Give Me Some Credit: A Capstone Project

Sean Anthony

Email: [sean.marc.anthony@gmail.com](mailto:sean.marc.anthony@gmail.com)

1. Introduction

One of the most important aspects of the banking industry is the issuance of credit. Banks determine which prospective borrowers are worthy of being financed based on a vast amount of data at their disposable. As is the case with most businesses, these banks seek maximum profit. In order to maximally profit, banks need to accurately predict the probability that these potential borrowers will pay back their loans in a timely manner. By studying a sample data set describing the history of past borrowers, my goal will be to more accurately asses the risk these banks face with each unique future lending opportunity.

1. Describing the Data

Two data sets were issued:

1. A training set consisting of 12 variables and 150,000 observations.
2. A test set consisting of the same 12 variables and 101,503 new observations.

A description of these variables:

1. Subject Identification (id): for organizational purposes, and will not provide any relevant information for the models.
2. Revolving Utilization (rev\_util): a ratio dividing the amount of credit currently being used by the maximum amount of credit this borrower currently has been provided, i.e. a subject with two lines of credit totaling $10,000 available with $2,000 currently being used will have a revolving utilization of .20.
3. Age (age): the age of the subject
4. Past Due by 30-59 days (past30): the number of times this subject has been past due by 30-59 days
5. Past Due by 60-89 days (past60): the number of times this subject has been past due by 60-89 days
6. Past Due by 90+ days (past90): the number of times this subject has been past due by 90 days or more
7. Debt Ratio (debt\_ratio): a ratio equaling the subject’s total amount of current debt divided by his current assets
8. Monthly Income (monthly\_inc): a subject’s current monthly income
9. Open credit lines and loans (open\_lines): a subject’s lines of credit and loans, excluding real estate
10. Real estate loans or lines (re\_lines): a subject’s number of real estate credit lines and loans
11. Dependents (depend): a subject’s number of current dependents
12. Serious Deliquency in 2 years (delinq)(dependent variable): the probability that this subject will be seriously delinquent, defined for this example as being greater than or equal to 90 days past due, within the next two years
13. Cleaning the Data

Immediately upon seeing the size of this data, it became probable that outliers and missing values would be an issue. Fortunately the situation could have been worse. In order to provide myself with more normalized distributions, the top .3% of many variables were filtered out. These variables included the revolving utilization, debt ratio, amount of open lines, and the amount of real estate lines. I still excluded this top .3% for the other variables, but also had to make additional changes. For age, one observation of ‘0’ was filtered. Additionally, all of the past due variables also had many observations of both ‘96’ and ‘98’. Considering all three of these variables were mistakenly marked for the same observations, I gave it a thought to change these to 0 considering this was overwhelmingly the most popular result. Instead, I decided to omit these due to their inconsequential amount.

Two variables had missing values, all marked as ‘NA’. For the number of dependents, I decided it was most likely that a subject’s ‘N/A’ for this variable meant that they had no dependents. Therefore, these values were changed to ‘0’. The monthly income variable was more challenging due to the range of values and the amount of missing data. My initial inclination was to omit these observations. However, I noticed this would leave me with about 20% less observations to work with. Even though the training set was large, I wanted to make use of as many observations as possible. Using the impute function, I randomly imputed these ‘N/A’ observations using the values from the rest of the training set. I then filtered out the top .3% as I had done previously. This left me with a more normalized graph, ranging from $0 monthly income to $45000. Finally, I took a look at the dependent variable to see the binary outcome ratio.

1. Regression Models

Due to the binary nature of the project’s dependent variable (1 – expected to be delinquent, 0 – not expected), I thought it best to begin with logistic regression. I split the training data into a further training (60% of the data set) and test group (40%). This was to become more confident in my Area Under the Curve numbers before submitting the probabilities for the actual test set. These were randomly assigned with the dependent variable being equally distributed within each subset.

I viewed the correlation between all variables, and decided to being my modeling with the one independent variable most correlated with the dependent variable – past due by more than 90 days. Reviewing the summary of this model, the large positive coefficient made me confident there this would be an improvement on random guessing. This was proven correctly as the AUC = .656937.

To improve upon this first model, I chose another relatively highly correlated variable - revolving utilization. I could have chosen either of other past due variables, but my fear of multicollinearity veered me in a different direction. After again checking the summary and seeing both variables demonstrating signs of strong significance, I hoped for a large improvement. I happily found an AUC = .807918 for the test set.

For the third model, I decided to take some time and realize whether my fear of multicollinearity was substantive. In addition to being similar to the naked eye, the three past due variables were extremely strongly correlated with each other (~.98). My third model consisted of the second model variables plus the other two past due variables. To my surprise, this gave me an AUC of .842166, another significant improvement.

Realizing the correlations of the other variables were much less significant, I decided my fourth and final model would disregard the least significant past due variable from the last model. I decided to filter past30 based on it’s less positive coefficient in relation to the others. Unfortunately, this model provided me with only an AUC = .822529, so the third model remained best.

1. Conclusion

My models have provided a pathway for banks to more accurately predict a future prospective borrower’s probability of serious delinquency over the next two years. Maybe more importantly, it’s provided a succinctness to the process. Banks should mainly focus on the history of a subject’s punctuality when paying their monthly statement. History seems to repeat itself in this regard. Additionally, banks should focus on the amount of credit a subject decides to use relative to the amount at their disposal. By using only four variables, with three closely related, a bank can more correctly allocate its research and development resources to focus on the data that matters the most.