# Predict Churning customers - Credit Card customers

Dhiraj Bankar

**Bellevue University**

## Abstract

Finding the hidden information and create a machine learning model is the goal of this project. As part of my analysis I took the credit card of a bank, where the a manager at the bank is disturbed with more and more customers leaving their credit card services. They would really appreciate if one could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction. End of this project I will create a model where it can predict the customers who is going to get churned.

## Context

A manager at the bank is disturbed with more and more customers leaving their credit card services. I would like to predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction.

Table of Contents

## Background

A manager at the bank is disturbed with more and more customers leaving their credit card services. They would really appreciate if one could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction

I got this dataset from a website with the URL as https://leaps.analyttica.com/home. I have been using this for a while to get datasets and accordingly work on them to produce fruitful results. The site explains how to solve a particular business problem.

Now, this dataset consists of 10,000 customers mentioning their age, salary, marital_status, credit card limit, credit card category, etc. There are nearly 18 features.

We have only 16.07% of customers who have churned. Thus, it's a bit difficult to train our model to predict churning customers.

## Data Understanding

Categorical data are commonplace in many Data Science and Machine Learning problems but are usually more challenging to deal with than numerical data. In particular, many machine learning algorithms require that their input is numerical and therefore categorical features must be transformed into numerical features before we can use any of these algorithms.

One of the most common ways to make this transformation is to **one-hot encode** the categorical features, especially when there does not exist a natural ordering between the categories (e.g. a feature ‘City’ with names of cities such as ‘London’, ‘Lisbon’, ‘Berlin’, etc.). For each unique value of a feature (say, ‘London’) one column is created (say, ‘City_London’) where the value is 1 if for that instance the original feature takes that value and 0 otherwise.

## Categorical Features

- Attrition_Flag (1: Existing Customer, 0: Attrited Customer): The Customer leave or not
- Gender (1: Male, 0: Female)
- Education_Level (Graduate , High School, Unknown, Uneducated, College, Post-Graduate, Doctorate)
- Marital_Status (Married, Single, Unknown, Divorced)
- Income_Category (Less than 40K, 40K - 60K, 80K - 120K, 60K - 80K, Unknown, 120K +) in dollar
- Card_Category (Blue, Silver, Gold, Platinum)

### *Count plot for all Categorial Features*

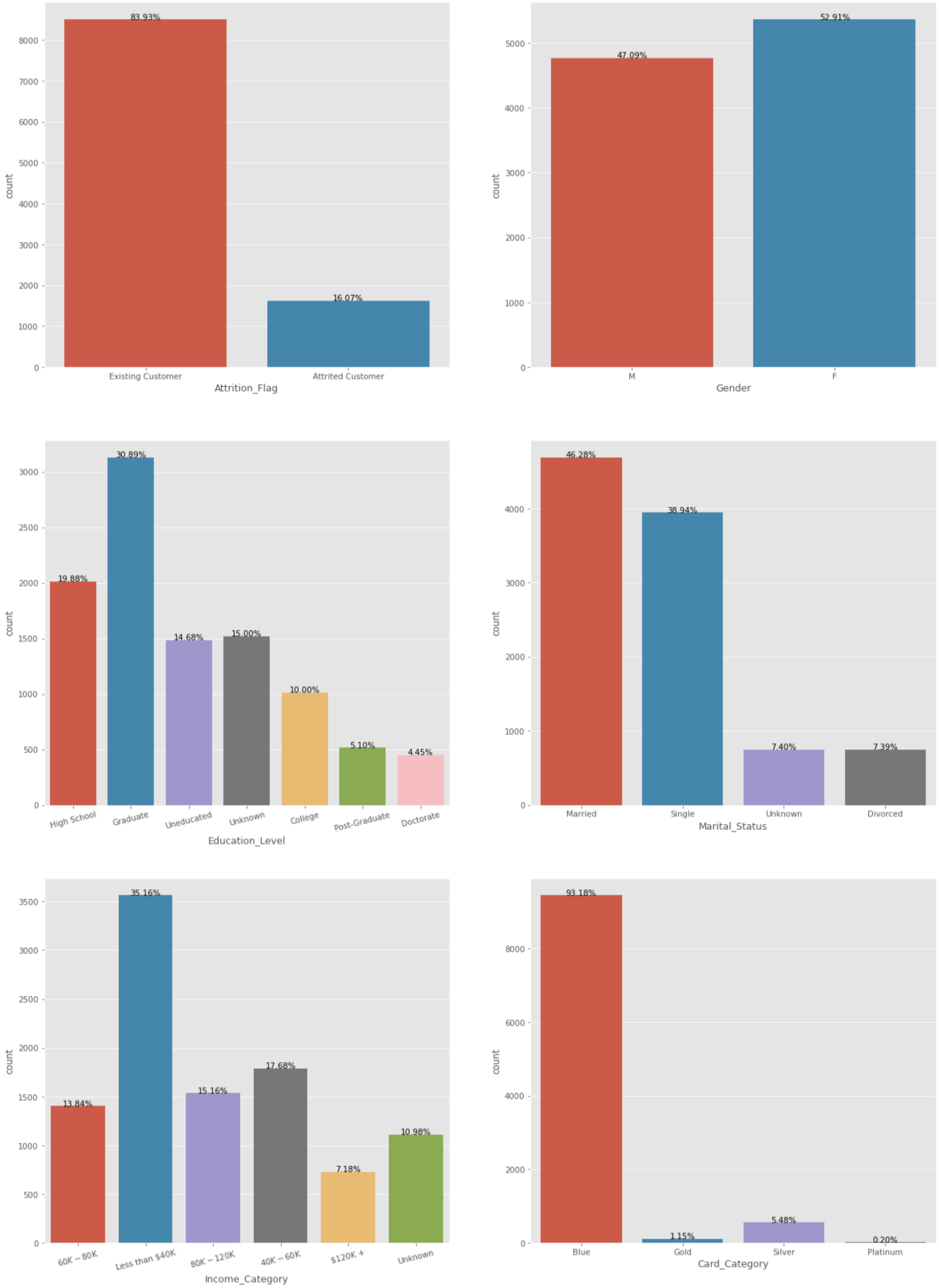


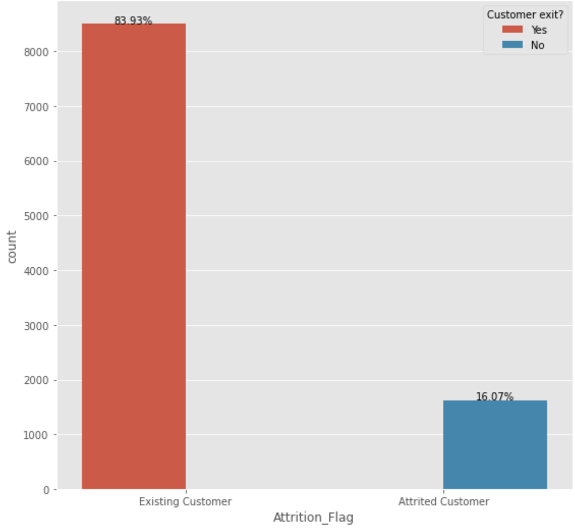
***Figure 1.*** *Count plot for all Categorial Features*

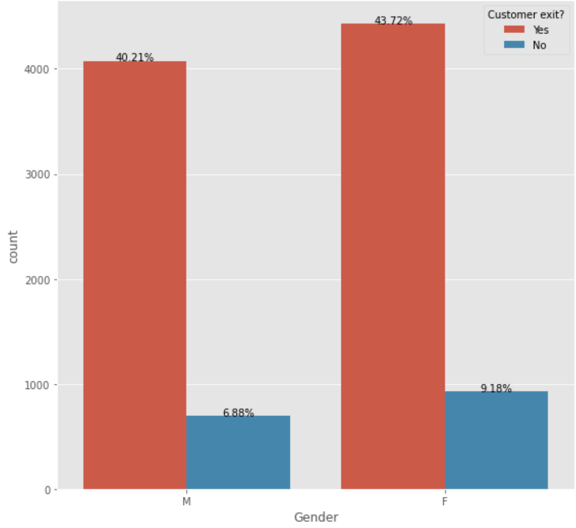
**Observations**

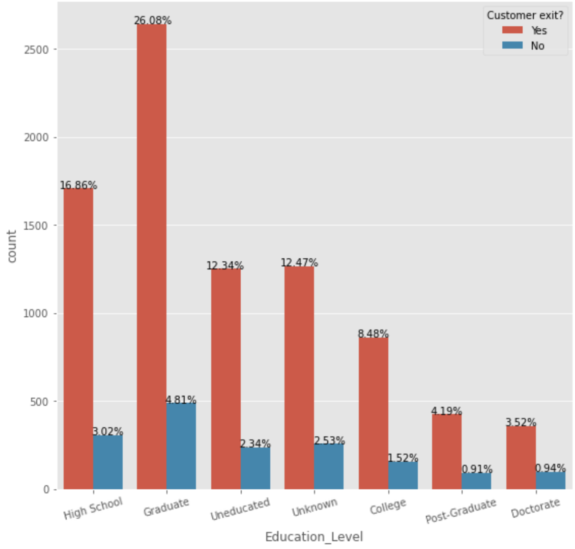
- We can see that the dataset is not equally distribute according to Attrition_Flag. We have samples which are mostly Existing.
- We can say that if education level is improved, using the credit card is decreasing.
- Generally people use blue card, it's must be correlated with income.

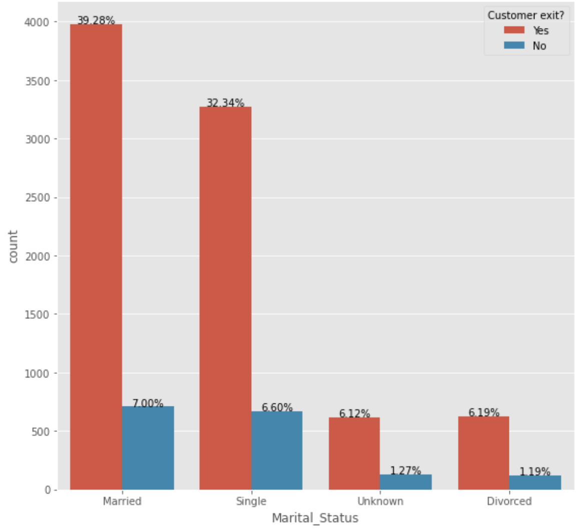
***Count plot for all Numerical Features***

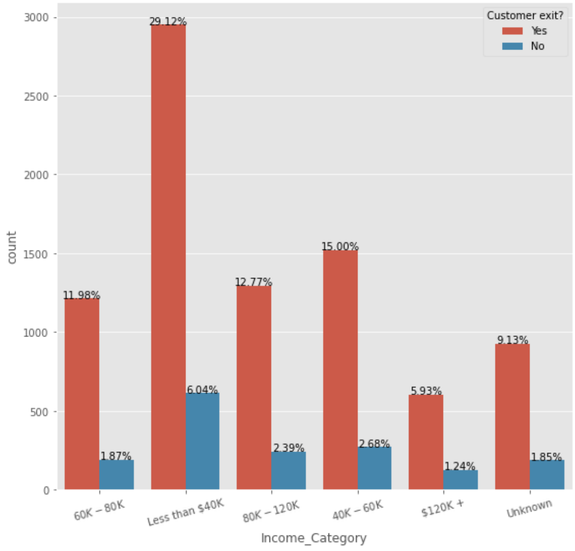
Features are usually numeric, but structural features such as strings and graphs are used in syntactic pattern recognition. ... The concept of "feature" is related to that of explanatory variable used in statistical techniques such as linear regression.


Customer exit?
🟥 Yes
🟦 No
count
8000
7000
6000
5000
4000
3000
2000
1000
0
83.93%
16.07%
Existing Customer
Attrited Customer
Attrition_Flag


Customer exit?
🟥 Yes
🟦 No
count
4000
3000
2000
1000
0
40.21%
6.88%
43.72%
9.18%
M
F
Gender


Customer exit?
🟥 Yes
🟦 No
count
2500
2000
1500
1000
500
0
16.86%
3.02%
26.08%
4.81%
12.34%
2.34%
12.47%
2.53%
8.48%
1.52%
4.19%
0.91%
3.52%
0.94%
High School
Graduate
Uneducated
Unknown
College
Post-Graduate
Doctorate
Education_Level


Customer exit?
🟥 Yes
🟦 No
count
4000
3500
3000
2500
2000
1500
1000
500
0
39.28%
7.00%
32.34%
6.60%
6.12%
1.27%
6.19%
1.19%
Married
Single
Unknown
Divorced
Marital_Status


Customer exit?
🟥 Yes
🟦 No
count
3000
2500
2000
1500
1000
500
0
11.98%
1.87%
29.12%
6.04%
12.77%
2.39%
15.00%
2.68%
5.93%
1.24%
9.13%
1.85%
60K − 80K
Less than $40K
80K − 120K
40K − 60K
$120K +
Unknown
Income_Category

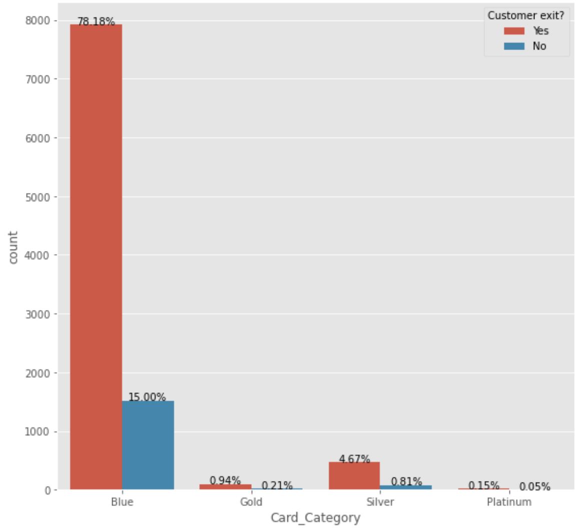

Customer exit?
🟥 Yes
🟦 No
count
8000
7000
6000
5000
4000
3000
2000
1000
0
78.18%
15.00%
0.94%
0.21%
4.67%
0.81%
0.15%
0.05%
Blue
Gold
Silver
Platinum
Card_Category

*Figure 2. Count plot for all Numerical Features*

**Numerical Features**

- Customer_Age: Customer's Age in Years
- Dependent_count: Number of dependents
- 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
- Total_Revolving_Bal: Total Revolving Balance 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

*Histogram for all numeric values*

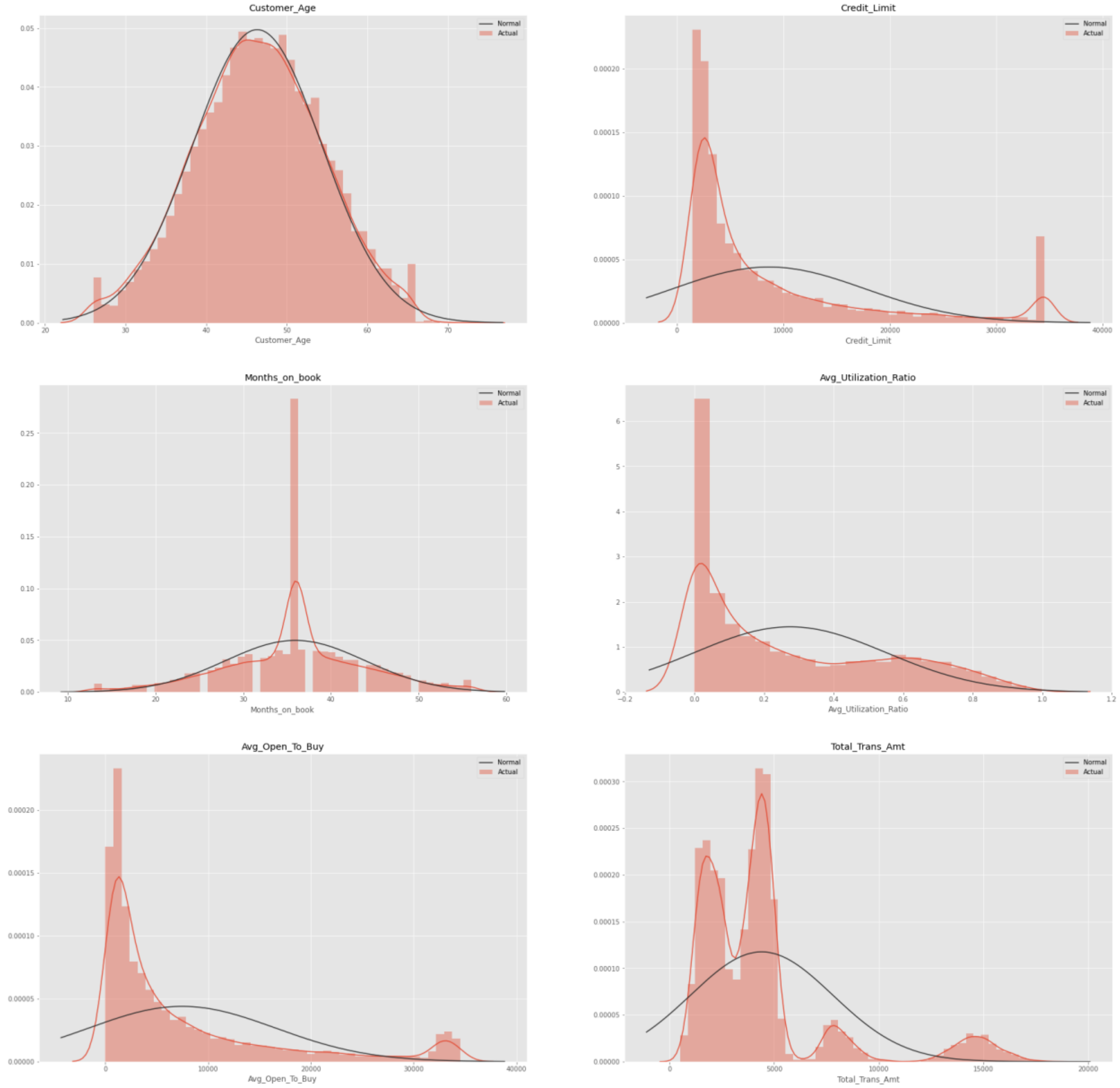
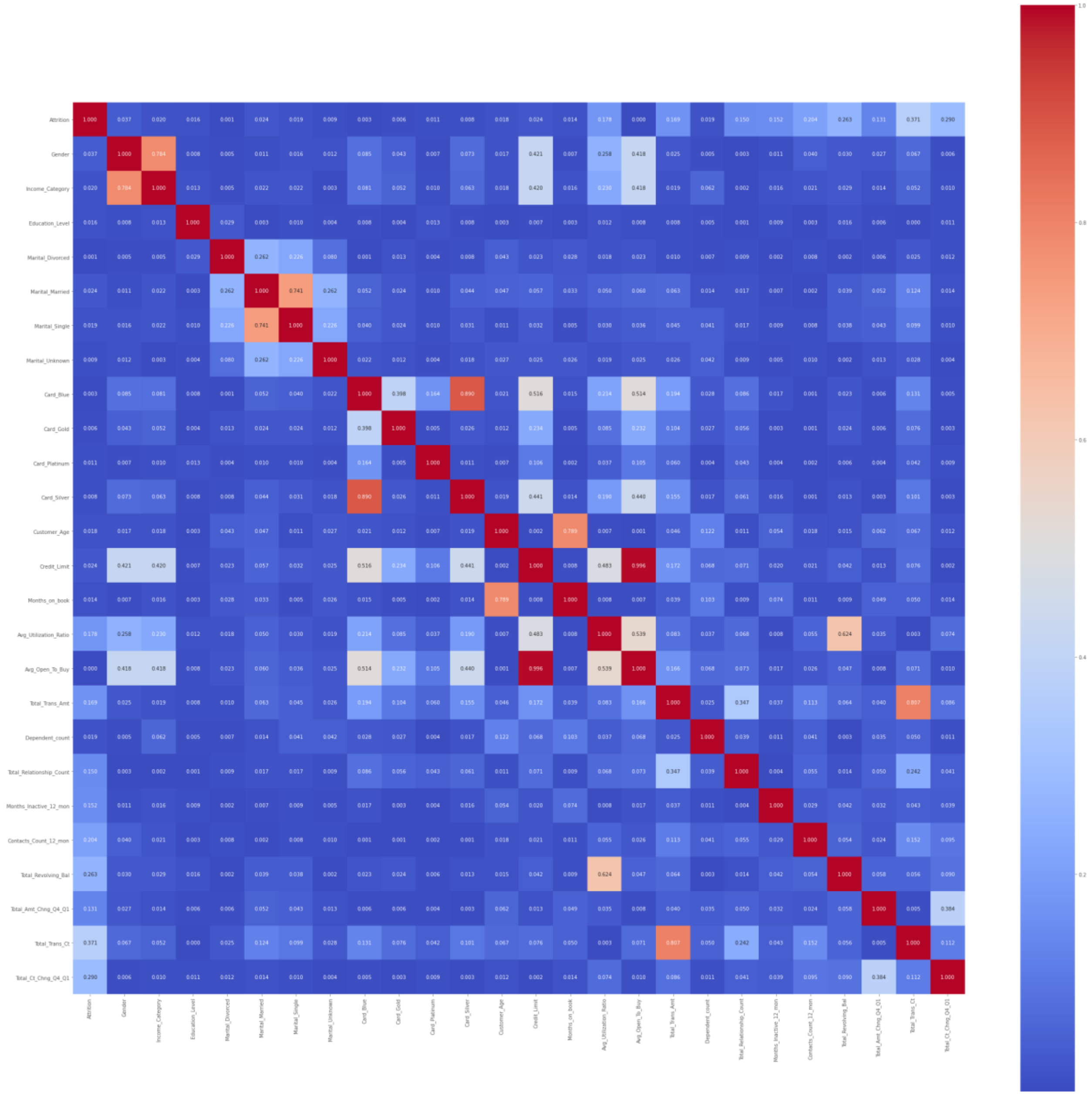


*Figure 3: Histogram for all numeric values*

## Defining the target variable

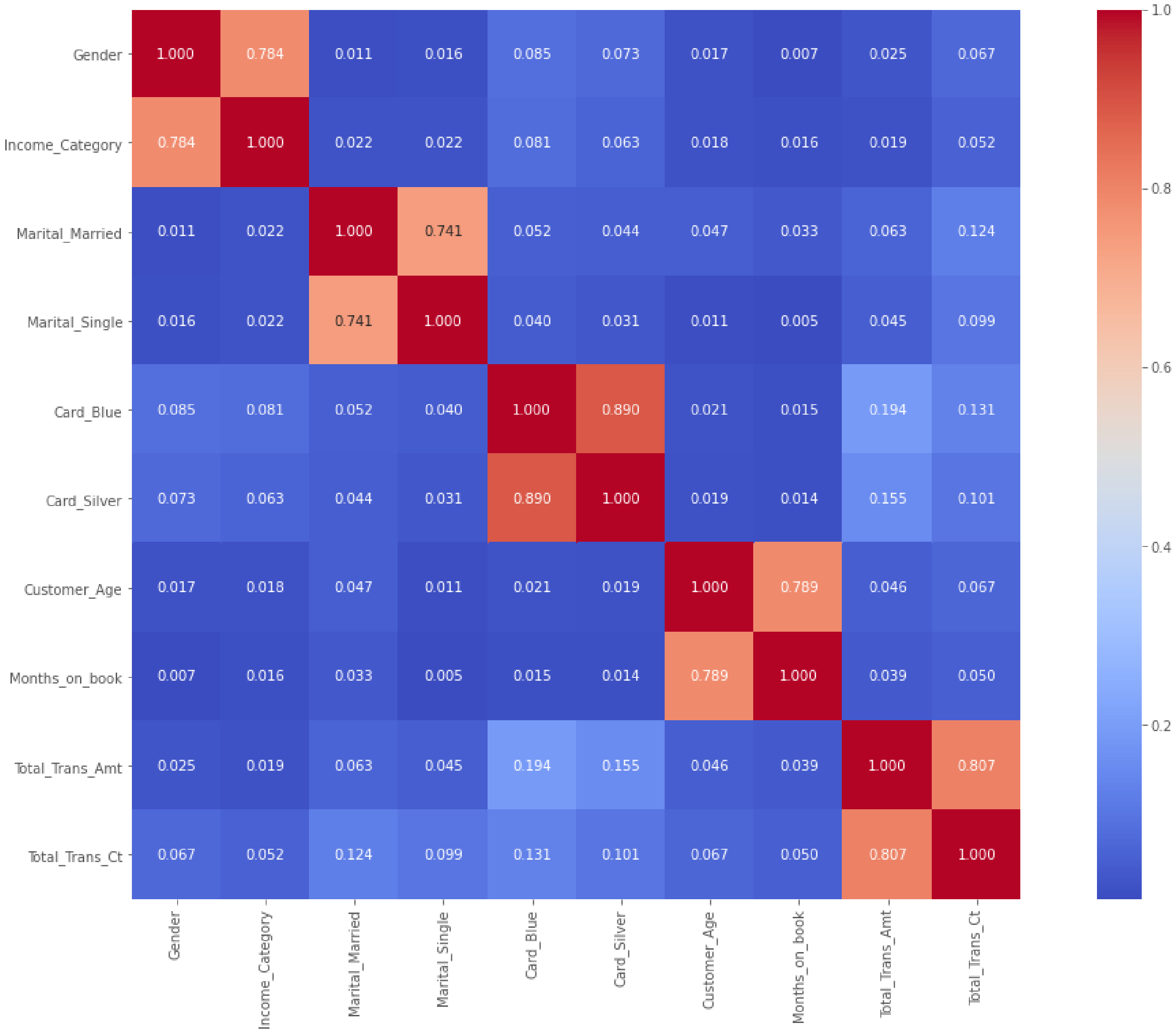
Before we start the Feature selection and ML algorithm. We should convert the works on numeric values. That's why we should transform Object, Category, etc. values to numeric values. I used correlation heat map method for feature selection.



From the above correlation heat map below are the list of attributes I am going to use for ML.

- Gender

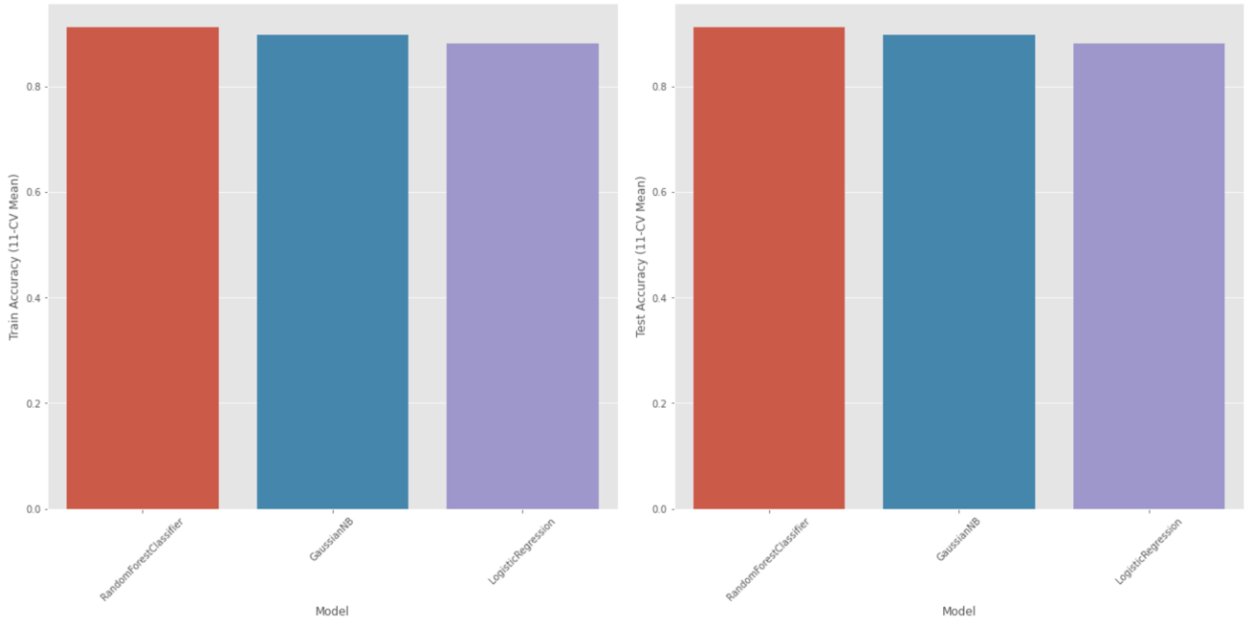
- Income_Category
- Marital_Married
- Marital_Single
- Card_Blue
- Card_Silver
- Customer_Age
- Months_on_book
- Total_Trans_Amt
- Total_Trans_Ct



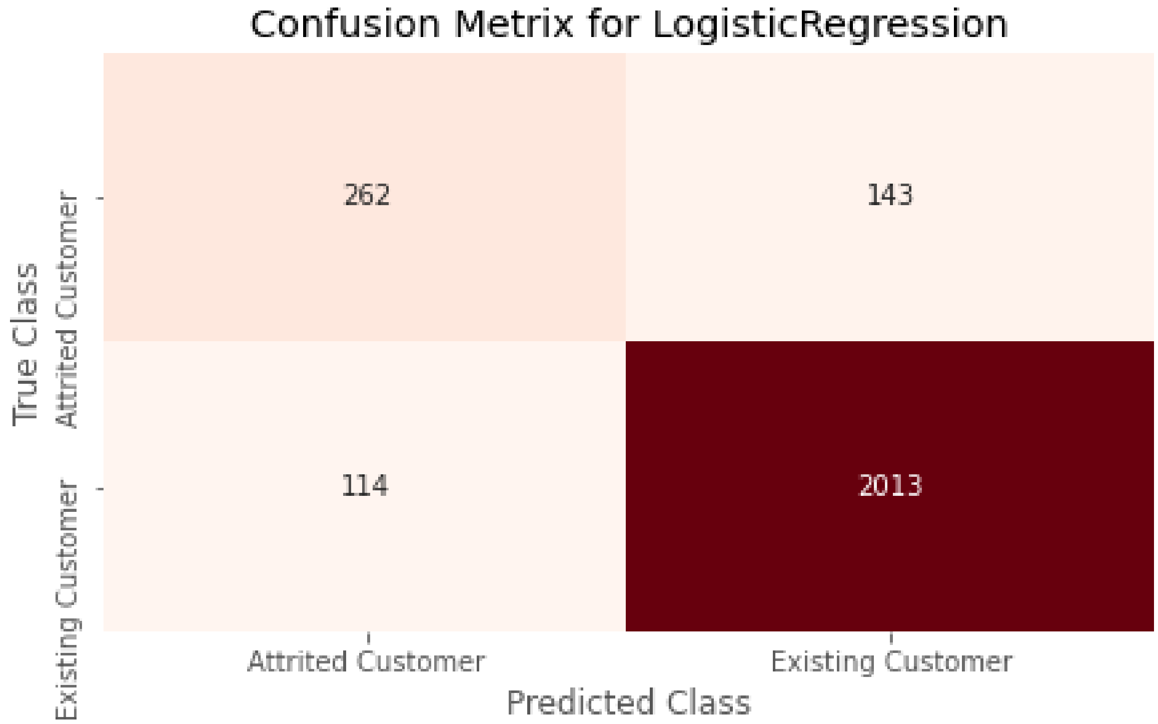
## Modeling

I performed model selection and model evaluation. Refer below results.

I selected LogisticRegression, RandomForestClassifier and GaussianNB models and compared the accuracy of each model for Train dataset and Test dataset and produces comparison results. From the bellow graphs the RenadomForestClassifer has highest accuracy and Logistic Regression has lowest accuracy. In next ween I will represent the feature selection by each module.



**V**isualizing a classifier  performance for LogisticRegression, RandomForestClassifier and GaussianNB.



## Results

The LogisticRegression model predicted the 114 existing customer as attrited customers and 143 attrited customers as existing customer

The RandomForestClassifier model predicted the 24 existing customer as attrited customers and 215 attrited customers as existing customer

The GaussianNB model predicted the 93 existing customer as attrited customers and 224 attrited customers as existing customer

As per accuracy the RandomForestClassifier should perform best but LogisticRegression model is performing better that other models. It could be due to feture selections which I did.

## Conclusion/Discussion

As part of this analysis I used multiple models like LogisticRegression, RandomForestClassifier and GaussianNB. Overall Random Forest Classifier seems more accurate. It is also more accurate in Existing customer findings and not recommended to find the attrited customers. However, Logistic Regression has more accuracy in terms of True or Recommended Scenarios. Definitely, I don't want to say my model is 100% accurate. I feel as a beginner model is really performing well. Definitely there is scope of improvement and best practices to get better results. I feel there is scope of improvement of feature selection. This could make model more tuned. Definitely we can use this model to reduce the attrited customers and save the credit card company problem.

## References

Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda

Machine Learning with Python Cookbook, Chris Albon

Dataset:

https://www.kaggle.com/sakshigoyal7/credit-card-customers

## Appendix A

Categorical data understanding

https://medium.com/hugo-ferreiras-blog/dealing-with-categorical-features-in-machine-learning-1bb70f07262d

https://en.wikipedia.org/wiki/Feature_(machine_learning)

LogisticRegression:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

RandomForestClassifier:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

GaussianNB:

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

## Questions

1) In the dataset the column credit limit what does it indicates whether it is monthly or average credit limit?

   Answer: Monthly credit limit.

2) What does it mean when the data have an Avg_Utilizatio_Ratio of 0?
3) Is it 0 because they can no longer use the credit card since they are already churned or
4) is it 0 because they havent used it for a few months?

Answer: It is the ratio of *(credit card spent + money withdrawal)/(Total available limit for credit card spends + Total money withdrawal limit)*

5) How would you treat 'Unknown' label?

   Answer: replace them with nulls first, since they don't provide any information. Thereafter imputed them with values based on K nearest neighbor to remove all nulls. However, all three have minimal effect on the dependent variable, so they can be left alone as well

6) Could you please elaborate this column? What do "Total_Trans_Amt" and "Total_Revolving_Bal " mean?

7) What is the difference?

   Answer: Total_Trans_amt is the sum of transactions one has done in the last 12 months. This basically tells us the total usage of a credit card by the user.

   Revolving balance is the unpaid amount that carries off on your next credit card's cycle. Total_revolving_bal would be its sum. This tells us the amount that they don't pay on time.

8) However, I'm not sure what the attributes "Total_Relationship_Count", "Contacts_count", "Total_Amt_Chng_Q4_Q1", "Total_Ct_Chng_Q4_Q1" represent for the costumers.

9) Furthermore, how was "Avg_Utilization_Ratio" calculated? Do you think it should be included in the training of a machine learning model?

   Answer:

   Total_Relationship_Count: total number of products held by the customers (cards, accounts, etc.)

Contacts_count_12_mont: I guess it holds the number of times the bank contacted the customer and/or viceversa. There doesn't seem to have any relationship with other fields (e.g. contact clients that left the bank, customers with revolving balance, etc). I guess it has to do with ad campaigns.

Total_Amt_Chng_Q4_Q1: represents how much the customer increased their expenditure when comparing the 4th quarter agains the 1st.

Total_Ct_Chng_Q4_Q1: similar to the previous but in number of transactions.

10) Is this dataset authenticated or somebody mad it for visualization purpose?

Answer: I don’t know the authenticity of the dataset. But personally I feel the most of the dataset available on Kaggle are good for learning purpose.