Churn Forecasting for Telco Using ML Models

Table of Contents

1.Abstract *2*

2.Introduction *2*

3.Method and Data *2*

3.1.Data Cleaning and Pre-processing 3

3.2.Data Description 3

4.Model *5*

4.1.Model selection 5

4.2.Model Evaluation 5

4.3.Features Importances 6

5.Conclusion *7*

1.Abstract

Customer churn is one of the major issues that companies, especially those in service-oriented businesses, have to deal with in order to retain customers in the modern competitive market. This paper addresses the issue by developing a predictive model based on the demographics and behaviors of customers. The model is designed to assist Telco, a telecommunications provider, to proactively identify customers at high risk of leaving. Hence, the company can make informed decisions to improve retention. The paper focuses on two primary objectives: identifying customer characteristics most associated with churn; and determining models that accurately predict customer churn. However, the model does not account for unmeasured variables such as service quality or customer sentiment, which may also influence churn decisions.

2.Introduction

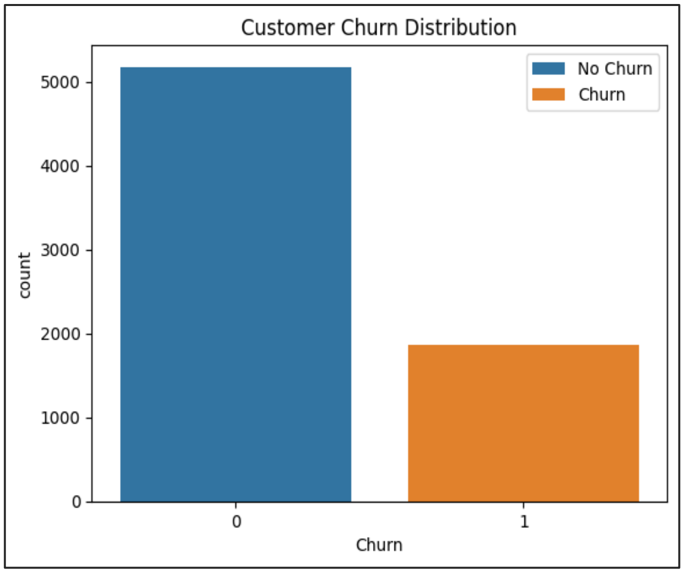
As the number of industries continues to grow, the market is getting more and more competitive. To attract customers, many companies flood the market with highly appealing and delectable offers leaving consumers overwhelmed by choices. As a result, customers tend to switch among providers in search of better deals or exclusive offers for new users. This rising trend of customer turnover has made acquiring new customers more challenging and more costly for businesses. In this context, retaining existing customers and keeping churn low are more cost effective for service-oriented businesses, Telco in particular.

The first step to reducing customer churn and keeping loyal customers is to understand the reason for churning. Therefore, this paper will focus on using machine learning to predict customer churn. The paper includes five classification techniques: decision tree, random forest, XGBoost, support vector machine (SVM), and logistic regression. These models were evaluated and compared using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) metrics to determine the most effective approach.

3.Method and Data

The original dataset contains 7043 rows (customers) and 21 columns(features), which can be obtained from this link here. The target variable is 'Churn,' indicating whether a customer had churned or not.

3.1.Data Cleaning and Pre-processing

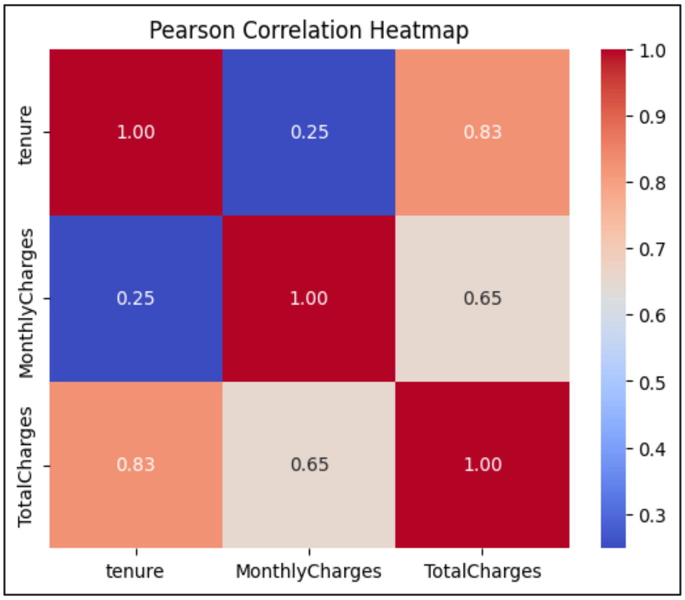


Following the initial analysis, the dataset was found to contain 7,043 entries. There are 11 missing values, which occur because some entries belong to new customers and therefore lack total charge information. For these cases, the total charge feature was set to 0. Additionally, the customer ID is just an identifier and does not contribute to the classification process, so it was excluded from the analysis.

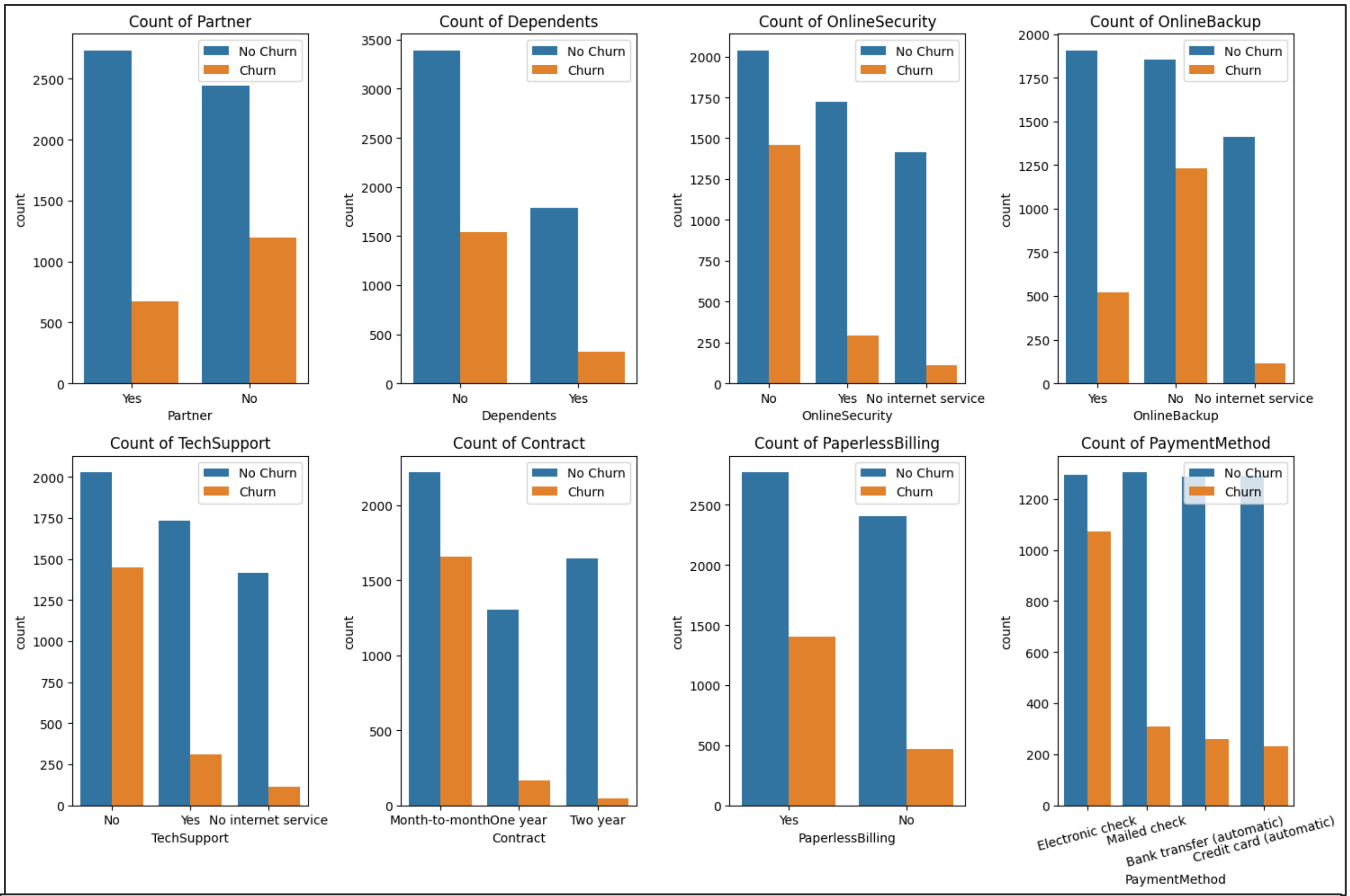
For the target variable, the study found that there were 5,174 non-churned and 1,869 churned customers. This distribution is expected, since only a small portion of customers typically churn. However, it results in an imbalanced dataset. To improve prediction accuracy, the Synthetic Minority Oversampling Technique (SMOTE) was used on training data to generate additional churned instances and balance the classes.

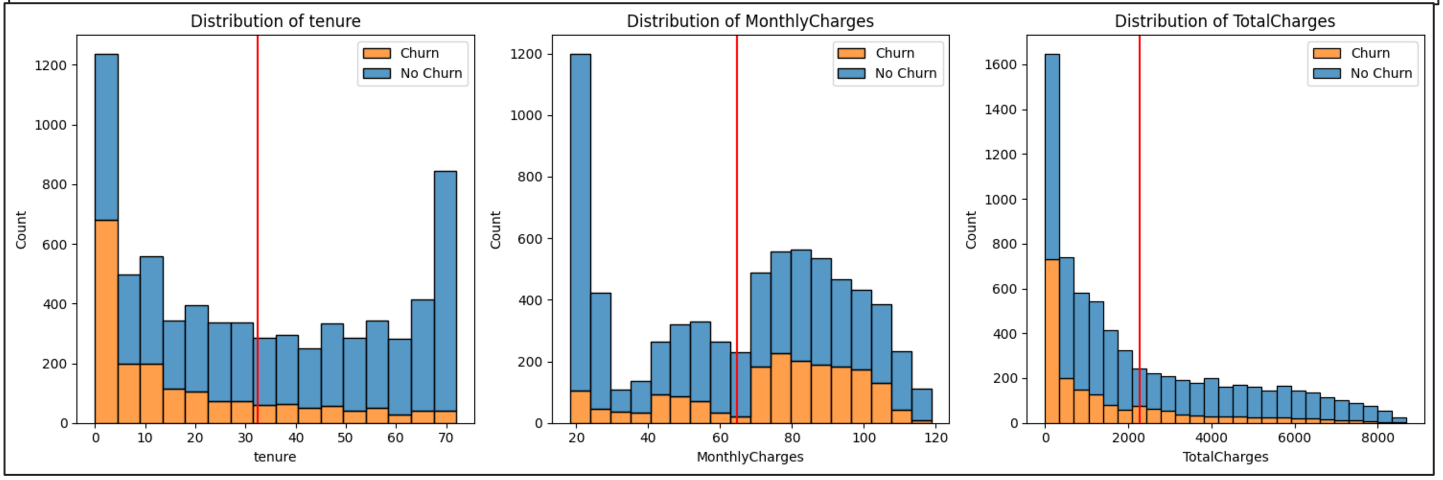
3.2.Data Description

The churn rate is highest among customers with short tenure, high monthly charges, and low total charges, indicating that new users who pay more upfront are more likely to leave early. The most churn-prone category is customers on month-to-month contracts, followed by those using fiber optic internet and electronic check as a payment method. Categorical features like no partner, no dependents, and lack of online security or tech support also correlate with higher churn. Among numerical features, TotalCharges has a strong positive correlation with tenure (0.83) and MonthlyCharges (0.65) which shows there's multicollinearity. Most features like tenure and TotalCharges are right-



skewed, with many customers in the lower range. These patterns suggest that early-stage users who don’t engage with bundled services or secure long-term plans are most at risk of churning.





4.Model

4.1.Model selection

To achieve more accurate predictions of customer churn, the study compared five different models using cross-validation to assess their F1 score. The models tested included Decision Tree, Random Forest, XGBoost, Logistic Regression, and SVM. For those linear classifiers (Logistic Regression and SVM), the TotalCharge feature will be droped due to high correlation with other featurers. The cross-validation results revealed that the ensemble methods outperformed the other models in terms of F1 score. Specifically, Random Forest achieved the highest performance with an average F1 score of 0.84, followed by XGBoost at 0.82, and Decision Tree at 0.79. Meanwhile, the linear classifiers—Logistic Regression and Support Vector Machine (SVM)—showed lower scores of 0.77 each.

Based on these findings, Random Forest, XGBoost, and Logistic Regression were selected for further investigation. Random Forest and XGBoost demonstrated superior predictive power, while Logistic Regression was included as a strong baseline model due to its interpretability and efficiency, despite its slightly lower F1 score. The next steps will involve hyperparameter tuning and feature selection analysis to further optimize and compare these models on a hold-out test set

4.2.Model Evaluation

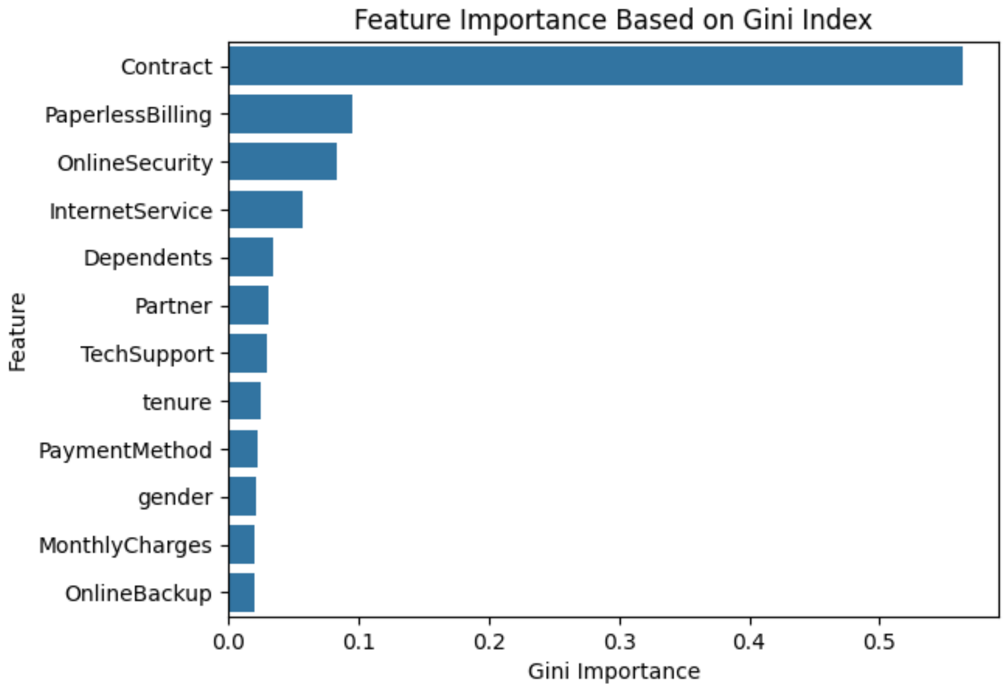
After conducting hyperparameter tuning and evaluating model performance on the test set, XGBoost was selected as the final model for predicting customer churn. It achieved the highest F1 score of 0.801 and the best accuracy of 79.8%, outperforming both Random Forest and Logistic Regression. These metrics indicate that XGBoost provides the best overall balance between correctly identifying churners and avoiding false alarms. In churn prediction, it's important to look beyond just accuracy due to class imbalance—where the number of churners is much smaller than non-churners.

The recall for the churn class (0.66) reflects how many actual churners were successfully identified, while the precision(0.61) shows how many of the predicted churners were truly going to churn. While Logistic Regression had the highest recall (0.81), it came with significantly lower precision (0.52) and F1 score (0.764), meaning it predicted many loyal customers as churners, which could result in wasted retention efforts. The F1 score, which balances precision and recall, is especially useful in this context because it provides a more realistic measure of performance on imbalanced datasets.

Compared to Random Forest (F1 score of 0.797), XGBoost slightly outperformed it across most metrics, which makes XGBoost the most suitable model for this customer churn prediction task.

Model	**F1 Score**	**Accuracy**	**Recall (Churn)**	**Precision (Churn)**
Random Forest	0,797	0,793	0,69	0,59
XGBoost	0,801	0,798	0,66	0,61
Logistic Regression	0,764	0,751	0,81	0,52

4.3.Features Importances



Feature importance refers to the technique used to identify which input variables contribute most to a model's predictions. In this study, feature importance was calculated using Gini importance from the tree-based model. The results show that Contract type is by far the most influential factor in predicting churn, contributing over half of the model's decision-making power. This suggests that customers on month-to-month contracts are far more likely to churn compared to those on long-term agreements. This suggests that offering more long-term contract options or incentives for customers to commit for longer periods might help in reducing churn. Other significant features include PaperlessBilling, OnlineSecurity, and InternetService, which reflect user behavior and perceived service value. Meanwhile, features like gender, MonthlyCharges, and OnlineBackup had relatively low importance, indicating limited impact on churn prediction in this dataset.

5.Conclusion

The study successfully developed a predictive model to identify customers at high risk of churning using machine learning techniques. Among the models evaluated, XGBoost emerged as the most effective, offering the best balance between precision, recall, and overall F1 score. Key findings revealed that contract type, billing preferences, and service engagement play crucial roles in customer retention. By leveraging these insights, Telco can implement targeted strategies to retain vulnerable customers, reduce churn rates, and enhance long-term profitability.