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.

© 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 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 **2**

1.1 Background . . . 2

1.2 Research Problem . . . 3

1.3 Research Objectives . . . 3

1.4 Research Questions . . . 3

1.5 Significance and Justification . . . 4

2 CHAPTER: LITERATURE REVIEW **6**

2.1 Theoretical Framework . . . 6

2.2 Empirical Studies . . . 6

2.3 Customer Churn in the Telecommunications Industry . . . 7

2.4 Industry Relevance and Contribution . . . 7

2.5 Gaps in Existing Literature . . . 7

2.6 Conceptual Framework . . . 8

3 CHAPTER: METHODOLOGY **9**

3.0.1 Proposed Workflow . . . 9

3.1 Business Understanding . . . 9

3.2 Data Understanding . . . 10

3.3 Data Collection . . . 10

3.4 Exploration Data Analysis . . . 11

3.5 Data Cleaning . . . 11

3.5.1 Treating Missing Data . . . 11

3.5.2 Treating Outliers . . . 12

3.5.3 Data Type Conversion . . . 12

3.5.4 Data Transformation . . . 12

3.6 Modeling . . . 13

3.6.1 Logistic Regression (LR) . . . 13

3.6.2 Decision Tree (DT) . . . 13

3.6.3 Random Forest (RF) . . . 14

3.6.4 Light Gradient Boosting Machine (LightGBM) . . . 14

3.6.5 Ensemble Model . . . 15

3.6.6 Handling Class Imbalance with SMOTE 15
3.7 Model Evaluation 16
3.8 Model Deployment 16

4 CHAPTER: EXPECTED OUTCOMES **18**

References **20**

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 contributes adapt to both the academic body of knowledge on churn prediction and the practical toolkit available to telecommunications providers.

By leveraging advanced ML techniques and embedding them into operational workflows, the research aims to create a sustainable, data-driven approach to customer retention in highly competitive markets.

2. CHAPTER: LITERATURE REVIEW

The telecommunications industry has undergone rapid transformation in recent years due to technological advancement, market liberalization, and shifting consumer expectations. While these trends have enhanced access and service delivery, they have also intensified competition, making customer retention a strategic priority. Customer churn, the phenomenon where subscribers discontinue their relationship with a provider, poses one of the most pressing challenges. Globally, annual churn rates in telecommunication is above 30%, resulting in substantial financial losses for service providers and reinforcing the importance of predictive approaches Khan et al. (2024).

2.1 Theoretical Framework

Customer churn has long been viewed as both an economic and relational problem. As noted by Zhang et al. (2022), modern service industries face churn risks driven by service quality and by customer experience and digital interactions. Rogers' Diffusion of Innovation theory remains relevant, explaining why customers adopt, continue, or abandon telecommunication services depending on perceived advantage and compatibility. Likewise, Customer Relationship Management (CRM) frameworks emphasize proactive retention through personalization, loyalty programs, and predictive insights . These theories collectively highlight the multidimensional nature of churn, spanning economic, behavioral, and technological perspectives.

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.

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.

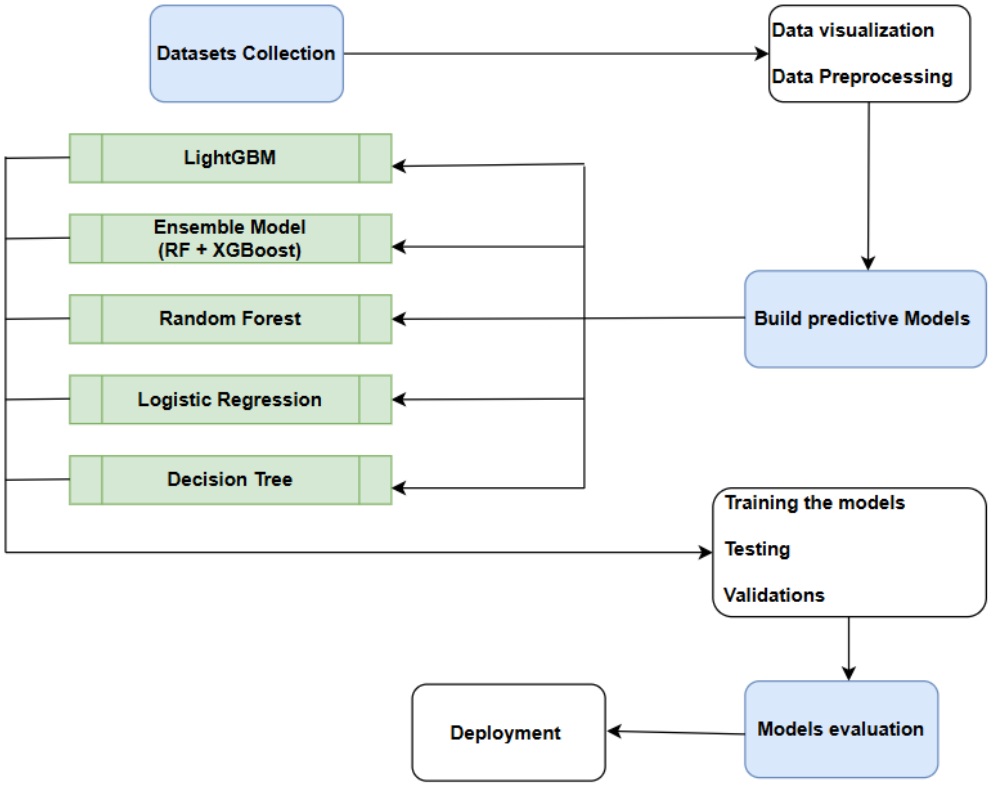

Figure 3.1: Proposed Workflow

3.1 Business Understanding

The telecommunications sector operates in an increasingly competitive environment characterized by market saturation, declining profit margins, and evolving consumer expectations. In such a landscape, customer retention has become a critical determinant of long-term profitability and business sustainability. The cost of acquiring new customers often exceeds that of retaining existing ones, making churn reduction a strategic priority for telecom operators. However, despite heavy investments in

marketing and customer engagement, many providers continue to face substantial churn rates that erode revenue and weaken market share.

In the Kenyan telecommunications market, competition among major service providers such as Safaricom, Airtel, and Telkom Kenya has intensified, driven by aggressive pricing strategies, rapid technological adoption, and diverse value-added services. Customers frequently switch between operators in search of better data plans, network quality, or customer support, leading to unpredictable churn behavior. Traditional approaches to churn management within these companies are largely reactive, relying on historical or static data that fail to capture socio-economic influences. This has resulted in inefficient resource allocation and delayed interventions that often occur after the customer has already defected.

3.2 Data Understanding

This study utilizes a secondary anonymized public dataset obtained from a Bulgarian telecommunication operator, Tokmakov (2024). The dataset provides rich information about customer characteristics, revenue generation, and subscription activity, enabling a comprehensive understanding of churn behavior in the telecommunications context.

3.3 Data Collection

This study employs secondary data obtained from a Bulgarian telecommunication operator. The dataset provides real-world customer-level information essential for churn prediction modeling, capturing multiple dimensions of user engagement, subscription activity, and financial performance. It includes key variables such as customer segmentation, number of active and inactive subscriptions, average mobile and fixed-line revenues, total revenue, and average revenue per user (ARPU). The target variable,CHURN, denotes whether a customer has discontinued service, enabling supervised machine learning analysis.

The dataset was compiled from the operator's Customer Relationship Management (CRM) system and anonymized prior to analysis to ensure confidentiality and compliance with the General Data Protection Regulation (GDPR, 2018). As an operational dataset, it reflects authentic business patterns, making it suitable for empirical analysis and model validation.

By utilizing this dataset, the study benefits from access to rich, high-quality records representing churn behavior in a competitive telecommunications market. The comprehensiveness of the data allows for in-depth exploration of revenue-driven churn determinants, cross-segmentation analysis, and development of interpretable predictive models. The use of actual operational data ensures both

methodological rigor and practical relevance, providing a strong foundation for deploying machine learning based churn prediction systems adaptable to emerging markets.

3.4 Exploration Data Analysis

Exploratory Data Analysis (EDA) will be performed to understand the distribution of numerical and categorical variables, assess potential class imbalance between churned and retained customers, and identify missing or inconsistent values. Statistical summaries and visualizations, such as correlation heatmaps, boxplots, and frequency distributions will be used to uncover relationships between revenue, activity, and churn. These insights will guide feature selection, data preprocessing, and model development to ensure the creation of a robust and interpretable churn prediction framework.

3.5 Data Cleaning

Data cleaning represents a crucial phase of the data pre-processing pipeline, aimed at improving the accuracy, reliability, and analytical value of the data. As the dataset is derived from operational CRM systems, potential sources of inconsistencies include missing values, incorrect numerical entries, duplicate customer records, and inconsistent categorical labels (e.g., customer segments or subscription status). The objective of this stage is to prepare a robust and standardized dataset suitable for statistical analysis and machine learning model development.

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 data handling will depend on the extent and nature of the missingness. Numerical features will be imputed using mean or median values, while categorical features will utilize mode or frequency-based imputation. In cases of significant missingness, the variable will be excluded or

replaced through regression-based imputation. The strategy ensures minimal data loss while maintaining dataset integrity.

3.5.2 Treating Outliers

Outlier detection will rely on statistical and visual diagnostics such as box plots and IQR analysis. Outliers will be treated through capping or log transformation to stabilize model performance.

3.5.3 Data Type Conversion

To maintain analytical consistency, all variables will be converted into appropriate data types. Categorical attributes will be converted to floating-point numbers. This step ensures compatibility with statistical procedures and machine learning frameworks.

3.5.4 Data Transformation

Feature Engineering

Feature engineering will enhance the dataset's predictive power by creating new features that capture behavioral and financial patterns. Derived metrics will be constructed to provide additional modeling context. These engineered features will help identify subtle churn patterns across customer segments.

Feature Scaling

Numerical features will be scaled using standardization (z-score normalization) or min–max normalization to ensure consistent variable ranges. This prevents features with larger numerical scales from dominating the model training process, particularly in algorithms such as Logistic Regression.

Feature Selection

Feature selection will be guided by correlation analysis, mutual information, and model-based importance ranking. Features with low variance or high collinearity will be removed to reduce redundancy and improve model efficiency. Importance metrics from tree-based models such as Random Forest and LightGBM will be used to prioritize key predictors of churn.

Data Encoding

Categorical variables will be transformed into numerical representations appropriate for the modeling algorithm. One-hot encoding will be applied to nominal variables, while ordinal encoding will be used

for hierarchical features. The target variable, will be binary-encoded (1 for churn, 0 for retention) to enable supervised learning.

This structured cleaning and transformation process ensures that the final dataset is accurate, consistent, and optimized for high-quality predictive modeling. By combining statistical rigor with domain-driven feature engineering, the cleaned dataset will provide a reliable foundation for churn prediction and business decision support.

3.6 Modeling

Customer churn prediction in the data is formulated as a supervised binary classification problem, where the goal is to classify each customer as either churned or retained. To achieve this, a comparative modeling framework will be employed using several machine learning algorithms, each offering distinct advantages in terms of interpretability, scalability, and predictive performance.

3.6.1 Logistic Regression (LR)

Logistic Regression will serve as the baseline model due to its interpretability and suitability for binary classification. It estimates the probability that a customer i will churn ($y_i = 1$) based on a vector of predictors $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$, using the logistic (sigmoid) function:

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

where $\mathbf{w}$ represents feature coefficients and b is the bias term. Coefficient signs indicate the direction of influence for example, negative coefficients for ARPU may suggest that higher revenue reduces churn likelihood. The model's simplicity and explainability make it ideal for stakeholder communication.

3.6.2 Decision Tree (DT)

Decision Trees recursively partition the dataset into homogeneous subsets based on feature thresholds that best separate churners from non-churners. At each node, the algorithm evaluates a splitting criterion such as the *Gini Impurity*, defined as:

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

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. This recursive process continues until a stopping criterion (e.g., minimum samples per leaf or maximum depth) is met. Decision Trees are particularly valuable for identifying interpretable churn drivers such as reduced Active_subscribers or declining ARPU, making them highly intuitive for business stakeholders.

3.6.3 Random Forest (RF)

Random Forests enhance Decision Trees by combining multiple trees trained on random subsets of the data and features, a process known as {bootstrap aggregation or bagging. 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^2_{RF} approximates:

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

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.

3.6.4 Light Gradient Boosting Machine (LightGBM)

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.

3.6.5 Ensemble Model

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.

3.6.6 Handling Class Imbalance with SMOTE

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.

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 operational predictive service via an Application Programming Interface (API). The deployment will be implemented using the FastAPI framework due to its high performance, asynchronous request handling, and automatic API documentation. The deployed service will accept customer data as input and return a churn probability score alongside key explanatory features derived from SHAP analysis.

Integration with the telecommunication operator's Customer Relationship Management (CRM) system will enable churn risk scoring and segmentation of at-risk customers. Business teams will access the predictions through a user-friendly dashboard displaying churn probabilities, top churn drivers (e.g., declining ARPU, inactive subscribers), and recommended retention actions.

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