Logistic Regression Report

**Introduction**

Logistic regression is a type of supervised learning technique used to predict the probability of a binary (yes/no) event occurring, or a classification problem (Chandrasekaran, 2021). Supervised learning is a technique where data is labeled or trained and the objective is to predict the correct label for new input data (Wilson, 2019). For example, we can use logistic regression to determine if a person is likely to have cancer or not (dependent variable) by giving training data with people who are both having and not having cancer, and the probability of a person having cancer could be based on age, tumor penetration, digital exam, antigen value, etc. called independent variables. Since we have two possible outcomes to this question – yes if they are having cancer, and no if they are not having cancer – this is called binary classification.

In this report, I will use logistic regression to predict whether a person will likely default on their credit card payments. Credit cards are physical or virtual cards used to pay for items or services using credit. Credit is the reputation to borrow money with the promise to repay it in the future (Barroso, 2022). Credit is determined by one’s financial behavior: the number of open accounts and duration, payment history including any late or missed payments, loans one has taken out and the remaining balances, and any financial disruptions like bankruptcy or foreclosure (Barroso, 2022). One can build credit over time; the better the credit score, the easier it is to borrow money. One can open a credit card, pay for their needs, and promise to repay the financial companies that lend them money. However, current delinquency rates are 2.8% for branded cards and 4% for retail services for Americans, with those aged 18-29 having a rate of 5.1% (Bautzer, 2023 & Peck, 2023). Thus, financial institutions need to consider who is more or less likely to default on credit cards before giving them one.

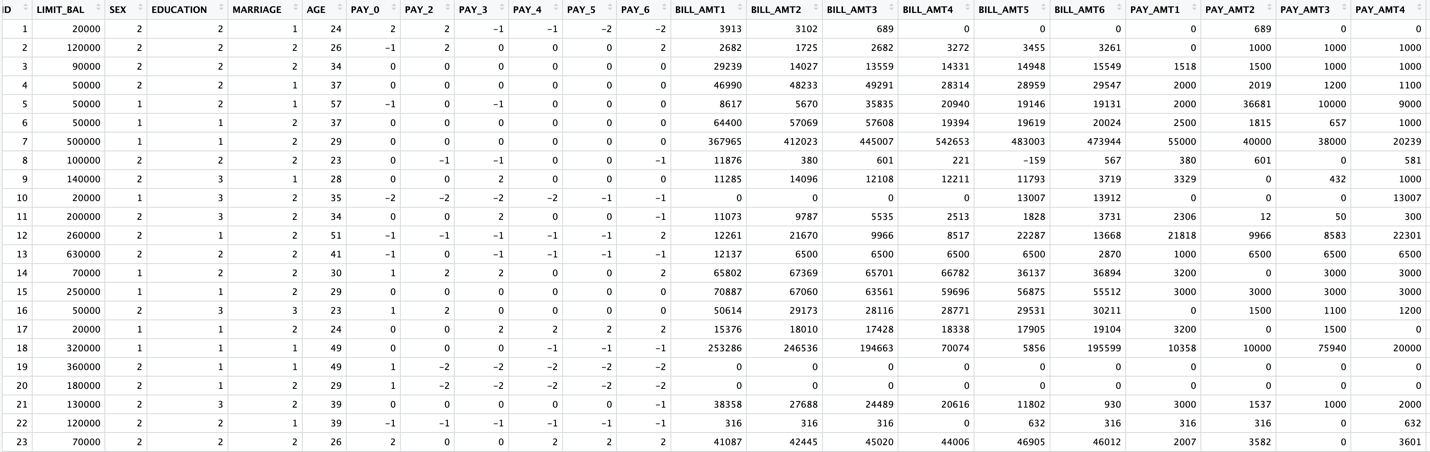
As mentioned above, in this report, I am going to use logistic regression to predict the probability of credit card default using the data of customer default payments in Taiwan. Since there are only two categorical outcomes, default and no default, as the dependent variable, the logistic regression algorithm is chosen to predict the likelihood of the event because it is a simple and efficient method for binary and linear classification problems.

**Analysis**

The default of credit card clients' data was obtained in 2016 on customer default payments in Taiwan (Yeh, 2016). It contains 30000 observations on 25 variables that provide customers’ personal information including ID, the amount of credit given, gender, education, marital status, age, history of past payment, amount of bill statement, amount of previous payment, and default payment next month (Yeh, 2016). The amount of given credit includes both the individual consumer credit and their family credit. The history of past payments, amount of bill statement, and amount of the previous payments were taken from April 2005 to September 2005.

Exploratory Analysis

To perform analysis, first I loaded the defaultcreditcardclients.csv file using R and start with exploratory analysis. Below is a screenshot of some important variables of the defaultcreditcardclients.csv file. The first row contains the attribute names. The remaining 30000 rows are the data, where each row corresponds to a record for a specific person.



PAY – history of past payment

Bill\_ATM: Amount of bill statement in NT dollar

Pay\_ATM: Amount of previous payment in NT dollar

Since the dataset contains a unique ID column, we are going to remove them during preprocessing. Next, I run the “str(credit)” command to display the dataset structure (Figure 1). We can see the number of observations and variables on the first line as well as the type of all variables that are integers. Next, I run the “summary(credit)” command to display the descriptive statistics for all variables (Figure 2). The output shows that the data do not have missing values. The dependent variable default.payment.next.month has values 0 and 1, which meet the requirement that the dependent variable be either categorical or have valid values 0 and 1, thus do not require preprocessing. Independent variables can be numeric, and thus also do not require preprocessing. However, some of the independent variables are categorical variables, therefore I will convert them to factor variables, including sex, education, marriage, age, and pay\_0, pay\_2 to pay\_6.

Preprocessing & Algorithm Intuition

As mentioned above, I used “credit$ID <- NULL” command to remove the unique identifier column. Next, I changed numeric variables to factor variables for all categorical variables (Figure 3). However, after checking back using “summary(as.factor())” command, I see that the data has many missing values (N/A’s) (Figure 4). Thus, I use “na.omit” command to remove data rows with missing values and run the “set.seed(1234)” command to generate the random number for the simulation and make sure that the result is reproducible (Figure 5).

To build a regression algorithm, I divide the data into training and test sets. The training set is used to build the classification tree model which contains 70% of the data, and the test set which contains 30% of the data is used to evaluate the accuracy of the model (Cognitive Class, 2017). In addition, I build and run the glm function and store the model in the model variable (Figure 6). GLM function stands for the general linear model, a generalization of ordinary linear regression used to fit generalized linear models by building a linear relationship between the response and predictors. Response equals the mean value of response when predictors take a value zero plus the change in the mean of the response per a unit change in predictors (Bati, 2018).

Using the print and summary command, we get the statistics for deviance residuals, including min, median, and max. A residual is a difference between the predicted value and an actual value. Next, I run the “exp(coef(model))” command to compute the intercept and coefficients for the odds ratio (Figure 7). The coefficient is the change in the log of odds if the variable is increased by 1 while the remaining variables remain constant. The correlation coefficient R^2, which is the most important factor in logistic regression, falls between 0 and 1 and is used to assess the strength of associations between data variables or determine whether a change in a predictor variable makes the event more or less likely. T-test and F-test are used to determine if the relationship is significant. A small p-value suggests that the relationship observed is unlikely to occur by chance.

After building the model on training data, we have to evaluate it by rounding the predicted values to the nearest integer and then run the confusion matrix command to show the predicted and actual predicted class (Figure 8). We also have to evaluate the model on test data after training data using the same rounding command and then store the values in a variable called mypredictions, followed by the matrix command (Figure 9). The classification accuracy is the sum of numbers on the diagonal divided by the sum of all numbers. After that, we are going to use the ROCR package to see how good the model is. The closer the curve, the better the model. Lastly, I use the minimal adequate mode or step function to remove non-significant variables. The step function removes a variable at a time until only significant variables remain.

**Result**

First, let’s evaluate the output of the train data. By looking at Figures 10 and 11, we can see the residual deviance, p-value as significant, standard error, and coefficients of all variables. Pay\_4payment delay for nine months and above decreases the most while pay\_5payment delay for nine months and above increases the most with the highest standard error. The coefficients are also statistically significant because p < 0.5. Some of the most significant variables are limit\_bal, female, education: university, single, marriage: others, pay\_0 payment delay for one month, three months, four months, and five months, and pay\_ATM1, 2. Next, we will evaluate the confusion matrix for training data in Figure 8. Since the classification accuracy is equal to the sum of numbers on the diagonal divided by the sum of all numbers, thus equal to 13304/16183 = 82%, which is a high accuracy. Next, I will do the same thing with the test data. The classification accuracy of the test data is 82%, which is close to the classification accuracy of the training data. We can also use the mean”(round(predict(model)))” command to get the values in R as in Figure 12. The classification error on the hand is equal to 18% for both test and train data, and the accuracy and error values add up to 100.

Next, let’s look at the ROCR plot in Figure 13. Since the line is close to 0.5 for a true positive rate, it means my model is not very liable. I’ll take a deeper look by using the residuals plot (Figure 14). My model indeed has a big distance from the predicted line, further supporting that it is not a good model. Now, I want to look at the effect that each individual variable has on the model. I use the effects package to plot the all-effect model, which shows me how the change in each variable has on the default on the payment variable (Figure 15). The plot also shows the 95% confidence interval for each variable. We can see here as the limit\_bal goes up, the default on payment probability goes down. Also, if the person is female, has other education, is single, or pays a higher amount for previous payments, are most likely to have lower default on payment rate. On the other hand, if the person is married or others, is older in age, or has higher bill statements, have a higher chance of default on payment.

After running the step function, there are only 13 significant variables left. I plot the effect model again and then look at the classification accuracy and confusion matrix for the reduced model (Figures 16 & 17). The result still shows the same accuracy of 82% and the same effects plot with fewer variables. Using the Naïve Bayes method, I get the same result for coefficients, p-value, and other values.

**Conclusion**

To summarize, by using the logistic regression technique, I was able to predict the probability of default on payment next month based on predictor variables. Many predictor variables have a significant impact on the probability of default on payment next month, including the amount of given credit, sex, education, gender, marital status, history of past payments, amount of bill statement, and previous payments. Despite the model’s high accuracy of 82%, the ROCR, residuals, and effects plots suggest that it is not a reliable model.

One limitation of this model is the inaccuracy of the metadata or variable description provided by the creator of the data. Some of the variables have more variation than the provided information, such as education or history of past payments. For example, education has 6 different categories while the author only provides four categories: graduate, university, high school, and others. Thus, while changing the variables to factor, I need to omit values that do not belong to any categories and that might affect the accuracy of the model due to sample size. Another limitation of logistic regression is that it is limited to binary classification problems, meaning that it can only be used to predict the probability of an outcome belonging to one of two classes. Additionally, it is sensitive to outliers and can be prone to overfitting if the data is not properly preprocessed. Overall, in order to achieve successful results with a logistic regression algorithm, it is essential to have high-quality data.

To ensure better results in the future, I will take the time to review the data description and quality before using it for logistic regression, in order to reduce overfitting and minimize the amount of time spent on preprocessing including removing outliers and incorrect values.

**References**

Barroso,Amanda (Jan 2022). *What is Credit and Why Do You Need It?* Retrieved from: <https://www.nerdwallet.com/article/finance/what-is-credit>

Bati, Firdu (Mar 2018). *Data 630 Lecture: Linear-Regression.* [Video: Youtube]. Retrieved from: https://www.youtube.com/watch?v=jqTDuhQR-no&t=107s

Bautzer, Tatiana (April 2023). *More US Consumers Are Falling Behind on Payments.* Retrieved from: https://www.reuters.com/markets/us/more-us-consumers-are-falling-behind-payments-2023-04-18/

Chandrasekaran, Maran (Nov 2021). *Logistic Regression for Machine Learning*. Retrieved from: <https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/>

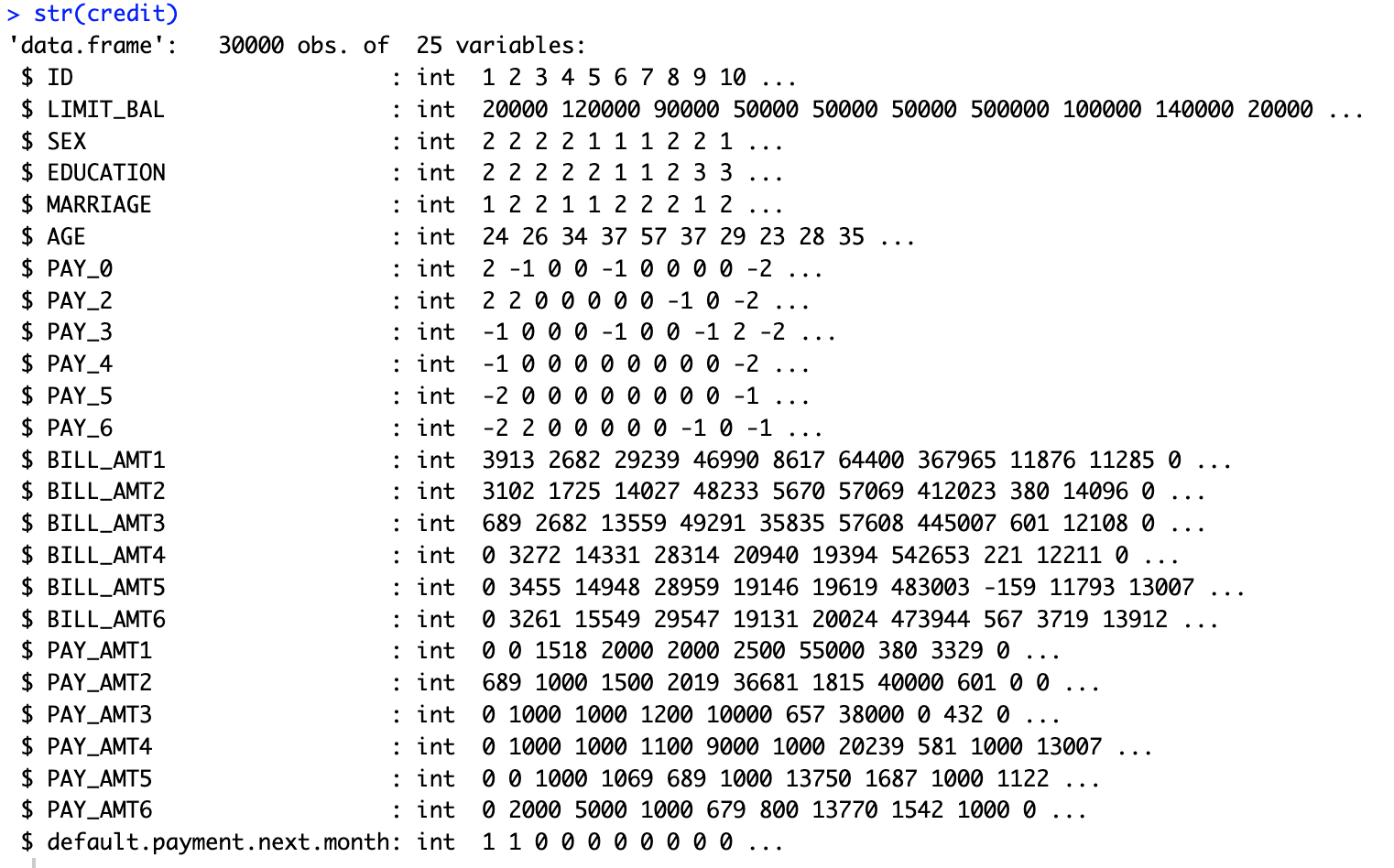
Cognitive Class (Apr 2017). *Machine Learning – Supervised Learning Model Evaluation.* [Video: Youtube]. Retrieved from: https://www.youtube.com/watch?v=VcSNDMBTE-s

Peck, Emily (May 2023). *Credit Card Delinquency Rate for Young Adults is Rising. Retrieved* from: https://www.axios.com/2023/05/16/credit-card-delinquency-young-people

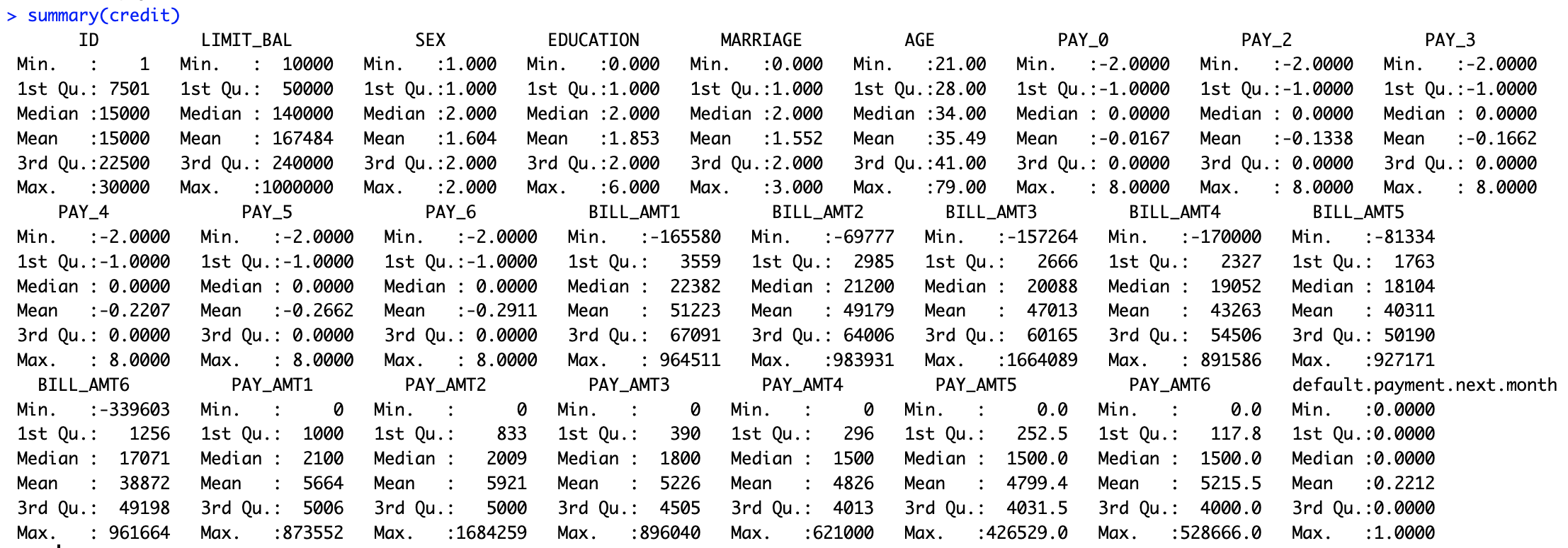
Wilson, Aidan (Sep 2019). *A Brief Introduction to Supervised Learning.* Retrieved from: <https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590>

Yeh, I-Cheng (Jan 2016). *Default of Credit Card Clients.* Retrieved from: <http://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>

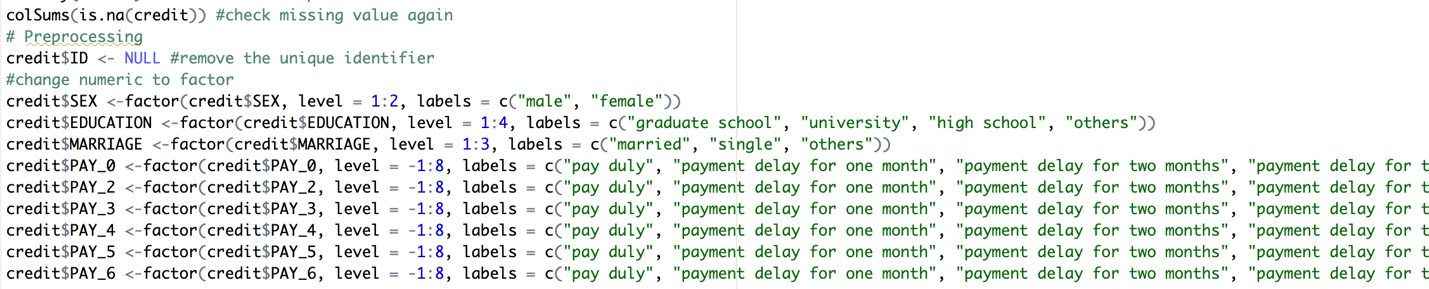
**Appendix**



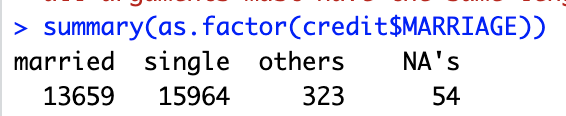
**Figure 1. Dataset structure**

****

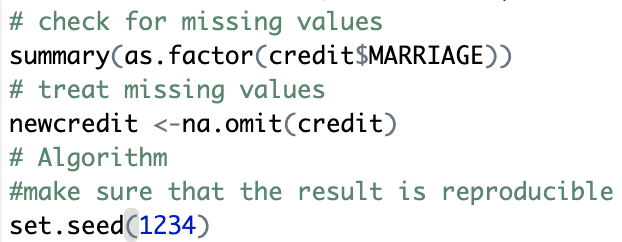
**Figure 2. Display descriptive statistics for all variables**

****

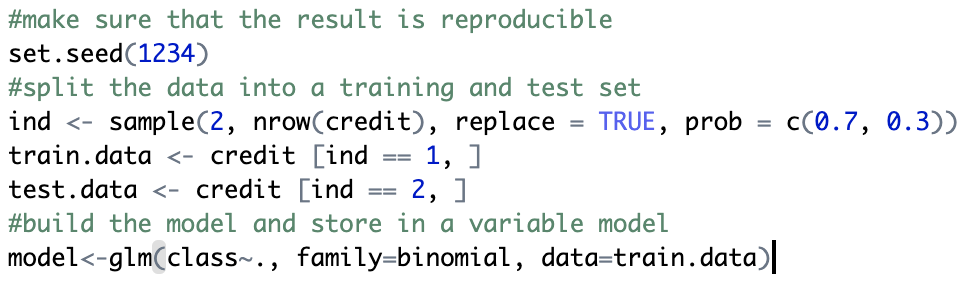
**Figure 3. Preprocessing**

****

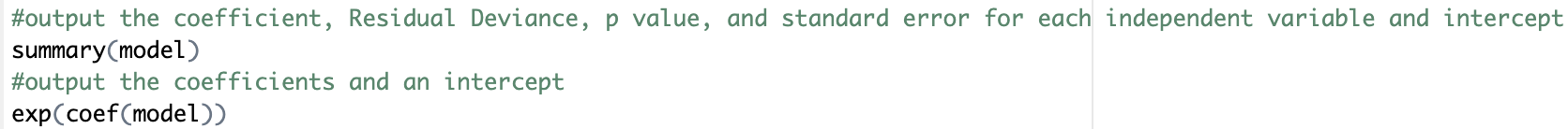
**Figure 4. Missing data after the change to factor**

****

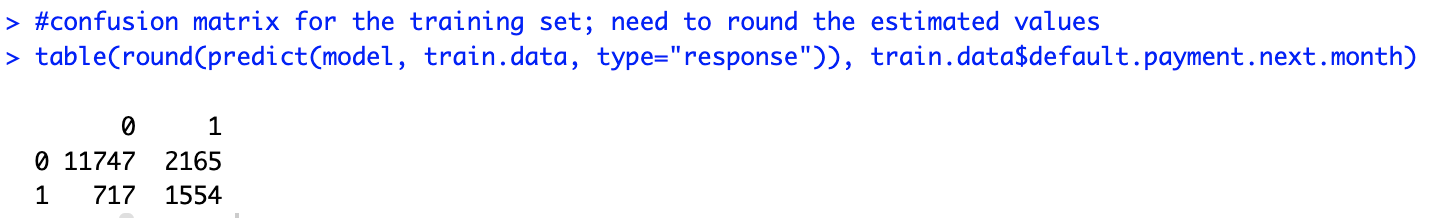
**Figure 5. Treat missing values and generate random numbers**

****

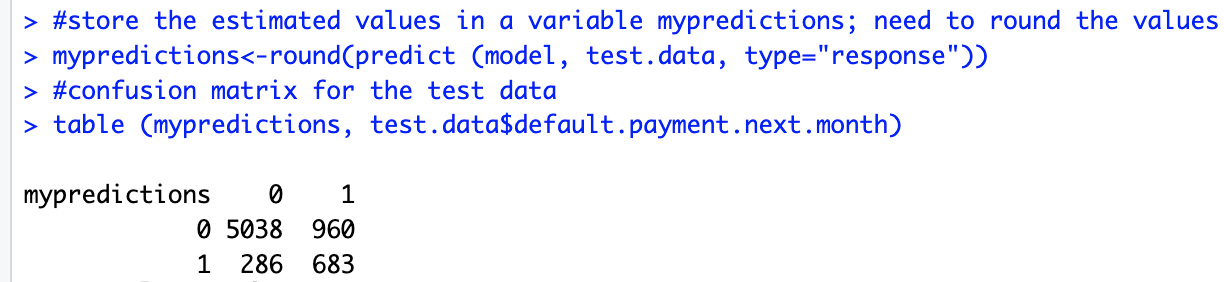
**Figure 6. Use the training set to build the model**

****

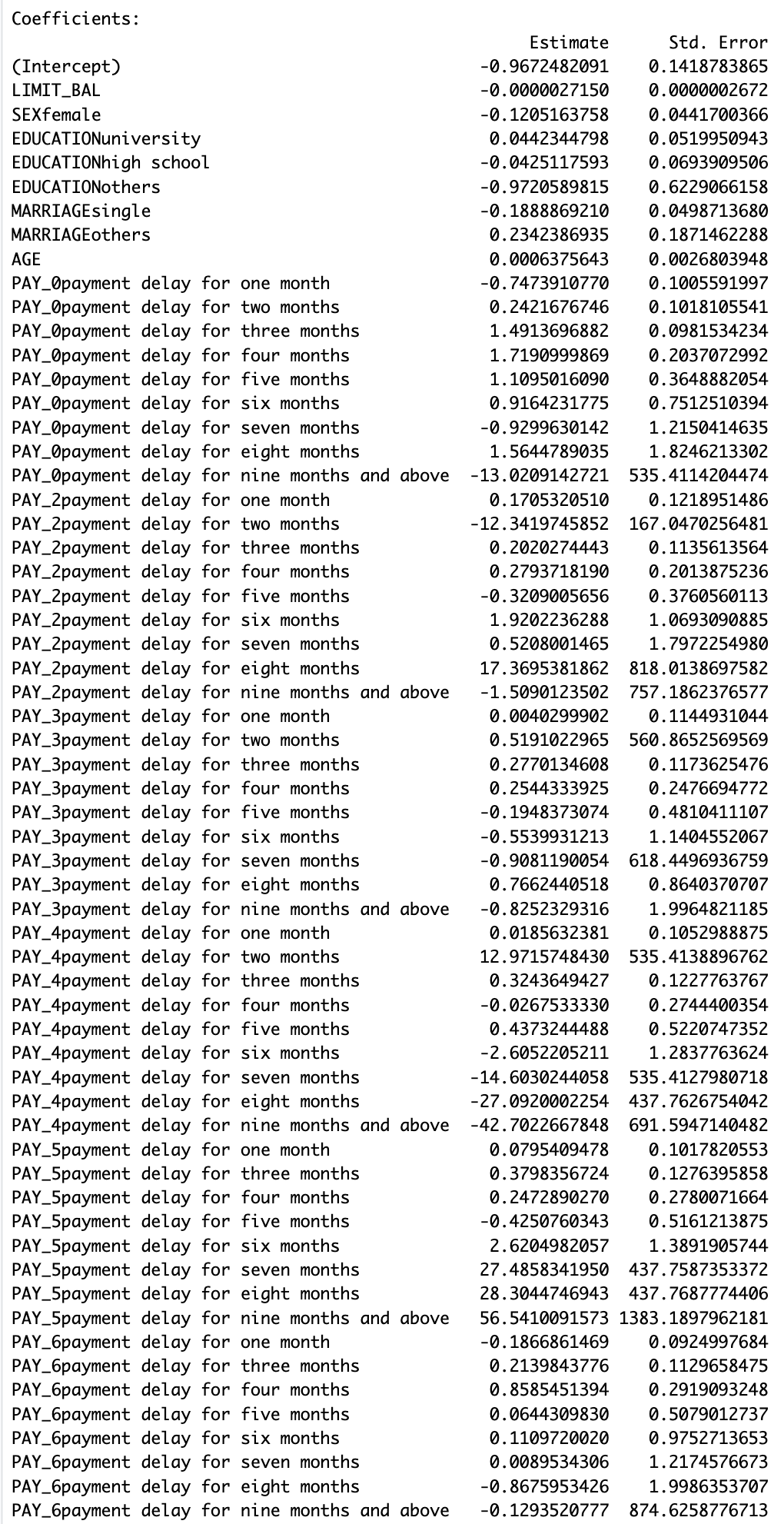
**Figure 7. Summary statistics and coefficients command**

****

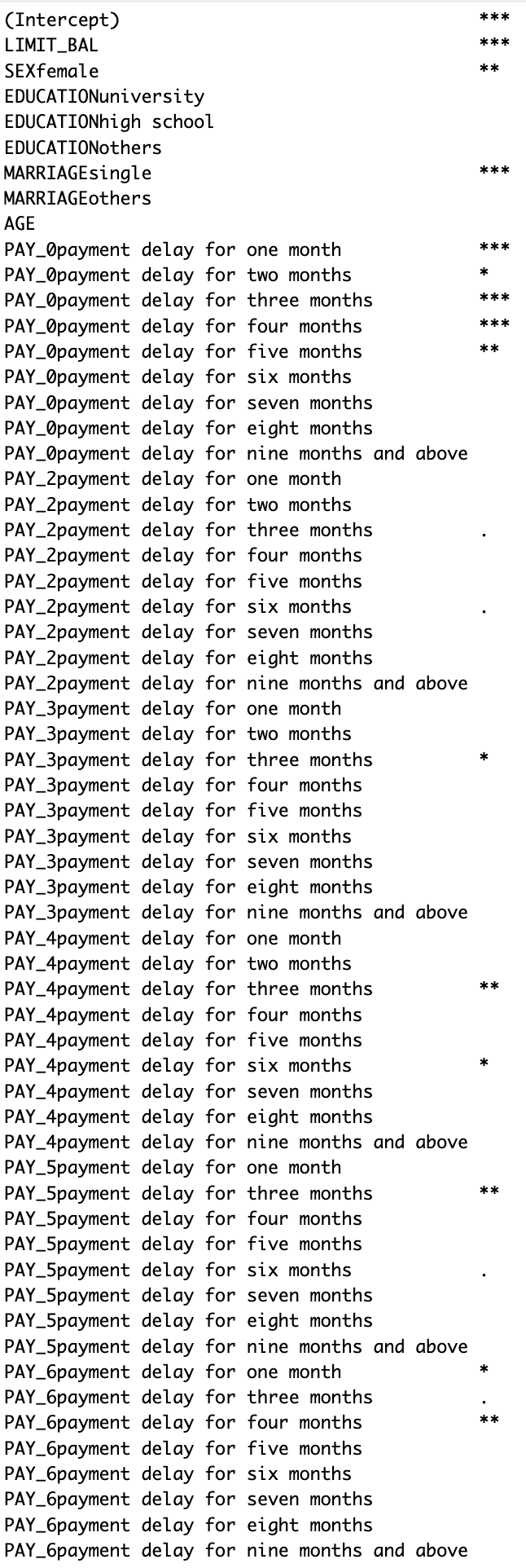
**Figure 8. Confusion matrix for the training Set**

****

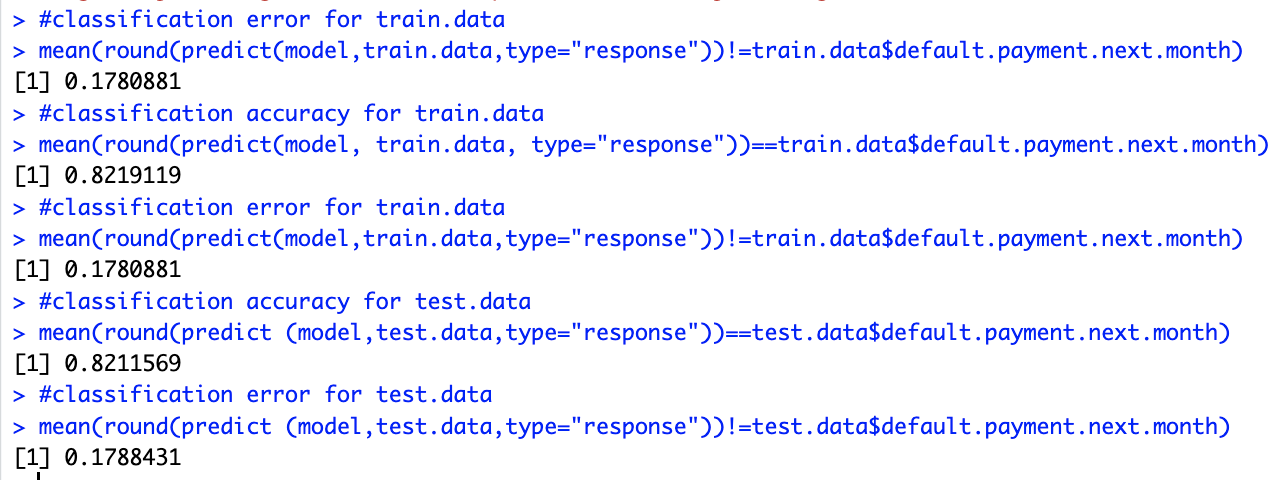
**Figure 9. Confusion matrix for the test data**

****

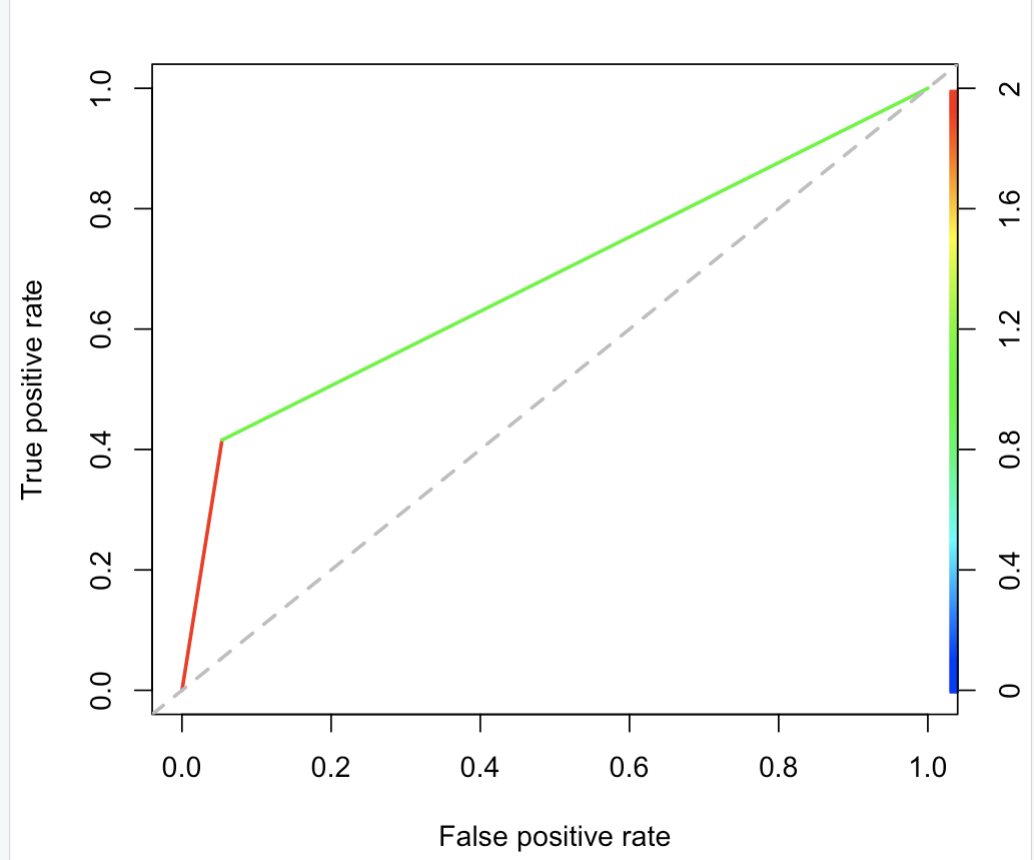
**Figure 10. Intercept, coefficients, and standard error for the odds ratio**

****

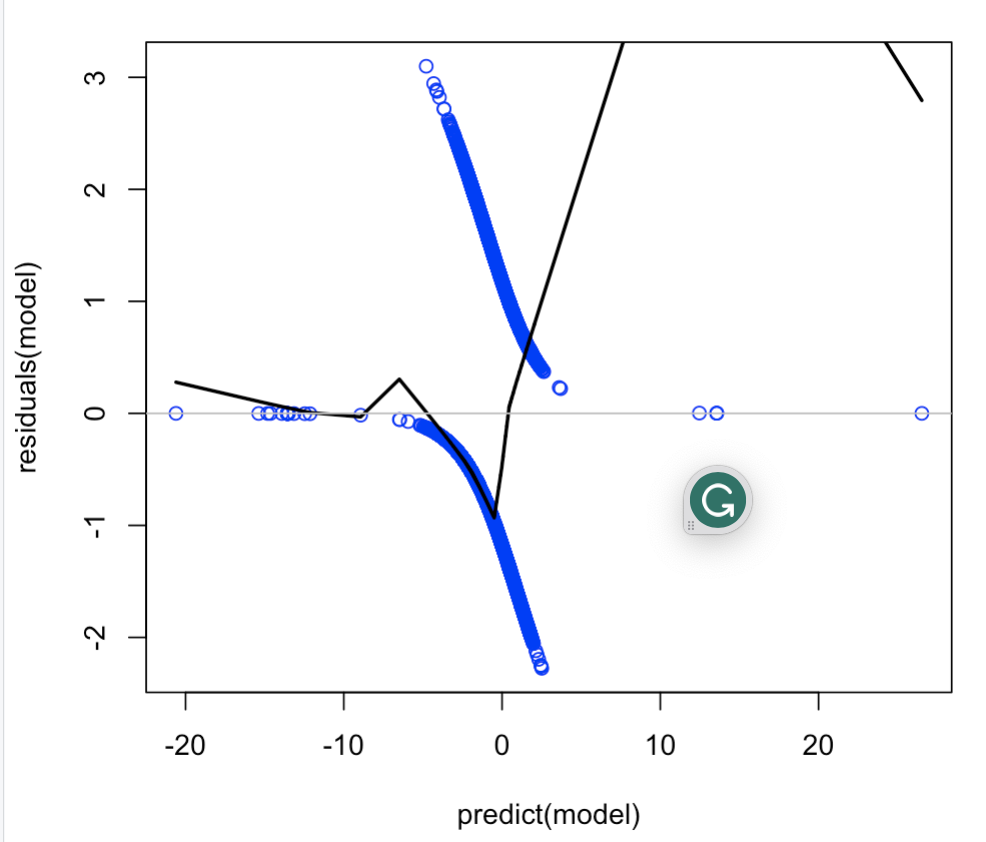
**Fig 11. Significant variables**

****

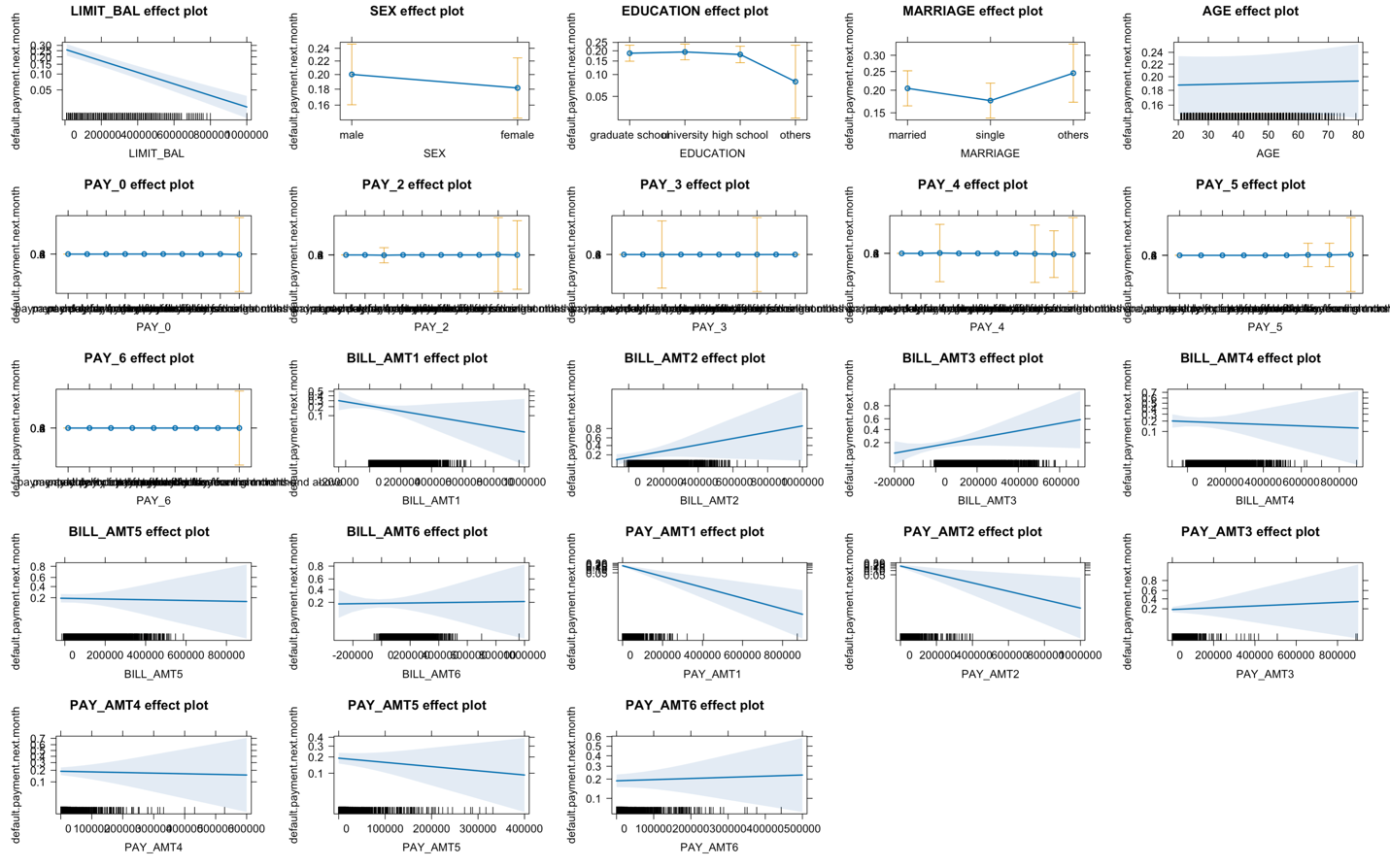
**Fig 12. Classification accuracy and error for train and test data**

****

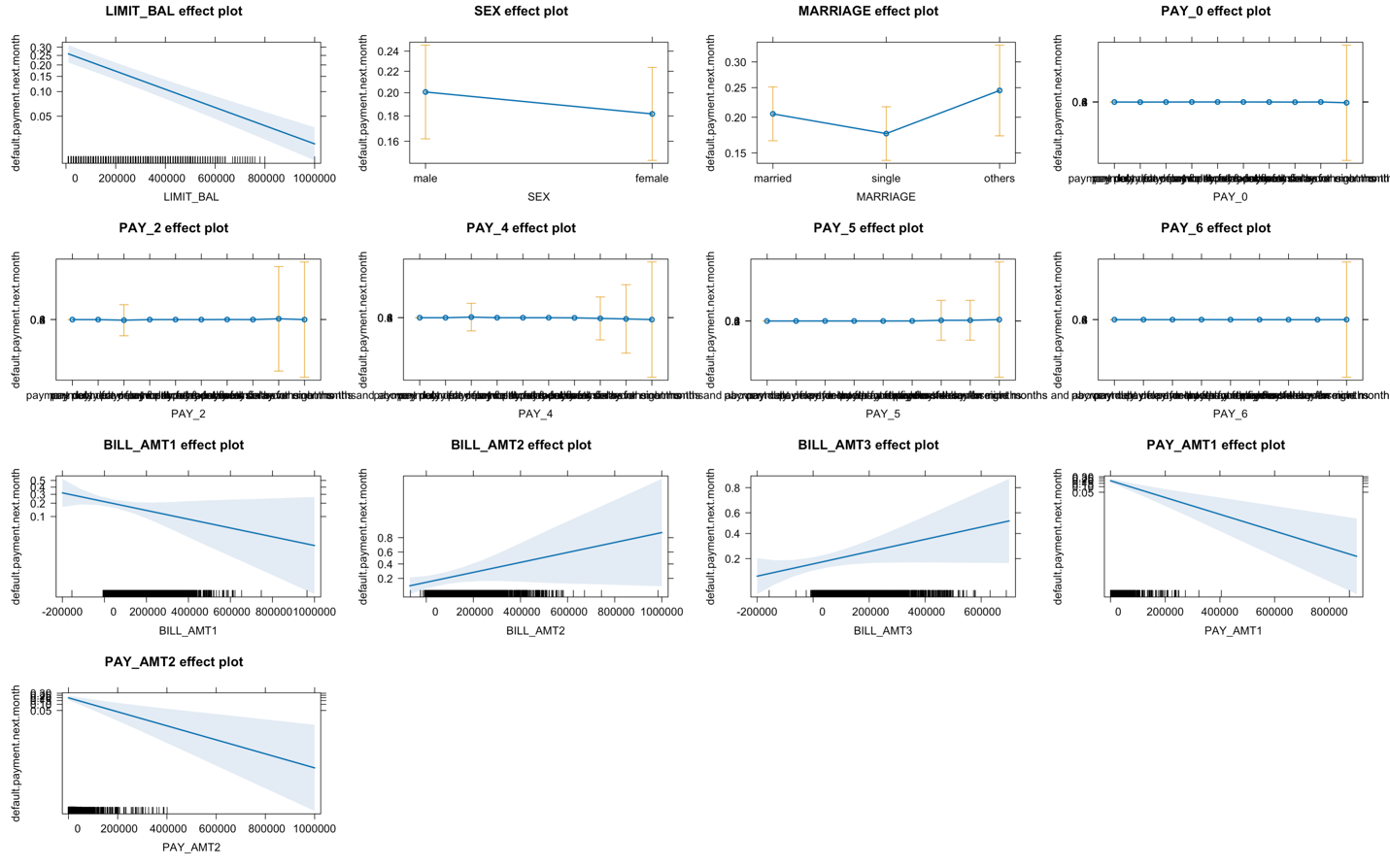
**Figure 13. ROCR plot**

****

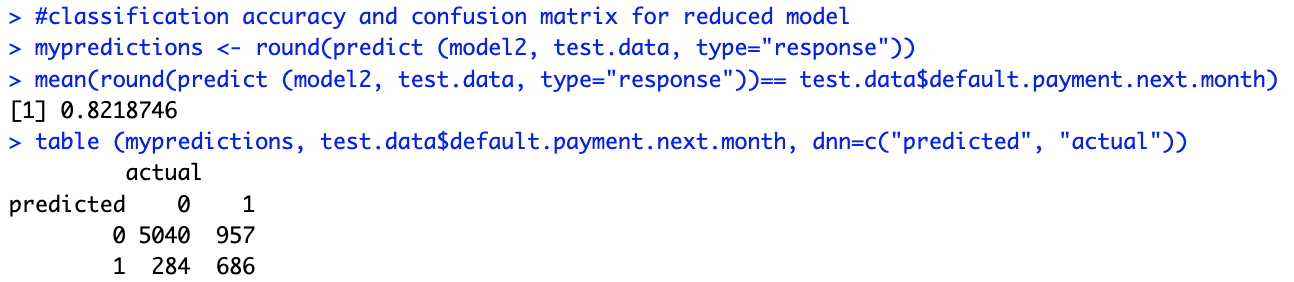
**Figure 14. Residuals plot**

****

**Figure 15. Effects plot**

****

**Figure 16. Effects plot for reduced model**

****

**Figure 17. Classification accuracy and confusion matric for reduced model**