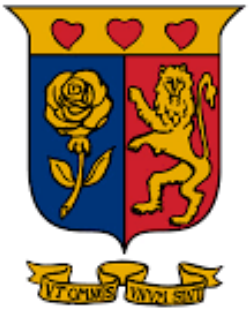


Strathmore
UNIVERSITY

Predicting Customer Churn in the Telecommunications Industry using Machine Learning Techniques

By
Adeline Makokha
191199

A Proposal Submitted in Partial Fulfillment of The Requirements for The Degree of Master of Science in Data Science and Analytics, @iLabAfrica, Strathmore University

Strathmore Institute of Mathematical Sciences
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.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University

Student's Name: **Adeline Makokha**

Sign:

27 September 2025

APPROVAL

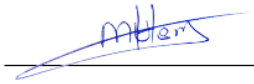
The proposal was reviewed and approved for defense by

Dr. Henry Muchiri

School of Computing and Engineering Sciences

Strathmore University

Sign:



13 October 2025

Abstract

In the recent telecommunications industry, predicting customer churn is a recurring issue 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 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 traditional models and modern models will be used. These includes algorithms like logistic regression, decision trees, naive Bayes (NB), support vector machine (SVM), random forest, k-nearest neighbors (KNN), and gradient boosting (GB). Interpretability approaches like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) will also 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

ANNs – Artificial Neural Networks

API – Application Programming Interface

AUC – Area Under the Curve

CNNs – Convolutional Neural Networks

CRISP-DM – Cross-Industry Standard Process for Data Mining

CRM – Customer Relationship Management

DL – Deep Learning

DT – Decision Tree

ERT – Extra Random Trees

FT – Functional Trees

GB – Gradient Boosting

k-NN – k-Nearest Neighbors

LightGBM – Light Gradient Boosting Machine

Lime – Local Interpretable Model-agnostic Explanations

LMT – Logistic Model Tree

LR – Logistic Regression

ML – Machine Learning

NB – Naive Bayes

RF – Random Forest

ROC – Receiver Operating Characteristic

ROI – Return on Investment

SHAP – SHapley Additive exPlanations

SMOTE – Synthetic Minority Over-sampling Technique

SVM – Support Vector Machine

Contents

CHAPTER ONE

1 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 rapidly growing areas within the broader telecommunications company. 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 which decided to increase their prices while others have reduced their prices, leading to a flexible shift of customers preferring other networks. At the core of the issue lies the financial burden of customer acquisition, as gaining a new subscriber typically requires significantly more investment than keeping an existing one, making effective churn management an essential business priority for telecommunication providers.
From a financial standpoint, keeping existing subscribers is far less costly than attracting new ones. Research indicates that the expense of acquiring new users may range between five to seven times higher than that of maintaining loyal 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 impact on mobile services, including calls and data usage, has really led to high competition in the telecommunications market. By combining demographic information, revenue information, and service usage, the customer can be used to intervene in the customer retention strategies, which will save on marketing expenditure and improve customer loyalty to one network.

1.2 Problem Statement

Industries have really intensified the increase in customer expectations, and customers are preferring affordable and reliable services. Most of the customers have left the network and prefer other telecommunications service providers that offer good and affordable service qualities. Most of the previous churn rate predictions remain reactive, that is, relying on static indicators and manual analysis, which can be inefficient, prone to human bias, and also time-consuming to come up with business decisions. This existing prediction also has limitations 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 have resulted in revenue leakage due to the high customer churn rate in the industry. Mwaura (2021).

1.3 Research Objectives

The purpose of this study is to address the existing gaps in customer churn prediction within the telecommunications industry through the following objectives:

1.3.1 Main Objectives

To develop a robust machine learning–based framework for predicting customer churn in the telecommunications industry.

1.3.2 Specific Objectives

i. To examine methodological and conceptual limitations in existing telecommunications churn prediction studies.
ii. To develop a customer churn prediction model using selected machine learning algorithms and telecommunication customer data.
iii. To evaluate the predictive performance of the developed models using standard classification metrics.
iv. To deploy the optimal churn prediction model within an interactive and interpretable dashboard to support managerial decision-making.

1.4 Research Questions

In alignment with the stated objectives, this study will be guided by the following research questions:
i. What methodological and data-related limitations characterize existing churn prediction models in the telecommunications industry?
ii. Which customer attributes, such as demographic, contractual, financial, or behavioral factors, have the strongest influence on churn likelihood?
iii. How can machine learning algorithms be developed, optimized, and validated to enhance the accuracy of churn prediction?
iv. What differences exist in the predictive performance of various machine learning approaches (e.g., logistic regression, random forest)?
v. In what ways can a predictive churn model be integrated into an interactive dashboard to support strategic decision-making and customer retention initiatives?

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 for customer retention have really reduced the high cost that is put in place to acquire new customers in the network. This has also stabilized the telecommunications revenue stream and has safeguarded the market share as well as enhanced a long-term customer value relationship with the industry.

The implications of customer churn have extended beyond the corporate world and are now also impacting the emerging markets. This is evident in our day-to-day lives through e-learning systems, e-health systems, and small businesses' operations, where telecommunication services have a major impact and are one of the drivers of the economy. There is 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: The 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 environments, the proposed framework offers a methodological blueprint that can be adapted for other telecommunication markets and industries facing similar churn-related challenges.

This research seeks to advance both practical and theoretical understanding of customer churn within the telecommunications sector. It aims to establish a sustainable, data-informed framework that integrates machine learning techniques into day-to-day business operations to enhance customer retention.

CHAPTER TWO

2 LITERATURE REVIEW

There has been a lot of shift over the past years in the digital evolution of technology; there are a lot of markets and startups that have opened up, and as a result, there is customer demand and more personalized quality services in the telecommunications industry. Most of these changes have really widened access to communication and led to the need for improvement. With this highly competitive ecosystem of customer churn, keeping the existing customers has been more critical than acquiring new customers into the network. One of the main challenges is the subscriber or customer making the decision to stop using the network, for 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 have significantly translated to revenue erosion in the industry; hence the need to develop and adopt a data-driven approach that can curb customer churn Nagarkar (2022).

2.1 Theoretical Framework

2.1.1 Customer Relationship Management

Shahabikargar et al. (2025) emphasized that maintaining customer loyalty is a strategic pillar of sustainable business growth. In today's competitive markets, organizations invest significant resources not only to attract new customers but also to retain existing ones. However, customer churn continues to undermine profitability, with even modest increases in churn rates resulting in major revenue losses. Effective Customer Relationship Management (CRM) requires data-driven insights into customer behavior, satisfaction, and cognitive patterns that indicate disengagement or dissatisfaction before actual churn occurs.

The authors highlighted that the success of any machine learning (ML) based churn prediction system largely depends on feature engineering, the process of selecting and transforming variables that best describe customer behavior. Traditional CRM churn studies have typically relied on demographic, product usage, and financial attributes (such as tenure, billing, and average revenue per user). While these features provide quantitative measures, they often overlook the qualitative dimensions of customer-company interaction, such as sentiment, tone, and intent expressed in text communications.

To address this limitation, Shahabikargar et al. (2025) proposed a Customer Churn-related Knowledge Base (ChurnKB) which is a domain-driven framework for enhancing feature engineering by incorporating unstructured textual data from customer-generated content like emails, chat transcripts, and feedback forms. This marked a significant shift from purely numeric features to text-derived behavioral indicators, expanding the predictive capability of traditional CRM systems.

ChurnKB leverages a combination of Natural Language Processing (NLP) techniques such as Term Frequency–Inverse Document Frequency (TF–IDF), cosine similarity, regular expressions, word tokenization, and stemming, to extract churn-relevant linguistic patterns. These textual features help uncover early signals of dissatisfaction or disengagement, such as complaints, negative sentiment, or cancellation intent.

To further enhance its adaptability, the authors integrated Generative AI, particularly Large Language Models (LLMs), into ChurnKB's pipeline. This integration enables the discovery of latent patterns in unstructured text, improving the system's ability to capture subtle emotional and cognitive cues related to churn risk. Additionally, ChurnKB includes feedback loops that refine the extracted features over time, ensuring continuous learning and alignment with evolving customer behavior.

The study tested the impact of knowledge-enhanced features on several ML algorithms, including Logistic Regression, Random Forest, Multilayer Perceptron (MLP), and XGBoost. When ChurnKB derived features were incorporated, performance improved significantly, most notably in the XGBoost model, where the F1-score increased from 0.5752 to 0.7891. This demonstrates how knowledge-based feature enrichment can substantially enhance CRM systems' ability to identify at-risk customers and enable earlier intervention.

Moreover, the framework's applicability extends beyond churn prediction, it can support related applications such as personalized marketing, sentiment tracking, cyberbullying detection, hate speech recognition, and mental health monitoring. This broad relevance underscores the versatility of combining CRM principles with AI-driven behavioral analytics.

Shahabikargar et al. (2025) concluded that while integrating LLMs into CRM-driven churn prediction enhances feature richness, it also raises challenges related to data governance, ethical AI use, and model interpretability. The authors noted that future research should focus on explainable AI (XAI) techniques to ensure that business users can understand and act upon AI outputs confidently. Furthermore, more work is needed to validate such models across sectors and regions, ensuring that CRM-driven insights remain contextually relevant in telecommunication markets.

2.1.2 Predictive Modeling Using Statistical and Segmentation

The telecommunications sector is among the industries most affected by customer attrition, as even a small rise in churn directly reduces profit margins. Jahromi and Sharifi (2020)underscored that retaining existing clients is considerably more cost-effective than acquiring new ones, emphasizing the importance of early churn detection through predictive modeling. Their study focused on understanding the behavioral and demographic characteristics that distinguish loyal subscribers from those likely to leave, providing a quantitative foundation for proactive customer retention strategies.

The research utilized customer datasets obtained from three major Chinese telecommunication operators, incorporating a range of demographic, usage, and service-related features. Before predictive modeling, the researchers applied customer segmentation, a crucial step for identifying homogeneous groups with similar churn tendencies. This segmentation process facilitated the development of differentiated marketing and retention interventions tailored to each customer group's unique behavioral profile.

The study found that the logistic regression–based model achieved superior predictive accuracy compared to the Fisher discriminant model. Specifically, the regression approach attained 93.94% accuracy, outperforming the discriminant analysis, which exhibited lower generalization performance. This result highlighted the logistic model's robustness in handling categorical predictors and non-linear relationships between variables. The findings confirm that even traditional statistical models can achieve strong performance when combined with effective segmentation and variable selection.

The research Zhang et al. (2022) demonstrated that integrating segmentation with predictive modeling enables telecommunication companies to identify high-risk customers with precision. Such insights allow targeted retention actions, including personalized offers, service upgrades, and loyalty incentives. From a business intelligence standpoint, the study reinforced that statistical models remain valuable for operational decision-making due to their interpretability, simplicity, and transparency qualities often sought in customer relationship management (CRM) systems.

While the study achieved high accuracy, it was based primarily on structured, numeric data and lacked behavioral variables such as text interactions or complaint history. This creates a methodological gap for future research to incorporate hybrid modeling approaches that combine traditional regression with

machine learning or AI-driven segmentation. Such integration would enable dynamic adaptation to changing customer behaviors in rapidly evolving telecommunication markets.

2.1.3 Forest Models and Ensemble Strategies

In recent years, ensemble-learning frameworks, particularly decision-forest models, have become the backbone of customer-churn prediction in telecommunications. Usman-Hamza et al. (2022) argued that single rule-based models, while interpretable, lack the scalability and robustness required for high-volume customer datasets. Their study highlighted that the imbalanced nature of churn data often limits the predictive strength of conventional algorithms such as Logistic Regression or single Decision Trees, motivating the use of ensemble strategies that combine multiple learners to enhance stability and accuracy.

The authors developed a suite of intelligent Decision-Forest (DF) architectures, including the Logistic Model Tree (LMT), Random Forest (RF), and Functional Tree (FT). Each model aggregates multiple weak learners to minimize variance and overfitting. This hybridization allows the ensemble to exploit the complementary strengths of diverse classifiers.

A central contribution of the work lies in addressing class imbalance, a common characteristic of telecommunication datasets where churners represent a small minority. The DF ensemble mitigates imbalance effects by combining heterogeneous trees that each learn different minority-class boundaries, yielding better recall for churn instances without over-penalizing non-churn predictions. This makes the model particularly suitable for large-scale, heterogeneous customer data environments.

Using publicly available benchmark datasets, the enhanced DF ensembles achieved superior predictive accuracy relative to traditional ML methods. Weighted stacking and soft-voting produced the most stable outcomes, outperforming baseline Logistic Regression, standalone Random Forest, and other benchmark classifiers in both precision and recall. The results confirmed that forest-based ensembles offer optimal solutions for churn-classification problems characterized by noise, non-linearity, and imbalance.

The study demonstrated that ensemble decision forests can generate interpretable and high-performing churn predictions at enterprise scale. By deploying these models, telecommunication companies can more effectively detect early churn signals, prioritize at-risk segments, and tailor retention campaigns. Furthermore, the research recommended extending DF-based models to other machine-learning tasks within telecommunication analytics, such as usage forecasting and cross-selling, due to their adaptability and robustness.

While the DF framework achieved high accuracy, Usman-Hamza et al. (2022) emphasized the need for explainable ensemble models to strengthen managerial trust. Future work should incorporate techniques such as SHAP or LIME for feature-importance interpretation and explore hybrid integration with deep learning to further improve sensitivity toward subtle churn indicator.

2.1.4 Machine Learning Approaches for Churn Prediction

Sohail et al. (2020)investigated the application of supervised machine learning algorithms to predict customer churn in the telecommunications sector. Their research emphasized that modern telecommunication companies generate extensive customer data—ranging from call records and billing transactions to service usage patterns, which can be leveraged to identify customers likely to discontinue services. The study compared traditional decision tree algorithms, namely Linear Decision Tree (LDT) and Unpruned Decision Tree (UDT), with more advanced methods such as Random Forest (RF) and Support

Vector Machine (SVM).

The authors observed that earlier systems relied heavily on decision tree algorithms that trained models on a wide range of features, many of which were irrelevant or redundant. This redundancy resulted in increased computational time, model overfitting, and reduced predictive accuracy, with performance peaking around 84%. These findings underline the weakness of unfiltered attribute selection and the need for feature optimization in predictive modeling.

To address these limitations, Sohail et al. (2020) proposed integrating Random Forest and Support Vector Machine classifiers. Random Forest was preferred for its ensemble nature and ability to perform automatic feature selection through Gini importance, thereby reducing the inclusion of irrelevant variables. On the other hand, SVM was utilized to maximize class separation and improve boundary precision between churners and non-churners. The hybrid model achieved 95% classification accuracy, representing a substantial improvement over traditional decision tree methods. This result demonstrated that the combination of ensemble and margin-based classifiers could significantly enhance churn prediction performance in telecommunication datasets.

The study concluded that effective churn prediction models directly support operational and strategic decisions within telecommunication companies. Higher model accuracy enables better identification of at-risk customers, allowing companies to implement proactive retention strategies such as targeted offers, service quality enhancements, and personalized communication. Moreover, the focus on selecting important attributes contributes to computational efficiency, which is critical for scaling predictive systems in large customer databases.

While Sohail et al. (2020) demonstrated improved accuracy, the study primarily relied on structured data and did not incorporate real-time analytics or unstructured sources such as customer feedback or sentiment data. This presents an opportunity for further research integrating Explainable AI (XAI) techniques and multimodal data fusion to enhance both interpretability and generalization across markets.

2.2 Empirical Studies

2.2.1 Handling Class Imbalance in Customer Churn Prediction

Nagarkar (2022) examined customer churn prediction in the telecommunications sector, focusing on how the imbalanced distribution of churn and non-churn cases affects the performance of machine learning models. The author noted that most telecommunication datasets contain far fewer churners compared to active customers, which biases models toward predicting the majority class and leads to misleading accuracy metrics. This imbalance poses a serious challenge to developing reliable and actionable churn prediction systems.

The study used real customer data from a major telecommunications operator in Nepal, consisting of 52,332 customer records, out of which 6,128 were churners and 46,204 were non-churners. This native dataset was supplemented with a public dataset of 3,333 subscribers for cross-validation and benchmarking against prior research. Such a design enabled the researcher to test algorithmic generalization across both proprietary and open data sources—an approach that strengthened the study's external validity.

To address the imbalance issue, Extreme Gradient Boosting (XGBoost) was employed. XGBoost is known for its robustness in handling complex, noisy, and skewed data. It enhances predictive accuracy through iterative learning, where each new tree corrects the residual errors from the previous ones. The author applied XGBoost both to the native and public datasets, demonstrating how gradient boosting can effectively minimize bias toward the majority class while maintaining high recall for churners.

The performance results were impressive: On the native Nepalese dataset, the model achieved an accuracy of 97% and an F1-score of 88%. On the public dataset, it achieved 96.25% accuracy and an F1-score of 86.34%, outperforming previously published models using the same data. These results validated the efficacy of XGBoost in learning from imbalanced data distributions while maintaining generalization capability across different datasets.

The study emphasized that models should not be evaluated solely on accuracy when dealing with imbalance. Metrics such as F1-score, recall, and ROC-AUC are more appropriate for assessing performance, as they reflect how well the model detects minority class instances (i.e., churners). Furthermore, the research highlighted the importance of context-specific datasets, showing that real-world telecommunication data often behaves differently from public datasets, underscoring the need for model calibration and domain adaptation.

While Nagarkar (2022) demonstrated the strong performance of XGBoost, the study primarily focused on a single model and did not compare results with oversampling or hybrid ensemble techniques (e.g., SMOTE combined with Random Forest). Additionally, the research did not incorporate explainability to interpret why customers churned, which is essential for operational use. Future work can address these limitations by integrating Explainable AI (SHAP/LIME) and resampling approaches such as SMOTE or ADASYN to further mitigate imbalance and improve interpretability.

2.2.2 Hybrid Approaches to Churn Prediction

Mishra et al. (2019) investigated a hybrid framework for predicting and analyzing customer churn in the telecommunications industry, emphasizing the complementary use of classification and clustering techniques. The authors argued that while many prior studies focused solely on classification accuracy, effective churn management also requires understanding why customers leave, enabling firms to design targeted retention strategies. Their study was grounded on the principle that retaining customers is far more cost-efficient than acquiring new ones, a consistent finding in CRM and telecommunications research.

Given the massive volume of customer data generated daily in telecommunication operations, Mishra et al. (2019) highlighted the need for effective feature selection to improve model performance and interpretability. They employed Information Gain and Correlation Attribute Ranking filters to identify the most influential variables driving churn. This preprocessing step reduced redundancy and noise, ensuring that only meaningful attributes such as service usage frequency, billing trends, and complaint history were used in the modeling process. This demonstrates the critical role of dimensionality reduction in improving both computational efficiency and model generalization.

For classification, the researchers compared multiple machine learning algorithms, including Random Forest (RF), Naïve Bayes, and Decision Tree (DT). Among these, the Random Forest classifier achieved the best performance, correctly classifying 88.63% of instances. The evaluation metrics used included Accuracy, Precision, Recall, F-score measure, and the ROC-AUC, ensuring a comprehensive assessment of predictive reliability. The authors concluded that ensemble models such as Random Forest outperform single classifiers in handling complex telecommunication datasets with mixed variable types and potential class imbalance.

Beyond classification, the study introduced a customer segmentation phase using k-means clustering based on cosine similarity. This approach grouped churned customers according to behavioral similarities, allowing the telecommunication operator to design group-specific retention offers. For example, one cluster might represent high-spending users sensitive to pricing, while another could include low-engagement customers needing service quality improvements. This dual-stage (classifica-

tion–clustering) method not only improved churn prediction accuracy but also enhanced actionable insight generation for marketing teams.

Mishra et al. (2019) underscored the necessity of integrating classification performance with post-prediction analysis to inform strategic interventions. However, the study's model did not include explainable AI (XAI) mechanisms such as SHAP or LIME, limiting interpretability for non-technical CRM users. Additionally, while segmentation improved marketing strategy design, it did not account for temporal behavior dynamics, suggesting a research gap in time-aware churn modeling. The authors' focus on structured data also leaves room for future studies incorporating unstructured or real-time sources, such as social media feedback or call-center transcripts.

2.2.3 Model Optimization and Deep Learning

Saha et al. (2023) explored the growing need for predictive systems capable of accurately identifying potential churners while maintaining business profitability in competitive telecommunications markets. The authors recognized that traditional churn models, while computationally efficient, often compromise either accuracy or interpretability when applied to large-scale, dynamic telecommunication data. To overcome this limitation, their research systematically evaluated multiple machine learning and deep learning paradigms to identify the optimal approach for churn prediction across diverse datasets.

The study examined three broad categories of models:

Traditional Machine Learning Algorithms, including Logistic Regression (LR), Decision Tree (DT), and k-Nearest Neighbors (kNN). Ensemble Learning Methods, such as AdaBoost, Random Forest (RF), Extreme Randomized Trees (ERT), XGBoost (XGB), Gradient Boosting Machines (GBM), Bagging, and Stacking. Deep Learning Models, specifically Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs).

This structured comparison allowed the researchers to test whether complex deep learning architectures could outperform established ensemble techniques without sacrificing computational efficiency or interpretability.

Two publicly available datasets were used: One representing a Southeast Asian telecommunications provider, and another from an American telecommunications operator. Both datasets included attributes such as contract type, monthly charges, tenure, and service usage, representing typical customer–operator interactions. Each dataset was split into training and test sets, and cross-validation was used to ensure model reliability. The study's design emphasized the balance between accuracy and profit-oriented performance, highlighting the importance of choosing algorithms that optimize both business and predictive metrics.

Among all models tested, deep learning approaches (CNN and ANN) demonstrated the highest predictive performance. On the Southeast Asian dataset, CNN achieved 99% accuracy, while ANN reached 98%. Similarly, on the American dataset, ANN slightly outperformed CNN, with 99% and 98% accuracy, respectively. These findings suggest that deep neural architectures can effectively capture complex nonlinear relationships within telecommunication data that simpler models may overlook.

The study also reaffirmed that ensemble models like Random Forest and XGBoost, while slightly less accurate, remained valuable due to their interpretability and generalization capabilities especially when computational resources are limited.

While CNNs and ANNs achieved near-perfect accuracy, the authors noted that deep models lack explain-

ability a crucial factor for business adoption. Telecommunication managers often require interpretable insights (e.g., which variables most influence churn decisions) rather than purely predictive outputs. Furthermore, the study did not address class imbalance, a persistent issue in churn datasets. These gaps open pathways for hybrid frameworks combining explainable AI (XAI) and ensemble deep learning, ensuring both high performance and transparency

2.2.4 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 feature 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 telecommunication 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 great improvements in predictive stability. In another strand of research, theySaha 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 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.

2.3 Customer Churn in the Telecommunications Industry

Customer churn remains one of the most pressing strategic and financial challenges in the telecommunications sector. According to Chang et al. (2024), the global telecommunications industry records one of the highest churn rates across all service-based industries, exceeding 30% annually. The constant pressure from competitors, coupled with low switching costs for consumers, compels operators to invest

heavily in churn management strategies. From a financial standpoint, customer acquisition is considerably more expensive than retention, making churn reduction critical to long-term profitability.

Chang et al. (2024) emphasized that understanding customer churn requires both behavioral insight and advanced predictive analytics. Their study demonstrated how ensemble learning models including Decision Trees, Boosted Trees, and Random Forests can effectively distinguish between customers likely to leave and those who remain loyal. The Random Forest model achieved the highest performance, attaining 91.66% accuracy, 82.2% precision, and 81.8% recall. This demonstrates its ability to identify a large proportion of at-risk customers while minimizing misclassification errors.

The study also integrated Explainable Artificial Intelligence (XAI) techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive Explanations), enabling model transparency and interpretability. These tools provided actionable insights into key churn drivers, such as contract type, call frequency, and payment history. Chang et al. concluded that integrating XAI with ensemble learning not only improves predictive accuracy but also supports managerial decision-making by explaining why certain customers are likely to churn.

Complementing these global insights, Mwaura (2021)conducted a study in the Kenyan telecommunications industry to understand how customer experience management influences churn mitigation. The study employed a descriptive survey design with 195 employees drawn from leading telecommunication companies. It explored the relationships between social environment, service interface, retail atmosphere, and price competitiveness in shaping customer loyalty.

The results revealed that the social environment and service interface had significant positive effects on churn mitigation—indicating that peer influence, quality interactions, and customer support directly enhance retention. Conversely, price sensitivity exhibited a negative relationship with churn control, suggesting that aggressive pricing by competitors increases customer switching. The study recommended that Kenyan telecommunication companies enhance personalization, service interaction quality, and office design, while implementing flexible and customer-friendly pricing strategies to foster loyalty.

Together, these studies underscore that churn is both a technical and behavioral phenomenon. From a technical perspective, machine learning and ensemble methods enable operators to predict churn with high accuracy using large-scale data analytics. From the behavioral standpoint, customer satisfaction, pricing structures, and service experiences remain decisive retention factors. The integration of both perspectives—predictive analytics for forecasting and experience management for intervention—creates a holistic approach to churn management.

For emerging markets, where competition among mobile service providers is intense, combining data-driven prediction with experience-driven strategies can yield sustainable customer loyalty. These insights form the conceptual basis for developing intelligent churn management systems that leverage both machine learning and human-centered relationship management.

2.4 Industry Relevance and Contribution

The telecommunications sector remains a highly dynamic and data-rich environment where customer churn directly affects organizational profitability and market stability. In this context, predictive modeling has evolved from traditional statistical methods toward machine learning and ensemble-based frameworks that can efficiently process large-scale, multidimensional data. Several studies (e.g., Chang et al. (2024); Usman-Hamza et al. (2022); Saha et al. (2023)) demonstrate that advanced learning techniques significantly enhance churn prediction accuracy, thereby offering a competitive edge in customer retention strategies.

Earlier statistical approaches, such as survival analysis and logistic regression, primarily focused on modeling the likelihood or timing of churn (e.g., Jahromi and Sharifi (2020)). While these models provided interpretable coefficients and inferential insights, they often struggled to capture nonlinear relationships and feature interactions inherent in telecommunication datasets. Recent studies highlight that

ensemble-based techniques, such as Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), deliver superior predictive accuracy and robustness (Saha et al. (2023)). These algorithms handle high-dimensional data and class imbalance more effectively, enabling telecommunication operators to detect subtle churn indicators early.

For instance, Chang et al. (2024) reported that ensemble models achieved over 91% accuracy and strong recall rates, while Usman-Hamza et al. (2022) found that enhanced decision-forest architectures with weighted soft voting and stacking yielded optimal results in imbalanced churn datasets. Similarly, deep learning models such as CNNs and ANNs achieved near-perfect accuracy in cross-regional datasets Saha et al. (2023), reinforcing their value in capturing complex behavioral dynamics.

Although advanced models achieve remarkable predictive performance, they introduce new challenges regarding interpretability and business integration. Telecommunication organizations require churn-prediction systems that are not only accurate but also transparent and explainable. The incorporation of Explainable AI (XAI) frameworks such as LIME and SHAP, addresses this issue by providing feature-level explanations that link model predictions to business metrics like revenue, loyalty, and customer lifetime value (Chang et al. (2024)).

This dual focus on predictive accuracy and interpretability ensures that decision-makers can translate technical model outputs into actionable business insights. Moreover, by linking churn probabilities to financial impact, firms can optimize retention strategies, allocate marketing resources more efficiently, and prioritize high-value subscribers for intervention.

The current study aims to fill these identified gaps by developing a scalable, interpretable, and context-aware churn prediction framework for the telecommunications sector. By integrating secondary datasets from diverse telecommunication markets and employing ensemble learning combined with XAI techniques, the study prioritizes both technical excellence and practical relevance. This approach contributes to academic literature by bridging predictive modeling with operational decision-making and offers telecommunication organizations a viable data-driven solution for proactive churn management and long-term customer retention.

2.5 Conceptual Framework

The conceptual framework underpinning this research integrates behavioral profiling, retention insights, and model evaluation as interrelated components for understanding and predicting customer churn in the telecommunications industry. This framework aligns with prior empirical studies that emphasize the combined importance of customer behavioral analysis, predictive modeling, and actionable retention strategies for sustainable business performance (Chang et al. (2024); Mwaura (2021); Usman-Hamza et al. (2022)). It represents a holistic approach that bridges machine learning–based prediction with customer relationship management (CRM) insights, ensuring both technical precision and managerial relevance.

2.5.1 Behavioral and Demographic Profiling

Behavioral profiling plays a pivotal role in churn prediction, as customer usage patterns, billing behaviors, and demographic characteristics provide early signals of dissatisfaction or disengagement. According to Jahromi and Sharifi (2020), key demographic and financial features such as tenure, total revenue, and service frequency, have consistently shown predictive significance in churn analysis. Similarly, Nagarkar (2022) demonstrated that behavioral indicators like recharge frequency and inactivity duration were the most critical determinants of churn in Nepal's telecommunication sector, with XGBoost achieving 97% accuracy. These findings support the integration of both quantitative (e.g., call duration, data

usage) and qualitative (e.g., customer satisfaction or complaint records) factors into predictive modeling frameworks. In this study, customer behavioral and demographic data will be analyzed to establish the key churn predictors, forming the foundation for the machine learning models to be trained and tested.

2.5.2 Retention Insights

Transforming predictive insights into actionable retention strategies is crucial for business sustainability. Mwaura (2021) found that improving customer experience through service interface quality, pricing strategies, and social environment, has a significant positive effect on retention within telecommunication sector. Meanwhile, Chang et al. (2024) emphasized that integrating machine learning–derived churn probabilities into CRM systems enables proactive retention campaigns.

2.6 Gaps in Existing Literature

Although considerable research has been conducted on customer churn prediction, several recurring gaps persist in both methodological and practical dimensions. Existing studies in the telecommunications industry primarily focus on predictive accuracy while overlooking the interpretability, financial relevance, and contextual diversity of the datasets used. As a result, while machine learning (ML) algorithms continue to advance, their business applicability and generalization across markets remain limited (Chang et al. (2024); Shahabikargar et al. (2025); Mwaura (2021)). Addressing these gaps is essential for developing scalable, transparent, and context-aware churn prediction frameworks that balance analytical performance with real-world usability.

2.6.1 Limited Feature Diversity

A recurring limitation across the literature is the overreliance on structured Customer Relationship Management (CRM) attributes, such as usage duration, billing frequency, and tenure, as primary churn predictors. Mwaura (2021) observed that while transactional and behavioral variables are useful, they often fail to incorporate deeper financial metrics such as Average Revenue per User (ARPU), Total Revenue, or service profitability indices, which are crucial for business forecasting and segmentation. Similarly, Nagarkar (2022) and Saha et al. (2023) emphasized that the absence of diversified features—like sentiment indicators, customer engagement scores, or network experience measures—can reduce model robustness and limit its interpretive value. Expanding feature representation to include financial, psychological, and contextual dimensions is therefore necessary for building models that align with both customer behavior and firm-level profitability.

2.6.2 Narrow Evaluation Practices

A second gap in churn prediction research lies in the evaluation of model performance. Most studies emphasize metrics such as accuracy, recall, or AUC, which are effective for assessing statistical performance but insufficient for understanding business implications. Shahabikargar et al. (2025) highlighted that accuracy-driven evaluation frameworks often fail to account for the financial cost of misclassifications, particularly false negatives (missed churners), which can translate to significant revenue losses. Moreover, very few studies integrate return-on-investment (ROI) or customer lifetime value (CLV) indicators into churn evaluation, despite their direct connection to managerial decision-making. Integrating financial metrics with predictive performance would enable a more comprehensive understanding of model impact, bridging the gap between data science outputs and strategic business outcomes.

2.6.3 Interpretability Limitations

Despite impressive advances in model accuracy through ensemble and deep learning methods, interpretability remains a major obstacle to operational adoption. Complex ML models such as Random

Forest, Gradient Boosting, and Convolutional Neural Networks often operate as opaque "black-box" systems, which makes it difficult for managers to understand how predictions are generated (Chang et al. (2024)). The lack of transparency in feature importance and decision logic reduces managerial trust and delays the integration of predictive models into everyday CRM operations. Recent efforts in Explainable Artificial Intelligence (XAI), including tools like SHAP and LIME, have improved post-hoc interpretability; however, their use in telecom churn literature remains limited (Usman-Hamza et al. (2022)). Therefore, a critical research need is to embed interpretability mechanisms into model design rather than as an afterthought, ensuring that predictive systems are both accurate and comprehensible.

In summary, existing churn prediction research remains constrained by three intertwined issues: feature homogeneity, narrow metric scope, and limited interpretability. These gaps not only restrict the scalability of predictive models across diverse telecommunication environments but also hinder their strategic adoption in telecommunication markets. The proposed study seeks to address these shortcomings by expanding the feature space to include financial and behavioral predictors, incorporating ROI-linked evaluation metrics, and prioritizing interpretable ML frameworks integrated with XAI for enhanced managerial trust and operational relevance.

CHAPTER THREE

3 METHODOLOGY

3.1 Research Design

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 1. This reproducible approach shows that it can be plugged into any other telecommunications industry.

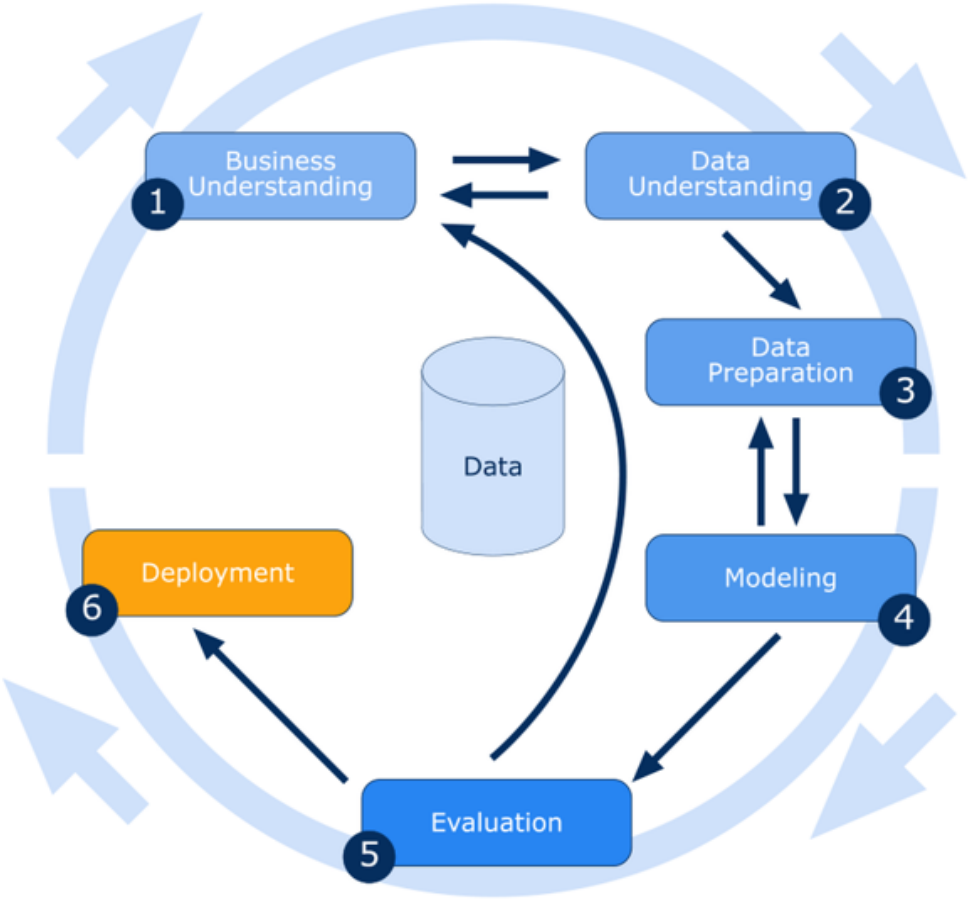


Figure 1: Cross-Industry Standard Process for Data Mining (CRISP-DM)

3.2 Proposed Workflow

The proposed workflow for this study is illustrated in Figure 2. It outlines the systematic process of building and evaluating machine learning models for customer churn prediction.

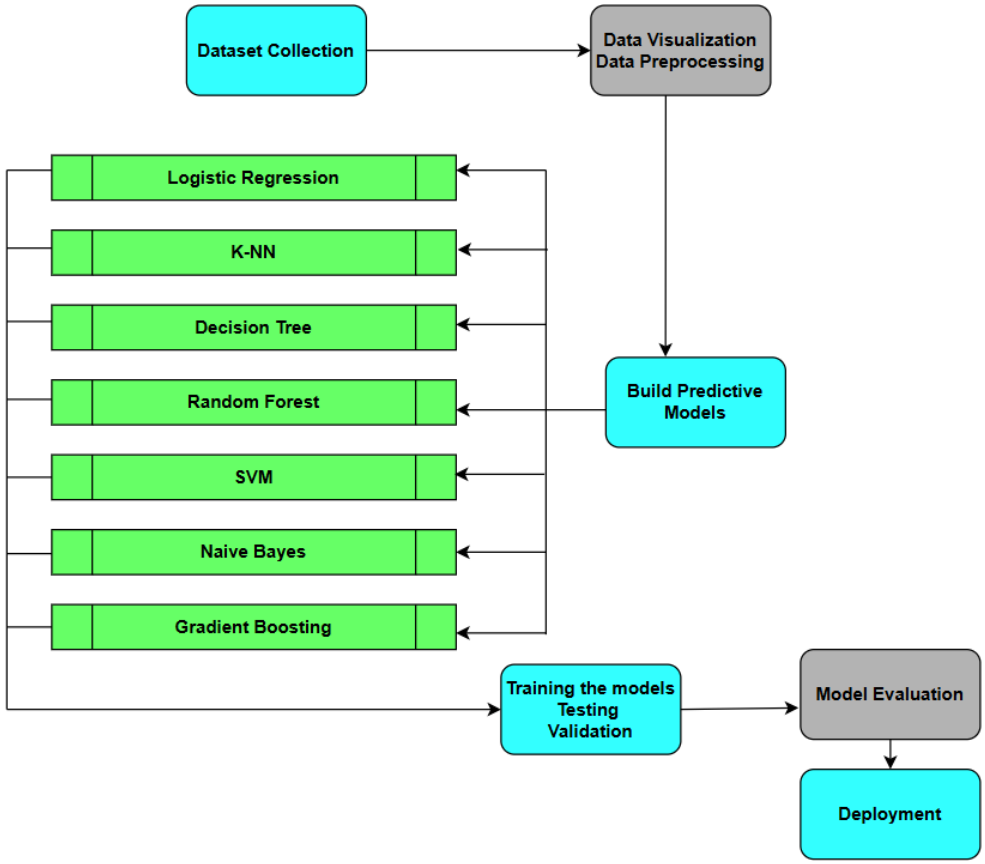


Figure 2: Proposed Workflow

3.3 Business Understanding

Telecommunications companies operate in a market where competition is intense, and customer expectations shift rapidly according to their preferences. As subscription growth slows and price competition increases, retaining existing users becomes a 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 operations. Despite continuous investment in marketing, loyalty programs, and service improvement initiatives, many operators still experience high churn levels that negatively affect revenue reliability and weaken long-term competitiveness. Understanding the drivers of churn and building a predictive model that can anticipate customer exit has therefore become a strategic necessity in this industry.

3.4 Data Understanding

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

3.5 Data Collection

This study will use secondary data sourced from a Bulgarian telecommunication operator, which is publicly available on the Mendeley data website Tokmakov (2024). This dataset comprises customers'

service usage, demographic information, and spending, among others, which will provide key indicators of customer churn, allowing us to anticipate the behaviors that contribute to customer retention and predict the main behavior that will help us retain customers in the network. The study will encourage iteration in model building and assessment until a good enough model is achieved.

3.6 Exploratory 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.

3.7 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.

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.

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.

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.

3.8 Data Transformation

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.

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 and min-max scaling, will bring features within a comparable scale, preventing dominant features from influencing the model outcome.

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 gradient boosting will guide feature prioritization.

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.

3.9 Model Selection

Customer churn prediction represents a supervised learning challenge where the goal is to classify customers into specific categories, typically churners or non-churners, based on a set of descriptive features.

In this research, seven modern classification algorithms are examined: Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Gradient Boosting (GB). The comparative analysis aims to determine the most accurate and interpretable approach suitable for telecommunication customer data. After model evaluation, the best-performing algorithm will be selected for final churn prediction.

3.9.1 Logistic Regression

Logistic regression estimates the likelihood that a customer will discontinue service (churn) based on their feature vector $\mathbf{x}i = [xi1, x_{i2}, \ldots, x_{ip}]$. The model computes this probability using the logistic function:

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x}_i + b)}} \tag{1}$$

Here, $\mathbf{w}$ represents the model coefficients and b is the bias term, both estimated through maximum likelihood optimization. Logistic regression is preferred for its interpretability, computational efficiency, and ability to quantify how individual customer attributes influence churn probability.

3.9.2 K-Nearest Neighbors

The KNN algorithm is a simple yet effective non-parametric method that classifies a new data instance based on the majority class among its k closest points in the feature space. The proximity between data points is typically measured using a distance metric such as Euclidean or Manhattan distance. KNN is advantageous for its intuitive design and adaptability to non-linear decision boundaries, though it can be computationally demanding for large-scale telecommunication datasets.

The predicted label is

$$\hat{y} = \text{mode}\{y_j \,|\, \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)\}, \tag{2}$$

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.

3.9.3 Decision Tree

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

$$\text{Gini}(t) = 1 - \sum_{c=1}^{C} p_c^2, \tag{3}$$

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.

3.9.4 Random Forest

The Random Forest algorithm constructs an ensemble of B independent decision trees, each trained on a randomly drawn subset (bootstrap sample) of the original dataset. For any given observation, the collective prediction is determined through a majority voting process among all individual trees:

$$\hat{y} = \text{mode} h_b(\mathbf{x})_{b=1}^{B} \tag{4}$$

This ensemble approach minimizes overfitting, enhances generalization, and effectively manages high-dimensional and heterogeneous data. Moreover, Random Forest inherently provides measures of feature importance, offering valuable insight into which attributes most influence customer churn.

3.9.5 Support Vector Machine

Support vector machines classify data by identifying an optimal separating hyperplane that maximizes the distance (margin) between different classes. The optimization problem can be expressed as

$$\min_{\mathbf{w},b} \frac{1}{2}|\mathbf{w}|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \ \forall i \tag{5}$$

SVMs are capable of modeling complex, non-linear decision boundaries by applying kernel functions such as polynomial, radial basis function (RBF), or sigmoid kernels. This makes them suitable for telecommunication churn prediction, where behavioral relationships between variables are often non-linear.

3.9.6 Naive Bayes

Naive Bayes is a probabilistic model that leverages Bayes' theorem under the simplifying assumption that input features are conditionally independent given the class label. The class posterior probability is computed as

$$P(y|\mathbf{x}) = \frac{P(y)\prod_{j=1}^{p} P(x_j|y)}{P(\mathbf{x})} \tag{6}$$

Although this independence assumption is rarely exact, the method performs remarkably well on large-scale, high-dimensional datasets. Its computational efficiency and simplicity make Naive Bayes particularly effective for rapid churn classification in telecommunication databases.

3.9.7 Gradient Boosting

Gradient boosting combines multiple weak learners, typically shallow decision trees, into a strong predictive model through a sequential learning process. At each stage m, the algorithm fits a new learner $h_m(\mathbf{x})$ to the pseudo-residuals of the previous model:

$$r_{im} = -\frac{\partial \mathscr{L}(y_i, F_{m-1}(\mathbf{x}i))}{\partial Fm-1(\mathbf{x}_i)} \tag{7}$$

The ensemble model is then updated according to

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta, h_m(\mathbf{x}) \tag{8}$$

where η represents the learning rate, controlling how much each new tree contributes to the model. Gradient boosting is widely recognized for its strong predictive accuracy, robustness to noise, and ability to uncover complex, non-linear dependencies prevalent in customer churn data.

3.10 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 a 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 which 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.

3.11 Handling Class Imbalance with SMOTE

In most telecommunications datasets, the number of customers who churn is typically much smaller than those who remain subscribed, resulting in a pronounced class imbalance. To mitigate this issue, the Synthetic Minority Over-sampling Technique (SMOTE) will be employed to enhance the representation of the minority (churn) class within the training data. SMOTE works by creating synthetic samples for the minority class rather than merely duplicating existing records. For each minority data point x_i, a new synthetic sample x_{new} is generated using the relationship:

$$x_{\text{new}} = x_i + \delta \times (x_{nn} - x_i), \quad \delta \sim U(0,1) \tag{9}$$

Here, x_{nn} denotes one of the k nearest neighbors of x_i, and δ is a random value drawn uniformly between 0 and 1. By interpolating between existing minority samples and their neighbors, SMOTE produces realistic synthetic observations, leading to a more balanced training distribution. This balancing improves the model's sensitivity to churners and prevents it from being overly biased toward the dominant non-churn class.

3.12 Model Evaluation

The performance of the predictive models will be assessed through an 80/20 train–test split, with stratified sampling applied to maintain the natural class ratio between churners and non-churners. This ensures that both subsets reflect the true population distribution, improving the reliability of model validation.

Model effectiveness will be quantified using key performance indicators, namely precision, recall, F1-score, and the area under the Receiver Operating Characteristic Curve (AUC–ROC), expressed as follows:

$$\text{Precision} = \frac{TP}{TP+FP}, \qquad \text{Recall} = \frac{TP}{TP+FN}, \qquad \text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

In this formulation, TP, FP, and FN represent true positives, false positives, and false negatives, respectively. Precision measures the proportion of correctly identified churners among all predicted churners,

whereas recall evaluates the proportion of actual churners successfully captured by the model. The F1-score serves as a harmonic mean of precision and recall, balancing the trade-off between the two, while the AUC–ROC assesses the model's capability to differentiate between churners and retained customers. In the business context, false negatives, cases where actual churners are misclassified as non-churners, carry a higher financial risk than false positives. Therefore, recall will be emphasized during model selection to ensure that high-risk customers are correctly detected. The final model will be chosen based on an optimal balance between recall and precision, complemented by interpretability analysis using SHAP (SHapley Additive exPlanations) to clarify the relative influence of input variables on churn predictions.

3.13 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 re-training 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.

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

Jahromi, A. T., & Sharifi, H. (2020). Analyzing customer churn prediction techniques: A review of the literature. *Future Internet*, *14*(3), 94. Retrieved from https://www.mdpi.com/1999-5903/14/3/94 doi: 10.3390/fi14030094

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

Mishra, R., Reddy, P. K., Nair, P., & Ranjan, R. (2019). Customer churn prediction in telecom using machine learning. In *2019 international conference on artificial intelligence and signal processing (aisp)* (pp. 1–5). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/8706988 doi: 10.1109/AISP.2019.8706988

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

Sohail, A., Khan, M. A., Kadry, S., & Nam, Y. (2020). Telecommunication customers churn prediction model using machine learning approach. In *2020 international conference on electrical, communication, and computer engineering (icecce)* (pp. 1–6). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/9262288 doi: 10.1109/ICECCE49384.2020.9262288

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

APPENDICES

A Project Appendix

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

A.1 Project Timeline: 4 Months

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

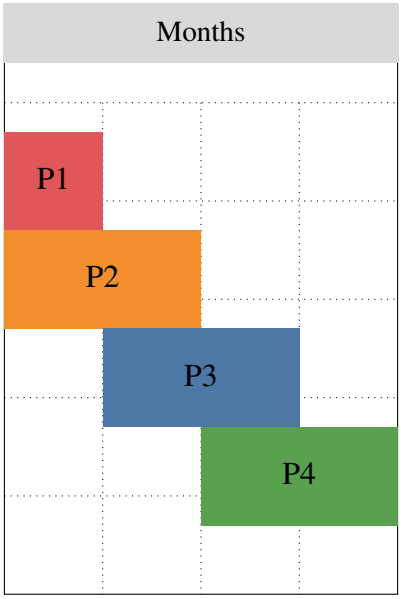


Figure 3: 4-Month Intensive Project Timeline

Legend:

- 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

A.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

A.2.1 Estimated Project Budget (KSh)

Table 1: Estimated Project Budget (KSh)

Category	**Basis of 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
B. Administration and Infrastructure		
Ethical and Institutional Review	Standard university ethics	10,000
Internet and Communication	high-speed internet for 4 months	20,000
Overleaf/LaTeX Subscription	Overleaf Premium plan	6,000
C. Data and Dissemination		
Data Access	Use of public dataset	0
Publication Fees	Open-access journal submission fees	40,000
D. Equipment and Materials		
External SSD Backup Drive	1 TB external drive	10,000
Stationery and Printing	Thesis printing and binding	8,000
	Subtotal	109,000
Miscellaneous/Contingency 10% of Subtotal	30,000	
Total Estimated Budget		**139,000**