Bank Credit Card Customer Churn Prediction and Analysis

Introduction

Customer churn is a major concern for big businesses. Losing a customer is awfully expensive for any business. The full cost of churn includes both lost revenue and marketing costs involved in replacing them with new ones. Identifying unhappy customers early on will help business win those customers with incentives, free trials promotions, and advertising.

This project uses Machine Learning to identify those who will churn so action can be taken well in advance to retain those customers. Also, it investigates the factors that affect customer churning

Data Wrangling

The data consists of bank customers who own credit card. It has about 10,000 customers and 23 features describing each user. It includes Client number, Attrition Flag, Customer Age, Gender, Dependent count, Education Level, Marital Status, Income Category, Card Category, Months on book, Total Relationship Count, Months Inactive 12 mon, Contacts Count 12 mon, Credit Limit, Total Revolving Bal, Avg Open to Buy, Total Amt Chng Q4 Q1, Total Trans Amt, Total Trans Ct, Total Ct Chng Q4 Q1 and Avg Utilization Ratio.

This data is obtained from Kaggle at the link https://www.kaggle.com/sakshigoyal7/credit-card-customers

Here is the description of few metrics in the data which are not trivial.

Revolving Balance: Portion of credit card spending that goes unpaid at the end of a billing cycle

Average Open to Buy: The difference between the credit limit assigned to cardholder account and the present balance on the account

Utilization Ratio: How much customer owes divided by credit limit expressed in percentage

Relationship Count: Number of products held by the customer

Amounts change Q4_Q1: Change in transaction amount in Q4 over Q1

Months on book: Number of months the customer stayed with the bank

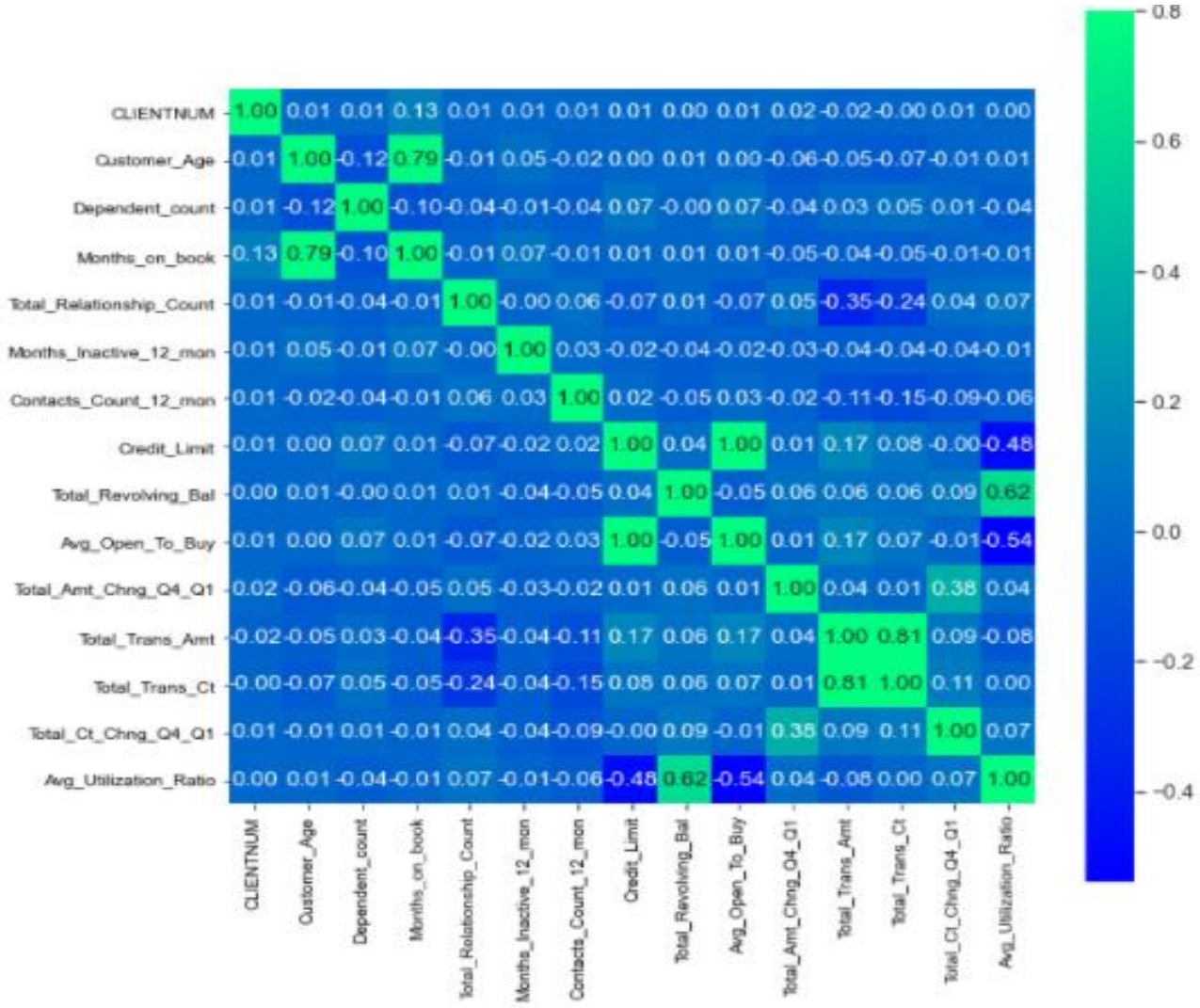
Exploratory Data Analysis

The churned customers are very few in the dataset. 83.93% of customers remain with the bank and attired customers make only 16.07%, so this is highly skewed.

Each customer is uniquely identified by Client Number. Customer Age ranges from 25 to 70, with most of them between 40 and 55. They are sorted into categories.

Outliers are noticed on these columns Customer_Age, Months_on_book, Months_Inactive_12_mon, Contacts_Count_12_mon, Credit_Limit, Avg_Open_To_Buy, Total_Trans_Amt, Total_Trans_Ct, and Total_Ct_Chng_Q4_Q1. There is a strong relation between gender and income category.

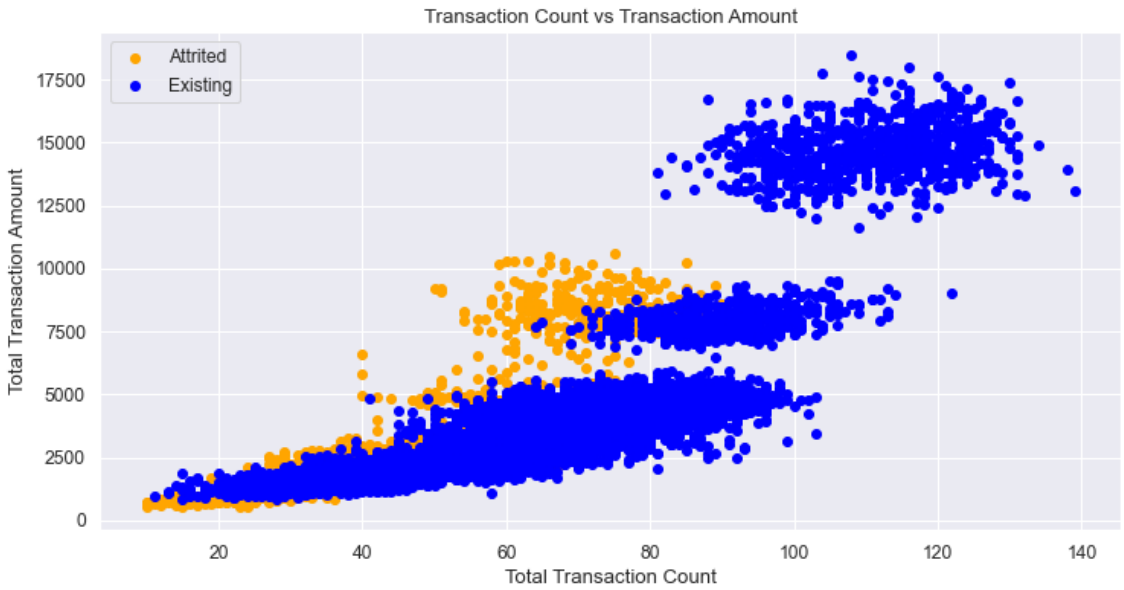
Heat Map indicating correlation of numerical attributes



Avg_open_to_buy and credit limit are highly correlated. Similarly, customer age and Months on book. Total_trans_count and Total_trans_amt are also highly correlated. Avg_utilization_ratio and Total_Revolving_bal also seem to be correlated, but we will verify further with the numbers and behavior.

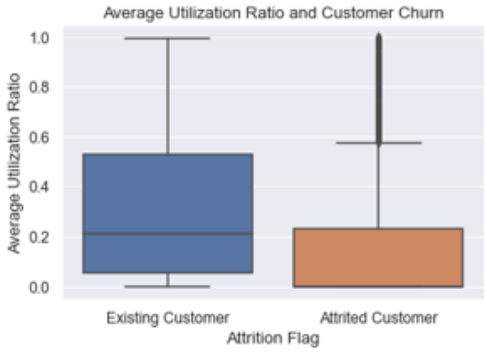
Transaction Count and Transaction Amount

Customer s who left the bank had fewer transaction counts than existing customers.



Customers who had higher transaction amount of at least 12000 seems to have stayed with the bank. They are satisfied customers who enjoy more operations with the bank. This seems to be a promising attribute to the model.

Revolving Balance and Average Utilization Ratio



There is more concentration of Revolving balance and utilization ratio at the lower levels for churned customers than existing customers. Existing customers have at least $500 as their minimum balance and their utilization is also exceptionally low.

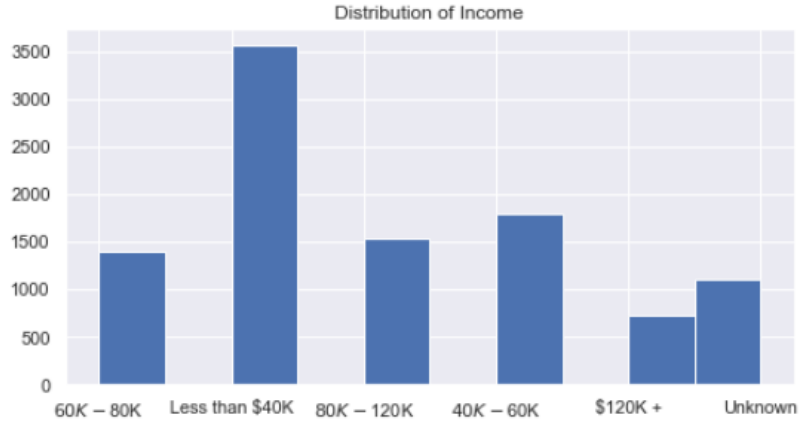
Education and Customer Churn

Customers who had college education and graduates have slightly higher rate of staying than doctorates. But the difference is extremely low. We cannot generalize any relation between churned customers and education. We can only conclude that there is higher percentage of customers who are graduates in the bank.

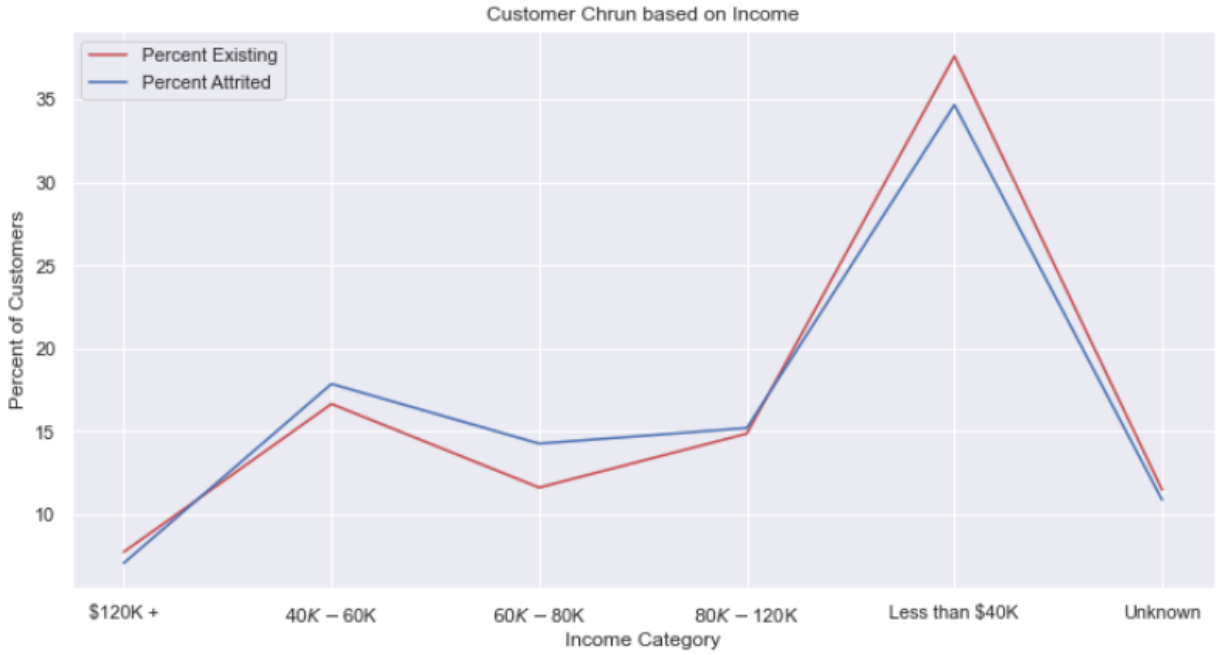
	Education	% Attrited Customer	%Existing Customer
0	College	9.47	10.11
1	Doctorate	5.84	4.19
2	Graduate	29.93	31.07
3	High School	18.81	20.08
4	Post-Graduate	5.65	4.99
5	Uneducated	14.57	14.71
6	Unknown	15.73	14.86

Income and Customer Churn

The bank has more customers who earn less than $40K than all other categories.



Hence, the impact of customer churn on the Income was compared based on the percentage. This was calculated based on the number of customers churned for a given Income category. For example, 7.74 customers belong to Income Category $120K if total number of churned customers is 100. Out of total churned customers 7.74% belong to category $120K+



This represents that 40K-80K Income range has higher percentage of staying with bank that lower income. Customers whose income is $120K+ has slight difference in percentage which is not reliable. So, the bank must try to attract customers in 40K to 80K range more.

Credit Limit and Utilization Ratio

The average utilization ratio is one of the indicators to show that customer is churning, most of the churned customers have lower credit limit and utilization ratio.

This plot shows the effect of churn on credit limit.

It implies that the median of credit limit of existing customer is almost the same that of churned customer. To check if this happened by chance, verified using hypothesis testing.

Null Hypothesis: Mean of credit limit of churned customers equal to mean of credit limit of existing customers.

Performed statistic test and obtained a p-value of 0.016. Since p-value is less than 0.5, we can reject the null hypothesis concluding that mean of credit limit of both sets are not equal.

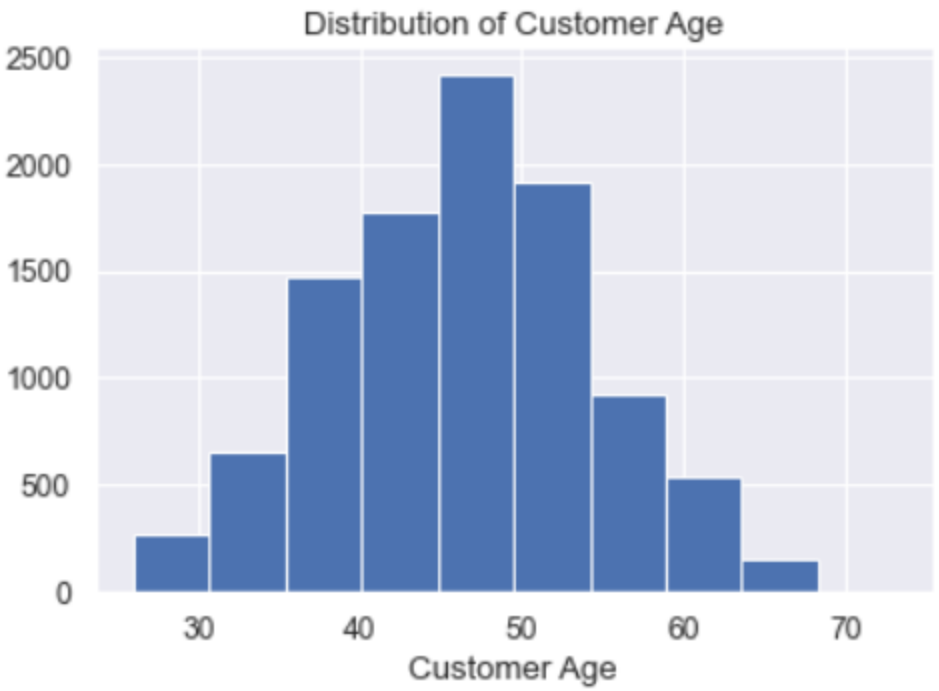
Age and Customer churn

The mean age of churned customer is slightly higher than that of existing customer.

```
Churned customer Mean Age:  46.65
Existing customer Mean Age:  46.26
```

Can we generalize this? Verified with a statistics test.

The distribution of customer age in the dataset is normal. Most of the values are between 40 and 55.

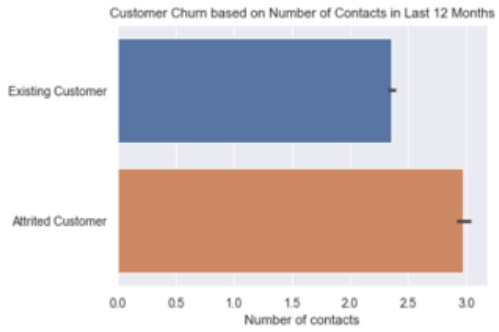


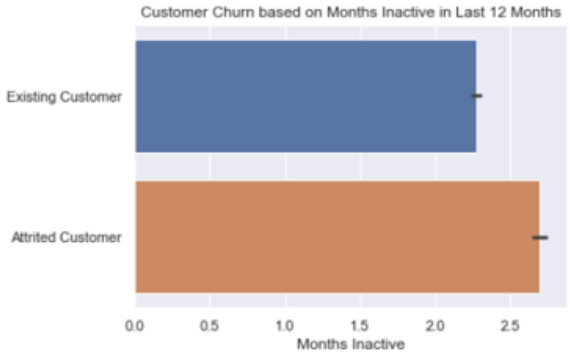
Null Hypothesis: Average age of people who have churned is equal to avg age of people who have not churned.

P-value obtained for this is 0.067, which indicated that we fail to reject the null hypothesis that the average age of 2 sets is equal.

Number of Contacts and Months Inactive

We would expect the customer contacts would be higher if they stay with bank. But there is an optimum level. As seen below, if the number of contacts exceed 2 the bank must be cautious and look for any churning signal. It could also be possible that they are frustrated and calling often about a problem.

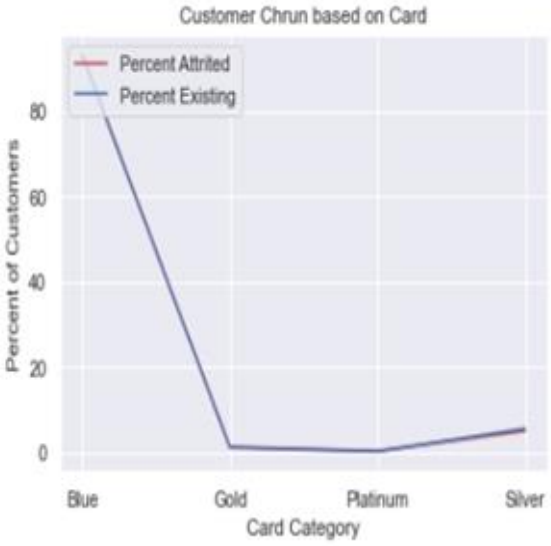


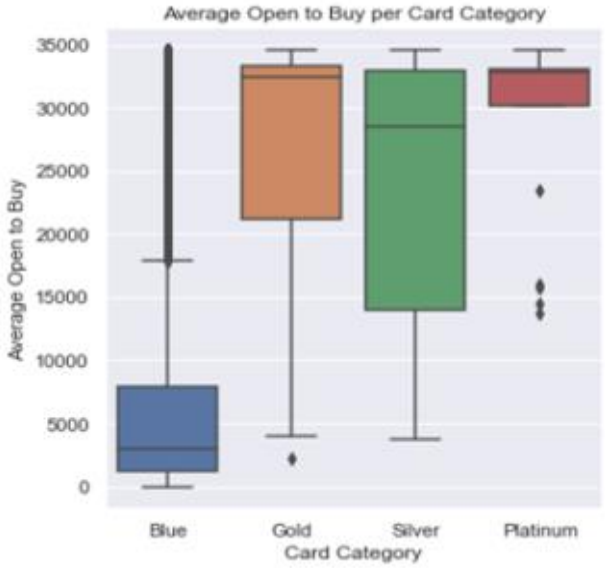


The same can be observed if they do not have any activity with the bank for more than 2 months.

Card Category and Average Open to Buy

Card Category did not have any effect on customer churning. But we can find that there are greater number of Blue card holders than Gold, Platinum or Silver.

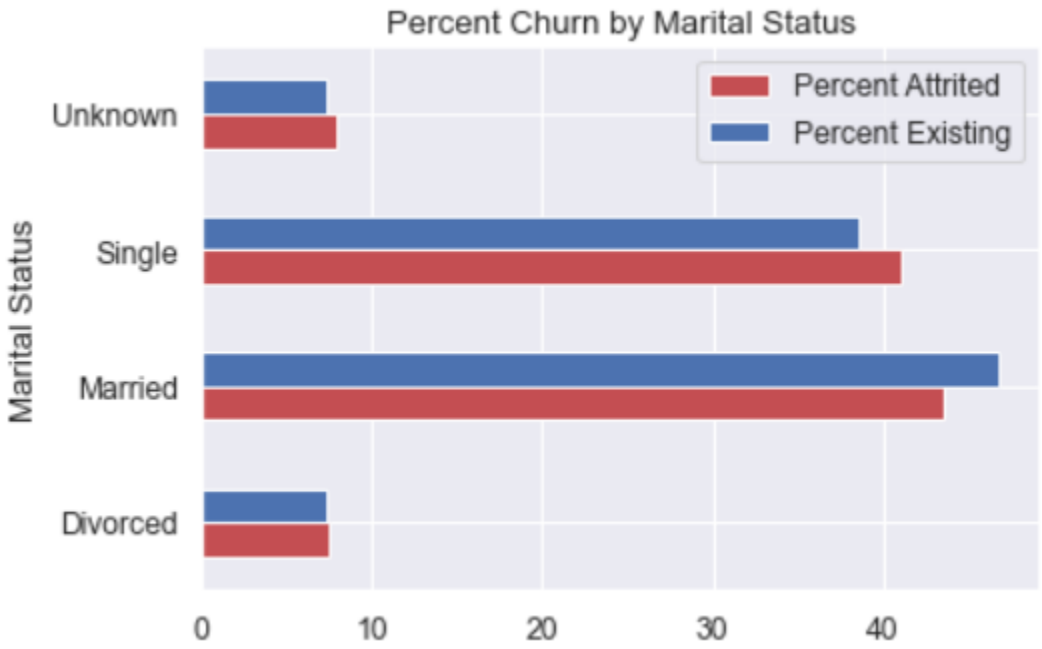




Average open to buy is higher for Gold, Silver and Platinum card members which is not surprising as they have higher Credit limit. Though there are more Blue card holders, Silver Card has higher variance which implies they have lower balance. So, they take more loans.

Marital Status and Customer Churn

There are more married customers in the bank than single and divorced.

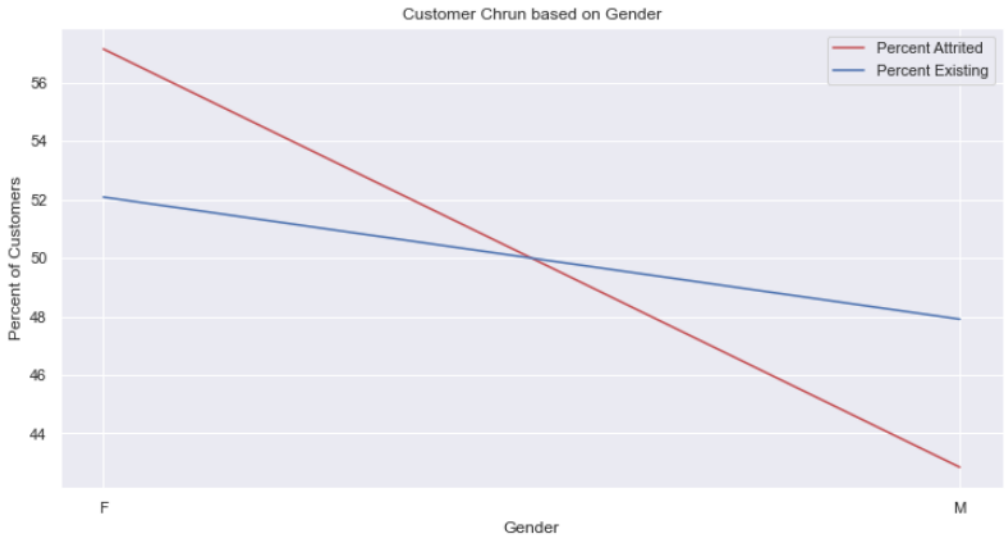


Large percentage of married customers are loyal to the bank. These customers probably have a stable income and stay in the same location. Whereas there us higher percentage of Single customers who have churned.

Gender and Customer Churn

There are more female customers than male customers in the bank.

```
F    5358
M    4769
Name: Gender, dtype: int64
```



There is higher percentage of female customers who churn compared to male customers. Here the percentage is calculated based on total churned customers. For example, 52% of customers who are churned are male.

Modelling

The categories Gender, Income, Card, Education and Marital status were all one hot encoded. The training data was standardized for some models. We cannot afford to lose the customer here. If the customer is predicted that he will churn, but in real if he does not churn, it is bearable compared to other way. If we predict that the customer will not churn and if he churns, then we will lose a customer which is expensive. So, false negative is costly we are more concerned with recall.

Implemented the model using Scikit library. More details are available in the notebook for each model. The performance metrics collected are confusion matrix and classification report for each model. The best model was further verified by ROC curve and AOC value.

The highest accuracy for all models is 0.97 and the highest recall is 0.90. Over sampling helped improve the recall value by a small percentage.

This table explains the outcomes of all models.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.89	0.81	**0.48**	0.60
Random Forest	0.96	0.92	**0.86**	0.88
SVM	0.91	0.78	**0.68**	0.72
KNN	0.86	0.70	**0.26**	0.37
Gradient Boosting	0.97	0.93	**0.89**	0.91
Over Sampling with Gradient Boosting	0.96	0.89	**0.90**	0.89

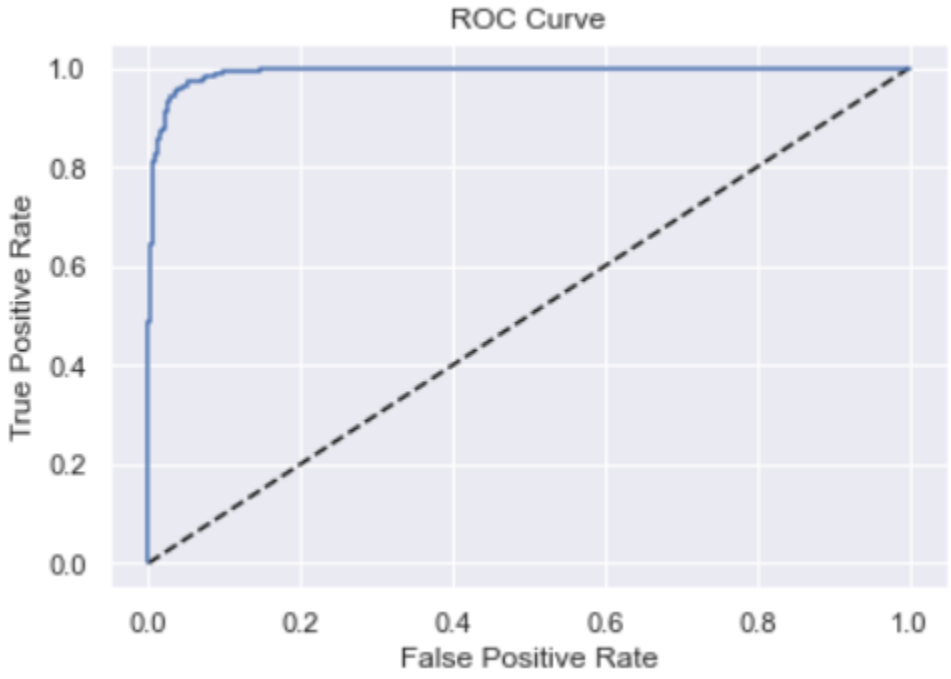
Gradient Boosting is the best model looking at the recall value. It is even better after over sampling using SMOTE as the data set was imbalanced. The hyper parameters that worked best for Gradient Boosting model are

learning_rate:0.1

max_depth:4

n_estimators:1000

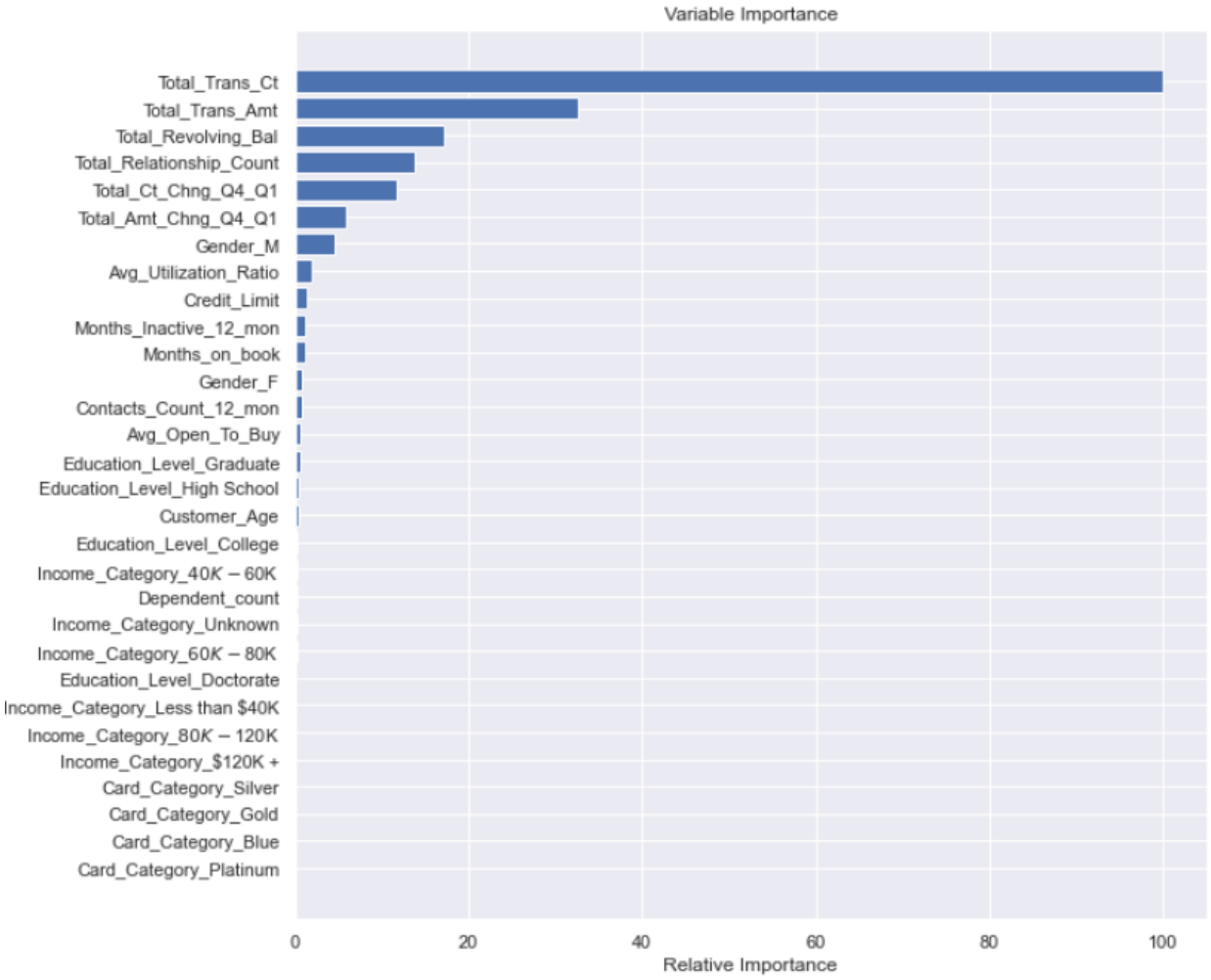
The ROC curve explains the relation between True positive and True Negatives for our model.



The area under the ROC curve is **0.99246**

Feature Importance

Important features were plotted for best model Gradient Boosting.



Total transaction Count and Amount seems the most important feature for predicting the customer churn.

Recommendation

The Bank can attract new male customers who earn between $40K and 80K annually by offering free trials as they tend to stay longer. Additional incentives can be given to them if they fall in the age range 45 to 55 and married. A new branch can be opened in the area where all the above criteria is met. These were observed when the data was explored.

But these factors contributed most to customer churn in order in our best model. Total Transaction Amount, Total Transaction Count, Total Revolving balance, Toral Relationship Count, Total Count change between Q4 and Q1, Total Amount change between Q4 and Q1, Gender, Avg Utilization ratio, Credit Limit.

Future Improvements

We can remove some more outliers and columns that are linear and investigate further. This can be combined with additional data of other products that the customers use in the same bank and cluster their behavior for further marketing.