**A brief on the approach used to solve the problem**

**Outline:**

1. Loading data
2. EDA
3. Feature Engineering:
4. Data Pre-processing
5. Handling imbalance
6. Model selection
7. Model Evaluation:

**EDA**

1. Understanding the Data:

* Different columns and its Datatypes
* Check for missing values- no missing values

Text

Description automatically generated

1. Correlation between variables:

Table

Description automatically generated

1. Did EDA using pandas profiling- refer file - 'output.html'. Understood the variables in the data. Checked for categorical and continuous variables.

* Continuous variables- Age, Balance
* Categorical variables- Gender, Income, Vintage, Transaction\_Status, Product\_Holdings, Credit\_Card, Credit\_Category

1. Separated the ID column from the data.
2. Used Seaborn to plot for making statistical graphics.

* Plotted the count plot of ever categorical variable to understand its effect on our target – customer churn. From this understood that ‘Gender’ is not contribution in to classification. All other variables are important.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

* Plotted boxplot with Vintage on train data – It showed uniform distribution with vintage.

Chart, box and whisker chart

Description automatically generated

* Plotted boxplot with Vintage on test data – It showed uniform distribution with vintage. Also similar distribution as train data.Chart, box and whisker chart

  Description automatically generated
* Checked the relationship between- Age, Balance, Vintage, Product holdings and Our target – Is\_churn.

Chart, scatter chart

Description automatically generated

From this we could see we have less datapoints for product holding 3+ so I also checked for this distribution in the test data and it gave same results as follows:

Chart, scatter chart

Description automatically generated

* Checked the relationship between – Age, Vintage, Transaction status, product holdings and target- Is\_Churn. We could see vintage is playing a key role and Trasaction status is also affecting the classification.

Chart

Description automatically generated

* Checked relationship between -Vintage, Age, Product holdings, Transaction status and its effect on our Target-Is\_Churn.

Chart

Description automatically generated

**Feature Engineering:**

Using EDA I noted that Gender was not contributing for classification. Hence we considered to drop the “Gender” column

I aimed at creating/working out more features like:

**1.Using Age as category and continuous both**

**2.Taking a ratio of Balance and age as a feature**

**3.Normalized balance within income group or age group . However this did not give me good result. So I did not experiment with it.**

**Data Pre-processing**

* Defined the categorical and numerical columns
* For categorical variables -One hot encoding using pandas ‘get\_dummies’
* For continuous variables we tried with -MinMaxScaler, StandardScaler and QuantileTransformer

However QuantileTransformer resulted in best score.

**Handling imbalance:**

From the target distribution it is apparent that the data is imbalanced. I have tried the solve this by undersampling technique like:

1. ClusterCentroid
2. NearMiss
3. EditedNearest Neighbours(we chose this one)

We also used **class weights** while fitting the data ,however it yielded poor results.

**Model Selection:**

I have experimented with SGD, NaiveBayes, Support vector and CART.

Adaptive boosting have given good performance.

**Text

Description automatically generated**

**Model Evaluation:**

For an imbalanced dataset, roc\_auc score is considered to be the most bankable.

ROC\_AUC score: 0.724

**On standard public test data:**

**Macro F1 score of 0.61066**

Graphical user interface, text, application, chat or text message

Description automatically generated