CHURN PREDICTION

Objective: To build a machine learning model to predict whether the customer will churn or not in the next six months.

Brief of data set:

- train Data (train.csv) contains the customer demographics and past activity with the bank(10 independent features). And also the target label representing whether the customer will churn or not.
- test.csv contains the customer demographics and past activity with the bank. And you need to predict whether the customer will churn or not.

Evaluation Metric:

- The evaluation metric for this hackathon is macro f1 score.

Public and Private Split

Test data is further divided into Public (40%) and Private (60%) data.

- Initial responses will be checked and scored on the Public data.
- The final rankings would be based on the private score which will be published once the competition is over.

Churn_prediction.ipynb description:

The given train and test data set didn't have any duplicate or null values.

The test data independent features were further classified as:

Product_Holdings, Credit_Category, Vintage, Income as ordinal data.

Gender, Transaction_Status, Credit_Card and Is_Churn as categorical data.

Age and Balance as numerical data.

EDA

Univariate analysis:

Categorical Variables -

- Target variable : Around 22% customers will churn after 6 months.
- Gender : Aound 55% of customers are male and 45% are female.
- Transaction_Status: There is not much difference between transaction status in tha last 3 months i.e. percentage of transaction(51%) and no transaction(48%) in past 3 months are similar.
- Credit_Card: Around 65% of the customers have a credit card.

Ordinal Variables -

- Product_Holdings: Customers holding 1 and 2 products with the bank are almost similar (around 49%).
- Credit_Category: Most of the customers have a poor credit category(poor credit score).
- Vintage: In Vintage graph ie number of years the customer is associated with the bank, least no of customers are associated with the bank for the past 5 years (around 4%), and around 14% are new customers (0 years), and rest having similar percentage of years of association with the bank(around 20%).
- Income:Most of the customers have an income range of 10L-15L and 5L-10L.

Numerical Variables:

- Age: The distribution of age is fairly normal, and the box plot shows that there are some outliers. Most of the customers are from the age group 30-40.
- Balance: The distribution of Balance is not normal, it has two peaks, and has some outliers.

Bivariate Analysis:

In this section the relation between Is_Churn (target variable) and remaining independent variables is being analyzed.

For eg:

- Females are more likely to churn as compared to males.
- Proportion of customers with no transaction in the past 3 months are more likely to churn in the next 6 months.
- Customers with 1 product holding are more likely to churn in next 6 months.

- Customers with average and poor credit category are more likely to churn in next 6 months.
- customers with Income more than 15L and 10L - 15L (ie with high icome range) are more likely to churn.
- It can be inferred that in the age group 61-72 has more churn percentage.
- Customers with high Balance(Average quarterly balance) are more likely to churn .

Multivariate Analysis:

Relation of two or more independent variables with the target variable is studied in this section.

Age and Is_Churn are positively correlated.

Feature Engineering:

Here, we seek to add features that are likely to have an impact on the probability of churning.

- Balance(average quarterly balance) and Vintage(No. of years associated with the bank) are functions of the year therefore combining them by dividing balance with vintage(when vintage=0 keep the balance as it is).
- balance_vintage which we created does not follow normal distribution and contains outliers, so to deal with this we apply Box Cox transformation on the balance_vintage variable.

Outlier Treatment on Age and Balance:

One of the way to remove skewness is to take log transformation, it does not affect smaller values but reduces the larger values

- Since Age and Balance variables contain outliers and do not follow normal distribution closely we apply log and boxcox transformations on them respectively.
- Boxcox transformation on Balance because log transformation did not gave the desired result.

* Applying the similar transformation and feature engineering on test data.

Model Building:

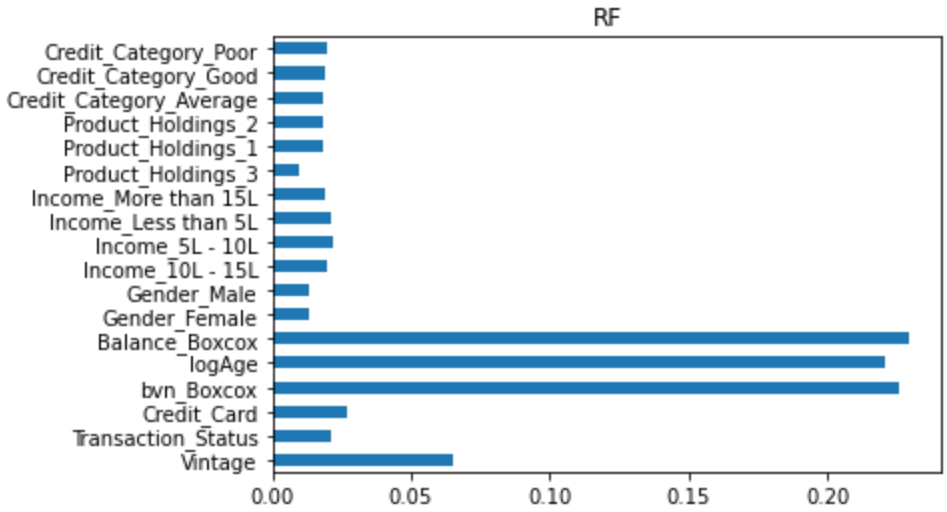
- Making the data model ready, by removing ID variable from train and test dataset, converting object data types using pd.get_dummies().
- Oversampling the train data to get the precise churn prediction since the data contains a higher percentage of 0's(not going to chun in next 6 months).
- Split the train data using train test split to check the score on unseen data, since the test set does not contain target variable .
- **Logistic Model**: using grid search cv we choose the parameters and fit and predict using the model.check the macro f1 score .
- **RandomForestClassifier**:
    - Using RandomizedsearchCV we select the parameters for this model called 'RF' and check the macro f1 score.
    - Using trial and error around the parameters from RandomizedsearchCV we fit this model called 'RF1' and check the macro f1 score.
- **XGBClassifier**:
    - Using trial and error choose the parameters for this model called 'model' and check the macro f1 score.

Best model :

- **RF model gave better performance than all the other 3 models on leaderboard.**

Feature Importance:

- **From the RF model the below graph shows the important features in prediction.**



This graph clearly shows that Balance_Boxcox, logAge and Bvn_Boxcox are the important features for predicting, these features are the ones which we created in the feature engineering and outlier removal stage. So our transformation was successful.