CHURN PREDICTION

APPROACH

I performed Churn analysis from the datasets given in this JOB-A-THON, which gives the customer information demographics and past activity with the bank and trying to better understand whether the customer will churn or not.

THE DATA

Here given 3 files – train.csv, test.csv and sample_submission.csv. In the training dataset, there are 6650 (each representing unique customer) rows with 11 columns: 9 features and 1 target feature(Is_churn). The data composed of both numerical and categorical features.

Target:

Is_churn – Whether the customer will churn or not

Numerical features:

Age – Age of customer

Balance – Average quarterly balance of customer

Vintage – Tenure of customer

Transaction Status – whether the customer has done any transaction in last 3 months or not.

Credit_card – Whether the customer has credit card or not.

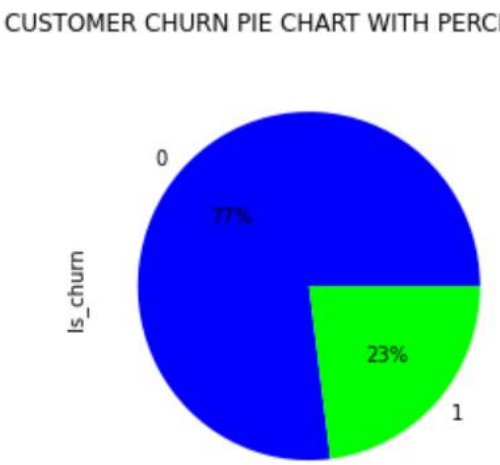
Categorical features:

Gender – Gender of customer

Income – Yearly income of the customer

Product holdings – No of product holdings with the bank

Credit_Category – category of customer based on credit score.



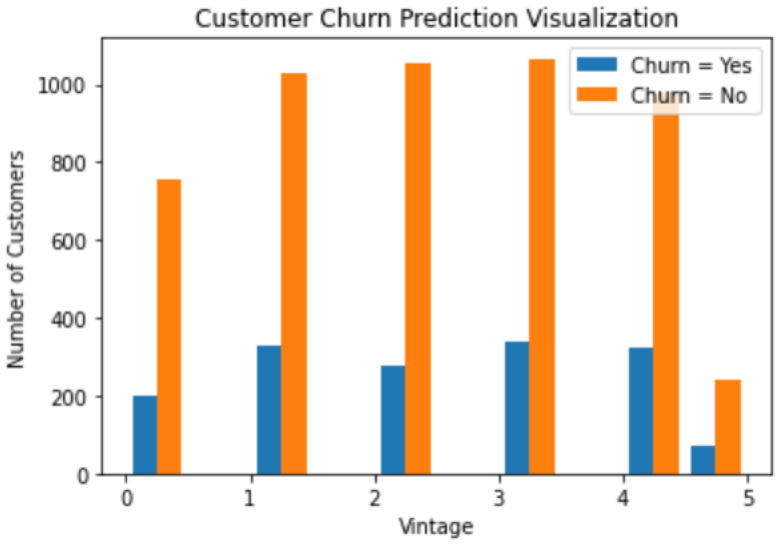
We can see from the pie chart that about 23% of bank customers from the given dataset end up churning.

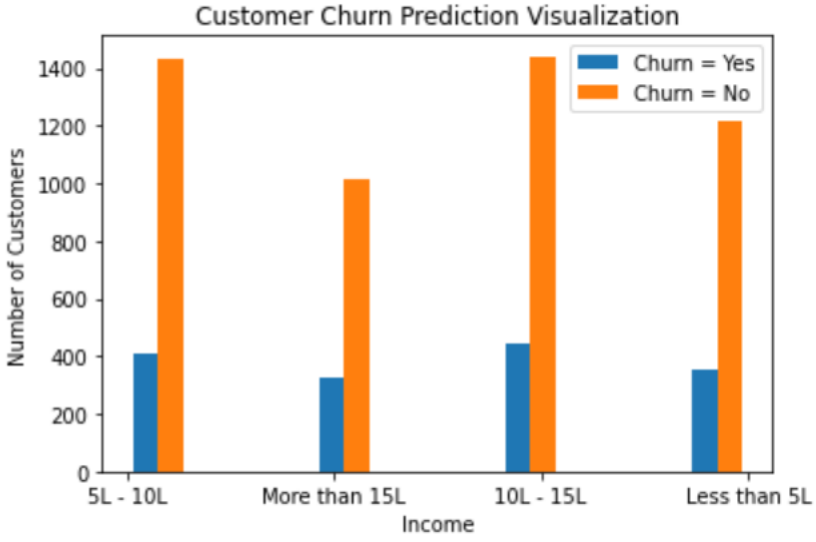
Analysing the dataset, we see that customer ID which is of object type and when we are going to build machine learning model customer ID doesn't help. So we drop the column.

Checked for datatypes of all variables and unique values of each variable.

Univariate analysis

Did univariate analysis of features.





Feature Selection

Used optbinning for selecting important features. Found out IV(Information Value). If IV> 0.5, then select those features otherwise drop the features. But it resulted with very low IV. So could not select features through this optbinning method. So I removed the code of it.

OneHotEncoding

Did OneHotEncoding of categorical features and applied MinMaxScaler to the features.

Accuracy

To test the accuracy and score of model, I splited the training set into train and test set, applied logistic regression, Naïve Bayes algorithm, Random Forest algo and checked for F1 score, did confusion matrix, found out precission, recall and AUC_ROC curve. Found that Naive Bayes model resulted better among all.

Then took the testset given and predicted the value for final submission.