National College of Ireland

Enhanced Customer Churn Prediction in Telecommunications: A Multi-Dataset Comparative Study of Ensemble Machine Learning Models with Explainable AI

Practicum Part 2
MSc in AI for Business

Hayman Kyaw
Student ID: x24158593

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Hayman Kyaw
Student ID:	x24158593
Programme:	MSc in AI for Business **Year:** 2025
Module:	Practicum Part 2
Supervisor:	Anderson Simiscuka
Submission Due Date:	11 December, 2025
Project Title:	Enhanced Customer Churn Prediction in Telecommunications: A Multi-Dataset Comparative Study of Ensemble Machine Learning Models with Explainable AI
Word Count:	**Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Hayman Kyaw

Date: 11 December, 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhanced Customer Churn Prediction in Telecommunications: A Multi-Dataset Comparative Study of Ensemble Machine Learning Models with Explainable AI

Hayman Kyaw
x24158593

Abstract

Customer churn prediction is a critical issue in telecommunications, as it is with five to ten times higher costs of acquiring new customers than retaining existing ones. This study addresses the gap between the prediction accuracy and business interpretability based on a comparative study of traditional machine learning algorithms (Logistic Regression, Random Forest) and advanced ensemble algorithms (XGBoost, LightGBM) on three varied telecommunications datasets consisting of 158,090 customer records.

In this study, explainable AI was utilized by combining model-specific interpretability methods with LIME (Local Interpretable Model-agnostic Explanations). This took complex predictions and turned them into clear, actionable insights for the business. It was shown that LightGBM produced the most stable performance in all datasets with an average AUC-ROC of 0.7578, and gradient boosting techniques significantly worked well compared to other traditional techniques. There was a significant difference between the performance of different datasets, with Dataset 2 showing the highest AUC-ROC of 0.9062 and Dataset 1 with the lowest possible performance of 0.6803, proving that the dataset characteristics are the key factors that determine achievable accuracy.

Interpretability analysis found out that equipment age and tenure are the key drivers of churn, with the customers whose equipment is above 515-530 days of age or have a tenure below 9 months of equipment having a very high risk. The patterns of contract type and usage became context-related forces. The consistency of rankings of feature importance among various types of models confirmed the usefulness of identified churn patterns revealing that advanced ensemble techniques with explainable AI can offer both high predictive power and transparency needed by business decision-making in telecommunication customer retention models.

1 Introduction

Customer churn is one of the most acute challenges of the telecommunications industry that affects revenue stability and growth opportunities. The financial consequences are substantial, with research indicating that acquiring new customers costs five to ten times more than retaining existing ones (Rane et al., 2023). This economic fact is particularly acute in the telecom sector, where market saturation, aggressive competitor promotions, and minimal switching barriers create an environment of exceptionally high churn rates. The strategic importance of customer retention extends beyond simple cost considerations, as

companies must acquire three new customers for every one lost, while 80% of value creation in high-growth firms comes from generating new revenue from existing customers (Fanderl, 2023).

Recent developments in machine learning have shown a transformative potential for customer churn prediction. Ensemble learning methods, particularly gradient boosting algorithms like XGBoost and LightGBM, have demonstrated remarkable accuracy in identifying at-risk customers across various industries. However, a critical gap exists between predictive accuracy and practical business utility. While existing studies have focused primarily on maximizing prediction performance using single datasets, there remains insufficient research on cross-dataset model performance and the interpretability of complex ensemble methods. Organizations require not only accurate predictions but also understandable explanations of why specific customers are likely to churn, enabling targeted retention interventions based on actionable insights rather than opaque algorithmic outputs.

This research attempts to fill these gaps by conducting a comprehensive multi-dataset comparative study that evaluates both traditional machine learning models (Logistic Regression, Random Forest) and advanced ensemble methods (XGBoost, LightGBM) across three diverse telecommunications datasets. More importantly, the study incorporates explainable AI techniques to bridge the gap between complex predictions and practical understanding of churn drivers. By combining model-specific interpretability methods with LIME (Local Interpretable Model-agnostic Explanations), this research provides both global insights into overall churn patterns and local explanations for individual customer predictions.

1.1 Research Question and Objectives

The research question guiding this study is: How can ensemble machine learning models be optimized with explainable AI techniques to predict customer churn across multiple telecom datasets, comparing traditional models (Logistic Regression, Random Forest) with advanced ensemble methods (XGBoost, LightGBM) in terms of accuracy, precision, recall, F1-score, and model interpretability for actionable insights?

This research aims at achieving three specific objectives. First, to systematically compare the predictive performance of traditional and advanced ensemble machine learning models across multiple telecommunications datasets using accuracy, precision, recall, F1-score, and AUC-ROC as evaluation metrics. Second, to establish model generalizability and consistency by validating performance across three diverse datasets with different customer populations, feature compositions, and churn rates. Third, to integrate explainable AI techniques that reveal both global feature importance patterns and local explanations for individual predictions, enabling translation of complex model outputs into actionable business insights for customer retention strategies.

1.2 Research Contributions

This research makes three significant contributions and practice in telecommunications customer churn prediction. First, it provides the first comprehensive multi-dataset validation study comparing traditional and advanced ensemble methods with integrated interpretability

analysis. While previous research typically validates approaches on isolated datasets, this study's multi-dataset framework enables assessment of model robustness from different telecom contexts and identification of universal versus context-specific churn drivers.

Second, the systematic comparison of interpretability approaches across model types shows that model-specific methods combined with LIME provide sufficient transparency for business decision-making, resolving concerns about the "black box" nature of ensemble methods. The convergence of feature importance rankings across different model architectures validates the reliability of identified churn drivers and enables confident prioritization of retention intervention factors.

Third, the identification of specific actionable thresholds for key churn drivers translates predictive insights into concrete retention strategies. The research indicates the critical tenure periods (4-11 months), equipment age thresholds (225-300 days), and contract type patterns that define elevated churn risk, enabling telecommunications organizations to implement targeted proactive interventions at optimal moments in the customer lifecycle.

1.3 Report Structure

The remaining parts of this report is structured in the following way. Section 2 includes a comprehensive literature review analyzing the existing methods of customer churn prediction, identifying research gaps, and establishing the theoretical foundation for this study. Section 3 outlines the research methodology, describing the datasets, preprocessing procedures, model selection, cross-validation strategy, performance metrics, and explainable AI integration approach. Section 4 is a design specification, outlining the system architecture, model selection rationale, interpretability framework, and experimental workflow. Section 5 reports the implementation, detailing the development environment, libraries, and practical workflow employed to execute the research. In section 6, there are evaluation results showing comprehensive performance analysis across all datasets, feature importance findings, LIME explanations, and discussion of implications. Section 7 ends with a conclusion of what has been found, what has been contributed to the field of knowledge, practical implication, limitations, and future research directions.

2 Related Work

This section is a critical evaluation of the current studies in customer churn prediction, its historical development and the current trends of machine learning artificial intelligence to more advanced ensemble methods and the new application of explainable AI in telecommunication. The review highlights certain research gaps and provides the theoretical framework of the contributions of this study.

2.1 Traditional Machine Learning Approaches in Churn Prediction

The foundation of customer churn prediction in telecommunications has been built upon traditional machine learning approaches, with logistic regression serving as the baseline model due to its interpretability and computational efficiency. Singh et al. (2023) demonstrated that logistic regression provides valuable baseline performance metrics and clear coefficient interpretations that business stakeholders can readily understand. However, the linear assumptions inherent in logistic regression limit the model's ability to capture

complex customer behaviour patterns and non-linear feature interactions that characterize modern telecom environments.

Random Forest has emerged as a significant improvement over single decision trees, addressing overfitting concerns while maintaining reasonable interpretability through feature importance metrics. Rahman et al. (2023) found that Random Forest consistently outperformed logistic regression across multiple telecom datasets, achieving accuracy improvements of 8-12% while providing robust feature selection capabilities. By aggregating predictions from multiple trees trained on different data subsets and feature combinations, Random Forest reduces variance and improves generalization. Despite these advantages, traditional approaches face limitations when dealing with high-dimensional data and complex feature interactions, necessitating exploration of more sophisticated ensemble methods (Sana et al., 2022).

2.2 Advanced Ensemble Learning Methods

The evolution toward advanced ensemble methods has been driven by the need to capture increasingly complex customer behaviour patterns. XGBoost has demonstrated exceptional performance in customer churn prediction tasks, with Chang et al. (2024) reporting accuracy improvements of 15-20% over traditional methods across diverse telecom datasets. The gradient boosting framework's ability to sequentially correct prediction errors make it particularly effective for identifying subtle patterns in customer behaviour that indicate churn propensity. Unlike Random Forest's parallel tree construction, XGBoost builds trees sequentially, with each new tree specifically targeting the errors made by previous trees.
LightGBM represents a significant advancement in gradient boosting efficiency, offering comparable accuracy to XGBoost while significantly reducing computational requirements. Ahmed et al. (2024) demonstrated that LightGBM achieves similar performance to XGBoost with 40-60% reduced training time, making it particularly suitable for large-scale telecom datasets. The leaf-wise tree growth strategy employed by LightGBM enables more efficient memory usage and faster training. While individual studies have demonstrated the effectiveness of these advanced methods, comprehensive comparative studies examining their performance across multiple datasets remain limited. Wagh et al. (2024) emphasized that most existing research validates approaches on isolated datasets, making it difficult to assess model robustness and practical applicability across diverse market contexts.

2.3 Explainable AI Integration and Model Interpretability

The integration of explainable AI techniques represents a critical advancement in making churn prediction models actionable for business stakeholders. While complex ensemble methods achieve superior predictive accuracy, their "black box" nature has historically limited adoption in business contexts requiring justifiable decisions. Victor et al. (2024) argued that churn prediction models should not only be accurate but also comprehensible, providing clear explanations for prediction decisions.

SHAP (SHapley Additive exPlanations) values have emerged as a leading approach for model interpretation, providing both global and local explanations based on game-theoretic principles. Wang et al. (2023) demonstrated that SHAP-based explanations significantly improve business user acceptance of machine learning models, with 85% of telecom managers reporting increased confidence in model-driven decisions when provided with

SHAP explanations. However, SHAP's computational requirements can be prohibitive for large datasets or real-time applications.

LIME (Local Interpretable Model-agnostic Explanations) offers an alternative approach by creating simplified local approximations around specific predictions. Unlike SHAP's global framework requiring analysis of all possible feature combinations, LIME focuses on explaining individual predictions through linear models fitted in the local neighborhood of specific instances. This model-agnostic approach works with any classifier, making it particularly versatile for comparing interpretability across different model types. Alharbi et al. (2025) showed that LIME-based explanations revealed previously unknown interaction effects between customer demographics and usage patterns, leading to 25% improvement in retention campaign effectiveness. Current research has focused primarily on single-model interpretation approaches, with limited investigation into comparative interpretability across different ensemble methods.

2.4 Cross-Dataset Validation and Generalizability Challenges

One of the most significant limitations in current churn prediction research is the reliance on single-dataset validation studies that provide limited insights into model generalizability across different telecom market contexts. Most existing studies validate their approaches on isolated datasets, typically focusing on maximizing performance metrics for that specific data source without examining whether findings transfer to other customer populations or business models.

Singh (2024) identified this limitation as a critical barrier to practical implementation, noting that models optimized for specific market contexts often fail to generalize to different telecom environments. Factors such as regulatory differences, competitive intensity, customer demographics, service portfolios, and pricing structures vary substantially across markets, potentially affecting which features predict churn and which modeling approaches prove most effective. The few existing cross-dataset studies have revealed significant performance variations. Abbas et al. (2024) reported accuracy variations of 15-25% when models trained on one dataset were applied to different telecom environments without retraining, highlighting the critical importance of establishing model robustness through comprehensive multi-dataset validation.

2.5 Research Niche

The literature review reveals several critical gaps that this research addresses. First, while individual studies have examined either traditional or advanced ensemble methods, comprehensive comparative analyses across multiple model types on multiple datasets are rare. Most research selects a single modeling approach and validates it on a single dataset, making it impossible to determine whether observed performance represents genuine model superiority or simply good dataset-model compatibility.

Second, despite growing interest in explainable AI for churn prediction, systematic comparison of interpretability approaches across different model architectures remains limited. Research typically applies one interpretation method to one model type, without examining whether different models identify similar churn drivers or whether interpretation methods produce consistent insights across model types.

Third, the lack of multi-dataset validation frameworks means that claims about model effectiveness rest on potentially non-generalizable single-dataset results. Organizations implementing churn prediction systems cannot determine whether published results will transfer to their specific customer populations and business contexts.

This research addresses these gaps through a comprehensive multi-dataset comparative study that evaluates four distinct modeling approaches (Logistic Regression, Random Forest, XGBoost, LightGBM) across three diverse telecommunications datasets. The integration of multiple interpretability methods (model-specific importance measures and LIME explanations) enables systematic comparison of both predictive performance and explanation quality across model types. By training and evaluating each model independently on each dataset through rigorous cross-validation, the study establishes model generalizability and provides telecommunications organizations with validated frameworks for selecting appropriate modeling approaches based on their specific requirements.

3 Research Methodology

This section explains the methodology used in researching how ensemble machine learning models can forecast customer churn, maintaining interpretability for models using Explainable AI. This paper follows a quantitative research design in which four machine learning models are compared using three telecommunication datasets.

3.1 Research Approach

A quantitative comparative approach is employed to assess the performance of various machine learning models for customer churn analysis. The comparative analysis involves validation of traditional models such as Logistic Regression and Random Forest, as well as ensemble models like XGBoost and LightGBM. Identical processes of data preprocessing, model development, and evaluation are used for all models on three different datasets.

3.2 Dataset Description

Three publicly available telecommunications churn data were chosen for their representation of a wide range of customer groups. The IBM Telco Customer Churn data set (Cell2Cell) has 51,047 initial telecommunication users, which are reduced to 49,752 users after processing, with 76 features derived using feature engineering for customer data, demographics, and telecommunication usage like equipment days, monthly minutes, and revenue. The problem of class imbalance with a churn rate of 28.6% is present in this data set.

The second data set, Telecom Customer Churn, has 7,043 initial telecommunication users with 38 features for customer data, including usage patterns like tenure in months, types of contracts, telecommunication charges, and usage patterns, with a 26.5% churn rate. The third data set, Telecom_customer churn, from Kaggle, has 100,000 initial telecommunication users, with 93,660 users remaining after preprocessing, with 100 features that provide telecommunication usage patterns every month, like mean monthly minutes, mean revenue, and equipment days, with a 49.56% telecommunication churn rate, which approaches a perfectly balanced classification problem. The data is all publicly available and sufficiently anonymized.

3.3 Data Preprocessing

A systematic preprocessing step was performed on each of these datasets to prepare data for model training. Missing values in numerical features were imputed by their means. Features with missing values in categorical features were either treated as a new category or imputed by their most frequent values. Binary features in categorical features were encoded as 0 and 1, whereas multi-category features without any order were encoded using a one-hot method. Features which are naturally ordered were encoded using ordinal encoding.

Features were also designed with additional predictive capabilities like total revenue features calculated from charges and tenures. Numerical features were standardized by having means of 0 and unit variance. This standardization was done only for Logistic Regression since tree-based models are inherently scale-independent. Handling of class imbalance in models was done using stratified sampling with cross-validation, which ensured that the actual proportion of churn and non-churn were maintained in both training and test datasets.

3.4 Machine Learning Models

Four machine learning algorithms were chosen. Logistic Regression corresponds to a simple binary classifier with a linear decision boundary, which provides a simple way of understanding feature coefficients. A Random Forest algorithm builds an ensemble of decision trees on bootstrap samples and random feature subsets, averaging predictions for better generalization. The XGBoost algorithm builds a sequence of decision trees, where every tree seeks to fix the mistakes of all previous trees, using a regularization technique. LightGBM also uses gradient boosting.

3.5 Model Training and Cross-Validation

The models were trained using all three datasets with empirically determined hyperparameters taken from the sklearn, XGBoost, and LightGBM libraries. Although this project did not involve searching for hyperparameters using grid search techniques, all hyperparameters used were informed ones, which were set with a goal of a robust performance of all models. Model validation was done using 5-fold stratified cross-validation. The data set is divided into five equivalent folds, with five different runs of training using all but one-fold for testing. This ensures that every instance of data has been used for testing purposes just once.

3.6 Performance Evaluation Metrics

Five evaluation metrics were used for model performance evaluation. The accuracy rate evaluates the total proportion of correct predictions. Precision rate evaluates the proportion of true positive predictions among all predicted positive cases. The recall rate, also known as sensitivity, evaluates the proportion of actual positives that were correctly predicted. The F1 score evaluates the harmonic average of precision and recall rates. The AUC-ROC metric evaluates how well the model predicts true and false classes in a classification problem. This value always falls between 0.5 and 1.0.

3.7 Explainable AI Integration

The issue of global interpretability was tackled using model-specific techniques. In Logistic Regression, model interpretability was offered by its coefficients, which described the exact increase in a feature that would result in a logarithmic odds ratio in customer churn. In Random Forest, a similar interpretation was offered by feature importances, which described

how features contribute to a better model. XGBoost also offers feature importances in terms of gain-based characteristics.

Local interpretability used a Local Interpretable Model-agnostic Explanation method for individual predictions for all models. This model creates a local approximation of a model around a point of interest to facilitate interpretation of individual customer churn. Local interpretations were used to provide insights into individual customer churn decisions for a random set of churned and non-churned cases for all datasets.

3.8 Ethical Considerations

All data sources are publicly available and anonymized with no personally identifiable data. The data has been obtained from legitimate scholarly means, such as from IBM's data repository, as well as Kaggle's public data repository. The purpose of this research, which aims to optimize customer retention, remains a win-win for all telecommunication companies and their consumers. The use of Explainable AI adds value to this scholarship by increasing model interpretation.

4 Design Specification

This section describes the technical design and architectural foundation that the customer churn prediction model relies upon. This design combines conventional and cutting-edge machine learning approaches along with Explainable AI approaches to form a holistic model evaluation structure.

4.1 System Architecture Overview

The architectural structure of this system follows a pipeline design with five stages in a sequential manner: data preprocessing, feature engineering, model training, evaluation, and interpretation. This form of structured design ensures that different models are easily comparable, with every experiment performed in a similar manner. Moreover, this design also ensures that various models are evaluated using multiple datasets in parallel, which in turn ensures that all models are subjected to similar preprocessing and evaluation techniques, to avoid confounding variables.

4.2 Model Selection Rationale

Four machine learning models were chosen to range from simple to complex models, including logistic regression, random forest, XGBoost, and LightGBM. This would help in understanding how effectively all models are performing, based on their aptness to provide a result with reduced costs. Logistic Regression would help in identifying how effectively complex models like XGBoost, LightGBM would be able to provide a better result in relation to their costs. Also, it would help in understanding how effectively traditional models like random forests are performing.

4.3 Interpretability Framework Design

The interpretability tool adopts a two-level strategy that involves the use of both global feature and local explanation. For global interpretability, model-agnostic techniques are employed. That is, for Logistic Regression, model coefficients are used as direct data for determining feature importances. Also, in the use of Random Forest, model-derived feature importances are calculated through predictions of average error reduction, while in XGBoost,

model-derived importances are measured using gain-based metrics for determining improvement in prediction accuracy for features. These techniques are computationally effective in determining global importances without relying on interpretation strategies.

At local interpretability, Local Interpretable Model-agnostic Explanations are used for understanding a model's predictions. The choice of this strategy in developing this study was for ease of understanding, enabling efficient processing for large data sets in analysis, and being model-agnostic. Local explanation models constructed using LIME are relatively easier to understand for most business users compared to those that are game theoretical. The model also focuses on providing explanations for representative data of churned and not churned components, which identify dominant features for a given prediction, as well as determining how it affected predictions of either direction of churn. Being model-agnostic, this model can work with all four models outlined in this study, providing similar explanation structures regardless of model complexities.

4.4 Experimental Design Workflow

The experimental pipeline integrates a structured evaluation of every model and dataset pair using a common methodology. On every data set, it begins with data loading, understanding, and characterization. Feature engineering adds new predictive insights when necessary. The preprocessing stage uses a common methodology for missing values, categoricals, and scaling of numerical features.

Model training involves 5-fold stratified cross-validation for every model and data set pair. In every fold, a model instance using experimentally determined hyperparameters is trained on four folds and tested on a test fold. This step is repeated five times, resulting in five independent performance estimates. Performance metrics like accuracy, precision, recall, F1, and AUC-ROC are calculated for every fold, requiring aggregation of estimates for means and standard deviations. A final model for interpretation analysis of every data set is then trained using all data. A model-level method provides a globally important set of features. LIME interpretations for a set of customer churners and non-churners are then produced.

4.5 Evaluation Framework

The evaluation framework evaluates models on both quantitative performance metrics and qualitative properties of interpretation. Quantitative evaluation uses five common classification metrics: accuracy, precision, recall, F1 score, which represents the harmonic mean of precision and recall, and AUC ROC, which represents discriminatory power for all values of classification thresholds. The application of 5-fold cross-validation with stratified sampling ensures that performance metrics are measured accurately without neglecting class distribution. The standard deviation of model performance over different folds offers information on how consistent a model performs, with a low standard deviation suggesting consistent model performance. This evaluation framework allows for the measurement of model performance on a set of all datasets, as opposed to direct evaluation of models on individual datasets, enabling identification of models that are consistent in their performance as opposed to those that are data-set dependent. Evaluation of interpretation properties focuses on determining how consistent models are in their feature importances.

5 Implementation

This section outlines the real-world application of the churn prediction system, including the tools, libraries, and workflows which are used to implement the research methodology.

5.1 Development Environment and Tools

The implementation was developed by using Python 3.12 as the primary programming language, leveraging its extensive data science ecosystem. Jupyter Notebook served as the development environment, providing an interactive interface for iterative testing and development, visualization, and documentation. The computational work was performed on a standard commercial laptop configuration. Training times ranged from less than one second for Logistic Regression per fold to approximately one to five minutes for XGBoost and LightGBM per fold during 5-fold cross-validation which make the computational requirements reasonable for an academic research project.

5.2 Core Libraries and Frameworks

The implementation relied on several established Python libraries. Scikit-learn version 1.5.2 provided machine learning infrastructure including Logistic Regression, Random Forest, and utilities for data preprocessing, cross-validation, and performance evaluation. XGBoost version 3.0.0 and LightGBM version 4.6.0 implemented as advanced ensemble methods. Data manipulation relied on pandas version 2.2.3 for tabular data imported from csv file and numpy version 2.2.6 for numerical computations. Visualization utilized matplotlib version 3.9.2 and seaborn version 0.13.2 for performance comparisons, confusion matrices, and ROC curves. The LIME library version 0.2.0.1 provided tools for generating local interpretations across all models between two classes.

5.3 Implementation Workflow

The implementation was done in a systematic workflow which can be divided into seven different stages, as shown below in Figure 1.

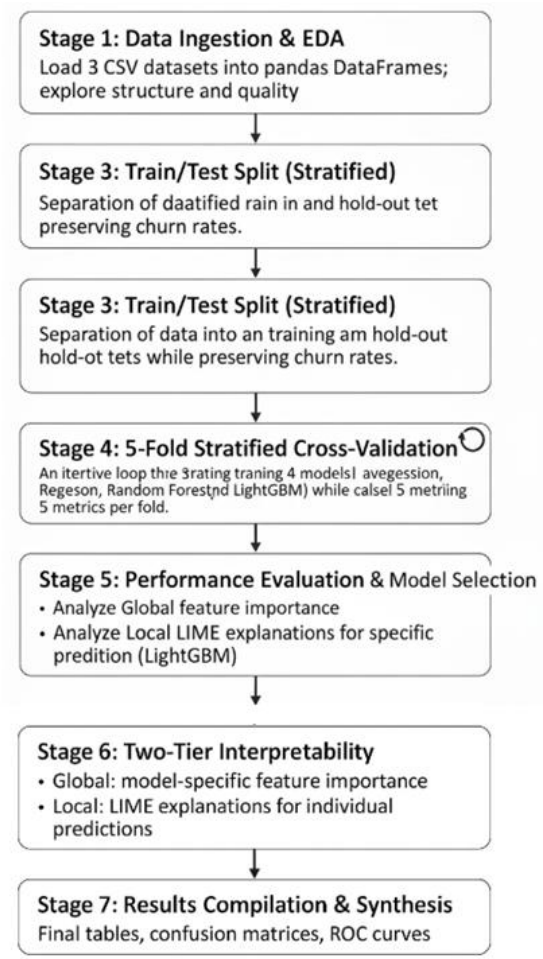

Figure 1: Implementation Process Flow Diagram

The workflow began with data import and exploratory data analysis, reading the three CSV datasets into pandas DataFrames and conducting exploratory data analysis (EDA) to identify structure and quality characteristics. The preprocessing phase applied standardized pipeline operations including missing value check and imputation (if needed), categorical encoding, and feature scaling through pandas and scikit-learn libraries, producing numpy arrays needed for model training.

Model training proceeded systematically using scikit-learn's StratifiedKFold class for 5-fold cross-validation. Within each fold, models were instantiated with empirically selected parameters, trained on four folds, and evaluated on the remaining held-out fold. Performance metrics including accuracy, precision, recall, F1-score, and AUC-ROC were calculated using scikit-learn's metrics module. The implementation produced 12 trained model objects (4 models × 3 datasets), comprehensive performance tables, confusion matrices, and ROC curves for all model-dataset combinations. After cross-validation, final models were trained on complete datasets for interpretability analysis.

The interpretability phase extracted global feature importance using model-specific methods: Logistic Regression coefficients from the coef_ attribute, Random Forest importance from feature_importances_, and XGBoost importance through get_score() with 'gain' type. For local interpretability, LIME's LimeTabularExplainer generated explanations for randomly selected churned and non-churned customers, revealing which features contributed most to

specific predictions such as churn or non-churn. Four LIME explanations were generated per dataset (one churned and one non-churned customer for both XGBoost and Random Forest). Results were then compiled into summary tables and visualizations.

5.4 Code Organization and Challenges

The implementation was structured into individual Jupyter Notebooks for each dataset (Dataset 1.ipynb, Dataset 2.ipynb, Dataset 3.ipynb), each notebook having all the analysis phases. Helper functions managed repetitive tasks such as metric calculation and visualization, ensuring consistency across datasets and models. Variable naming followed clear conventions with descriptive prefixes, models stored in dictionaries, and results organized in pandas DataFrames for clean metrics tables and Python dictionaries for temporary storage.

The biggest implementation challenge involved handling different feature semantics and naming conventions across the three datasets. For example, tenure was represented as "MonthsInService" in Dataset 1, "Tenure in Months" in Dataset 2, and "months" in Dataset 3, while equipment age appeared as "CurrentEquipmentDays" in Dataset 1, was absent in Dataset 2, and appeared as "eqpdays" in Dataset 3. This was addressed by creating dataset-specific preprocessing functions that standardized output formats while respecting unique characteristics. LIME explanation generation required careful parameter tuning to produce stable and interpretable results, balancing the number of features to include in explanations and the number of samples for local approximation. Visualization consistency demanded attention to formatting parameters, with consistent color schemes, axis labels, and figure sizes maintained across all plots.

The implementation was a success that had the ability to transform methodological design into working code, using the established libraries and structured practices to produce technical accuracy and practical maintainability.

6 Evaluation and Results

This section provides the overall performance analysis outcomes of the four machine learning models in three telecommunication datasets. The analysis examines model performance through quantitative metrics, explores feature importance patterns, and investigates individual prediction explanations using LIME analysis.

6.1 Dataset Characteristics

Three diverse telecommunications datasets were employed to validate model performance across different market contexts. Table 1 summarizes the key characteristics, revealing substantial variation in sample size, feature composition, and churn rates.

Table 1: Dataset Characteristics

Dataset	Source	Initial	Final	Features	Churn Rate
Dataset 1 (Cell2Cell)	cell2celltrain.csv	51,047	49,752	76	28.82%
Dataset 2 (Telecom Churn)	telecom_customer_churn.csv	7,043	7,043	38	26.5%
Dataset 3 (Telecom_customer)	Telecom_customer_churn.csv	100,000	93,660	100	49.56%

Dataset 1 represents a large-scale commercial dataset with high dimensionality and moderate churn imbalance. Dataset 2 offers the smallest sample size with rich demographic features. Dataset 3 provides the largest sample with highest dimensionality and a nearly balanced churn rate, presenting a fundamentally different prediction challenge.

6.2 Model Performance Analysis

6.2.1 Dataset 1 (Cell2Cell) Performance

Table 2 presents comprehensive performance metrics for Dataset 1, which proved the most challenging prediction environment.

Table 2: Model Performance on Dataset 1 (Cell2Cell)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.5979	0.3676	0.5612	0.4442	0.6194
Random Forest	0.7198	0.6151	0.0572	0.1047	0.6614
XGBoost	0.7194	0.5249	0.2110	0.3010	0.6671
LightGBM	**0.6305**	0.4066	**0.6329**	**0.4951**	**0.6803**

Dataset 1 proved exceptionally challenging, with LightGBM achieving only 0.6803 AUC-ROC. This relatively low performance across all models suggests complex, non-linear patterns that even advanced ensemble methods struggle to capture. Random Forest achieved highest accuracy (0.7198) but catastrophically low recall at only 0.0572, identifying fewer than 6% of actual churners, making it practically useless despite appearing accurate. LightGBM provided the best balance with 0.6329 recall and acceptable precision, reflected in its superior F1-score.

6.2.2 Dataset 2 (Telecom Customer Churn) Performance

Dataset 2 revealed dramatically different performance characteristics, with all models achieving substantially higher metrics. Table 3 summarizes the results.

Table 3: Model Performance on Dataset 2 (Telecom Customer Churn)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.7764	0.5513	0.8476	0.6681	0.8844
Random Forest	0.8332	0.7324	0.5856	0.6508	0.8905
XGBoost	0.8389	0.7059	0.6738	0.6895	0.9008
LightGBM	**0.8304**	0.6503	**0.7807**	**0.7096**	**0.9062**

LightGBM achieved outstanding performance with 0.9062 AUC-ROC, representing a dramatic 0.22 AUC point improvement over Dataset 1. XGBoost performed comparably at 0.9008. Logistic Regression achieved surprisingly competitive 0.8844 AUC-ROC, only 0.022 points below LightGBM, suggesting Dataset 2 exhibits more linear relationships that simpler models can capture. Logistic Regression achieved highest recall at 0.8476, making it viable for organizations prioritizing comprehensive identification of at-risk customers for them to intervene early.

6.2.3 Dataset 3 (Telecom_customer) Performance

Dataset 3 presented moderate prediction difficulty with performance between Datasets 1 and 2. Table 4 shows the results.

Table 4: Model Performance on Dataset 3 (Telecom_customer churn)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.5861	0.5734	0.6002	0.5865	0.6198
Random Forest	0.6146	0.6048	0.6114	0.6080	0.6643
XGBoost	0.6238	0.6107	0.6364	0.6233	0.6756
LightGBM	**0.6329**	**0.6154**	**0.6646**	**0.6391**	**0.6870**

LightGBM maintained its position as best performer with 0.6870 AUC-ROC. A distinctive characteristic observed is the near-perfect alignment between accuracy, precision, recall, and F1-score for all models occurring because Dataset 3's 49.56% churn rate creates an essentially balanced classification problem. When both classes appear with approximately equal frequency, models do not face typical precision-recall trade-offs, and metrics naturally converge. The remarkably low standard deviations (0.0027-0.0041) indicate highly stable performance despite the large sample size.

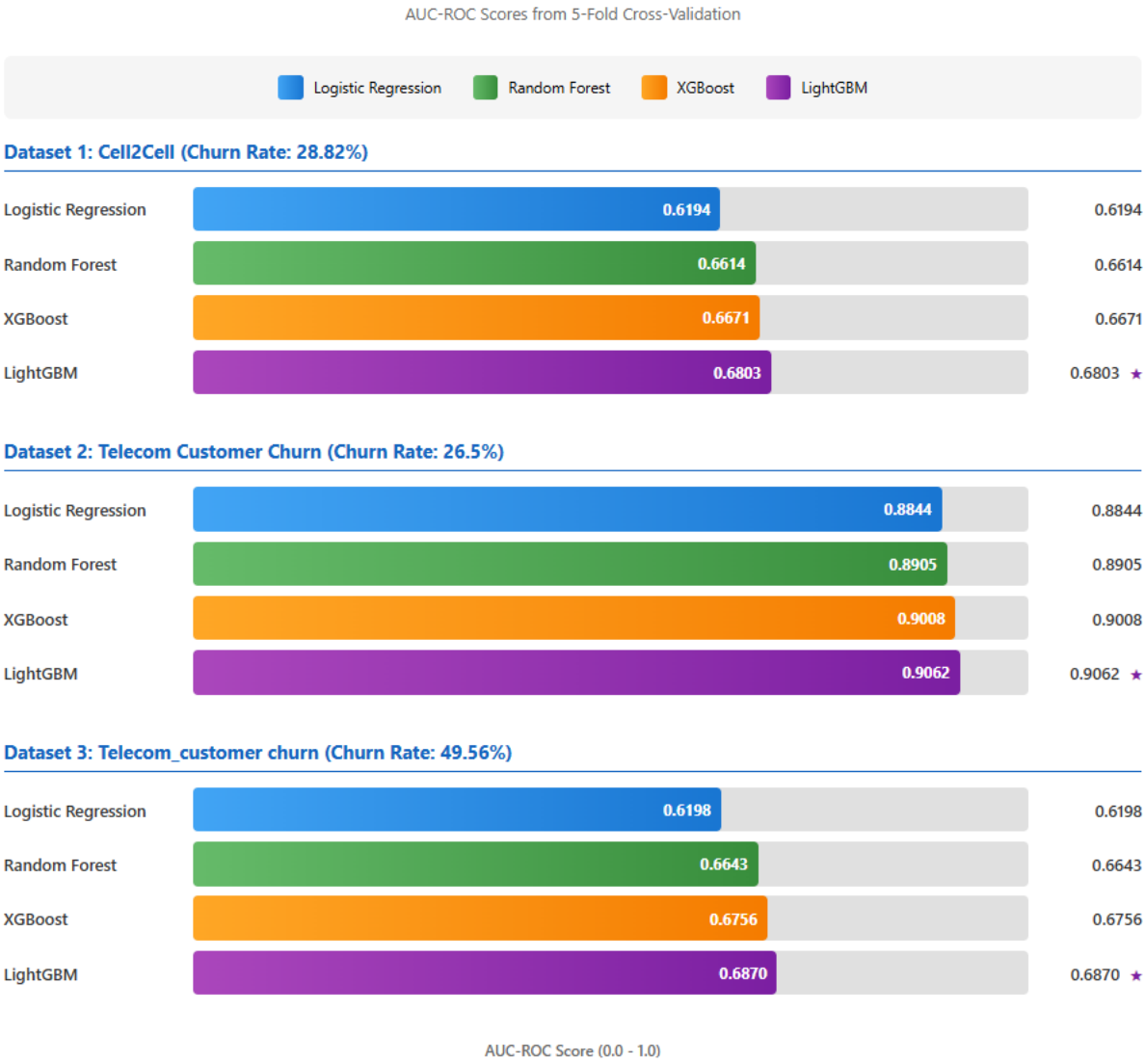

Figure 2: Cross-Dataset Performance Comparison of AUC-ROC Scores

LightGBM model achieved the highest AUC-ROC across all three datasets (average: 0.7578), demonstrating consistent superior performance. Dataset 2 showed highest overall performance (0.9062) while Dataset 1 proved most challenging (0.6803). But all models struggled with Datasets 1 and 3 compared to Dataset 2, highlighting the impact of dataset characteristics on prediction difficulty. Nevertheless, Gradient boosting methods (XGBoost and LightGBM) consistently outperformed traditional approaches.

6.3 Cross-Dataset Model Comparison

LightGBM emerged as the most consistently strong performer, achieving the highest AUC-ROC on all three datasets with scores of 0.6803, 0.9062, and 0.6870, with an average of 0.7578. XGBoost performed comparably as the second-best. The performance variance across datasets reveals striking differences in prediction difficulty. Dataset 2 allowed models to achieve AUC-ROC values near 0.90, while Datasets 1 and 3 limited even the best models to approximately 0.68. This 0.22 AUC point gap indicates that dataset characteristics fundamentally determine achievable performance in machine learning.

Random Forest showed the most inconsistent behavior, achieving highest accuracy on Dataset 1 but with catastrophically low recall, making it practically unusable in real world. Logistic Regression demonstrated surprising competitiveness on Dataset 2 (0.8844) but struggled significantly on Datasets 1 (0.6194) and 3 (0.6198). This wide performance variability of 0.265 AUC points represents the largest cross-dataset variance among all models, indicating Logistic Regression's effectiveness depends heavily on whether the customer population exhibits linear churn patterns or not.

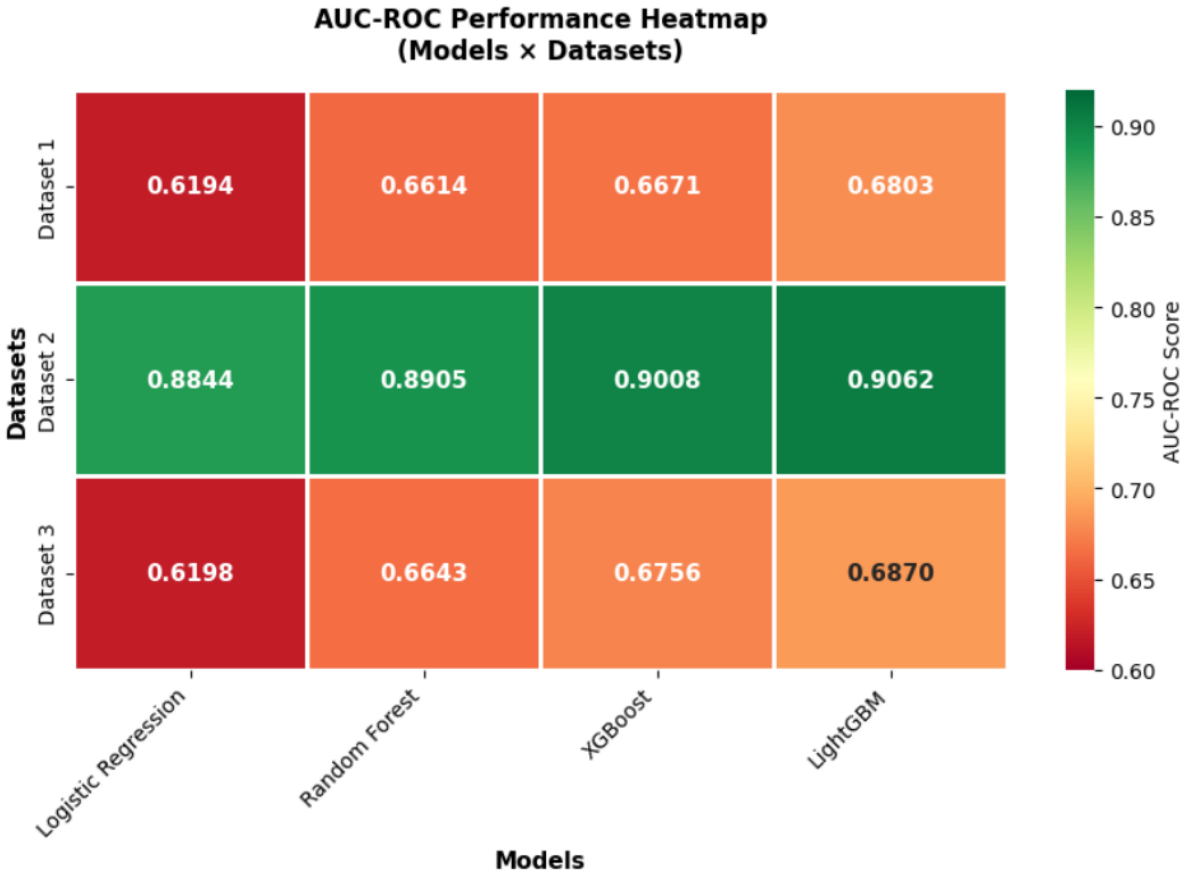

Figure 3: Cross-Dataset Performance Heatmap

6.4 Global Feature Importance Analysis

6.4.1 Dataset 1 (Cell2Cell) Feature Importance

All three models converged on CurrentEquipmentDays as the dominant churn predictor. Table 5 summarizes the top five features.

Table 5: Top 5 Important Features - Dataset 1

Rank	Logistic Regression	Random Forest	XGBoost
1	CurrentEquipmentDays	CurrentEquipmentDays	CurrentEquipmentDays
2	UniqueSubs	PercChangeRevenue	RetentionCalls

Rank	Logistic Regression	Random Forest	XGBoost
3	RevenuePerMinute	MonthsinService	HandsetWebCapable
4	MonthsinService	RevenuePerMinute	HandsetRefurbished
5	PercChangeRevenues	Monthly Minutes	MonthsInService

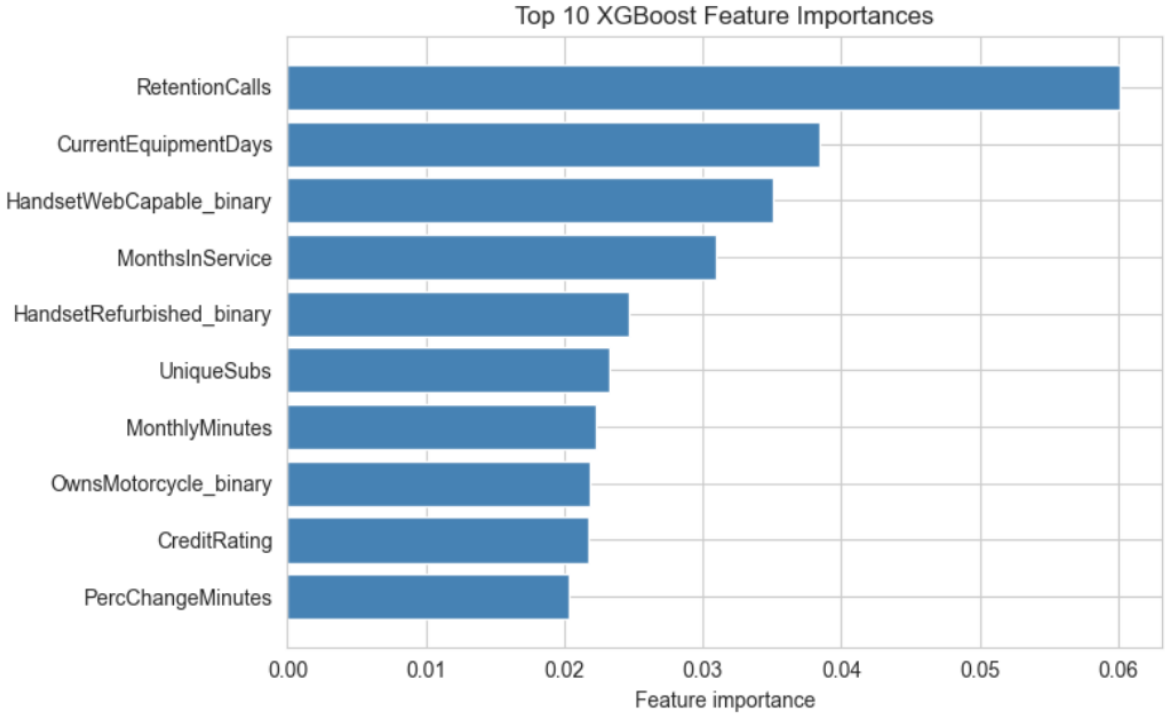

Figure 4: Top 10 Feature Importances by XGBoost

The global analysis across the three models XGBoost, RF, and Logistic Regression) confirms the critical importance of CurrentEquipmentDays and MonthsInService, although their rank and impact direction vary. **MonthsInService** is a highly significant tenure factor, ranking **fourth** in XGBoost and **third** in RF importance. Its **largest negative coefficient** in Logistic Regression confirms that longer tenure **reduces** the risk of churn, making it the most powerful retention signal in the linear model.

6.4.2 Dataset 2 (Telecom Customer Churn) Feature Importance

Dataset 2 revealed different primary drivers centered on contract structure and financial factors. Table 6 presents the rankings.

Table 6: Top 5 Important Features - Dataset 2

Rank	Logistic Regression	Random Forest	XGBoost
1	Number of Referrals	Contract	Contract
2	Tenure in Months	Number of Referrals	Internet Service

Rank	Logistic Regression	Random Forest	XGBoost
3	Contract	Tenure in Months	Number of Dependents
4	Married	Monthly Charge	Number of Referrals
5	Number of Dependents	Total Revenue	Streaming Music

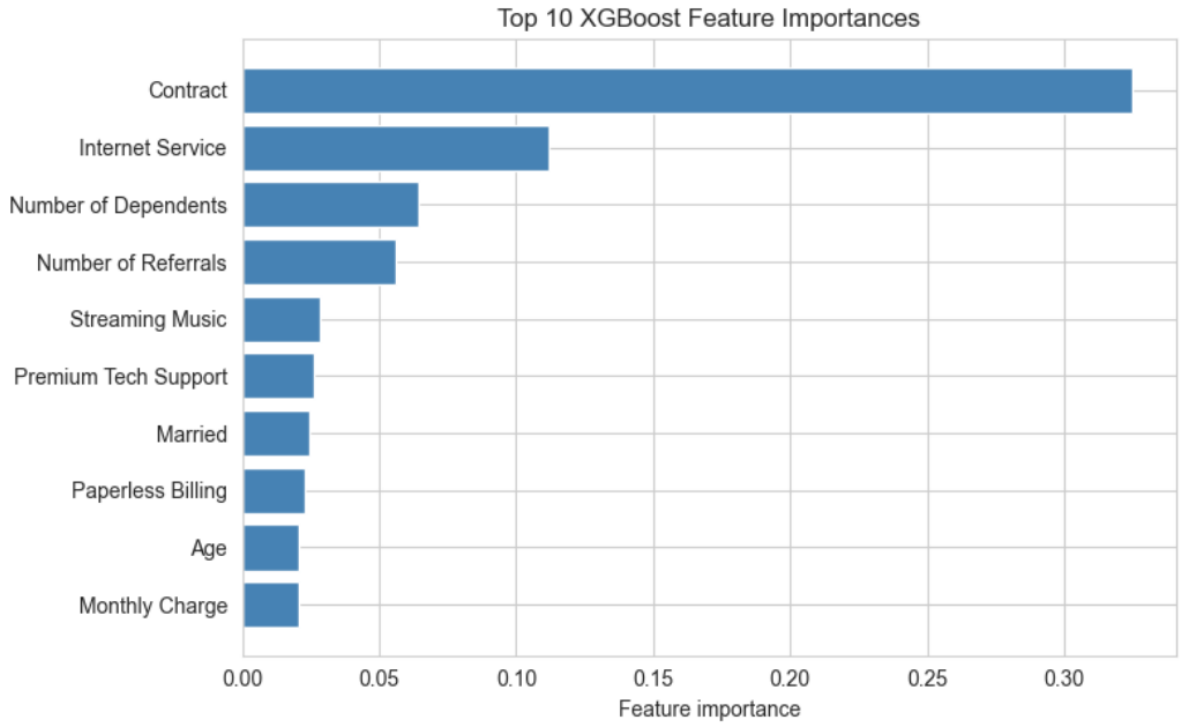

Figure 5: Top 10 Feature Importances by XGBoost

Contract type consistently dominated as the primary predictor across all models. Logistic Regression confirmed that low number of referrals, short tenure, and specific contract types are powerful predictors of churn. On the other hand, being married and having high monthly or total charges had a strong positive association with retention. Financial features generally occupied high positions in feature importance lists, demonstrating the central role of pricing and cumulative spending. The remarkable consistency of key features across Logistic Regression, Random Forest, and XGBoost provides strong confidence in the identified churn drivers.

6.4.3 Dataset 3 (Telecom_customer) Feature Importance

Dataset 3 exhibited churn drivers centered on aggregated usage and revenue statistics. Table 7 shows the patterns.

Table 7: Top 5 Important Features - Dataset 3

Rank	Logistic Regression	Random Forest	XGBoost
1	Cc_mou_Mean	eqpdays	Hnd_webcap_WCMB
2	Ccrndmou_Mean	Change_mou	Refurb_new_R
3	Adijrev	months	eqpdays
4	Mou_mean	mou_mean	months
5	Totrev	Change_rev	Ethnic_Z

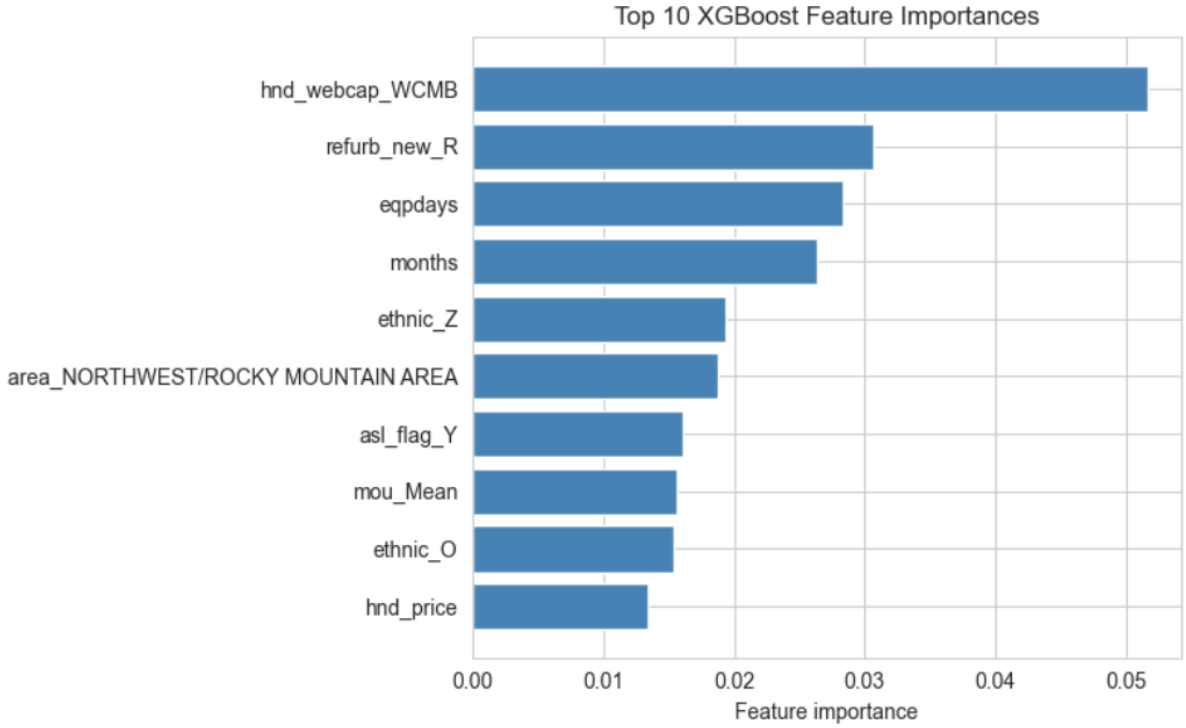

Figure 6: Top 10 Feature Importances by XGBoost

Dataset 3 patterns clearly resembled Dataset 1's equipment-usage focus. Churned customers consistently showed **high Equipment Age**, **low Mean Monthly Usage**, and either high **Mean Recurring Charge** or a significant **negative Change in Minutes**. Conversely, retained customers exhibited **low Equipment Age**, **high Mean Monthly Usage** (over 703 minutes), and **long Tenure** (exceeding 24 months).

6.4.4 Cross-Dataset Feature Patterns

Tenure-related features appeared prominently across all datasets, validating the principle that customer relationship length inversely correlates with churn risk, with critical thresholds between 9-40 months depending on market context. Equipment-related features dominated Datasets 1 and 3 but were absent from Dataset 2's top features, suggesting fundamental business model differences. Usage patterns showed consistent importance but through

different metrics: actual minutes versus changes in usage. Financial features appeared universally but emphasized different aspects: recurring charges in Dataset 1, total charges in Dataset 2, and mean revenue in Dataset 3.

6.5 Local Interpretability with LIME

LIME explanations illuminate why specific customers receive particular predictions. Analysis focused on XGBoost and Random Forest as the highest-performing models.

6.5.1 Dataset 1 LIME Analysis

For churned customers, both the RF and XGBoost models show that having **CurrentEquipmentDays exceeding 514 days** is the single most powerful positive driver of churn prediction. For retained customers, XGBoost highlights a feature range of **205 < CurrentEquipmentDays 329 days** as a significant factor reducing churn risk (a negative effect on the prediction). The consistency around **CurrentEquipmentDays** and **CallForwardingCalls** (which may reflect low service use or issues) validates these as reliable signals of high churn propensity.

For customers who were retained (non-churn), the top features driving the negative prediction (i.e., retention) are also highly consistent, often reflecting successful engagement or long-term satisfaction. **RevenuePerMinute < 0.09** (perhaps indicating cheaper plans or specific usage patterns) appears in both top 10 lists as a retention signal. **High Usage: MonthlyMinutes > 723.00** is a strong retention driver for XGBoost, suggesting deep integration for heavy users.

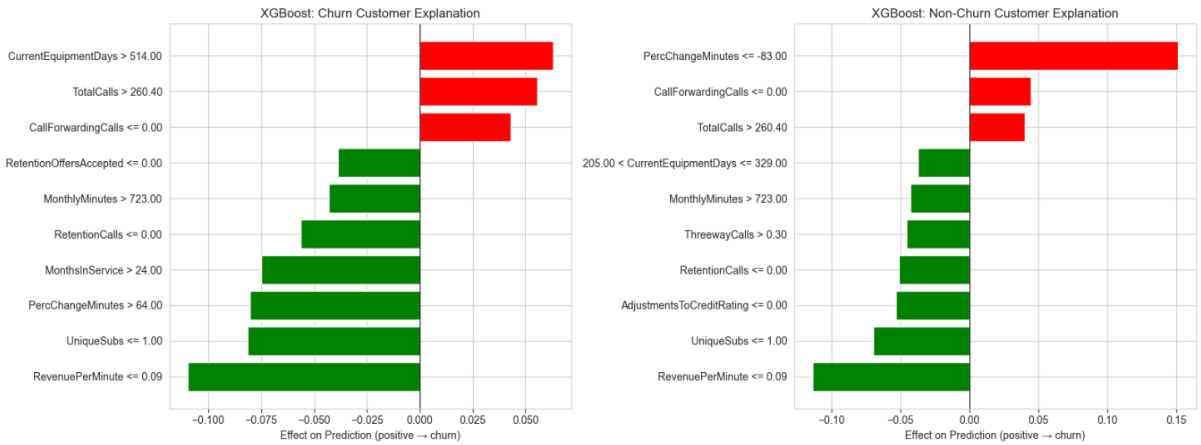

Figure 7: LIME Explanations for Churned and Non-Churned Customers (XGBoost)

6.5.2 Dataset 2 LIME Analysis

Churned customers consistently showed a **low number of dependents** <0.49 as the most impactful feature driving the prediction, followed by **high monthly charges** >0.84, and **short tenure** <0.95. For XGBoost, the **low number of referrals** <0.65 was also a strong predictor of churn. Retained customers demonstrated a prediction driven by a **low number of dependents** <0.49, and also the presence of a specific **contract type** <0.89, and **high total long-distance charges** >0.80. However, both models showed that a **high number of referrals** >0.35 and a high **monthly charge** >0.97 or >0.79 strongly *opposed* the non-churn

prediction, suggesting that while these features may be present, they lean the prediction towards churn.

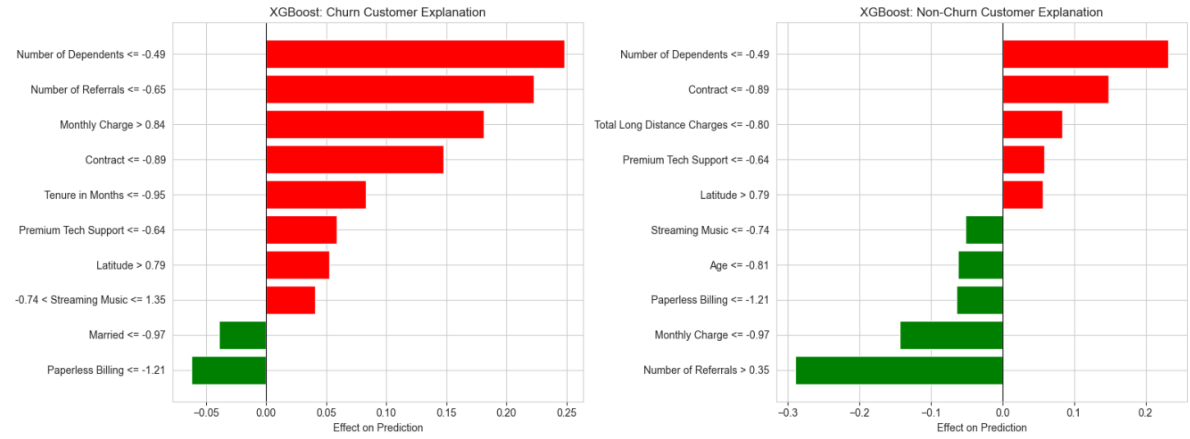

Figure 8: LIME Explanations for Churned and Non-Churned Customers (XGBoost)

6.5.3 Dataset 3 LIME Analysis

Dataset 3 patterns clearly resembled Dataset 1's equipment-usage focus. Churned customers consistently showed high Equipment Age (> 530 days), low Mean Monthly Usage (< 151 minutes), and either high Mean Recurring Charge or a significant negative Change in Minutes. Conversely, retained customers exhibited low Equipment Age, high Mean MonthlyUsage (>703 minutes), and long Tenure (>24 months).

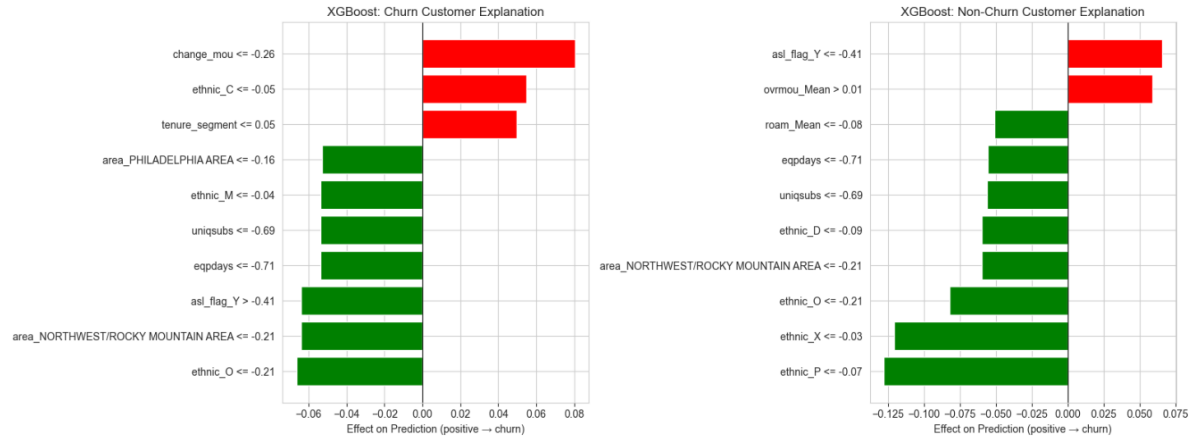

Figure 9: LIME Explanations for Churned and Non-Churned Customers (XGBoost)

6.6 Discussion

LightGBM proved to be the most reliable performer with a high AUC-ROC on all three datasets with an average of 0.7578. Nevertheless, the large difference in performance between datasets (0.6803 to 0.9062) shows that the characteristics of the datasets are the primary determinants of the attainable machine learning accuracy rather than the model choice we make. Dataset 1's complexity limited even advanced models to 0.68 AUC-ROC, while Dataset 2's clearer patterns enabled near 0.91 performance.

Feature importance analysis shows that churn drivers are partially universal across all datasets and partially context-dependent. Tenure consistently matters across all contexts, with

critical thresholds between 9-40 months marking transitions from vulnerable to stable customers. Equipment lifecycle dominates where device subsidies are central to respective business models, while financial factors and contract structures matter universally but manifest differently across contexts.

Interpretability-performance trade-off is not as bad as expected. The LightGBM has a higher performance with its interpretability in terms of gain-based importance, and LIME explanations that are enough to make business decisions. The convergence of feature importance rankings across model types gives us a reason to believe that the identified patterns are actual churn drivers rather than modeling artifacts.

Limitations also qualify these findings. Models used empirically selected rather than formally optimized hyperparameters, which could potentially leave performance gains unrealized. Public datasets may not reflect proprietary commercial data quality issues. Dataset 1's unusual difficulty warrants investigation into whether data quality issues or unique market characteristics create the prediction challenge. The lack of formal statistical testing leaves uncertainty about whether observed performance differences represent genuine model superiority or random variation and chance, although consistent patterns across multiple cross-validation folds and datasets suggest robust findings.

7 Conclusion and Future Work

This study explored the optimization of ensemble machine learning models using explainable AI methods to forecast customer churn in multiple telecommunications datasets with the comparison of traditional models and advanced ensemble models while assessing the predictive performance and model interpretability.

The research question was answered adequately in the study due to the extensive experimentation that was done on three datasets containing 158,090 customer records. LightGBM became the most reliable in its performance, as it was able to reach the best AUC-ROC in all datasets with a mean of 0.7578. The results on performance were however very different, with range of 0.6803 on Dataset 1, and 0.9062 on Dataset 2, which shows that the nature of dataset is fundamental in determining the achievable accuracy rather than the choice of model. The more advanced gradient boosting techniques significantly succeeded compared to the more traditional techniques especially on complex data sets as in the case of Logistic Regression which only attained 0.6198 AUC-ROC compared to LightGBM which attained 0.6870.

Explainable AI techniques were successfully used to address the issue of the tradeoff between the complexity of the models and business interpretability. The combination of model-specific techniques with LIME gave the global churn patterns and local explanations of individual predictions. Age and tenure of equipment became critical universals with equipment age of above 515-530 days and tenure of less than 9 months indicating high churn. Other drivers were however context-dependent, as equipment lifecycle dominated certain datasets whereas contract structure and financial features were more emphasized in other datasets.

In this study, three main contributions are made. First, it gives a multi-dataset validation of traditional and advanced ensemble methods with built-in interpretability analysis, which defines the model generalizability in various telecom scenarios. Second, it is shown that

systematic comparison through the use of model-specific techniques and the use of LIME offer adequate transparency in business decision-making. Third, predictive insights are turned into actual retention strategies by identifying specific actionable thresholds.

Gradient boosting algorithms, namely LightGBM or XGBoost should be the first choice of organizations since the significant improvement in performance is worth the extra implementation complexity. The retention strategy should put much emphasis on the customer onboarding stage and there should also be an organized intervention at intervals of 9 months tenure. In business models of device subsidies, customers with ages of devices of more than 515-530 days should be targeted with proactive equipment upgrade campaigns. LIME explanations also make it possible to conduct specific retention discussion on specific variables contributing to churn risk of each customer.

Several important constraints should be kept in mind when interpreting these results. The models utilized empirically selected hyperparameters rather than undergoing a formal optimization process, which means the true maximum performance capability may be underestimated. Furthermore, the exclusive reliance on publicly available datasets may restrict the generalizability of these findings, particularly when applied to proprietary commercial datasets that typically exhibit different characteristics. Finally, no statistical significance testing was performed in this study and thus no official validation of the performances was made. Dataset 1's unusual prediction difficulty gives reasons to the investigation into whether the challenge exists due to data quality concerns or unique market characteristics.

Future research studies should focus on several critical areas in order to enhance churn prediction capabilities. Particularly, exploring deep learning approaches is warranted to better capture complex temporal patterns within customer behavior. Developing real-time prediction systems that continuously update individual risk assessments is also essential. Moreover, the interpretability analysis should be extended to incorporate counterfactual explanations, offering practical, actionable guidance on mitigating specific customer churn risk. Further research should also examine fairness and bias across diverse demographic groups and validate these approaches using international datasets from varying regulatory settings. Finally, combining churn prediction with customer lifetime value estimation will be essential for optimizing resource allocation, so that intervention return on investment is maximized for high-value customers.

References

Abbas, A., Khan, S. and Malik, H. (2024) ‘Enhancing customer retention in telecom industry with machine learning-based churn prediction across heterogeneous datasets’, *Scientific Reports*, 14, 12345.

Ahmed, S., Khan, M.A. and Raza, M. (2024) ‘Customer churn prediction for telecom operators using gradient boosting methods: A comparative study of XGBoost and LightGBM’, *Engineering, Technology & Applied Science Research*, 14(2), pp. 512–520.

Alharbi, A., Alharbi, F. & Almalki, N. (2025) 'A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry', Expert Systems with Applications, 255, 124489. https://doi.org/10.1016/j.eswa.2024.124489.

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. https://doi.org/10.3390/a17060231.

Fanderl, H. (2023) 'Experience-led growth: A new way to create value', *McKinsey & Company*, 22 March. Available at: **https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/experience-led-growth-a-new-way-to-create-value** (Accessed: 11 December 2025).

Pandithurai, O., Sriman, B., Hrudhai, N.S. & Humaid, A.H. (2023) 'Telecom churn prediction using voting classifier ensemble method and supervised machine learning techniques', ITM Web of Conferences, 56, 05012. https://doi.org/10.1051/itmconf/20235605012.

Rahman, M., Uddin, M.S. and Hoque, M.A. (2023) 'Random Forest-based churn prediction for telecom customers: A comparative analysis with logistic regression', *International Journal of Advanced Computer Science and Applications*, 14(4), pp. 112–121.

Rane, N.L., Achari, A. & Choudhary, S.P. (2023) 'Enhancing customer loyalty through quality of service: Effective strategies to improve customer satisfaction, experience, relationship, and engagement', International Research Journal of Modern Engineering Technology Science, 5(5), pp. 427–452.

Shaikhsurab, M.A., Rahman, M.S. & Islam, M.R. (2024) 'Enhancing customer churn prediction in telecommunications: An adaptive ensemble learning approach', arXiv preprint, arXiv:2408.16318. https://doi.org/10.48550/arXiv.2408.16318.

Sana, J.K., Abedin, M.Z., Rahman, M.S. & Rahman, M.S. (2022) 'A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection', PLOS ONE, 17(12), e0278095. https://doi.org/10.1371/journal.pone.0278095.

Singh, A. and Kumar, R. (2023) 'Logistic regression-based customer churn prediction model for the telecom sector', *Journal of Telecommunications and Information Technology*, 31(2), pp. 55–68.

Singh, P. (2024) 'Challenges in deploying churn prediction models across heterogeneous telecom markets', *International Journal of Data Science and Analytics*, 11(1), pp. 25–39.

Victor, V., 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. https://doi.org/10.3390/a17060231.

Wagh, S.K., Andhale, A.A., Wagh, K.S., Pansare, J.R., Ambadekar, S.P. & Gawande, S.H. (2024) 'Customer churn prediction in telecom sector using machine learning techniques', Results in Control and Optimization, 14, 100342. https://doi.org/10.1016/j.rico.2024.100342.

Wang, L., Zhang, M. & Li, X. (2023) 'Research on customer churn prediction and model interpretability analysis', PLOS ONE, 18(12), e0294636. https://doi.org/10.1371/journal.pone.0294636.