**INTERIM REPORT- CAPSTONE PROJECT**

**CLASSIFICATION PROBLEM: CREDIT DEFAULT (BASED ON TAIWAN DATASET)**

**TEAM MEMBERS:**

1. ADITYA JALAL
2. CHETAN GUGLANI
3. DHARMENDRA PRATAP SINGH
4. RAJAT ARORA
5. SUMEDHA OBEROI

**Under the mentorship of** : Ms ANJANA AGRAWAL

**ABSTRACT:**

**INDEX:**

**INTRODUCTION:**

Credit card is type of financial tool which is used as a form of credit extended to complete transactions in absence of cash. The basic format of credit card functioning relates towards ‘consumption before payment’. Along with this, credit card users aren’t liable to repay the consumed amount at one go. Through revolving credit scheme there is flexibility in credit payments. Revolving credit involves required minimum payment of amount for the account to remain active. If a customer opts for paying through revolving credit, he/she is liable to pay an additional interest charge on the outstanding amount post paying the minimum amount.

In most scenarios, upon delayed payment of even minimum amount a penalty is levied post a 20-25 days’ extension period.

For a bank, a customer who opts revolving credit is a good customer in comparison to the one who pays his/her dues on time, because banks earnings are made through the interest charged.

The dataset contains information relating to the credit card clients in Taiwan from April 2005 to September 2005. The data set comes from the [UCI Machine Learning Repository Irvine](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients).

The aim of the project is to analyze and predict whether the customer would default payment of credit card.

(Establish a baseline and compare several machine learning models on a performance metric)For closest accurate prediction we will first perform work on Logistic Regression –Classification Algorithm to prepare a robust model. We will also work towards understanding the important features in our predictive model.

Will discuss on Logistic Regression baseline model, if its assumptions are not met then we have to apply some other model and treat as baseline.

**EXPLORATORY DATA ANALYSIS:**

* **Data Information:**

The first and foremost step in any data analysis process is to understand the nature and breakdown of the data and attributes. The below attached is the summary of the dataset.

Table : Data Information:

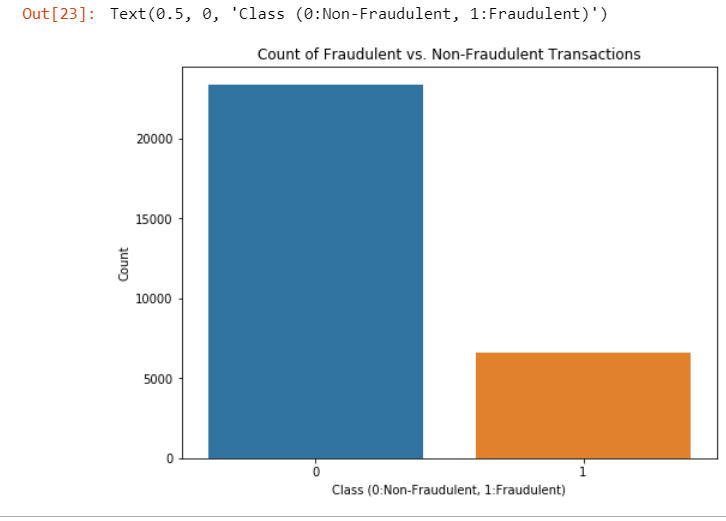
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RangeIndex:** | **30000** |  |  |  |
| **columns** | **25** |  |  |  |
| **Data columns** | **Rows** | **Description** | **Null values** | **Data Type** |
| ID | 30000 | Customer ID | non-null | Integer |
| LIMIT\_BAL | 30000 | Credit Limit | non-null | Integer |
| SEX | 30000 | Gender of the customer | non-null | Integer |
| EDUCATION | 30000 | Education status of the customer | non-null | Integer |
| MARRIAGE | 30000 | Maritial Status of the customer | non-null | Integer |
| AGE | 30000 | Age of the customer | non-null | Integer |
| PAY\_1 | 30000 | Repayment September | non-null | Integer |
| PAY\_2 | 30000 | Repayment August | non-null | Integer |
| PAY\_3 | 30000 | Repayment July | non-null | Integer |
| PAY\_4 | 30000 | Repayment June | non-null | Integer |
| PAY\_5 | 30000 | Repayment May | non-null | Integer |
| PAY\_6 | 30000 | Repayment April | non-null | Integer |
| BILL\_AMT1 | 30000 | Bill amount September | non-null | Integer |
| BILL\_AMT2 | 30000 | Bill amount August | non-null | Integer |
| BILL\_AMT3 | 30000 | Bill amount July | non-null | Integer |
| BILL\_AMT4 | 30000 | Bill amount June | non-null | Integer |
| BILL\_AMT5 | 30000 | Bill amount May | non-null | Integer |
| BILL\_AMT6 | 30000 | Bill amount April | non-null | Integer |
| PAY\_AMT1 | 30000 | Previous month payment in September | non-null | Integer |
| hPAY\_AMT2 | 30000 | Previous month payment in August | non-null | Integer |
| lPAY\_AMT3 | 30000 | Previous month payment in July | non-null | Integer |
| PAY\_AMT4 | 30000 | Previous month payment in June | non-null | Integer |
| PAY\_AMT5 | 30000 | Previous month payment in May | non-null | Integer |
| PAY\_AMT6 | 30000 | Previous month payment in April | non-null | Integer |
| DEFAULT | 30000 | Will customer default ? | non-null | Integer |

**From the above we can draw the following below inferences:**

* Dataset has 25 columns and 30,000 rows
* The entire dataset is of integer data type. Although, we understand Sex, Education, Marriage, Pay\_’X’ columns and Default are categorical in nature.
* The dataset provided is for a 6-month bill cycle period.
* The target column is already in 1-0 number format, wherein 0- represents No and 1- represents Yes, with respect to whether a customer will default or not. And the other understood categorical columns (Sex, Marriage, Education) are in Numerical Categories\*. Pay\_’X’ columns contain (-2=no consumption- inactive account, -1=paid duly, 0=the use of revolving credit, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above) subclasses. Values are till 8th Month
* At an overview, data does not contain any null values. But by further investigating, we have observed that 2 categorical columns namely (Marriage and Education) contain ‘unknown’ class which are nothing but undocumented values. So, as of now we haven’t replaced the ‘unknown’ entries but have further evaluated their relevance as discussed ahead.
* There were 23364 non-fraudulent transactions (77.880%) and 6636 fraudulent transactions (22.120%).

The given target column is slightly imbalanced, but doesn’t seem to require any data balancing so far, even though the chances of Fraud are relatively low, we still need to drill down and examine further pointers that contribute towards this 22% distribution.

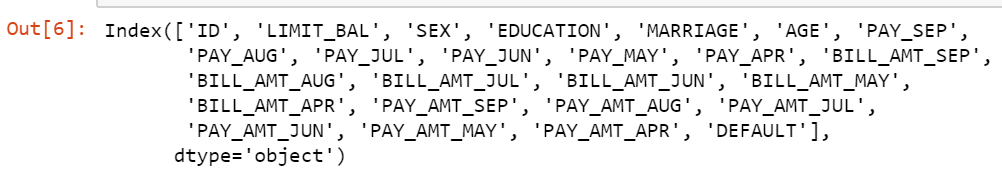
**Figure 1 DEFAULT COLUMN BREAKDOWN**



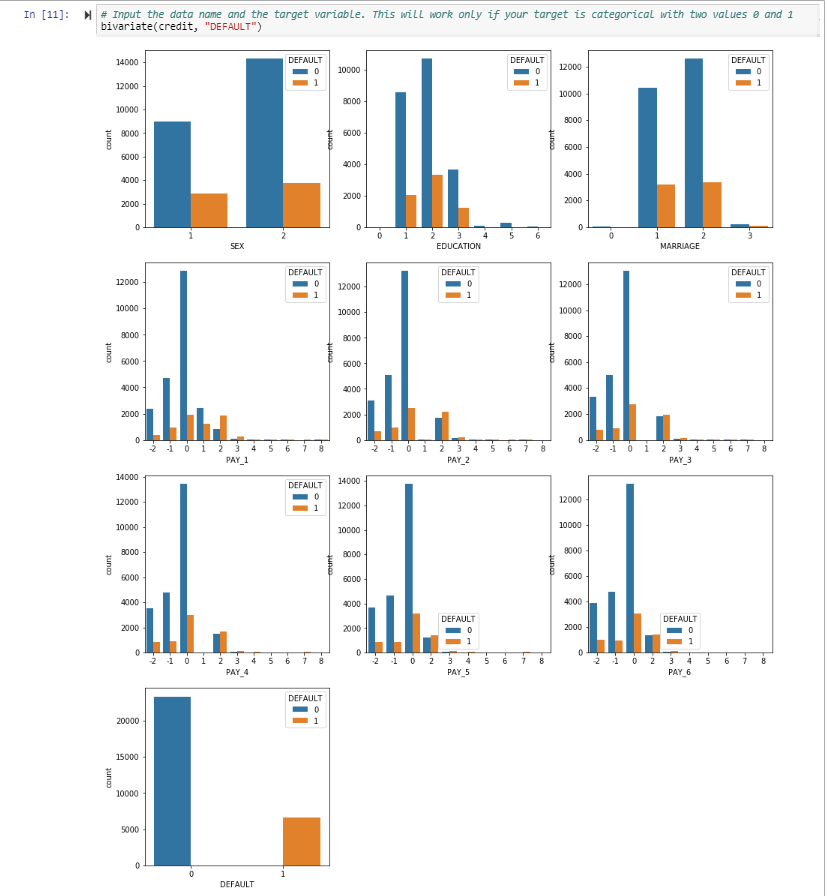
* **Data Cleaning:**

For simplicity in understanding, we altered the naming convention for the columns Pay\_’X’, Bill\_AMT\_’X’ and PAY\_AMT\_’X’, (where X stands for months 1-6) as they appeared to be in puzzling manner, to their respective months as below:

**Figure 2 Renamed Columns**



Post renaming, we also looked into subclasses of columns MARRIAGE and EDUCATION. We noted, that column MARRIAGE included 4 subclasses – [0,1,2,3] and EDUCATION included subclasses-[0,1,2,3,4,5,6]. To better under what these subclasses represent and in general the breakdown of rest of the columns, we did a univariate and bivariate analysis on all the columns (below).

**Figure 3: UNIVARIATE COUNT ANALYSIS** **Figure 4: BIVARIATE COUNT ANALYSIS- WITH DEFAULT**

In the data description we were given only 3 categories in the Marriage (1 = Married, 2 = Single and 3 = others) even though there existed a 0 category. This could happen because someone just forgot to mention their Marital Status or they deliberately did not mention that. From the graphical analysis, we noted that 0 category holds some relevance, thus we renamed the subclasses as per below:

1. 'Unknown',
2. 'Married',
3. 'Single',
4. 'Others'

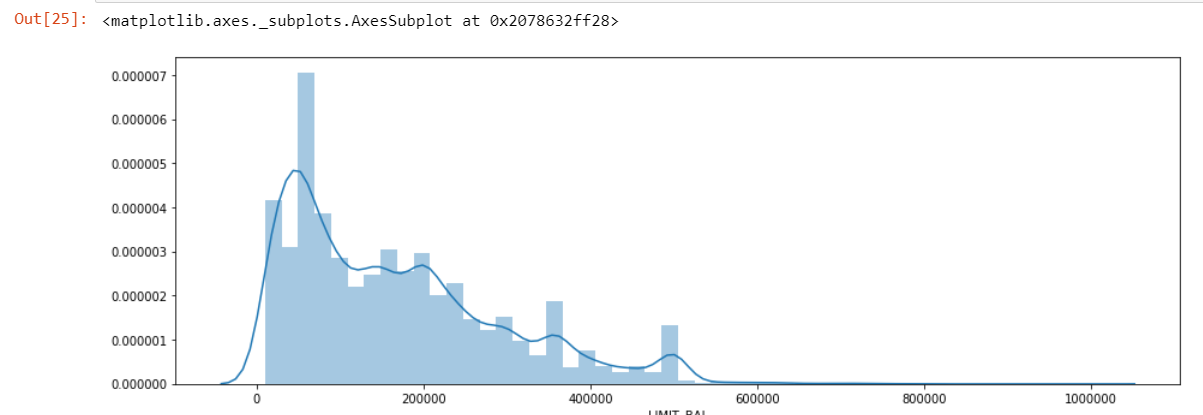
We withheld clubbing 0 category with 3 before further analyzing the feature, since Others category existed, it would be unwise to club the Unknown with Others before further analyzing the same.

Similarly, for EDUCATION we were not given any description for 0 subclass. Thus, renamed the subclasses as per below:

1. 'Unknown',
2. 'Graduate',
3. 'University',
4. 'High School',
5. 'Others',
6. 'Unknown',
7. 'Unknown'

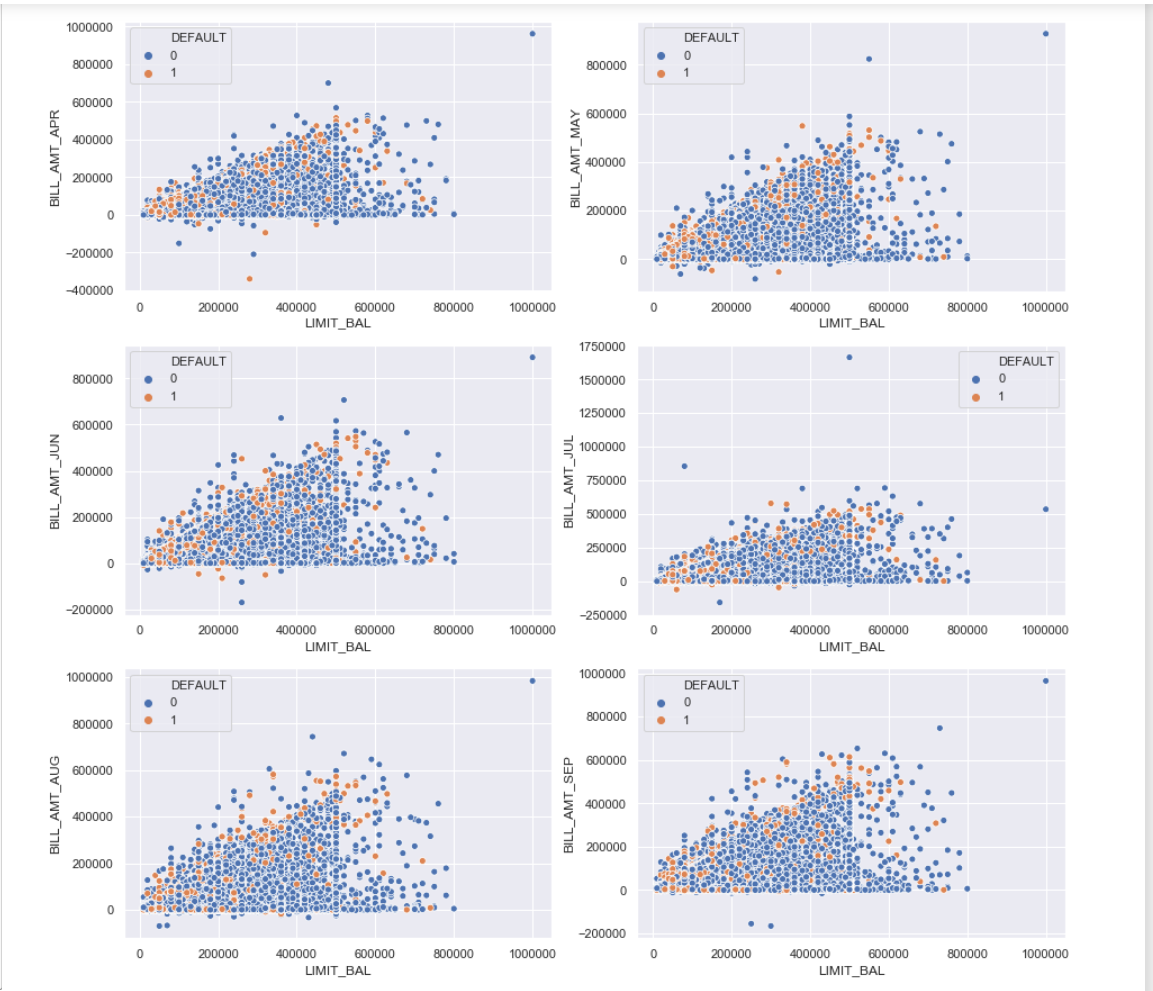
* **OBSERVATIONS:**

1. LIMIT\_BAL column – It is fairly understood the relevance credit limit implies on credit card function. The Limit balance defines the entitled credit limit to every customer. We drew a distplot of the same to understand the nature of the values within.

**Figure 5 LIMIT\_BAL Distplot**

* Limit\_Bal is rightly skewed, median = 140,000, mean = 1,674,848. Mean > Median, implies the data is heavier on the right tail, implying significant presence of outliers.
* SKEWNESS=0.99 and Kurtosis = 0.53. Data is moderately skewed towards right, although within the ideal range.
* There are few outliers specifically around 600000,800000,1000000 limits. We need to check if these are our outliers or are extreme values?. The same is further explored in our analysis
* There appears to be Credit limit less than 0, ie, we have negative credit limit, which entails that there appears to be credit extended to customers- maybe a customer made extra payment in one month and the difference is credited in next month.
* There are modalities/peaks in histogram hinting towards hidden possible clusters. Major Credit balance appears to be 200000-300000 range.
* These modalities can be clustered on the basis of binning to identify whether our extreme values are our outliers. Also, the clustering can help us identify any patterns in the defaulters.

1. To understand nature of our customers spending capacity, we plotted a Bivariate analysis between Bills paid and Limit Balance:

**Figure 6: BIVARIATE ANALYSIS BETWEEN BILL\_PAY , LIMIT\_BAL w.r.t to DEFAULT**

* From the scatter plot above, we can understand it is a heteroscedastic plot, wherein as the Credit Limit balance increases, our Bill Amount also increases. This talks about the spending capacity of the customer. A Higher Limit denotes a Higher Spending Capacity.
* Thus, our question from Observation 1 , whether our pointers on the right tail in Limit Balance our Outliers or Extreme Values, we can safely comment that the extreme points aren't my Outliers, instead are our extreme values that needs to be included in the analysis.
* The trend which is visualised is, maximum Defaulters lie at the extreme left end of the x-axis. We can majorly see them cluster around lower ranges of LIMIT\_BAL. That means, customers with lower credit limit tends to default more. This point is further analyzed.

1. The correlation coefficients between Limit Bal- Bill and Limit Bal- Pay amount for respective months is below :

APR: 0.29038895064794834 0.21959536860441858

MAY: 0.2955623376582321 0.21720243239549658

JUN: 0.2939876237159871 0.20324241022458217

JUL: 0.2832357835816838 0.21016674772338997

AUG: 0.2783143639977619 0.1784079536837043

SEP: 0.2854298649649899 0.19523591523220946

* The variables appear to display a positive relationship between each other. Using Cohen’s standard, we can say that since our variables range lie between 0.10 – 0.29, there appears to be a weak association between the two variables.

**RESEARCH WORK:**

To further understand the breakdown of our dataset, we explored a bit on the background of the same. We referred to different research papers\*. To understand the nature of our anomalies