Customer Attrition Prediction using Machine Learning

Capstone Project
Data Science Career Track, Springboard

Thanks to mentor Julian Jenkins III

How could a manager retain his customer leaving or discontinuing the services provided to him. So that business can stay healthy and add more services to customers to enhance annual profit. Build a predictive analysis on the credit card dataset to understand whether the customer churn or not, and identify the reasons for them to leave.

Customer churn means a customer's ending their relationship with a bank/company for any reason. Although churn is inevitable at a certain level, a high customer churn rate is a reason for failing to reach the business goals. So identifying customers who would churn is very important for business.

I have used a credit card datset.Credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees,balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

Objectives:

- Explore the dataset and visualize the same
- Build a model to predict the customer is going to churn or not
- Optimize the 3 models with appropriate techniques
- Generate a set of insights and recommendations that may help the bank

Data Source

From google data search
https://www.kaggle.com/sakshigoyal7/credit-card-customers
OR
https://www.kaggle.com/c/1056lab-credit-card-customer-churn-prediction/data

Data Dictionary

- CLIENTNUM: Client number. Unique identifier for the customer holding the account
- Attrition_Flag: if the account is closed then "Attrited Customer" else "Existing Customer"

- Customer_Age: Age in Years
- Gender: Gender of the customer
- Dependent_count: Number of dependents
- Education_Level: Educational Qualification -Graduate, High School, Unknown, Uneducated, College, Post-Graduate, Doctorate.
- Marital_Status: Marital Status
- Income_Category: Annual Income Category
- Card_Category: Type of Card
- Months_on_book: Time frame with the Bank
- Total_Relationship_Count: Total no. of products held by the customer
- Months_Inactive_12_mon: No. of months inactive in the last 12 months(one year)
- Contacts_Count_12_mon: No. of Contacts between the customer and bank in the last 12 months(one year)
- Credit_Limit: Credit Limit on the Credit Card
- Total_Revolving_Bal: The balance that carries over from one month to the next is the revolving balance
- Avg_Open_To_Buy: Open to Buy refers to the amount left on the credit card to use (Average of last 12 months)
- Total_Trans_Amt: Total Transaction Amount in Last 12 months((one year))
- Total_Trans_Ct: Total Transaction Count in Last 12 months((one year))
- Total_Ct_Chng_Q4_Q1: Ratio of the total transaction count in 4th quarter and the total transaction count in 1st quarter
- Total_Amt_Chng_Q4_Q1: Ratio of the total transaction amount in 4th quarter and the total transaction amount in 1st quarter
- Avg_Utilization_Ratio: Represents how much of the available credit the customer spent

---

The customer churn, also known as customer attrition, refers to a customer ending relationship with a company for some reasons.Identify and visualize which factors contribute to customer churn.

Data Science Problem

A manager at the bank would like to know the reason why more and more customers are leaving their credit card services. They would really appreciate it if one could predict for them who will get churned so they can proactively improve the service, so that the bank can run profitably.Because it is well known that getting new customers is more costlier than retaining customers.

Credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees, balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances. Customers' leaving credit card services would lead banks to loss, so the bank wants to analyze the data of customers and identify the customers who will leave their credit card services and reason for the same – so that the bank could improve upon those areas.

Data Wrangling

There are 10127 rows and 23 columns. Here are the steps performed to clean and oraganize the data.

- Checked for missing values
- Checked for Null
- Checked for missing values
- Checked for unique values
- Weird values or filled, for example “unknown”
- Corrected weirdly formatted values(income category)
- From the initial look on data, found below information.

And initial analysis as as follows-

- Average customer age is ~46 and min and max customer age is 70 to 82.
- Average period of relationship with the bank is ~36 months with a minimum of 13 and max as 56.
- Average Total number of products with the customer ~4 and maximum is 6.
- Mean of Credit_limit 8631 while median is 4549 ,data may have outliers.
- Total_Revolving_Bal(unpaid portion) has mean as 1162 while median is 1276, No outliers.
- Avg_Open_To_Buy(amount left on the credit card) has mean 7469 and max as 34516.Appears some outliers.
- Total_Trans_Amt has an average of 4404 and median of 3899. This indicates outliers.
- CLIENTNUM appears to be unique value.This also tells there are no duplicate values

My target variable is :Attrition_Flag: if the account is closed then "Attrited Customer" else "Existing Customer"

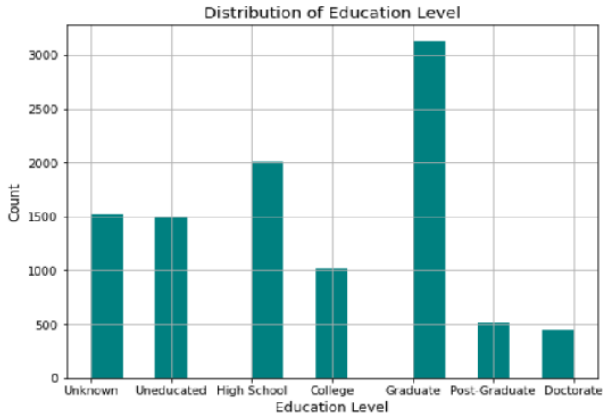
After finding unique values on each features,a casual inspection by eye reveals following

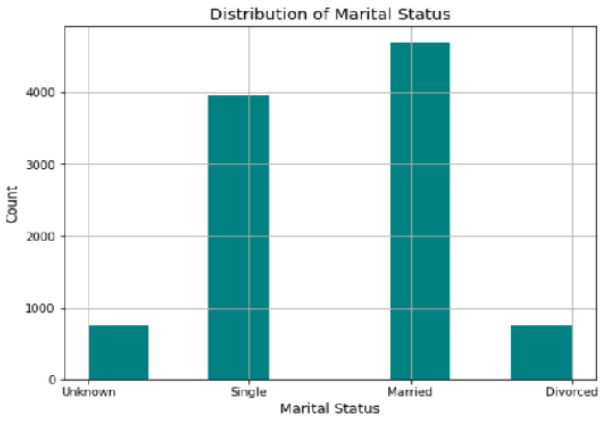
- Total of 16% customers have attrited against 84% are still with the bank.
- Bank has both Male and Female are almost same number.
- More customers are with education "Graduate", followed by "unknown", with "Uneducated" and "College".
- Most of them are married or single.
- Less than 40K income customers are more.
- Customers with 2 or 3 dependents top the list.Maximum dependency is 5.
- Blue cards have more customers.
- Maximum 6 products are held by customers, people with 3 products tops the list.Followed with 4 and 5.
- Customers are in contact with Bank for 0 times at the least, followed by 6 times.Customer with 2 and 3 times contacted by Bank are topping the list.
- More customers stayed inactive for 2-3 months.
- Education level,Income category,marital status has an "Unknown" category , this will have to be treated as missing value and will have to be imputed.

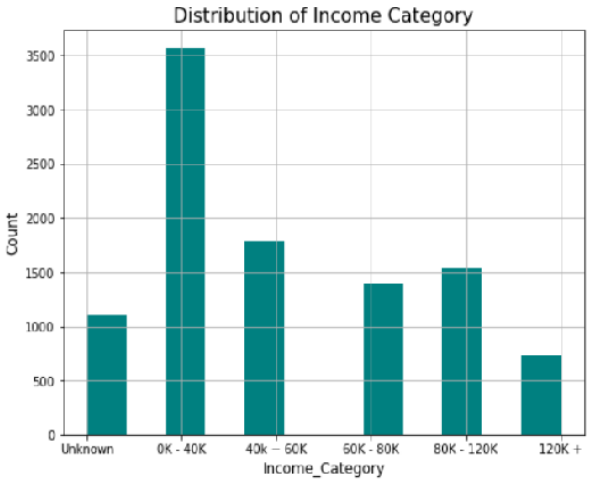
"Unknown" was fixed with respective modes for these and arranged the order as well.

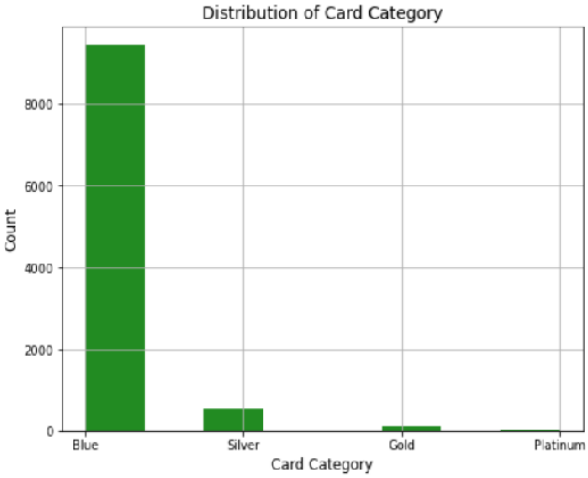
Card category as arranged according to highest to lowest

Here is the look -









Feature Scaling and Distribution:

4 different types scaling is applied -

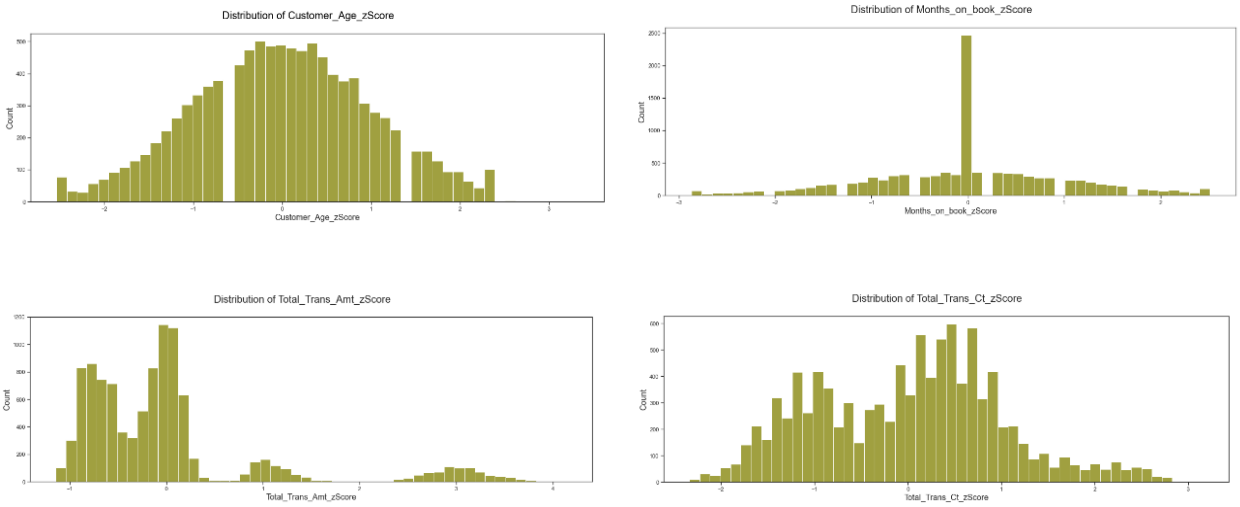
1. z-score scaling : Approximately normally distributed
2. Divided by Median: related by magnitude or right skewed
3. Log scaling:related by magnitude
4. Square root : for counts

Z-Score Scaling - number of standard deviations away from the mean

Formula for Z-score = (Observations - Mean)/Standard Deviation

Applied zscore scaling on to 4 features,"Customer_Age","Months_on_book","Total_Trans_Amt","Total_Trans_Ct" and created all 4 new columns with "_zscore" suffix .

Here is the dist plot



Analysis from zscore plot

- Customer Age appears to be normally distributed. May have negligible outliers or it should be good to train the model with these outliers.
- Months on book : Has the spike in the middle for 0 value.
- Total_Trans_Amt : This is interesting, with right skewed and bimodal, also with outliers.
- Total_Trans_Ct : Appears to be bimodal, may have outliers.
- Months_Inactive : Appears as left skewed. May have outliers
- Contacts_Count_12_mon : Distributed normally.

Outliers were checked with threshold value 3 for these features and found -

- 391 outliers for Total_Trans_Amt (Total Transaction Amount in Last 12 months)
- 124 outliers for Months_Inactive (No. of months inactive in the last 12 months)
- 54 outliersfor Contacts_Count (No. of Contacts between the customer and bank in the last 12 months)
- 2 outliers for Total_Trans_Ct (Total Transaction Count in Last 12 months)
- 1 outliers for Customer_Age
- and none for Months_on_book (Time frame with the Bank)

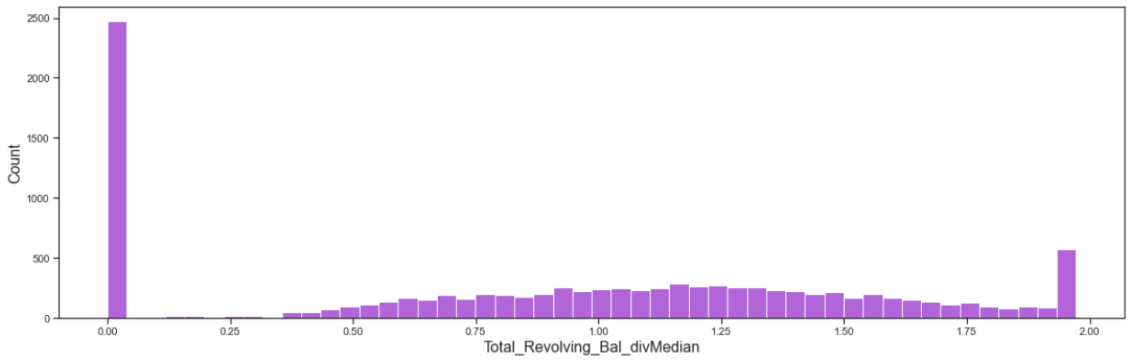
Median Scaling

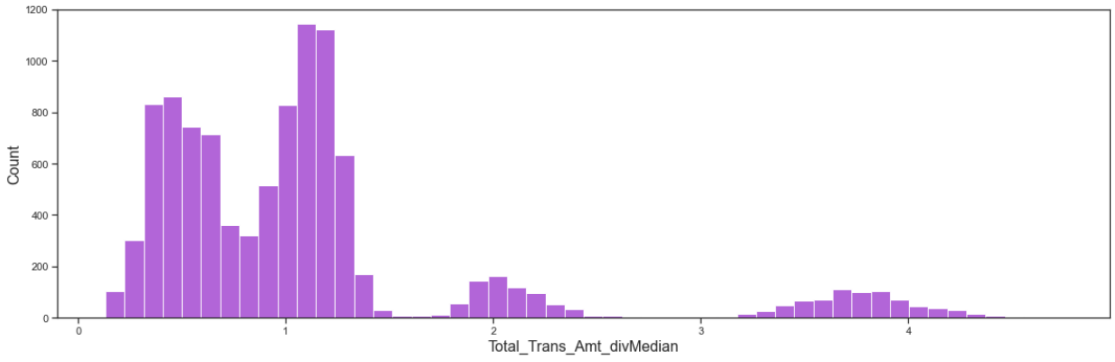
Applied dividing median scaling on 6 features, and those are "Credit_Limit" and new columns created with"_divMedian" suffix.

- Total_Revolving_Bal
- Avg_Open_To_Buy
- Total_Trans_Amt
- Total_Amt_Chng_Q4_Q1
- Total_Ct_Chng_Q4_Q1

Found these features Skewed to the left.

- Credit Limit -Credit Limit on the Credit Card
- Total_Trans_Amt -Total Transaction Amount in Last 12 months
- Total_Amt_Chng_Q4_Q1 - Ratio of the total transaction count in 4th quarter and the total transaction count in 1st quarter
- Total_Trans_Amt - Very interesting, right skewed, bimodal with outliers
- Avg_Open_To_Buy - The amount left on the credit card to use





Log transformation

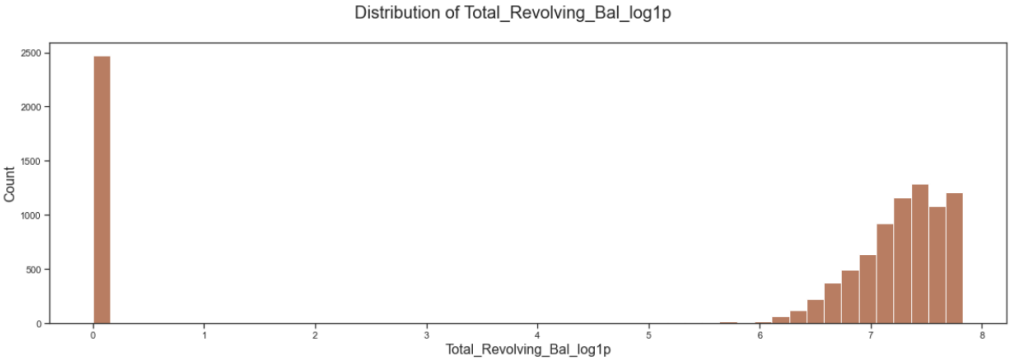
Log transformation is applied for these 3 column and new columns are created with suffix

"_log"

- Credit_Limit
- Avg_Open_To_Buy
- Total_Trans_Amt

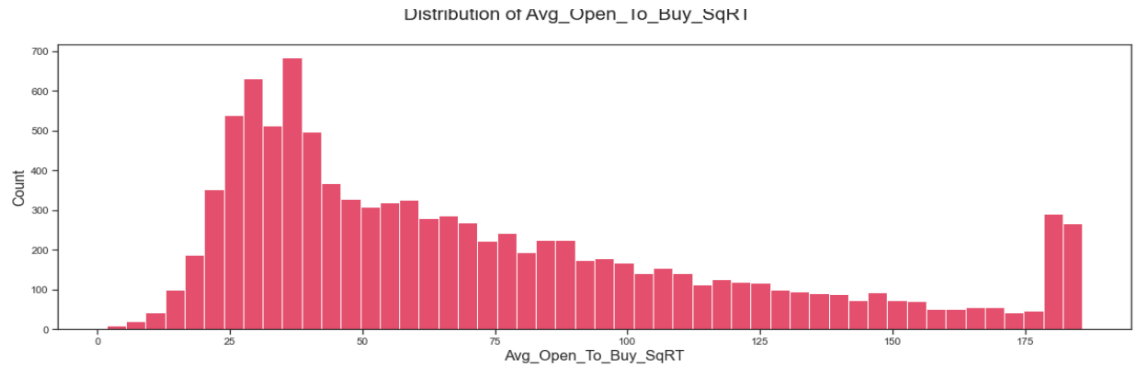
for Total_Revolving_Bal as this has the "0" values, log1p scaling is applied.

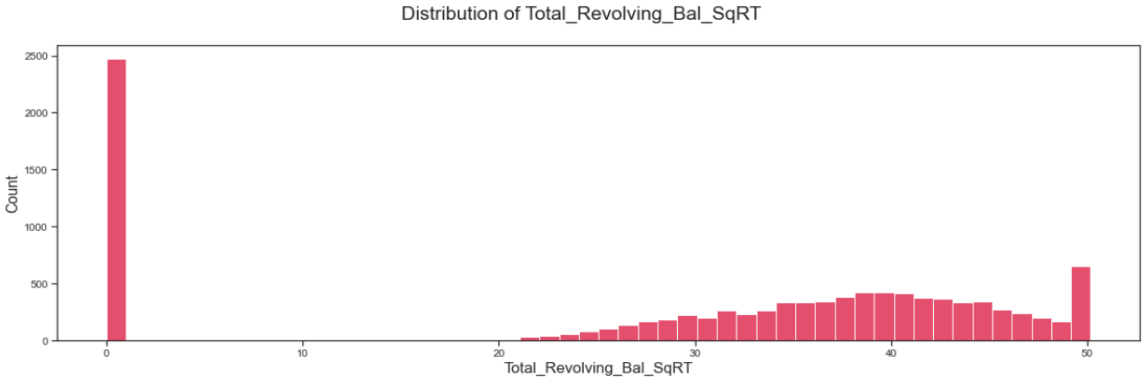
- Credit_Limit: Credit Limit on the Credit Card - slightly left skewed, with spike on extremes.
- Avg_Open_To_Buy: The amount left on the credit card to use - right skewed with outliers.



Square root scaling

This is applied to "Total_Trans_Amt","Credit_Limit","Total_Revolving_Bal","Avg_Open_To_Buy","Total_Trans_Ct" and new columns are created with suffix "_SqR"





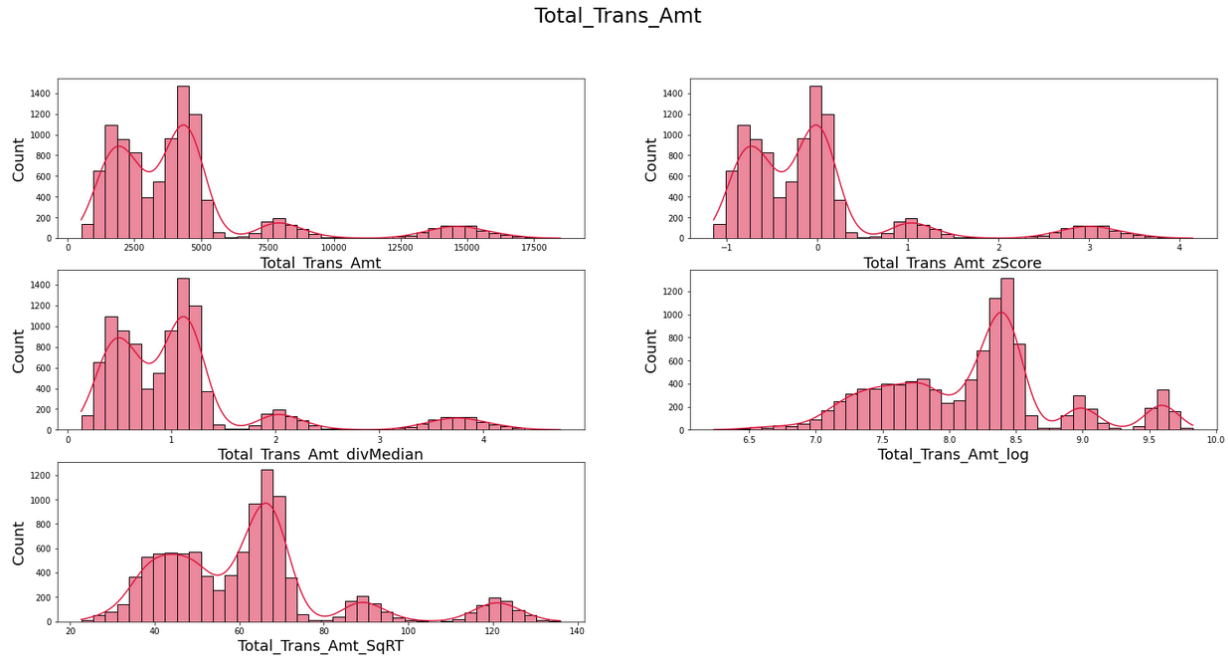
All together 25 new columns are added, making 48 columns in total. These may be useful in knowing better insights like outliers on values of these columns in EDA. No columns are dropped.Over all there are no missing values, however Education level and Marital status have unknown values. These unknown values are fixed with their "Mode". The target is identified as "Attrition_Flag".

Feature Gender is binary encoded as well.

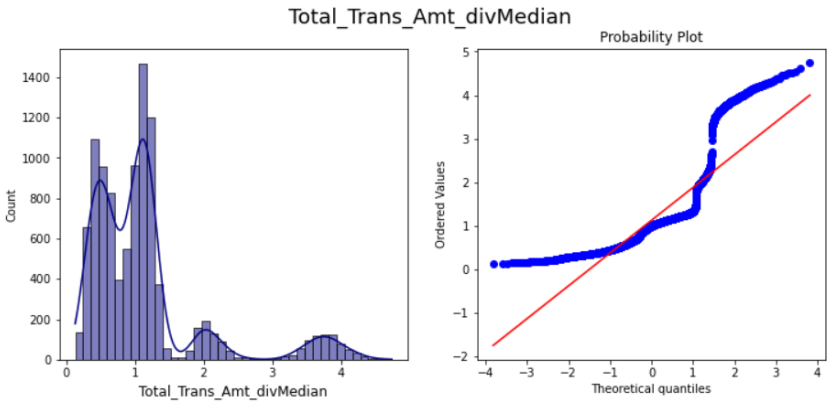
Exploratory Data Analysis

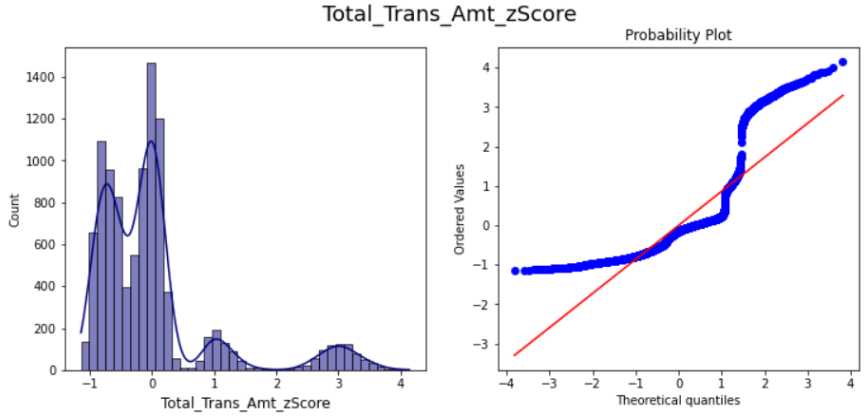
At this Exploratory Data analysis step feature relationship is evaluated.The features that are likely to have the most impact in modeling based on relationships between the features and the response variable are identified. SCaled features are compared with original features with help of hist plot. Pearson correlation coefficients were used to identify statistical relationship strengths.

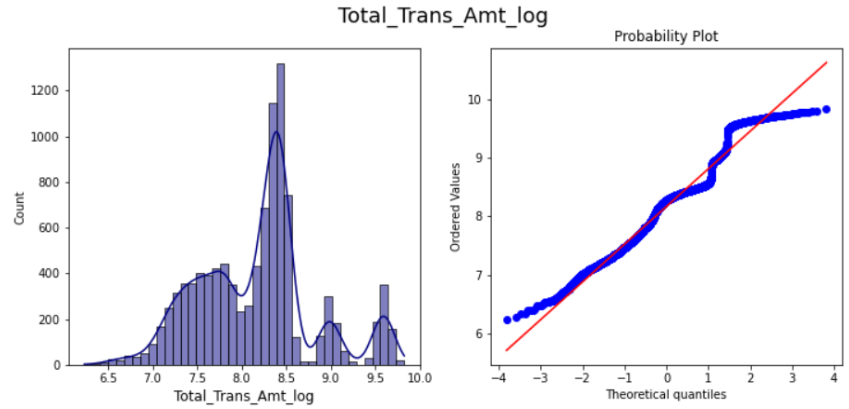
Scaling is visualized with distribution plot one sample is below -

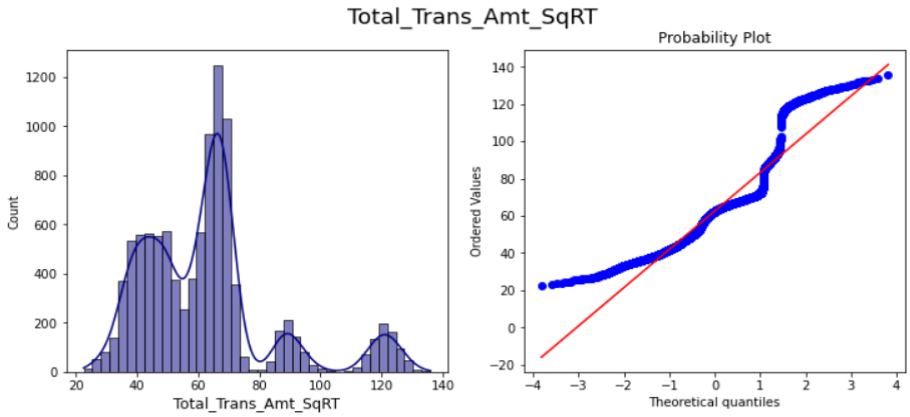


Features are verified for Gaussian or Normally Distributed using Q-Q plot. Here is the plots for all scaled features on total_trans_amt









After Comparing each scaled feature with the original and with the help of Q-Q plot, below scaled features are selected for further exploration.

- Customer_Age_zScore
- Credit_Limit_log or div Median
- Total_Revolving_Bal_divMedian
- Avg_Open_To_Buy_log or div Median
- Total_Amt_Chng_Q4_Q1_divMedian
- Total_Trans_Amt_log or zScore
- Total_Trans_Ct_zScore
- Total_Ct_Chng_Q4_Q1_divMedian
- Months_on_book_zScore
- Total_Amt_Chng_Q4_Q1_divMedian

Rest of the features are kept as is. No scaling is applied And those are

- Gender_Encoded
- Dependent_count
- Education_Level_sorted
- Marital_Status_sorted
- Income_Category_sorted
- Card_Category_sorted
- Months_on_book_zScore
- Total_Relationship_Count

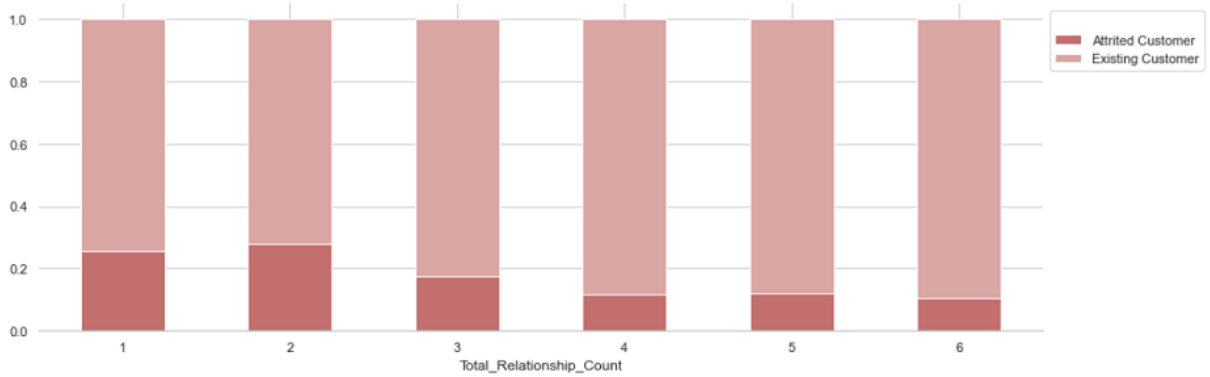
- Months_Inactive_12_mon
- Contacts_Count_12_mon
- Avg_Utilization_Ratio
- naive_cls1
- naive_cls2
- CLIENTNUM

Stacked plot : Used this analysis the Attrited and existing customer ratio on each feature.

With stacked plot found

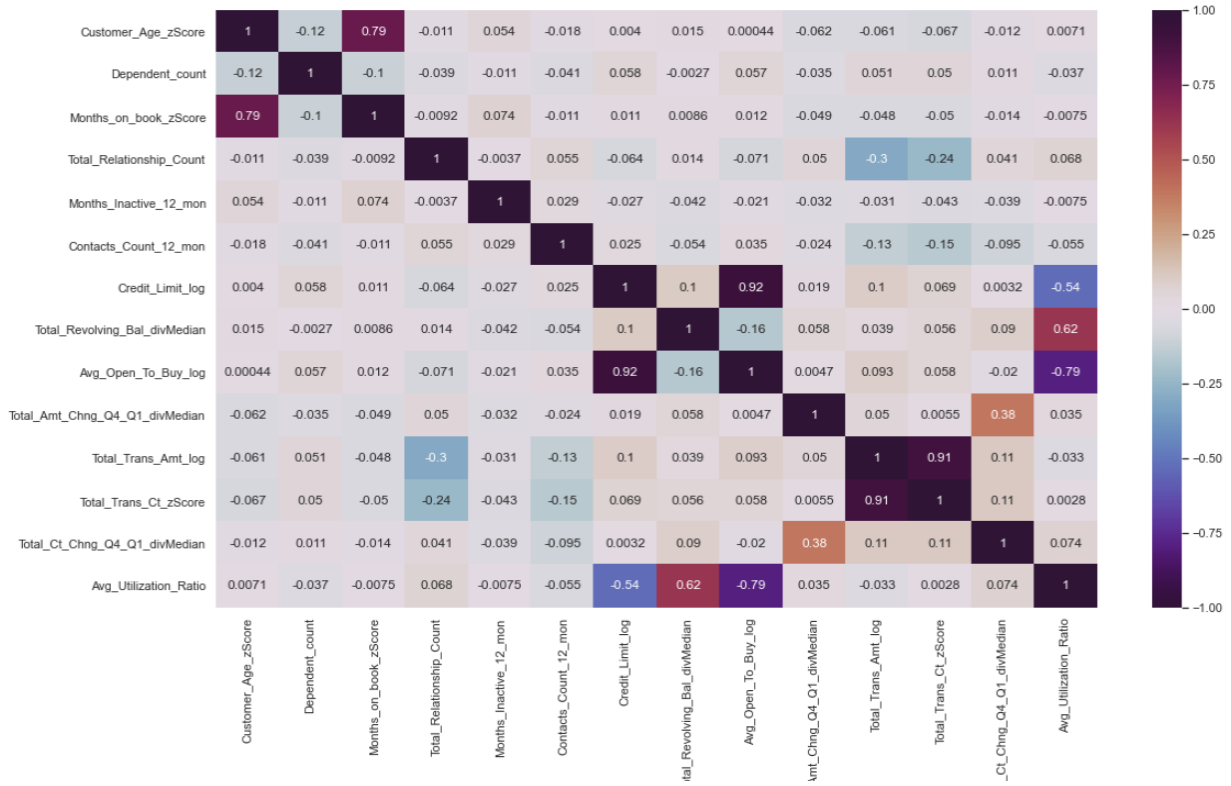
- Customers who earned more than 120k and less than 40k.
- Customers with 3 dependent attrited more.
- Customers who were single attrited more
- Customers having 1 or 2 bank products churned more customers compared to customers with more bank products.
- Customers who were doctorate or postgraduate attrited most.
- Customers with platinum cards churned more but there are only 20 samples so this is inclusive.
- Customers who were never inactive churned most.Customers who were inactive for 4 months attrited most followed by 3 month and 5 month.
- customers who were contacted most in the last 12 months attrited.Did bank had any information about there attrition which was a reason bank was contacting those customers so many times.? or so much contact from the bank leads to attrition.

Sample plot



Attrition_Flag	Attrited Customer	Existing Customer	All
Months_Inactive_12_mon			
0	15	14	29
1	100	2133	2233
2	505	2777	3282
3	826	3020	3846
4	130	305	435
5	32	146	178
6	19	105	124
All	1627	8500	10127

Feature Correlation heat map



From the heat map -

- As expected there is very high correlation between total transfer amount and total transfer count. Total transfer count is dropped.
- Credit limit and Average open to buy is fully correlated, Average open to buy is dropped..
- Total_Trans_Ct and Total_Trans_Amt are close for fully correlation,.91, Total trans count is dropped.
- It is also logical that Total_Trans_Amt is correlated to Total_Amt_Chng_Q4_Q1,total ct_change_Q4_Q1. Total_Amt_Chng_Q4_Q1,total ct_change_Q4_Q1 are dropped.

Customer age and Time frame with bank are highly correlated.Though Female customer attrited more compared to male, but not much difference.Customers who were doctorate or postgraduate attrited most.Graduate and high schoolers stayed.On The contrary Customers who earned more than 120k attrited more. As expected there is very high correlation total transfer amount and total transfer count.Customer having 2 or 3 bank product attrited more compared to other customers with more bank products.Total transaction Amount has different distribution with data between 0 -2500 , 2500-5000, 750-10000 and then 12500-17500. To a surprise, customers who were contacted most in the last 12 months attrited.

After feature scaling the final list of features for the pre-processing looked follows and Target is 'Attrition_Numeric'

- Customer_Age_zScore
- Credit_Limit_log
- Total_Revolving_Bal_divMedian
- Total_Trans_Amt_log
- Avg_Utilization_Ratio
- Gender_Encoded
- Dependent_count
- Education_Level_sorted
- Marital_Status_sorted
- Income_Category_sorted
- Card_Category_sorted
- Months_on_book_zScore
- Total_Relationship_Count
- Months_Inactive_12_mon
- Contacts_Count_12_mon

Pre-Processing

In the Pre-processing I applied the scaling best found from the EDA as follows

- Z scoring for "Customer_Age","Months_on_book"
- Div median for 'Total_Revolving_Bal'
- Log scaling for "Credit_Limit","Total_Trans_Amt"
- Test and Train sets are created with 80:20 ratio.
- One hot encoding is applied

Pair plot

Modeling

Steps followed to build model are -

1. Set the model object, Pipeline, cross-validator, etc.
2. Evaluate the fit on the training data (make sure everything is working,
3. is the metric acceptable?)
4. Pick the threshold (using the training data)
5. Evaluate on the test data
6. Make sure to compare train and test results (generally perform worse on test)

Model used are

1. naive model
2. Logistic Regression
3. Random Forest Classification
4. KNeighbours Classification
5. XGBoost Classification

Evaluation Metrics

Thinking of customers will not exit but he does, this means income loss for the Bank. Banks need to take action steps for this scenario. Banks are looking for recall to be maximized, greater the recall lesser the chances of false negative means lesser chances of predicting customers will not exit where in reality they do. Since classes in the data are unbalanced, I can make use of the confusion matrix to examine the outcome of the model. Recall and precision metrics can be calculated from the confusion matrix, and this would help me assess the models.

This project aims to predict potential churn customers, and it is realized that the client cost of mistakenly classifying non-churn customers as churn may be high in practice because banks would not want to lose valuable customers, and the banks would like to identify churners at their best efforts as well. Thus, it would be useful to consider Recall. To complement this Receiver Operating Characteristic curve (ROC) is used. ROC is a plot of True Positive Rate (TPR) against False Positive Rate (FPR). This means I can consider TPR and FPR simultaneously, by making use of the area under the curve (AUC) of ROC.

TPR or Recall or Sensitivity tells us what proportion of the positive class got correctly classified. TPR = TP/(TP+FN)

FPR tells us what proportion of the negative class got incorrectly classified by the classifier. FPR = FP/(TN+FP)

All models are used to predict the actual class of the data point by predicting its probability of belonging to different classes. This gives us more control over the result.And ROC_AUC threshold is used to interpret the result of the classifier.

Model used are

1. Logistic Regression
2. Random Forest Classification
3. KNeighbours Classification
4. XGBoost Classification

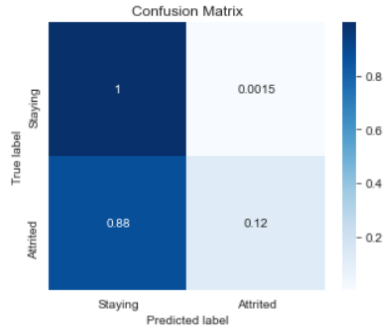
Steps followed to build model are

1. Set the model object, Pipeline, cross-validator, etc.
2. Evaluate the fit on the training data (make sure everything is working,
3. is the metric acceptable?)
4. Pick the threshold (using the training data)
5. Evaluate on the test data
6. compare train and test results

Logistic Regression: Metrics Evaluation

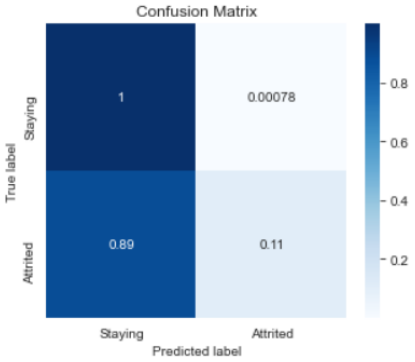
LG classification report on train set

	precision	recall	f1-score	support
Staying	0.86	1.00	0.92	5949
Attrited	0.94	0.12	0.21	1139
accuracy			0.86	7088
macro avg	0.90	0.56	0.57	7088
weighted avg	0.87	0.86	0.81	7088



LG classification report on test set

	precision	recall	f1-score	support
Staying	0.85	1.00	0.92	2551
Attrited	0.96	0.11	0.20	488
accuracy			0.86	3039
macro avg	0.91	0.56	0.56	3039
weighted avg	0.87	0.86	0.81	3039

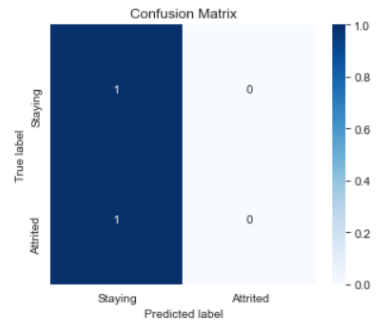


This model is generalized better on train set and test set.However roc_AUC score 0.75, means it can 75% chance of identifying default and non-default class.Recall perfect for defaults, precision is 0.85. only. Indicating this model can not predict the defaults correctly, but non-deaults are predicted as defaults. Meaning those who are actually attrited are predicted as existing. However it is better than naive model.Can identity majority class much better, but still not to the expectation.

Random Forest Classification: Metric Evaluation

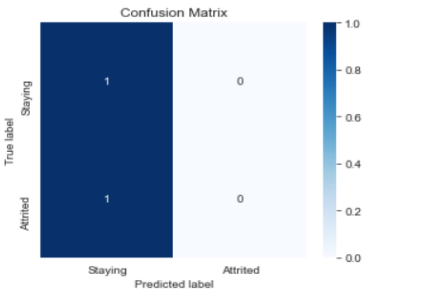
RF classification report on train set

	precision	recall	f1-score	support
Staying	0.84	1.00	0.91	5949
Attrited	1.00	0.00	0.00	1139
accuracy			0.84	7088
macro avg	0.92	0.50	0.46	7088
weighted avg	0.87	0.84	0.77	7088



RF classification report on test set

	precision	recall	f1-score	support
Staying	0.84	1.00	0.91	2551
Attrited	1.00	0.00	0.00	488
accuracy			0.84	3039
macro avg	0.92	0.50	0.46	3039
weighted avg	0.87	0.84	0.77	3039



This model performed poorly in comparison to logistic regression. Same as naive model.

Grid search may yield better results with hypertuning params, can also try feature importance and apply top 3 or 5 features to get better results. Though recall is 1, +ve predictive rate precision is 0.84. Indicating 84% chance of correct prediction on defaults to non-defaults.

Hyper param tuning did not play the role, it is the same as naive model.

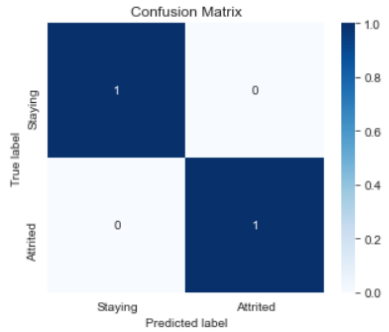
KNeighbours Classification: Metric Evaluation

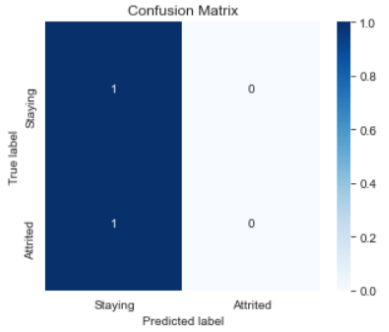
KN classification report on train set

	precision	recall	f1-score	support
Staying	1.00	1.00	1.00	5949
Attrited	1.00	1.00	1.00	1139
accuracy			1.00	7088
macro avg	1.00	1.00	1.00	7088
weighted avg	1.00	1.00	1.00	7088

classification report

	precision	recall	f1-score	support
Staying	0.84	1.00	0.91	2551
Attrited	1.00	0.00	0.00	488
accuracy			0.84	3039
macro avg	0.92	0.50	0.46	3039
weighted avg	0.87	0.84	0.77	3039





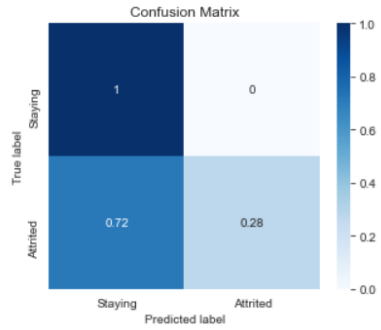
KNeighbours is performing same as Random Forest on Test set. However Max roc_auc is perfect 1 on train set,meaning model can identify between all the defaults and the non-defaults points correctly. This is supported with confusion matrix of train set. However on applying to test set it is visible False -ve is 1, and true -ve is 0. can be seen with precision 0.84 Hyper param tuning did not play the role, it is same as naive model.

It is potential that all may be predicted as true +ves meaning, there is attrition, Actually customers are leaving.

XGBoost Classification: Metric Evaluation

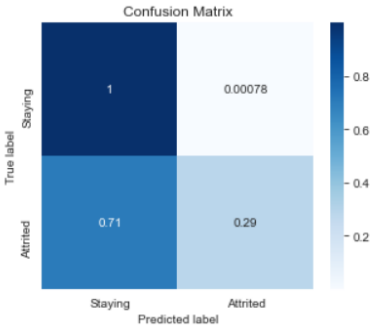
classification report

	precision	recall	f1-score	support
Staying	0.88	1.00	0.94	2551
Attrited	0.99	0.29	0.45	488
accuracy			0.89	3039
macro avg	0.93	0.64	0.69	3039
weighted avg	0.90	0.89	0.86	3039



classification report

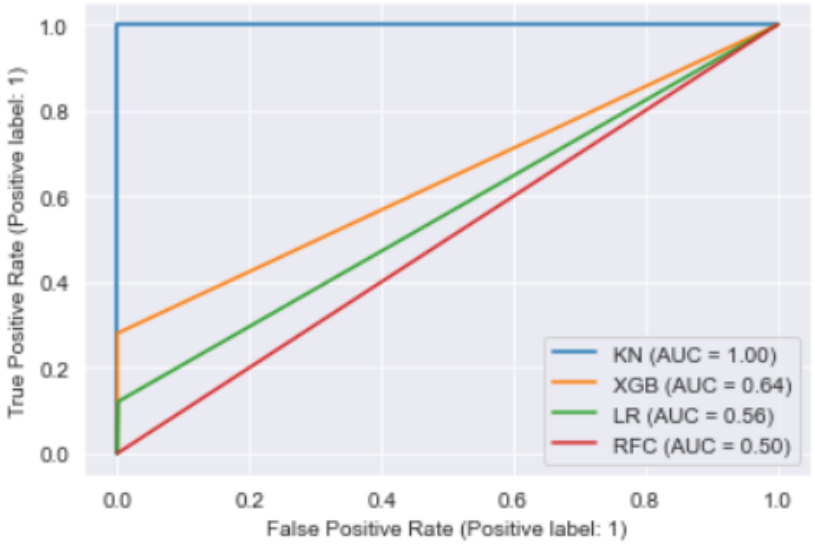
	precision	recall	f1-score	support
Staying	0.88	1.00	0.94	2551
Attrited	0.99	0.29	0.45	488
accuracy			0.89	3039
macro avg	0.93	0.64	0.69	3039
weighted avg	0.90	0.89	0.86	3039



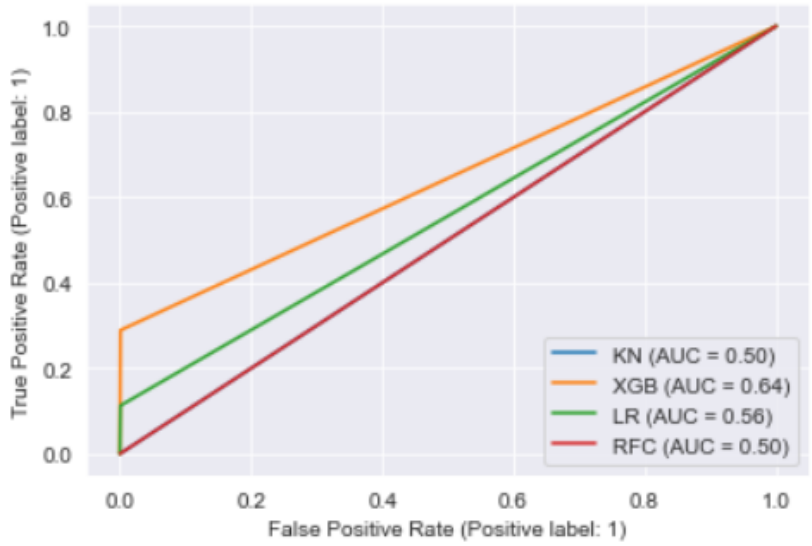
This is generalized better compared to all modeland performed fairly well, as false -ve rate is 0.7 compared with logistic and random forest on both on train and test set. And I can see true +ve rate stayed 1, with recall 1 and improved precision 88%.generalized. This brings a model that can fairly make mistakes in unseen data, in comparison to other models.

ROC_AUC Curve Evaluation

On train set



On test set



- Logistic regression is performing well on both training and test sets. With a score of 0.55 This is generalized better.
- KNeighbors performed very badly on train is 1.0 and on test data with roc_auc 0.71. It is overfitting.Meaning it can perfectly identify between all the Positive and the Negative class points correctly on train set.However on test data there is a chance KN identify the defaults from the non-defaults class values
- RandomForest has both train and test roc_auc is 0.5. It is the same as naive. Model does not have the ability to predict defaults and non-defaults.
- XGboost has done fairly well in terms of roc_auc performance with slight difference between train and test roc_auc, those are 0.64 and .64.scores are almost close.Same as Logistic regression, generalized better.

Summary

- XGBoost and Logistic regression is performing well on both training and test set.
- RandomForest is the same as naive. Model does not have the ability to predict defaults and non-defaults.
- KNeighbors is overfitting

Additional Models

- It would be interesting to see the results of other models, meaning trying different algorithms
- Making combination or hybrid models, e.g. RF + TensorFlow
- Trying feature crosses like dmatrix from patsy
- Clustering or association rules