University Of Hertfordshire

School of Physics, Engineering and Computer Science

Advanced Computer Science Masters Project

7COM1039-0901-2025

Title: Customer Churn Prediction using Hybrid Ensemble and Neural Network Models with Class Imbalance Handling

Name: Srinivasa Reddy Pulyala

Student ID: 23096195

Supervisor: Szilvia Csaki Istvanne Biro

Proof-Reading and Quality Assurance Declaration

I confirm that I have critically proof-read and quality-checked this report. I have ensured that it is free from grammatical, spelling, and formatting errors and that it meets a high standard of clarity, coherence, and presentation.

Declaration

This report is submitted in partial fulfilment of the requirements for the degree of Master of Science in Advanced Computer Science at the University of Hertfordshire (UH).

I hereby declare that the work presented in this project and report is entirely my own, except where explicitly stated otherwise. All sources of information and ideas, whether quoted directly or paraphrased, have been properly acknowledged and referenced in accordance with academic standards. I understand that any failure to properly acknowledge the work of others may constitute plagiarism and could result in academic penalties.

I confirm that no human participants were involved in this MSc project.

I hereby give permission for this report to be made available on the University of Hertfordshire website, provided that the source is appropriately acknowledged.

Acknowledgements

I would like to express my sincere gratitude to my academic supervisor for her valuable guidance, constructive feedback, and continuous support throughout the completion of this project. Her expertise and insights significantly contributed to the quality and direction of this work.

I am also grateful to the wider data science community for making datasets, tutorials, and learning resources openly available, which supported my understanding of key concepts. In addition, I would like to acknowledge the developers of the Python ecosystem, whose tools and libraries provided a strong foundation for implementing and evaluating the machine-learning models used in this project.

Finally, I would like to thank my peers, friends, and family for their encouragement, discussions, and ongoing support throughout this journey.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Full Form**
SMOTE	Synthetic Minority Over-sampling Technique
LR	Logistic Regression
RF	Random Forest
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
ANN	Artificial Neural Network
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
EDA	Exploratory Data Analysis
ROC–AUC	Receiver Operating Characteristic – Area Under the Curve

Table 1 List of Abbreviations

Abstract

Customer churn represents a significant challenge for banks as the loss of existing customers reduces revenue and increases the cost of customer acquisition. Accurately predicting customer churn enables banks to implement targeted retention strategies and improve long-term profitability this project focuses on the design implementation and evaluation of a customer churn prediction framework using machine-learning and deep-learning techniques.

The study is based on the Kaggle Bank Customer Churn Modelling dataset which contains demographic, financial and behavioural information for 10,000 anonymised banking customers. A structured modelling workflow was followed including data preprocessing, exploratory data analysis and class imbalance handling using the Technique SMOTE. Five predictive models were implemented and evaluated: Logistic Regression, Random Forest, XGBoost, LightGBM and an Artificial Neural Network (ANN). Model performance was assessed using accuracy, precision, recall, F1-score, ROC–AUC and confusion matrices.

The results demonstrate that class imbalance significantly reduces the ability of models to correctly identify churned customers when left unaddressed. After applying SMOTE, improvements in recall and F1-score were observed across all models indicating more effective detection of the minority churn class. Among the evaluated approaches XGBoost achieved the strongest overall performance recording an accuracy of 0.8555 precision of 0.6569, recall of 0.6069, F1-score of 0.6309 and the highest ROC–AUC of 0.8696. While LightGBM and the ANN also demonstrated competitive results XGBoost consistently provided the most balanced and reliable performance across evaluation metrics.

To support interpretability, SHAP analysis was applied to examine the factors influencing churn predictions. Customer age, account balance, geography and the number of products held were identified as the most influential features. Overall, the findings indicate that combining class balancing techniques advanced ensemble models and explainable machine-learning methods leads to more accurate and practically useful churn prediction supporting informed decision-making in the banking sector.

2. Introduction

Customer churn has become a major concern for the banking sector as the loss of existing customers directly impacts revenue and increases the cost of acquiring new ones. Retaining current customers is widely recognised as more cost-effective than attracting new customers making churn prediction an important strategic task for financial institutions (Reichheld and Sasser, 1990) with the rapid growth of digital banking services customers now expect faster, more reliable, and more personalised experiences as a result, understanding customer behaviour and identifying those who are likely to leave has become increasingly important for banks operating in highly competitive markets (Gupta and Zeithaml, 2006).

In recent years, machine-learning techniques have been increasingly applied to customer churn prediction due to their ability to model complex patterns in customer data more effectively than traditional statistical approaches. Machine-learning models can capture non-linear relationships between demographic, financial, and behavioural features, leading to improved predictive performance (Verbeke et al., 2012). As banks collect large volumes of customer data, these data-driven approaches offer strong potential for supporting proactive customer retention strategies.

Despite these advances, several challenges limit the practical effectiveness of churn prediction systems. One of the most significant issues is class imbalance where the number of customers who churn is much smaller than the number of customers who remain. This imbalance often causes predictive models to favour the majority class, resulting in poor detection of high-risk churn customers (He and Garcia, 2009). In addition, customer behaviour in banking environments is complex and influenced by multiple interacting factors, making it difficult for traditional models to accurately capture churn patterns (Neslin et al., 2006).

Another important challenge is model interpretability. Many advanced machine-learning models, including ensemble methods and neural networks, operate as "black boxes", making it difficult for banks to understand how predictions are generated or to justify automated decisions to regulators and stakeholders (Ribeiro, Singh and Guestrin, 2016). In regulated financial environments, transparency, fairness, and accountability are essential and models that lack interpretability are often difficult to deploy in practice. Furthermore, many existing studies focus heavily on overall accuracy while giving less attention to evaluation metrics that are more suitable for imbalanced datasets, such as recall and F1-score, which are critical for reliably identifying churned customers.

This project aims to address these limitations by developing a structured machine-learning framework for customer churn prediction that balances predictive performance with interpretability and practical relevance. Multiple machine-learning and deep-learning models are compared within a single framework to evaluate their strengths and limitations under consistent experimental conditions. Class imbalance is addressed using SMOTE, which has been shown to improve minority-class detection by generating synthetic churn samples (Chawla et al., 2002). To improve transparency and trust in model predictions, explainability techniques based on SHAP are applied to identify the key features influencing churn decisions (Lundberg and Lee, 2017).

Beyond technical performance, this project also considers ethical, commercial and economic factors that influence the real-world adoption of churn prediction systems. Predictive models must be designed and evaluated carefully to avoid biased decision-making and to ensure that automated systems support fair and responsible use of customer data (O'Neil, 2016). By integrating robust modelling techniques, class balancing and explainable artificial intelligence this project seeks to provide a practical and trustworthy approach to churn prediction that aligns with both business objectives and regulatory expectations.

The remainder of this report is structured as follows. The Literature Review examines existing research on churn prediction, highlighting key methods and research gaps. The Methodology chapter describes the dataset preprocessing steps, model selection, class balancing approach and evaluation metrics. The Quality and Results chapter presents and analyses the experimental findings. Finally, the Evaluation and Conclusion chapter discusses the project's contributions, limitations, and potential directions for future work.

2.1 Problem Overview

In the modern banking landscape, customer churn has become one of the most persistent and costly challenges, as switching between providers has become easier than ever (Gupta and Zeithaml, 2006). Banks operate in highly competitive markets where customers are influenced by service quality, digital experience and personalised engagement when a customer leaves the institution not only loses future revenue but must also incur higher acquisition costs making customer retention a more economical strategy than acquisition (Reichheld and Sasser, 1990). The rapid growth of digital banking has further intensified this challenge with customers increasingly expecting faster, smoother, and more reliable services (Arner, Barberis and Buckley, 2017). As a result, identifying customers who are likely to churn and understanding the factors driving their decisions is essential for maintaining financial stability and long-term customer relationships.

Beyond commercial impact churn prediction also has ethical and economic implications. Predictive systems must be designed with fairness and transparency to avoid reinforcing bias or discrimination in automated decision-making (O'Neil, 2016). In addition, banks must consider the cost of misclassification as incorrect predictions can lead to unnecessary retention efforts or missed intervention opportunities (Verbeke et al., 2012). These considerations highlight that churn prediction is not merely a technical modelling task, but a broader strategic challenge involving responsible data use and informed decision-making.

2.2 Current Issues

Although customer churn prediction has been widely researched, several challenges continue to limit the effectiveness of existing approaches. One of the most persistent problems is class imbalance where the number of customers who churn is far smaller than those who remain. This imbalance makes it difficult for models to learn meaningful minority-class patterns and often leads to biased predictions that overlook high-risk customers (He and Garcia, 2009). In addition, traditional analytical methods and simple rule-based approaches struggle to capture the complex non-linear behaviour that characterises modern banking customers particularly in digital environments (Neslin et al., 2006).

Another major challenge is the lack of interpretability in many machine-learning models. Complex algorithms such as ensemble methods and neural networks often operate as “black boxes” making it difficult for financial institutions to understand or justify automated decisions to regulators and stakeholders (Ribeiro, Singh and Guestrin, 2016). Furthermore, many existing studies place excessive emphasis on overall accuracy while giving limited attention to commercial relevance, operational practicality and fairness in decision-making. These limitations highlight the need for churn prediction systems that balance predictive performance with transparency, ethical responsibility and business value the potential solutions to these challenges are examined in later chapters of this report.

2.3 Project Details

This project focuses on developing a machine-learning framework for predicting customer churn in the banking sector. The primary objective is to identify customers who are at risk of leaving by analysing patterns in historical customer data. To achieve this multiple machine-learning models are compared to assess their predictive performance and practical suitability for real-world banking applications.

The project incorporates key stages including data preprocessing, exploratory data analysis class balancing using SMOTE, model training and result interpretation using SHAP. These components are designed to address key challenges in churn prediction particularly class imbalance and the need for transparent and explainable decision-making.

The work is conducted in a structured and systematic manner, progressing from data preparation to model development, evaluation, and interpretation. Detailed descriptions of the modelling approach experimental procedures, evaluation strategy, and practical considerations are presented in the Methodology and Quality and Results chapters.

2.4 Aims and Objectives

Aim

The aim of this project is to predict customer churn in the banking sector using machine-learning and deep-learning models, compare their performance using appropriate evaluation metrics, and identify the key factors influencing churn through explainable artificial intelligence techniques.

Objectives

- To prepare the customer churn dataset by carrying out data cleaning, categorical feature encoding, feature scaling, and the removal of non-relevant attributes.
- To conduct Exploratory Data Analysis (EDA) to analyse feature distributions, relationships between variables, and patterns related to churn behaviour.
- To design and evaluate both baseline and advanced predictive models, including Logistic Regression, Random Forest, XGBoost, LightGBM, and an Artificial Neural Network.
- To mitigate the effects of class imbalance by applying SMOTE and to evaluate its influence on churn prediction performance.
- To assess and compare model performance using multiple evaluation metrics, including accuracy, precision, recall, F1-score, confusion matrices, and ROC–AUC.

- To employ SHAP-based explainability methods to identify and interpret the key features that contribute to customer churn.
- To identify the most appropriate predictive model by considering performance, interpretability, and practical suitability for real-world use.
- To generate actionable insights and recommendations that can support effective churn-reduction strategies within the banking sector.

2.5 Research Question and Novelty

2.5.1 Research Questions:

1. Which demographic, financial, and behavioural features most influence customer churn?
2. Which of the selected models achieves the most effective balance between accuracy and recall in predicting customer churn?
3. How does class imbalance affect model performance, and to what extent does SMOTE improve minority-class detection?

2.5.2 Novelty and Research Gap

Although customer churn prediction has been widely studied, many existing works focus on a single model or prioritise overall accuracy without adequately addressing class imbalance or model interpretability. For example, Verbeke et al. (2012) demonstrated the effectiveness of ensemble methods for churn prediction but placed limited emphasis on explainability. Similarly, Neslin et al. (2006) explored churn drivers using traditional analytical techniques, which struggle to capture complex non-linear customer behaviour.

More recent studies have shown the effectiveness of advanced machine-learning models such as XGBoost for churn prediction however, these studies often lack a systematic comparison across multiple model families under consistent experimental conditions (Chen and Guestrin, 2016; Kumar and Ravi, 2019). In addition, explainability techniques such as SHAP are still rarely integrated into churn prediction pipelines using publicly available banking datasets.

The novelty of this project lies in its unified comparative framework, where classical, ensemble, and deep-learning models are evaluated both before and after the application of SMOTE. By combining performance evaluation with SHAP-based explainability and a strong focus on recall, F1-score, and ROC–AUC, this study addresses an important research gap by delivering a more balanced, interpretable and business-oriented churn prediction analysis.

2.5.3 Feasibility, Commercial Context, and Risk

The project is highly feasible due to the use of a publicly available, anonymised dataset and widely adopted open-source machine-learning libraries, ensuring manageable technical complexity and reproducibility. From a commercial perspective churn prediction systems offer substantial economic value as even small improvements in customer retention can significantly reduce operational costs and increase customer lifetime value (Reichheld and Sasser, 1990; Gupta and Zeithaml, 2006).

However, commercial adoption also involves several risks. Predictive models may introduce bias particularly when trained on imbalanced datasets potentially leading to unfair or discriminatory outcomes (O’Neil, 2016). Organisations may also face challenges related to over-reliance on automation integration costs and compliance with regulatory requirements for transparency and accountability in financial decision-making (Ribeiro, Singh and Guestrin, 2016). These technical, ethical, and economic risks are examined in greater detail in the Evaluation and Conclusion chapter

2.6 Report Structure

This report is organised into several chapters that reflect the key stages of the project. The Introduction presents the project context, objectives and research questions. The Literature Review reviews existing research and identifies the research gaps addressed in this study. The Methodology chapter describes the dataset modelling approach and evaluation strategy. The Quality and Results chapter presents and analyses the experimental findings while the Evaluation and Conclusion chapter summarises the outcomes, limitations, and future work. Supporting materials are provided in the Appendices.

3. Literature Review

3.1 Introduction

Academic research on customer churn prediction has evolved significantly as organisations increasingly rely on data-driven methods to support retention strategies in competitive industries such as banking and telecommunications early studies in this domain predominantly applied traditional statistical techniques including logistic regression to estimate the likelihood of customer defection. While these methods offered transparency and ease of interpretation their capacity to represent complex customer behaviour was limited particularly when applied to large and high-dimensional datasets (Neslin et al., 2006).

As data availability and computational capabilities expanded research attention shifted towards machine-learning approaches that are better suited to modelling non-linear relationships within customer data. Prior studies demonstrate that ensemble and tree-based models can capture interactions between demographic, financial, and behavioural variables more effectively than conventional techniques leading to measurable improvements in churn prediction performance (Verbeke et al., 2012). Consequently, recent literature increasingly explores the comparative performance of classical, ensemble, and neural network models under varying experimental conditions.

A key challenge consistently highlighted in churn prediction research is the issue of class imbalance where churn events constitute a small proportion of the overall customer population. This imbalance has been shown to bias learning algorithms towards the majority class resulting in models that achieve high overall accuracy but perform poorly in identifying churned customers (He and Garcia, 2009). To mitigate this limitation researchers have proposed several solutions including resampling strategies and cost-sensitive learning. Among these approaches resampling methods such as SMOTE are widely discussed due to their effectiveness in enhancing minority-class representation without removing existing observations (Chawla et al., 2002).

In addition to addressing imbalance recent studies emphasise the importance of evaluation metrics that better reflect business priorities. Rather than relying solely on accuracy researchers increasingly advocate the use of recall, F1-score, ROC–AUC and confusion matrices to assess churn models as these measures provide clearer insight into the costs associated with misclassification in practical retention scenarios (Verbeke et al., 2012).

More recently model interpretability has emerged as a central concern particularly in regulated environments such as financial services. Although complex models often outperform simpler alternatives their lack of transparency presents challenges for regulatory compliance and stakeholder trust. As a result, explainable artificial intelligence (XAI) techniques have gained prominence as mechanisms for improving model transparency and supporting responsible decision-making (Ribeiro, Singh and Guestrin, 2016).

Overall, the literature reflects substantial progress in churn prediction through the adoption of advanced modelling techniques class-balancing strategies and improved evaluation practices. However, many studies examine individual methods in isolation or prioritise predictive performance without integrating interpretability as a core component. These limitations indicate a need for structured comparative analyses that jointly evaluate multiple model families explicitly address class imbalance and incorporate explainable methods. The following sections review key studies in greater detail and identify the research gaps that inform the methodological approach adopted in this project.

3.2 Key Studies and Works

Research on customer churn prediction reflects a clear methodological shift from traditional statistical techniques to more advanced machine-learning approaches driven by the need to model complex customer behaviour and improve predictive performance. Early studies frequently employed Logistic Regression due to its interpretability and ease of implementation. However, empirical investigations by Hadden et al. (2007) and Idris, Khan and Lee (2012) demonstrate that Logistic Regression often performs poorly when faced with non-linear relationships and imbalanced churn datasets limiting its effectiveness in identifying high-risk customers in competitive and data-intensive environments.

To address these limitations later research increasingly adopted ensemble learning methods. Breiman (2001) showed that Random Forest improves predictive stability by aggregating multiple decision trees thereby reducing variance compared to single-model approaches. Similarly gradient boosting techniques such as XGBoost and LightGBM have been widely reported to outperform classical models by effectively capturing complex feature interactions and non-linear patterns (Chen and Guestrin, 2016; Ke et al., 2017). While these studies report strong performance gains many focus on individual ensemble models in isolation providing limited comparative insight across different model families. This gap motivates the comparative framework adopted in this project which evaluates classical, ensemble, and neural network models under consistent experimental conditions.

Neural networks represent another significant direction in churn prediction research. Studies such as Zhang, Zhao and LeCun (2021) indicate that neural architectures are capable of modelling highly complex behavioural patterns and can achieve competitive predictive performance when sufficient data and appropriate tuning are available. However, compared to tree-based ensemble methods neural networks typically require greater computational resources and offer lower interpretability. Including an Artificial Neural Network in this project enables a balanced comparison between predictive performance, interpretability and practical feasibility across different modelling paradigms.

Class imbalance remains a persistent challenge across churn prediction studies as churned customers typically represent a small minority of the overall dataset. Weiss (2004) and Ling and Li (1998) demonstrate that imbalanced data can produce misleading accuracy scores with models favouring the majority non-churn class while failing to identify churners effectively. To mitigate this issue researchers have proposed resampling and cost-sensitive learning techniques. Among these SMOTE introduced by Chawla et al. (2002) is widely adopted due to its ability to enhance minority-class representation without discarding existing data. Despite its popularity many studies apply SMOTE without systematically evaluating model performance before and after balancing. This project directly addresses this limitation by explicitly analysing the impact of SMOTE across all models.

Beyond predictive performance, model interpretability has become increasingly important particularly in regulated domains such as banking. Although ensemble models and neural networks often achieve superior accuracy their black-box nature limits transparency and stakeholder trust. Ribeiro, Singh and Guestrin (2016) argue that the absence of explainability can hinder real-world adoption, even for highly accurate models. More recently SHAP has emerged as a unified framework for explaining complex predictions (Lundberg and Lee, 2017) with studies such as Ou (2023) demonstrating its practical value in identifying churn-related features in financial datasets. However, interpretability is still frequently treated as a secondary consideration rather than an integrated component of model evaluation. This project addresses this gap by embedding SHAP-based explainability directly within the comparative modelling framework.

Overall existing literature provides strong evidence that advanced machine-learning model, class-balancing techniques, and alternative evaluation metrics can enhance churn prediction performance. Nevertheless, limitations persist due to fragmented model comparisons, inconsistent handling of class imbalance, and insufficient integration of explainability. By systematically comparing multiple modelling approaches explicitly evaluating the effects of SMOTE and incorporating SHAP-based interpretation this project builds on prior research in a structured and practically relevant manner directly informing the research questions and methodological choices adopted in this study.

3.3 Identification of Research Gaps

Despite extensive research on customer churn prediction in banking and related service sectors several important limitations persist in the existing literature. A key gap is the frequent over-reliance on overall accuracy as the primary evaluation metric. Many studies report strong accuracy results without adequately considering metrics better suited to imbalanced datasets, such as recall, F1-score, and ROC–AUC, which can lead to misleading conclusions and poor identification of high-risk churn customers (Verbeke et al., 2012; He and Garcia, 2009).

A second limitation concerns the narrow scope of model comparison. Existing studies often focus on a single algorithm or a small subset of models, making it difficult to assess how classical machine-learning methods, ensemble techniques, and deep-learning approaches perform relative to one another under consistent experimental conditions (Hadden et al., 2007; Chen and Guestrin, 2016). This limits the ability of organisations to select models that best align with operational and business requirements.

Model interpretability represents another significant research gap. Many churn prediction studies prioritise predictive performance while treating interpretability as a secondary concern. Only a limited number of works integrate explainable artificial intelligence techniques to provide transparent feature-level insights despite their importance in regulated environments such as banking (Ribeiro, Singh and Guestrin, 2016; Lundberg and Lee, 2017).

Finally, class imbalance remains an inadequately addressed challenge. Although several studies acknowledge the disproportionate representation of churned customers many either overlook the issue or apply balancing techniques without systematic evaluation leading to biased predictions that underestimate churn risk (Weiss, 2004; Chawla et al., 2002).

This project addresses these gaps through a structured and comparative framework that evaluates classical, ensemble, and deep-learning models using consistent data and evaluation metrics appropriate for imbalanced classification. Class imbalance is explicitly handled using SMOTE and SHAP-based explainability is integrated to enhance transparency and practical relevance directly responding to limitations identified in prior research.

3.4 Relation to the Current Study

The literature reviewed in this chapter directly informs the development of this study's aims objectives and hypothesis prior research shows that customer churn is influenced by a mixture of demographic financial and behavioural characteristics, yet many existing studies do not examine these factors in sufficient depth or use modern interpretability techniques. This limitation supports the focus of the first research question which aims to explore and understand which features have the strongest influence on customer churn.

The reviewed literature also highlights several methodological limitations in previous work many studies rely heavily on accuracy as the main evaluation metric even though churn datasets are typically imbalanced and accuracy alone can hide poor detection of the minority class. In addition classical machine learning models ensemble algorithms and deep learning methods are rarely compared within the same framework making it difficult to determine which approach performs best under consistent conditions these gaps directly relate to the second research question which examines which of the selected models provides the best balance between accuracy and recall placing emphasis on metrics that more accurately reflect minority-class performance.

Another recurring issue in the literature is the challenge of class imbalance. Many models tend to favour the majority class, and several studies fail to apply appropriate balancing techniques this observation leads to the third research question which investigates how class imbalance affects model performance and whether the use of SMOTE improves the detection of churn cases. By examining model behaviour before and after the application of SMOTE the project aims to provide clearer evidence of how balancing techniques influence predictive outcomes.

Overall, the literature review identifies important gaps in feature interpretability, evaluation metrics, model comparison and imbalance handling. These gaps directly shape the research questions justify the chosen methodology and support the overall aim of developing a robust interpretable and practically meaningful approach to churn prediction within the banking sector.

4. Methodology

4.1 Methodology Introduction

This chapter outlines how the project was designed, implemented, tested and validated to address the research aims and questions. It outlines the data-science methodology adopted for customer churn prediction and justifies the selection of preprocessing techniques predictive models, class-balancing methods and evaluation strategies used in the study (CRISP-DM, 2000; Provost and Fawcett, 2013).

The methodology covers data preparation model training and testing class imbalance handling using SMOTE and the integration of interpretability techniques to support transparent decision-making. Emphasis is placed on ensuring the reliability and validity of results through a consistent experimental design and the use of evaluation metrics appropriate for imbalanced classification problems (He and Garcia, 2009; Verbeke et al., 2012).

Ethical, practical and professional considerations related to the use of customer data and predictive modelling in the banking domain are also addressed, including issues of fairness, transparency and responsible deployment (O'Neil, 2016; Ribeiro, Singh and Guestrin, 2016). By clearly documenting methodological choices and validation procedures this chapter provides a reproducible and well-justified foundation for the results and analysis presented in subsequent chapters.

Figure 1 presents an overview of the methodology workflow used in this project. The workflow begins with importing the bank customer dataset, followed by exploratory data analysis and data preprocessing. Class imbalance is addressed using SMOTE before the data is split into training and testing sets. Multiple machine-learning models are then trained and evaluated, and SHAP is applied to explain model predictions and identify the key factors influencing customer churn.

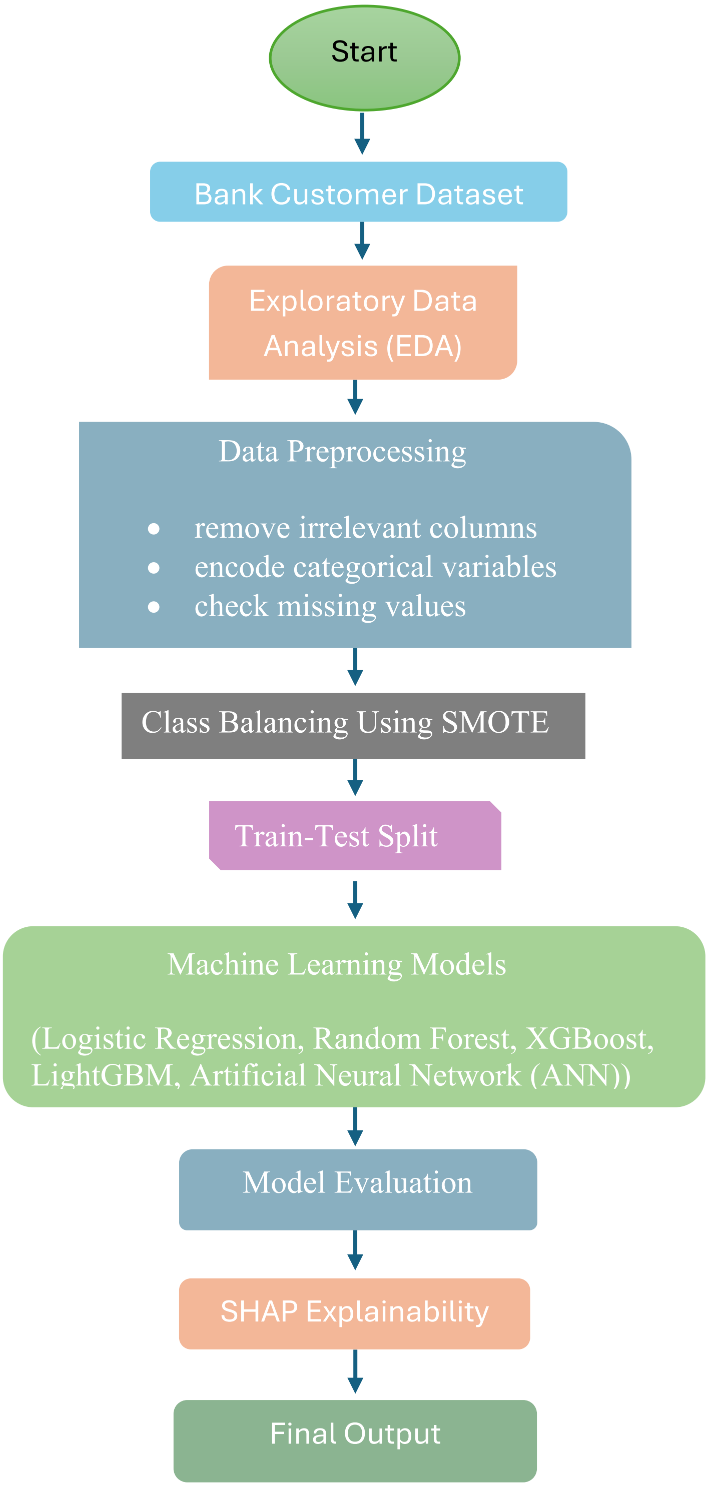

Start
Bank Customer Dataset
Exploratory Data Analysis (EDA)
Data Preprocessing
remove irrelevant columns
encode categorical variables
check missing values
Class Balancing Using SMOTE
Train-Test Split
Machine Learning Models
(Logistic Regression, Random Forest, XGBoost, LightGBM, Artificial Neural Network (ANN))
Model Evaluation
SHAP Explainability
Final Output

Figure 1 Methodology Workflow

4.2 Choice of Methods

This project adopts a data-science workflow to support the prediction of customer churn and the identification of influential factors driving churn behaviour. The selected methods are designed to provide a structured progression from data preparation and exploration to predictive modelling and interpretation ensuring alignment with the research aims and questions (CRISP-DM, 2000; Provost and Fawcett, 2013).

The methodological process starts with data preparation which involves removing non-informative attributes, transforming categorical features into numerical form and applying feature scaling to support effective model training. Exploratory Data Analysis (EDA) is subsequently performed to analyse feature distributions relationships between variables and the severity of class imbalance. The dataset is then divided into training and testing subsets using a stratified split to maintain class proportions and support reliable performance evaluation (Kuhn and Johnson, 2013).

Given the imbalanced nature of churn data, SMOTE is employed to generate synthetic minority-class samples and improve the learning of churn patterns (Chawla et al., 2002; He and Garcia, 2009). A diverse set of predictive models is selected to enable meaningful comparison across different modelling paradigms.

Logistic Regression is used as a baseline model while Random Forest, XGBoost, and LightGBM are applied to capture complex non-linear relationships. An Artificial Neural Network is also included to assess the trade-off between predictive performance, interpretability, and computational complexity (Breiman, 2001; Chen and Guestrin, 2016; Ke et al., 2017).

Model evaluation is performed using accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices, providing a balanced assessment of performance on imbalanced data (Verbeke et al., 2012) finally SHAP-based explainability is applied to interpret model predictions and identify the most influential churn-related features supporting transparency and practical decision-making in a banking context (Lundberg and Lee, 2017).

4.3 Justification and Support of Choices

The methods selected for this project were chosen to align with the characteristics of the dataset and the objectives of customer churn prediction. Data preprocessing and exploratory data analysis were essential for understanding feature distributions identifying class imbalance and ensuring data suitability for modelling (Kuhn and Johnson, 2013).

Given the imbalanced nature of churn datasets, SMOTE was selected to enhance minority-class representation. Prior studies demonstrate that oversampling techniques can improve churn detection without discarding valuable data (Chawla et al., 2002; He and Garcia, 2009). Logistic Regression was employed as a baseline model due to its simplicity and interpretability providing a reference point for evaluating the benefits of more complex approaches (Neslin et al., 2006).

Ensemble methods including Random Forest, XGBoost, and LightGBM, were incorporated because they are well suited to structured tabular data and have consistently demonstrated strong performance in churn prediction particularly in capturing non-linear relationships and feature interactions (Breiman, 2001; Chen and Guestrin, 2016; Ke et al., 2017). An Artificial Neural Network was included to assess whether increased model complexity offers meaningful performance gains compared to classical and ensemble methods (Zhang, Zhao and LeCun, 2021).

Multiple evaluation metrics were used to provide a balanced assessment of model performance. As accuracy alone can be misleading for imbalanced datasets metrics such as recall, F1-score, and ROC–AUC were prioritised to better reflect churn detection effectiveness (Verbeke et al., 2012). Finally, SHAP-based explainability was selected to enhance transparency and interpretability, addressing the practical need for accountable decision-support systems in regulated banking environments (Lundberg and Lee, 2017; Ribeiro, Singh and Guestrin, 2016).

4.4 Project Design

The project is designed as a structured machine-learning system that follows a clear sequence of stages from data acquisition to model interpretation. The dataset used is the publicly available Bank Customer Churn dataset hosted on Kaggle which contains 10,000 customer records with demographic, financial, and behavioural attributes (Kaggle, 2023). As the data is pre-collected and anonymised it is suitable for supervised learning and does not require primary data collection.

The overall project design follows a layered architecture comprising data exploration, preprocessing, class balancing, model development and evaluation. Each stage feeds logically into the next enabling a systematic progression from understanding the dataset to training multiple machine-learning models and interpreting their outputs. This structured design ensures that the workflow remains organised, reproducible, and aligned with the project's research aim of predicting customer churn using data-driven techniques (Provost and Fawcett, 2013).

4.5 Preliminary Exploratory Analysis for Methodological Decisions

A limited set of exploratory visualisations is included in this chapter to support key methodological decisions rather than to provide detailed behavioural insights. These visuals are used to identify class imbalance, examine basic feature relationships, and justify preprocessing choices such as encoding, class balancing, and model selection. More detailed exploratory analysis and behavioural interpretation are presented later in the Quality and Results chapter

4.5.1 Churn Distribution

Before model development, the distribution of the target variable was examined to assess whether the dataset was balanced. Understanding the proportion of churned and non-churned customers is critical, as class imbalance can significantly affect model performance and evaluation.

Figure 2 illustrates the distribution of churned and non-churned customers in the dataset. The data is highly imbalanced, with a substantially larger proportion of non-churn customers. This imbalance directly informed the decision to apply SMOTE during preprocessing to improve minority-class representation during model training.

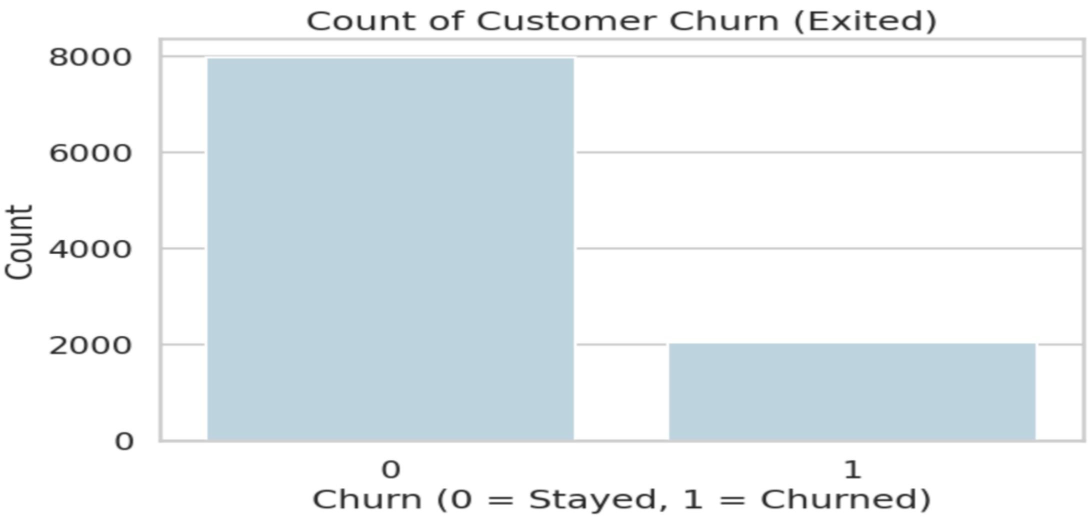


Figure 2 Customer Churn Distribution

4.5.2 Correlation Heatmap

A correlation analysis was conducted to examine linear relationships between numerical and encoded features and the churn variable. This analysis helps determine whether simple linear models are sufficient or whether more complex modelling techniques are required.

Figure 3 presents the correlation heatmap for the dataset. Most features exhibit weak linear correlations with churn, suggesting that churn behaviour is influenced by multiple interacting factors rather than single variables. Age shows a moderate positive association with churn, while balance and geographical indicators also display some relationship. These observations support the use of non-linear machine-learning models and feature-level explanation techniques such as SHAP.

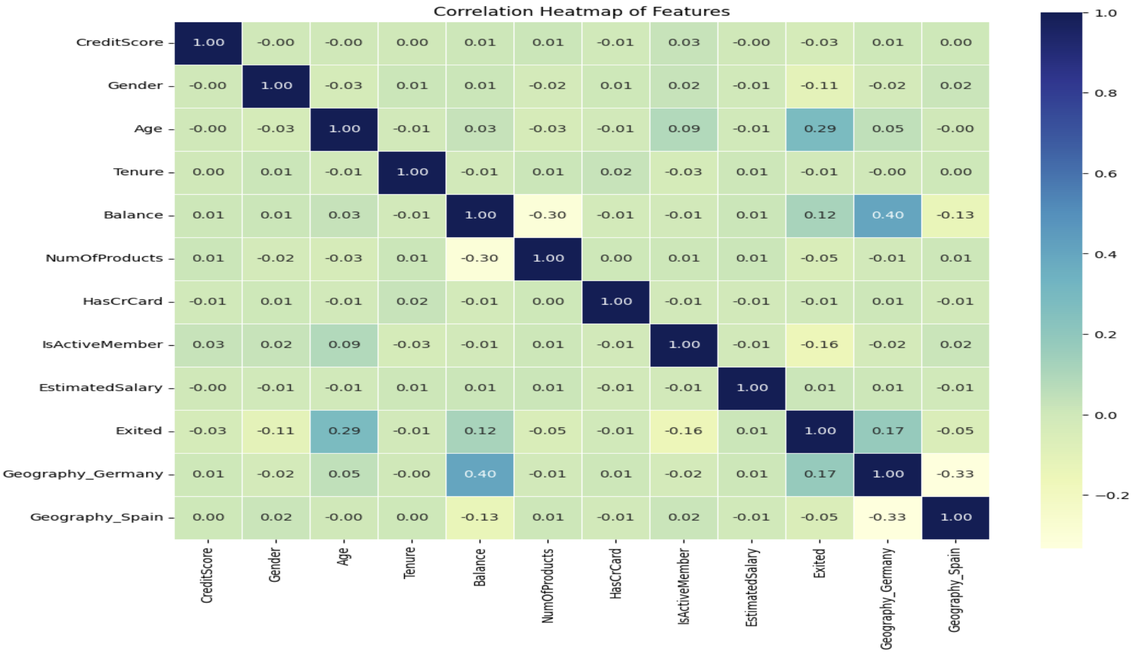


Figure 3 Correlation Heatmap

4.6 System Architecture of the Churn Prediction Framework

The system architecture illustrates how data flows through each stage of the churn prediction framework implemented in this project. The process begins by loading the Kaggle Bank Customer Churn dataset into a Pandas DataFrame which serves as the primary structure for data analysis and model development.

In the data processing layer, the dataset is cleaned and prepared for modelling. This includes removing irrelevant identifier attributes such as RowNumber, CustomerId and Surname which do not contribute to churn prediction. Categorical features including Gender and Geography are encoded into numerical form to enable processing by machine-learning algorithms. The dataset is also verified to ensure that no missing values are present. Basic exploratory checks are conducted at this stage to confirm feature distributions and class imbalance before further processing.

This layered architecture ensures a clear separation between data preparation, model training, evaluation, and interpretation stages, supporting a structured, reproducible and scalable churn prediction system

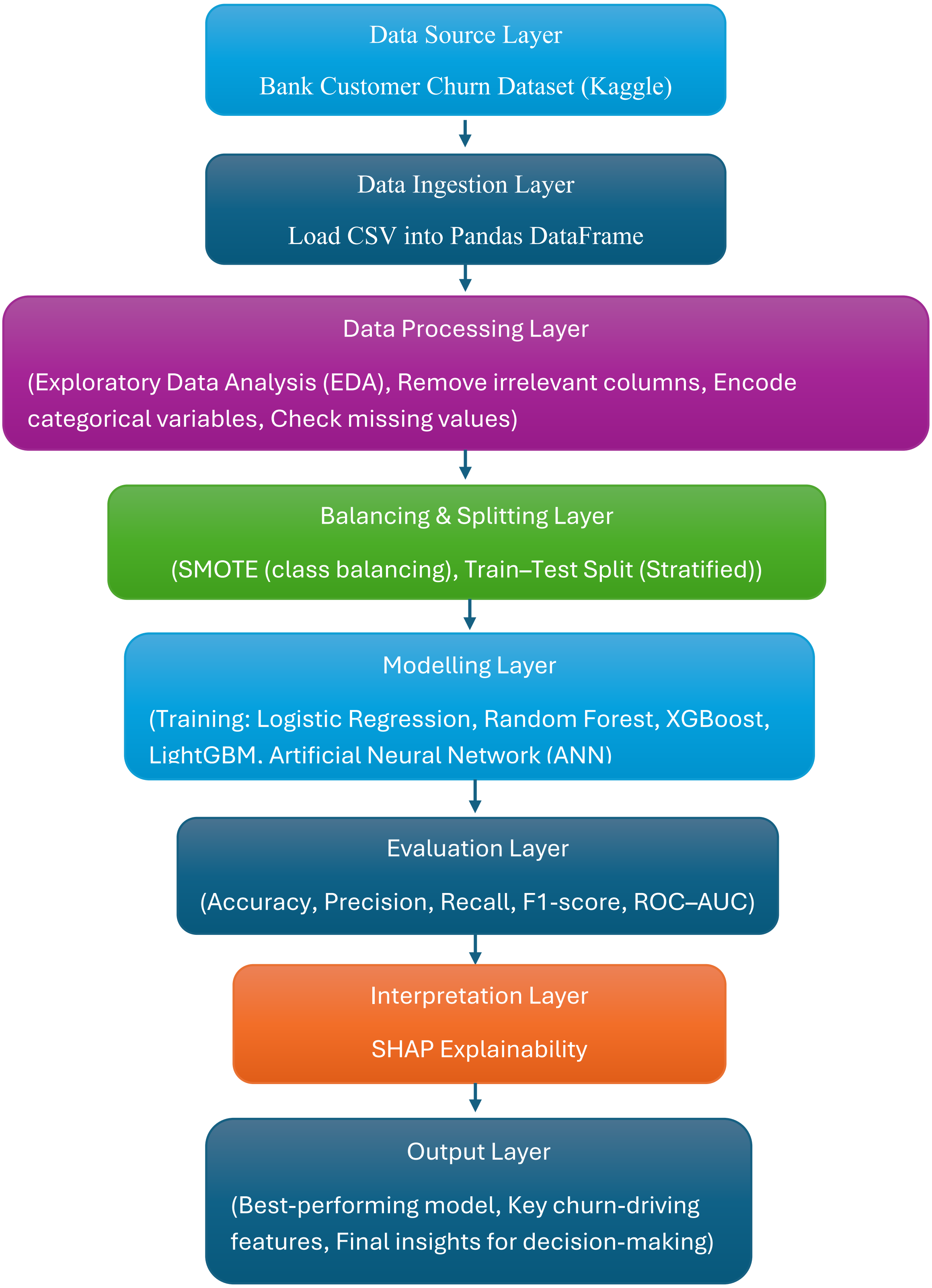

Data Source Layer
Bank Customer Churn Dataset (Kaggle)
Data Ingestion Layer
Load CSV into Pandas DataFrame
Data Processing Layer
(Exploratory Data Analysis (EDA), Remove irrelevant columns, Encode categorical variables, Check missing values)
Balancing & Splitting Layer
(SMOTE (class balancing), Train–Test Split (Stratified))
Modelling Layer
(Training: Logistic Regression, Random Forest, XGBoost, LightGBM, Artificial Neural Network (ANN)
Evaluation Layer
(Accuracy, Precision, Recall, F1-score, ROC–AUC)
Interpretation Layer
SHAP Explainability
Output Layer
(Best-performing model, Key churn-driving features, Final insights for decision-making)

Figure 4 Overall Architecture of the Churn Prediction System

4.7 Use of Tools and Techniques

Python is adopted as the primary programming language for this project because it offers a comprehensive ecosystem for data analysis and machine-learning development (Van Rossum and Drake, 2009). Libraries such as Pandas and NumPy are utilised for efficient data handling and numerical computation while Matplotlib and Seaborn are employed to support data exploration and visual analysis. Machine-learning tasks including preprocessing, feature encoding stratified train–test splitting performance evaluation and the implementation of baseline models such as Logistic Regression and Random Forest are carried out using the Scikit-learn library (Pedregosa et al., 2011).

Class imbalance is mitigated by applying SMOTE from the imbalanced-learn library exclusively to the training dataset, thereby enhancing minority-class representation during model learning. Ensemble-based approaches including XGBoost and LightGBM are employed due to their suitability for structured tabular data and their proven effectiveness in customer churn prediction (Chen and Guestrin, 2016; Ke et al., 2017). In addition, an Artificial Neural Network is implemented using TensorFlow and Keras to enable a comparative assessment of deep-learning methods against traditional and ensemble models. To support model interpretability SHAP is utilised to provide transparent insights into the relative importance of features influencing churn predictions (Lundberg and Lee, 2017).

Following preprocessing, the dataset is divided into training and testing sets using a stratified strategy to maintain class distribution across both subsets. Feature scaling is performed with the StandardScaler which is fitted exclusively on the training data to avoid information leakage. SMOTE is then applied to the training set prior to model development. Model performance is assessed using accuracy, precision, recall, F1-score, and ROC–AUC. SHAP analysis is subsequently employed to determine the most influential features, and the results identify the best-performing model together with key interpretability insights. This end-to-end workflow supports a structured, reliable, and reproducible approach to customer churn prediction. Selected code examples illustrating preprocessing, class balancing, model training, evaluation, and explainability are included in Appendix C.

4.8 Machine Learning Models Used

To evaluate customer churn prediction under consistent experimental conditions this project employs five machine-learning models representing classical, ensemble, and deep-learning approaches. Logistic Regression is used as a baseline model due to its simplicity and interpretability, providing a reference point for assessing the benefits of more advanced techniques (Hosmer et al., 2013). As a linear model it offers insight into feature influence but is limited in capturing complex non-linear relationships.

To address these limitations, ensemble learning methods are included. Random Forest combines multiple decision trees to improve predictive performance and reduce overfitting making it well suited for churn prediction on structured business data while also providing feature importance information (Breiman, 2001). Gradient boosting techniques are further explored using XGBoost and LightGBM. XGBoost incorporates regularisation and efficient optimisation strategies enabling it to model complex feature interactions while controlling overfitting and has demonstrated strong performance in real-world churn prediction tasks (Chen and Guestrin, 2016). LightGBM is designed for computational efficiency and scalability using a leaf-wise tree growth strategy that allows faster training and effective handling of large feature sets making it suitable for detecting subtle churn-related patterns (Ke et al., 2017).

In addition to the tree-based approaches, an Artificial Neural Network (ANN) is implemented to examine the effectiveness of a deep-learning solution. Neural networks are able to model complex non-linear patterns that are often difficult for traditional machine-learning techniques to capture. The ANN is constructed using TensorFlow and Keras and trained on the preprocessed dataset allowing a direct comparison with classical and ensemble models to determine whether greater model complexity results in meaningful improvements in churn prediction performance (Goodfellow et al., 2016).

4.9 Test Strategy

The testing strategy for this project focuses on evaluating the performance reliability and generalisation capability of the machine-learning models developed for customer churn prediction. As the project is centred on data analysis and predictive modelling rather than software engineering testing is conducted through systematic model validation rather than traditional unit or integration testing (Provost and Fawcett, 2013).

The dataset is partitioned into training and testing sets using a stratified train–test split to maintain the original distribution of churn and non-churn customers. To avoid data leakage and support reliable performance assessment, SMOTE is applied exclusively to the training data. Logistic Regression, Random Forest, XGBoost, LightGBM, and the Artificial Neural Network are then trained on the balanced training set and evaluated using unseen test data.

Model performance is evaluated using a range of metrics, including accuracy, precision, recall, F1-score and ROC–AUC, to provide a balanced assessment that is appropriate for imbalanced classification tasks (He and Garcia, 2009; Saito and Rehmsmeier, 2015). Confusion matrices are further analysed to examine classification errors and the detection of minority-class instances. Alongside quantitative evaluation SHAP-based explainability is applied to ensure that model predictions are consistent with domain knowledge and to enhance transparency in the decision-making process (Lundberg and Lee, 2017).

This testing strategy ensures that model performance is evaluated consistently across all approaches, results are comparable, repeatable and predictions can be interpreted reliably within a real-world banking context.

4.10 Evaluation Metrics

The performance of the customer churn prediction models is evaluated using multiple classification metrics. This approach is required because churn prediction represents an imbalanced classification task in which accuracy alone may lead to misleading interpretations of model effectiveness (He and Garcia, 2009). These evaluation measures are computed from the confusion matrix which is composed of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Accuracy represents the ratio of correctly predicted instances to the total number of observations and is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

While accuracy provides a general overview of performance, it is insufficient in churn prediction because the majority class (non-churn) dominates the dataset.

Precision measures the proportion of correctly predicted churn customers among all customers predicted as churn:

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

Precision reflects the reliability of churn predictions and highlights the cost of false positives, where non-churn customers are incorrectly targeted

Recall measures the proportion of actual churned customers correctly identified by the model:

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

Recall is particularly important in churn prediction, as failing to identify churn-prone customers can result in lost revenue and missed retention opportunities (Verbeke et al., 2012).

The F1-score is calculated as the harmonic mean of precision and recall, expressed as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

This metric offers a balanced evaluation of model performance when both false positives and false negatives are of concern which is common in imbalanced datasets.

The Receiver Operating Characteristic (ROC) curve illustrates a model's ability to differentiate between churned and non-churned customers across a range of classification thresholds and is based on the true positive rate (TPR) and false positive rate (FPR):

- **True Positive Rate (TPR):**

$$TPR = \frac{TP}{TP + FN}$$

- **False Positive Rate (FPR):**

$$FPR = \frac{FP}{FP + TN}$$

The **ROC–AUC** summarises the ROC curve into a single value between 0 and 1, where higher values indicate stronger class separation. ROC–AUC is particularly useful as it assesses model discrimination independently of a fixed threshold (Fawcett, 2006).

Finally, the **confusion matrix** provides a detailed breakdown of prediction outcomes, enabling direct analysis of misclassification patterns and supporting practical interpretation of model performance.

Overall, the combined use of accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices ensures a robust and reliable evaluation framework for customer churn prediction, where no single metric alone can adequately capture model effectiveness.

4.11 Ethical, Legal, Social and Professional Issues

Although this study is based on a publicly available and anonymised dataset a range of ethical, legal, social and professional considerations remain relevant when developing machine-learning models for customer churn prediction in practical banking settings churn prediction systems typically process sensitive personal and financial data requiring organisations to comply with the General Data Protection Regulation (GDPR) which mandates the lawful fair and transparent handling of personal information (European Parliament and Council, 2016). In addition, principles such as data minimisation and purpose limitation must be adhered to in order to ensure that customer data is used responsibly and solely for legitimate business objectives.

Bias and fairness represent critical ethical concerns in predictive modelling machine-learning algorithms can unintentionally reproduce or amplify existing social biases leading to discriminatory outcomes (Mehrabi et al., 2021). Features such as age, geography or gender may influence predictions in ways that disadvantage certain customer groups to mitigate these risks fairness monitoring bias detection techniques and transparent reporting are essential professional practices when deploying churn prediction systems in industry.

Social implications must also be considered as churn prediction systems influence how customers are targeted for retention strategies. Organisations must ensure that retention actions are ethical non-intrusive and respectful of customer autonomy transparency regarding how customer data is used helps build trust and aligns with professional standards of accountability and integrity in data science practice (ACM, 2018).

Professional considerations further include the responsible development evaluation and communication of predictive systems. Good documentation reproducibility and clear acknowledgement of model limitations are essential explainability techniques such as SHAP support these professional standards by improving transparency and enabling stakeholders to understand the reasoning behind model predictions.

Although this academic project does not involve direct user data collection or real-world deployment these ethical, legal, social and professional considerations remain highly relevant for any future implementation of churn prediction systems in practical banking environments.

4.12 Project Management and Planning

The project was planned and managed using a structured Gantt chart to organise tasks manage dependencies and ensure timely completion of each project phase. The Gantt chart outlines key activities including data preparation exploratory analysis, model development, evaluation and final reporting. The complete project schedule is provided in Appendix A.

4.13 Dataset Description (Feature table)

A detailed description of the dataset features used in this study is provided in Appendix B (Table 8).

5. Quality and Results

5.1 Overview of Experimental Results

This chapter reports the experimental results obtained from the customer churn prediction models developed in this study. After completing data preprocessing, feature encoding, feature scaling, and class balancing using SMOTE, multiple machine-learning models were trained and evaluated on the bank customer churn dataset.

The evaluated models include Logistic Regression, Random Forest, XGBoost, LightGBM, and an Artificial Neural Network, allowing a consistent comparison of classical, ensemble, and deep-learning approaches. Model performance is analysed using accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices, with particular emphasis on metrics that are appropriate for imbalanced classification problems.

The chapter analyses individual model performance examines the impact of SMOTE compares results across models and uses SHAP to explain feature importance. Findings are discussed in relation to project objectives, existing literature and practical feasibility.

5.2 Customer Characteristics and Churn Patterns

This section presents selected observations from the customer dataset that are relevant to understanding churn behaviour and interpreting the predictive results presented later in this chapter. Unlike the exploratory analysis in the Methodology chapter, which focused on supporting preprocessing and modelling decisions, the analysis here highlights key customer characteristics that provide contextual insight into model performance and explainability outcomes. These observations support the interpretation of SHAP-based feature importance discussed in subsequent sections.

5.2.1: Gender vs Customer Churn

Figure 5 illustrates the relationship between gender and customer churn by comparing the number of customers who remained with the bank and those who exited across male and female groups. Churn is observed in both gender categories although the proportion of churned customers differs slightly between them. This suggests that gender may have some influence on churn behaviour however, it is unlikely to be a dominant factor in isolation. The role of gender is therefore interpreted alongside other demographic and financial variables and is examined further using SHAP analysis to assess its relative importance.

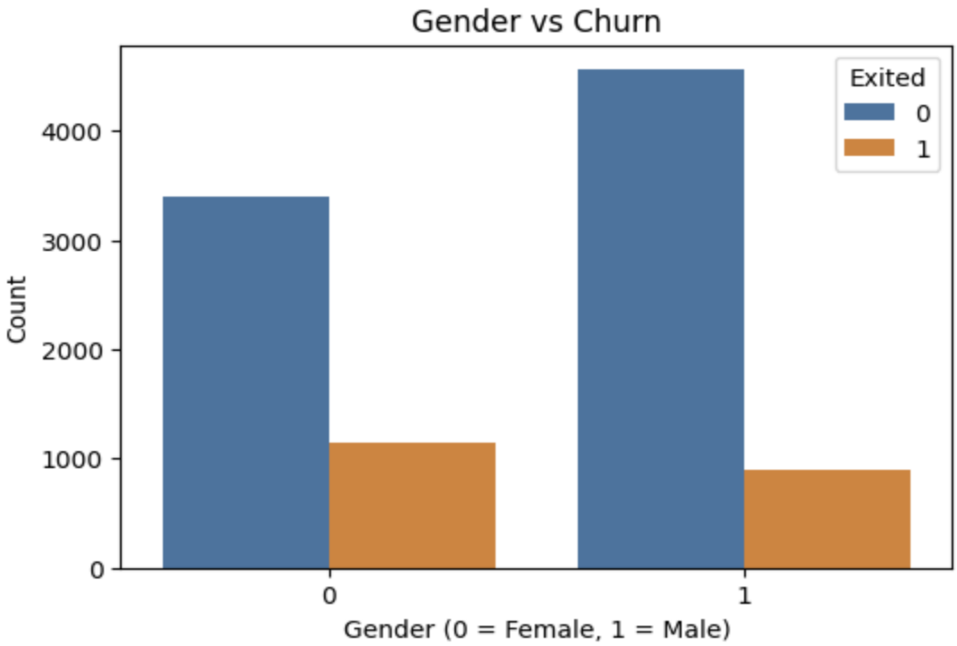


Figure 5 Gender vs Customer Churn

5.2.2: Age Distribution of Churned and Non-Churned Customers

Figure 6 presents the age distribution of churned and non-churned customers using a boxplot. The figure indicates that customers who exited the bank tend to be older on average than those who remained. This suggests that age plays an important role in influencing churn behaviour. However, age alone does not fully explain churn and should be interpreted alongside other demographic and financial attributes. Its relative contribution is examined further through SHAP-based explainability in later sections.

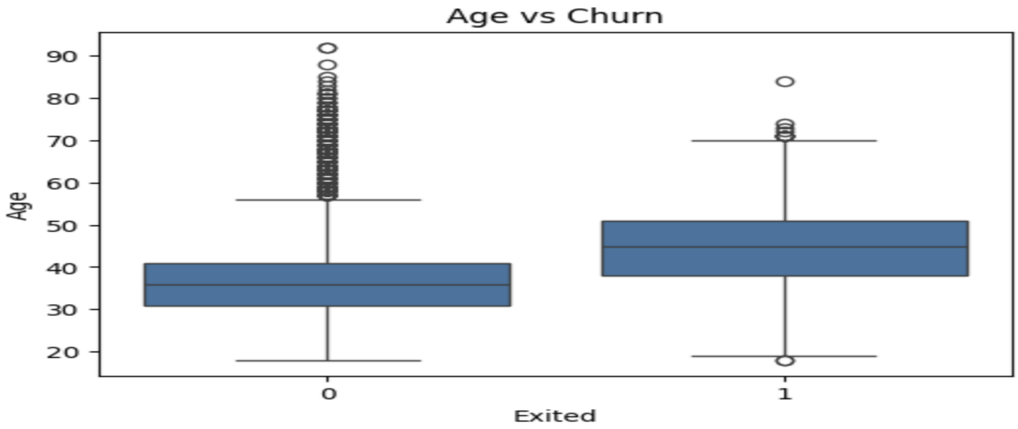


Figure 6 Age Distribution of Churned and Non-Churned Customers

5.3 Model Performance Evaluation.

5.3.1 Logistic Regression

Logistic Regression is employed as a baseline model to establish a reference for comparing the performance of more advanced churn prediction techniques. The model is trained on the preprocessed and SMOTE-balanced training dataset and evaluated using previously unseen test data. While it achieves reasonable performance in identifying churned customers, its ability to model complex non-linear customer behaviour remains limited.

Table 1 summarises the evaluation results for the Logistic Regression model. The model records an accuracy of 0.7155 and a recall of 0.6978 indicating that a large proportion of churned customers are successfully identified. However, the precision value of 0.3890 reveals a relatively high rate of false positive predictions meaning that many customers predicted to churn ultimately remain with the bank. This trade-off between recall and precision is reflected in the F1-score of 0.4996. In addition, the ROC–AUC score of 0.7769 indicates a good overall capacity to discriminate between churned and non-churned customers.

Figure 7 shows the confusion matrix for the Logistic Regression model. The model correctly classifies a large proportion of non-churn customers while also identifying many churned customers explaining its relatively high recall. However, the presence of a notable number of false positives highlights the trade-off between recall and precision observed for this linear model. These limitations motivate the use of ensemble and deep-learning approaches examined in the following sections.

Accuracy	Precision	Recall	F1-score	ROC-AUC
0.7155	0.3890	0.6978	0.4996	0.7769

Table 2 Evaluation Metrics for Logistic Regression

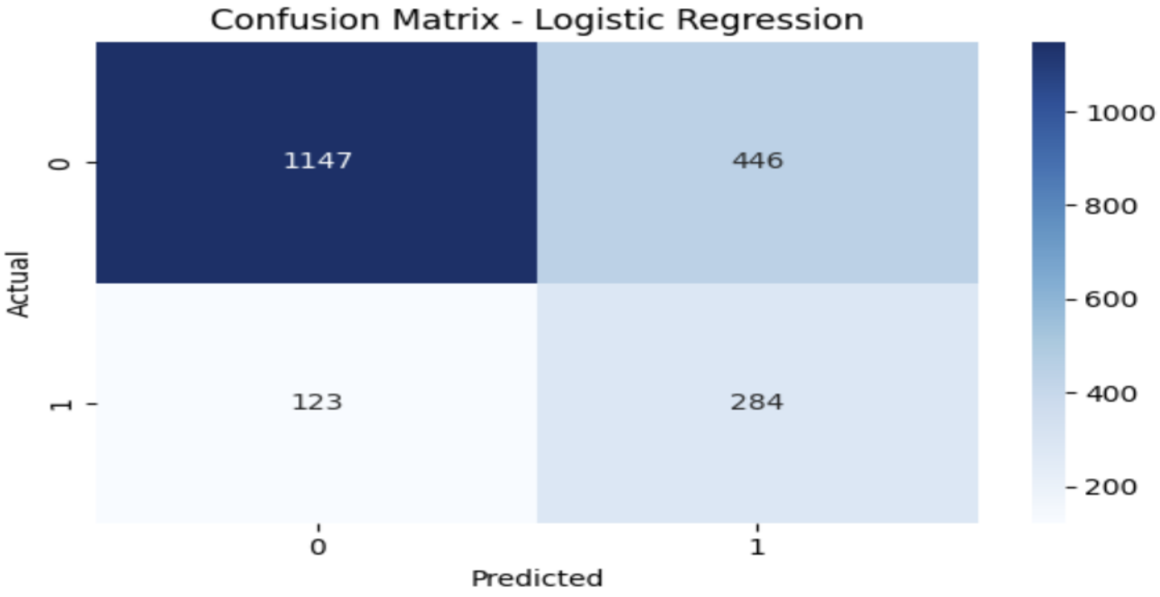


Figure 7 Confusion Matrix for Logistic Regression

5.3.2 Random Forest

Random Forest was applied to evaluate whether an ensemble-based approach could improve churn prediction performance compared to the baseline Logistic Regression model. Trained on the SMOTE-balanced dataset the model demonstrates improved ability to capture non-linear relationships between customer attributes and churn behaviour.

As shown in Table 2, Random Forest achieved an accuracy of 0.8420, indicating strong overall classification performance. The precision score of 0.6140 represents a notable improvement over the baseline model reflecting fewer false churn predictions the recall value of 0.6020 indicates that the model correctly identified a substantial proportion of churned customers resulting in a balanced F1-score of 0.6079. The ROC–AUC score of 0.8501 further confirms strong class-separation capability.

Figure 8 presents the confusion matrix for the Random Forest model. Compared to Logistic Regression, the model shows fewer false positives and improved precision while maintaining stable churn detection. Overall, these results demonstrate that Random Forest achieves a better balance between identifying churned customers and reducing misclassification supporting its effectiveness for churn prediction.

Accuracy	Precision	Recall	F1-score	ROC-AUC
0.8420	0.6140	0.6020	0.6079	0.8501

Table 3 Performance Metrics of the Random Forest Model

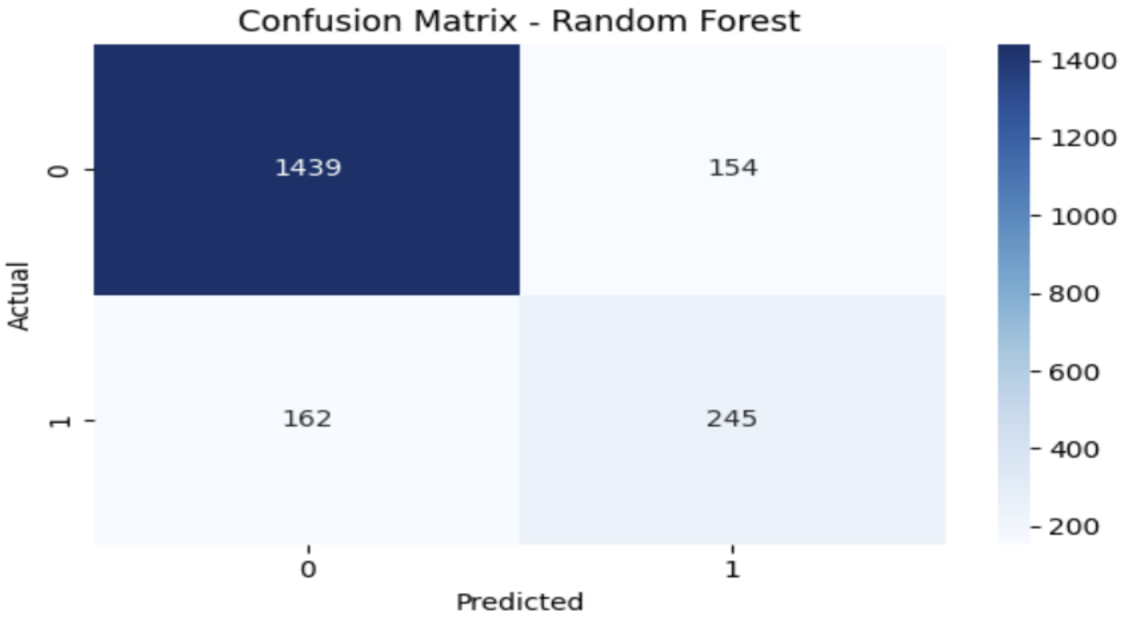


Figure 8 Confusion Matrix for Random Forest

5.3.3 XGBoost

XGBoost was evaluated to assess the effectiveness of gradient boosting for customer churn prediction. Trained on the SMOTE-balanced dataset the model demonstrates strong predictive performance reflecting its ability to capture complex feature interactions and focus on difficult-to-classify churn cases.

Table 3 presents the evaluation metrics for the XGBoost model. The model achieved an accuracy of 0.8555 and a precision of 0.6569 indicating fewer false churn predictions compared to previous models. The recall value of 0.6069 shows that a substantial proportion of churned customers were correctly identified resulting in a balanced F1-score of 0.6309. The ROC–AUC score of 0.8696 further confirms strong class-separation capability.

Figure 9 shows the confusion matrix for the XGBoost model. The model correctly classifies a large number of non-churn customers while also identifying many churned customers, contributing to its balanced recall and precision. Compared to Logistic Regression and Random Forest XGBoost achieves a stronger trade-off between churn detection and misclassification reinforcing its effectiveness for churn prediction.

Accuracy	Precision	Recall	F1-score	ROC-AUC
0.8555	0.6569	0.6069	0.6309	0.8696

Table 4 Performance Metrics of the XGBoost Model

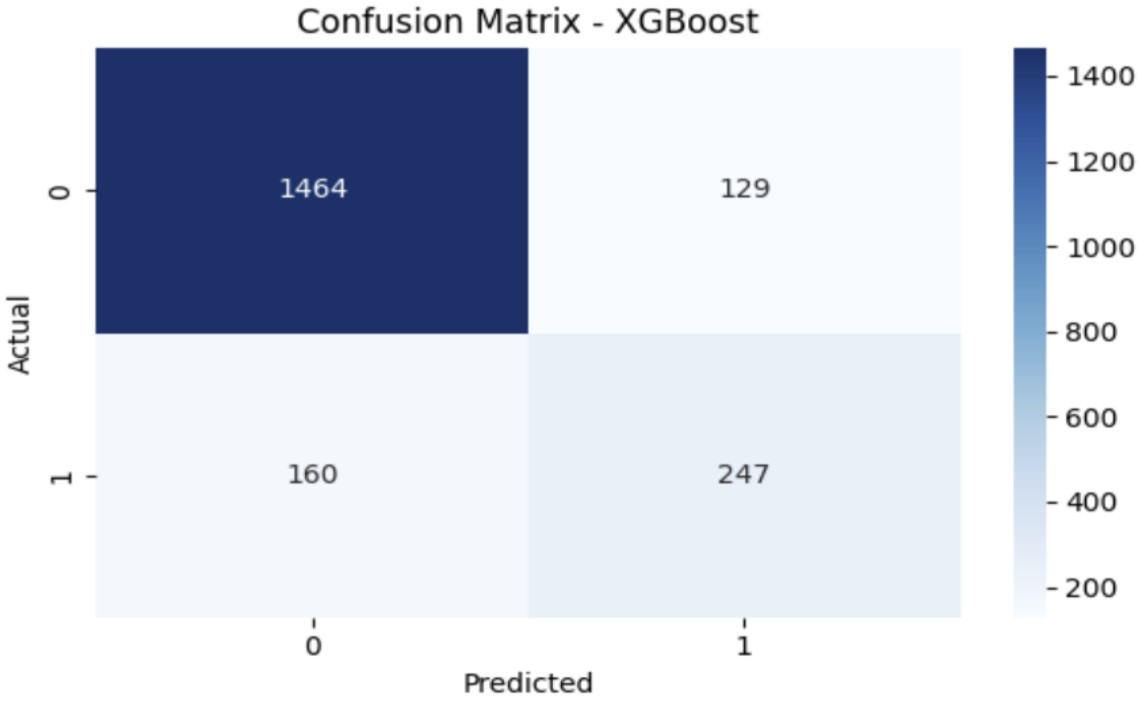


Figure 9 Confusion Matrix for XGBoost

5.3.4 LightGBM

LightGBM was evaluated to assess a more efficient gradient boosting framework under the same experimental conditions. Trained on the SMOTE-balanced dataset the model demonstrates strong and stable performance indicating its effectiveness in handling structured customer data and capturing subtle churn-related patterns.

Table 4 presents the evaluation metrics for the LightGBM model. The model achieved an accuracy of 0.8585 and a precision of 0.6925 reflecting a low rate of false churn predictions however, the recall value of 0.5479 indicates that a smaller proportion of churned customers were identified compared to other boosting models. This trade-off is reflected in the F1-score of 0.6118 the ROC–AUC score of 0.8552 confirms strong overall class-separation capability.

Figure 10 shows the confusion matrix for the LightGBM model. The model correctly classifies a large majority of non-churn customers contributing to high accuracy and precision. However, a higher number of false negatives explains the lower recall indicating a more conservative prediction strategy. Overall, the results highlight a clear trade-off between precision and recall for LightGBM in the churn prediction task.

Accuracy	Precision	Recall	F1-score	ROC-AUC
0.8585	0.6925	0.5479	0.6118	0.8552

Table 5 Performance Metrics of the LightGBM Model

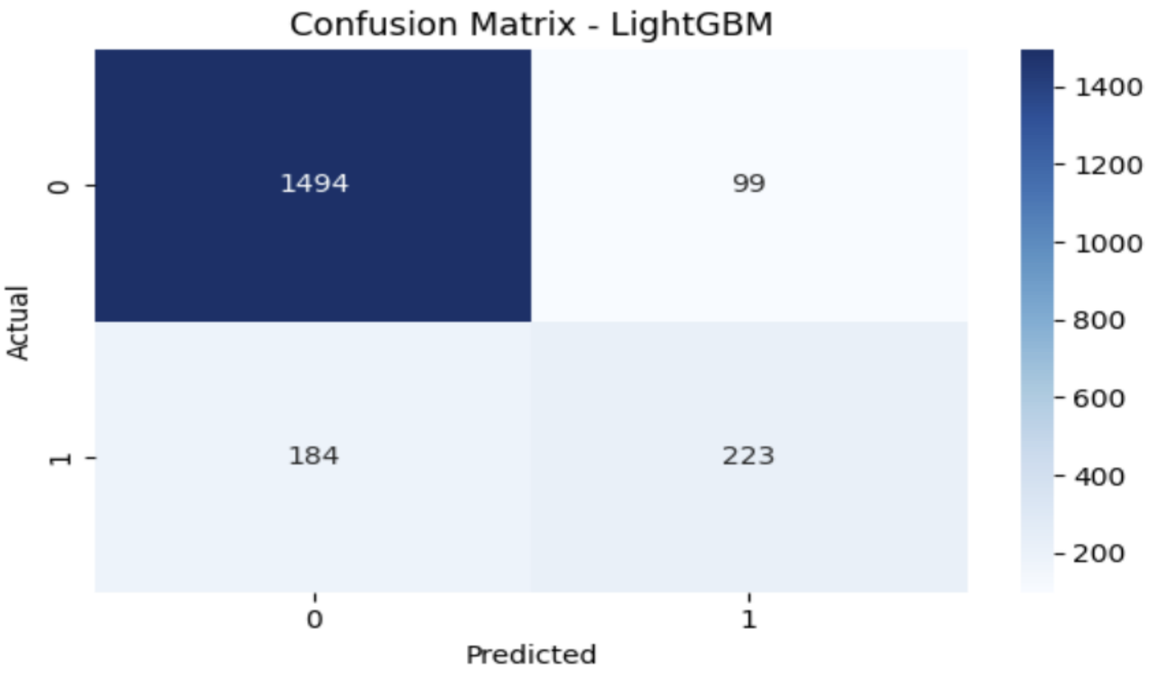


Figure 10 Confusion Matrix for LightGBM

5.3.5 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) was evaluated to assess whether a deep-learning approach could capture complex non-linear patterns in customer behaviour. Trained on the SMOTE-balanced dataset the ANN demonstrates competitive performance indicating its ability to learn richer feature representations compared to linear models.

Table 5 presents the evaluation metrics for the ANN. The model achieved an accuracy of 0.8145 and a recall of 0.6609 showing strong sensitivity to churned customers. However, the precision score of 0.5359 indicates a higher number of false positives compared to tree-based models resulting in an F1-score of 0.5919 the ROC–AUC value of 0.8546 confirms strong overall class-separation capability.

Figure 11 shows the confusion matrix for the ANN. The model correctly identifies a substantial number of churned customers, explaining its high recall but also produces more false positives, which lowers precision. Overall, the ANN prioritises churn detection and performs well in recall-focused scenarios though its performance is less balanced than that of boosting-based models.

Accuracy	Precision	Recall	F1-score	ROC-AUC
0.8145	0.5359	0.6609	0.5919	0.8546

Table 6 Performance Metrics of the Artificial Neural Network Model

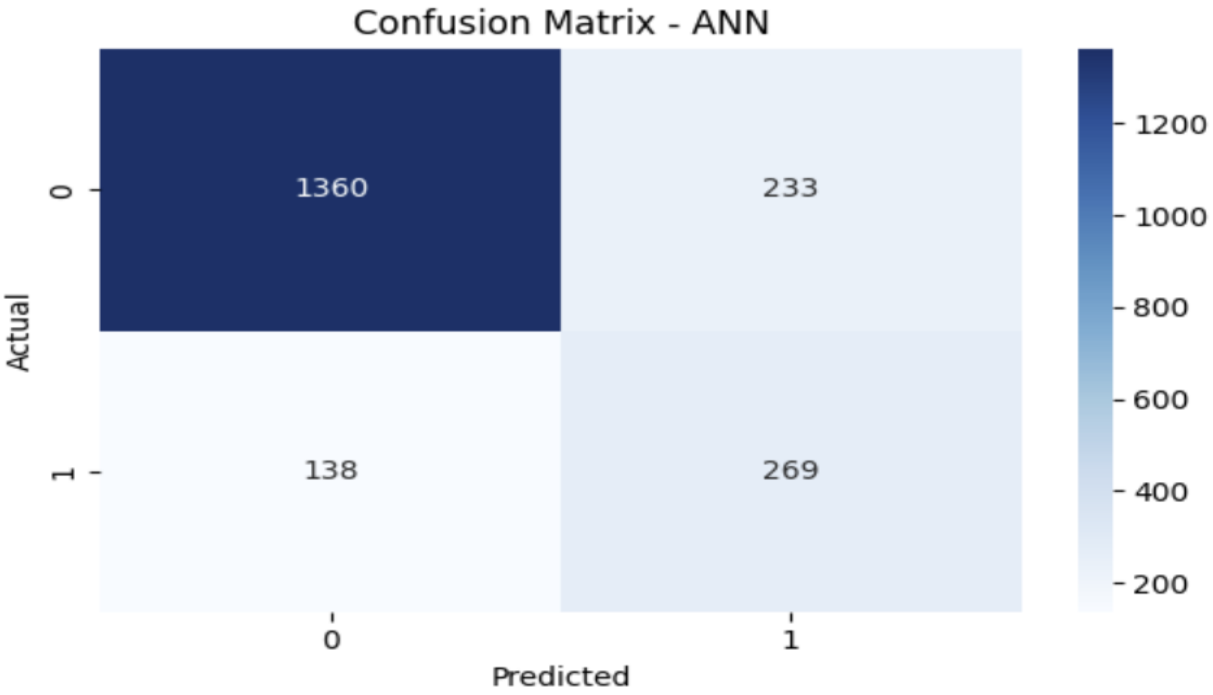


Figure 11 Confusion Matrix for ANN

5.4 Impact of Class Imbalance and SMOTE

Customer churn prediction represents an imbalanced classification task because churned customers form a much smaller proportion of the dataset compared with non-churned customers. When predictive models are trained on such imbalanced data, they often become biased toward the majority class, leading to high overall accuracy but weak detection of churned customers. This behaviour was evident during the initial evaluation of models trained on the original, unbalanced dataset.

To mitigate this challenge SMOTE was applied exclusively to the training data in order to enhance the representation of churned customers. Following the application of SMOTE the models exhibited notable improvements in recall and F1-score, reflecting more effective identification of the minority churn class. These performance gains were especially pronounced for XGBoost, LightGBM, and the Artificial Neural Network, which benefited from being trained on a more balanced dataset.

Overall, the application of SMOTE enhanced churn detection performance without significantly compromising overall classification accuracy. These findings are consistent with existing research which highlights the importance of handling class imbalance when developing reliable and practically useful churn prediction models (Chawla et al., 2002; He and Garcia, 2009).

5.5 comparative Analysis of Machine Learning Models

A comparative analysis of the machine-learning models highlights notable differences in performance across accuracy, precision, recall, F1-score, ROC–AUC and confusion matrix results. Logistic Regression records an accuracy of 0.7155 alongside a relatively high recall of 0.6978, demonstrating its ability to identify a large proportion of churned customers. Nevertheless, the model's low precision value of 0.3890 and F1-score of 0.4996, driven by a high rate of false positive predictions reduce its practical reliability.

Random Forest demonstrated a substantial improvement achieving an accuracy of 0.8420 precision of 0.6140 and F1-score of 0.6079. The confusion matrix shows a reduction in false positives compared to Logistic Regression while maintaining reasonable churn detection. Its ROC–AUC score of 0.8501 confirms strong class discrimination and balanced performance.

Among all evaluated models XGBoost demonstrates the strongest overall performance the model achieves an accuracy of 0.8555 a precision of 0.6569 and a recall of 0.6069 while also recording the highest F1-score (0.6309) and ROC–AUC value (0.8696). Analysis of the confusion matrix further indicates a favourable balance between correctly identifying churned customers and limiting false churn predictions supporting XGBoost as the most reliable model in this study.

LightGBM achieved the highest accuracy (0.8585) and precision (0.6925) indicating excellent identification of non-churn customers however its lower recall (0.5479) shows that a higher proportion of churn cases were missed reflecting a conservative prediction strategy that prioritises precision over recall.

The Artificial Neural Network achieved an accuracy of 0.8145 and recall of 0.6609 demonstrating effective churn detection but its lower precision (0.5359) resulted in more false positives. While the

ANN captured complex patterns its overall performance remained slightly below that of the boosting models.

Overall, the results demonstrate that XGBoost achieves the most effective balance between precision and recall, as evidenced by its highest F1-score and ROC–AUC values. These outcomes align with previous research highlighting the strong performance of ensemble and gradient boosting approaches in customer churn prediction (Verbeke et al., 2012).

5.6 Explainability and Feature Importance Analysis

Figure 12 displays the SHAP summary plot which highlights the relative importance of features and their average influence on churn predictions. The analysis shows that Age is the most influential feature indicating that customer age has a substantial effect on churn behaviour. Number of Products emerges as the second most important factor suggesting that customers who hold fewer banking products are more likely to leave the bank.

Tenure and IsActiveMember also exhibit strong influence, highlighting the importance of long-term engagement and account activity in customer retention. Features such as Gender, Balance, and Geography (Germany) have a moderate impact on churn predictions, while Estimated Salary, Credit Score, Geography (Spain), and HasCrCard contribute relatively little to the model’s decisions.

Overall, the SHAP results demonstrate that churn is driven primarily by behavioural and engagement-related factors rather than purely financial attributes. This confirms the reliability of the model and shows that SHAP provides transparent and interpretable insights that are meaningful for real-world banking decision-making (Lundberg and Lee, 2017).

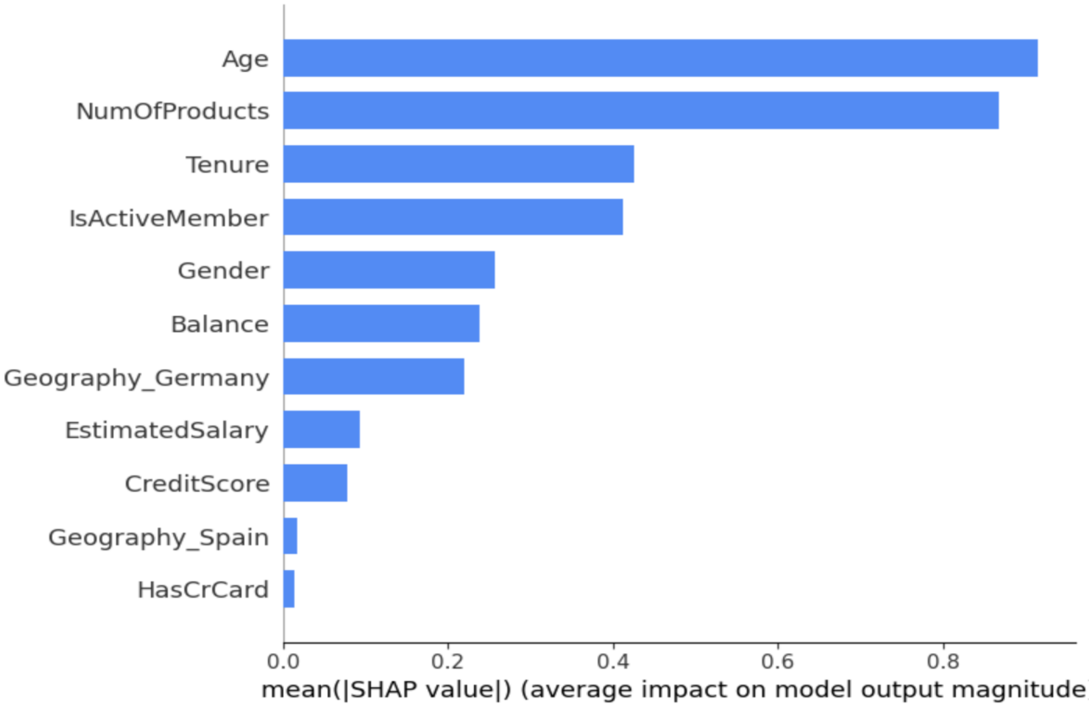


Figure 12 SHAP Summary Plot Showing Feature Importance

5.7 Hyperparameter Tuning

Hyperparameter tuning was applied to the XGBoost model to examine whether optimisation could further improve predictive performance and generalisation capability. XGBoost was selected for tuning because its performance is sensitive to parameter configuration and it demonstrated strong baseline results prior to optimisation (Chen and Guestrin, 2016).

GridSearchCV is employed to systematically evaluate combinations of key hyperparameters, including the learning rate, maximum tree depth, number of estimators, and regularisation settings. Cross-validation is performed on the training dataset to mitigate overfitting and enhance model robustness, which is particularly important given the imbalanced characteristics of the churn dataset (Bergstra and Bengio, 2012).

Table 6 compares the performance of the XGBoost model before and after hyperparameter tuning. The tuned model achieved slightly higher precision (0.6618) indicating improved control over false positive churn predictions. However, the default configuration achieved higher recall, F1-score and ROC–AUC demonstrating stronger overall churn detection capability. This suggests that the tuned parameter set adopted a more conservative prediction strategy prioritising precision and stability over minority-class detection.

These findings highlight that hyperparameter optimisation does not always guarantee improved performance across all evaluation metrics. Instead, its effectiveness depends on dataset characteristics and the relative importance of precision versus recall. This reinforces the importance of empirical evaluation rather than assuming automatic performance gains through tuning (Verbeke et al., 2012).

Model Version	Accuracy	Precision	Recall	F1-score	ROC–AUC
XGBoost (Tuned)	0.8520	0.6618	0.5577	0.6053	0.8469
XGBoost (Default)	0.8555	0.6569	0.6069	0.6309	0.8696

Table 7 XGBoost Performance Before and After Hyperparameter Tuning

5.8 Discussion in Relation to Objectives and Literature

The findings of this study directly support the stated project objectives and are consistent with conclusions reported in existing customer churn prediction research. A primary objective was to identify a suitable predictive model for churn detection within an imbalanced banking dataset. The experimental results demonstrate that ensemble-based approaches, particularly XGBoost, achieve the most effective balance between precision and recall, as reflected by the highest F1-score and ROC–AUC values. This result is in line with previous studies, which highlight the suitability of gradient boosting methods for churn prediction due to their ability to model complex non-linear relationships and interactions among customer features (Verbeke et al., 2012; Chen and Guestrin, 2016).

Another important objective was to examine the impact of class imbalance and the effectiveness of SMOTE. The results clearly indicate that applying SMOTE to the training data improved recall and

F1-score across all models confirming that class imbalance significantly affects churn detection performance. This finding supports prior studies which highlight that imbalanced datasets can lead to biased predictions and poor minority-class identification if not properly addressed (Chawla et al., 2002; He and Garcia, 2009). The improvements observed after balancing validate the methodological choices adopted in this project.

6. Evaluation and Conclusion

This project aimed to design implement and evaluate a robust and interpretable customer churn prediction framework for the banking sector using machine-learning and deep-learning techniques. The primary objectives were to identify customers at risk of churn compare the performance of multiple predictive models under consistent experimental conditions address the challenges associated with class imbalance and provide transparent explanations for model decisions overall the project successfully achieved its objectives and delivered strong predictive performance alongside meaningful practical insights.

From a technical perspective the project demonstrated that customer churn prediction is best approached as an imbalanced classification problem requiring careful preprocessing appropriate evaluation metrics and advanced modelling techniques Logistic Regression provided a useful and interpretable baseline but showed clear limitations due to its linear decision boundary particularly in handling complex customer behaviour and reducing false churn predictions which is consistent with prior churn studies (Neslin et al., 2006; Verbeke et al., 2012) Ensemble and boosting-based models significantly improved performance, confirming that non-linear models are better suited to structured banking data (Breiman, 2001; Chen and Guestrin, 2016).

Among all models evaluated XGBoost achieved the most balanced and reliable performance producing the highest F1-score and ROC–AUC while maintaining stable precision and recall. This supports existing research that identifies gradient boosting methods as highly effective for churn prediction due to their ability to capture complex feature interactions and control overfitting (Chen and Guestrin, 2016; Ke et al., 2017). LightGBM achieved the highest accuracy and precision indicating strong confidence in churn predictions, but at the expense of recall a behaviour previously noted in precision-oriented boosting models (Ke et al., 2017) The Artificial Neural Network demonstrated strong churn detection capability with higher recall but increased false positives, reflecting the known trade-off between sensitivity and stability in neural models applied to tabular data (Goodfellow et al., 2016; Zhang, Zhao and LeCun, 2021) These outcomes confirm that no single metric is sufficient and that model selection must align with business priorities rather than accuracy alone (Verbeke et al., 2012).

The handling of class imbalance was a critical factor influencing model effectiveness. Initial observations confirmed that imbalanced data leads to biased predictions favouring non-churn customers even when overall accuracy appears high the application of SMOTE to the training data significantly improved minority-class learning and increased recall and F1-score across all models this confirms that addressing imbalance is essential for practical churn prediction systems, particularly in business environments where failing to identify churned customers directly impacts revenue (Chawla et al., 2002; He and Garcia, 2009). These findings strongly support prior research which emphasises imbalance-aware learning strategies for reliable churn detection.

Explainability played a central role in strengthening the practical value of this project by integrating SHAP into the evaluation process the project moved beyond pure predictive accuracy to provide transparent and interpretable insights into model behaviour SHAP analysis consistently identified Age, Number of Products, Tenure, and IsActiveMember as the most influential features driving churn predictions these results align closely with established churn literature and domain knowledge reinforcing confidence in the model outputs (Neslin et al., 2006; Lundberg and Lee, 2017; Ou, 2023). Importantly this explainability component addresses a key limitation of many churn studies where high-performing models lack transparency and are difficult to justify in regulated financial environments (Ribeiro, Singh and Guestrin, 2016).

Hyperparameter tuning was explored through optimisation of the XGBoost model while tuning is often assumed to improve performance the results showed that the tuned model did not outperform the default configuration in terms of recall, F1-score, or ROC–AUC this highlights an important empirical insight optimisation does not automatically lead to better results and must always be validated against strong baselines similar observations have been reported in applied machine-learning research where default configurations can already be well-optimised for general use cases (Chen and Guestrin, 2016).

From a research perspective, the findings of this project strongly align with and extend existing literature. Previous studies consistently report that ensemble and gradient boosting models outperform traditional classifiers in churn prediction tasks due to their ability to model non-linear relationships and feature interactions (Breiman, 2001; Verbeke et al., 2012). The superior performance of XGBoost and LightGBM in this project confirms these conclusions. Furthermore, the emphasis on recall, F1-score, and ROC–AUC rather than accuracy alone reflects best practice in imbalanced learning research (He and Garcia, 2009). By systematically comparing classical, ensemble, and neural models within a single framework and integrating explainability, this project addresses several gaps identified in prior work.

In terms of feasibility and commercial relevance the project demonstrates a realistic and deployable churn prediction approach the use of open-source tools a structured workflow and interpretable outputs make the solution practical for real-world adoption. From a business perspective effective churn prediction enables banks to prioritise retention efforts reduce customer acquisition costs and improve long-term profitability reinforcing established findings on the economic value of customer retention (Reichheld and Sasser, 1990; Gupta and Zeithaml, 2006). However, real-world deployment would require additional considerations including fairness assessment, continuous model monitoring, and compliance with data protection regulations such as GDPR (European Parliament and Council, 2016; Mehrabi et al., 2021).

Several limitations should be acknowledged the dataset used is static and does not include real-time behavioural or transactional data typically available in commercial banking systems Hyperparameter tuning was applied only to one model and external validation on additional datasets would be required to confirm generalisability. Despite these limitations the project demonstrates strong methodological rigour and awareness of real-world constraints.

Future work could extend this research by incorporating time-series data cost-sensitive learning approaches fairness-aware modelling and real-time churn prediction systems exploring automated hyperparameter optimisation and hybrid modelling strategies could further enhance robustness and performance.

In conclusion, this project delivers a comprehensive, interpretable, and business-oriented customer churn prediction framework by combining class balancing, advanced ensemble models, and explainable AI, the study provides both strong predictive performance and actionable insights. The results confirm that gradient boosting models, particularly XGBoost offer the most reliable balance between churn detection and prediction accuracy making them highly suitable for deployment in the banking sector. Overall, the project makes a meaningful contribution to churn prediction research and provides a solid foundation for future academic and commercial applications.

7 References

- Arner, D.W., Barberis, J. and Buckley, R.P. (2017)
  FinTech and RegTech in a nutshell, and the future in a sandbox. North Carolina Banking Institute, 21(1), pp. 81–105.
- Association for Computing Machinery (ACM) (2018)
  ACM Code of Ethics and Professional Conduct. Available at: https://www.acm.org/code-of-ethics (Accessed: day Month year).
- Bergstra, J. and Bengio, Y. (2012)
  Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, pp. 281–305.
- Breiman, L. (2001)
  Random forests. Machine Learning, 45(1), pp. 5–32.
- 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.
- Chen, T. and Guestrin, C. (2016)
  XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA: ACM, pp. 785–794.
- CRISP-DM Consortium (2000)
  CRISP-DM 1.0: Step-by-step data mining guide. Available at: https://www.the-modeling-agency.com/crisp-dm.pdf (Accessed: day Month year).
- European Parliament and Council (2016)
  Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union, L119, pp. 1–88.
- Fawcett, T. (2006)
  An introduction to ROC analysis. Pattern Recognition Letters, 27(8), pp. 861–874.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016)
  Deep Learning. Cambridge, MA: MIT Press.
- Gupta, S. and Zeithaml, V. (2006)
  Customer metrics and their impact on financial performance. Marketing Science, 25(6), pp. 718–739.
- Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2007)
  Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), pp. 2902–2917.
- He, H. and Garcia, E.A. (2009)
  Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), pp. 1263–1284.

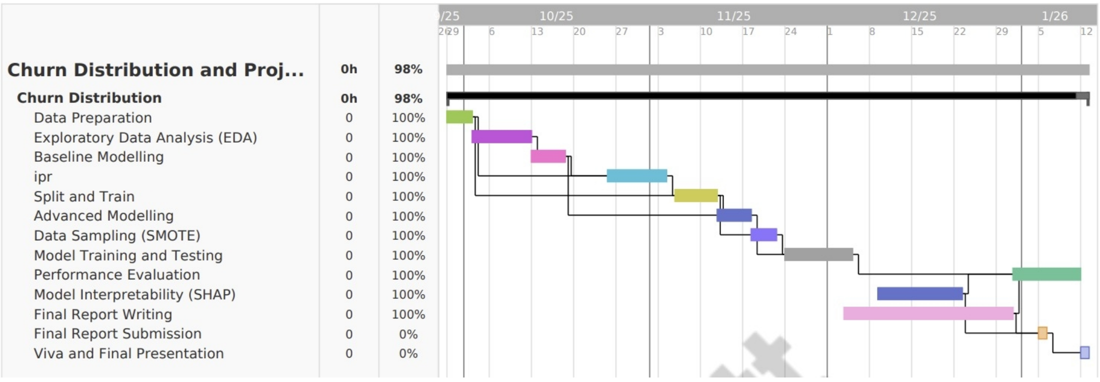
- Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013)
  Applied Logistic Regression. 3rd edn. New York: Wiley.
- Idris, A., Khan, A. and Lee, Y.S. (2012)
  Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification. Applied Intelligence, 39(3), pp. 659–672.
- Kaggle (2023)
  Bank Customer Churn Dataset. Available at: https://www.kaggle.com/ (Accessed: day Month year).
- Kuhn, M. and Johnson, K. (2013)
  Applied Predictive Modeling. New York: Springer.
- Kumar, V. and Ravi, V. (2019)
  Predicting customer churn using machine learning: A systematic literature review. Computers & Industrial Engineering, 131, pp. 1–16.
- Ling, C.X. and Li, C. (1998)
  Data mining for direct marketing: Problems and solutions. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD ’98). New York: AAAI Press, pp. 73–79.
- Lundberg, S.M. and Lee, S.-I. (2017)
  A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, pp. 4765–4774.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021)
  A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), pp. 1–35.
- Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. and Mason, C.H. (2006)
  Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), pp. 204–211.
- O’Neil, C. (2016)
  Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing Group.
- Ou, C. (2023)
  Explainable artificial intelligence for customer churn prediction in the banking sector. Expert Systems with Applications, 213, 118899.
- Pedregosa, F. et al. (2011)
  Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, pp. 2825–2830.
- Provost, F. and Fawcett, T. (2013)
  Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. Sebastopol, CA: O’Reilly Media.
- Reichheld, F.F. and Sasser, W.E. (1990)
  Zero defections: Quality comes to services. Harvard Business Review, 68(5), pp. 105–111.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) ‘Why should I trust you?’ Explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. San Francisco, CA: ACM, pp.

1135–1144.Saito, T. and Rehmsmeier, M. (2015)
The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.

- Van Rossum, G. and Drake, F.L. (2009)
  Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
- Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2012)
  New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach. European Journal of Operational Research, 218(1), pp. 211–229.
- Weiss, G.M. (2004)
  Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsletter, 6(1), pp. 7–19.
- Zhang, Y., Zhao, J. and LeCun, Y. (2021)
  Deep learning for customer behaviour modelling: A review. IEEE Transactions on Neural Networks and Learning Systems, 32(7), pp. 2861–2876.

8. Appendices

8.1 Appendix A: Project Planning and Gantt Chart



Churn Distribution and Proj...	**0h**	**98%**
Churn Distribution	**0h**	**98%**
Data Preparation	0	100%
Exploratory Data Analysis (EDA)	0	100%
Baseline Modelling	0	100%
ipr	0	100%
Split and Train	0	100%
Advanced Modelling	0	100%
Data Sampling (SMOTE)	0	100%
Model Training and Testing	0	100%
Performance Evaluation	0	100%
Model Interpretability (SHAP)	0	100%
Final Report Writing	0	100%
Final Report Submission	0	0%
Viva and Final Presentation	0	0%

Figure 13 Project Gantt Chart Showing Project Phases, Dependencies, and Milestones

8.2 Appendix B: Dataset Description and Feature Definitions

Feature Name Description	Description
CreditScore	The score determines whether a customer has a good relationship with the bank.
Geography	Customers place of residence
Gender	Customer gender
Age	Age of a person
Tenure	Person being with same bank for years
Balance	Amount that person has in the account
NumOfProducts	Services provided by bank
HasCrCard	Whether the customer has a credit card
IsActiveMember	Person using services or not
EstimatedSalary	Annual salary earned by one person
Exited	Indicating customer churn

Table 8 Description of Features in the Bank Customer Churn Dataset

8.3 Appendix C: Representative Code Snippets

Snippet C1: Import libraries and load dataset

```
import pandas as pd
import numpy as np
df = pd.read_csv("Churn_Modelling.csv")
```

Snippet G2: Drop irrelevant identifier columns

```
df.drop(columns=["RowNumber", "CustomerId", "Surname"], inplace=True, errors="ignore")
```

Snippet G3: Encode categorical variables (Gender + Geography)

```
# Gender (Female=0, Male=1)
df["Gender"] = df["Gender"].astype(str).str.strip().str.title()
df["Gender"] = df["Gender"].map({"Female": 0, "Male": 1})
# One-hot encode Geography (France becomes reference when drop_first=True)
df = pd.get_dummies(df, columns=["Geography"], drop_first=True)
# Ensure dummy columns are integers (good practice)
geo_cols = [c for c in df.columns if c.startswith("Geography_")]
df[geo_cols] = df[geo_cols].astype(int)
```

Snippet G4: Feature/target split + stratified train/test split

```
from sklearn.model_selection import train_test_split
X = df.drop(columns=["Exited"])
y = df["Exited"]
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Snippet G5: Standard scaling (fit on train only)

```
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
X_train_s = scaler.fit_transform(X_train)
X_test_s = scaler.transform(X_test)
feature_names = X.columns
```

Snippet G6: Apply SMOTE on training data only (avoid leakage)

```
from imblearn.over_sampling import SMOTE
smote = SMOTE(random_state=42)
X_train_sm, y_train_sm = smote.fit_resample(X_train_s, y_train)
```

Snippet G7: Unified evaluation function (accuracy, precision, recall, F1, ROC–AUC)

```
import numpy as np
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score
def evaluate_model(model, X_test, y_test):
    # Probabilities
    if hasattr(model, "predict_proba"):
        y_prob = model.predict_proba(X_test)[:, 1]
    else:
        y_prob = model.predict(X_test, verbose=0).ravel()
    # Predictions (threshold 0.5)
    y_pred = (y_prob >= 0.5).astype(int)
    return {
        "Accuracy": accuracy_score(y_test, y_pred),
        "Precision": precision_score(y_test, y_pred, zero_division=0),
        "Recall": recall_score(y_test, y_pred, zero_division=0),
        "F1": f1_score(y_test, y_pred, zero_division=0),
        "ROC_AUC": roc_auc_score(y_test, y_prob),
    }
```

Snippet G8: One representative model training (XGBoost example)

```
from xgboost import XGBClassifier
xgb_model = XGBClassifier(
    n_estimators=300,
    max_depth=4,
    learning_rate=0.05,
    subsample=0.8,
    colsample_bytree=0.8,
    eval_metric="logloss",
    random_state=42
)
xgb_model.fit(X_train_sm, y_train_sm)

xgb_results = evaluate_model(xgb_model, X_test_s, y_test)
print(xgb_results)
```

Snippet G9: SHAP explainability (XGBoost)

```
import shap
# Create DataFrame with correct feature names
X_test_shap = pd.DataFrame(X_test_s, columns=feature_names)
explainer = shap.TreeExplainer(xgb_model)
shap_values = explainer.shap_values(X_test_shap, check_additivity=False)
# If SHAP returns a list for binary classification, choose the positive class
if isinstance(shap_values, list):
    shap_values = shap_values[1]
shap.summary_plot(shap_values, X_test_shap)
shap.summary_plot(shap_values, X_test_shap, plot_type="bar")
```