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 telecom industry, predicting customer churn is a recurring problem that has a big impact on long-term viability, profitability, and competitiveness. The context and research challenge are first described in the paper, which also highlights the sector's growing rivalry and the shortcomings of conventional churn control techniques. It emphasizes the goals of finding important churn drivers, creating a prediction model based on machine learning, via assessing various algorithms, and creating a deployable dashboard to help in decision-making. Recent empirical research on churn prediction is combined with theoretical frameworks like Technology Adoption Theory and 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 shortcomings in terms of interpretability, profit-sensitive evaluation, and a narrow concentration on other telecommunication sectors. The methodology suggests using secondary anonymized public data from a Bulgarian 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, and LightGBM—will be compared. Interpretability approaches like SHAP and hyperparameter tuning will be used. 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. In addition to providing methodological insights that can be applied to other telecoms, this research offers a deployable and replicable methodology that solves the urgent commercial demand for churn reduction.

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

ACRONYMS

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

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, represents a significant and ongoing concern. Business profitability, operational sustainability, and long-term market competitiveness are all significantly impacted by churn. Annual telecommunication churn rates are more over 30% worldwide. Khan et al. (2024), highlights the need for proactive prediction and retention measures and causing significant annual revenue losses for service providers.

Retaining customers is significantly more cost effective compared to acquiring new customers in an economic perspective. Studies has shown that acquiring new customer can be five to seven times more expensive compare to retaining customers Usman-Hamza et al. (2022). Increase in churn rate can cause reduction in revenue stability in the market, constraining the organizational competitiveness.

From a social and cultural perspective, telecommunication is important in accessing digital inclusion services, and also very critical for new emerging economies in the Sub-Saharan Africa *Mobile Economy Report: Sub-Saharan Africa 2022* (2022).

There are many ongoing investments strategies that have been put in place by many telecommunications companies but still most of this companies uses traditional methods to analysis and predict churn. This traditional methods mainly include using static datasets that doesn't capture the dynamic shifts in customer preferences, their social and economic behaviors or even their daily spent Chang et al. (2024). Most of this approaches has result to insufficient decision making when it comes to resources allocation to reduce churn rate within the company limiting the predictive accuracy of churn rate.

A lot of study has been done include both the data science and machine learning spaces, just to address these limitations. Machine learning algorithms like Logistic Regression, Random Forest and Decision Tress have shown an exemplary potential of classifying data as either churn or non churn hence identifying the customers who can leave the network before the actually leave. Rahman et al. (2024). These advanced models can be incorporated in large datasets, homogenous datasets as well as heterogenous datasets just to capture the underlying patterns that can predict customer churn.

The rapid digital adoption in the market and socio- ecoomic on mobile services on calls and data usages has really led 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 service providers that provide good service qualities. Most of the prvous churn rate prediction 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 has also limitation when it comes to bridging the gap between 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 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 limitations in the existing customer churn literature 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, LightGBM) compare in predicting customer churn?
5. How can a predictive churn model be integrated into a CRM system through 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: 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 environment, the proposed framework offers a methodological blueprint that can be adapted for other markets and industries facing similar churn-related challenges.

This study tend to contribute the practical knowledge available and the academic concept on customer churn in 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

The telecommunications sector has experienced a profound shift over the last decade as digital technologies evolve, markets open up, and customers demand more personalized and reliable services. These changes have widened access and improved communication options, but they have also intensified industry rivalry. In this highly competitive environment, keeping existing customers has become just as important as acquiring new ones. One of the most persistent challenges is customer churn—the decision by a subscriber to stop using a network's services. High churn rates remain a global issue, with many operators losing more than a third of their customer base every year. Such losses translate to significant revenue erosion and mounting pressure on companies to adopt data-driven approaches that can anticipate and prevent customer exits Khan et al. (2024).

2.1 Theoretical Framework

Customer churn can be understood through multiple lenses because it affects both the financial stability of a firm and the quality of its relationship with subscribers. According to Zhang et al. (2022), shifts in how customers interact with digital platforms and how they perceive service quality play a central role in influencing whether they decide to remain or leave. The long-standing ideas in Rogers' Diffusion of Innovation theory also provide useful insight: users tend to adopt or drop telecommunication services based on how well the service fits their needs, how beneficial they believe it is, and how seamlessly it aligns with their daily routines. Complementing this perspective, Customer Relationship Management (CRM) approaches stress that companies must anticipate customer needs through tailored communication, targeted engagement, and data-driven retention strategies. Together, these frameworks show that churn is shaped by intertwined economic pressures, user behaviors, and technological experiences.

2.2 Empirical Studies

Recent years have witnessed significant advances in churn modeling using machine learning and deep learning. For instance, Usman-Hamza et al. (2022) demonstrated that ensemble decision forests with weighted voting strategies outperformed baseline classifiers in imbalanced telecommunication churn datasets. Similarly, Saha et al. (2023) highlighted the superiority of convolutional neural networks (CNNs) in hybrid datasets from India and the U.S., emphasizing that ROI-sensitive evaluation matters as much as predictive accuracy.

Context-specific models remain essential. In Nepal, Nagarkar (2022) applied XGBoost to 52,332 telecom customers, showing that recharge frequency and inactivity duration are strong predictors of churn, enabling targeted retention campaigns. In Syria, integrated social network analysis (SNA) features, improving AUC from 0.84 to 0.933, demonstrating the value of relational data. More recently, Poudel and Sharma (2024) introduced explainer-aware churn models that combined interpretability methods with predictive algorithms, showing the importance of stakeholder trust in adopting churn systems.

At a broader level,Shahabikargar et al. (2025) reviewed churn prediction across industries and highlighted the growing importance of integrating external data, as well as the need to balance performance with transparency. Their findings reinforce that churn solutions must be both accurate and interpretable to be actionable in practice.

2.3 Customer Churn in the Telecommunications Industry

Customer churn remains a major challenge for the telecom industry, reducing revenue stability and profitability. Studies confirm that global churn rates can exceed 30%, even in mature markets, highlighting the urgency for robust retention strategies Chang et al. (2024). In Sub-Saharan Africa, where telecoms are a lifeline for financial transactions, e-learning, and small business operations, the stakes are particularly high Mwaura (2021). While predictive modeling has advanced significantly, few studies focus directly on African contexts, revealing a critical gap in geographic representation.

2.4 Industry Relevance and Contribution

This review highlights key insights:

Ensemble and deep learning approaches consistently outperform traditional models.Interpretability and business ROI are increasingly important evaluation criteria, and other markets remain underexplored.

The proposed framework provides a scalable and practical approach tailored to the realities of telecommunication churn management in emerging markets.

2.5 Gaps in Existing Literature

- Feature Limitations: Available datasets mainly include structured CRM variables (billing, tenure, usage), but exclude important signals such as Total Revenue, Average Mobile Revenue, Average Fix Revenue, and ARPU. This narrows the feature space for building robust churn models Mwaura (2021).

- Evaluation Metrics Bias: Existing research often prioritizes accuracy and AUC, while overlooking profit-sensitive or ROI-based metrics like ARPU (Average Revenue Per User) Shahabikargar et al. (2025). This creates a gap between academic results and actionable business insights.
- Interpretability Challenges: Many ML models offer limited transparency, reducing business stakeholders understanding and trust in the results Chang et al. (2024) and translate them into real retention strategies.

2.6 Conceptual Framework

The conceptual framework for this study comprises the following components:

- User Behavior Analysis: Examination of customer demographics, usage patterns, contractual terms, and payment behaviors to identify churn predictors.
- Machine Learning Models: Implementation of algorithms including Logistic Regression, Random Forest, Decision Trees, and Gradient Boosting to predict churn, with comparative performance analysis.
- User Retention Strategies: Translation of model outputs into actionable retention initiatives.
- Performance Evaluation: Use of metrics such as precision, recall, F1-score, and ROC-AUC to assess predictive performance.

2.7 Empirical Studies

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 telecom 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 attrition, 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.

2.8 Customer Churn in the Telecommunications Industry

Churn continues to pose a significant threat to revenue certainty and long-term profitability in the telecommunications sector. 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 payments, remote education, and microenterprise operations Mwaura (2021). Despite the growing body of churn research, very few studies directly address African market conditions. This lack of contextual evidence highlights an important gap that limits the applicability of existing predictive approaches across the region.

2.9 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. At the same time, organizations increasingly require models that balance predictive power with clarity, enabling business teams to interpret outputs and integrate them into decision processes. Another recurring insight is the disproportionate focus on markets outside Africa, 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 telecom providers in developing regions.

2.10 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 ARPU, revenue components, and service-specific income streams Mwaura (2021). This restricts the richness of models and can weaken 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. ROI-linked indicators, including ARPU and 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. Limited transparency often reduces trust and slows the adoption of churn systems as practical decision-support tools Chang et al. (2024).

2.11 Conceptual Framework

The conceptual model guiding this research integrates four complementary components:

- **Behavioral and Demographic Profiling:** Analysis of user characteristics, consumption habits, billing routines, and contractual terms to establish key indicators associated with churn risk.
- **Predictive Modeling:** Deployment of machine learning techniques—including Logistic Regression, Random Forests, Decision Trees, and Gradient Boosting—to identify high-risk customers and compare algorithmic performance using consistent criteria.
- **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.

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 telecommunications sector.

3.0.1 Proposed Workflow

The proposed workflow for this study is illustrated in Figure 3.1. It outlines the systematic process of building and evaluating machine learning models for customer churn prediction.

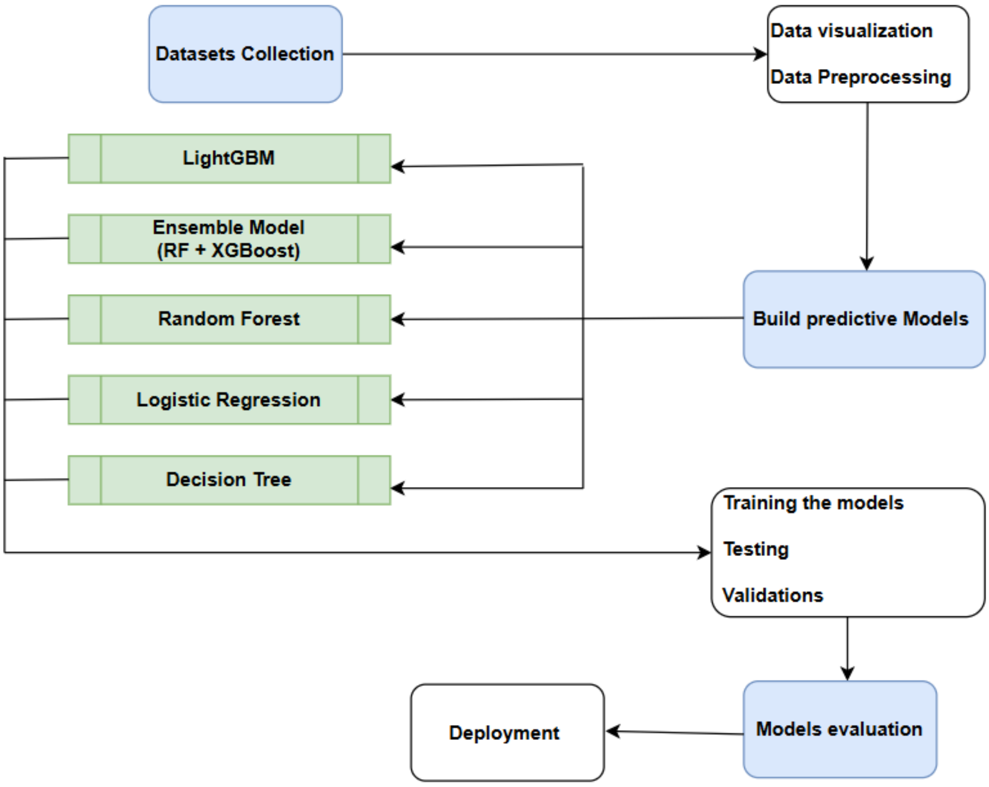

Figure 3.1: Proposed Workflow

3.1 Business Understanding

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

3.2 Data Understanding

The analysis draws on a publicly available anonymized dataset from a Bulgarian telecommunications provider Tokmakov (2024). The dataset includes customer attributes, subscription patterns, and revenue indicators, offering a broad representation of consumer behavior within an operational telecom environment. Its structure allows for comprehensive churn investigation, from behavioral patterns to financial signals.

3.3 Data Collection

This research relies on secondary data sourced from the CRM system of a Bulgarian telecom operator. The dataset contains anonymized customer-level information capturing demographic attributes, subscription activity, and multiple revenue metrics essential for modeling churn. Core variables include customer market segment, counts of active and inactive subscriptions, mobile and fixed-line revenues, total revenue, and Average Revenue Per User (ARPU). The target label, CHURN, identifies whether a subscriber discontinued service, enabling supervised learning.

The dataset was anonymized before release to meet GDPR requirements and ensure customer privacy. Because the data originates from a real operational environment, it reflects authentic business dynamics such as usage variability, seasonal behavior, and revenue fluctuations. Its granularity and diversity provide a strong foundation for exploring churn determinants and building predictive models that can generalize to similar telecom markets, especially in emerging economies.

3.4 Exploration Data Analysis

Exploratory Data Analysis (EDA) will be applied to understand variable distributions, identify class imbalance between churn and non-churn cases, and diagnose missing or inconsistent values. Graphical tools—including boxplots, histograms, and correlation matrices—will be used to examine relationships between revenue measures, subscription activity, and churn outcomes. These insights will inform preprocessing decisions, feature engineering strategies, and the selection of modeling techniques to ensure a reliable and interpretable predictive framework.

3.5 Data Cleaning

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

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 based on unique identifiers will be detected and removed to prevent double counting.

Categorical attributes will be standardized to ensure uniform labeling and eliminate typographical inconsistencies. Outliers in financial indicators such as will be identified through z-score and interquartile range (IQR) analysis, and extreme anomalies will be evaluated before applying appropriate transformations or capping to maintain distributional integrity.

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

New features will be generated to enhance predictive strength—for example, ratios, aggregated revenue metrics, or indicators capturing usage dynamics. Feature engineering helps uncover subtle behavioral signals associated with churn.

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.

Feature Selection

Correlation patterns, mutual information scores, and model-driven importance rankings 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 one-hot or ordinal techniques depending on their hierarchy. The target variable will be set to a binary format for supervised classification.

Through this systematic transformation pipeline, the dataset becomes standardized, analytical, and ready for high-quality modeling.

3.6 Modeling

The churn prediction task is framed as a binary classification problem. Several machine learning algorithms will be implemented to compare interpretability, computational efficiency, and predictive performance.

3.6.1 Logistic Regression (LR)

Logistic Regression serves as the baseline classifier due to its transparency and ease of interpretation. It estimates churn probability through a sigmoid transformation of a linear combination of input variables: $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$,

$$P(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x}_i + b)}} \tag{3.1}$$

Coefficient signs and magnitudes will illuminate how features such as ARPU or activity levels influence churn risk.

3.6.2 Decision Tree (DT)

Decision Trees segment data using recursive splits that maximize separation between classes. Splits are evaluated using Gini Impurity:

$$\text{Gini}(t) = 1 - \sum_{k=1}^{K} p_k^2 \tag{3.2}$$

where p_k is the proportion of class k instances at node t. The optimal split is chosen by minimizing the weighted impurity across child nodes:

$$\Delta\text{Gini} = \text{Gini}(t) - \sum_{j \in \text{children}} \frac{N_j}{N_t} \text{Gini}(j) \tag{3.3}$$

where N_j and N_t denote the number of samples in the child and parent nodes, respectively. Trees are easy to visualize and interpret, making them valuable for identifying intuitive churn drivers..

3.6.3 Random Forest (RF)

Random Forests aggregate predictions from multiple decision trees built on bootstrapped samples. The ensemble prediction is determined through majority voting: Given B trees $\{h_b(\mathbf{x})\}_{b=1}^{B}$, each trained on a bootstrap sample, the final prediction for an instance $\mathbf{x}$ is obtained via majority voting:

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

The expected generalization error of the Random Forest is minimized through variance reduction, as the ensemble prediction variance σ_{RF}^2 approximates:

$$\sigma_{\text{RF}}^2 = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{3.5}$$

where ρ is the correlation between trees and σ^2 is the variance of individual trees. This demonstrates that increasing B (the number of trees) and reducing ρ (via feature randomness) improves model stability and accuracy.

The method reduces overfitting and enhances predictive robustness through variance reduction across trees.

3.6.4 Light Gradient Boosting Machine (LightGBM)

LightGBM utilizes gradient boosting and leaf-wise tree growth to optimize predictive accuracy. Residuals guide the sequential training process:

LightGBM is a gradient boosting framework that sequentially adds weak learners (typically shallow trees) to minimize a differentiable loss function $\mathcal{L}(y_i, F(\mathbf{x}_i))$. At each iteration m, the model fits a new learner $h_m(\mathbf{x})$ to the negative gradients (residuals) of the loss function with respect to the current model prediction $F_{m-1}(\mathbf{x})$:

$$r_{im} = -\frac{\partial \mathcal{L}(y_i, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)} \tag{3.6}$$

The ensemble is then updated as:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta\, h_m(\mathbf{x}) \tag{3.7}$$

where η is the learning rate controlling the contribution of each weak learner. LightGBM differs from traditional boosting by employing a *leaf-wise tree growth strategy*, which selects the leaf with the largest loss reduction at each step. Formally, it chooses the split that maximizes the *information gain IG*:

$$IG = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{3.8}$$

where G and H represent the first and second-order gradients of the loss function, and λ, γ are regularization parameters. This optimization approach allows LightGBM to efficiently handle large-scale, high-dimensional telecom datasets. Its interpretability via SHAP (SHapley Additive exPlanations) values further enables decomposition of each prediction into feature-level contributions, enhancing explainability and decision transparency for churn prediction.

Its efficiency and compatibility with SHAP interpretability make it suitable for high-dimensional telecom datasets.

3.6.5 Ensemble Model

To leverage the strengths of multiple algorithms, an ensemble will combine predicted probabilities: To improve robustness and generalization, an ensemble model will combine the probabilistic outputs from multiple base learners Logistic Regression, Random Forest, and LightGBM through a weighted averaging mechanism:

$$\hat{y}_{\text{ensemble}} = \sum_{m=1}^{M} \alpha_m \, \hat{y}_m, \quad \sum_{m=1}^{M} \alpha_m = 1 \tag{3.9}$$

where α_m represents the contribution weight of each model. This ensemble strategy balances interpretability and predictive strength, resulting in a more resilient churn prediction framework.

Weighted blends improve generalization and provide balanced performance across metrics.

3.6.6 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_{new} as:

$$x_{\text{new}} = x_i + \delta \times (x_{nn} - x_i), \quad \delta \sim U(0, 1) \tag{3.10}$$

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):

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.11}$$

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 CRM 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 containing financial, behavioral, and subscription attributes such as ARPU, Active subscribers, and Effective Segment, the study will generate actionable insights that can be adapted to emerging telecommunications markets, including Kenya. 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 financial (e.g., revenue, ARPU), behavioral (e.g., activity rate), and categorical (e.g., customer segment) variables to create a multidimensional representation of churn risk. Feature engineering and class rebalancing 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 integrated into Customer Relationship Management (CRM) systems. 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 African and global 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

Mobile economy report: Sub-saharan africa 2022. (2022). https://www.gsma.com. (Accessed 9 June 2025)

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