

AI-Based Customer Churn Prediction with Explainability

Omkar Ravindra Salunkhe¹, Aarti Sanjay Gawai²

¹ Student, Department of Information Technology and Computer Science, D.G Ruparel College of Arts, Science and Commerce, Mumbai, Maharashtra, India

² Assistant Professor, Department of Information Technology and Computer Science, D.G. Ruparel College of Arts, Science and Commerce, Mumbai, Maharashtra, India

¹ omsalunkhe694@gmail.com, ² aartigawai1611@gmail.com

ABSTRACT

Customer churn is a significant concern for organizations because losing existing customers can affect revenue, customer relationships, and long-term business growth. Machine learning (ML) techniques are increasingly used to identify customers likely to churn; however, high-performing models may be difficult for business users to understand. This study investigates whether machine learning models can effectively predict customer churn while providing reliable and understandable explanations for their predictions. Five machine learning models are evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC. Explainable Artificial Intelligence (XAI) techniques, particularly SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), are used to identify factors influencing global and individual churn predictions. The study also compares explanations across different models and XAI techniques to examine their consistency and investigates whether the best-performing model provides useful explanations. The experimental results show that Logistic Regression achieved the strongest overall performance according to the selected criteria, with a ROC-AUC of 0.8520 and Recall of 0.7219, while sharing the highest Accuracy of 0.8028 with LightGBM. The explainability analysis showed partial consistency in important customer factors across models, although their relative importance varied. Overall, the findings indicate that predictive effectiveness and explainability can be combined to support transparent and data-driven customer retention decisions.

Key Words: Customer churn prediction, machine learning, explainable artificial intelligence, XAI, SHAP, LIME, Customer retention, model interpretability.

1. INTRODUCTION

Customer churn refers to the situation in which a customer stops using a product or service. For organizations operating in competitive markets, customer loss can have a direct effect on

revenue and make it more difficult to maintain long-term customer relationships. As a result, businesses increasingly analyze customer information to identify individuals who may be at risk of leaving and to take appropriate retention measures before churn occurs. Churn prediction has consequently become an important application of data analytics and artificial intelligence across areas such as telecommunications, banking, financial services, subscription platforms, and digital services. Earlier approaches to studying customer churn relied largely on statistical techniques for identifying customer characteristics associated with service discontinuation. The increasing availability of customer data has encouraged the use of machine learning (ML) methods for this task. ML algorithms can process a large number of customer attributes and learn relationships from historical records that may be difficult to identify through conventional approaches. Algorithms including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost) have therefore been applied to churn prediction. These models use historical customer information to estimate the likelihood that an individual will discontinue a service.

Obtaining a prediction, however, does not necessarily reveal the reasoning behind it. This issue becomes more relevant when complex models such as Random Forest, XGBoost, gradient boosting methods, and neural networks are used. Although such models can capture complex relationships in customer data, their decision-making process may not be straightforward for business users to interpret. For a churn prediction system, identifying a customer as being at high risk is only one part of the problem. Decision-makers may also need to understand the factors contributing to that prediction before determining an appropriate retention response. Characteristics such as contract type, customer tenure, monthly charges, service usage, payment method, and customer support experience may affect the estimated churn risk.

Explainable Artificial Intelligence (XAI) can help address this interpretability issue by providing information about how input features contribute to machine learning predictions. Rather than considering the model output independently, XAI techniques can be used to examine the factors associated with a particular prediction. Among the approaches used for this purpose, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are relevant to churn analysis. SHAP can be used to investigate feature contributions at both an overall and individual level, whereas LIME focuses on explaining the behavior of a model for a specific instance. These techniques can therefore provide

additional information about the factors associated with a customer's predicted churn.

Existing research has demonstrated the application of machine learning and XAI to customer churn prediction. However, many studies primarily concentrate on predictive performance or demonstrate the use of a particular explanation technique. Comparatively less attention has been directed toward determining whether explanations remain consistent when different machine learning models and XAI methods are considered together. A model that achieves strong predictive results may not necessarily produce explanations that are equally useful or consistent. This raises an important question concerning the relationship between predictive performance and interpretability when churn predictions are used to support customer retention decisions.

The present study therefore evaluates whether machine learning models can provide effective churn predictions while also generating understandable and reliable explanations. Five machine learning models are compared using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC. SHAP and LIME are applied to examine the factors associated with churn predictions from global and local perspectives. The study also compares explanations across different models and XAI techniques to determine whether important customer characteristics remain consistent. In addition, the study examines whether the model showing the strongest predictive performance can also provide useful explanations for understanding churn predictions.

The study makes three main contributions. First, it evaluates multiple machine learning models using several performance measures rather than relying on Accuracy alone. Second, it combines SHAP and LIME to investigate feature influence at both the global level and the level of an individual customer prediction. Third, it examines the explanations produced by different predictive models to assess the consistency and usefulness of the identified churn-related factors. Through this approach, the study considers predictive performance and explainability together when analyzing customer churn prediction.

2. LITERATURE REVIEW

Customer churn prediction has become an important area of machine learning research because organizations need to identify customers who may discontinue their products or services. Previous studies have examined churn using a wide range of customer information, including demographic characteristics, financial details, service usage, behavioral patterns, and account-related attributes. These studies have explored different stages of the churn prediction process, from data preparation and feature selection to model development, evaluation, and the use of predictions for business decisions.

Manzoor et al. reviewed more than 200 studies on machine learning-based customer churn prediction and examined the different stages involved in developing churn prediction systems. Their review covered areas such as feature engineering, model development, evaluation, and the use of prediction results in business decision-making. Telecommunications, financial services, and online services

were identified as important application areas. The study also reported an increasing use of ensemble and deep learning methods and pointed out that conventional performance measures alone may not fully reflect the practical value of a churn prediction system.

Identifying the customer characteristics associated with churn is another important part of this research area. Freire et al. conducted a systematic review of customer departure in the telecommunications sector and identified several factors associated with churn, including service quality, pricing, customer service, billing, and service consumption. Their findings suggest that customer churn is not generally determined by a single characteristic. Instead, several customer and service-related factors can interact with one another, making machine learning useful for capturing relationships within churn data.

A number of studies have investigated ensemble and boosting algorithms for improving churn prediction performance. Methods such as Random Forest, XGBoost, LightGBM, and CatBoost have been applied in different churn prediction settings and have reported strong predictive results. These models can capture more complex relationships in customer data; however, their increased complexity can also make it more difficult to understand the reasons behind individual predictions. This creates a practical challenge when prediction results are intended to support business decisions.

Predictive performance alone may therefore not be sufficient when churn models are used in customer retention applications. Business users may need to know why a particular customer has been classified as a potential churner before deciding how to respond. Explainable Artificial Intelligence (XAI) has gained importance in this context because it can provide information about the relationship between input features and model predictions. Freire et al. highlighted that prediction algorithms have received greater attention than explainability in much of the churn prediction literature, indicating the need to better understand the reasons behind churn classifications.

SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are two commonly used approaches for interpreting machine learning predictions. LIME explains an individual prediction by creating a simpler approximation of model behavior around a particular customer instance. SHAP, on the other hand, estimates the contribution of individual features to a prediction and can be used to study both overall feature importance and individual predictions. Their use allows churn prediction systems to provide more information than a probability or class label alone.

Recent studies have increasingly combined machine learning with XAI techniques for customer churn analysis. Mahin et al. investigated an XAI-based neural-machine learning ensemble for business customer churn prediction, while Chaudhary et al. examined explainable machine learning approaches for customer retention in the telecommunications sector using techniques such as SHAP and LIME. These studies reflect a broader shift in churn prediction research from

focusing only on predictive accuracy toward providing information that can help users understand model decisions.

Despite this progress, some important issues remain. Many existing studies evaluate churn models mainly through measures such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC, while comparatively less attention is given to whether the explanations produced by these models are consistent and reliable. A model may perform well on predictive metrics but still produce explanations that are difficult to interpret or unstable across different situations. In addition, several studies apply a single explanation technique to one selected model, which limits the ability to determine whether different models and explanation methods identify similar churn-related characteristics.

The comparison of XAI techniques is therefore relevant when assessing the reliability of churn explanations. Oprea and Bâra compared a unified perturbation-based XAI approach with established techniques such as SHAP and LIME, considering aspects including explanation fidelity, consistency, and computational runtime. Their findings suggest that different explanation approaches can produce similar feature-importance patterns while differing in efficiency and fidelity. This indicates that XAI methods should not automatically be considered interchangeable. Instead, their outputs need to be examined to determine whether the important customer factors remain reasonably consistent across different predictive models and explanation techniques.

Overall, the existing literature demonstrates substantial progress in machine learning-based customer churn prediction and the application of explainability methods. However, the consistency of explanations across multiple predictive models and XAI techniques remains an area that requires further examination. This study addresses this issue by comparing five machine learning models and using SHAP and LIME to investigate the factors influencing churn predictions. The analysis focuses not only on predictive performance but also on whether the important customer characteristics identified through different models and explanation techniques show meaningful consistency.

3. RESEARCH GAP

Existing research demonstrates that machine learning can be effectively applied to customer churn prediction, with ensemble and boosting algorithms frequently being investigated for improving predictive performance. However, a considerable portion of the existing work concentrates on achieving better prediction results through model selection, feature engineering, ensemble techniques, and hyperparameter optimization. Comparatively fewer studies examine whether the resulting predictions can be accompanied by explanations that remain understandable and reliable across different models.

A further gap concerns the evaluation of Explainable Artificial Intelligence (XAI) methods in churn prediction. Techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are increasingly used to interpret machine learning predictions. However, these techniques are often applied to an

individual model or investigated separately. As a result, there is limited evidence regarding whether different XAI approaches identify similar customer characteristics when they are applied to multiple predictive models.

Generating feature-importance values alone is also not sufficient to establish that an explanation is reliable. For a more complete assessment of an explainable churn prediction system, it is necessary to examine whether the important customer characteristics identified by different models and explanation techniques show reasonable consistency.

Variations in feature importance can occur because different algorithms learn different relationships from the same customer data. Examining these variations is therefore important for determining whether the identified factors represent recurring patterns associated with churn or are strongly dependent on the selected prediction model or explanation method.

Based on these observations, the main gap addressed by this study is the limited systematic examination of the relationship between predictive performance and explanation reliability across multiple machine learning models and XAI techniques. To address this gap, the study evaluates multiple machine learning models using common performance measures and applies SHAP and LIME to interpret their predictions. The resulting explanations are compared across models and XAI techniques to identify influential customer characteristics and examine their consistency. The study also considers whether the model demonstrating the strongest predictive performance can provide explanations that are useful and understandable in the context of customer retention decisions.

4. PROBLEM STATEMENT

Customer churn is an important business concern because organizations need to identify customers who are likely to discontinue their services and take appropriate retention measures. Machine learning models can learn patterns from historical customer data and provide predictions about churn risk. However, a prediction indicating that a customer is likely to leave does not explain the reasons behind that prediction. This lack of transparency can make it difficult for business decision-makers to evaluate the prediction and select suitable customer retention strategies.

Many existing churn prediction studies focus on improving predictive performance through machine learning algorithms, ensemble methods, feature engineering, and hyperparameter optimization. More recent studies have introduced Explainable Artificial Intelligence (XAI) techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to improve the interpretability of model predictions. However, applying an XAI technique does not by itself establish that an explanation is reliable, consistent, or equally useful across different predictive models. Different models may identify different customer characteristics as important, and different XAI techniques may also provide different explanations for the same prediction.

This creates a challenge in developing trustworthy customer churn prediction systems. A model may achieve strong

predictive results while producing explanations that are difficult to interpret or inconsistent across models and explanation techniques. Therefore, evaluating a churn prediction system only through measures such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC may not provide a complete view of its suitability for practical customer retention applications.

The problem addressed in this study is to determine whether machine learning models can provide both effective churn predictions and reliable, understandable explanations. The study evaluates multiple machine learning models, identifies the customer characteristics influencing their predictions, compares SHAP- and LIME-based explanations, and examines the consistency of explanations across different models. It also investigates whether the model demonstrating the strongest predictive performance can provide explanations that are useful for supporting customer retention decisions.

5. OBJECTIVE

5.1 Model Development

To develop multiple machine learning models for predicting customer churn. The selected models are trained on the prepared churn dataset to identify patterns associated with customer departure and provide a basis for comparing different prediction approaches.

5.2 Performance Comparison

To compare the predictive performance of different machine learning models using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC. These measures are used to examine model performance from different perspectives, particularly because the dataset contains more non-churn than churn cases.

5.3 Churn-Related Factors

To identify the key customer factors that influence churn predictions. The study examines characteristics related to customer tenure, charges, contracts, internet services, and other service-related attributes through the explainability analysis.

5.4 XAI-Based Explanations

To apply SHAP and LIME to generate understandable explanations for customer churn predictions. SHAP is used to examine feature contributions, while LIME is applied to explain an individual customer prediction.

5.5 Explanation Comparison

To compare the explanations generated by SHAP and LIME across different machine learning models. The comparison

focuses on the customer characteristics identified by the two techniques and the differences in their relative importance.

5.6 Explanation Consistency

To evaluate the consistency and reliability of the generated explanations across models and XAI techniques. The study examines whether important customer characteristics are repeatedly identified and how their relative contributions vary.

5.7 Performance and Explainability

To examine the relationship between predictive performance and explainability. The study also investigates whether the model with the strongest predictive performance can provide useful explanations for its churn predictions.

5.8 Practical Usefulness

To assess the potential usefulness of explainable churn predictions for supporting customer retention and decision-making. The study considers how information about the factors behind churn predictions can provide additional context for customer-related decisions.

6. RESEARCH METHODOLOGY

This study adopts a quantitative experimental research methodology to investigate whether machine learning models can effectively predict customer churn while providing understandable explanations for their predictions. The methodology consists of several stages, including dataset selection, data preprocessing, feature transformation, data splitting, machine learning model development, hyperparameter tuning, model evaluation, and explainability analysis using SHAP and LIME. These stages were applied systematically to ensure a consistent comparison of the selected machine learning models and their explanations.

6.1 Model Performance Results

Dataset selection: The IBM Telco Customer Churn dataset was selected for this study. The dataset contains customer information covering demographic characteristics, account details, subscribed services, contract information, billing details, and customer churn status. The original dataset consists of 7,043 customer records and 21 attributes, with Churn serving as the target variable indicating whether a customer discontinued the service. The dataset includes attributes such as gender, senior citizen status, partner and dependents, customer tenure, phone and internet services, online security, technical support, contract type, paperless billing, payment method, monthly charges, and total charges. The customerID attribute was treated as a unique identifier and removed from the analysis because it does not provide meaningful information for predicting customer churn.

6.2 Performance Comparison

Data Preprocessing: The dataset was examined for missing values, duplicate records, and inappropriate data types before model development. No duplicate records were identified during this process. The TotalCharges attribute was initially stored as an object data type and was therefore converted into numerical format using numeric conversion. During this conversion, 11 invalid or blank values were identified and treated as missing values. Since these records represented only a small proportion of the overall dataset, they were removed from the analysis. After completing the preprocessing stage, the final dataset consisted of 7,032 customer records and 20 attributes. The target variable, Churn, was converted into binary numerical values, where No was encoded as 0 and Yes was encoded as 1, representing non-churn and churn customers, respectively. The input features were then separated from the target variable before model development.

6.3 SHAP Analysis

Feature Transformation and Data Splitting: The dataset contained both numerical and categorical features, requiring appropriate transformation before applying the machine learning algorithms. Categorical variables were transformed using One-Hot Encoding, while numerical features were standardized using StandardScaler. Feature transformation was implemented through a preprocessing pipeline to ensure that the same preprocessing procedures were consistently applied during model training and testing. The processed dataset was divided into training and testing subsets using an 80:20 train-test split. Stratified sampling based on the Churn variable was applied to preserve the distribution of churn and non-churn customers in both subsets. A random state of 42 was used to ensure reproducibility of the experimental results.

6.4 LIME Analysis

Machine Learning Models: Five machine learning classification models were selected to compare different predictive approaches: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, and LightGBM Classifier. Logistic Regression was selected as a linear baseline model, while the Decision Tree was included because of its rule-based structure and relatively simple interpretability. Random Forest, XGBoost, and LightGBM were selected as ensemble-based approaches capable of identifying more complex relationships within customer data. Initially, baseline versions of the selected models were trained to establish their initial predictive performance. Hyperparameter tuning was subsequently performed to identify improved model configurations.

6.5 Explanation Consistency

Hyperparameter Tuning: Hyperparameter optimization was performed using RandomizedSearchCV with 5-fold Stratified Cross-Validation. Stratification was used to maintain the distribution of churn and non-churn customers across the validation folds. ROC-AUC was selected as the optimization metric for the hyperparameter search. The parameters considered during optimization included regularization strength for Logistic Regression; maximum tree depth and minimum sample requirements for Decision Tree; the number

of estimators and tree depth for Random Forest; and learning rate, tree depth, subsampling, and feature sampling parameters for XGBoost. For LightGBM, the search included parameters such as the number of estimators, learning rate, number of leaves, and tree depth. The best-performing configuration for each model was selected and subsequently evaluated using the unseen test dataset.

6.6 Model Evaluation

Model Evaluation: The tuned machine learning models were evaluated using multiple performance metrics because the churn dataset contains a larger proportion of non-churn customers than churn customers. Therefore, relying solely on Accuracy may provide an incomplete assessment of model performance. The evaluation included Accuracy, Precision, Recall, F1-Score, Receiver Operating Characteristic–Area Under the Curve (ROC-AUC), and Precision-Recall Area Under the Curve (PR-AUC). Using multiple metrics enabled a comprehensive comparison of the predictive performance of the five machine learning models. The model demonstrating the strongest overall performance was identified by considering the combined results of these evaluation measures, with particular emphasis on ROC-AUC and Recall because of their relevance to customer churn prediction.

6.7 Explainability Analysis

Explainability Analysis: To address the explainability component of the study, two Explainable Artificial Intelligence (XAI) techniques were applied: SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). SHAP was used to provide global explanations by identifying the features that had the greatest influence on churn predictions across the evaluated models. SHAP analysis was performed on the selected Logistic Regression, LightGBM, and XGBoost models. In contrast, LIME was used to provide a local explanation for an individual customer prediction. A test instance with a high predicted churn probability was selected, and LIME was applied to identify the features contributing to the predicted churn outcome. The combination of SHAP and LIME enabled the study to examine explainability from both global and local perspectives and to compare the factors influencing customer churn predictions.

7. RESULT AND DISCUSSION

This section presents and discusses the experimental results obtained using the IBM Telco Customer Churn dataset. Five machine learning models, namely Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM, were trained and evaluated after hyperparameter tuning. The final models were tested on an unseen test dataset using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC. In addition to predictive performance, SHAP and LIME were used to examine the explainability of the selected models from global and local perspectives

7.1 Model Performance Comparison

Model Performance Comparison: Table I presents the performance of the tuned machine learning models on the test dataset.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Logistic Regression	0.8028	0.6381	0.7219	0.6037	0.852	0.6174
LightGBM	0.8028	0.6549	0.5267	0.587	0.841	0.6557
Random Forest	0.7946	0.6431	0.5107	0.5693	0.8347	0.6408
XGBoost	0.7946	0.642	0.516	0.5719	0.8408	0.6591
Decision Tree	0.7903	0.6037	0.615	0.6093	0.8267	0.6074

Table -1: Performance Comparison Of Tuned Machine Learning Models

The results show that all five models achieved reasonable predictive performance, although their strengths varied across the evaluation metrics. Logistic Regression achieved the highest ROC-AUC of 0.8520, indicating the strongest discrimination between churn and non-churn customers among the evaluated models. It also achieved the highest Recall of 0.7219, which is particularly relevant to churn prediction because identifying a larger proportion of potential churners can provide organizations with greater opportunities to initiate retention actions. Logistic Regression and LightGBM achieved the same highest Accuracy of 0.8028.

LightGBM achieved the highest Precision of 0.6549, indicating that it produced the highest proportion of correctly identified churners among the customers predicted as churners. XGBoost achieved the highest PR-AUC of 0.6591, demonstrating competitive performance in evaluating the model's ability to distinguish the positive churn class, particularly in the presence of class imbalance. The Decision Tree achieved the highest F1-Score of 0.6093, reflecting a relatively balanced performance between Precision and Recall for this model.

Considering the overall evaluation, Logistic Regression was selected as the strongest-performing model in this experiment, primarily because it achieved the highest ROC-AUC and Recall while also achieving the joint-highest Accuracy. This selection does not imply that Logistic Regression performed best on every metric; rather, it reflects the importance assigned to discrimination and churn identification in the context of this study. An important observation is that the more complex ensemble models did not substantially outperform Logistic Regression on the selected dataset. This suggests that increased model complexity does not necessarily result in superior predictive performance for this particular churn prediction problem.

7.2 Predictive Performance Discussion

Discussion of Predictive Performance: The experimental results demonstrate that model selection should not be based solely on algorithmic complexity. Although tree-based ensemble models such as Random Forest, XGBoost, and LightGBM are capable of learning complex and nonlinear relationships, Logistic Regression achieved the highest ROC-AUC and Recall in the present experiment. This indicates that the relationships captured by the linear model were sufficient to provide competitive, and in some respects superior, churn prediction performance on the selected dataset.

Recall is particularly important in customer churn prediction because a false negative occurs when a customer who is likely to churn is incorrectly classified as a non-churn customer. Such a customer may not be identified for a potential retention intervention. The Recall value of 0.7219 achieved by Logistic Regression indicates that the model identified a larger proportion of actual churners than the other evaluated models.

Precision provides a different perspective because incorrectly identifying non-churning customers as potential churners may result in unnecessary retention efforts. LightGBM achieved the highest Precision of 0.6549, suggesting that it may be preferable in situations where reducing false positive churn predictions is more important. Similarly, the highest PR-AUC obtained by XGBoost indicates competitive performance when the focus is placed on the positive churn class. These differences demonstrate that there is no single metric that completely represents model usefulness. Therefore, model selection should consider the practical objectives of the intended customer retention application.

7.3 SHAP-Based Explainability

SHAP-Based Global Explainability: To investigate the factors influencing churn predictions, SHAP analysis was applied to the selected Logistic Regression, LightGBM, and XGBoost models. SHAP provides feature-level explanations by quantifying the contribution of individual features to model predictions.



This analysis was used to examine which customer characteristics had a greater influence on the predicted churn outcomes and to compare the influential features across the selected models.

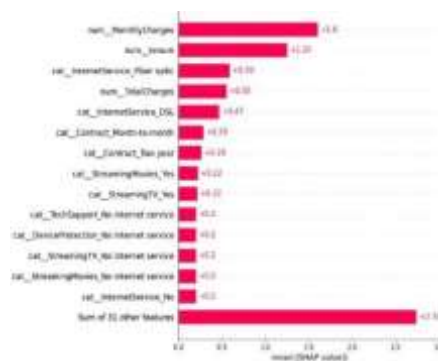


Figure1: SHAP feature importance for Logistic Regression.

The SHAP analysis identified customer characteristics related to customer tenure, monthly charges, total charges, contract type, internet service type, online security, and other service-related features as influential factors in the churn predictions. These findings indicate that customer churn is associated with a combination of customer duration, financial characteristics, contractual conditions, and subscribed services. The presence

of related feature categories across the evaluated models suggests that some customer characteristics contribute consistently to the prediction of churn.

However, the relative importance and magnitude of individual features may differ between models because each algorithm learns patterns from the data in a different manner. Therefore, the SHAP analysis provides not only information about influential customer characteristics but also a basis for examining the consistency of explanations across predictive models. This is particularly relevant to the research objective because consistent identification of important factors can provide greater confidence in the interpretation of model predictions.

7.4 LIME-Based Explainability

LIME-Based Local Explainability: While SHAP was used primarily to examine global feature contributions, LIME was applied to explain an individual customer prediction generated by the selected Logistic Regression model. A test instance with a high predicted churn probability was selected to demonstrate how a machine learning prediction can be interpreted at the individual customer level.

num_MonthlyCharges	0.99
num_tenure	-1.20
cat_InternetService_Fiber optic	1.00
cat_InternetService_DSL	0.00
num_TotalCharges	-0.86
cat_Contract_Two year	0.00
cat_StreamingTV_Yes	1.00
cat_StreamingMovies_No internet service	0.00
cat_OnlineSecurity_No internet service	0.00
cat_InternetService_No	0.00

Figure 2: LIME explanation for the selected high-risk customer.

For the selected test instance, the model predicted a churn probability of approximately 0.85, compared with a no-churn probability of approximately 0.15. This indicates that the selected customer was classified as having a high likelihood of churn. The LIME explanation identified characteristics related to monthly charges, customer tenure, fiber-optic internet service, contract-related characteristics, and other service-related features as contributing to the prediction.

7.5 SHAP and LIME Comparison

Comparison of SHAP and LIME Explanations: The combined use of SHAP and LIME provides complementary perspectives on the explainability of the churn prediction models. SHAP was used to examine the broader influence of customer characteristics across the selected models, whereas LIME was used to explain the prediction of an individual customer. SHAP therefore provides a global perspective on model behavior, while LIME provides a local perspective on a specific prediction.

The explanations generated using these techniques indicate that factors such as customer tenure, charges, contract characteristics, and service-related variables can contribute to

churn predictions. However, the importance and contribution of individual features may vary depending on the predictive model and explanation technique. This variation is important when assessing explanation consistency, as the presence of common feature categories across explanations provides evidence of partial consistency, while differences in feature ranking indicate that the explanations are not completely identical.

Consequently, SHAP and LIME should be viewed as complementary rather than interchangeable techniques. SHAP can help organizations understand the major factors influencing churn predictions at a broader level, while LIME can provide a more specific explanation for an individual customer. Their combined use therefore supports the objective of developing a churn prediction framework that considers both predictive performance and interpretability.

7.6 Answers to the Research Questions

Which model predicts customer churn best?

Based on the evaluation criteria used in this study, Logistic Regression demonstrated the strongest overall predictive performance. It achieved the highest ROC-AUC of 0.8520 and the highest Recall of 0.7219, while also achieving the joint-highest Accuracy of 0.8028. Although other models achieved better results on individual metrics, Logistic Regression provided the strongest overall performance according to the criteria emphasized in this study.

Which customer factors influence churn predictions?

The SHAP analysis identified factors related to customer tenure, monthly charges, total charges, contract type, internet service, and other service-related characteristics as influential in churn predictions. These findings indicate that customer financial conditions, contractual arrangements, duration of service, and subscribed services can contribute to the prediction of customer churn.

How do SHAP and LIME explain predictions?

SHAP provides a global perspective by identifying features that contribute to predictions across the evaluated models, whereas LIME provides a local explanation for an individual customer prediction. In this study, SHAP was used to analyze the selected Logistic Regression, LightGBM, and XGBoost models, while LIME was used to explain an individual high-risk customer prediction generated by Logistic Regression.

Are the explanations consistent across models?

The SHAP analysis indicates partial consistency across the evaluated models, as several broad feature categories, including tenure, charges, contract characteristics, and service-related variables, appeared to influence churn predictions. However, the relative importance and contribution of individual features varied between models. Therefore, the results do not indicate complete agreement across models, but they provide evidence that several important customer characteristics are repeatedly associated with churn prediction.

Does the best-performing model also provide useful explanations?

The results indicate that Logistic Regression, which demonstrated the strongest overall predictive performance according to the selected criteria, can also provide useful explanations through SHAP and LIME. SHAP was used to identify influential factors at a global level, while LIME provided an explanation for an individual high-risk prediction. Therefore, the findings suggest that strong predictive performance and useful explainability can coexist in the selected churn prediction setting.

8. LIMITATIONS

Dataset/generalizability

Although the proposed study provides useful insights into customer churn prediction and explainability, several limitations should be acknowledged. First, the experimental analysis was conducted using the IBM Telco Customer Churn dataset. Therefore, the findings may not be directly generalizable to other industries, such as banking, insurance, e-commerce, or healthcare, where customer behavior and the factors influencing churn may differ. In addition, the dataset represents a specific set of customer characteristics and service attributes and may not capture all real-world factors that influence customer decisions.

Model and scope limitations

Second, the study evaluated five machine learning models, namely Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. Although these models provide a useful comparison between linear, tree-based, and ensemble approaches, other advanced machine learning and deep learning techniques were not included due to the defined scope of the study and available time. Furthermore, model performance may vary depending on the dataset characteristics, feature engineering techniques, class distribution, and selected hyperparameter configurations. Consequently, the performance observed in this study should not be considered representative of all possible machine learning approaches or datasets.

XAI limitations

Another limitation concerns the explainability analysis. SHAP was used to provide global feature-level explanations, while LIME was used to explain an individual prediction. Although these techniques can improve the transparency of machine learning predictions, their explanations may vary depending on the selected model, data distribution, background data, and individual instances. Therefore, the explanations generated in this study should be interpreted as supporting insights into model behavior rather than as evidence of direct causal relationships between customer characteristics and churn.

Business/application limitations

Finally, the study focused primarily on predicting whether a customer is likely to churn and explaining the factors associated with that prediction. It did not examine the timing of churn, the financial value or lifetime value of individual customers, or the effectiveness of specific retention strategies. Consequently, while the developed models can assist organizations in identifying customers who are at higher risk of churn, additional analysis would be required to determine the most appropriate and effective business action for retaining individual customers.

9. FUTURE SCOPE

Future research can extend the proposed study by evaluating the methodology on customer churn datasets from multiple industries. Applying the approach to domains such as banking, e-commerce, insurance, telecommunications, and subscription-based services would help determine whether the observed predictive performance and explainability findings remain consistent across different business environments. Using datasets with different customer characteristics and behavioral patterns could also provide a broader assessment of the generalizability of the proposed approach.

Further research could investigate additional machine learning and deep learning approaches, including neural networks, stacking ensembles, and hybrid models. Advanced feature engineering techniques and automated machine learning methods could also be explored to improve predictive performance while reducing the manual effort required for model selection and hyperparameter optimization. Such approaches may provide opportunities to examine whether more advanced models can achieve improved predictive performance without significantly reducing interpretability.

An important direction for future work is the development of a real-time customer churn prediction system. Instead of relying primarily on historical customer data, future systems could continuously analyze customer interactions, service usage, complaints, payment behavior, and other behavioral information to identify customers who may be at immediate risk of churn. This could enable organizations to respond to changing customer behavior more quickly and support timely retention interventions.

Future studies could also integrate churn prediction with customer segmentation and retention recommendation systems. After identifying a customer as being at high risk of churn, the system could recommend personalized retention actions, such as discounts, service upgrades, targeted customer support, or customized offers. Such an approach would extend the application of the proposed system from simply predicting customer churn to supporting actionable decision-making and personalized customer retention.

Finally, future research can investigate additional Explainable Artificial Intelligence techniques and conduct more comprehensive evaluations of the consistency, stability, and usefulness of explanations across different machine learning models and datasets. Integrating global and local explainability with user-friendly visualization dashboards could further improve the accessibility of model explanations for business managers and decision-makers. This could contribute to the development of more transparent, practical, and trustworthy AI-based customer churn prediction systems.

10. CONCLUSION

This study examined the use of machine learning and Explainable Artificial Intelligence (XAI) for customer churn prediction, with a focus on both predictive performance and the interpretability of model decisions. Five machine learning

models—Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM—were evaluated using multiple performance measures, including Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC. SHAP and LIME were additionally used to understand the factors contributing to churn predictions at global and individual levels.

Among the evaluated models, Logistic Regression provided the strongest overall performance according to the criteria considered in this study. It achieved a ROC-AUC of 0.8520 and a Recall of 0.7219, while its Accuracy of 0.8028 was equal to that of LightGBM. LightGBM produced the highest Precision, XGBoost achieved the highest PR-AUC, and Decision Tree recorded the highest F1-Score. These results demonstrate that model performance can vary depending on the evaluation measure, making it important to consider multiple metrics rather than relying on a single measure.

The explainability analysis provided additional insight into the factors associated with churn predictions. SHAP identified customer-related characteristics such as tenure, monthly charges, total charges, contract type, internet service, and other service-related attributes as important contributors to model predictions. LIME further demonstrated how these factors influenced an individual high-risk customer prediction. Although several important factors appeared across different models and explanation techniques, their relative contribution was not identical, indicating partial rather than complete consistency.

The findings therefore suggest that selecting a churn prediction model should not depend only on predictive performance. Interpretability is also valuable when model predictions are intended to support customer retention decisions. A model may provide strong predictive results while another model or explanation technique may offer different insights into the reasons for churn. Combining predictive evaluation with SHAP and LIME can provide a broader view of both what the model predicts and why it makes a particular prediction.

Overall, this study demonstrates that machine learning-based churn prediction can be complemented by XAI to produce more transparent and informative decision support. The combined use of predictive performance measures and explanation techniques can help organizations better understand customer churn patterns and make more informed retention decisions. Further research using additional datasets, models, and XAI techniques can help assess the generalizability and stability of these findings across different customer and business environments.

11. REFERENCES

[1] K. Peng, Y. Peng, and W. Li, "Research on customer churn prediction and model interpretability analysis," *PLOS ONE*, vol. 18, no. 12, p. e0289724, Dec. 2023, doi: 10.1371/journal.pone.0289724.

[2] U. Tiwari, S. Ashwani, A. J. Tripathy, and K. D. Kumar, "A two-stage ensemble approach for analysis of optimizing customer churn with LIME interpretability," in *2024 IEEE International Conference on Computing, Power and*

Communication Technologies (IC2PCT), 2024, pp. 1540–1545, doi: 10.1109/IC2PCT60090.2024.10486713.

[3] "Explaining customer churn prediction in telecom industry using tabular machine learning models," *Machine Learning with Applications*, vol. 17, p. 100567, Sep. 2024, doi: 10.1016/j.mlwa.2024.100567.

[4] "Hybrid black-box classification for customer churn prediction with segmented interpretability analysis," *Decision Support Systems*, vol. 181, p. 114217, Jun. 2024, doi: 10.1016/j.dss.2024.114217.

[5] A. Manzoor, M. A. Qureshi, E. Kidney, and L. Longo, "A review on machine learning methods for customer churn prediction and recommendations for business practitioners," *Journal of Business Research*, 2024.

[6] "Prediction of customer churn behavior in the telecommunication industry using machine learning models," *Algorithms*, vol. 17, no. 6, p. 231, 2024.

[7] "A comprehensive evaluation of machine learning and deep learning models for churn prediction," *Information*, vol. 16, no. 7, p. 537, 2025.

[8] A. El Attar and R. El-Hajj, "Explainable AI-driven customer churn prediction: A multi-model ensemble approach with SHAP-based feature analysis," *Frontiers in Artificial Intelligence*, vol. 9, 2026, doi: 10.3389/frai.2026.1748799.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[11] Kaggle, "IBM Telco Customer Churn Dataset," dataset used for experimental evaluation.

[12] A. Asif, M. S. Arif, and A. Mukheimer, "A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry," *Results in Engineering*, vol. 26, Art. no. 104629, 2025, doi: 10.1016/j.rineng.2025.104629.

[13] P. Jiang, Z. Liu, M. Z. Abedin, J. Wang, W. Yang, and Q. Dong, "Profit-driven weighted classifier with interpretable ability for customer churn prediction," *Omega*, vol. 125, Art. no. 103034, 2024, doi: 10.1016/j.omega.2024.103034.

[14] Y. Li and K. Yan, "Prediction of bank credit customers churn based on machine learning and interpretability analysis," *Data Science for Financial Econometrics*, 2025, doi: 10.3934/DSFE.2025002.

[15] "Telecom customer churn prediction with explainable machine learning," in *Proc. 2025 2nd International*

Conference on Modeling, Natural Language Processing and Machine Learning, 2025, doi: 10.1145/3757110.3757152.

BIOGRAPHIES

	Mr Omkar Ravindra Salunkhe M.Sc. Information Technology Student, Dept. of IT/CS, D.G.Ruparel College
	Ms. Aarti Sanjay Gawai M. Sc. Information Technology Asst. Prof., Dept. of IT/CS, D. G. Ruparel College