**Customer Churn Model**

**Objective:**

 To analyze the past data of employee churn and predict whether the customer will churn or not in the next 6 months

**Importing the libraries and reading the data**

1. Necessary libraries are imported.

Libraries used are numpy, pandas, seaborn, sklearn, model\_selection, train\_test\_split, GradientBoostingClassifier

1. Data sets are loaded into data frames as “train” for training the model and “test” for testing the model.

**Exploratory Data Analysis**

1. The shape and dimensions of the data are checked.

The train data set as 6650 rows and 11 columns, while the test data has 2581 rows and 10 columns.

1. Upon checking the information of the data, it is observed that the data has

‘ID', 'Age', 'Gender', 'Income', 'Balance', 'Vintage','Transaction\_Status', 'Product\_Holdings', 'Credit\_Card','Credit\_Category', 'Is\_Churn' columns.

1. ID is the unique representation of each row. And Is\_Churn is whether the customer has churned or not. 1 represents that the customer has churned.
2. It is observed that only Balance in float data type. And other columns are in integer and object data types.
3. There are no null values in both the train and test datasets. The data appears to be clean to work on.
4. Upon checking for the number of unique values in each of the columns. It is observer that the data in the columns Age and Balance have a high number.
5. The other columns except ID, Age, Balance are categorical.
6. Checked for the unique data in each of the categorical columns.

Gender : ['Female', 'Male']

Income : ['5L - 10L', 'Less than 5L', 'More than 15L', '10L - 15L']

Vintage : [4, 2, 0, 1, 3, 5]

Transaction\_Status : [0, 1]

Product\_Holdings : ['1', '2', '3+']

Credit\_Card : [0, 1]

Credit\_Category : ['Average', 'Poor', 'Good']

Is\_Churn : [1, 0]

**Analysis of each of the columns with respect to churn rate**

1. Gender :

#It is observed that the ratio if Males in the data set is higher that the females

#The churn rate difference wrt gender is not very high

1. Income:

#The data set is slightly dominanat with customers with income of 10L - 15L and 5L - 10L

#The ratio difference of Churn wrt Income however is almost same across all categories

1. Vintage:

#The dataset follows a slight normal distribution wrt to Vintage

#There is no significant difference in churn ration wrt vintage

1. Transaction Status:

#It is observed that slightly higher number of customers whose transaction status is zero have churned as compared to those whose transaction status was one

1. Product Holdings:

# The number of customers with more than 3 product holdings is very less as compared to the number of customers who have 1 or 2 product holdings.

# The difference churn ratio is not significantly high but more number of customers with only one product holdings have churned more

1. Credit Card:

# Half of the customers do not have a credit card

# But the churn rratio is same in between the customers who have credit card and who don't

1. Credit Category:

#The credit category is poor for most of the customers.The number follows by Average and then Good

#The customer churn is high for customers with poor credit category followed by Average and good

1. Age and Balance

# The average age of the customers who churned is 44

#These two columns are very significant in building the model

1. Is\_Churn:

#The number of customers who have churned is very less as compared to those who haven't

#It is also observed that there is imbalance in the dataset

**Data Preprocessing**

1. Label Encoding the Income, Product\_Holdings, Credit\_Category as they are categorical and have some order.
2. The columns are mapped as below:

gender\_mapper = {"Female":0, "Male":1}

income\_mapper = {"Less than 5L":0,"5L - 10L":1, "10L - 15L":2,"More than 15L":3}

product\_holdings\_mapper = {"1":0,"2":1, "3+":2}

credit\_category\_mapper = {"Poor":0,"Average":1, "Good":2}

1. The above mappings are done both on train and test data.
2. Dropped ID from both train and test data.

**Building the Model**

1. Train data is split for building the model and validating it.Since the data is imbalanced i.e,it has very less number of customers who have churned and high number of customers who haven’t , we use stratify with respect to Is\_Churn so that the data is sampled and split in such a way that the data used to build the model and evaluate the model have same ratio of each of the labels.
2. Since the target variable that we have to predict is categorical, we use a classification algorithm.
3. Logistic Regression, KNN, Decision Tree and Random Forest had a decent accuracy but the f1 score was low.
4. GradientBoostingClassifier is used as the machine learning model.

**Validating, testing and tuning the model**

The model is validated based on macro f1 score.

With the GradientBoostingClassifier’s default parameters the model was doing fine.

The performance of the model was decent on validation data set, however the f1 score on the test dataset was less. Since the outputs were predicted as probabilities, changing the threshold to convert the probabilities to labels has helped in increasing the model performance on the test data set.The threshold has been set to 0.3 for conversion. The number of estimators is set to 40 based on trial and error.

The model has a macro f1 score of 0.58 on the validation data and 0.60 in the test data.