Credit card customer churn in a bank

Ranidu Fernando
S14429
6/14/22

Abstract

With this economic development, more and more people start using credit cards. In the USA, it has been identified as the second most popular non-cash instrument in 2003. The use of credit cards has already become a convenient way to expand purchasing power. Credit cards provide benefits to customers and merchants that are not provided by other payment instruments. But many banks eventually face situations where their customers have decided to leave the credit card service. This will lead to a loss for the bank. Therefore credit card customer churn analysis is taken an important role. By the results, banks can take corresponding actions to retain customers. In this study, descriptive analysis and implementation of a supervised machine learning model to predict the credit card customer churn were carried out. Here the fitted supervised machine learning model is Logistic regression. Further data preprocessing techniques, oversampling techniques, and feature selection techniques were used and trained the dataset for the above model and test its accuracy. The findings of this study are the majority of cardholders are female customers, most popular credit card type is the blue card, most of the customers' education level is Graduate, married customers tend to have a credit card more than single customers and the majority of credit card holders' income is less than $40K, Attrited customers' total transaction count is less than the existing customers' total transaction count. Half of the attired customers' total revolving balance is zero. Other findings are gender, number of dependents, education level, total relationship count, number of months inactive and number of contacts, total transaction count and amount, card type, and marital status are the variables that are most associated with credit card customer churn. An interesting finding is customer age and their income level are not much associated with credit card customer churn.

Contents

Introduction

Background of the study

Nowadays with increasingly intense competition in the market, major banks pay more attention to customer relationship management. With this economic development, more and more people start using credit cards. A real-time effective credit card holders' churn analysis is important and helpful for bankers to maintain credit cardholders. If bankers could predict who is going to get churned, they can proactively go to the customer to provide them a better service and turn customers' decisions in the opposite direction. From this analysis, expecting to conduct a descriptive data analysis and an advance analysis to predict the credit card customer attrition.

Objectives

To identify social, economic, and demographical attributes of credit card customers.

To identify the factors associated with credit card customer churn.

To implement a machine learning model to predict the credit card customer churn.

Significance of the study

There are many inactive customers who rarely or stop using their credit cards, or rather, they are attrition accounts. Customer churn represents the loss of customers or clients as they stop using some products or services. An important reason for customer churn analysis is that the cost of developing a new customer is much higher than that of retaining an existing customer. Typically, it costs up to five times as much to make a new sale to a new customer as it does to make an additional sale to an existing customer. Analyzing this helps bank managers to develop a sound marketing strategy to retain quality customers.

Literature review

As the economy develops a large number of credit cards are issued. But many of the card holders are not active (or called churn holders). With increasing bank competition customers are able to choose among multiple service providers and easily exercise their right of switching from one service provider to another. If banks can predict future behaviors before the customers close their account or stop using the card to pay, they can market to retain these customers. The main purpose of the research paper was not to provide a new data mining algorithm that predicts credit card customer churn, but to focus on the application of the churn prediction, to provide a framework for understanding the knowledge of the card holders' hidden patterns using a dataset of a Chinese bank. 41 variables are selected for the study including basic information of cardholders, basic information related to the credit cards, risk related variables and transaction variables. The researcher used a balanced random sample of fifty percent of the churners (215 samples) and five percent of the non-churners (244 samples) to be a training set. 4997 customers would be a test set to validate the model and test the predicted performance of the model. Two models were built using Logistic regression and decision tree algorithm. In order to find the power of different variables, the researcher build models with different variable combinations because all 41 variables are there and it would mislead the model if put all the designed variables into the algorithms.

The best fitted model of this paper is based on logistic regression, which includes two customer personal information variables, four card basic information variables, three risk information variables and six transaction information variables. The test results of the model have shown that the logistic regression performs better than the decision tree.

The study has taken a small sample to train the dataset but better if it would take a large sample.

From the result, the logistic regression model will give better predictions for the analysis.

Theory and methodology

Data wrangling is a process that should be conducted before starting an analysis or model fitting. One of the data wrangling techniques is data cleaning. Data cleaning can be applied to remove noise and correct inconsistencies in data by filling in missing values, smoothing noisy data, and identifying or removing outliers. A common method of filling missing values is using the mean, median, or mode of the data.

Data visualization simply is the creation of visual representations of data. The purpose of these representations is to clearly communicate insights from data through charts and graphs. A graph can be used to present data that are too numerous or complicated to be described adequately in the text and in less space.

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. A 100% stacked bar chart is a chart type designed to show the relative percentage of multiple data series in stacked bars, where the total (cumulative) of each stacked bar always equals 100%. Side-By-Side boxplots are used to display the distribution of several quantitative variables or a single quantitative variable along with a categorical variable.

Class Imbalance

When observation in one class is higher than the observation in other classes then there exists a class imbalance. Example: To detect fraudulent credit card transactions, the fraudulent transaction count is around 400 when compared with the non-fraudulent transaction count of around 90000. The problem of class imbalance is most machine learning algorithms work best when the number of samples in each class is about equal. This is because most algorithms are designed to maximize accuracy and reduce errors. If the data set is imbalanced then getting a pretty high accuracy just by predicting the majority class, but fails to capture the minority class, which is most often the point of creating the model in the first place.

A widely adopted technique for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling). SMOTE: Synthetic Minority Oversampling Technique is a technique that uses in machine learning to deal with class imbalance datasets.

Dimensionality reduction

Dimension reduction simplifies complex high-dimensional data. It summarizes data with lower-dimensional data with minimal loss of information. There are several ways to do dimensionality reduction, one is feature selection. Feature selection is the process of selecting the most important variables for fitting the model among a set of independent variables. In python, this technique is carried out using the Recursive Feature Elimination (RFE).

Train and Test data split

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves dividing the data set into two subsets, one is for training the model and the other one is for testing the model.

Logistic regression

Logistic regression is used when we have a dichotomous variable (Two levels categorical variable Ex- Yes/ No) as the response variable. In that case, cannot use linear regression for the predictions. Logistic regression estimates the probability of an event occurring, Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. When predicting some variable using this model a score will be given through this model and the probability should be found and then with that probability according to some cutoff value, the prediction will be done. EX- $P \geq$ Some cutoff $\dashrightarrow$ Assign to group 01

$P <$ Some cutoff $\dashrightarrow$ Assign to group 02

Generally, this cutoff value is 0.5.

Model evaluation can measure using some metrics like Miss Classification Error (MCE) , Accuracy(1-MCE), F1 Score and AOC value.

Methodology

The analysis was carried out in a sequence of steps using SPSS and Python. Data preprocessing using SPSS wa s the first step. The dataset contained some unneeded columns such as client number and Naïve Bayes Classi fier so removed those columns from the dataset. Missing values are identified in Education Level, Marital Sta tus, and Income Category variables. Filled those missing values with the mode of those variables. There were no duplicates in the dataset.

After data preprocessing descriptive analysis including summary statistics and graphical visualization was car ried out to identify the social, economic, and demographical attributes of credit card customers.

To do the regression analysis categorical variables were converted into dummy variables using the One-hot e ncoding process in python. The data set was imbalanced and therefore resampled using the over-sampling te chnique SMOTE in python. Dataset was split into two parts train and test set. Then feature selection was carr ied out to select the most important variables for fitting the model. Finally, a model was built using the Logis tic regression to predict the credit card customer churn using sklearn in python and checked the accuracy of the model.

Data

The dataset was obtained from Kaggle, and it consists of 10,127 customers mentioning their age, salary, marital status, credit card limit, credit card type, etc. All the data are uniquely identified by the customer's ID number. There are nearly 18 features. The dataset consists of 1617 customers churned (16.05%) and 8500 existing customers (83.92%). Table 1 shows the variables that were used in the analysis.

Table 1

Variable name	Variable type	Description
Attrition_flag	Categorical	customer activity status
Customer_age	Continuous	Customers' age in years
Gender	Categorical	M=Male, F=Female
Dependent_count	Continuous	Number of dependents
Education_Level	Categorical	Educational Qualification of the account holder
Marital_Status	Categorical	Whether the customer is Married, Single, Divorced or Unknown
Income_Category	Categorical	Annual Income Category of the account holder
Card_Category	Categorical	Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book	Continuous	Period of relationship with the bank
Months_Inactive_12_mon	Continuous	No. of months inactive in the last 12 months
Credit_Limit	Continuous	Credit Limit on the Credit Card
Total_Trans_Ct	Continuous	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Continuous	Change in Transaction Count (Q4 over Q1)
Total_Trans_Amt	continuous	Total Transaction Amount (Last 12 months)
Total_Amt_Chng_Q4_Q1	Continuous	Change in Transaction Amount (Q4 over Q1)
Total_Relationship_Count	Continuous	Total no. of products held by the customer
Total_Revolving_Bal	Continuous	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Continuous	Open to Buy Credit Line (Average of last 12 months)
Contacts_Count_12_mon	Continuous	No. of Contacts in the last 12 months
Avg_Utilization_Ratio	Continuous	Average Card Utilization Ratio

- Missing values are identified in Education Level, Marital Status, and Income Category columns. Filled those missing values with the mode of those variables.

variable	Mode of the variable that filled the missing values
Education Level	Graduate
Marital status	Married
Income category	Less than $40k

- Encoded the categorical variables

Attrition flag – 1 as Attrited Customer 0 as Existing Customer

Gender – 1 as Male 0 as Female

Marital status – Single status is defined as the reference level

MS_Divorced	MS_Married
1	0
0	1

Card type – Gold credit card as the reference level

CC_Blue	CC_Platinum	CC_Silver
1	0	0
0	1	0
0	0	1

Education level and Income category encode as order

'Uneducated'- 0 , 'High School'- 1 , 'College' -2 , 'Graduate' – 3 , 'Post-Graduate' – 4 , 'Doctorate' – 5

'Less than $40K' – 0 , '$40K - $60K' -1 , '$60K - $80K' – 2 , '$80K - $120K' – 3 , '$120K +' – 4

- Variables categorized into independent variables as X and dependent variable as Y
- Class imbalance problem was solved using the over-sampling technique SMOTE.

Exploratory data analysis

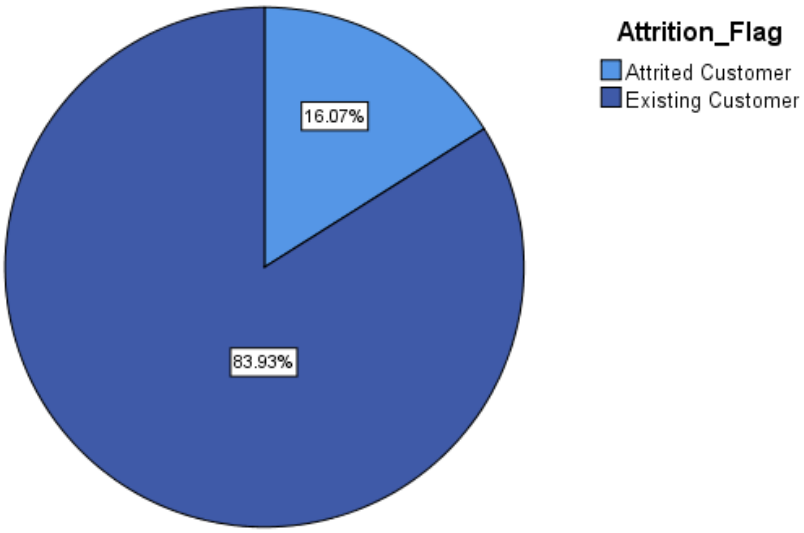


Figure 1

The original data set consists of 16.07% Attrited Customers and 83.93% Existing Customers. Usually, credit card customer churn is a small proportion of the whole customer.

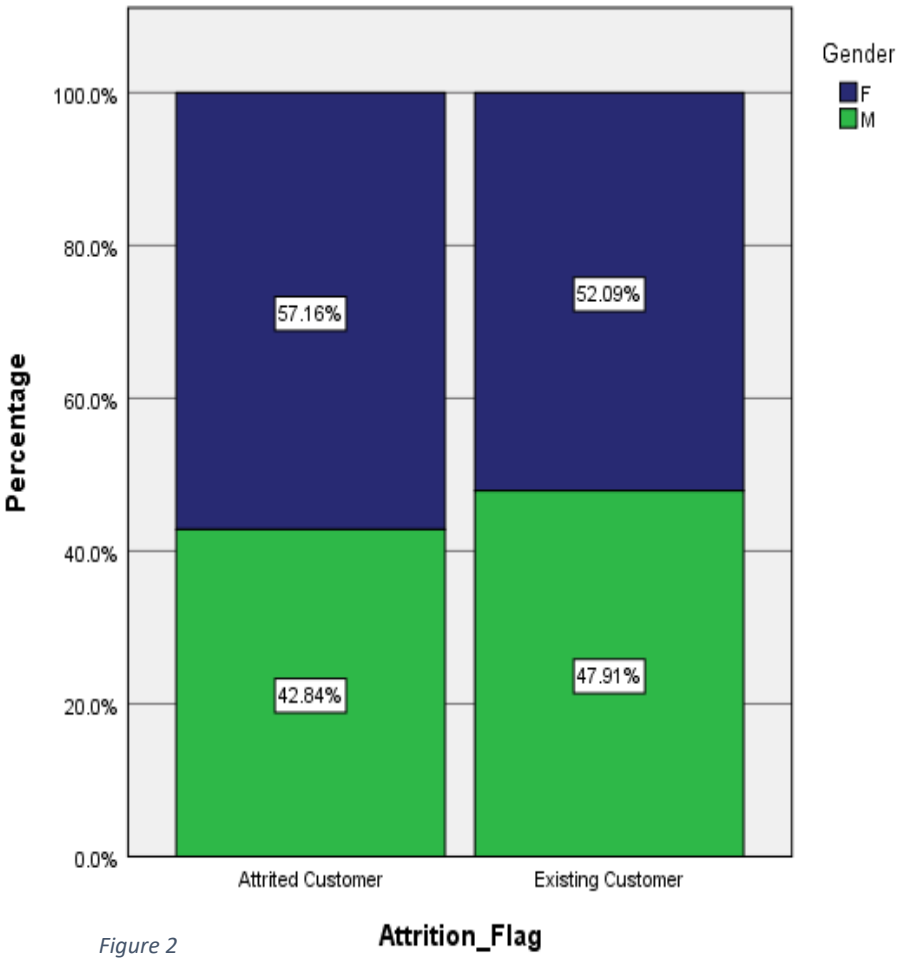


Figure 2

Figure 2 graph shows the gender distribution in existing customers and attrited customers. 57.16% of attrited customers and 52.09% of existing customers are female customers.

Attrited customer female proportion is larger than the existing customer's female proportion.

Therefore, more female customers are getting churned among the attrited customers.

The majority of credit card customers are female.

47% of male customers are among the existing customers.

There are many types of credit cards that come with different benefits and features. The credit card types are Blue, Silver, Gold and Platinum as the order of rewards. Platinum credit cards can be considered as the most exclusive and elite credit card type and Blue credit card is an entry-level credit card type. In order to identify the most popular credit card type among the credit card holders, the pie chart shows the majority are using Blue credit cards which is 93.18% of total customers. The second-highest credit card type is the Silver credit card. Very few customers are using platinum cards.

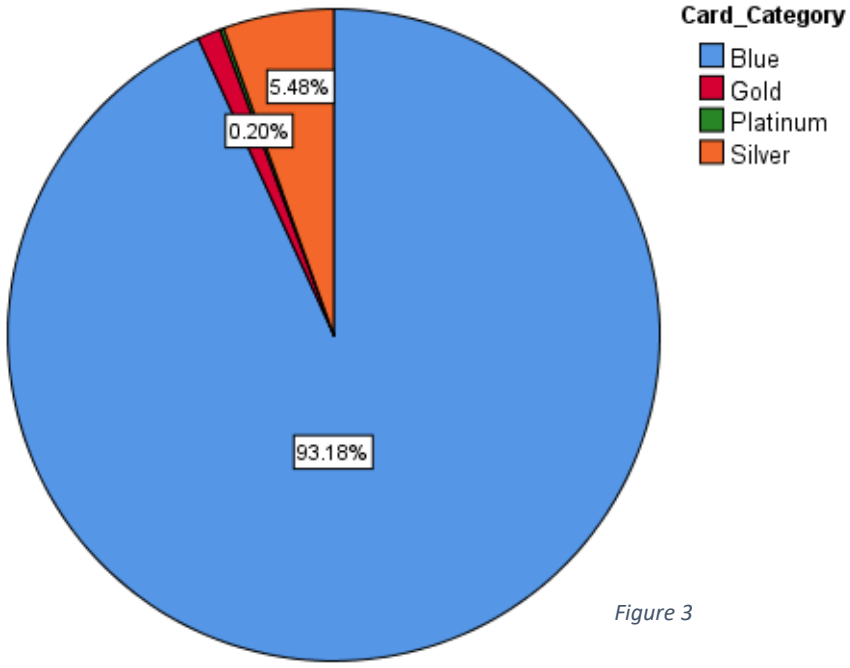


Figure 3

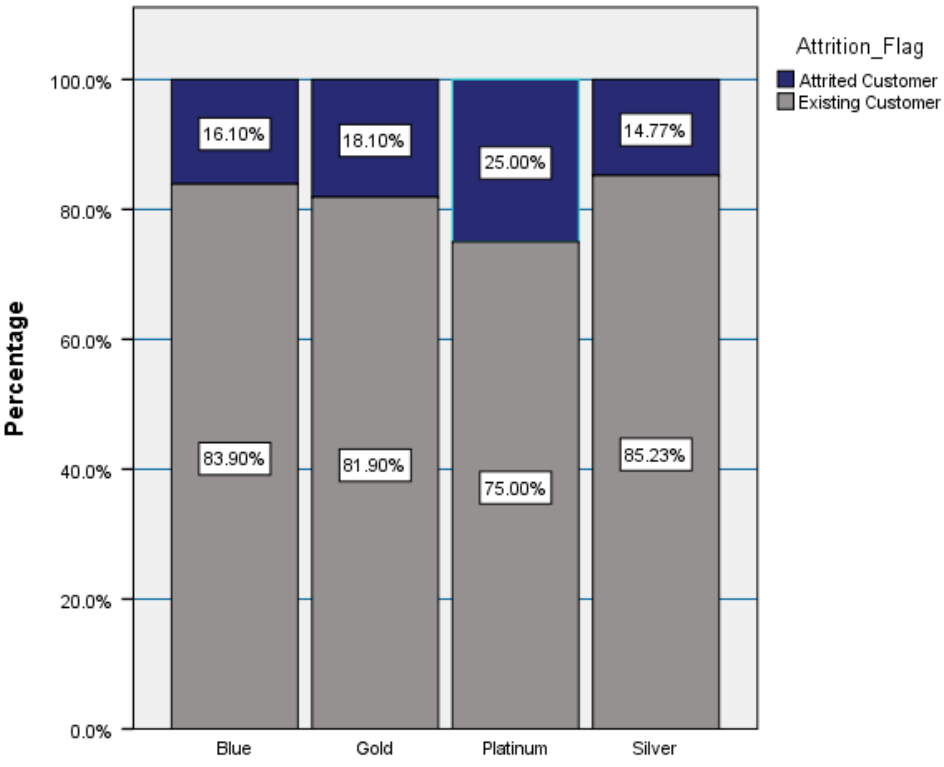


Figure 4

Platinum cardholders' attrited customer proportion is higher than the other card types. which is 25% of platinum card holders got churned. Gold cardholders' attrition percentage is also higher than the Blue and Silver cards. Silver cardholders' customer retention is higher than the other card types. Blue cardholders' customer retention is the second-highest retention percentage which is 83.90%.

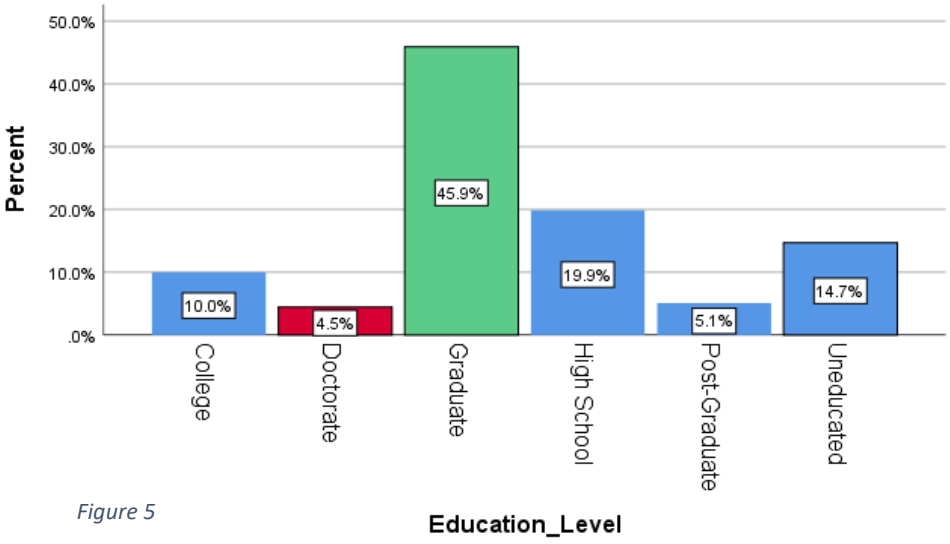


Figure 5

The dataset is consist with customers who have different education levels. They are Uneducated, high school, college, graduate, post-graduate and doctorate. The majority of the customers' education level is Graduate which is 45.9%. Doctorate customers are the least among the customers. The second highest education level among the customers is high school.

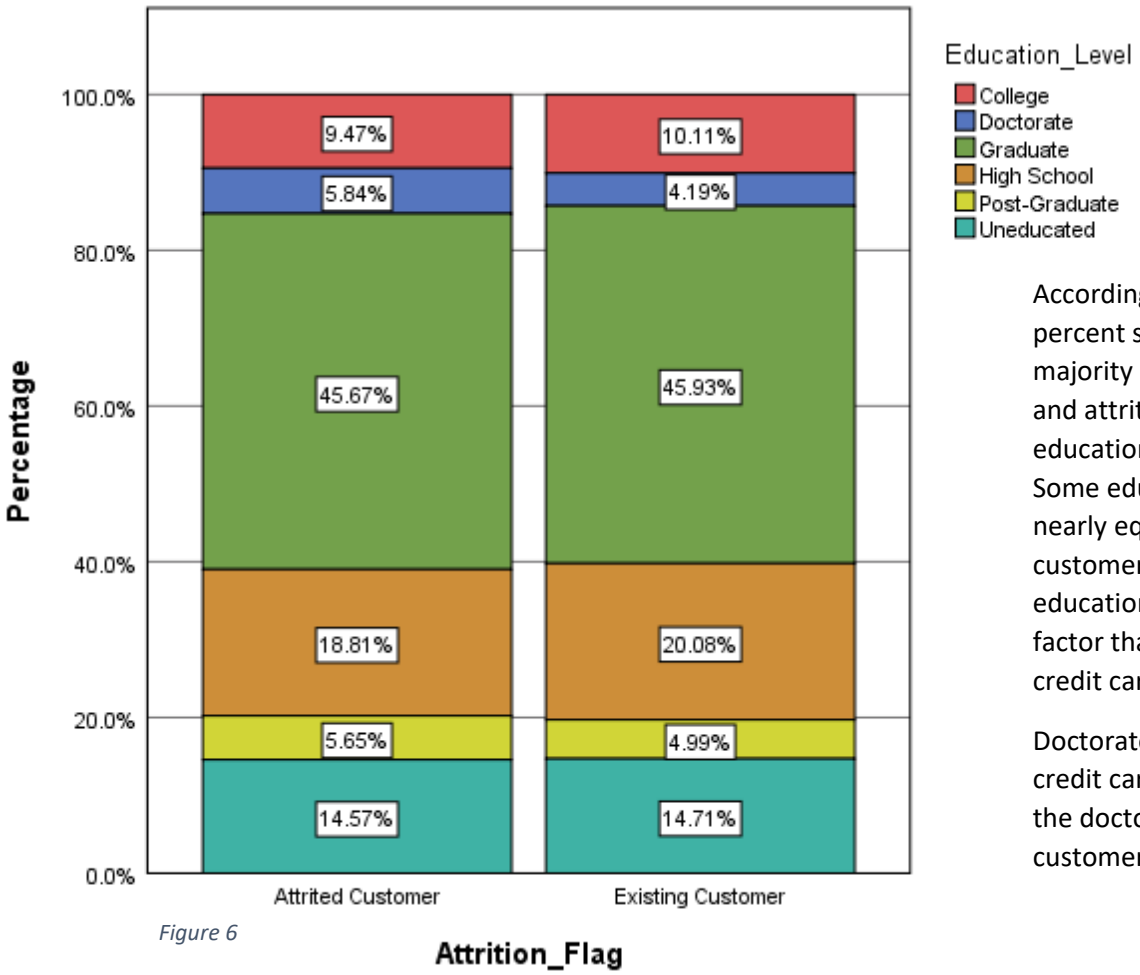


Figure 6

According to this hundred percent stack bar chart majority of existing customers and attrited customers' education level is graduate. Some education levels are nearly equal for both customers. Therefore, education level might not be a factor that associates with credit card customer churn.

Doctorate attrited customers' credit card churn is higher than the doctorate existing customers.

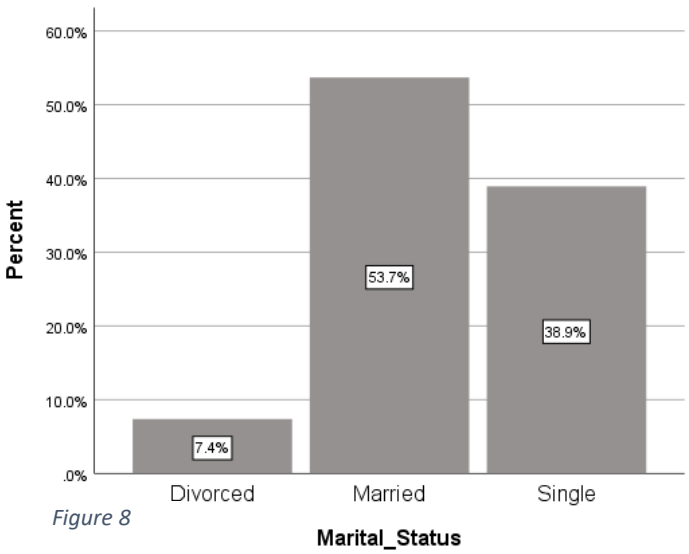


Figure 8

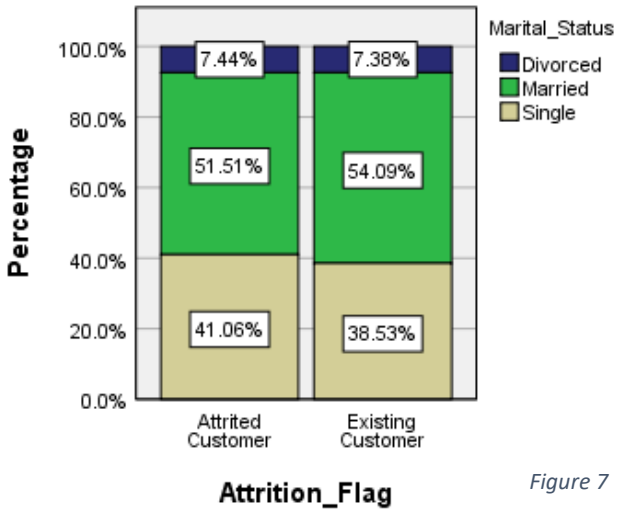


Figure 7

By considering the figure7, 53.7% of customers are married. And also married customers are more tend to have a credit card than single customers. Only about 7.4% of the customers are divorced.
The attrited customers' single and the divorced status percentages are higher than the existing customers' single and divorced status percentage.

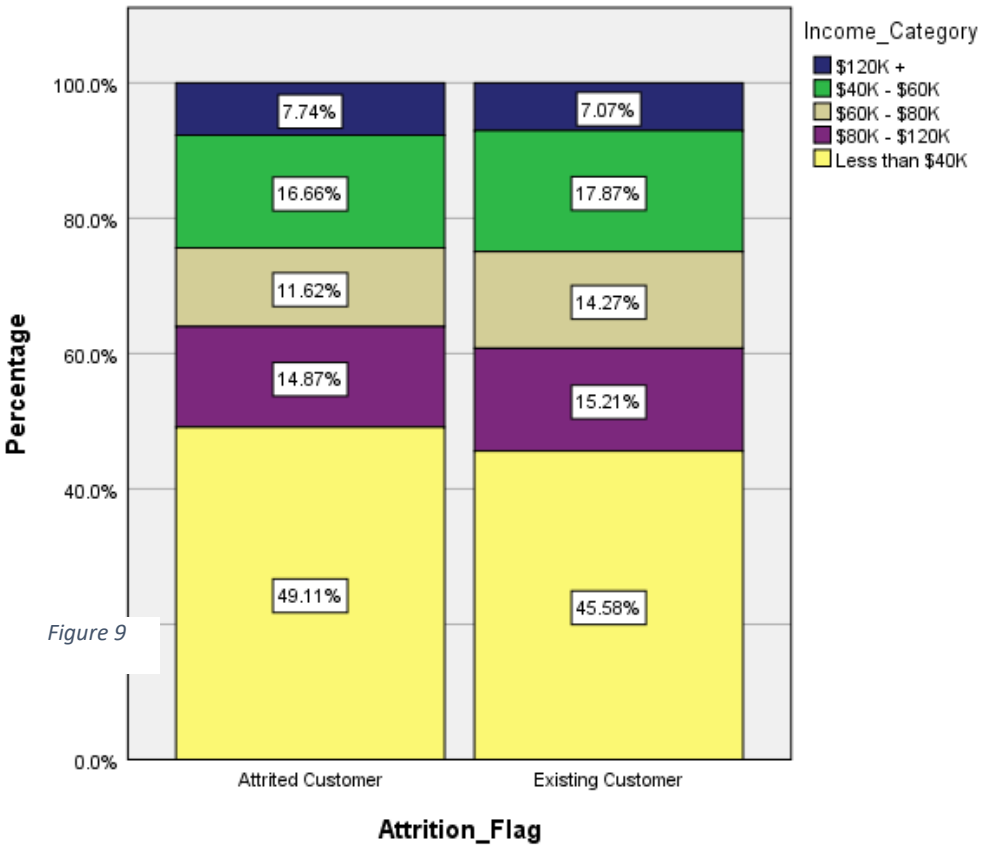


Figure 9

Figure 9 shows income category proportions among existing customers and attrited customers.

49.11% of Attrited customers are in the less than $40k income category. A larger proportion of customers whose income is $120+ are in the attrited customer category. Generally, customers who have a higher income will continue the credit card service.
But here customers who have higher income got churned. The reason can be they are not satisfied with the service.

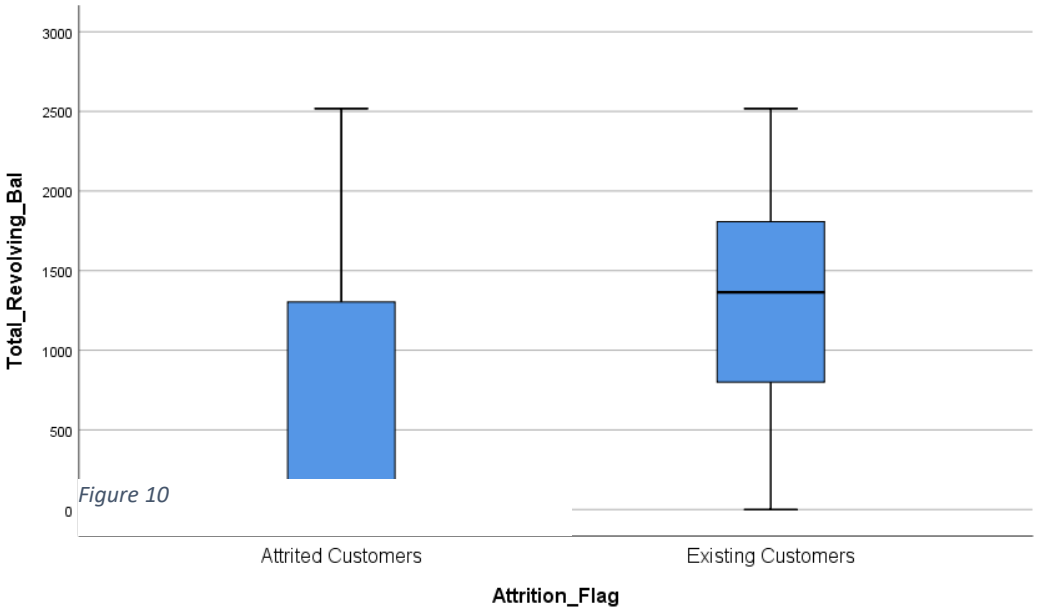


Figure 10

Figure 10 shows the total revolving balance of the attrited and existing customers. A revolving balance is the portion of credit card spending that goes unpaid at the end of a billing cycle.
Revolving credit is best when someone wants the flexibility to spend on credit month over month and it can be beneficial to earn rewards points and cash back.

The graph shows that attrited customers' total revolving balance spread is more towards the zero. This means most of the attrited customers' total revolving balance is low. The reason can be attrited customers don't know the benefits of revolving credit.

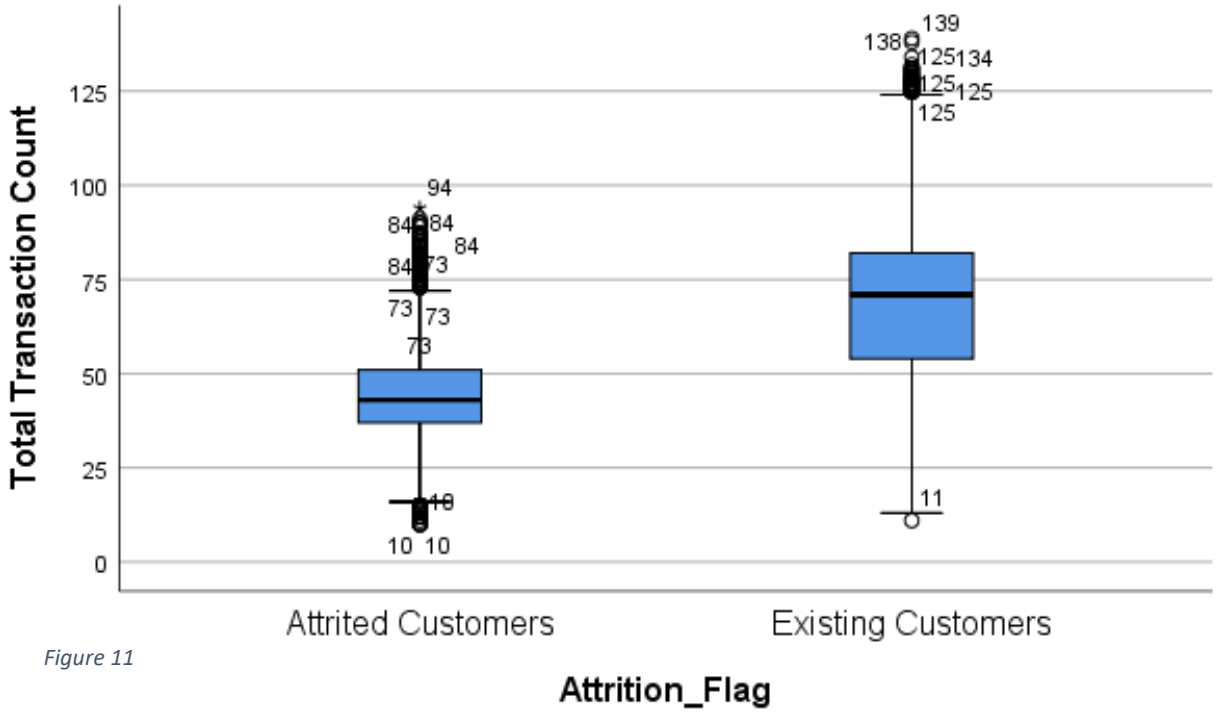


Figure 11

Figure 11 shows that 75% of attrited customers had a number of transactions less than 51 and their maximum number of transactions is less than 73. According to the figure attrited customers' total transaction count is less than existing customers. Attrited customers rarely used credit cards for their transactions. They might not interest or not have a benefit from using credit cards.

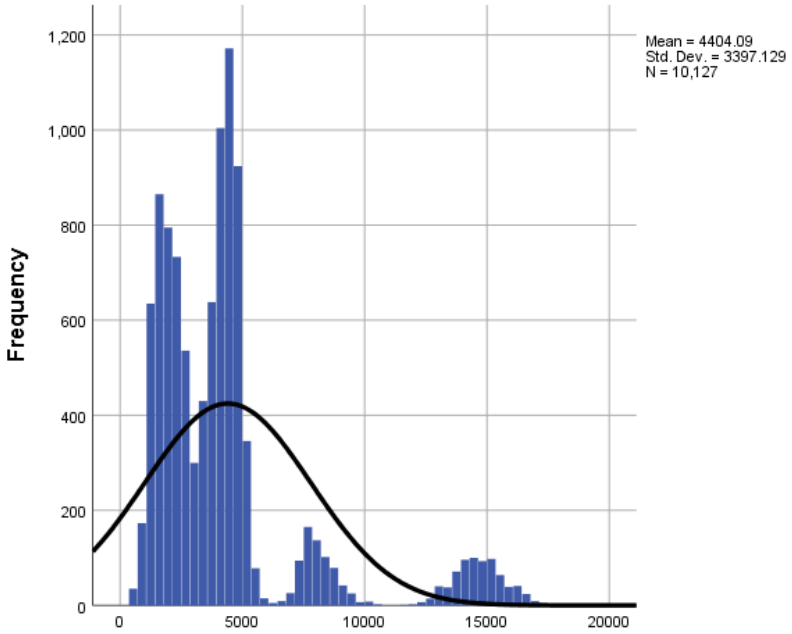


Figure 12

Figure 12 shows the distribution of total transaction amount of last 12 months of the customers. The majority of cardholders spend less than $5000 per year on their transactions. The reason can be more credit card holders are from income category less than $40000. A few number of customers are spending more than $15000 per year. When the transaction amount is high the bank will have the benefit.

Advanced analysis

- First conducted the feature selection procedure to select the most important (most significant) variables to fit the model.

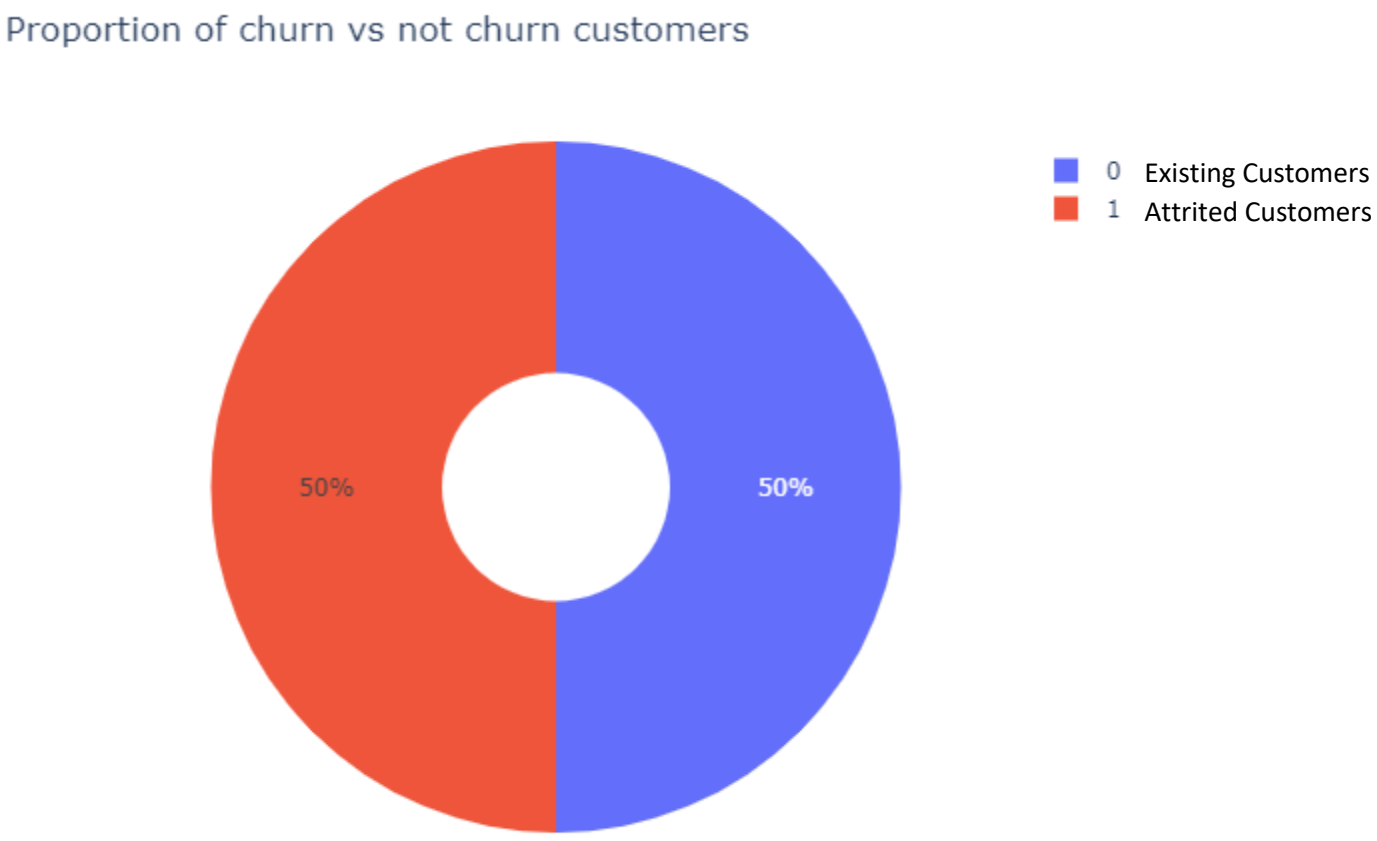
The selected variables are,

```
['Gender', 'Dependent_count', 'Education_Level','Total_Relationship_Count',
'Months_Inactive_12_mon','Contacts_Count_12_mon', 'Total_Amt_Chng_Q4_Q1', 'To
tal_Trans_Ct','Total_Ct_Chng_Q4_Q1', 'Avg_Utilization_Ratio', 'MS_Divorced',
'MS_Married', 'CC_Blue', 'CC_Platinum', 'CC_Silver']
```

Therefore selected only the above variables and dropped the other variables. Those dropped variables are,

['Customer_Age','Income_Category','Months_on_book','Credit_Limit','Total_Trans_Amt','Total_Revolving_Bal','Avg_Open_To_Buy']

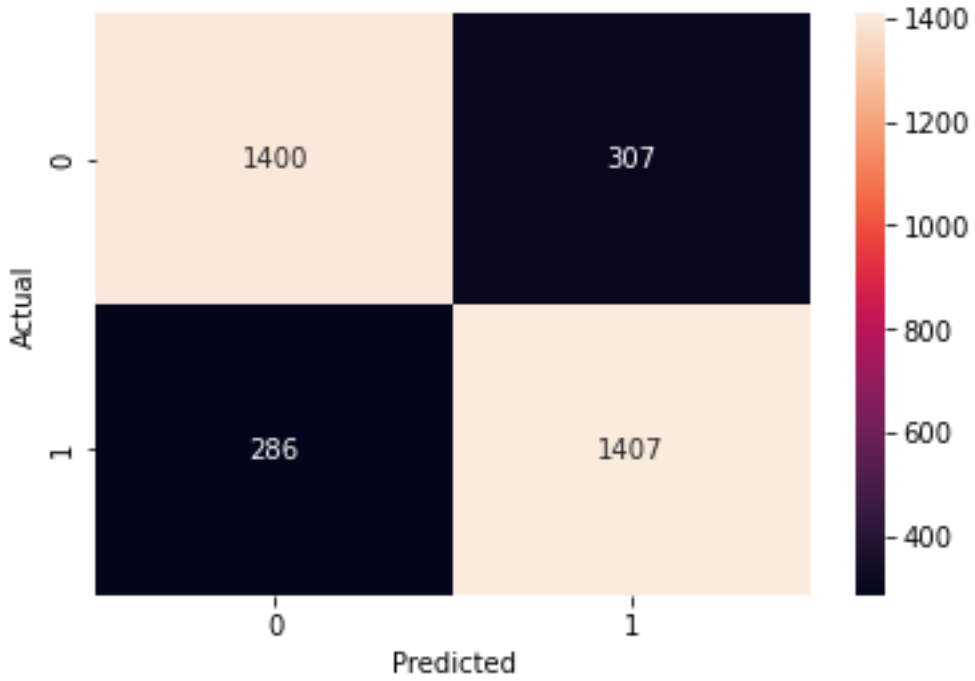
- Using the selected variables analysis was carried out.
- After solving class imbalance using SMOTE, the proportion of customers are as follows.



The fitted logistic regression equation is,

$$\hat{Y} = 5.559648 - 1.08445\,(\text{Gender}) + 0.103891\,(\text{Dependent\_count}) + 0.108353\,(\text{Education\_Level}) - 0.49426\,(\text{Total\_Relationship\_Count}) + 0.3606861\,(\text{Months\_Inactive\_12\_mon}) + 0.31726269\,(\text{Contacts\_Count\_12\_mon}) + 1.15216683\,(\text{Total\_Amt\_Chng\_Q4\_Q1}) - 0.06608787\,(\text{Total\_Trans\_Ct}) - 2.37313089\,(\text{Total\_Ct\_Chng\_Q4\_Q1}) - 3.64874064\,(\text{Avg\_Utilization\_Ratio}) - 1.76215383\,(\text{MS\_Divorced}) - 0.79481382\,(\text{MS\_Married}) + 0.10382682\,(\text{CC\_Blue}) - 0.01426835\,(\text{CC\_Platinum}) - 0.36707535\,(\text{CC\_Silver})$$

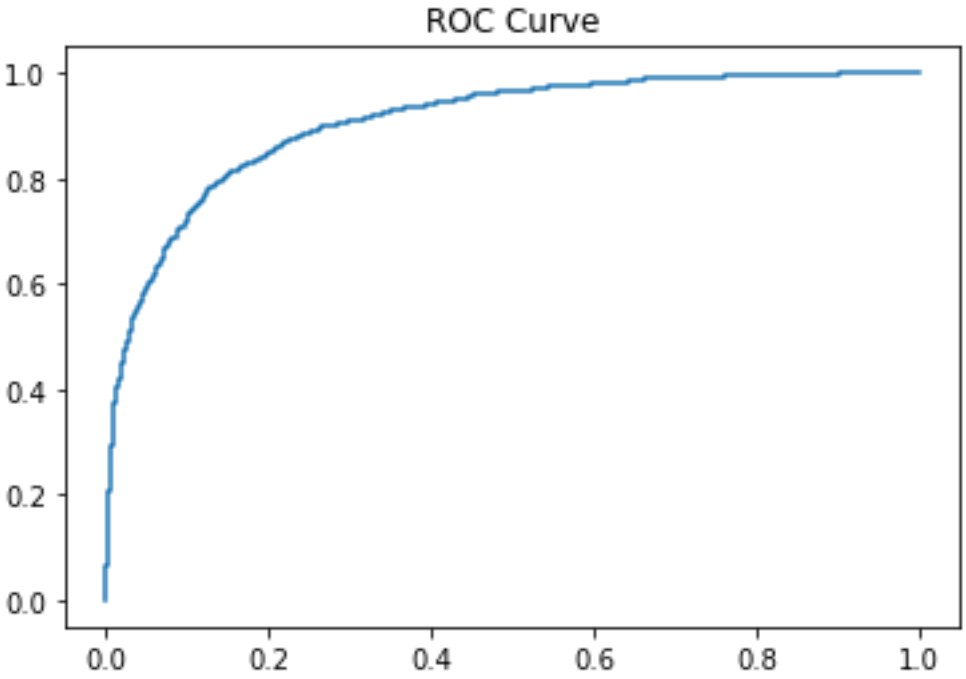
From the confusion matrix out of 3500 testing observations 2807 observations have predicted correctly by the fitted model. Therefore, the overall model accuracy is 82.5% and Miss classification error is 17.5%.



Classification report

	precision	recall	f1-score	support
0	0.83	0.82	0.83	1707
1	0.82	0.83	0.83	1693
accuracy			0.83	3400
macro avg	0.83	0.83	0.83	3400
weighted avg	0.83	0.83	0.83	3400

Class accuracy is shown by F1-Score. F1 score is 83%. This is also a good score.



Area under the curve value is 0.9068

Figure shows the Receiver Operating Characteristics (ROC) curve. The area under the ROC curve also shows that the accuracy of the model is good.

The fitted regression equation coefficients show that,

- female customer's attrition possibility is higher than the male customers.
- When the dependent count is high, the attrition possibility will increase.
- Doctorate customer churn possibility is higher when compared to other education levels.
- When the total relationship count (Total no. of products held by the customer) is high the customer will retain.
- If the customer is not active for several months, then there is a probability that the customer will churn.

General Discussion and Conclusion

- Usually, credit card customer churn is a small proportion of all customers.
- Majority of cardholders are female in both attrited customers and existing customers. And also, female customers got churned more than male customers. Therefore, need to be more focused on females when promoting credit cards.
- Most popular credit card type is the blue card. A small percentage of customers have platinum cards and platinum cardholders' retention is less when compared to other cardholders. Bank managers need to be more focused on platinum card holders. Blue credit card holders' customer churn is higher than the other types of credit card holders'.
- Most of the customers' education level is Graduate. More than 70% of customers are having a formal education level.
- Married customers tend to have a credit card more than single customers. The reason can be married customers spendings are more than those who are single.
- Majority of credit card holders' income is less than $40000 and also the majority of cardholders spend less than $5000 per year for their credit card transactions.
- Attrited customers' total transaction count is less than the existing customers' total transaction count. Attrited customers have rarely used credit cards for their transactions.
- 50% of attired customers' total revolving balance is zero. Therefore when the customer is having a lower revolving balance they are most likely to leave. This could be indicating that customer have found another bank with lower interest rates.
- 'Gender', 'Dependent_count', 'Education_Level','Total_Relationship_Count', 'Months_Inactive_12_mon','Contacts_Count_12_mon', 'Total_Amt_Chng_Q4_Q1', 'Total_Trans_Ct','Total_Ct_Chng_Q4_Q1', 'Avg_Utilization_Ratio', 'MS_Divorced', 'MS_Married', 'CC_Blue', 'CC_Platinum', 'CC_Silver' are the variables that most associated with credit card customer churn.
- Customer age and income level are not much associated with credit card customer churn.
- When the number of dependent count is increasing the customer churn probability will increase.
- If the customer is inactive for several months that customer's churn possibility is high.
- Banks should contact the customers regularly and check whether they are satisfied with the service and what the bank could do to improve their service.
- Banks should give promotions for their customers.

References

[1] What is Logistic regression? | IBM

[2] Imbalanced Classification | Handling Imbalanced Data using Python (analyticsvidhya.com)

[3] Guangli Nie; Wei Rowe; Lingling Zhang; Yingjie Tian; Yong Shi (2011). *Credit card churn forecasting by logistic regression and decision tree. , 38(12), 15273–15285.* doi:10.1016/j.eswa.2011.06.028

[4] Credit Card customers | Kaggle