**Credit Card Project Writeup**

**Summary Statistics and Data Cleaning:**

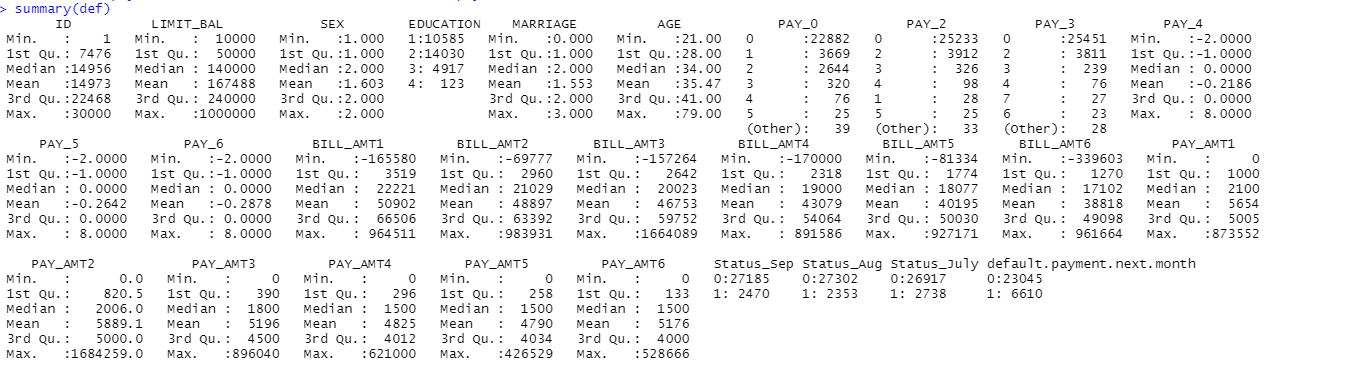
At the first glance of the data, the variables sex, marriage status, and age seemed irrelevant to whether the next payment in credit card will default or not. Also, the process of collecting those data could have embedded some biases, so I decided to exclude them from the analysis. There were a few assumptions I made for the remaining data cleaning process:

1. For the variable “education”, as data description only specified what values 1 to 4 mean, I assumed any number out of that range was an error, thus was assigned as “NA”. In total, there were 345 NAs, and I removed the rows of those data with missing values as 345 is just a small portion of the data set and does not necessarily change the data structure.
2. In real life, if card holders missed paying the credit cards for three months, their status and the next payment would be automatically considered as default. That means the credit history in the most recent three months could have the strongest correlation with default payment. Therefore, I picked the payment history in the last three months (X6 to X8) to be my predictors. Similarly, If I were using the bill statement and payment amount, I would only consider those of September, August, and July.
3. The variables payment history (X6-X11) have some values other than those specified in the data description. There are more than 10000 data with value of 0, and we are not sure about what it means. As the amount of the data with value of 0 is significantly large, taken more than 1/3 of the data set, I thought it might be caused by a clerical error where researchers may have put a wrong Excel formula and the data has been assigned to a value higher than the original value by 1. In other words, the data with the value of 0 in fact means they have paid the balance on time. In data cleaning section, I assigned the value of 0 to all of the payment paid duly as R sometimes cannot process negative numbers for categorical variables.
4. I assumed payment often occurs one month later than the bill statement, so I decided not to use the bill statement data in September. More specifically, if I want to know if the payment in August is paid on time, I will probably use the bill statement in August and payment amount in September.
5. Normally speaking, if card holders are able to pay at least 2% of their billing statement every month, the payments will not be considered as default.

When I looked further into the data, I noticed that there might be some relationship between the amount of bill statement and the amount of previous payment. If there is a higher balance in the bill statement, the card holders probably need to pay more to satisfy the minimum amount of the payment (roughly 2%). From a finance perspective, I was wondering if I could create a new variable by using bill statement and previous payment to further explain their relationship. In the Excel sheet, three columns are added and labeled as “Status\_Sep”, “Status\_Aug”, and “Status\_July” in representing the payment status in September, August, and July (referring to the second assumption). I set the outcome to be binary with value 0 and 1. If the payment amount is equal to or greater than 2% of the bill statement, it returns 0, meaning the payment does not default; otherwise, it returns 1, suggesting the payment defaults for that month.

I also decided to drop the bill statement and previous payment for my predictors because it seems like it is simply exploring if a higher or lower amount of bill statement or payment could have impacted the default payment, but it cannot explain the relationship between the two variables.

After applying log function to the only numeric variable “LIMIT\_BAL” and converting the categorical variables I am going to use into factors, the statistics should be summarized as below:

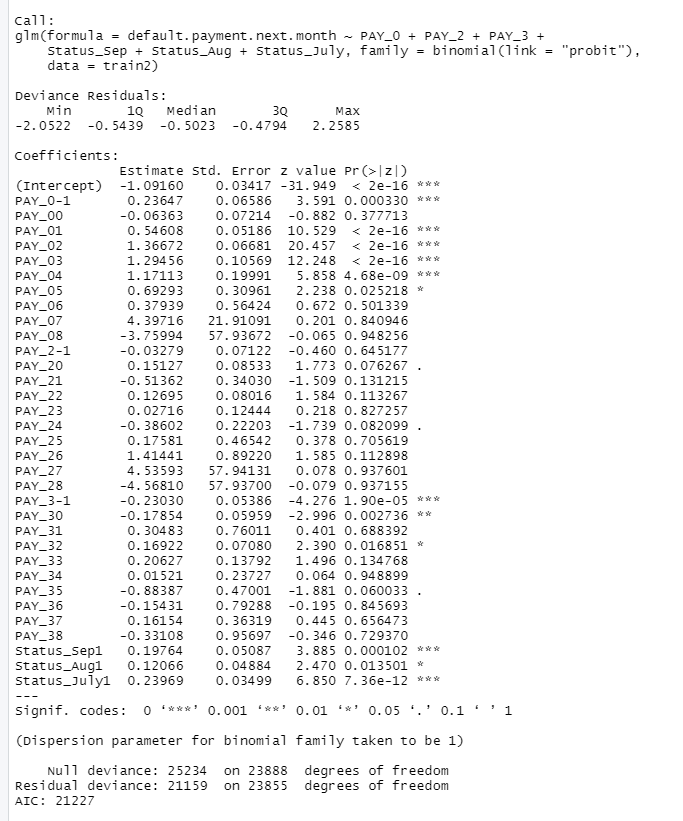
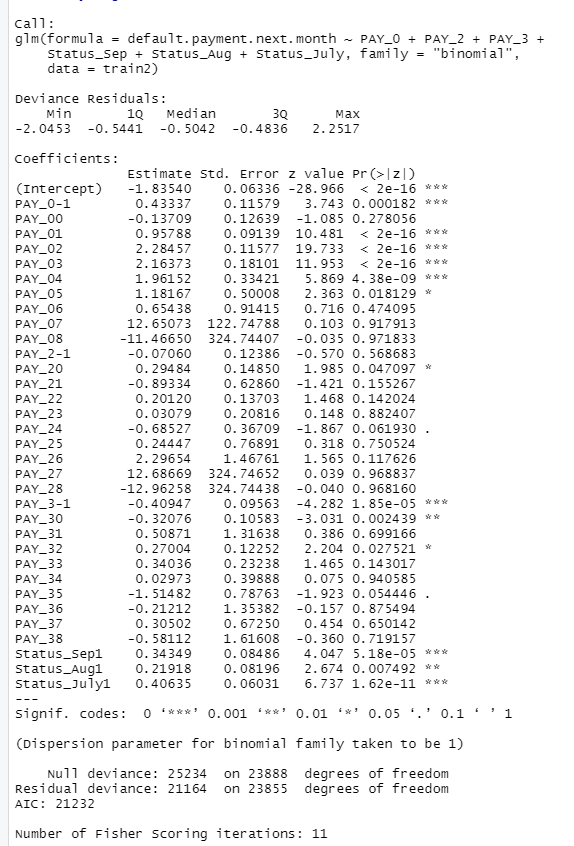


**Exploratory Analysis:**

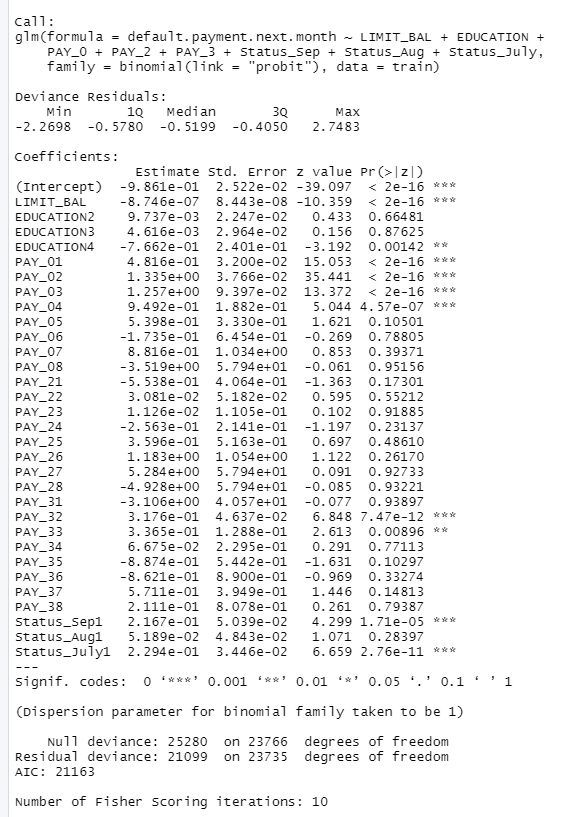
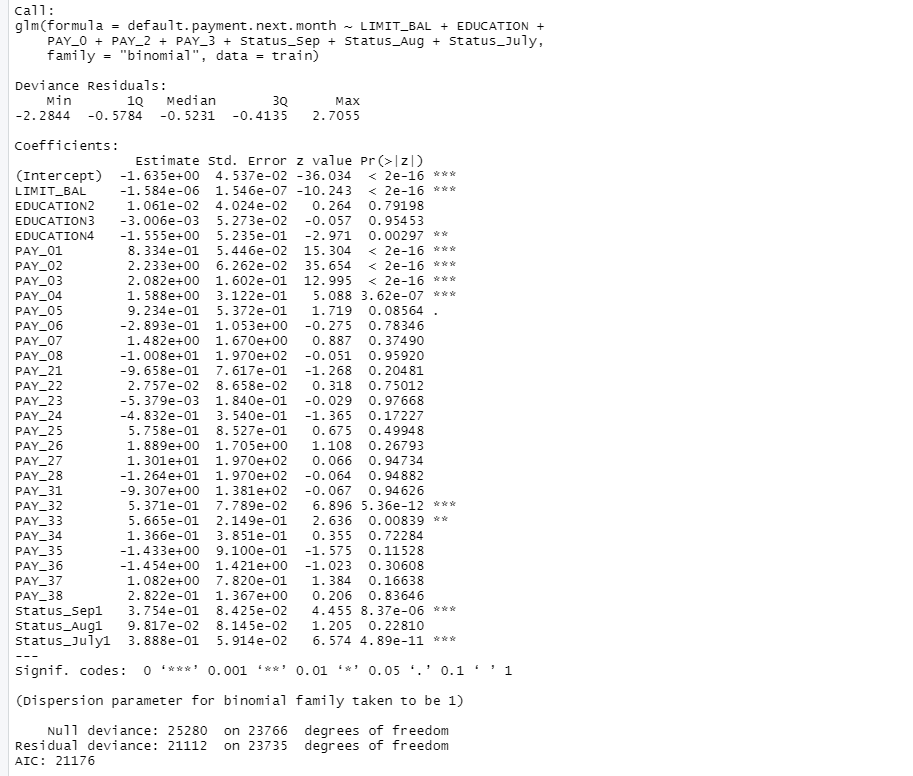
It is likely that education could help predict the default payment for next month because it may affect the levels of understanding on credits and how to manage credits in general. It may also have an impact on the levels of income, which can have an indirect relationship with default payment. Moreover, the amount of credit may be a good predictor because credit score may reflect past default status. More specifically, as the numbers of default payment increase, the credit score will probably decrease. In order to test out whether these two variables have statistical significance in building the model, I decided to compare the results with and without them.

For my first attempt, I used six variables mentioned above for my logistic and probit regression models: “Pay\_0”, “Pay\_2”, “Pay\_3”, “Status\_Sep”, “Status\_Aug”, and “Status\_July”. The logistic model returned an AIC of 21232, and probit model had a slightly lower AIC of 21227, referring to the regression tables of the two models below. Both logistic and probit models gave a very high accuracy level of 81.96% and 81.82% respectively. The result also suggested that the repayment status in the past months as predictors of the next default payment are effective with a low p-values. This occurs because they record the past payment behavior of the card holders , which can help predict the future performance. For example, if the past three payments are paid duly, the variable suggests the next payment probably will not default either.

After running the data for several times, it was found that probit model always has a lower AIC score, though in very small difference, and the accuracy of both models is within the range of 80% to 82%.



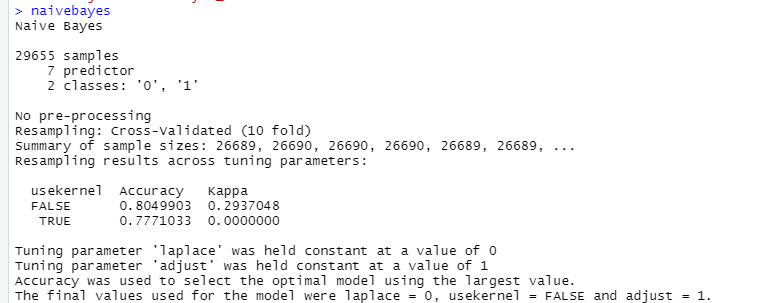
In order to find out whether education and the amount of credit can be the good predictors on the default rate or not, I tried to incorporate the two variables, “EDUCATION” and “LIMIT\_BAL”, with the previous six variables to build a new model. As shown in the table on the left corner, the amount of credit has a p-value smaller than 5%, suggesting it is statistically significant in being a predictor of the probability in default of the next payment. Again, the AIC score of logit regression is slightly higher than that of probit regression. However, the overall accuracy improves to 82.00% and 82.03% respectively. As this group of predictors has a higher accuracy on average compared to the previous attempt, the probit model, which takes the impact of education and credit score into account, seems to be the model that fits the best with the data. Although education has a p-value larger than the significance level, I decided to keep this predictor because the model still maintains a higher accuracy, and the variable itself has a higher practical significance than its statistical significance.



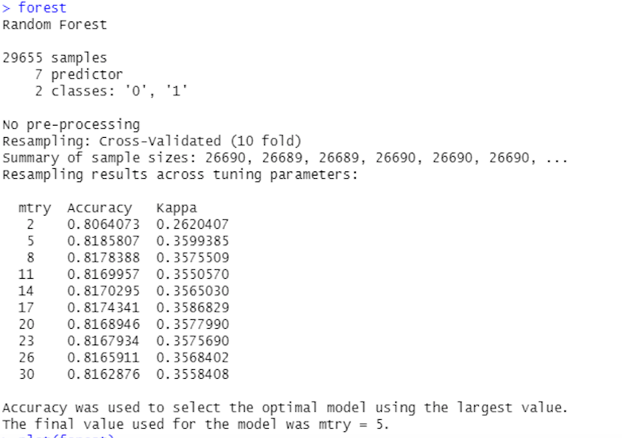
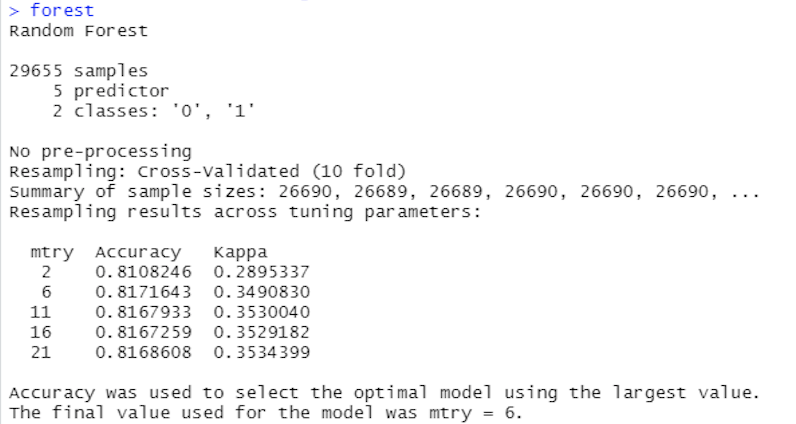
**Cross-Validations:**

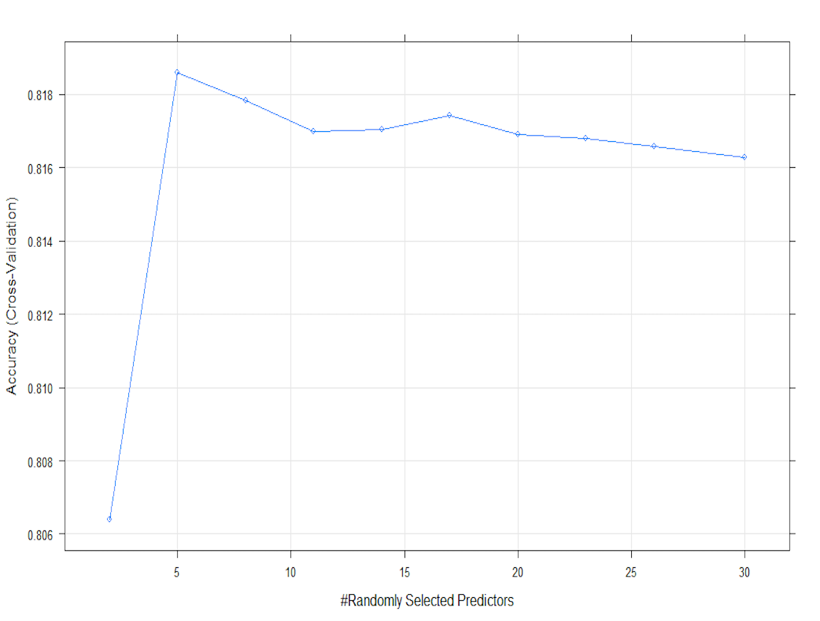
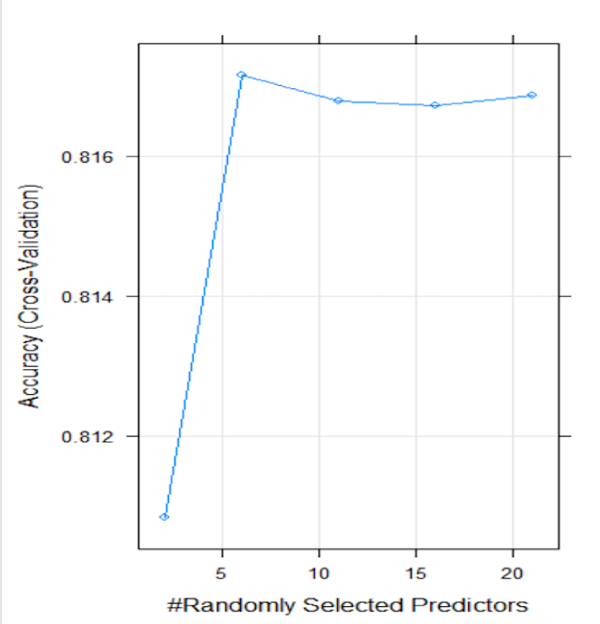
I tried KNN, Decision Trees, Naive Bayes, Random Forest, and SVM Linear model to see which one is the best cross-validation models. When I was trying to run KNN and Decision Trees Model, 50 error messages came up and I turned to TA for help. We tried to incorporate different variables and change the folds, while none turned out to be working, so I decided to use other cross validation methods.

For Naive Bayes Model, I set the folds to be 10 and I used all of the categorical variables mentioned above. I removed the numeric variablethe amount of credit, as it has some incompatibility with R. It turned out the accuracy to be 77.71%, which was acceptable, though it did not achieve the high accuracy shown in our logistic and probit regression. I removed some variables that seemed less significant out of the model, such as “Pay\_2” and “Status\_Aug”, but the accuracy level remained exactly the same with the original one.



For Random Forest Model, I set the folds to be 10. In my first attempt, I changed the tunelength to be 3, and used only three variables “Pay\_0”, “Status\_Sep”, and “Status\_July”, which are the ones with the highest statistical significance according to the probit model. The best accuracy for this attempt is 81.63%, which is already significantly higher than the Naïve Bayes Model. Next, I experimented and increased the tunelength to 5, and it achieved a higher accuracy of 81.72%, as shown in the table on the left. I came up with the hypothesis that a higher tunelength may bring us a higher accuracy, so I decided to further increase the tunelength to 10 while using all of the variables. It turned out this hypothesis was correct and I got an 81.858% of accuracy as shown in the table on the right, which is the best figure among all of the attempts I had for this model.





As running an SVM Linear Model took 5 hours on my device, I only ran it for once. Because the data size of 30000 is large and it should take a while to run, I reduced both the folds and tunelength to be 5. At the same time, I used only 5 variables in the model with the highest statistical or practical significance, including “education”, “Pay\_0”, “Pay\_3”, “Status\_Sep”, and “Status\_July”. The accuracy appeared to be the highest among all of the cross-validation models, showing as 81.864%. This figure is slightly better than the previous Random Forest Model with 10 folds, though using less variables and folds. In this sense, it seems like SVM Linear Model has the best prediction power and accuracy just behind logit and probit model. It is possible for us to get an even better result if I can increase the number of folds, tunelength, and predictors in the future.

