# Churn Prediction

## Understanding Problem Statement

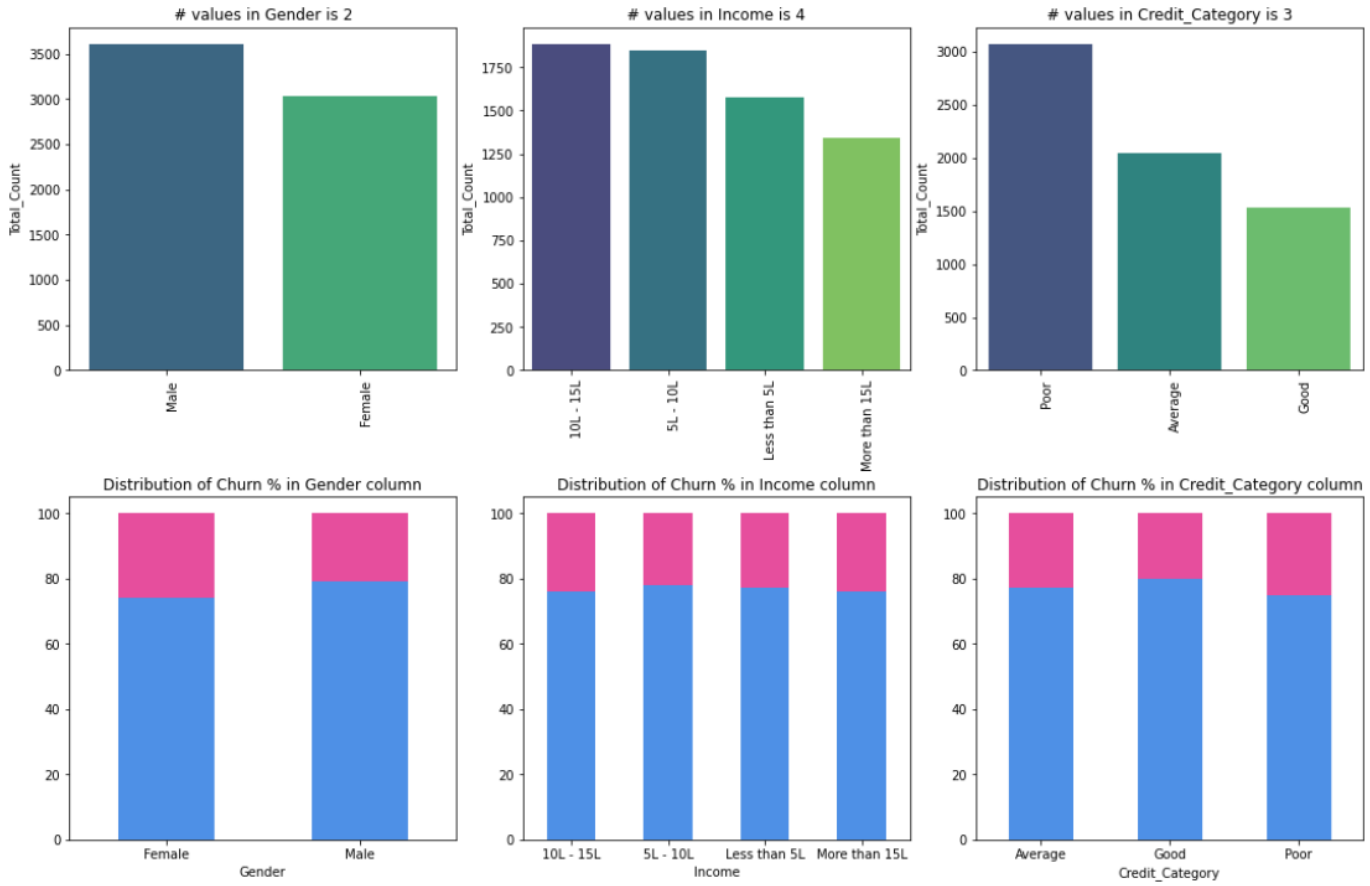
Decreasing the Customer Churn is a key goal for any business. Predicting Customer Churn (also known as Customer Attrition) represents an additional potential revenue source for any business. Customer Churn impacts the cost to the business. Higher Customer Churn leads to loss in revenue and the additional marketing costs involved with replacing those customers with new ones.

In this challenge, as a data scientist of a bank, I was asked to analyze the past data and predict whether the customer will churn or not in the next 6 months. This would help the bank to have the right engagement with customers at the right time.

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

I have done Extensive EDA to understand the data. Performed the following data manipulations to prepare the data for model building.

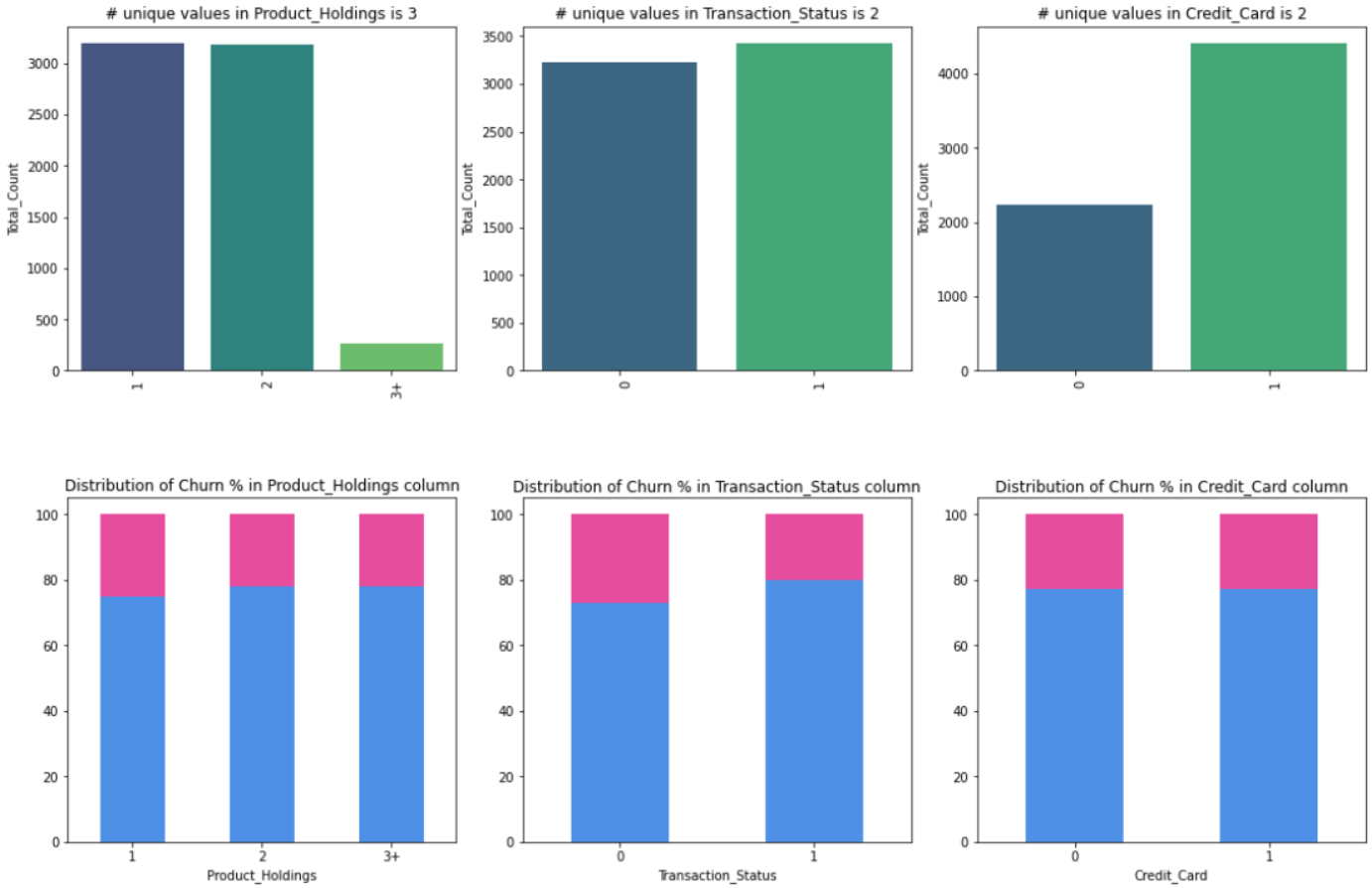
## Exploratory Data Analytics




We note the following:

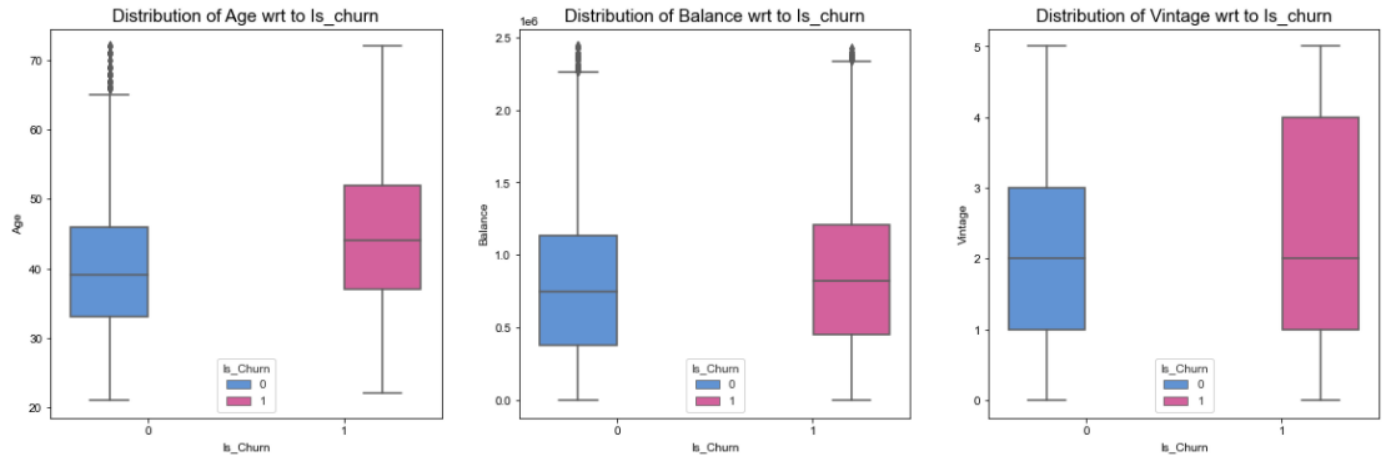
- The proportion of female customers churning is greater than that of male customers. Bank should rollout promotions to retain female customers.
- Those with 5L-10L salary churn less compared to other income bands.

Naturally belonging to poor credit category are more likely to churn




- Customers with more product holdings are less likely to churn. so, the Bank needs to focus on encouraging more customers to have more product holdings which will discourage the customers to leave the service.
- Unsurprisingly those who did not make any transaction in the last 3 months are likely to churn
- Interestingly, no difference in the churn rate for those having credit card and not.

A cause of concern, overall proportion of inactive members is quite high suggesting that the bank may need a program implemented to turn this group to active customers as this will definitely have a positive impact on the customer churn.




We note the following:

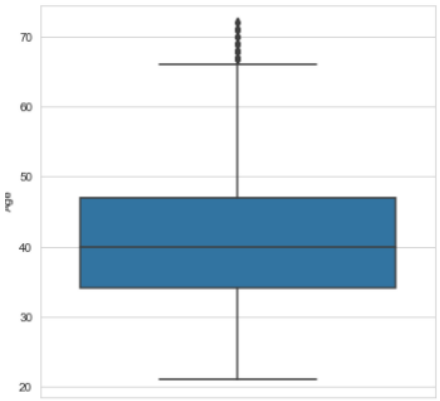
- The older customers are churning more than the younger ones. The bank may need to review their target market or review the strategy for retention between the different age groups.

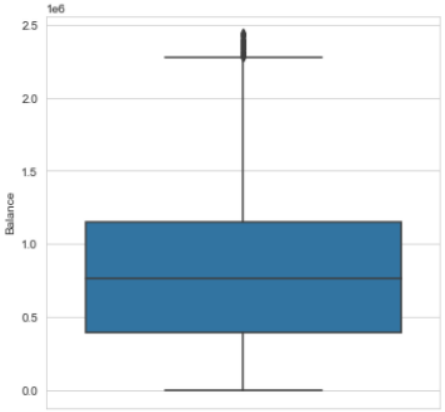
- Even customers have been associated with the bank for about 4 years are leaving.

# Data Manipulations/ Feature Engineering

### *Checking for Outliers*

Though the values in Age and Balance seem to have outliers in them upon careful observation of the percentile values we can we can rule out that there are any outliers.




**Percentile values In Age column**
80th percentile value 49.0
85th percentile value 52.0
90th percentile value 55.0
95th percentile value 59.0
99th percentile value 66.0

**Percentile values In Balance column**
80th percentile value 1248321.5820000004
85th percentile value 1368751.6124999998
90th percentile value 1511154.5219999999
95th percentile value 1721617.5554999998
99th percentile value 2154397.1733000013

### *Convert variables*

Convert categorical variables into numeric values and Binary variables.

Credit category has classes poor, average and good. I converted them into numeric values(ordinal) giving them clear order as well. For variables such as gender, Income, Product_Holdings where ordering doesn't matter, I have converted them into binary values and creating dummy variables.

### *Normalization of variables/ feature scaling*

For 'Age', 'Balance', 'Vintage','Credit_Category', the values belong to a wide range.

Machine learning algorithms tend to perform better or converge faster when the different features (variables) are on a smaller scale. Therefore it is common practice to normalize the data before training machine learning models on it. Normalization also makes the training process less sensitive to the scale of the features. This results in getting better coefficients after training.

Normalization makes the features more consistent with each other, which allows the model to predict outputs more accurately.

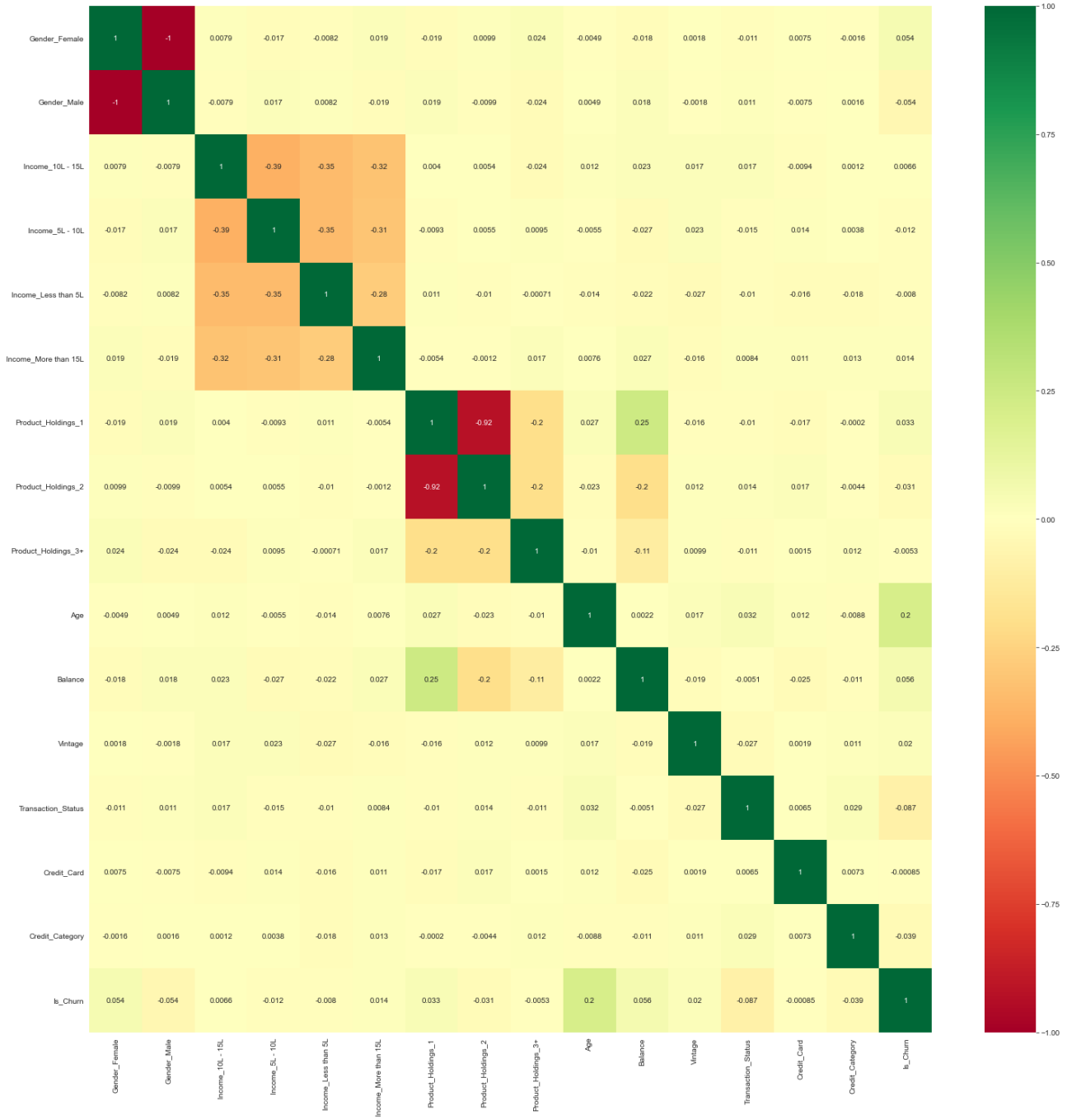
I used below given formula and normalized the variables. And now values are between the range 0 to 1.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Normalization

### *Checking correlation between variables*

There are some variables that have extreme positive or negative correlation, such variables are deleted. Remove unimportant features: We drop the features 'Gender_Male','Product_Holdings_2','Income_5L - 10L' that do not add value for modeling.




## *Check Multicollinearity using VIF:*

I also looked into multicollinearity using Variable Inflation Factors (VIF).

Unlike Correlation matrix, VIF determines the strength of the correlation of a variable with a group of other independent variables in a dataset.

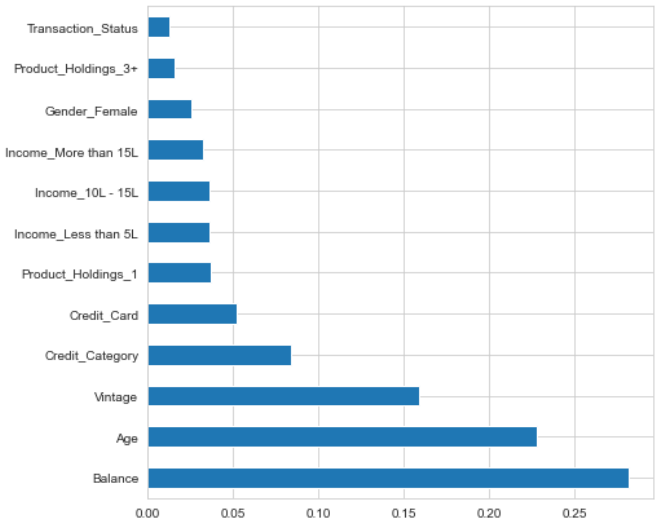
VIF starts usually at 1 and anywhere exceeding 10 indicates high multicollinearity between the independent variables.

VIF for all variables is <5. Age has high VIF amongst the variables.

| | variables | VIF |
|---|---|---|
| 6 | Age | 4.167285 |
| 7 | Balance | 3.255376 |
| 8 | Vintage | 2.914928 |
| 10 | Credit_Card | 2.645338 |
| 4 | Product_Holdings_1 | 2.068797 |
| 9 | Transaction_Status | 1.962048 |
| 1 | Income_10L - 15L | 1.841648 |
| 11 | Credit_Category | 1.831220 |
| 0 | Gender_Female | 1.756595 |
| 2 | Income_Less than 5L | 1.661755 |
| 3 | Income_More than 15L | 1.604634 |
| 5 | Product_Holdings_3+ | 1.075955 |

### *Checking Feature Importance:*

I used ExtraTreesRegressor() to check the feature importance and Balance, Age, Vintage and Credit Category stood out to be important features.




### *Recursive Feature Elimination*

I have also used Recursive Feature Elimination (RFE) to choose either the best or worst performing features

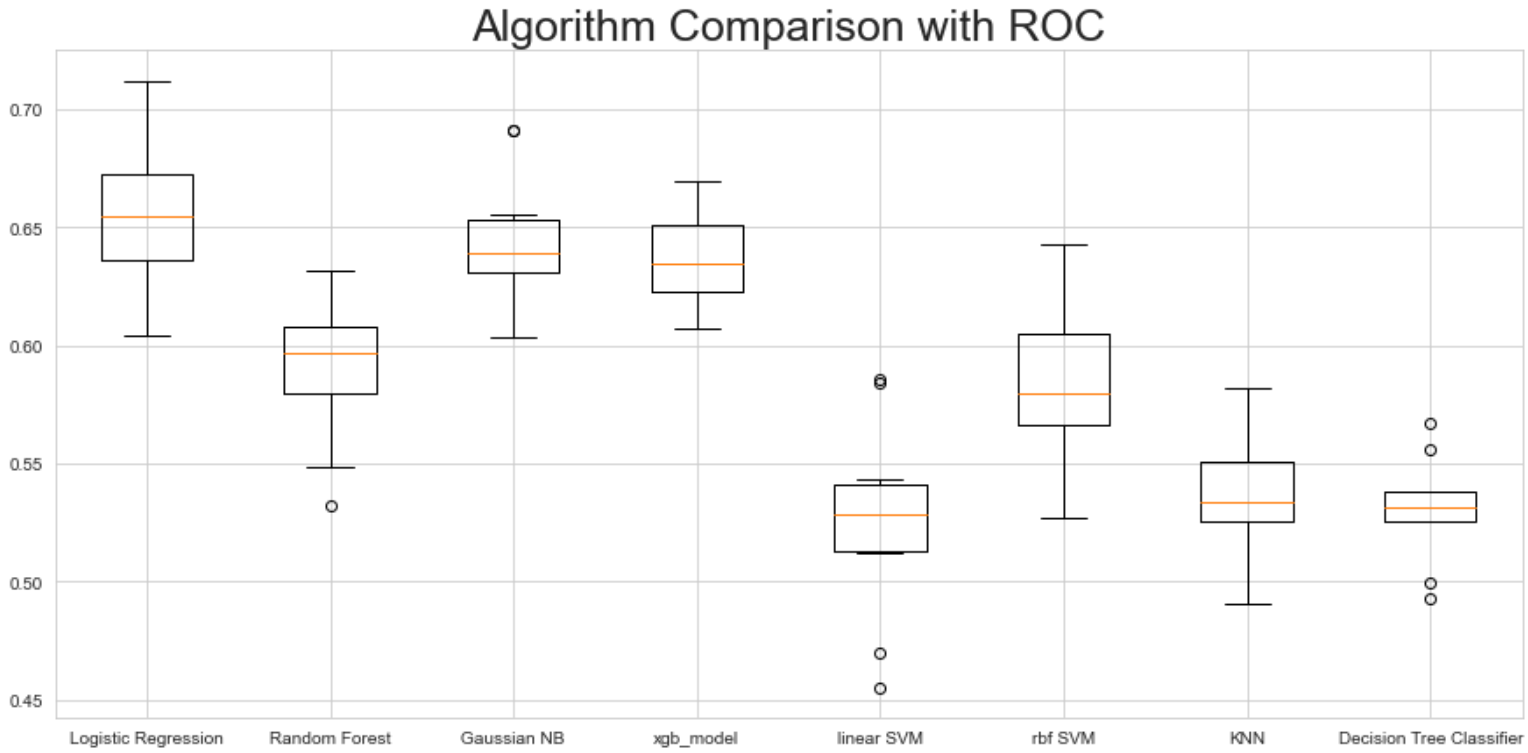
```
[1 4 6 2 3 5 1 1 1 1 7 1]
['Gender_Female' 'Income_10L - 15L' 'Income_Less than 5L'
 'Income_More than 15L' 'Product_Holdings_1' 'Product_Holdings_3+' 'Age'
 'Balance' 'Vintage' 'Transaction_Status' 'Credit_Card' 'Credit_Category']
```

The ranking of the features is as shown above. The RFE has helped us select the following features as important features giving them rank 1:

'Gender_Female' 'Age' 'Balance' 'Vintage' 'Transaction_Status' 'Credit_Category'

## Choosing baseline model for our dataset

I used test train split of 25% and 75% and tried several base models including Logistic Regression, Random Forest, Gussian Naive Bayes , XGBoost, linear SVM, rbf SVM, k-Nearest Neighbors and Decision Tree with 10-fold Cross-Validation to understand how each model is performing.

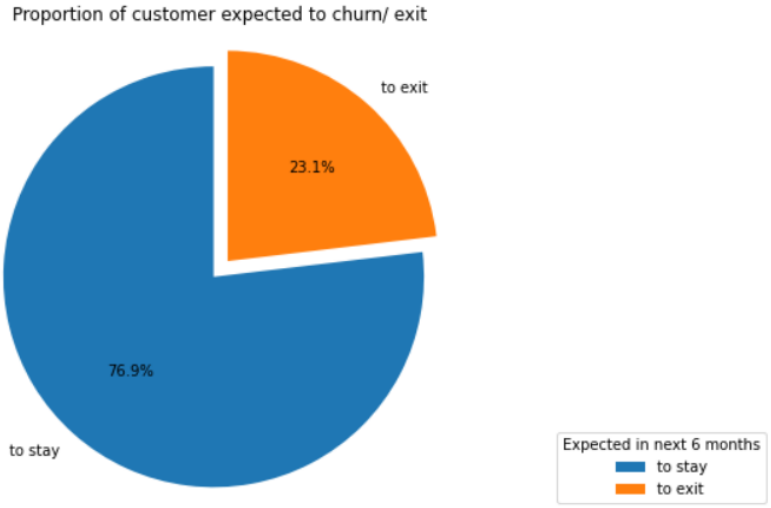


The base logistic regression model is giving an accuracy of 0.77 and macro avg F1 score of 0.48. This model is classifying True Negatives with high accuracy. To evaluate performance of all classes I checked the F1 score which is very low.

On way to address model performance is by training the model with Balanced classes in target variable.

## Case of Imbalanced Data:

Distribution of classes in target variable Is_Churn is imbalanced i.e. the number of customers who churned was significantly smaller than the number of customers who did not.




I addressed the class imbalance using (SMOTE)Synthetic Minority Oversampling: Since the dataset was small, oversampling the positive class increased my data. As a result, though there is a fear of model over fitting, macro f1 recall score on hold out sample sinificantly improved.

It is also doing a good work in predictingg the False Negatives.

## Summarization

In summary, our best submission was by training a Logistic regression model, with over samplin data and with the feature engineering steps described above.