Customer Churn Problem Definition

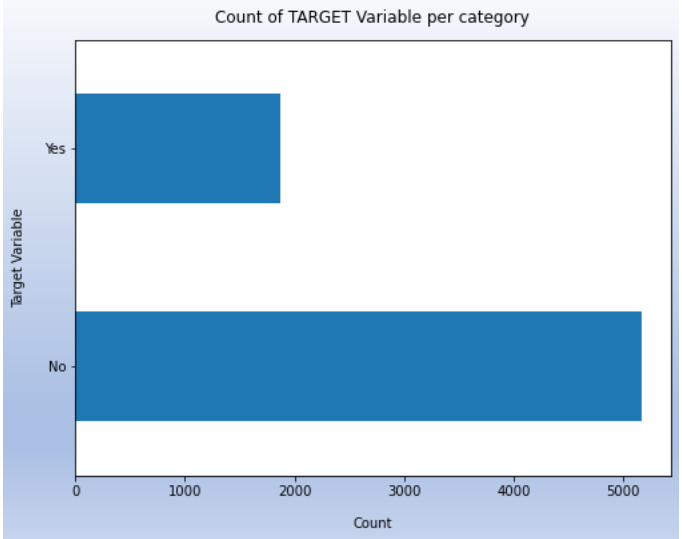
Problem Definition

Customer churn is when a company's customers stop doing business with the company. It is a critical metric for any business because it costs more to acquire new customers than to retain existing customers.

Customer retention can be achieved with good customer service and products. The best way to avoid it is to truly know its customers. Here, we will examine customer data from IBM Sample Data Sets with the aim of building a prediction model that predicts customer churn.

Data Analysis and EDA

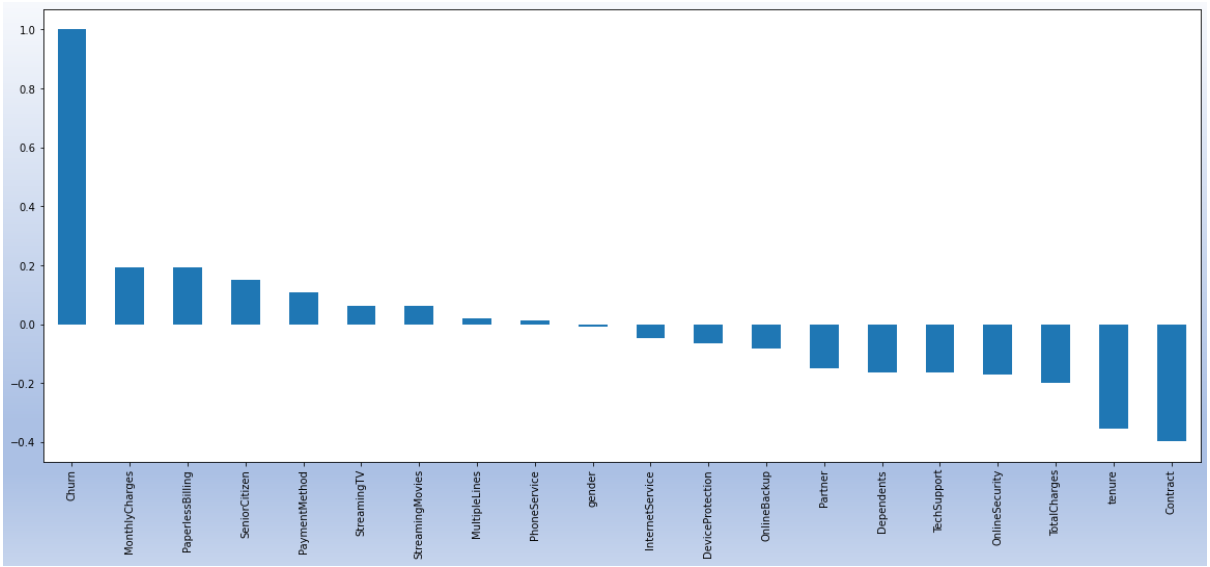
- Identifying the ratio of churners



The churners are less as compared to churners.

- There are 7043 rows and 21 columns in the dataset.
- On checking the data types of the column, some are floats and some are objects. By using the .describe() we get insights about mean, standard deviation and percentile values of the variables. As Senior citizen is a categorical variable so the 25%,50%,70% distribution does not seem to be proper. Also 75% customers have tenure less than 55 months. Average Monthly Charges are USD 64.76 whereas 25% customers pay more than USD 89.85 per month.
- On checking the data types of the column, some are floats and some are objects. Though Total Charges column should be float or integer but by using .info() the data type is an object which means there seems to be something to looked into to convert it to float or integer.
- Removed the NaNs and spaces from the total charges column.

- After analyzing the data, found that some columns like “Online Backup”, “Streaming Movies”, “Tech Support” etc have three categories as “NO”, “YES” & “No Internet Service”. As No Internet Service” itself means No, so replaced it with No for all the columns.
- Count plot for the various categorical column was plotted and information derived says, No Online Security, No Tech Support, No Device Protection, No Online Backup, No Streaming Movies and No Streaming Tv are high churners.
- Also Non Senior Citizens are high churners, having partners churn more and no dependents churn less.
- Scatter plot between Monthly Charges and Total Charges says Total Charges increase as Monthly Charges increase which is obvious.
- Scatter plot between Total Charges and tenure says, tenure of 0 to 10 have paid maximum amount of Total Charges.
- Heatmap analysis to find out the correlation between the independent and dependent variable was plotted.
    - The heatmap revels, Monthly Charges and Paperless Billing have same correlation. Same with Streaming TV and Streaming Movies, Dependants and Tech Support.
    - High Churn seen in case of Month to month contracts**,** No online security**,** No Tech support**,** First year of subscription and Fibre Optics Internet
    - Low Churn is seen in case of Long term contracts**,** Subscriptions without internet service**.**
    - Factors like Gender**,** Availability of PhoneService have alomost no impact on Churn



- The similarly correlated columns were dropped but the accuracy score did not seem to increase a lot. So, considered not to drop the columns which would rather lead to data loss. Skewness was checked using skew().
- Though Total Charges was skewed, did not apply any transformation technique as it is negatively correlated with our target variable “churn” and applying transformation methods to negative values will result in NaN values. On checking for outliers, the dataset looked to be very clean with no outliers. There existed class imbalance. Since it is a classification problem the class needs to be balanced before training the model to avoid biasedness. Here, the class imbalance was dealt using oversampling technique.

Building Machine Learning Models

As the dataset is imbalanced. So, used SMOTE oversampling technique to balance the dataset.

I used accuracy as metrics to measure the model, as it is a classification model. Also checked recall, precision & f1 score.

Found out the best Random state to be 85.

Below are the scores for the respective classifiers-

Logistic Regression	83
Decision Tree Classifier	82
Random Forest Classifier	85
SVC	85

The least difference in accuracy and cv score is for Random Forest Classifier. So, we found our best model as Random Forest Classifier.

With RF Classifier, also we are able to get quite good results, in fact better than Decision Tree.

EDA concluding Remarks

- Electronic check medium are the highest churners
- Monthly customers are more likely to churn because of no contract terms.
- No Online security, No Tech Support category are high churners. Non senior Citizens are high churners