National College of Ireland

A Comparison of Feature Selection Algorithms with Explainable AI

MSc Research Project
Data Analytics

Pattamaporn Sanluang
Student ID: X21122466

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Pattamaporn Sanluang
Student ID:	X21122466
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	16/09/2024
Project Title:	A Comparison of Feature Selection Algorithms with Explainable AI
Word Count:	7,180
Page Count:	22

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:	Pattamaporn Sanluang
Date:	16th September 2024

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):	

A Comparison of Feature Selection Algorithms with Explainable AI

Pattamaporn Sanluang
X21122466

Abstract

In the telecommunication sector, losing a customer has a direct negative impact on revenue and long-term growth prospect, making customer churn prediction a major business priority. To develop successful retention strategies, an understanding of the causes of customer churn is required. However, traditional black-box models often lack the transparency needed to interpret insights into actionable decisions. This research employed Explainable AI (XAI) techniques with classification models, namely Feature Importance, SHAP, and LIME to select the top features influencing churn. This approach enhances the interpretability of churn prediction models without compromising accuracy. Results show that XAI-based feature selection models were easier to interpret (smaller models), are easier to interpret from business point of view, and were able to obtain a higher true positive rate of customer churn. While the baseline Random Forest model achieved the highest F1-score of 79% with all the features, a subset of features selected using SHAP only decreased F1-score to 77%, yet still provided valuable insights into feature contribution. Selected features were able to correctly detect a higher true positive, which in this context is the correct customers that will leave the company. In other words, the confusion matrix revealed that the retrained XGBoost model produced more true positive predictions, which confirms the benefits of targeted feature selection. Demonstrates the potential of XAI to balance model accuracy with transparency offering telecommunications companies a more informed approach to customer retention strategies.

1 Introduction

Explainable Artificial Intelligent (XAI) has become an important research field with the purpose of providing clarity to typically opaque machine learning models (Longo et al.; 2024). Although machine learning (ML) algorithms are proficient at identifying patterns and making predictions, their black-box nature often presents significant challenges. For instance, neural networks and ensemble models, while capable of delivering high accuracy but often lack the transparency needed to identify and analyze the specific features influencing model outcomes. Existing regulatory requirements in strict industries such as insurance, finance and healthcare demand transparency (Hui et al.; 2022), making black-box approaches unsuitable in these contexts. Due to these regulatory demands, data analysts normally focus on traditional methods that prioritize interpretability to ensure that decision-making processes are both transparent and understandable to humans.

There is also the growing complexity and competitiveness of many industries, including telecommunications, that have created a need for solutions that balance accuracy with transparency. Hence, the present research report focuses on the application of Explainable Artificial Intelligence (XAI) within the context of the telecommunications data, particulary in predicting customer churn. XAI techniques enable data analysts to understand feature importance behind black box model and identify key factors driving customer churn, ensuring that these predictions are actionable. With this approach, it can help companies enhance decision-making and concentrate their efforts on areas with the most potential impact, ultimately improving customer satisfaction and reducing churn rates.

Telecommunication companies are facing rapid growth driven by advancement in areas such as network function virtualization, software defined networking and the deployment of 5G technology (Papavassiliou; 2020). These innovations bring competition and continuous pressure to reduce expenses while maintaining high service quality and customer satisfaction. This environment drives them to depend on customers who are using or subscribing to their services, as long-term customers tend to generate more revenue over time compared to new customers (Ahmad A K and Kadan; 2019). Therefore, many telecom companies are increasingly relying on ML algorithms in their daily operations, analysing customer data to better understand their behavior and try to predict who are at risk of churning as retaining current customers requires less cost than acquiring new ones. Given this challenge, it is important to develop models beyond accurately predict churn but also provide clear insights into the underlying factors driving these outcomes to support companies in prioritizing customer retention strategies. (Razak and Wahid; 2021).

1.1 Research Gap

From the literature review it was seen an opportunity to explore Explainable AI (XAI) for feature selection. In particular, little is known about how good feature selection is with such a type of XAI compared to existing methods (e.g. using Random Forest).

In other words, while many studies have explored churn prediction models and techniques, there is still a gap in understanding if feature selection using Explainable AI (XAI) is any better than existing ones, with the added benefit of providing model transparency and trust. This research aims to fill the gap of XAI integration by incorporating interpretable machine learning models and XAI methods. These methods have an ability to explain the rationale behind the decision-making process made by the prediction models (Peng et al.; 2023). Which will help us identify and explain the specific features and customer behaviours that influencing customer churn.

This approach seeks to develop a feature selection framework based on XAI that provides accurate churn predictions while offering explanations on features and insights. As an example of such a framework, the report shows how companies could use it to prioritize their customer retention strategies more effectively.

1.2 Research question

The research gap points to the usage of XAI for feature importance. To make fair comparisons with and without XAI, models will be created in the context of customer churn. In other words,

- To what extent and how do interpretable ML algorithms improve our understanding of customer churn models?

There is a degree of subjective evaluation when using feature selection with XAI. To quantify XAI contribution to the models, a subset of features will be retained and model re-trained. Model predictions will then be used to quantitative evaluate whether those XAI-based models are any better than standard, out-of-the-box feature importance algorithms (such as Random Forest).

1.3 Research objectives

The objective of this research is to examine the potential of interpretable Machine Learning (ML) models focus on binary classification problems for Telco customer churn prediction using the dataset obtained from an open-sourced, Kaggle[1] which derived from public data module of IBM Cognos Analytics[2]. The proposed machine learning models including Random Forest, Extreme Gradient Boosting (XGBoost) and Logistic Regression. To go beyond identifying at-risk customers and address research question, these models will be combined with XAI methods namely Feature Important, SHAP (SHapley Additive exPlanations)[3] and LIME (Local Interpretable Model-agnostic Explanations)[4] to access feature importance and local explanations on customer behaviours. The top five features from these techniques will be identified based on their contribution to the model's prediction and interpretability. These key features will then be used for feature selection, followed by retraining the models to ensures a focused analysis of the most critical factors influencing customer churn.

In order to address the research question, the following three research objectives are developed in the present report.

1. Develop a baseline model – Review and implement interpretable machine learning models for churn prediction. The implementation will serve as a benchmark for evaluating the effectiveness of XAI-integrated models.

2. Feature ranking and selection using XAI – Design and implement an XAI-Integrated framework incorporating XAI techniques such as Feature Importance, SHAP and LIME. This framework will be used to identify and select the most important features influencing customer churn which will be included in the final model.

3. Evaluate the subset of selected features – Train proposed machine learning models then implement feature selection through XAI techniques. Compare the performance of these models using traditional evaluation metrics.

1.4 Outline

Following in this paper, Section 2 reviews and discusses of related work on Telco customer churn prediction models and XAI applications. Section 3 details the research methodology

[1] `https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset.`

[2] `1https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=samples-telco-customer-churn`

[3] `https://shap.readthedocs.io/en/latest/`

[4] `https://www.geeksforgeeks.org/introduction-to-explainable-aixai-using-lime/`

describing process of data understanding, data collection, data preprocessing, modeling and evaluation are presented. Section 4 and Section 5 describe the design specification focusing on feature selection through XAI techniques and implementation of the proposed framework respectively. Section6 presents evaluation of model performance and discussion of results and Section 7 concludes the research and discusses future work.

2 Related Work

The use of machine learning models in telco customer churn prediction is receiving increased attention due to the fast-growing environment of this industry. Studies on telco customer churn prediction, interpretable machine learning, and Explainable AI (XAI) methods were critically evaluated and reviewed in this section.

2.1 Telco Customer Churn Prediction using Machine Learning Models

In marketing practice, customer retention is a critical goal, particularly in the competitive telecommunications industry. Predicting customer churn accurately allows companies to take proactive measures to retain their customers and reduce turnover rates. Machine learning models have become essential tools in this effort, offering sophisticated techniques to analyze customer behavior and predict churn. the use of machine learning techniques to improve churn prediction accuracy has been widely explored. A study done by (Raja and Jeyakumar; 2019) investigates the application of ML to predict customer churn in the telco industry. Utilizing the IBM Watson dataset, they aimed to enhance the accuracy of churn prediction models and identify which algorithm performs best in this context by employing K-Nearest Neighbors (KNN), Random Forest (RF), and XGBoost classifiers. The results indicated that XGBoost outperformed KNN and RF, achieving an accuracy score of 94% and an F1-score of 0.94, particularly highlighting that Fiber Optic customers with higher monthly charges are more likely to churn.

Another telco dataset sourced from public were utilised by many researchers to perform binary classification task to predict telco customer churn using both ensemble and traditional machine learnings model. A study by (Tyagi and Manjunath; 2022) found that among six ML models, XGBoostClassifier yields the highest accuracy along with the maximum true positives (969)and minimum false positives (83) in confusion matrix, proving the accurate ability to predict true churn. Furthermore, the research shows that Customers mainly decide to churn based on financial factors such as monthly and total charges. Suggests that lowering service costs is key but must be balanced to avoid increasing company expenses while attracting more customers. substantial influence on churn.

Several studies have explored alternative techniques beyond traditional methods, (Pejić Bach et al.; 2021) presented a three-stage approach to predict telecom customer churn by incorporating clustering and classification method. In the first stage, telco customer churn dataset was prepared and the next stage, market segmentation was implemented using decision trees and cluster analysis such as k-means clustering. They were able to identify six clusters. In the final stage, the chi-squared automatic interaction detector (CHAID) decision tree algorithm was used to develop classification models to identify customer churn. The study found that the decision tree for Cluster 3 was the

most successful at predicting churners with 81.4% accuracy. Consequently, it is recommended that companies should pay more attention to this group of customers for their retention strategies. A study that combining churn prediction and customer segmentation have also explored by (Wu et al.; 2021), the authors proposed an integrated framework that combines churn prediction and customer segmentation. They utilized factor analysis, customer segmentation, and customer behavior analytics which were incorporated after the prediction for the development of retention strategies. In their study, three datasets were utilised and they found that Dataset 2 achieved the highest accuracy of 93.6% and an F1-score of 77.20% from Random Forest. The results emphasize the importance of using appropriate evaluation metrics and handling class imbalance to improve churn prediction models.

Another study by (Senthan et al.; 2021), the authors focus on creating a churn prediction model for the telecommunication sector in Sri Lanka utilizing various supervised machine learning algorithms, Artificial Neural Networks (ANN) and ensemble techniques have also been considered. The study involves collecting a dataset of 10,000 postpaid customers from a local telecom company with 20 customer attributes. The model performance of the XGBoost was selected due to its high efficiency of 83.13% outperformed all other models. Additionally, another study with a closely related approach by (Razak and Wahid; 2021) which evaluates multiple models, including linear regression, random forest, support vector machine (SVM), K-nearest neighbor (KNN), and decision tree on open-sourced Telco customer churn dataset including 21 variables. SMOTE technique was used to handle class imbalance of churn and not churn customers. The models were then compared based on their accuracy, precision, recall, F1-score, and area under the curve (AUC). Among the tested algorithms, the random forest model surpassing others and achieved an accuracy of 95.5% confirms its suitability for churn prediction in this context.

2.2 Interpretable Machine Learning Model with Explainable AI

Several studies focus on integrating Explainable AI (XAI) techniques to understand how features interact or influence to produce predictive outcomes. This approach mainly aims to make machine learning models more transparent and interpretability of their predictions.

The effectiveness of interpretable machine learning techniques such as LIME and SHAP was incorporated by (Kaushik et al.; 2024). They trained several ML models on METABRIC dataset. Logistic Regression was found to perform best among all models. Later, SHAP and LIME were applied on top of the trained model. From the visualization, they found that features such as Lymph nodes examined positive, Tumor size, Nottingham Prognostic Index (NPI), HER2 status, and Menopausal state were consistently important in the prediction across different models. Suggesting an important of interpretability.

To further utilize the same approach, (Haneesha Samudrala et al.; 2024) classified Parkinson's Disease (PD) diagnosis with feature selection methods using LIME and SHAP. A dataset of voice recordings whom 21 people have PD was trained on several ML models and Random Forest performed best with the highest accuracy and F1-score of 97%. SHAP analysis identified a subset of 7 features that increased the model's accuracy to 98%, indicating a more efficient feature selection process. Similarly, LIME analysis also achieved an accuracy of 98% using 8 features, demonstrating the effectiveness of both feature selection methods in improving model interpretability and performance. Another study explored

an application of XAI in healthcare area by (Sharma and Midhunchakkaravarthy; 2023) employed the XGBoost algorithm on Dementia Prediction dataset with an accuracy of 93.33%, while XGBoost performs well in classification, LIME and SHAP offer different insights into feature importance. LIME identifies MMSE and age as significant factors for predicting dementia,while SHAP highlights the importance of CDR, suggesting better clinical coherence. The comparative analysis shows that SHAP provides more reliable and clinically relevant explanations than LIME, as it correctly prioritizes clinical features like MMSE and CDR over less relevant features such as socioeconomic status (SES) and the number of visits. The study's findings emphasize the importance of using XAI techniques to demystify the decision-making process of AI models, thereby increasing trust among medical practitioners.

The CNN-based Intrusion Detection Systems (IDS) was built by (Oseni et al.; 2023) they then developed SHAP based Explainable Framework, combining high-performance IDS with explainability for detecting and interpreting attacks in the Internet of Vehicle Networks. They validated their model using the ToN_IoT dataset which achieved a high performance with 99% accuracy and a 98% F1 score and and was able to distinguish between normal and various types of attack traffic effectively. The used SHAP and LIME to enhance the transparency and interpretability of the AI model's decisions on AI-powered cyber security system was developed in the studies of (Lundberg et al.; 2022) and (Kaur and Gupta; 2024) for "Anomaly-based In-Vehicle Intrusion Detection System (IV-IDS)" and the Internet of Things (IoT) respectively. After model training on Survival data, (Lundberg et al.; 2022) created a visualization explanation tool called "VisExp" to explain the behavior of the AI-based IV-IDS. And then surveys with experts in the field. The survey results indicated that expert's trust in the AI-based IV-IDS significantly increased when provided with VisExp, compared to the rule-based explanation because of better interpretability with help of SHAP and LIME. On the other hand, (Kaur and Gupta; 2024) utilized XGBoost applied to the CICIoT 2023 dataset to classify IoT network traffic into malicious and non-malicious categories. The model achieved an accuracy of 95.59%. SHAP and LIME were used to explain the model's predictions and gain insights into the most influential features for detecting attacks which help increased to 97.02% after recursive feature elimination.

3 Methodology

In this research, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was utilized to explore the telecommunications sector, with a specific focus on customer churn with the objective to understand customer characteristics and underlying factors that influence churn. Which can be actionable business insights that could inform strategic decision-making and customer retention initiatives. The methodology involved several steps, as illustrated in Figure 1 below.

3.1 Data Understanding

The Telco Customer Churn dataset obtained from an open-sourced, Kaggle[5] were used. This dataset is a cleaned and processed version of the original dataset provided by IBM

[5]`https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset.`

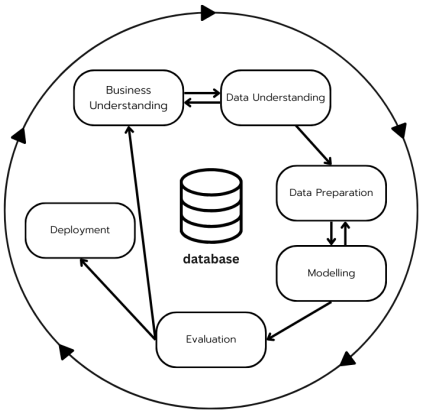


Figure 1: Cross-industry standard process for data mining (CRISP-DM) methodology

Cognos Analytics[6] which is ready for analysis. However, further data preprocessing and transformation were still required to make the dataset suitable for this research.

The dataset contains data from a fictional telecommunications company that offered home phone and internet services to 7,043 customers in California, USA. There are 20 features including detailed information about customer demographics, service usage, payment information and churn status. The dataset also indicates whether customers have left, stayed, or signed up for the company's services.

3.2 Data Preprocessing

The dataset used in this research required several preprocessing steps to ensure the data was ready for analysis. First, the dataset in CSV file was imported from the file path, defined new dataframe and checked for null values, which revealed no missing values initially. However, upon further inspection, it was identified that the 'TotalCharges' column contained 11 null values. These rows were removed from the dataset to maintain the integrity of the analysis.

The 'customerID' column, which served no analytical purpose, was dropped from the dataset, reducing the dataset to 20 columns and 7,043 rows. The next step involved converting suitable features into categorical data types. The columns including 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', and 'Churn' were transformed into categorical features. This conversion is crucial as it allows for more efficient data manipulation and better interpretation during the modeling phase.

A check for duplicate rows revealed 22 duplicates, which were also removed, ensuring the uniqueness and consistency of each record in the dataset. This preprocessing step refined the dataset to 7,010 rows.

[6] 1https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=samples-telco-customer-churn

3.3 Exploratory Data Analysis

Once the dataset was preprocessed, a detailed exploratory data analysis was conducted on the telco customer churn dataset to examine the distribution, patterns, and relationships among all 20 features. This exploratory data analysis process helped identify significant trends, outliers and potential correlations for further analysis and model building.

Bar charts were unitilized to visualize the distribution of categorical columns. These features appeared within expected ranges where the majority of customers subscribe to phone service as it is a dominant product for a telecom company. While these visualizations provided initial insights into the dataset, determining the impact of categorical features on customer churn is still challenging. XAI techniques will be utilized to reveal relationships between categorical features and churn prediction in this research.

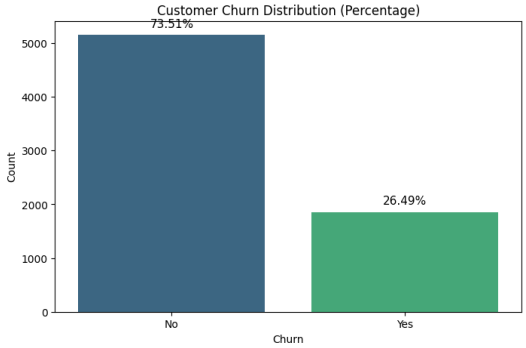


Figure 2: Customer Churn Distribution (original class imbalance)

Figure 2 illustrates the distribution of customer churn within the dataset. We can observe a class imbalance, with 73.51% of customers classified as non-churn and 26.49% as churn. We trained the models on the imbalanced dataset prior to balancing and it was clear that imbalanced data led to bias, particularly with the minority class (churners). For instance, Logistic Regression produced a recall score of 0.91 for non-churners and only 0.55 for churners which indicates that model struggles to identify the minority class accurately even though it has high overall accuracy.

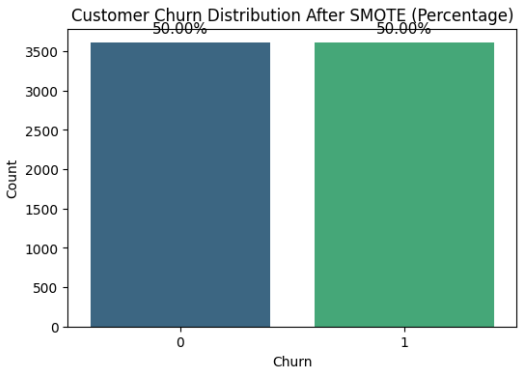


Figure 3: Customer Churn Distribution After SMOTE (class balance)

To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset by generating synthetic samples for the minority class based on the existing ones, thereby improving the balance between the classes (Joloudari et al.;

2023). As a result, both the churn and non-churn classes are now represented equally at 50%, as illustrated in Figure 3. This balanced distribution is expected to enhance the performance of predictive models and ensure more accurate and reliable outcomes.

To assess the impact of balancing the dataset, we retrained the Logistic Regression model. The recall for non-churners improved to 0.74 indicating less bias towards the majority class and the recall for churners increased significantly to 0.80. This improvement demonstrates that balancing the dataset helps the model better identify churners.

Although XAI techniques like SHAP and LIME are model-agnostic and can be applied directly to imbalanced datasets. In this research, we chose to apply them on the balanced dataset as the performance of XAI methods is highly dependent on the quality of the model to produce reliable interpretability and explanations (Liu et al.; 2022). Models that trained on imbalanced data tend to be biased towards the majority class, which can lead to inaccurate or misleading explanations when using XAI and potentially misguiding business decisions.

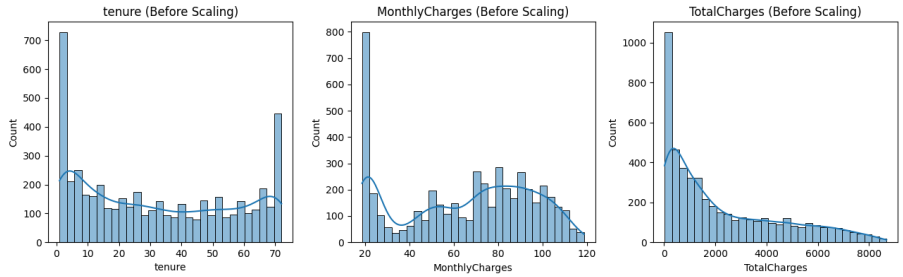


Figure 4: Distribution of numerical features before scaling

As per visualisation of three key numerical features, Figure 4 above presents the distributions of Tenure, MonthlyCharges, and TotalCharges. The tenure representing the customer's duration (in months) with the service, shows a right-skewed distribution indicating a high amount of customers with shorter tenures. Next plot shows the distribution of MonthlyCharges which represent the monthly fee, is comparatively consistent with slightly right-skewed. It is understood that there is a wide range of charges across the customer base. Lastly, TotalCharges also displays a right-skewed pattern indicating a concentration of customers with lower total charges.

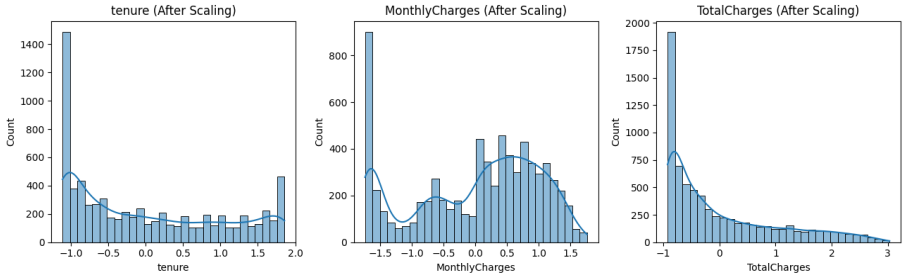


Figure 5: Distribution of numerical features after scaling (shape of distributions preserved after scaling)

As process of skewness handling, Figure 5 illustrates the distributions of the three key numerical features after the application of StandardScaler[7] which is a function in Scikit-Learn, an open-source Python library. This scaling technique was considered for its ability

[7] `https://www.pythonprog.com/sklearn-preprocessing-standardscaler/`

to center the data around zero and standardize the variance to ensure fairness in model performance. Tenure, while still presenting a slight right skew but it is now centered around zero. MonthlyCharges and TotalCharges have found that the distributions are more symmetrical and centered around zero.

3.4 Modeling

In this research, we employed three distinct supervised machine learning algorithms namely Random Forest, Extreme Gradient Boosting (XGBoost) and Logistic Regression to predict customer churn in the telecommunications sector. Prior to model fitting, the dataset was preprocessed to ensure optimal performance, involving the removal of null values and duplicates, transforming suitable features into categorical features. We initially faced challenges with class imbalance and skewed numerical features, which were mitigated by implementing SMOTE techniques to balance the churn status which now equal at 50% and the StandardScaler was used for normalisation. The categorical features were then encoded using the one-hot encoding techniques and later split into training and testing sets with a 70/30 ratio. Feature selection was implemented by incorporating XAI techniques, specifically Feature Important, SHapley Additive exPlanations (SHAP) and Local Interpretable Model agnostic Explanations (LIME). These techniques were used to identify and select the most important features influencing customer churn, and the models will be retrained with these features to assess their impact on predictive performance.

3.5 Evaluation

Evaluating the performance of machine learning models is crucial for understanding their predictive capabilities. Traditional metrics such as Accuracy, Precision, Recall, and F1-Score were employed to assess a model's performance. In the context of customer churn prediction, a high precision score is important to avoid unnecessary retention efforts, while a high recall score is crucial to minimize the loss of valuable customers. F1-Score provides a balance between Precision and Recall, combining them into a single metric. With these four metrics, we ensure to obtain a comprehensive understanding of model performance and make informed decisions.

4 Design Specification

In this section, we outline the proposed approach for enhancing customer churn prediction in the telecommunications sector using a combination of machine learning models and Explainable AI (XAI) techniques. This design specification focuses on the functional description of the models, the rationale behind the selected techniques and the expected outcomes.

Explainable AI

Many algorithms, the prediction ability would be damaged by removing or changing predictors. Thus, we employed three commonly used machine learning models to predict the Telco customer churn and interpreted the results using Feature Importance, SHAP and LIME. Figure 6 below represents the concept of black-box versus XAI-enhanced interpretable ML models in the context of customer churn prediction.

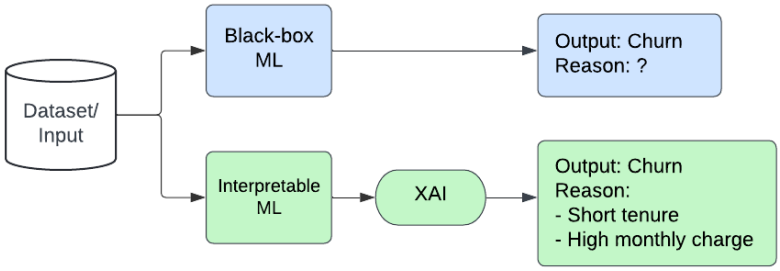


Figure 6: Black-box model versus interpretable and XAI models (Hui et al.; 2022)

To enhance model interpretability, this research used a combination of global and local explainability methods[8]. Feature importance is the most widely used technique for explaining model predictions for a black box model. It calculates the contribution of features or customer behavior towards the churn prediction output by changing the values of each feature and observing the increase in model prediction errors (such as log-loss). Features with high importance score implies that increasing the feature value would control or induce churn. This can clarify why the model makes a particular prediction and make the model more interpretable. SHAP provides both global and local perspective on feature importance based on game-theoretic approach by assigning each feature a value representing its contribution to the model's output. In contrast, LIME, which is also an agnostic model like SHAP, it focuses on explaining individual predictions by building locally approximations of point of interest of the observation that is being explained in the model. This approach offers insights into how customer behaviour combinations influence churn.

XAI-Integrated Feature Selection

In this research, XAI methods were utilized not only to identify the importance of features influencing prediction outcomes but also as a feature selection process. By applying the insights provided by XAI libraries, we were able to pinpoint the most important features contributing to customer churn predictions. This feature selection approach enabled us to reduce the dimensionality of the dataset, thereby enhancing computational efficiency and focusing our analysis on the features with the highest impact on predictive performance.

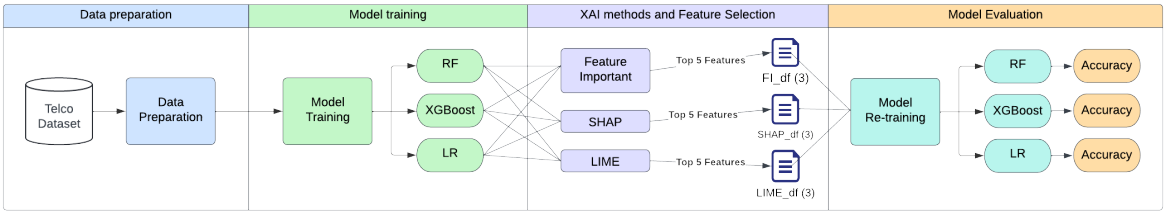


Figure 7: Methodology for Telco customer churn analysis

Figure 7 illustrates the methodology to the research objective, involved several key steps. Initially, three proposed machine learning models, Random Forest, XGBoost,

[8]https://www.markovml.com/blog/lime-vs-shap/

and Logistic Regression were trained on the telco dataset. Followed by XAI-integrated feature selection process, three XAI methods as previously mentioned which are Feature Importance, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations) were applied to assess and rank features based on their important or contribution to churn prediction.

The top five features identified by these XAI methods were then used to retrain with baseline machine learning models. This process aimed to evaluate the impact of these selected features on model accuracy and overall performance while ensuring that the resulting models were more transparent and understandable. And to determine which model delivers the best performance and interpretability when combined with XAI-based feature selection.

While the choice of five features was made to enhance the analysis and focus on the most influential predictors in this research. Other systematic methods like Recursive Feature Elimination (RFE), threshold selection (e.g., SHAP values at 0.5), cross-validation, or factor analysis could have been employed to determine the optimal number of features to be selected for model building.

The results of these experiments were compared against the baseline performances of the models to assess the effectiveness of the XAI-integrated feature selection process. This evaluation provided valuable insights between feature importance, model performance, and interpretability, ultimately contributing to the development of more transparent and effective churn prediction models in the telecommunications industry.

5 Implementation

5.1 Model Training and Baseline Performance

The initial phase involved training three black box machine learning models Random Forest, XGBoost, and Logistic Regression using Python to establish baseline performance metrics. The dataset was cleaned and preprocessed before training. The Random Forest and XGBoost models were configured with their default hyperparameters, while the Logistic Regression model used standard settings. Performance metrics, including accuracy, precision, recall, and F1-score, were calculated using the scikit-learn library which provided the tools for model training and evaluation.

5.2 Feature Selection with XAI Techniques

After establishing baseline performance, XAI techniques were employed to identify the top five features influencing each model's predictions. The process and tools used for each model are detailed below:

Random Forest: Feature Importance determined using the `.feature_importances_` a built-in function the scikit-learn module of Python library. This function measures how useful each feature is in a Random Forest, it calculates how much a feature improves the model's accuracy across all decision trees, with higher scores indicating more influential features. For global feature importance, SHAP was employed with the TreeExplainer from the SHAP library, which provides an additive feature importance score for tree-based models. Figure 8 illustrates the visualisation of 5 features with the highest important score obtained from Feature Important and LIME that applied on Random Forest model.

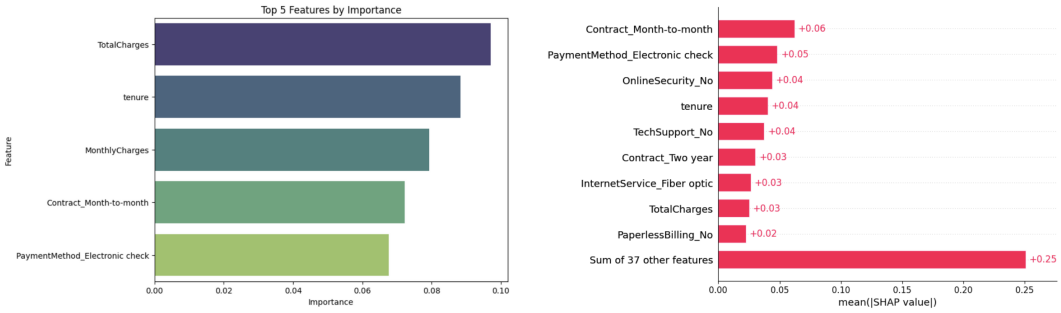


Figure 8: Top 5 Features from Feature Important applied on Random Forest Model

```
Feature: Contract_Month-to-month, Value: 0.0, Importance: -0.15871729995420142
Feature: 0.00 < OnlineSecurity_No, Value: 1.0, Importance: 0.07406466827632015
Feature: 0.00 < PaymentMethod_Electronic check, Value: 1.0, Importance: 0.07371659136043589
Feature: 0.00 < TechSupport_No, Value: 1.0, Importance: 0.0691805379820325
Feature: 0.00 < PaperlessBilling Yes, Value: 1.0, Importance: 0.04915116042295682
```

Figure 9: Top 5 Features from LIME applied on Random Forest Model

LIME was used with the `LimeTabularExplainer` from the LIME library in Python to generate local explanations for this classification task, highlighting feature importance for individual predictions. These different approaches were necessary because Random Forest models involve complex interactions among trees, and each XAI tool provides a unique perspective on feature significance.

While LIME offers visualisation of features important, it is better to investigate the result by printing them out as we can observe feature's important score as shown in Figure 9 above. In this case, features with negative important scores suggest they associate with lower likelihood of churn but still represent significant factors in the prediction. However, LIME ranks features with the highest absolute importance scores based on their contribution to the model's prediction not just their absolute value, regardless of whether the score is positive or negative. Hence, why we can observe Contract_Monthtomonth on LIME's top features rank.

XGBoost: Feature Importance was assessed using `.booster.get_score()` from the xgboost library in Python, which offers feature importance scores tailored for gradient boosting models. Similar to Random Forest, SHAP analysis was performed using the `TreeExplainer` from the SHAP library to provide a comprehensive view of feature contribution to the churn prediction. LIME's `LimeTabularExplainer` was employed to offer local interpretability for individual predictions. The different functions and tools were needed due to XGBoost's gradient boosting approach, which necessitates methods specifically designed for its boosting framework.

Logistic Regression: Feature Importance was evaluated using the `.model.coef_ attribute` from the scikit-learn library, which measures the impact of each feature in a linear model. SHAP analysis for Logistic Regression utilized the `Explainer` from the shap library to provide global feature importance in a linear context. LIME's `LimeTabularExplainer` was used to produce local explanations, consistent with its use in other models. Different tools and methods were employed because Logistic Regression models are linear and require specific approaches for interpreting coefficients and feature contributions.

5.3 Model Retraining with Selected Features

With the top five features identified through each XAI technique, each model was retrained using these features to assess the impact of feature selection. Python was utilized to implement this process, with the models Random Forest, XGBoost, and Logistic Regression being retrained using the features selected by Feature Importance, SHAP, and LIME. The retraining involved using the scikit-learn and xgboost libraries from Python to adjust the models according to the selected features. Performance metrics, including accuracy, precision, recall, and F1-score, of the retrained models were compared to the baseline results. This comparison aimed to determine the effect of XAI-based feature selection on model performance and interpretability, highlighting how feature selection influences predictive capability and model clarity.

6 Evaluation

The results of the experiments were analyzed and presented in two main parts. The first part focuses on the performance of the proposed machine learning models in predicting customer churn to act as a benchmark. The second part assesses the effectiveness of incorporating Explainable AI (XAI) techniques into traditional black box models and understands how these techniques impact model performance and interpretability. The evaluation begins with an analysis of the baseline performance of the Random Forest, eXtreme Gradient Boosting (XGBoost), and Logistic Regression models. This is followed by a detailed examination of each experiment and a comprehensive discussion of the findings.

Table 1: Baseline Models (without XAI)

Predictive Models	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	80	79	80	79
XGBoost	78	77	78	77
Logistic Regression	76	81	76	77

Table 1 above presents the performance comparison of three proposed predictive models as measured by different evaluation metrics as baseline before implementing XAI methods. It can be observed that the Random Forest model achieved the highest accuracy of 80%, outperforming the other models. This can be attributed to the ability to identify complex relationships and patterns within the dataset by combining the insights of multiple decision trees which results in a more reliable and accurate prediction. In terms of the F1-score, which balances precision and recall, the Random Forest model also achieved the highest score of 79% indicating a balanced performance in predicting both the positive and negative classes. The XGBoost model followed closely with an F1-score of 77% along with Logistic Regression model at the same score despite the differences in their algorithmic approaches.

Table 2 summarizes the performance of baseline models follows by retrained models using features selected by different XAI methods. In this research, the weighted average F1-score is used as the evaluation metric. As it offers balanced performance measure that considers both precision and recall, and closely aligns with accuracy, given the equal distribution of classes after SMOTE application. Each experiment was detailed below.

Table 2: F1-score comparison of retrained model across XAI methods (%)

Predictive Models	Baseline	Feature Important	SHAP	LIME
Random Forest	79	75	75	77
XGBoost	77	75	77	75
Logistic Regression	77	75	75	75

6.1 Experiment 1: Random Forest

In this experiment, XAI techniques were applied on top of the Random Forest model to extract the most important features influencing customer churn. These identified features were then used to retrain the model to assess the impact on performance. Figure 10 below displays visualisation of top features obtained from each XAI methods.

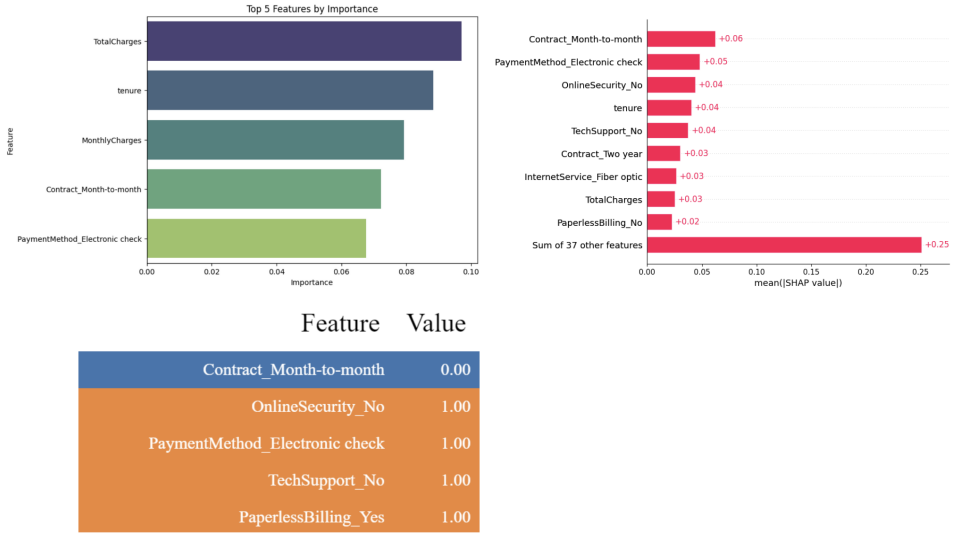


Figure 10: Top 5 Features from Feature Important(Top left), SHAP(Top right) and LIME (Bottom left) applied on Random Forest Model

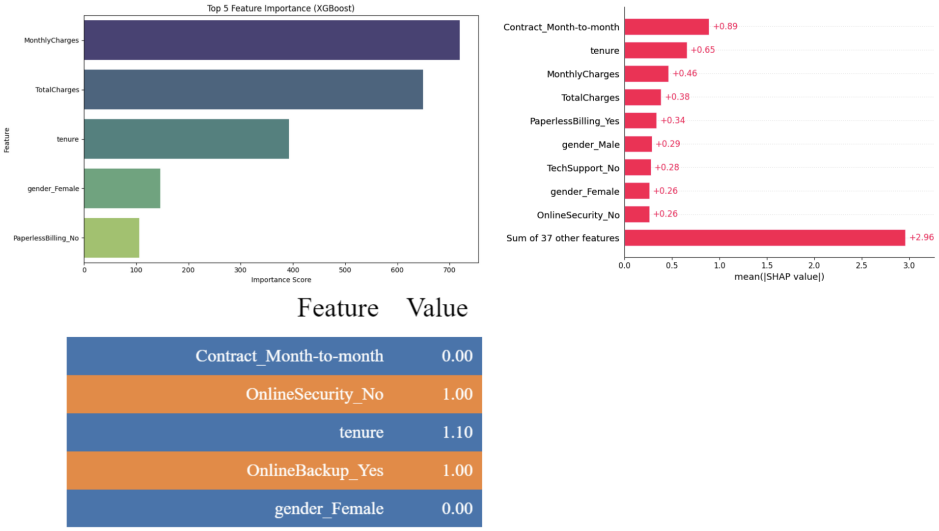
Feature Importance: Initially, the top five features identified through Feature Importance were TotalCharges, Tenure, MonthlyCharges, Contract_Month-to-month, and PaymentMethod_Electronic check. When the model was retrained using these selected features, the F1-score decreased to 75%. This reduction in performance suggests that while Feature Importance can highlight key features but excluding additional relevant features may detract from the model's performance.

SHAP: SHAP identified Contract_Month-to-month, PaymentMethod_Electronic check, OnlineSecurity_No, tenure and TechSupport_No as the top features, overlapping with Feature Importance on several counts. After retraining the Random Forest model with these SHAP-selected features, the F1-score was maintained at 75%, but still lower than the baseline model. This marginal decrease in performance highlights SHAP's strength in enhancing interpretability through global feature importance analysis. And it also shows that SHAP's selected features might not fully capture all predictive elements necessary for maintaining high performance or accuracy. The consistent identification of Contract_Month-to-month, PaymentMethod_Electronic check and tenure reinforces their importance in churn prediction.

LIME: LIME's identified Contract_Month-to-month, OnlineSecurity_No, PaymentMethod_Electronic check, TechSupport_No, and PaperlessBilling_Yes as the most influential features. When the Random Forest model was retrained with these features, the F1-score was at 77%, outperformed the Feature Important and SHAP-enhanced model. The decrease in performance compared to the base model suggests that LIME's local explanations might lead to feature selections that are suboptimal for the global model's performance.

6.2 Experiment 2: XGBoost

This experiment investigated the impact of feature selection using XAI techniques on XGBoost model.



Feature	Value
Contract_Month-to-month	0.00
OnlineSecurity_No	1.00
tenure	1.10
OnlineBackup_Yes	1.00
gender_Female	0.00

Figure 11: Top 5 Features from Feature Important(Top left), SHAP(Top right) and LIME (Bottom left) applied on XGBoost Model

Feature Importance: Applying Feature Importance to the XGBoost model identified MonthlyCharges, TotalCharges, tenure, gender_Female, and PaperlessBilling _No as the top five features. Interesting insight gain from this model is the gender where it usually not consider a critical feature to churn outcome. When the model was retrained using these selected features, the F1-score decreased to 75%. This decline suggests that although these features are correlated with churn, they may not fully capture the complex patterns or interactions driving customer churn behavior.

SHAP: SHAP analysis identified Contract_Month-to-month, tenure, MonthlyCharges, TotalCharges and PaperlessBilling _Yes as the most important features. This closely alignment with the features identified by Feature Importance reinforces their significance in predicting churn. After retraining the XGBoost model with these SHAP-selected features, the F1-score stabilized at 77%, slightly lower than the base model. This stable performance indicates that SHAP effectively captured the primary drivers of churn for the XGBoost model, suggesting that SHAP might be more suitable for identifying globally important features in complex models like XGBoost.

LIME: LIME method led to the selection of Contract_Month-to-month, OnlineSecurity_No, tenure, OnlineBackup_Yes and gender_Female as the most influential features. Retraining the model with these features resulted in a substantial drop in F1-score to 75%. This significant decrease highlights the limitations of LIME in providing globally relevant feature insights. While some overlap with features identified by other methods was observed. The repeat inclusion of features like gender_Female suggests that LIME might be overfitting to local patterns within the data. This behavior is characteristic of LIME and emphasizes the need for caution when using it for global feature importance.

6.3 Experiment 3: Logistic Regression

Lastly, XAI methods were applied on Logistic Regression model to access feature important done by different methods which detailed below.

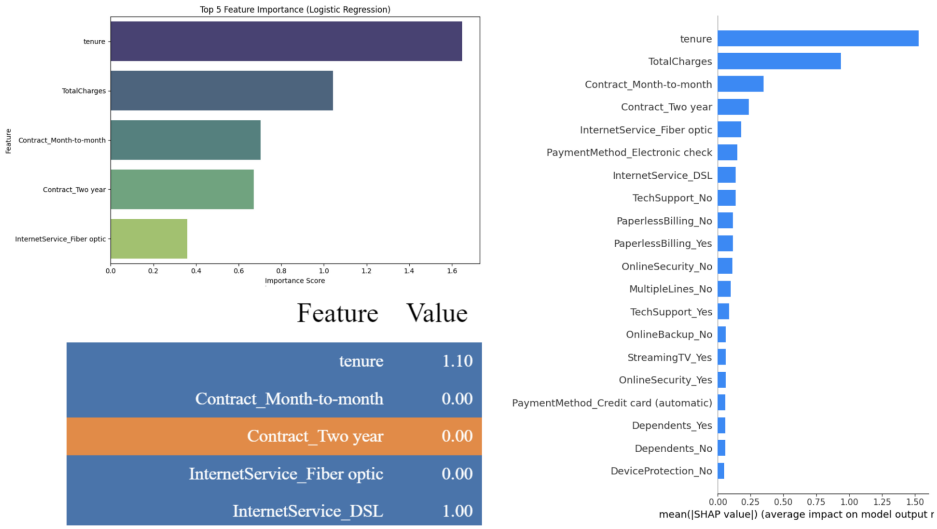


Figure 12: Top 5 Features from Feature Important(Top left), SHAP(Top right) and LIME (Bottom left) applied on Logistic Regression Model

Feature Importance: Retraining the Logistic Regression model with features selected based on Feature Importance resulted in an F1-score of 75%. The selected features included tenure, TotalCharges, Contract_Month-to-month, Contract_Two year and InternetService_Fiber optic. This reduction in F1-score suggests that reducing the feature set might limit the model's predictive capabilities, even for a simpler model like Logistic Regression.

SHAP: SHAP-selected features, which are tenure, TotalCharges, Contract_Month-to-month, Contract_Two year and InternetService_Fiber optic. The F1-score also remained at 75%. This exact alignment in both top features and F1-score with the Feature Importance results indicates that SHAP effectively identified features that did not significantly impact the overall model accuracy. However, the consistent inclusion of features like tenure and TotalCharges in both SHAP and Feature Importance highlights their crucial role in predicting churn. SHAP's inclusion of tenure, TotalCharges and Contract_Month-to-month aligns with its previous identification of important features in more complex models like XGBoost which reflecting SHAP's ability to capture globally significant features across different model types.

LIME: LIME's selected features were tenure, Contract_Month-to-month, Contract_Two year, InternetService_Fiber optic, and InternetService_DSL. The F1-score with these features was also 75%, showing no change. The feature InternetService_DSL, which was not selected by Feature Importance or SHAP, suggests that LIME might highlight features based on specific instances rather than global relevance. This characteristic of LIME highlight its tendency to focus on local patterns which may not always align with features identified as globally significant by other XAI methods.

6.4 Discussion

In this section, we interpret the results obtained from our experiments and summarize how the incorporation of XAI techniques influences model performance and interpretability and their impact within the telecommunications context. The baseline Random Forest model achieved a weighted average F1-score of 79%. However, when re-trained with the top features identified by XAI methods, the F1-score decreased to 77% for Feature Important and 75% on both SHAP and LIME, respectively. This decline suggests a trade-off between interpretability and predictive performance. While Feature Importance and SHAP successfully identified key features such as PaymentMethod_Electronic check and tenure, excluding other relevant features likely impacted the model's overall performance.

The XGBoost model initially achieved F1-score of 77%, which was maintained when retrained with SHAP-selected features, but dropped to 77% and 75% with Feature Importance and LIME, respectively. This outcome highlights SHAP's strength in preserving model performance while enhancing interpretability. Although LIME is useful for individual predictions, it might lead to feature selections that do not generalize well, resulting in a reduced F1-score. And XGBoost's complexity requires a more fine-tuned approach to feature selection as different XAI techniques might emphasize different aspects of the data. The Logistic Regression model began with a baseline F1-score of 77%. When re-trained, the F1-score decreased to 75% across all XAI methods. This consistent drop suggests that Logistic Regression is particularly sensitive to changes in the feature set. The results indicate that while these XAI techniques can enhance interpretability by focusing on key features, they may exclude important features, leading to a slight reduction in model performance. This experiment highlights the balance that must be struck between interpretability and predictive performance, especially in simpler models like Logistic Regression.

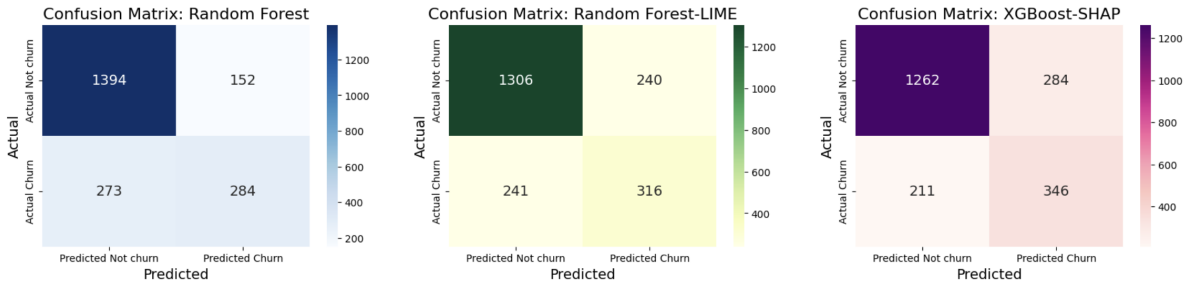


Figure 13: Confusion Matrices of a baseline Random Forest (left), LIME-enhanced Random Forest (middle) and SHAP-enhanced XGBoost (right). Notice XAI-based models were able to correctly classify higher churn numbers than a baseline model.

Figure 13 above presents the confusion matrices for both the baseline Random Forest model, the retrained Random Forest with features selected from LIME's explainer library and the retrained XGBoost model using features selected from the SHAP library.

The Random Forest model achieved the highest F1-score of 79%, serving as the benchmark for comparison. Both the LIME-enhanced Random Forest and the SHAP-enhanced XGBoost models achieved slightly lower F1-scores of 77%. Despite this minor reduction in F1-score, the confusion matrix presents a valuable observation where the retrained XGBoost model demonstrated greater performance in correctly identifying customers that likely to churn, evidenced by a higher True Positive count (346) and a lower False Negative count (211) compared to both the baseline Random Forest and the LIME-enhanced Random Forest models. This suggests that although the XGBoost model has less overall performance and accuracy, it was more effective at correctly identifying customers who were likely to churn and reducing the risk of false negative cases where churn is incorrectly predicted as non-churn.

From the business perspective, the different subsets of highly important features generated by XAI methods reveals frequent appearance of several features across all models, such as tenure and Contract_Month-to-month emphasizes their influence in customer churn. High importance values of these features suggest that the duration of the customer contract and its flexibility are indeed very important factors. It reinforces the reliability of these predictors and is something that business can be aware of when implementing target retention strategies.

Additionally, SHAP and LIME methods show how these features impact individual predictions and help provide local interpretability that can be used to personalized customer engagement. Features like OnlineSecurity_No and TechSupport_No identified by SHAP in the Random Forest model suggests that customers who do not subscribe to additional online services and lack of technical support service are at higher churn risk. Hence, the company can see that this is a potential area for upselling or bundling services and improve their customer services. And the difference in InternetService (Fiber vs. DSL) also offers insights into how types of internet service impact customer churn. We can observe that customers with Fiber are more likely to churn compared to those with DSL, it could possibly be a few reasons. For example, customers who opt for Fiber Optic internet are likely seeking high-speed and high-stability internet service. When these expectations aren't met or the service quality is not better than DSL, especially when Fiber Optic typically has higher cost, they may be more inclined to churn when they do not feel it is worth the value.

In conclusion, the application of XAI in identifying key features relevant to churn. XAI not only provides transparency to the model, but also ensures that the model's outcome aligns with business strategic priorities of the telecommunications sector. With this process, the model becomes more actionable which allows companies to focus their efforts on the factors that matter most. For instance, strategies could be developed to increase customer tenure, introduce more competitive pricing, or simplify payment processes, all of which are crucial to reducing customer churn.

6.5 Limitations

Despite the valuable insights provided by our experiments, there are several limitations to consider that may influence the interpretation and generalizability of the findings.

Dataset Size and Features: The dataset used in this study consisted of 7,043 rows

and 20 features. While this size is adequate for preliminary analysis, a larger dataset could provide a more comprehensive view of customer churn and improve model performance and stability. Additionally, the relatively small number of features might have constrained the models' ability to capture all relevant factors affecting churn. Expanding the feature set could lead to more nuanced insights and potentially enhance prediction accuracy.

Feature Selection Techniques: This study utilized three XAI techniques namely Feature Importance SHAP and LIME to assess feature importance as they were considered the most widely used methods. While SHAP offered stable global insights it may miss local patterns that LIME can capture. The choice of only these three XAI methods limits the scope of feature selection analysis. Exploring additional XAI techniques like Partial Dependence Plots (PDP) or Individual Conditional Expectation (ICE) could provide a more comprehensive understanding of model performance.

Model Complexity: The experiments compared Random Forest, XGBoost, and Logistic Regression models, each with different complexities. Random Forest and XGBoost, being ensemble-based methods, can capture intricate patterns and interactions but are also more challenging to interpret. Logistic Regression, while simpler and more interpretable, may not model complex relationships as effectively. The varying performance across models highlights the trade-offs between complexity and interpretability in feature selection and prediction accuracy.

7 Conclusion and Future Work

Results show that XAI-based feature selection models were easier to interpret (smaller models) from business point of view and were able to obtain a higher true positive rate of customer churn. This research has demonstrated that while XAI techniques like SHAP, LIME, and Feature Importance has proven an ability to provide significant insights into feature important behind black box model that accessible to users, they can also present a trade-off between interpretability and model performance. The Random Forest and XGBoost models, when enhanced with XAI-selected features, showed that while interpretability increased, predictive accuracy sometimes decreased. This highlights the importance of carefully selecting and combining features to trade-off accuracy and interpretability.

These findings have broad implications not only for churn prediction but for any modeling task that requires a balance between accuracy and transparency. This approach enables telco companies to adopt more advanced models that better capture complex, non-linear relationships and leading to more informed business decisions aim at reducing customer churn. The research also suggests that the use of XAI can bridge the gap between model interpretability and business actionability. By identifying the most impactful features related to churn, companies can focus on strategy that are likely to have the greatest effect such as enhancing customer engagement or marketing efforts.

Future work could be explore expanding the feature set or applying more XAI methods to provide a more comprehensive view of feature importance. Additionally, testing these approaches in real-time or with larger datasets could help validate their applicability in diverse scenarios. It would be valuable to assess whether these methods can be applied successfully to a broader range of predictive modeling tasks. If proven effective, this could open up new opportunities for deploying explainable AI in areas where complex models have traditionally been avoided due to concerns about transparency and trust.

References

Ahmad A K, J. A. and Kadan, A. (2019). Customer churn prediction in telecom using machine learning in big data platform, *Journal of Big Data* **6**(28).

Haneesha Samudrala, S. S., Thambi, J., Vadluri, S. R., Mahalingam, A. and Pati, P. B. (2024). Enhancing parkinson's disease diagnosis using speech analysis:a feature subset selection approach with lime and shap, *2024 3rd International Conference for Innovation in Technology (INOCON)*, pp. 1–5.

Hui, A., Ahn, S., Lye, C. and Deng, J. (2022). Ethical challenges of artificial intelligence in health care: A narrative review, *Ethics in Biology, Engineering and Medicine: An International Journal* **12**.

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S. and Hussain, S. (2023). Effective class-imbalance learning based on smote and convolutional neural networks, *Applied Sciences* **13**(6).
URL: *https://www.mdpi.com/2076-3417/13/6/4006*

Kaur, N. and Gupta, L. (2024). Enhancing iot security in 6g environment with transparent ai: Leveraging xgboost, shap and lime, *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*, pp. 180–184.

Kaushik, A., Madhuranath, B., Rao, D., Dey, S. R. and Sampatrao, G. S. (2024). Interpreting breast cancer recurrence prediction models: Exploring feature importance with explainable ai, *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*, pp. 1–6.

Liu, M., Ning, Y., Yuan, H., Ong, M. and Liu, N. (2022). Balanced background and explanation data are needed in explaining deep learning models with shap: An empirical study on clinical decision making.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T. and Stumpf, S. (2024). Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, *Information Fusion* **106**: 102301.
URL: *https://www.sciencedirect.com/science/article/pii/S1566253524000794*

Lundberg, H., Mowla, N. I., Abedin, S. F., Thar, K., Mahmood, A., Gidlund, M. and Raza, S. (2022). Experimental analysis of trustworthy in-vehicle intrusion detection system using explainable artificial intelligence (xai), *IEEE Access* **10**: 102831–102841.

Oseni, A., Moustafa, N., Creech, G., Sohrabi, N., Strelzoff, A., Tari, Z. and Linkov, I. (2023). An explainable deep learning framework for resilient intrusion detection in iot-enabled transportation networks, *IEEE Transactions on Intelligent Transportation Systems* **24**(1): 1000–1014.

Papavassiliou, S. (2020). Software defined networking (sdn) and network function virtualization (nfv), *Future Internet* **12**(1).
URL: *https://www.mdpi.com/1999-5903/12/1/7*

Pejić Bach, M., Pivar, J. and Jaković, B. (2021). Churn management in telecommunications: Hybrid approach using cluster analysis and decision trees, *Journal of Risk and Financial Management* **14**(11).
URL: *https://www.mdpi.com/1911-8074/14/11/544*

Peng, K., Peng, Y. and Li, W. (2023). Research on customer churn prediction and model interpretability analysis, *PLOS ONE* **18**(12): 1–26.
URL: *https://doi.org/10.1371/journal.pone.0289724*

Raja, B. and Jeyakumar, P. (2019). An effective classifier for predicting churn in telecommunication, *Journal of Advanced Research in Dynamical and Control Systems* **11**: 221–229.

Razak, N. I. A. and Wahid, M. H. (2021). Telecommunication customers churn prediction using machine learning, *2021 IEEE 15th Malaysia International Conference on Communication (MICC)*, pp. 81–85.

Senthan, P., Rathnayaka, R., Kuhaneswaran, B. and Kumara, B. (2021). Development of churn prediction model using xgboost - telecommunication industry in sri lanka, *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–7.

Sharma, V. and Midhunchakkaravarthy, D. (2023). Xgboost classification of xai based lime and shap for detecting dementia in young adults, *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–6.

Tyagi, R. and Manjunath, S. (2022). *Customer Churn Analysis Using Machine Learning*, IJCACI, pp. 495–507.

Wu, S., Yau, W.-C., Ong, T.-S. and Chong, S.-C. (2021). Integrated churn prediction and customer segmentation framework for telco business, *IEEE Access* **9**: 62118–62136.