CREDIT CARD CUSTOMER CHURN PREDICTION

BY
VENKATA SAI KUMAR RAMBHA (MUID 11033313)
SARAVIND REDDY SAMA (MUID 11033265)

Abstract:

We see that there is a fast development of technology and financial institutions are growing accordingly. Also with the increase in technology, we see that people tend to adapt to these technological changes where the spending ratio of people has increased over time. Where financial institutions need to mainly focus on customer retention. As we know that the retention of existing customers is more important for any company rather than attracting new customers as the profit margin plays a major role in the Businesses. We also see that the customer churn rate is higher in the financial sector than in other sectors. Credit card is the major profit margin for banks and financial services. This paper mainly focuses on predicting churn using various linear models and feasible solutions to the business to focus on what type of customer can churn over a period. We have applied certain models such as Logistic, Random Forest, naive Bayes, and Stepwise analysis to predict which model can predict the highest accuracy.

NOTE: ALL FIGS ARE IN THE APPENDIX

1. INTRODUCTION:

We wanted to know the importance of churning for financial institutions as customer churn increased in the financial sector over a period. The competition for financial institutions has increased tremendously. The corporate credit card market is estimated to be 14.1 billion [1]. Where we see that there are multiple offers that they provide to customers at a low price and better quality from different financial institutions here is where the customer leaves from bank to bank. The banks have begun to realize that customer relationship management is important to retain their customers [3]. When we observe the past findings in the early 2000s the service industries should actively learn the behavior of their customers. the main involvement is that communicating with the customer is highly important to retain a long time of customer relationship with the company. [4]. There are two churn periods identified: the initial years of customer joining and the second is when the customer spends almost 20 years with the company [6]. According to some arguments, a 5% improvement in client retention might lead to an 18% decrease in operational expenditures [7].

The focus is that what are the main factors where the banks or financial institutions need to focus and retain their customers. we use the data mining processes to analyze and develop a model to anticipate customer churn. The data mining process has become the main key factor in industries in recent times where there is an enormous amount of data generated and this data needs to be converted to use full insights. Using various techniques of data mining such as linear regression, Random Forest, Stepwise, Naive Bayes, and Classification trees we get to decide which model gives us better results.

2. LITERATURE REVIEW:

The service industries to find new customers is expensive such that we decide to analyze the existing customer's behavioral patterns [3]. Customer attrition is the main aspect of the service industries different methods and models have been used from the past matrices and most of the show the solution of retention. Many studies have proven fact that the financial institution can increase its profits by 85% when it can retain at least 5% of its customers [3]. Every research has its own domain many researchers have developed different models using logistic regressions and t-tests for customer churn prediction and loyalty of customers. these helped the companies to have a strong relationship with customers [5]. The studies proposed analysis that helps financial industries to anticipate the customer who is more likely to churn. The system uses different techniques to measure the efficiency of the model. The model has worked with 4 methodologies: Decision Tree, Random Forest, Gradient Boosted Machine Tree(GBM), and Extreme Gradient Boosting(XGBOOST). The big data platform was decided upon as the Hortonworks Data Platform (HDP). Almost all stages of the product's development, including data analysis, function development, training, and software testing, utilized Spark engines. K-fold cross-validation was used to optimize the method hyper-parameters. The sample for learning is rebalanced by taking a sample of data to balance the two classes because the target class is unbalanced. The churn class was multiplied to fit with the other class in the study's initial oversampling step. In order to compare the broad class with the second class, a random under-sampling strategy was also applied, which reduces the sample size of the broad class. Training on the Decision Tree method and hyper-parameter optimization began [8]. As compared to simple linear regression and LR models, RF offered a better fit for the estimate and validation sample. developed a decision tree operator-based model that forecasts the propensity for a consumer to leave [6]. Few researchers worked with LR model, ensemble decision tree variants, and decision trees. They concluded that ensemble learning typically increases the predictability of flexible models like decision trees, which results in better predictions. The ensemble models were also found to outperform individual decision trees and LR. had experience with RF, neural networks, LR, and Automatic Relevance Determination (ARD). Their findings demonstrate that RF regularly outperformed the competition. used self-organizing neural networks, Hopfield neural networks, and multilayer feed-forward neural networks to handle churn concerns. using decision trees, I built a churn model that has an accuracy rate of 85%. Using the rough sets technique, achieved a 90% overall classification accuracy rate [7].

NOTE: ALL FIGS ARE IN THE APPENDIX

3. DATA SOURCES AND CHARACTERISTICS:

We have obtained a data set from Kaggle where the data consists of 10000 observations and 21 variables where we feel that the Total transaction amount in the last 12 months, total transaction count in the last 12 months, and total revolving balance are the top three important features. We do the exploratory analysis to see the visualizations and then split the data set into training and validation and do the further modeling part. We have observed that the Random Forest method gives us better results of all the models used. We have ranked the top 8 important features that can tell us whether the customer can churn or not.

3.1 DATA CHARACTERISTICS:

Where they are 21 variables we can see the top features which can accurately predict customer churn based on the analysis. We have identified the top features: **Total_Relationship_Count** - Total no. of products held by the customer, **Months_Inactive_12_mon** - No. of months inactive in the last 12 months, **Total_Revolving_Bal**- Total Revolving Balance on the Credit Card, **Total_Amt_Chng_Q4_Q1**- Change in Transaction Amount (Q4 over Q1), **Total_Trans_Amt** -Total Transaction Amount (Last 12 months), **Total_Trans_Ct** -Total Transaction Count (Last 12 months), **Total_Ct_Chng_Q4_Q1** - Change in Transaction Count (Q4 over Q1), **Customer Age** - Demographic variable - Customer's Age in Years.

3.2 CLEAN AND PREPROCESS:

We have 21 factors where we first start with the EDA process by plotting the variables and checking if there are any outliers, missing values, duplicates, etc., and find the relationships between the dependent variable and the other variables to identify whether the customer gets churned or not. Some of the important features for consideration are:

Attrition flag: From this plot, we can interpret that the dataset consists of 2000 attired customers and 8000 existing customers.

NOTE: ALL FIGS ARE IN THE APPENDIX

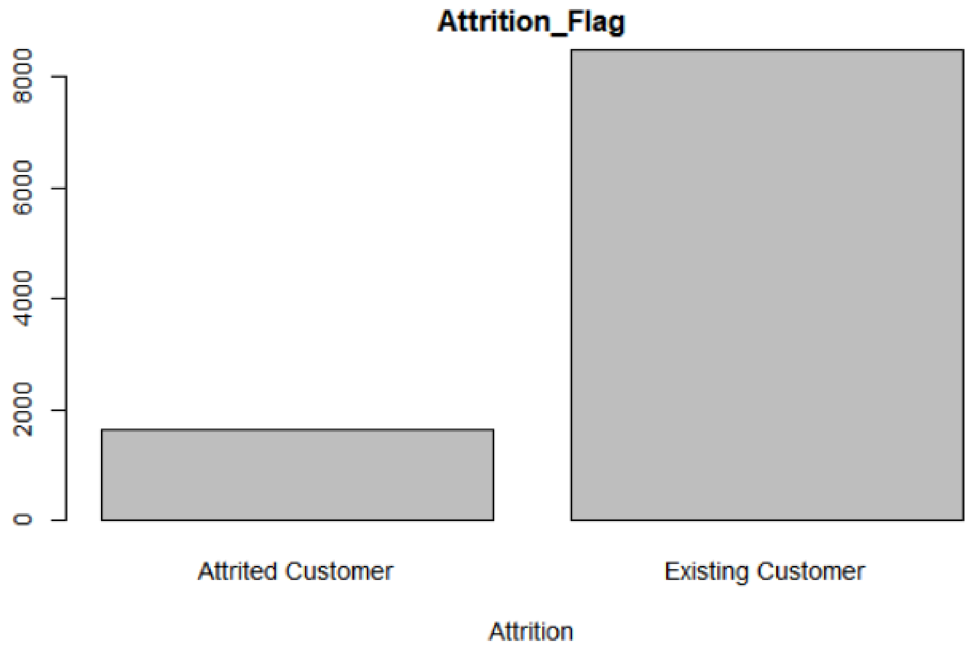


Figure 1: Attrition Flag of the customers.

Income Category: From this, we can see that this dataset has the most number of customers from the income category of less than 40k.

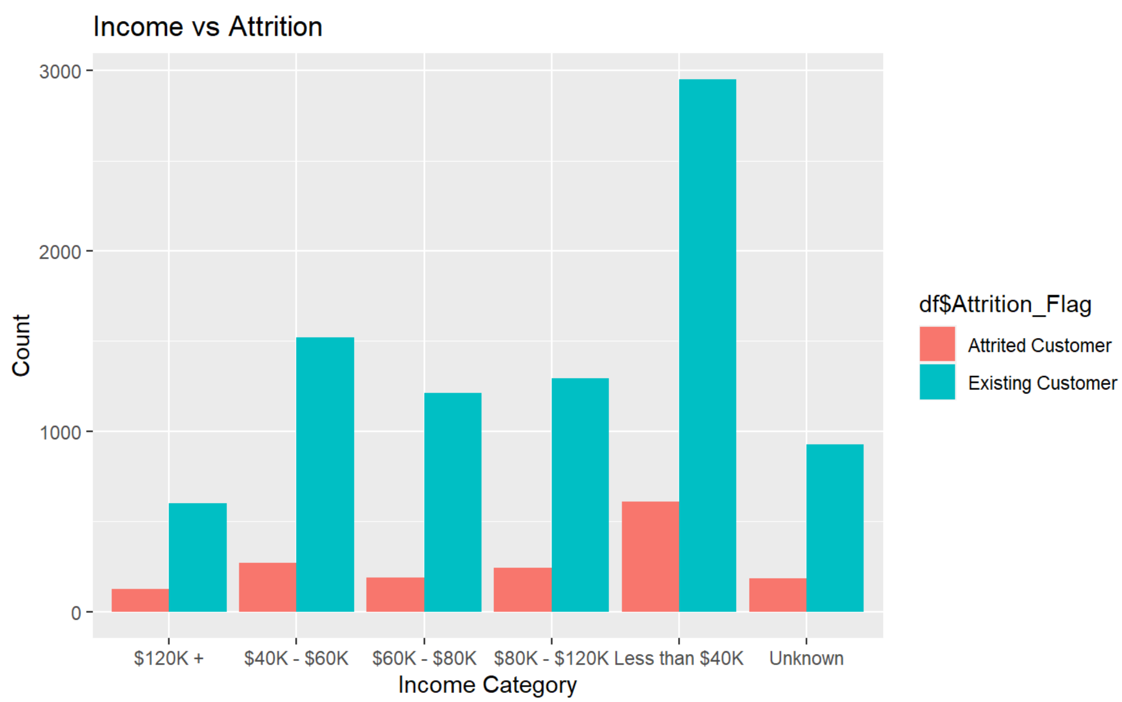


Figure 2: Income category of the customers.

NOTE: ALL FIGS ARE IN THE APPENDIX

Education level: From this, we can see that this dataset has the most number of customers from the graduate degree.

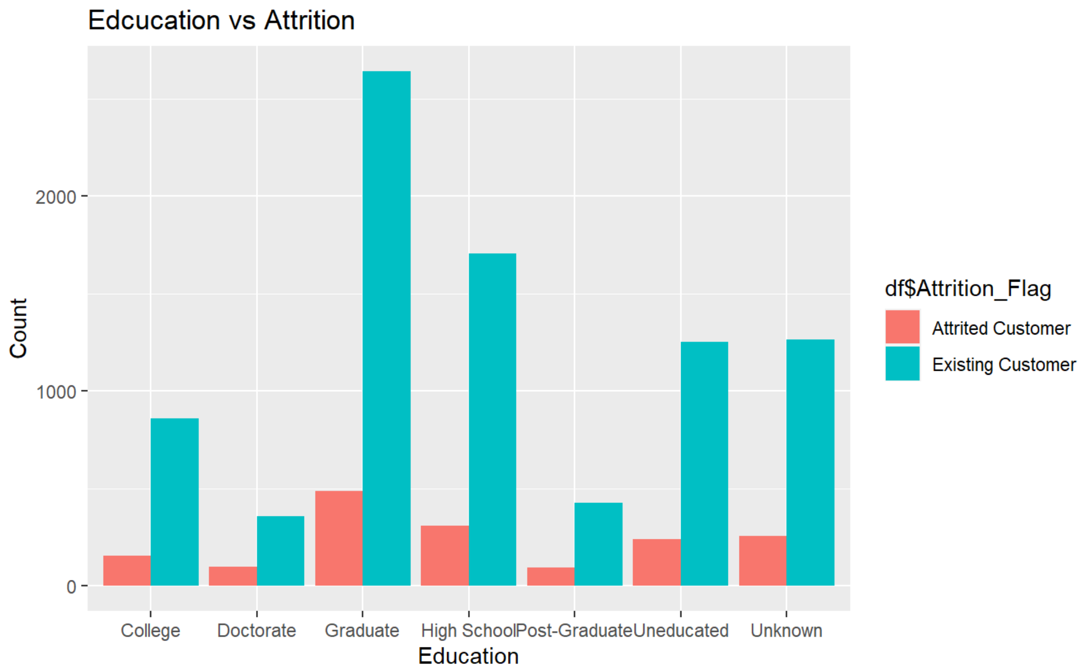


Figure 3: Education level of the customers.

Card Category: From this, we can see that this dataset has the most number of customers from the blue card.

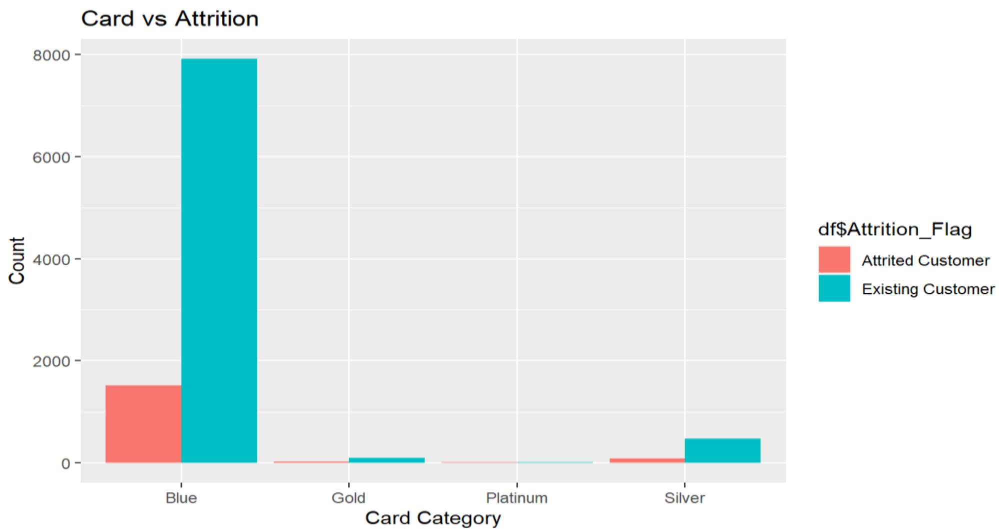


Figure 3: Card Category of the customers.

NOTE: ALL FIGS ARE IN THE APPENDIX

3.3 REDUCE DIMENSION:

As we can observe the data in the summary(fig1), we found that there are 15 numerical variables and 6 categorical variables. Where we observed that there are no duplicates or missing values but there are certain unknown values which are **3,327** values in the data set(fig2,3,4). Also, there are certain outliers which are **127** values in the data where we have dealt with outliers by deleting the extreme outliers, we wanted to check the model functioning, so we have performed both ways by keeping the outliers and by removing them. The Dimension reduction is made by keeping all the variables we have plotted the correlation plots against each other and then we performed the regression analysis to know the importance of the variables (figs 5& 6).

3.5 PARTITION OF DATA:

After outliers and having a distinguished set of data. The data set has been partitioned to 60% of the data for Training and 40% of the data is partitioned for validating the data but the challenge here is that the data set is imbalanced as the entire data set has only 20% of Attired Customers and 80 % has existing customers. When we run the data, with the 20 and 80% the results will be biased hence the oversampling technique is been used to balance the data set (fig 7). We have Perferformed the multicollinearity test to reduce the variable we find that the Credit limit, average open to buy, and total revolving are highly correlated with each other. We have delt with the problem in the modeling part.

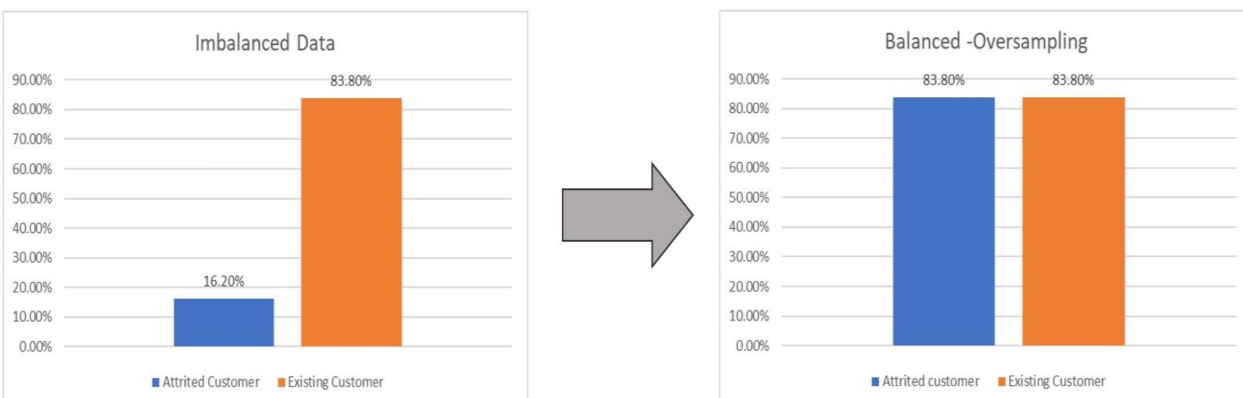


Fig4: Data balancing

4. METHODOLOGIES:

Logistic Regression

Logistic regression is the linear model which identifies the conditional probabilities between the variables where we also see that the dependent variable is categorical and also we are focusing mainly on the Attired customers and trying to predict the retention of existing customers based on the factors of the Attired customers. Based on the results of the logistic regression we have identified that the Months on book and credit limit are the ones that are not impacting customer attrition, so we tried to omit the variables from our regression. (fig:9)

Stepwise Regression:

The Stepwise regression can be done where we can use it to build the regressors which are highly important for the dependent variable where this is statistically valid for keeping or removing variables. Here we see that the regression has provided 12 variables by omitting a few (fig10).

Random Forest

Random forest is a supervised learning model. It was proposed by Breiman and Cutler in 2001, and is based on decision tree and ensemble learning [9]. As we know that the random forest is the best technique used for the prediction as it does a bootstrapping where it creates a bubble amount of the dependent variable and gives us accurate results which can be accessed and related. We see that when we ran a random forest for our attrition on all the variables we found the variable importance of the variables which are highly important for our model to predict the attrition. As observed in (Fig 11) the important variables are been provided so we choose the first eight variables which are having a major impact on attrition.

Decision trees

Classification Tree: This are the algorithm that can perform the classification where we can fit the complex structures to easily interpret the results as we observe the (fig12) we can tell that the most important variables towards attrition.

Naive Bayes

In the other set of classifications of independent towards the decision variable, we see that these set of variables are collectively important for the attrition of the customers (fig 13).

NOTE: ALL FIGS ARE IN THE APPENDIX

5. EMPIRICAL RESULTS:

We observed that the random forest has the highest predictive performance in apart from all the models. Where we see that in both with and without outliers the Accuracy of the Random Forest is high and the main factors for the company to look at are the revolving balance, total transactions count, and total transaction amount. The results and the graphs can be viewed in the appendix.

Regression	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
Logistic	82.41	85.89	78.77	0.823	80.83	88.44	73.01	0.807
Random forest	92.47	96.5	88.25	0.924	93.25	95.4	91.05	0.932
Naïve bayes	79.87	75.06	84.9	0.8	78.96	76.46	81.53	0.8
Classification Tree	89.99	86.8	93.28	0.9	90.39	90.11	90.69	0.904
Stepwise	82.41	85.81	78.77	0.823	80.83	88.44	73.01	0.807

Fig 5: Results of the regression

NOTE: ALL FIGS ARE IN THE APPENDIX

6. CONCLUSION:

This paper aimed at predicting the churn of credit card customers, the dataset provided to us consists of 10,127 observations containing of customer age, revolving balance, total transaction count ...etc. and does research and analysis based on it.

We did preprocess the data and as the dataset was imbalanced, we used the oversampling technique to balance the data and then applied the regression methods logistic regression, stepwise, naïve Bayes, classification tree, and Random Forest. We modified the hyperparameters in each model to increase accuracy and evaluate model performance using ROC & AUC and the confusion matrix.

Random forest gave the best results compared to all other methods and we have identified 8 features that best predict the attrition rate of customers with the highest accuracy. Random forest gave us an accuracy of 92.47 and a sensitivity of 96.5 without including the outliers in the model. The classification tree was the second-best model with 90% accuracy and 87% sensitivity. The main features that we identified to predict the model are **Customer age, Total Relationship Count, Months Inactive 12 months, Total Revolving Bal, Total Amt Change Q4-Q1, Total Trans Amt, Total Trans Ct, Total Ct Change Q4-Q1** these have a significant impact on the model forecasting. It is seen that the total transaction count in the last 12 months and the total revolving balance of the customer are the most important features to predict Attrition. The Blue Card users have attired the highest.

7. RECOMMENDATIONS:

- If the customer transaction count is decreased compared to the previous quarter the business needs to motivate the customer to keep using the card by providing some incentives. Ex- Cashback offers, Reward points for every transaction made.
- Develop marketing strategies targeting the blue card category.
- The business needs to mainly focus on customers who have not been using their cards for more than 2 months.
- The target customers for the business would be the customers who have income less than 40K and has a graduate degree.

NOTE: ALL FIGS ARE IN THE APPENDIX

9. REFERENCES:

[1] N. X. Hong, and L. Yi, “Standing at the crossroadscredit card,” Reporters' Notes, vol. 5, pp. 41-43, 2020. (in Chinese)

[2] R. Rajamohamed, and J. Manokaran, “Improved credit card churn prediction based on rough clustering and supervised learning techniques,” Cluster Computing, vol. 21, pp. 65-77, June 2017.

[3] G. L. Nie, W. Rowe, L. L. Zhang, Y. J. Tian, and Y. Shi, “Credit card churn forecasting by logistic regression and decision tree,” Expert Systems with Applications, vol. 38, pp. 15273-15285, 2011.

[4] Bolton, R.N. (1998) ‘A dynamic model of the duration of the customer’s relationship with a continuous service provider: the role of satisfaction’, Marketing Science, p.45.

[5] Bolton, R.N., Kannan, P.K. and Bramlett, M.D. (2000) ‘Implications of loyalty program membership and service experiences for customer retention and value’, Journal of the Academy of Marketing Science, Vol. 28, pp.95–108.

[6] Lariviere, B. and Van den Poel, D. (2004a) ‘Customer attrition analysis for financial services using proportional hazard models’, European Journal of Operational Research, Vol. 157, pp.196–217.

[7] Karakostas, B., Kardaras, D. and Papathanassiou, E. (2005) ‘The state of CRM adoption by the financial services in the UK: an empirical investigation’, Information & Management, Vol. 42, pp.853–863.

[8] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” Journal of Big Data, vol. 6, no. 1, p. 28, 2019.

[9] Y. A. Amrani, M. Lazaar, and K. E. E. Kadiri, “Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis,” Procedia Computer Science, vol. 127, pp. 511-520, 2018.

[10] Miao, Xinyu, and Haoran Wang. “Customer Churn Prediction on Credit Card Services Using Random Forest Method.” Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022), 2022.

[11] Nie, Guangli, et al. “Credit Card Churn Forecasting by Logistic Regression and Decision Tree.” Expert Systems with Applications, vol. 38, no. 12, 2011, pp. 15273–15285.,

[12] Kumar, Dudyala Anil, and V. Ravi. “Predicting Credit Card Customer Churn in Banks Using Data Mining.” International Journal of Data Analysis Techniques and Strategies, vol. 1, no. 1, 2008, p. 4.,

[13] Benton, W., Bengtson, J., & Technical, S. (n.d.). “Accelerating customer churn prediction” Retrieved December 16, 2022,

[14] Rahman, Manas & Vasimalla, Kumar. (2020). Machine Learning Based Customer Churn Prediction In Banking. 1196-1201. 10.1109/ICECA49313.2020.9297529.

NOTE: ALL FIGS ARE IN THE APPENDIX

APPENDIX

DATA DICTIONARY:

Credit card customers (predicting bank churners)

Source: leaps.analyttica.com

Attribute	**Description**
CLIENTNUM	Client number. Unique identifier for the customer holding the account
Attrition_Flag	Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age	Demographic variable - Customer's Age in Years
Gender	Demographic variable - M=Male, F=Female
Dependent_count	Demographic variable - Number of dependents
Education_Level	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.
Marital_Status	Demographic variable - Married, Single, Divorced, Unknown
Income_Category	Demographic variable - Annual Income Category of the account holder (< 40K, 40K-60K, 60K-80K, 80K-120K, > 120k)
Card_Category	Product Variable - Type of Card (Blue, Silver, Gold, Platinum
Months_on_book	Period of relationship with 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
Contacts_Count_12_mon	No. of Contacts in the last 12 months
Credit_Limit	Credit Limit on the Credit Card

Avg_Open_To_Buy	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Average Card Utilization Ratio
Total_Revolving_Bal	Total Revolving Balance on the Credit Card

Fig1: Summary of the data set:

```
   CLIENTNUM         Attrition_Flag      Customer_Age      Gender         
 Min.   :708082083   Length:10127       Min.   :26.00   Length:10127      
 1st Qu.:713036770   Class :character   1st Qu.:41.00   Class :character  
 Median :717926358   Mode  :character   Median :46.00   Mode  :character  
 Mean   :739177606                      Mean   :46.33                     
 3rd Qu.:773143533                      3rd Qu.:52.00                     
 Max.   :828343083                      Max.   :73.00                     
 Dependent_count Education_Level    Marital_Status     Income_Category   
 Min.   :0.000   Length:10127       Length:10127       Length:10127      
 1st Qu.:1.000   Class :character   Class :character   Class :character  
 Median :2.000   Mode  :character   Mode  :character   Mode  :character  
 Mean   :2.346                                                           
 3rd Qu.:3.000                                                           
 Max.   :5.000                                                           
 Card_Category      Months_on_book  Total_Relationship_Count Months_Inactive_12_mon
 Length:10127       Min.   :13.00   Min.   :1.000            Min.   :0.000         
 Class :character   1st Qu.:31.00   1st Qu.:3.000            1st Qu.:2.000         
 Mode  :character   Median :36.00   Median :4.000            Median :2.000         
                    Mean   :35.93   Mean   :3.813            Mean   :2.341         
                    3rd Qu.:40.00   3rd Qu.:5.000            3rd Qu.:3.000         
                    Max.   :56.00   Max.   :6.000            Max.   :6.000         
 Contacts_Count_12_mon  Credit_Limit   Total_Revolving_Bal Avg_Open_To_Buy
 Min.   :0.000         Min.   : 1438   Min.   :   0        Min.   :    3  
 1st Qu.:2.000         1st Qu.: 2555   1st Qu.: 359        1st Qu.: 1324  
 Median :2.000         Median : 4549   Median :1276        Median : 3474  
 Mean   :2.455         Mean   : 8632   Mean   :1163        Mean   : 7469  
 3rd Qu.:3.000         3rd Qu.:11068   3rd Qu.:1784        3rd Qu.: 9859  
 Max.   :6.000         Max.   :34516   Max.   :2517        Max.   :34516  
 Total_Amt_Chng_Q4_Q1 Total_Trans_Amt Total_Trans_Ct   Total_Ct_Chng_Q4_Q1
 Min.   :0.0000       Min.   :  510   Min.   : 10.00   Min.   :0.0000     
 1st Qu.:0.6310       1st Qu.: 2156   1st Qu.: 45.00   1st Qu.:0.5820     
 Median :0.7360       Median : 3899   Median : 67.00   Median :0.7020     
 Mean   :0.7599       Mean   : 4404   Mean   : 64.86   Mean   :0.7122     
 3rd Qu.:0.8590       3rd Qu.: 4741   3rd Qu.: 81.00   3rd Qu.:0.8180     
 Max.   :3.3970       Max.   :18484   Max.   :139.00   Max.   :3.7140     
 Avg_Utilization_Ratio
 Min.   :0.0000       
 1st Qu.:0.0230       
 Median :0.1760       
 Mean   :0.2749       
 3rd Qu.:0.5030       
 Max.   :0.9990       
```

Fig 2: Missing Values

```
               CLIENTNUM           Attrition_Flag             Customer_Age 
                       0                        0                        0 
                  Gender          Dependent_count          Education_Level 
                       0                        0                        0 
          Marital_Status          Income_Category            Card_Category 
                       0                        0                        0 
          Months_on_book Total_Relationship_Count   Months_Inactive_12_mon 
                       0                        0                        0 
   Contacts_Count_12_mon             Credit_Limit      Total_Revolving_Bal 
                       0                        0                        0 
         Avg_Open_To_Buy     Total_Amt_Chng_Q4_Q1          Total_Trans_Amt 
                       0                        0                        0 
          Total_Trans_Ct      Total_Ct_Chng_Q4_Q1    Avg_Utilization_Ratio 
                       0                        0                        0 
```

Fig 3: Unknown variables:

variable <chr>	total_unknown <int>
Education_Level	1495
Income_Category	1096
Marital_Status	736
CLIENTNUM	0
Attrition_Flag	0
Customer_Age	0
Gender	0
Dependent_count	0
Card_Category	0

Fig 4: Box plot for outliers:

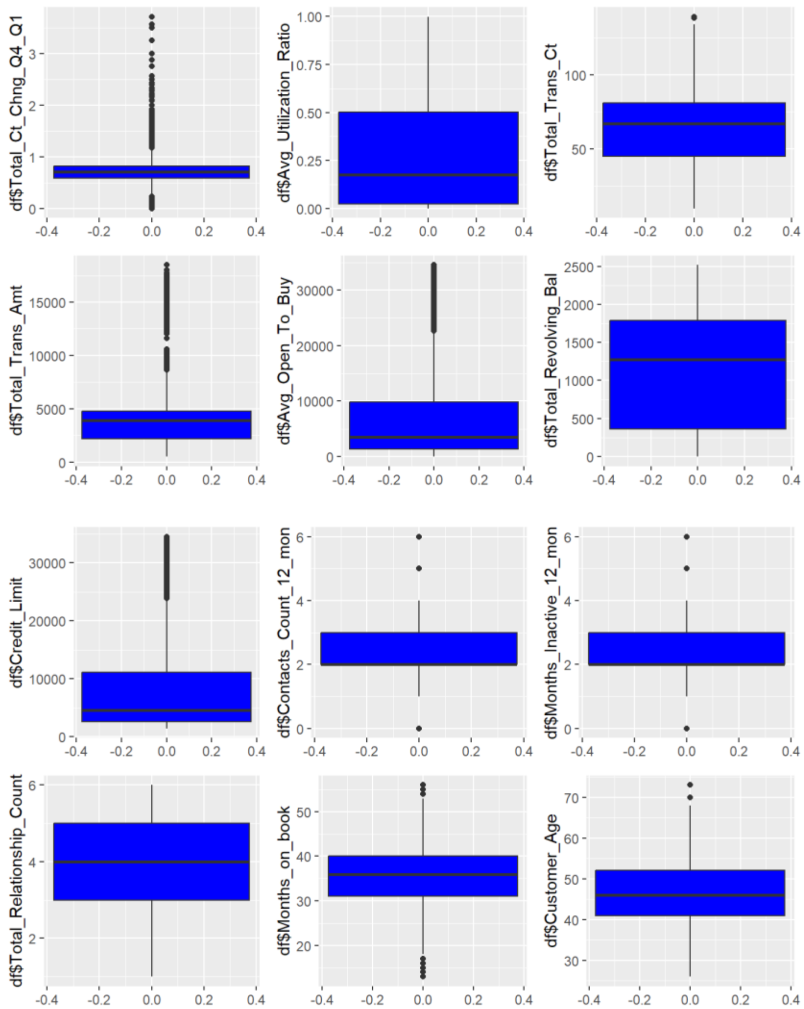


Fig5: Correlation Matrix Numerical:

Note: All the fig are attached in the appendix

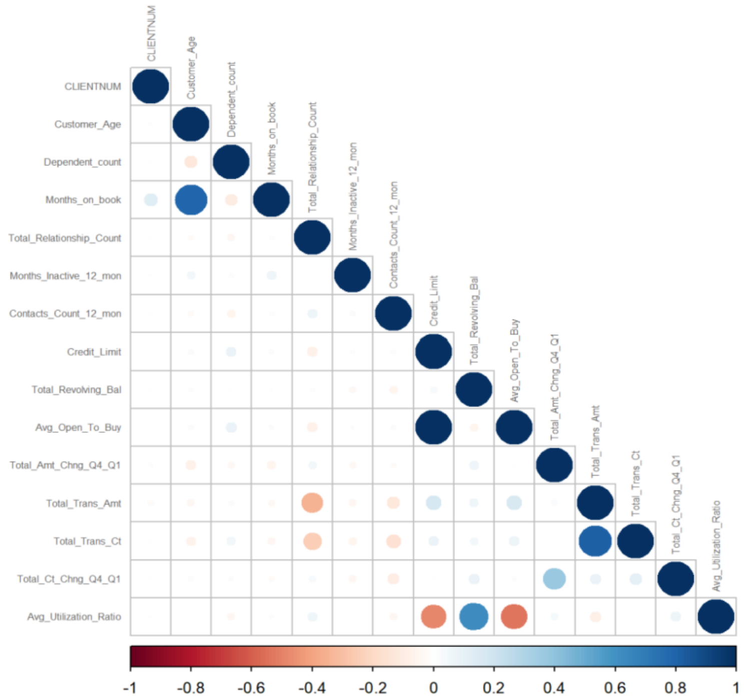


Fig6: Correlation Matrix Categorical :

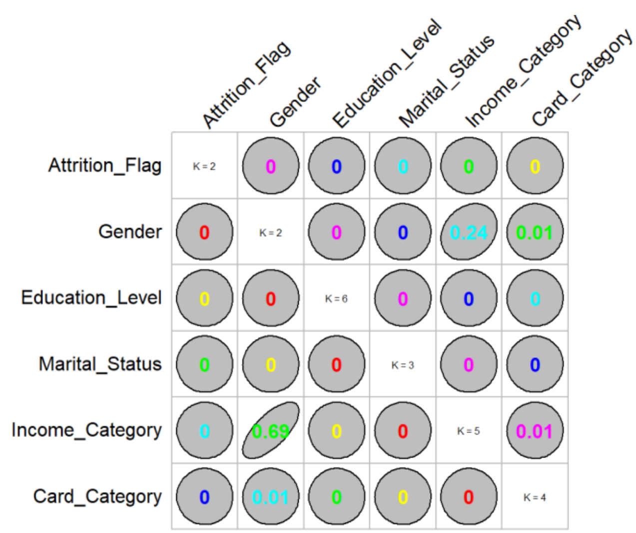
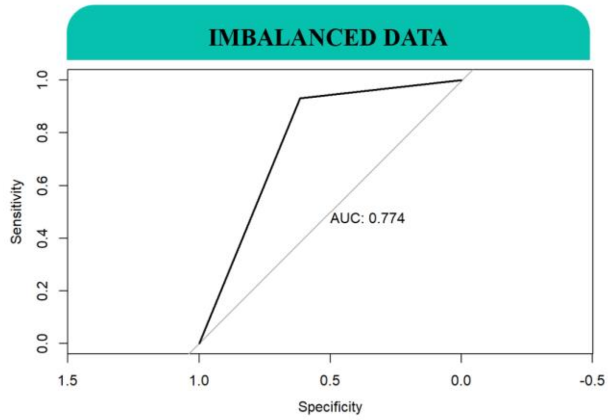
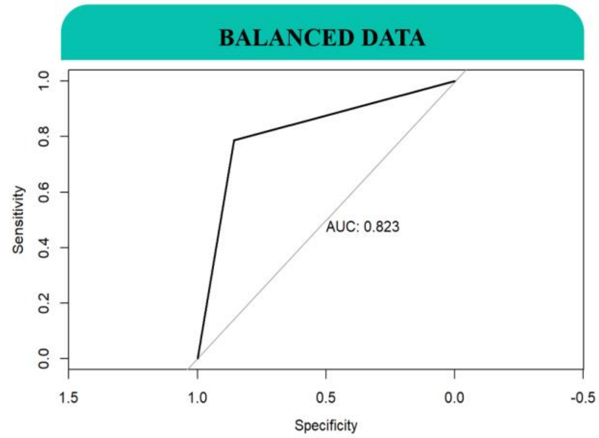


Fig 7: Balancing the data set:

Note: All the fig are attached in the appendix





Increased by 5%

Fig 8 : VIF TABLE

Description: df [29 × 3]

Variables <chr>	Tolerance <dbl>	VIF <dbl>
Customer_Age	0.3704030	2.699763
GenderM	0.2889652	3.460624
Dependent_count	0.9497183	1.052944
Education_LevelDoctorate	0.7368842	1.357065
Education_LevelGraduate	0.3991738	2.505174
Education_LevelHigh School	0.4514740	2.214967
Education_LevelPost-Graduate	0.6822788	1.465676
Education_LevelUneducated	0.5432321	1.840834
Marital_StatusMarried	0.2750280	3.635993
Marital_StatusSingle	0.2773115	3.606053
Income_Category$40K - $60K	0.2259263	4.426223
Income_Category$60K - $80K	0.3563375	2.806328
Income_Category$80K - $120K	0.3558969	2.809803
Income_CategoryLess than $40K	0.1217886	8.210946
Card_CategoryGold	0.8759183	1.141659
Card_CategoryPlatinum	0.9642617	1.037063
Card_CategorySilver	0.6781636	1.474571
Months_on_book	0.3733620	2.678366
Total_Relationship_Count	0.8908960	1.122466
Months_Inactive_12_mon	0.9670202	1.034105
Contacts_Count_12_mon	0.9264119	1.079433
Credit_Limit	0.0000000	Inf
Total_Revolving_Bal	0.0000000	Inf
Avg_Open_To_Buy	0.0000000	Inf
Total_Amt_Chng_Q4_Q1	0.8025399	1.246044
Total_Trans_Amt	0.2828846	3.535011
Total_Trans_Ct	0.2909406	3.437128
Total_Ct_Chng_Q4_Q1	0.7358777	1.358921

Note: All the fig are attached in the appendix

Fig9 : Linear regression with all variables:

```
Call:
glm(formula = Attrition_Flag ~ ., family = "binomial", data = train_clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7415  -0.4638   0.0706   0.4730   2.5893

Coefficients: (1 not defined because of singularities)
                                 Estimate   Std. Error z value           Pr(>|z|)
(Intercept)                   8.879122503  0.616670750  14.398 < 0.0000000000000002 ***
Customer_Age                 -0.020653245  0.009924163  -2.081             0.037424 *
GenderM                      -0.964371003  0.178209299  -5.411         0.0000000625 ***
Dependent_count               0.063452878  0.037394514   1.697             0.089725 .
Education_LevelDoctorate      0.853263292  0.253054782   3.372             0.000747 ***
Education_LevelGraduate       0.506480003  0.159392099   3.178             0.001485 **
Education_LevelHigh School    0.480089025  0.169289438   2.836             0.004570 **
Education_LevelPost-Graduate  0.429737944  0.233297876   1.842             0.065473 .
Education_LevelUneducated     0.408546071  0.186823634   2.187             0.028757 *
Marital_StatusMarried        -0.482575782  0.185522624  -2.601             0.009291 **
Marital_StatusSingle          0.002309729  0.186303731   0.012             0.990108
Income_Category$40K - $60K   -1.073394332  0.248720151  -4.316         0.0000159119 ***
Income_Category$60K - $80K   -0.468780397  0.219077604  -2.140             0.032372 *
Income_Category$80K - $120K  -0.369187979  0.203356543  -1.815             0.069452 .
Income_CategoryLess than $40K -0.880716124  0.274914944  -3.204             0.001357 **
Card_CategoryGold             1.234676783  0.407146068   3.033             0.002425 **
Card_CategoryPlatinum         2.120038099  1.359970496   1.559             0.119024
Card_CategorySilver           0.592127939  0.218300018   2.712             0.006679 **
Months_on_book               -0.000144722  0.009816488  -0.015             0.988237
Total_Relationship_Count     -0.374553532  0.033075337 -11.324 < 0.0000000000000002 ***
Months_Inactive_12_mon        0.617835334  0.055519984  11.128 < 0.0000000000000002 ***
Contacts_Count_12_mon         0.488849644  0.045654182  10.708 < 0.0000000000000002 ***
Credit_Limit                 -0.000013472  0.000008497  -1.585             0.112863
Total_Revolving_Bal          -0.000839336  0.000081527 -10.295 < 0.0000000000000002 ***
Avg_Open_To_Buy                        NA           NA      NA                   NA
Total_Amt_Chng_Q4_Q1         -0.955021377  0.238782336  -4.000         0.0000634636 ***
Total_Trans_Amt               0.000578645  0.000029662  19.508 < 0.0000000000000002 ***
Total_Trans_Ct               -0.133983856  0.004947051 -27.084 < 0.0000000000000002 ***
Total_Ct_Chng_Q4_Q1          -2.389736851  0.217761616 -10.974 < 0.0000000000000002 ***
Avg_Utilization_Ratio        -0.315376028  0.288625512  -1.093             0.274533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5826.2  on 4202  degrees of freedom
Residual deviance: 2886.0  on 4174  degrees of freedom
AIC: 2944

Number of Fisher Scoring iterations: 6
```

Fig10: Step Wise regression

Note: All the fig are attached in the appendix

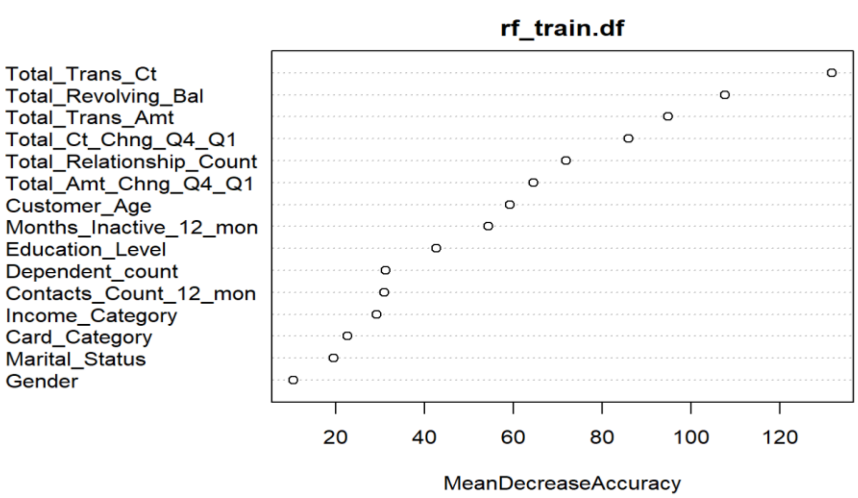
```
Step:  AIC=2940.85
Attrition_Flag ~ Customer_Age + Gender + Dependent_count + Education_Level +
    Marital_Status + Income_Category + Card_Category + Total_Relationship_Count +
    Months_Inactive_12_mon + Contacts_Count_12_mon + Total_Revolving_Bal +
    Total_Amt_Chng_Q4_Q1 + Total_Trans_Amt + Total_Trans_Ct +
    Total_Ct_Chng_Q4_Q1

                           Df Deviance    AIC
<none>                          2888.9 2940.9
+ Credit_Limit              1   2887.2 2941.2
+ Avg_Open_To_Buy           1   2887.2 2941.2
- Dependent_count           1   2891.7 2941.7
+ Avg_Utilization_Ratio     1   2888.5 2942.5
+ Months_on_book            1   2888.8 2942.8
- Education_Level           5   2903.8 2945.8
- Card_Category             3   2904.0 2950.0
- Customer_Age              1   2900.2 2950.2
- Income_Category           4   2908.7 2952.7
- Total_Amt_Chng_Q4_Q1      1   2905.2 2955.2
- Marital_Status            2   2913.9 2961.9
- Gender                    1   2918.4 2968.4
- Contacts_Count_12_mon     1   3011.9 3061.9
- Total_Ct_Chng_Q4_Q1       1   3015.7 3065.7
- Months_Inactive_12_mon    1   3022.8 3072.8
- Total_Relationship_Count  1   3030.3 3080.3
- Total_Revolving_Bal       1   3197.0 3247.0
- Total_Trans_Amt           1   3326.1 3376.1
- Total_Trans_Ct            1   4092.8 4142.8
```

Fig 11: Random Forest;

```
Call:
 randomForest(formula = Attrition_Flag ~ ., data = train_clean,      ntree = 500, mtry = 4,
nodesize = 5, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 1.33%
Confusion matrix:
                  Existing Customer Attrited Customer class.error
Existing Customer              2038                43 0.020663143
Attrited Customer                13              2109 0.006126296
```



Note: All the fig are attached in the appendix

Fig 12: Classification Tree

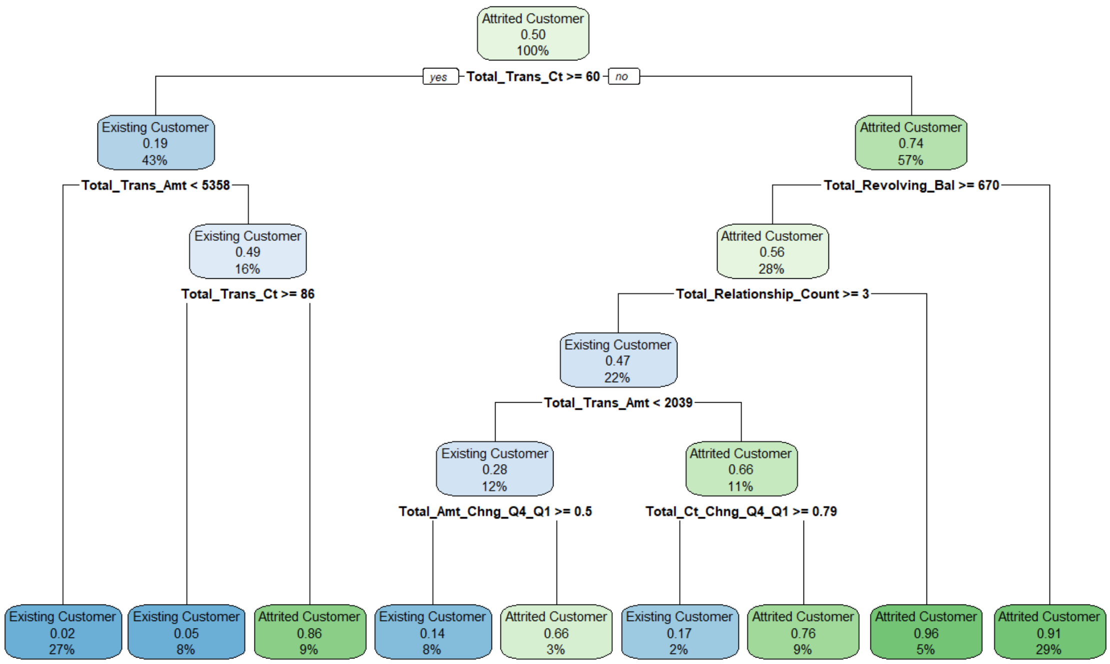


Fig 13: Navie Bayes

Note: All the fig are attached in the appendix

```
Naive Bayes Classifier for Discrete
Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace
= laplace)

A-priori probabilities:
Y
Existing Customer Attrited Customer
        0.4951225          0.5048775

Conditional probabilities:
                   Customer_Age
Y                       [,1]     [,2]
  Existing Customer 46.37818 7.993668
  Attrited Customer 46.45759 7.706261


Total_Relationship_Count
Y                       [,1]     [,2]
  Existing Customer 3.852955 1.553687
  Attrited Customer 3.313855 1.574037

                   Months_Inactive_12_mon
Y                      [,1]      [,2]
  Existing Customer 2.25036 0.9365860
  Attrited Customer 2.61263 0.8267058

                   Total_Revolving_Bal
Y                       [,1]     [,2]
  Existing Customer 1261.472 771.1422
  Attrited Customer  676.181 916.4297

                   Total_Amt_Chng_Q4_Q1
Y                        [,1]      [,2]
  Existing Customer 0.7705805 0.2232258
  Attrited Customer 0.6914562 0.2140632

                   Total_Trans_Amt
Y                       [,1]     [,2]
  Existing Customer 4810.320 3740.677
  Attrited Customer 3182.377 2448.297

                   Total_Trans_Ct
Y                       [,1]     [,2]
  Existing Customer 69.13936 23.34266
  Attrited Customer 44.73563 15.96068

                   Total_Ct_Chng_Q4_Q1
Y                        [,1]      [,2]
  Existing Customer 0.7440927 0.2282483
  Attrited Customer 0.5648134 0.2432939
```

REGRESSION RESULTS:

LOGESTIC:

Note: All the fig are attached in the appendix

```
   Existing Customer Attrited Customer
  0              1230                291
  1               202               1080
```

```{r}
accuracy1 <- sum(cm1[1], cm1[4]) / sum(cm1[1:4])
accuracy1
```

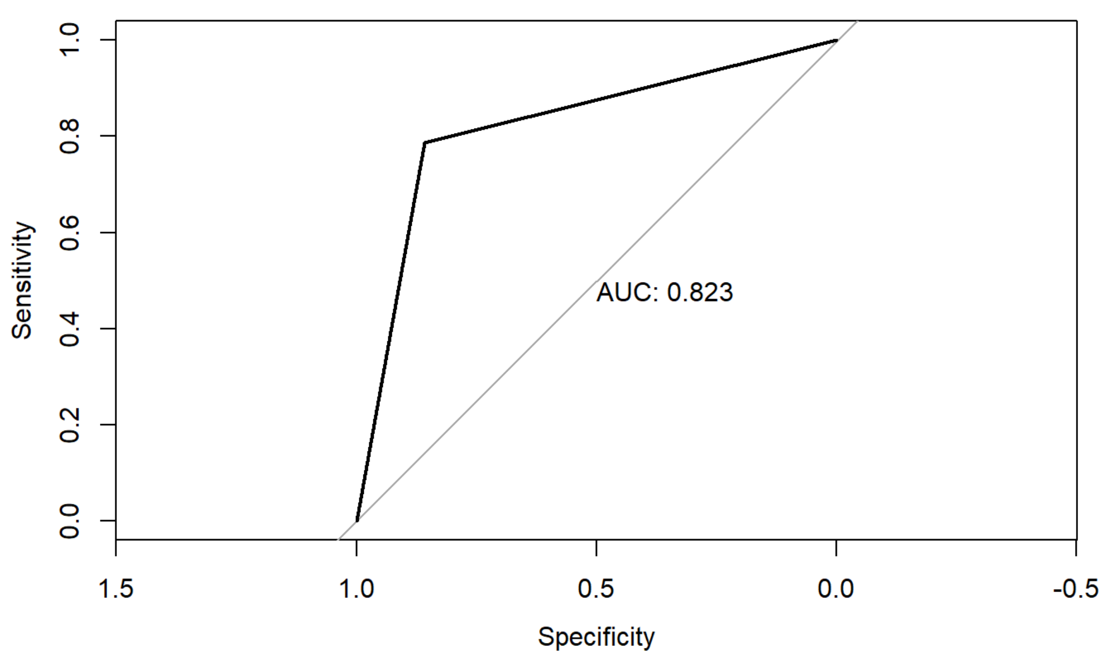
```
[1] 0.824117
```

```{r}
Sensitivity1 <- cm1[1] / sum(cm1[1:2])
Sensitivity1
```

```
[1] 0.8589385
```

```{r}
Specificity1 <- cm1[4] / sum(cm1[3:4])
Specificity1
```

```
[1] 0.7877462
```



Note: All the fig are attached in the appendix

RANDOM FOREST:

```
                   Existing Customer Attrited Customer
Existing Customer               1382               161
Attrited Customer                 50              1210
```

```{r}
accuracy3 <- sum(cm3[1], cm3[4]) / sum(cm3[1:4])
accuracy3
```

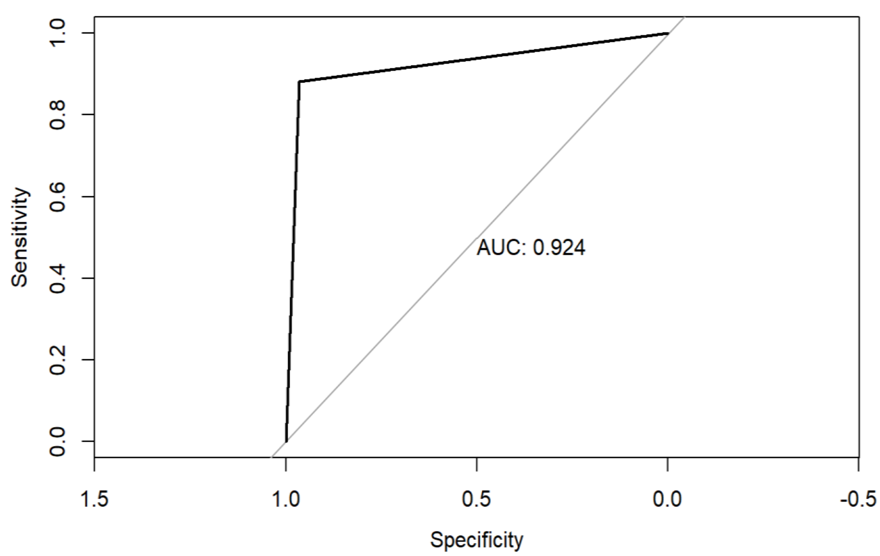
```
[1] 0.9247235
```

```{r}
Sensitivity3 <- cm3[1] / sum(cm3[1:2])
Sensitivity3
```

```
[1] 0.9650838
```

```{r}
Specificity3 <- cm3[4] / sum(cm3[3:4])
Specificity3
```

```
[1] 0.8825675
```



Note: All the fig are attached in the appendix

NAIVE BAYES:

```{r}
accuracy4 <- sum(cm4[1], cm4[4]) / sum(cm4[1:4])
accuracy4
```

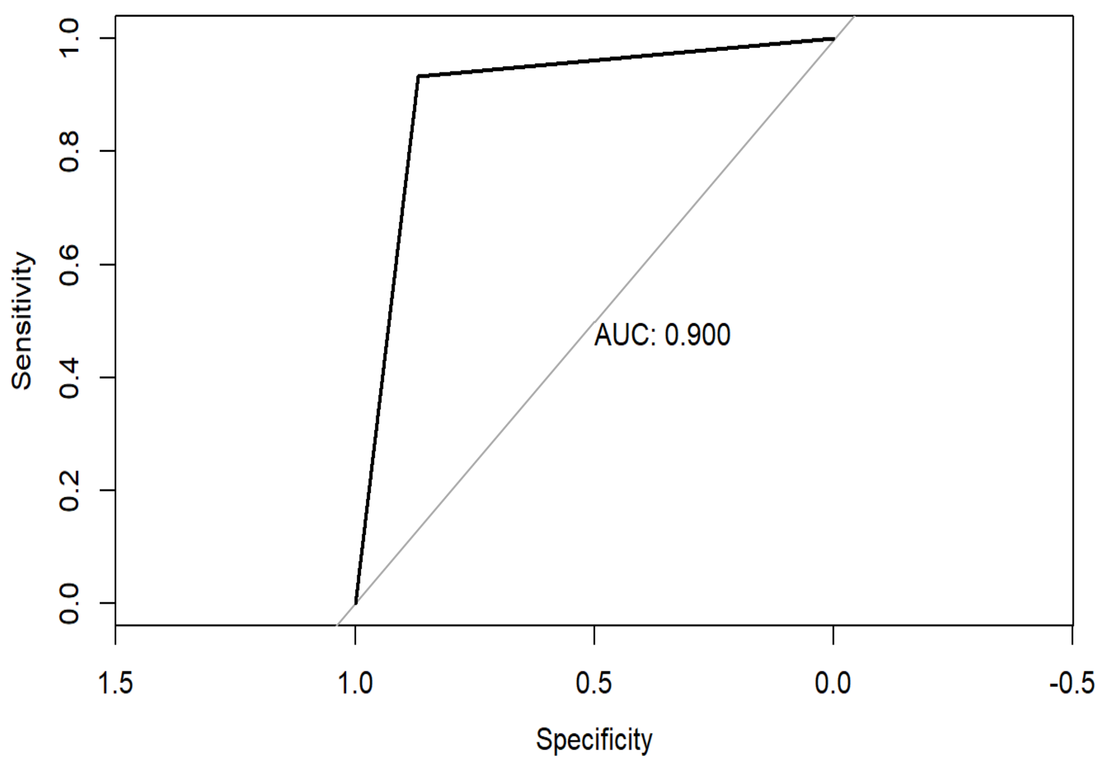
```
[1] 0.8997503
```

```{r}
Sensitivity4 <- cm4[1] / sum(cm4[1:2])
Sensitivity4
```

```
[1] 0.8680168
```

```{r}
Specificity4 <- cm4[4] / sum(cm4[3:4])
Specificity4
```

```
[1] 0.9328957
```



Note: All the fig are attached in the appendix

DECISION TREE

```
                   Existing Customer Attrited Customer
  Existing Customer              1075               207
  Attrited Customer               357              1164
```

```{r}
accuracy5 <- sum(cm5[1], cm5[4]) / sum(cm5[1:4])
accuracy5
```

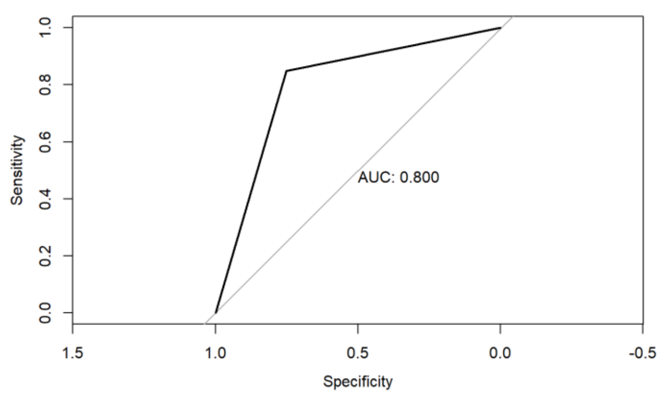
```
[1] 0.798787
```

```{r}
sensitivity5 <- cm5[1] / sum(cm5[1:2])
sensitivity5
```

```
[1] 0.7506983
```

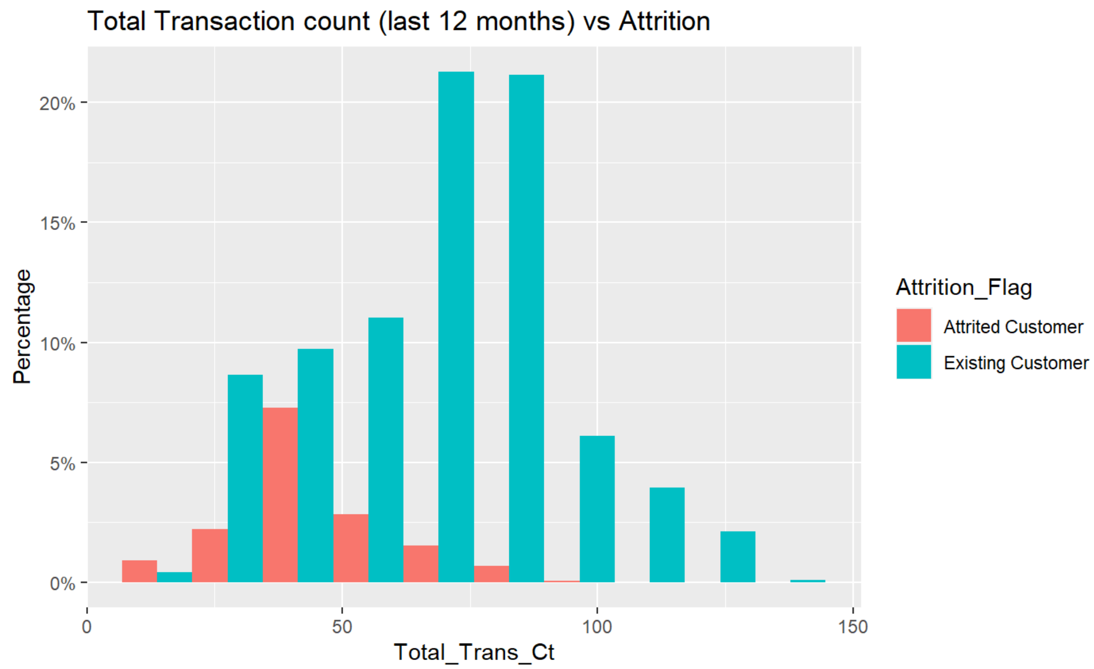
```{r}
specificity5 <- cm5[4] / sum(cm5[3:4])
specificity5
```

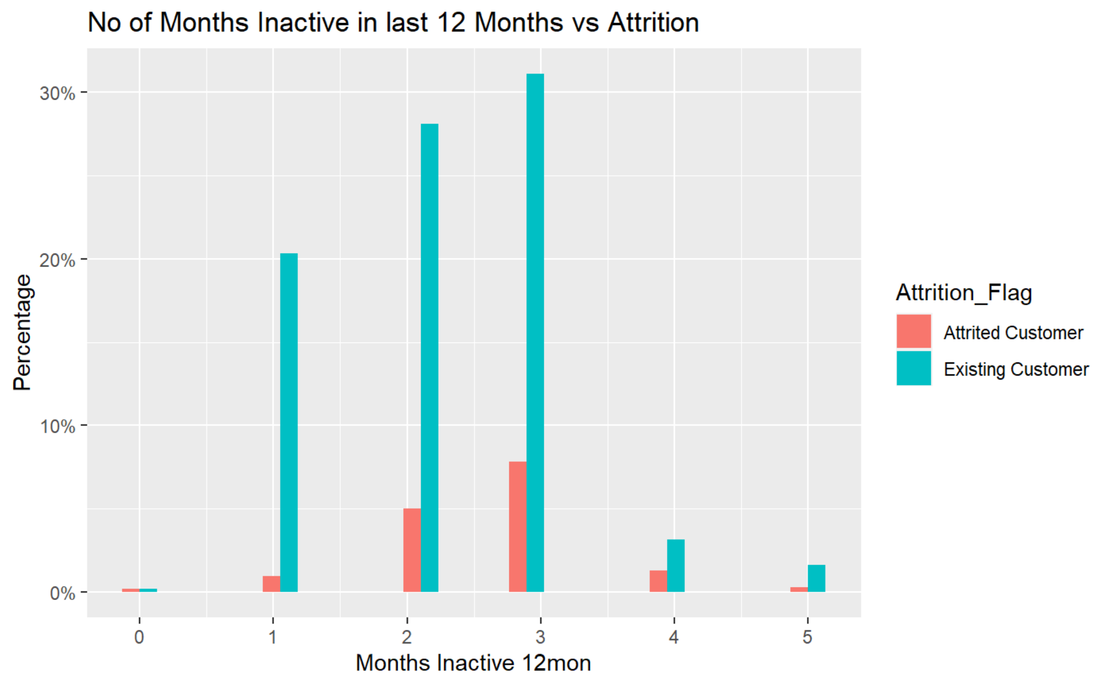
```
[1] 0.8490153
```



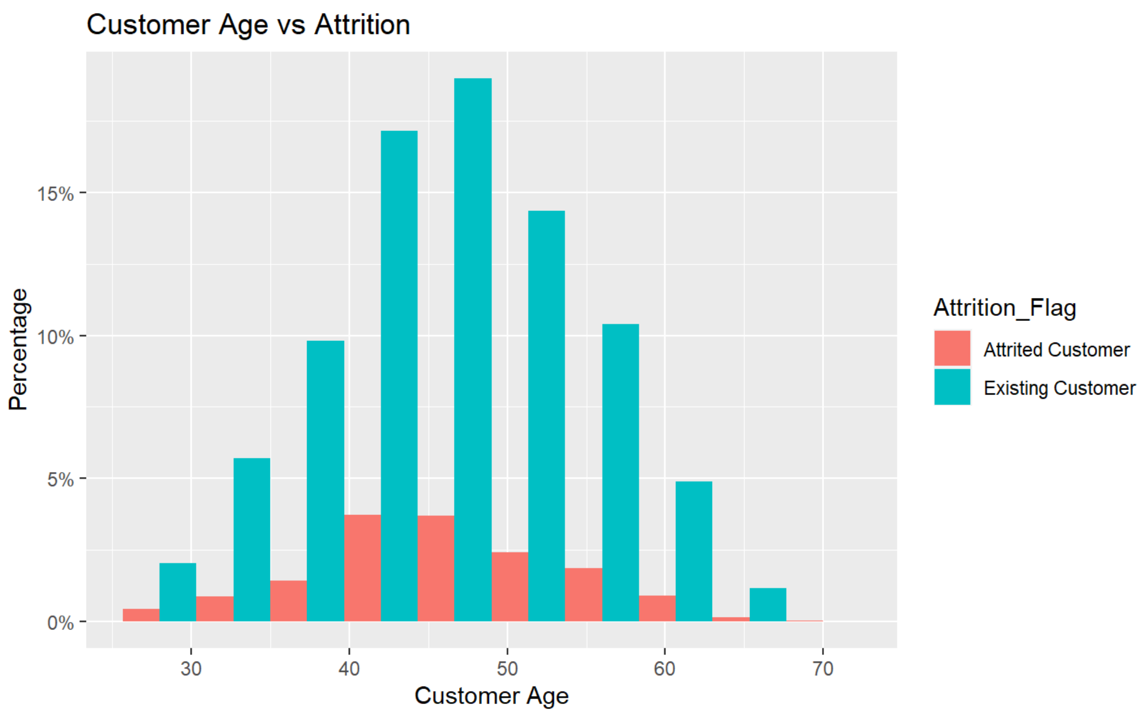
Note: All the fig are attached in the appendix

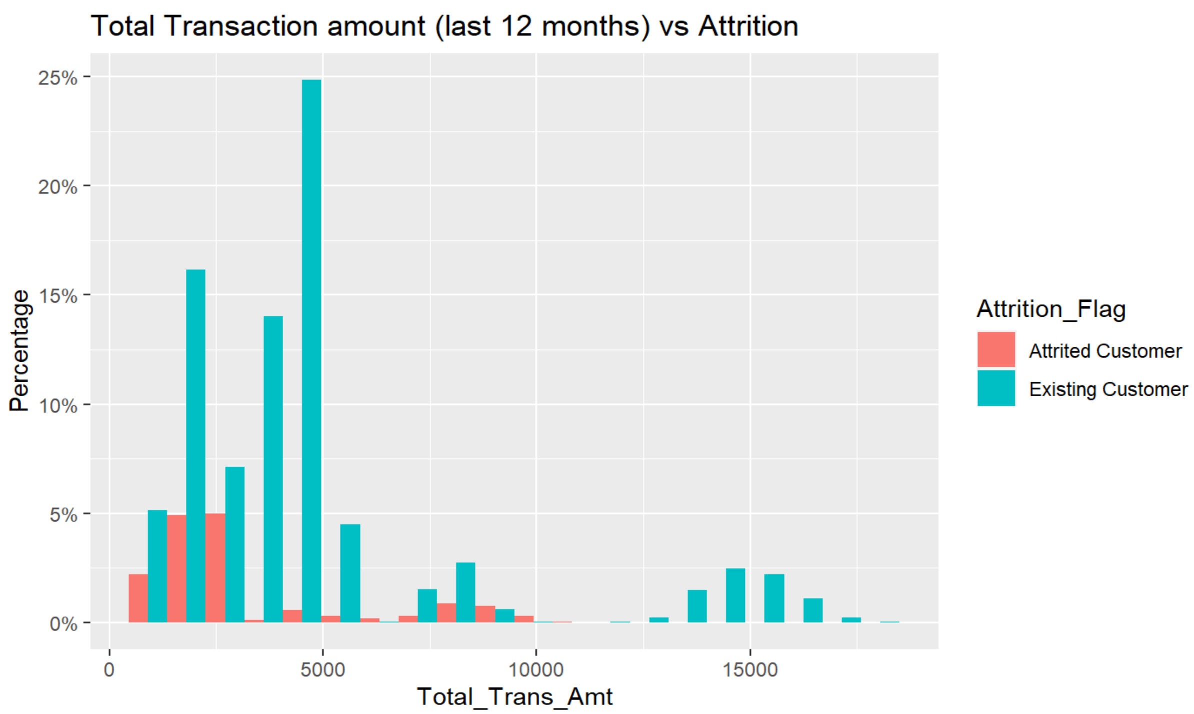
OTHER VISUALIZATIONS:



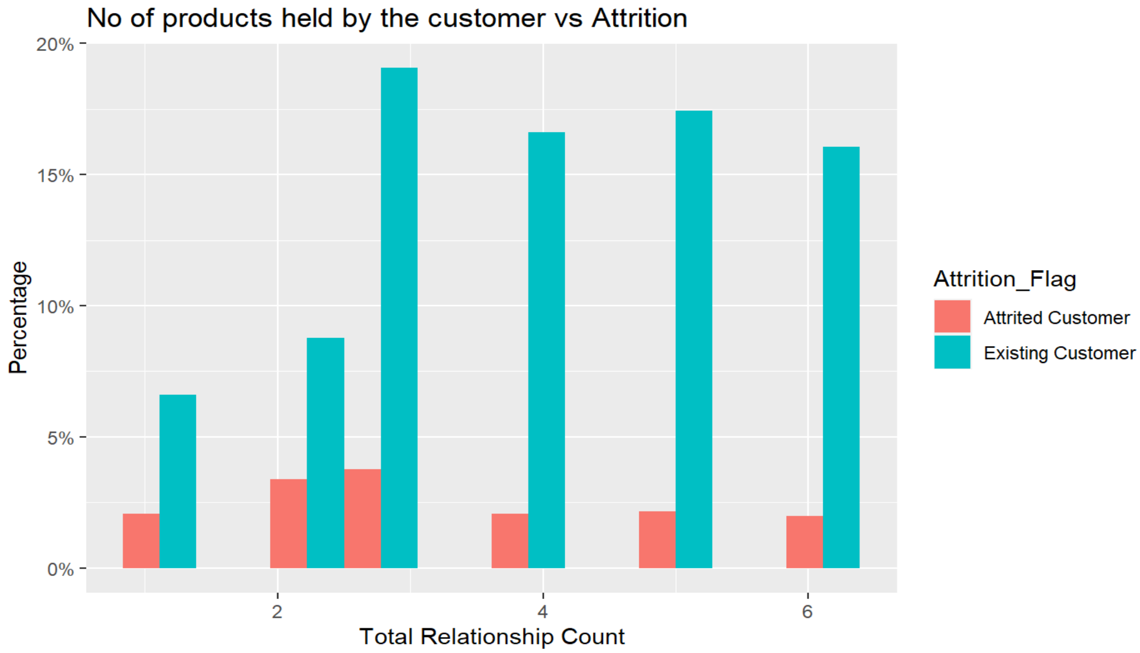


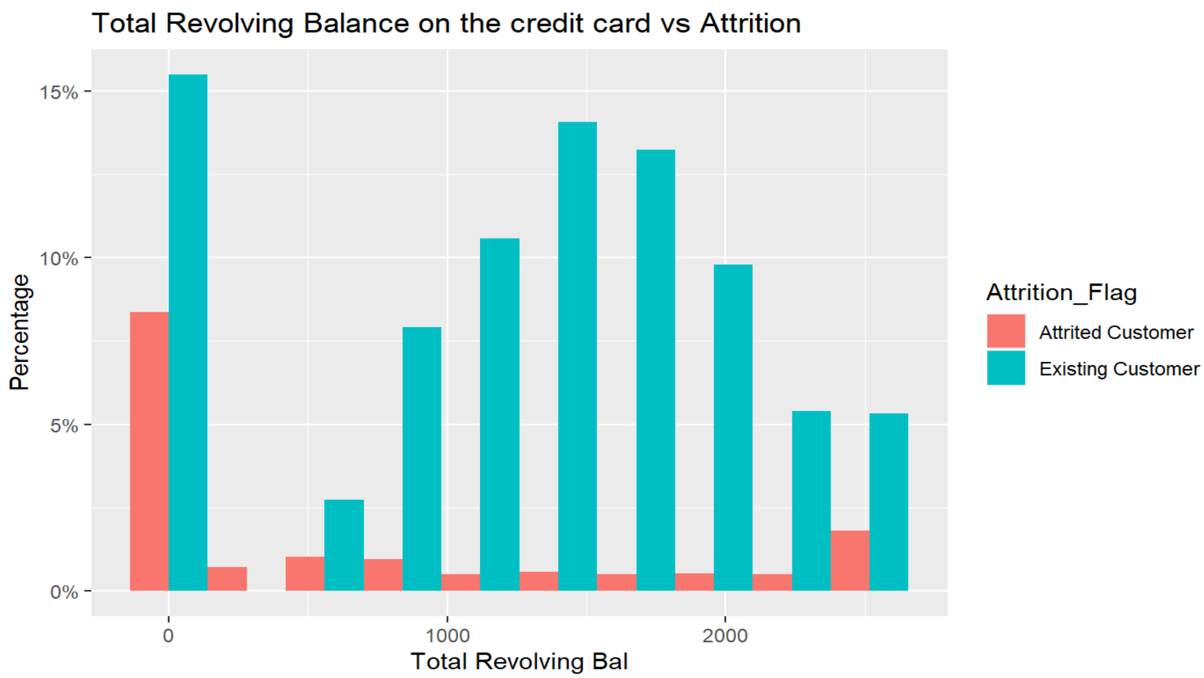
Note: All the fig are attached in the appendix



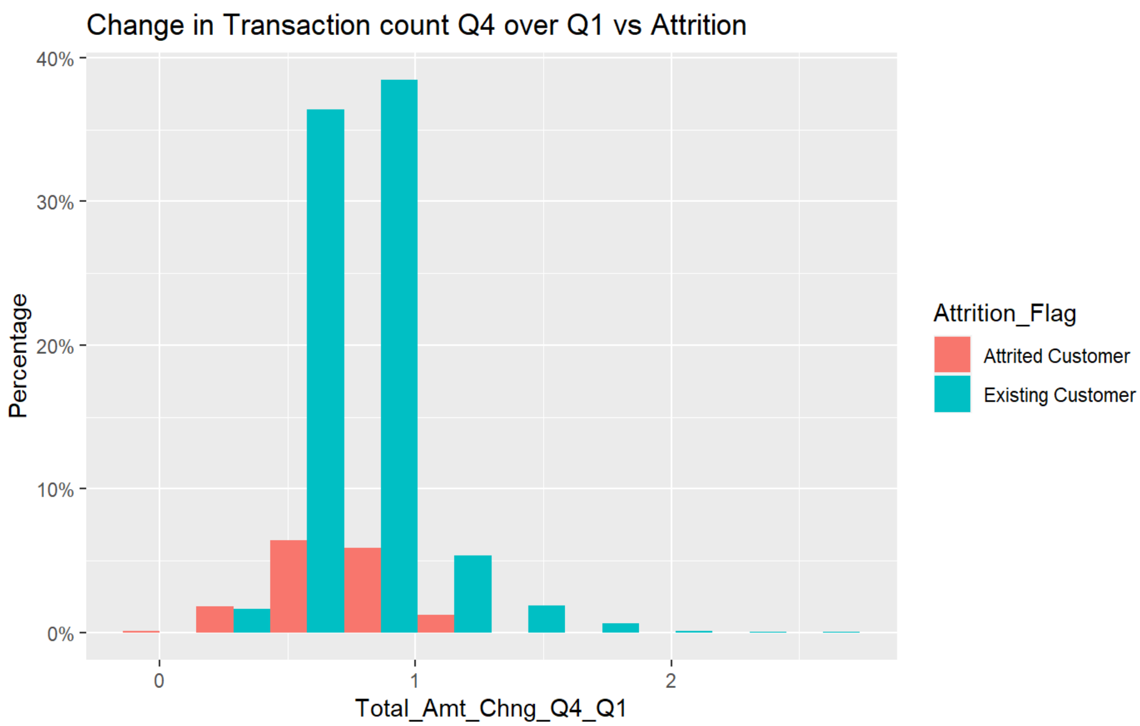


Note: All the fig are attached in the appendix





Note: All the fig are attached in the appendix



Note: All the fig are attached in the appendix