Enhancing Customer Churn Prediction in Telecom: A Comparative Analysis of Ensemble Algorithms

Gowrishangar KP
Indiana University

Yashwanth Vijayaragavan
Indiana University

project-yavijay-gokovi

Abstract

One of the biggest problems that companies in a variety of industries confront is customer attrition. It affects income in addition to posing issues with client loyalty and happiness. This project uses three different binary classification algorithms—Random Forest, Gradient Boosting, and Support Vector Machines—to investigate predictive analysis on the "Telco Customer Churn" dataset that was acquired through Kaggle. The dataset is a useful tool for figuring out what factors affect customer attrition because it includes a variety of consumer attributes like gender, contract type, senior citizenship, and more. In this report, we present the methodology and results of our analysis, shedding light on the performance of each algorithm in predicting customer churn. The findings provide insights that can assist telecom service providers in understanding and mitigating churn, ultimately improving customer retention.

Keywords

Churn, Random Forest, Gradient Boosting, Support Vector Machines, Algorithms, Attrition, Customer retention, Feature importance

1 Introduction

In the highly competitive telecommunications sector, controlling customer attrition is essential to keeping your firm successful. The rate at which consumers cancel their contracts or subscriptions, or customer churn, directly affects a business's earnings. For telecom service providers to create successful customer retention strategies, they must have a thorough understanding of the elements that lead to client turnover. In this study, we examine the analysis of the "Telco Customer Churn" real-world dataset from Kaggle. The dataset includes a variety of consumer factors that may impact a customer's choice to leave, such as contract specifics, service usage, and demographic data. This project's main objective is to use machine learning techniques to forecast client attrition using the data at hand.

Previous work

From [6], we learn about the different data preparation techniques that are used on large data set to prepare the data for complex data mining algorithms. Also, we take inspiration from

[1] and [2] to carry out Gradient Boost algorithm on complex data set for binary classification. Cutler, D. [3] and Breiman, L.[4] works throw insight on random forest technique in classifying the data set across different domains.

2 Methods

2.1 Data Preprocessing

In the initial phase of our analysis, we began with the loading of the dataset and a succinct display of its initial rows, offering a preliminary glimpse. Our attention then turned to the identification and conversion of non-numeric values within the "TotalCharges" column. Subsequently, we delved into generating summary statistics for numeric columns and leveraged visualizations, including bar plots and count plots, to unveil relationships between features and the "Churn" label.

2.2 Exploratory Data Analysis (EDA)

The exploration of numeric feature relationships with the "Churn" label took center stage in this phase. We employed bar plots, highlighting means, and count plots for categorical columns. Additionally, a pie chart provided a comprehensive overview of the percentage distribution of churn labels. A crucial facet of our analysis involved scrutinizing Pearson's correlation matrix, unraveling intricate feature correlations.

2.3 Missing Data Handling

Addressing missing values in the "TotalCharges" column constituted a pivotal step in our data preprocessing journey. Employing iterative imputation, we meticulously handled missing entries. Simultaneously, we tackled skewness in numeric columns through a judicious application of logarithmic transformation.

2.4 Feature Engineering

A proactive approach to enhancing our model's understanding involved the introduction of a new feature—the logarithm of "TotalCharges." This feature engineering step aimed to provide a nuanced perspective on existing feature relationships.

2.5 Data Modeling and Evaluation

Transitioning into the modeling arena, we meticulously split our dataset into training and testing subsets. A suite of essential preprocessing steps, including feature scaling and encoding, preceded the selection of models—Support Vector Classifier (SVC), Random Forest Classifier (RFC), and Gradient Boosting (GB) Classifier. Rigorous evaluation ensued, leveraging k-fold cross-validation to scrutinize accuracy, precision, recall, and F1-score metrics.

2.6 Feature Importance Analysis

A meticulous examination of feature importance unfolded through the Feature Importance Checker. This analysis, revealing the top 20 features for each model, played a crucial role in the subsequent feature selection process.

3 Results

7043 instances were analysed across 21 features using different data exploration methods. Machine learning models like Random Forest and Gradient Boosting helped in retrieving key features that could help the company in customer retention and result in increased profits.

Figure 1 displays the different unique values of each feature against the churn label. It gives important information about the company's internet providing services, customers' bill cycle behavior against churn label and customers response against other key services that the organization offers.

Figure 2 displays the average value of the numerical column against the churn values. It gives useful insight into the number of months and the total cost accumulated by a customer during their lifetime.

Figure 3 displays the correlation between different numerical data.

Figure 4 displays the results of different machine learning models that were used in this experiment. It displays the results of average F1 score, accuracy, precision, recall and f1 value of support vector machine, random forest and gradient boosting from left to right. A balanced tradeoff between precision and recall resulted in using random forest and gradient boosting models for retrieving the top features of the dataset.

Figure 5 displays the results of random forest and gradient boosting model from left to right. It shows the top 20 features from the dataset according to their F1 score. Monthly charges, Total charges and Tenure month being the most important features of the set.

4 Discussions

These are the following recommendation for the telecommunication company from our analysis,

- Tenure, Monthly Fees, and Internet Service Type are recognized by Random Forest Classifier as important variables.
- Tenure, Total Charges, and Monthly Charges are deemed by Gradient Boosting to be the most significant features.
- Telecom companies should concentrate on customers with high monthly charges, particularly those with fiber optic internet service, in order to reduce churn.
- Customer retention is significantly impacted by contract terms, payment options, and tenure, highlighting the significance of customized retention strategies.

From the nature of the dataset displayed in Figure 6, we get to know that the data is slightly imbalanced. This is due to the high difference between the churn labels. If we were to reduce this imbalance and potentially arrive at a normal and balanced data. We could further explore the interactions between these important features to develop more effective churn prevention strategies.

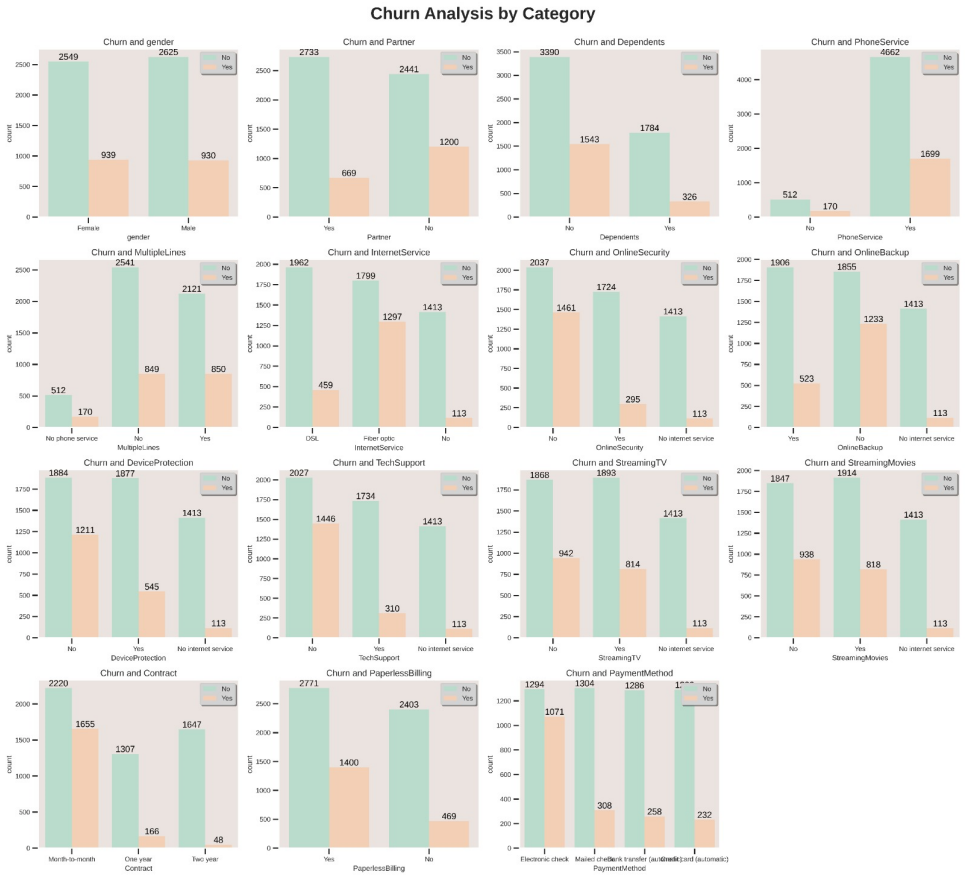

Figure 1: Unique values of features against churn label

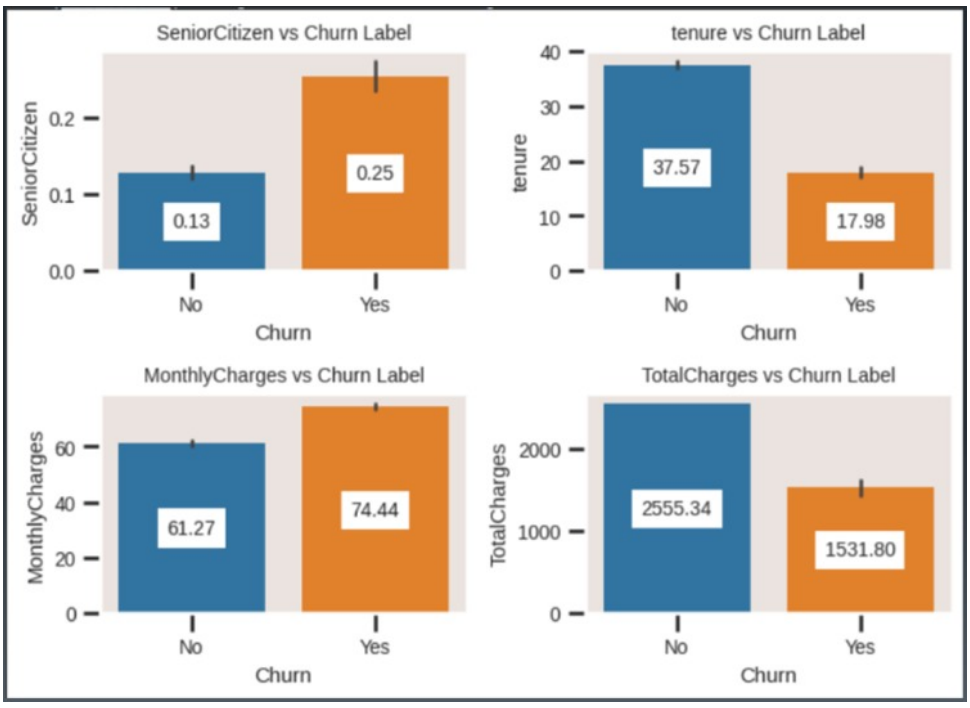

Figure 2: Numerical data against churn label.

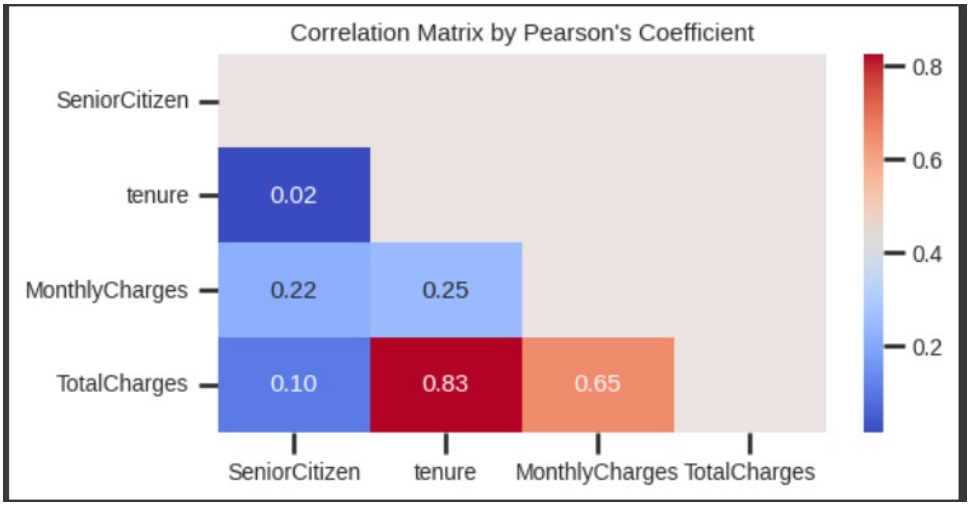

Figure 3: Correlation Matrix.

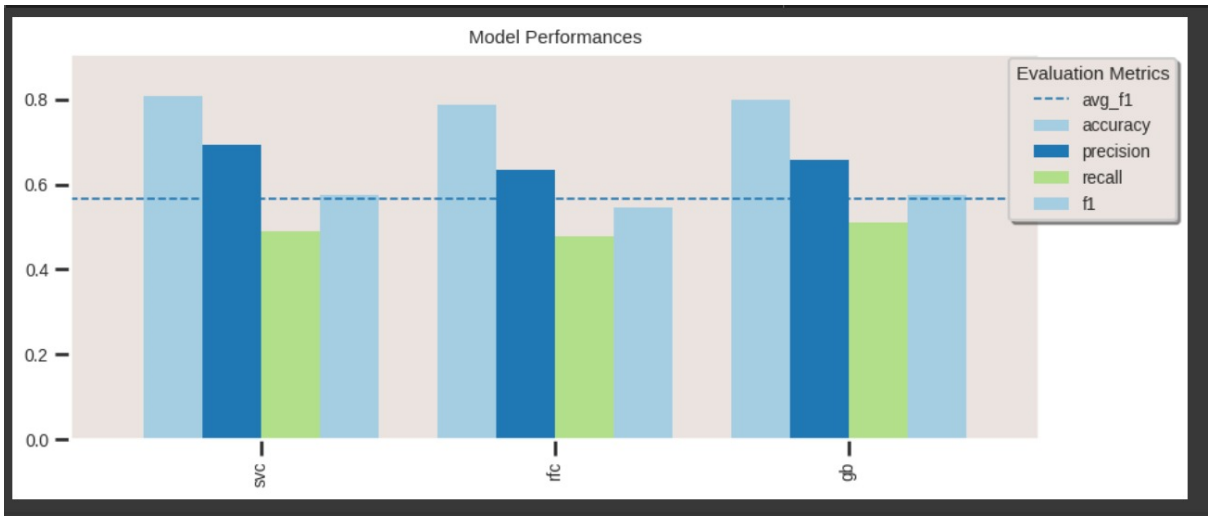

Figure 4: Performance of models.

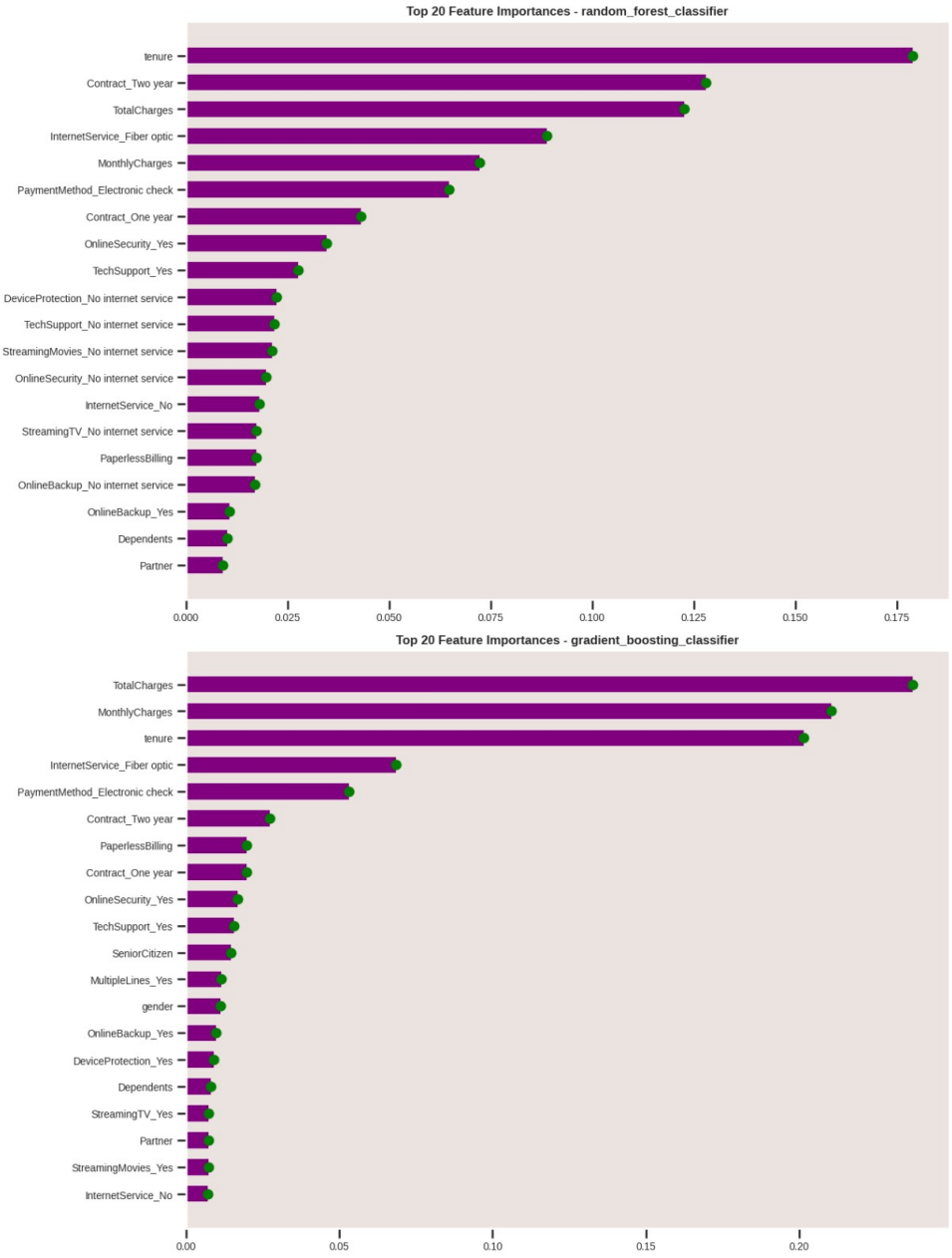

Figure 5: Top 20 features.

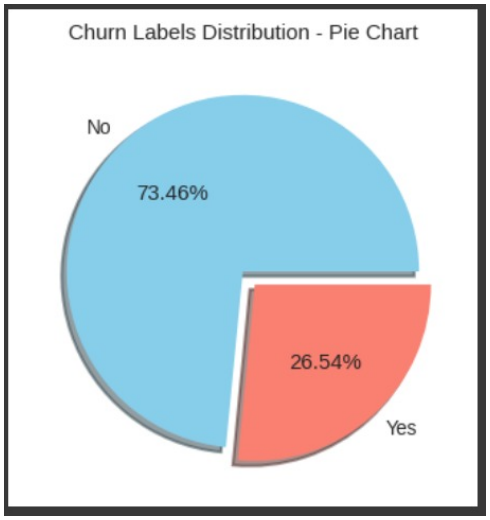

Figure 6: Churn distribution.

5 Author Contribution

- Gowrishangar Kovilpalayam Punithan (Exploratory Data Analysis - EDA): In charge of cleaning up the data, putting it into a visual format, and doing preliminary research to identify patterns and abnormalities.
- Yashwanth Vijayaragavan (Model Training): Specialized in training Random Forest Classifier, Support Vector Classifier, and Gradient Boosting models, including algorithm selection, feature importance, and performance evaluation.

6 References

[1] Friedman, J. H. (2002). "Stochastic Gradient Boosting." Computational Statistics Data Analysis, 38(4), 367-378.

[2] Chen, T., Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[3] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J. (2007). "Random Forests for Classification in Ecology." Ecology, 88(11), 2783-2792.

[4] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32.

[5] Cortes, C., Vapnik, V. (1995). "Support-Vector Networks." Machine Learning, 20(3), 273-297.

[6] Pyle, D. (1999). "Data Preparation for Data Mining." The Morgan Kaufmann Series in Data Management Systems.