Customer Churn Classification
DATA 1030 Midterm Report
Yuetian Li
Data Science Initiative at Brown University
GitHub repository: https://github.com/yuetianli22/DATA1030_FinalProject

Introduction

While customer churn is a widespread problem across almost every industry nowadays, it is especially relevant to the telecommunications market. The competition is at an all-time high as many believe its market has reached the saturation point. To retain customers effectively, it is essential for the telecom companies to accurately predict customer churn. The current project attempts to create a machine learning tool to classify whether a customer would leave the platform based on customers' information.

The dataset used for this project was originally from IBM Cognos Analytics and was published in Kaggle. The dataset is well-documented. It contains customer IDs, the target variable *Churn* with 2 categories indicating whether customers left within the last month (Yes=Class 1), and 19 features. Four categorical features indicate demographic information of customers including gender, senior citizen status and whether they had partners and dependents. Nine categorical features indicate whether costumers signed up for phone service, multiple phone lines, internet service, online security, online backup, device protection, tech support, streaming TV, and streaming movies. Six features indicate the customer account information including how long they had been a customer (i.e., *tenure*), contract type, payment method, monthly charges, total charges, and whether they used paperless billing. In total, there are 7043 customers.

Several projects published in Kaggle have used this dataset to predict customer churn. In one of the projects, the author used logistic regression and achieved an accuracy of 81% [2]. In another project, the author used XGBoost and achieved a classification accuracy of 83% [3]. Based on the published projects, the highest accuracy achieved is around 83%. Such accuracy could be due to the limited information provided by the features in dataset and could be addressed through additional data collection. In the current project, I aim to explore different machine learning approaches and further improve the performance of classifier.

Exploratory Data Analysis

The exploratory data analysis (EDA) investigated the target variable *Churn* and found it imbalanced with 26.5% of customers leaving the platform and 73.5% of customers staying. In the whole dataset, there were 11 missing values in feature *TotalCharges.*

The relation between the target variable and every feature was visualized. The following section contains several of the figures created during EDA.

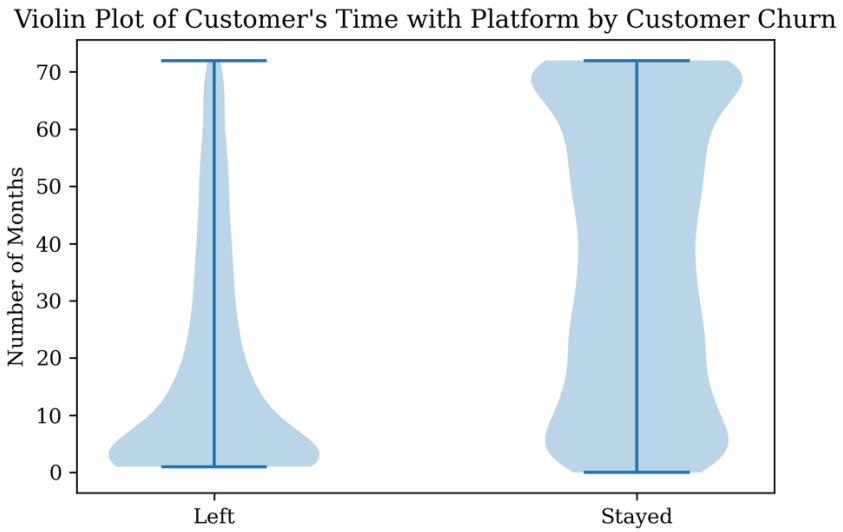

Figure 1. The kernel density plots for the number of months that customer sstayed with the company, segregated by customer churn. Most of customers who ended up leaving the platform stayed less than 10 months.

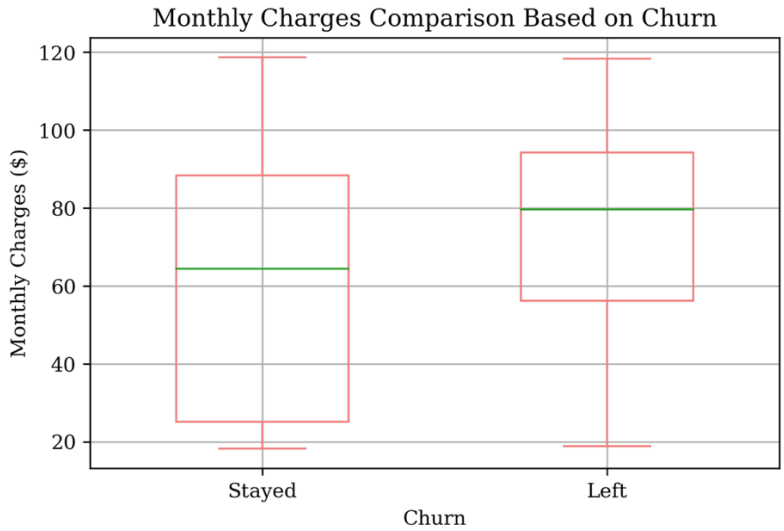

Figure 2. The boxplot of the monthly charges grouped by customer churn. The monthly charges mean of customers who left were higher compared to staying ones.

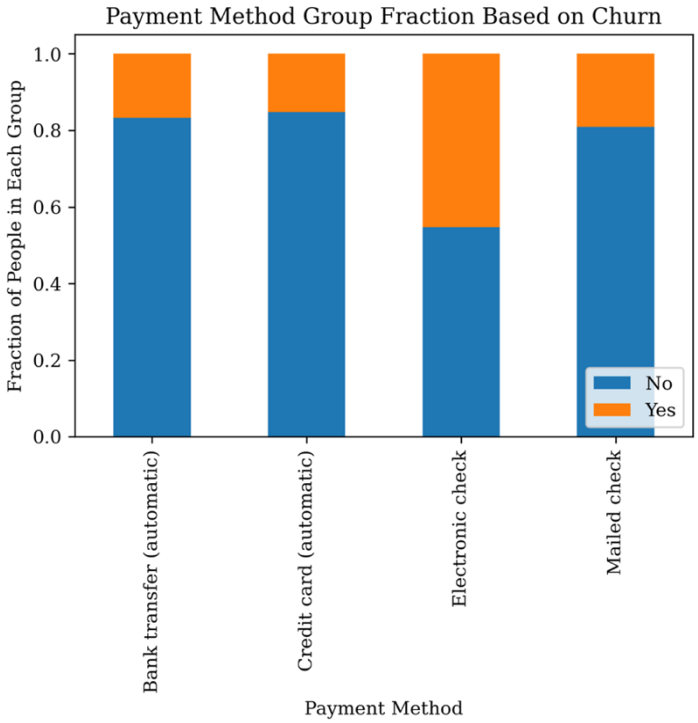

Figure 3. The figure displays the fraction of customer churn for each payment method. Among the four payment methods, people who paid with electronic check had highest churn rate.

Methods

Data preprocessing

The EDA suggests that the target variable *Churn* is imbalanced with 26.5% of customers leaving the platform and 73.5% of customers retained. Thus, the current project applied stratified splits to this imbalanced dataset to ensure that the percentage of two categories in *Churn* is consistent in the training, validation, and test datasets. The stratified splitting step allocated 20% of the observation to testing, 60% to training, and 20% to validation.

Given each observation represents one individual customer, and each customer only has one observation in the dataset, the data is assumed to be independent and identically distributed without group or time-series structures.

The exploratory data analysis found 11 missing values in feature *TotalCharges.* Given such small size of missingness (0.16%), I decided to delete these 7 datapoints with missing values.

The preprocessor applied StandardScaler on the continues features *tenure*, *MonthlyCharges*, and *TotalCharges* as the histograms of these features suggest none of them are reasonably bounded and suitable for the MinMaxEncoder. OneHotEncoder was applied on the unordered categorical features that are either demographic information or contain unordered categories (e.g., payment methods). The preprocessor applied OrdinalEncoder on the categorical features related to services signed up by customers. The categories in these features can be ranked as they indicate different levels of service that customers rely on the platform to provide. As a result, the final preprocessed dataset has 27 features. The target variable was label encoded as Class 1 (Churned) and Class 0 (No Churned).

Machine Learning Model Selection

Seven machine learning models were trained and compared: a logistic regression with L1 regularization or Lasso logistic regression, a logistic regression with L2 regularization or Ridge logistic regression, a logistic regression with Elastic Net, a support vector machine classifier, a K-nearest neighbors classifier, and an XGBoost classifier. The hyperparameters that were tuned in each of seven models were listed in Table 1. Brute-force grid search method was used to determine the optimal combination of hyperparameters. This process was repeated for 10 different data splits and at 10 different random states of models.

Models	Hyperparameters	Values
L1 Logistic	C	1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4
L2 Logistic	C	1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4
EN Logistic	C	1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4
	L1 Ratio	0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99
SVC	gamma	1e-2, 1e-1, 1e0, 1e1, 1e2, 'auto', 'scale'
	C	0.01, 0.1, 0.5, 1, 5, 10, 20
KNN	n_neighbors	1, 2, 3, 5, 10, 30, 50, 100, 200
	weights	uniform, distance
RF	max_features	1, 3, 5, 10, 20, None
	max_depth	1, 3, 5, 7, 10, 15, 20, None
XGBoost	gamma	0, 0.1, 0.2, 0.3, 0.4
	min_child_weight	1, 3, 5, 7
	subsample	0.4, 0.5, 0.65, 0.75, 1
	colsample_bytree	0.3, 0.4, 0.5, 0.7, 1
	learning_rate	0.05, 0.1, 0.2, 0.3

Table 1. Hyperparameters used to tune machine learning models

During tuning, F1 score was used to evaluate model performance. As the goal of this project is to predict customer churn, I intend to minimize the costs of retaining a costumer that model inaccurately predict to be leaving and the costs of losing a customer that model inaccurately predict to be staying. Thus, it is important to consider both recall and precision. Since F1 score takes recall and precision both into consideration, it is chosen as the evaluation metric.

After tuning, the best model hyperparameters at each grid search were extracted and the corresponding F1 scores were used for model comparison. The averaged and standard deviations (SDs) of F1 scores for each model across 10 random states were summarized in Figure 4.

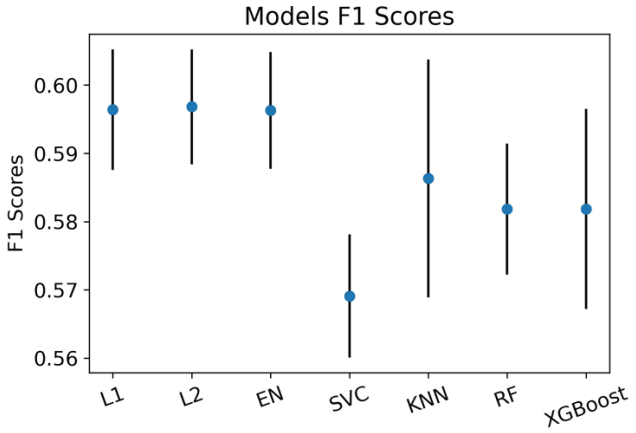

Figure 4. Averaged F1 scores with SDs across 10 random states of seven ML models

The logistic regression models with L2 regularization had the highest averaged test set F1 scores (mean F1=0.59) with a second lowest standard deivation (SD=0.008) and thus was chosen as the model of choice. For this model, the inverse regularization parameter C=10.0 achieved the highest test score for four out of ten random states. Once the model was finalized, examinations of global and local features were conducted to better interpret the findings.

Results

Evaluation of Models

The baseline F1 score was calculated through a classifer that always predicted class 1 and rsulted value is 0.42. Compared to the baseline F1 score, the logistic regression models achieved an averaged F1 score 0.59, which is 21 SDs above the baseline F1 score. It was also noted that the averaged recall is 0.54 with SD at 0.013, which suggests that the ability of model to identify *all* customer churn is not that ideal.

Global Feature Importances

Permutation Feature Importance. Over 10 random states of logistic L2 model, the 10th model with the inverse regularization parameter C=10 was chosen to conduct feature permutation. For each feature, 10 random shuffles were conducted. The mean and std of test set F1 scores over 10 random shuffles were summarized in Figure 5. Based on the permutation test, random shuffles of five features including *tenure, InternetService, Contract, MultipleLines* and *StreamingMovies* would disturb F1 scores most tremendously, suggesting that these features are the most important to produce stable F1 scores in the current model.

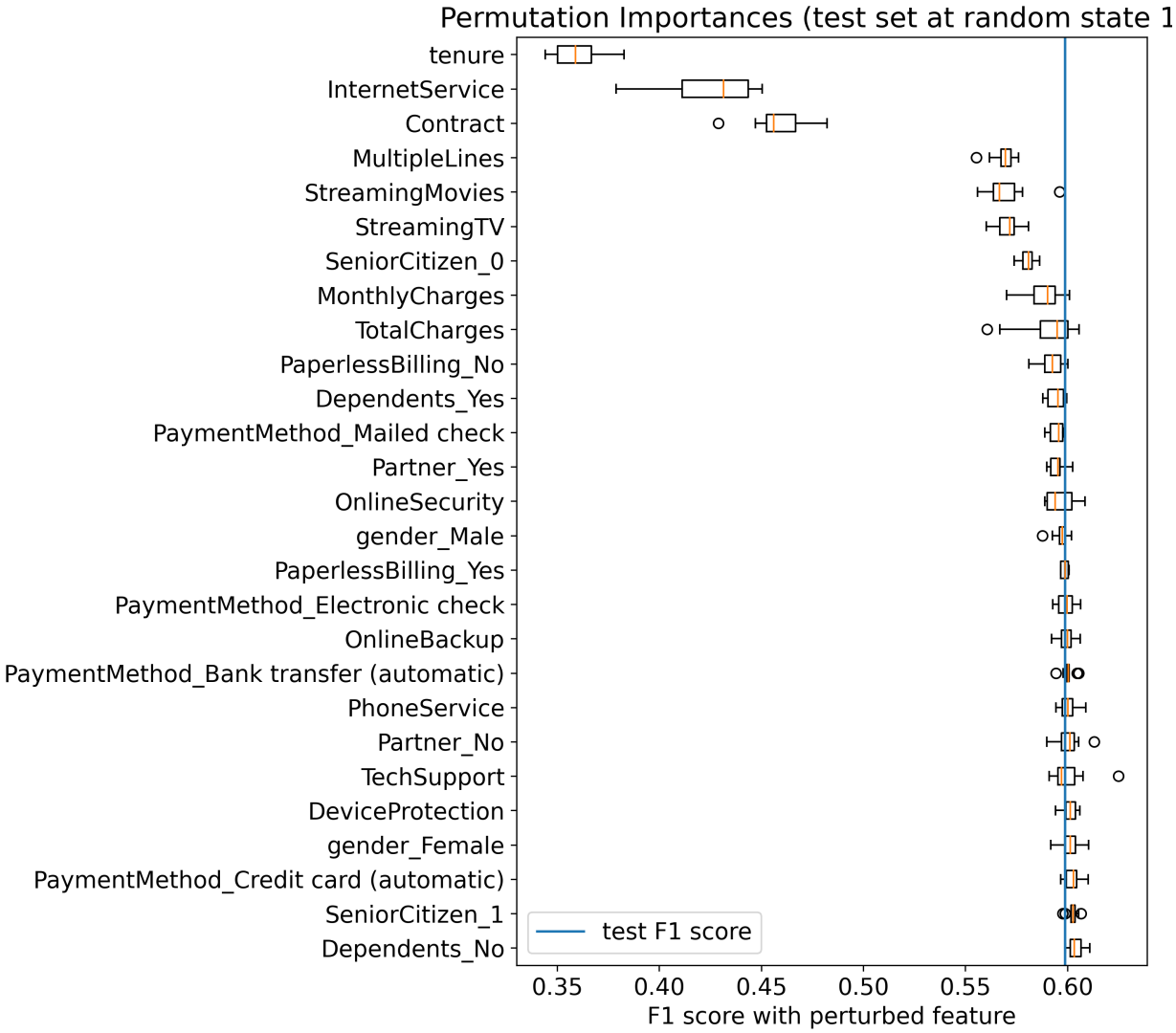

Figure 5. Features were ranked based on how disturbed F1 scores of test set were during permutation feature importance calculation.

Coefficients. Over 10 random states of Logistic L2 model, all features were standard-scaled, and the resulted coefficient values were used to measure global feature importance. The mean and SDs of coefficient values over 10 random states were summarized in Figure 6. Features with largest coeffcients include *tenure, InternetService, Total Charges* and *Contract.* Demographic features such as *gender* and *partner* are the leaset important. A large negative coefficient value of *tenure* suggests that longer customers had stayed with the company, less likely they would churn. The positive coefficient value of *InternetService* suggests that if customers signed up for more adviacned internet service (e.g., fiber optic), they were more likely to churn. The positive coefficient value of *Totalcharges* suggests that with a larger total bill charge, the customers were more likely to leave the platform.

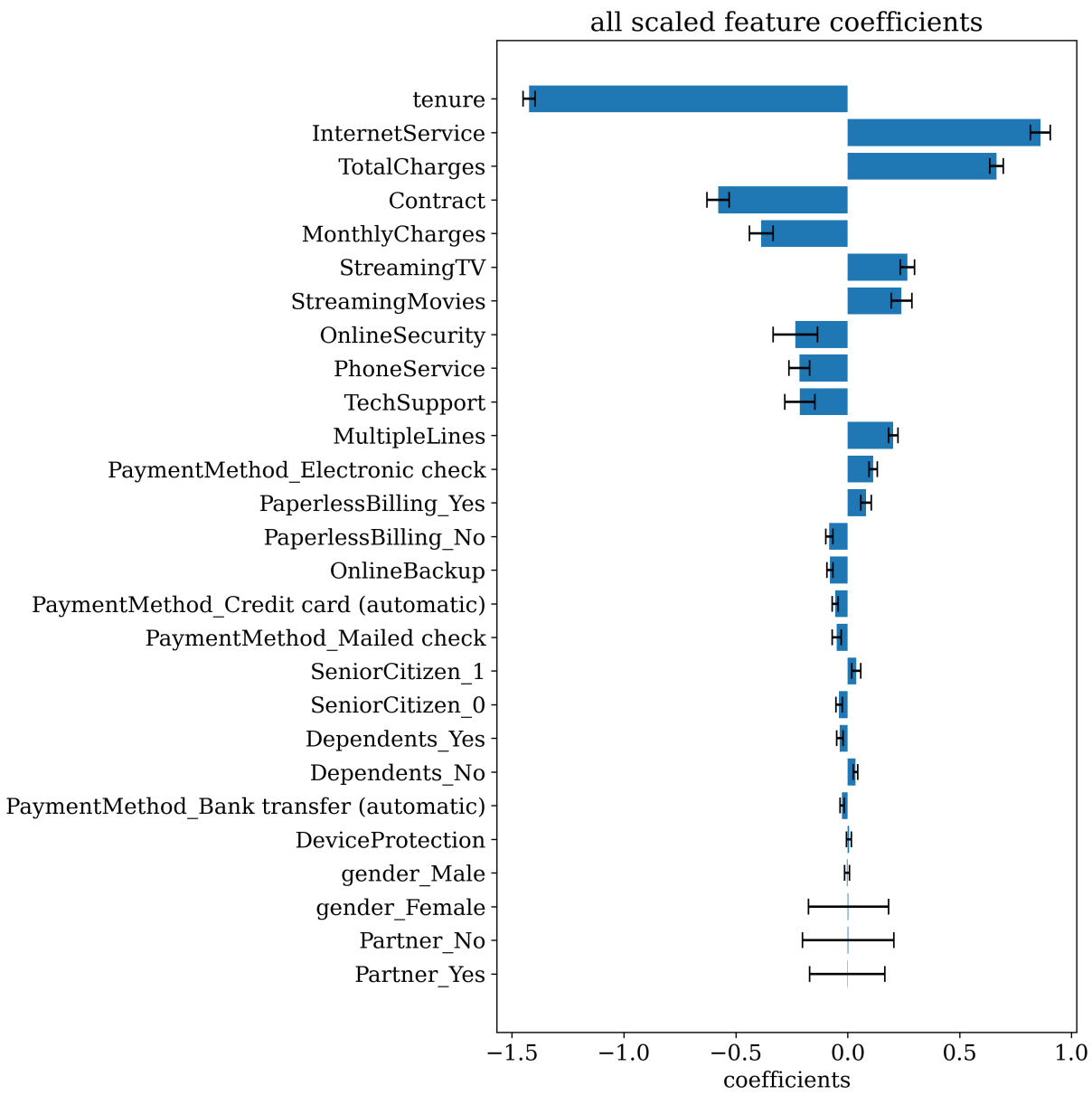

Figure 6. Standard-scaled features were ranked based on coefficient values over 10 random states of Logisitc L2 model.

SHAP Values. Over 10 random states of logistic L2 model, the 10th model with the inverse regularization parameter C=10 was chosen to calculate SHAP values. The mean absolute SHAP valus for features over all rows of the test set indicated the global contribution of features to the prediction. Based on SHAP values, the top five features including *tenure, InternetService, TotalCharges, MonthlyCharges,* and *Contract* have largest impact on the prediction and least important features include *PaymentMethod*, *SeniorCitizen* and *Partner* (Figure 7.).

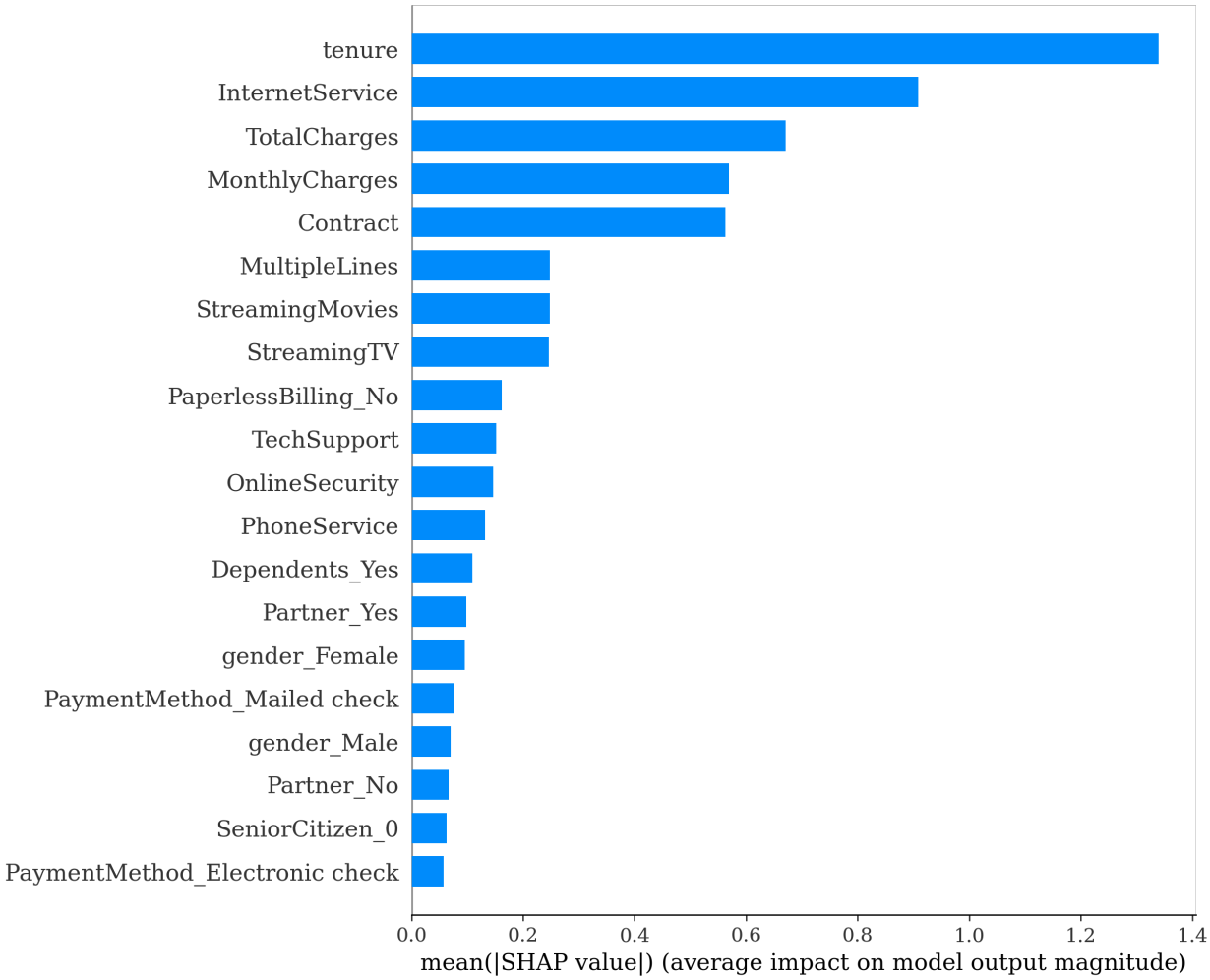

Figure 7. Features were ranked based on the mean absolute SHAP values over all rows of the test data.

Consistent across three approaches to calculate global importance, *tenure, InternetSerivce, TotalCharges, Contract* and *MontlyCharges* are the most important features. Features such as *gender*, *Partner*, *Dependents*, and *deviceProtection* are the least important features.

Local Feature Importances

As SHAP values were calculated as described above, two datapoint were examined to show how each feature contributed to their respective prediction (Figure 8.). For example, for customer #7, the prediction is churn, represented by that f(x) is greater than 0. Features such as *tenure* and *InternetService-FiberOpitic* contributed to increase of prediction of churn (class 1) wheares features such as *TotalCharges* contributed to decrease of such prediction. The largest local impact comes from feature *tenure*. For customer #33, similarly, features such as tenure, contract, and MonthlyCharges contributed to increase of prediction of churn wheares features such as *TotalCharges* and *InternService* contributed to decrease of such prediction.

Consistently with global feature importance and across these two datapoints, *tenure, TotalCharges, InternetSerivce, Contract* are among the most important features.

Figure 8. Based on SHAP values, this figure shows how each feature contributes to the prediction in two sample datapoints.

Based on the global and local feature importances, as expected, longer the customers had stayed with the company, less likely they would churn. Additionally, cutomers who were on month-to-month contract were more likely to churn compared to those with annual contract. Such result suggests that to retain a costumer, it is more useful to persuade customers to sign a longer contract in the first place. Unexpectedly, customers who signed up for more advanced internet service such as fiber optic were more likely to leave. Such unexpected result suggests an improvement is needed especially in more advanced internet service to keep customers satisfactory.

Outlook

Given that the model is a logistic regression, the results are more interpretable compared to past models that used XGboost. Additionally, the current project uses F1 score as evaluation metric, which considers both recall and precision in a biased dataset compared to previous models that only used accuracy score. One issue of current model is low recall. It suggests that the ability of model to identity all churned customers is not so ideal. To improve the classifier, more data should be collected and features such as customers' satisfaction, data plans, and monthly data used are helpful to address why customers leave and what companies can do to retain them.

Reference

[1]: BHARTI PRASAD. (2021, July). *CUSTOMER CHURN PREDICTION*. Kaggle. https://www.kaggle.com/bhartiprasad17/customer-churn-prediction

[2]: ATINDRABANDI. (2018). *Telecom Churn Prediction.* Kaggle. https://www.kaggle.com/bandiatindra/telecom-churn-prediction