MMA860 Final Project

Technical Appendix

Team Alfred

# Overview

This technical appendix describes the method used to analyze credit card customer risk level from the dataset given. We, AlfredExpress, as a credit card company would like to predict the likelihood of customer default based on the information supplied in their applications.

We are given two datasets containing customer information from credit card applications and their credit history after obtaining credit cards from us. Based on the information given, we categorized customers into high-risk and low-risk. This will be helpful for us, as a credit card company, to make decisions on the approval of an application from potential customers.

# Datasets

The first dataset consists of information such as gender, age, income, marital status, family size, mortgage information, occupation, etc. The second dataset provides information on repayment history. Naturally, the second dataset is the one we would like to focus on as these are the customers with approved credit cards applications.

# Data Cleaning

We first start our analysis by cleaning the data. As mentioned above, we base our analysis on the customers from the first dataset. We need to filter out the customers not found in the second dataset from the first dataset to merge the two. We did that by left joining the two datasets, which left us with only the rows with IDs found in the second dataset. We also noticed that some of the unique IDs are duplicates as all other information is the same, so we removed the duplicates.

We then adjusted the birthday to age in years as a positive integer as we feel the actual date of birth should not affect their credit risk, and age should suffice. The same logic applies to years worked, and we made the same changes to that variable.

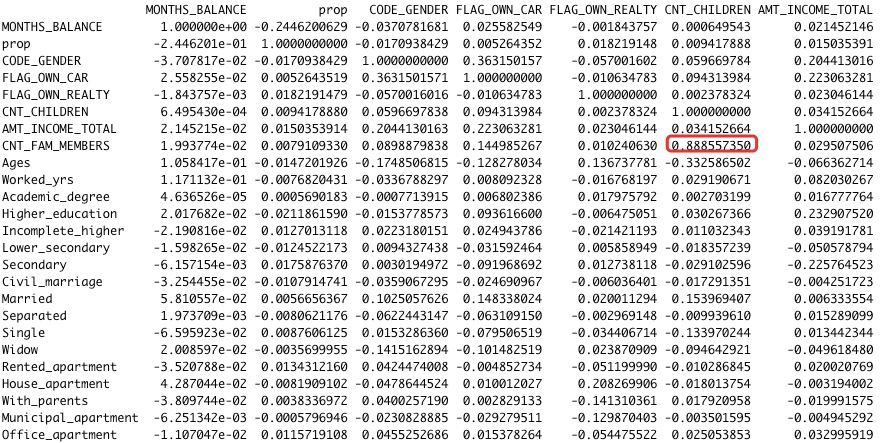
We checked the Mobile phone variable and decided to drop it as all observations have a value of 1 and it does not have predictive power. We also dropped the variables which indicate whether a customer has a work phone, landline, or email as the information is supplied by the customers and the accuracy cannot be guaranteed. Also, intuitively, the credit risk level of a customer should not be affected by whether they have an email address or a landline. We made a business decision to exclude them.

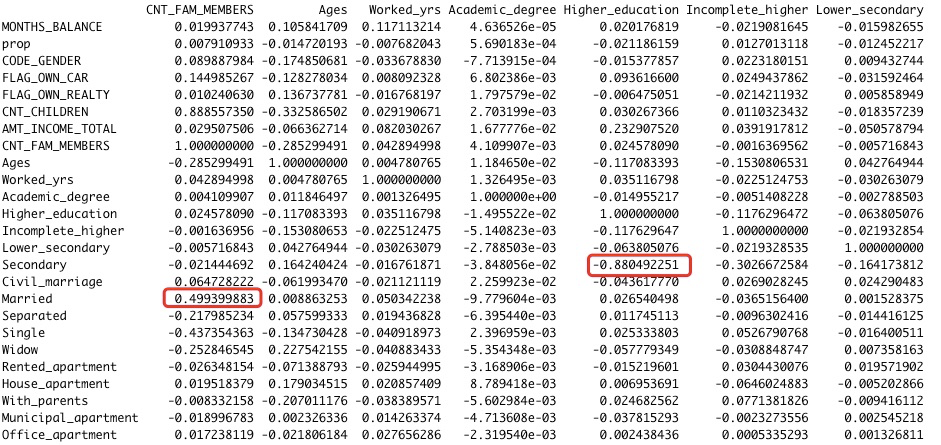
# Data Exploration and Feature Engineering

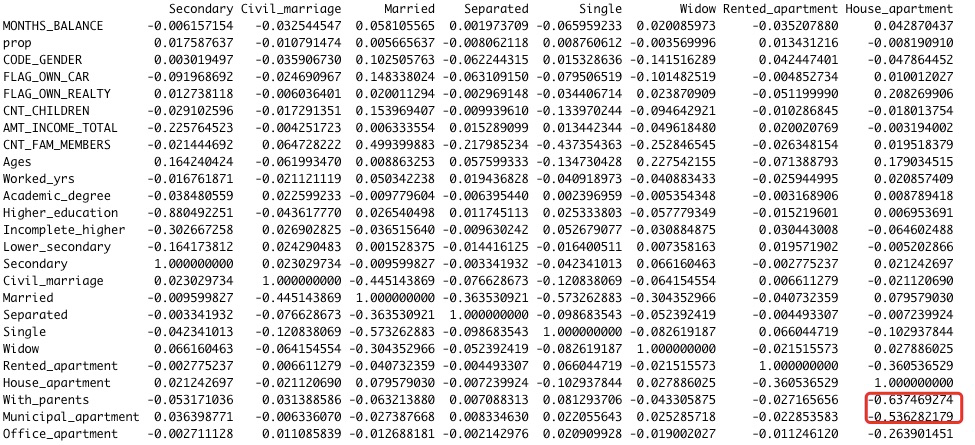
We assigned dummy variables to categorical variables for easier analysis. For gender, car ownership, and house ownership, we assigned 0 and 1 for the two categories. For education, marital status, and house types, we created n-1 dummy variables than the number of categories for each variable.

For modeling purposes, we presume that customers with a credit card history of fewer than 6 months would not provide us with useful insight and decided to exclude them from the dataset. We decreased the size of our dataset by about 15%, but we believe this will help us improve the accuracy of our model.

Now that we have a clean dataset with only numerical values and dummy variables, we can run a correlation analysis between the independent variables. From the results, we see that number of children and family size have a correlation of 0.89, married and family size have a correlation of 0.50, and that secondary and higher education have a correlation of -0.88. As for living arrangements, house apartment also has a high negative correlation with both living with parents and municipal apartment at –0.64 and –0.54 respectively. All these highly correlated variables are what we expected because each pair is in the same category. We take note of these variables and will identify and correct any collinearity issues once we get to the modeling process.

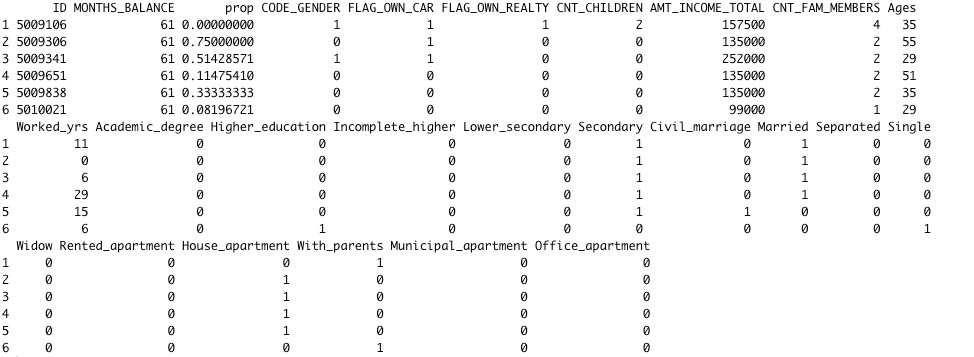






We also identified the outliers by calculating the 0.001 and 0.999 intervals for all the variables. If either interval has a value that does not make sense in the real world, we exclude the observation from the dataset. For example, the higher bound for number of children is 8. We dropped the observations with 11, 14, and 20 children as they are more likely to be typos.

To determine the customer risk level, we need to have a measure for their credit card payment history. We divided the number of months that a customer makes payment late by the total number of months that a customer has held the card to get a proportion of late payments. This gives us a continuous numerical value to measure the credit risk for any customer.



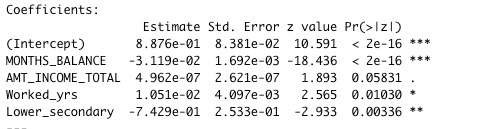
# Predictive Modelling

We considered using linear regression to assign a categorical value of “good,” “bad,” and “ambiguous” for each customer. However, the challenge is that linear regression is not the best tool for predicting categorical values. We would need to subjectively set a threshold for the three categories which could affect our model’s performance if the thresholds are not accurately selected.

We then decided to use logistic regression as it is good at predicting categorical outcomes. We also changed the outcome to the binary of high risk and low risk for simplicity, so we only need to use the basic form of logistic regression.

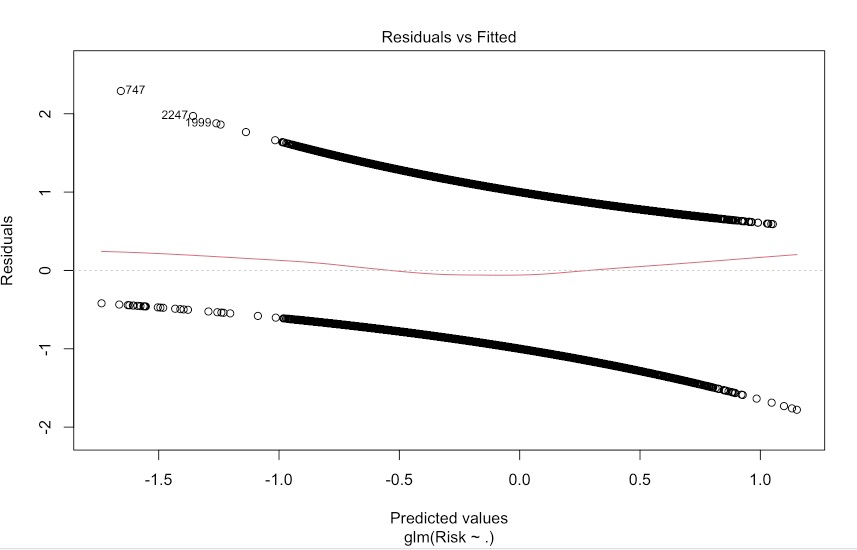
From the last step, we have a proportion of late payments. To categorize a customer as high risk or low risk, we selected a threshold of 0.4. We usually see a penalty in the credit score when there are more than 3 months of late payments in a year, which translates to a threshold of 3/12 = 0.25. However, our data is on customers who already have a credit card, meaning that they have been approved for credit before. Therefore, we would give them the benefit of the doubt and increase the threshold to 0.4.

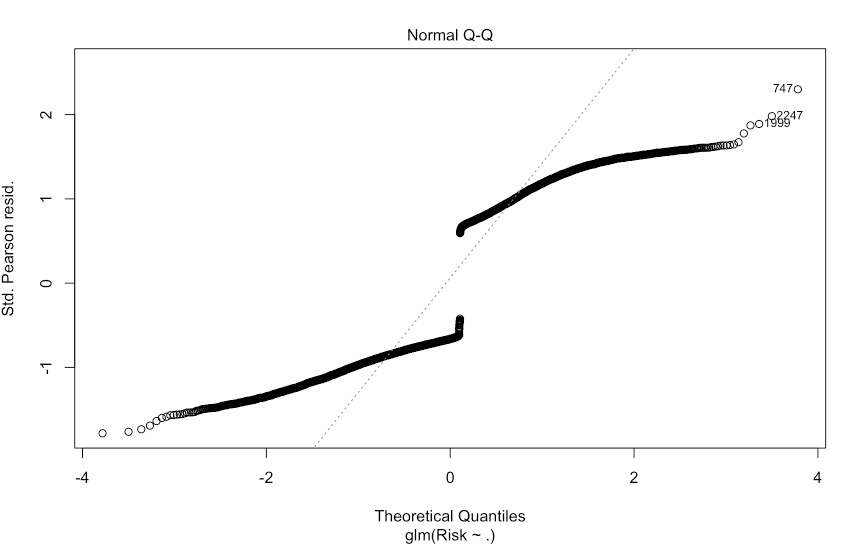
We used the TTT framework and went from the top down to filter out variables. In the full model, we included all the independent variables and only the variable months balance is statistically significant. We then started removing independent variables from the most insignificant ones. With the collinear variables, we tried leaving either one of the pair in the model but none of them had explanatory power, so we are comfortable with removing all of them. In the end, we were left with account history in months, years worked, and one of the dummy variables - lower secondary school as well as income which had a p-value of 0.05831. Using the usual 0.05 threshold, this would not be statistically significant, but we felt from a business perspective, income is one of the most important criteria we look at when approving credit card applications. Given that the p-value is very close to 0.05, we decided to include it in our model.

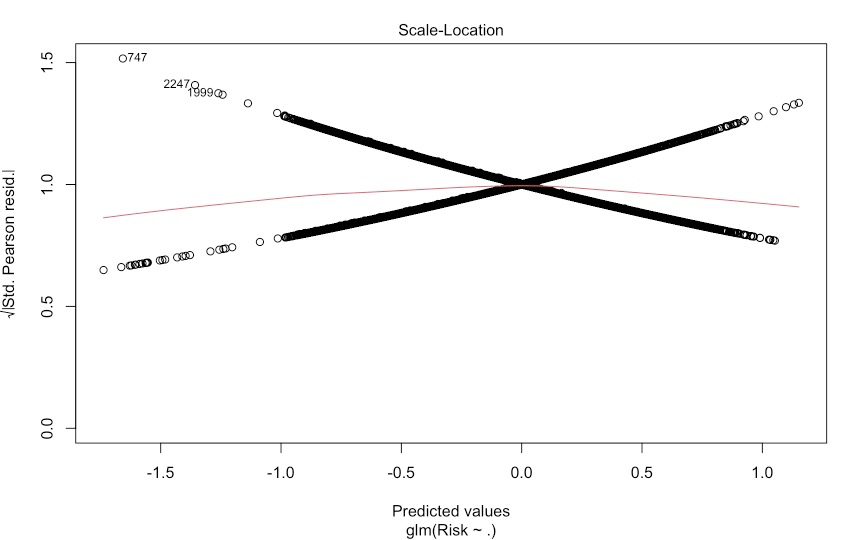


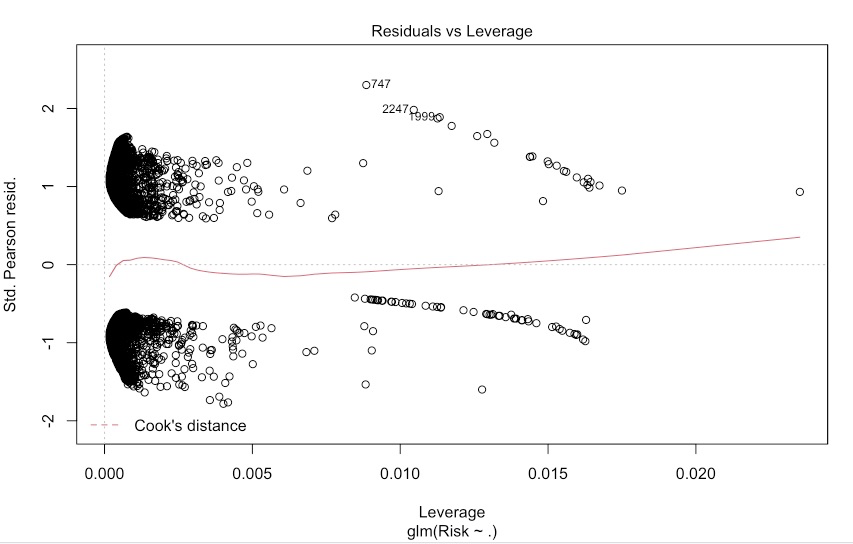
# Hypothesis Testing

Our null hypothesis is that none of the independent variables have explanatory power on whether a customer would default on their credit card loan. From the summary of our regression above, we can see that the account history in months, years worked, and lower secondary school all have a p-value of 0.05 and below, suggesting that they are statistically significant, allowing us to reject the null hypothesis in favor of the alternative hypothesis. If we relax the 0.05 threshold a little bit, income also has predictive power. This also aligns with our expectations from a business perspective. It is plausible that income, years worked which speaks to stability of someone’s job, and lower secondary school which means very low education level are all correlated with someone's credit risk level.









# Challenges

* Given that logistic regression is outside the scope of this course, we had a difficult time understanding the concept and how to apply it to our problem. We could have made some mistakes without realizing them when applying logistic regression to the dataset. We also have difficulty interpreting the residual plots as they look different from the ones in linear regression. We decided on a threshold of 0.4 which is subjective and could influence the accuracy of our model. By using this threshold, we circumvented the problem of class imbalance. However, we know that in real life, it is more likely that most people would fall into the low-risk category. Also, the use of this measure may not be the best solution to this problem.
* The technique we applied to the problem may be inadequate as this problem requires more complicated solutions. We did not touch on techniques that are better suited for vintage analysis. Machine learning may be better suited for the problem using algorithms like random forests, decision trees, and KNN.
* The dataset we were given is from a real credit card company. Like any other dataset from real life, there are missing and wrong values in the dataset. The imputation technique we learned in class does not apply here as the values missing are not numerical. We have also eliminated some of the wrong values by looking at outliers. Despite our best efforts to clean the data, there are still problems in the dataset that we were unable to identify and correct.
* Furthermore, the dataset itself does not contain much useful information except for the ones we identified as statistically significant. We know that in real life, credit card companies look at credit scores which are influenced by credit history, loans, spending habits, number of accounts open, etc. A lot of the data are not available in this dataset. For example, we only know whether the applicant has a car or has a house but there is no information on the amount of their car lease/loan or mortgage. The exact amount of their past due payments is not given, only the number of months an account is behind on payments. For a credit card company, $10 past due for half a year is less likely to be problematic than $10000 for two months. Without this information, it is difficult to make meaningful predictions about customer risk level.