greatlearning
Learning for Life

THE USAGE OF MACHINE LEARNING TECHNIQUES TO PREDICT THE PROBABLITY OF DEFAULT OF CREDIT CARD CLIENTS.

INTERIM REPORT

G3-PGP DSE FT CHN CAPSTONE PROJECT – INTERIM REPORT

PROJECT SUMMARY

BATCH DETAILS	PGPDSE-FT-CHENNAI SEP21
TEAM MEMBERS	1. GREASH K, 2. JISMY JOHN, 3. PRUTHIV RAJAN K, 4. SHIBIKANNAN TM, 5. VIGNESH PRABAKARAN
DOMAIN OF PROJECT	FINANCE
PROPOSED PROJECT TITLE	THE USAGE OF MACHINE LEARNING TECHNIQUES TO PREDICT THE PROBABLITY OF DEFAULT OF CREDIT CARD CLIENTS.
GROUP NUMBER	3
TEAM LEADER	SHIBIKANNAN TM
MENTOR NAME	P.V.SUBRAMANIAN

MENTOR SIGN

TEAM LEADER SIGN

greatlearning
Learning for Life

TABLE OF CONTENTS

S.NO	TITLE	PAGE NO
1	CHAPTER 1 - REVIEW	1
1.1	INDUSTRY REVIEW	1
1.1.1	CURRENT PRACTICES AND BACKGROUND RESEARCH	1
1.2	LITERRATURE SURVEY	3
1.2.1	PUBLICATIONS	3
1.2.2	APPLICATION PAST AND UNDERGOING RESEARCH	5
2	CHAPTER 2 – DATASET AND DOMAIN	8
2.1	DATA DICTIONARY	8
2.2	DATA SET INFORATION	9
2.3	VARIABLE CATEGORIZATION	9
2.4	PRE PROCESSING ANALYSIS	10
2.4.1	INFO FUNCTION	10
2.4.2	REDUNDANT	11
2.5	PROJECT JUSTIFICATION	12
2.5.1	AIM	12
2.5.2	COMPLEXITY INVOLVED	12
2.5.3	OUTCOMES	12
3	CHAPTER 3 – DATA EXPLORATION	13
3.1	RELATIONSHIP BETWEEN VARIBALES	13
3.1.1	BIVARIATE ANALYSIS	13
3.1.2	MULTIVARIATE ANALYSIS	19
3.2	OUTLIERS	19
3.3	STATISTICAL SIGNIFICANCE OF VARIABLE	20
3.3.1	CHI- SQUARE TEST	20
3.3.2	NON- PARAMETRIC TEST	23
4	CHAPTER 4 – FEATURE ENGINEERING	24
4.1	FEATURE SELECTION	24
5	CHAPTER 5 – CLASSIFICATION	26
5.1	BASE MODEL BUILDING	26

5.2	**CLASSIFICATION REPORT**	**26**
5.3	**CONFUSION MATRIX**	**27**

greatlearning
Learning for Life

LIST OF FIGURES

FIG.NO	TITLE OF FIGURE	PAGE NO
2.1	JUPYTER NOTEBOOK INFO FUNCTION	10
2.2	BAR PLOT OF X3	11
2.3	BAR PLOT OF X4	11
3.1	BOX PLOT OF CREDIT LIMIT VS SEX	14
3.2	BOX PLOT OF DISTRUBUTION OF PEOPLE VS CREDIT CARD FACILITY	15
3.3	DENSITY PLOT OF MAXIMUM LIMIT OF CREDIT CARD LIMIT AMOUNT	15
3.4	BAR PLOT OF BILL AMOUNT VS. DEFAULT CREDIT CARD CUSTOMERS	16
3.5	BAR PLOT OF AUGUST BILL AMOUNT VS. DEFAULT CREDIT CARD CUSTOMERS	16
3.6	BAR PLOT OF JULY BILL AMOUNT VS. DEFAULT CREDIT CARD CUSTOMERS	17
3.7	BAR PLOT OF JUNE BILL AMOUNT VS. DEFAULT CREDIT CARD CUSTOMERS	18
3.8	BAR PLOT OF MAY BILL AMOUNT VS. DEFAULT CREDIT CARD CUSTOMERS	18
3.9	BAR PLOT OF APRIL BILL AMOUNT VS. DEFAULT CREDIT CARD CUSTOMERS	19
3.10	HEAT MAP FOR ALL NUMERIC VARIABLES	20
3.11	BAR PLOT FOR Y	24
3.12	NON-PARAMETRIC TEST IN JYUPTER NOTEBOOK	24
4.1	FEATURE SELECTION IN JYUPTER NOTEBOOK.	25
5.1	CLASSIFICATION REPORT IN JYUPTER NOTEBOOK.	26
5.2	CONFUSION MATRIX	27

CHAPTER - 1

1.1 INDUSTRY REVIEW

1.1.1 CURRENT PRACTICES AND BACKGROUND RESEARCH:

1. **Title: Credit card industry analysis.**
   Link:
   https://corporatefinanceinstitute.com/resources/knowledge/credit/credit-card-industry-analysis
   Summary:
   Credit card industry analysis helps determine the current state of credit card companies and their latest products and services. The credit card industry relies on constant innovations in marketing and technology, which has resulted in increasing competition among credit card companies. Credit card companies invest billions of dollars in marketing activities to acquire new customers and expand their customer base. Innovative marketing activities, such as rewards programs, discounts, loyalty points and zero interest, are implemented with the aim of attracting more customers to the program.
2. **Title: Changes landscape of credit's card industry.**
   Link:
   https://www.pwc.in/industries/financial-services/fintech/dp/the-changing-landscape-of-indias-credit-industry.html
   Summary:
   India has historically been a debit card market. However, the boom in credit score card issuance within side the remaining decade has modified this narrative and credit score playing cards are getting used prominently. This boom is similarly extended via way of means of the numerous services and products being presented via way of means of FIs, and such merchandise are being an increasing number of utilized by customers, particularly the millennial population. Credit card issuance has grown considerably in India at a compound annual boom charge (CAGR) of 20% within side the remaining 4 years. The quantity of credit score cardholders multiplied from 29 million in March 2017 to sixty two million in March 2021. It has similarly grown via way of means of 26% and 23% respectively in 2019 and 2020. However, the COVID-19 pandemic affected the boom charge of India`s credit score card enterprise and it grew via way of means of most effective 7% in 2020–21. The boom charge is predicted to enhance marginally in FY21–22 however will continue to be gradual because of the regulations on card issuance via way of means of a few massive banks and bills networks. Similarly, credit score card transactions had been developing at a CAGR of 16% until 2019–20 however went again to the 2018–19 tiers in FY20–21, as depicted within side the determine above. The boom charge became low all through the primary 1/2 of of 2020–21 aleven though it received momentum all through the second one 1/2 of.

3. **Title: Machine-learning algorithms for credit-card applications.**
   Link:
   https://academic.oup.com/imaman/articleabstract/4/1/43/656001?redirectedFrom=PDF
   Summary:
   Credit checks include forecasting applicant credibility and profitability. The purpose of this paper is to apply a set of algorithms to credit card scoring. Little is known about the strengths and weaknesses of their comparisons, despite the fact that many numbers and connectionist learning algorithms address the same problem of learning from classified examples. An experiment comparing top-down guided learning algorithms (G & T and ID3) with perceptron, pocket, and back propagation multi-layer neural learning algorithms is an approved Scottish Bank credit card whose decision-making process is primarily credit. Implemented using a set of applications. Rating system. Overall, they all work with the same level of classification accuracy, but training the neural algorithm takes much longer. This white paper describes the motivations for using machine learning algorithms for credit card scoring, details the algorithms, and compares the performance of these algorithms in terms of accuracy.

1.2 LITERATURE SURVEY

1.2.1 PUBLICATIONS:

1. **Title: Credit scoring using the hybrid neural discriminant technique.**
   Source:
   Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications, 23(3), 245–254.
   Summary:
   Credit scoring has become a very important task as the credit industry has experienced double-digit growth over the last few decades. Artificial neural networks are becoming a very popular alternative to credit scoring models due to their associated memory properties and generalisability. However, decisions about network topology, the importance of potential input variables, and the lengthy training process have long been criticised, limiting their application to the handling of credit scoring issues. The purpose of the proposed study is to investigate the performance of credit scoring by integrating neural networks and back propagation into the traditional approach of discriminate analysis. Including the credit score results from the discriminate analysis simplifies the network structure and improves the accuracy of the credit score of the designed neural network model. Credit score against the bank's credit card dataset. The task will be executed. As the results show, the proposed hybrid approach converges much faster than traditional neural network models. In addition, the accuracy of the credit score associated with the proposed methodology has improved, which is superior to traditional discriminant analysis and logistic regression approaches.

2. **Title: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.**
   Source:
   Lee, Y. S., Yen, S. J., Lin, C. H., Tseng, Y. N., Ma, L. Y. (2004). A data mining approach to constructing probability of default scoring model. In Proceedings of 10th conference on information management and implementation (pp. 1799–1813).
   Summary:
   This study covers the default cases of customer payments in Taiwan and compares the prediction accuracy of default probabilities among six data mining methods. From a risk management perspective, the default estimation probability prediction accuracy results are more valuable than the binary results that classify trusted or untrusted customers. Since the actual failure probability is unknown, this study presented a new "sort smoothing method" to estimate the actual failure probability. Using the actual failure probability (Y) as the response variable and the predicted failure probability (X) as the independent variable, the results of simple linear regression (Y = A + BX) show an artificially created prediction model. I am. The coefficient of determination of the neural network is the highest. Its regression intercept (A) is close to zero and its regression coefficient (B) is close to 1. Therefore, of the six data mining techniques, only artificial neural networks can accurately estimate the probability of actual failure.

3. **Title: Using neural network rule extraction and decision tables for credit-risk evaluation.**
   Source: Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science, 49(3), 312–329.
   Summary:
   Due to their universal approximation properties, neural networks have recently received considerable interest in the development of credit risk assessment models. However, most of this work is primarily focused on developing predictive networks without trying to explain how classification is done. In application domains such as credit risk assessment, it is imperative that credit risk managers have a compact and easy-to-understand set of rules. In this article, we evaluated and contrasted three neural network rule extraction methods (Neurorule, Trepan, and Nefclass) for credit risk assessment. The propositional rules derived from Neurorule were particularly concise and very easy to understand. We also explained how to represent the rules extracted using DT. DT displays rules in an intuitive graphic format that is easy for human experts to see. In addition, it enables simple and user-friendly advice in daily work. The rules and tree DTs extracted by Neurorule and Trepan have shown to be compact and powerful. In conclusion, extracting rules from neural networks and DTs is an effective and powerful management tool that can create sophisticated and user-friendly decision support systems for credit risk assessment. In addition, it would be interesting to apply the proposed approach to other interesting issues in business science. B. Forecast churn, customer retention and bankruptcy.

4. **Title: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.**
   Source:
   Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635.
   Summary:
   Interest in predicting scorecard development is uninterrupted. However, there have been some advances, such as new learning methods, performance measurements, and techniques for reliably comparing different classifiers that are not considered in the credit assessment literature. To fill the gap in these studies, we will update the study by Baesens et al. Compare some new classification algorithms with the state of your credit score. In addition, we will investigate how the evaluation of alternative scorecards differs between established and new indicators of predictive accuracy. Finally, find out if a more accurate classifier makes sense for management. Our research provides professionals and scholars with valuable insights into credit scoring. This helps practitioners keep up with the technological advances in predictive modelling. From an academic point of view, this study provides an independent assessment of modern scoring methods and provides a new basis for comparing future approaches.

5. **Title: The art and science of customer relationship management.**
   Source:
   Berry, M., & Linoff, G. (2000). Mastering data mining: The art and science of customer relationship management. New York: John Wiley & Sons, Inc.
   Summary:
   They cover more advanced topics such as preparing data for analysis and creating the necessary infrastructure for data mining at your company. Features significant updates since the previous edition and updates you on best practices for using data mining methods and techniques for solving common business problems. Covers a new data mining technique in every chapter along with clear, concise explanations on how to apply each technique immediately. Touches on core data mining techniques, including decision trees, neural networks, collaborative filtering, association rules, link analysis, survival analysis, and more. Provides best practices for performing data mining using simple tools such as Excel.

1.2.2 APPLICATION PAST AND UNDERGOING RESEARCH:

6. **Title: Cash and credit card crisis in Taiwan.**
   Source: Chou, M. (2006). Cash and credit card crisis in Taiwan. Business Weekly, 24–27.
   Summary:
   In the past, banks issued credit and debit cards to students even if they were unemployed. As a result, many banks sent credit card vendors to college campuses, persuaded students to apply for credit and debit cards, and seduced them at low interest rates. Due to the lack of work experience and financial skills of these students, young people did not know how much interest they would have to pay by using credit and debit cards vigorously. The main function of AMC is to accept bad debts from banks. By taking on some of the bad debts, AMC helps manage the risk of financial activity and improve the bank's balance sheet. In 2000, the Taiwan Parliament passed the Financial Institution Merger Act. Under the law, foreign companies such as Lone Star, Merrill Lynch and Lehman Brothers are allowed to fund AMC. At the same time, many Taiwanese banks jointly funded several AMCs.

7. **Title: Analysis of financial credit risk using machine learning**
   Source:
   https://arxiv.org/ftp/arxiv/papers/1802/1802.05326.pdf
   Summary:
   Bankruptcy can have a devastating effect on the economy. Bankruptcy of multinational corporations can disrupt the global financial ecosystem as more companies expand abroad and harness foreign resources. Recent advances in communications and information technology have made it increasingly difficult to collect and store business-related data. Using published datasets, we applied a variety of machine learning techniques to determine the relationship between the company's current state and its near-future fate. The results show that predictions with an accuracy of over 95% can be

achieved with any machine learning technique when using useful features such as expert scoring. However, the correlation is not very strong when using pure financial factors to predict whether a company will go bankrupt. More features are needed to better explain the data, but this is a higher dimension where data from thousands of public companies is not enough to fill this space with sufficient density. Leads to problems. Due to this "curse of dimensionality", flexible nonlinear models tend to overfit the training sample and therefore cannot be generalised to invisible data. For high-dimensional Polish bankruptcy datasets, simpler models such as logistic regression can predict a company's bankruptcy one year later with an accuracy of 66.4.

8. **Title: The Importance of Credit Risk Management in Banking**
   Source:
   https://blog.crifhighmark.com/the-importance-of-credit-risk-management-in-banking/#:~:text=They%20need%20to%20manage%20their,reserves%20at%20any%20given%20time
   Summary:
   Following were the advantages of credit card risk management learnt from the article. It helps in predicting and/ or measuring the risk factor of any transaction. t helps in planning ahead with strategies to tackle a negative outcome. It helps in setting up credit models which can act as a valuable tool to determine the level of risk while lending.

9. **Title: The importance of machine leaning in risk management.**
   Source:
   https://www.cqf.com/blog/importance-of-machine-learning-for-risk-management#:~:text=One%20prominent%20use%20case%20for,is%20fraud%20detection%20and%20prevention.&text=As%20in%20the%20case%20of,questionable%20behavior%20has%20been%20discerned
   Summary:
   For those seeking a broad and deep background in quant finance, from models and methods to machine learning in today's financial and economic environment, the CQF program is timely, flexible and high quality with a focus on wealth management and risk management. Provides professional training in data science and machine learning. For those on this journey, CQF provides the foundation for pursuing the most interesting and challenging opportunities in the current and future quantitative financial industry.

10. **Title: Credit Card Risk Assessment Based on Machine Learning.**
    Source:
    https://www.researchgate.net/publication/333871592_Credit_Card_Risk_Assessment_Based_on_Machine_Learning
    Summary:
    This article provides an example of a construction bench in Beijing. SSMTE is used to oversample the data. Handling outliers and missing values. And I standardized the variables. Make sure that the range of values for is within the same range of Finally, logistic regression and GridSearchCV regression are set up. The simulation also compares the search rates of logistic regression and GridSearchCV regression. The

feasibility and effectiveness of the logistic algorithm is checked. Since this paper studies a bank credit card model assessment . The ultimate meaning is to determine whether the user will overdue or not repay. So identifying a positive sample is crucial. So the bigger the number of positive samples is the better. The over sampled Logistic Regression model is superior to the GridSearchCV model before oversampling. A credit card risks assessment model based on logistic regression, which has good reliability and validity.User non-default rate can be predicted based on user data.It is fully stated that logistic regression can be applied to the construction of credit card risk assessment models.

greatlearning
Learning for Life

CHAPTER - 2

DATASET AND DOMAIN

Dataset: Default of credit card clients
Domain: Finance
Aim :
This research aimed at the case of customers default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods.

2.1 DATA DICTIONARY:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:
X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
X6: the repayment status in September, 2005.
X7: the repayment status in August, 2005;
X8: the repayment status in July , 2005;
X9: the repayment status in June, 2005;
X10: the repayment status in May, 2005;
X11:the repayment status in April, 2005.
The measurement scale for the repayment status is:
-1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; 3 = payment delay for three months. 4 = payment delay for four months; 5 = payment delay for five months; 6 = payment delay for six months; 7= payment delay for seven months ;8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar).
X12: amount of bill statement in September, 2005;
X13: amount of bill statement in August, 2005;
X14: amount of bill statement in July, 2005;
X15: amount of bill statement in June, 2005;
X16: amount of bill statement in May, 2005;
X17: amount of bill statement in April, 2005;
X18-X23: Amount of previous payment (NT dollar).
X18: amount paid in September, 2005;
X19 :amount paid in August, 2005;
X20: amount paid in July, 2005;

X21: amount paid in June, 2005;
X22: amount paid in May, 2005;
X23: amount paid in April, 2005.

2.2 DATA SET INFORMATION:

This research aimed at the case of customers default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, this study presented the novel Sorting Smoothing Method to estimate the real probability of default. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

2.3 VARIABLE CATEGORIZATION (COUNT OF NUMERIC AND CATEGORICAL):

A variable can be defined as something that is subject to change. There are two types of variables, namely:

- Numeric
- Categorical

Numeric variables:
The numbers of numeric features are: 14

X1,X5,X12,X13,X14,X15,X16,X17,X18,X19,X20,X21,X22,X23

Categorical variables:
The numbers of categorical features are: 9
The numerical features are:

X2,X3,X4,X6,X7X8,X9,X10,X11

2.4 PRE PROCESSING DATA ANALYSIS:

The process of converting raw data into a comprehensible format is known as data preparation. Before using machine learning or data mining methods, make sure the data is of good quality.

1. **Info function:**
   Info function is used to understand the non-null value count and the dtype of the data.
   Inference:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30000 entries, 1 to 30000
Data columns (total 24 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   X1      30000 non-null  object
 1   X2      30000 non-null  object
 2   X3      30000 non-null  object
 3   X4      30000 non-null  object
 4   X5      30000 non-null  object
 5   X6      30000 non-null  object
 6   X7      30000 non-null  object
 7   X8      30000 non-null  object
 8   X9      30000 non-null  object
 9   X10     30000 non-null  object
 10  X11     30000 non-null  object
 11  X12     30000 non-null  object
 12  X13     30000 non-null  object
 13  X14     30000 non-null  object
 14  X15     30000 non-null  object
 15  X16     30000 non-null  object
 16  X17     30000 non-null  object
 17  X18     30000 non-null  object
 18  X19     30000 non-null  object
 19  X20     30000 non-null  object
 20  X21     30000 non-null  object
 21  X22     30000 non-null  object
 22  X23     30000 non-null  object
 23  Y       30000 non-null  object
dtypes: object(24)
memory usage: 5.7+ MB
```

Fig:2.1 Jupyter notebook info function

Inferences:
Total number of columns in the dataset : 24
Number of dependent variables: 23
No of independent values : 1
Dtype: Object (conversion of dtype required)

2. **Redundant :**

Study performed to understand the redundant features present in each columns

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

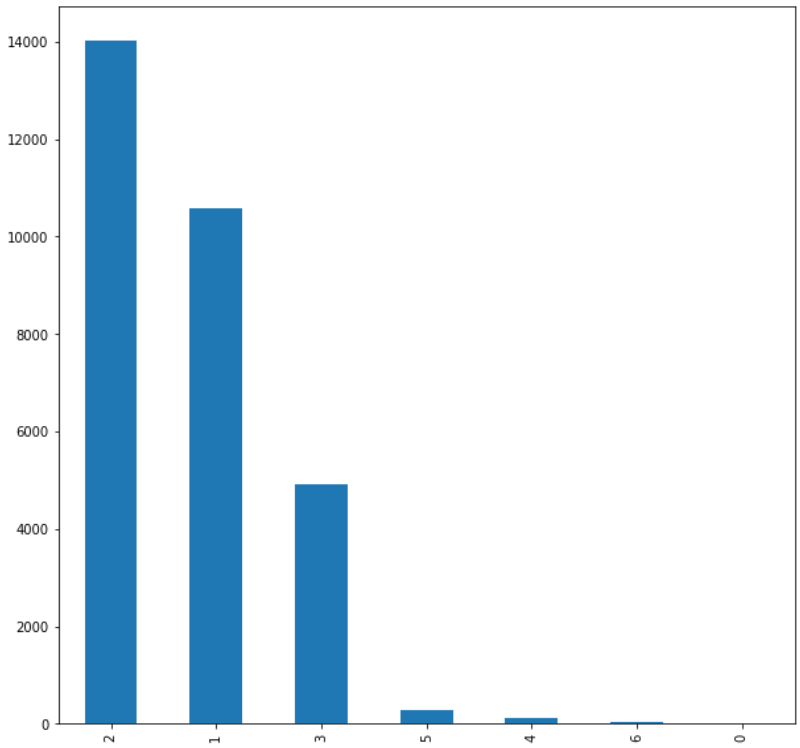

Fig:2.2 X3 Bar plot

Inferences: This bar chart states that 5, 6, 0 can be removed in the column.

X4: Marital status (1 = married; 2 = single; 3 = others).

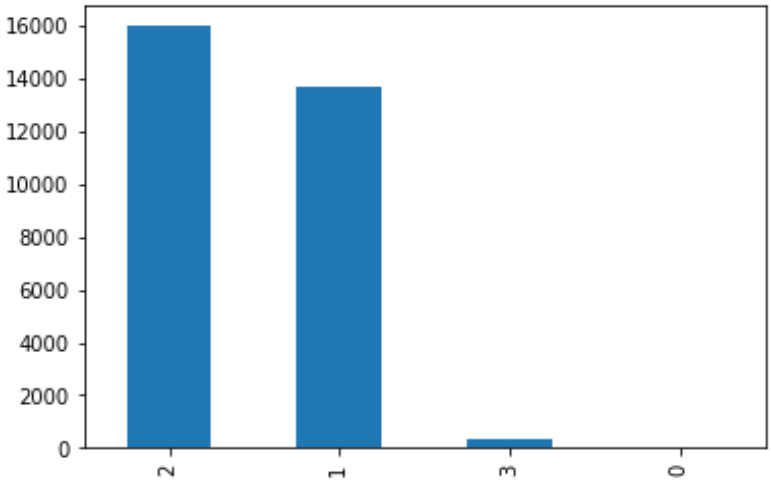

Fig:2.3 X4 Bar plot

Inferences: This bar chart states that 0 can be removed in the column.

2.5 PROJECT JUSTIFICATION

Aim :

This research aimed at the case of customers default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods.

Complexity involved:

- Fixing outlier,
- understanding the nature of the clients,
- To avoid false clients detections.

Outcomes:

- This will be a business value problem to avoid the loss for the client.
- The significant factors affecting the failure of repayment of credit card payment is analyzed based on the dataset and is used to create a model to understand the probable number of defaulters.
- This will also help the card issuer to decrease the credit limit amount , hence reduce the credit risk in the coming months.

CHAPTER – 3
DATA EXPLORATION – EDA

3.1 RELATIONSHIP BETWEEN VARIABLES

The dataset consists of numerical variables which are significant in determining the default credit card customers. Hence, Bivariate technique is deployed to see the relationship between numerical verses categorical variable as well as numerical v/s numerical.

3.1.1 Bivariate Analysis:

Bivariate analysis is one of the statistical analyses where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes that occurred between the two variables and to what extent.

1. **CREDIT LIMIT WITH SEX**

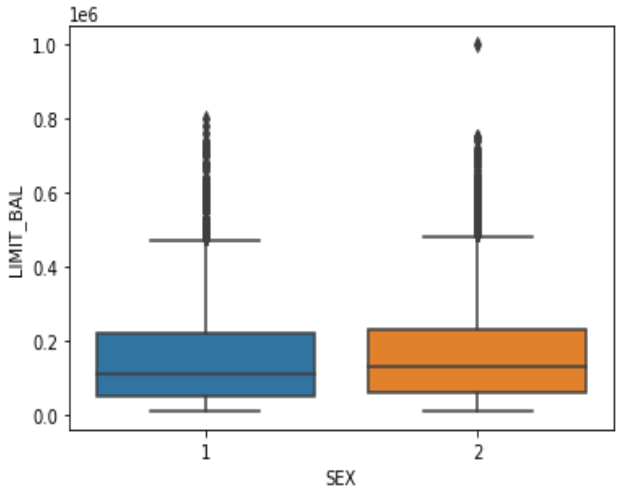

Fig:3.1 Box plot of credit limit vs. sex

Inferences: Credit Limit is evenly distributed among males and females.

2. **Distribution of people using credit card facility:**

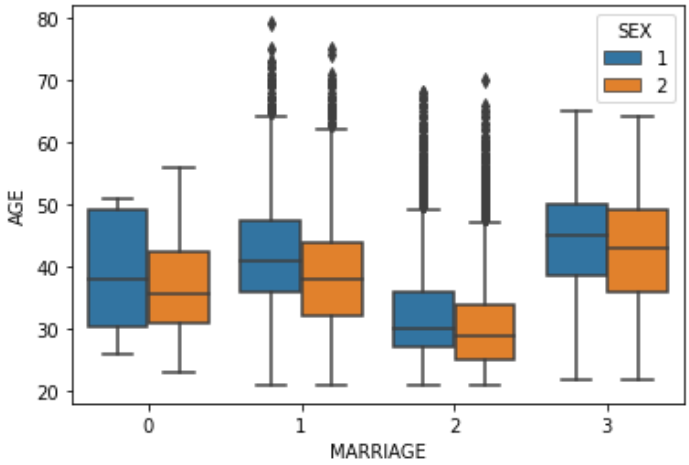

Fig:3.2 Boxplot of Distribution of people VS credit card facility

Inferences: The dataset mostly contains couples in their mid-30s to mid-40s and single people in their mid-20s to early-30s.

3. **The maximum limit of Credit card limit amount:**

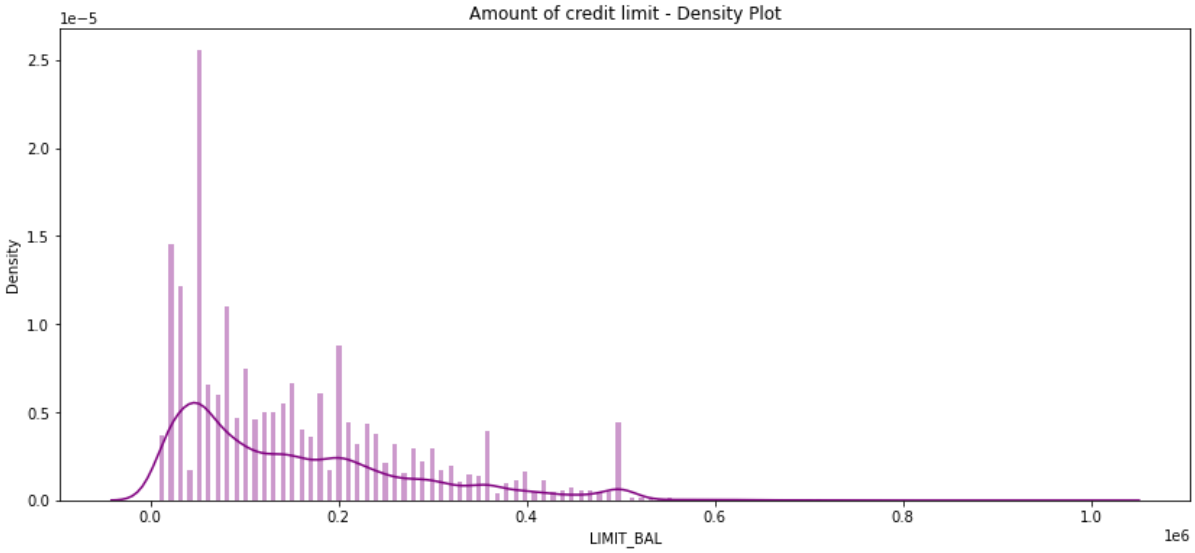

Fig: 3.3 Density plot of maximum limit of Credit card limit amount

Inferences : The largest number of credit cards are with limit of 50,000

4. BILL AMOUNT Vs DEFAULT CREDIT CARD CUSTOMERS:

We use group by function to analyse the percentage of default customers with respect to each bill amount in the month of September, August, July, June, May, April respectively.
Relationship between September bill amount and default credit card customer

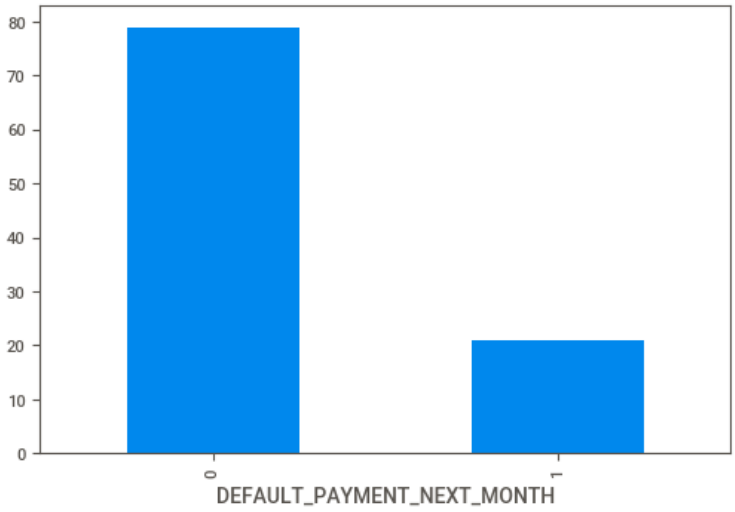

***Fig: 3.4 Bar plot of* bill amount vs. default credit card customers**

DEFAULT_PAYMENT_NEXT_MONTH
0 79.246859%
1 20.753141%

Inferences: From this, we can infer that nearly 79.24 percent of the people will not be in default for the next month and 20.75 percent has the chance of becoming default in the coming month.

5. Relationship between August bill amount and default credit card customers.

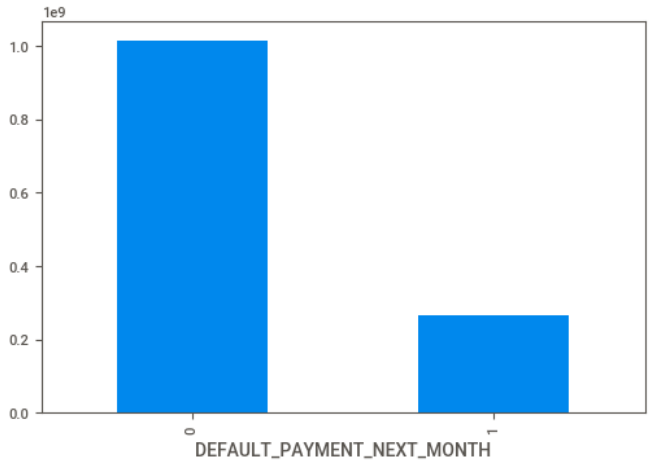

***Fig: 3.5 Bar plot of August* bill amount vs. default credit card customers**

DEFAULT_PAYMENT_NEXT_MONTH
0 78.962973
1 21.037027

Inferences: From this, we can infer that nearly 78.96 percent of the people will not be in default for the next month and 21.03 percent has the chance of becoming default in the coming month.

6. **Relationship between July bill amount and default credit card customers.**

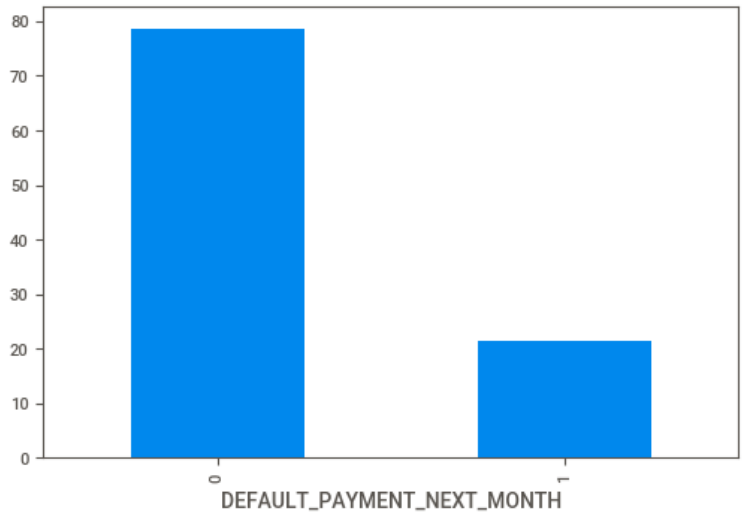

***Fig: 3.6 Bar plot of July* bill amount vs. default credit card customers**

DEFAULT_PAYMENT_NEXT_MONTH
0 78.873112
1 21.126888

Inferences: From this, we can infer that nearly 78.87 percent of the people will not be in default for the next month and 21.12 percent has the chance of becoming default in the coming month.

7. **Relationship between June bill amount and default credit card customers.**

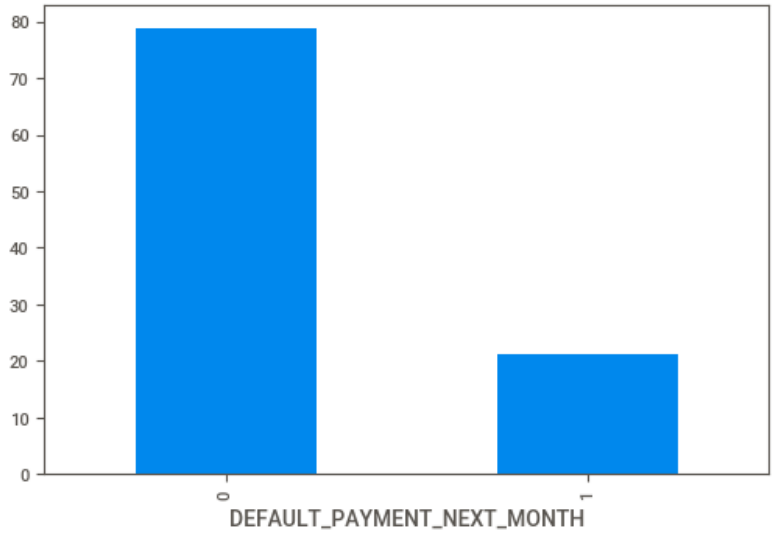

***Fig: 3.7 Bar plot of June* bill amount vs. default credit card customers**

DEFAULT_PAYMENT_NEXT_MONTH
0 78.661353
1 21.338647

Inferences : From this, we can infer that nearly 78.66 percent of the people will not be in default for the next month and 21.33 percent has the chance of becoming default in the coming month.

8. **Relationship between May bill amount and default credit card customers.**

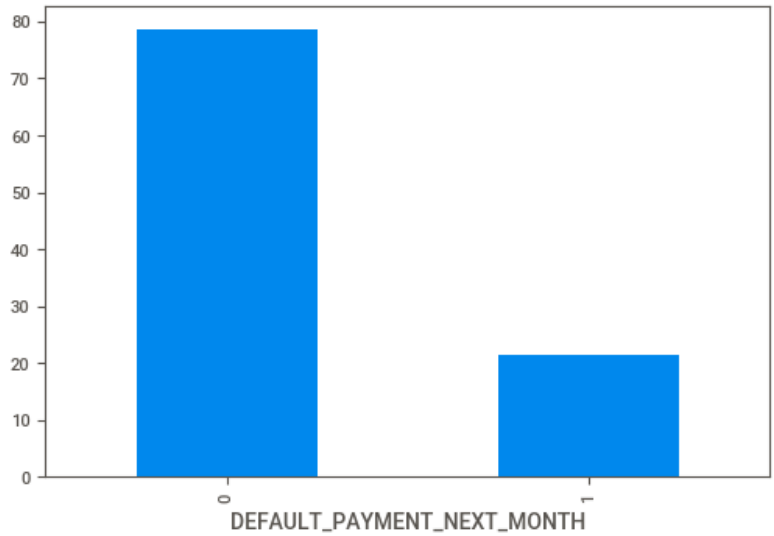

***Fig: 3.8 Bar plot of May* bill amount vs. default credit card customers**

DEFAULT_PAYMENT_NEXT_MONTH
0 78.5252
1 21.4748

Inferences : From this, we can infer that nearly 78.52 percent of the people will not be in default for the next month and 21.47percent has the chance of becoming default in the coming month.

9. **Relationship between April bill amount and default credit card customers.**

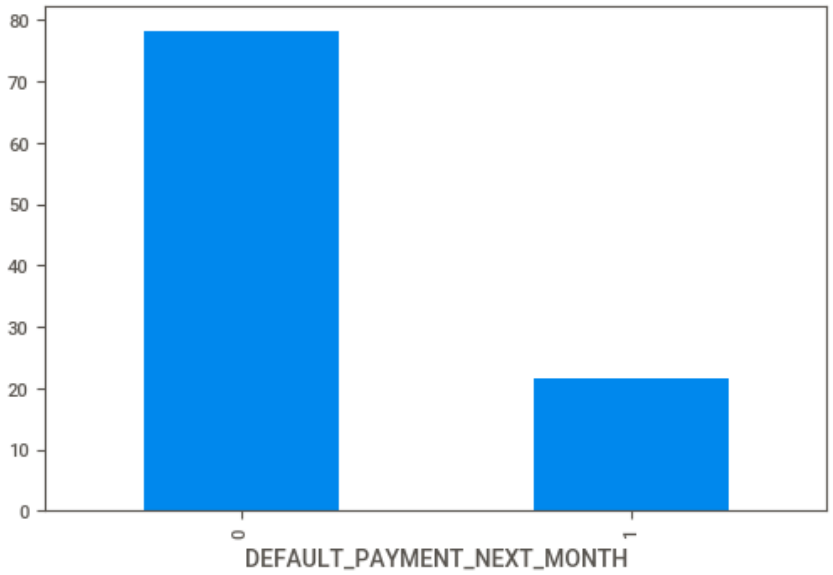

***Fig: 3.9 Bar plot of April* bill amount vs. default credit card customers**

DEFAULT_PAYMENT_NEXT_MONTH
0 78.281992
1 21.718008

Inferences : From this, we can infer that nearly 78.28 percent of the people will not be in default for the next month and 21.71 percent has the chance of becoming default in the coming month.

3.1.2 MULTIVARIENT ANALYSIS

10. Relationship between numerical variables using Correlation Matrix

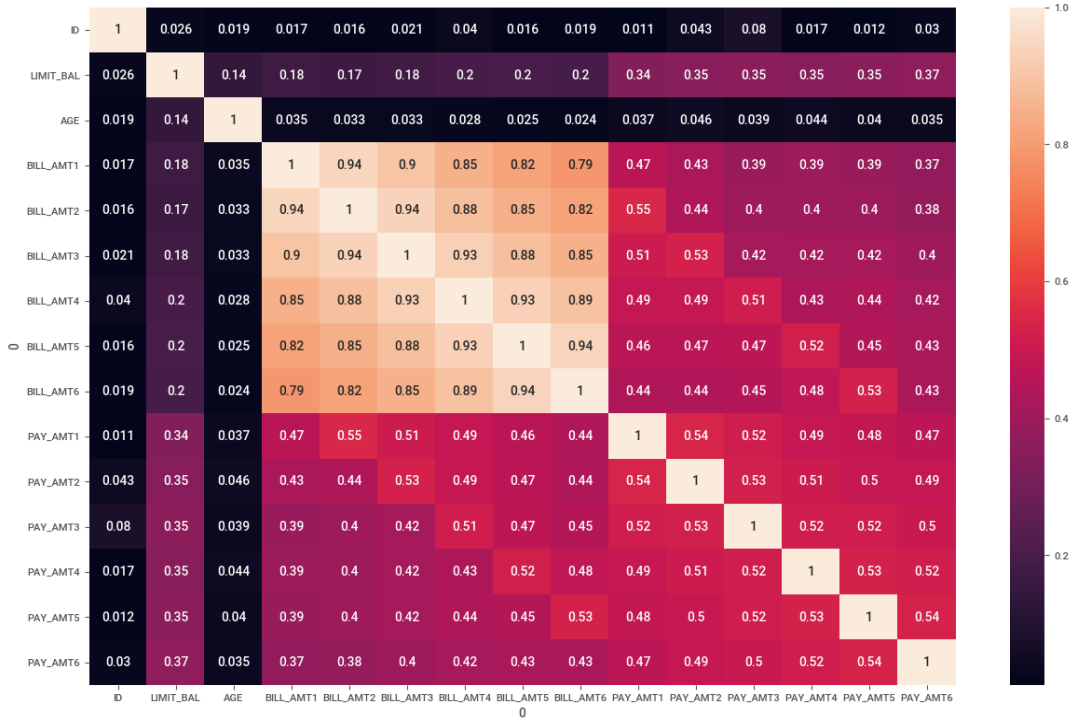

Fig: 3.10 Heat map for all independent variables

Inferences : Correlation is decreasing with distance between months.

11. Presence of outliers and its treatment :

Outliers are unusual data points in the dataset and they can distort the statistical analysis. But all outliers cannot be removed as such because it can be significant. We have outliers in the aged, limit amount, bill amount, pay amount column and in fact it can happen as a normal part of the process and with respect to the domain. In the dataset, outlier is not part of a error measurement, hence we are not removing the outliers

The number of outliers in ID is 0
The number of outliers in LIMIT BAL is 768
The number of outliers in AGE is 208
The number of outliers in BILL_AMT1 is 1683
The number of outliers in BILL_AMT2 is 1680
The number of outliers in BILL_AMT3 is 1710
The number of outliers in BILL_AMT4 is 1742
The number of outliers in BILL_AMT5 is 1769
The number of outliers in BILL_AMT6 is 1835
The number of outliers in PAY_AMT1 is 2087

The number of outliers in PAY_AMT2 is 2254
The number of outliers in PAY_AMT3 is 1972
The number of outliers in PAY_AMT4 is 1835
The number of outliers in PAY_AMT5 is 1841
The number of outliers in PAY_AMT6 is 1862

12. Statistical Significance of Variables

● Surprisingly, determining which variable is the most important is more complicated than it first appears.

● Statistical significance tests are designed to address this problem and quantify the likelihood of the samples of skill scores being observed given the assumption that they were drawn from the same distribution.

● If this assumption, or null hypothesis, is rejected, it suggests that the difference in skill scores is statistically significant.

● If the p-value for a variable is less than the significance level, the sample data provide enough evidence to reject the null hypothesis for the entire population. Data will favour the hypothesis that there is a non-zero correlation.

● Changes in the independent variable are associated with changes in the dependent variable at the population level. This variable is statistically significant and probably a worthwhile addition to the model.

● On the other hand, a p-value that is greater than the significance level indicates that there is insufficient evidence in the sample to conclude that a non-zero correlation exists.

3.3.1 Chi-Square Test for Independence:

Chi-Square Test for Independence is done to evaluate the statistical significance of categorical variables against the target/dependent variable.

Hypothesis Formation:

Null Hypothesis (Ho): Sex and default payment next month are independent
Alternate Hypothesis (Ha): Sex and default payment next month are dependent

Statistical Significance of relationship between sex and default payment next month:

Test Statistics: 47.70879689062111

pValue: 4.944678999412044e-12

Degrees of freedom: 1

Hypothesis Formation:

Null Hypothesis (Ho): Education and default payment next month are independent
Alternate Hypothesis (Ha): Education and default payment next month are dependent

Statistical significance of relationship between education and default payment next month:

Test Statistics: 163.21655786997073

pValue: 1.2332626245415605e-32

Degrees of freedom: 6

Hypothesis Formation:
Null Hypothesis (Ho): Marriage and default payment next month are independent
Alternate Hypothesis (Ha): Marriage and default payment next month are dependent

Statistical significance of relationship between marriage and default payment next month:

Test Statistics: 35.66239583433609

pValue: 8.825862457577375e-08

Degrees of freedom: 3

Hypothesis Formation:
Null Hypothesis (Ho): Pay 1 and default payment next month are independent.
Alternate Hypothesis (Ha): Pay 1 and default payment next month are dependent.

Statistical significance of relationship between pay 1 and default payment next month:

Test Statistics: 5365.964977413581
pValue: 0.0
Degrees of freedom: 10

Hypothesis Formation:
Null Hypothesis (Ho): Pay 2 and default payment next month are independent
Alternate Hypothesis (Ha): Pay 2 and default payment next month are dependent

Statistical significance of relationship between pay 2 and default payment next month:

Test Statistics: 3474.4667904168564

pValue: 0.0

Degrees of freedom: 10

Hypothesis Formation:
Null Hypothesis (Ho): Pay 3 and default payment next month are independent.
Alternate Hypothesis (Ha): Pay 3 and default payment next month are dependent.

Statistical significance of relationship between pay 3 and default payment next month:

Test Statistics: 2622.4621276828025

pValue: 0.0

Degrees of freedom: 10

Hypothesis Formation:
Null Hypothesis (Ho): Pay 4 and default payment next month are independent.
Alternate Hypothesis (Ha): Pay 4 and default payment next month are dependent.

Statistical significance of relationship between pay 4 and default payment next month:

Test Statistics: 2341.469945438205

pValue: 0.0

Degrees of freedom: 10

Hypothesis Formation:
Null Hypothesis (Ho): Pay 5 and default payment next month are independent.
Alternate Hypothesis (Ha): Pay 5 and default payment next month are dependent.

Statistical significance of relationship between pay 5 and default payment next month:

Test Statistics: 2197.694900930992

pValue: 0.0

Degrees of freedom: 9

Hypothesis Formation:
Null Hypothesis (Ho): Pay 6 and default payment next month are independent.
Alternate Hypothesis (Ha): Pay 6 and default payment next month are dependent.

Statistical significance of relationship between pay 6 and default payment next month:

Test Statistics: 1886.835309001187

pValue: 0.0

Degrees of freedom: 9

Inferences: From the above results, we see that all the p value is less than the significance level which shows that the alternate hypothesis is selected and we can say that the target variable is dependent on all the categorical variable

Class imbalance:

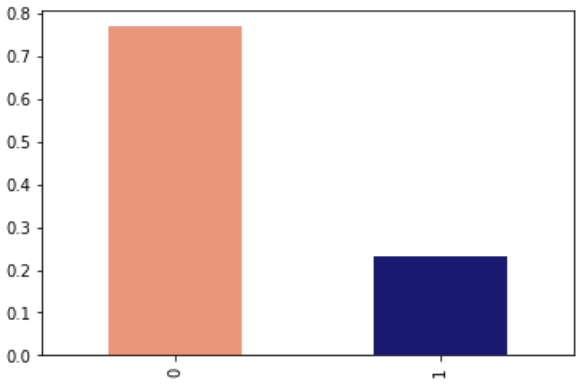

Fig: 3.11 Bar plot for Y.

0 76.863
1 23.137

Inferences: There is imbalance in the target variable and but since it's not that huge, there is no need to treat the same.

3.3.2 NON-PARAMETRIC TEST (KRUSKAL TEST)

Hypothesis Formation:
Null Hypothesis: All medians are equal.
Alternate Hypothesis (Ha): At least one median is different.

```
[ ]  # Hypothesis for Kruskal:

     # Ho: All medians are equal
     # Ha: Atleast one median is different
```

```
stats.kruskal(a0,a1,b0,b1,c0,c1,d0,d1,e0,e1,f0,f1,g0,g1,h0,h1,i0,i1,j0,j1,k0,k1,l0,l1,m0,m1,n0,n1)
```

```
KruskalResult(statistic=157311.5087661885, pvalue=0.0)
```

Inference:

```
==> pvalue of Kruskal Result for scores of different adverse effects < 0.05 (sig. lvl)

==> Hence, Ho is rejected and all medians are not equal.
```

Fig: 3.12 Non-parametric test in jyupter notebook.

Inference : pvalue of Kruskal Result for scores of different adverse effects < 0.05 (sig. lvl) . Hence, Ho is rejected and all medians are not equal.

CHAPTER - 4

FEATURE ENGINEERING

4.1 FEATURE SELECTION:

- Since 'ID' column does not affect our model, it is dropped for feature engineering.
- The most important features are PAY_1_2, PAY_AMT1, BILL_AMT1, BILL_AMT2, BILL_AMT6, BILL_AMT3

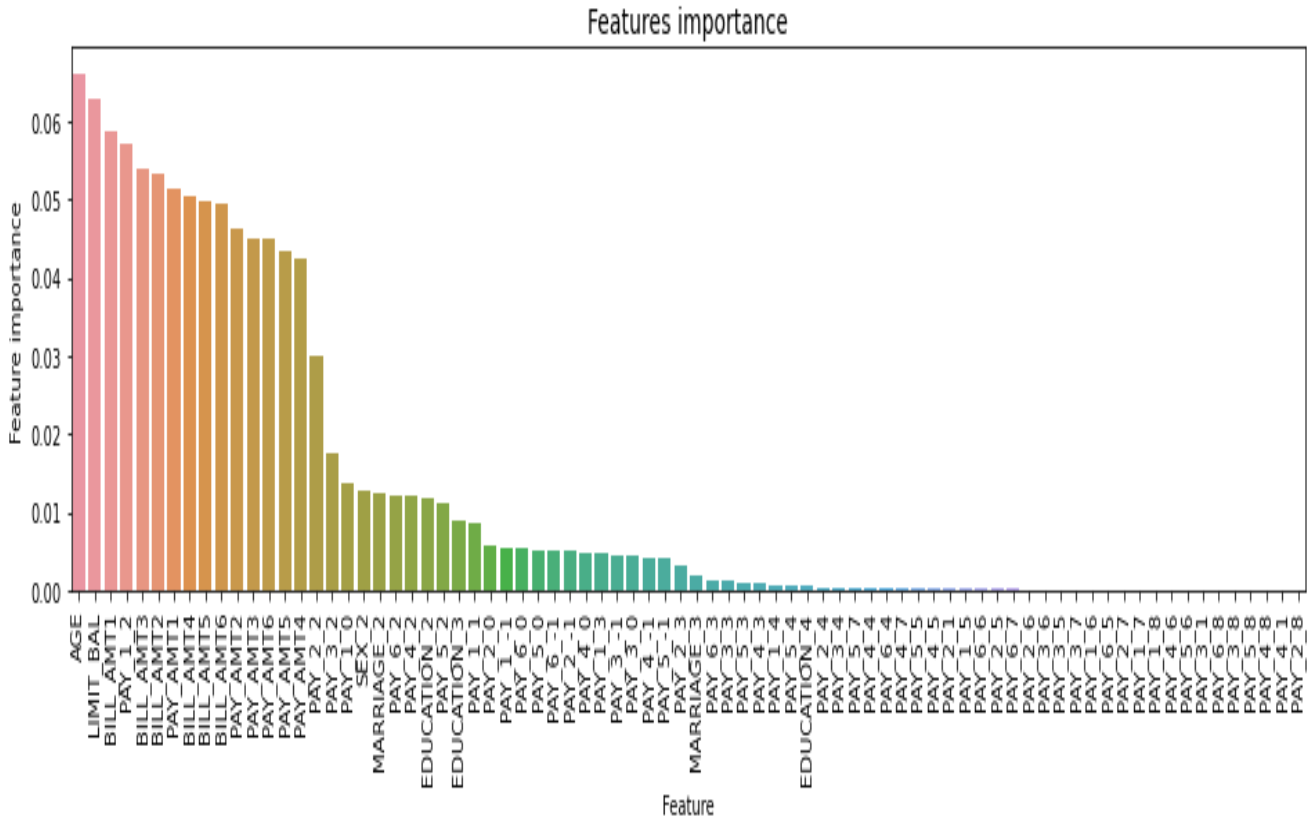

Fig: 4.1 Feature selection in jyupter notebook.

CHAPTER - 5

CLASSIFICATION

5.1 BASE MODEL

- Random forest method has been followed.
- As the name implies, a random forest is made up of a huge number of individual decision trees that work together as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model.
- Any of the individual constituent models will outperform a large number of reasonably uncorrelated models (trees) working as a committee.
- When creating each individual tree, it employs bagging and feature randomization in order to generate an uncorrelated forest of trees whose committee prediction is more accurate than that of any one tree.
- We need features that have at least some predictive power. After all, if we put garbage in then we will get garbage out.
- The trees of the forest and more importantly their predictions need to be uncorrelated (or at least have low correlations with each other). While the algorithm itself via feature randomness tries to engineer these low correlations for us, the features we select and the hyper-parameters we choose will impact the ultimate correlations as well.

5.2 CLASSIFICATION REPORT

- To get even more insight into model performance, we should examine other metrics like precision, recall, and F1 score.

```
========================= Confusion Matrix =========================

[[6485  389]
 [1250  757]]

========================= Classification Report =========================

              precision    recall  f1-score   support

           0       0.84      0.94      0.89      6874
           1       0.66      0.38      0.48      2007

    accuracy                           0.82      8881
   macro avg       0.75      0.66      0.68      8881
weighted avg       0.80      0.82      0.80      8881


========================= All AUC Scores =========================

[0.74045846 0.74086576 0.75046212 0.73368445 0.76223465 0.78360639
 0.79499045 0.76866535 0.78300132 0.785778  ]

========================= Mean AUC Score =========================

Mean AUC Score - Random Forest: 0.7643746935277562
```

Fig: 5.1 Classification report in jyupter notebook.

5.3 CONFUSION MATRIX:

A confusion matrix is a table that shows how well a classification model (or "classifier") performs on a set of test data for which the true values are known.

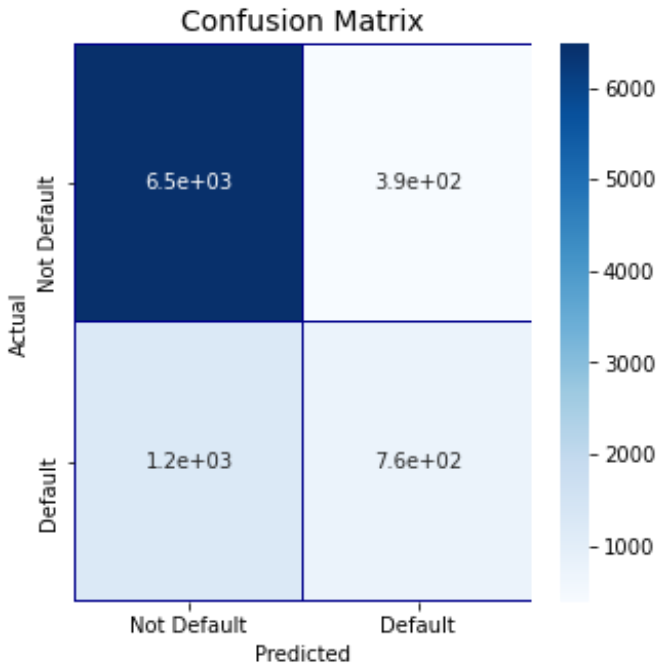

Fig: 5.2 Confusion Matrix

Greetings Dear Moderator ,

Our Mentor has given us the changes us to do which we will incorporate with and resubmit it later.

With Regards

SHIBIKANNAN .T . M
TEAM LEADER