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 borrowers that default on their home equity loans, which happens when a loan is 90+ days outstanding. The value of 90+ day delinquencies is 2.23 billion in 2016 (2). Predicting defaults for home equity line of credits are extremely important because the lender is always second in line to recoup their loan behind any 1st mortgage that may exist. In this study we seek to uncover variables that classify a borrower’s propensity to default. This research study dives into a collection of potential variables and delinquency indicators 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 including a review of the summary statistics, addressed missing values, standardized data, engineered additional features and more. Second, we ran the data through a logistic regression model and determined which values had the highest probabilities of class membership. We established a baseline logistic regression model and recorded its performance measures on predicting the binary response accurately. Next using the results of the logistic model as a baseline, we built additional models to see if we could determine a predictor that performs better than the logistic regression model. Five other prediction models were created to compete with our baseline model in hopes of selecting the best performing 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 borrower’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 Summary Statistics. The response variable “BAD” is binary, where a 1 means the borrower 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** | **MIN** | **MAX** | **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 Summary Statistics 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 indicator variables to be prepared for subsequent modeling efforts. The data dictionary can be found in Table 3 Data Dictionary in the appendix.

**Exploratory Data Analysis**

From the initial exploration, we can see an issue with incomplete records as shown in Figure 17 Incomplete Data. We have to determine what fields need to be populated that are required for our 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 (mean, median, regression, etc.) 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 as shown in Figure 18 Removing Incomplete Records. The reason for that decision is we can build the model based on the 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 Figure 1 Evidence of multicollinearity in "MORTDUE". This is important to reduce redundancy and isolate variable importance. 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 600 records we feel we are covered under the central limit theorem.

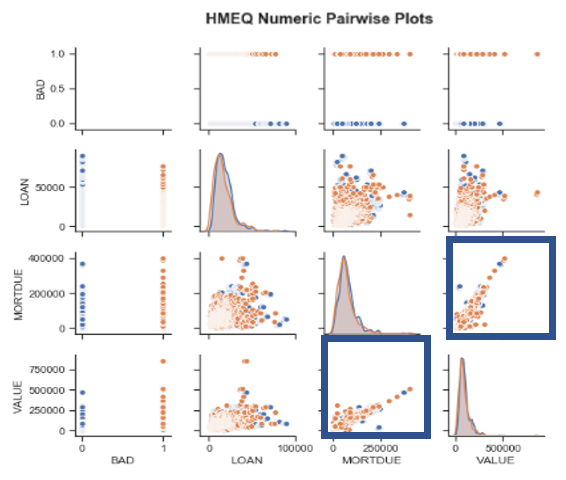
We see evidence of multicollinearity with "VALUE" and "MORTDUE", because as home values increase, so does the total mortgage due. There is clearly a linear relationship between two explanatory variables of which we do not want. We do not need both variables, but we will assess that in later modeling efforts. The full scatterplot matrix can be found in Figure 4 EDA Scatterplot in the appendix.

Figure 1 Evidence of multicollinearity in "MORTDUE"

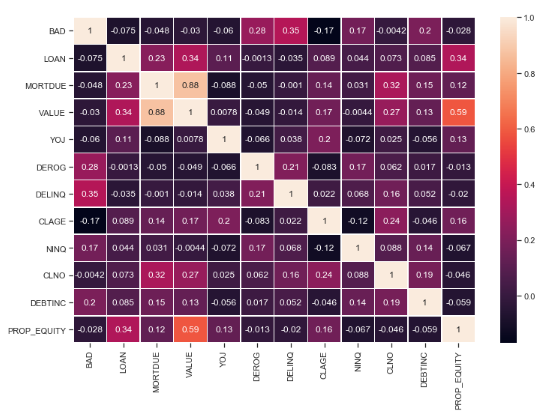
The Figure 2 EDA Heatmap confirms the high multicollinearity of “VALUE” and “MORTDUE.”

Figure 2 EDA Heatmap

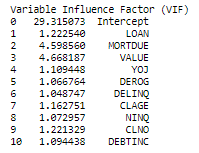
Although the scatterplot matrix and heat map provide visual evidence of possible multicollinearity, Figure 4 EDA Variable Inflation Factor (VIF) list quantifies the severity of multicollinearity. As suspected, 'VALUE' and 'MORTDUE' are borderline, and fall just under the acceptable threshold (VIF >5). In an attempt to reduce the number of variables and complexities in our model, we chose to remove the 'VALUE' variable given it has a VIF closest to 5. This will allow us to reduce model complexity as much as possible before our model/feature selection efforts. We chose to remove “JOB” for reasons explained below.

Figure 3 EDA Variable Inflation Factor (VIF)

Like other generalized linear models, logistic regression is not robust to multicollinearity. Now that we have removed a variable (VALUE) 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 as shown in Figure 4 Imbalanced Data. This is something we need to address in our analysis.

A screenshot of a cell phone

Description automatically generated

Figure 4 Imbalanced data

A screenshot of a cell phone

Description automatically generatedWe 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. Figure 5 Reason Code Split and Default split

Figure 5 Reason Code Split and Default split

A screenshot of a video game

Description automatically generated

Figure 6 Default Split by "YOJ" and "CLAGE"

A box and whisker plot in Figure 6 Default Split by "YOJ" and "CLAGE" 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. In Figure 7 Terms with evidence of Interaction we show fields we will consider in the future for interaction terms in a model.

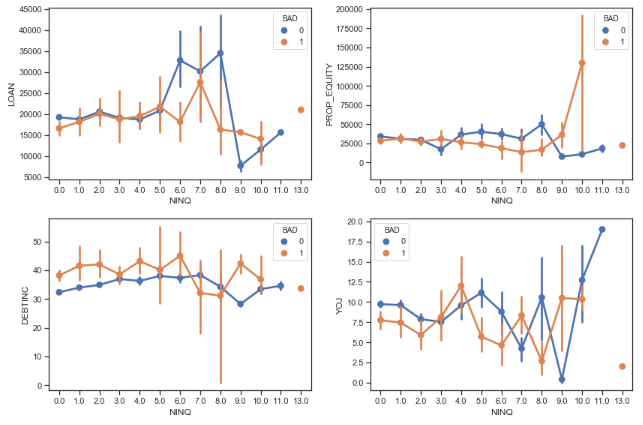


Figure 7 Terms with evidence of Interaction

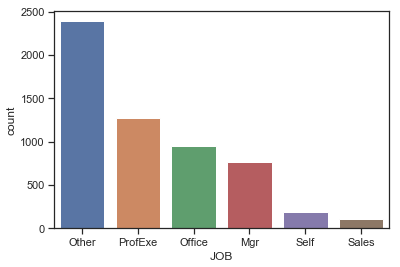
The “JOB” category and mainly the distribution makes this field of little practical value. Almost half, 42% of the jobs fall into an “Other” category as shown in Figure 8 Job Categories splits. This kind of general classification bucket could contain a variety of entry-level and higher paying positions and in it would be difficult and of little practical value to predict based on a field that is used as a catch-all. As a result, this variable will be not be included in our modeling efforts. That makes two total variables (“JOB” and “VALUE”) eliminated as explanatory variables in our prediction models.

Figure 8 Job Categories splits

#### **Objective 1- LOGISTIC REGRESSION**

**Restatement of Problem**

Our objective is to classify borrowers who will default or pay their home mortgage equity loan. Since the target variable is binary, we will use logistic regression as our primary baseline model.

Dummy variables were created to ingest the categorical variables into our logistic regression model.

Although we prefer to have more observations, up-sampling would require sampling with replacement. Instead, we decided to down-sample so the majority class (0) would match the minority class (1). In other words, the data will be down-sampled to get an even distribution of records that are paid and defaulted. Using this method meant reducing our total data size to 602, however it meant we had balanced, complete data that we were comfortable to use for model creation. Also, our dataset consists of values on different scales. We do not want our prediction model giving higher importance to variables just because they aren't scaled. The RobustScaler will standardize all continuous variables except the indicator and response variables. This particular scaler is robust to outliers and uses Figure 9 RobustScaler Formula to scale the explanatory variables.

A close up of a logo

Description automatically generated

Figure RobustScaler Formula

**Model Selection**

As we consider models, Forward, Backward and Stepwise modeling selection techniques require that you have some idea of 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 for our first model we wanted to see what success we would have using all of the features with a logistic regression.

We also built and fitted a LASSO regression with the same variables as the baseline regression model in an attempt to reduce model complexity even further. As shown below, LASSO did its job as a feature selection tool by removing 'MORTDUE' and 'REASON\_HomeImp' from the model.

A close up of text on a white background

Description automatically generated

Figure Lasso model summary

We chose to compare our models based primarily on accuracy and precision, as there are major consequences to misidentifying high risk borrowers that tend to default (False Negative). Misclassifying people who will pay (False Positive) as defaulters is not as detrimental. We included AUC, accuracy, sensitivity and specificity as our primary performance measures to better understand the strengths of this and future models. We can also compare misclassification rate (1-Accuracy) to further evaluate our models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODEL** | **PRECISION** | **AUC** | **ACCURACY** | **SENSITIVITY** | **SPECIFICITY** |
| Logistic Model | 0.777 | 0.762 | 0.752 | 0.860 | 0.677 |
| LASSO Model | 0.770 | 0.833 | 0.743 | 0.857 | 0.667 |

**Checking Assumptions**

The assumptions of logistic regression are binary, which is met as our response variable “BAD” is binary. The data must be independent, and as stated in our data exploration, the data is assumed to be independent as typically a home equity or other loan would be one per household, and each row is a home equity loan for a person. It would be exceptional to have multiple home equity loans per house at the same bank. Next, we have the assumption of multicollinearity which we have explored in the Figure 1 Evidence of multicollinearity in "MORTDUE" and Figure 2 EDA Heatmap below- and resolved by removing the “VALUE” column. We also have the assumption of continuous independent variables being linearly related to the log odds. Although the log shape is subtle for some continuous variables, we verified this was in place. Figure 11 DEBTINC Log Odds Linear Plot

A screenshot of a cell phone

Description automatically generated

Figure DEBTINC Log Odds Linear Plot

We did perform some fit diagnostics based on our logistic regression model and found 1 record of concern with the index value of 3679 as shown in the Figure 10 Leverage vs. Normalized Residuals Squared, Figure 11 Influence Plot, and Figure 12 Cook's D plots.

A picture containing screenshot

Description automatically generated

Figure Leverage vs. Normalized Residuals Squared

A close up of a logo

Description automatically generated

Figure Influence Plot

A screenshot of a cell phone

Description automatically generated

Figure Cook's D

This index row was removed from the data frame given its significant Cook’s D value and high influence.

**Parameter Interpretation and CONFIDENCE INTERVALS**

The Logistic Regression Model

We built and fit the model in python using the statsmodel package for logistic regression using 10 features. We eliminated variables such as “JOB” based on intuition after conducting EDA.

Building the Model

Predicted(BAD) = β0 + β1\*(LOAN) + β2\*(MORTDUE\_0) + β3\*(YOJ) + β4\*(DEROG) + β5\*(DELINQ) + β6\*(CLAGE) + β7\*(NINQ) + β8\*(CLNO) + β9\*(DEBTINC) + β10\*(REASON\_HomeImp))

Fitting the Model

log⁡((𝑃(𝑥))/(1−𝑃(𝑥)))= −0.555+−.0731(𝐿𝑂𝐴𝑁)+.0533(𝑀𝑂𝑅𝑇𝐷𝑈𝐸)−.3296(𝑌𝑂𝐽)+.08722(𝐷𝐸𝑅𝑂𝐺)+.8963(𝐷𝐸𝐿𝐼𝑁𝑄)+.4040(𝐶𝐿𝐴𝐺𝐸)+.3518(𝑁𝐼𝑁𝑄)−.4495(𝐶𝐿𝑁𝑂)+.6442(𝐷𝐸𝐵𝑇𝐼𝑁𝐶)+.1164(𝑅𝐸𝐴𝑆𝑂𝑁\_𝐻𝑂𝑀𝐸𝐼𝑀𝑃)

Parameter interpretation

The parameter results make logical sense, as having a longer job history decreases your odds of default, while having derogatory information, delinquencies and a high debt to income ratio will increase the odds of default. For our model, we took the Logit Odds Ratio for parameter interpretation.

A 1 unit increase or decrease in a variable affects the odds of defaulting. For example, 1 unit increase in 'YOJ' would account for an 26% decrease in the odds of defaulting.

A 1 unit increase or decrease in a variable affects the odds of defaulting. For example, 1 unit increase in ‘DEROG’ would account for an 162% increase in the odds of defaulting.

A 1 unit increase or decrease in a variable affects the odds of defaulting. For example, 1 unit increase in ‘DELINQ’ would account for an 166% increase in the odds of defaulting.

A 1 unit increase or decrease in a variable affects the odds of defaulting. For example, 1 unit increase in ‘DEBINC’ would account for an 80% increase in the odds of defaulting.

Confidence Interval Interpretation

The odds of defaulting with YOJ are 26% less with the true effect between -47% and .9%. We are 95% confident the odds ratio is between -.47 and 0.009.

The odds of defaulting with DEROG are 162% more with the true effect between 65% and 318%. We are 95% confident the odds ratio is between 0.66 and 3.18.

The odds of defaulting with DELINQ are 166% more with the true effect between 87% and 281%. We are 95% confident the odds ratio is between 0.87 and 2.81.

The odds of defaulting with DEBTINC are 80% more with the true effect between 38% and 133%. We are 95% confident the odds ratio is between 0.38 and 1.33.

A screenshot of a cell phone

Description automatically generated

Figure Odds ration of Baseline Model Interpretation

**Final conclusions**

For our first analysis we used the logistic regression and LASSO regression model and predicted a borrower’s risk of default with a precision of 77% and using 10 features that are available in the dataset. This is a first step in developing the best model, and ideally this would be run and validated against additional datasets with more data demographics known, such as geography and year. The average real estate loan can vary within and across geographic regions, and there are certain years, such as the great recession in 2008 that will have data that is different than subsequent years- having +that information can help us control possible confounding variables.

The significant predictors that we found and included in this model are: “DEROG”, “DELINQ” and “DEBINC”, however there are borrowers that paid that had some data in these fields. What this means is although these fields can indicate default, you can mis-classify payers if too much weight are attributed to these variables. Using these parameters to establish a predictive model for the default rate for a given borrower, we established the following fitted model:

log⁡((𝑃(𝑥))/(1−𝑃(𝑥)))= −0.555+−.0731(𝐿𝑂𝐴𝑁)+.0533(𝑀𝑂𝑅𝑇𝐷𝑈𝐸)−.3296(𝑌𝑂𝐽)+.08722(𝐷𝐸𝑅𝑂𝐺)+.8963(𝐷𝐸𝐿𝐼𝑁𝑄)+.4040(𝐶𝐿𝐴𝐺𝐸)+.3518(𝑁𝐼𝑁𝑄)−.4495(𝐶𝐿𝑁𝑂)+.6442(𝐷𝐸𝐵𝑇𝐼𝑁𝐶)+.1164(𝑅𝐸𝐴𝑆𝑂𝑁\_𝐻𝑂𝑀𝐸𝐼𝑀𝑃)

The 3 fields ‘DEROG’, ‘DELINQ’ and ‘DEBTINC’ have both statistical and practical significance, they are strong classification indicators to default. Knowing this, if this were a customer/borrower data set, these would be fields I would use to trigger proactive messaging about credit counseling with the intention to reduce defaults.

There were limited variables provided, so there are additional fields that are not included in the data set that likely affect a borrower's risk of default and that the parameters of significance here are not significant when accounting for other significant parameters excluded in this data.

Although the results are preliminary, this data could be used to identify current borrowers with identified risk factors and target them with messaging around using credit responsibly, and other resources to prevent defaults. The results of that targeting could be used as a feedback loop to further refine the model.

#### **Objective 2- ADDITIONAL MODELING**

We created a logistic regression model using intuition, but non-parametric models are frequently used for classification. Our objective is to create a reliable classification model for identifying borrowers hat default, but the baseline model has 10 features- we would like to see if additional modeling can yield the same or better precision but with less complexity. Again, fewer features or less complexity makes the required data fields fewer and lends to a more flexible model than a model with many features.

We opted for a LASSO regression model (Lasso) because it is a good feature reduction tool that allows us to reduce model complexity. Controlled by L1 regularization and alpha, coefficients converged until it becomes a zero and removed from our model.

We chose Random forest for our next model, as it a supervised learning algorithm that is flexible and commonly used for classification and feature selection. A random forest is comprised of decision trees, which are created on randomly selected samples within the dataset. It iterates through the sample, calculating predictions for each tree and identifies the best solution.

Linear Discriminant Analysis, which is closely related to Quadratic Discriminant Analysis that classify based on a linear and a quadratic combination of features that separates two or more classes of objects or events. Both models have minimal tuning parameters and can work well when there is separation in the data. Our data has a good amount of overlap, so there was concern that it would be a more challenging model in our dataset.

The K nearest neighbors (KNN) model is a non-parametric method that can be used for both classification and regression. We designated to look at the features associated with the nearest 6 neighbors to determine classification.

For our complex logistic regression model we created calculated variable that represents property equity (VALUE-LOAN). We included this engineered feature to our data frame based on the coefficient output of our LASSO Regression model.

**Main Analysis Content**

We feel precision will be the most useful to choose the best model for our purposes, but we also wanted to see other measures of error, such as accuracy. Out of all of our models, Random Forest was our best classifier, based on precision, accuracy and AUC. This can be visually appreciated as the green line in Figure 15 Visual ROC Curves of the multiple models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODEL** | **PRECISION** | **AUC** | **ACCURACY** | **SENSITIVITY** | **SPECIFICITY** |
| Logistic Model(Base) | 0.777 | 0.762 | 0.752 | 0.860 | 0.677 |
| Lasso | 0.770 | 0.833 | 0.743 | 0.857 | 0.667 |
| Random Forest | 0.853 | 0.916 | 0.851 | 0.887 | 0.814 |
| LDA | 0.815 | 0.835 | 0.768 | 0.932 | 0.675 |
| K Nearest Neighbor | 0.790 | 0.861 | 0.702 | 0.941 | 0.609 |
| LRC | 0.771 | 0.832 | 0.743 | 0.857 | 0.667 |

Table 2 Comparing Multiple Models

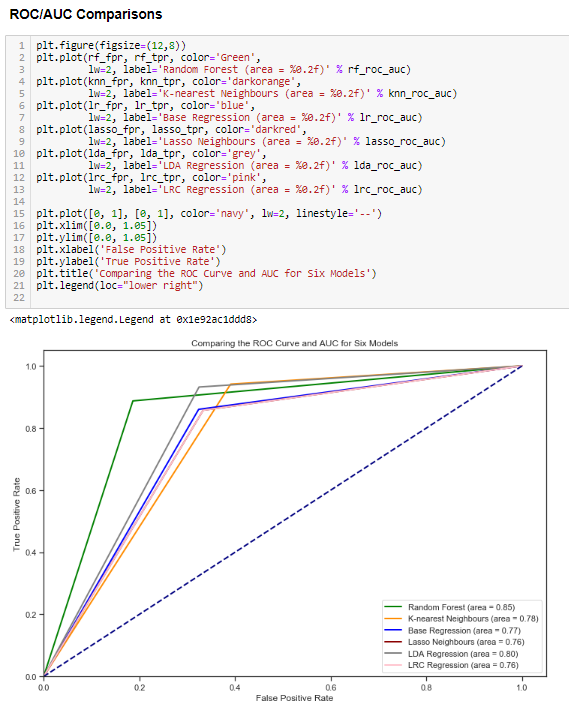


Figure 16 Visual ROC Curves of the multiple models

**Conclusion/Discussion**

Our objective was to classify borrowers as people who will default or pay their home mortgage equity loan. Our best performing model was random forest, which we attribute to the nature of the model iterating through the different combinations of variables versus other models which had more limited tuning. As we are not subject matter experts in home equity loans or banking, the algorithm was able to run through more scenarios to find the best fit. Performance measures increased as we increased the number of iterations or trees. We found the sweet spot to be 60 iterations. A better classifier would be in the 90% range, but for this analysis, we feel 85% is a good basis for additional research.

Our data had many incomplete records which we removed from the data set, our intention was to get the best model based on complete records and ideally with fewer features to make the model portable and flexible enough to apply to other data sets. One issue that is common when dealing with real estate values is a long right tail and outliers. These are natural and valid data points and shouldn’t be removed. One future consideration could be splitting out loans by size, and evaluating them separately, and seeing how the modeling would change.

**APPendix**

Data Dictionary

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Description** | **Levels** | **Dummy\_Column?** | **Class** | **Unique values** | **Missing** |
| BAD | 1 = applicant defaulted on loan or seriously delinquent; 0 = applicant paid loan | 2 | N | integer | 2 | 0% |
| LOAN | Amount of the loan request |  | N | integer | 540 | 0% |
| MORTDUE | Amount due on existing mortgage |  | N | numeric | 5054 | 8.69% |
| VALUE | Value of current property |  | N | numeric | 5382 | 1.88% |
| REASON | DebtCon = debt consolidation; HomeImp = home improvement | 2 | Y | factor | 3 | 0% |
| JOB | Occupational categories | 6 | Y | factor | 7 | 0% |
| YOJ | Years at present job |  | N | numeric | 100 | 8.64% |
| DEROG | Number of major derogatory reports |  | N | integer | 12 | 11.88% |
| DELINQ | Number of delinquent credit lines |  | N | integer | 15 | 9.73% |
| CLAGE | Age of oldest credit line in months |  | N | numeric | 5315 | 5.17% |
| NINQ | Number of recent credit inquiries |  | N | integer | 17 | 8.56% |
| CLNO | Number of credit lines |  | N | integer | 63 | 3.72% |
| DEBTINC | Debt-to-income ratio |  | n | Numeric | 4694 | 21.26% |

Table 3 Data Dictionary

Incomplete Data

Here we see that many variables have incomplete data, such as 21% of records have no data in the “DEBTINC” field

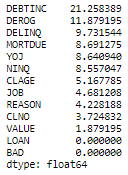


Figure 17 Incomplete Data

Removing Incomplete Records

We remove data with incomplete rows, which is 41% of the dataset

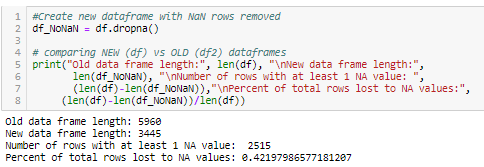


Figure 18 Removing Incomplete Records

EDA Scatterplot Matrix

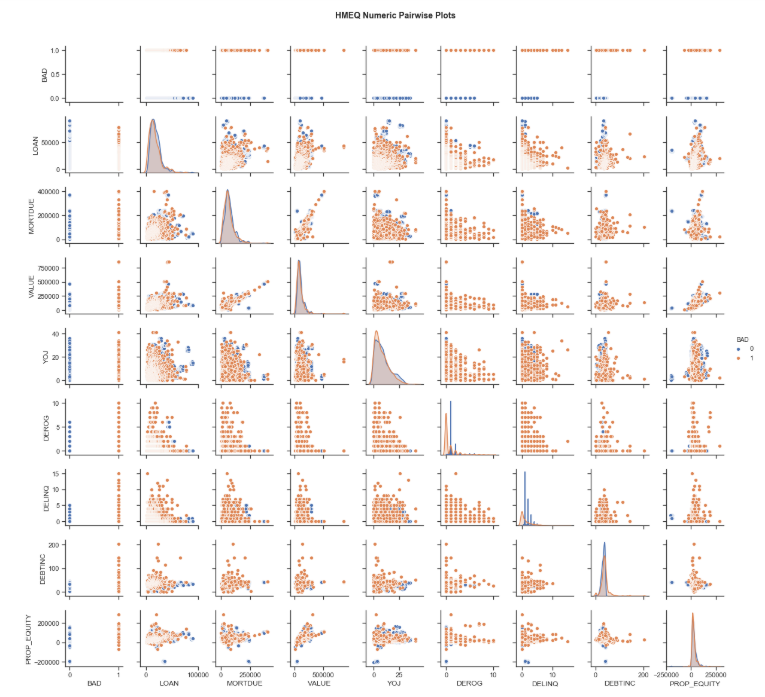


Figure 19 EDA Scatterplot

Outlier removal

|  |
| --- |
| Figure 20 Leverage versus Normalized Residual Plot  Figure 21 The leverage point  Figure 22 The influence plot with leverage point    Figure 23 The outlier record 3679  Figure 24 Confirming the removal of outlier 3679 |

Baseline Logistic Regression Model

|  |
| --- |
|  |

LASSO Regression Model

|  |
| --- |
|  |

Random Forest Model

|  |
| --- |
|  |

Linear Discriminant Analysis Model

|  |
| --- |
|  |

K Nearest Neighbors Model

|  |
| --- |
|  |

Logistic Regression (Complex) Model

|  |
| --- |
|  |

### **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>

### **CODE**

Readable PDF version of Python code can be found and downloaded here:

<https://github.com/bbal20/stats2_project2/blob/master/stats2_project2_python_code.pdf>

Python Jupyter Notebook code can be found and downloaded here:

<https://github.com/bbal20/stats2_project2/blob/master/stats2_project2_code.ipynb>