WPI

Statistical Methods for Data Science
(DS 502/MA 543)

Group 5 Project Final Report

Telco Customer Churn Prediction

Team Members:



Vandana Anand **Kratika Agrawal** **Jiaju Shi**

Introduction

With the increasingly fierce competition in the telecommunications industry, retaining customers becomes a key business indicator for better operation and competitiveness of the company. This key point requires understanding the characteristics of the customer that stopped using the company's services, analyzing the reasons for the customer leaving, predicting whether customers will leave, determining the retention of target users, and making effective plans to retain those target users.

The data are from a **Kaggle dataset** that comes from the IBM database. IBM is an American multinational technology and consulting company headquartered in Armonk, New York. Their primary business model is to produce and sell computer hardware and software as well as supply consulting services for system architecture and network hosting.

The Telecom Customer Churn dataset from IBM is used to determine the behavior of the employer to show the major causes of telecom customer churn and develop a targeted customer retention plan. A customer could churn due to many factors that could include anything from their own demographics to which services they have signed up to use. Our project focuses on pinpointing these factors to see which features cause customers to churn as well as finding the best model to use on the dataset.

Statistical Data Pre-processing

The original data was processed by handling inconsistent values and deleting or replacing the abnormal part of the data so that the input data becomes normal and reliable. Sampling, analyzing, pre-processing the existing data, and then modeling and evaluating the data can potentially reduce the problem of potential customer loss.

Descriptive Statistics

- The entire data size is about 955MB that consists of 21 columns and 7044 rows
- Each row stands for a customer and each column, or feature, holds customer's attributes described on the column metadata.
- Customers who left within the last month – the column called Churn (Target Column).
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

- Customer account information – how long they have been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

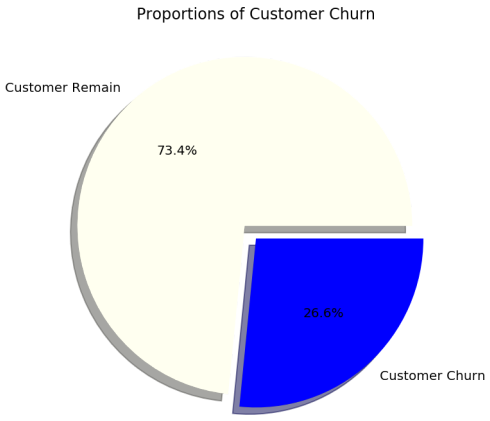
Data Cleaning

We first focused on cleaning and preprocessing the data so that the dataset is ready to apply the classification models. We performed the following tasks:

- Removed duplicate records
- Handled missing data
- Managed extreme values or outliers
- Checked and convert data to the proper data type (ie. numerical)
- Normalized the data using standard scaler
- Analyze initial data visualizations

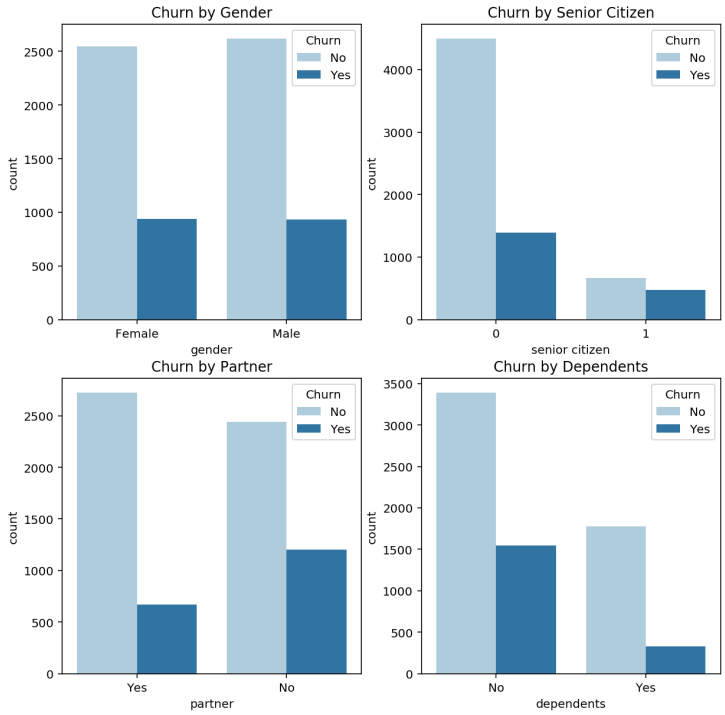
Data Visualizations

In order to get an understanding of the data and the impacts of churn, we used python to graph each of the categorical and numerical features versus churn. The following are some of the first data visualizations we used to get a gist of the dataset:

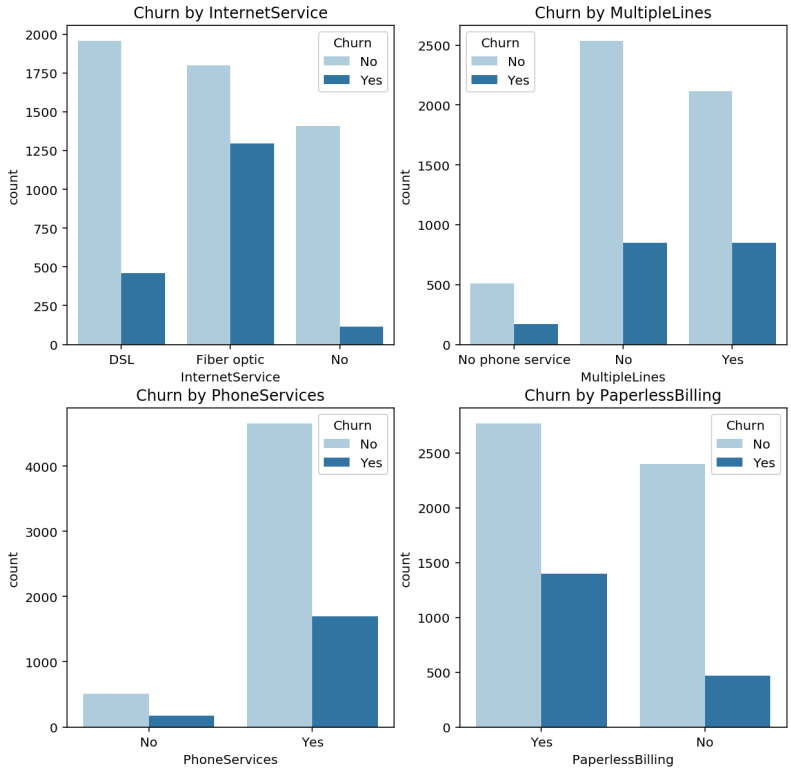


The pie chart tells us that overall, 26.6% of customers churned while the other 73.4% of customers continued to use the company's services. We wanted to find out what services or features made the 26.6% of customers churn.

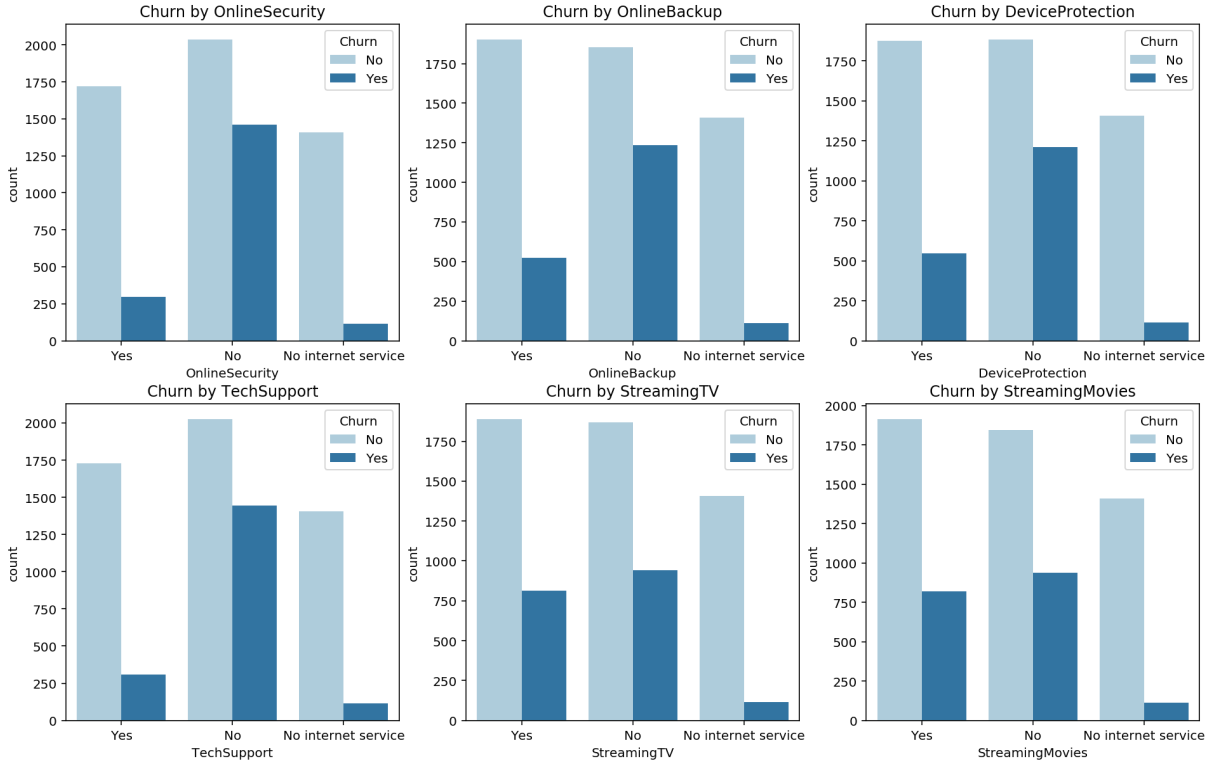
Churn by Demographic information:



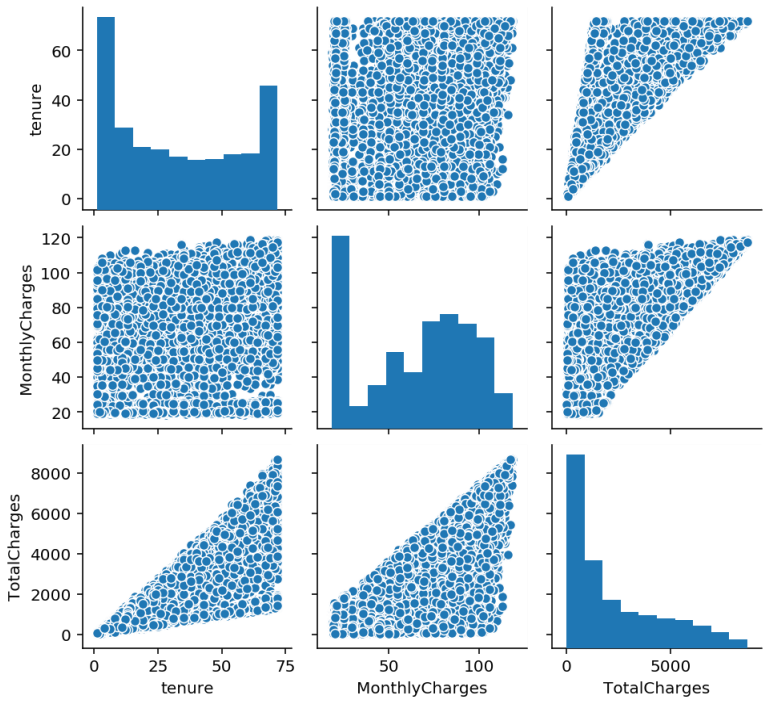
Churn by Phone Services:



Churn by Additional Offered Services:



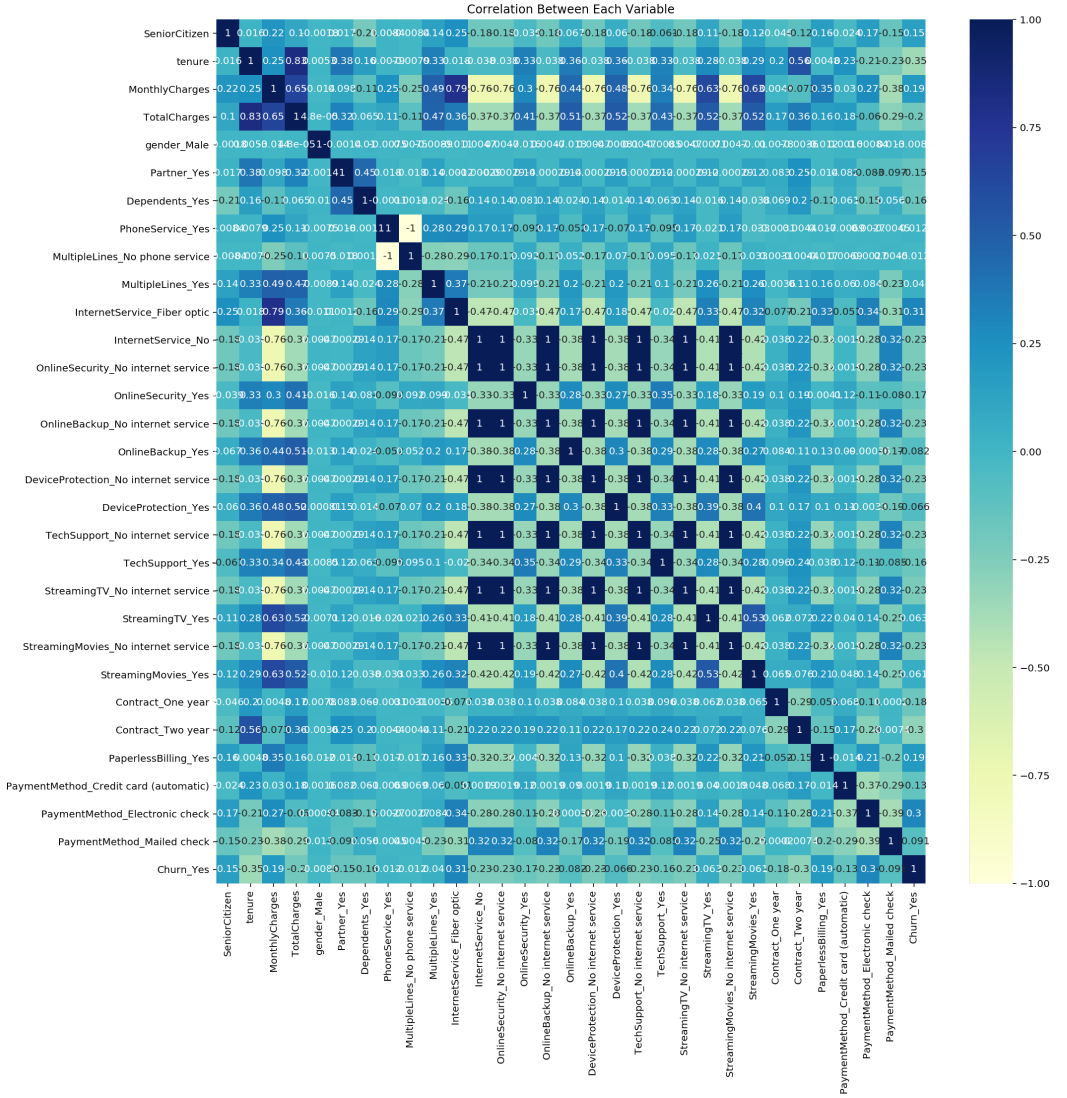
Churn vs Numerical Features:



From the bar plots, we found many interesting pieces of information from our initial analysis of the dataset such as:

- The churn percentages are almost equal for males and females, but higher if the customer is not a senior citizen.
- Customers who have a partner and dependents have a lower churn rate than those without a partner and dependents.
- The churn rate is higher for customers who opt to have phone services and paperless billing.
- Customers that don't take the additional offered services have a churn percentage that is a lot higher than those that do take the service.
- Customers that don't have internet service have the lowest churn rate.
- As tenure and total charges increase, the churn rate decreases.

Correlation Heat Map:

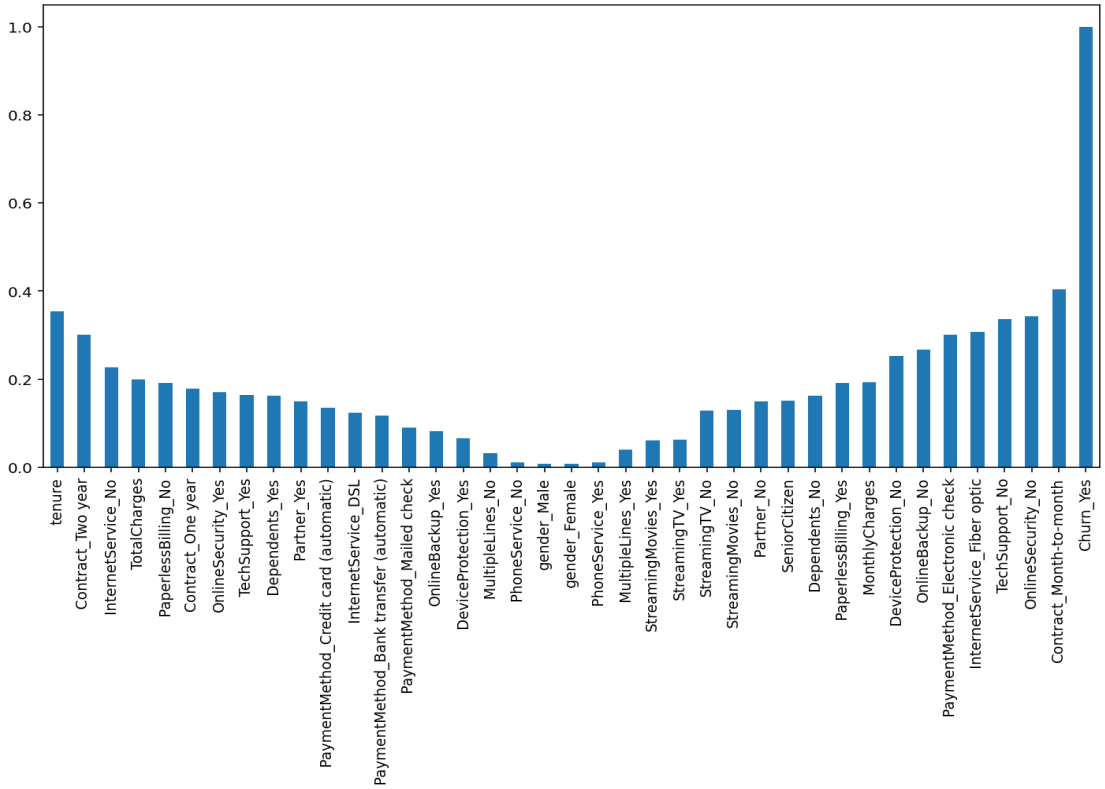


We noticed high correlations between various terms in the heatmap and identified all the attributes that have a correlation higher than 0.95 among each other:

['MultipleLines_No phone service',
 'OnlineSecurity_No internet service',
 'OnlineBackup_No internet service',
 'DeviceProtection_No internet service',
 'TechSupport_No internet service',
 'StreamingTV_No internet service',
 'StreamingMovies_No internet service']

Dropping these terms provides a list of features that better and uniquely describe Churn.

Correlation Bar Plot of Churn vs Features:



From the correlation bar plot, it is observed that customers with month-to-month contracts, high tenure, two-year contracts, no online security or no technical support tend to churn more. Features such as gender, phone service, multiplelines_no are least related to Churn. We need to keep checking for these features that mostly cause Churn when

performing the models as well in order to recommend to the Telecom company the areas they need to work on and how they can retain customers.

Feature Engineering

- Recognized the datatype of each column
- Changed the Object type columns (Target Charges) to Float
- Searched for null values in the dataset and remove related rows (9 rows in our case)
- Divided the dataframe into feature data and target
- Identified all categorical columns and created dummy variables for them

We also performed feature choice techniques to extract key features for evaluation. For instance, this was based on different services that the customer signed up for or demographic information to see which areas or types of customers would churn. We also removed unnecessary feature columns that wouldn't be useful for our analysis, such as the 'customerID' and 'const' columns.

```
telcoData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
```

Selected Features

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7032 entries, 0 to 7042
Data columns (total 32 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   SeniorCitizen                          7032 non-null   int64
 1   tenure                                 7032 non-null   float64
 2   MonthlyCharges                         7032 non-null   float64
 3   TotalCharges                           7032 non-null   float64
 4   gender_Male                            7032 non-null   uint8
 5   Partner_Yes                            7032 non-null   uint8
 6   Dependents_Yes                         7032 non-null   uint8
 7   PhoneService_Yes                       7032 non-null   uint8
 8   MultipleLines_No phone service         7032 non-null   uint8
 9   MultipleLines_Yes                      7032 non-null   uint8
 10  InternetService_Fiber optic            7032 non-null   uint8
 11  InternetService_No                     7032 non-null   uint8
 12  OnlineSecurity_No internet service     7032 non-null   uint8
 13  OnlineSecurity_Yes                     7032 non-null   uint8
 14  OnlineBackup_No internet service       7032 non-null   uint8
 15  OnlineBackup_Yes                       7032 non-null   uint8
 16  DeviceProtection_No internet service   7032 non-null   uint8
 17  DeviceProtection_Yes                   7032 non-null   uint8
 18  TechSupport_No internet service        7032 non-null   uint8
 19  TechSupport_Yes                        7032 non-null   uint8
 20  StreamingTV_No internet service        7032 non-null   uint8
 21  StreamingTV_Yes                        7032 non-null   uint8
 22  StreamingMovies_No internet service    7032 non-null   uint8
 23  StreamingMovies_Yes                    7032 non-null   uint8
 24  Contract_One year                      7032 non-null   uint8
 25  Contract_Two year                      7032 non-null   uint8
 26  PaperlessBilling_Yes                   7032 non-null   uint8
 27  PaymentMethod_Credit card (automatic)  7032 non-null   uint8
 28  PaymentMethod_Electronic check         7032 non-null   uint8
 29  PaymentMethod_Mailed check             7032 non-null   uint8
 30  Churn_Yes                              7032 non-null   uint8
```

Some of the selected features have the word Yes or No, because most of the categorical columns are categorized in these two Yes or No classes. If the customer has this feature, the feature selection selects the feature with Yes, or No otherwise. Then, we can use this to see whether or not a customer has said feature will affect the churn.

Modeling Techniques

The overall steps that were followed for applying the models were to split the training and testing data set, select the classifier algorithm, construct the training model, evaluate the training model, and then develop and evaluate the testing model. Finally, we compared all the models to see which one provided the best performance accuracy for the dataset. Each model that was used will be detailed in the following sections.

Model Selection

We explored various classification techniques and selected the hyperparameters for better accuracy in predicting what factors will lead to customer loss. To process and analyze our data sets, we used the following models:

- Lasso Cross Validation
- Random Forest Classifier
- K-Nearest Neighbor
- Support Vector Classifier
- XGBoost
- Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes

We are utilizing all of these models because one of our goals includes finding the most appropriate algorithm model by comparing accuracy scores for the current dataset to solve the Telecom business situation.

Resampling Techniques

Before training our models, we had to apply some resampling techniques to the dataset since some of the data were imbalanced due to some categories having more

data than others. For example, the Churn column has 73% of the data belonging to class ‘No’ and 27% belonging to class ‘Yes’.

- We use Stratified k-fold Cross Validation so as to make sure that the ratio of both classes’ data is even across various folds.
- Also, we over sampled the minority, such as Churn_Yes, and used the Synthetic Minority Over-Sampling Technique (SMOTE). After performing this procedure, the data belonging to each class was 50-50% signifying a balance.

Summary of Model Analysis

The models started learning from the training data and then were applied to the testing data to obtain a prediction. After the model was exercised ten times as per the k=10 stratified sampling, an output of ten ROC-AUC numbers were displayed that described the accuracy of the model. Then, the best percentage was chosen. We compared the predictions and ROC-AUC scores obtained from each modeling technique to find the best model. After this process, we analyzed the impact of various attributes over customer churn.

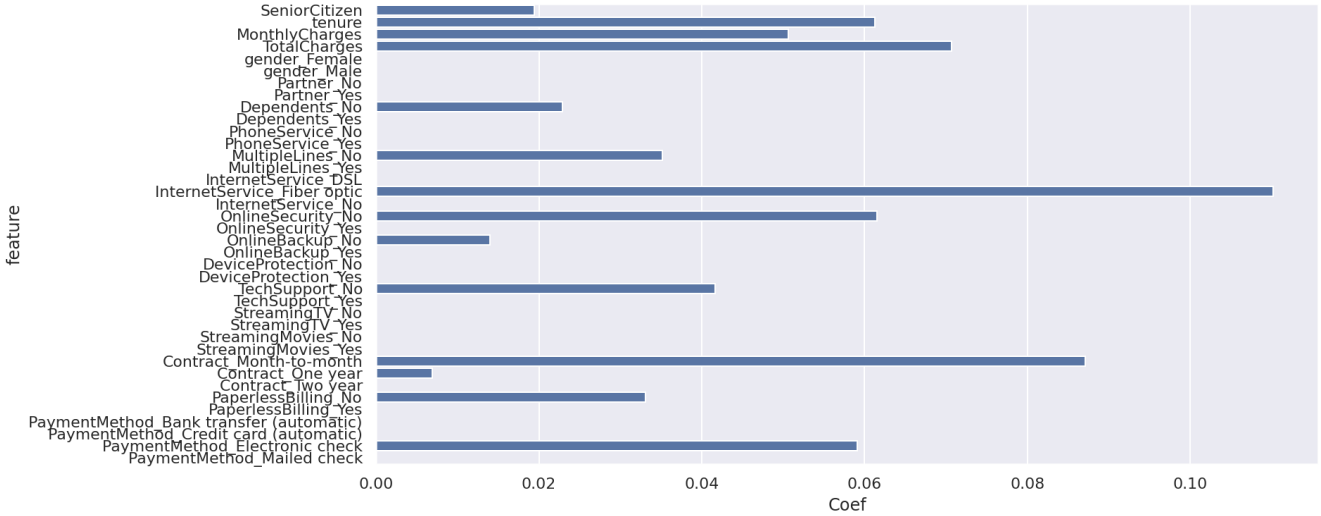
1. Lasso regression:

Lasso (Least Absolute Shrinkage and Selection Operator) is a linear regression method that adopts L1-regularization. The adoption of L1 regularization will make the feature weight acquired by part to be 0, so as to achieve the purpose of sparse and feature Selection.Different from ridge regression, the first order penalty function of absolute value is used to replace the second order function of sum of squares. Although the form is slightly different, the results are quite different. In LASSO, when lambda is small, some of the coefficients go to zero and the ridge regression makes it difficult to get a coefficient to zero exactly.

In the mathematical definition, the target equation to be calculated by lasso is:

$$\min_{\beta_0,\beta}\left\{\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j|\le t.$$[2]

By implementing lasso regression, we found that lasso regression has a lower accuracy than other machine learning algorithms, with an accuracy of 82.96% in our experiment.



The diagram above shows the correlation relationship between target feature "Churn "with other features. By comparing it with other feature data, it shows that Internet Service Fiber Optics, Monthly/Total charges, tenure, Month-to-Month Contract, and electronic check payments contribute relatively high effects on the target.

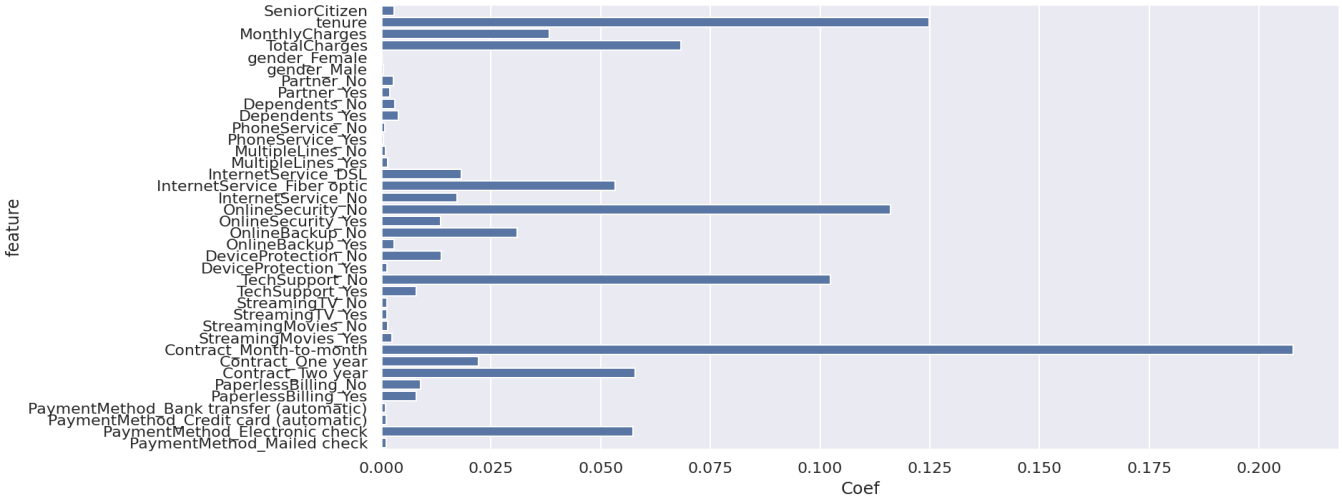
2. Random Forest Classifier:

We applied the Random Forest Classifier on the dataset which is a Bagging technique. The algorithm was used with various hyper-parameter settings, however the combination that yielded best result of AUC-Score 83.15% is:

n_estimators=200

max_depth=4

max_features=6



Random Forest Classifier chooses Month-to-Month Contract, tenure, No-Online Security, No Technical Support as Important Features that are highly related to the Target.

3. K-Nearest Neighbors

KNN arithmetic is a common method of machine supervised learning. For a given test sample, the K training samples closest to the training set were found, and then the future prediction was made based on the information of these training samples.

In KNN, the distance between objects is calculated as the non-similarity index between objects to avoid the matching problem between objects. In this case, Euclidean distance or Manhattan distance are generally used:

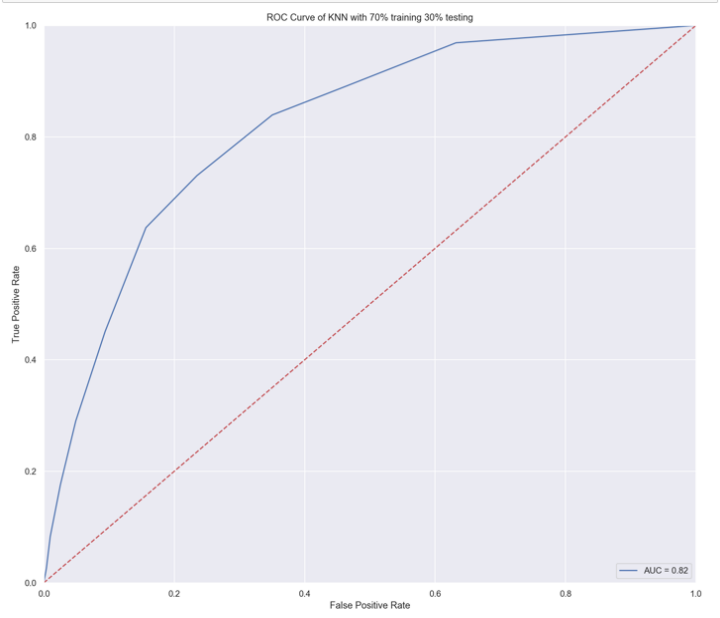
$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}\ , \quad d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k-y_k|}$$

In the experiment, the training set and the test set are divided by using two different resampling methods (validation and K-fold resampling method) and using a training set and test set to estimate the accuracy of the model by KNN algorithm. Then taking the ROC curve in both cases to compare the performance of each algorithm.

1. Split the data by validation model.

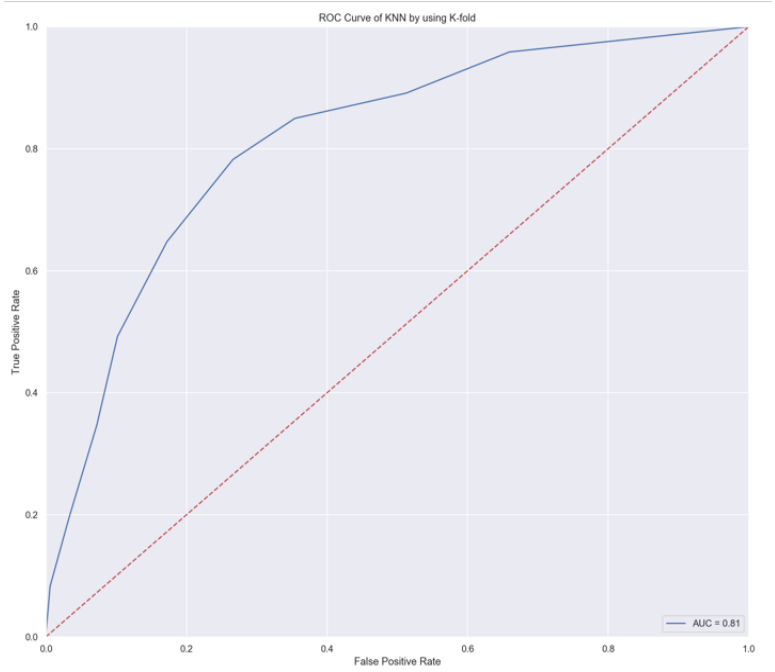
Confusion Matrix

	True	False
True	1389	172
False	290	259



2. Split data by using the K-fold model, with K = 10.

	True	False
True	459	52
False	98	95



The result shows that the difference between the two algorithms was not particularly significant, with AUC-ROC of 80.63%. Compared with other algorithms, KNN is not the most accurate. In the future design, in order to improve the performance of the KNN algorithms, the training and testing data set may be need re-adjusted by other statistical algorithms such as Multiple dimensional scaling (MDS) or PCA.

k=17 works best in classifying customers to Churn or not. High value of k signifies a simpler model, i.e. the model defines a more linear decision boundary.

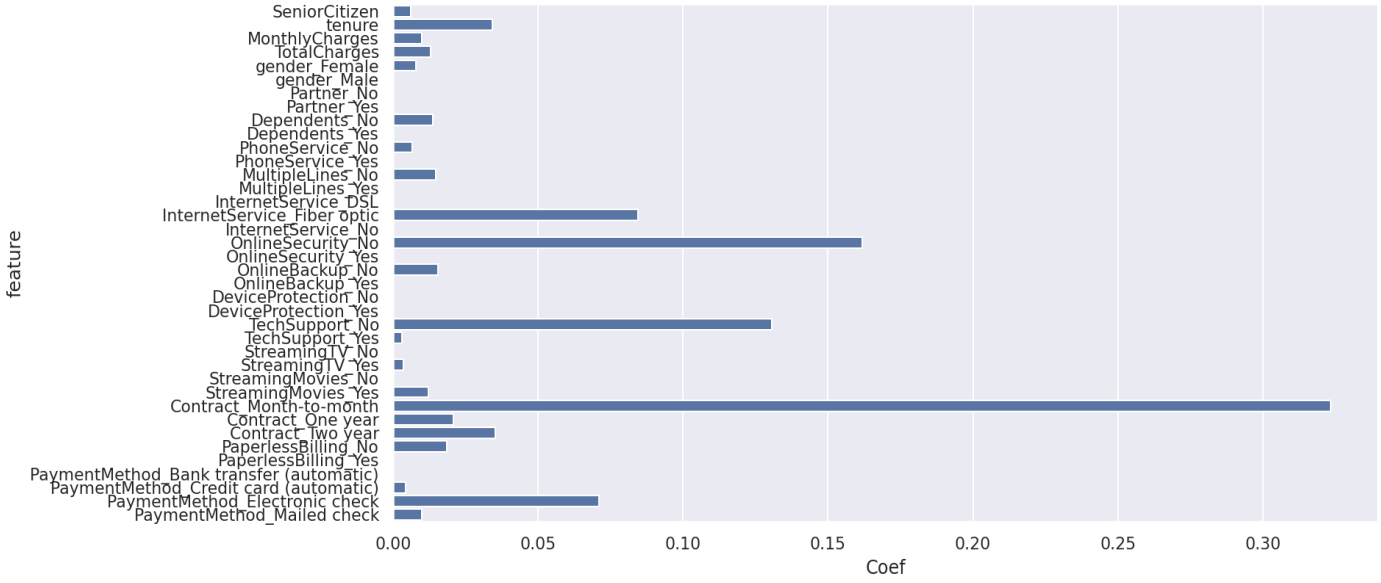
4. Support Vector Classifier

Support Vector Machine is a very efficient algorithm for classification. It can identify linear and non-linear decision boundaries with good accuracy.
We implemented SVC to classify two classes on Churn dataset. The model yielded better results for the 'Linear' kernel, which signifies that the decision boundary is somewhat linear for two classes.

Model yielded a 10-fold Validation AUC Score= 71.94%
And Test AUC Score= 70.23%

5. XGBoost

Boosting algorithms generally prove to perform better because of their ensembling mechanism. XGBoost is a Boosting technique that we used to predict Churn for our model. XGBoost used various Decision Trees to build ensemble models.
Various hyper-parameter setting that yielded the best results:
max_depth=2, max_features=6, learning_rate =0.09
Model 10-fold Validation AUC Score = 85.08%
Test AUC-Score = 83.75%



XGBoost identifies Contract Month-to-Month, No Online Security, No Technical Support as top three features that cause Churn.

6. Logistic Regression

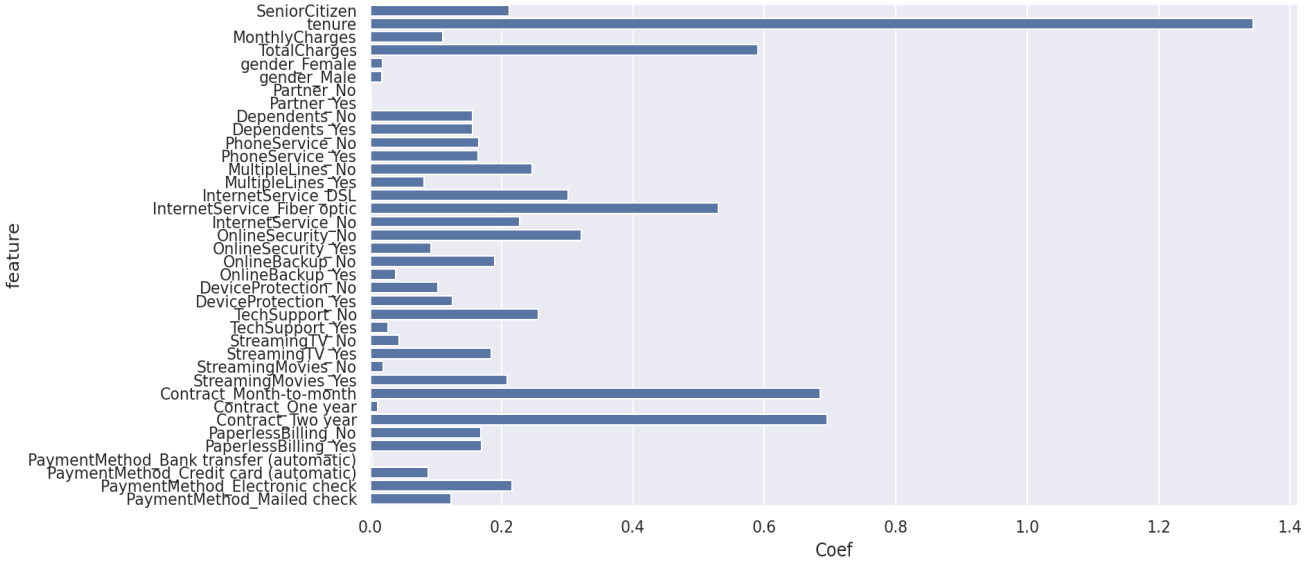
Logistic Regression is a regression predictive analysis method usually used for two-class classification problems. It is used to describe the data and explain the relationship between a dependent categorical variable and one or more nominal, ordinal, interval, or ratio-level independent variables. In this case, logistic regression is an appropriate method to use because the target, Churn, is a categorical variable represented by two classes, 0's and 1's for no and yes, respectively.

Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$\text{logit}(p) = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$

for i = 1…n .

The logistic model that we used utilized this concept and obtained ten ROC-AUC scores from each run of the model. The best test score was 83.59%. Although this is not a very high score overall, it gave an average performance compared to the other models that we tried.



The plot above that shows the coefficients of each variable displays how much of an impact each feature has on Churn. According to logistic regression, tenure, two year contracts, month-to-month contract, total charges and no internet service are the top variables that affect Churn the most. On the other hand, customers with a partner, device protection, and male genders have the least effect on Churn.

7. Linear Discriminant Analysis

Linear Discriminant Analysis, or LDA is a linear classification predictive modeling technique used for dimensionality reduction. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order to avoid overfitting

as well as reduce computational costs. It is especially useful to see the separation between multiple classes in the Churn data.

To use LDA, estimates are obtained for π_k and $f_k(X)$ that are used to approximate the Bayes' theorem, which classifies an observation to the class 'k' for which $p_k(X)$ is largest:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

LDA is calculated by the following equation:

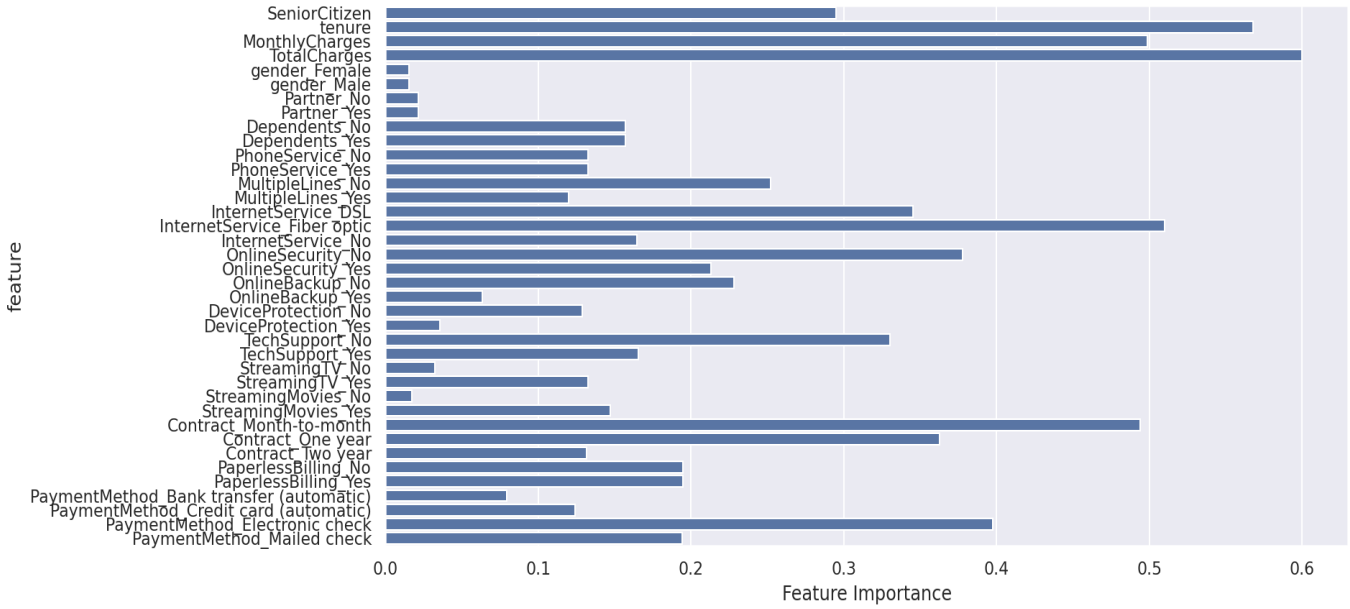
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log(\pi_k)$$

$$\hat{G}(x) = \arg \max_k \delta_k(x) \quad \text{--> Classification Rule}$$

Given any x value, the numbers are plugged into the formula and the 'k' class that maximizes the value is picked. Since the number of classes is small, and often only two classes, an exhaustive search over the classes is effective. The covariance matrix, Σ, is identical, so the quadratic term is dropped and LDA provides a linear boundary.

In addition, there were some collinearity issues in the data so we also had to compute the Variance Inflation Factor (VIF) for each variable, meaning some independent variables (the features) could influence other independent variables. Variables that had a VIF above 5 were eliminated. These variables include OnlineSecurity_No internet service, OnlineBackup_No internet service, DeviceProtection_No internet service, TechSupport_No internet service, StreamingTV_No internet service. The variables with the name 'No internet service' were the same features that were dropped when looking at the correlation plots.

After adjusting the model, the best test ROC-AUC score was 82.99%. This model performance was on the lower side, but was still in the 80% test score range compared to other models.



The correlation plot shows that, according to LDA, the top five features that impact Churn the most are customers that are Total Charges, Tenure, Monthly Charges, Fiber optic Internet Service, Contract Month-to-Month.

8. Quadratic Discriminant Analysis

Quadratic Discriminant Analysis, or QDA is very similar to LDA except it introduces a second degree, allowing for more complex terms of classification. QDA uses a quadratic decision surface, calculated by a multivariate Gaussian Distribution, to separate input vectors into classes. In addition, the covariance matrix is assumed to be different for each class. Because QDA allows for more flexibility in the covariance matrix, it can fit the data better than LDA.

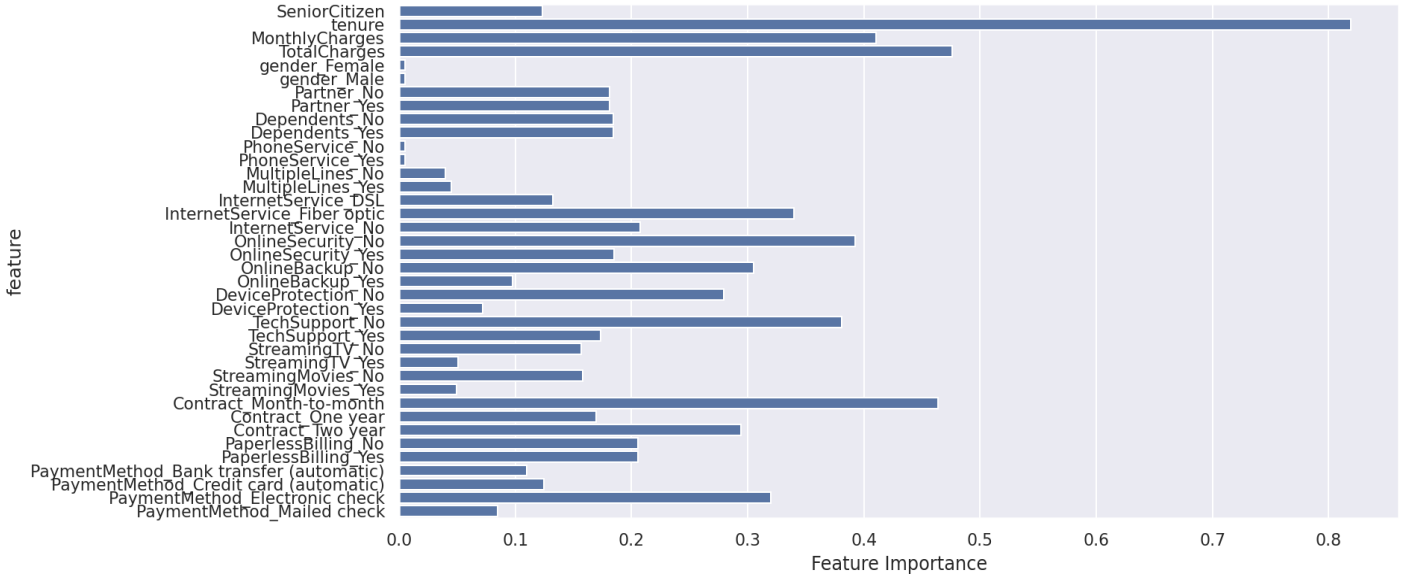
Mathematically, the QDA function is defined as:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

The quadratic discriminant function is similar to the linear discriminant function except that the covariance matrix, Σ_k is not identical and the quadratic terms cannot be discarded. Thus, the function is quadratic and contains second order terms. The classification rule is equivalent to the one in LDA as well. The only difference is the class

'k' that maximizes the quadratic discriminant function is found and the decision boundaries are quadratic equations.

The best test ROC-AUC score was 78.37%. This model performance was on the lower side, and was in the 70% test score range. This may be because the data is not quadratic and there are many features or parameters to estimate. However, it was still interesting to compare QDA with other models. QDA gave a lower performance compared to the other models that we tried.



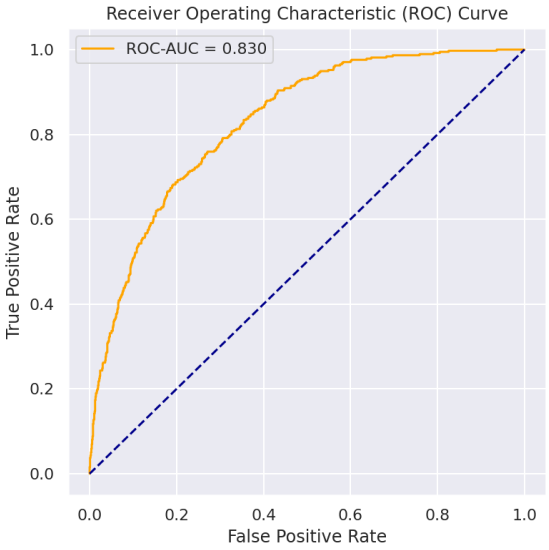
The correlation plot shows that according to QDA, tenure, Total Charges, Contract Month-to-Month, Fiber Optic Internet Service, No Technical Support are few of the features that mostly cause Churn.

9. Naive Bayes

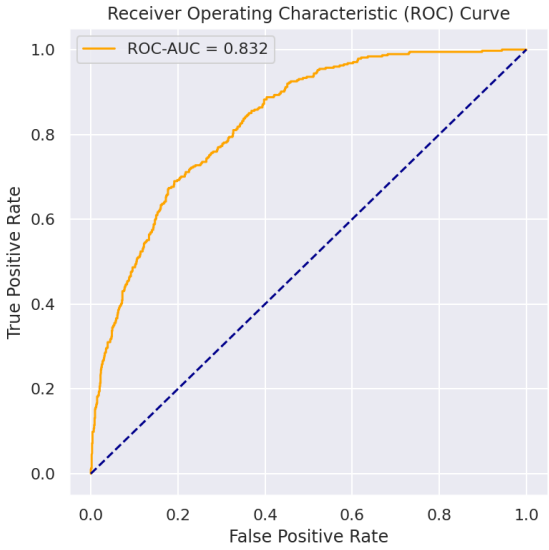
Naive Bayes Classifier is a technique based on Bayes Theorem. It holds the assumption that the presence of any particular feature in a class is independent and unrelated to the presence of other features of the class.

Naive Bayes Validation AUC-Score=83.21%
Test AUC-Score=81.85%

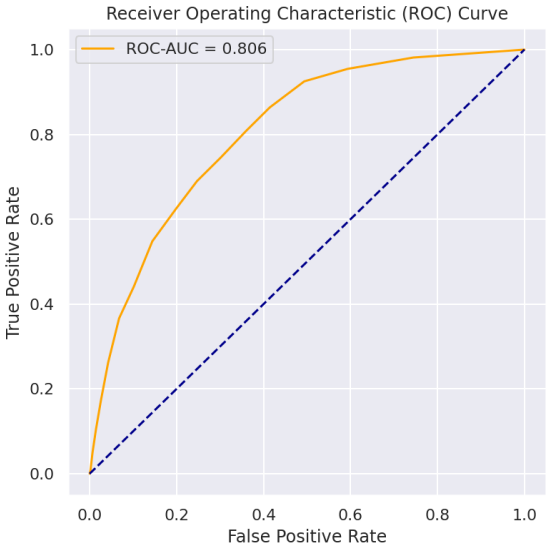
Model Performance:



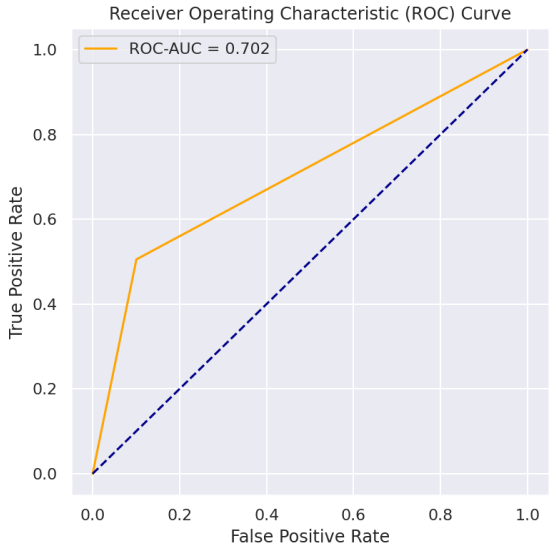
LASSO Regression



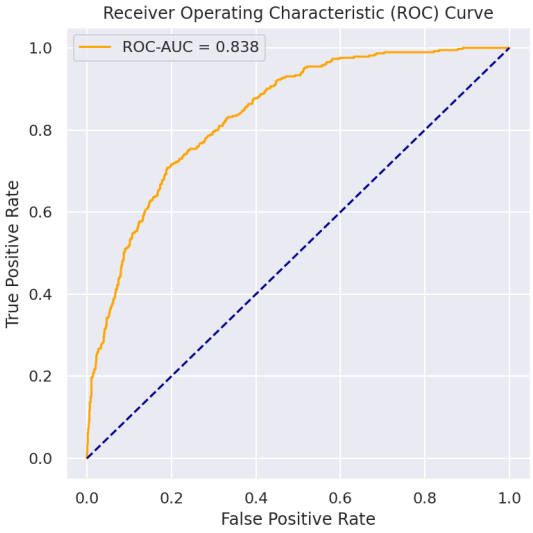
Random Forest Classifier



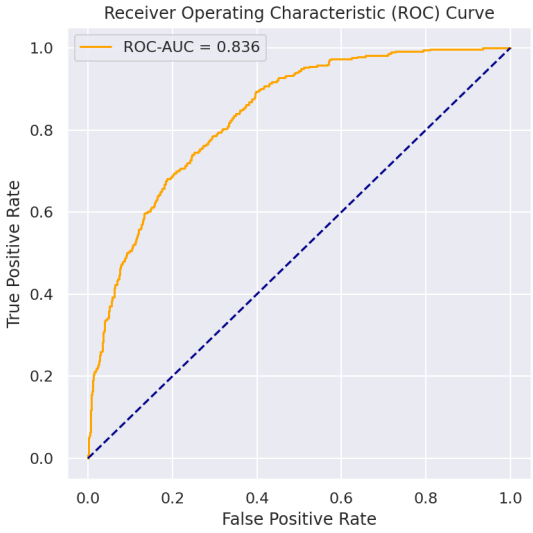
kNN(k=17)



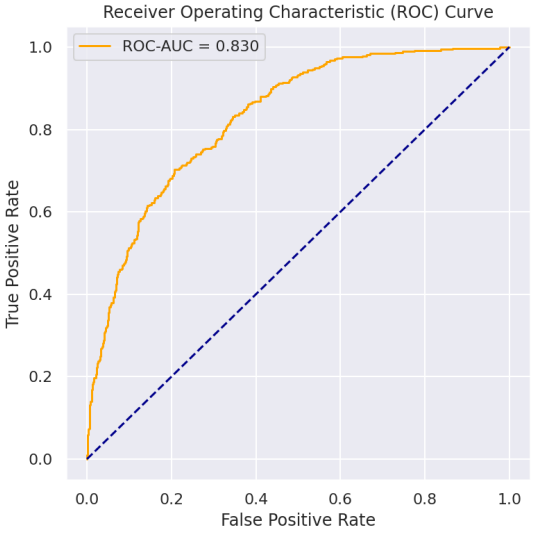
Support Vector Classifier



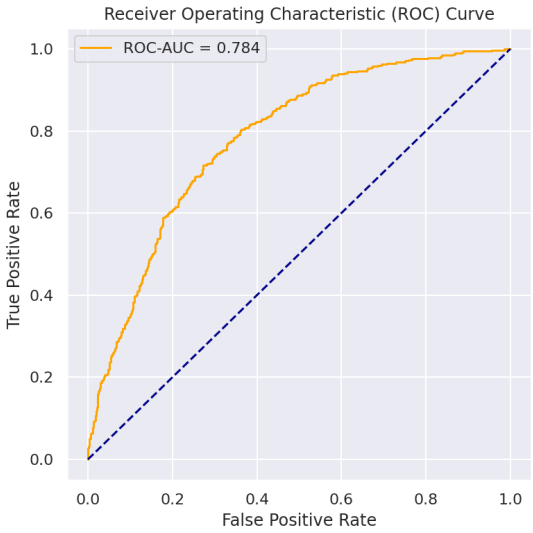
XGBoost



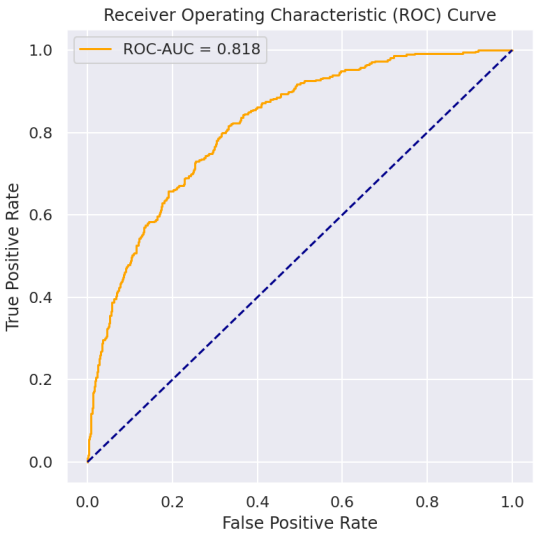
Logistic Regression



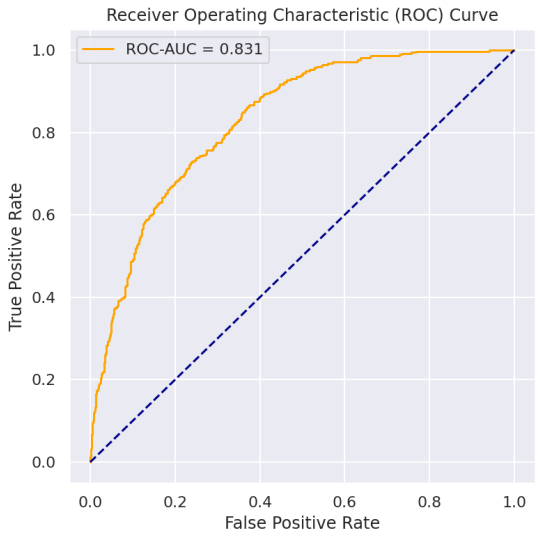
LDA



QDA



Naive Bayes



Ensemble Model

Model Evaluation

Since the dataset is unbalanced, we evaluated the result using the Confusion Matrix and calculated the sensitivity and specificity of our result. Then, we compared the performance of the various algorithms using the AUC score and ROC curve.

The model that had the highest ROC-AUC score was XGBoost with 83.75%. This high result occurred because XGBoost is an ensemble technique to boost weights that have higher importance and build ensemble trees to produce the output. This technique reduces the model bias and variance, subsequently tending to produce better results for classification.

Summary of Results

Using the results obtained from the model evaluation, we created a table to organize the model accuracies, ranked from highest to lowest, and important selected features:

Model	ROC-AUC Score	Features Highly Correlated with Churn
XGBoost	**0.8375**	Contract_Month-to-Month, Online Security_No, TechSupport_No, InternetService_FiberOptic, PaymentMethod_Electronic check
Logistic Regression	0.8359	Tenure, Contract_TwoYear, Contract_Month-to-Month, TotalCharges, InternetService_FiberOptic
Random Forest	0.8315	Contract_Month-to-Month, Tenure, Online Security_No, TechSupport_No, TotalCharges
LDA	0.8299	TotalCharges, Tenure, InternetService_FiberOptic, Contract_Month-to-Month, PaymentMethod_Electronic check
Lasso	0.8296	Online Security_No, Contract_OneYear, gender_Female, MonthlyCharges, Online Security_Yes
Naive Bayes	0.8185	N/A
KNN	0.8063	N/A
QDA	0.7837	Tenure, TotalCharges, Contract_Month-to-Month, MonthlyCharges, Online Security_No
SVC	0.7023	N/A

We found that the features that affected customer churn from most to least were:

- Month-to-Month Contracts
- Longer tenure
- Higher Total Charges
- No Online Security
- Fiber Optic Internet Service
- No Technical Support
- Payment Method: Electronic Check

Month-to-month contracts could influence churn because customers may not be satisfied in the first month that they sign up, but have to keep going with the service because of their contract. Longer tenure could be a reason because customers may want to try other services or want change after they are subscribed to the same service for a long time. In addition, higher total charges could cause customers to leave as they seek better deals at other services.

Recommendations for other services provided by the company include improving online security, fiber optic internet services, and technical services as well as securing

electronic check payments. This would ensure higher customer satisfaction. In addition, the company can offer discounts and promotions that would retain customer loyalty.

Conclusion

For the future, if we had more time to expand on the Telecom customer churn problem, we could try Principal Component Analysis (PCA). PCA is a technique for reducing the dimensionality of datasets by increasing interpretability and minimizing information loss. Its process is to create new uncorrelated variables that successively maximize variance. Models can be implemented using the PCA results, which can potentially improve their performance. In addition, we could also use the Neural Networks model, a set of algorithms that can recognize patterns, to automatically learn data features and yield better results.

Overall, we learned many important concepts from doing this project. This includes applying theoretical knowledge and models that we learned in class, as well as models that we researched, to a practical business problem. We were able to perform data cleaning to remove inconsistencies, feature analysis to see which features would be important, and provide recommendations that could help the business become even more successful. We also understood when and how to apply resampling techniques, different classification models, as well as bagging and boosting to achieve better accuracy. We also had the opportunity to extract the most important features and best fit a model to make predictions of customer churn.

References

"What Is Logistic Regression?" *Statistics Solutions*, www.statisticssolutions.com/what-is-logistic-regression/.

"Sklearn.linear_model.Lasso¶." *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.

"Linear Discriminant Analysis." *Dr. Sebastian Raschka*, 3 Aug. 2014, sebastianraschka.com/Articles/2014_python_lda.html.

"9.2.3 - Optimal Classification: STAT 508." *PennState: Statistics Online Courses*, online.stat.psu.edu/stat508/lesson/9/9.2/9.2.3.

orange1orange1 45711 gold badge33 silver badges99 bronze badges, et al. "How to Systematically Remove Collinear Variables in Python?" *Cross Validated*, 1 Feb. 1965, stats.stackexchange.com/questions/155028/how-to-systematically-remove-collinear-variables-in-python.

9.2.8 - Quadratic Discriminant Analysis (QDA), online.stat.psu.edu/stat508/book/export/html/696.

Jolliffe, Ian T., et al. "Principal Component Analysis: a Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 13 Apr. 2016, royalsocietypublishing.org/doi/10.1098/rsta.2015.0202.

Jnyh. "JNYH/Project-McNulty." *GitHub*, 8 Oct. 2019, github.com/JNYH/Project-McNulty/blob/master/Churn_Prediction.ipynb.

"Sklearn.linear_model.Lasso¶." *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.

蓝蝶.css-1cd9gw4{margin-left:.3em;} . "电信用户流失分析-Python." 知乎专栏, zhuanlan.zhihu.com/p/88439986.

默言.css-1cd9gw4{margin-left:.3em;} "电信客户流失预测." 知乎专栏, zhuanlan.zhihu.com/p/58414385.

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/