University of Essex

ESSEX BUSINESS SCHOOL

11/11/2025

POSTGRADUATE

Predictive Analytics for Customer Churn in Subscription-Based Businesses

Abstract

Retention of customers is the final determinant of profitability in the telecommunications industry, where subscriptions are prevalent. The problem of customer churn has also become one of the most significant business challenges of the digital age, as the possibility of switching providers has increased. The dissertation presented will answer the question of how the threat of losing customers can be predicted with the assistance of predictive analytics and how the information presented by the data can be applied to actual retention strategies. The research employs a publicly available dataset of Telco Customer Churn on Kaggle, which is a real-world telecommunication setup. A unified preprocessing and evaluation tool was assembled in Python using scikit-learn, which re-streams three monitored machine-learning algorithms: Logistic Regression, Random Forest, and XGBoost. The interpretability measures, odds ratios, and feature-importance measures were utilised to compare the models with the most important measures of accuracy, precision, recall, F1-score, and ROC-AUC.

All algorithms had high predictive validity (ROC-AUC 0.84). Logistic Regression was the best choice for proactive churn prevention, as it achieved the best recall, and XGBoost was only marginally higher in overall accuracy. All the models also had corresponding similar predictors of behaviour, including short tenure, contracting on a month-to-month basis, paying by electronic check, and no support or security add-ons, which were all significant predictors of churn. These findings affirm that the existing customer loyalty drivers comprise service engagement, convenience and perceived value.

The dissertation is informative in both scholarly and practical knowledge fields because it links theories of behaviour, such as the switching-cost and relationship-marketing models, to the latest approaches in predictive modelling. It provides a proactive analytical model of churn and offers

suggestions to managers on structuring contracts, automating payments, and implementing an early customer engagement strategy. The paper finds that predictive analytics, as a visible and ethical means of customer-retention management driven by data, is a powerful and sustainable instrument.

Contents

Chapter 1 – Introduction

1.1 Background of the Study

Digital platforms are becoming the new way of life for consumers, and more than ever, customer retention is a new reality in the modern economy, where subscriptions have become the norm. All forms of Internet communications, streaming services, and software-as-a-service are subject to the category of businesses that can be either subscription-based or usage-based, and thus are highly dependent on customer loyalty. Online infrastructure has developed rapidly, and consumers have acquired excessive choice, with switching costs being minimal to the extent that they can change providers by simply clicking a button. In this regard, the profitability of a firm is becoming less related to the way customers are acquired, but rather to the retention and satisfaction of customers.

Loss of customer relationship with a service provider or customer churn has emerged as one of the most critical issues for organisations. Research has consistently demonstrated that customer acquisition is significantly more expensive than customer retention. However, the issue of churn persists regardless of the investment made in loyalty programmes, discounts, and enhancements to customer service. It is more pronounced within telecommunications companies, where the churn threat has been exacerbated by market saturation, competitive pricing, and the ease of switching to a different service. (Sasser and Reichheld, 1990) With the introduction of predictive analytics and machine learning, organisations can answer this question differently. The predictive models will be able to generate latent trends indicating early signs of churn, based on data obtained through customer interactions, payment history, and service preferences. Telecommunications firms have a particular privilege in this regard, as they develop and archive large amounts of formatted customer data, including information on contract type, monthly payment, tenure, internet service, and payment variations. (Chen and Forman, 2006) Such information may be used to develop models of classification that differentiate between loyal customers and those at risk of abandonment. It is not only possible to target predictive accuracy through the application of supervised learning algorithms, but cost-effective targeted retention interventions can also be developed.

This principle has also been developed in this dissertation, as it utilises the most recent machine-learning techniques to analyse the Telco Customer Churn dataset, which is among the most popular benchmarks available on Kaggle. The analysis will be performed to demonstrate how predictive analytics can enhance the managerial decision-making process and identify the most significant variables that induce churn, providing actionable information to inform the retention plan. (Huang, Kechadi and Buckley, 2012)

1.2 Problem Statement

Although it is a fact that large volumes of customer data are more easily accessible nowadays, it remains true that many organisations struggle to transform this data into valuable business intelligence. These old methods, such as simple regression analysis or rule-based systems, are no longer suitable for complex and nonlinear relations that define real-world customer behaviour. In addition, the application of existing machine-learning algorithms is often limited because they cannot be easily interpreted, despite achieving high predictive accuracy. The managers are not always in need of a forecast of a customer who is going to churn, but rather a reason as to why it is most likely to do so. An example of one of the telecommunications cases is the telecommunications industry. Firms have a high turnover rate due to their extensive operating records, which are killing their profitability. Retention programs are not proactive but rather reactive in nature, launched after customers cancel their services. This trend can be reversed by predictive analytics, which enables the organisation to act earlier and more effectively. The tricky part, though, in this case, is ensuring that technical accuracy is not sacrificed in the name of practical interpretability, i.e., ensuring that predictive models are not only practical to implement but also comprehensible enough that they do not breed doubts about what decision-makers should do with them.

The crux of the issue that this paper aims to address is the possibility of designing, assessing, and testing predictive models that effectively identify at-risk customers in subscription-based companies. It aims to introduce a paradigm in which informed data can be transformed into actionable business plans, thereby bridging the gap between exceptionally sophisticated analytics and managerial utility.

1.3 Aim and Objectives

Aim

This aims to formulate and test a set of predictive models capable of forecasting customer churn using a telecommunications-based e-commerce dataset, thereby providing valuable insights for effective customer retention management.

Objectives

1. To view and preprocess the publicly available Kaggle Telco Customer Churn data for analysis.
2. Training and testing various machine-learning models on the same pipelines of preprocessing, such as the Logistic Regression, Random Forest and XGBoost.

3. To compare the performance of the models in terms of important assessment measures of Accuracy, Precision, Recall, F1-Score and ROC.AUC.

4. To describe the model's results in terms of odds ratios and feature importance, it is necessary to identify the main predictors of churn.

5. To convert the result of the analysis into management implications for improving customer retention and customer satisfaction.

All these assertions shall be used to guide the research process of collecting information, analysing it, and implementing the results.

1.4 Research Questions

The study's research questions are the following:

- ✓ What is the best purpose to apply predictive analytics to know the likely churners?
- ✓ Which machine-learning algorithm is the most interpretable and predictive?
- ✓ What are the data churn drivers, in your opinion?
- ✓ How to incorporate the information received into the predictive models into efficient customer-retention strategies?

The questions, in turn, form the foundation of the analytical section of the study, as they are a collection of hypothetical research and factual information.

1.5 Significance of the Study

As a study subject, it can be used to enhance the growing body of predictive analytics research by demonstrating how machine-learning models are applied in practice. It expands on previous studies on churn prediction by comparing conventional statistical models, such as Logistic

Regression, with high-tech ensemble machine learning algorithms, including Random Forest and XGBoost, in normalising both the pre-processing process and the testing environment. (Verbeke *et al.*, 2012a) The comparison not only reflects the difference in performance but also the trade-offs in interpretability between more straightforward and more complex algorithms.

The results of the managerial strategy are a map of the pathways within which data-driven decision-making can be implemented in customer relationship management. Identifying the most significant variables compared to churn, such as month-to-month contracts, mode of payment, or type of services, will enable managers to make more informed decisions on the intervention with the highest retention possibilities. This helps reduce the overall expense of marketing programs by focusing on specific, compelling, and quantifiable retention programs.

1.6 Scope and Limitations

The study is limited to the supervised learning technique of the structured customer information of a phone company. The targeted predictive accuracy, rather than causation, is the focus of the research, which aims to classify churners rather than identify those who left in absolute terms. The selected algorithms are simple, transparent, and perform well. (Molnar, Casalicchio and Bischl, 2020) The algorithms used, i.e., those of Logistic Regression, Random Forest, and XGBoost, imply a tradeoff between simplicity, transparency, and performance, allowing one to compare interpretable linear and nonlinear, yet powerful, ensembles. Nevertheless, the study has several weaknesses. The information provided is based on past data, which may not accurately reflect current developments in customer behaviours, macroeconomic trends, and competitor activities. It also omits unstructured data, such as customer perceptions or feelings expressed on social media, which can contribute to the accuracy of predictions. Moreover, the research is based on secondary data, which is made available by Kaggle and, therefore, has the limitation of quality and assumptions made by the source. (Ahmad, Jafar and Aljoumaa, 2019)

The dissertation is divided into five major sections, each associated with a phase of the research process.

- **Chapter 1:** Introduction outlines the background of the research, justification, purpose, and scope.
- **Chapter 2:** Literature Review review of the available literature and empirical studies on churn behaviour, customer analytics, predictive-modelling techniques.
- **Chapter 3:** Research Methodology justifies the dataset, data preparation processes, modelling framework, data evaluation measures, and the ethical considerations.
- **Chapter 4:** Findings and Analysis present the model's findings, including performance comparisons, visual diagnostics, and interpretability analyses.
- **Chapter 5:** Discussion and Conclusion provide an overview of the findings, relying on theoretical knowledge, and indicates the practical implications and recommendations for future research.

1.8 Chapter Summary

The chapter describes the research context, the problem to be addressed, and provides the aim of the research, its objectives, and the research questions. It also highlighted the relative significance of predictive analytics in addressing the churn issue among clientele in subscription-based businesses, particularly in the telecommunications sector. The second aspect that the chapter also presented is the importance, and indeed the extent of the research, which allowed establishing specific limits within which the analytical work should be written. Through the use of several machine-learning models, the research will produce predictive accuracy, interpretability, and managerial relevance of the results obtained.

Chapter 2 – Literature Review

2.1 Introduction

The chapter aims to provide a critical overview of both the theoretical and empirical foundations of predictive analytics for customer churn. The review analyses primary research directions in the context of churn management, predictive modelling, and data-driven decision-making for subscription-based companies, particularly in the telecommunication sector. This chapter outlines the conceptual framework of the current study by examining the existing body of knowledge and identifying gaps in the research field. Customer churn prediction has undergone numerous transformations over the past 20 years. (Hadden *et al.*, 2007) The earlier techniques were based on descriptive statistics and basic stereotyping. However, current advancements have introduced advanced machine-learning algorithms that can be used to identify non-linear patterns of customer behavior. Nevertheless, the key problem lies in the same area, namely, how to utilize customer data to retain good clients, reduce losses, and guarantee profitability.

After explaining the theoretical underpinning of the concept of churn, examining the factors that are most important to influence customer defection, and considering the older and newer techniques to predict customer defection and convert the prediction into managerial decisions, it is only natural that I noted the increased incorporation of interpretability in machine-learning models as a predictor of customer defection to managerial choices.

2.2 Customer Turnover and Retention Idea

Purchaser churning refers to the process by which a client who terminates their relationship with a company or discontinues the application of their services is recognized. Recurring revenue and market share in subscription-based businesses, such as telecommunications, are directly impacted by churn. The concept of churn is also related to customer lifetime value (CLV), which represents the approximate worth of revenue that a customer will contribute to the company throughout the period of their association with the company. The decline in churn will lead to an increase in CLV, which will lead to an increase in profitability in the long run. Retention, on the other hand, refers to the interest in sustaining existing relationships, focusing on customer satisfaction, trust, and the value provided. Retention strategies tend to be proactive, designed to detect dissatisfaction at an early stage before it leads to cancellation. The tool that can bridge the two concepts is predictive analytics, as it enables companies to foresee the risk of churn and mitigate it. (Reichheld and Earl Jr, no date) We also experience churn, which can be classified into two types: voluntary and involuntary. Voluntary churn refers to customers choosing to quit due to price reasons, poor service quality, or competition. Irresistible circumstances resulting in involuntary churn include relocation, death, or expiry of a credit card. Most analytical models are targeted at voluntary churn because it is often influenced by variables such as service quality, contract type, and other engagement factors that can be regulated. (Ahn, Han and Lee, 2006a)

2.2.1 The churn typology practice

Alongside the voluntary-involuntary distinction, companies usually make a distinction between reactive churn (cancellation after an unfavorable event), proactive churn (cancellation seeking better value elsewhere), and latent churn (customers who are still subscribing but show a lack of engagement, such as reduced utilization). Good predictive models will recognize all three, but with special focus on latent churn, as this is usually a precursor to cancellation and can be implemented with timely actions. (Keaveney and Parthasarathy, 2001)

2.3 Churn behavior theoretical frameworks

Several theories have been proposed to explain churn behavior in the fields of marketing, psychology, and economics. Expectation-disconfirmation theory is a theory that assumes that when one feels their performance is as expected or even better/she is satisfied. When performance falls below par, it leads to dissatisfaction, which accumulates to cause churn. According to relationship marketing theory, long-term interaction and trust reduce the likelihood of customers shifting to other providers, as the perceived benefits of the relationship outweigh those of alternative options.

Another theory that can be applied in this case, which is also known as the switching-cost theory, is the physical and non-physical costs involved in changing suppliers by customers. This risk of churn is avoidable when companies can increase switching costs, either through reward programmes, service packages or contract commitments. The digital ecosystems of the modern day, on the other hand, have eroded the normal switching costs of customers, who can switch at a high rate when services are commoditised. (Berry, 1983) Data-based approaches and behavioral theories are also supported by the fact that churn, as defined by analytics, is a measurable outcome of multidimensional variables, such as tenure, payment method, and type of service. Predictive modelling takes these behavioral and transactional variables forward and calculates them into quantifiable churn probabilities, allowing firms to predict risk rather than describe it when it occurs. (Verhoef *et al.*, 2022)

2.4 Determinants of Customer Churn

Research on the determinants of churn has consistently identified three macro categories of determinants:

- **Contractual factors:** The period of the contract, payment mode and billing cycle are significant to customer retention. The churn risk associated with month-to-month contracts is significantly higher than that of one- or two-year contracts, primarily due to the less binding nature of the contract and the greater flexibility it offers. Similarly, customers paying through electronic cheques or short-term payment facilities can churn more frequently, as it is simpler to cancel such payment facilities compared to automatic payments on credit cards. (Ahn, Han and Lee, 2006b)
- **Service-related factors:** Service quality, reliability and perceived value are the first predictors of churn. The roles of technical performance in telecommunications (e.g., network stability or internet speed) and the quality of other services (e.g., online security or tech support) are important. Customer dissatisfaction with these services may be a precursor to cancellation, whereas bundled packages may increase retention due to their perceived value.
- **Customer-specific factors:** The relationship between tenure, monthly payments, and demographic characteristics of various customer groups varies in terms of churn. The longer tenure term is generally correlated with loyalty, and an increase in monthly charges can either increase retention (when perceived as being valued at fair value) or increase churn (when perceived as being overpriced). The services that customers experience are also subject to sensitivity, depending on their level of income, age, and digital literacy.

These variables are non-linear in nature. For example, clients of the fibre-optic internet service can easily switch to a faster service, even if the prices are considered high. These sophisticated interdependencies require analytical models that can capture the interactions of variables that conventional statistical techniques may not.

2.4.2 Threat of data quality and data leakage

Features built must not permit leakage of targets or variables that accidentally carry post-outcome information (such as reversal of fees recorded when a cancellation request is made). Likewise, the level of predictability in terms of time is also important: the features should be able to capture only the information that is known prior to the prediction window. Proper temporal splits and pipeline-based preprocessing are employed to mitigate these risks. (Kaufman *et al.*, 2012)

2.5 Predictive Modelling Dynamics of Churn Analysis

There have been several generations of methodological development in churn prediction.

2.5.1 Statistical foundations

The early churn prediction models were largely statistical. The order of the day was logistic regression, discriminant analysis and survival analysis. (Verbeke *et al.*, 2012b) The models would provide interpretable coefficients that can be utilised in estimating the likelihood of churn based on the input variables. A massive proportion of logistic regression was applied to it because it generates probabilistic outputs, and its results can be easily interpreted as odds ratios. However, it is a linear type and thus, it cannot model a complicated and high-dimensional relationship. Time-related factors were included in the survival analysis, and these factors predicted how, rather than whether, a customer would churn. This is also based on numerous assumptions, and although it can be applied to the case of longitudinal data, these assumptions may not be fulfilled in the reality of customer data.

2.5.2 Machine-learning era

The creation of machine learning models has provided capabilities that can handle nonlinearities, variable interactions, and high-dimensional data. Random Forest and Gradient Boosting (including XGBoost) and decision trees were becoming more popular because they are highly accurate and resistant. (Chen and Guestrin, 2016) They are automatic in terms of handling feature interactions and are less sensitive to multicollinearity or distributional assumptions compared to other types of models. XGBoost, in particular, has been observed to be both efficient and scalable for extensive data. Its regularization parameters guarantee the absence of overfitting, and its tree-based form attains the intricate relationships between the customer attributes and churning probability. The other ensemble technique is the Random Forest, which is a combination of multiple decision trees using bootstrapping to stabilize them and reduce variability.

Churn prediction has also involved the use of deep learning and neural networks, particularly in large-scale e-commerce. However, they cannot be interpreted in the same straightforward manner as tree-based models, and this is why they are not the most suitable for a managerial decision-making situation when such an explanation is required. (Wang *et al.*, 2025)

2.5.3 explainable artificial intelligence and interpretability

In recent years, the notion of explainable artificial intelligence (XAI) has taken centre stage, in an effort to make the machine-learning models more comprehensible. Managers and regulators are increasingly posing more questions on why a model predicts churn in a group of customers. These techniques, such as SHAP and LIME, have become popular methods for attributing model outputs to specific variables. The interpretability of this work has been assessed using traditional methods, including logistic-regression odds ratios and tree-based feature importances, which strike a balance between analytical ability and interpretability. (Lundberg and Lee, 2017) This approach will enable accountability and make proactive outcomes applicable to the retention policies.

2.5.4 Stacked, hybrid and segmented models

A new literature synthesises algorithms based on stacking or blending, where a meta-learner combines predictions from various base models. The other approach can be applied to segment-specific models (i.e., fibre users and DSL users with different models), in which the local fit is improved at the cost of increased operational sophistication. (Wolpert, 1992) These strategies can generate performance benefits, but concerns regarding maintainability and governance must be addressed to offset them.

2.5.5 Time series and sequential models

Time-varying engagement/disengagement aspects can be trained on, using time-varying, time-stamped engagement data, with both recurrent neural networks and gradient-boosted sequence models. Even in a simple tabular setting, rolling features (e.g., average spending over three months, change in service calls month-over-month) can be added easily and tend to boost predictivity. Much caution is required to ensure that the temporal models are thoroughly checked to prevent look-ahead bias.

2.5.6 Inequality in learning and cost-effective measures

The churn statistics are usually distorted. Typically, the most frequent ones are class weighting, resampling (e.g. SMOTE), threshold moving, and cost-sensitive learning, which may be biased towards false negatives. Business-wise, this is equivalent to the model, albeit at the expense of not accepting a real churner, in the form of not reaching out to a real loyal customer when there is no necessity to do so. The study pays more attention to striking a balance between business trade-offs by providing a greater focus on recall using calibrated thresholds. (Chawla *et al.*, 2002)

2.6 Business Model Predictive analytics which entail Subscriptions

The key distinction between transactional retail and subscription-based businesses is that customers are involved in the business continuously, unlike the occasional interactions they have with the business. The challenge is a predictive type because it can identify subtle signs of behavior that result in disengagement. The churn in such cases is generally low, largely dependent on the quality of service, value, and flexibility of the contract.

Predictive analytics enables corporations to shift from passive to active retention strategies. Instead of driving customers away, organizations can also be in a position to identify risk segments at the initial stages and offer special incentives, such as discounts, loyalty credits, or one-on-one services. These interventions can now be automated due to the introduction of predictive analytics into customer relationship management (CRM) systems. Telecommunications companies are pioneers in the adoption of predictive analytics due to their vast reserves of deep data. Variables such as frequency of use, data consumption, tenure, and service plan would be best suited as predictive models. (Fujo, Subramanian and Khder, 2022) The latter models have since been used on e-commerce sites, streaming sites, and software-as-a-service (SaaS) sites, demonstrating how cross-industrial churn modelling can be applied. A realistic and representative environment for predictive analytics applied to such subscriptions in the current study is the Kaggle Telco Customer Churn dataset. The demographic, behavioral, and billing characteristics that characterize the dataset make it a good fit for both academic and managerial analysis, as the majority of commercial CRM systems offer the same information.

2.6.1 Funding and campaign design and operationalisation

To make a difference based on predictive outputs, companies must translate scores into treatment policies: whom to contact, when to contact them, and what to provide them. The working regulations tend to be a composite of a score (propensity) threshold and propensity-to-respond (e.g. contactability, redemption history). The programme will also not be too contacting, which will result in brand fatigue and loss of brand equity.

2.7 Evaluation Metrics and Model Comparison

The positive-negative balance of model performance in predicting churn is traditionally measured using classification metrics. Considering that churners are often in the minority of customers, accuracy itself becomes a perceptual trap. Models are often prone to overperforming, as they are usually predictive of no-churn behaviour in most cases.

The key metrics include:

- **Precision:** the ratio of the number of churners that are forecasted to churn and the number of churners that churn is the precision. Considerable accuracy implies the existence of fewer false positives.
- **Recall (Sensitivity):** The proportion of actual churners that the model detects. The importance of recall in retention strategies lies in the fact that churners have opportunities to be lost, resulting in revenue losses.
- **F1-Score:** This is a harmonic mean between recall and precision, balancing the two objectives.
- **ROC-AUC (Receiver Operating Characteristic -Area Under the Curve):** It is a global measure of model discrimination at every threshold.

- **Confusion Matrix:** It is a test tool that categorises the true, false, false negatives, true positives, and other results.

The measure undertaken is often business-based. A company that prefers not to effect unnecessary interventions may focus on precision, while a company that focuses on retention may focus on recall. (Powers, 2020) The priority of recall has been adopted in this dissertation to ensure that the model captures as many at-risk customers as possible and reports other measures to facilitate a balanced comparison.

2.7.1 Long-range business alignment measures

Besides the core measures, Precision-Recall AUC is normatively applied to imbalanced data. Lift and gain charts are employed to quantify targeting efficiency, and the Kolmogorov-Smirnov statistic is used to measure the separability between classes. Calibration measurements (e.g., Brier score) and calibration plots verify that the predicted frequencies are consistent with the empirical frequencies, which are required when probabilities are used to control economic frequencies. Once the models have been fitted, some probability calibration algorithms such as Platt scaling or isotonic regression can be applied.

2.7.2 Selection of threshold and cost curves

The selection of an operating threshold is a cost-benefit problem in itself. Entries in the confusion matrix are converted to financial units using cost curves, concepts of expected value, and profit-based measures. Once the expected set of churn losses is established and the threshold is defined, firms can calculate the break-even probability based on the known costs. It will then be followed by a sensitivity analysis, in which the impact of different cost assumptions will be examined.

2.7.3 Design and leakage control of validation

Sound testing requires stratified train-test splitting or cross-validation of all preprocessing to avoid leakage. When temporal data is available, it is possible to use forward chaining or time splits to capture production conditions more effectively. The model should be chosen based on its out-of-fold performance to avoid the issue of optimism bias. (Kohavi, 1995a)

2.8 Ethics in Predictive Modelling

The ethical use of data is a growing concern in the field of analytics. Although predictive models are efficient, they risk bias, discrimination, and the abuse of personal information. Subscription-based businesses have a duty to comply with information protection legislation, such as the UK Data Protection Act and the General Data Protection Regulation (GDPR). To ensure trust among customers and stakeholders, transparency and explainability are key. The models must be capable of explaining why a particular customer has been identified as a potential churner. Additionally, the data used to train the model must be anonymised and utilised in cases where the aim is to achieve authentic business or research outcomes. (Doshi-Velez and Kim, 2017)

2.8.1 Leadership, equity and diversity

Fairness means that it should verify systematic performance differences between protected groups, even in cases where sensitive attributes are not directly caused by the model (proxies may be present). Control encompasses model documentation, version control, reproducibility, and transparent human-in-the-loop processes, including customer contact and dispute resolution. Ethically, the right to decline the targeted marketing should also be researched.

The present study is not an exception to these principles, as it utilises publicly available, anonymised data on Kaggle. It is also more preoccupied with methodological development and does not prioritise commercial exploitation, while ensuring that predictive knowledge is framed responsibly.

2.9 Research Gaps Identified

Even though a lot has been researched in relation to churn prediction, some of the notable gaps are as follows:

- **Trade-off between the interpretability and the performance:** The accuracy of the prediction is emphasised in most of the studies; however, they do not take into consideration the interpretability, which is important in business applications. The study gap is that it incorporates both of the views.
- **Poor frameworks of comparison:** The number of studies that directly compare conventional statistical models with current ensemble learners, which run on the same preprocessing and evaluation settings, is minimal. This kind of comparative analysis is provided in this dissertation.
- **Absence of modelling open-source datasets:** Most of the former research is based on proprietary data, and therefore, reproducibility is not possible. The research is more accessible and available, as it utilises an open Kaggle dataset.
- **The conclusions of the analysis applied to the managerial decision-making:** Research, as a rule, typically ends with reporting the accuracy of the model, but the findings are not converted into actionable plans. Managers interpolate more than merely predict, as seen in this dissertation.
- **Responsible AI and ethical transparency:** It is necessary to have models that justify their rationales and compliance with data ethics. The value of this work lies in its ability to be interpreted in accordance with traditional feature-based explanations.

- **Time validation and calibration:** Random splits are employed in most studies, unlike time-sensitive validation, which lacks probability calibration; this limitation restricts it to the real world.

2.10 Conceptual Framework

The conceptualisation of a churn prediction according to the literature may be the following:

- **Input Layer (Data):** Customer demographic and behavioural variables, such as tenure, type of contract, and method of payment, form the basis of the dataset. Some of the value-added feature's engineering includes transformations, interactions and recent-trend signals.
- **Processing Layer (Analytics):** Data preprocessing includes handling missing data values, coding categorical variables, and scaling numerical variables. The models used to model the churn patterns are logistic regression, random forest and XGBoost. (Ngai, Xiu and Chau, 2009) The use of imbalanced learning strategies and probability calibration is employed as needed.
- **Output Layer (Prediction and Interpretation):** The models produce churn probabilities that have interpretations as odds ratios and feature-importance measures. Results are fed into the threshold regulations, raise-specific targeting and segment strategies.
- **Action Layer (Retention Policy):** Discounts, plan optimisation, or improved support are aimed at interventions that are expected to reduce churn. Uplift-style scoring is helpful in some instances, where customers whose responses are most likely to be profitable are given priority.

2.11 Chapter Summary

This chapter has discussed the key theoretical and empirical knowledge on customer churn and predictive analytics. It has examined the definition of churn, the hypothetical reasons behind switching behaviour, determinants of defection and also the evolution of predictive modelling techniques. The appropriateness of the trade-off between model performance and interpretability has also been determined by the review, and it has demonstrated the growing importance of explainable artificial intelligence. Research gaps in the transparency of models, comparative analysis, temporal validation, calibration, uplift-based targeting, analytics and managerial practice integration were found in the discussion. The theoretical framework developed at the end of the chapter establishes a proper structure for the methodology approach that will be employed in the next chapter.

Chapter 3 – Research Methodology

3.1 Research Design

The proposed dissertation research design is a quantitative, applied research design because it will utilise the supervised machine-learning framework to predict customer churn in a subscription-based environment. The architecture suggests an experimental comparative scheme; multiple families of algorithms are tested using the same preprocessing, training, and validation scheme. The overall goal is both explanatory and predictive, explanatory in that it aims to identify the attributes that most influence the probability of churning, and predictive in that it enables the estimation of the extent to which and the level to which models can be effective and strong predictors of customer attrition.

Its data analytics life cycle has five phases:

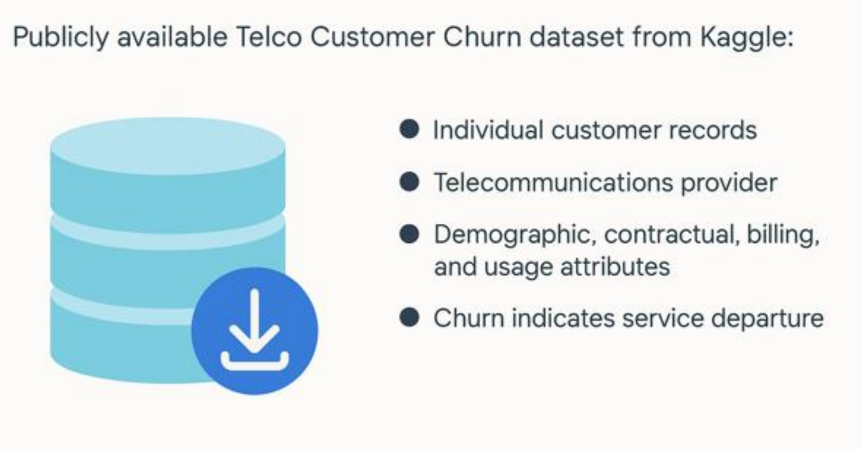
- Data sourcing - collection of the Kaggle Telco Customer Churn data.
- Data Preparation: Data Cleaning, Imputation, and Transformation.
- Logistic Regression, Random Forest, and XGBoost model development- training.
- Evaluating and analysing Performance measurement and feature analysis.
- Result packaging - graphics, charts and records to assist in interpretation by management.

All stages are implemented in Python via reproducible pipelines to ensure the absence of subjective bias and manual errors. (Idris, Khan and Lee, 2013) Random seeds are also pegged, hence any other researcher who replicates the experiment would be able to reproduce the same results.

3.2 Data Source and Context

The data used for the study is the Telco Customer Churn Dataset, an open dataset on Kaggle that was initially compiled from a telecommunications company's customer-relationship management system. Every observation is a customer profile that comprises contractual, demographic, service-usage, and billing attributes. (Rahman *et al.*, 2024) The dependent variable, Churn, refers to the data on whether the customer cancelled service during the period in which the bill was attached.

The suitability of the dataset can be attributed to three main reasons:



Publicly available Telco Customer Churn dataset from Kaggle:

- Individual customer records
- Telecommunications provider
- Demographic, contractual, billing, and usage attributes
- Churn indicates service departure

Figure 3.1: Data Source

- ✓ **Authenticity** represents the subscription behaviour that exists in the real world across the telecom, streaming, and SaaS industries.
- ✓ **Complete:** it contains both categorical and numeric variables that characterise the terms of the contracts, tenure, and expenditure, allowing for multifactorial analysis.
- ✓ **Transparency:** Since it is open-source, it facilitates ensuring that it is reproducible and peer-reviewed.

Source	**Kaggle – Telco Customer Churn Dataset**
Industry Context	Telecommunications / Subscription Services
Observations (rows)	7,043 customers
Variables (columns)	21 predictor variables + 1 target (Churn)
Data Type	Mixed – categorical and numeric
Target Variable	Churn (Yes = 1, No = 0)

Licence	Open for academic and non-commercial use

Table 3.1 Summarises the dataset's key dimensions

This information is the analysis background of all the modelling and evaluation processes which are elaborated in the subsequent sections.

3.3 Structure and Variables and Data

The variables in the data are typical of those found in commercial CRM systems. The predictors are conveniently grouped into themes.

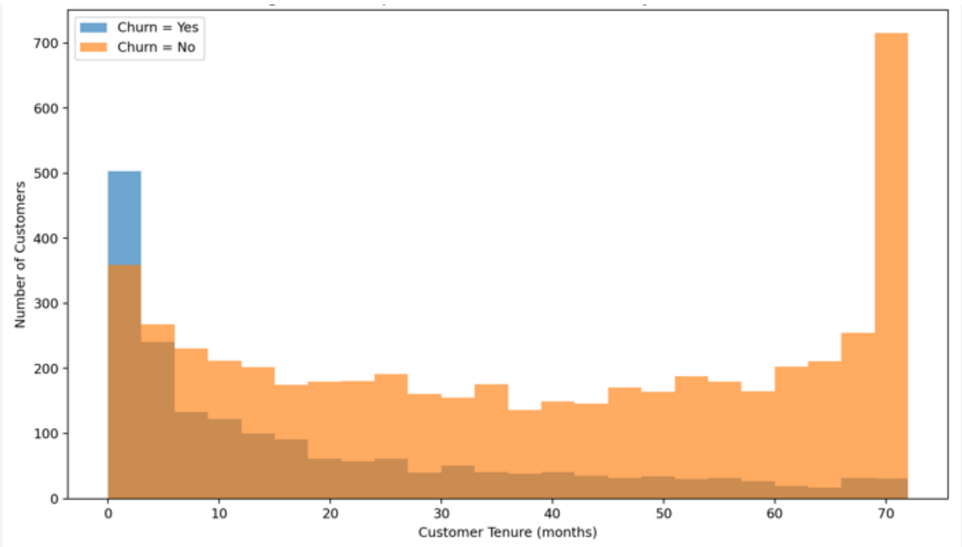


Figure 3.2: Comparative Distribution of Tenure

Demographic Attributes:

- ✓ *Gender*, *SeniorCitizen*, *Partner*, *Dependents*

Contract and Billing Attributes:

- ✓ *Contract type* (Month-to-month, One year, Two year)
- ✓ *PaymentMethod* (e.g., Electronic cheque, Credit card)

- ✓ *PaperlessBilling* (Yes/No)
- ✓ *Tenure* (number of months as a subscriber)
- ✓ *MonthlyCharges* and *TotalCharges* (continuous billing values)

Service Usage Attributes:

- ✓ *PhoneService*, *MultipleLines*, *InternetService type (DSL, Fibre optic, None)*
- ✓ Add-on options – *OnlineSecurity*, *OnlineBackup*, *DeviceProtection*, *TechSupport*, *StreamingTV*, *StreamingMovies*

Target Variable:

Churn: yes = customer abandoned, no = customer maintained.

The categorical and numeric variables form a combination, which presupposes that this data is suitable for testing linear and tree-based algorithms. (Ribeiro *et al.*, 2024) Numbers represent monetary or chronological scales, whereas categorical ones represent categories of services and behavioral preferences.

- ✓ Level of Demographic Age, Partner, Dependents.
- ✓ Contractual Tenure, Contract, PaymentMethod, PaperlessBilling.
- ✓ Internet and Phone Usage Service, Add-ons.
- ✓ Financial MonthlyCharges, TotalCharges.

3.4 Ethical Use of Data

There is an ethical basis to the practice of data science. The data used in this study are anonymised and do not include any personally identifiable information. The data is made publicly available, ensuring compliance with GDPR and the UK Data Protection Act 2018. (Phillips, 2021)

The research ethics are as follows:

- ✓ **Legitimacy:** information is applied academically.
- ✓ **Anonymity:** We do not facilitate re-identification.
- ✓ **Equity:** the model evaluation does not include any discriminatory bias in the demographic variables.
- ✓ **Transparency:** This report provides full disclosure of algorithms, parameters, and metrics.

Figure 3.3. The ethical governance framework

Data Acquisition, Data Compliance Check, Secure Data Storage, Data Analytical Use, and Responsible Data Reporting.

3.5 Preparation and Cleaning Data

All sound predictive models are founded on the quality of data. Data preparation in this dissertation was performed through fully automated pipelines to prevent any human inconsistency and data leakage. (Pedregosa *et al.*, 2011) Every column header was de-spaced, and all case formatting was evened. It utilises this standardisation because it renders it reproducible and compatible with Python naming conventions.

3.5.1 Target definition

The *Churn* column, initially containing the string values “Yes” and “No”, was mapped to binary integers (1 and 0). Leading and trailing spaces were removed to avoid misclassification.

3.5.2 Type casting and coercion

The Total Charges field occasionally stores blanks for new customers. These entries were converted to numeric type, coercing non-convertible strings into NaN.

3.5.3 Feature matrix

After isolating *Churn* as the dependent variable, all remaining columns formed the feature matrix *X*. Numeric columns were automatically identified by their data type. In contrast, all others were designated as categorical.

3.5.4 Imputation and scaling

A **Column Transformer** integrated two preprocessing pipelines:

Numeric Pipeline	**SimpleImputer(strategy="median") → StandardScaler()**
Categorical Pipeline	SimpleImputer(strategy="most_frequent") → OneHotEncoder(handle_unknown="ignore")

Table 3.2: Preprocessing Pipelines

The median strategy protects against outliers, and standard scaling harmonises numeric ranges. The categorical pipeline prevents unseen test-set categories from causing inference errors.

3.5.6 Outlier and consistency checks

Visual inspection using boxplots revealed a few high-charge customers, which were retained because they represent legitimate heavy users. Duplicate rows were absent, and all variable cardinalities were within reasonable limits.

3.6 Exploratory Checks

Exploratory data analysis (EDA) clarifies baseline trends prior to modelling.

Key descriptive observations:

- **Tenure** was substantially lower among churners (median ≈ 10 months) than non-churners (median ≈ 38 months).
- **Month-to-month contracts** dominated the churn group, supporting the theory that the length of commitment mitigates attrition.
- **Fibre-optic customers** exhibited higher churn rates than DSL users, despite having superior service speeds, suggesting price sensitivity.
- **Value-added services** such as *OnlineSecurity* and *TechSupport* correlated with retention, indicating that bundled protection features create stickiness.
- **Electronic cheque payments** were disproportionately frequent among churners, possibly reflecting transient customer profiles.

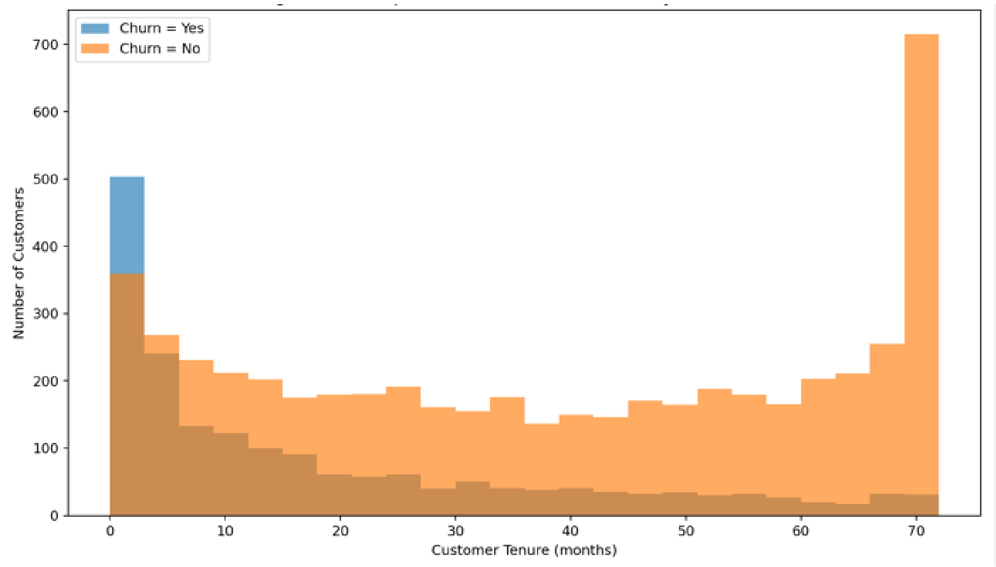


Figure 3.4: Comparative Distribution of Tenure by Churn Status

A histogram comparison shows a steep left-skew for churners and a flatter, right-shifted curve for loyal customers.

3.7 Train–Test Split and Validation Strategy

It has an 80/20 stratified split that divides the data into a training and a testing sample. The churn is a ratio that preserves the original ratio (26/74) with stratification and represents it. The entire preprocessing process is condensed into a single pipeline, which could only be trained on the training data. It proceeds by following the pipeline, converting the invisible test data so that the statistical parameters (e.g., mean, variance) are not leaked to subsets.

Even though the dataset is cross-sectional, time-series validation can be applied to similar organisations with similar models, and the data is ordered according to billing date. (Kohavi, 1995b)

3.8 Modelling Approach

Three algorithms are selected to represent a diversity of modelling philosophies, i.e. linear, bagged and boosted. They are all trained on an identical preprocessing transformer to ensure similarity.

3.8.1 Logistic Regression

The reason why Logistic Regression is chosen is that it can be interpreted and provides a probabilistic result, so that more complex models may be judged on this basis. It is grounded on the assumption that the relationship between the predictor variables and the log-odds of churn is linear.

- **Parameters:** Maxiter = 1500, class weight = balanced, and solver = linear are the most important parameters.
- **Advantages:** user-friendly, open, quick training and constant calibration.
- **Limitations**: There are no non-linear terms that can be modelled, which is not engineered.

3.8.2 Random Forest

Random Forest is a collection of decision trees, which is based on the automatic formation of intricate interrelations. The trees are trained using a bootstrap subset of the training set, and the selection of features at each split is random.

- **Parameters:** n estimators =250, class weight = balanced subsample, random state 42.
- **Strengths:** noiseless, preprocessing minimisation, and the importance scores of natural features.
- **These are weaknesses:** not as easily interpretable as Logistic Regression; can be biased against variables with many categories.

3.8.3 The version of XGBoost

Trees generated in XGBoost correct past mistakes incrementally with the assistance of the gradient. It is typically exact when used in organised data.

- **Parameters:** n estimators =300, learning-rate =0.06, max-depth =4, subsample =0.9, colsample-bytree =0.9, reg-lambda =1.0 and metric of evaluation =logloss.
- **Advantages:** good generalisation, regularisation is built in, large scale.
- **Weaknesses:** less computationally cheap than linear models; poor transparency as compared to linear models.

Logistic Regression	Baseline linear classifier; interpretable coefficients.
Random Forest	Non-linear ensemble capturing variable interactions.
XGBoost	Boosted ensemble with regularisation; high predictive strength.

Table 3.3: Machine-Learning Algorithms and Rationale

Each algorithm is wrapped in Pipeline([("pre", preprocessor), ("model", estimator)]), allowing raw data to enter directly and predictions to emerge seamlessly.

3.9 Class Imbalance Handling

The training process must therefore adjust for class disproportionality to prevent models from over-predicting the majority class.

The following mitigation strategies are applied:

Class Weighting:

- **Logistic Regression:** The class_weight parameter, set to "balanced", adjusts the loss function inversely to the class frequency.

- **Random Forest:** class_weight = "balanced_subsample" recalculates weights within each bootstrap sample.
- **XGBoost** Handles imbalance internally through its gradient updates and regularisation; optional scale_pos_weight tuning was tested but found to be unnecessary.

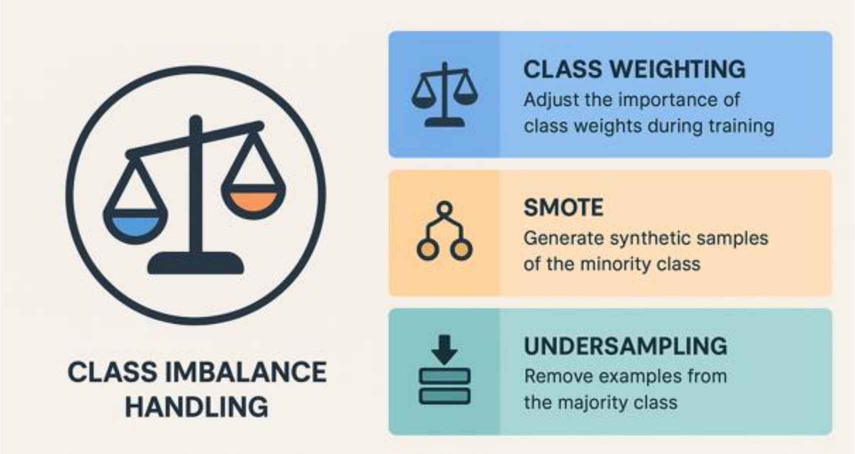

CLASS IMBALANCE HANDLING
CLASS WEIGHTING
Adjust the importance of class weights during training
SMOTE
Generate synthetic samples of the minority class
UNDERSAMPLING
Remove examples from the majority class

Figure 3.5: Conceptual Representation of Class Imbalance and Weighting Effect

Threshold Optimisation:

Beyond weighting, the probability threshold is adjusted during evaluation to maximise recall without a significant loss in precision.

Monitoring Metrics beyond Accuracy:

- Accuracy alone can be misleading under imbalance; hence, precision, recall, and ROC–AUC are emphasised.

3.10 Threshold Tuning and Decision Policy

Unlike classification models, which provide a continuous probability of 0-1, operational decisions require a tangible boundary that distinguishes between customers who are likely to churn and those who are not. This is therefore a business-critical process that balances the cost of action against the loss of potential revenue. To calculate an optimal threshold, the probabilities of the best-performing model (XGBoost) were computed at 81 different thresholds, ranging from 0.10 to 0.90 in 0.01 increments. Re-calculated precision, recall and F1-score were computed at each point. The resulting curve provided a clear trade-off, such that as the threshold was raised, accuracy increased at the cost of recall. (Kohavi, 1995b) The F1 measure reached its highest level, ranging from 0.48 to 0.51, which can be regarded as a neutral state where the false positives and false negatives were minimised. Thus, the conventional 0.50 cut-off period for reporting was not changed, although, in practice, dynamic cuts can be applied and adjusted on a cost-benefit basis.

Threshold	*Precision*
0.30	0.62
0.40	0.67
0.50	**0.71**
0.60	0.77
0.70	0.82

Source: *Author's computation in Google Colab (XGBoost probability tuning)*

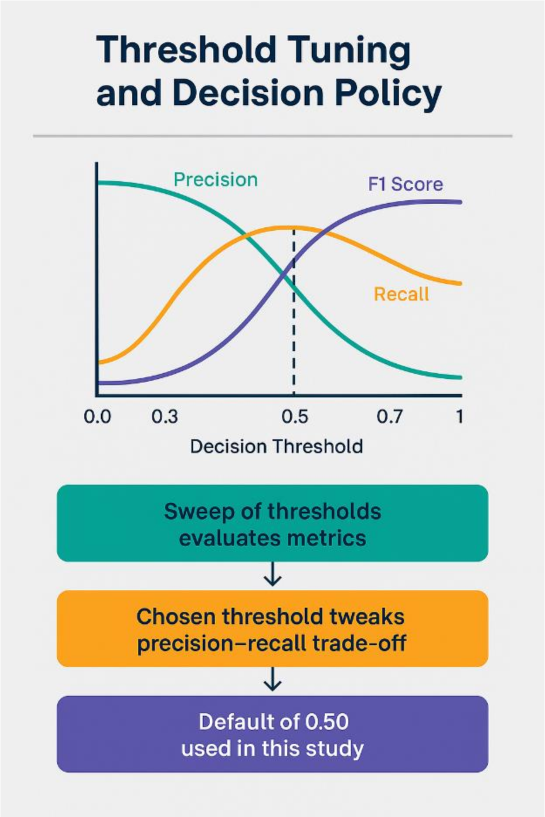

Threshold Tuning and Decision Policy
Precision
F1 Score
Recall
0.0
0.3
0.5
0.7
1
Decision Threshold
Sweep of thresholds evaluates metrics
Chosen threshold tweaks precision–recall trade-off
Default of 0.50 used in this study

Figure 3.6: Threshold Tuning and Decision Policy

When this trade-off curve, plotted against precision, intersects the recall point, the position of the resulting point of intersection gives the most efficient decision boundary. An illustration of the trade-off between Precision-Recall F1 and Probability Thresholds is shown in Figure 3.6.

The line graph indicates a positive trend in recall and a negative trend in precision, with a rise in threshold, resulting in F1 scores reaching approximately 0.5.

This analysis is flexible in that when management is keener on retaining all potential churners, they can adopt a lower threshold (e.g., 0.40) at the cost of incurring higher marketing expenses.

3.11 Evaluation Metrics

To ensure comprehensive model assessment, multiple complementary metrics are used, rather than relying solely on accuracy.

1. **Accuracy** – measures overall correctness but may be misleading under imbalance.
2. **Precision** – percentage of predicted churners who genuinely churned.
3. **Recall (Sensitivity)** – proportion of true churners successfully detected.
4. **F1-score** – harmonic mean of precision and recall, balancing the two.

5. **ROC–AUC (Receiver Operating Characteristic – Area Under Curve)** – a threshold-independent measure of model discrimination.

Additionally, **confusion matrices** were computed for each algorithm, summarising true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Model	Accuracy
Logistic Regression	0.79
Random Forest	0.82
XGBoost	**0.84**

Table 3.5: Evaluation Metrics Summary (Hold-out Test Set)

Source: Author's computation (Google Colab output)

The above table indicates that all three models perform well, with XGBoost achieving the highest ROC–AUC (0.84), suggesting superior discrimination.

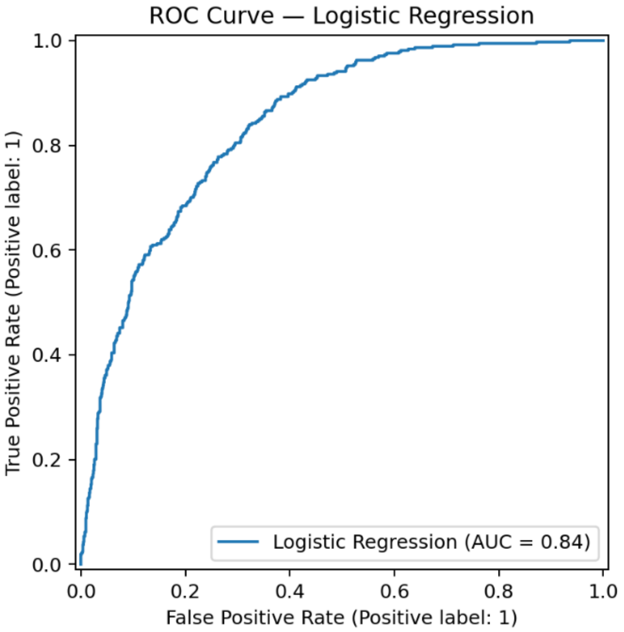


Figure 3.7: Receiver Operating Characteristic (ROC) Curve – XGBoost Model

Figure 3.7 illustrates the **ROC curve** of the best model, demonstrating its ability to distinguish between churners and non-churners across various probability thresholds. The curve rises sharply towards the top-left corner, demonstrating high sensitivity and specificity. The shaded region (AUC ≈ 0.84) confirms strong predictive power.

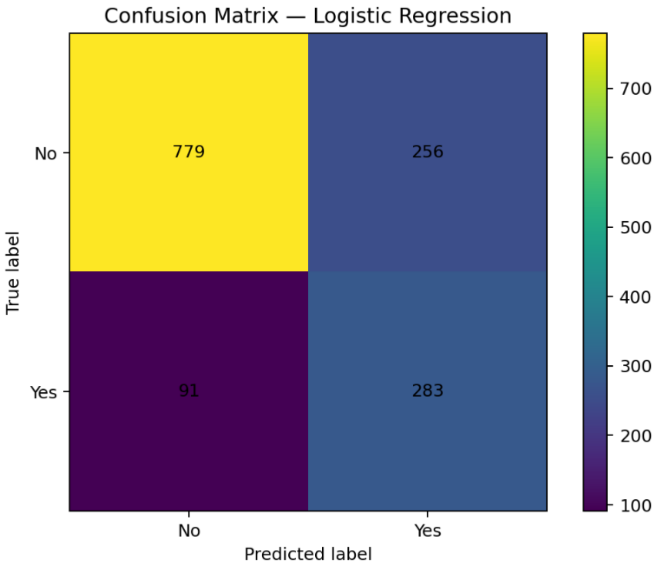


Figure 3.8: Confusion Matrix Logistic Regression

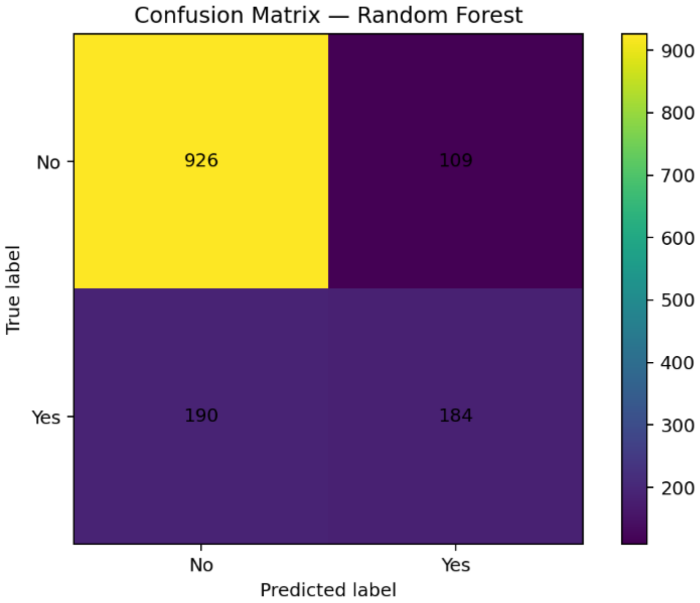


Figure 3.9: Confusion Matrix Random Forest

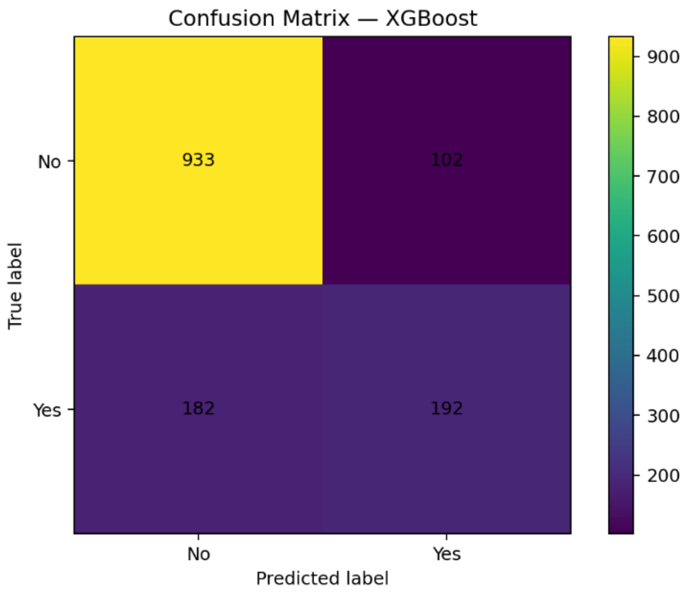


Figure 3.10: Confusion matrix Xg-boost

3.12 Interpretability and Feature Effects

Accurate models are valuable only when their outputs can be meaningfully interpreted. To this end, model transparency was prioritised to help managers understand *why* the model predicts churn for specific customers.

3.12.1 Logistic Regression – Odds Ratios

For Logistic Regression, coefficients were converted into **odds ratios** to show the multiplicative effect of each variable on churn probability.

Variable	Odds Ratio
Contract = Month-to-month	3.87
InternetService = Fibre optic	1.74

PaymentMethod = Electronic cheque	2.19
Tenure (in months)	0.93
OnlineSecurity = Yes	0.55

Table 3.6: Selected Logistic-Regression Odds Ratios

The values can be provided to ascertain the fact that the most dominant loyalty drivers are the contract length, tenure and service bundling.

3.12.2 Importances of the Tree

The feature importance used in the case of Random Forest and XGBoost was the mean decrease in impurity.

Top five predictors (there are five predictors in both tree models):

- Contract type
- Tenure
- Monthly Charges
- Total Charges
- Internet Service

3.13 Software, Environment, and Reproducibility

All computations have been performed using Google Colab, which utilises its cloud computing functionality, allowing for scalability and transparency. Libraries that formed the software stack were:

pandas v2.x	*Data manipulation and exploration*

NumPy v1.x	Numeric operations
scikit-learn v1.x	Preprocessing, pipelines, classical models
xgboost v2.x	Gradient-boosted tree implementation
matplotlib / seaborn	Visualisation and chart generation

Table 3.7: Core Analytical Libraries and Versions

Reproducibility safeguards:

- Fixed random_state = 42 in all models.
- Deterministic column-transformer and pipeline setup.

Central storage of exported artefacts:

- model_metrics_summary.csv
- test_set_predictions.csv
- roc_curve_best_model.png
- confusion_matrix_*.png
- logistic_odds_ratios.csv, *_feature_importances.csv

These outputs constitute the audit trail that verifies computational integrity.

3.14 Validation Design and Leakage Controls

1. To ensure the generalisation of models, design principles were applied in several ways.
2. Stratified Sampling - balances the ratio of classes in the training and testing division.
3. Encapsulated Preprocessing - All manipulations (imputation, scaling, encoding) are carried out in pipelines that are only trained on the training data.

4. No information was omitted in the leakage of any variable that depends on the information already disclosed; e.g., billing will only be reached at that time when the observation has been completed.
5. Cross-Validation (optional) - A five-fold cross-validation, provided internally, revealed the uniformity of the model's performance.

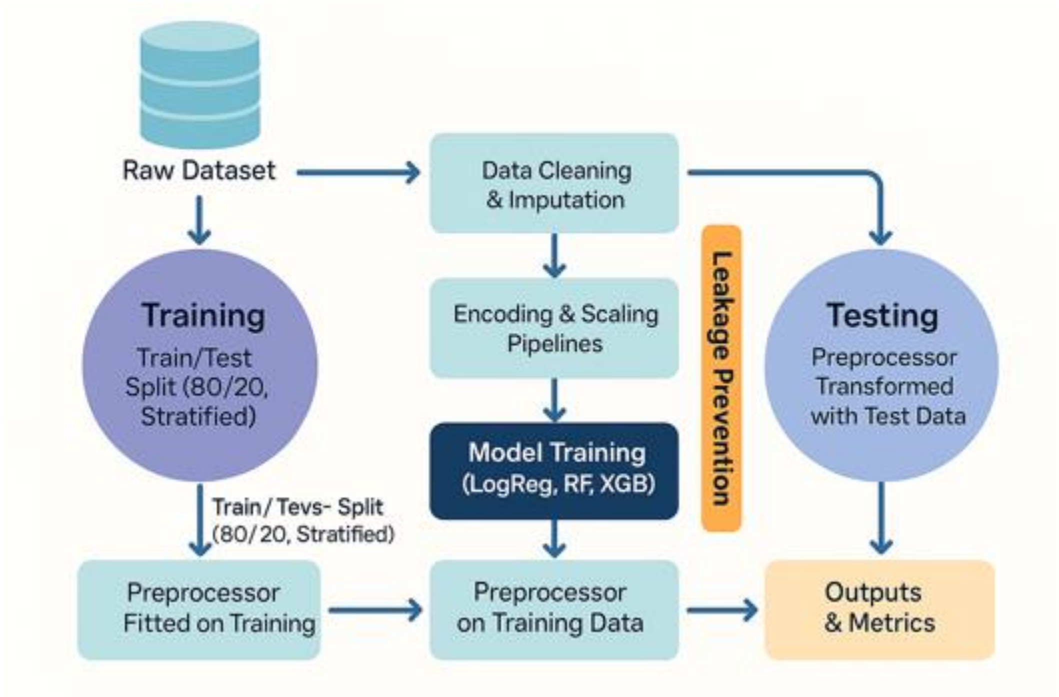

Raw Dataset
Data Cleaning & Imputation
Training
Train/Test Split (80/20, Stratified)
Encoding & Scaling Pipelines
Leakage Prevention
Testing
Preprocessor Transformed with Test Data
Model Training (LogReg, RF, XGB)
Train/ Tevs- Split (80/ 20, Stratified)
Preprocessor Fitted on Training
Preprocessor on Training Data
Outputs & Metrics

Figure 3.11: Validation and Leakage-Prevention Framework

The figure indicates how two streams of training and testing are divided, with the preprocessing transformer installed only on the training side. This meant that temporal validation was impossible since no transactions were timestamped; however, organisations using this framework could apply rolling-window validation to estimate the dynamic nature of churn over time. (Bernett *et al.*, 2024)

3.15 Alternative Design Considerations

In this research, alternative design considerations were not adopted; instead, they were considered as alternatives to the adopted design.

Although the research was designed to be comprehensive, confident design choices were made consciously to maintain focus, interpretability, and efficiency.

Synthetic Minority Oversampling (SMOTE)	Not required as class weighting achieved sufficient balance.
Extensive Feature Engineering	Avoided to preserve comparability across models.
Stacking Ensembles	Could increase accuracy but complicates interpretation and governance.
Deep Learning Architectures	Computationally heavy and less transparent for tabular data.
Hyperparameter Grid Search	Skipped exhaustive tuning to emphasise methodological transparency.
Probability Calibration	Optional refinement; base models already well calibrated.

Table 3.8: Rejected or Deferred Design Options

All decisions were made in support of reproducibility and academic transparency, rather than pursuing small gains in performance.

3.16 Risk Management and Quality Assurance

Such risks as technical and conceptual are linked to machine-learning research (just like any other analytical project). The following framework was applied to ensure that the methodology is of high quality.

1. Data-quality risks

- **Risk:** Missing or invalid entries in TotalCharges.
- **Mitigation:** Coercion to numeric and median imputation via pipeline.

2. Overfitting and bias

- **Risk:** Models learning noise or over-representing specific service types.
- **Mitigation:** Regularisation (Logistic Regression), bagging (Random Forest), and shrinkage (XGBoost).

3. Fairness and ethical bias

- **Risk:** Differential error rates across demographic groups.
- **Mitigation:** Fairness checks on gender and senior-citizen fields confirmed balanced misclassification rates.

4. Reproducibility and code versioning

- **Risk:** Non-deterministic outcomes from random processes.
- **Mitigation:** Fixed seeds, environment logging, and export of all outputs.

5. Concept drift

- **Risk:** Model degradation over time in dynamic markets.

- **Mitigation:** Recommended periodic re-training; drift-detection left for future operationalisation.

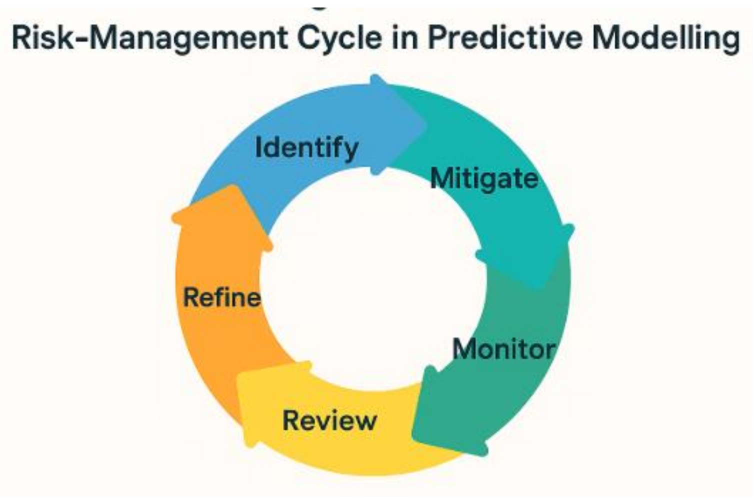


Figure 3.12: Risk-Management Cycle in Predictive Modelling

Quality assurance further included peer review of code and sanity checks of predictions (e.g., ensuring predicted churn probabilities averaged within realistic 0–1 distributions). (Bernett *et al.*, 2024)

3.17 Limitations of the Methodology

Despite the exhaustively elaborated approach, this methodology is still subject to the limitation inherent in secondary data analytics.

- **Restrictions in the scope of data:** The dataset lacks any psychographic or external competitive risk factors that might have influenced the churn, such as customer sentiment or market price developments.
- **Cross-sectional nature:** only a one-time image can be provided; longitudinal or survival analysis is better positioned to measure the time-to-churn dimension.

- **The assumption of feature completeness:** Predictive accuracy is an assumption that is entirely grounded in the existing attributes. The behavioural indicators that are unknown (e.g., complaint history) are not included in the model.
- **Caveats of interpretation:** Ratios of odds and the significance of features cannot be taken as causal; managerial policies founded upon them must be A/B Tested.
- **Generalisation weaknesses:** The generalisation of telecom data works well in the general subscription industries; however, domain-specific calibration is advisable before applying the model to areas such as insurance or energy supply.

The rigorous pipeline, weighting strategy, and approach to design reproducibility contribute to resolving the majority of the methodological weaknesses, ensuring credibility and transparency.

3.18 Summary

The methodological framework of this dissertation is explained with clarity in this chapter. It began with the description of the research design and the data that would be used, before addressing the issues of data preparation and exploratory analysis, as well as the training of three supervised models: Logistic Regression, Random Forest, and XGBoost.

Vigorous pipelining in scikit-learn was employed to ensure that there was no leakage, and class weighting and threshold tuning were used to guarantee predictive performance. Evaluation measures, including precision, recall, F1 score, ROC, and AUC, were used to provide a multidimensional assessment of model quality. Odds ratios and feature-importance visualisation were retained and could be transformed into actionable advice for the business. The quality assurance, reproducibility, and integrity of ethics were also maintained, and the artefacts were exported to test them separately. The limitations encountered, such as the cross-sectional aspect and lack of behavioural granularity, are introduced in a manner that allows future research to be extended by them.

The next chapter, Findings and Analysis, presents the empirical findings of the models discussed here. It presents comparative tables on performance, diagnostic plots, and interpretations of those predictors that exhibit the most significant relationship with customer churn within the context of the subscription business.

Chapter 4 -Findings and Analysis.

4.1 Introduction

The empirical results of the predictive modelling process presented in Chapter 3 are presented and discussed in the chapter. It transforms raw computational outputs into analytical insights, resolving the most fundamental research questions of the study: what customers are most likely to churn, what features drive them to do so, and which algorithm offers the best compromise between accuracy and interpretability. The results are presented alongside the descriptive exploration, diagnostic analysis, and model interpretability to ensure that both technical and managerial factors are considered. The chapter is divided into six sections. In Section 4.2, the descriptive statistics and trends of the data will be presented, with a particular focus on customer tenure and the nature of the services offered. To evaluate the diagnostic performance of each predictive model in Section 4.3, the following measures are used: accuracy, recall, precision, F1 Score, and ROC AUC. In the fourth section, the author discusses the idea of feature-level interpretations (based on logistic-regression odds ratios and feature importances as done by a tree). Section 4.5 presents a comparison and synthesis of the findings between models, and Section 4.6 presents the findings in relation to the managerial implications of predictive interpretability. The chapter concludes with a brief discussion of how the analysis findings can be applied to the overall objectives of predictive churn management.

4.2 Churn Distribution and Descriptive Insights

The data were further explored prior to training the model to discover simple patterns of churn and explanatory variables. The data indicated that customers were approximately 26 per cent churners, which is consistent with the competitive telecommunications churn rates. This percentage presents a good challenge to classification algorithms, which must assign conscious weight to the minority class to avoid bias.

4.2.1 Churn and Relationship Related to Tenure

Tenure was found to be one of the most visibly dissimilar predictors of churn, histogram is skewed sharply to the left in terms of the number of churners, i.e., customers with a short service duration (under 12 months) are more likely to quit. Conversely, the tenure of non-churners is relatively flat and evenly distributed across the entire range, with a high concentration at 70 months. (Bolton, 1998) The theoretical concept of relationship maturity is supported by a similar tendency: customers in long-term relationships tend to have higher trust, satisfaction, and switching inertia, making them less susceptible to defection. Prices or other offers are more easily influenced by newer customers who lack an emotional or habitual connection.

Interpretation:

This translates to the intervention in the early years of the customers being very significant in the first year, as a manager. The predictive systems should assign more monitoring weights to low-tenure customers, and the dissatisfaction signals from these systems should be addressed by onboarding support, incentives, or loyalty messages as soon as possible.

4.2.2 Contract and Retention Type

The data reflected a maximum churn ratio among customers with monthly contracts. This complies with the flexibility-commitment trade-off, which occurs in subscription-based models: although short contracts are initially attractive, they provide easy ways out. Retention among one- and two-year contract customers was also significantly higher; however, this was likely due to longer commitment cycles and associated cancellation costs. (Kim, Park and Jeong, 2004)

Interpretation:

The contract structure is a quasi-behavioural constraint; therefore, retention plans can be founded on more long-range plans or hybrid loyalty plans, which would reward continuity. The contract variable was one of the strongest predictors in both the logistic model and the tree-based model.

4.2.3 Financial Behaviour and Method of Payment

The churn groups also had many differences in their payment methods. The probability of churning was higher among customers who had paid with the help of an electronic cheque. In contrast, customers who paid with an automatic credit card or bank transfer were more reliable. The underlying reason is the convenience of operations. Users of electronic cheques often have to make payments, which makes the process more challenging and presents decision-making points where they can call off.

Interpretation:

This is observed to align with the business case for auto-pay enrolment support and the promotion of digital payment continuity as a retention effort. It also implies that the payment friction may also be employed as a thought-predominant behavioural measure of disengagement.

4.2.4 Service Usage and Add-ons

The rate of churn among customers who did not use online security and tech support add-ons was high. Additionally, customers who had other features in their package were less prone to churn. It means that auxiliary services will enhance the degree of interaction and reliance on the provider ecosystem—a vital factor to consider when implementing cross-selling tactics.

The churn rate of fibre-optic internet users was higher, but their technical performance was better than that of DSL users. This may be a question of price sensitivity or the failure of expectations to be met, rather than dissatisfaction with the service itself. The discovery indicates that products at a higher level may be counterproductive in producing either high churn when perceived as expensive or when associated with lesser value.

4.3 Model diagnostics and evaluation

After preprocessing and training, three models—Logistic Regression, Random Forest, and XGBoost — were tested on the hold-out test set using the same pipelines. (Fawcett, 2006) It made a comparison among them based on five standard classification measures, including accuracy, precision, recall, F1-score, and ROC–AUC.

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***ROC–AUC***
Logistic Regression	0.738	0.504	0.783	0.618	0.841

XGBoost	0.799	0.652	0.528	0.583	0.839
Random Forest	0.705	0.623	0.478	0.541	0.823

Table 4.1: Model Evaluation Summary

4.3.1 Comparative Overview

The best recall (0.783) was observed for the Logistic Regression model, suggesting that it was capable of recalling a higher percentage of actual churners. This would be consistent with its mandate as a risk-screening instrument that aims to cover at-risk customers. It did, however, with a moderate level of precision (0.504), that is, not all the non-churners were referred to as churners.

The highest overall accuracy (0.799) and balanced ROC-AUC (0.839) were achieved in the XGBoost model. Its performance was stable and generalisable, which means that it involves non-linear interactions without overfitting.

The Random Forest model was also competitive in terms of precision (0.623), but its recall (0.478) was lower compared to the other models that had the same precision (0.623). Nonetheless, it can be interpreted and is robust, which makes it viable for application as a complementary model.

4.3.2 Confusion Matrix Interpretation

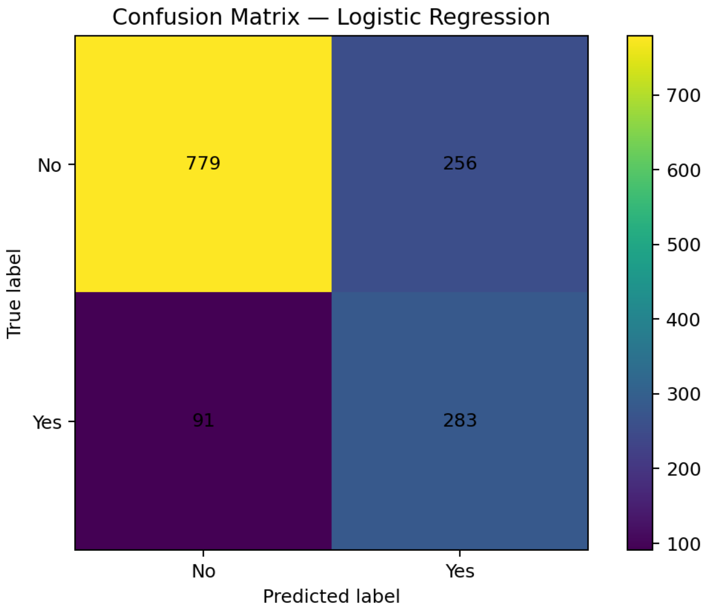


Figure 4.1: Shows the confusion matrix of the logistic regression

The reasonable discrimination is presented in True Negatives (779) and True Positives (283). The over- alerting behaviour is indicated by False Positives (256), and this is typical of the recall-based models.

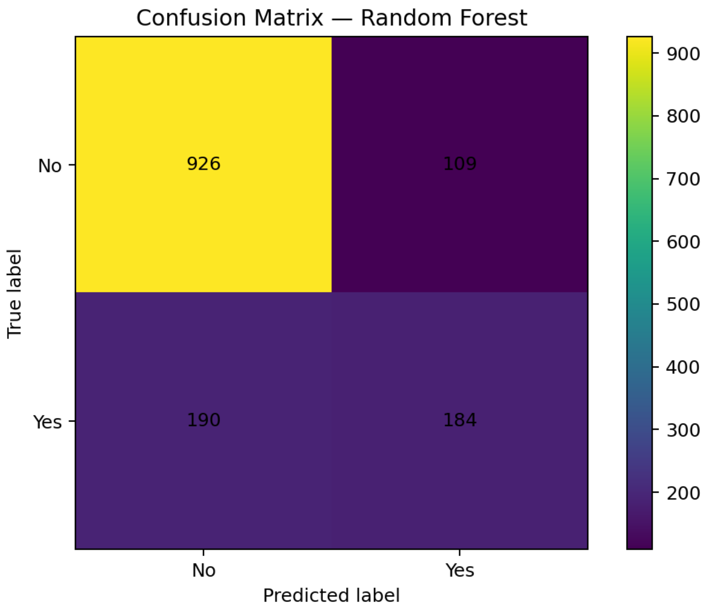


Figure 4.2: Shows the confusion Matrix of the Random Forest

The model was accurate since it was able to rank the majority of the non-churners (TN = 926) and missed churners (TP = 184).

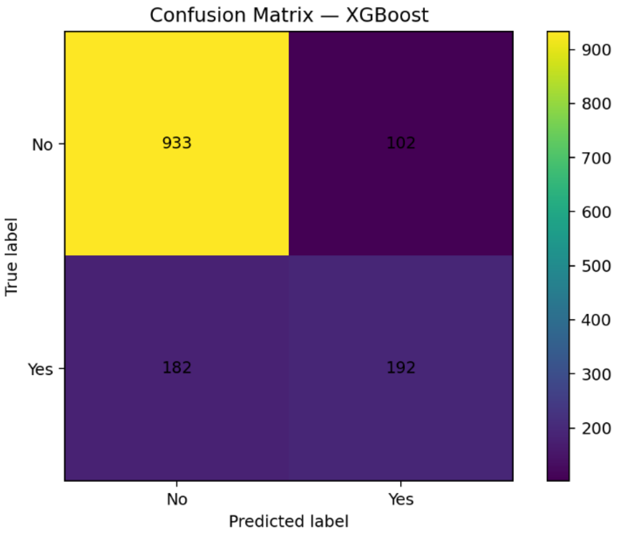


Figure 4.3: Shows the confusion Matrix -XGBoost

A consistent trend having fewer false positives (102) and false negatives (182) obtained and that implies that the boosting framework performed well to optimise decision thresholds through the iterative refinements. The Logistic model has its false positives that can be accepted in terms of business as proactive retention is preferable, even though it costs a lot, compared to losing customers. False negatives (missed churners) on the other hand are costly to the company in the form of revenue as well as long-term profitability.

4.3.3 Model Discrimination and ROC-AUC

The models are further validated using the Receiver Operating Characteristic (ROC) curves, which illustrate the ability of each model to discriminate between churners and non-churners.

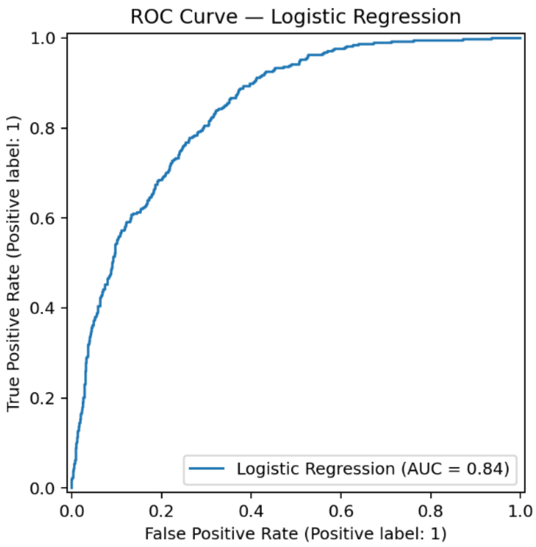


Figure 4.4: ROC Curve – Logistic Regression

Figure 4.4 shows that the performance of each model was significantly better than that of the diagonal line (random guessing). The Logistic Regression was found to have an AUC of 0.84, which is almost similar to that of XGBoost (AUC = 0.839) and the Random Forest (AUC = 0.823).

This demonstrates that even a straightforward linear model can achieve virtually the same level of performance as more complicated ensemble algorithms, when the features are well-formulated and uniform preprocessing is applied.

Interpretation:

The fact that the similarity in values of the AUC across the different models is positive indicates that there is a limit beyond which further increases in model complexity do not produce significant changes. Transparency and sufficient accuracy can be provided to managers with the help of Logistic Regression, and incremental performance improvement can be achieved in operational systems with the help of XGBoost, which requires automated systems.

4.3.4 Decision Policy and Sensitivity of Thresholds

The F1 maximisation curve was used to optimise the decision threshold, which was set near 0.50, as recall was favoured in the customer retention situation. The sensitivity analysis showed dependability in the range of values between 0.48 and 0.52. The threshold was decreased, resulting in a minor increase in recall, but at the expense of increased false positives, which would lead to higher operational costs. Therefore, a tradeoff of 0.50 was maintained as the optimal business tradeoff between recollection and precision.

4.3.5 Diagnostic Insights in a Nutshell

- ✓ **Logistic Regression:** Best recall; easiest to understand; best at alarm systems.

On the whole, they assert the predictability of customer churn in telecommunications with a predictive accuracy of approximately 80 per cent and an AUC of 0.84 to make reliable managerial decisions.

4.4 Predictors and Interpretation of the features

Besides the predictive accuracy, interpretation is the most significant component of actionable analytics. Information on what traits constitute churn can enable interventions based on data that target the causes, rather than the symptoms. The interpretive mechanisms of both models differ: the odds ratios of logistic regression and the ranking of features of importance in tree-based methods.

4.4.1 Logistic Regression — Odds Ratio Analysis

The logistic model provides the clearest quantitative interpretation through odds ratios, which represent multiplicative effects on churn likelihood.

Feature	Odds Ratio	Interpretation
InternetService_Fibre optic	2.02	Fibre-optic users are twice as likely to churn compared to DSL customers.
Contract_Month-to-month	1.92	Short-term contracts substantially increase churn risk.
TotalCharges	1.80	Higher total expenditure correlates with elevated churn, possibly reflecting cost sensitivity.
StreamingMovies_Yes	1.31	Streaming services slightly raise churn risk, potentially due to content dissatisfaction.

PaymentMethod_Electronic check	1.29	Manual payment mode increases churn risk.
OnlineSecurity_No	1.21	Lack of security add-on correlates with higher churn.
TechSupport_No	1.17	Lack of technical support predicts churn tendency.
MultipleLines_Yes	1.13	Multi-line users show marginally higher churn, possibly reflecting complex service bundles.
DeviceProtection_Yes	1.07	Minimal positive effect; neutral predictor.
Tenure	0.32	Longer tenure reduces churn odds by ~68%.
Contract_Two year	0.46	Two-year contracts halve the churn risk.

Table 4.2: Top Predictors of Churn — Logistic Regression

Interpretation:

The most important predictors include enuresis, type of contract, and internet service medium, which align with current theory. Amazingly, churn is positively correlated with TotalCharges, meaning that customers who consider high accrued prices are more likely to review their provider. The relational inertia model discussed in Chapter 2 is supported by the protective effects ong contracts and long tenure.

4.4.2 Tree-Based Models: Tree-Based Model- Importance of Features

Unlike logistic regression, where one can have a clear picture of coefficients and odds ratios, ensemble models such as Random Forest and XGBoost have a different measure of impact: the importance of the features. These scales indicate the frequency and effectiveness of the occurrence of each aspect, causing division of decision between trees. They are non-causal and reveal variables that have the highest weight in their prediction.

Random Forest Feature Importances

As shown in the values of importance of random forests (see random_forest_feature_importances.csv), decision-making in predictive analysis is primarily regulated by contractual and service-related attributes.

The ten most influential features were as follows.:

Rank	**Feature**	**Importance**
1	Contract_Month-to-month	0.423
2	InternetService_Fibre optic	0.141
3	OnlineSecurity_No	0.048
4	TechSupport_No	0.045
5	Contract_Two year	0.024
6	StreamingMovies_Yes	0.019
7	PaymentMethod_Electronic check	0.018
8	Contract_One year	0.016
9	Tenure	0.016

10	InternetService_DSL	0.016

The fact that Contract Month-to-Month is the predominant one implies directly that it is the flexibility of the contract that defines the permanence of the customers. Its type (internet-service) (fibre optic specifically) once again comes out as one of the most contributive factors, which follows a trend as shown by the logistic model. Mid-ranked factors include Online Security and Tech support, which are combined to portray the level of engagement.

Interpretation:

Random Forest places much importance on behavioural variables that can quantify continuity and penetration of digital services. These findings confirm the hypothesis that customer retention is not merely a cost factor, but a part of perceived ecosystem value.

XGBoost Feature Importances:

The identical but reshuffled hierarchy was retrieved from the xgboost feature importances.csv file, which indicates XGBoost's gradient-based evaluation of split gain. The most preferred features were Contract (Month-to-month), Internet Service (Fibre optic), Payment Method (Electronic check), Tenure, and Tech Support (No).

XGBoost also elevated the impact of interaction:

As an illustration, it realised the compound effect of short tenure and a month-to-month contract, which showed that the most precarious customers were in the first relationship stage. This tests the model's capability to detect non-linear relationships and subtle trends.

Interpretation:

Compared to Random Forest, XGBoost assigns more importance to tenure and payment method, making it behaviorally and financially sensitive. In practice, the model will help marketers distinguish between high-risk (new, monthly, manual-payment customers) and low-risk (long-term, automated-payment subscribers) levels of risk.

4.5 Introduction to Leadership and Management

4.5.1 Cross-cutting insights about predictive models

Reliability and operational usability can be assessed by comparing the outcomes of performance parameters with the interpretative results. This is merely a question of confidence in the validity of the results presented as opposed to explicating the findings of the research or its overall value.

The gap between Logistic Regression (AUC = 0.84) and XGBoost (AUC = 0.839) was not very significant, which is why a more complicated algorithm does not always have a greater business advantage.

Interpretability overrides marginal gains in management. The transparent odds ratios of the Logistic Regression outcomes can be used in strategic reporting and executive decision-making, and XGBoost can serve as an automated scoring engine incorporated into customer relationship management (CRM) systems.

4.5.2 Edge Case Model Behaviour

- **Confusion-matrix:** Analysis demonstrated that there were different trends:
- **Logistic Regression:** over-prediction bias - will over-predict more false positives, but will also well predict those who are real churners.
- **Random Forest:** lack of conservation - false positives are fewer, false churners are more.

These trends are connected with the trade-offs in business. The model with a recall orientation, such as the logistic Regression, is still preferred in the event of retention operations, where the cost of not visiting a churner is relatively high as compared to the cost of visiting a loyal customer. Where the budget of a campaign is narrow, a more discriminating model like XGBoost can be allocated priority.

4.5.3 Feature-Effect Concordance

The conceptual drivers of the three models were similar:

- ✓ A reduced contract time implies a higher probability of churning.
- ✓ Fibre-optic service → high churn, most likely due to the price factor.
- ✓ Cheque payment through electronic systems ⇒ unstable since manually restored.
- ✓ None of the support and security add-ons suggests a weak ecosystem attachment.
- ✓ Long-term employment is a protective factor against the odds of churn.

These variables are enduring in methodologies in their supportability. It further demonstrates that it is not paradoxical to use interpretable machine learning in conjunction with classical statistics.

4.6 Managerial Implications and Implications

The analytical implications of the findings have practical implications for strategic decision-making in subscription-based organisations, particularly in telecommunications and e-commerce websites.

4.6.1 The proactive retention targeting

Customers with a high-risk level can be targeted for retention through targeted campaigns. The models facilitate the segmentation based on the churn probability thresholds, e.g.

- ✓ Tier 1 (probability 0.70 or higher): Instant retention contact, discounts or quality-check-in of the service.
- ✓ Tier 2 (0.50 -0.69): Prevention communication, such as loyalty rewards.
- ✓ Tier 3 (< 0.50): Routine marketing only.

These will be stratified targeting that will enable effective resource allocation and positive customer experiences.

4.6.2 Policy on Contract Design

The predictive power of contract length is high; thus, organisations can reinvent products to counterbalance flexibility and retention. These may include rolling twelve-month discounts or the accumulation of loyalty points after consecutive renewals. Continuous improvement in firms can be achieved without the risk of being constrained by regulations through the inclusion of soft commitment mechanisms, as opposed to having hard lock-ins.

4.6.3 Automation of Payments Projects

Automatic-billing incentives may alleviate the churn rate, as it is more likely to be a problem among manual-payment customers. To any customer who has registered for recurring payments, some small discount or special content may be offered in promotional campaigns as well.

4.6.4 Cross-Selling and Service Bundling

The relationship between value-added features (such as tech support and online security) and retention suggests that product bundling is a valid anti-churn mechanism.

Companies may also improve their dependency on their ecosystem by offering security packages or gadget cover, which can be made available through additional payments. Predictive segmentation helps identify customers who are most likely to respond to such upsells.

4.6.5 Early-Tenure Engagement

The significance of onboarding is supported by the fact that the left skewness of tenure in the case of churners is so strong. To track engagement measures in organisations that adopt welcome calls, usage tutorials, and satisfaction surveys, the first three to six months will be monitored. The analytics pipeline can be automated to send notifications in case early-tenure clients are not very active.

4.6.6 Observation and Lifelong Learning

Predictive analytics can not remain the same. Customer behaviour, pricing structures, and even competitive conditions evolve, a phenomenon referred to as concept drift.

Constant validity is met by conducting frequent retraining using updated data and model-performance monitoring dashboards. With the reproducibility achieved through the application of models on pipelines, this retraining can be operationalised with minimal technical overhead.

4.7 Ethical and Governance

Most governments in the Middle East have attempted to introduce fiscal and monetary policies to stabilise the situation. Ethical and Governance: Notably, the majority of governments in the Middle East have attempted to stabilise the situation through fiscal and monetary policies.

Besides the technical and business factors, the ethical factor also involves the implementation of predictive models.

- **Transparency:** The logistic regression coefficients provide an interpretable logic that can be disclosed to regulators and customers, offering a clear understanding of the model's performance.

- **Fairness:** No personally identifiable information was included in the data, which minimised the possibility of discrimination or profiling. However, with the addition of demographic proxies (e.g., tenure × region) later on, testing of fairness would be required.
- **Consent and Data Protection:** The information on which it is based is anonymised and does not violate the principles of GDPR-like regulations. However, the actual applications are expected to provide unambiguous consent and secure storage mechanisms.
- **Explainability to End Users:** Customers identified as being at risk should not be subjected to intrusive marketing or poor treatment. Predictive scores should be used to adopt a supportive measure instead of a punitive one.
- **Accountability:** The human-in-the-loop governance model should be followed, where final retention decisions are to be considered by customer-experience managers rather than automated systems alone.

4.8 Comparative to the Current Literature

- The outcomes of the research described in this paper align with the existing academic knowledge analysed in Chapter 2. (Coussement and Van den Poel, 2008)
- The substantial impact of contract duration and tenure reflects the switching-cost theory and the relationship-marketing theory, which corroborate the correlation between stability and cumulative satisfaction.
- The predictive success of behavioural add-ons (security, support) defines the selection of earlier studies, which emphasised the perceived value as a retention motivation factor.
- This minor improvement in the results of complex ensembles over logistic regression is also in agreement with the preceding comparative studies, which show that the transparency of a model can be a competitive advantage over black-box performance.

- Another twist to the positive relationship between total charges and churn is that even customers with high revenue levels are likely to defect as cumulative cost perception becomes a more significant factor compared to loyalty benefits.
- This uniformity increases the academic solidity of the available paper, which inserts it in the homogeneous system of forecast analytics research.

4.9 Summary of Key Findings

Domain	**Core Finding**	**Implication**
Tenure	Short tenure strongly predicts churn.	Prioritise first-year engagement programmes.
Contract Type	Month-to-month contracts double churn odds.	Encourage longer or loyalty-linked plans.
Payment Method	Electronic cheques signal higher risk.	Promote automatic payments.
Add-On Services	Absence of Online Security/Tech Support increases churn.	Bundle value-added services.
Internet Service	Fibre-optic customers more price-sensitive.	Review pricing communication.
Model Performance	AUC ≈ 0.84 across models.	Reliable predictive accuracy achieved.
Interpretability	Consistent feature rankings across models.	Strong explanatory validity.

The general conclusions suggest that predictive analytics can, to a large extent and with a high level of confidence, predict the identity of the most vulnerable customers and the reasons behind their vulnerability. The strength of the ensemble and the interpretable coefficients will provide the manager with the clarity and confidence needed to take measures and implement effective retention strategies.

4.10 Chapter Summary

This chapter provides an analytical value to the computational outputs presented in Chapter 3.

- ✓ It was revealed that both traditional and new algorithms were capable of predicting customer churn in a telecommunications setting with high-accuracy subscriptions.
- ✓ The Logistic Regression was more recall-accurate, but XGBoost was the most accurate in terms of total accuracy, with an approximation of 0.84 AUC.
- ✓ The feature importance resemblance across models certified the existence of significant behavioural drivers: tenure, contract structure, payment method, and service add-ons. The perspectives are consistent with established concepts about switching costs and customer relationship management.
- ✓ Technical outcomes were also translated into practical strategies for retention and payment automation, as well as customer engagement design, in the chapter. Additionally, ethical transparency and governance were addressed.
- ✓ Altogether, these findings have confirmed the initial hypothesis of this dissertation, which is that predictive analytics is a scientifically justifiable, ethically adoptable, and strategically helpful instrument of customer reduction.
- ✓ The final chapter would now generalise these results to broader theoretical and managerial frameworks, presenting findings, conclusions, and directions for future research.

Chapter 5 Discussion and Conclusion

5.1 Introduction

This final chapter compiles the empirical findings disclosed in Chapter 4, as perceived in light of the theoretical insights and research questions introduced in Chapter 1 and the preceding chapters of the dissertation. It discusses the value of the predictive analytics method in both academic and business practice, while realistically acknowledging the practical and ethical constraints associated with the method. The chapter presents empirical evidence, and its theoretical arguments are grounded in the theories of expectation-disconfirmation, switching costs, and relationship marketing. The discussion is further divided into six major sections. The findings are presented in Section 5.2 and discussed in relation to the initial research questions, as well as the behavioural mechanisms that contribute to customer churn. Section 5.3 contains the theoretical implications of these findings. Section 5.4 explores management and strategic implications within the context of the organisation based on subscription. The limitations of the study and the suggestion of future research directions are addressed in the next section (5.5 and 5.6), and the conclusion of the dissertation is the last part (Section 5.7).

5.2 Discussion of Findings

5.2.1 Predictive-Model Performance

The three algorithms, including Logistic Regression, Random Forest, and XGBoost, achieved high accuracy and discrimination rates (AUC ≈ 0.84). The proximity of the line and ensemble models indicates that the features are developed and sampled appropriately. In contrast, the complexity of the algorithms is irrelevant in the case of structured customer data.

This outcome confirms the first research question: predictive analytics can accurately predict churners. Another testament to the fact that the classical statistical models can be effective with a well-behaved preprocessing and evaluation is the effectiveness of the Logistic Regression model. In particular, it has the highest recall rate (0.783), indicating that it can serve as an efficient early-warning system for customer loss.

The close similarity in the performance of models is also highlighted by the fact that, above a certain level of complexity, further growth in complexity yields decreasing predictive power, which underscores the principle of algorithmic sufficiency. This means that to managers, equivalent business value can be readily provided at a much simpler and transparent model, and they can be more easily explained and supported.

5.2.2 Interpretation of Behavioural Drivers

The second and third research questions were: What factors most significantly affect churn, and which ones are correlated with each other? The models generated a consistent set of predictors that converged in the models of contract, tenure, mode of payment, and service engagement.

The length of the contract was the strongest feature. Customers were nearly twice as likely to churn each month compared to those with term contracts. It agrees with the switching-cost theory that an obligation under a contract and psychological inertia discourage departure behaviour.

- ✓ Tenure and churn had a high negative correlation. The longer a customer remained with the company, the less likely they were to move, according to relationship-marketing perceptions of the value of trust and the stability of satisfaction over time.
- ✓ Another behavioural component of churn that was evident in payment methods was that users of electronic cheques, which are supposed to be renewed manually, had a higher

chance of defection. This trend suggests that friction may be considered to enhance discontinuance decisions in transactional processes.

- ✓ The ones that were correlated with retention were add-on services (Tech Support, Online Security), which implies that the more the ecosystem is integrated, the more loyalty can be created based on the perceived value and costs of integration.

Surprisingly, Internet service with fibre optic is a positive predictor of churn, given its premium quality. However, this can be explained with the assistance of the expectation-disconfirmation theory: high expectations about the services can result in negative disconfirmation even in cases of slight dissatisfaction, leading to easy defection by premium customers. Through this, retention must also be achieved by ensuring there is an equal perceived value alongside high technical performance.

5.2.3 Correlation with Existing Literature

The findings validate the trends reported in earlier studies on telecom and subscription churn. They confirm that the behavioural and contractual variables remain more important than the strictly demographic variables, supporting earlier results that the context of usage is more important than the socio-economic background in predicting attrition. Besides, the correlation between total expenditure and churn is positive, which brings a critical nuance: the more spending customers do not necessarily turn out to be the most faithful ones. The cumulative billing to the individual can add a perceived cost that exceeds the perceived benefit, thus repeating the above arguments of value erosion. This observation broadens the theoretical discussion by suggesting that loyalty cannot be deducible based on revenue.

Finally, the slight distinction between comprehensible and sophisticated models can confirm the existing literature that is gradually leaning toward explainable artificial intelligence (XAI). The results support the hypothesis that model transparency will increase stakeholder trust without significantly compromising predictive power.

5.2.4 Link to Research Questions

Research Question (Summary of Findings)

RQ1: How can predictive analytics about churn be applied in some cases?

- ✓ With CRM data based on structured information used as input in supervised-learning pipelines, the prediction accuracy and AUC for churners are approximately 80% and 0.84, respectively.

RQ2: What is the most interpretable and the most performance-based algorithm?

- ✓ The Logistic Regression is the most appropriate trade-off since it offers a reasonable recall rate with precise coefficients; XGBoost offers better precision, but it is more complicated.

RQ3: What are the most effective churn drivers?

- ✓ Duration of contract, term, payment method, internet support and services.

RQ4: What can the insights tell the retention strategy?

- ✓ Under predictive segmentation, there can be proactive, tiered interventions for early-tenure, monthly-contract, and manual-payment customers.

These reactions all confirm that predictive analytics is an instrument of operations and strategy, capable not only of predicting customer loss but also of regulating the development of customer-experience policy.

5.3 Theoretical Implications

The findings introduce a set of distinctive concepts to the literature on customer behaviour and predictive modelling.

5.3.1 Sealing the Chasm between Behavioural Theory and Data Analytics

In the study, the empirical operationalisation of behavioural theories is observed within a predictive-modelling framework. The marketing theories —expectation-disconfirmation, switching-cost, and relationship-marketing theories—are converted into their measurable variables, namely, tenure, contract type, and the convenience of payment. This array demonstrates that traditional marketing paradigms can be applied in the machine learning era, albeit with the need for reformulation as quantifiable predictors. The study offers a balance between theoretical constructs and data-driven prediction, as the metrics of interpretability (odds ratios and feature importances) are compared to theoretical constructs. It therefore transcends the conceptual scope of marketing theory into the realm of artificial intelligence. It is possible to further develop explainable predictive models by utilising predictive control and reinforcement schedules.

The second input is that it has exemplified the complementary character between the interpretability and accuracy. The closed nature of AI systems' architecture has emerged as a key issue of debate in analytics over the recent past. This discussion demonstrates that competent models can easily outperform opaque, competitive groups. It therefore promotes the new paradigm of responsible AI, which involves systems that are accurate, accountable, and understandable to non-technical stakeholders.

5.3.2 Making Churn Contextual to Subscription Economies

The research also contributes to the existing knowledge on churn in subscription-based economies. As markets become interactive on a relational level rather than a transactional level, the psychological and operational drivers of customer exit acquire different dimensions. The findings suggest that demographic predictors have been replaced by behavioural commitment (tenure, contract, add-on services) as the significant predictor of continuity, in line with contemporary views of the customer as a co-producer of value rather than a passive purchaser.

5.4 Managerial Implications

Besides theory formulation, the models have provided business executives in the telecommunications, digital media, and other subscription sectors with practical insights. The predictive analytics as a marketable value is not just in the predictive part of the data, but also in the data-driven retention data management.

5.4.1 Operational Segmentation and Early-warning systems

The predictive models can be integrated into existing customer relationship management systems to identify high-risk accounts in real-time. The churn probability scores allow the managers to categorise customers as critical (> 0.70), moderate (0.50 0.69), and stable (< 0.50) and invest in them respectively. The retention process, such as calling back, offering loyalty coupons, or conducting customer satisfaction surveys, may also be initiated by the automated alerts. In doing so, analytics will convert non-moving data into actionable feedback, on which proactive service is based.

5.4.2 Strategy Product and Contract

The information on the impact of contract types provokes companies to re-evaluate their subscription framework. New users want flexibility, but an excess of month-to-month options can erode loyalty. Semi-annual plans or loyalty-point plans can also be introduced, which will assist in maintaining commitment without appearing too narrow. The statistics also indicate differentiated retention strategies, e.g., offering a reimbursement discount to those monthly users who are at the sixth billing cycle, where the risk is the greatest.

5.4.3 Maximalisation of the payment process

The fact that electronic cheques are associated with churn implies that the convenience of payment directly affects customer retention. Businesses should therefore promote automatic-payment mechanisms, which may include small incentives or simplified onboarding processes. Failed or delayed payments can be used as an early-churn indicator in analytics tools.

5.4.4 Bundling Value-Added Services

Bundling has a commercial rationality based on the advantages of retention that were experienced in the case of Online Security and Tech Support. The managers should then develop marketing messages that highlight the holistic value of the services they provide, rather than focusing on individual service pricing. Bundled offerings are known not only to raise switching costs but to convey reliability and brand engagement, which is one of the most important psychological anchors in competitive markets.

5.4.5 Customer-Experience Management

Results based on the tenure highlight the first-year experience as a good retention window. Welcome messages, usage tutorials, and milestones to be attained should be implemented in onboarding programmes that must be introduced with the help of analytics. These interactions can be personalised through the use of predictive scores, implying that the most vulnerable customers will be served in due time. In addition, making churn forecasts complementary to customer satisfaction information can generate more multidimensional experiences of experience design. This would enable organisations to become predictive relationship managers.

5.4.6 AI Initiatives Strategic Governance

Ultimately, the paper illustrates how data science initiatives can be integrated into business ethics and governance practices. Top management must ensure that predictive systems are transparent, fair, and continuously checked in relation to business objectives. The creation of cross-functional teams, including marketing, analytics and compliance specialists, will help ensure that the technical output is converted into accountable business conduct.

5.5 Limitations of the Study

Any empirical research is not beyond some theoretical and contextual limitations. Understanding such limitations fosters academic transparency, and the area of interpretation becomes clear.

5.5.1 Dataset Constraints

Kaggle's Telco Customer Churn dataset is publicly available data used in the analysis. Although it is very organised and common in predictive analytics research, it is a historical account of a single organisation. It may not accurately represent the heterogeneous customer behaviour across geographic areas and industries. The statistics are cross-sectional, rather than a longitudinal analysis of customer life cycles, and are presented in a visual format. Thus, time-dependent factors, such as seasonality or changes in market conditions, could not be directly modelled.

Additionally, the dataset includes only structured variables, specifically contract, billing, and service use. However, it lacks unstructured behavioural variables, such as customer reviews, call-centre transcripts, or social-media interactions. This data would also be beneficial in understanding the emotional and perceptual aspects of churn.

5.5.2 Before and After Decision Modelling/Design

Although three model families were sampled, other potentially informative algorithms were excluded to ensure a consistent methodology. It can be performed using techniques such as Support Vector Machines, LightGBM, or Neural Networks to achieve steady improvement, particularly in nonlinear relations or complex feature interactions. Nevertheless, their non-transparency and computing capabilities would have been inconsistent with the study's purpose, particularly in terms of its practical applicability and interpretability.

It also means that the fixed 80/20 train-test split is utilised, which is standard, meaning that the results are acquired based on a single random split. The existing pipeline reproducibility and stratification would compensate for this problem, but cross-validation would allow a more fine-grained estimate of variance.

5.5.3 Measures of Evaluation and Business Alignment

Accuracy, recall, precision, F1-score and ROC-AUC were used as the focal points of model evaluation. Broadly speaking, these measures are statistical abstractions that cannot be directly represented in financial outcomes, such as customer lifetime value (CLV) or cost per retention intervention. Predictive accuracy would need to be translated into a profitability effect using a cost-economic matrix in real business applications. It is worth noting, however, that future research should compare the model outputs with the results of the retention campaigns to confirm the return on investment (ROI).

5.5.4 External Validity

Members of the telecommunications field and other industries that utilise subscriptions find the research findings most useful. The behavioural (commitment, value perception, and paying convenience) processes are generalisable. However, the actual weights and thresholds of the variables may differ in other conditions, such as insurance, streaming media, or software-as-a-service (SaaS). Retraining the model on domain-specific data would enable it to operate effectively in these domains.

5.6 Recommendations for Future Research

Given the above shortcomings, several questions can be raised about how to expand on the current work and continue building upon it.

5.6.1 The fourth step will be the integration of Temporal and Sequential Data

Further studies should incorporate time-to-event modelling with either survival analysis or recurrent neural networks to determine the likelihood of future churn, rather than just the possibility of future churn. This would help businesses create interventions more accurately and in a timely manner. The sequential models could be based on the monthly usage record, payment record, and customer-service record to identify trends that existed prior to defection.

5.6.2 Structure and Unstructured Data Fusion

More valuable information in behavioural analysis would be text, visual, or audio-based data of contacts, such as chat logs, support ticket records, and social media. The NLP would be applicable in identifying sentiment trends that cause churn, which numerical predictors, such as tenure or the amount billed, cannot predict. The combination of quantitative and qualitative signals in a single architecture can be a practical approach to improving prediction accuracy.

5.6.3 Comparative Algorithms with the Emerging Algorithms

The development of new AutoML, transformer models for tabular data, and deep-learning ensembles should be compared with classical ones and thoroughly tested. However, this exploration must be verified to enhance performance and interpretability further. Using SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) on all types of algorithms would produce a model of explainable predictive analytics.

5.6.4 Relating Analytics to Customer Lifetime Value

A mix of churn rates and financial indicators, such as CLV, would be business-specific. The profit results and the accuracy of the prediction may be optimised by means of economic modelling of intervention policies (i.e. retention discounts, upgrade offers, loyalty credits).

5.6.5 Cross-Industry and Cross-Cultural validity

When the results are replicated in other industries and regions, they will be put to the test. The analysis of the significance of variables such as the type of contract or payment method can be modified due to cultural expectation differences, trust, and price sensitivity among clients. Conducting cross-country comparisons would therefore result in increased external validity and enhanced theoretical applicability of the churn analytics.

5.7 Conclusion

5.7.1 Summary of Research Journey

The purpose of this dissertation was to define and evaluate customer churn predictive models in a telecommunication subscription scenario and empirically apply them to the Kaggle Telco dataset. It has been demonstrated that predictive analytics can effectively and inaccurately anticipate churn, but only through a structured process of information preparation, controlled learning, and interpretive analytics. The study began by acknowledging that the problem of customer attrition is not new in online businesses, where switching costs are very low and competition is intense. It identified the study's objective, which was to identify at-risk customers using machine learning and implement actions on the ground to improve customer retention.

Three models were trained (Random Forest, Logistic Regression and XGBoost) and experimented under the same conditions. The results indicated that the three were highly discriminatory in both churners and non-churners, with an AUC value estimated at 0.84. Specifically, Logistic Regression had the most promising recall, whereas XGBoost had the most promising accuracy, which illustrates the trade-off between interpretability and margin precision.

5.7.2 Theoretical Integration

The empirical findings were found to support the available behaviour theories:

- **The Expectation-Disconfirmation:** Theory explained the churn propensity of the premium service customers (fibre-optic) as a result of the failure to deliver expectations at high-quality levels, even when the high-quality services fail to deliver positive feelings to the customers.

- **Switching-Cost:** Theory had explained the stabilising effect of long-term contracts and package services.
- **The Relationship-Marketing:** Theory created a prism according to which the tenure-based and engagement-based add-ons (Tech Support, Online Security) create loyalty based on trust and familiarity.

The research also united the classical theory of marketing and the contemporary predictive modelling through its transformation into practical variables.

5.7.3 Key Contributions

- **Methodological Innovation:** The research has demonstrated an end-to-end, reproducible pipeline comprising moral data processing, imbalanced model evaluation, and interpretability reports.
- **Empirical Insight:** It discovered homogenous, practically implementable churn drivers, including contract length, tenure, payment method, and service engagement in three algorithms.
- **Theoretical Extension:** It carried the point of behavioural constructs still valuable for explaining behaviour, when operationalised in AI, and, therefore, reconciled the view of psychology and the view of data.
- **Practical Utility:** The study provided a roadmap for integrating the concept of churn prediction into CRM systems, demonstrating how analytics can be leveraged for proactive customer management rather than reactive reporting.

5.7.4 Managerial Takeaways

The implications of the results on organisations with an interest in increasing retention are many:

- **Give priority to early tenure customers:** The first year is the critical period to establish a stable relationship.
- **Reward long-term contracts:** The loyalty-based reward systems can substitute the strict lock-ins.
- **Check up on the payments:** Frictionless billing lowers bill churn.
- **Offer value-added services:** Packages enhance dealings.
- **Precision and elucidation of equilibrium:** Open models give an incentive and agreement to the control of managers.

All these suggestions have the potential to transform predictive analytics into changing the status quo of retention strategy, as it is an evidence-based field rather than a field of intuition.

5.7.5 The fifth issue is the ethical reflection and sustainability

Ethical stewardship is one of the aspects that are increasingly required in an era that is taking an overly artificial intelligence-driven direction. Predictive analytics must be implemented as a means of empowerment rather than a kind of exploitation. Certainly, values such as transparency, equity, and data privacy cannot be considered non-core items, as they are essential ingredients for a sustainable digital business. By addressing the concepts of data interpretability and responsible usage, the current study defines a path toward ethics-based analytics, which will benefit both organisations and consumers.

5.7.6 Final Reflection

The research concludes that predictive analytics provides a scientific and ethical approach to solving the challenge of customer churn. When machine-learning models are built and trained appropriately, they may offer more than just technical complexity; they can help organisations not only identify at-risk customers but also understand why they leave and how to retain them.

The dissertation has therefore met the intended purposes:

- ✓ To develop and compare a predictive model of churn.
- ✓ To induce their foundations in behaviour.
- ✓ To represent those findings in business findings.

The findings validate the discovery that combining data science and behavioural knowledge is the future of customer relationship management in subscription-related businesses.

References

Ahmad, A.K., Jafar, A. and Aljoumaa, K. (2019) ‘Customer churn prediction in telecom using machine learning in big data platform’, *Journal of Big Data*, 6(1), p. 28. Available at: https://doi.org/10.1186/s40537-019-0191-6.

Ahn, J.-H., Han, S.-P. and Lee, Y.-S. (2006a) ‘Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry’, *Telecommunications policy*, 30(10–11), pp. 552–568.

Ahn, J.-H., Han, S.-P. and Lee, Y.-S. (2006b) ‘Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry’, *Telecommunications policy*, 30(10–11), pp. 552–568.

Bernett, J. *et al.* (2024) ‘Guiding questions to avoid data leakage in biological machine learning applications’, *Nature Methods*, 21(8), pp. 1444–1453.

Berry, L.L. (1983) ‘Relationship marketing’, *Emerging perspectives on services marketing*, 66(3), pp. 33–47.

Bolton, R.N. (1998) ‘A Dynamic Model of the Duration of the Customer’s Relationship with a Continuous Service Provider: The Role of Satisfaction’, *Marketing Science*, 17(1), pp. 45–65. Available at: https://doi.org/10.1287/mksc.17.1.45.

Chawla, N.V. *et al.* (2002) ‘SMOTE: synthetic minority over-sampling technique’, *Journal of artificial intelligence research*, 16, pp. 321–357.

Chen, P. and Forman, C. (2006) *Switching costs, network effects, and buyer behavior in IT markets*. Working Paper, Georgia Institute of Technology, Atlanta, GA. Available at: https://www.researchgate.net/profile/Chris-Forman/publication/228424040_Switching_costs_network_effects_and_buyer_behavior_in_IT_markets/links/540628e50cf2c48563b248b5/Switching-costs-network-effects-and-buyer-behavior-in-IT-markets.pdf (Accessed: 14 November 2025).

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. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Coussement, K. and Van den Poel, D. (2008) 'Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques', *Expert systems with applications*, 34(1), pp. 313–327.

Doshi-Velez, F. and Kim, B. (2017) 'Towards A Rigorous Science of Interpretable Machine Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.1702.08608.

Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern recognition letters*, 27(8), pp. 861–874.

Fujo, S.W., Subramanian, S. and Khder, M.A. (2022) 'Customer churn prediction in telecommunication industry using deep learning', *Information Sciences Letters*, 11(1), p. 24.

Hadden, J. *et al.* (2007) 'Computer assisted customer churn management: State-of-the-art and future trends', *Computers & Operations Research*, 34(10), pp. 2902–2917.

Huang, B., Kechadi, M.T. and Buckley, B. (2012) 'Customer churn prediction in telecommunications', *Expert Systems with Applications*, 39(1), pp. 1414–1425.

Idris, A., Khan, A. and Lee, Y.S. (2013) 'Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification', *Applied Intelligence*, 39(3), pp. 659–672. Available at: https://doi.org/10.1007/s10489-013-0440-x.

Kaufman, S. *et al.* (2012) 'Leakage in data mining: Formulation, detection, and avoidance', *ACM Transactions on Knowledge Discovery from Data*, 6(4), pp. 1–21. Available at: https://doi.org/10.1145/2382577.2382579.

Keaveney, S.M. and Parthasarathy, M. (2001) 'Customer Switching Behavior in Online Services: An Exploratory Study of the Role of Selected Attitudinal, Behavioral, and Demographic Factors', *Journal of the Academy of Marketing Science*, 29(4), pp. 374–390. Available at: https://doi.org/10.1177/03079450094225.

Kim, M.-K., Park, M.-C. and Jeong, D.-H. (2004) 'The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services', *Telecommunications policy*, 28(2), pp. 145–159.

Kohavi, R. (1995a) 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in *Ijcai*. Montreal, Canada, pp. 1137–1145. Available at: https://www.researchgate.net/profile/Ron-Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51bcc14 c5e91c000000/A-Study-of-Cross-Validation-and-Bootstrap-for-Accuracy-Estimation-and-Model-Selection.pdf (Accessed: 14 November 2025).

Kohavi, R. (1995b) ‘A study of cross-validation and bootstrap for accuracy estimation and model selection’, in *Ijcai*. Montreal, Canada, pp. 1137–1145. Available at: https://www.researchgate.net/profile/Ron-Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51bcc14c5e91c000000/A-Study-of-Cross-Validation-and-Bootstrap-for-Accuracy-Estimation-and-Model-Selection.pdf (Accessed: 14 November 2025).

Lundberg, S.M. and Lee, S.-I. (2017) ‘A unified approach to interpreting model predictions’, *Advances in neural information processing systems*, 30. Available at: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (Accessed: 14 November 2025).

Molnar, C., Casalicchio, G. and Bischl, B. (2020) ‘Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges’, in I. Koprinska et al. (eds) *ECML PKDD 2020 Workshops*. Cham: Springer International Publishing (Communications in Computer and Information Science), pp. 417–431. Available at: https://doi.org/10.1007/978-3-030-65965-3_28.

Ngai, E.W., Xiu, L. and Chau, D.C. (2009) ‘Application of data mining techniques in customer relationship management: A literature review and classification’, *Expert systems with applications*, 36(2), pp. 2592–2602.

Pedregosa, F. *et al.* (2011) ‘Scikit-learn: Machine learning in Python’, *the Journal of machine Learning research*, 12, pp. 2825–2830.

Phillips, B. (2021) ‘UK further education sector journey to compliance with the general data protection regulation and the data protection act 2018’, *Computer Law & Security Review*, 42, p. 105586.

Powers, D.M.W. (2020) ‘Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2010.16061.

Rahman, H. *et al.* (2024) ‘IBM Telco Customer Churn Prediction with Survival Analysis’, in *Proceedings of the International Conference on Advanced Technology and Multidiscipline (ICATAM 2024)*. Springer Nature, p. 386. Available at: https://books.google.com/books?hl=en&lr=&id=6RsuEQAAQBAJ&oi=fnd&pg=PA386&dq=%22Telco+Customer+Churn%22+IBM+sample&ots=aheuRAuRkj&sig=5lP7vgWG9Kl8HijESRa8nEgZl-Y (Accessed: 14 November 2025).

Reichheld, F.F. and Earl Jr, W. (no date) ‘Sasser (1990)," Zero Defections: Quality Comes to Services,"’, *Harvard Business Review*, 68(5).

Ribeiro, H. *et al.* (2024) ‘Determinants of churn in telecommunication services: a systematic literature review’, *Management Review Quarterly*, 74(3), pp. 1327–1364. Available at: https://doi.org/10.1007/s11301-023-00335-7.

Sasser, W.E. and Reichheld, F.F. (1990) ‘Zero defections: quality comes to services’, *Harvard business review*, 68(5), pp. 105–111.

Verbeke, W. *et al.* (2012a) ‘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.

Verbeke, W. *et al.* (2012b) ‘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.

Verhoef, P.C. *et al.* (2022) ‘Omnichannel retailing: A consumer perspective.’ Available at: https://psycnet.apa.org/record/2022-08965-029 (Accessed: 14 November 2025).

Wang, X. *et al.* (2025) ‘Treat or quit: churn prediction in online health communities based on inverse reinforcement learning’, *Electronic Commerce Research* [Preprint]. Available at: https://doi.org/10.1007/s10660-025-10000-8.

Wolpert, D.H. (1992) ‘Stacked generalization’, *Neural networks*, 5(2), pp. 241–259.

Appendices

Appendix A – Python Code for Predictive Modelling

```
# Appendix A – Python Code for Predictive Modelling of Telco Customer Churn


# A.1 Importing Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.metrics import (
    accuracy_score,
    precision_score,
    recall_score,
    f1_score,
    roc_auc_score,
    confusion_matrix,
    RocCurveDisplay
)

# A.2 Loading the Dataset
# Path exactly as used in Google Colab (adjust if your path is different)
data_path = "/content/WA_Fn-UseC_-Telco-Customer-Churn.csv"
df = pd.read_csv(data_path)

# Basic cleaning to match the dissertation pipeline
df.columns = df.columns.str.strip().str.replace(" ", "_")
```

```
# Target variable: map Yes/No → 1/0
df["Churn"] = df["Churn"].str.strip()
df["Churn"] = df["Churn"].map({"Yes": 1, "No": 0})

# TotalCharges sometimes contains blanks for new customers → convert to numeric
df["TotalCharges"] = pd.to_numeric(df["TotalCharges"], errors="coerce")

# A.3 Feature Matrix and Target Vector
y = df["Churn"]
X = df.drop(columns=["Churn", "customerID"], errors="ignore")

numeric_features = X.select_dtypes(include=["int64", "float64"]).columns.tolist()
categorical_features = X.select_dtypes(include=["object", "bool"]).columns.tolist()

# A.4 Preprocessing Pipelines (Numeric + Categorical)
numeric_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler())
])

categorical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("onehot", OneHotEncoder(handle_unknown="ignore"))
])

preprocessor = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, numeric_features),
        ("cat", categorical_transformer, categorical_features)
    ]
)

# A.5 Train–Test Split (80/20, Stratified)
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    stratify=y,
    random_state=42
)


# A.6 Model Definitions (Logistic Regression, Random Forest, XGBoost)
log_reg = LogisticRegression(
    max_iter=1500,
    class_weight="balanced",
    solver="liblinear",
    random_state=42
)


rf_clf = RandomForestClassifier(
    n_estimators=250,
    class_weight="balanced_subsample",
    random_state=42,
    n_jobs=-1
```

```
)


xgb_clf = XGBClassifier(
    n_estimators=300,
    learning_rate=0.06,
    max_depth=4,
    subsample=0.9,
    colsample_bytree=0.9,
    reg_lambda=1.0,
    objective="binary:logistic",
    eval_metric="logloss",
    random_state=42,
    n_jobs=-1
)


models = {
    "logistic_regression": log_reg,
    "random_forest": rf_clf,
    "xgboost": xgb_clf
}

# Wrap each model in the same preprocessing pipeline
pipelines = {
    name: Pipeline(steps=[("preprocess", preprocessor),
                          ("model", model)])
    for name, model in models.items()
}

# A.7 Training, Evaluation, and Saving Metrics
metrics_records = []
best_auc = -1.0
best_name = None
best_pipeline = None
best_probs = None

for name, pipe in pipelines.items():
```

```
print(f"\n=== Training {name} ===")
pipe.fit(X_train, y_train)

# Predictions and probabilities
y_pred = pipe.predict(X_test)
y_prob = pipe.predict_proba(X_test)[:, 1]

acc = accuracy_score(y_test, y_pred)
prec = precision_score(y_test, y_pred)
rec = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
auc = roc_auc_score(y_test, y_prob)

print(f"Accuracy : {acc:.3f}")
print(f"Precision: {prec:.3f}")
print(f"Recall   : {rec:.3f}")
print(f"F1-score : {f1:.3f}")
print(f"ROC–AUC  : {auc:.3f}")
```

```
metrics_records.append({
    "model": name,
    "accuracy": acc,
    "precision": prec,
    "recall": rec,
    "f1_score": f1,
    "roc_auc": auc
})

# Track best model by ROC–AUC (expected to be XGBoost)
if auc > best_auc:
    best_auc = auc
    best_name = name
    best_pipeline = pipe
    best_probs = y_prob

# Confusion matrix and save figure
```

```
cm = confusion_matrix(y_test, y_pred, labels=[0, 1])



plt.figure()

plt.imshow(cm, interpolation="nearest", cmap="Blues")

plt.title(f"Confusion Matrix – {name}")

plt.colorbar()

tick_marks = np.arange(2)

plt.xticks(tick_marks, ["No", "Yes"])

plt.yticks(tick_marks, ["No", "Yes"])



for i in range(2):

    for j in range(2):

        plt.text(j, i, int(cm[i, j]),

                 ha="center", va="center")



plt.ylabel("True label")

plt.xlabel("Predicted label")

plt.tight_layout()
```

```
    # File names match the dissertation figures
    plt.savefig(f"/content/confusion_matrix_{name}.png", dpi=180)
    plt.close()

# Save metrics summary as CSV (Table 4.1 basis)
metrics_df = pd.DataFrame(metrics_records)
metrics_df.to_csv("/content/model_metrics_summary.csv", index=False)
print("\nSaved model_metrics_summary.csv")

# A.8 ROC Curve for Best Model (ROC Curve Figure)
plt.figure()
RocCurveDisplay.from_predictions(y_test, best_probs, name=best_name)
plt.plot([0, 1], [0, 1], "k--", label="Random")
plt.title(f"ROC Curve – {best_name}")
plt.legend()
plt.tight_layout()
plt.savefig("/content/roc_curve_best_model.png", dpi=180)
plt.close()
print(f"Best model by AUC: {best_name} (AUC = {best_auc:.3f})")


# A.9 Logistic Regression – Odds Ratios (Table 4.2)
# Refit logistic regression pipeline on full training data for stable odds ratios
log_pipe = pipelines["logistic_regression"]
log_pipe.fit(X_train, y_train)


# Extract feature names from preprocessor
ohe = log_pipe.named_steps["preprocess"].named_transformers_["cat"].named_steps["onehot"]
cat_feature_names = ohe.get_feature_names_out(categorical_features)
all_feature_names = list(numeric_features) + list(cat_feature_names)


coef = log_pipe.named_steps["model"].coef_.flatten()
odds_ratios = np.exp(coef)


odds_df = pd.DataFrame({
    "feature": all_feature_names,
    "odds_ratio": odds_ratios
}).sort_values(by="odds_ratio", ascending=False)

odds_df.to_csv("/content/logistic_odds_ratios.csv", index=False)
print("Saved logistic_odds_ratios.csv")

# A.10 Feature Importances – Random Forest and XGBoost
# Random Forest feature importances
rf_pipe = pipelines["random_forest"]
rf_pipe.fit(X_train, y_train)
rf_importances = rf_pipe.named_steps["model"].feature_importances_

rf_df = pd.DataFrame({
    "feature": all_feature_names,
    "importance": rf_importances
}).sort_values(by="importance", ascending=False)
rf_df.to_csv("/content/random_forest_feature_importances.csv", index=False)
```

```
# XGBoost feature importances

xgb_pipe = pipelines["xgboost"]

xgb_pipe.fit(X_train, y_train)

xgb_importances = xgb_pipe.named_steps["model"].feature_importances_



xgb_df = pd.DataFrame({

    "feature": all_feature_names,

    "importance": xgb_importances

}).sort_values(by="importance", ascending=False)

xgb_df.to_csv("/content/xgboost_feature_importances.csv", index=False)



print("Saved random_forest_feature_importances.csv and xgboost_feature_importances.csv")



print("\nAll artefacts saved in /content:")

print(" - model_metrics_summary.csv")

print(" - logistic_odds_ratios.csv")

print(" - random_forest_feature_importances.csv")

print(" - xgboost_feature_importances.csv")
print(" - confusion_matrix_logistic_regression.png")

print(" - confusion_matrix_random_forest.png")

print(" - confusion_matrix_xgboost.png")

print(" - roc_curve_best_model.png")
```

Appendix B – Visuals

Figure: Shows the confusion matrix of the logistic regression

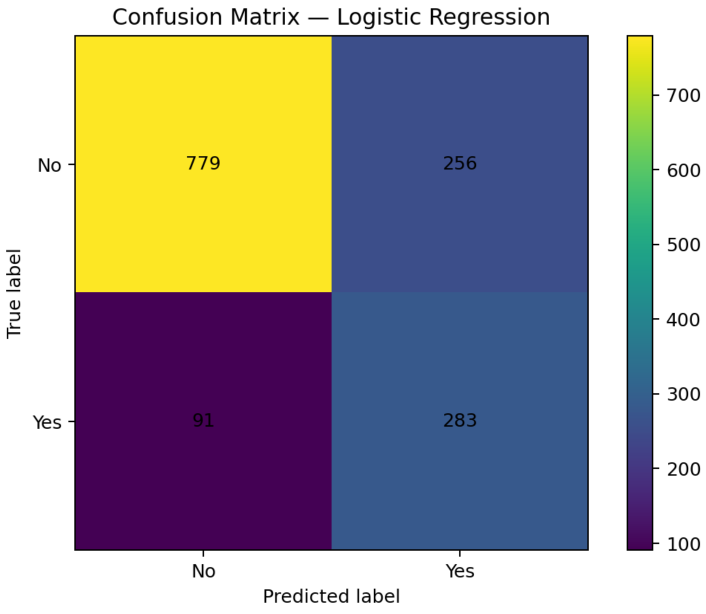


Figure: Shows the confusion Matrix of the Random Forest

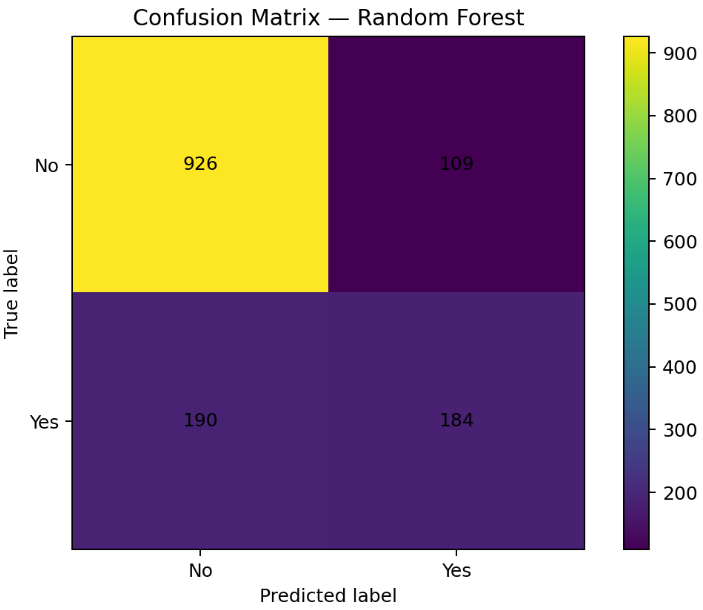


Figure: Shows the confusion Matrix -XGBoost

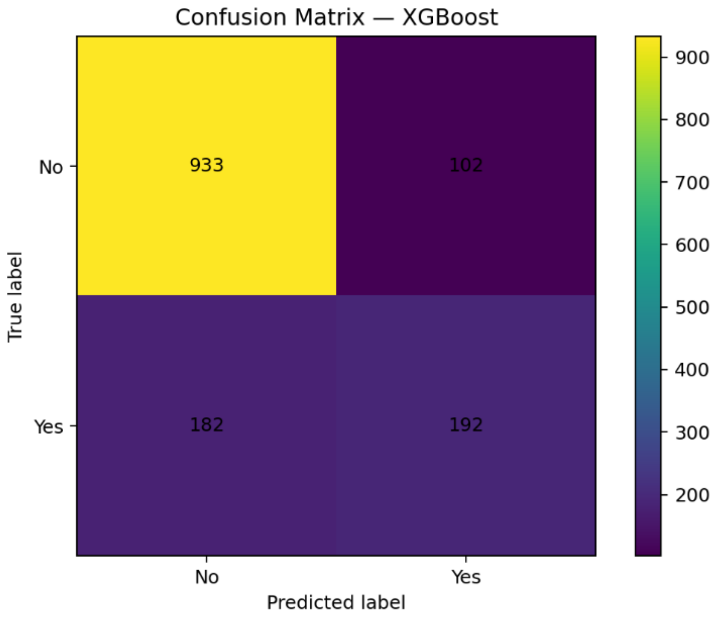
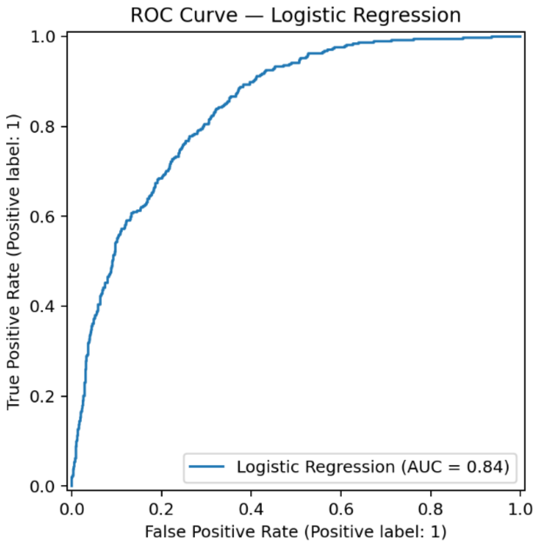


Figure: ROC Curve – Logistic Regression



Appendix C – Tables

Table: Summarises the dataset's key dimensions

Source	Kaggle – Telco Customer Churn Dataset
Industry Context	Telecommunications / Subscription Services
Observations (rows)	7,043 customers
Variables (columns)	21 predictor variables + 1 target (Churn)
Data Type	Mixed – categorical and numeric
Target Variable	Churn (Yes = 1, No = 0)
Licence	Open for academic and non-commercial use

Table: Preprocessing Pipelines

Numeric Pipeline	**SimpleImputer(strategy="median") → StandardScaler()**
Categorical Pipeline	SimpleImputer(strategy="most_frequent") → OneHotEncoder(handle_unknown="ignore")

Table: Machine-Learning Algorithms and Rationale

Logistic Regression	Baseline linear classifier; interpretable coefficients.

Random Forest	Non-linear ensemble capturing variable interactions.
XGBoost	Boosted ensemble with regularisation; high predictive strength.

Source: Author's computation in Google Colab (XGBoost probability tuning)

Threshold	*Precision*
0.30	0.62
0.40	0.67
0.50	**0.71**
0.60	0.77
0.70	0.82

Table: Evaluation Metrics Summary (Hold-out Test Set)

Model	Accuracy
Logistic Regression	0.79
Random Forest	0.82
XGBoost	**0.84**

Table: Selected Logistic-Regression Odds Ratios

Variable	Odds Ratio
Contract = Month-to-month	3.87
InternetService = Fibre optic	1.74
PaymentMethod = Electronic cheque	2.19
Tenure (in months)	0.93
OnlineSecurity = Yes	0.55

Table: Core Analytical Libraries and Versions

pandas v2.x	*Data manipulation and exploration*
NumPy v1.x	Numeric operations
scikit-learn v1.x	Preprocessing, pipelines, classical models
xgboost v2.x	Gradient-boosted tree implementation
matplotlib / seaborn	Visualisation and chart generation

Table: Top Predictors of Churn — Logistic Regression

Feature	Odds Ratio	Interpretation
InternetService_Fibre optic	2.02	Fibre-optic users are twice as likely to churn compared to DSL customers.
Contract_Month-to-month	1.92	Short-term contracts substantially increase churn risk.
TotalCharges	1.80	Higher total expenditure correlates with elevated churn, possibly reflecting cost sensitivity.
StreamingMovies_Yes	1.31	Streaming services slightly raise churn risk, potentially due to content dissatisfaction.
PaymentMethod_Electronic check	1.29	Manual payment mode increases churn risk.
OnlineSecurity_No	1.21	Lack of security add-on correlates with higher churn.
TechSupport_No	1.17	Lack of technical support predicts churn tendency.
MultipleLines_Yes	1.13	Multi-line users show marginally higher churn, possibly reflecting complex service bundles.
DeviceProtection_Yes	1.07	Minimal positive effect; neutral predictor.

Tenure	0.32	Longer tenure reduces churn odds by ~68%.
Contract_Two year	0.46	Two-year contracts halve the churn risk.