**CUSTOMER’S ATTRITION BY USING LOGISTIC REGRESSION IN BANKING INDUSTRY**

A DISSERTATION

SUBMITTED TO

DEPARTMENT OF STATISTICS

SVKM’S NMIMS (DEEMED -TO-BE UNIVERSITY)

IN PARTIAL FUFILLMENT FOR THE DEGREE OF

**MASTER OF SCIENCE**

**IN STATISTICS AND DATA SCIENCE**

BY

SUKHADA SAKHALKAR

PRANAJAL GAHANKARI

SAURABH BHADALE

SNEH SHAH

HRISHIKESH GURAV

UNDER THE SUPERVISION OF

PROF. PRIYESH TIWARI

**SUNANDAN DIVATIA SCHOOL OF SCIENCE**

SVKM’S Narsee Monjee Institute of Management Studies

(Deemed-To-Be University)

Vile Parle West, Mumbai, Maharashtra 400056

APRIL, 2022

**Group information:**

|  |  |  |
| --- | --- | --- |
| Name | Sap Id | Year |
| Sukhada Sakhalkar | 75322100032 | 2021-22 |
| Pranjal Gahankari | 75322100041 | 2021-22 |
| Saurabh Bhadale | 75322100046 | 2021-22 |
| Sneh shah | 75322100059 | 2021-22 |
| Hrishikesh Gurav | 75322100061 | 2021-22 |

**GLOSSARY AND TERMINOLOGIES**

**Precision:**

Precision is the ratio of the properly classified cases to the total number of misclassified cases and properly classified cases. The equations of precision can be explained as follows:

𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 = 𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒 /𝑇𝑟𝑢𝑒𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒+𝐹𝑎𝑙𝑠𝑒𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒

**Recall:**

Recall is the proportion of correctly classified cases to total number of correctly classified cases and unclassified ones. Recall is represented mathematically by the following equation: 𝑅𝑒𝑐𝑎𝑙𝑙 = 𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒/ 𝑇𝑟𝑢𝑒𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒+𝐹𝑎𝑙𝑠𝑒𝑁𝑒𝑔𝑎𝑡𝑖𝑣𝑒

**F-Score:**

F-score integrates the precision and recall measure which is regarded as a good indicator of relationship between them. It can be represented as given below:

𝐹𝑠𝑐𝑜𝑟𝑒 = 2 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 ∗𝑅𝑒𝑐𝑎𝑙𝑙 / 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 +𝑅𝑒𝑐𝑎𝑙𝑙

**Accuracy:**

Like-wise Accuracy gives the ratio of the total number of predictions that have been calculated properly. It is mathematically represented as shown below:

𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒+𝑇𝑟𝑢𝑒 𝑁𝑒𝑔𝑎𝑡𝑖𝑣𝑒 𝑇𝑟𝑢𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒+𝐹𝑎𝑙𝑠𝑒 𝑁𝑒𝑔𝑎𝑡𝑖𝑣𝑒+𝐹𝑎𝑙𝑠𝑒 𝑃𝑜𝑠𝑖𝑡𝑖𝑣𝑒+𝑇𝑟𝑢𝑒 𝑁𝑒𝑔𝑎𝑡𝑖𝑣𝑒.

**Hyperparameter tuning by gridsearch:**

In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values

**ACKNOWLEDGEMENT**

We would like to express our warmest thanks of gratitude to our supervisor Prof Priyesh Tiwari for providing us the opportunity to do this project on Customer’s attrition by using logistic regression in banking industry. Their invaluable guidance and advice helped us to carry out the project successfully.

We would like to thank our course coordinator Dr. Pradnya Khandeparkar for providing necessary facilities required for completion of this project.

We would also like to extend our gratitude to the whole Department of Statistics for guiding and supporting us at different stages of our project.

Finally, with great support of our team members and excellent guidance of mentors we are able to complete our project within specific time.

**ABSTRACT**

The customer churn prediction (CCP) is one of the challenging problems in the banking industry. With the advancement in the field of machine learning, the possibilities to predict customer churn has increased significantly. Our proposed methodology, consists of five phases. In the first phase, data pre-processing is performed. In the second phase, Exploratory data analysis is done. In the third phase, data analysis part is completed. Next, the data has been split into two parts train and test set in the ratio of 80% and 20% respectively. In the prediction process, most popular predictive models have been applied, namely, logistic regression, KNN, random forest on train set. Odds ratio was also used to see which variables affects the most on the churn rate. Random forest outperforms other model with accuracy of

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Sr.No | Topic | Page No |
| 1 | Introduction |  |
| 2 | Motivation |  |
| 3 | Objective |  |
| 4 | Literature Review |  |
| 5 | Data Description |  |
| 6 | Data Pre-processing |  |
| 7 | Exploratory Data analysis |  |
| 8 | Methodology |  |
| 9 | Results |  |
| 11 | Conclusion |  |
| 12 | Limitations |  |
| 13 | Future Scope |  |
| 14 | References |  |

**Background of the study**

* Banking industry face major challenge with customer churn, as customers switch to the service due to various reasons like credit limit, multi service offerings, marketing promotions by competitors, etc.
* Identifying these potential customers early on who may voluntarily churn and providing them retention incentives in form of discounts & combo offers will help the organization to retain those customers and reduce revenue loss.
* The banks can also internally study any possible operational causes and improve its product offerings.
* Proactive actions will prevent the loss of revenue for the banks and will improve / retain the market share among the industry peers in terms of the number of active customers.

**INTRODUCTION**

The churn rate, also known as the rate of attrition or customer churn, is the rate at which customers stop doing business with an entity. It is most commonly expressed as the percentage of service subscribers who discontinue their subscriptions within a given time period. A high customer churn rate indicates that the company is losing customers at an alarming rate. Customer churn can be attributed to myriad reasons and it is the company that needs to discover these reasons via patterns and trends that are present in customer data. Modern businesses nowadays employ complex algorithms to predict customers that are most likely to churn, i.e., move away from the company. By using such algorithms, companies can know in advance, the customers that are most likely to give up the company’s services and therefore, come up with customer retention strategies to mitigate the losses that the company might face.

**CHURN IN BANKING**:

With the growing competition in banking industry, banks are required to follow customer retention strategies while they are trying to increase their market share by acquiring new customers. It is shown that improving the retention rate by up to 5 % can increase a bank’s profit up to 85 %. Additionally, attracting new customers costs more to any company rather than retaining the old ones who are likely to produce more profit. Thus, banks should maintain their competitive advantage by taking the advantage of machine learning models to predict customer churn.

We have randomly sampled population of 10000 customers from three European-based banks, this project intends to propose an efficient predictive model for customer churn in banking industry, using different supervised classification techniques. Model performance, goodness of fit, feature selection, class imbalance and dealing with outliers will be discussed in the following sections.

**MOTIVATION**

Acquiring a new customer can be five to 25 times more expensive than holding on to an existing one. Keeping your current customers happy is generally more cost-effective than acquiring first-time customers. You do not need to spend big on marketing, advertising or sales outreach. It is easier to turn existing customers into repeating ones, since they already trust your brand from previous purchases. New customers, however, often require more convincing when it comes to that initial sale. Customer loyalty will not just give you repeat business. Loyal customers are more likely to give free recommendations to their colleagues, friends and family. Creating that cycle of retained customers and viral marketing is one way your company can cultivate customer loyalty for long-term success.

**ROLE OF DATA ANALYTICS:**

Customer Churn data analysis can help the company understand the underlying reasons why

customers may choose to leave the company. By implementing predictive analytics techniques and applying them to existing customer churn data from records, it is possible to understand the likelihood of customers switching or discontinuing the service. Customers with high a probability of switching can then be worked with to make sure they remain with the current provider.

**OBJECTIVE OF THE PROJECT:**

* The objective is to predict the customers who may attrite from the existing bank service in near future and prevent the business loss.
* Analyze using standard SEMMA (Sample, Explore, Modify, Model and Assess) approach and choose the best model
* Recommend strategies to bank based on analysis of service they are offering that will help in retaining the customer based on available data.

**PROBLEM STATEMENT:**

The attrition and acquisition of users are the major concerns in banking industry. The fast growth of marketplace in every business is giving rise to increased customer base. Accordingly, banks have recognized the significance of retaining the customers who is on hand. It has become necessary for banks to reduce the churn rate of customers since the inattention might negatively influence profitability of the bank. Churn prediction contributes to identify those users who are likely to switch the service. Banking is enduring the problem of ever-increasing churn rate. Accordingly, the current study employs machine learning algorithm on big-data platform. Machine learning algorithm techniques facilitate these to be protected with efficient approaches for lessening the rate of churn. Silent churn is one type which is considered complicated to predict since there might have such kind of users who might probably churns in the near future. It must be the aim of the decision-maker and advertisers to lessen the churn ratio since it is a recognized fact that comparatively.

**LITERATURE REVIEW**

In [1] a brief idea is given about Customers churn data analysis in Telecom industry by using Logistic regression.

In this paper [2] customers churn analysis done by using machine learning algorithms like KNN, Naive bayes model, logistic regression.

Xia and Jin [3], Presented that the Customer Churn forecast was estimated to be based on Support Vector Machine (SVM). The analysed data sets have been obtained from the University of California machine learning and home telecommunications database. Compared to the BPANN, Decision Tree. the logistic regression, and classification system in Naïve Bayes, SVM has good predictions, high generation capabilities, and a good match.

In [4] Hemlata Jain, Ajay Khotta, Sumit Srivastava predicted customer churn in telecommunication industry by using logistic regression and Logit boost.

**DATA COLLECTION**The data that is feasible for analysis in banking dataset has been used and the prediction has been carried out for the same.

**DATA DESCRIPTION**

This is Bank customers dataset. In our dataset we have 10128 rows. Variables included in the dataset are described below. Out of 21 variables. We also replace binary values of the outcome variable (Exited) with “Exiting customers” and “attrited customers” labels to have a better representation of outputs when visualizing results and discussing the performance.

In this dataset we have 5 categorical features like Gender, Education level, marital status, income category, card category.

And 14 numerical features ustomer\_Age'.

'Dependent\_count',

'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',

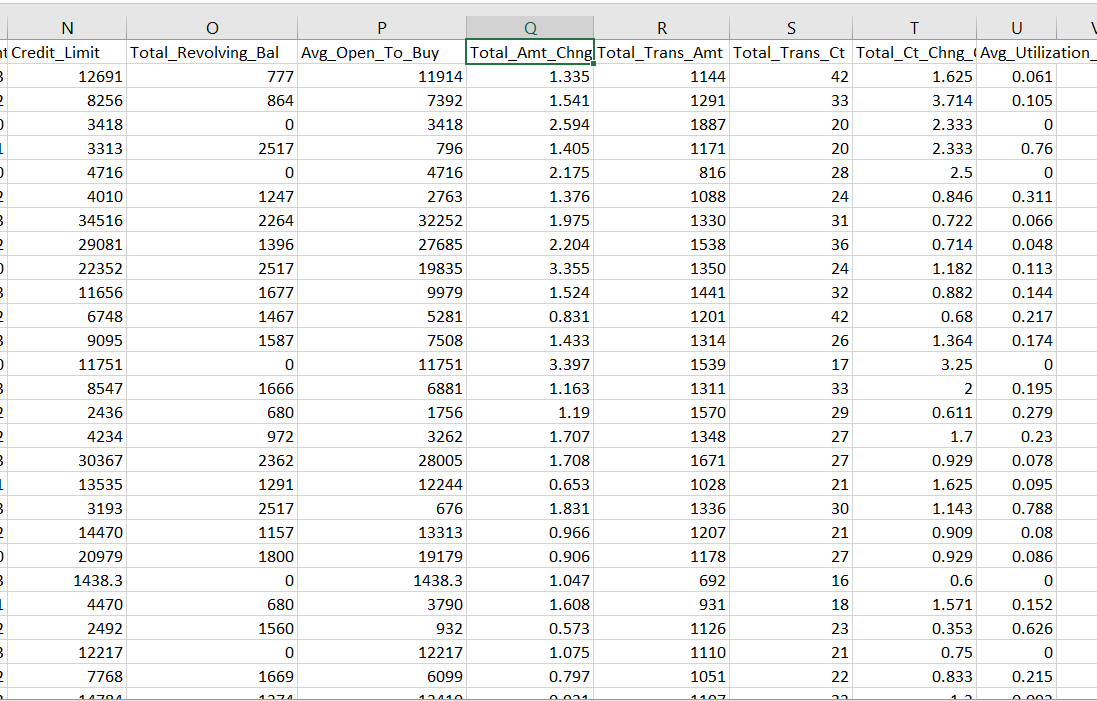
'Total\_Trans\_Ct',

'Total\_Ct\_Chng\_Q4\_Q1',

'Avg\_Utilization\_Ratio'

**SNAPSHOT OF DATA:**



****

**DESCRIPTION OF DATA ATTRIBUTES:**

* Client number: Unique identifier for the customer holding the account.
* Attrition flag: if account is closed then 1 otherwise 0
* Customer age: customer age in years.
* Dependent count: Number of dependents
* Education levels: educational qualification of the account holder.
* Marital Status: married, single, divorced, Unknown
* Income category: Annual income category of the account holder.
* Card\_ category: Type of card (blue, silver, Gold, platinum)
* Months on book: Period of relationship with banks
* Total relationship count: Total no. of products held by the customer
* Months inactive (12 months): No. of months inactive in the last 12 months
* Contact count: No. of Contacts in the last 12 months
* Credit limit: Credit Limit on the Credit Card
* Total revolving balance: Total Revolving Balance on the Credit Card
* Avg open to buy: Open to Buy Credit Line (Average of last 12 months)
* Total amount of change: Change in Transaction Amount (Q4 over Q1)
* Total trans amount: Total Transaction Amount (Last 12 months)
* Total transaton ct: Total Transaction Count (Last 12 months)
* Total ct chng Q: Change in Transaction Count (Q4 over Q1)
* Avg\_Utilization\_Ration: Average Card Utilization Ratio

**DATA PREPROCESSING:**

* **Unnecessary feature removal:**

Before processing further, we need to remove features that cannot be used on modelling. The CLIENTNUM feature contains unique ID for every customer so it cannot be used as Predictor (input) for the Target (output), and should be dropped.

 let's define which features are numerical and categorical and assign each to a variable.

* **Data Type of the variables:**

In this dataset we have 5 catogorical features like Gender, Education level, marital staus, income category, card category.

And 14 numerical features customer\_Age'.

'Dependent\_count', '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',

'Avg\_Utilization\_Ratio'

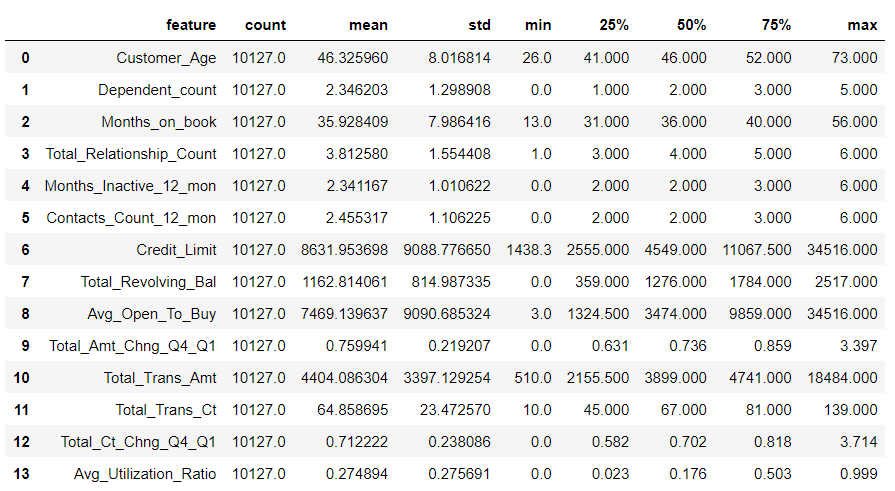
* **Missing Values Check:**

There is no missing value found on the dataset.

* **Predictor feature Exploration**

First, define the numerical predictors, which are all the numerical features we defined earlier. Assign it to a variable, as it will be used later for slicing the data frame.

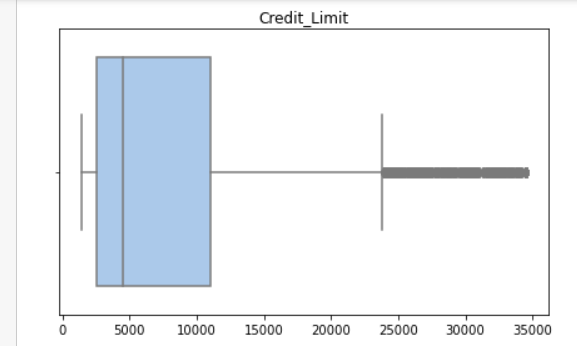
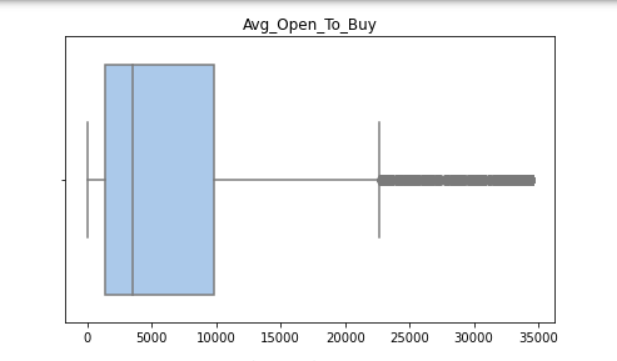
**DESCRIPTIVE STATISTICS:**

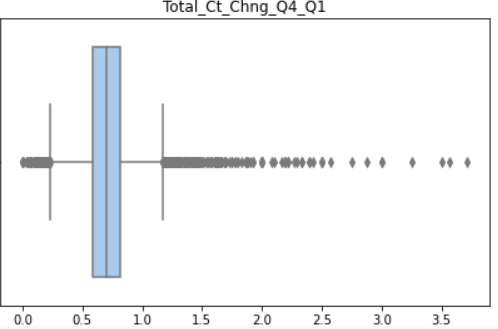
****

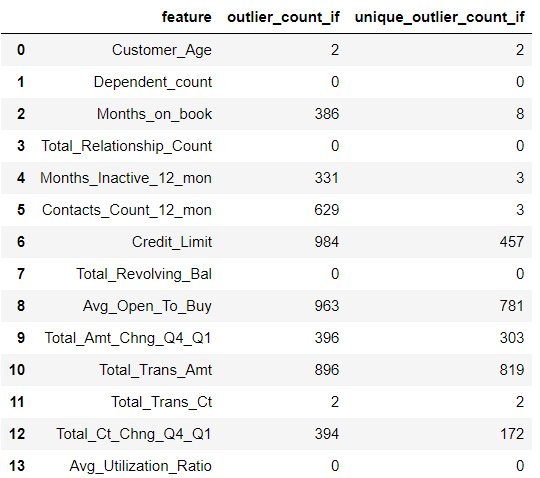
From above descriptive statistics we get an idea about minimum, maximum, mean, standard deviation, 25 percentile, 50th percentile, 75th percentile.

**OUTLIER:**

By plotting the Boxplot for each variable, the outliers are identified. The below variables have outliers.





**Inner fence and outer fence:**

Fences can be used to illustrate extreme values(outliers) in boxplot. Sometimes you might see reference to “inner fences” and “outer fences”. These are defined as:

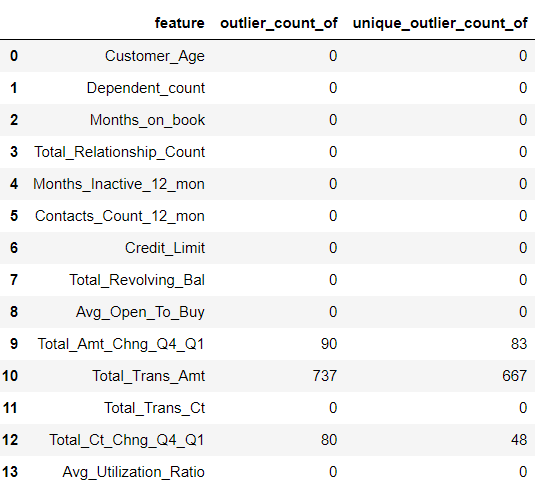
* Lower inner fence: Q1 – (1.5 \* IQR)
* Upper inner fence: Q3 + (1.5 \* IQR)
* Lower outer fence: Q1 – (3 \* IQR)
* upper outer fence: Q3 + (3 \* IQR)

Outer Fence The lower boundary of this method is defined by Q1 - (3 \* IQR), and the upper boundary is defined by Q3 + (3 \* IQR)

We remove this outlier from our dataset using inner Fence and outer Fence.

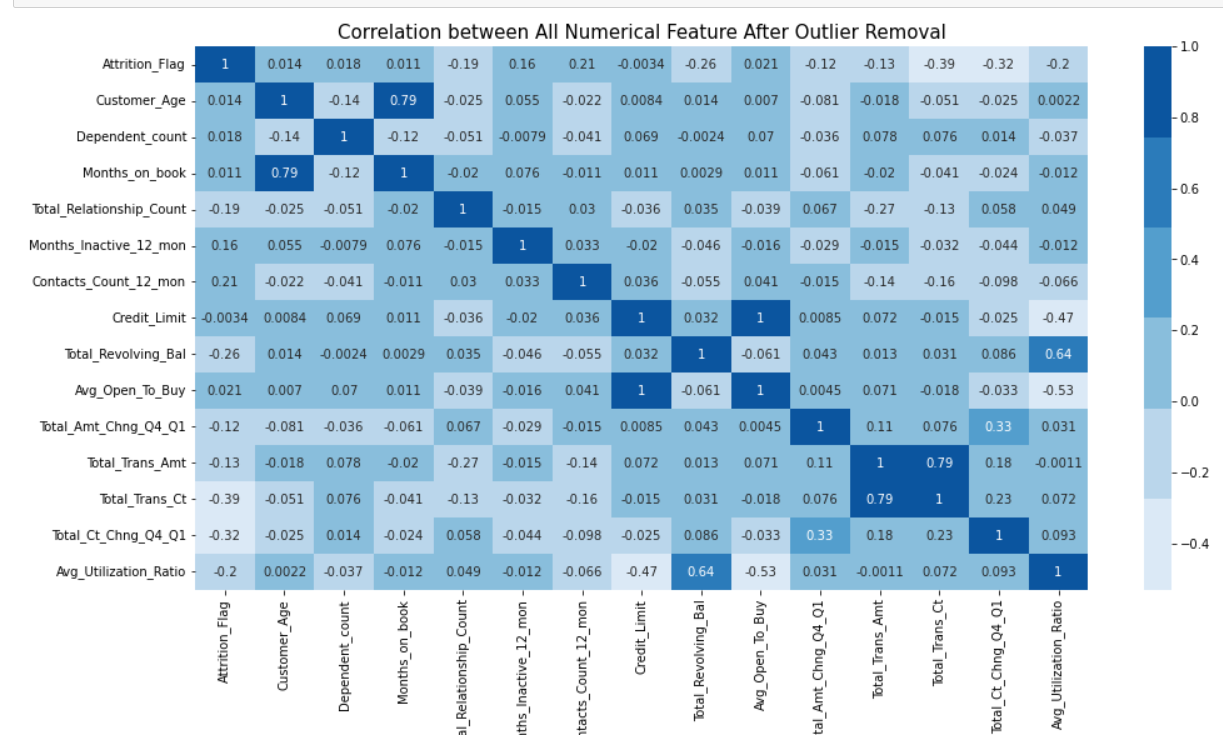
Inner Fence, rows removed: 32.84%

Outer Fence, rows removed: 8.77%



Removing too much rows would make our data not representative of the actual dataset. Normally, it should only be around 5% to 10%. Therefore, we will use the Outer Fence method to find and remove the outliers as it only removes 8.77% or the rows.

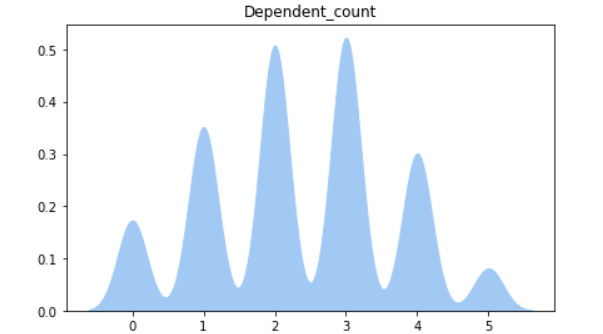
**CORRELATION Plot for all numerical features:**

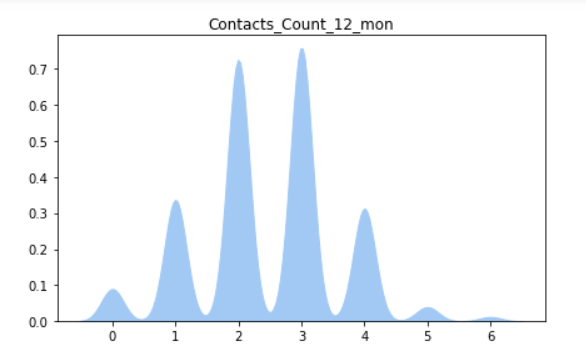


From correlation plot we conclude that Avg\_open to buy is highly correlated and hence it lead to be cause for multicollinearity. Therefore, we drop this variable it.

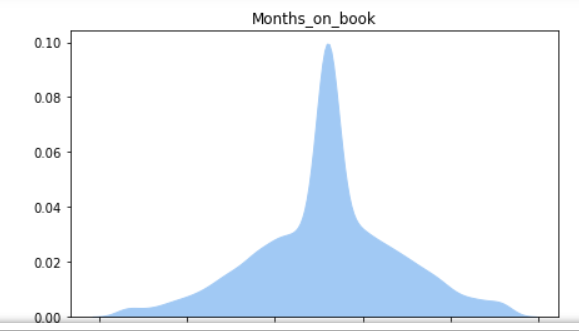
* **To Check the distributions of the numerical features**

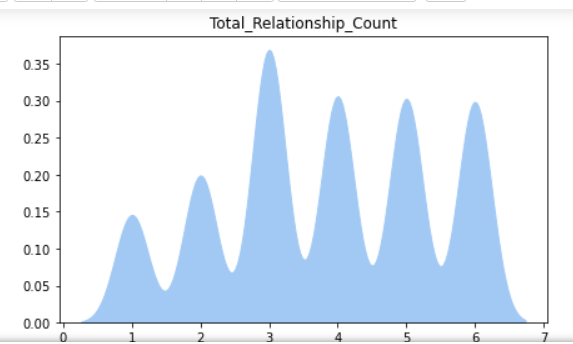
****

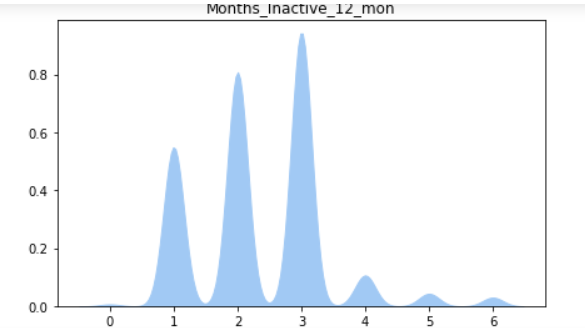
****

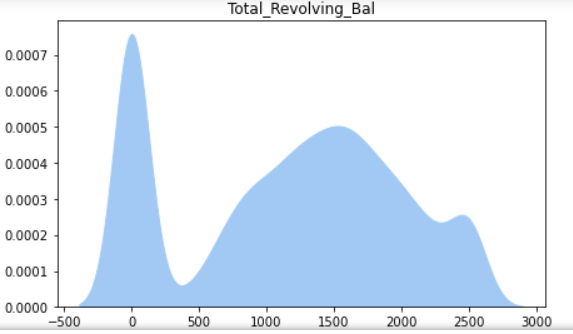
****

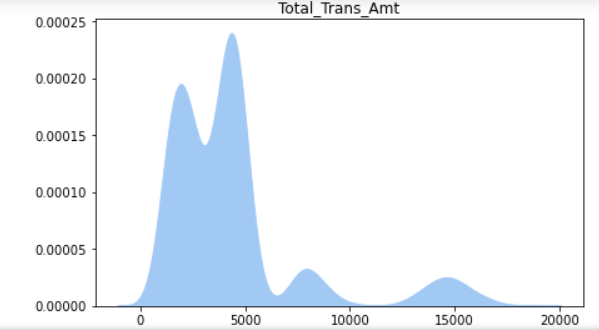
The above features Customer\_age, Dependent count, contact couny 12 months have gaussian distribution. Rest of other doesn’t have normal distribution.

****

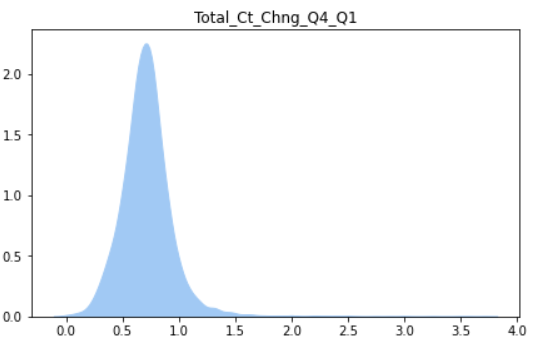
****

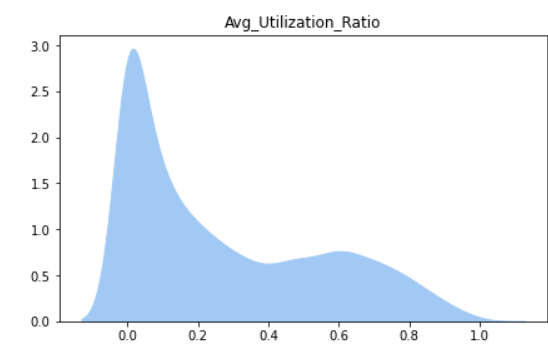
****

****

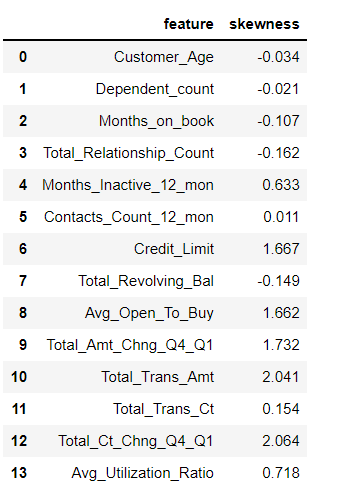
****

****

****

****

Another approach of seeing the distribution of the data is by checking the skewness values.

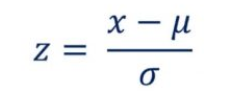
****

Features that have skewness between -0.05 and 0.05 are assumed to have gaussian distribution, which are Customer\_age, Dependent count, contact\_count\_12 mon

**STANDARDIZATION:**

Standardization is a scaling method where the values are cantered around mean with a unit standard deviation. It means if we will calculate mean and standard deviation of standard scores it will be 0 and 1 respectively.

The formula for standardized values:



Where,

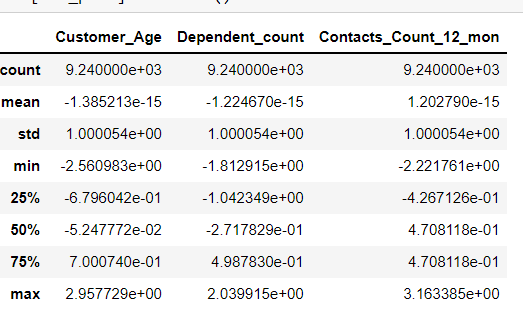
µ= mean of the given distribution

σ = standard deviation of the given distribution

This Z is called standard score and it represents the number of standard deviations above or below the mean that a specific observation falls.

Hence these variables are normally distributed we use standard scalar for standardization.

Features to be standardized are Customer\_Age, Dependent\_count, Contacts\_Count\_12\_mon



**NORMALIZATION:**

It is a scaling technique method in which data points are shifted and rescaled so that they end up in a range of 0 to 1. It is also known as **min-max scaling**.

The formula for calculating normalized score:

**X new = (X — X min)/ (X max — X min)**

Here, Xmax and Xmin are the maximum and minimum values of the feature respectively, the

data doesn't have a gaussian distribution, so we use Normalization.

['Months\_on\_book',

'Total\_Relationship\_Count',

'Months\_Inactive\_12\_mon',

'Credit\_Limit',

'Total\_Revolving\_Bal',

'Total\_Amt\_Chng\_Q4\_Q1',

'Total\_Trans\_Amt',

'Total\_Trans\_Ct',

'Total\_Ct\_Chng\_Q4\_Q1',

'Avg\_Utilization\_Ratio'

This are the features which needs to be normalized.

**ENCODING:**

**Label Encoding** refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

We encoded Education l;evel as 'Unknown':0 , 'Uneducated':1, 'High School':2, 'College':3, 'Graduate':4, 'Post-Graduate':5, 'Doctorate':6

Encode Income\_Category as 'Unknown':0, Less than $40K':1, $40K - $60K':2, '$60K - $80K':3, '$80K - $120K':4, $120K +':5

Encode Card\_Category as Blue':0, 'Silver': 1,’Gold’:2, 3 'Platinum':

* Train Test split: By using sklearn library we split the data into train and test

Train set shape: (7392, 20)

Test set shape: (1848, 20)

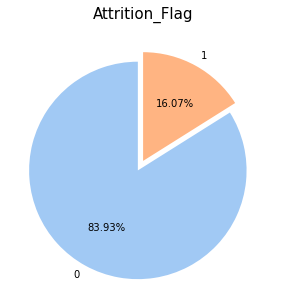
**EXPLORATORY DATA ANALYSIS**

* **Target Feature Exploration:**

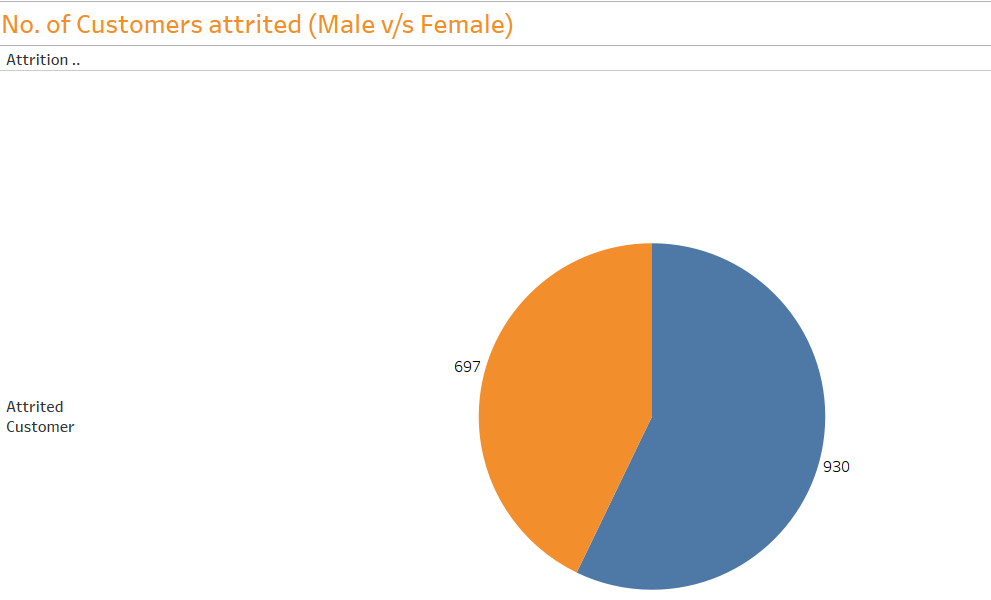
The Target Feature (output) that we want to predict is Attrition Flag, which has two unique values, that is 'Existing Customer' and 'Attrited Customer'. Before doing the exploration, encode those values into 0 and 1.

Existing Customer': 0, Attrited Customer:1

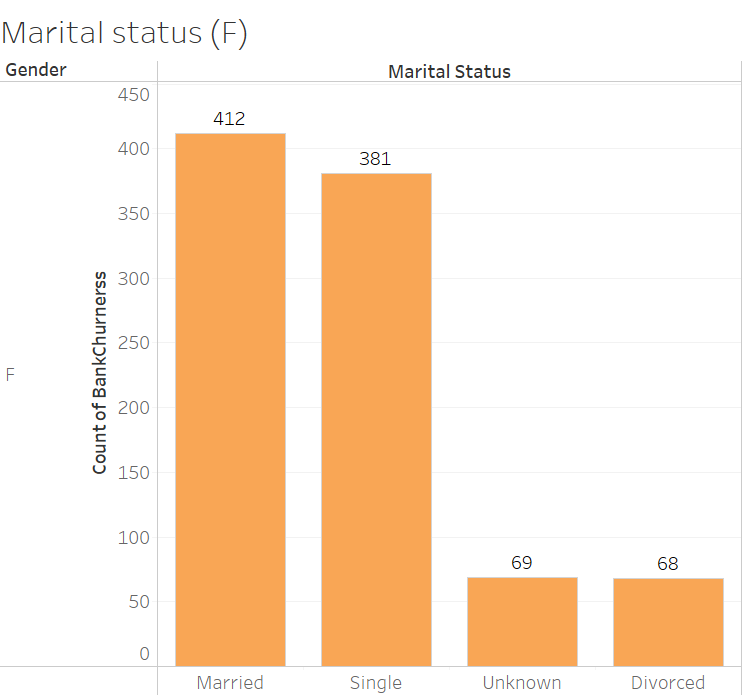
* **Target feature Distribution:**

****

Here we can see that we have 16.07% of customers that attrited /churned, or 1627 out of 10127.here is high percentage of peoples who will continues the service. Now we see among this churned people how many males and females discontinues the service.

****

From above pie chart it is observed that the large number of female customers are churning as compare to male.



Married women are churned more as compare to Single, Unknown, Divorced. It may cause due to after marriage they migrate from one place to another

Education status:



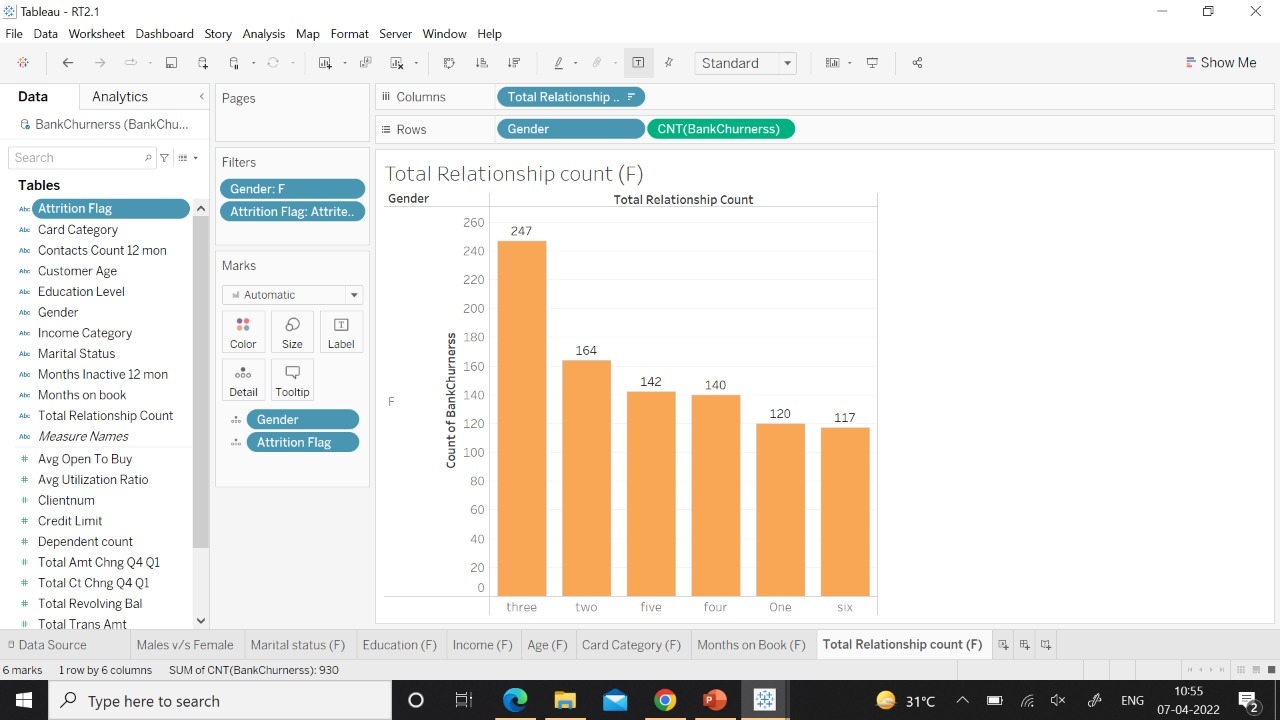
Those females are graduated and studying in high school churned more. It may cause due to they don’t have substantial money to keep in the bank.

Income:



Those female have monthly income less than 40k they churned more.They discontinues the service because less income.

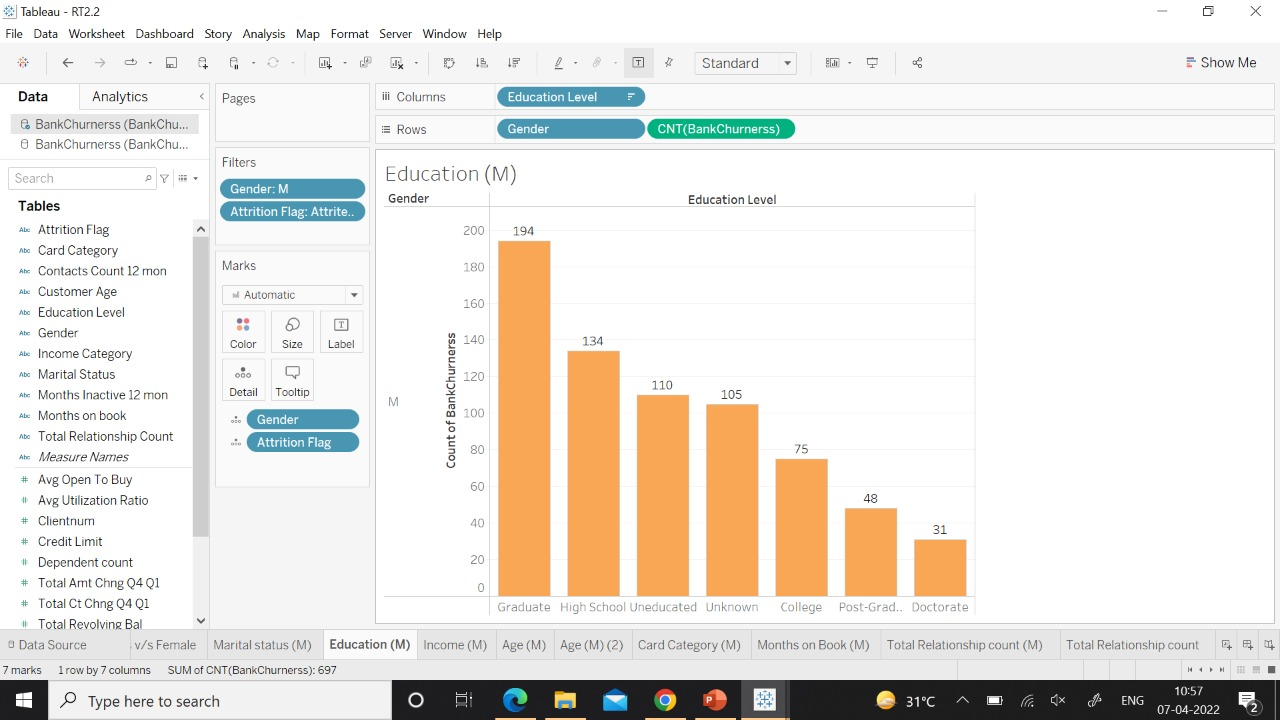
Total Relationship Count:



Those female customers having 3 products like credit card, debit card etc they more likely to churn

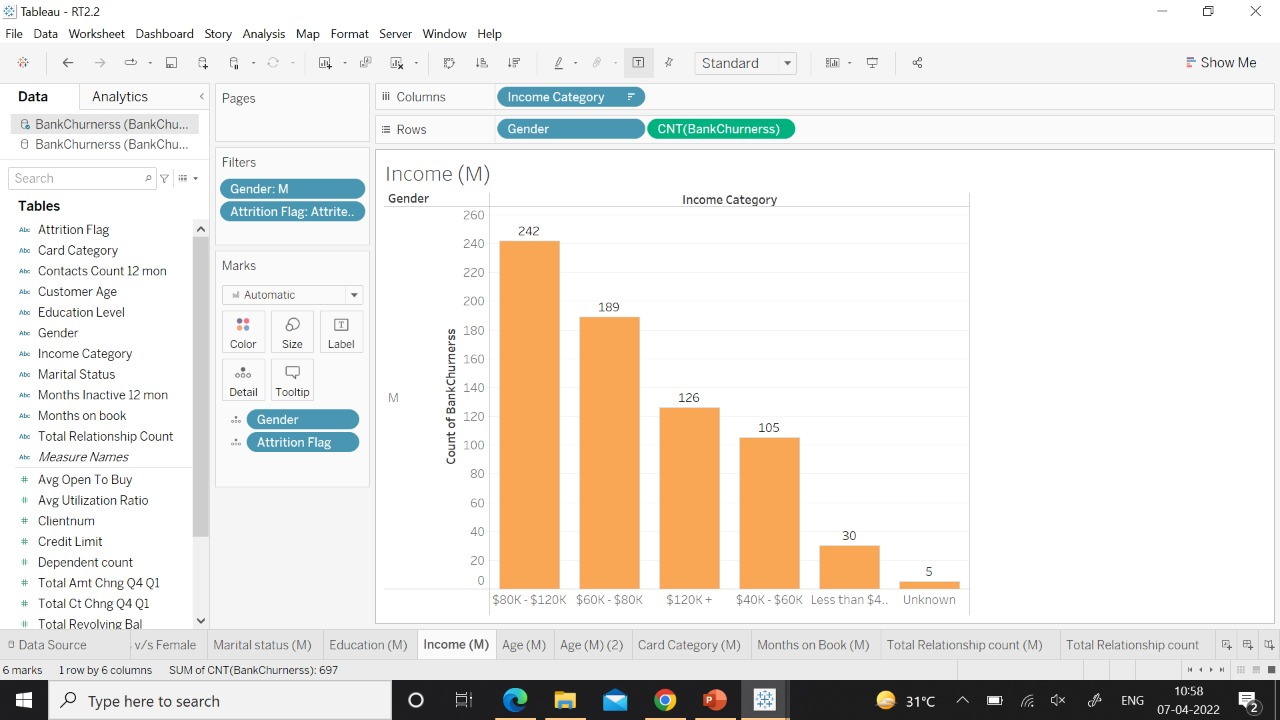
Now for Male:

Education:



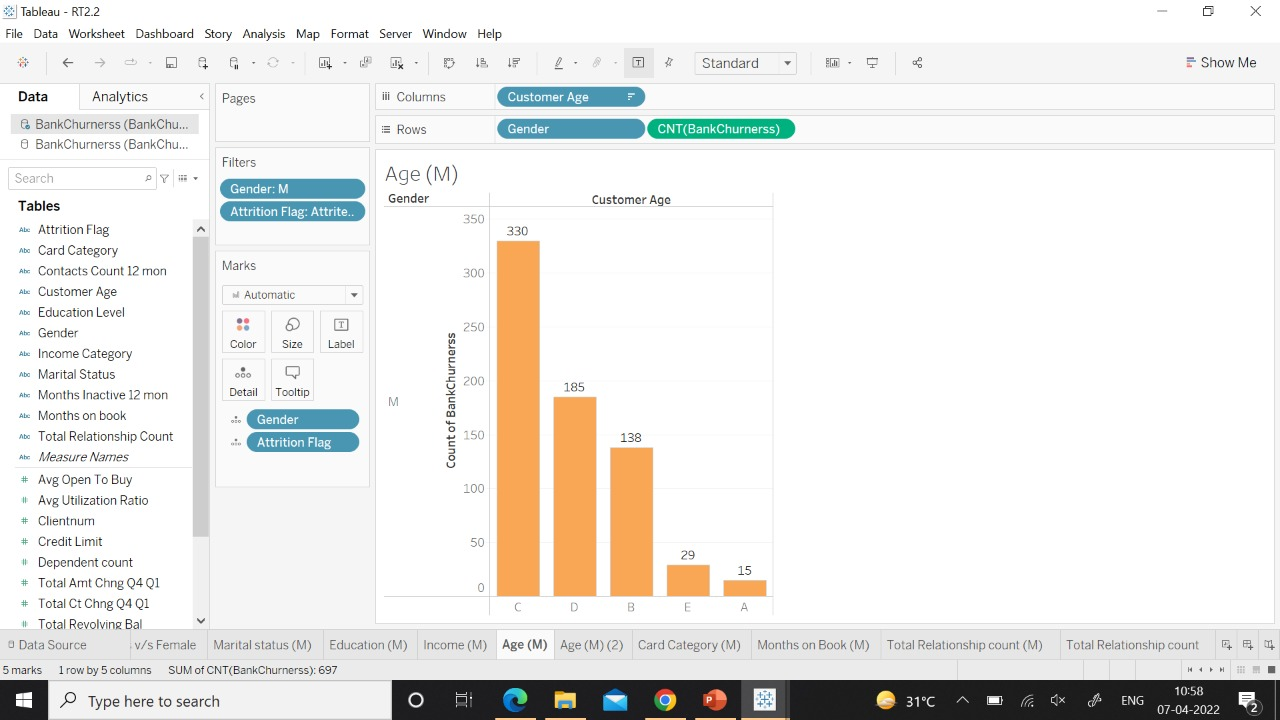
Those males are graduate and in high school have high number of churners. From above graph ewe conclude that 194 and 134 males are churning from graduate and high school respectively.

Income:



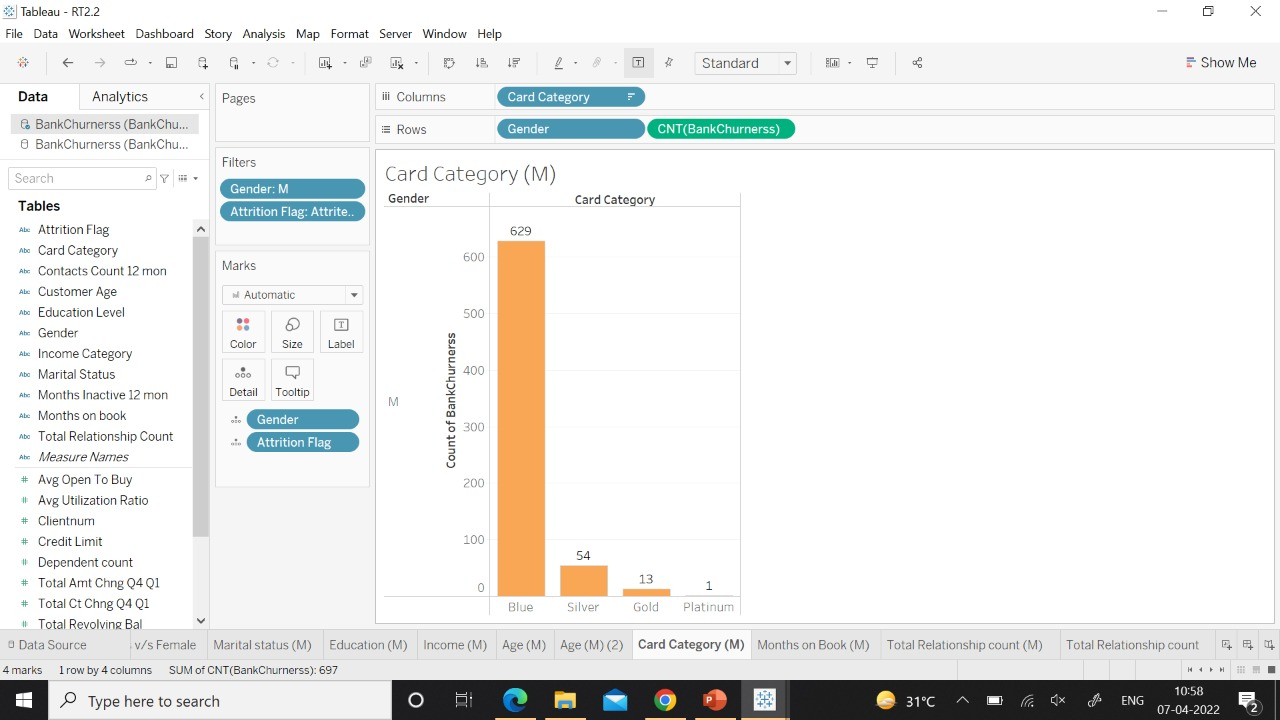
Those males have monthly income in between 80k and 120k are churning more. It may cause due to they getting better offer and service from another bank as they are earning more.

Age:



We divided males into 6 categories A-25 to 30, B-31 to 40, C-41 to 50, D-51 to 60 , E-61 to 70, F-71 to 80. From above graph we conclude that those males have age group between 41 to 50 they are churning more. It may cause due extra expenses that time.

Card Category:



Males having blue card are churning more as compare to silver, gold and platinum. Blue card is debit card payment system operating in France. It may have happened due to less credit limit.

**METHODOLOGY**

**To treat Imbalanced data:**

**Synthetic Minority Oversampling Technique** or **SMOTE**is technique to oversample the minority class. Simply adding duplicate records of minority class often don’t add any new information to the model. In SMOTE new instances are synthesized from the existing data. If we explain it in simple words, SMOTE looks into minority class instances and use *k* nearest neighbour to select a random nearest neighbour, and a synthetic instance is created randomly in feature space.

**MODEL FITTING:**

**LOGISTIC REGRESSION**

 Logistic regression is a binary classification algorithm belonging to the generalized linear regression model. It can also be used to solve problems with more than 2 classes. It is possible to use logistic regression to create a model using the customer churn data and use it to predict if a particular customer of a set of customers will discontinue the service.

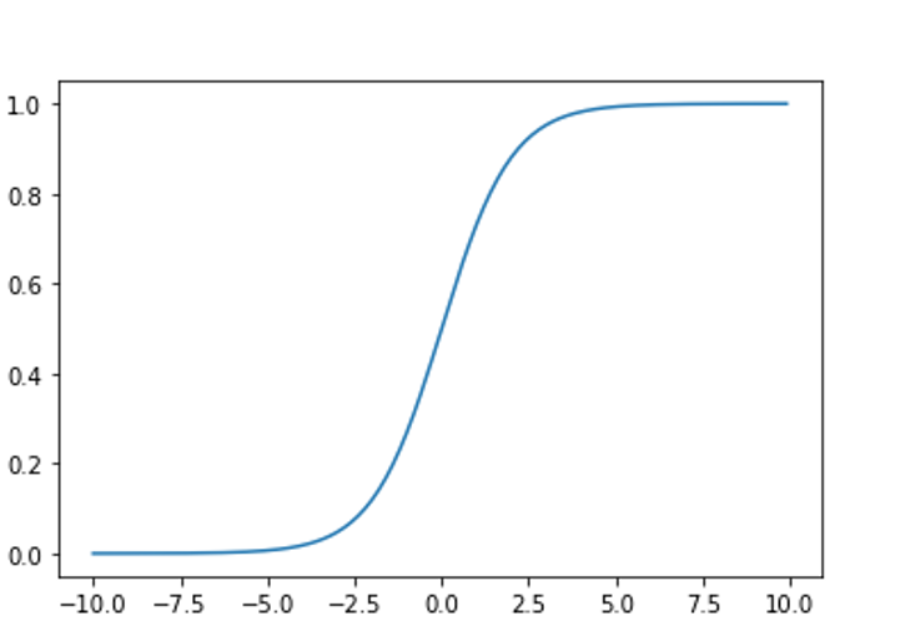
**Logistic Regression Equation and Sigmoid Function**

The Logistic Regression function:

Where,

**: intercept coefficients**

**= slope coeffients**  
X0 to Xn are the independent variables impacting the dependent variable.And P (Y =1 | X) is the probability of a positive outcome. Notice the exponent in the function. This is where linear regression plays in (β0+β1X1+β2X2+β3X3+β4X4+⋯+βnXn). The logistic regression function is a sigmoid function.



As can be seen in the graph, the function has output values between 0 and 1 with a transition between the levels. This characteristic of the function helps with the prediction of binary outcomes. Based on the value of variables the output can be either at level 1 or 0, which corresponds to the probability of the customer leaving the company or continuing with it.

**Assumptions of logistic regression:**

Following assumptions are made for Logistic regression:

1. Binary logistic regression requires the dependent variable to be binary and to follow a binomial distribution (e.g. will a customer discontinue service or not, Yes or No).
2. Observations should be independent of each other (e.g. data of one customer should not depend on data of another customer, or the same customer should not be used repeatedly in the data)
3. Multicollinearity among the independent variables should not exist.
4. The linearity of independent variables with respect to log odds of the dependent variable.
5. Large sample size.

Initial model:

For Feature selection we have used lasso regression. Lasso regression will automatically select those features that are useful, discarding the useless or redundant features. In first iteration.  The lasso regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets. This type of regression is used when the dataset when you want to automate variable elimination and feature selection.

SELECTED FEATURES:

Dependent\_count

Education\_Level

Income\_Category

Total\_Relationship\_Count

Months\_Inactive\_12\_mon

Contacts\_Count\_12\_mon

Total\_Revolving\_Bal

Total\_Trans\_Ct

Total\_Ct\_Chng\_Q4\_Q1

Gender\_M

Marital\_Status\_Married

This are the features selected among all independent features.

Similarly for 2nd iteration

Dependent\_count

Education\_Level

Total\_Relationship\_Count

Months\_Inactive\_12\_mon

Contacts\_Count\_12\_mon

Total\_Revolving\_Bal

Total\_Trans\_Amt

Total\_Trans\_Ct

Total\_Ct\_Chng\_Q4\_Q1

Gender\_M

Marital\_Status\_Married

And third iteration,

Customer\_Age

Dependent\_count

Education\_Level

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

Avg\_Utilization\_Ratio

Gender\_M

Marital\_Status\_Married

Marital\_Status\_Single

Marital\_Status\_Unknown

If we select all features and fit the model the accuracy, F score and AUC is as shown below

Logistic Regression Accuracy (all): 0.8560606060606061

Logistic Regression F-Score (all): 0.6649874055415618

Logistic Regression AUC (all): 0.9169008535784635

Similarly, accuracy for all three iterations:

For first iteration:

Logistic Regression Accuracy (1st lasso): 0.8203463203463204

Logistic Regression F-Score (1st lasso): 0.5921375921375922

Logistic Regression AUC (1st lasso): 0.8663286024546695

For 2nd iteration:

Logistic Regression Accuracy (2nd lasso): 0.8533549783549783

Logistic Regression F-Score (2nd lasso): 0.6608260325406757

Logistic Regression AUC (2nd lasso): 0.9189858073640083

For 3rd iteration:

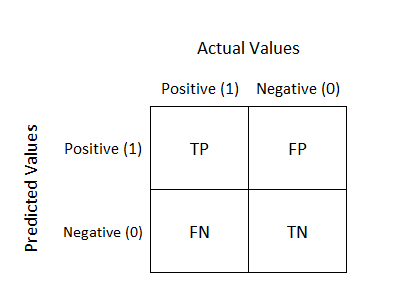
Logistic Regression Accuracy (3rd lasso): 0.8544372294372294

Logistic Regression F-Score (3rd lasso): 0.6633291614518146

Logistic Regression AUC (3rd lasso): 0.9168503459770696

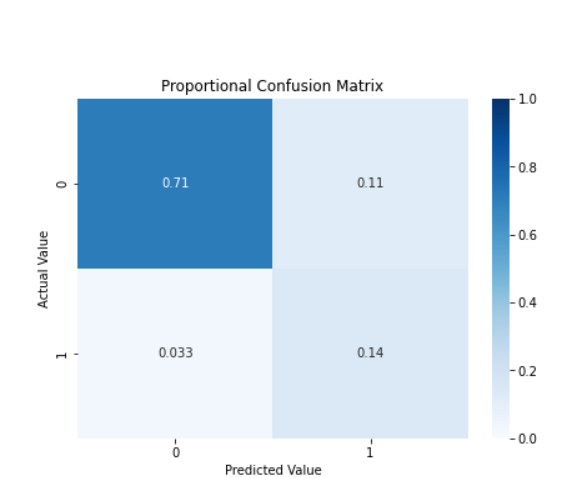
In 2nd iteration selected features have high accuracy, F score, AUC. Hence, we fit model with these features.

**Confusion matrix:**



Logistic Regression Confusion Matrix:

|  |  |
| --- | --- |
| 1313 | 210 |
| 61 | 264 |



From confusion matrix we conclude that

Correctly identified exiting customer:1313

Correctly identified churning customer:264

Total missclasifications:210 ,61

Logistic Regression Accuracy: 0.853

Logistic Regression F-Score: 0.661

Logistic Regression AUC: 0.919

**Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. If we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same

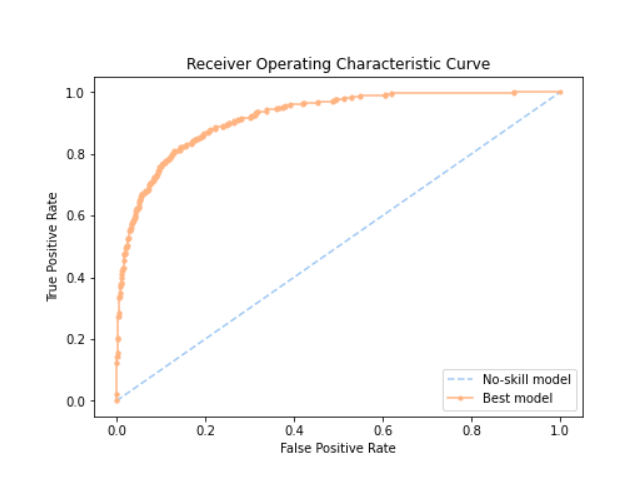
Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.853 which means our model is approx. 85.3% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

**F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall. In our case, F1 score is 0.661. F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

**AUC curve** -

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. Area under the curve is 91.9%



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | **0.96** | **0.86** | **0.91** |
| **1** | **0.56** | **0.81** | **0.66** |

We conclude that model gives good accuracy in test data also hence the model is good fit..

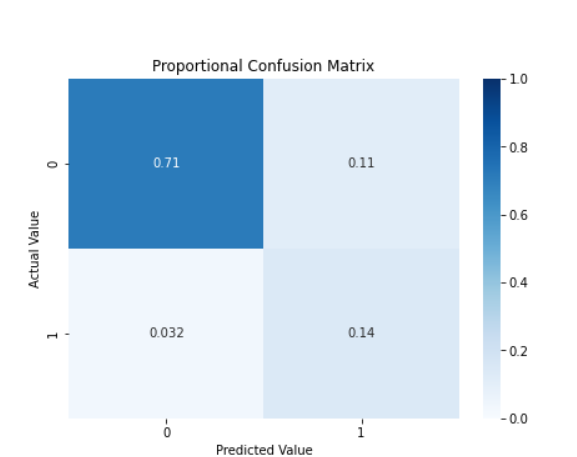
Now for better accuracy we use **Stepwise regression model.**

Stepwise model selection can have different criteria for including/excluding different regressors. If you use the p-values for the specific model parameters' Wald test or the resultant model R2, you will not do well, mostly because of internal validation bias. AIC or BIC are much better criteria for model selection

The selected features are ['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', 'Gender\_M']

Logistic Regression Confusion Matrix

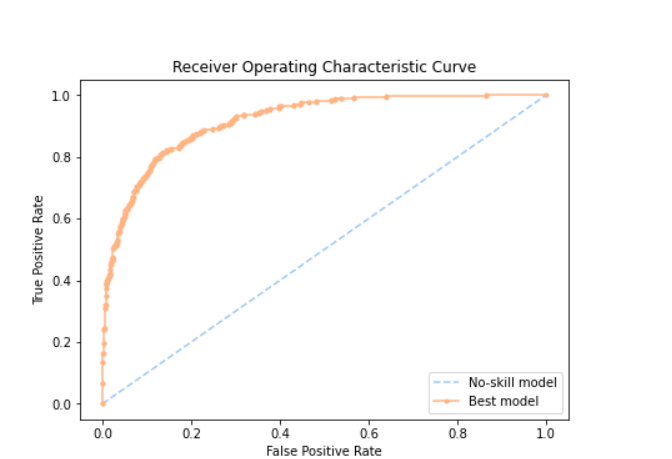
|  |  |
| --- | --- |
| 1315 | 208 |
| 60 | 265 |



Logistic Regression Accuracy: 0.855

Logistic Regression F-Score: 0.664

Logistic Regression AUC: 0.918



Training Accuracy: 0.8816674872804858

Testing Accuracy: 0.854978354978355

We conclude that when we use our accuracy is increased by 3%

**K-NEAREST NEIGHBOURHOOD**

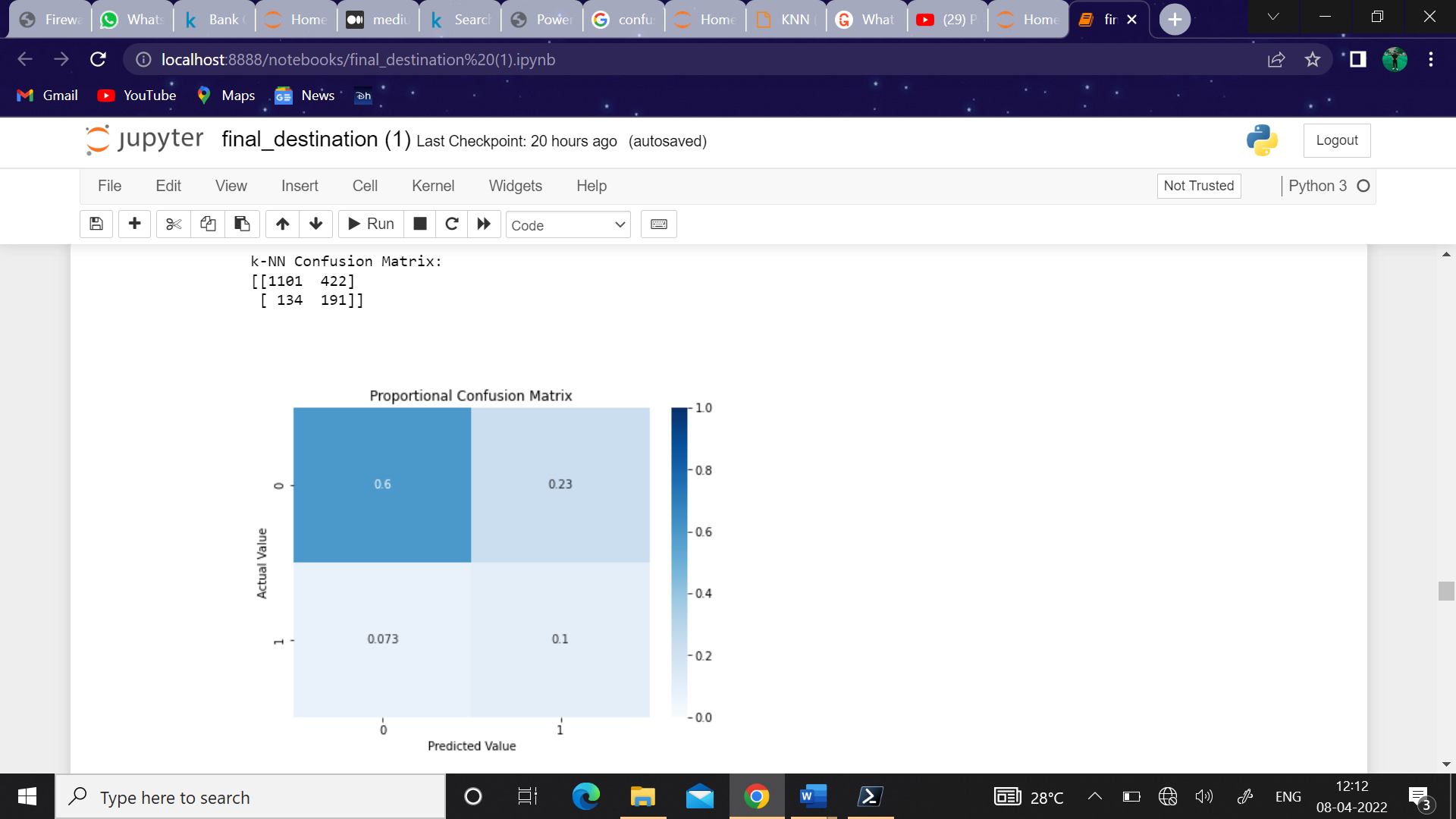
K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. When a new customer is presented, the algorithm looks through the database for customers who are most similar to the target customer. It then predicts if the customer would churn based on whether those similar customers churned or not.

The KNN machine learning method works on the distance between the data points and is based on the basic premise that similar data points are close to each other. If your neighbours are categorized in a certain way, you will likely fall under the same category too based on the virtue of you being near to them. Conversely, those who are not near but farther away are likely to fall under a different classification than yours. There can be exceptions (points belonging to a different category being near to points in a different category) and consequently, there can be errors in the predictions based on distance, especially around the decision boundaries. We can implement the KNN model and check how accurately it can predict the classification of points. Think of these data points as points in an n-dimensional space and the KNN algorithm predicting their categories based on the distance between the uncategorized data points and their k nearest neighbours, where k is a parameter.

**Model fitting:**

K-NN Confusion Matrix

|  |  |
| --- | --- |
| 1101 | 422 |
| 134 | 191 |



From confusion matrix we conclude that

Correctly identified exiting customer:1101

Correctly identified churning customer:191

k-NN Accuracy: 0.699

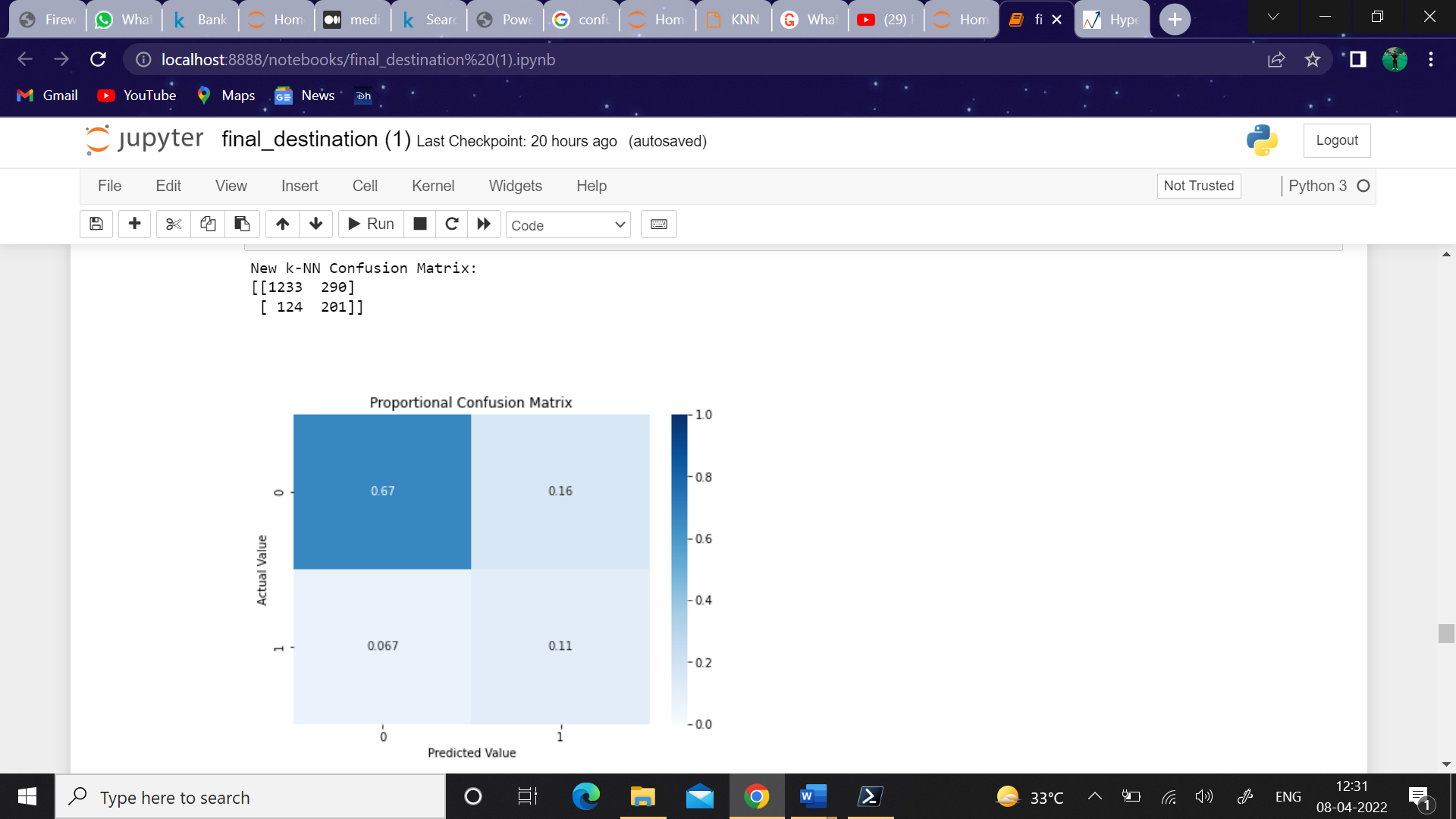
k-NN F-Score: 0.407

k-NN AUC: 0.685

As we can see the model gives 69% accuracy with 40% of F-score and AUC as 68%. Now this above confusion matrix is obtained by fitting the model without any hyperparameter tuning done. Hyperparameter tuning is done as model parameters are learned from data and hyper-parameters are tuned to get the best fit. Searching for the best hyper-parameter can be tedious, hence search algorithms like grid search and random search are used. Grid search picks out a grid of hyperparameter values and evaluates all of them. Guesswork is necessary to specify the min and max values for each hyperparameter. Random search randomly values a random sample of points on the grid. We will be going with Grid search for our hyperparameter tuning.

K-NN Confusion Matrix

|  |  |
| --- | --- |
| 1233 | 290 |
| 124 | 201 |



From confusion matrix we conclude that

Correctly identified exiting customer:1233

Correctly identified churning customer:201

New k-NN Accuracy: 0.776

New k-NN F-Score: 0.4926

New k-NN AUC: 0.7802

So, after hyperparameter tuning the accuracy increased up to 77% and F-score with 49% and AUC with 78%.

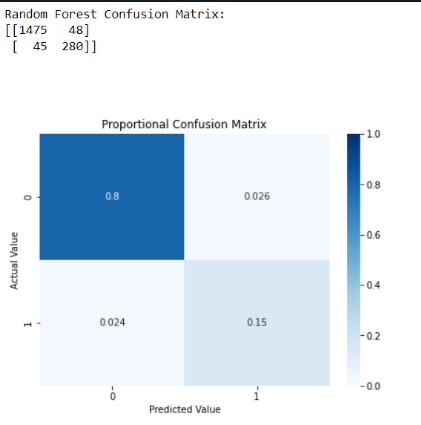
**RANDOM FOREST:**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction rather than relying on individual decision trees. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting

by aggregating the result. In Random Forests, the random subsets are selected in a procedure called ‘bagging’, in which each data point has an equal probability of being selected for each new random subset.

Initial Model:

|  |  |
| --- | --- |
| 1475 | 48 |
| 45 | 280 |



From confusion matrix we conclude that

Correctly identified exiting customer:1475

Correctly identified churning customer:280

Random Forest Accuracy: 0.95

Random Forest F-Score 0.858

Random Forest AUC: 0.982

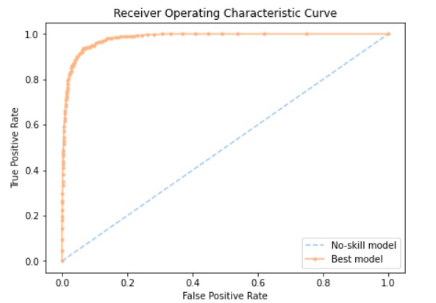
**Accuracy** - It is simply a ratio of correctly predicted observation to the total observations. If we have high accuracy then our model is best.

For our model, we have got 0.95 which means our model is approx. 95% accurate.

**FI Score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy. In our case, F1 score is 0.858

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

**AUC Curve** - The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. Area under the curve is 98.2%



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | **0.97** | **0.97** | **0.97** |
| **1** | **0.85** | **0.86** | **0.86** |

Hyperparameter tuning in Random Forest:

While model parameters are learned during training such as the slope and intercept in a linear regression hyperparameters must be set by the data scientist before training. In the case of a random forest, hyperparameters include the number of decision trees in the forest and the number of features considered by each tree when splitting a node. Hyperparameter tuning relies more on experimental results than theory, and thus the best method to determine the optimal settings is to try many different combinations evaluate the performance of each model. After hyperparameter tuning we get the confusion matrix as

|  |  |
| --- | --- |
| 1475 | 49 |
| 40 | 285 |



From new confusion matrix we conclude that

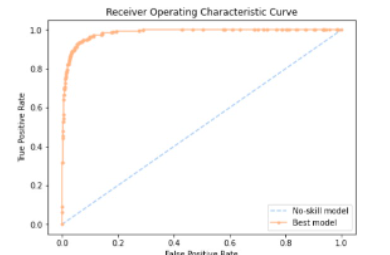
Correctly identified exiting customer:1474

Correctly identified churning customer:285

New Random Forest Accuracy: 0.952

New Random Forest F-Score: 0.865

New Random Forest AUC: 0.984



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | **0.97** | **0.97** | **0.97** |
| **1** | **0.85** | **0.88** | **0.86** |

So we can observe that after hyperparameter tuning the accuracy increased up to 95.2% and F-score with 86.5% and AUC with 98.4%.

**ODDS RATIO:**

In logistic regression the odds ratio represents the constant effect of a predictor X, on the likelihood that one outcome will occur. The key phrase here is constant effect. In regression models, we often want a measure of the unique effect of each X on Y. If we try to express the effect of X on the likelihood of a categorical Y having a specific value through probability, the effect is not constant.

Odds ratio

The odds ratio can be interpreted as the estimated increase in the probability of

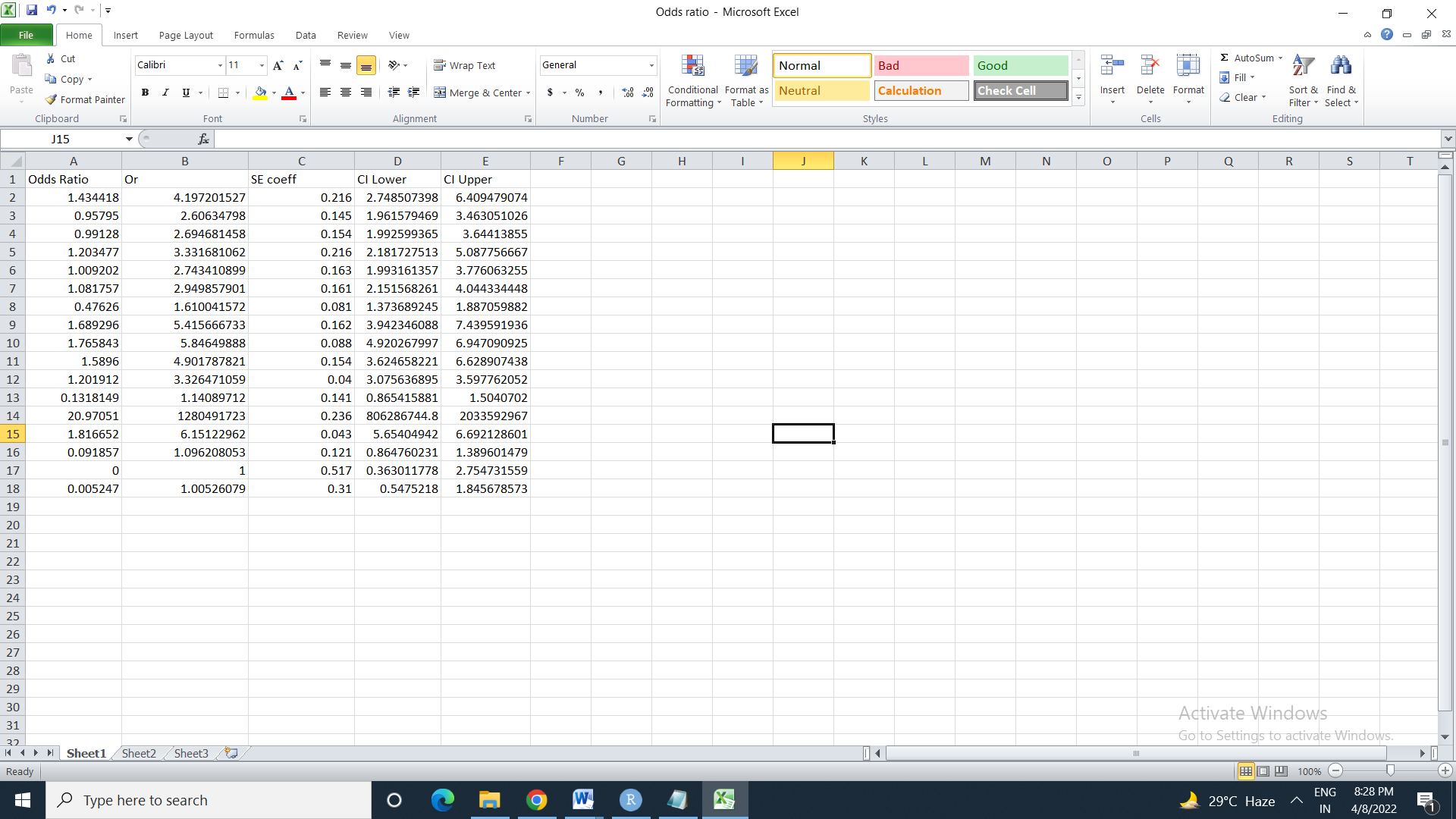
success associated with a one - unit change in the value of the predictor variable. In

general, the estimated increase in the odds ratio associated with a change of d units

in the predictor variable is exp(d.

The 100(1-α)% confidence interval for odds ratio is given by

OR = exp()



This is the excel file for confidence interval of odds ratio.

From the above output we can see that 4 variables i.e. Total\_Ct\_Chang\_Q4\_Q1,Total\_trans\_Ct,Total\_Trans\_Amt,Total\_Relationship\_Count are insignificant while others are significant.

**Interpretations:**

* Customers with Doctorate degree are 2 times more likely to leave than other customers.
* Customers with Post Graduate degree are 1.5 times more likely to leave than other customers.
* Married customers are more likely to leave 2 times than divorced customers.
* Married customers are more likely to leave 2 times than single customers.
* As 1 dependent increases, the chances that customer leaves increase by 1.5 times.
* If customer is inactive for 1 month the chances that customer will churn increases by 21 times.
* If the person is contacted by bank every months , the chances that customers leaves will increases by 2 times.

**CONCLUSION:**

1. We conclude that amongst the 3 model which we fitted i.e., Logistic regression, K-nearest neighbourhood algorithm and Random Forest algorithm, the accuracy for Random Forest is much better than Logistic and KNN. Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. The reason why KNN has least accuracy then the other two is because of the randomness in the dataset, If the data is a jumble of all different classes, then KNN will fail because it will try to find k nearest neighbours but all points are random.

**LIMITATION:**

1. The dataset didn’t consist of feedback variable due to which it was quite difficult to get the perfect reasons for a customer being churned.
2. There was absence of date & time period in the dataset due to which it was unable to get the reasons of churning based on trends.

**FUTURE SCOPE:**

1. This research and model building was done on a bank dataset, but it can also be carried out on various other sectors.
2. Based on customer sentiments too we can get the customer churn rate for a sector and suggest some strategies to retain them.

**BIBLIOGRAPHY:**

1)<https://medium.com/data-science-on-customer-churn-data/customer-churn-data-analysis-using-logistic-regression-3861e2d4d1f3#:~:text=Logistic%20regression%20is%20a%20binary,the%20generalized%20linear%20regression%20model.&text=It%20is%20possible%20to%20use,customers%20will%20discontinue%20the%20service>.

2) [customer-churn-data-a-project-based-on-logistic-regression.pdf](file:///C:\Users\user\Downloads\customer-churn-data-a-project-based-on-logistic-regression.pdf)

3) <https://towardsdatascience.com/predicting-customer-churn-using-logistic-regression-9543c60f6d47>.

