Data 610 – Fall 2017

Assignment 2 – Exploratory Data Analysis of Mortgage Defaulters

Kenneth Lulie

Dr. Moretto

University of Maryland University College

October 15, 2017

**Initial Review of Data Set**

The data set selected for this exploratory data analysis was a database of mortgages provided as a class resource. This data set has 15,154 observations with a total of 16 variables per observation, and appears to have been created using data available at the time of a mortgage’s creation, along with the current status of the mortgage (active, paid-off, or defaulted). This data set was presumably created to assist a mortgage lender in understanding likely causes of mortgage defaults, or to assist an investment bank in understanding what mortgages would be safest to purchase.

The variables that could be used in a predictive model include information regarding the borrower, such as age, total monthly income, total monthly debt expense, credit score, and if the borrower was a first time home buyer (Geffner, 2015). They also include characteristics of the mortgage, such as the original loan amount, the original appraisal value of the house, the purchase price of the house, the state the house is located in, and the current status of the mortgage as either paid off, defaulted, or active. Additionally, the data set includes data derived from other fields, such as the ratio of the value of the loan to the value of the house (LTV). A significant limitation in the data set was lack of information regarding dates, which is important when considering how economic trends may have impacted the data.

**Error & Mistake Correction**

When reviewing the data for mistakes, bad records, data entry errors and outliers it was noted that the credit score for 11 observations was over 850. This appears to be an error as the maximum credit score assigned by any credit bureau in the US is 850 (Detweiler, 2016). The data also contained 19 records with an appraisal value of 0. As the data set has over 15,000 records these 30 observations were removed from the data set since the number is de minimis compared to the total (Osborne, J., & Overbay, 2004).

Additionally, the dataset included an untitled column that was populated as a formula. To improve the data set, this column’s format was changed to values, and the column was titled appropriately. A blank column in the table was also removed, along with one cell containing a calculation. The rest of the data was appropriately formatted for the data type. The final data set after editing has 16 variables and 15,124 observations (see Table 1).

**Initial Exploration**

The purpose of this data set is presumably to gain insight into which factors known at the time of a mortgage application lead to a higher risk of mortgage default. The initial questions posed by Watson Analytics did not address this purpose (see Figure 1), as the questions focused on derived fields in the data with none of the questions relating to loan defaults. For example, the lead insight asked, “What are the values of Credit\_score by state”, and used the sum of the credit scores. The default aggregation was changed to average resulting in a heat map of average credit scores by state (see Figure 2). This heat map appears to show geographic trends, such as higher average credit scores in the Northwestern United States.

To gain insight into the purpose of the data exploration (understanding and potentially predicting loan defaults from original application data), Watson Analytics was asked to “Show me distribution of loan status” (see Figure 3). Watson provided a bar chart, which was an appropriate visualization for comparing these groups. The bar chart showed that, out of a total of 15,123 observations, 6,849 mortgages had already been paid off, 7,873 mortgages were still active, and only 402 mortgages had defaulted. This data was then filtered using a local filter to show only first time homebuyers ( see Figure 4), and the bar chart showed that 2,347 active mortgages, 1,699 paid off mortgages, and 116 defaulted mortgages belonged to first time homebuyers. From this, it was calculated that the default rate for the general population was 2.65%, whereas the default rate for first time home buyers defaults was 2.79%.

**Understanding Mortgage Status**

Understanding that the data is divided into three mortgage status categories of paid off, defaulted, and active, it’s a natural progression to ask questions about the differences of the mortgages in each category. Since credit scores are heavily relied on in the mortgage industry (Geffner, 2015), Watson Analytics was asked, “Show me Credit Score by Status” (see Figure 5). Watson Analytics provided the answers in a standard bar chart, which was an appropriate visualization for comparing these groups. Interestingly, the bar chart showed that active mortgages and paid-off mortgages had almost identical average credit scores, with an average of 688.57 for active mortgages and 688.41 for paid off mortgages. Defaulted mortgages had an average credit score of 654.23, which indicates that credit score should be a good variable to use in predicting defaults, as the difference in averages is quite striking in a large data population.

To explore the role of credit further, a global filter was created to show observations where the credit score was under 630, which is considered a poor credit rating (Detweiler, 2016). Using this filter, it was calculated that 125, or 4.81%, of the 2,594 mortgages associated with poor credit scores had defaulted compared to 2.65% of the general population. Next, Watson Analytics was used to produce a chart comparing mortgage status with credit score grouping (See Figure 6). The grouping was calculated by equally splitting the range of credit scores into five groups which showed the heaviest distribution of mortgages was in the range of 594 to 747. Improvements to this grouping could be explored by using a count of observations rather than equally splitting the range of the variable.

The use of the hierarchy function in Watson Analytics was also explored, but ultimately decided against, as there was no information on the types of mortgages or any time related data. To improve the outcome of the hierarchy feature, new fields could be added to the data.

**What Drives Defaults**

The next stage in the data exploration was to gather information on what combinations of variables could help predict mortgage defaults. Watson Analytics has built in analytic tools that create decision trees, which use rules to categorize data (IBM, 2017). Watson was asked “What is a predictive model for status”, and provided a decision tree to assist in understanding the rules developed (see Figure 7). In its first attempt, the decision tree simply used the “outcome” variable to predict defaults, and then compared paid off mortgages to active mortgages. While these results were interesting, they were not in the focus of the data exploration which is to understand what causes loan defaults, as active mortgages should not be used for predictions because their outcome is still unknown.

The inclusion of active mortgages in the data was shown to be a distraction in this data exploration. To get more relevant insights into what causes defaults, the data set was filtered to exclude observations where the status of the mortgage was active. Additionally, the Outcome variable was set to be excluded in predictions, and the Status variable was set to be the default prediction target. Using this new data set, Watson Analytics was asked “What is a predictive model for Status” (see Figure 8) and provided a decision tree model to predict between paid-off versus defaulted loans.

Due to the extreme imbalance in the number of observations between paid off loans (6,849) and defaulted loans (402), the decision tree maximized accuracy by predicting a paid-off status in every category. Useful insights were nevertheless gained from the accuracy statistics for each node, as the nodes with the lowest accuracy led to insights on the riskiest combinations of variables for mortgages, and the highest accuracy gives insights on the least risky combination of variables for mortgages which could be useful for understanding causes of mortgage defaults. One insight gained is that the median income of the state was the most powerful predictive factor, possibly indicating that wealthier states were better poised to handle the period of general economic decline observed during the late 2000s and early 2010s.

To add a calculation to the data set, a new column was added and titled “IncomeToLoanMultiplier”, where the loan value of the mortgage was divided by the borrower’s monthly income. This could give a quicker indication to the mortgage company if the borrower would need outside help to make payments on their mortgage.

**Data Quality**

The data quality of the data set used was 64% as calculated by Watson Analytics. The data metrics for the variables were reviewed by the quality report function (IBM, 2017), and showed no missing values in any of the columns. However, many of the number variables had skewed data, generally with positive skew lowering the quality score. The categorical variables such as status and outcome also had poor scores due to the uneven distribution towards one category. To improve the quality of the data, the columns in number format could be transformed to correct the skew towards a normal distribution.

References

Geffner, M (2015). Mortgage Applications: A Borrower’s Guide. Retrieved October 15, 2017, from http://www.bankrate.com/finance/mortgages/mortgage-application-borrowers-guide.aspx

Multimedia: September 16, 2016). Overview of Watson analytics [Web]. Retrieved from https://www.ibm.com/communities/analytics/watson-analytics-blog/video-overview-of-watsonanalytics/

Detweiler, G (2016, December 8). What Is a Good Credit Score? Retrieved October 15, 2017, from https://www.credit.com/credit-scores/what-is-a-good-credit-score/

Osborne, J., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research & Evaluation, 9(6). Retrieved December 2, 2015 from http://pareonline.net/getvn.asp?v=9&n=6

Appendix A

Table 1. Example of 'MortgageDefaulters-mod1' data set after editing

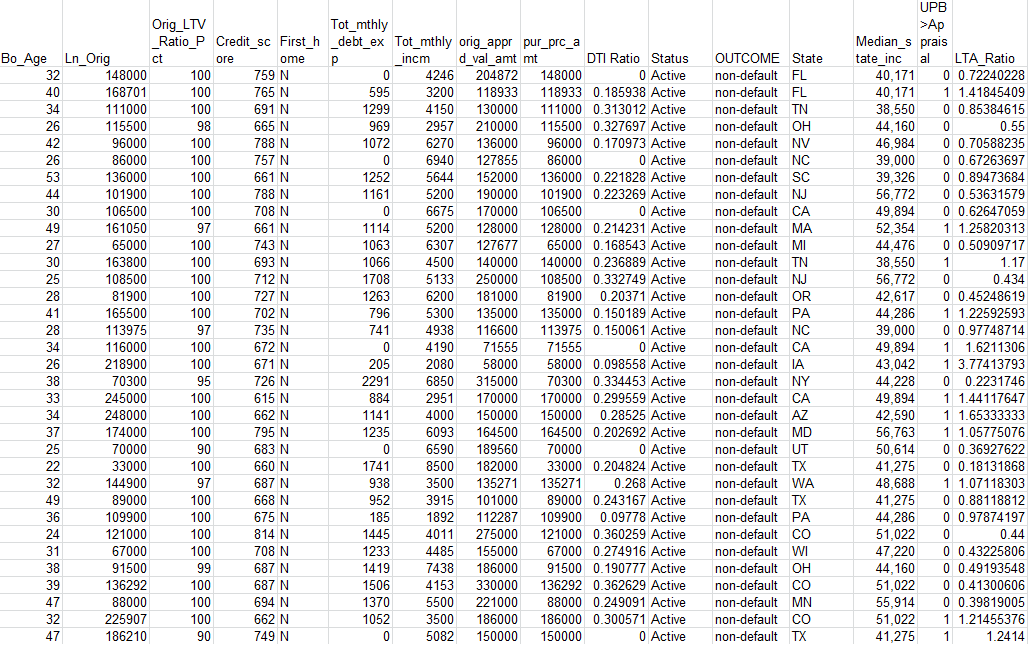
****

Figure 1. Initial Starting Points provided by Watson Analytics

