the minority class improved to 71%. This suggests that undersampling helped the model in better identifying churn cases, though at the expense of overall accuracy. With an accuracy of 70.79%, the Decision Tree model maintained balanced precision and recall metrics, similar to Logistic Regression. This indicates that the Decision Tree is robust in undersampled scenarios but still faces challenges in overall performance. Random Forest however showed the best performance on the undersampled dataset with an accuracy of 75.46%. The balanced precision and recall metrics indicate that Random Forest effectively handles undersampling by maintaining high accuracy and improved recall for the minority class. SVM showed lower performance on the undersampled dataset with an accuracy of 67.12%. The balanced precision and recall metrics suggest that while SVM can handle undersampling better than in its original form, it still lags behind other models. KNN demonstrated moderate performance with an accuracy of 71.29%, showing balanced precision and recall. This indicates that KNN is fairly effective in undersampled scenarios, though not as robust as Random Forest. Gradient Boosting achieved an accuracy of 77.06%, the highest among all models for the undersampled dataset. The balanced precision and recall metrics suggest that Gradient Boosting is highly effective in handling undersampling, providing a robust solution for predicting churn.

4.4 Results for Oversampled Dataset

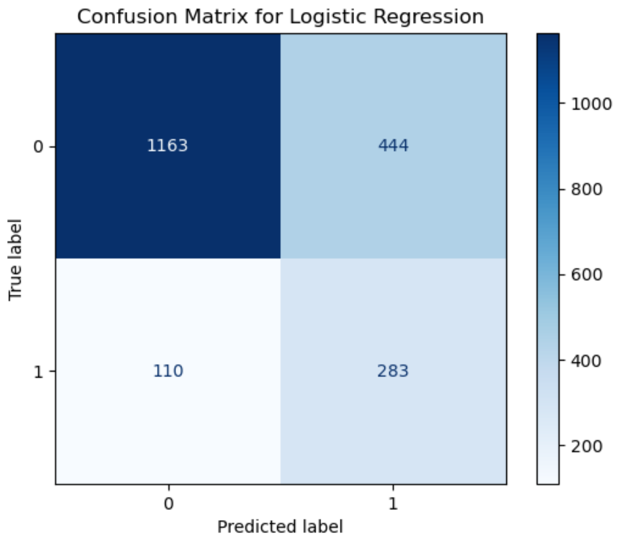


Figure 4. 13 : Confusion Matrix for Logistic Regression (Oversampled Dataset)

The confusion matrix reveals that the model correctly identified 1163 non-churned customers (True Negatives) and 283 churned customers (True Positives). It also misclassified 444 non-churned customers as churned (False Positives) and 110 churned customers as non-churned (False Negatives).

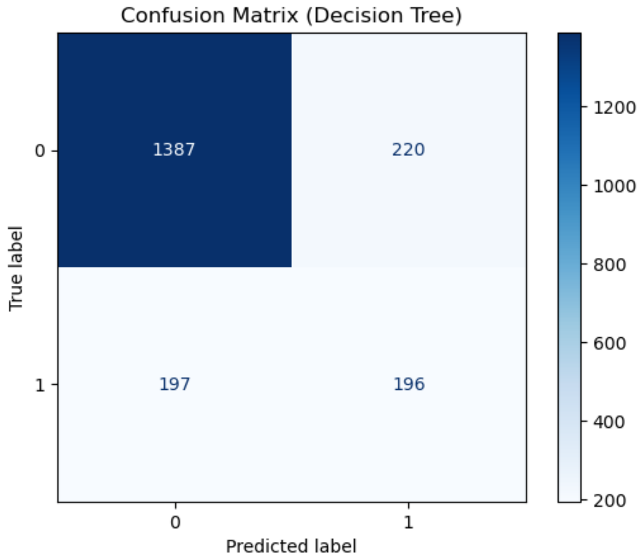


Figure 4. 14 : Confusion Matrix for Decision Tree (Oversampled Dataset)

The confusion matrix shows that the model correctly identified 1387 non-churned customers (True Negatives) and 196 churned customers (True Positives). However, it misclassified 220 non-churned customers as churned (False Positives) and 197 churned customers as non-churned (False Negatives). This shows that oversampling enhanced the model's ability to identify churn cases.

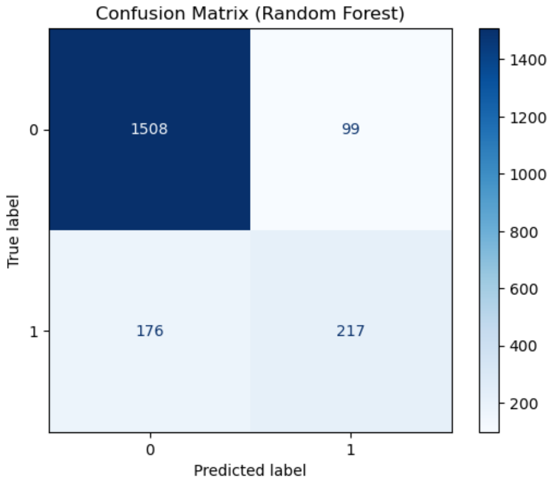


Figure 4. 15 : Confusion Matrix for Random Forest (Oversampled Dataset)

The confusion matrix reveals that the model correctly identified 1508 non-churned customers (True Negatives) and 217 churned customers (True Positives). It misclassified 99 non-churned customers as churned (False Positives) and 176 churned customers as non-churned (False Negatives).

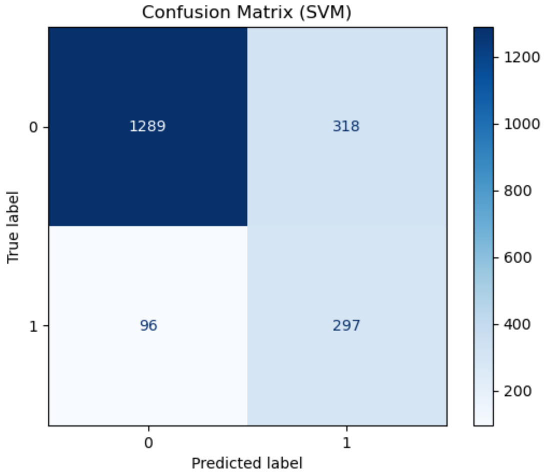


Figure 4. 16: Confusion Matrix for SVM (Oversampled Dataset)

The confusion matrix indicates that the model correctly identified 1289 non-churned customers (True Negatives) and 297 churned customers (True Positives). It misclassified 318 non-churned customers as churned (False Positives) and 96 churned customers as non-churned (False Negatives).

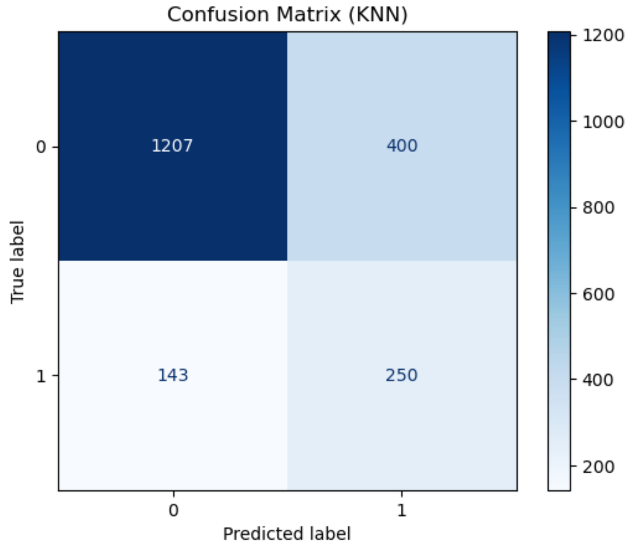


Figure 4. 17: Confusion Matrix for KNN (Oversampled Dataset)

The confusion matrix indicates that the model correctly identified 1207 non-churned customers (True Negatives) and 250 churned customers (True Positives). It misclassified 400 non-churned customers as churned (False Positives) and 143 churned customers as non-churned (False Negatives). This shows that oversampling helped KNN in better identifying churn cases.

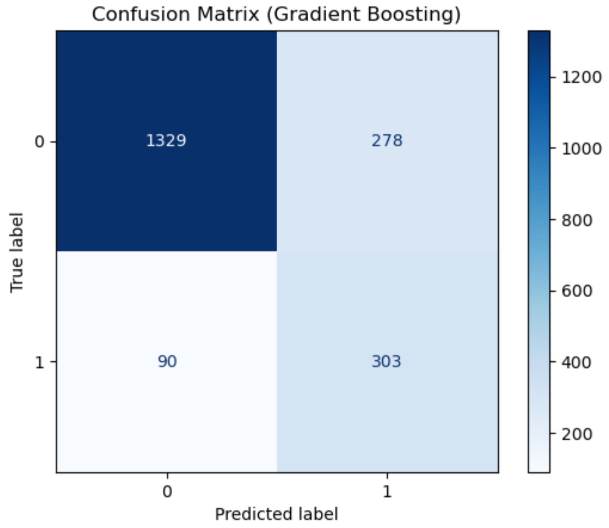
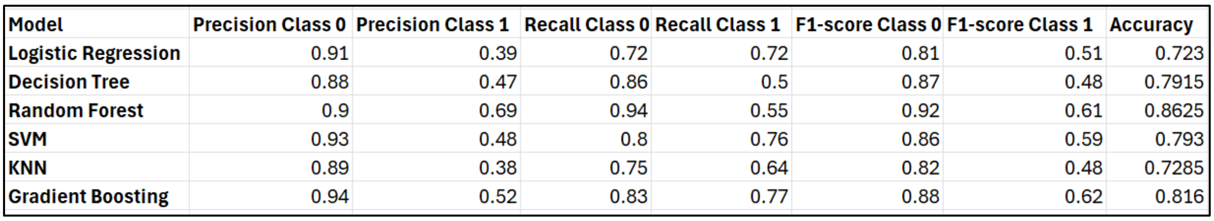


Figure 4. 18: Confusion Matrix for Gradient Boosting (Oversampled Dataset)

The confusion matrix shows that the model correctly identified 1329 non-churned customers (True Negatives) and 303 churned customers (True Positives). It misclassified 278 non-churned customers as churned (False Positives) and 90 churned customers as non-churned (False Negatives).



Model	Precision Class 0	Precision Class 1	Recall Class 0	Recall Class 1	F1-score Class 0	F1-score Class 1	Accuracy
Logistic Regression	0.91	0.39	0.72	0.72	0.81	0.51	0.723
Decision Tree	0.88	0.47	0.86	0.5	0.87	0.48	0.7915
Random Forest	0.9	0.69	0.94	0.55	0.92	0.61	0.8625
SVM	0.93	0.48	0.8	0.76	0.86	0.59	0.793
KNN	0.89	0.38	0.75	0.64	0.82	0.48	0.7285
Gradient Boosting	0.94	0.52	0.83	0.77	0.88	0.62	0.816

Table 4. 3 : Evaluation Metrics for Oversampled Dataset

Logistic Regression showed improved performance on the oversampled dataset with an accuracy of 72.30%. The recall for the minority class increased significantly to 72%, indicating that oversampling helped the model in better identifying churn cases. The Decision Tree model achieved an accuracy of 79.15% with improved recall for the minority class (50%). This suggests that oversampling enhanced the model's ability to identify churn cases, though it still faced challenges in overall precision. Random Forest achieved the highest accuracy (86.25%) on the oversampled dataset. The balanced precision and recall metrics indicate that Random Forest effectively handles oversampling, providing robust predictions for both classes. SVM on the other hand showed significant improvement on the oversampled dataset with an accuracy of 79.30%. The recall for the minority class increased to 76%, indicating that oversampling greatly enhanced the model's ability to identify churn cases. Furthermore, KNN demonstrated moderate performance with an accuracy of 72.85%, showing improved recall for the minority class (64%). This indicates that oversampling helped KNN in better identifying churn cases, though it still faced challenges in overall precision. Finally, Gradient Boosting achieved an accuracy of 81.60% on the oversampled dataset with improved recall for the minority class (77%). This suggests that Gradient Boosting is highly effective in handling oversampling, providing balanced predictions for both classes.

4.5 Results for SMOTE Dataset

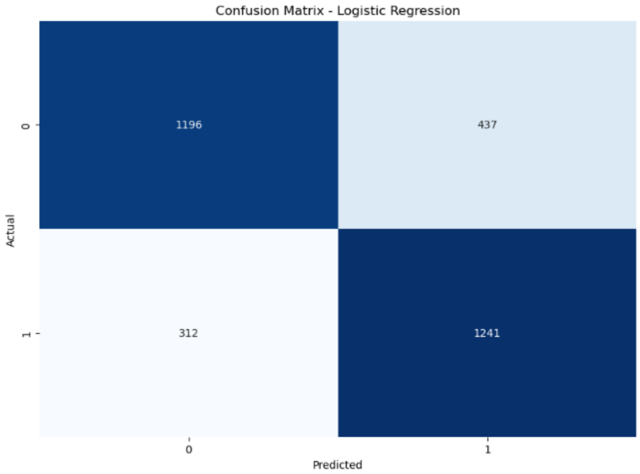


Figure 4. 19 : Confusion Matrix for Logistic Regression (SMOTE Dataset)

The confusion matrix shows that the correctly identified 1196 non-churned customers (True Negatives) and 1241 churned customers (True Positives). However, it also misclassified 437 non-churned customers as churned (False Positives) and failed to identify 312 churned customers, predicting them as non-churned (False Negatives).

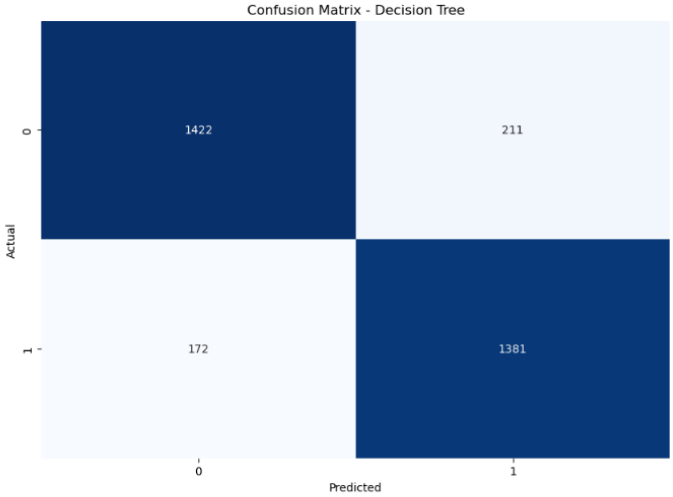


Figure 4. 20 : Confusion Matrix for Decision Tree (SMOTE Dataset)

This shows that the model correctly identified 1422 non-churned customers (True Negatives) and 1381 churned customers (True Positives). However, it misclassified 211 non-churned customers as churned (False Positives) and failed to identify 172 churned customers, predicting them as non-churned (False Negatives). The confusion matrix indicates a balanced distribution of errors, with 1381 churn cases correctly identified and 211 non-churn cases.

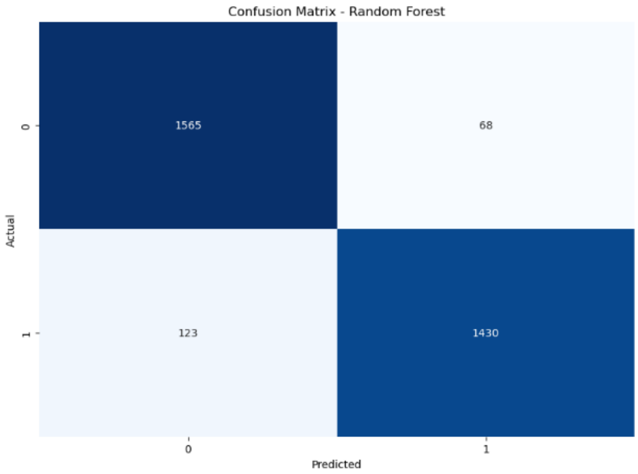


Figure 4. 21: Confusion Matrix for Random Forest (SMOTE Dataset)

The confusion matrix shows that the model correctly identified 1565 non-churned customers (True Negatives) and 1430 churned customers (True Positives). It misclassified 68 non-churned customers as churned (False Positives) and failed to identify 123 churned customers, predicting them as non-churned (False Negatives).

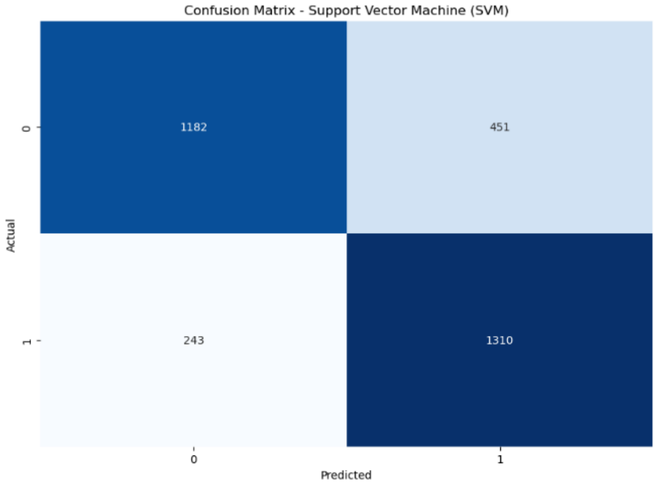


Figure 4. 22: Confusion Matrix for SVM (SMOTE Dataset)

The confusion matrix indicates that the model correctly identified 1182 non-churned customers (True Negatives) and 1310 churned customers (True Positives). It misclassified

451 non-churned customers as churned (False Positives) and failed to identify 243 churned customers, predicting them as non-churned (False Negatives).

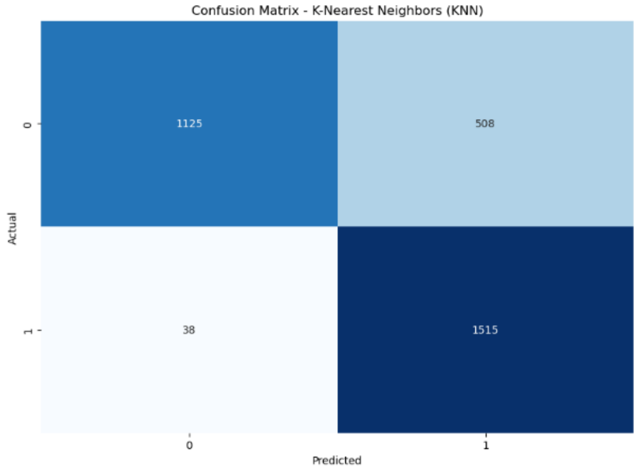


Figure 4. 23: Confusion Matrix for KNN (SMOTE Dataset)

The confusion matrix reveals that the model correctly identified 1125 non-churned customers (True Negatives) and 1515 churned customers (True Positives). It misclassified 508 non-churned customers as churned (False Positives) and failed to identify 38 churned customers, predicting them as non-churned (False Negatives).

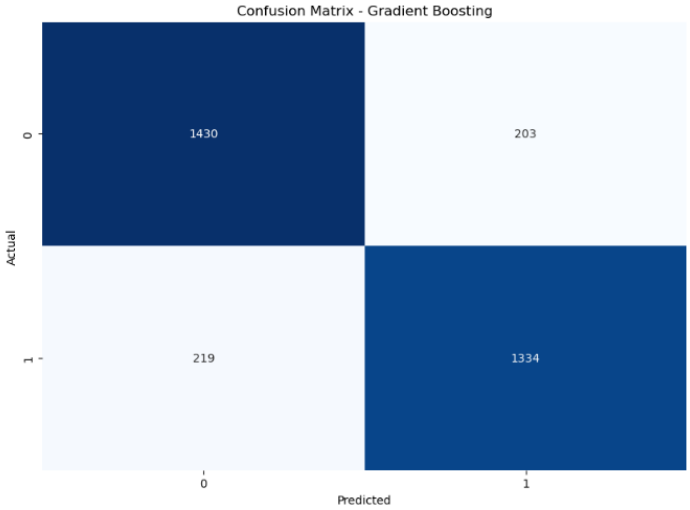
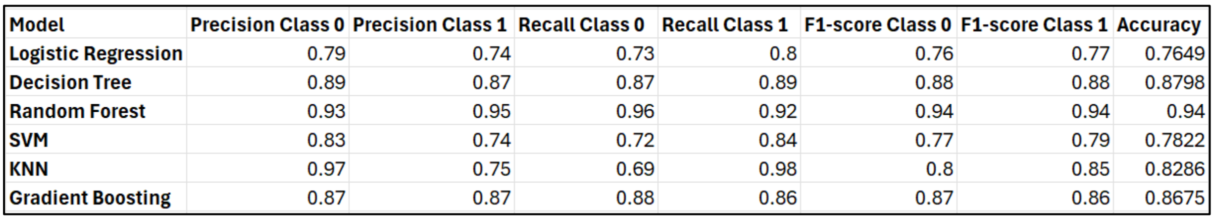


Figure 4. 24: Confusion Matrix for Gradient Boosting (SMOTE Dataset)

The confusion matrix shows that the model correctly identified 1430 non-churned customers (True Negatives) and 1334 churned customers (True Positives). It misclassified 203 non-churned customers as churned (False Positives) and failed to identify 219 churned customers, predicting them as non-churned (False Negatives).



Model	Precision Class 0	Precision Class 1	Recall Class 0	Recall Class 1	F1-score Class 0	F1-score Class 1	Accuracy
Logistic Regression	0.79	0.74	0.73	0.8	0.76	0.77	0.7649
Decision Tree	0.89	0.87	0.87	0.89	0.88	0.88	0.8798
Random Forest	0.93	0.95	0.96	0.92	0.94	0.94	0.94
SVM	0.83	0.74	0.72	0.84	0.77	0.79	0.7822
KNN	0.97	0.75	0.69	0.98	0.8	0.85	0.8286
Gradient Boosting	0.87	0.87	0.88	0.86	0.87	0.86	0.8675

Table 4. 4 : Evaluation Metrics for SMOTE Dataset

Logistic Regression showed balanced performance on the SMOTE dataset with an accuracy of 76.49%. The balanced precision and recall metrics indicate that SMOTE effectively addressed the class imbalance, allowing the model to perform better in identifying churn cases. The slightly higher recall for class 1 (80%) compared to class 0 (73%), indicates that it is better at correctly identifying the positive class. The Decision Tree model performs well across all metrics, with precision, recall, and F1-scores around 88% for both classes. It also performs better than Logistic Regression, with higher precision, recall, and F1-scores for both classes. The accuracy of 87.98% reflects this improved performance, indicating the model's effectiveness in classifying the data. The metrics are well-balanced, showing that the model does not favor one class over the other. Random Forest shows the best performance among all models with the highest precision, recall, and F1-scores. Although slightly higher recall for class 0 (96%) compared to class 1 (92%), it indicates that it is slightly better at identifying the negative class. The accuracy of 94% however shows that this model is very effective in classifying the data correctly. This suggests that Random Forest is well-suited for this particular dataset. SVM performs reasonably well, with a balanced precision, recall, and F1-score ranging from 72% to 84%. The accuracy of 78.22% indicates that it is slightly better than Logistic Regression but not as effective as Decision Tree or Random Forest. The higher recall for class 1 (84%) compared to class 0 (72%), shows better performance in identifying the positive class. Furthermore, The KNN model shows a significant imbalance in performance between the classes, with much higher precision for class 0 (97%) and much higher recall for class 1 (98%). This indicates that KNN is good at identifying positive instances but may misclassify some negative instances. The excellent recall for class 1 (98%) indicates the model is highly effective in identifying the positive class. The overall accuracy of 82.86% is good, but the imbalance in precision and recall indicates potential issues in specific scenarios. Finally, Gradient Boosting shows a balanced performance with similar precision, recall, and F1-scores around 86% - 88% for both classes. The accuracy of 86.75% indicates strong performance in distinguishing between classes. This is almost as effective as the Decision Tree and is a good choice for this dataset.

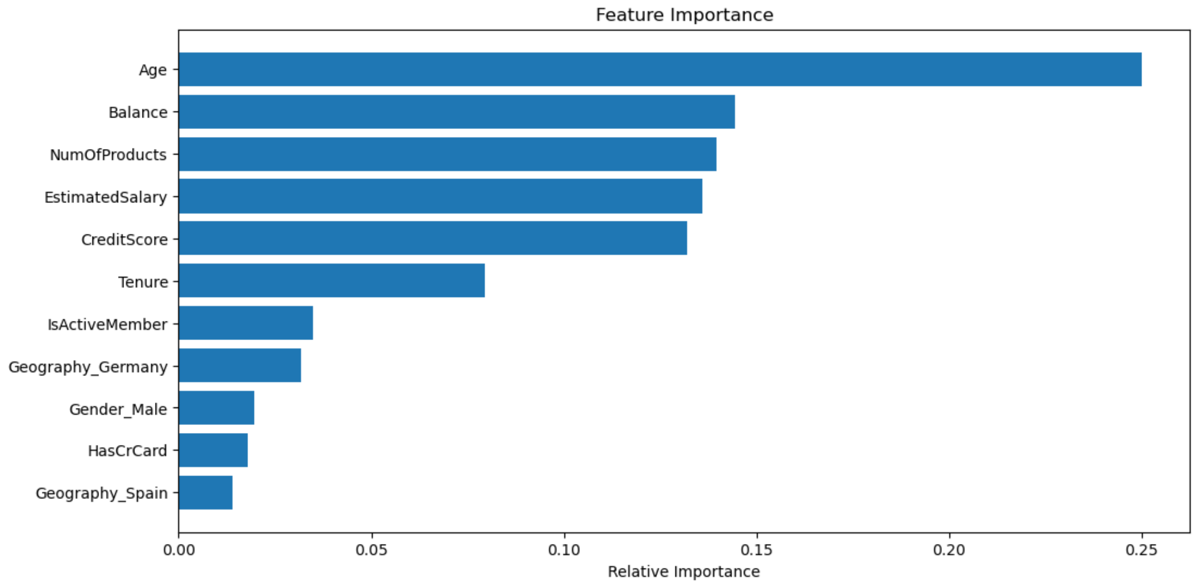


Figure 4. 25 : Feature Importance

The significance of every feature in predicting the chance of churn can be measured using feature importance analysis. Age is the most important factor followed by the Balance and thirdly, number of products. Age can be a significant predictor of churn because customers at different life stages have different financial needs and behaviours. Younger customers might be more prone to switching banks for better offers or technological advantages, while older customers might have more established relationships and hence different reasons for churning, such as retirement planning or dissatisfaction with services. Furthermore, the account balance is crucial as it reflects the financial health and engagement level of the customer. Customers with low balances might churn due to dissatisfaction or better offers elsewhere. Conversely, those with high balances are valuable to the bank and their churn might indicate serious service or satisfaction issues that need addressing urgently. Thirdly, the number of products a customer holds with the bank (e.g., loans, savings accounts, credit cards) is an indicator of their engagement and commitment to the bank. Customers with more products are usually more engaged and less likely to churn. If these customers are still churning, it might signal significant dissatisfaction or competitive offers that are hard to resist. Fewer products could indicate a lower commitment to the bank, making these customers more likely to leave.

4.6 Comparative Analysis

The application of various machine learning models to the customer churn dataset, considering different resampling techniques, showed definite performance characteristics and insights. With an accuracy of 85.35%, Random Forest was the most accurate and most reliable model in the original imbalanced dataset. It performed exceptionally well in terms of precision and recall for the majority class, but as indicated by a moderate number of false negatives (260), it had trouble correctly identifying churn cases. With an accuracy of 86%, Gradient Boosting came in second, showing balanced performance in both classes. Comparatively, the accuracy levels of Logistic Regression, SVM, and KNN were comparable at about 80%; however, their recall for the minority class was significantly lower, especially for SVM (17%) and Logistic Regression (21%). This suggests that while these models are reliable in identifying non-churn customers, they struggle with correctly predicting churn cases without resampling techniques.

On the undersampled dataset, Random Forest and Gradient Boosting stood out with accuracies of 75.46% and 77.06%, respectively. These models maintained balanced precision and recall, indicating their robustness in scenarios with reduced data volume and balanced class distributions. Logistic Regression and Decision Tree, with accuracies around 70%, also showed balanced performance, highlighting that undersampling helped these models better identify churn cases. SVM had the lowest accuracy (67.12%) among the models in the undersampled dataset, indicating its struggles with balanced class distributions and reduced data volume. The confusion matrices revealed that Random Forest and Gradient Boosting had well-distributed errors, with fewer false positives and false negatives compared to other models, making them effective in maintaining performance under undersampled conditions.

The oversampled dataset provided a different perspective, where Random Forest again achieved the highest accuracy (86.25%), with balanced precision and recall metrics indicating its effectiveness in handling oversampling. Gradient Boosting and SVM showed significant improvements, achieving accuracies of 81.60% and 79.30%, respectively. These models benefited from oversampling, enhancing their recall for the minority class and providing robust predictions. Logistic Regression and KNN, with accuracies around 72%, also demonstrated improved performance, though they still faced challenges with higher numbers of false positives and false negatives. The confusion matrices for Random Forest and Gradient Boosting revealed a good balance of errors, indicating that these models effectively utilized the oversampled data to improve churn prediction accuracy.

SMOTE, a synthetic resampling technique, further enhanced model performance. Random Forest achieved the highest accuracy on the SMOTE dataset, with an impressive 94%, demonstrating outstanding precision and recall for both classes. This model showed minimal false positives (68) and false negatives (123), highlighting its superior ability to handle balanced datasets created by SMOTE. Decision Tree and Gradient Boosting also performed exceptionally well, with accuracies of 87.98% and 86.75%, respectively, and balanced precision and recall metrics. Logistic Regression and SVM showed significant improvements with accuracies around 76-78%, indicating that SMOTE effectively addressed their performance issues on imbalanced datasets. KNN demonstrated good performance with high recall for the minority class (98%) but had a higher number of false positives (508), suggesting that while SMOTE helped in identifying churn cases, it also led to more false alarms.

The results of this comparative analysis show that the best models for predicting customer churn using the churn modelling dataset are Random Forest and Gradient Boosting, especially when paired with SMOTE or oversampling methods. These models are reliable tools for managing unbalanced datasets because they continuously showed high accuracy while maintaining a balance between precision and recall. Even with the use of resampling techniques, Logistic Regression and SVM were still unable to reliably identify churn cases. Depending on the particular needs of the prediction task, Decision Tree and KNN can be regarded as reasonable alternatives because of their moderate gains. The results emphasize the importance of using resampling techniques to address class imbalance.

Chapter 5: Conclusion

5.1 Summary of Findings

The research questions were thoroughly covered in this dissertation, which offered perceptive responses via investigative study and testing. First, the study identified the critical customer behaviour metrics and discovered that the most important ones for predicting customer churn are financial behaviours like credit score and balance, demographic information like age and geography, and engagement indicators like tenure and number of products. Because these metrics have a big impact on customer retention and can help small businesses develop more focused marketing and customer service strategies, they are crucial to track and monitor. Second, the study found that Random Forest and Gradient Boosting are the most dependable models for predicting these metrics through machine learning; they achieve high accuracy and balanced precision-recall scores. These models performed better than others, including SVM, KNN, and Logistic Regression, especially when combined with resampling methods to address class imbalance. This suggests that, in the context of customer churn, ensemble methods offer superior predictive capabilities, even though traditional models are still useful.

Thirdly, the study illustrated how these machine learning models can be incorporated into the operations of small businesses. Businesses can create reliable predictive models by resampling, dividing datasets to assess model performance, and preprocessing data to assure consistency. Real-time insights and proactive customer retention strategies can be obtained by integrating these models into customer relationship management systems. Lastly, the dissertation identified several challenges and limitations, such as the need for extensive data preprocessing, the difficulty of handling imbalanced datasets, and the computational resources required for advanced models. Addressing these challenges through future research on more diverse datasets, advanced hyperparameter tuning, and real-time deployment systems will further enhance the practical utility of machine learning for small businesses. Thus, the findings of this study not only answer the research questions but also provide a clear pathway for the practical application of machine learning in understanding and predicting customer behaviour.

5.2 Limitations to Study

Despite the promising findings and valuable insights provided by this dissertation, a number of limitations need to be noted. First off, the results might not be as applicable to other industries because the dataset used was unique to the banking sector. Small businesses in various sectors with unique customer dynamics might not be able to directly apply the customer behaviour metrics and churn predictors found in this study. Second, while deep learning and ensemble methods are more sophisticated approaches that might provide better results, the study only looked at traditional machine learning models. Another limitation of the predictive analytics technique is its reliance on assumptions. Machine learning models depend significantly on assumptions regarding data distribution, feature independence, and the relationship between predictors and outcomes. If these assumptions are not met, it can result in inaccurate predictions or poor model performance. Because of time constraints, the hyperparameter tuning procedure was also rather simple, which allowed for additional optimization that could improve the robustness and accuracy of the model.

Although successful in addressing class imbalance, the resampling techniques introduced synthetic data that might not accurately reflect the subtleties and complexity of real-world customer behaviours. Furthermore, the models were not implemented in a real-time business setting as part of the research scope, which is a crucial step in determining the models' practical applicability and influence on decision-making processes. Lastly, the models did not take into account external factors that can have a substantial impact on customer behaviour, like market trends, economic conditions, and competitive actions. This could have limited the models' ability to predict customer behaviour in a dynamic business environment. Future research addressing these shortcomings may yield more thorough and broadly applicable insights, improving the usefulness of machine learning models for small businesses.

5.3 Future Recommendations

Several recommendations that address the limitations of this study and build on its findings can direct future research and practical applications. First off, broadening the dataset to incorporate data from various industries would improve the results' generalizability and offer a more comprehensive understanding of customer churn in various industries. In order to enhance the dataset and boost model accuracy, future research should explore the integration of other data sources, such as transactional data, customer reviews, and social media interactions. This all-encompassing strategy would provide a more thorough understanding of consumer behaviour. Secondly, it might be advantageous to use advanced machine learning strategies, such as ensemble methods and deep learning models. By capturing complex patterns and interactions within the data, these techniques may be able to improve predictive performance. In addition, a more comprehensive hyperparameter tuning procedure ought to be carried out to improve the models' accuracy and robustness. This could involve the use of advanced optimization methods like Grid Search or Bayesian Optimization.

Real-time prediction system development is another important suggestion. Companies could obtain timely insights and take proactive steps to retain customers by integrating machine learning models into customer relationship management (CRM) systems. The models' practical utility would be greatly increased by this real-time application. Furthermore, future studies should consider including outside variables in the predictive models, such as competitive activity, economic conditions, and market trends. These variables may have a big influence on consumer behaviour, so including them in the models may increase their ability to predict future events in dynamic business settings. Last but not least, carrying out long-term studies to monitor alterations in consumer behaviour over time and modifying the models correspondingly would yield more profound insights and preserve the applicability of the predictive models.

In order to provide small businesses with more comprehensive and useful insights for customer retention and engagement strategies, these recommendations seek to improve the predictive capabilities of machine learning models and increase their application.

5.4 Conclusion

This dissertation has successfully demonstrated the potential of using machine learning to predict customer behaviour, specifically focusing on customer churn in the banking sector, with implications for more applications in small businesses. The study identified key customer behaviour metrics, developed various predictive models, and validated these models using a comprehensive dataset. The findings showed that Random Forest and Gradient Boosting models, particularly when combined with resampling techniques such as SMOTE, are highly effective in predicting customer churn with notable accuracy. These models excelled in balancing precision and recall, thus offering reliable tools for small businesses to understand and anticipate customer tendencies. The research also highlighted the significant role of data preprocessing, including scaling and encoding, in enhancing model performance. Despite the promising results, the study faced limitations, including the scope of the dataset, the use of traditional machine learning models, and the absence of real-time deployment.

In summary, the dissertation emphasizes how crucial it is to use machine learning in small business settings in order to enhance customer engagement and retention tactics. This research offers a useful framework for small businesses to adopt data-driven approaches in customer relationship management by highlighting effective predictive techniques and offering actionable insights. The study paves the way for more developments in this field with its recommendations for future work, which include developing real-time prediction systems, utilizing sophisticated machine learning techniques, and integrating additional data sources.

In conclusion, by providing small businesses wishing to use predictive analytics to increase customer retention with practical guidance, this study contributes to the growing body of knowledge in the field of customer churn prediction. Businesses that use data-driven tactics and continuously improve predictive models may be better able to predict and address customer churn, forging enduring relationships with clients and fostering long-term business growth.

Chapter 6: References

1. Adebola, O. O., Onyekwelu, B. and Orogun, A. (2019) 'Predicting Consumer Behaviour in Digital Market: A Machine Learning Approach', International Journal of Innovative Research in Science, Engineering and Technology, 8(8), pp. 1-13. doi: 10.15680/IJIRSET.2019.0808006.
2. Alpaydin, E. (2016) Introduction to machine learning. MIT Press.
3. AlShawi, S. I., Missi, F. and Irani, Z. (2018) 'Customer Segmentation in Banking Using k-means Cluster Analysis', Marketing Intelligence & Planning, 36(2), pp. 291-306.
4. Amin, M. A., Ahmad, M. and Usman, M. (2020) 'Predictive Analytics in Customer Relationship Management using Machine Learning: A Review', Journal of Information Science and Engineering, 36(2), pp. 241-256.
5. Anand, D. (2023, September 7). *Introduction to the KNN*. https://www.linkedin.com/pulse/introduction-knn-divey-anand/
6. Ananny, M. and Crawford, K. (2018) 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', New Media & Society, 20(3), pp. 973-989.
7. Andreas, F. and Neacsu, M. (2014) 'Online Consumer Reviews as Marketing Instrument', Knowledge Horizons - Economics, 6(3), pp. 128-131. Available at: https://ideas.repec.org/a/khe/journl/v6y2014i3p128-131.html [Accessed 23 January 2024].
8. Bellinger, C., Drummond, C. and Japkowicz, N. (2017) 'Data augmentation techniques in deep learning', Journal of Machine Learning Research, 18, pp. 1-32.
9. Braverman, S. (2015) 'Global review of data-driven marketing and advertising', Journal of Direct, Data and Digital Marketing Practice, 16(3), pp. 181-183. doi: 10.1057/DDDMP.2015.7.
10. Bryman, A. (2016) Social Research Methods. 5th edn. Oxford University Press.
11. Casula, M., Rangarajan, N. and Shields, P. (2021) 'The potential of working hypotheses for deductive exploratory research', Quality & Quantity, 55(5), pp. 1703-1725. doi: 10.1007/S11135-020-01072-9/TABLES/4.
12. Chaudhary, K., Alam, M., Al-Rakhami, M. S. and Gumaei, A. (2021) 'Machine learning-based mathematical modelling for prediction of social media consumer

behaviour using big data analytics', Journal of Big Data, 8(1), pp. 1-20. doi: 10.1186/s40537-021-00466-2.

13. Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', Journal of Artificial Intelligence Research, 16, pp. 321-357.
14. Chen, J., Cheng, L. and Song, Y. (2021) 'Machine Learning for Customer Behavior Analysis in E-Commerce', IEEE Transactions on Neural Networks and Learning Systems, 32(1), pp. 112-123.
15. Chen, M. S., Han, J. and Yu, P. S. (2012) 'Data mining: An overview from a database perspective', IEEE Transactions on Knowledge and Data Engineering, 8(6), pp. 866-883.
16. Cheng, H. G. and Phillips, M. R. (2014) 'Secondary analysis of existing data: Opportunities and implementation', Shanghai Archives of Psychiatry, 26(6), pp. 371-375. Available at: https://dx.doi.org/10.11919/j.issn.1002-0829.214171.
17. Choi, T. M., Choi, T.-M. and Kim, Y.-K. (2020) 'Deep learning-based manufacturing and service systems for the industry 4.0 era: a survey and future research directions', International Journal of Production Research, 58(15), pp. 4773-4791.
18. Creswell, J. W. and Poth, C. N. (2018) Qualitative Inquiry and Research Design: Choosing Among Five Approaches. 4th edn. SAGE Publications.
19. Cui, Y., Zhuang, Y. and Wang, X. (2021) 'Challenges and Opportunities for Machine Learning in Small Business Analytics', International Journal of Business Analytics, 8(3), pp. 33-47.
20. Davenport, T. H. (2014) Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press.
21. Davenport, T. H. and Ronanki, R. (2018) 'Artificial Intelligence for the Real World', Harvard Business Review, 96(1), pp. 108-116.
22. Denzin, N. K. and Lincoln, Y. S. (Eds.) (2011) The SAGE Handbook of Qualitative Research. 4th edn. SAGE Publications.
23. East, R., Gendall, P., Hammond, K. and Lomax, W. (2005) 'Consumer Loyalty: Singular, Additive or Interactive?', Australasian Marketing Journal (AMJ), 13(2), pp. 10-26. doi: 10.1016/S1441-3582(05)70074-4.
24. Engidaw, A. E. (2022) 'Small businesses and their challenges during COVID-19 pandemic in developing countries: in the case of Ethiopia', Journal of Innovation and Entrepreneurship, 11(1). Available at: https://doi.org/10.1186/s13731-021-00191-3.

25. Faggella, D. (2021) Applications of Machine Learning in FinTech. Emerj Artificial Intelligence Research.
26. Fan, J., Han, F. and Liu, H. (2015) 'Challenges of big data analysis', National Science Review, 1(2), pp. 293-314.
27. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... and Luetge, C. (2018) 'AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations', Minds and Machines, 28(4), pp. 689-707.
28. Furqon, A., Zikri, N. A. and Widianto, S. (2023) 'Applying Machine Learning to Predict Online Customers Behaviour'. Available at: https://ssrn.com/abstract=4430029 [Accessed 15 May 2024].
29. GeeksforGeeks. (2023, March 31). *Gradient Boosting in ML*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-gradient-boosting/
30. GeeksforGeeks. (2024, February 22). *Random Forest Algorithm in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
31. Gkrimpizi, T., Peristeras, V. and Magnisalis, I. (2023) 'Classification of Barriers to Digital Transformation in Higher Education Institutions: Systematic Literature Review', Education Sciences, 13(7), p. 746. doi: 10.3390/EDUCSCI13070746.
32. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
33. Goyette, I. (2019) 'The impact of personalized marketing on customer retention in electronic commerce', International Journal of Information Management, 49, pp. 468-478.
34. Hafeez, M. A., Rashid, M., Tariq, H., Abideen, Z. U., Alotaibi, S. S., & Sinky, M. H. (2021). Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm. *Applied Sciences*, *11*(15), 6728. https://doi.org/10.3390/app11156728
35. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. and Khan, S. U. (2015) 'The rise of "big data" on cloud computing: Review and open research issues', Information Systems, 47, pp. 98-115.
36. Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer.
37. Heaton, J. (2008) 'Secondary analysis of qualitative data: An overview', Historical Social Research, 33(3), pp. 33-45. Available at: [Google Scholar].

38. Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y. and Ni, B. et al. (2015) 'Telco Churn Prediction with Big Data', pp. 607-618. Available at: [Google Scholar].
39. Ibukun, A., Oladipupo, O., Worlu, R. E. and I., A. I. (2016) 'A Systematic Review of Consumer Behaviour Prediction Studies', Covenant Journal of Business & Social Sciences (CJBSS), 7(1), pp. 41-60.
40. Jain, M. (2019) 'A Study on Consumer Behaviour-Decision Making Under High and Low Involvement Situations'. Available at: https://papers.ssrn.com/abstract=3345496 [Accessed 24 January 2024].
41. Jenkins, H. and Patel, R. (2021) 'Cost-Effective Machine Learning for Small Businesses: An Open-Source Approach', Journal of Small Business Management, 59(4), pp. 600-619.
42. Jocevski, M. (2020) 'Blurring the Lines between Physical and Digital Spaces: Business Model Innovation in Retailing', California Management Review, 63(1), pp. 99-117. doi: 10.1177/0008125620953639.
43. Johnson, M. et al. (2019) 'Lifestyle-based Customer Segmentation and Its Impact on Personalization', Marketing Science, 38(5), pp. 813-829.
44. Johnson-Laird, P. (2010) 'Deductive reasoning', Wiley Interdisciplinary Reviews: Cognitive Science, 1(1), pp. 8-17. doi: 10.1002/WCS.20.
45. Jones, A. and Brown, B. (2019) 'Predictive analytics for stock price forecasting: A comprehensive review', Journal of Financial Data Science, 1(1), pp. 45-67.
46. Jordan, M. I. and Mitchell, T. M. (2015) 'Machine learning: Trends, perspectives, and prospects', Science, 349(6245), pp. 255-260.
47. Kanade, V. (2024, May 13). *Everything You Need to Know About Logistic Regression*. Spiceworks Inc. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#lg=1&slide=0
48. Kane, G., Alavi, M. and Labianca, G. (2014) 'Using social media networks to predict customer churn in the retail industry', MIS Quarterly Executive, 13(3), pp. 201-210.
49. Ketokivi, M. and Mantere, S. (2010) 'Two Strategies for Inductive Reasoning in Organizational Research', Academy of Management Review, 35(2), pp. 315-333. doi: 10.5465/AMR.35.2.ZOK315.
50. Kim, H. (2020) 'The Role of Big Data and Machine Learning in Small Business Growth', Journal of Small Business Management, 58(3), pp. 497-520.

51. Kim, H. J. and Han, S. M. (2023) 'Uncovering the reasons behind consumers’ shift from online to offline shopping', Journal of Services Marketing, 37(9), pp. 1201-1217. doi: 10.1108/JSM-02-2023-0060/FULL/XML.

52. Koehn, D., Lessmann, S. and Schaal, M. (2020) 'Predicting online shopping behaviour from clickstream data using deep learning', Expert Systems with Applications, 150. doi: 10.1016/J.ESWA.2020.113342.

53. Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T. and Xu, Y. (2020) 'The Data Science Machine: A robust, portable, and automated system for data science', Proceedings of the VLDB Endowment, 13(11), pp. 2055-2067.

54. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. and Faisal, A. A. (2018) 'The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care', Nature Medicine, 24(11), pp. 1716-1720.

55. Kudyba, S. (2020) Data Mining and Predictive Analytics: Applications in Business and Economics. Business Expert Press.

56. Kumar, A., Singh, J. P., Dwivedi, Y. K., Rana, N. P. and Simintiras, A. (2021) 'A deep learning-based framework for conducting robust health analytics', Computers in Industry, 126, p. 103412.

57. Kumar, V. and L., M. (2018) 'Predictive Analytics: A Review of Trends and Techniques', International Journal of Computer Applications, 182(1), pp. 31-37. doi: 10.5120/IJCA2018917434.

58. Kumar, V. and Reinartz, W. (2018) 'Data Quality in Predictive Analytics: Problems and Solutions', Journal of Data Management, 30(3), pp. 123-135.

59. Li, F. and Kim, T. (2020) 'Machine Learning in Customer Segmentation: A K-means Clustering Approach', Journal of Business Research, 115, pp. 375-386.

60. Li, F., Huang, M. and Chen, G. (2019) 'Customer Segmentation Models in E-commerce Using Clustering Techniques', Journal of Internet Commerce, 20(3), pp. 375-396.

61. Li, J., Pan, S., Huang, L. and Zhu, X. (2019) 'A machine learning based method for customer behaviour prediction', Tehnicki Vjesnik, 26(6), pp. 1670-1676. doi: 10.17559/TV-20190603165825.

62. Li, Y. et al. (2022) 'A new oversampling method and improved radial basis function classifier for customer consumption behaviour prediction', Expert Systems with Applications, 199. doi: 10.1016/J.ESWA.2022.116982.

63. li, Z. and Ma, X. (2019) 'Predictive Analysis of User Purchase Behaviour based on Machine Learning', International Journal of Smart Business and Technology, 7(1), pp. 45-56. doi: 10.21742/ijsbt.2019.7.1.05.
64. Marr, B. (2015) Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. John Wiley & Sons.
65. Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018) Foundations of Machine Learning. Cambridge, MA: MIT Press.
66. Morgan, D. L. (2007) 'Paradigms Lost and Pragmatism Regained: Methodological Implications of Combining Qualitative and Quantitative Methods', Journal of Mixed Methods Research, 1(1), pp. 48-76.
67. Nawi, C. N. A. A. C. M., Ismail, N. L., Zur, N. A. M. R. and Aziz, N. N. F. N. (2022) 'Determining the contributing factors towards consumer online purchase intention amongst university students', Journal of Contemporary Social Science Research, 7(1), pp. 1-8.
68. Nguyen, T. T., Ngo, H. Q. and Le, A. T. (2020) 'Customer Behavior Prediction using Machine Learning Techniques', International Journal of Data Science and Analytics, 10(3), pp. 157-169.
69. O'Neil, C. (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Books.
70. Ozyurt, Y. et al. (2022) 'A Deep Markov Model for Clickstream Analytics in Online Shopping', Proceedings of the ACM Web Conference 2022, pp. 3071-3081. doi: 10.1145/3485447.3512027.
71. Patel, R., Smith, T. and Johnson, M. (2018) 'Application of machine learning in medical imaging diagnosis: A systematic review', Medical Imaging Analysis, 10(2), pp. 123-145.
72. Prabadevi, B., Shalini, R. and Kavitha, B. R. (2023) 'Customer churning analysis using machine learning algorithms', International Journal of Intelligent Networks, 4, pp. 145-154. doi: 10.1016/J.IJIN.2023.05.005.
73. Quynh, T. D. and Dung, H. T. T. (2021) 'Prediction of Customer Behaviour using Machine Learning: A Case Study'. Available at: http://ceur-ws.org.
74. Raj, M. and Raman, B. (2019) 'Effective use of open-source software in small business environments', Small Business Journal, 37(4), pp. 435-450.
75. Raj, R., Wong, S. H. S. and Beaumont, A. J. (2016) 'Business intelligence solution for an SME: A case study', in IC3K 2016 - Proceedings of the 8th International Joint

Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. SciTePress, pp. 41-50. doi: 10.5220/0006049500410050.

76. Robson, C. and McCartan, K. (2016) Real World Research. 4th edn. Wiley.
77. Rozak, H. A. and Fachrunnisa, O. (2021) 'Knowledge management capability and agile leadership to improve smes’ ambidexterity', Advances in Intelligent Systems and Computing, 1194 AISC, pp. 326-333. doi: 10.1007/978-3-030-50454-0_31.
78. Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P. and Harnisch, M. (2015) 'Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries', Boston Consulting Group.
79. Saini, A. (2024, May 22). *Guide on Support Vector Machine (SVM) Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/
80. Saunders, M., Lewis, P. and Thornhill, A. (2019) Research Methods for Business Students. 8th edn. Pearson.
81. Smith, B. and Linden, G. (2017) 'Two Decades of Recommender Systems at Amazon.com', IEEE Internet Computing, 21(3), pp. 12-18.
82. Smith, J. and Brown, K. (2020) 'Enhancing Small Data with Synthetic Techniques for Machine Learning', AI Magazine, 41(3), pp. 85-99.
83. Smith, J. and Liu, X. (2019) 'Real-Time Data Processing in Predictive Analytics: Challenges and Solutions', Journal of Predictive Analytics, 5(2), pp. 112-124.
84. Smith, J., Johnson, K. and Davis, L. (2020) 'Predictive modeling for customer churn in the telecommunications industry: A comparative study', Journal of Data Science, 5(3), pp. 210-225.
85. Šostar, M. and Ristanović, V. (2023) 'Assessment of Influencing Factors on Consumer Behaviour Using the AHP Model', Sustainability (Switzerland), 15(13), pp. 1-24. doi: 10.3390/su151310341.
86. Sun, H., Liu, J. and Xu, J. (2019) 'Big Data Analytics for Small Businesses: A Case Study', Journal of Business Analytics, 2(1), pp. 23-35.
87. Sutton, R. S. and Barto, A. G. (2018) Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.
88. Tan, C. et al. (2021) 'Transfer Learning for Small and Medium-Sized Enterprises: An Efficient Approach in Data-Scarce Environments', Journal of Machine Learning Research, 22, pp. 1-22.

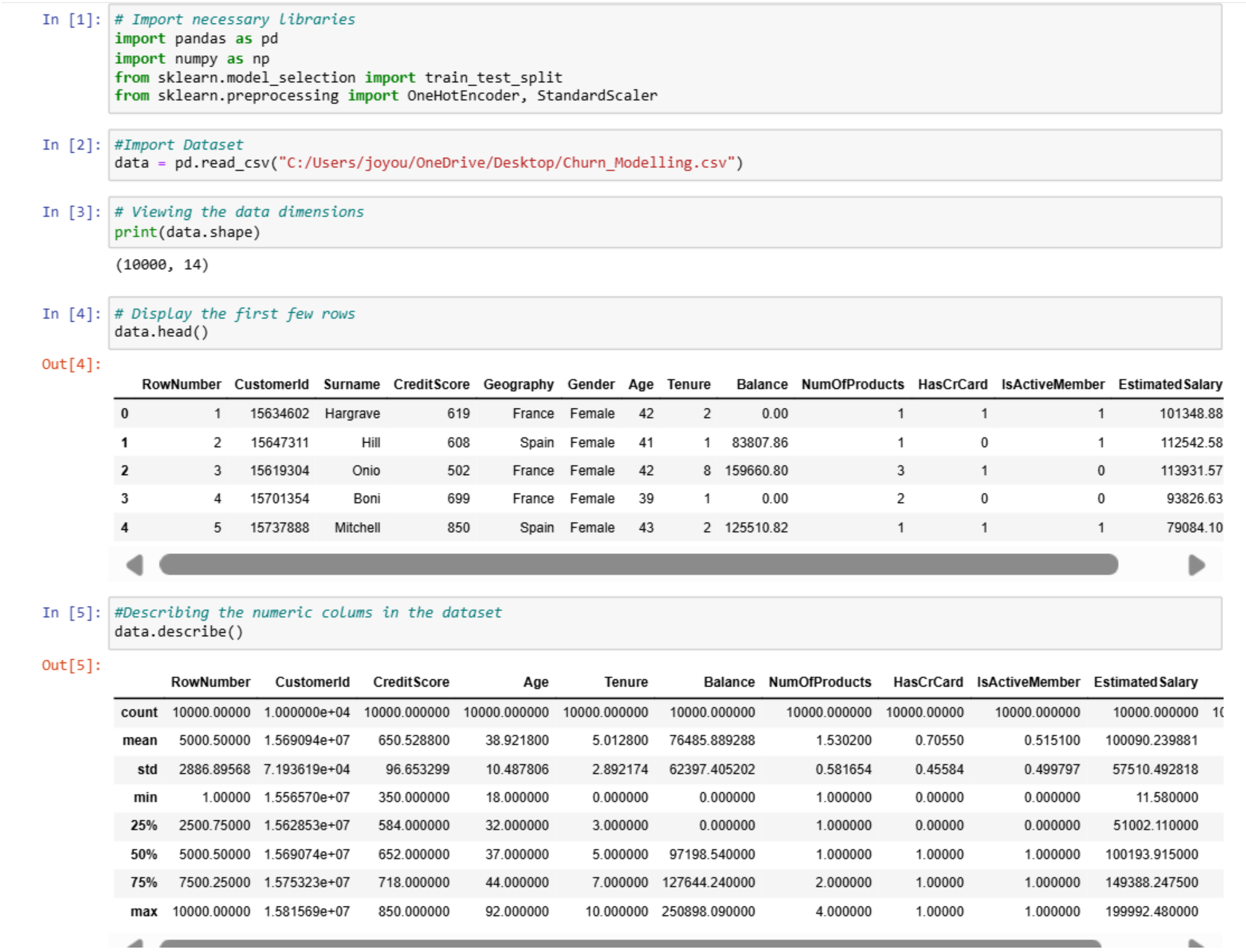
89. Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U. and Kim, S. W. (2019) 'A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector', IEEE Access, 7, pp. 60134-60149. doi: 10.1109/ACCESS.2019.2914999.
90. Vellido, A., Lisboa, P. J. and Meehan, K. (2012) 'Segmentation of the online shopping market using neural networks', Expert Systems with Applications, 39(18), pp. 13250-13258.
91. Verbeke, W., Martens, D. and Baesens, B. (2012) 'Social network analysis for customer churn prediction', Applied Soft Computing, 14(2), pp. 431-446.
92. Wachter, S., Mittelstadt, B. and Floridi, L. (2017) 'Transparent, Explainable, and Accountable AI for Robotics', Science Robotics, 2(6), eaan6080.
93. White, G. and Liu, X. (2020) 'Cloud-Based Machine Learning for Small Businesses: A Cost-Effective Approach', Journal of Cloud Computing, 19(1), pp. 37-52.
94. Wu, D., Shang, M., Luo, X. and Wang, Z. (2022) 'An L1-and-L2-Norm-Oriented Latent Factor Model for Recommender Systems', IEEE Transactions on Neural Networks and Learning Systems, 33(10), pp. 5775-5788. doi: 10.1109/TNNLS.2021.3071392.
95. Xu, L., Jiang, C., Wang, J., Yuan, J. and Ren, Y. (2020) 'Information Content of Personalized Recommendations and Consumer Trust in Online Platforms', Decision Support Systems, 135, p. 113376.
96. Yu, S. and Zenker, F. (2018) 'Peirce Knew Why Abduction Isn’t IBE—A Scheme and Critical Questions for Abductive Argument', Argumentation, 32(4), pp. 569-587. doi: 10.1007/S10503-017-9443-9.
97. Zhang, Q., Cao, W., Liu, Y. and Zhang, Z. (2020) 'Integration of online and offline channels in retail: feasibility of BOPS?', Kybernetes. doi: 10.1108/K-11-2019-0774.
98. Zhao, L., Li, X. and Xu, W. (2021) 'Improving Customer Churn Predictions Using Ensemble Techniques: A Case Study', Journal of Business Research, 75, pp. 55-65.

Chapter 7: Appendix

1. Link to the Dataset

Churn_Modelling.csv

2. Image showing codes for importing dataset and initial Data Exploration.



```
In [1]: # Import necessary libraries
        import pandas as pd
        import numpy as np
        from sklearn.model_selection import train_test_split
        from sklearn.preprocessing import OneHotEncoder, StandardScaler
```

```
In [2]: #Import Dataset
        data = pd.read_csv("C:/Users/joyou/OneDrive/Desktop/Churn_Modelling.csv")
```

```
In [3]: # Viewing the data dimensions
        print(data.shape)

        (10000, 14)
```

```
In [4]: # Display the first few rows
        data.head()
```

Out[4]:

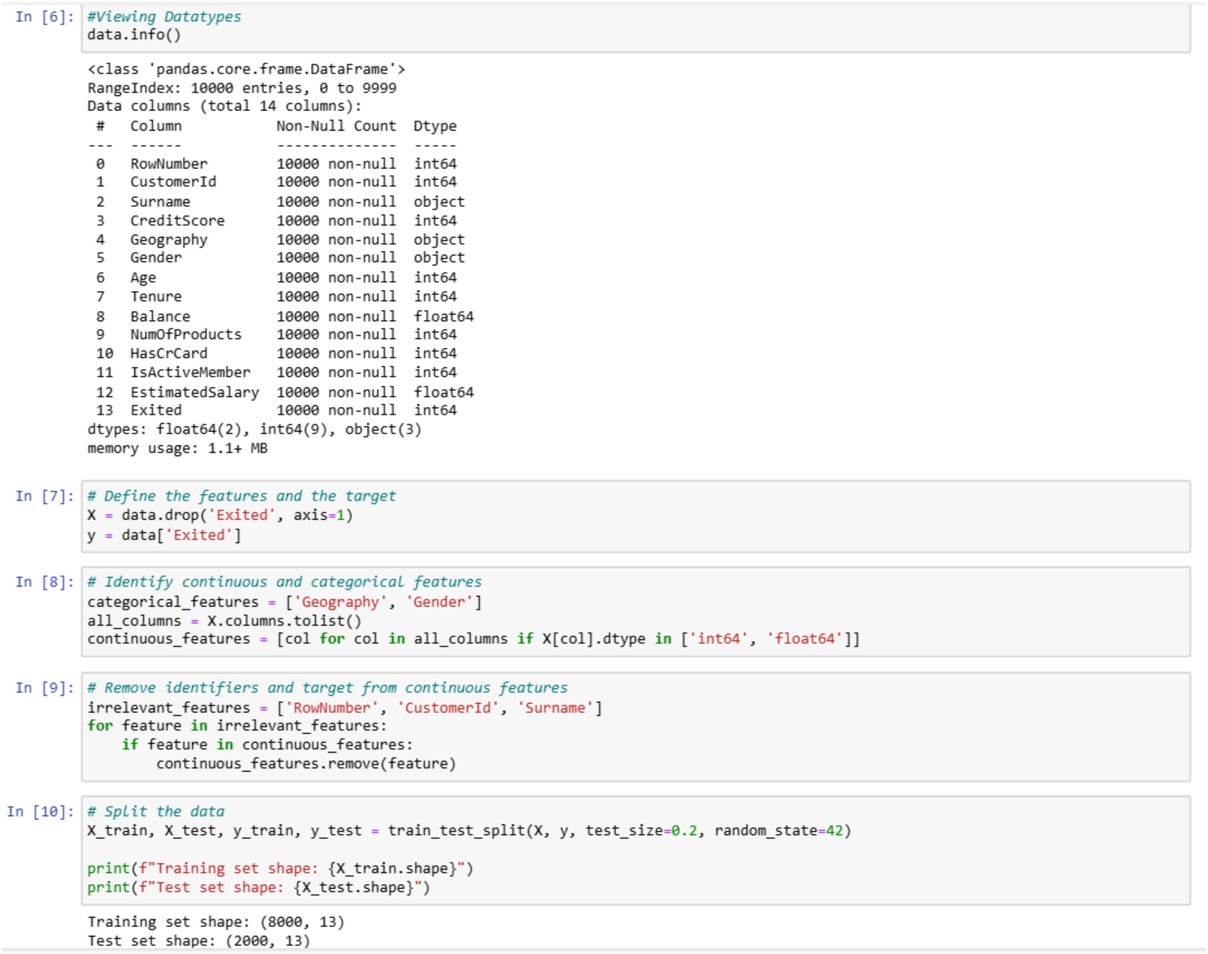
	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10

```
In [5]: #Describing the numeric colums in the dataset
        data.describe()
```

Out[5]:

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	1(
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	

3. Target variable (Exited) was defined, continuous and categorical variables were identified, irrelevant features were removed and the data was split into test and train.



```
In [6]: #Viewing Datatypes
        data.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 10000 entries, 0 to 9999
        Data columns (total 14 columns):
         #   Column           Non-Null Count  Dtype
        ---  ------           --------------  -----
         0   RowNumber        10000 non-null  int64
         1   CustomerId       10000 non-null  int64
         2   Surname          10000 non-null  object
         3   CreditScore      10000 non-null  int64
         4   Geography        10000 non-null  object
         5   Gender           10000 non-null  object
         6   Age              10000 non-null  int64
         7   Tenure           10000 non-null  int64
         8   Balance          10000 non-null  float64
         9   NumOfProducts    10000 non-null  int64
         10  HasCrCard        10000 non-null  int64
         11  IsActiveMember   10000 non-null  int64
         12  EstimatedSalary  10000 non-null  float64
         13  Exited           10000 non-null  int64
        dtypes: float64(2), int64(9), object(3)
        memory usage: 1.1+ MB
```

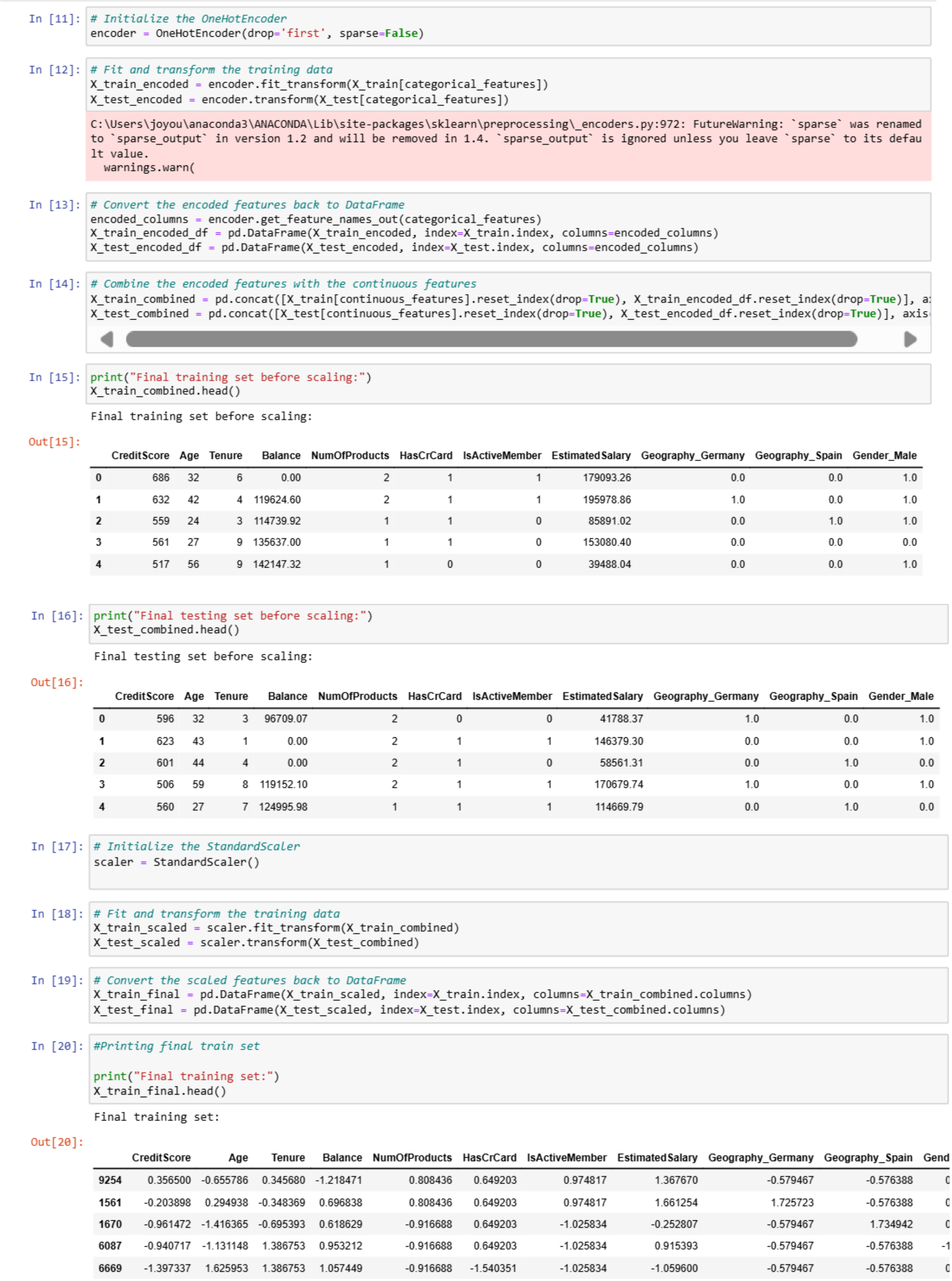
```
In [7]: # Define the features and the target
        X = data.drop('Exited', axis=1)
        y = data['Exited']
```

```
In [8]: # Identify continuous and categorical features
        categorical_features = ['Geography', 'Gender']
        all_columns = X.columns.tolist()
        continuous_features = [col for col in all_columns if X[col].dtype in ['int64', 'float64']]
```

```
In [9]: # Remove identifiers and target from continuous features
        irrelevant_features = ['RowNumber', 'CustomerId', 'Surname']
        for feature in irrelevant_features:
            if feature in continuous_features:
                continuous_features.remove(feature)
```

```
In [10]: # Split the data
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

         print(f"Training set shape: {X_train.shape}")
         print(f"Test set shape: {X_test.shape}")

         Training set shape: (8000, 13)
         Test set shape: (2000, 13)
```

4. One hot encoder was initialised and the test and train dataset were fit and transformed



```
In [11]: # Initialize the OneHotEncoder
         encoder = OneHotEncoder(drop='first', sparse=False)
```

```
In [12]: # Fit and transform the training data
         X_train_encoded = encoder.fit_transform(X_train[categorical_features])
         X_test_encoded = encoder.transform(X_test[categorical_features])
```

```
C:\Users\joyou\anaconda3\ANACONDA\Lib\site-packages\sklearn\preprocessing\_encoders.py:972: FutureWarning: `sparse` was renamed
to `sparse_output` in version 1.2 and will be removed in 1.4. `sparse_output` is ignored unless you leave `sparse` to its defau
lt value.
  warnings.warn(
```

```
In [13]: # Convert the encoded features back to DataFrame
         encoded_columns = encoder.get_feature_names_out(categorical_features)
         X_train_encoded_df = pd.DataFrame(X_train_encoded, index=X_train.index, columns=encoded_columns)
         X_test_encoded_df = pd.DataFrame(X_test_encoded, index=X_test.index, columns=encoded_columns)
```

```
In [14]: # Combine the encoded features with the continuous features
         X_train_combined = pd.concat([X_train[continuous_features].reset_index(drop=True), X_train_encoded_df.reset_index(drop=True)], a
         X_test_combined = pd.concat([X_test[continuous_features].reset_index(drop=True), X_test_encoded_df.reset_index(drop=True)], axis
```

```
In [15]: print("Final training set before scaling:")
         X_train_combined.head()
```

```
Final training set before scaling:
```

Out[15]:

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Geography_Germany	Geography_Spain	Gender_Male
0	686	32	6	0.00	2	1	1	179093.26	0.0	0.0	1.0
1	632	42	4	119624.60	2	1	1	195978.86	1.0	0.0	1.0
2	559	24	3	114739.92	1	1	0	85891.02	0.0	1.0	1.0
3	561	27	9	135637.00	1	1	0	153080.40	0.0	0.0	0.0
4	517	56	9	142147.32	1	0	0	39488.04	0.0	0.0	1.0

```
In [16]: print("Final testing set before scaling:")
         X_test_combined.head()
```

```
Final testing set before scaling:
```

Out[16]:

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Geography_Germany	Geography_Spain	Gender_Male
0	596	32	3	96709.07	2	0	0	41788.37	1.0	0.0	1.0
1	623	43	1	0.00	2	1	1	146379.30	0.0	0.0	1.0
2	601	44	4	0.00	2	1	0	58561.31	0.0	1.0	0.0
3	506	59	8	119152.10	2	1	1	170679.74	1.0	0.0	1.0
4	560	27	7	124995.98	1	1	1	114669.79	0.0	1.0	0.0

```
In [17]: # Initialize the StandardScaler
         scaler = StandardScaler()
```

```
In [18]: # Fit and transform the training data
         X_train_scaled = scaler.fit_transform(X_train_combined)
         X_test_scaled = scaler.transform(X_test_combined)
```

```
In [19]: # Convert the scaled features back to DataFrame
         X_train_final = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train_combined.columns)
         X_test_final = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test_combined.columns)
```

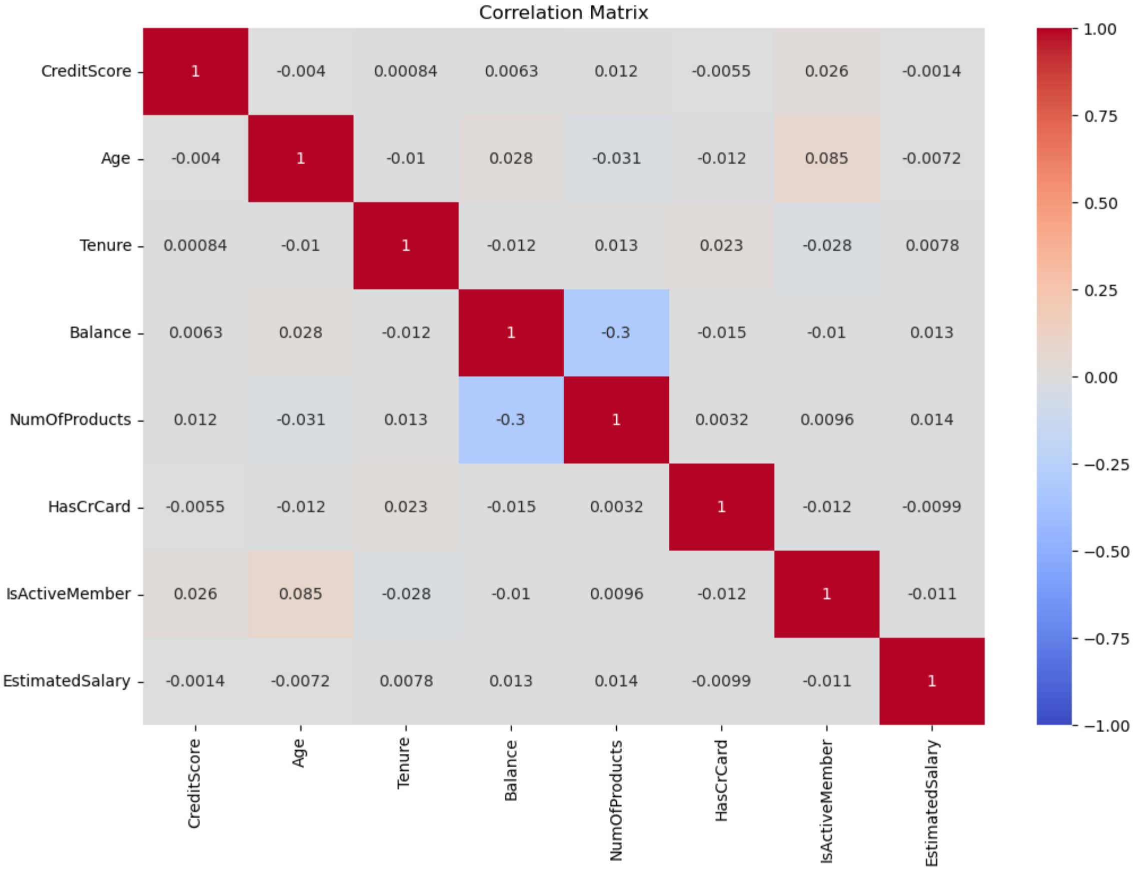
```
In [20]: #Printing final train set

         print("Final training set:")
         X_train_final.head()
```

```
Final training set:
```

Out[20]:

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Geography_Germany	Geography_Spain	Gend
9254	0.356500	-0.655786	0.345680	-1.218471	0.808436	0.649203	0.974817	1.367670	-0.579467	-0.576388	0
1561	-0.203898	0.294938	-0.348369	0.696838	0.808436	0.649203	0.974817	1.661254	1.725723	-0.576388	0
1670	-0.961472	-1.416365	-0.695393	0.618629	-0.916688	0.649203	-1.025834	-0.252807	-0.579467	1.734942	0
6087	-0.940717	-1.131148	1.386753	0.953212	-0.916688	0.649203	-1.025834	0.915393	-0.579467	-0.576388	-1
6669	-1.397337	1.625953	1.386753	1.057449	-0.916688	-1.540351	-1.025834	-1.059600	-0.579467	-0.576388	0

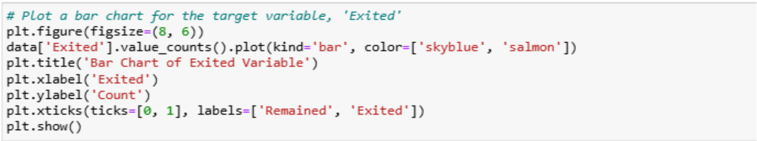
5. Correlation Matrix



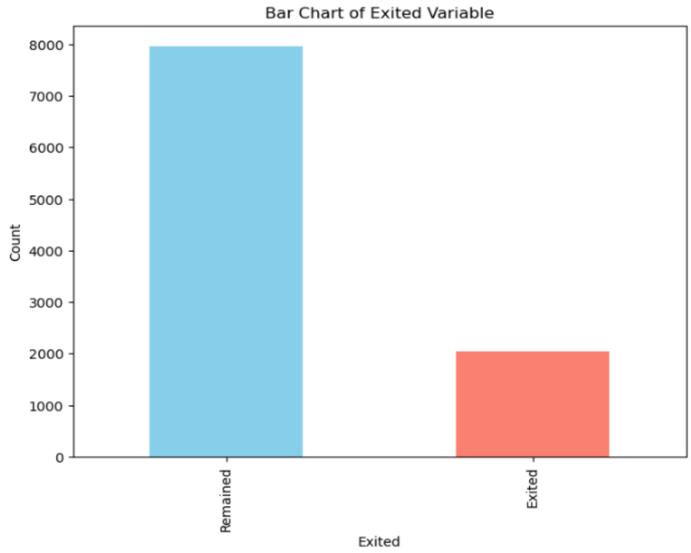
Correlation Matrix

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
CreditScore	1	-0.004	0.00084	0.0063	0.012	-0.0055	0.026	-0.0014
Age	-0.004	1	-0.01	0.028	-0.031	-0.012	0.085	-0.0072
Tenure	0.00084	-0.01	1	-0.012	0.013	0.023	-0.028	0.0078
Balance	0.0063	0.028	-0.012	1	-0.3	-0.015	-0.01	0.013
NumOfProducts	0.012	-0.031	0.013	-0.3	1	0.0032	0.0096	0.014
HasCrCard	-0.0055	-0.012	0.023	-0.015	0.0032	1	-0.012	-0.0099
IsActiveMember	0.026	0.085	-0.028	-0.01	0.0096	-0.012	1	-0.011
EstimatedSalary	-0.0014	-0.0072	0.0078	0.013	0.014	-0.0099	-0.011	1

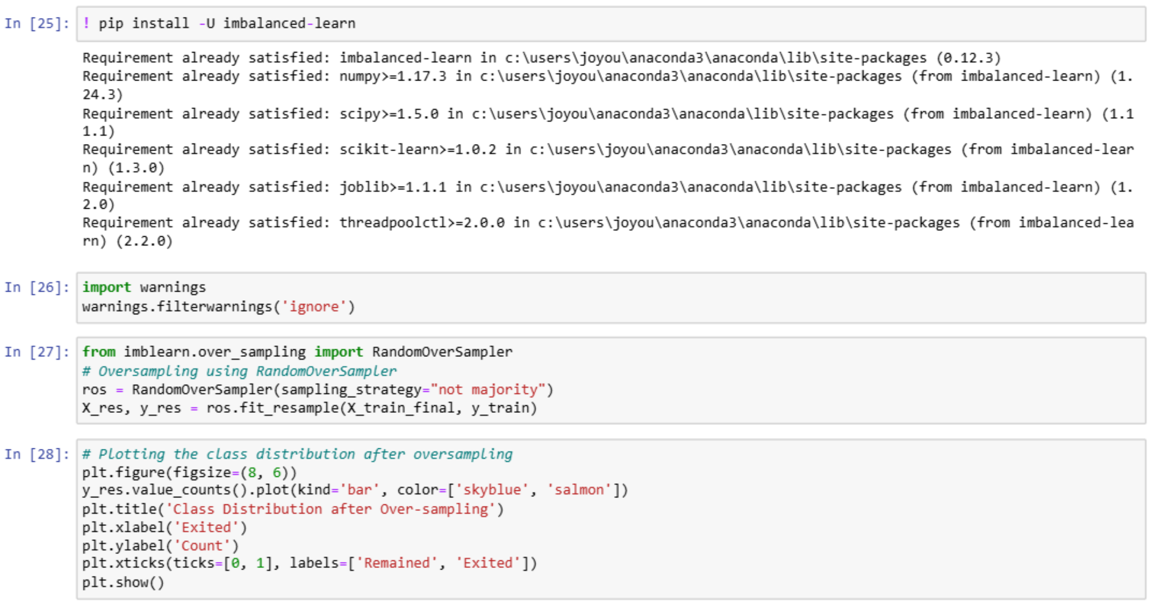
6. Visualising the target variable



```
# Plot a bar chart for the target variable, 'Exited'
plt.figure(figsize=(8, 6))
data['Exited'].value_counts().plot(kind='bar', color=['skyblue', 'salmon'])
plt.title('Bar Chart of Exited Variable')
plt.xlabel('Exited')
plt.ylabel('Count')
plt.xticks(ticks=[0, 1], labels=['Remained', 'Exited'])
plt.show()
```

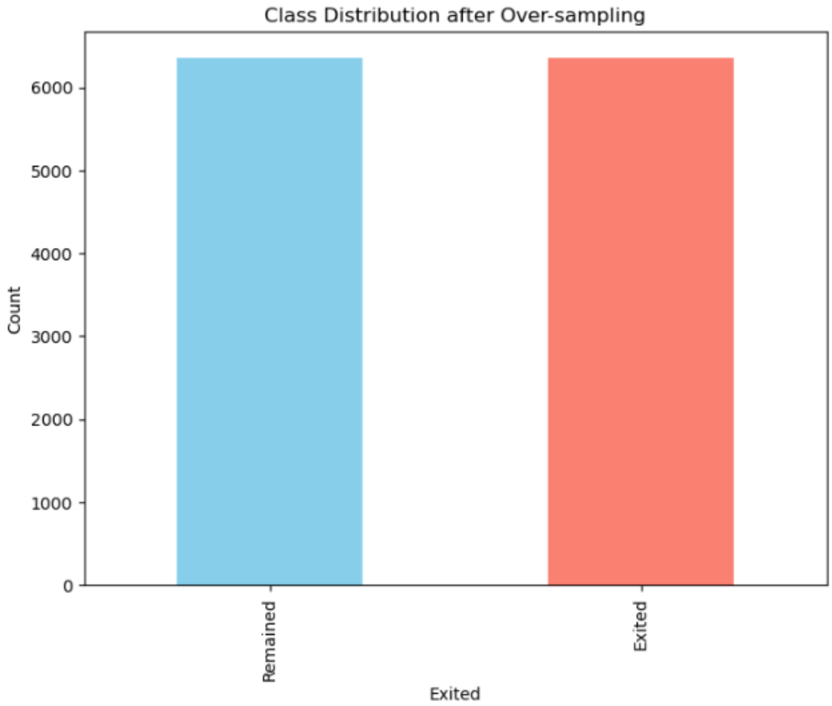


7. Using Oversampling technique to balance dataset

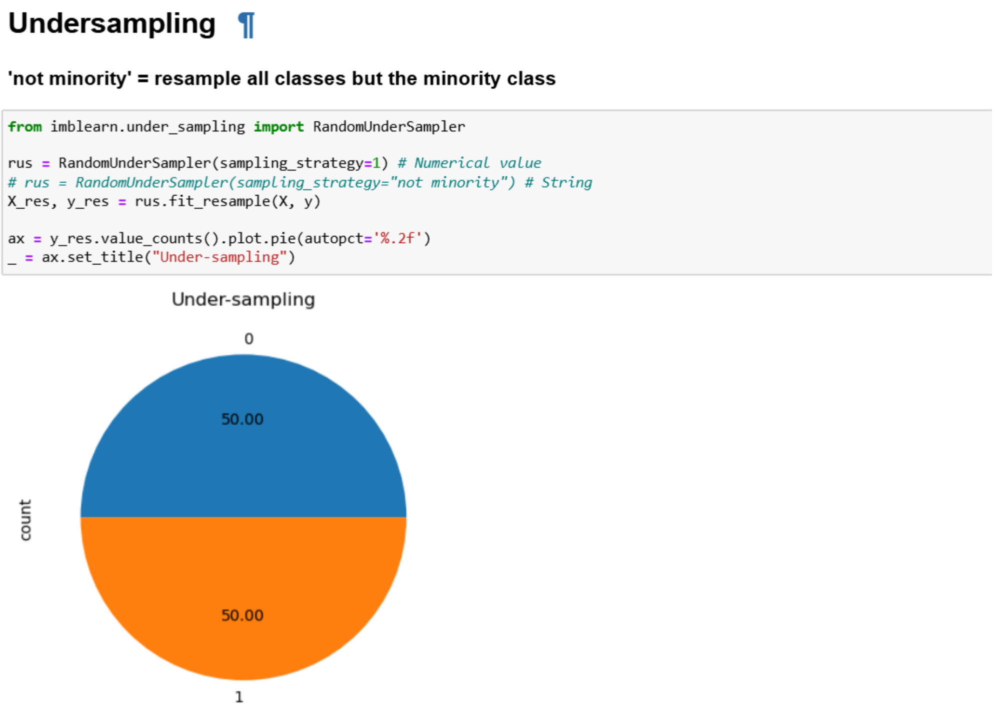


```
In [25]: ! pip install -U imbalanced-learn

         Requirement already satisfied: imbalanced-learn in c:\users\joyou\anaconda3\anaconda\lib\site-packages (0.12.3)
         Requirement already satisfied: numpy>=1.17.3 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-learn) (1.
         24.3)
         Requirement already satisfied: scipy>=1.5.0 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-learn) (1.1
         1.1)
         Requirement already satisfied: scikit-learn>=1.0.2 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-lear
         n) (1.3.0)
         Requirement already satisfied: joblib>=1.1.1 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-learn) (1.
         2.0)
         Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-lea
         rn) (2.2.0)

In [26]: import warnings
         warnings.filterwarnings('ignore')

In [27]: from imblearn.over_sampling import RandomOverSampler
         # Oversampling using RandomOverSampler
         ros = RandomOverSampler(sampling_strategy="not majority")
         X_res, y_res = ros.fit_resample(X_train_final, y_train)

In [28]: # Plotting the class distribution after oversampling
         plt.figure(figsize=(8, 6))
         y_res.value_counts().plot(kind='bar', color=['skyblue', 'salmon'])
         plt.title('Class Distribution after Over-sampling')
         plt.xlabel('Exited')
         plt.ylabel('Count')
         plt.xticks(ticks=[0, 1], labels=['Remained', 'Exited'])
         plt.show()
```

8. Results of oversampling visualised



9. Code and result for balancing using undersampling method

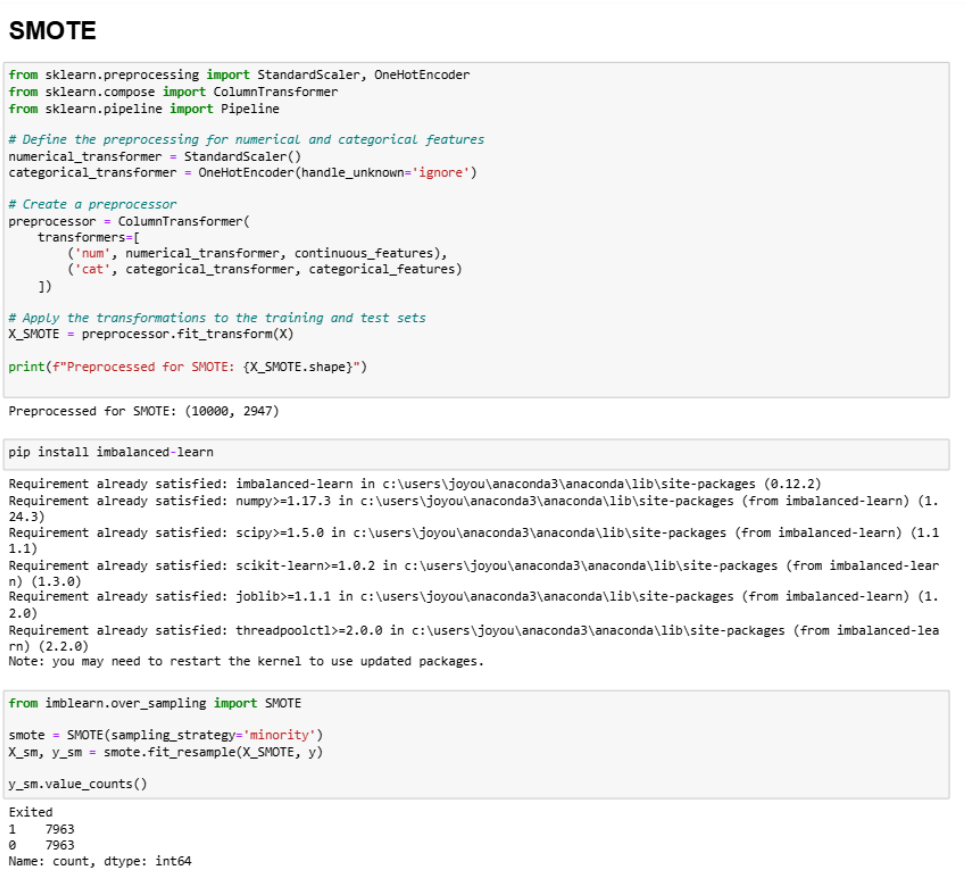


Undersampling ¶

'not minority' = resample all classes but the minority class

```
from imblearn.under_sampling import RandomUnderSampler

rus = RandomUnderSampler(sampling_strategy=1) # Numerical value
# rus = RandomUnderSampler(sampling_strategy="not minority") # String
X_res, y_res = rus.fit_resample(X, y)

ax = y_res.value_counts().plot.pie(autopct='%.2f')
_ = ax.set_title("Under-sampling")
```

10. Code and result for balancing using SMOTE



SMOTE

```
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Define the preprocessing for numerical and categorical features
numerical_transformer = StandardScaler()
categorical_transformer = OneHotEncoder(handle_unknown='ignore')

# Create a preprocessor
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, continuous_features),
        ('cat', categorical_transformer, categorical_features)
    ])

# Apply the transformations to the training and test sets
X_SMOTE = preprocessor.fit_transform(X)

print(f"Preprocessed for SMOTE: {X_SMOTE.shape}")
```

```
Preprocessed for SMOTE: (10000, 2947)
```

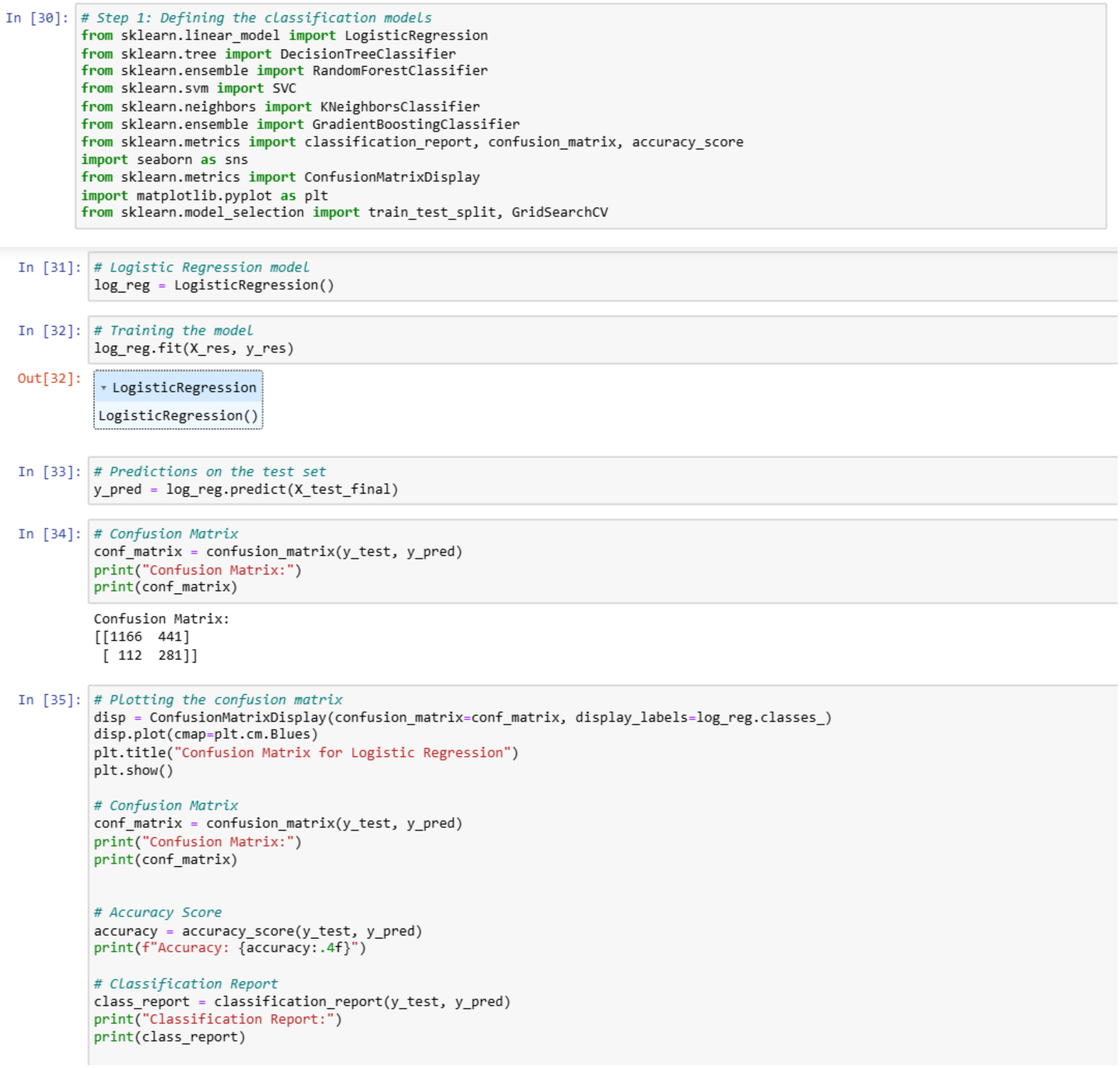
```
pip install imbalanced-learn
```

```
Requirement already satisfied: imbalanced-learn in c:\users\joyou\anaconda3\anaconda\lib\site-packages (0.12.2)
Requirement already satisfied: numpy>=1.17.3 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-learn) (1.
24.3)
Requirement already satisfied: scipy>=1.5.0 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-learn) (1.1
1.1)
Requirement already satisfied: scikit-learn>=1.0.2 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-lear
n) (1.3.0)
Requirement already satisfied: joblib>=1.1.1 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-learn) (1.
2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\joyou\anaconda3\anaconda\lib\site-packages (from imbalanced-lea
rn) (2.2.0)
Note: you may need to restart the kernel to use updated packages.
```

```
from imblearn.over_sampling import SMOTE

smote = SMOTE(sampling_strategy='minority')
X_sm, y_sm = smote.fit_resample(X_SMOTE, y)

y_sm.value_counts()
```

```
Exited
1    7963
0    7963
Name: count, dtype: int64
```

11. Codes showing how the models were defined, trained and how confusion and evaluation metrics were plotted. This snapshot only shows for Logistic regression, however, same steps were followed for all other models.



```
In [30]: # Step 1: Defining the classification models
         from sklearn.linear_model import LogisticRegression
         from sklearn.tree import DecisionTreeClassifier
         from sklearn.ensemble import RandomForestClassifier
         from sklearn.svm import SVC
         from sklearn.neighbors import KNeighborsClassifier
         from sklearn.ensemble import GradientBoostingClassifier
         from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
         import seaborn as sns
         from sklearn.metrics import ConfusionMatrixDisplay
         import matplotlib.pyplot as plt
         from sklearn.model_selection import train_test_split, GridSearchCV
```

```
In [31]: # Logistic Regression model
         log_reg = LogisticRegression()
```

```
In [32]: # Training the model
         log_reg.fit(X_res, y_res)
```

```
Out[32]: ▾ LogisticRegression
         LogisticRegression()
```

```
In [33]: # Predictions on the test set
         y_pred = log_reg.predict(X_test_final)
```

```
In [34]: # Confusion Matrix
         conf_matrix = confusion_matrix(y_test, y_pred)
         print("Confusion Matrix:")
         print(conf_matrix)
```

```
Confusion Matrix:
[[1166  441]
 [ 112  281]]
```

```
In [35]: # Plotting the confusion matrix
         disp = ConfusionMatrixDisplay(confusion_matrix=conf_matrix, display_labels=log_reg.classes_)
         disp.plot(cmap=plt.cm.Blues)
         plt.title("Confusion Matrix for Logistic Regression")
         plt.show()

         # Confusion Matrix
         conf_matrix = confusion_matrix(y_test, y_pred)
         print("Confusion Matrix:")
         print(conf_matrix)


         # Accuracy Score
         accuracy = accuracy_score(y_test, y_pred)
         print(f"Accuracy: {accuracy:.4f}")

         # Classification Report
         class_report = classification_report(y_test, y_pred)
         print("Classification Report:")
         print(class_report)
```