Project Report

Customer Churn Prediction System

1. Problem Statement:

The ability to accurately predict the customers who have a higher probability to churn is critical in customer retention. Careful analysis of available data on customers can reveal any trends or potential issues that needs to be addressed. Using this, we can also build machine learning models that can predict probability of customer churn. This helps an organization to devise suitable retention measures as churn rate is an important factor that directly affects the revenue.

2. Data

Telco is a fictional telecommunications company that provided home phone and Internet services to customers in California in Q3. The company is looking to identify any interesting patterns among the customers who left the company in Q3 so that they can implement strategies to retain more customers in the next quarter. The following data are used in the analysis and modelling.

- IBM Accelerator catalog This dataset includes demographics information about customer such as gender, dependents, etc. as well as specific information on type of services the customer have with the company. The target column contains the churn info. It has the value 1 if the customer has left the company. Otherwise, it has the value of 0.
- United States zip codes database This dataset maps the zip codes to county name. This helps in better management of categorical feature without losing much information on customer's geographic location

3. Method

3.1 Data Cleaning

Data Cleaning Report

Data preprocessing Report

The dataset includes 7043 observations about telecommunication customers from California. Out of the 7043 customers, 27% of the customers left the company in the end of Q3. Each observation contains various columns related to the customer and the type of services they use.

Handling of missing values:

Only the columns `Churn Reason` and `Total Charges` columns have missing values.

- `Total Charges` column has 11 missing values. This column is correlated with `Monthly Charges` and `Tenure Months` column. Hence, we can safely drop this column.
- `Churn Reason` values are missing only for those observations with `Churn Label` = No. Thus, it represents the customers that are still with the company, and it definitely makes sense for those customers to have null value for `Churn Reason` column. Thus, it is an attribute whose values cannot be obtained at the time of prediction. To avoid data leakage, we drop this column as well.

Preprocessing and feature engineering

- **Preprocessing of numerical columns:** The identified numerical columns are `Tenure Months`, `Monthly Charges`, `Total Charges`, `Churn Score` and `CLTV`. Out of these, Monthly Charges and Total Charges are highly correlated. Hence, we keep only one of these columns. Also, as all the columns are in different scales, they are normalized to be on the same scale. This prevents columns with high magnitude such as CLTV from dominating over other low magnitude columns.
- **Preprocessing of categorical columns:** The low cardinality categorical columns that are important to predict customer churn identified are `Gender`, `Senior Citizen`, `Partner`, `Dependents`, `Phone Service`, `Multiple Lines`, `Internet Service`, `Online Security`, `Online Backup`, `Device Protection`, `Tech Support`, `Streaming TV`, `Streaming Movies`, `Contract`, `Paperless Billing` and `Payment Method`. The categorical values in these columns are encoded using one-hot encoding technique.

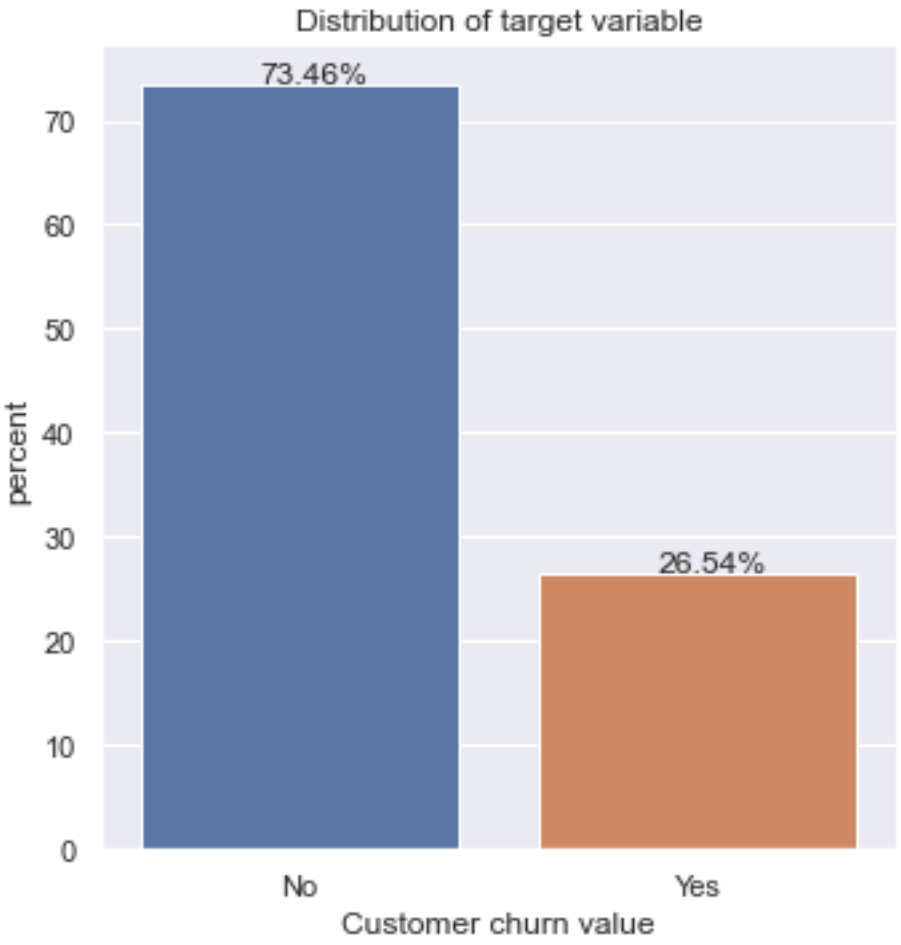
- **Handling of customer location information:** The data set includes customers from the state of California spread over 1652 unique postal codes and 1129 unique cities. As they are high cardinality values, these columns are dropped. The zip code values are mapped to corresponding county values using the United States Zip Codes database. The county values are then encoded using Label Encoder. As we can represent customer's geographic information in this way, we also delete the Latitude and Longitude columns.

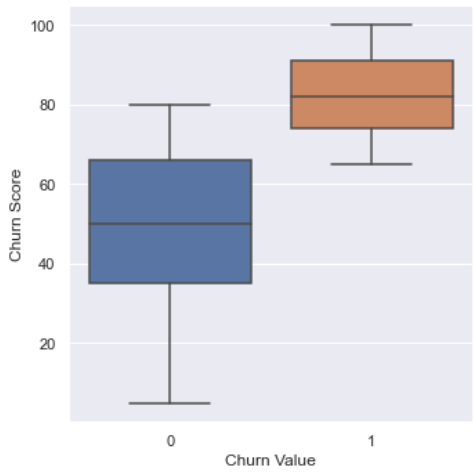
3.2 EDA

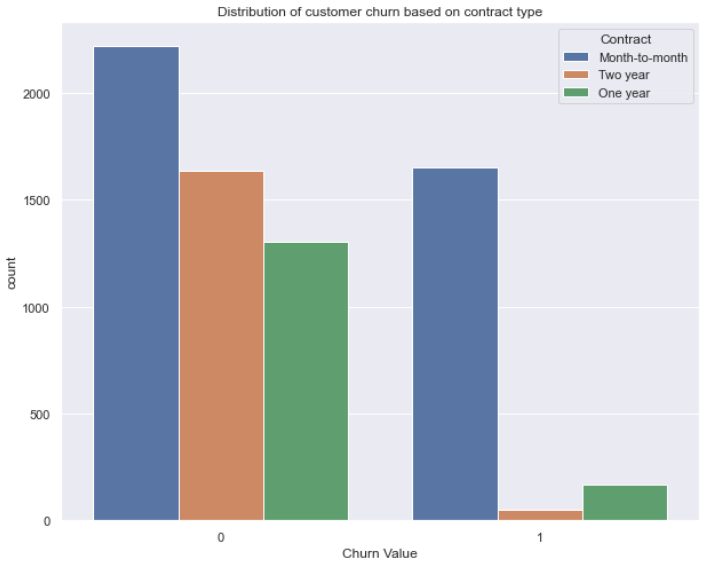
EDA Report

Analysis of target variable

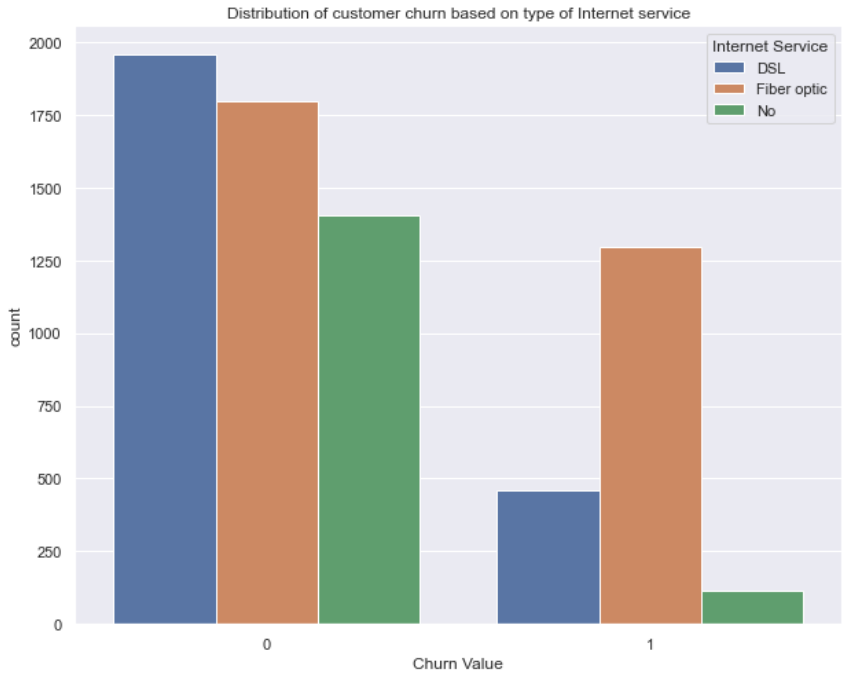
It appears that we have an imbalanced target class with approximately 27% customers with churn = 1 class and 73% customers with churn = 0 class. Also, customers who churn have relatively high churn score value or on month-to-month contract as depicted below.

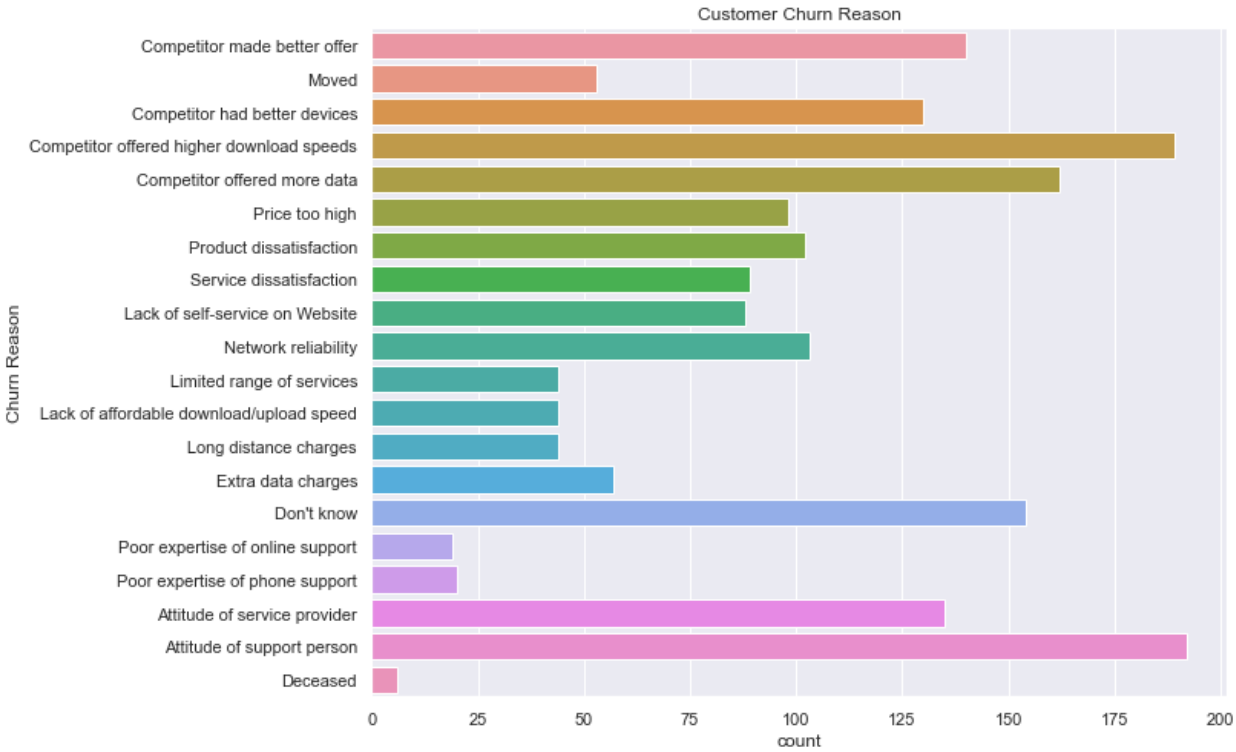

Distribution of target variable
73.46%
26.54%
percent
0
10
20
30
40
50
60
70
No
Yes
Customer churn value


Churn Score
20
40
60
80
100
0
1
Churn Value


Distribution of customer churn based on contract type
Contract
Month-to-month
Two year
One year
count
0
500
1000
1500
2000
0
1
Churn Value

Some other interesting relationships identified during EDA are shown below.

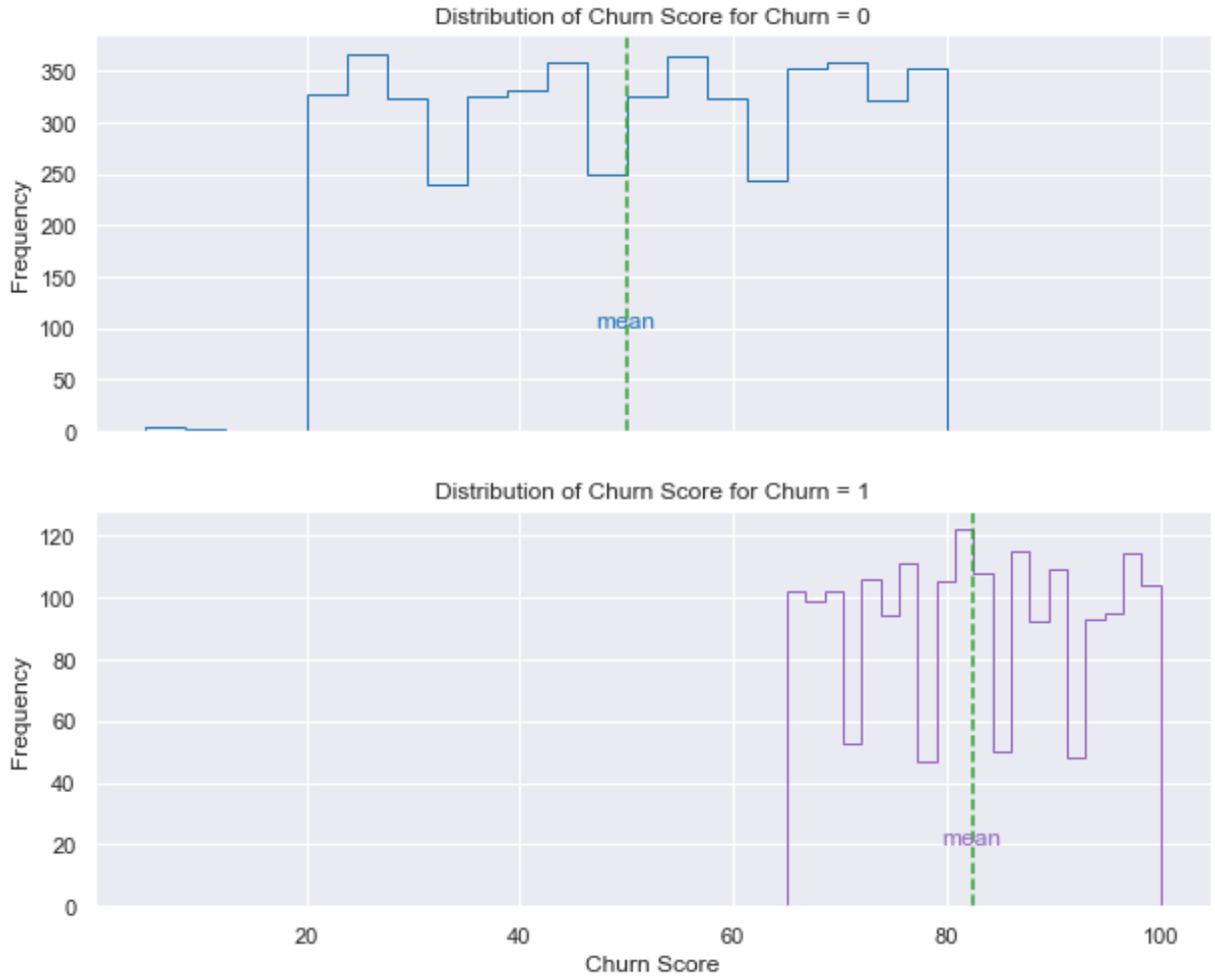




Approximately 70% of the customers who left the company had Fibre Optic Internet service. This along with above data collected from customers who left the company indicates that better internet service from competitor could be a significant factor leading to customer churn.

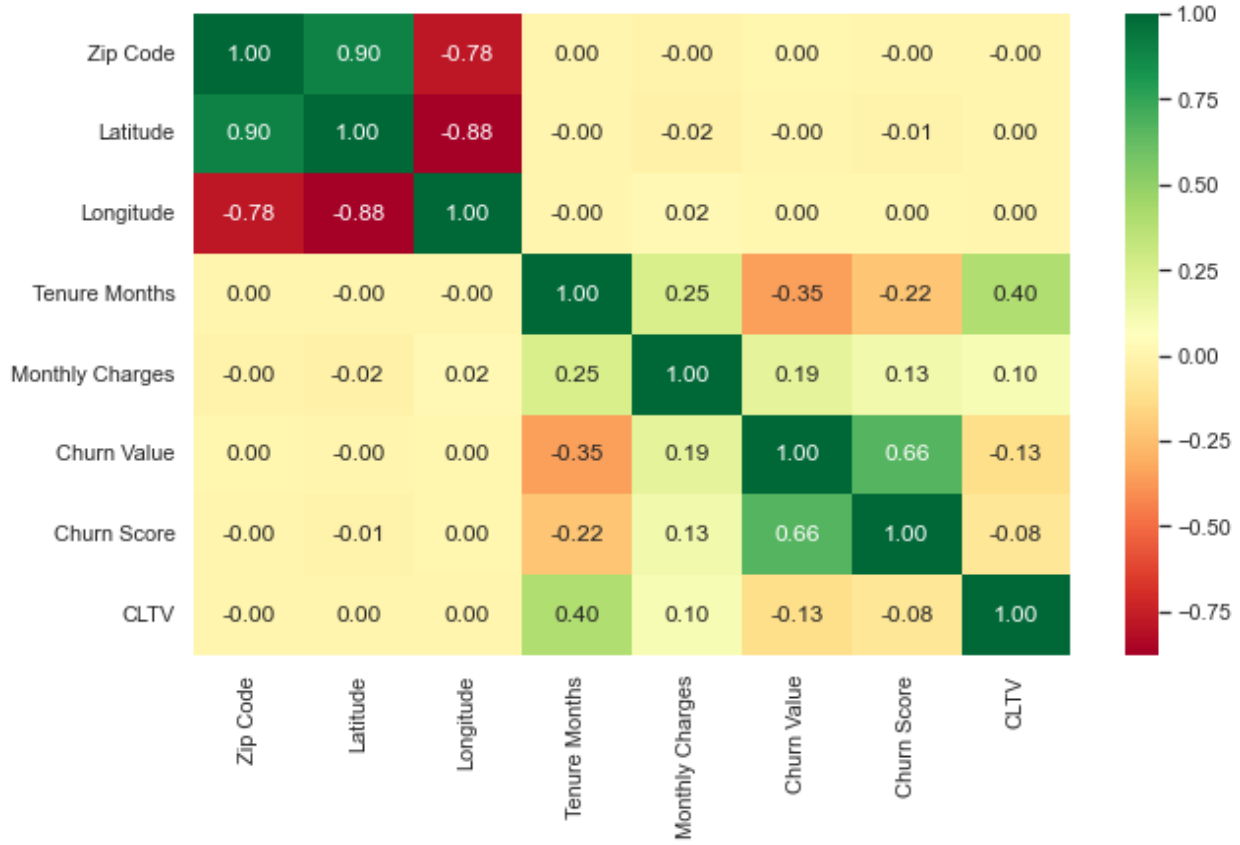
Statistical estimation of average churn score of customers who churn

To determine how useful is the feature churn score in predicting the probability of customer churn, we look at the distribution of churn score and mean churn score for customers who churn and who do not churn.



It is evident that customers who churn have higher churn score. Bootstrapping technique is applied to statistically estimate the confidence interval for the difference in mean churn score for customers who churn and customers who do not churn. The 95% confidence interval of the difference between mean churn scores of customers from both categories is estimated to be between 31.74 and 33.01.

Correlation Analysis



There is strong correlation between columns related to geography such as Zip Code, Latitude and Longitude. Also, target variable `Churn Value` is **negatively correlated** with `Tenure months` and **positively correlated** with `Churn Score`.

3.3 Algorithms & Machine Learning

ML Notebook

Overview:

To predict whether a customer will churn given various attributes of the customer, we build a machine learning model. We split the available data into train (80% of data) and test sets (remaining 20% of data). We train different machine learning algorithms using the training set. Also perform hyperparameter tuning of the models using cross validation. Then compare the results to find the best model that suits our problem context. I have selected 3 machine learning algorithms for my initial review.

1 . Ada Boost

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

2. Linear Discriminant Analysis

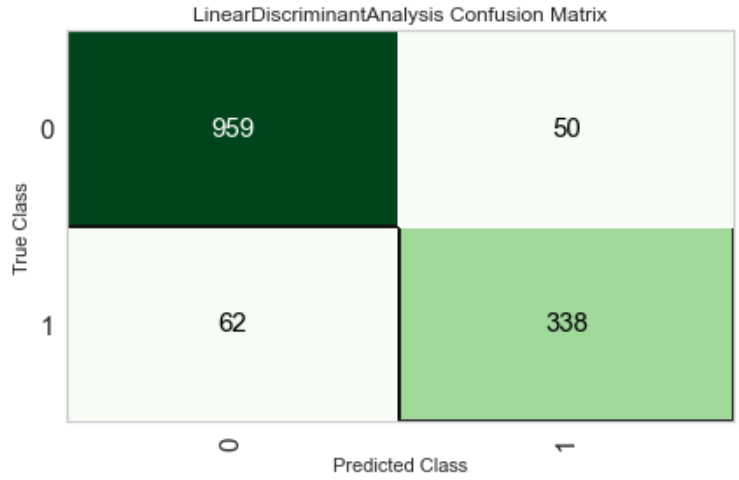
A classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule. The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix.

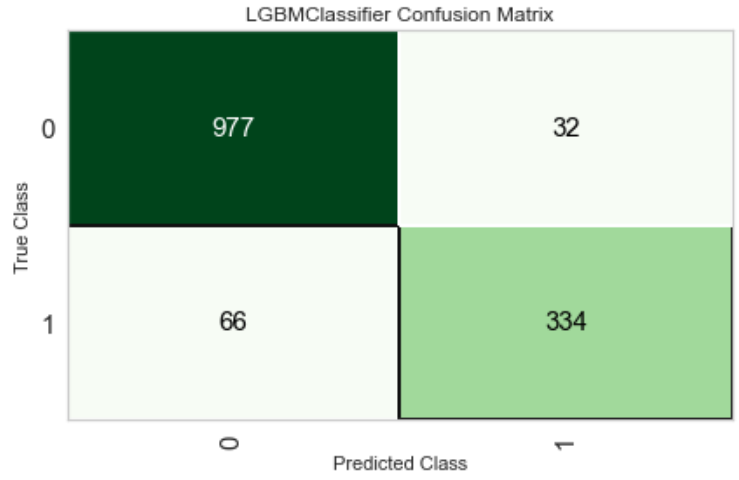
3. Light Gradient Boosting Machine

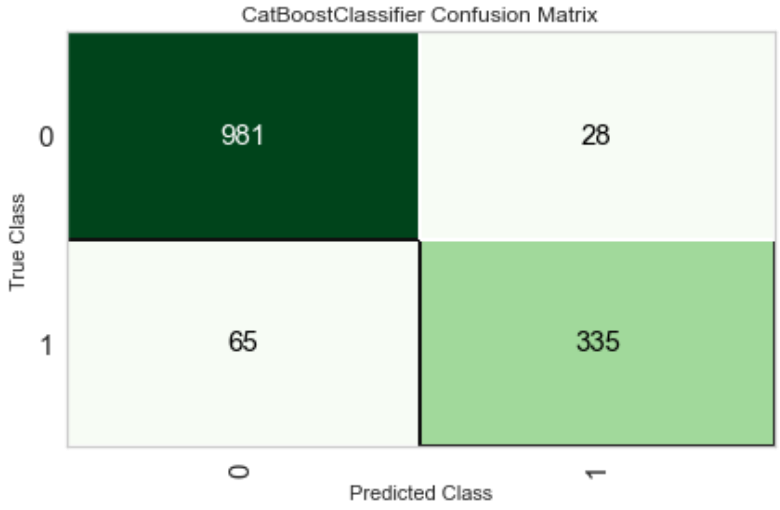
LightGBM, short for Light Gradient Boosting Machine, is a free and open source distributed gradient boosting framework for machine learning originally developed by Microsoft. It is based on decision tree algorithms and used for ranking, classification and other machine learning tasks.

Model Evaluation:

Here, accurately predicting churn = 1 is critical. If a customer is mis predicted as churn = 0, the company fails to apply necessary retention measures to retain that customer and thereby increases the chances of the customer leaving the country. Hence, our aim is to reduce the False Negatives and increase the True Positives. Thus, we give more importance to the Recall score and rank the models based on best Recall score.





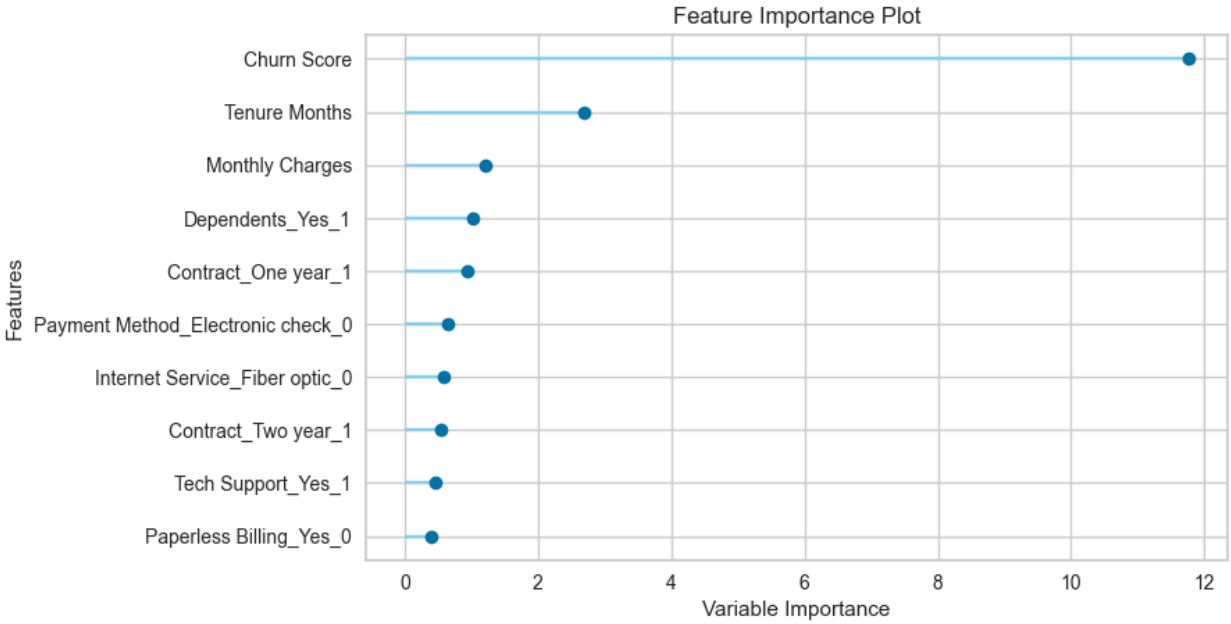


Best Recall score (Class: 1) = 0.845

Model : Linear Discriminant Analysis

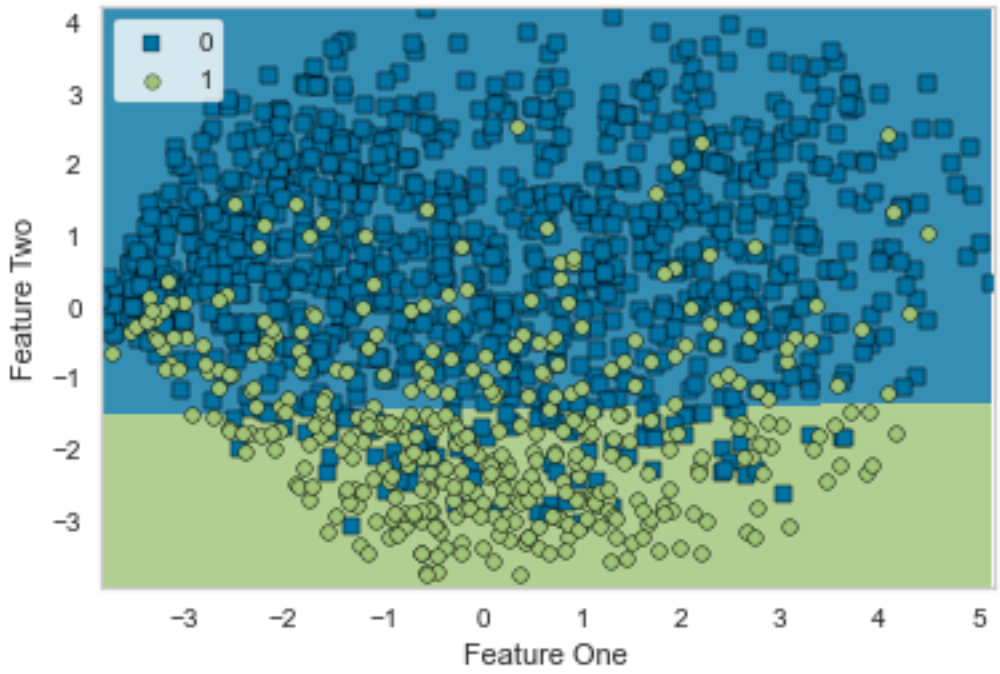
From the confusion matrix, it is evident that Linear Discriminant Analysis achieves the best recall score and minimum False Negative number. Hence we choose this model for our final prediction.

The contribution of different features in the prediction is identified as shown below.



Churn Score is a major contributor in predicting whether customer will churn. Tenure Months and Monthly Charges are few other important features identified as significant to prediction process.

The boundary plot of the data is shown below.



4. Takeaways

- EDA revealed that customers who are on month-to-month contract have high probability to churn compared to those on 2-year contract.
- Churn Score is a good predictor of customer churn. The difference between average churn score of those who churn and do not churn is between 31.74 and 33.01.
- Among the customers who churned, 70% of them had Fibre Optic Internet Service with the company. Analysis of churn reason also reveals that 33% of the customers who left the company were offered with better internet or devices by competitors.
- Among the customers who left the company, 37% were not satisfied with the services provided by the company.

5. Future Extensions

Availability of transaction data of customers can further assist in finding time sensitive trends in customer behavior.

More research on competitors in the market and the services they offer can reveal any attention needed on the services and packages offered by the company. It also helps in understanding where the company stands in the competitive market.

6. References

Not all the work in this notebook is original. Some parts were borrowed from online resources. I take no credit for parts that are not mine. They were solely used for illustration purposes.

1. https://www.statisticshowto.com/welchs-test-for-unequal-variances/

2. https://medium.com/@sosterburg/mapping-data-with-folium-356f0d6f88a9