**BIA 652**

**FINAL PROJECT PROPOSAL**

|  |  |
| --- | --- |
| **NAME** | **CWID** |
| **DISHANT NAIK** | **10454341** |
| **VIREN GHORI** | **10456691** |

**INTRODUCTION**

The advancement in technology has brought in convenience and made our life simpler. Today we are just one swipe away to make any kind of payment. Now a days we do not need to carry bunch of notes and coins whenever we go out for shopping. All you need is a thin plastic card to make a payment. After the invention of payment methods like Apple pay and Samsung pay you don’t even need to carry a card as well.

A credit card is a payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services based on the cardholder's promise to the card issuer to pay them for the amounts plus the other agreed charges.

While making it easy for us to do shopping while sitting on our couch, it has on the other hand made it easy for people to buy unnecessary items they don’t need or just buy things which they cannot repay. And this raises concerns of credit card holders defaulting to the credit card landers. So instead of dealing it after credit card holder default it is better to do some analysis and get alerted when credit card holder shows sign of defaulting in advance.

**Project statement:**

Defaulting credit card is one of the major concerns in finance world. Banks and other credit card landers always need to be alerted to detect such person and take appropriate actions against them to avoid defaulting. And that is what motivated us to build our project around it. In this project we will try to classify a credit card user as **“Credible clients”** or **“Not credible clients”.** To do this first we will closely analyze data in hand and will look for any strong or strange relationship between dependent and independent variables. Then we will use **“Logistic Regression”** to classify the dependent variable.

**Data set description:**

**Data set source:** <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

**About data set (Used for the project) source:**

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

**About Data set:**

This data set aimed at the case of customer default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

**Description of response variable Y:**

“default.payment.next.month” is the response variable for our analysis. Its datatype is “int64” and it only contains 0s and 1s. This variable indicates 1 when client is not credible and 0 when client is credible.

**Description of predictor variables X:**

There are in total 23 unique predictors. Description of each is as below.

1. LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit [Data type: float64]
2. SEX: Gender (1=male, 2=female) [Data type: int64]
3. EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown) [Data type: int64]
4. MARRIAGE: Marital status (1=married, 2=single, 3=others) [Data type: int64]
5. AGE: Age in years [Data type: int64]
6. PAY\_0: Repayment status in September 2005 (-1 = pay duly, 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) [Data type: int64]
7. PAY\_2: Repayment status in August 2005 (scale same as above) [Data type: int64]
8. PAY\_3: Repayment status in July 2005 (scale same as above) [Data type: int64]
9. PAY\_4: Repayment status in June 2005 (scale same as above) [Data type: int64]
10. PAY\_5: Repayment status in May 2005 (scale same as above) [Data type: int64]
11. PAY\_6: Repayment status in April 2005 (scale same as above) [Data type: int64]
12. BILL\_AMT1: Amount of bill statement in September 2005 (NT dollar) [Data type: float64]
13. BILL\_AMT2: Amount of bill statement in August 2005 (NT dollar) [Data type: float64]
14. BILL\_AMT3: Amount of bill statement in July 2005 (NT dollar) [Data type: float64]
15. BILL\_AMT4: Amount of bill statement in June 2005 (NT dollar) [Data type: float64]
16. BILL\_AMT5: Amount of bill statement in May 2005 (NT dollar) [Data type: float64]
17. BILL\_AMT6: Amount of bill statement in April 2005 (NT dollar) [Data type: float64]
18. PAY\_AMT1: Amount of previous payment in September 2005 (NT dollar) [Data type: float64]
19. PAY\_AMT2: Amount of previous payment in August 2005 (NT dollar) [Data type: float64]
20. PAY\_AMT3: Amount of previous payment in July 2005 (NT dollar) [Data type: float64]
21. PAY\_AMT4: Amount of previous payment in June 2005 (NT dollar) [Data type: float64]
22. PAY\_AMT5: Amount of previous payment in May 2005 (NT dollar) [Data type: float64]
23. PAY\_AMT6: Amount of previous payment in April 2005 (NT dollar) [Data type: float64]

**Data set summary statistics:**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Distribution of “Credible clients”** **and** **“Not credible clients” in Dataset:**

**Chart, bar chart

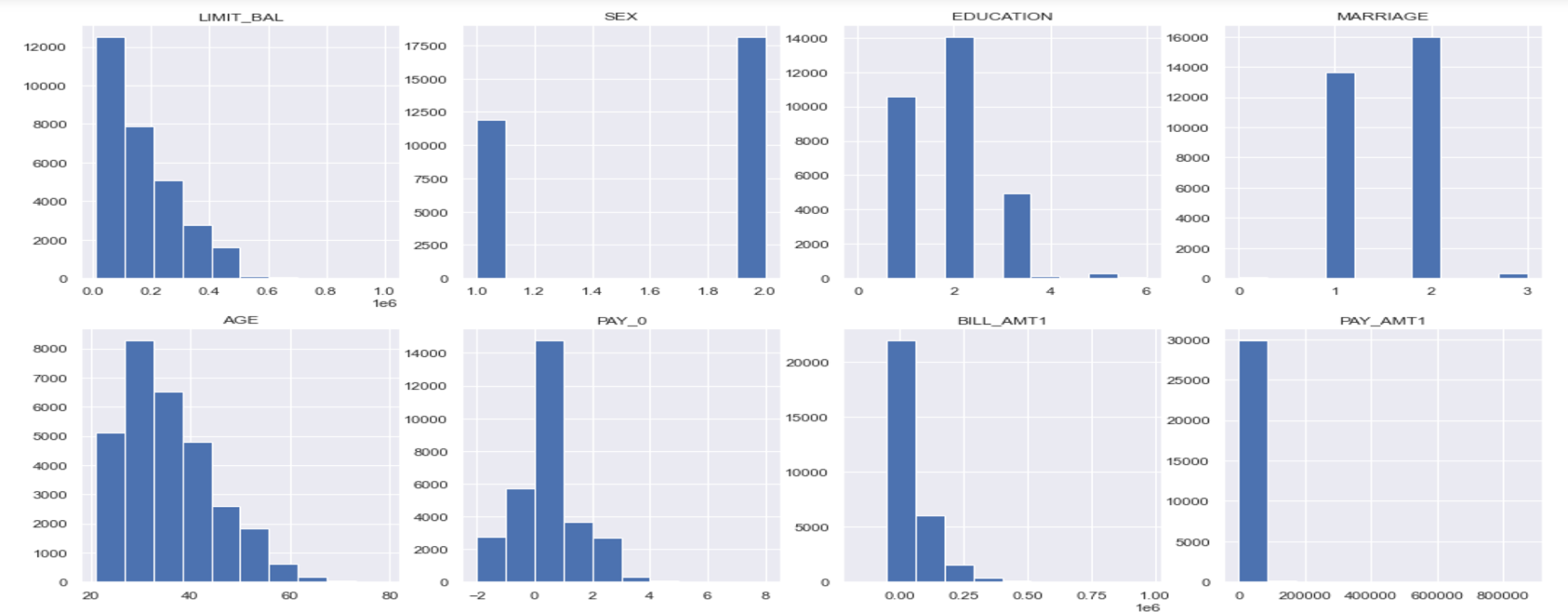
Description automatically generated**

**Correlation matrix of dataset:**

**Chart

Description automatically generatedHistograms of independent variables:**

**Note:** Not all independent variables are included in this plot.

****

**Regression method:**

Logistic regression is used when dependent variable is binary variable in nature or in simpler words when outcome of dependent variable is either success or failure. As dependent variable of this dataset is binary, we will be using Logistic regression. Logistic regression for 2 class classification can be represented by following formula.

]