Strathmore
UNIVERSITY

# Predicting Customer Churn in the Telecommunications Industry using Machine Learning Techniques


**By**

**Adeline Makokha**

**191199**


**A Research Proposal Submitted in Partial Fulfillment of the Requirements for Completion of Master of Science in Data Science and Analytics (MSc. DSA)**


**iLab Africa**

**Strathmore Institute of Mathematical Sciences (SIMS)**

**Strathmore University**

**Nairobi, Kenya**


**November, 2025**

## DECLARATION AND APPROVAL

### DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the proposal contains no material previously published or written by another person except where due reference is made in the proposal itself.



Student's Name: **Adeline Makokha**

Sign:

*27 September 2025*

### APPROVAL

The proposal was reviewed and approved for defense by

Dr. Henry Muchiri

School of Computing and Engineering Sciences

Strathmore University

Sign:

*13 October 2025*

**ABSTRACT**

In the telecommunications industry, predicting customer churn is a recurring problem that has a big impact on long-term viability, profitability, and competitiveness. The context and research challenges are first described here, which also highlights the sector's growing rivalry and the shortcomings of conventional churn control techniques. This research emphasizes the goals of finding important churn drivers, creating a prediction model based on machine learning models, via assessing various algorithms, and creating a deployable dashboard to help in decision-making. Recent research on churn prediction is combined with theoretical framework like Customer Relationship Management (CRM) in a review of relevant literature. Previous research shows the promise of deep learning and machine learning models, but it also highlights insufficient information in terms of interpretability, profit-sensitive evaluation, and a narrow concentration on telecommunication sectors hence the huge gap between technological customer churn prediction and business insights to inform decision-making. The methodology suggests using secondary public data from a telecommunication provider and adheres to the CRISP-DM framework. Cleaning, feature engineering, addressing class imbalance, and encoding are examples of data preprocessing procedures to be used. In order to improve performance and stakeholder trust, a number of models including logistic regression, decision trees, random forest, k-nearest neighbors (KNN), support vector machine (SVM), naive bayes (NB) and gradient boosting (GB) will be compared. Interpretability approaches like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) will be used for analysis to get results. The creation of a reliable churn prediction system with excellent predictive accuracy, interpretability, and scalability is one of the anticipated results. From an operational point of view, the system will facilitate constant monitoring, efficient resource allocation, and proactive customer retention tactics and strategies. In addition to providing methodological insights that can be applied to other telecommunications, this research offers a deployable and replicable methodology that solves the urgent commercial demand for customer churn reduction.

Keywords: Random Forest, Customer Churn, Telecommunication, Customer Retention

## ACRONYMS

1. ANNs – Artificial Neural Networks
2. API - Application Programming Interface
3. AUC - Area Under the Curve
4. CNNs – Convolutional Neural Networks
5. CRISP-DM - Cross-Industry Standard Process for Data Mining
6. CRM- Customer Relationship Management
7. DL – Deep Learning
8. DT – Decision Tree
9. ERT – Extra Random Trees
10. FT – Functional Trees
11. GB – Gradient Boosting
12. k-NN – k-Nearest Neighbors
13. LightGBM - Light Gradient Boosting Machine
14. Lime - Local Interpretable Model-agnostic Explanations
15. LMT – Logistic Model Tree
16. LR – Logistic Regression
17. ML – Machine Learning
18. NB - Naive Bayes
19. RF – Random Forest
20. ROC - Receiver Operating Characteristic
21. ROI - Return on Investment
22. SHAP - SHapley Additive exPlanations
23. SMOTE - Synthetic Minority Over-sampling Technique
24. SVM - Support Vector Machine

# Contents

# 1. CHAPTER: INTRODUCTION

## 1.1 Background

Within the worldwide telecommunications business, customer churn is defined as customers leaving their subscription or services in favor of competing offers, which represents a significant and ongoing concern in the industry. Customer churn has significantly impacted the telecommunication business's profitability, operational sustainability, and long-term market competitiveness. Annual telecommunication churn rates are above 30% Khan et al. (2024) in the global telecommunication ecosystem. This has raised the need for proactive prediction and retention measures and causing significant annual revenue losses for service providers.

Mobile communication continues to dominate global connectivity, with mobile devices both smartphones and feature phones, accounting for the majority of voice and data traffic worldwide. Despite this saturation, the mobile segment remains one of the fastest-growing areas within the broader telecommunications industry. As competition intensifies, the strategic focus has shifted from acquiring new subscribers to retaining existing ones, a pattern seen across many mature markets. In this context, churn represents the proportion of customers who discontinue a service in favor of alternative providers as referenced by Khan et al. (2024). For telecommunication operators, tracking churn has become a core performance indicator, similar to how long-term service industries monitor client loyalty. The study further suggests that annual customer losses in the post-COVID period range between 30% and 35% due to hardship circumstances. These rates may escalate further as new entrants and larger competitors reshape the market, where some telecommunications companies decided to increase their prices while others have reduced their prices, leading to a flexible shift of customers preferring other networks. Compounding the challenge is the financial reality that bringing in a new subscriber often costs several times more than retaining an existing one, making churn reduction a financially critical priority for cellular service firms.

From an economic perspective, retaining customers is significantly more cost-effective compared to acquiring new customers. Studies show that acquiring new customers can be five to seven times more expensive compared to retaining customers Usman-Hamza et al. (2022). An increase in the number of churners can cause a reduction in revenue stability in the market, constraining the organizational competitiveness. The rapid growth and technological evolution of the telecommunications industry have broadened the number of companies operating in the sector, intensifying competitive pressure. As a result, customer churn has become a persistent challenge, driven by saturated markets, aggressive competitive offerings, and frequent release of attractive service bundles. In such a flexible environment, telecommunication operators must continually find ways to protect and grow their revenue streams while also considering the dynamic factors in the industry. Strategies commonly recommended include attracting new users, increasing the value of existing customers, and prolonging subscriber loyalty. This makes churn prediction an essential component of modern telecommunication operations.

A study by Usman-Hamza et al. (2022) shows that Customer Churn Prediction (CCP) provides organizations with the ability to anticipate customer loss and design targeted retention initiatives that strengthen revenue performance and competitive position. Telecommunication firms now possess extensive datasets on subscriber behavior ranging from voice, data, and messaging records to demographic profiles and billing histories. These rich information sources are valuable for identifying patterns that signal a likelihood of churn. The challenge for operators is to use these insights proactively, detecting early signs of disengagement or inactivity before a customer decides to leave.

Globally, service-oriented industries particularly telecommunications continue to grapple with customer churn as digitalization accelerates and competition intensifies. International studies show that telecommunication markets generate significant revenue, especially in developing regions, yet face persistent challenges as operators introduce new technologies and service offerings to retain increasingly mobile customers. With monthly churn rates reported in several markets and evidence that retaining a subscriber is far more cost-effective than acquiring a new one, customer loyalty has become a strategic priority worldwide. Over the years, machine learning research has played a critical role in supporting these efforts by enabling firms to identify subscribers likely to leave and by revealing behavioral patterns embedded in customer databases. While global literature has extensively explored churn prediction through techniques such as feature extraction, ensemble learning, and gradient boosting, a growing number of local studies highlight the need to move beyond simple binary classification and incorporate contextual understanding of why customers churn. Recent contributions, such as Rahman et al. (2024), demonstrate that models like Logistic Regression, Random Forests, LightGBM, and ensemble workflows can handle diverse datasets and accurately flag high-risk customers before they exit. However, for regions, where telecommunication services underpin financial transactions, education, and everyday communication, combining predictive accuracy with interpretable insights is essential for designing effective retention strategies that reflect local user behavior and market realities

The study highlighted by Mwaura (2021) demonstrates that customer experience management plays a central role in reducing churn within Kenya's telecommunications industry. Using survey data from 195 telecommunication staff, the research found that social factors, quality of service interfaces, and the atmosphere of retail outlets all contribute positively to customer retention, while pricing pressures increase the likelihood of churn. These findings emphasize that customer loyalty in Kenya is shaped not only by service performance but also by relational, experiential, and affordability dimensions. As the sector becomes more competitive and essential to daily digital activity, telecommunication operators must prioritize customer-centered strategies—leveraging social engagement, enhancing user interactions, improving retail environments, and offering price-sensitive plans—to strengthen retention and maintain market stability.

The rapid digital adoption in the market and socio- economic on mobile services on calls and data usage has really lead to high competition in the telecommunications market. By combining demographic information, revenue information and service usage for the customer can be used to intervene the customer retention strategies, which will save on marketing expenditure and improve customer loyalty to one network.

## 1.2 Research Problem

With increase in the customer expectation and customers preferring affordable and reliable services has really intensify the competition in the industries. Most of the customers have leave the network and prefer other telecommunications service providers that provide good and affordable service qualities. Most of the previous churn rate predictions remain reactive, that is relying on static indicators and also manual analysis that can be inefficient, prone to human bias and also time-consuming to come up with business decisions. This existing prediction also has limitation when it comes to bridging the gap between the technological concept of customer churn and the business insights. These methods have struggled to predict customer churn and as a result leading to revenue leakage due to the high customer churn rate in the industry. Mwaura (2021).

## 1.3 Research Objectives

This study seeks to solve the identified gaps through the following research objectives:

**Main Objective**

To predict customer churn in the telecommunications industry.

**Specific Objectives**

i. To describe the challenges and limitations in the existing customer churn literature in the telecommunications industry.

ii. To develop a churn prediction model that is machine learning-based using telecommunications data.

iii. To evaluate and compare the machine learning algorithms performance.

iv. To deploy machine learning models in an interactive dashboard.

## 1.4 Research Questions

To address the above stated objectives, the study is guided by the following research questions:

1. What challenges and limitations exist in current churn prediction approaches within the telecommunications industry?
2. Which demographic, contractual, financial, and usage-related factors are the most significant predictors of customer churn?
3. How can machine learning models be developed and optimized to predict churn in the telecommunications sector?
4. How does the performance of different ML algorithms (e.g., logistic regression, random forest ) compare in predicting customer churn?
5. How can a predictive churn model be deployed in an interactive dashboard to support decision-making and interventions?

## 1.5 Significance and Justification

Globally, effective management of churn is mandatory in order to ensure sustainability, profitability and competitiveness in the telecommunications industry. The strategies to customer retentions has really reduce the high cost that is put in place to acquire new customers in the network. This has also stabilize the telecommunications revenue stream, has safeguard the market share as well as enhance a long customer value relationship with the industry.

The implications of customer churn has extended beyond the corporate world and now also impacting the emerging markets. This is evident in our day to day lives through e learning systems, e- health systems, small businesses operation where telecommunication services has major impact and one of the drivers of the economy. There is a slow progress towards digital inclusion and also economic development as a result of high churn rates.

This research is justified on three key fronts:

- Economic efficiency: By identifying at-risk customers before they leave, providers can target retention efforts more precisely, thereby reducing costs and maximizing return on marketing investment.
- Operational effectiveness: Th e integration of the predictive model into CRM systems ensures actionable insights are delivered to decision-makers enabling timely and effective interventions.
- Scalability and replicability: While tailored to socio-economic and competitive environment, the proposed framework offers a methodological blueprint that can be adapted for other telecommunication markets and industries facing similar churn-related challenges.

This study tend to contribute to the practical knowledge available and the academic concept on customer churn in the telecommunications industry. The research aims to create a sustainable and data-driven approach that will leverage the Machine learning techniques into an operational workflow just to retain customers.

# 2. CHAPTER: LITERATURE REVIEW

There has been a lot of shift over the past years in the digital evolution of technology, lot of markets and startups have open up and as a result leading to customer demand and more personalized quality services in the telecommunications industry. Most of this changes has really widened the access of communication and led to the need of improvement. With this high competitive ecosystem of customer churn, keeping the existing customer has been more critical just as acquiring new customers in to the network. One of the main challenges is the subscriber or customer making the decision to stop using the network due to one reason or another. Many telecommunications operators have lost a third of their customer base to other operators making customer churn a critical global issue. These losses has significantly translated to revue erosion in the industry hence the need to develop and adopt a data driven approaches that can curb customer churn Nagarkar (2022).

## 2.1 Theoretical Framework

### 2.1.1 Customer Relationship Management

Shahabikargar et al. (2025) provide an important contribution to Customer Relationship Management (CRM) by demonstrating how deeper insights into customer cognition and emotional signals can significantly strengthen churn prevention strategies. Traditional CRM systems often rely on demographic attributes, usage behavior, or billing history; however, this study shows that much richer information lies within customer–company interactions, particularly unstructured text such as emails or service requests. Importantly, the study highlights that this approach not only enhances customer retention but can also extend to broader CRM applications such as personalized engagement, early detection of customer dissatisfaction, and proactive service intervention—positioning ChurnKB as a powerful tool for building stronger, more responsive customer relationships.

### 2.1.2 Predictive Modelling of Churn Using Statistical and Segmentation-Based Approaches

Zhang et al. (2022) study provides an important theoretical contribution to understanding churn behavior by integrating customer segmentation with statistical prediction techniques. Using data from three major Chinese telecommunication operators, the research employed Fisher discriminant analysis and logistic regression to construct predictive models capable of identifying customers at risk of leaving. The findings indicated that logistic regression delivered superior performance, achieving a prediction accuracy of 93.94%, outperforming the discriminant approach. This demonstrates the effectiveness of regression-based models in capturing behavioral differences across customer segments and translating them into actionable churn insights. Zhang et al. (2022) work reinforces the theoretical premise that combining segmentation strategies with interpretable statistical models enhances the ability of telecommunication firms to anticipate churn and implement targeted retention interventions, ultimately supporting more profitable and customer-centric decision-making.

### 2.1.3 Forest Models and Ensemble Strategies for Churn Prediction

Usman-Hamza et al. (2022) research provides a strong theoretical foundation for understanding how intelligent decision forest models can enhance customer churn prediction in the telecommunications sector, where competitive pressures and high acquisition costs make retention a strategic priority. The study emphasizes that traditional rule-based approaches lack scalability and that conventional machine learning techniques often struggle with the inherent imbalance between churn and non-churn classes. To address these limitations, the authors developed and evaluated several decision forest variants—including Logistic Model Trees, Random Forests, and Functional Trees—alongside enhanced ensemble versions that incorporate weighted soft voting and stacking mechanisms. Using publicly available benchmark datasets, the study demonstrated that these decision forest models consistently outperformed baseline machine learning methods and showed strong resilience when dealing with imbalanced data. The findings highlight the theoretical and practical value of ensemble-based decision forests in producing more stable and accurate churn predictions, supporting their adoption for telecommunications CCP and broader machine learning applications.

### 2.1.4 Ensemble Learning and Explainable AI Approaches for Telecom Churn Prediction

Recent work examining churn dynamics in the telecommunications sector—an industry known to record annual churn levels exceeding 30%—has highlighted the value of ensemble learning in developing accurate and interpretable prediction models. The study evaluated multiple ensemble-based and tree-driven algorithms, including Decision Trees, Boosted Trees, and Random Forests, to determine their suitability for forecasting customer exit in large-scale telecom datasets. Among the tested models, Random Forest delivered the strongest performance, achieving 91.66% accuracy, 82.2% precision, and 81.8% recall, demonstrating its ability to correctly identify a large proportion of at-risk subscribers. In addition to its predictive strength, the model's interpretability was enhanced through the integration of explainable artificial intelligence techniques such as LIME and SHAP, which provided insight into the feature contributions influencing churn outcomes. The study's findings emphasize that combining ensemble learning with XAI tools not only improves prediction accuracy but also increases transparency—an essential requirement for telecom organizations seeking actionable, trustworthy decision-support systems for customer retention.

## 2.2 Conceptual Framework

The conceptual models guiding this research is mainly integrated in four components that include:

- **Behavioral and Demographic Profiling:** This mainly entails customer characteristics like service usage, billing routine and revenue impact just to establish most of the key predictors in customer churn.
- **Predictive Modeling:** With an aim to deploy machine learning algorithms which include, Logistic Regression, Decision Trees, Random Forest, Light GBM and Ensemble Learning, in order to compare the performance of the various models for consistency.
- **Retention-Oriented Insights:** Translation of churn probabilities into actionable strategies such as personalized offers, targeted communication, and segmented interventions that align with customer needs.
- **Model Assessment:** Evaluation using precision, recall, F1-score, and ROC-AUC to ensure a balanced understanding of predictive behavior across churn and non-churn classes.

## 2.3 Empirical Studies

### 2.3.1 Handling Class Imbalance in Telecommunication Churn Prediction

Recent research in telecommunications churn analysis emphasizes that revenue loss stemming from customer exit remains one of the sector's most pressing challenges. As retention strategies increasingly outweigh acquisition-focused approaches in their cost-effectiveness, predictive modeling has become central to identifying customers at risk. However, many traditional machine learning models struggle with the highly imbalanced structure of telecom datasets, where churners represent only a small fraction of the customer base. Addressing this limitation, Nagarkar (2022) evaluated a large, operator-specific dataset from Nepal comprising 52,332 subscribers and demonstrated that XGBoost could effectively capture churn behavior despite the skewed class distribution. The model achieved strong performance, recording 97% accuracy and an 88% F1-score on the native dataset, which included 6,128 churners and 46,204 non-churners. When applied to a smaller publicly available dataset of 3,333 records for comparison, XGBoost also delivered improved results, achieving 96.25% accuracy and an 86.34% F1-score. These findings highlight the robustness of gradient-boosting algorithms in handling real-world telecommunication data and reinforce the importance of using representative datasets to validate churn prediction models.

### 2.3.2 Comparative Performance of Traditional, Ensemble, and Deep Learning Models

Saha et al. (2023) conducted one of the most comprehensive comparative evaluations of churn prediction models by testing a wide range of learning strategies, from traditional classifiers to advanced deep learning architectures. The study examined ensemble techniques such as AdaBoost, Random Forest, Extreme Randomized Trees, XGBoost, Gradient Boosting, and bagging and stacking methods alongside conventional approaches including Logistic Regression, Decision Trees, and k-Nearest Neighbors. Artificial Neural Networks and Convolutional Neural Networks were also included to assess the performance gap between classical machine learning and deep learning. Using two public datasets representing the Southeast Asian and American telecommunications markets, Saha et al. (2023) found that deep learning models delivered the strongest results. CNNs achieved accuracy scores of 99% and 98% on the Southeast Asian and American datasets respectively, while ANNs produced similarly high accuracies of 98% and 99%. These findings demonstrate that deep learning can capture complex behavioral patterns more effectively than ensemble or traditional classifiers, positioning it as a powerful approach for telecom operators seeking high-precision churn prediction in highly competitive markets.

### 2.3.3 Knowledge-Based and Text-Driven Feature Engineering

Shahabikargar et al. (2025) recent work offers an important advancement in churn prediction by shifting the focus from traditional demographic or usage-based variables to richer, knowledge-driven features extracted from customer interactions. The study emphasizes that understanding customer cognition, emotions, and behavioral signals is essential for anticipating churn more accurately than relying solely on structured CRM records. To address this gap, the researchers introduced the Customer Churn-related Knowledge Base (ChurnKB), a feature engineering framework that incorporates domain expertise, textual data mining techniques such as TF-IDF, cosine similarity, tokenization, and stemming, as well as generative AI models capable of interpreting unstructured text like emails. By integrating these knowledge-based features into machine learning models—including Random Forests, Logistic Regression, Multilayer Perceptrons, and XGBoost—the study demonstrated substantial performance gains. Notably, the F1-score of XGBoost improved from 0.5752 to 0.7891, illustrating the value of incorporating cognitive and behavioral indicators alongside conventional features. Beyond telecom churn management, the authors highlight that the same knowledge-based feature extraction approach can support broader applications such as personalized marketing, online harm detection, and mental health monitoring, signaling its potential as a versatile tool for business intelligence and digital safety

The last decade has seen remarkable progress in how customer churn is modeled, particularly with the rise of machine learning and deep learning approaches. Usman-Hamza et al. (2022) showed that decision-forest ensembles equipped with weighted voting can handle imbalanced telecommunication datasets more effectively than baseline classifiers, demonstrating strong improvements in predictive stability. In another strand of research, Saha et al. (2023) experimented with convolutional neural networks on mixed datasets from India and the United States, concluding that financial metrics such as return on investment should carry equal importance to raw predictive performance when evaluating model usefulness.

Geographical context continues to shape model behavior. In Nepal, Nagarkar (2022) used XGBoost on more than fifty thousand mobile subscribers and observed that recharge regularity and prolonged inactivity were the strongest signals of customer retention, enabling more targeted interventions. In the Syrian context, the introduction of social network analysis features pushed AUC values from 0.84 to 0.933, illustrating the influence of relational information on churn outcomes. More recently, Poudel and Sharma (2024) proposed models that integrate explainability tools directly into the prediction workflow, noting that transparency is essential for organizational acceptance of churn analytics.

A broader review by Shahabikargar et al. (2025) emphasized the rapid increase in external data integration and the need for interpretability alongside accuracy. Their synthesis underscores a growing shift toward models that not only perform well but can also justify their predictions to decision-makers.

## 2.4 Customer Churn in the Telecommunications Industry

Churn continues to pose a significant threat to revenue certainty and long-term profitability in the telecommunications sector. The industry statistics show that annual churn rates in some regions exceed 30%, even in highly developed markets Chang et al. (2024). The challenge is intensified in Sub-Saharan Africa, where mobile networks support essential services such as digital services like payments, remote education, remote jobs and online sales Mwaura (2021) where most of the communications relies on online platform hence the need to buy resources like data bundles for communication. Despite the growing body of churn research, very few studies directly address market conditions. This lack of contextual evidence highlights an important gap that limits the applicability of existing predictive approaches of machine learning models across the region.

## 2.5 Industry Relevance and Contribution

The review of prior research points to several consistent themes. Advanced machine learning techniques especially ensemble and deep learning methods tend to outperform traditional statistical models like survival analysis which mainly focuses on predicting the timing of churn rather than just whether churn happens. This analysis can be more complex to implement and also interpret compared to the model models like Random Forest or Light GBM when it comes to churn prediction. At the same time, organizations increasingly require models that balance predictive power with financial clarity, enabling business teams and stakeholders to interpret outputs and integrate them into decision processes as per as revenue is concerned. Another recurring insight is the disproportionate focus on other markets, leaving emerging economies underrepresented in current literature.

The framework proposed in this study responds directly to these gaps by prioritizing scalable modeling approaches, interpretability, and practical alignment with the operational realities of telecommunication providers.

## 2.6 Gaps in Existing Literature

**Limited Feature Diversity:** Many datasets rely heavily on transactional CRM variables such as usage volume, payment history, and subscription tenure while overlooking high-value indicators like financial components, Mwaura (2021). This restricts the richness of models and can weaken financial prediction quality.

**Narrow Evaluation Practices:** There is a strong preference for accuracy-based metrics, yet these do not always reflect the financial consequences of churn. Return on Investment linked indicators, including lifetime value measures, are rarely incorporated despite their relevance for managerial decision-making Shahabikargar et al. (2025).

**Interpretability Limitations:** Although advanced algorithms perform well, their outputs are not always intuitive for business users. The limited transparency often reduces trust in business stakeholders and slows the adoption of churn systems as practical decision-support tools Chang et al. (2024).

# 3. CHAPTER: METHODOLOGY

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to ensure a structured, repeatable, and scalable approach to customer churn prediction in the telecommunications sector, as illustrated in Figure 3.1. This reproducible approach shows that it can be plugged into any other telecommunications industry.

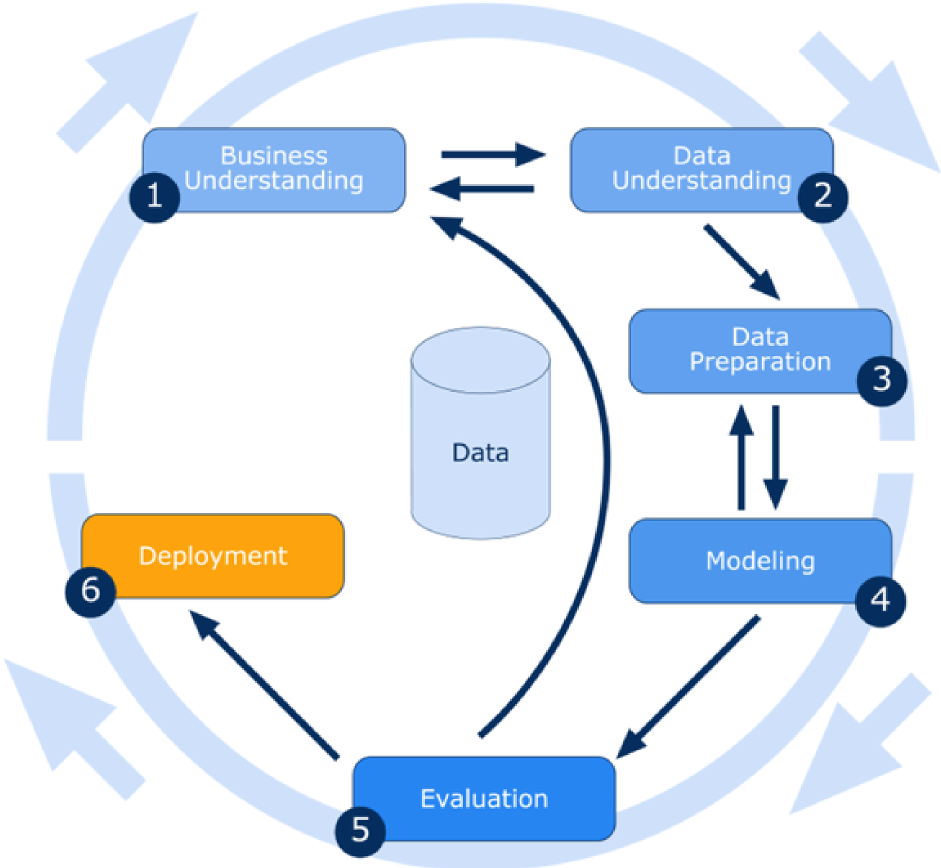


**Figure 3.1: Cross-Industry Standard Process for Data Mining (CRISP-DM)**

### 3.0.1 Proposed Workflow

The proposed workflow for this study is illustrated in Figure 3.2. It outlines the systematic process of building and evaluating machine learning models for customer churn prediction.

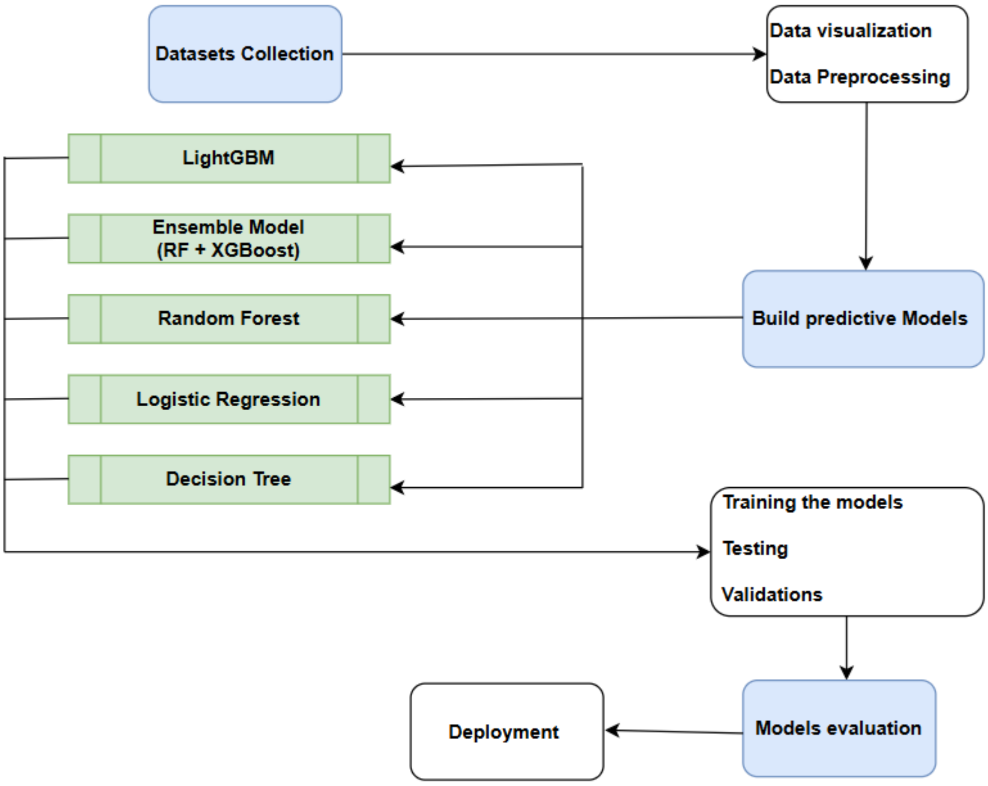


**Figure 3.2: Proposed Workflow**

## 3.1 Business Understanding

Telecommunications companies operate in a market where competition is intense and customer expectations shift rapidly as per their preferences. As subscription growth slows and price competition increases, retaining existing users becomes more important and cheaper method than acquiring new ones. For many operators, the financial burden of attracting new subscribers is far higher than maintaining current customers, making churn reduction central to business sustainability and operational. Despite continuous investment in marketing, loyalty programs, and service improvement initiatives, many operators still experience high churn levels that negatively affects revenue reliability and weaken long-term competitiveness. Understanding the drivers of churn and building predictive model that can anticipate customer exit has therefore become a strategic necessity in this industry.

## 3.2 Data Understanding

This phase will majorly concentrates on identifying, collecting, and analyzing datasets essential for achieving project goals. Tasks include collecting initial data, describing data, exploring data, and verifying data quality.

## 3.3 Data Collection

This study will use secondary data sourced from Bulgarian telecommunication operator which is public available on Mendeley data website Tokmakov (2024). This dataset comprises of customer's service usage, demographic information, pending amont others in which it will provide key indicators of customer churn allowing us to anticipate the behaviors that contribute to customer retentions and predict the main behaviour that will help us to retain customers into the network. The study will encourage iteration in model building and assessment until a good enough model is achieved.

## 3.4 Exploration Data Analysis

Exploratory Data Analysis (EDA) will be applied to understand variable distributions, identify class imbalance cases, and diagnose missing or inconsistent values. Graphical tools—including box plots, histograms, and correlation matrices will be used to examine relationships between various attributes. These insights will inform preprocessing decisions, feature engineering strategies, and the selection of modeling techniques to ensure a reliable and interpretable predictive framework that will be used.

## 3.5 Data Cleaning

Data cleaning is a fundamental step that ensures the dataset is consistent, accurate, and suitable for modeling. Given that the information is extracted from telecommunication systems, it may contain missing entries, duplicated records, and inconsistent categorical labels. The cleaning process will follow a structured approach: identifying missing values, applying appropriate imputation strategies, identifying duplicates, standardizing categorical fields, and reviewing outliers in the dataset.

A structured data cleaning workflow will be followed to ensure quality and consistency. Missing values in key numerical fields will be assessed for extent and pattern. For variables with minimal missingness, mean or median imputation will be applied, while variables with substantial or non-random missingness will undergo domain-informed estimation or exclusion. Duplicate records in the dataset based on unique identifiers will be identified and removed to prevent double-counting.

### 3.5.1 Treating Missing Data

Missing values will be handled based on their type and extent. Numerical attributes will be imputed using mean or median statistics, while categorical fields will use mode-based imputation. Variables with excessive or patterned missingness may undergo regression-based imputation or exclusion if they contribute minimal analytical value.

### 3.5.2 Treating Outliers

Outliers will be identified using statistical techniques such as z-scores and IQR-based thresholds. Depending on their origin, extreme values will be capped, transformed, or retained where they reflect genuine business behavior.

### 3.5.3 Data Type Conversion

All variables will be converted to appropriate data formats to ensure compatibility with analysis tools. Categorical variables will be encoded numerically, enabling their use in machine learning algorithms.

### 3.5.4 Data Transformation

#### Feature Engineering

Raw dataset will be refined to enhance the performance of machine learning models. This process will involve the creation of new features or modification of existing ones to extract meaningful information and patterns. Different techniques will contribute to the generation of features that better capture the underlying complexities of the data.

#### Feature Scaling

Numerical variables will undergo standardization or min–max scaling to ensure that models relying on distance metrics or gradient-based optimization behave consistently. The standardization methods like z-score normalization, Min-max scaling ,will bring features within a comparable scale, preventig dominant features from influeincing the model outcome.

#### Feature Selection

This step will involve correlation patterns, mutual information scores, and model-driven importance rankings which will be used to remove redundant or weak predictors. Tree-based algorithms such as Random Forests and LightGBM will guide feature prioritization.

#### Data Encoding

Categorical attributes will be encoded using either one-hot encoding, label encoding or binary encoding techniques depending on their hierarchy. This process will involve converting categorical data into a numerical format, allowing algorithms to interpret and utilize this information effectively.

## 3.6 Model Selection

Customer churn prediction is a supervised classification task in which the objective is to categorize instances into predefined classes based on their features. This study evaluates seven distinct modern algorithms Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Gradient Boosting (GB) to identify the most accurate and interpretable model for telecommunication data. Subsequently , the most suitable model is selected for predicting customer churn.

### 3.6.1 Logistic Regression (LR)

Logistic Regression estimates the probability that a customer will churn given an input vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$:

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp[-(\mathbf{w}^\top \mathbf{x}_i + b)]}. \tag{3.1}$$

The model parameters $(\mathbf{w}, b)$ are obtained by maximizing the log-likelihood of observing the training data. LR is selected for its simplicity, interpretability, and ability to quantify the effect of each attribute churn probability.

### 3.6.2 K-Nearest Neighbors (KNN)

KNN is a non-parametric classifier that assigns a new observation the majority class among its $k$ nearest samples. The predicted label is

$$\hat{y} = \text{mode}\{y_j \mid \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)\}, \tag{3.2}$$

where $\mathcal{N}_k(\mathbf{x}_i)$ denotes the $k$ closest neighbors according to Euclidean distance. KNN is robust to noisy data and effective when class boundaries are well separated.

### 3.6.3 Decision Tree (DT)

Decision Trees split the dataset recursively using an impurity metric such as the Gini index:

$$\text{Gini}(t) = 1 - \sum_{c=1}^{C} p_c^2, \tag{3.3}$$

where $p_c$ is the fraction of samples of class $c$ at node $t$. DTs offer visual transparency and help managers trace the rules leading to churn predictions.

### 3.6.4 Random Forest (RF)

Random Forest builds an ensemble of $B$ independent trees, each trained on a bootstrap sample of the data. The final prediction is given by majority voting:

$$\hat{y} = \text{mode}\{h_b(\mathbf{x})\}_{b=1}^{B}. \tag{3.4}$$

RF mitigates overfitting, handles large heterogeneous datasets, and provides internal estimates of feature importance.

### 3.6.5 Support Vector Machine (SVM)

SVM constructs a hyperplane that maximizes the margin between classes:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1,\ \forall i. \tag{3.5}$$

Its kernel functions allow modeling non-linear decision boundaries, making it suitable for complex churn relationships.

### 3.6.6 Naive Bayes (NB)

Naive Bayes applies Bayes' theorem under the assumption of feature independence:

$$P(y|\mathbf{x}) = \frac{P(y)\prod_{j=1}^{p} P(x_j|y)}{P(\mathbf{x})}. \tag{3.6}$$

Despite its simplicity, Naive Bayes performs well on high-dimensional datasets and offers rapid classification, an advantage for large telecommunication databases.

### 3.6.7 Gradient Boosting (GB)

Gradient Boosting iteratively builds a strong learner by combining multiple weak decision trees. At each iteration $m$, residuals are computed as

$$r_{im} = -\frac{\partial \mathcal{L}(y_i, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)}, \tag{3.7}$$

and a new tree $h_m(\mathbf{x})$ is fitted to these residuals. The model update follows

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta\, h_m(\mathbf{x}), \tag{3.8}$$

where $\eta$ is the learning rate. Gradient Boosting excels in predictive accuracy and adapts to complex nonlinear patterns often present in churn data.

### 3.6.8 Justification for the Selected Classifiers

The decision to employ five classification algorithms—Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Gradient Boosting (GB) was made to ensure a balanced assessment of both interpretable and high-performing predictive models. Each algorithm contributes unique analytical value and addresses different aspects of the customer churn phenomenon, which is inherently multidimensional in nature.

Logistic Regression is adopted as the baseline model because of its simplicity, transparency, and effectiveness in binary classification problems. It estimates the likelihood of a customer discontinuing service based on explanatory variables. LR provides interpretable coefficient weights that help managers understand the marginal influence of each variable on churn probability, making it highly suitable for initial benchmarking.

KNN offers an instance-based, non-parametric approach where each new observation is classified by the majority label among its nearest neighbors in feature space. This method captures local behavioral similarities among customers for example, subscribers with comparable usage or spending patterns tend to exhibit similar churn tendencies. Although computationally intensive on large datasets, KNN's flexibility in modeling non-linear decision boundaries adds valuable comparative insight.

Decision Trees are selected for their interpretability and intuitive structure. The model recursively partitions data using feature-based thresholds, creating a visual hierarchy of decision rules. In a churn context, a DT might reveal that customers who are most at risk. Its rule-based nature enables straightforward explanation to non-technical stakeholders and supports managerial decision-making.

The Random Forest algorithm, an ensemble of multiple Decision Trees, is incorporated to improve predictive accuracy and reduce overfitting. By aggregating the outputs of many trees trained on random subsets of data and features, RF captures complex, non-linear relationships while maintaining robustness to noise. Furthermore, its feature importance scores provide quantitative measures of the most influential churn drivers.

SVM is included for its strong generalization capacity, particularly in high-dimensional feature spaces. By constructing an optimal hyperplane that maximizes the margin between classes, SVM performs well even with overlapping class boundaries. Kernel functions further enhance its ability to model non-linear relationships in customer behavior, offering an advanced benchmark against tree-based models.

Naive Bayes contributes a probabilistic perspective grounded in Bayes' theorem. Despite its simplifying assumption of feature independence, it is computationally efficient and performs remarkably well with large-scale categorical data. Its ability to quickly estimate churn probabilities makes it ideal for initial screening of customers at risk, particularly when rapid, low-cost predictions are desirable.

Gradient Boosting is chosen for its capacity to deliver state-of-the-art predictive accuracy by combining multiple weak learners in a sequential, error-correcting manner. Each new tree in the sequence focuses on the residual errors of the previous ensemble, allowing the model to learn complex patterns. Although parameter tuning is essential to prevent overfitting, GB's flexibility and performance make it indispensable in churn analytics. Moreover, its compatibility with interpretability frameworks such as SHAP enables a clear understanding of feature contributions.

These seven algorithms collectively span the major paradigms of supervised learning—linear, probabilistic, instance-based, tree-based, and boosting models—allowing for a comprehensive comparison of predictive behavior. The inclusion of both interpretable (LR, DT) and high-performing (RF, GB, SVM) algorithms ensures a balanced evaluation of accuracy, computational efficiency, and managerial usability. Through this comparative analysis, the study aims to select the model that best captures customer behavior dynamics while maintaining practical relevance for operational decision-making in the telecommunications sector.

### 3.6.9 Handling Class Imbalance with SMOTE

Given that churners form a smaller proportion of the dataset, SMOTE will generate synthetic minority samples: In the dataset, the number of churners is expected to be substantially lower than non-churners, leading to class imbalance. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) will be applied to the training data. For each minority instance $x_i$, SMOTE generates a synthetic instance $x_{\text{new}}$ as:

$$x_{\text{new}} = x_i + \delta \times (x_{nn} - x_i), \quad \delta \sim U(0,1) \tag{3.9}$$

where $x_{nn}$ is one of the $k$ nearest neighbors of $x_i$. This process ensures balanced class representation, improving the model's ability to identify churners without bias toward the majority class.

This ensures more balanced training and reduces model bias toward the majority class.

## 3.7 Model Evaluation

The models will be evaluated using an 80/20 train–test split, employing stratified sampling to preserve the proportion of churn and non-churn classes. Model performance will be assessed using key classification metrics, including Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC–ROC):

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.10}$$

where $TP$, $FP$, and $FN$ denote true positives, false positives, and false negatives, respectively. Precision indicates how many predicted churners are actual churners, while recall measures how many actual churners were correctly identified. The F1-score balances both, and a high AUC–ROC value indicates effective discrimination between churners and non-churners.

Given that the business cost of false negatives (failing to identify actual churners) is higher than that of false positives, recall will be prioritized in the final model selection. The model with the optimal trade-off between recall and precision ideally supported by explainability techniques such as SHAP will be selected for deployment.

## 3.8 Model Deployment

The best-performing model will be deployed as an API-driven service using FastAPI. This setup enables high-speed inference and seamless integration with systems. When customer data is submitted, the API will return a churn probability and accompanying interpretability insights derived from SHAP values.

These predictions will feed into a dashboard that segments customers by risk level and highlights key churn drivers. The modular design supports regular model retraining to reflect emerging behavioral trends, ensuring that the churn detection system remains adaptive and actionable for business teams.

The API-driven architecture ensures modularity, scalability, and continuous learning, allowing for retraining with new customer data as behavioral trends evolve. This approach bridges predictive analytics and business decision making, enabling proactive interventions that enhance customer retention and revenue stability in the telecommunications sector.

# 4. CHAPTER: EXPECTED OUTCOMES

The proposed study seeks to develop a robust, machine learning–driven framework for predicting customer churn using the Bulgarian telecommunication dataset. By leveraging real-world customer data, the study will generate actionable insights that can be adapted to any telecommunications market. The outcomes will specifically address key research gaps related to feature diversity, model interpretability, and business integration of predictive analytics.

In relation to the first objective, which involves identifying challenges and limitations in existing churn prediction literature, the study will provide a comprehensive analysis of prior approaches, highlighting deficiencies such as inadequate handling of class imbalance, low interpretability of black-box models, and limited generalization across market contexts. This review will establish the foundation for a more transparent and transferable modeling framework.

Aligned with the second objective, the research will develop predictive models capable of capturing customer churn behavior through advanced machine learning algorithms. The models will integrate variables to create a multidimensional representation of churn risk. Feature engineering and class re-balancing using SMOTE will further enhance the dataset's predictive strength and fairness.

In fulfilling the third objective, the study will rigorously evaluate and compare multiple machine learning algorithms Logistic Regression, Decision Tree, Random Forest, and LightGBM based on performance metrics such as recall, precision, F1-score, and ROC-AUC. Emphasis will be placed on maximizing recall to ensure that high-risk churners are correctly identified, while maintaining interpretability through model explainability tools such as SHAP (SHapley Additive exPlanations).

The best performing model will be deployed through a FastAPI-based predictive dashboard. The dashboard will display churn probabilities, key explanatory features, and visual analytics for decision support, enabling managers to design proactive retention campaigns grounded in data-driven insights.

Overall, the study is expected to produce a scalable, interpretable, and business oriented churn prediction framework that enhances strategic decision making and customer retention in the telecommunications sector. By combining predictive accuracy with explainable AI and seamless system integration, the research will minimize revenue leakage, improve marketing efficiency, and strengthen customer relationships. Beyond Bulgaria, the framework offers a replicable model for other telecommunication markets seeking to transition toward evidence based, customer centric operations.

# References

Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. *Algorithms*, *17*(6), 231. Retrieved from https://research.tees.ac.uk/files/86885815/algorithms-17-00231.pdf doi: 10.3390/a17060231

Khan, M. T., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Customer churn behaviour in the telecommunication industry. *Algorithms*, *17*(6), 231. Retrieved from https://doi.org/10.3390/a17060231 doi: 10.3390/a17060231

Mwaura, E. T. (2021). *Effects of customer experience management on customer churn in the telco industry in kenya.* Project Report. Retrieved from https://afribary.com/works/effects-of-customer-experience-management-on-customer-churn-in-the-telco-industry-in-kenya (Accessed: 2025-09-27)

Nagarkar, P. (2022). A customer churn prediction model using xgboost for the telecommunication industry in nepal. *Procedia Computer Science*, *215*, 652–661. Retrieved from https://www.sciencedirect.com/science/article/pii/S187705092202138X doi: 10.1016/j.procs.2022.12.083

Poudel, S., & Sharma, S. (2024). Explaining customer churn prediction in telecom industry using interpretable machine learning techniques. *Information Sciences Letters*, *36*, 100043. Retrieved from https://www.sciencedirect.com/science/article/pii/S2666827024000434 doi: 10.1016/j.isl.2024.100043

Rahman, M., Bista, R., & Poudel, K. (2024). Explaining customer churn prediction in the telecom industry using interpretable machine learning techniques. *Information Sciences Letters*, *36*, 100043. Retrieved from https://www.sciencedirect.com/science/article/pii/S2666827024000434 doi: 10.1016/j.isl.2024.100043

Saha, L., Tripathy, H. K., Gaber, T., El-Gohary, H., & El-Kenawy, E. S. M. (2023). Deep churn prediction method for telecommunication industry. *Sustainability*, *15*(5), 4543. Retrieved from https://www.mdpi.com/2071-1050/15/5/4543 doi: 10

.3390/su15054543

Shahabikargar, M., Beheshti, A., Mansoor, W., Zhang, X., Foo, E. J., Jolfaei, A., ... Shabani, N. (2025). Churnkb: A generative ai-enriched knowledge base for customer churn feature engineering. *Algorithms*, *18*(4), 238. Retrieved from `https://www.mdpi.com/1999-4893/18/4/238` doi: 10.3390/a18040238

Tokmakov, D. (2024). *Customer churn dataset containing real records from a leading bulgarian telecom operator, specifically for business customers (version 1).* Mendeley Data. Retrieved from `https://data.mendeley.com/datasets/nrb55gr66h/1` (Licensed under CC BY 4.0) doi: 10.17632/nrb55gr66h.1

Usman-Hamza, F. E., Balogun, A. O., Capretz, L. F., Mojeed, H. A., Mahamad, S., Salihu, S. A., ... Salahdeen, N. K. (2022). Intelligent decision-forest models for customer churn prediction. *Applied Sciences*, *12*(16), 8270. Retrieved from `https://www.mdpi.com/2076-3417/12/16/8270` doi: 10.3390/app12168270

Zhang, T., Moro, S., & Ramos, R. F. (2022). A data-driven approach to improve customer churn prediction based on telecom customer segmentation. *Future Internet*, *14*(3), 94. Retrieved from `https://www.mdpi.com/1999-5903/14/3/94` doi: 10.3390/fi14030094

# 5. APPENDICES

## 5.1 Project Appendix

This appendix provides essential details regarding the project's condensed execution schedule and the financial resources required for its successful completion.

### 5.1.1 Project Timeline: 4-Month Chart

The proposed research will follow an intensive, structured four-month schedule, emphasizing rapid progression from methodology design to the final dissertation submission.

P1: Phase I: Design and Literature Review P2: Phase II: Data Preprocessing and Feature Extraction P3: Phase III: Model Development and Training P4: Phase IV: Evaluation, Analysis, and Final Drafting

## 5.2 Project Budget

The project budget below outlines the estimated costs required for the development, analysis, and dissemination of the machine learning based telecommunication customer churn prediction framework. All estimates are presented in Kenya Shillings (KSh) and are based on current institutional and market rates.

### 5.2.1 Estimated Project Budget (KSh)

| Category | Estimate | Estimated Cost (KSh) |
|---|---|---|
| A. Computational Resources | | |
| Data Storage and Access | 250 GB cloud storage | 5,000 |
| Software Licenses | Annual license for data visualization | 10,000 |
| Subtotal A | | 15,000 |
| B. Administration | | |
| Ethical and Institutional Review | Standard university ethics | 10,000 |
| Internet and Communication | high-speed internet @ KSh 4,000/month | 20,000 |
| Overleaf/LaTeX Subscription | Overleaf Premium plan | 6,000 |
| Subtotal B | | 36,000 |
| C. Data and Dissemination | | |
| Data Access | Use of public dataset | 0 |
| Publication Fees | Open-access journal submission fees | 40,000 |
| Subtotal C | | 40,000 |
| D. Equipment and Materials | | |
| External SSD Backup Drive | 1 TB external drive | 10,000 |
| Stationery and Printing | Report printing | 8,000 |
| Subtotal D | | 18,000 |
| Subtotal (A + B + C + D) | | 109,000 |
| Miscellaneous / Contingency (10) | unforeseen costs | 30,000 |
| Total Estimated Budget | | 139,000 |

**Table 5.1: Proposal budget**