DATA REPORT

(ABRACA-DATA)

Relevant Links:

[Powerpoint](https://docs.google.com/presentation/d/1jTs8l7sdQZoxH8AOs28x2fpAX99CwcxukW_NqclO--Y/edit?ts=601256d0#slide=id.gb96210f68b_0_101)

[Github](https://github.com/allanstar-byte/Default-of-Credit-Card-Clients..git)

[Rmd / Html / Dataset](https://drive.google.com/drive/folders/1GlCMa_aoLxXAtiyqfkQj-VVndzJx1_Hf?usp=sharing)

**1.BUSINESS UNDERSTANDING.**

**Business Overview**

The health of the credit card industry is best measured not by the number of people with cards, but rather the number who pay their bills. When the bank issues a credit card to a client the client agrees to certain terms such as minimum payment by the due date listed on the credit card statement.

Credit card default happens when a customer becomes severely delinquent on their credit card payments say if a client misses the minimum payment six months in a row, then the credit card will be in default and the bank will more than likely close the account.

For lenders, an increase in credit card delinquencies is an expensive proposition, since the more delinquent an account becomes, the smaller the chance it will be repaid at all. Statistically speaking, financial institutions will see repayment from only 20% of accounts that have been delinquent for more than 180 days, according to [The Credit Research Foundation](https://www.crfonline.org/).

These institutions have limited options for collecting on overdue payment and often end up incurring losses such as reduced reimbursement for losses and Lost interest and fee income.

Therefore predicting delinquencies is an important objective for these lending institutions.

**Business Objective**

To estimate the probability of default payment by a credit card client based on the historical data provided using supervised and unsupervised machine learning models.

**Business Success Criteria**

For this analysis to be successful, this research aimed at the case of customers’ default payments in Taiwan and compares the predictive accuracy of probability of default among supervised and unsupervised machine learning methods.

The evaluation metric for our Supervised problem was the confusion matrix .

**Assessing the Situation**

1. **Resource Inventory.**

* The Dataset: ((See [here](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients) ))

1. **Assumptions**

* The data was collected over a period of six months and is complete; it has no null values.

1. **Constraints**

* The data was collected on a consistent schedule and was therefore accurate.

**Data Mining Goals**

By carrying out this analysis we hope to:

* To estimate the probability of default payment by a credit card client based on the historical data provided using supervised and unsupervised machine learning models.

**Data Mining Success Criteria**

This research aimed at the case of customers’ default payments in Taiwan and compares the predictive accuracy of probability of default among supervised and unsupervised machine learning methods.

**2.DATA UNDERSTANDING**

The dataset used in this analysis was sourced from the Machine Learning Repository of University of California, Irvine. (See [here](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients) ). It was recorded on excel worksheets and has 30,000 rows and 24 columns.

**Data Description**

The datasets used in this analysis had the following attributes:

1. **Default payment** (Yes = 1, No = 0), a binary variable, as the response variable.
2. **X1: Amount of the given credit (NT dollar):** it includes both the individual consumer credit and his/her family (supplementary) credit.
3. **X2: Gender** (1 = male; 2 = female).
4. **X3: Education** (1 = graduate school; 2 = university; 3 = high school; 4 = others).
5. **X4: Marital status** (1 = married; 2 = single; 3 = others).
6. **X5: Age** (year).
7. **X6 - X11: History of past payment**.

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; . . .;
* 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; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

1. **X12-X17: Amount of bill statement** (NT dollar).

* X12 = amount of bill statement in September, 2005;
* X13 = amount of bill statement in August, 2005; . . .;
* X17 = amount of bill statement in April, 2005.

1. **X18-X23: Amount of previous payment** (NT dollar).

* X18 = amount paid in September, 2005;
* X19 = amount paid in August, 2005; . . .;
* X23 = amount paid in April, 2005.

**Verifying the Data Quality**

The data had no missing entries and no null entries.

There were 35 duplicated entries which were dropped.

**3.DATA PREPARATION**

**Loading the Libraries**

The first step in this process included loading our libraries into our R Markdown File. For this analysis we used the tidyverse and ggplot2 libraries.

**Loading the Datasets**

We then loaded our datasets on our environment and merged them. Next we previewed our dataset and accessed information on the data type of our respective columns.

**Data Cleaning.**

We cleaned our data in the following steps.

1. **Validity**

We checked for the relevance of the columns in our dataset.

The ID column from our dataset was found to be redundant and was dropped.

1. **Accuracy**

In checking for the accuracy of the dataset some of the values in our

1. **Completeness**

We checked for the presence of null values within our dataset.

The dataset did not have any null values.

1. **Consistency**

We checked for the presence of duplicated values in our dataset.

There were thirty five (35) duplicated entries which we dropped from the dataset.

1. **Uniformity**

Ultimately after undertaking the data cleaning the resultant columns used in our analysis were:

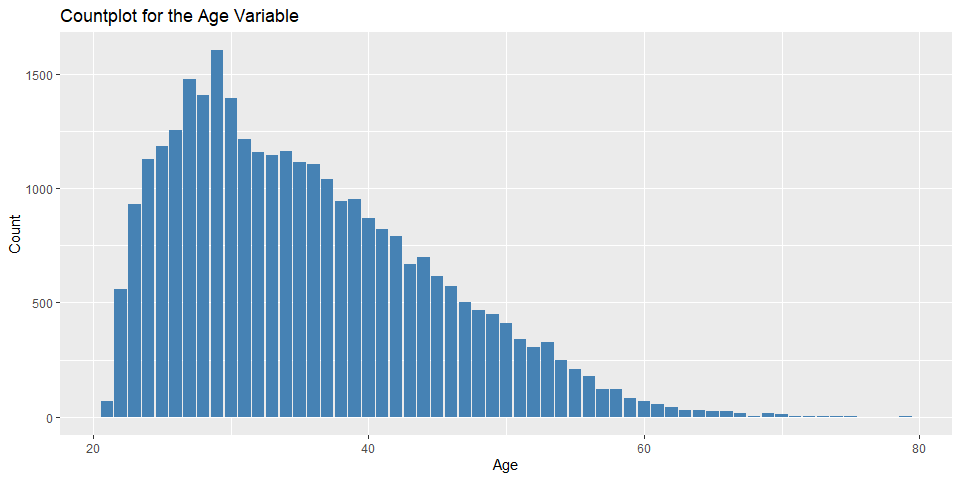
1. LIMIT\_BAL: Amount of given credit in NT dollars.
2. SEX: Gender (1=male, 2=female).
3. EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
4. MARRIAGE: Marital status (1=married, 2=single, 3=others)
5. AGE: Age in years
6. PAY\_0: Repayment status in September, 2005 (0=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)
7. PAY\_2: Repayment status in August, 2005 (scale same as above)
8. PAY\_3: Repayment status in July, 2005 (scale same as above)
9. PAY\_4: Repayment status in June, 2005 (scale same as above)
10. PAY\_5: Repayment status in May, 2005 (scale same as above)
11. PAY\_6: Repayment status in April, 2005 (scale same as above)
12. BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)
13. BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)
14. BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)
15. BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)
16. BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)
17. BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)
18. PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)
19. PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)
20. PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)
21. PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)
22. PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)
23. PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)
24. default.payment.next.month: Default payment (1=yes, 0=no)

**4.DATA ANALYSIS**

We then carried out our exploratory data analysis , starting with the univariate , then the bivariate and ultimately the multivariate data analysis.

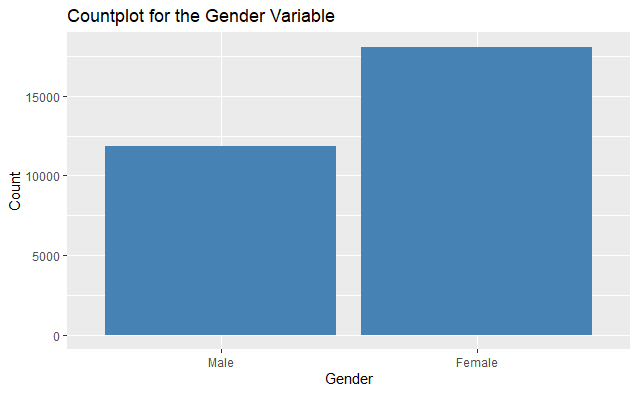
**4.1.Univariate Data Analysis**

We carried out our univariate analysis , plotting count plots for the categorical plots and histograms for the continuous variables.

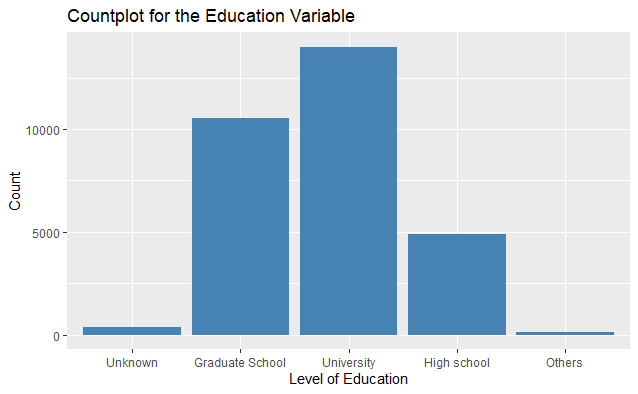


We noted that many of the clients were aged between 23-42 years of age .

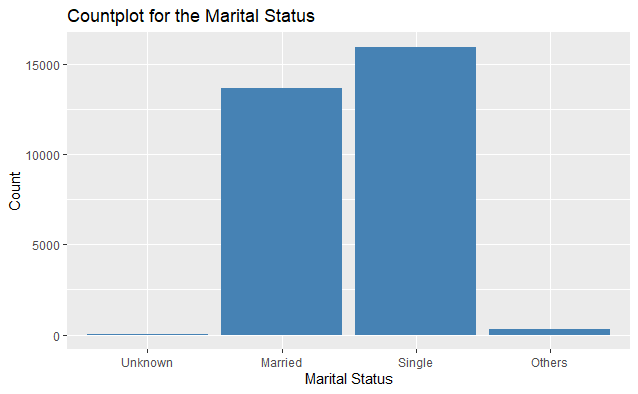
The mean age of the clients was 35.49 , while the youngest client was aged 21 and the oldest was 79.



We also noted that many of the clients , 18091 were female compared to the almost 11874 male clients .



We also noted that 14,019 of the clients had attained a University level education while 10563 had Graduate school education level , while 4915 had attained a high school level education.



15945 clients were single while the number of married clients was approximately 13643 and the ones in the others category were 323 while the unknown was 54 .

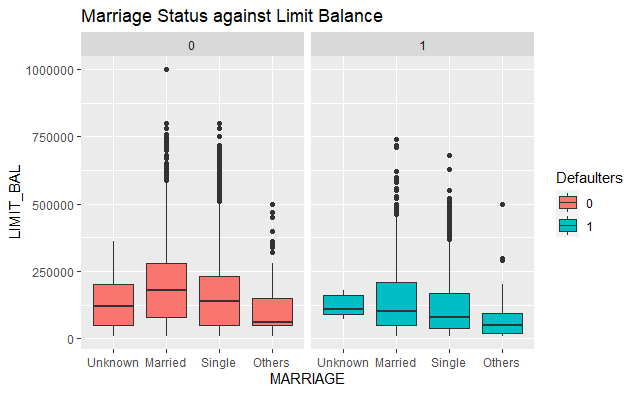
Summary

From the univariate analysis we noted that :

1. There were more single people than married people within the population.
2. Majority of the people in the dataset had a tertiary education , either at a university level or a graduate level.
3. There were more female customers than male customers and the majority of the population was made up of people aged between 20-40 years.

**4.2.Bivariate Data Analysis**

Next we carried out our bivariate data analysis , we compared the variables within the dataset with the response variable i.e. whether or not a client would default from their credit card payments.

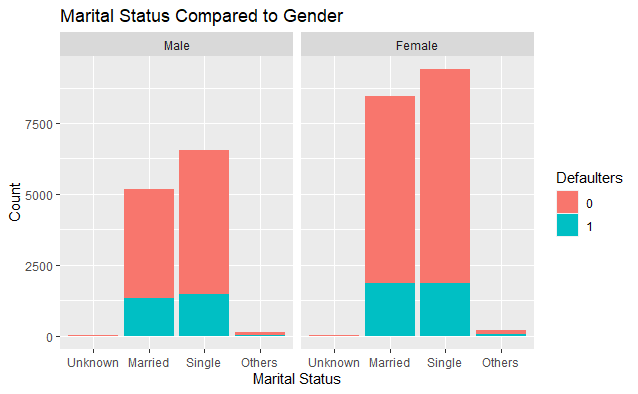


We plotted the outliers between the marital status and the limit\_balance between those who defaulted and those who did not. We can see that the limit balances for those who did not default is higher .

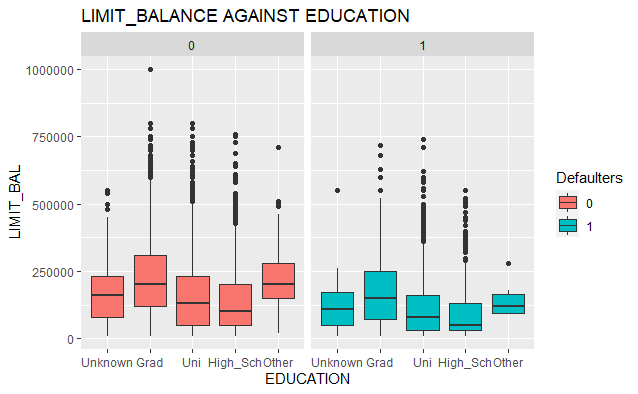
The highest limit balance among the people who defaulted on their payments was 750,000 while that for those who did not was 1,000,000 NT.

Married people also seem to have higher limit balances than their single counterparts.

The people who defaulted on their payments , regardless of their marital status also had lower limit balances than their counterparts that paid their credit card bill.



The proportion of delinquency for both men and women , regardless of their marital status is equal.

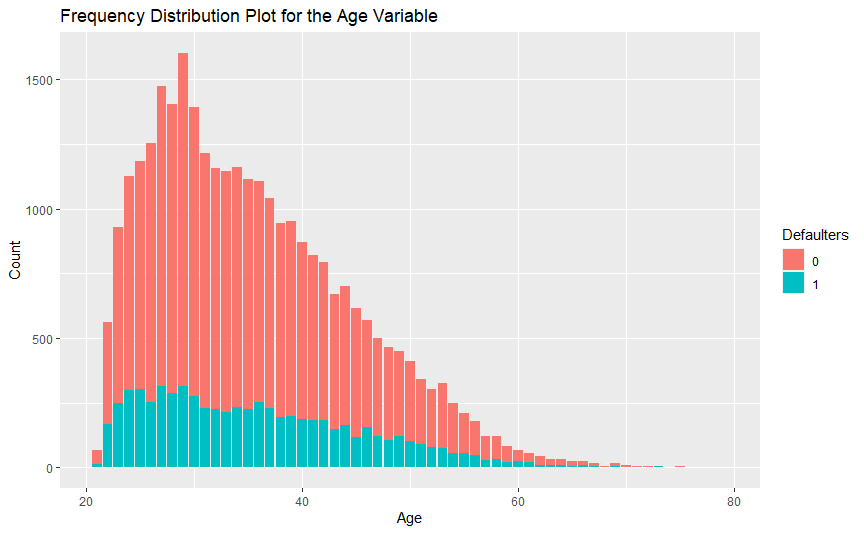


We can see from these boxplots that the clients with the higher limit balance are those with a graduate degree.

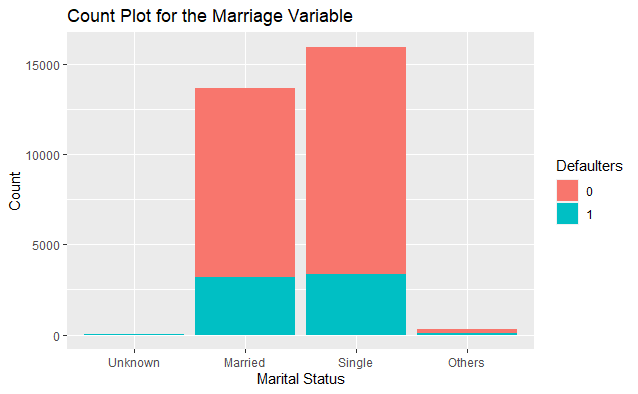
There were very few people with a limit balance of over 500,000 and with graduate degrees that defaulted from paying their credit card fees.

People with university degrees that defaulted had lower limit balances than those that did not , this can also be said for those with a High school education.

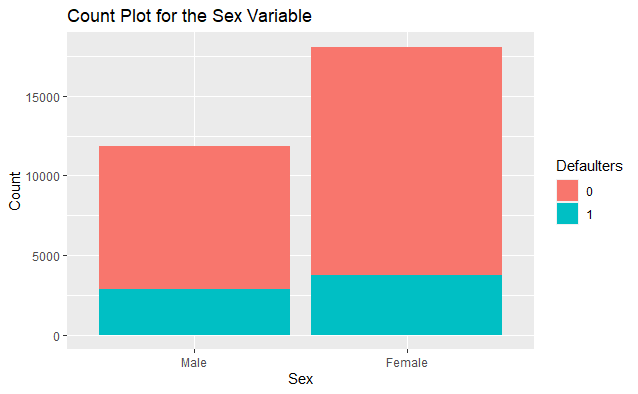
Those with a high school education have a significantly lower limit compared to all the other education levels.



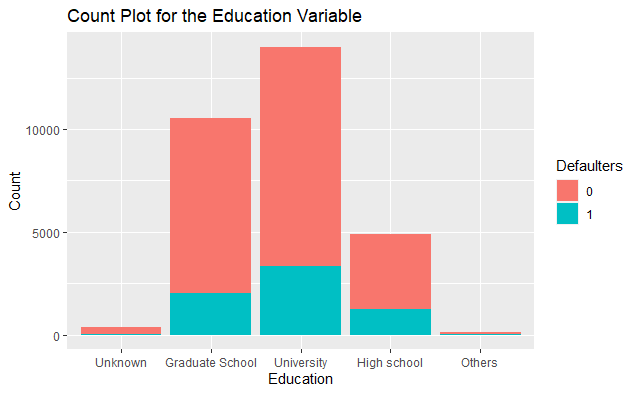
From the plot we can see that the number of people who defaulted was more between the age of 22 all the way to 40 , however these people also made up the majority of the dataset.



The number of people who defaulted on their credit card payments is almost equal among the single and the married people in the dataset , this is despite the fact that there are more single people within this population.



The number of women who defaulter from their payments is higher than that of men. However we should also take into account that there are 18,091 women within this population and only 111,897 men.



The demographic with a university education , who also make up the bigger share of the education column can be seen to have defaulted at their payments more than those with other education levels.

Summary

From our univariate and bivariate analysis , we made many notable observations.

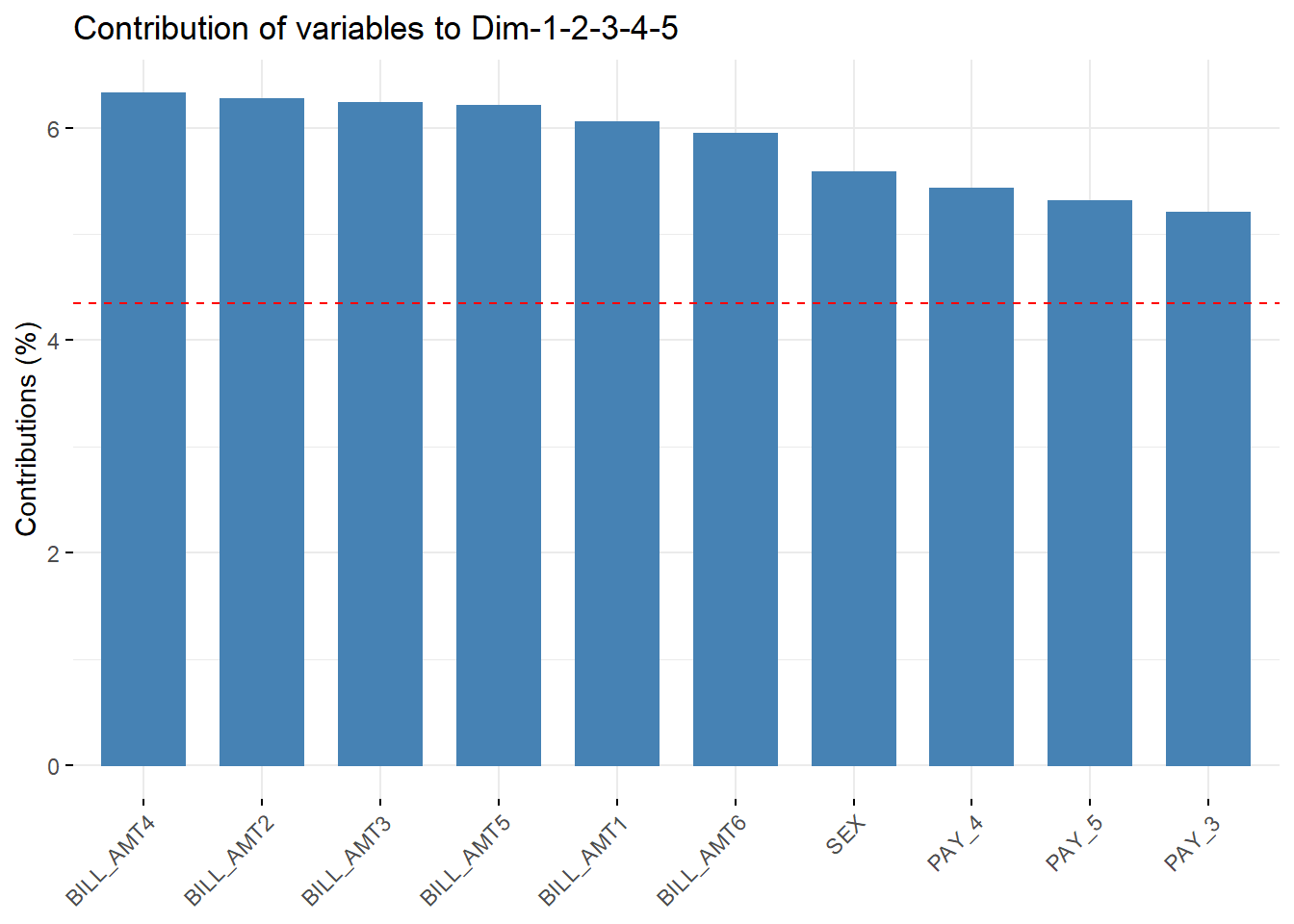
1. Irregardless of factors such as marital status and education , the clients who defaulted on their payments had lower limit balances than the clients who paid their credit card bill.
2. Despite there being more single clients , the number of married people who defaulted was almost equal to that of single people.
3. More women defaulted from their payments than men however women made up more than half the entire population.
4. Those with a highschool education regardless of whether or not they paid their credit card fees had lower limit balances than any other demographic.
5. Those aged between 20-40 made up the largest demographic of both the defaulters and the dataset in general.

**4.3.Multivariate Data Analysis**

Given that our dataset had over 20 columns , we applied some dimensionality reduction techniques in order to get which variables were most suitable for building our model.

**4.3.1. Principal Component Analysis**

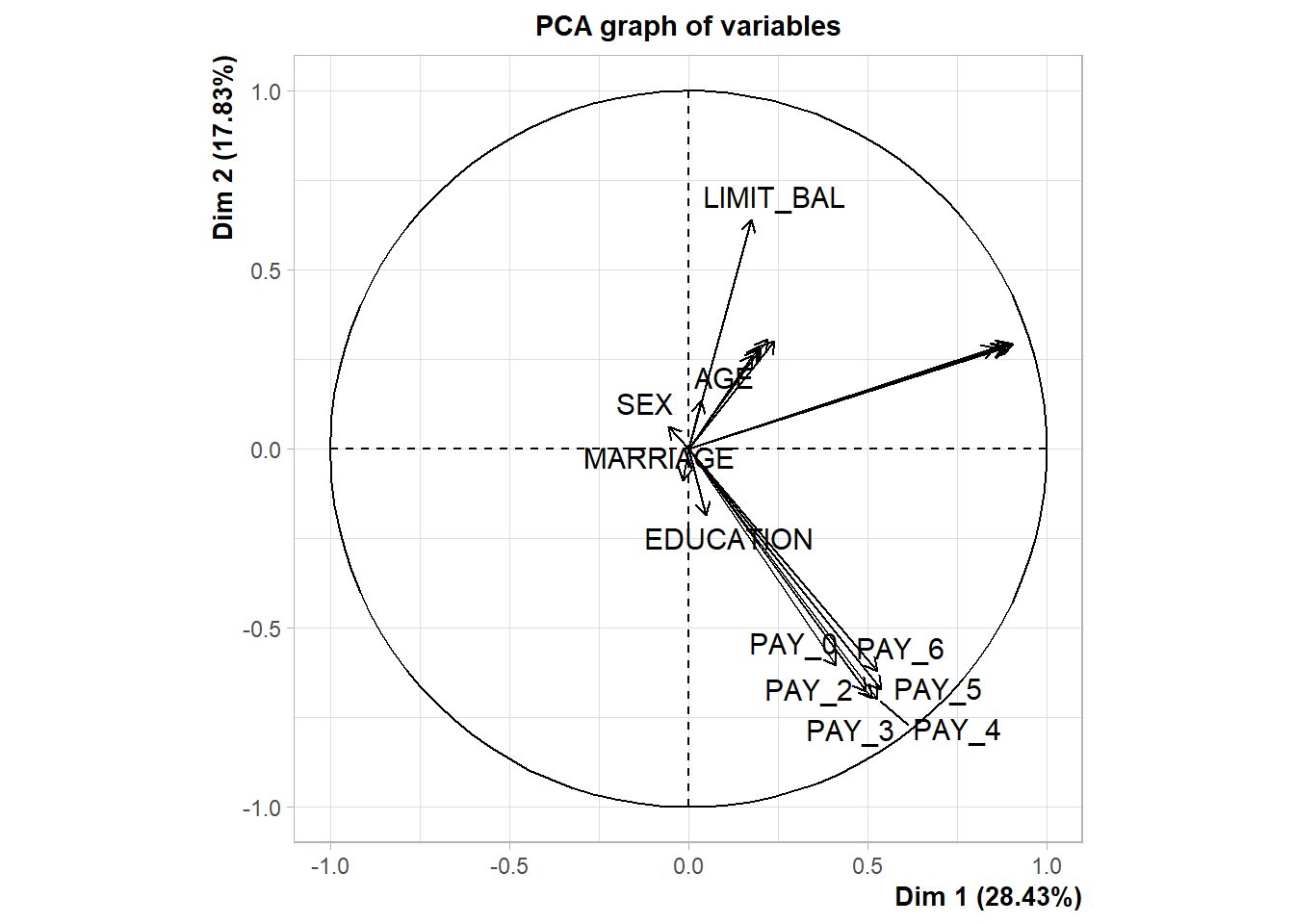
We used the Principal Component Analysis to get the best features for our model.



**4.3.2. Feature Selection**

There were 8 variables found to be highly correlated. These were dropped and the PCA process was repeated again to identify the most contributing variables amongst the remaining variables.

The final selected variables are shown in the diagram below:



**5. CONCLUSION**

In this study, we employed a unique, very large dataset consisting of bank customers’ credit card repayment data to build and test decision trees, logistic regression, and KNN models for predicting credit card delinquency. The performance of the models is summarised below:

|  |  |
| --- | --- |
| **MODEL** | **ACCURACY** |
| Logistic Regression | 81.1% |
| Decision Trees | 82% |
|  |  |

|  |  |
| --- | --- |
| **Model** | **Optimal clusters** |
| K- Means | 2 |

We find that decision trees outperform logistic regression in correct prediction of credit card delinquencies.

From our analysis, we concluded the following two major factors

be observed keenly of whether someone is a defaulter or not are:

**Limit Balance** - From our analysis demonstrates that the clients who defaulted on their payments had lower limit balances than the clients who paid their credit card bill.

**Repayment Status** - which means that just by looking at how late someone is on their payments in 1 month, we can tell if they will default or not. The pay delay by almost one month is the one outstanding followed by those who pay duly as their credit card as of and when it falls due as exhibited by the various count plots for different months.

**6. RECOMMENDATIONS**

In general , in order to mitigate losses and prevent people from defaulting or falling behind on their credit card payments the bank should put measures that incentivise timely payment. For instance the bank should charge late fees for any payment made later than the day it was due.

Given that those aged between 20-40 made up the largest demographic of both the defaulters and the dataset in general the bank could put up measures that would help them prevent losses. For instance they could ask for proof of employment before issuing credit cards.

We recommend the Decision Trees model which performed better than the Logistic Regression Model.

If considering clustering, we recommend KMeans clustering model as it was able to correctly identify 2 clusters.