**CREDIT RISK ANALYTICS**

## Post Graduate Program in Data Science & Engineering

# Location: Chennai Batch: PGPDSE\_MAR’22

## Submitted by:

**Rahul Thukaram**

**Priyadarshini B**

**Lakshmi Narayana S R**

**Sai Dileep M R**

**Vangara Harish**

**Mentored by**

**Mr. Jatinder Bedi**

**INTRODUCTION:**

Consumer credit is the personal debt taken on to purchase goods and services. Although any type of personal loan could be labeled consumer credit, the term is more often used to describe unsecured debt that is taken on to buy everyday goods and services. However, consumer debt can also include collateralized consumer loans like mortgage and car loans.

Consumer Credit is one of the main sources of income for a bank. But that does not mean there is no inherent risk in the consumer credit business, as it is dependent on various factors apart from the consumer’s credit history, like the current Economic climate, Government debt and various other trickle down effects that may put pressure on the clients ability to pay back their debts. Our goal is to reduce this outside risk by creating a model that will only allow us to provide a line of credit to the consumers who are less likely to default on their obligations in tougher times than to provide loans to people who are more likely to default under tougher circumstances.

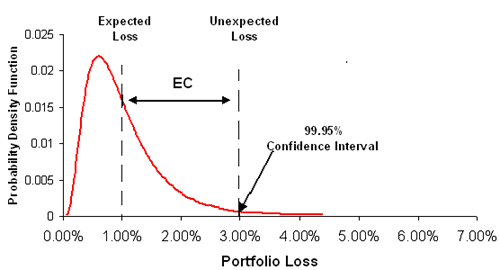
The current solution to this problem is to build a model which predicts the consumer’s creditworthiness based on their past banking history, thus classifying them into ‘GOOD/BAD’ investments. But this can be optimized in a way such that, we can make use of the past data to build models that can help classify new bank clients into good/bad investments based on the trends, characteristics and similarities from the previous client data, thus helping us predict the creditworthiness of the new client with no banking history.

**DATASET INFORMATION:**

This dataset is from a midsized Brazilian bank which contains the banking details of its clients who are applying for a loan. The dataset contains of 50,000 client records and 54 information variables that explains the attributes of each client. The target variable is a categorical variable which consists of two classes 1/0, indicating that the client is either GOOD/BAD investments based on their creditworthiness determined by their characteristics and trends from their history of transactions.

**PROBLEM STATEMENT:**

Credit risk or credit default risk is a type of risk faced by the lenders. Credit risk arises because a debtor can always default on their debt payments, thus causing the lending institutions to write off the loan or to bear the loss on its balance sheet. Financial Institutions do credit risk analytics to prevent this to a great extent.The majority of a Bank’s income comes in the form of interest payments and fees from its loans and credit lending programs. The objective is to reduce the possibility of credit risk so that the institutions can make money on their loans and credit programs which will help them pay their interest payments on the bank deposit liabilities and to become a more profitable business.



***Fig-1: Explanation of the portfolio risk caused by mass credit defaults***

One of the approaches to this type of problem is to classify the bank clients who are coming in for a loan into good or bad investments based on their creditworthiness. This means taking into account various personal, banking and credit details and using them to form a model to predict if the future customer is creditworthy or not. Credit risk analysis is assessing the possibility of the borrower’s repayment failure and the loss caused to the financial institution when the borrower does not for any reason repay the loan obligations. The cash flow of the institution is impacted when the interest accrued and principal amounts are not paid. Though, there is a grey area in guessing who and when will default on borrowings, with the help of credit analysis we can help mitigate the severity of complete loss of the borrowings and aid in the principal recovery process.

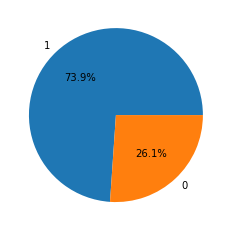
**VARIABLE CATEGORIZATION WITH DESCRIPTION:**

The dataset used in the project consists of 54 variables. Out of these 54 variables, we have our target variable which is a categorical variable, Binary class classified to be precise. There are 10 numeric variables and 44 categorical variables which also includes the encoded categories.

|  |  |  |
| --- | --- | --- |
| **Var\_Id** | **Var\_Title** | **Var\_Description** |
| 1 | ID\_CLIENT | Sequential number for the applicant (to be used as a key) |
| 2 | CLERK\_TYPE | Not informed |
| 3 | PAYMENT\_DAY | Day of the month for bill payment, chosen by the applicant |
| 4 | APPLICATION\_SUBMISSION\_TYPE | Indicates if the application was submitted via the internet or in person/posted |
| 5 | QUANT\_ADDITIONAL\_CARDS | Quantity of additional cards asked for in the same application form |
| 6 | POSTAL\_ADDRESS\_TYPE | Indicates if the address for posting is the home address or other. Encoding not informed. |
| 7 | SEX |  |
| 8 | MARITAL\_STATUS | Encoding not informed |
| 9 | QUANT\_DEPENDANTS |  |
| 10 | EDUCATION\_LEVEL | Educational level in gradual order not informed |
| 11 | STATE\_OF\_BIRTH |  |
| 12 | CITY\_OF\_BIRTH |  |
| 13 | NATIONALITY | Country of birth. Encoding not informed but Brazil is likely to be equal 1. |
| 14 | RESIDENCIAL\_STATE | State of residence |
| 15 | RESIDENCIAL\_CITY | City of residence |
| 16 | RESIDENCIAL\_BOROUGH | Borough of residence |
| 17 | FLAG\_RESIDENCIAL\_PHONE | Indicates if the applicant possesses a home phone |
| 18 | RESIDENCIAL\_PHONE\_AREA\_CODE | Three-digit pseudo-code |
| 19 | RESIDENCE\_TYPE | Encoding not informed. In general, there are the types: owned, mortgage, rented, parents, family etc. |
| 20 | MONTHS\_IN\_RESIDENCE | Time in the current residence in months |
| 21 | FLAG\_MOBILE\_PHONE | Indicates if the applicant possesses a mobile phone |
| 22 | FLAG\_EMAIL | Indicates if the applicant possesses an e-mail address |
| 23 | PERSONAL\_MONTHLY\_INCOME | Applicant's personal regular monthly income in Brazilian currency (R$) |
| 24 | OTHER\_INCOMES | Applicant's other incomes monthly averaged in Brazilian currency (R$) |
| 25 | FLAG\_VISA | Flag indicating if the applicant is a VISA credit card holder |
| 26 | FLAG\_MASTERCARD | Flag indicating if the applicant is a MASTERCARD credit card holder |
| 27 | FLAG\_DINERS | Flag indicating if the applicant is a SINERS credit card holder |
| 28 | FLAG\_AMERICAN\_EXPRESS | Flag indicating if the applicant is an AMERICAN EXPRESS credit card holder |
| 29 | FLAG\_OTHER\_CARDS | Despite being label "FLAG", this field presents three values not explained |
| 30 | QUANT\_BANKING\_ACCOUNTS |  |
| 31 | QUANT\_SPECIAL\_BANKING\_ACCOUNTS |  |
| 32 | PERSONAL\_ASSETS\_VALUE | Total value of the personal possessions such as houses, cars etc. in Brazilian currency (R$). |
| 33 | QUANT\_CARS | Quantity of cars the applicant possesses |
| 34 | COMPANY | If the applicant has supplied the name of the company where he/she formally works |
| 35 | PROFESSIONAL\_STATE | State where the applicant works |
| 36 | PROFESSIONAL\_CITY | City where the applicant works |
| 37 | PROFESSIONAL\_BOROUGH | Borough where the applicant works |
| 38 | FLAG\_PROFESSIONAL\_PHONE | Indicates if the professional phone number was supplied |
| 39 | PROFESSIONAL\_PHONE\_AREA\_CODE | Three-digit pseudo-code |
| 40 | MONTHS\_IN\_THE\_JOB | Time in the current job in months |
| 41 | PROFESSION\_CODE | Applicant's profession code. Encoding not informed |
| 42 | OCCUPATION\_TYPE | Encoding not informed |
| 43 | MATE\_PROFESSION\_CODE | Mate's profession code. Encoding not informed |
| 44 | EDUCATION\_LEVEL | Mate's educational level in gradual order not informed |
| 45 | FLAG\_HOME\_ADDRESS\_DOCUMENT | Flag indicating documental confirmation of home address |
| 46 | FLAG\_RG | Flag indicating documental confirmation of citizen card number |
| 47 | FLAG\_CPF | Flag indicating documental confirmation of tax payer status |
| 48 | FLAG\_INCOME\_PROOF | Flag indicating documental confirmation of income |
| 49 | PRODUCT | Type of credit product applied. Encoding not informed |
| 50 | FLAG\_ACSP\_RECORD | Flag indicating if the applicant has any previous credit delinquency |
| 51 | AGE | Applicant's age at the moment of submission |
| 52 | RESIDENCIAL\_ZIP\_3 | Three most significant digits of the actual home zip code |
| 53 | PROFESSIONAL\_ZIP\_3 | Three most significant digits of the actual job zip code |
| 54 | TARGET\_LABEL\_BAD=1 | Target Variable: BAD=1, GOOD=0 |

**TARGET VARIABLE:**

The target variable of the above data set is TARGET\_LABEL\_BAD=1, which is a binary classification variable (1/0). Our objective is to predict whether the client is creditworthy or not.



***Fig-2: This image depicts the amount of YES vs No (1/0) in the target variable***

In the above dataset, 73.9% of the clients are credit worth and 26.1% of the clients are not applicable for the loan approval. **We observe that there is a presence of moderate amount of class imbalance.**

**DATA-PRE PROCESSING:**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

The data consists of 50,000 rows and 54 columns. Out of these we have 44 categorical columns and the rest as numerical.

**Data types of the variables are as follows:**

CLERK\_TYPE object

PAYMENT\_DAY int64

APPLICATION\_SUBMISSION\_TYPE object

POSTAL\_ADDRESS\_TYPE int64

SEX object

MARITAL\_STATUS int64

QUANT\_DEPENDANTS int64

STATE\_OF\_BIRTH object

CITY\_OF\_BIRTH object

NATIONALITY int64

RESIDENCIAL\_STATE object

RESIDENCIAL\_CITY object

RESIDENCIAL\_BOROUGH object

FLAG\_RESIDENCIAL\_PHONE object

RESIDENCIAL\_PHONE\_AREA\_CODE object

RESIDENCE\_TYPE float64

MONTHS\_IN\_RESIDENCE float64

FLAG\_MOBILE\_PHONE object

FLAG\_EMAIL int64

PERSONAL\_MONTHLY\_INCOME float64

OTHER\_INCOMES float64

FLAG\_VISA int64

FLAG\_MASTERCARD int64

FLAG\_DINERS int64

FLAG\_AMERICAN\_EXPRESS int64

FLAG\_OTHER\_CARDS int64

QUANT\_BANKING\_ACCOUNTS int64

QUANT\_SPECIAL\_BANKING\_ACCOUNTS int64

PERSONAL\_ASSETS\_VALUE float64

QUANT\_CARS int64

COMPANY object

PROFESSIONAL\_STATE object

FLAG\_PROFESSIONAL\_PHONE object

PROFESSIONAL\_PHONE\_AREA\_CODE object

MONTHS\_IN\_THE\_JOB int64

PROFESSION\_CODE float64

OCCUPATION\_TYPE float64

PRODUCT int64

FLAG\_ACSP\_RECORD object

AGE int64

RESIDENCIAL\_ZIP\_3 object

PROFESSIONAL\_ZIP\_3 object

TARGET\_LABEL\_BAD=1 int64

CLIENT\_ID int64

The above shows the data types of the various variables, but there are also categorical variables that are already encoded for us during the data collection process, hence we can use them as they are for our model building, but the other categorical variables need to be encode to their numeric form before the model building process.

**NULL VALUE IMPUTATION:**

The next step of data pre-processing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset. The following are the null value percentages:

PAYMENT\_DAY 0.000

APPLICATION\_SUBMISSION\_TYPE 0.000

SEX 0.000

MARITAL\_STATUS 0.000

QUANT\_DEPENDANTS 0.000

STATE\_OF\_BIRTH 0.000

CITY\_OF\_BIRTH 0.000

NATIONALITY 0.000

RESIDENCIAL\_STATE 0.000

RESIDENCIAL\_CITY 0.000

RESIDENCIAL\_BOROUGH 0.000

FLAG\_RESIDENCIAL\_PHONE 0.000

RESIDENCE\_TYPE 2.698

MONTHS\_IN\_RESIDENCE 7.554

FLAG\_MOBILE\_PHONE 0.000

FLAG\_EMAIL 0.000

PERSONAL\_MONTHLY\_INCOME 0.000

OTHER\_INCOMES 0.000

FLAG\_VISA 0.000

FLAG\_MASTERCARD 0.000

FLAG\_DINERS 0.000

FLAG\_AMERICAN\_EXPRESS 0.000

FLAG\_OTHER\_CARDS 0.000

QUANT\_BANKING\_ACCOUNTS 0.000

QUANT\_SPECIAL\_BANKING\_ACCOUNTS 0.000

PERSONAL\_ASSETS\_VALUE 0.000

QUANT\_CARS 0.000

COMPANY 0.000

FLAG\_PROFESSIONAL\_PHONE 0.000

MONTHS\_IN\_THE\_JOB 0.000

PROFESSION\_CODE 15.512

OCCUPATION\_TYPE 14.626

PRODUCT 0.000

FLAG\_ACSP\_RECORD 0.000

AGE 0.000

RESIDENCIAL\_ZIP\_3 0.000

PROFESSIONAL\_ZIP\_3 0.000

TARGET\_LABEL\_BAD=1 0.000

CLIENT\_ID 0.000

Here there are only 4 variables with null values, hence we try and impute them where ever possible. We can impute this using mean, median, mode, b fill, f fill or based on the dataset, you can impute the best possible value such that there is not much of a deviation in the summary of the variable from before the imputation.

* Here, for the variable, Residence type, we have imputed the value using the mode value because the value for the mode is substantially high from the other categories as shown below:

**>> X['RESIDENCE\_TYPE'].value\_counts()**

|  |
| --- |
| 1.0 41572 |
| 2.0 3884 |
| 5.0 1983 |
| 0.0 760 |
| 4.0 311 |
| 3.0 141 |

* For the variable, Months in residence, we have imputed the value with the median as it is a numeric variable.
* For the variable, Profession code, we have again imputed the null value with the mode as there is a large disparity in the mode value comparing to the rest of the classes.

**>> X['PROFESSION\_CODE'].value\_counts()**

|  |
| --- |
| 9.0 30092 |
| 11.0 3545 |
| 0.0 3540 |
| 2.0 2827 |
| 12.0 489 |
| 10.0 425 |
| 16.0 344 |
| 13.0 313 |
| 7.0 216 |
| 8.0 144 |
| 6.0 136 |
| 15.0 63 |
| 17.0 35 |
| 4.0 27 |
| 3.0 18 |
| 5.0 12 |
| 14.0 9 |
| 1.0 8 |
| 18.0 1 |

* For the variable occupation type, we do the same null imputation with mode as there is a disparity in the mode V/s the other classes.

**>> X['OCCUPATION\_TYPE'].value\_counts()**

|  |
| --- |
| 2.0 16947 |
| 1.0 8742 |
| 4.0 7000 |
| 5.0 6891 |
| 0.0 2788 |
| 3.0 319 |

**CHECKING FOR OUTLIERS:**

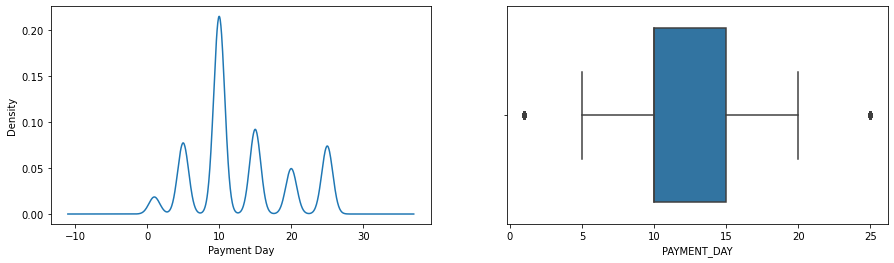
Data has outliers present in each of the numerical columns. For making the base model, we should not perform any outlier treatment and retain all the rows present in the data. The outlier treatment is done after the base model building.

**EXPLORATORY DATA ANALYSIS:**

### Univariate Analysis:

*For Numerical Variables: -* We plot the distribution curve and box plot to study the variation of the numerical data.

1. Payment day:

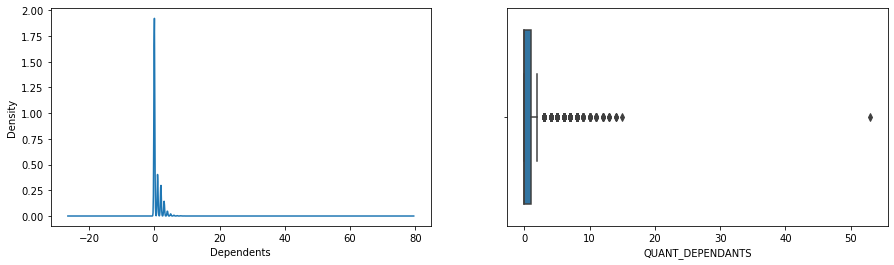


Skew: 0.5071180550490278

Kurtosis: -0.5988538279497719

* + Payment day is almost normally distributed.
  + It is mesokurtic
  + IQR of payment day lies from 5-20. Outliers are present.

1. Quantity of Dependents:

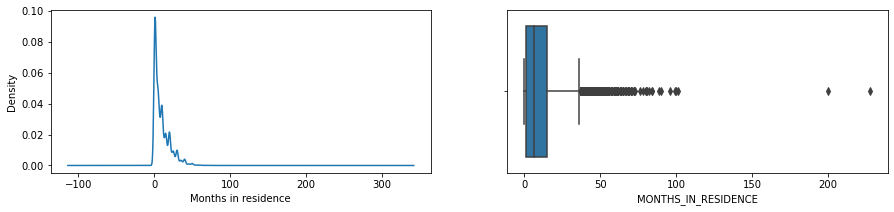


Skew: 4.077756812678044

Kurtosis: 83.1914311606446

* + Quant\_dependents is right skewed.
  + It is leptokurtic
  + IQR of lead dependents lies from 0-1. Outliers are present.
  + Highest frequency is at 0

1. Months in Residence:

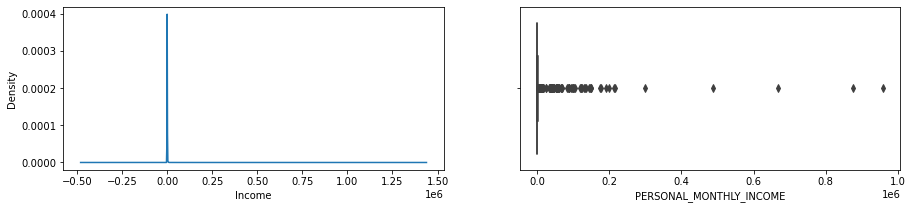


Skew: 1.9036703408954152

Kurtosis: 9.129723205274011

* + Months in residence are right skewed.
  + It is leptokurtic
  + IQR of month’s in residence lies from 0-15. Outliers are present.

1. Personal monthly Income:

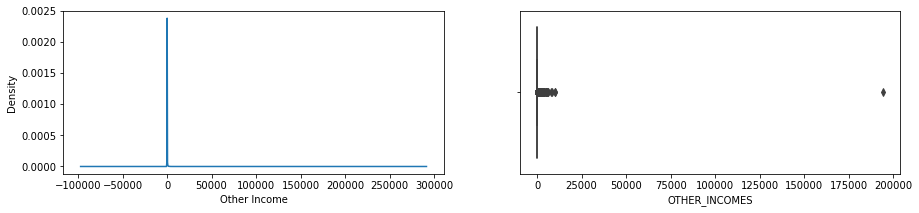


Skew: 85.7057480035924

Kurtosis: 8975.879000318488

* + Personal monthly Income is right skewed.
  + It is leptokurtic
  + IQR of monthly income lies from 0-800. Outliers are present.

1. Other Income:



Skew: 207.2713503426558

Kurtosis: 45136.46300246626

* + Other Income is right skewed.
  + It is leptokurtic
  + Outliers are present.

1. Age:



Skew: 0.4731456810100421

Kurtosis: -0.4050732444053988

* + Age is right skewed.
  + It is mesotokurtic
  + IQR of lead time lies from 0-53. Outliers are present.

*For Categorical Variables –* We plot a combination of bar graph and pie chart to understand the distribution of categorical data in the dataset.

1. Application Submission Type:

There are 3 different category classes in the types of submissions, they are web, carga and offline. We see that there are more web based applications compared to the others.

1. Sex:

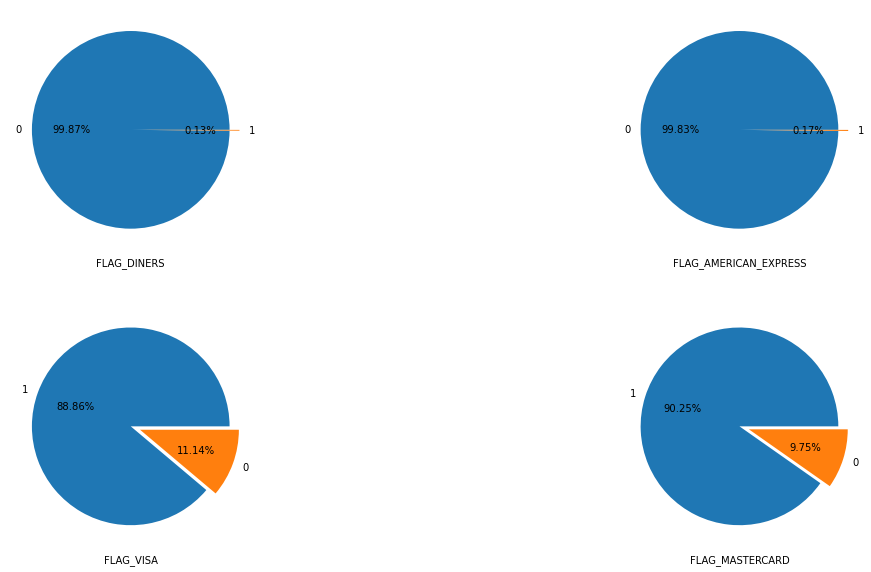
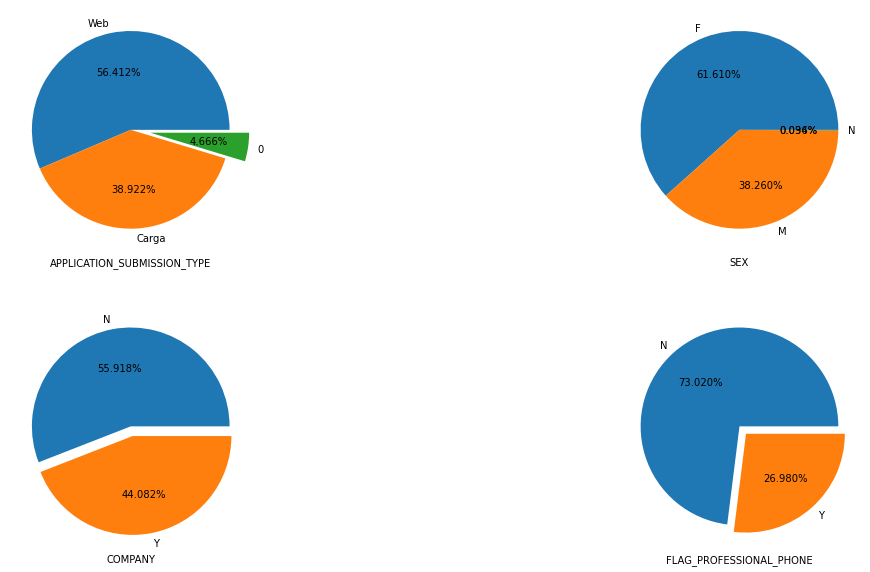
There are 3 different categories in the gender class, Male, Female and Not mentioned. We see that there are more female applicants than there are male applicants.

1. Company:

This is a binary class (yes/no). We can see that there are more unemployed people coming in for a loan than people with a job.

1. Flag professional phone:

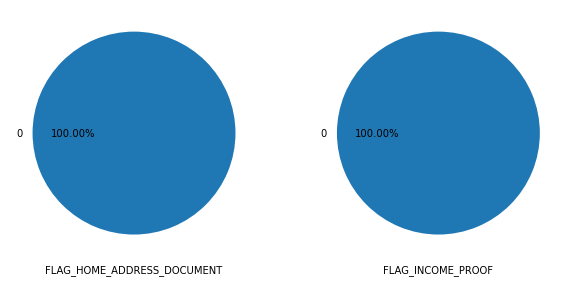
This is also a binary class variable. We can see that the number of No’s exceed the number of Yes’, which means that the professional phone Is not flagged in many cases.



1. Flagged Cards:

From the above images, we see that there is more number of flagged cards compared to non flagged card holders with diners coming first, closely followed by Amex. Visa and MasterCard’s have comparatively less flagged cards compared to Diners and Amex, none the less, they also have high number of flagged cards.

1. Flagged Proof:



From the image above, we can see that there are no flagged documents during verification, which means there is no variance in these columns. Lets also cross check the claim with the help of the count plots of the document verification.



Here we can cross verify that there is no additional class in the variables.

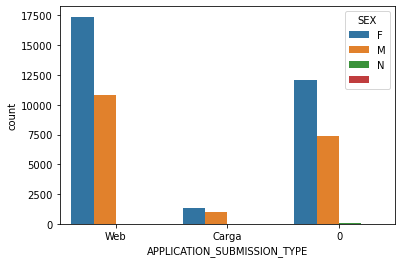
1. Product:



The above image shows us that there are 3 different types of product in the variable. The product 1 is having the most amount of client count compared to product 2 and product 7.

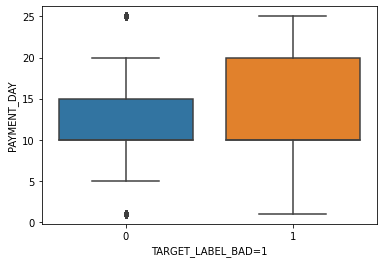
## Multi-Variate Analysis:

1. Application Type VS Gender:



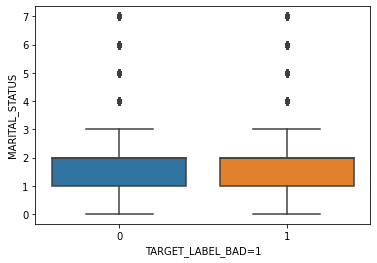
This shows that Web application has the most count and female applicants have the overall highest count.

1. Payment Day VS Target:



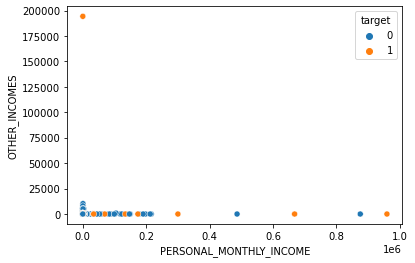
This shows us that there is almost no difference in the payment days for good and bad clients, but the 75th percentile for the bad clients is 5 days away from the good clients.

1. Marital Status VS Target:



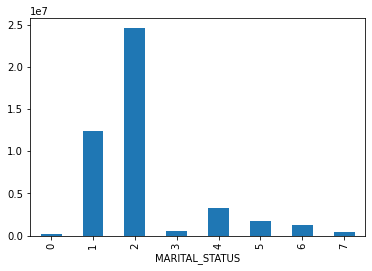
From the above image, we can see that the marital status has little to no effect on the target variable.

1. Income VS Other Incomes with Target as Hue:



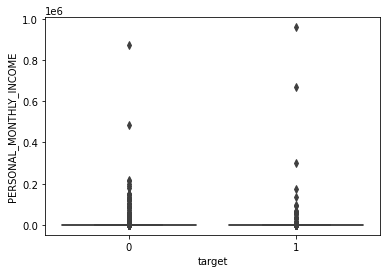
This shows that people with almost no personal income have some other sources of income and as the monthly income increases, there are almost no other sources of income. It also shows that there is a large amount of personal income lies between 0-25000.

1. Marital Status VS Personal Income:



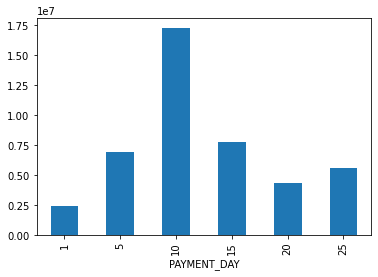
Here we can see that the class 2 has the highest amount of personal income vs the class 1. The rest have significantly less personal income.

1. Income VS Target:



There is some discrepancies in the data for bad clients and income.

1. Payment Day VS Personal Income:



From the above image, we see that the people getting paid on the 10th day of the month have relatively higher incomes.

**CORRELATION MATRIX:**

Heat-Map - Pearson Correlation Matrix

(Assumption : For the Pearson correlation, both variables should be normally distributed. Other assumptions include linearity and homoscedasticity)

It gives a measure of how much two numeric variables are linearly correlated. It tries to obtain a best fit line between two numeric variables and how close the points are to a fitted line.

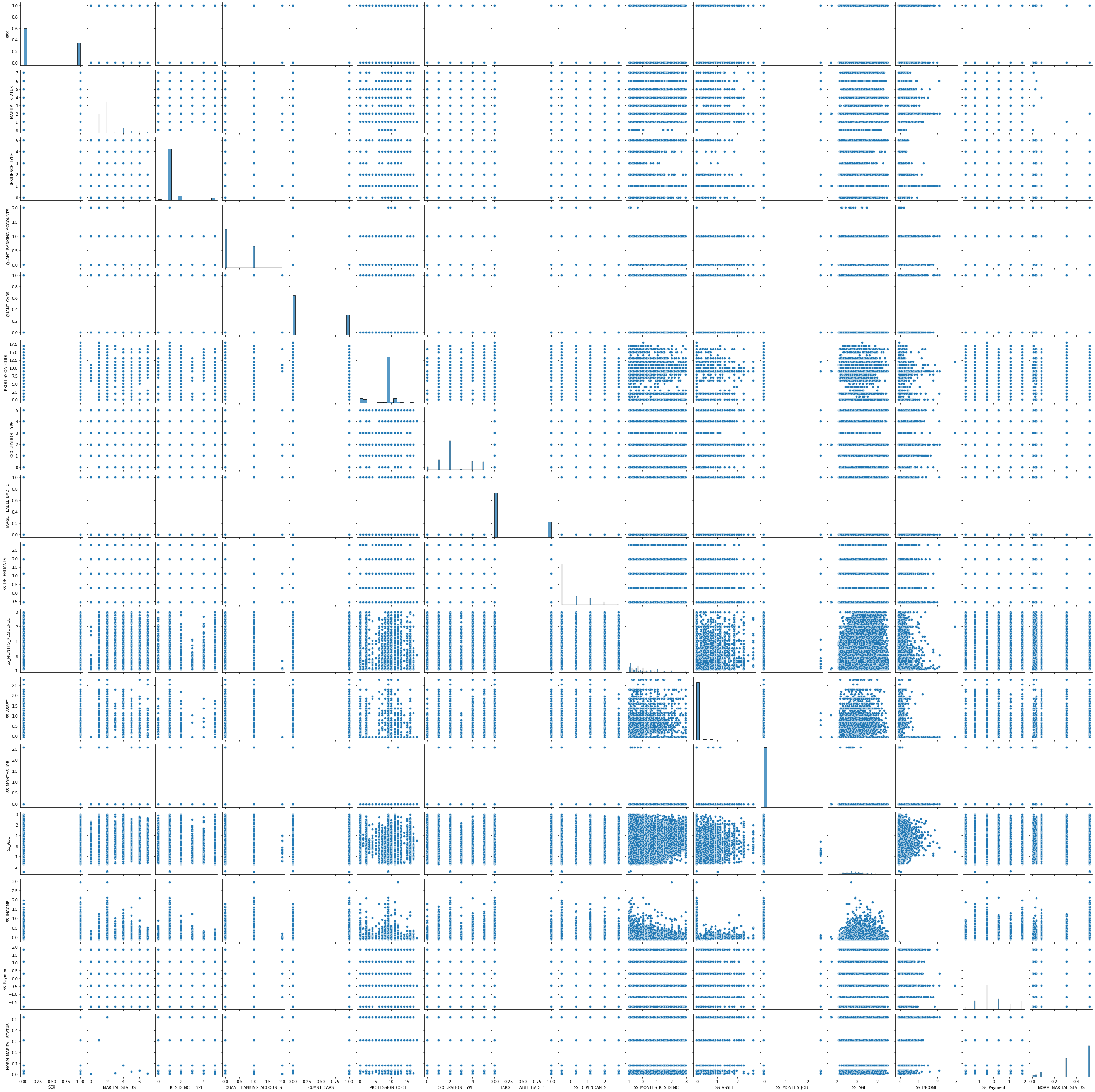


***Fig- Heat map: Pearson Correlation Matrix***

From the heat map we can understand:

* There is high correlation positive between quantity of cars and the quantity of bank accounts.
* Flagged email and quantity of cars have strong negative correlation.
* Flagged email and quantity of bank accounts also have a strong negative correlation

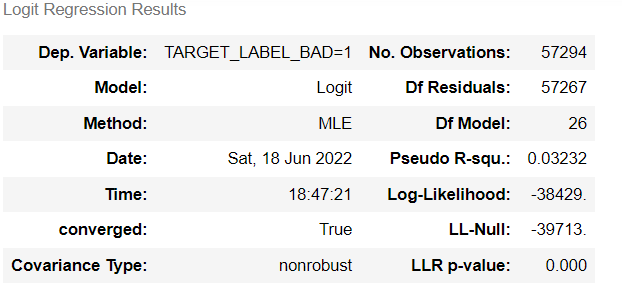
**PAIR PLOT:**



***Fig- Pair plot of all the numeric variables***

**STATISTICAL TEST:**

Model Summary:



Variable Summary:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **SEX** | 0.1270 | 0.018 | 6.952 | 0.000 | 0.091 | 0.163 |
| **NATIONALITY** | 0.2226 | 0.049 | 4.513 | 0.000 | 0.126 | 0.319 |
| **FLAG\_RESIDENCIAL\_PHONE** | -0.5676 | 0.024 | -23.730 | 0.000 | -0.615 | -0.521 |
| **RESIDENCE\_TYPE** | -0.0062 | 0.010 | -0.625 | 0.532 | -0.026 | 0.013 |
| **FLAG\_EMAIL** | 0.0333 | 0.033 | 1.009 | 0.313 | -0.031 | 0.098 |
| **FLAG\_VISA** | -0.1025 | 0.032 | -3.226 | 0.001 | -0.165 | -0.040 |
| **FLAG\_MASTERCARD** | -0.4142 | 0.034 | -12.044 | 0.000 | -0.482 | -0.347 |
| **FLAG\_DINERS** | 0.0058 | 0.328 | 0.018 | 0.986 | -0.637 | 0.649 |
| **FLAG\_AMERICAN\_EXPRESS** | -0.6690 | 0.311 | -2.148 | 0.032 | -1.279 | -0.059 |
| **FLAG\_OTHER\_CARDS** | -0.6058 | 0.237 | -2.560 | 0.010 | -1.070 | -0.142 |
| **QUANT\_BANKING\_ACCOUNTS** | -0.1089 | 0.042 | -2.597 | 0.009 | -0.191 | -0.027 |
| **QUANT\_CARS** | 0.1164 | 0.046 | 2.513 | 0.012 | 0.026 | 0.207 |
| **COMPANY** | -0.1606 | 0.020 | -8.200 | 0.000 | -0.199 | -0.122 |
| **PRODUCT** | 0.0225 | 0.008 | 2.668 | 0.008 | 0.006 | 0.039 |
| **Carga** | -0.3586 | 0.057 | -6.255 | 0.000 | -0.471 | -0.246 |
| **Web** | -0.1589 | 0.041 | -3.894 | 0.000 | -0.239 | -0.079 |
| **NORM\_STATE** | -0.5364 | 0.171 | -3.136 | 0.002 | -0.872 | -0.201 |
| **SS\_DEPENDANTS** | 0.0142 | 0.010 | 1.354 | 0.176 | -0.006 | 0.035 |
| **SS\_MONTHS\_RESIDENCE** | -0.0228 | 0.010 | -2.173 | 0.030 | -0.043 | -0.002 |
| **SS\_ASSET** | -0.4105 | 0.048 | -8.491 | 0.000 | -0.505 | -0.316 |
| **SS\_MONTHS\_JOB** | -12.5647 | 3.628 | -3.463 | 0.001 | -19.675 | -5.454 |
| **SS\_AGE** | -0.3485 | 0.010 | -33.805 | 0.000 | -0.369 | -0.328 |
| **SS\_INCOME** | -0.1876 | 0.104 | -1.795 | 0.073 | -0.392 | 0.017 |
| **SS\_Payment** | 0.1365 | 0.009 | 15.871 | 0.000 | 0.120 | 0.153 |
| **NORM\_PROFESSION\_CODE** | 0.2863 | 0.032 | 8.883 | 0.000 | 0.223 | 0.350 |
| **NORM\_MARITAL\_STATUS** | -0.3538 | 0.053 | -6.615 | 0.000 | -0.459 | -0.249 |
| **NORM\_OCCUPATION\_TYPE** | 0.0356 | 0.059 | 0.601 | 0.548 | -0.081 | 0.152 |

The above two images shows the summary statistics from building the Logistic regression model form the stats models library. Here we keep note of the pseudo R squared value

from the model summary image. The variable summary image shows us the various statistics of the variables. Here we can see there are almost 6 variables that have p values greater than 0.05, which means they are insignificant, just as we concluded from the variable relationship with the target variable.

1. Categorical columns – For categorical columns we perform chi-square test to check for the significance of the categorical column with respect to the target column.

|  |
| --- |
| *Hypothesis of Chi-square test* |
| *H0 : Attributes are independent* |
| *H1 : Attributes are dependent* |

We observe that the p\_values of diner, occupation type and residence type columns are greater than 0.05. Hence, we accept the null hypothesis in these scenarios. Therefore, we conclude that all the other categorical features are significant.

1. Numerical columns – We perform parametric and non-parametric tests for the numerical s. Under parametric test we perform ANOVA and under non-parametric test we perform Mann Whitney U test.

|  |
| --- |
| *Hypothesis for numerical tests* |
| *H0 : Two samples have the same mean (i.e insignificant)* |
| *H1 : Two samples have different mean (i.e significant)* |

* + ANOVA test & Mann Whitney U test

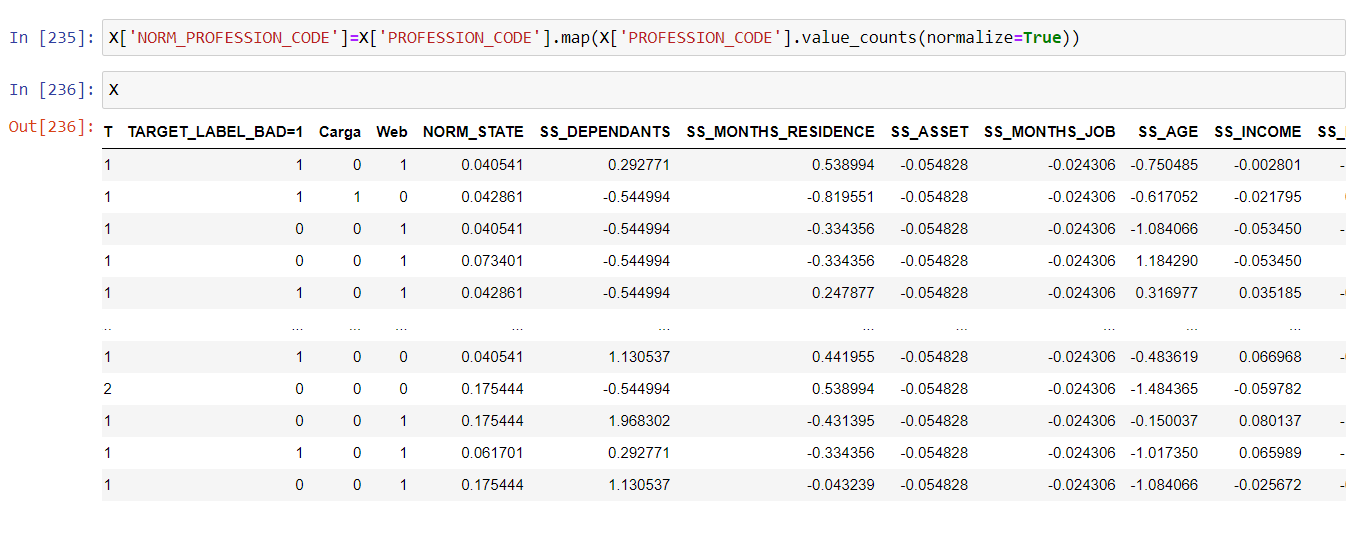
We observe that the p\_values for some variables like income and number of dependents are greater than 0.05. Hence these variables are insignificant.

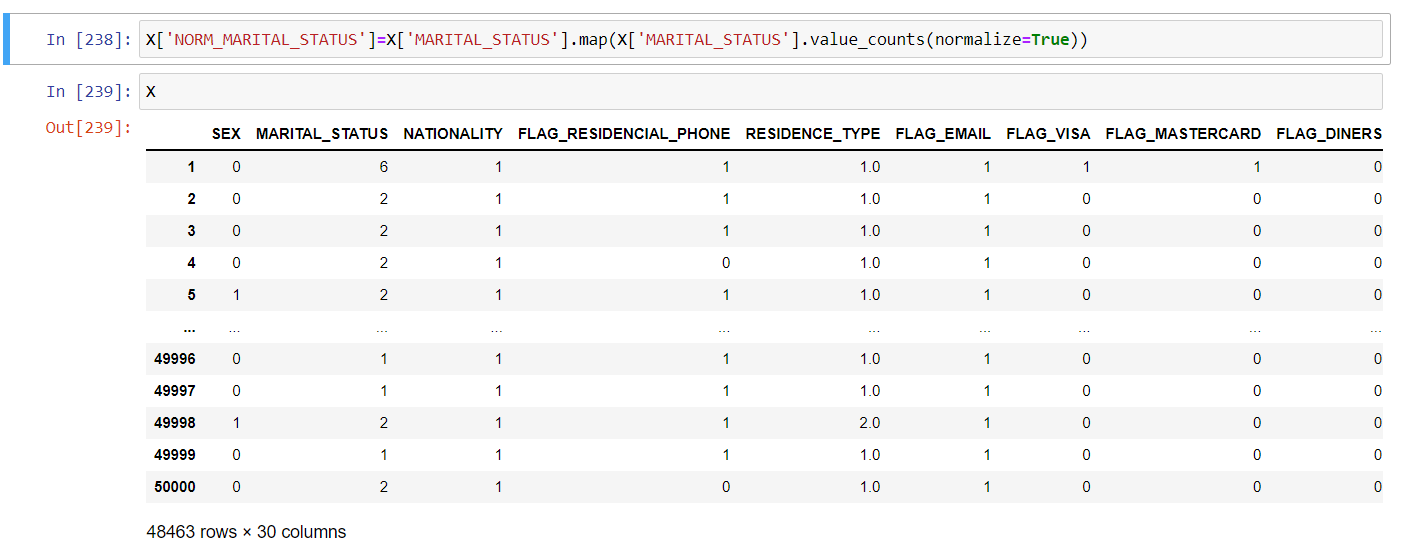
# BASEMODEL

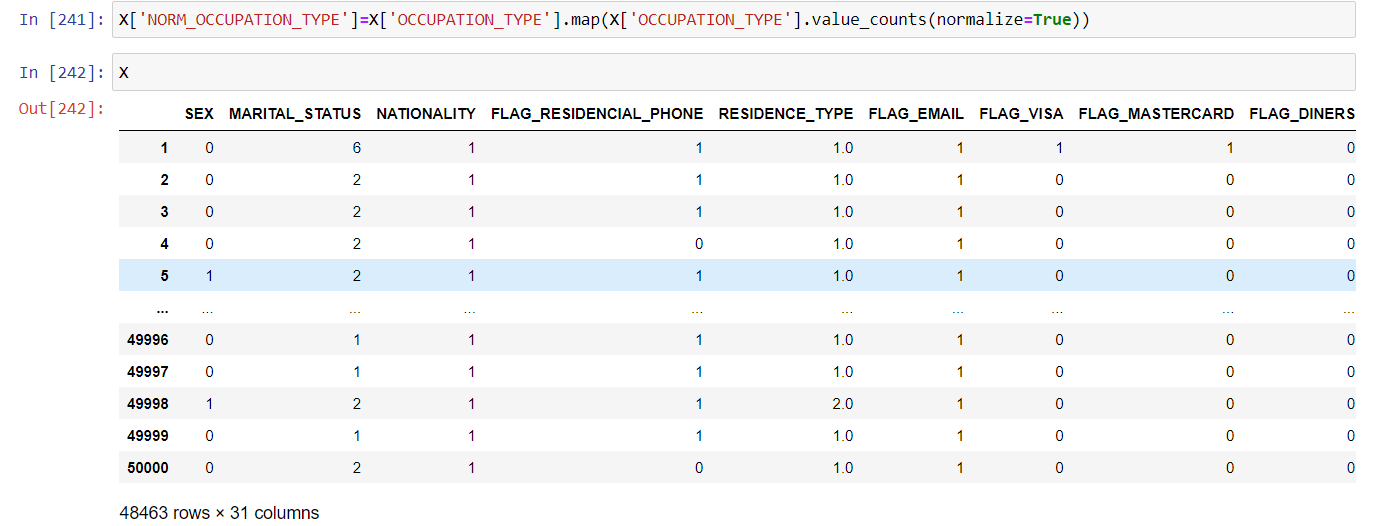
## Logistic Regression Model:

We have select Logistic Regression as our base model. For this we have encode all the categorical variables using frequency and dummy encoding and have used the numerical column as it is.

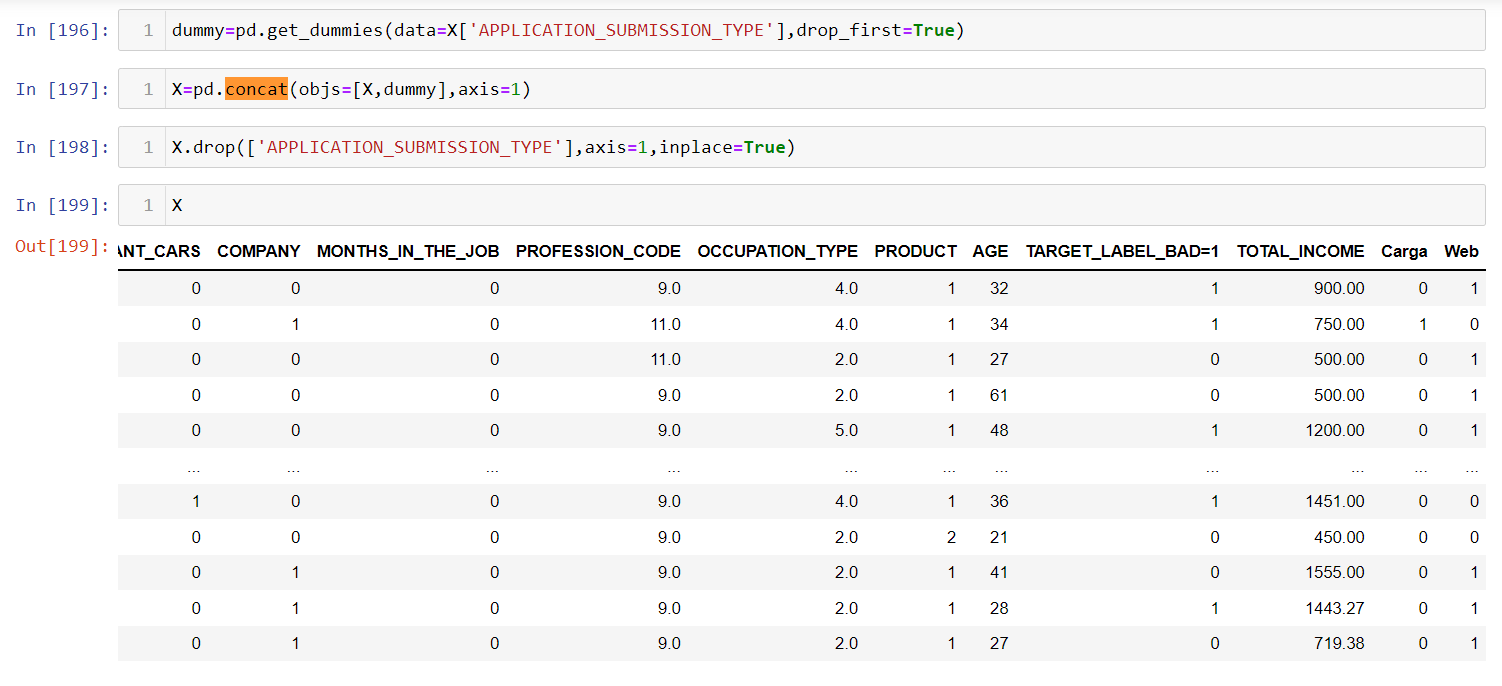
Encoding:





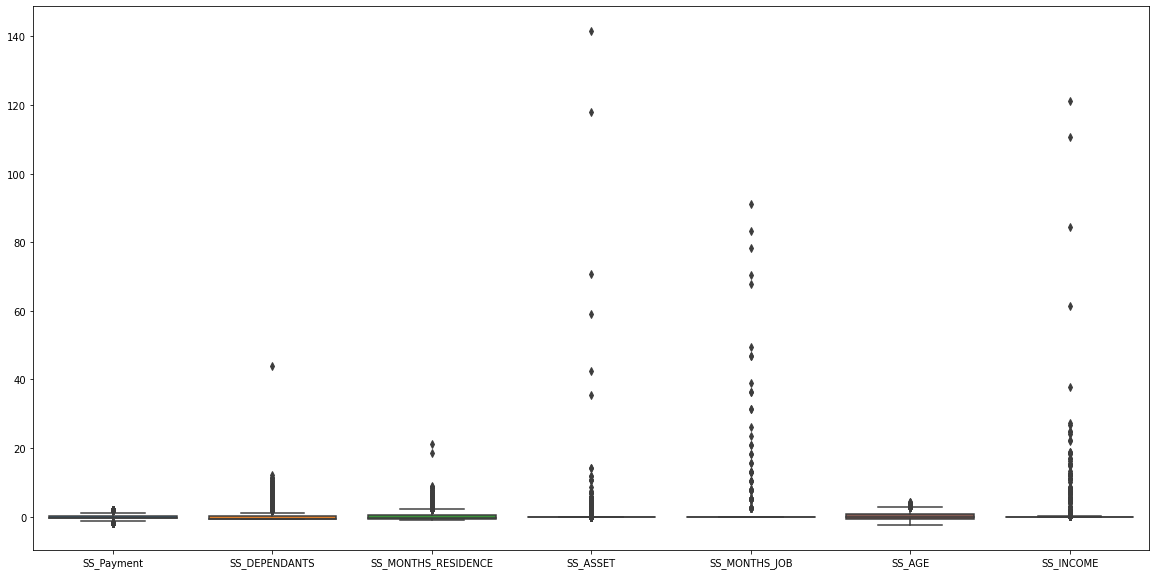


In the above scenarios we have used the frequency encoder for encoding the categorical variables into numeric variables such that the model will be able to interpret the values. We should keep in mind that the categorical variables can only be converted to numeric using frequency encoder when the number of counts for each class in the variable is unique and not repetitive. If this is not the case we go for dummy encoding.

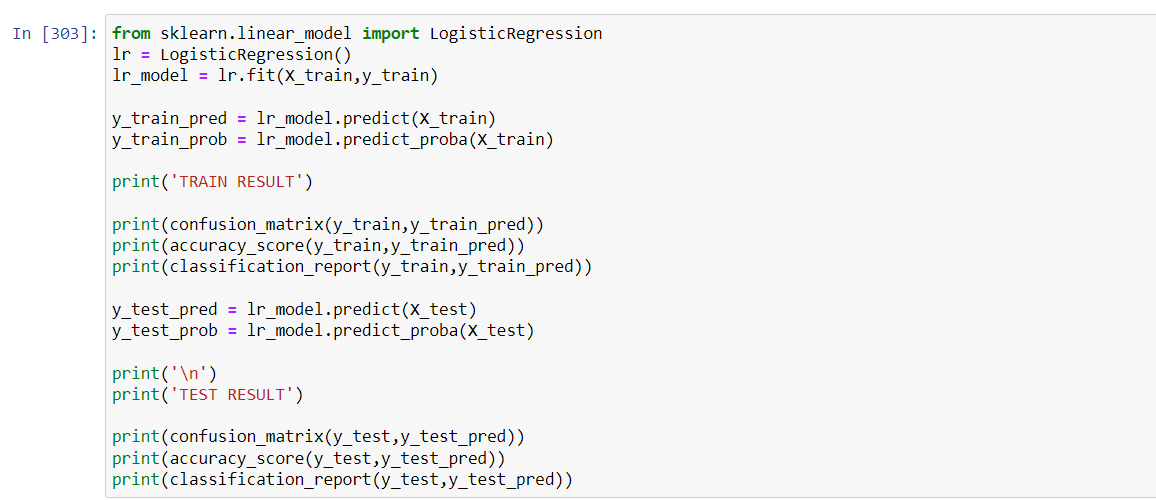


The above image shows the dummy encoding of the application type variable. This was possible only because there we minimal amount of classes. We have dropped the first column as in dummy encoding; we have an option to drop the first column without changing the function of the encoder.

Variable Distribution after Encoding:

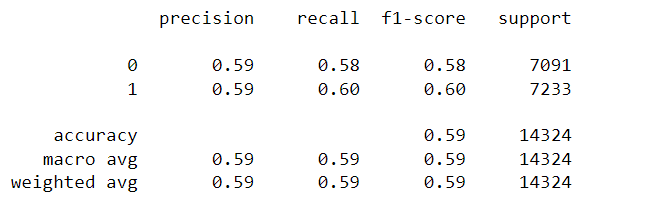


Logistic Regression Model:



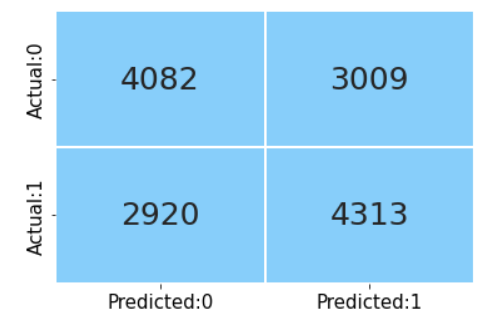
The above model is built after training the data from the train test split. The model is built from the Sci-Kit learn library.

Classification Report:



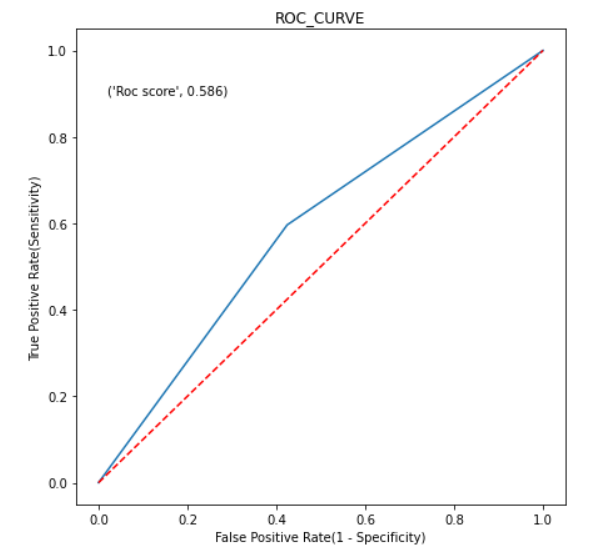
The above shows the classification report from building the base model using logistic regression. The accuracy is 59% where as the other metrics are also similar to the accuracy. It is better to have the metrics close to 1. This is only possible by tuning the model. This is what needs to be done after building the base model such that we get the best possible accuracy and metrics for the classification of the target variable.

Confusion Matrix:



The above image shows the True Positive, True Negative, False Positive and False Negative values predicted by our base model. From the accuracy score, we already know that the model will predict the class with 60% accuracy and that is shown here in the confusion matrix.

ROC Curve:



From the ROC curve, we can see the area under the curve which also depicts the accuracy of our model.

Future Work:

Further after building the base model, we will proceed with building non-linear models followed by feature selection and hyper parameter tuning. Also, from the EDA we learnt that various categorical columns have high cardinality. Cardinality means presence; there is presence of multiple unique values in a categorical column. Encoding such columns with One Hot Encoding leads to increase in the dimension of the dataset and higher computation. Hence, we need to explore methods to deal with such columns. Along with this according to our business problem we have to select an evaluation metric to compare the results of various models which we will have to build.

**REFERENCES:**

* CFI-Institute: <https://corporatefinanceinstitute.com/resources/knowledge/credit/>
* PAKKD Data Mining: <https://github.com/deepanshu88/Datasets/blob/master/CreditData/PAKDD%202010.zip>
* Brandon Foltz: https://www.youtube.com/c/BrandonFoltz/playlists