Telecom Customer Churn Prediction by Using ML

Li-Ru Hong
University of Colorado
Boulder, USA
li-ru.hong@colorado.edu

Abstract

The report analyzes data from 7,043 customers of a telecommunications company in California for Q2 2022. Each record corresponds to one customer and includes demographic and location details, tenure, subscribed services, and the status for the quarter.

Initial exploratory data analysis revealed variations in churn rates depending on the promotional offer. For instance, customers using offer E have a churn rate 7.86 times higher than those using offer A. Significant churn rate differences were also observed based on the use of internet services, online security, premium tech support, and contract type. Further analysis of feature importances highlighted key numerical features such as the number of dependents, referrals, tenure in months, and monthly charges. This information is critical for recalibrating marketing and sales strategies.

Various machine learning methods, including logistic regression, k-nearest neighbors, random forest, and naïve Bayes, were employed to develop predictive models. Among these, gradient boosting methods demonstrated the most superior performance.

After balancing the dataset, feature selection, Bayesian Optimization, and cross-validation, the final LightGBM model achieved an AUC score of 0.919 on unseen test data. The combination of feature importances and data analysis provided applications for adjusting business marketing strategies.

Introduction

The telecom industry faces significant challenges in retaining customers due to the high level of competition. Customer churn, the process where customers discontinue their subscription to a telecom service, is a critical issue that impacts revenue and growth. The objective of this project is to develop an effective predictive model that can forecast customer churn based on various features such as demographic information, service usages, and customer behavior. By identifying potential churners, the telecom company can take proactive measures to retain customers, thereby reducing churn rates and increasing customer loyalty.

Related Work

Previous research on telecom customer churn has utilized various features such as customer demographics, service usage patterns, and customer interactions, applying machine learning techniques like logistic regression, decision trees, and k-nearest neighbors to build predictive models.

However, despite the availability of large datasets, telecom companies often face issues with incomplete or noisy data, which limits model performance. There is still room for further research in addressing data imbalance, selecting the most relevant features, and interpreting model results. Additionally, since customer behavior changes with market and technological shifts, it is essential for models to quickly adapt to these changes.

Proposed Work

The proposed project work commenced with data analysis, followed by suitable data processing and feature engineering. This phase includes addressing imbalanced data and selecting pertinent features. Initially, eight different models will be comparatively evaluated to select the most appropriate one. Following the selection, Bayesian Optimization will be employed to fine-tune the chosen model.

Subsequent phases will delve into the analysis of feature importances. By exploring relevant features, we aim to uncover the underlying reasons for customer churn, providing critical insights that will enable the company to enhance its marketing strategies.

Exploratory Data Analysis

Among the 7,043 customers, the proportion of churned customers is 26.5%, the joined is 6.4%, and the stayed is 67.0%.

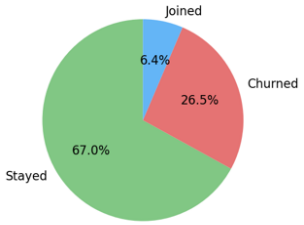

Figure 1: **Customer Status Proportion**

Categorical Features

Figure 2 uses bar charts to display the status distribution across different categories by various features. The comparison of quantities and the distribution of statuses among different categories, as well as the usage of various services, can be clearly observed. The differences between some feature categories are quite significant, such as the number of customers using phone service being 9.33 times that of those not using it; customers using internet service are 3.62 times more than those not using it; and customers subscribing to unlimited data are 6.15 times more than those not subscribing. In terms of Payment Method, the most commonly used method is bank withdrawal, followed by credit card, with only a very small number of customers using mailed checks.

Figure 3 shows the proportion of churned customers across different categories for each feature. In this analysis, it can be observed that for certain features, the proportion of customers labeled as churned varies significantly among different categories. This indicates that these features are among the key factors influencing whether a customer churns. For example, the type of offer a customer uses shows distinct differences in churn rates: the churn rate for customers using Offer A is 6.73%, for Offer B it's 12.26%, for Offer C it's 22.89%, for Offer D it's 2.74%, and over half of the customers using Offer E churn, at 52.92%. This is 7.86 times higher than the churn rate for Offer A! For features with such significant distinctions, the differences between the

offers can be examined, allowing us to identify customer preferences and key product characteristics that matter to them. This insight can then be used to improve and re-strategize in order to retain these churned customer groups.

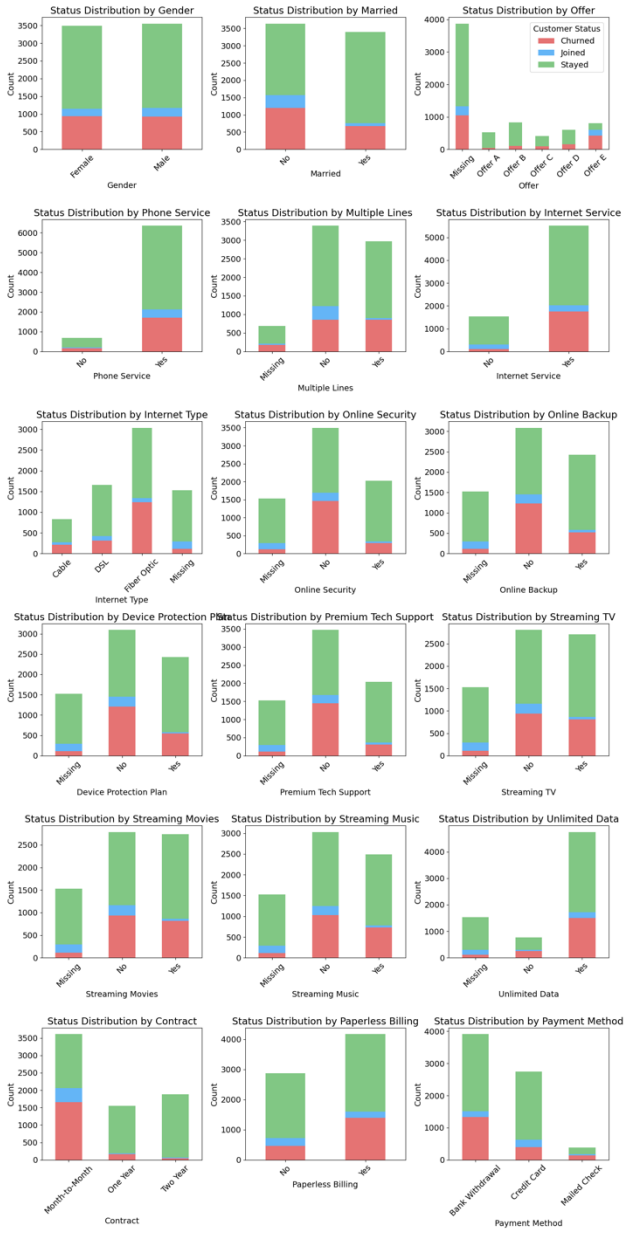

Figure 2: **Status Distribution by features**

Offer	Offer A	Offer E	Offer E / Offer A
Churned	6.73%	52.92%	7.86
Internet Service	No	Yes	Yes / No
Churned	7.40%	31.83%	4.30
Online Security	Yes	No	No / Yes
Churned	14.61%	41.77%	2.86
Tech Support	Yes	No	No / Yes
Churned	15.17%	41.64%	2.75
Contract	Two Year	Monthly	Monthly / Two Year
Churned	2.55%	45.84%	17.98

Table 1: **Summary of features with significant differences in churned customer proportions**

In addition to the type of offer, it was also found that in certain specific features, the proportion of churned customers differs significantly. For example, among those using internet service, the churn rate is 31.83%, whereas for those not using it, the churn rate is 7.40%, a difference of 4.30 times.

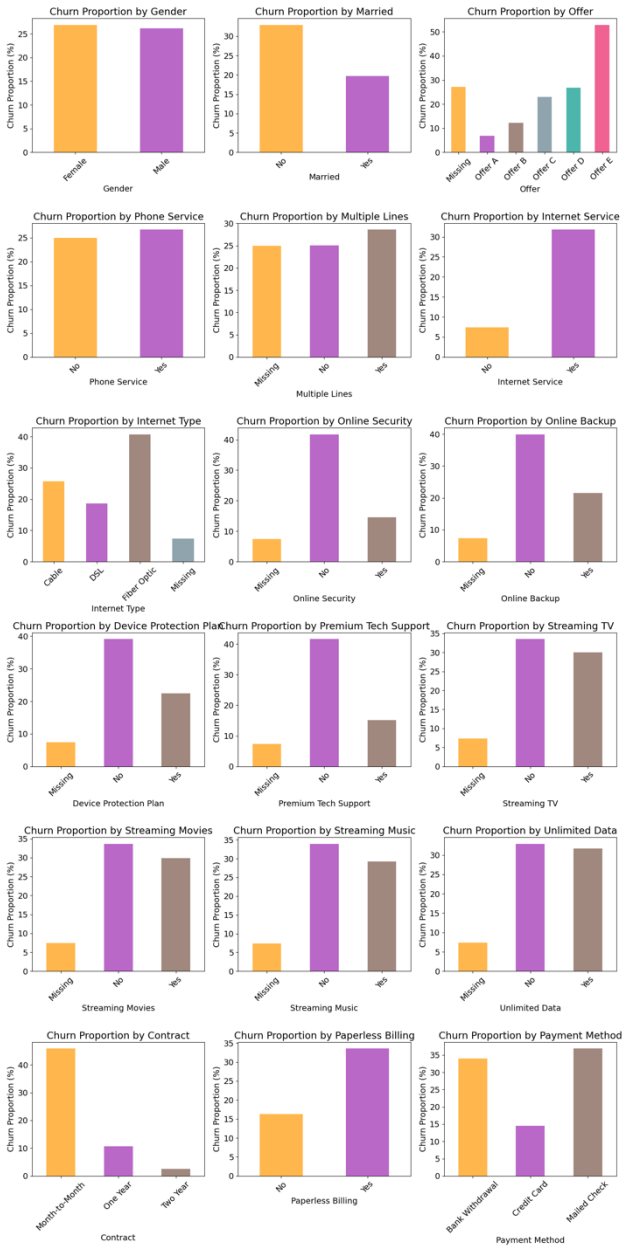

Figure 3: **Churn Proportion by features**

Numerical Features

Figures 4 to 8 display the distributions of customer status and churn proportion for numerical features: Age, number of dependents, number of referrals, tenure in months, and Zip code.

The analysis clearly indicates that the number of users is higher among younger and middle-aged groups. However, the proportion of churn is significantly higher among users aged 65 and above.

Users without any dependents constitute the largest group, and also show a higher churn rate. Notably, as the number of dependents increases, so does the churn rate, highlighting this as a key factor in customer attrition. The figure for the number of

referrals reveal a clear trend as well, where customers with fewer referrals are more likely to churn.

The figure for churn proportion by tenure in months clearly demonstrate that customers with longer tenure tend to have higher loyalty, which aligns with expectations.

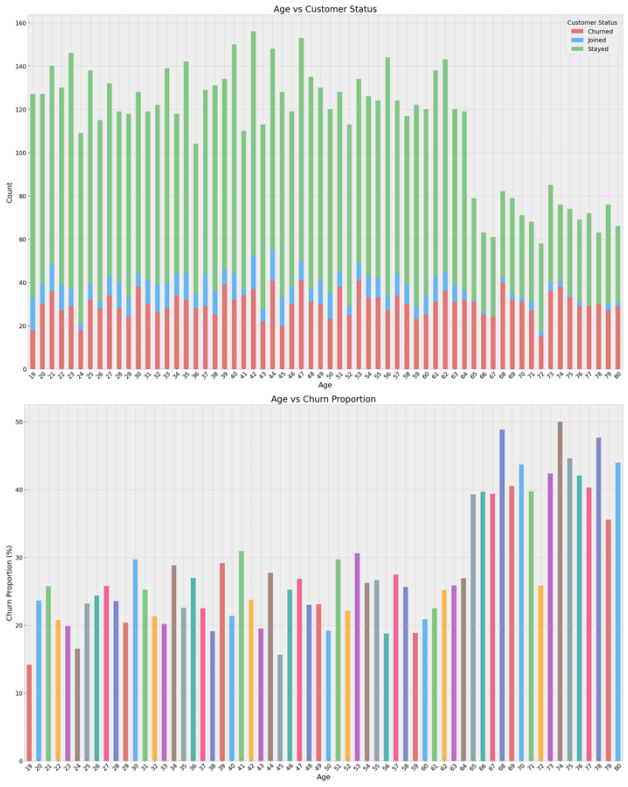

Figure 4: **Status Distribution and Churn Proportion by Age**

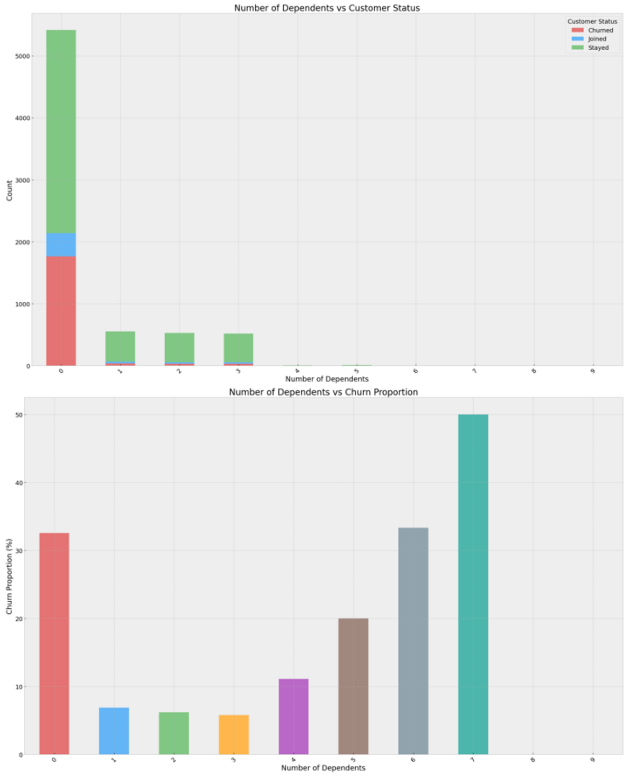

Figure 5: **Churn Status and Proportion by number of Dependents**

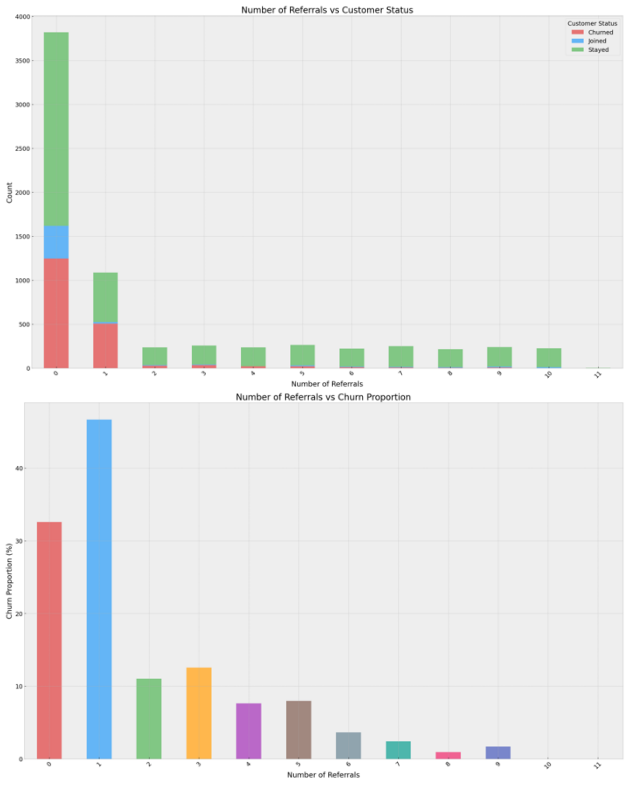

Figure 6: **Churn Status and Proportion by number of Referrals**

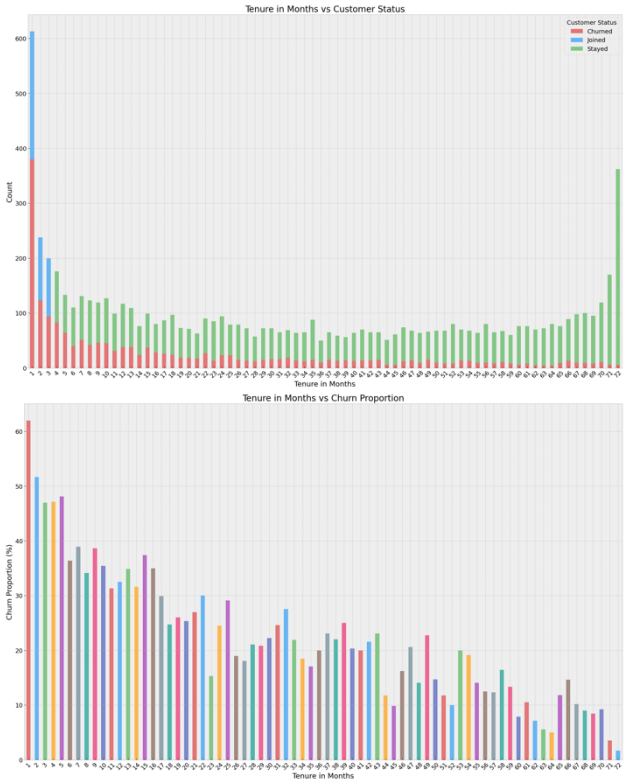

Figure 7: **Churn Status and Proportion by Tenure in Months**

The figures for Zip code reveal noticeable phenomenon in certain areas with high population densities, where the churn proportion is correspondingly higher. This observation suggests a need for a more in-depth investigation into these areas and the development of tailored offers to address the higher churn rates.

Figure 8: **Churn Status and Proportion by Zip Code**

Basic Models and Evaluation

The predictive performance of the models is evaluated using a variety of metrics, including recall, F1, AUC score, and confusion matrix. Recall measures the model's ability to identify churned customers, while the F1 score assesses the overall classification performance by balancing precision and recall. AUC score provides a quantitative evaluation of the model's ability to distinguish between positive and negative classes. The confusion matrix offers a detailed view of the model's predictions across different classes, providing insights into its practical performance.

Eight models, as shown in Table 2, including logistic regression, KNN, Decision Tree, Random Forest, XGBoost, LightGBM, and others, were experimented with. After evaluating metrics, gradient boosting demonstrated the most outstanding overall performance.

Models	Precision	Recall	F1	AUC
Logistic Regression	0.7335	0.6551	0.6921	0.9027
KNN	0.6393	0.5829	0.6098	0.8308
Decision Tree	0.6189	0.6123	0.6156	0.7380
Random Forest	0.7544	0.5668	0.6473	0.8959
Extra Trees	0.7481	0.5401	0.6273	0.8885
XGBoost	0.7614	0.6230	0.6853	0.9147
LightGBM	0.7642	0.6497	0.7023	0.9184
Naïve Bayes	0.5458	0.8128	0.6531	0.8555

Table 2: **Classification Report for unseen test dataset**

However, a detailed examination of the confusion matrix, as shown in Figure 9, reveals that the model excessively predicts the target as class 0, which means it fails to identify potential churn customers and too frequently predicts customers as non-churn. This is not the desired outcome. Therefore, after selecting two gradient methods: XGBoost and LightGBM, the subsequent phases of feature engineering and Bayesian Optimization will address this issue.

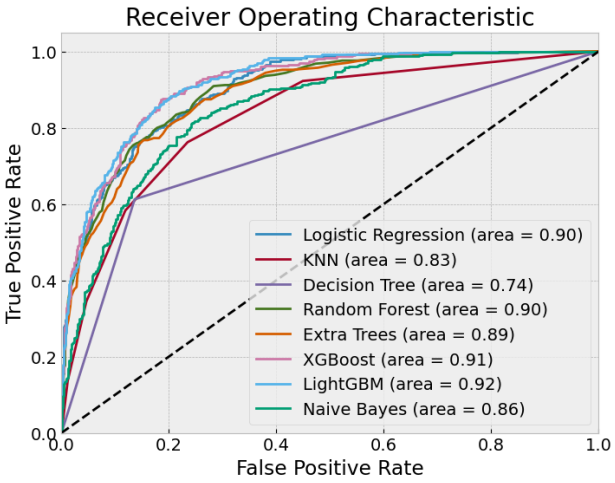

Figure 10: **ROC Curve**

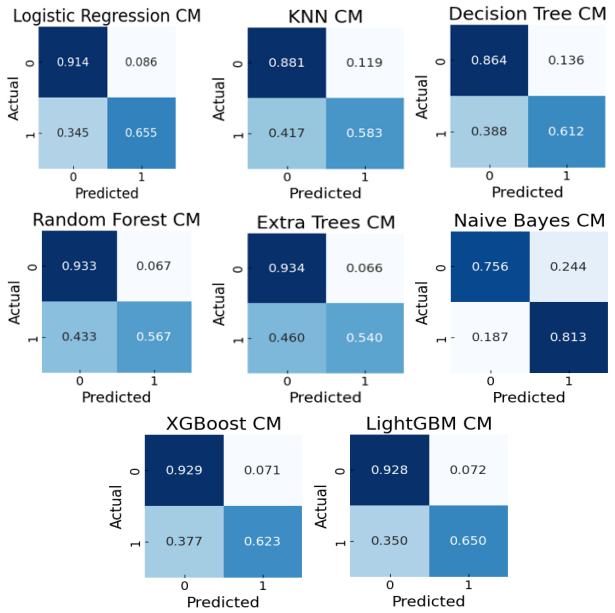

Figure 9: **Confusion Matrix**

Feature Engineering

In the original dataset, we had 7,043 users, each with 38 features. After performing data cleaning, preprocessing, and calculating the correlation matrix for all features, we proceeded with feature selection, ultimately using only 21 features to build the model.

Although this approach reduced the number of features by nearly half, the model's performance only slightly decreased. However, after fine-tuning, the model's performance improved, surpassing that of the original model which used 38 features.

This method of first removing correlated features and then fine-tuning helps prevent overfitting and enhances the models robustness.

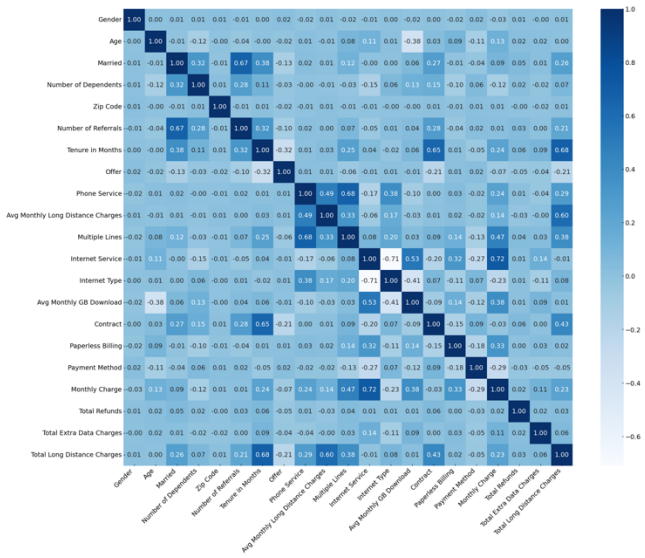

Figure 11: **Correlation Matrix after feature selection**

Bayesian Optimization

In the Bayesian optimization process, the goal was to optimize the hyperparameters of the classifiers to enhance model performance. The evaluation metric used was the mean ROC-AUC score from 3-fold cross-validation.

The hyperparameters adjusted included the number of trees, depth, minimum loss reduction required at a node, the proportion of features used per tree, regularization terms, and learning rate, etc. Additionally, the proportion of imbalanced classes was balanced during the training process.

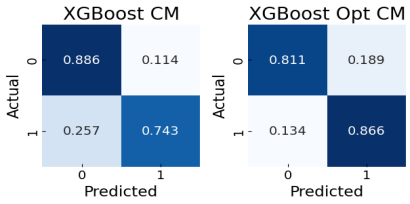

Figure 12: **Confusion Matrix for XGBoost**

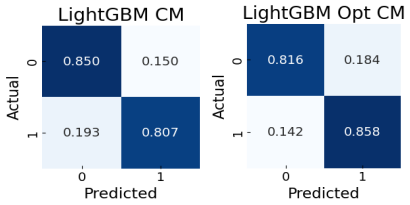

Figure 13: **Confusion Matrix for LightGBM**

Figure 12 and 13 show the confusion matrices before and after optimization. Table 3 presents the classification report for unseen test data, and Figure 14 displays the ROC curves for XGBoost before and after fine-tuning. The red curve represents the optimized model, which exhibits slight performance improvements.

Models	Precision	Recall	F1	AUC
XGBoost fs Reduced	0.7020	0.7433	0.7221	0.9080
XGBoost Optimized	0.6231	0.8663	0.7248	0.9184
LightGBM fs Reduced	0.6608	0.8075	0.7268	0.9142
LightGBM Optimized	0.6282	0.8583	0.7254	0.9190

Table 3: **Classification Report for unseen test dataset**

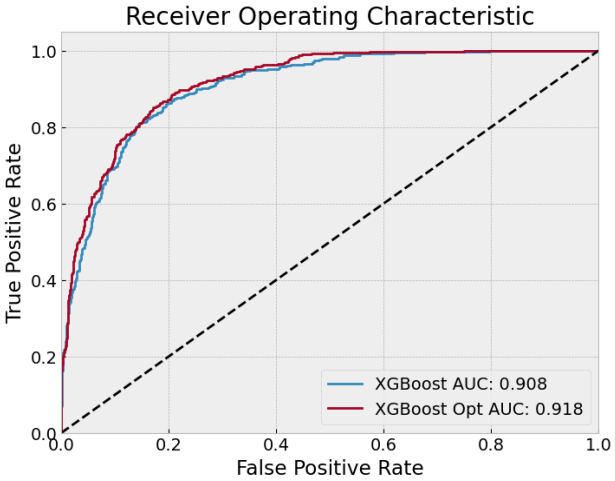

Figure 14: **ROC Curves before and after Optimization**

Feature Importances

After calculating feature importances, we identified the key features influencing the prediction outcomes. Among them, the type of contract had the greatest impact, followed by whether the customer uses internet service, the type of internet service, the number of dependents, and the number of referrals, among others. All of which are significant factors.

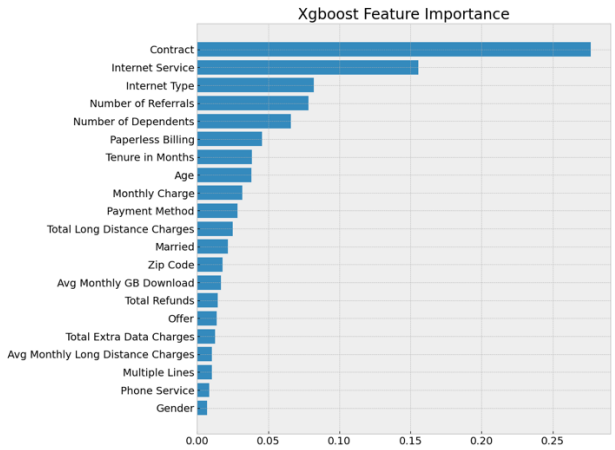

Figure 15: **Feature Importances**

Application

Through the modeling and analysis conducted so far, we are not only able to predict whether a customer is a potential churn risk but also to address the reasons contributing to churn. This enables us to enhance customer retention and develop marketing strategies that attract new customers.

For instance, considering the type of contract, we can incentivize customers to commit to two-year long-term contracts by offering complementary services such as online backup, device protection, and premium tech support, or by offering affordable options for unlimited data download or streaming services like TV, music, and movies. According to our analysis, this would be an effective strategy.

Another significant factor, the number of dependents, suggests that creating family and friend-specific promotional plans that offer more benefits as more connections are added could potentially address the high churn risk among customers with multiple dependents. For the number of referrals, implementing schemes that increase referral counts and provide additional discounts is also one of the commonly known strategies to date.

Zip Code is also an important feature. Previous analysis has shown that certain high-population areas experience a high rate of customer churn. Targeted investigations and tailored attractive offers for these regions could be beneficial.

Furthermore, a notable group is identified: customers over the age of 65 exhibit high churn rates, yet customers who do not use internet services demonstrate high retention rates. Therefore, it is also worth considering designing a specific plan for this special group to address their high churn rates. For example, offering an exclusive, highly advantageous plan for users over 65 that only includes telephone services without internet.

In future work, further analysis to identify different special groups, their influencing factors, and solutions, could be a useful and intriguing project topic.

Conclusion

In summary, this project aims to establish a comprehensive and robust predictive model for customer churn in the telecom industry using machine learning methods. With limited information, it identifies potential churn customers and key factors influencing churn rates. From these insights, tailored marketing strategies are formulated for different scenarios to reduce churn rates, ultimately enhancing customer retention and loyalty.

Timeline

Week 1: Project Initialization and Data Analysis

During these weeks, the project scope and key objectives are defined, followed by customer data collection and preprocessing. Exploratory data analysis is then performed to gain insights into data characteristics, with feature engineering focused on creating new features and addressing data imbalance issues.

Week 2: Model Development and Evaluation

These weeks focused on developing predictive models using machine learning algorithms like gradient boosting and KNN, accompanied by basic hyperparameter tuning to enhance initial model performance. Model performance is evaluated using metrics such as recall, F1 score, ROC AUC, and confusion matrix.

Week 3: Final Model Refinement and Application

During this period, the most appropriate machine learning method was determined, and Bayesian Optimization was used for fine-tuning, with K-fold cross-validation implemented for robustness testing. Key features were identified using feature importances, which were then utilized to adjust and recommend marketing sales strategies.

Week 4: Final Evaluation and Presentation

In the final weeks, the report is prepared, summarizing the methods, results, insights, and recommendations. Necessary adjustments to the project will be made, followed by the presentation of the final report.

References

[1] Sharmila K Wagh, Aishwarya A Andhale, Kishor S Wagh, Jayshree R Pansare, Sarita P Ambadekar, S H Gawande
Customer churn prediction in telecom sector using machine learning techniques (2023)
10.1016/j.rico.2023.100342

[2] Varun E, Pushpa Ravikumar, Chandana S, Spandana K M
An efficient technique for feature selection to predict customer churn in telecom industry (2019)
10.1109/ICAIT47043.2019.8987317

[3] A A Q Ahmed, D Maheswari
Churn prediction on huge telecom data using hybrid firefly based classification (2017)
10.1016/j.eij.2017.02.002