Telco Customer Churn

Data Science Project by Raja Ridgway

Problem

Executives at Telco, a small telecommunications company in California, were interested in understanding how to develop programs to retain customers. Given that the cost of acquiring new customers is more expensive than retaining existing customers, **company executives wanted to more accurately target specific existing customers that might want to leave**. Specifically, company executives wanted a predictive model for which customers would churn.

The data team at Telco developed a dataset of 7,043 customers from Q3 of 2021. The dataset included demographic information and tenure, services used, contracts and payments methods, monthly and total charges, and whether the customer churned. The data team wanted to use the dataset, which was provided as a .csv file, to answer the questions posed by the company executives. No additional data sources were used.

Approach

To answer the questions, several steps were taken:

- **Access and Clean the Dataset.** The data was downloaded from Kaggle and processed through a Jupyter notebook. Features names were compared to descriptions; data types, outliers, and duplicate records were addressed; and irrelevant features were dropped. A brief review of summary statistics was also conducted using pandas profiler to develop a better sense of the customers in the dataset. A clean version of the dataset was exported for use in the exploratory data analysis.

- **Conduct Exploratory Data Analysis.** The clean version of the dataset was visualized to identify which variables might have an impact on customer churn. Specifically, the data was visualized using count plots (categorical features), histograms, and boxplots (numerical features). Several features appeared to have some impact on customer churn - those features were tested for statistically significant association with customer churn using the Chi-Square Test (given that churn is categorical yes/no).

- **Preprocess and Model the Data.** The clean version of the data set was used to train and fit the data to several models. First, one-hot encoding was used to convert the categorical features into numerical values. Then, standardization was used for the numerical features given that they were on different scales (i.e., tenure and monthly/total charges). Note that standardization was used instead of min/max scaling as new data might have higher values for all numeric values. An eventual 67/33 train/test split was eventually used for five machine learning models: logistic regression, K-nearest neighbor, random forest, gradient boosting, and cat boost. These models were chosen because of the categorical target feature.

- **Evaluate the Models.** Models were compared using accuracy, F1-scores, lift, and mean cross-validation scores. Hyper parameter tuning was then used to differentiate between the best models.

Findings

Exploratory Data Analysis

When comparing the distributions to customer churn, there were several interesting findings:

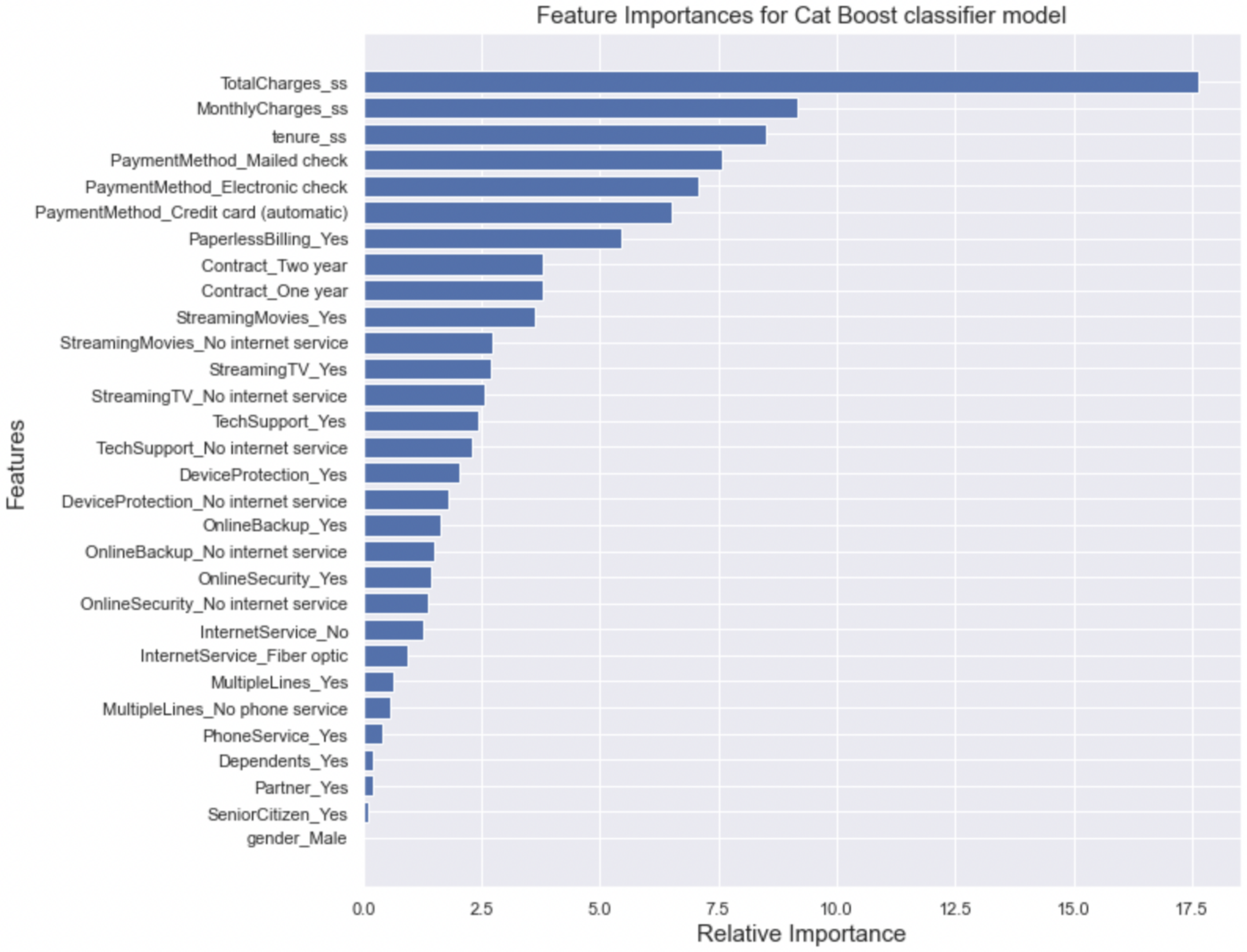
- **Demographics.** While gender does not appear to impact churn, senior citizen status does (senior citizens are more likely to churn than non-senior citizens). Additionally, customers without partners or dependents are more likely to churn than those with partners or dependents. Finally, the customers with the shortest tenure are more likely to churn than those with longer tenure.
- **Phone and Internet Service.** Having phone service does not appear to impact churn, and neither does multiple lines or not. The type of internet service does appear to impact churn, however, with those customers having fiber optic being more likely to churn than those with DSL or without internet service.
- **Additional Services.** Paying for additional services when having internet service does appear to impact churn. Customers paying for online security, online backup, device protection, or tech support are less likely to churn than those that don't pay for those services. Paying for streaming TV or streaming movies, however, does not appear to have the same impact on churn.
- **Billing Options and Charges.** Customers paying month-to-month and those with paperless billing appear to be more likely to churn than those paying on longer contracts and those without paperless billing. Additionally, customers who pay with an electronic check are more likely to churn than any of the other payment methods. Customer churn is the highest amongst those paying around $70-80 and customer churn decreases as total charges increase.

To test if the variables have a statistically significant association, hypothesis testing was done. The Chi-Square test was run for the 11 categorical variables that demonstrated some level of impact on customer churn. The Chi-Square test was chosen as the churn variable is also categorical. **All 11 categorical variables were shown to have a significant association with customer churn.**

Model Findings

Logistic regression (F1-score), random forest (balanced), and cat boost (lift) had the best results based on the evaluation metrics of accuracy, F1-scores, lift, and mean cross-validation scores. After hyper parameter tuning was implemented, Cat Boost provided the best overall model.

Additionally, the analysis highlighted which of the features were the most important:



Recommendations

Based on this analysis, two action steps are recommended:

- **Incentivize contracts.** The customers least likely to churn are those with the longest tenure who have higher monthly charges. Additionally, as can be seen in the feature importances graph, both one year and two year contracts were important for keeping customers - and likely led to longer tenures.

- **Market toward single people without dependents, but don't focus on gender.** Customers with partners and dependents were less likely to churn, so creating incentives for that demographic might not retain as many customers. Gender did not appear to be a factor, so developing incentives based on gender will likely not have an impact.

Future Research

There are several potential ways to expand up on this research:

- **Increase the Number of Customer Records.** The analysis completed for this project was with 7,043 records from one location. To improve the model, more records from additional locations and across a larger period of time could be used to improve the accuracy of the model.

- **Learn more about monthly charges and tenure.** Monthly charges and tenure obviously have an impact on whether a customer churns. Identifying the monthly charge threshold above which a customer is much more likely to stay or the number of months of tenure before they are much more likely to stay would help provide

- **Add probabilities of customer churn.** The current model only predicts if a specific customer will churn or not. A next step could be to identify the probability that a customer will churn which would allow for better action planning to be taken around retention methods.