**Prediction of Credit Card Attrition**

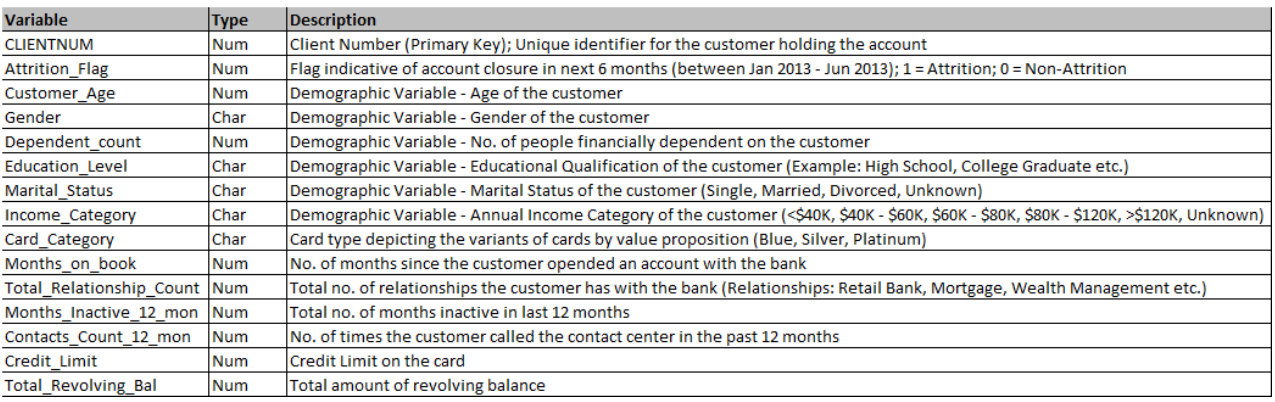
*(Vigneshwar K - 270403)*

**AIM:**

To build a model that would be able to determine the probability of attrition of an account holder within the next six months using Logistic Regression.

**THEORY:**

We are given with the Banking\_CreditAttrition dataset with more than 10,000 data containing the following information:



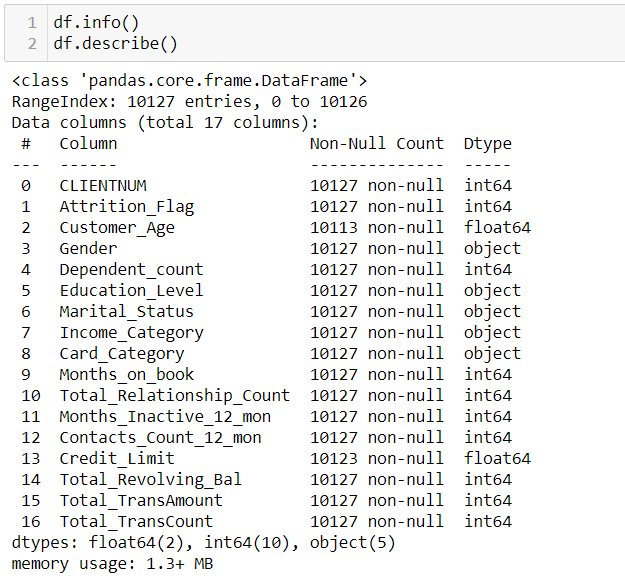
**IMPLEMENTATION:**

Steps followed:

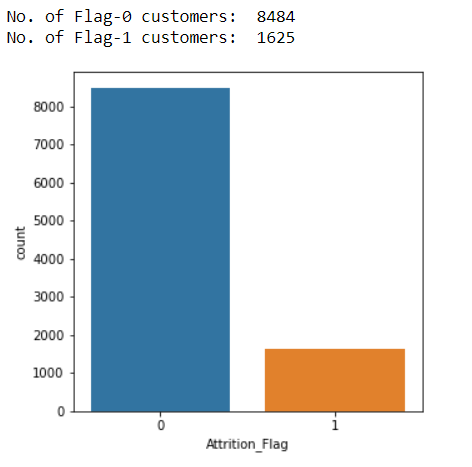
* Import required libraries and read the csv files.
* Analyse the data in the datasets.
* Pre-process the data in the datasets.
  + Remove unwanted columns and data
  + Check and remove duplicates.
  + Check Missing values
  + Check outliers
  + Treat outliers
  + Checking Multi-collinearity.
* Fit the data into suitable ML model.
* Predict the results of test dataset.

**ANALYSE DATA:**

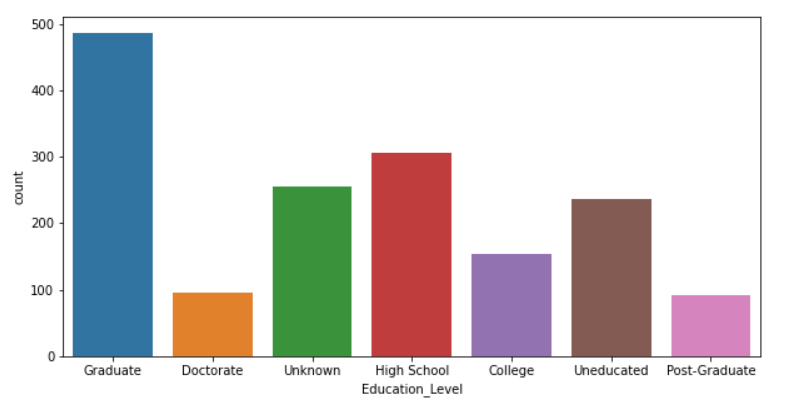
Two new features are engineering using the past six months data of Transaction amount and transaction count, to help reduce the data spread and simply data analysis.



**Distribution of customers:**

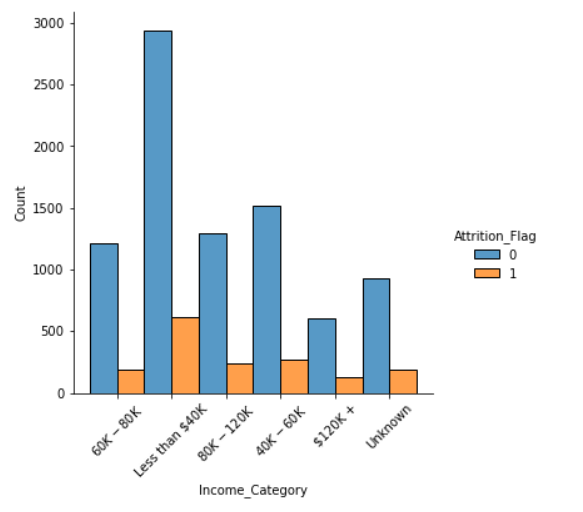


**Distribution of Attrition based on education:**

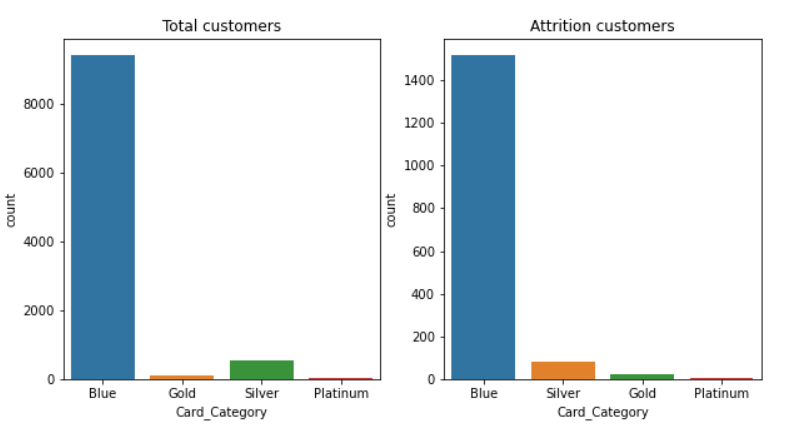


We can not infer a lot form this data as the total distribution of all data is similar to this, so we cannot take education as predictor variable.

**Distribution of Attrition based on Income:**



**Distribution of attrition based on card category:**



This information seems to be less helpful as the number of attritions corresponds to the total number of card holder in that category. So, it is highly unlikely that we would get useful information to predict the outcome with this data.

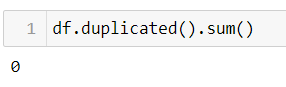
**DATA PRE-PROCESSING:**

1)Remove unwanted data and columns:

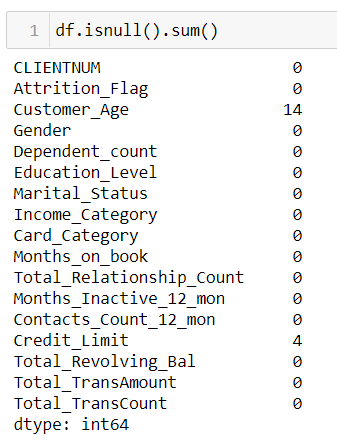
The columns like Card\_category, Income\_category, Eduction seems insignificant while predicting the outcome, so such columns and data must be removed to simply analysis.

2)Check and remove duplicates:

The duplicates in the customerID columns and other columns need to be removed so that there is no redundant data present which would cause bias and result in wrong predictions.



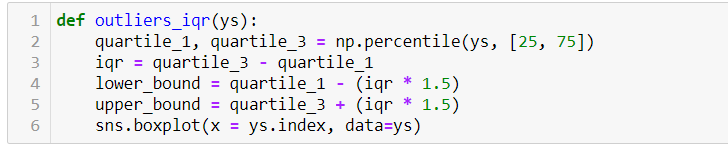
3)Check missing values:

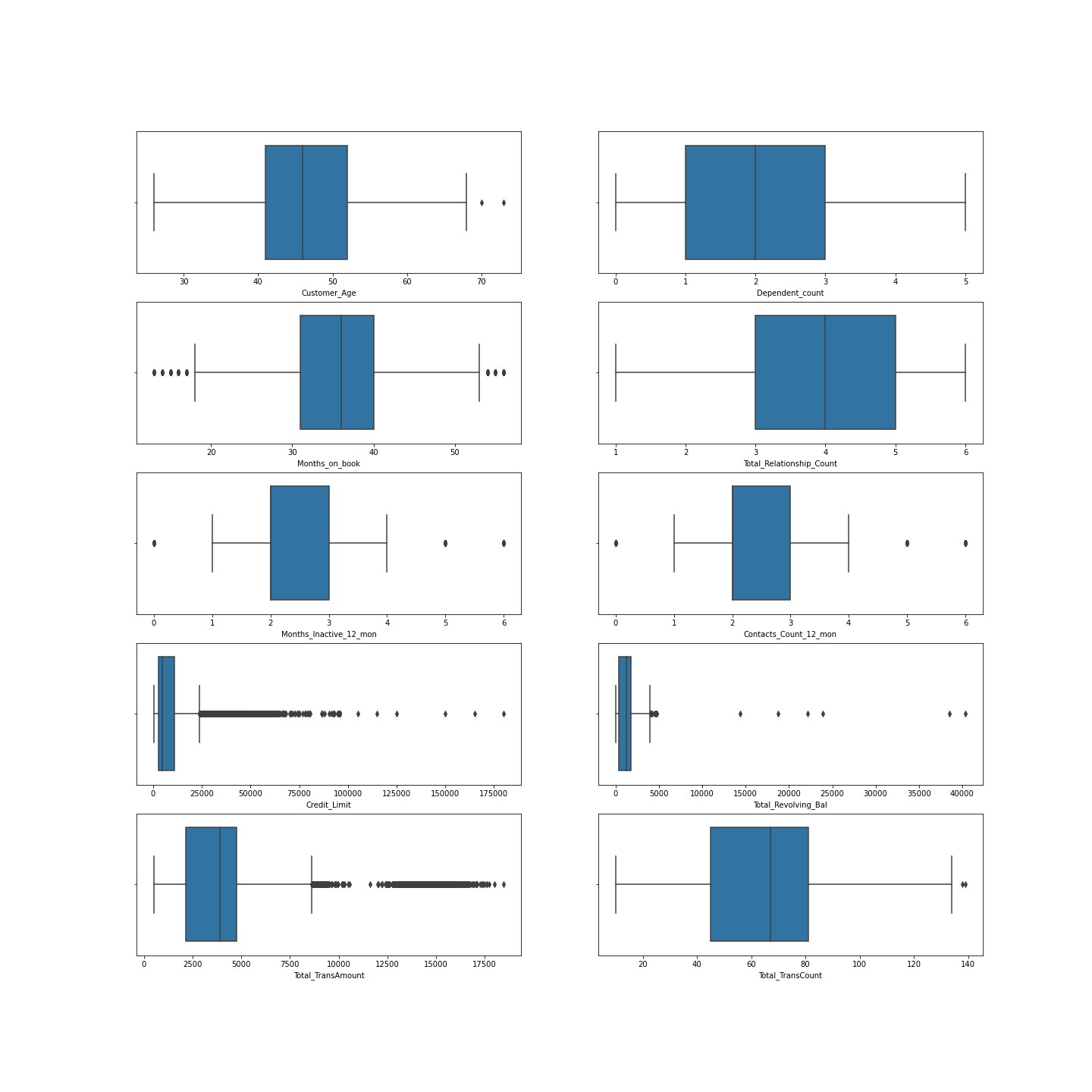


As there are very few missing data, it is ideal to remove those data.

4)Check outliers:

Presence of outliers in data will reduce the efficiency of the model and may lead to wrong predictions. So, outliers need to be taken care of. We can use Box plot to visualize the outliers present in the data.





5) Treat outliers:

Here, we chose to treat the outliers by using Floor-Cap method. In which, the values greater than a cap(max) value or lesser than a floor(min) value are replaced by user-specified cap and floor values. Here, the cap and floor are assigned as follows:

Cap: Q3+ (1.5xIQR); Floor: Q1 - (1.5 x IQR)

Cap – 95th percentile

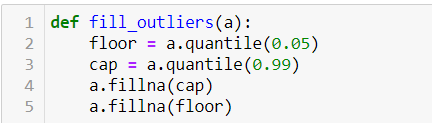
Floor – 5th percentile

Where,

Q3 – 3rd quartile

Q1 – 1st quartile

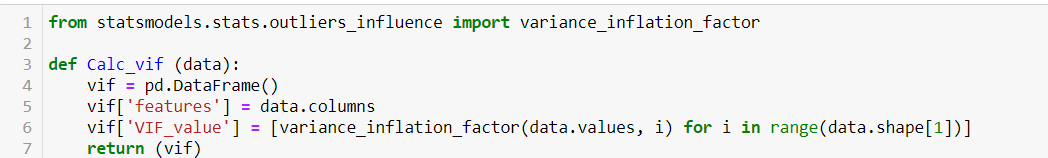
IQR – Inter quartile region.



6) Checking Multi-collinearity:

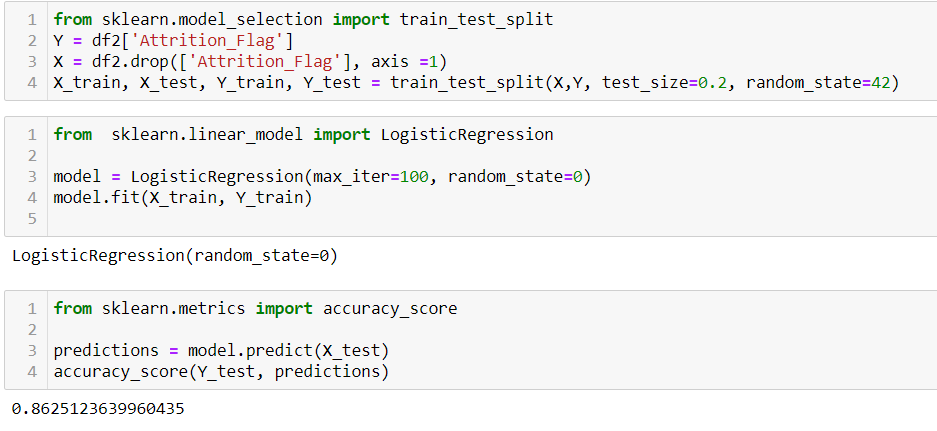
Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

One way to measure multicollinearity is the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated.



**PREDICTION OUTCOMES:**

Once the data is processed, it is time to fit the data into our model. As it is classification problem, we need to use classification algorithms.



This model gives an accuracy of 86%