Using Logistic Regression and other Models to predict Defaults

Authors: Kito Patterson and Hayley Horn

**INTRODUCTION**

Home Equity loans are an alternative to credit card debt for home owners because it leverages a secured asset, and lower risk can mean a lower interest rate. Home equity lines of credit account for approximately 4% of debt among consumers (1). While the lower interest rate may be a benefit, there are still many default on their home equity loans, which happens when a loan is 90+ days outstanding. The value of 90+ day delinquencies is 2.23B in 2016(2). In this study we seek to uncover factors that contribute to these defaults so that we can predict an applicant’s propensity to default. This research study dives into a collection of potential variables and delinquency indicator on home equity loans. Ideally, the information gleaned from this research can help identify risk factors prior to default and giving some chance for intervention.

Our approach to this problem had three components. First, we performed exploratory data analysis and reviewed the data gathered for data filtering, transformation, and feature analysis. Second, we ran the data through two logistic regression models and determined which values had the highest probabilities of class membership. Next using the results of the logistic model as a baseline, we performed additional models to see if we could determine a predictor that performs better than the logistic regression model.

**DATA DESCRIPTION**

The data set came from a book “*Credit Risk Analytics: The R Companion, Scheule Roesch Baesens, 2017.”* (3) The data set called HMEQ contains anonymized characteristics, and delinquency information for 5,960 home equity loans. In the data set, each row represents a single consumer’s information. We will assume independence, as multiple persons within a household would not typically have multiple home equity loans on the same home. There are 13 variables, and the summary statistics for each variable are captured in Table 1. The target variable “BAD” is binary, where a 1 means a person defaulted on a loan and a 0 means the loan was paid.

There are some shortcomings in the data that limits our ability to make inferences using the data set. There is no time period for this data set, no indication on geography covered and no demographic information. That being said, the data set size is sufficient to get an understanding of variables that impact consumer defaults.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **VARIABLE** | **N** | **TRANSFORMATION** | **MINIMUM** | **MAXIMUM** | **MEAN** | **STD DEV** |
| BAD | 5960 |  | 0.00 | 1.00 | 0.20 | 0.40 |
| LOAN | 5960 |  | 1100.00 | 89800.00 | 18596.01 | 11170.30 |
| MORTDUE | 5442 |  | 2063.00 | 399550.00 | 73765.40 | 44460.41 |
| VALUE | 5848 |  | 8000.00 | 855909.00 | 101778.25 | 57390.44 |
| REASON | 5708 |  | 0.00 | 0.00 | N/A | N/A |
| JOB | 5681 |  | 0.00 | 0.00 | N/A | N/A |
| YOJ | 5445 |  | 0.00 | 41.00 | 8.92 | 7.57 |
| DEROG | 5252 |  | 0.00 | 10.00 | 0.25 | 0.85 |
| DELINQ | 5380 |  | 0.00 | 15.00 | 0.45 | 1.13 |
| CLAGE | 5652 |  | 0.00 | 1168.23 | 179.76 | 85.82 |
| NINQ | 5450 |  | 0.00 | 17.00 | 1.19 | 1.73 |
| CLNO | 5738 |  | 0.00 | 71.00 | 21.30 | 10.14 |
| DEBTINC | 4693 |  | 0.52 | 203.31 | 33.78 | 8.60 |

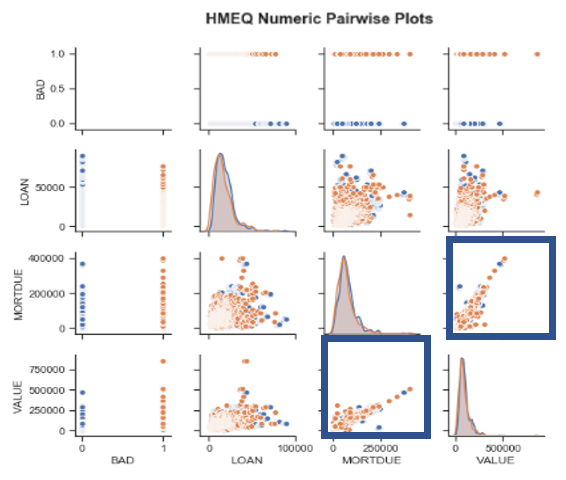
*Table 1 Summary statistics*

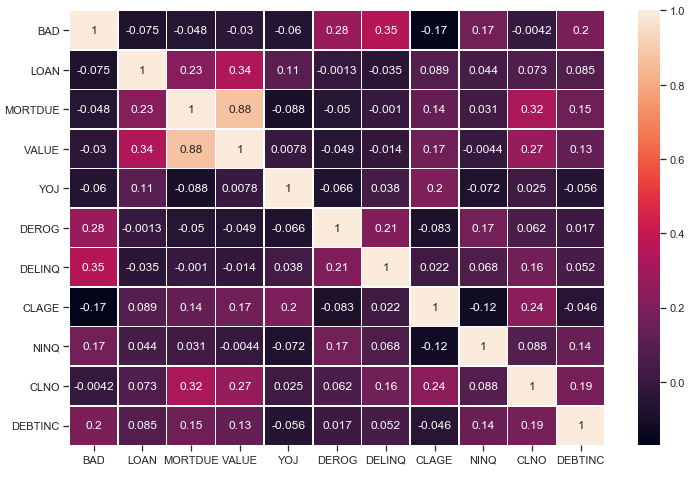
Table 1 shows the summary statistics of the final data set, comprising of 5960 observations. The fields “LOAN”,” MORTDUE” and “VALUE” have a large range of values and standard deviations. This is expected for these fields, as loan and mortgage sizes can vary within areas, and across geographies. The fields “REASON” and “JOB” are categorical and were converted to numeric values. The data dictionary can be found in Table x in the appendix.

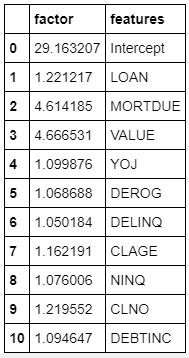
**Exploratory Data Analysis**

From the initial exploration, we can see an issue with incomplete records. We have to determine what fields need to be populated that are needed in a model. One approach is to remove all rows with at least 1 NaN as incomplete records, in which case 44% of the dataset would be excluded. Some fields, such as “DEBTINC” is missing in 21% of the records, and imputations seem a less favorable option without knowing more about how the data was populated. We chose to remove the incomplete rows, which the remaining complete records still accounted for 3,364 records. The reason for that decision is we can built the model based on the most complete information, and in later iterations include records that were formerly excluded as incomplete if they contain the fields required in the best performing model.

Next we look for multicollinearity using a scatter plot. In the scatterplot matrix, we color-coded the defaults with blue and loans that were paid as orange. “LOAN” has a long right tail, which is expected based on the summary statistics, but with a data set size of over 3,000 we are covered under the central limit theorem. We see evidence of multicollinearity with "VALUE" and "MORTDUE", because as home values increase so does the total mortgage due. We may not need both variables, but we will assess that in later modeling efforts. ”. The full scatterplot matrix can be found in Table x in the appendix.

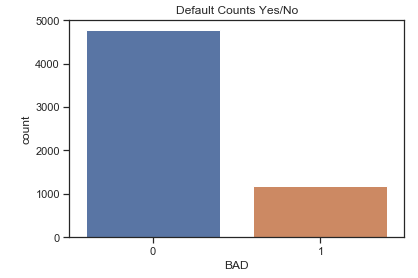


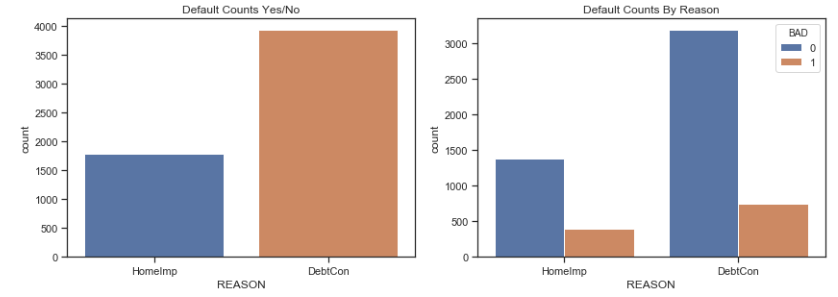
The heatmap confirms the high multicollinearity of “VALUE” and “MORTDUE.”

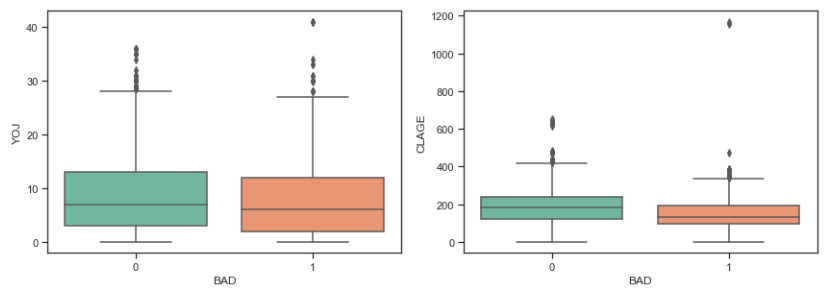


Although the scatterplot matrix and heat map provide visual evidence of possible multicollinearity, VIF values quantifies the severity of multicollinearity. As suspected, 'VALUE' and 'MORTDUE' are borderline, but they fall just under the acceptable threshold (VIF >5). In an attempt to reduce the number of variables and complexities in our model, we chose to remove the 'VALUE' variable given it is has a VIF closest to 5. This will allow us to reduce model complexity as much as possible before our model/feature selection efforts.

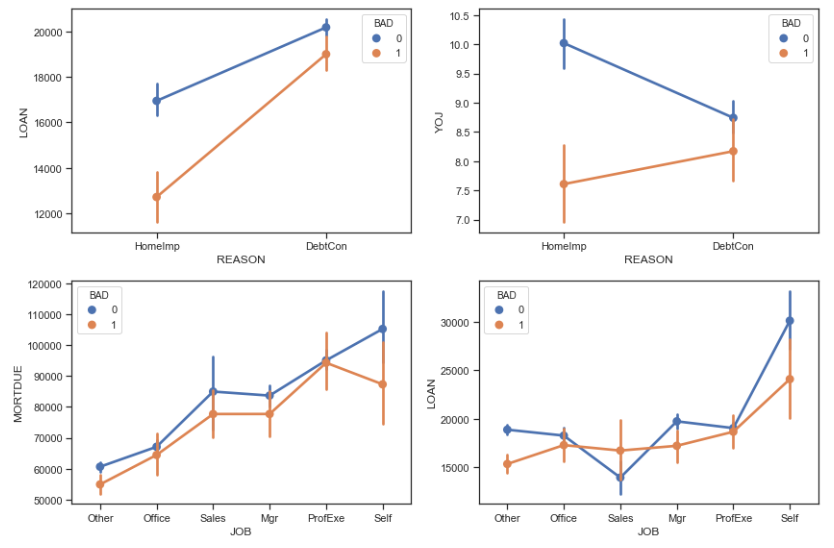
Logistic regression, like other generalized linear models is not robust to multicollinearity. Now that we have removed a variable to address multicollinearity, we continue our analysis of the data set.

Next, we look at our response variable, and it is clear we have an unbalanced data set. This is something we will need to up or down balance in our analysis.

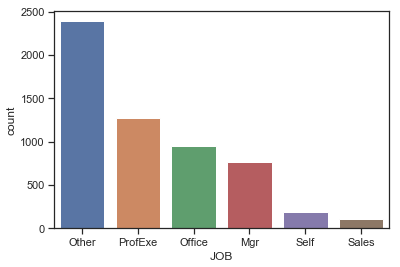
We also look at the data split by our response variable. Most borrowers use home equity loans for debt consolidation, and this is true for both the defaults and paid borrowers.

A box and whisker plot in figure X shows us that borrowers who default tend to have fewer years on the job than borrowers who don't default. This could mean that defaulted borrowers are likely less established with lower salaries that may signal more defaults. There also seems to be some separation between both response populations and "CLAGE", which is the age of oldest credit line in months. This could indicate that less credit history can potentially lead to higher defaults. It is important to note the mean for defaults are skewed by a few outliers.

There is evidence of interaction between the factor levels. This may suggest subsequent models may perform better with interaction terms included.



The distribution of the categories of job is less helpful than it could be, as most jobs fall into an “other” category which could contain a variety of entry level and higher paying positions.



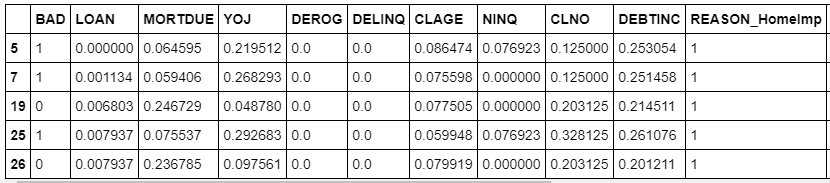
#### **Addressing Objective 1:**

**Restatement of Problem and the overall approach to solve it**

Our objective is to identify factors that are associate with defaulted loans in order to predict an applicant’s propensity to default. Since the target variable is binary, we will use binary logistic regression. We will create dummy variables for our 2 categorical explanatory variables, “JOB” and “REASON” so they can be included in our modeling.

The data will be rescaled using Min-Max so the data is scaled to a fixed range of 0 to 1. The benefit to this approach is smaller standard deviations, which can suppress the effect of outliers- which will be helpful in this dataset, which offsets the downside of the data set being bounded.

Here we can see the results of the Min-Max scaling, all values are between 0 and 1. The dummy variable created for “REASON” is also visible. The table is truncated for readability. Since all of the variables are between 0 and 1, there are no outliers that need to be removed.



DO WE NEED TO balance our data CHANGE LINE31

**Model Selection**

As we consider models, Forward, Backward and Stepwise modeling selection techniques require that you have some variable order and importance. Given this isn't our domain expertise, we sought a better alternative than arbitrary or best guess ordering. In this data set we have only 13 variables, and theorize only a few predictors actually influence the response- we opted for a LASSO regression model (LRM) as better starting place. The LRM worked as expected. 2 of the variable coefficients were shrunk toward zero, essentially removing them from the model. The Ridge model was considered, but it is best used for data that has many variables and multicollinearity, which we felt wasn’t a big issue in our data set and we had mitigated by removing a value. As a secondary model, we used Elastic Net (ENR), which is a combines the penalties of RIDGE and LASSO, but is less influenced by data than LASSO. (4)

After running both models, we determined the best model was XXX based on the metrics of Accuracy and AUC values.

|  |  |  |
| --- | --- | --- |
| **MODEL** | **Accuracy Score** | **AUC** |
| **LRM** | 93.16 | 0.803 |
| **ENR** |  |  |

Table 4 Model Comparison Data

**Checking Assumptions**

The assumptions of logistic regression is X.

We have eliminated the “VALUE” field to address multicollinearity based on the variable inflation factors (VIFs). The data are assumed to be independent.

Lack of fit test

Influential point analysis (Cook’s D and Leverage)

Residual Plots **Optional**

**Parameter Interpretation and CONFIDENCE INTERVALS**

**Final conclusions**

For our first analysis we determined the X model was best and are able to associate with defaulted loans with an accuracy of X%.

#### Objective 2: 2-Way ANOVA

#### **Addressing Objective 2:**

Make sure it is clear how many models were created to compete against the one in Objective 1. Make note of any tuning parameters that were used and how you came up with them (knn and random forest logistics) **Required**

**Main Analysis Content** Required

Overall report of the error metrics on a test set or CV run. Also if the two best models have error rates of .05 and .045, can we really say that one model is outperforming the other? What other tools that we learned in the second half of this class that could help us get at that?

**Conclusion/Discussion** Required

The conclusion should reprise the questions and conclusions of objective 2 with recommendations of the final model, what could be done to help analysis and model building in the future, and any insight as to why one method outshined all the rest if that is indeed the case. If they all are similar why did you go with your final model?

**Appendix**

Data Dictionary

This data dictionary provides details on the data used in our study from the HMEQ datasource.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Description** | **Levels** | **Recode** | **Class** | **Unique values** | **Missing** |
| BAD | 1 = applicant defaulted on loan or seriously delinquent; 0 = applicant paid loan | 2 |  | integer | 2 | 0% |
| LOAN | Amount of the loan request |  |  | integer | 540 | 0% |
| MORTDUE | Amount due on existing mortgage |  |  | numeric | 5054 | 8.69% |
| VALUE | Value of current property |  |  | numeric | 5382 | 1.88% |
| REASON | DebtCon = debt consolidation; HomeImp = home improvement | 2 |  | factor | 3 | 0% |
| JOB | Occupational categories |  |  | factor | 7 | 0% |
| YOJ | Years at present job |  |  | numeric | 100 | 8.64% |
| DEROG | Number of major derogatory reports |  |  | integer | 12 | 11.88% |
| DELINQ | Number of delinquent credit lines |  |  | integer | 15 | 9.73% |
| CLAGE | Age of oldest credit line in months |  |  | numeric | 5315 | 5.17% |
| NINQ | Number of recent credit inquiries |  |  | integer | 17 | 8.56% |
| CLNO | Number of credit lines |  |  | integer | 63 | 3.72% |
| DEBTINC | Debt-to-income ratio |  |  | Numeric | 4694 | 21.26% |

EDA Scatterplot MAtrix



LRM Model

|  |  |
| --- | --- |
|  |  |
| Figure 1 Confusion Matrix for LRM | Figure 2 AUC for LRM |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | LOAN | MORTDUE | YOJ | DEROG | DELINQ | | -0.93095113 | 0 | -0.1935585 | 6.26522199 | 6.33414015 | |  |  |  |  |  | | CLAGE | NINQ | CLNO | DEBTINC | REASON\_HomeImp | | -5.93777339 | 1.02862774 | -0.34175919 | 10.91766626 | -0.05152882 | |  |  |  |  |  | | JOB\_Office | JOB\_Other | JOB\_ProfExe | JOB\_Sales | JOB\_Self | | -0.60534424 | 0 | -0.17423549 | 0.93146919 | 0.39305138 |   Figure 3 LRM Coefficients | |

### **REFERENCES**

1. <https://www.statista.com/statistics/944954/personal-debt-source-usa/>
2. <https://www.statista.com/statistics/681709/value-of-90-delinquent-heloc-balances-usa/>
3. Data <http://www.creditriskanalytics.net/datasets-private.html>
4. Ridge, lasso and Elastic: <https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net> <used, make sure it is in my words>

**CODE**:

**POSSIBLE referenceS**

<https://jupyter.brynmawr.edu/services/public/dblank/CS371%20Cognitive%20Science/2016-Fall/Simple%20logistic%20regression.ipynb>

<https://git.generalassemb.ly/BAH-DC-1/logistic-regression/blob/master/logistic-regression-starter.ipynb> <-yas

This could also benefit consumers, if they have high indicators for default, it could be an opportunity for consumer debt counseling or other intervention to give them an opportunity to course correct before they find themselves in default

####FEATURE SCALING AND BALANCING

<https://sebastianraschka.com/Articles/2014_about_feature_scaling.html>

std\_scale **=** preprocessing**.**StandardScaler()**.**fit(X\_train)

X\_train **=** std\_scale**.**transform(X\_train)

X\_test **=** std\_scale**.**transform(X\_test)

<http://carmenlai.com/2016/11/12/user-churn-prediction-a-machine-learning-workflow>