**MGT 6203 Project Progress Report Team 15 - Credit Card Default**

**Team Members:** Liqing Jing(ljing32), Juan Zhang(jzhang3410), Qinghua Lin(qlin81)

**Github link:** <https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-15>

**Background**

People are using credit card all over the world, the risk of default payments is high and predictive accuracy of probability of default would be helpful for banks. How can we prevent the risk of default payments and predict the probability of default? That would be the research we will focus on.

**Research Questions**

1. Is there a correlation between credit limit and defaulting on credit card payment?
2. Which age group (young, middle, senior) is more likely to default on credit card payment?
3. Are married or single people more likely to default on credit card payment?
4. Is there a correlation between delayed payment history and defaulting on credit card payment? Are people who constantly delay on payment more likely to default on credit card payment?
5. Is there a correlation between education level and defaulting on credit card payment? Which education level (graduate school, university, high school, others) is more likely to default?
6. Is there a higher probability of defaulting if the individual delayed in payment in the previous month?

**Overview of Data**

The original raw dataset(refer to [https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients](https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients%20)) has 30000 rows of data and it includes 25 columns. There are 24 variables plus an ‘ID’ column. The definition of the variables in the dataset are as followings:

|  |  |
| --- | --- |
| Variable Name | Definition |
| Y: default payment next month | (1 = default; 0 = not default ) |
| X1: Amount of the given credit (NT dollar) | Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. |
| X2: Gender | Gender (1 = male; 2 = female). |
| X3: Education | Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). |
| X4: Marital status | Marital status (1 = married; 2 = single; 3 = others). |
| X5: Age | Age (year). |
| X6 - X11: History of past payment | 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; . . .;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; . . .; 9 = payment delay for nine months and above. |
| X12-X17: Amount of bill statement (NT dollar) | Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005;. . .; X17 = amount of bill statement in April, 2005. |
| X18-X23: Amount of previous payment (NT dollar) | Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005. |

The sample data is shown as below:

A screenshot of a computer

Description automatically generated

**Data Cleaning and Feature Engineering**

A few steps were taken to clean and manipulate the data including:

1. Removed the first row (X1, X2,….)
2. Set the next row as header
3. Renamed the column from "default payment next month" to “default”
4. Removed rows with unexplained value in variables EDUCATION (0,5,6) and MARRIAGE (0). After removal, 29601 rows of data left in the dataset.

A screenshot of a computer program

Description automatically generated

1. Based on the requirements from our research questions, steps were taken to add new variables and transform variables.

* Added variable num\_months\_delay to count number of months had delays in the past
* Added variable max\_delay\_months to get the maximum number of delays in past months
* Added dummy variables to indicate if there was delay in past 1-6 months, naming delay\_month1, delay\_month2, delay\_month3, delay\_month4, delay\_month5, delay\_month6
* Transformed ‘AGE’ variable to 1-young (0-40), 2-middle(41-60), 3-senior(61+)

1. Removed columns and variables that are not required for the model. 13 independent variables left and one dependent variable.

A screenshot of a computer code

Description automatically generated

**Imbalance Data Resolution**

An issue of imbalanced data is shown in the plot where 78% are nondefault and 22% are default.

A graph of a number of objects

Description automatically generated with medium confidence

To resolve the imbalanced data, the dataset was split into cross validation dataset(train + validation )(70%) and test dataset(30%). Then the oversampling approach was applied to the cross validation dataset(train + validation ) dataset. Below is size of each set after oversampling:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nondefault | Default | Total |
| cross validation dataset (train + validation ) | 16167 | 16167 | 32334 |
| Test dataset | 6829 | 2009 | 8838 |
| Total |  |  | 41172 |

**Data exploration and analysis**

1. Bar plots of each variable allow us the first claims on the data.

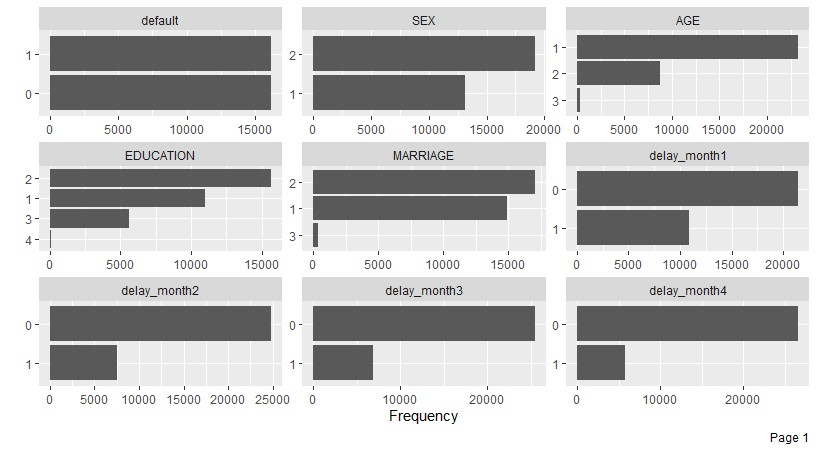
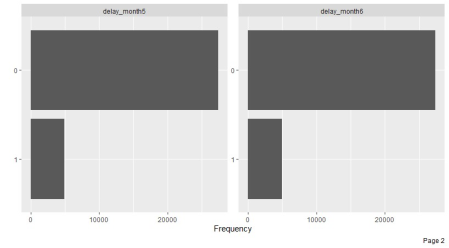
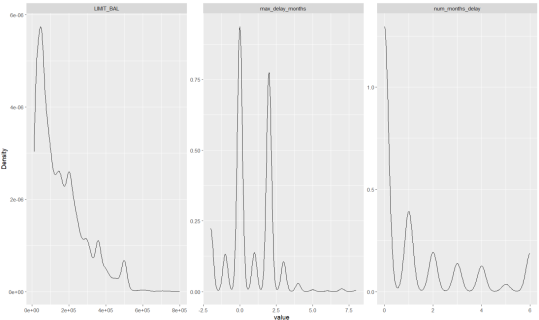


Figure 1

1. Explore distribution and density plots. We can see that Given Credit(limit\_balance), max\_delay\_months and num\_months\_delay are not normally distributed and more likely to be positively skewed distribution. So we may try the non-parametric test and may need transformation when testing for different models in the next period of this project.

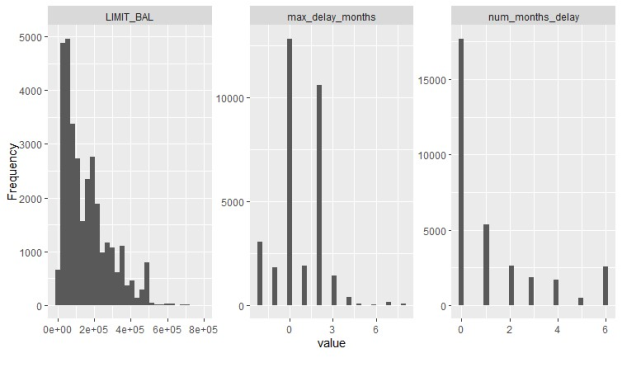


Figure 2

1. Visualization with statistical details(especially plot each variable with the proportions of different groups)

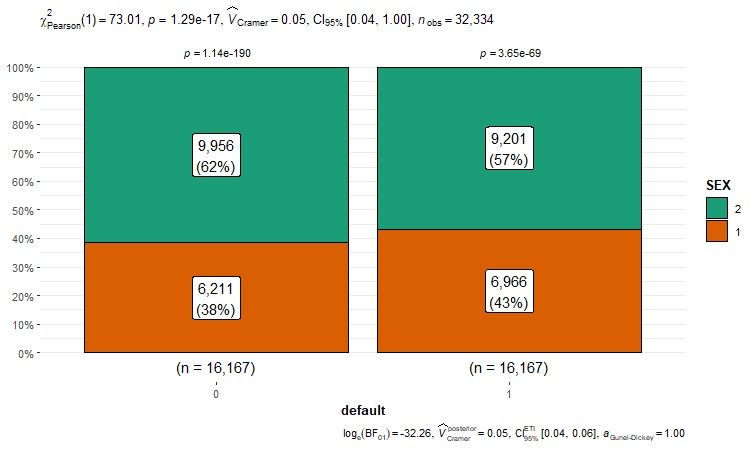
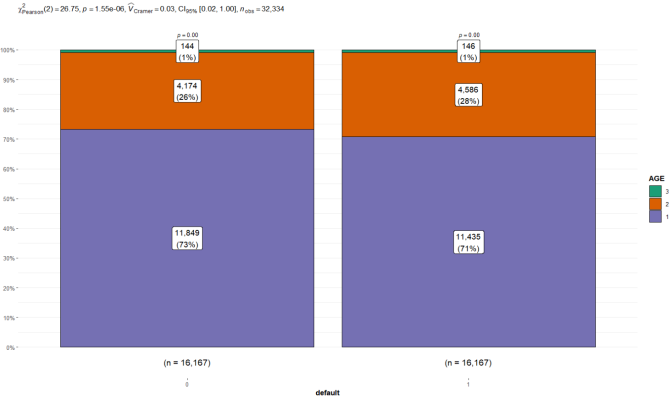


Figure 6 (left) & Figure 7(right)

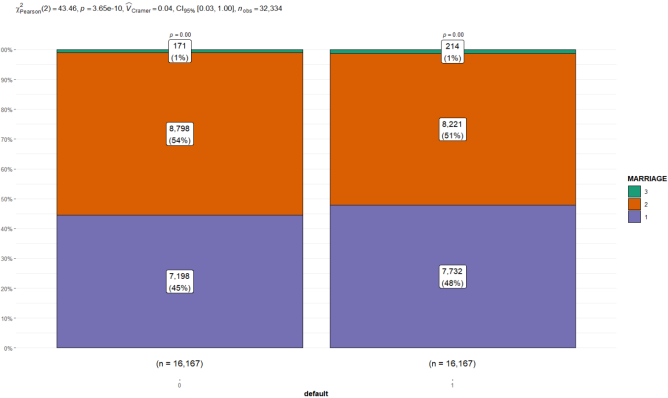
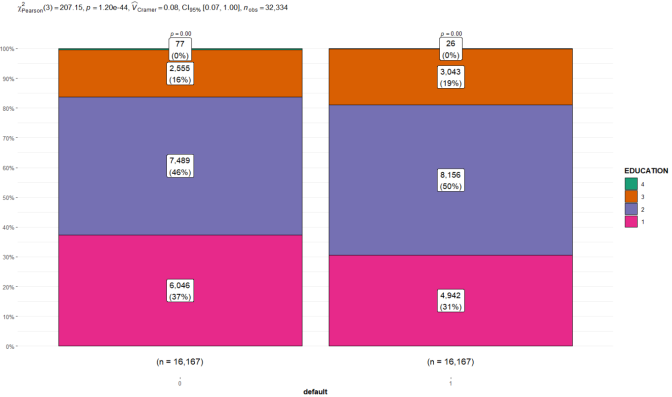


Figure 8 (left) & Figure 9(right)

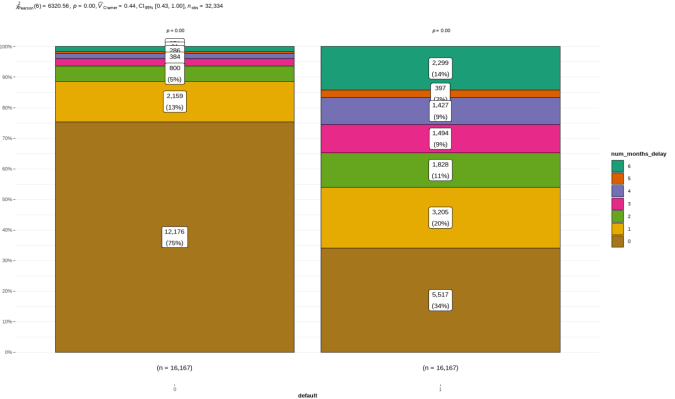
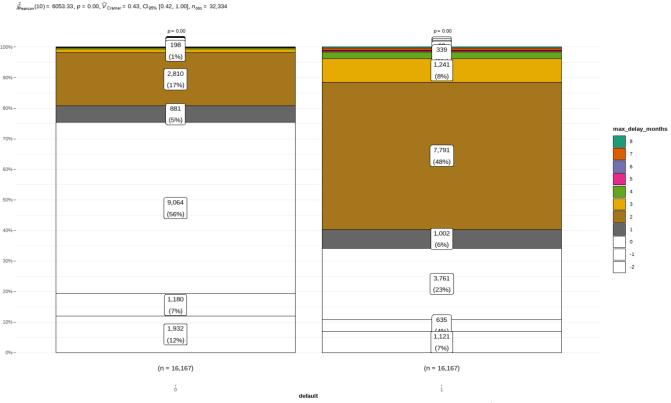


Figure 10(left) & Figure 11(right)

According to the plots, we count and calculate percentages for every category and visualizes frequency table in the form of stacked bars as well as provides numerous details, which allows us to conclude that:

* The gender could be associated with the default, the males may more likely to be default. About 53%(6966/13177) males would be default while 48% (9201/19157) in females( Figure 6)
* The middle-age groups are more likely to be default compared with the young group and senior. 52.35% of middle age groups would be default while 49.11% in young age group and 50.34% in senior age group.(Figure7) ;
* The education level is strongly associated with the default, namely the more educated they get, the more likely they would not be default. 44.98% of graduate school group would be default while 52.13% in university group and 54.36% in high school group.(Figure8);
* The married group are more likely to be default, they may have more pressure from family. 51.79% of married group would be default while 48.30% in single group.(Figure9)
* From the plot, it is obviously that the more number of months had delays in prepayment, the more likely they would be default(Figure 10 & Figure 11). So we could hypothesis that the past delay history correlate with default in next month. People who have bad history may be more likely to be default;

1. Boxplot of categorical variables and numeric variables.

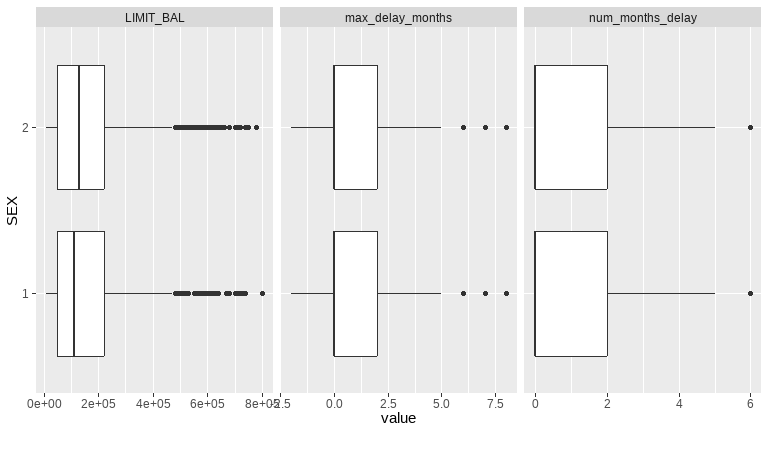
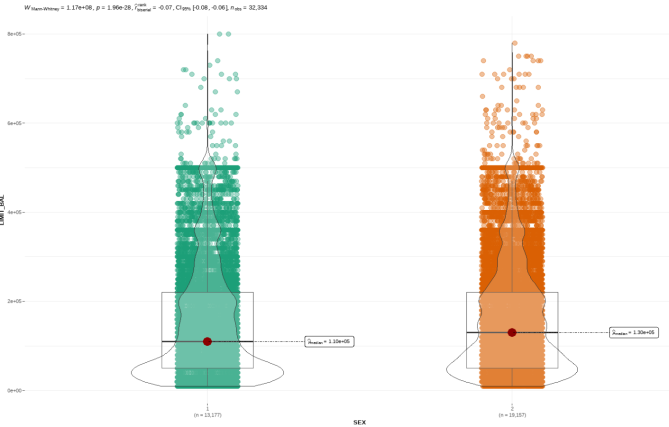
 

Figure 12

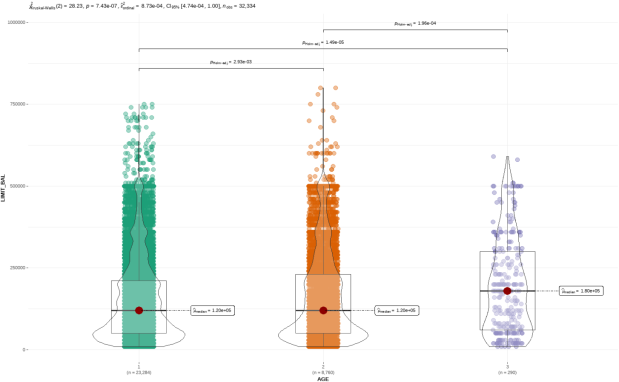
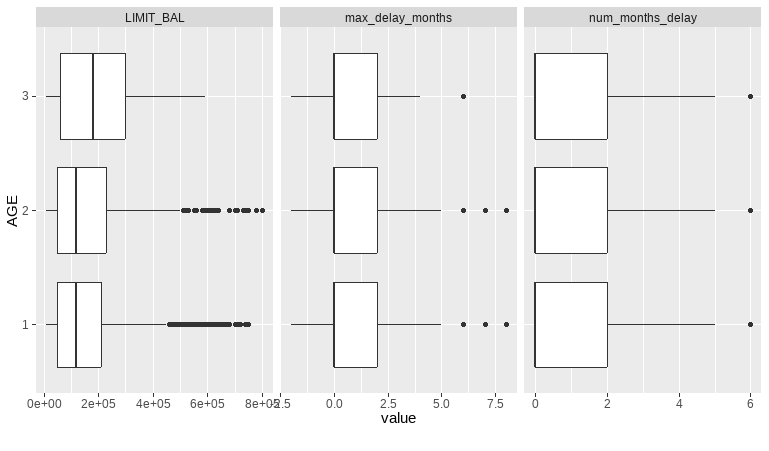


Figure 13

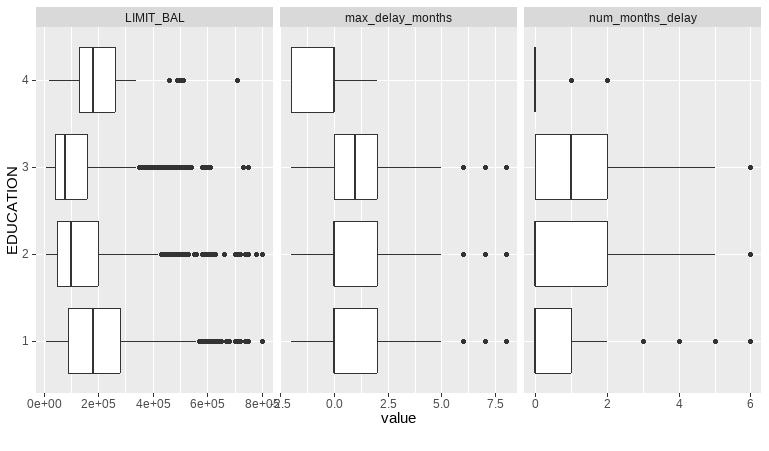
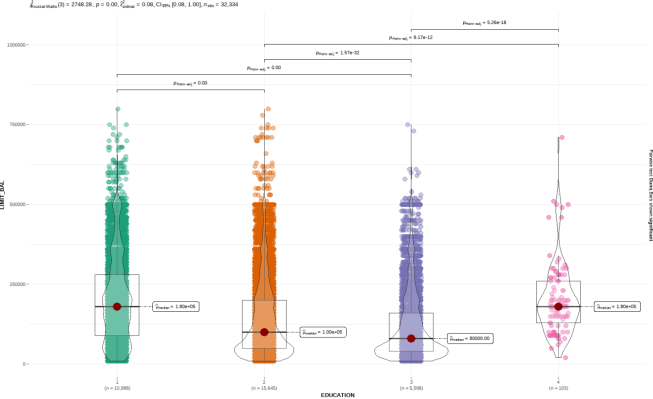
 

Figure 14

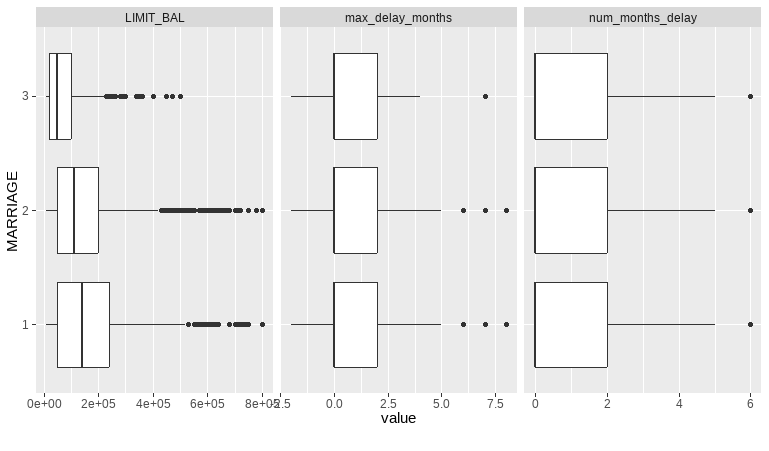
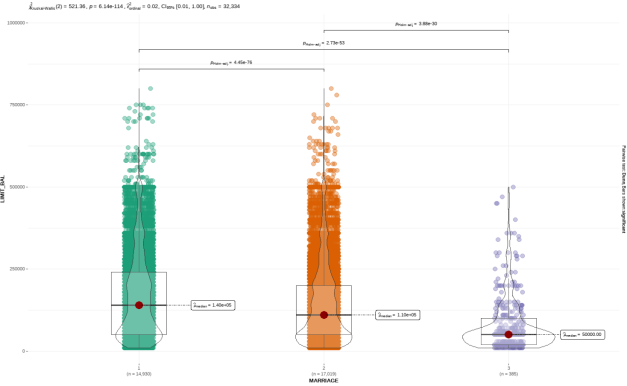
 

Figure 15

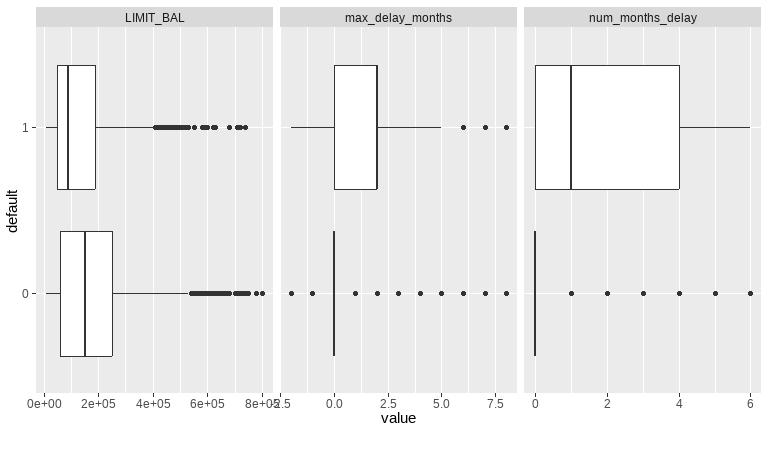
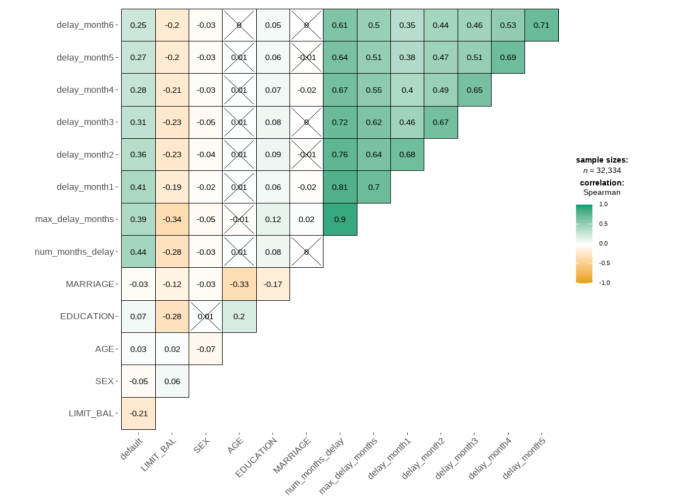


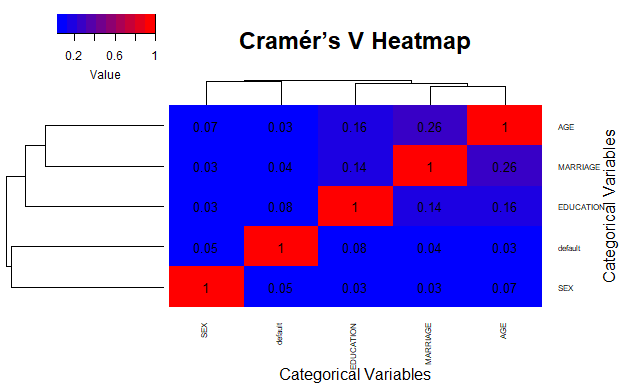
Figure 16

* Averagely, the females tend to have more given credit than the males, and the p-value of pairwise test is significantly less than 0.001.(Figure 12)
* It looks like the senior age groups would have more given credit than the other two while the median of middle age and young age almost the same. the p-value of pairwise test is significantly less than 0.001(Figure 13)
* The education level seems to be associated with the given credit, the more educated they get, the more likely they have the credit. The median of graduate school is significantly higher than the university and high school groups. (Figure 14)
* The married groups has higher given credit than single group, the p-value of pairwise test is significantly less than 0.001(Figure 15)
* The less given credit they would have, the more likely they may be default. Usually, the less likely they could pay duly, the less given credit would be. So that makes sense.(Figure16)

1. Explore Correlation



Non-parametric spearman is appropriate for not normally or not very linearly distribution. The plot displays correlation coefficients and shows the strength of the correlation while the color shows the direction where green is positive and orange is negative correlation. And if it is not significant correlations, it is simply crossed out.



Cramér’s V is a measure of association between categorical variables, and it is an extension of the chi-squared test. The above is the Cramér’s V heatmap for the categorical variables of "SEX", "EDUCATION", "MARRIAGE", "AGE" and "default". It ranges from 0 to 1, where 0 indicates no association, and 1 indicates a perfect association.

**Overview of Modeling**

We will implement the following three modeling in the next period of this project:

1. GLM

We try this traditional classification GLM model and used 10-fold cross validation and calculated the accuracy for each fold. Then we explore the most significant factors on default and find the model goodness of fit and predictive power.

1. XGB

Similar to random forests, XGBoost uses additive methods to build trees one at a time with gradient boosting to learn the optimal discriminative model for prediction. We use 10-folds cross validation to get the average cross validation accuracy.

1. KNN

We chose this model because it is easy to interpret, understand, and implement. We used a loop with 10-fold cross validation to find the optimal K with the highest model accuracy in the range from 1 to 20. We will plot the K-Value vs. Cross Validation Accuracy chart.

**References**

1. Dataset: [https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients](https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients%20)