An Analysis in Situational Factors on Determining Short Term Credit Default Rates During the Peak of the Mortgage Crisis

Authors

Spencer Hanson
spha0827@colorado.edu

Benjamin Miller
bemi0682@colorado.edu

Maxim Moghadam
mamo5089@colorado.edu

William Brickowski
wibr5703@colorado.edu

ABSTRACT

In this project we take the Fannie Mae and Freddie Mac mortgage data during the peak quarters of the housing bubble, the last two quarters of 2008 and the first two quarters of 2009 to analyze which attributes correlated highly to default rates. The goal of the project is two fold. We would aim to come up with an accurate prediction for default rates as well as analyzing attributes that led to the growth and ultimate implosion of the housing bubble. The knowledge obtained from such analysis will prove useful in several settings. It can be applied in the credit rating sector to come up with more accurate predictions of credit risk. it can also be used to classify particular lending practices and particular attributes of loans that led to the

1 PROBLEM STATEMENT

Credit scores are an important factor in many financial decisions, especially in determining whether or not to give out a loan. Financial institutions use credit scores to determine eligibility, taking into account income, loan principle, loan purpose, number of borrowers, and postal code among other factors. By integrating and analyzing macroeconomic data in conjunction with loan-level mortgage data we aim to develop a more accurate prediction of whether or not a person is likely to default on their loan. This would enable more intelligent investment decisions by banks or other entities, and could prevent mass defaults such as those that caused the 2008 financial crisis.

2 PREVIOUS WORK

Since its inception in the 1950s, credit scoring has been one of the fastest growing fields in analytical statistics. Because of this there is an abundance of work in the formulation of statistical credit scoring models with much of it published relatively recently. We look specifically to three types of literature in the field: Classic credit analysis, models for credit default, and statistical shortcomings of credit analysis.

2.1 Classic credit analysis

When deciding whether to give a loan, bankers frequently look at the "Five C's": character, capital, capacity, conditions, and collateral.

Character, or a person's history, is part of the system in which credit is gained by methods such as paying back loans, credit cards, and making payments consistently and on time, and is lost by being late for, missing or defaulting on payments.

Capital, capacity, and collateral are all measures of the wealth of a person and how much they will bet on themselves paying off the loan: generally, these are down payments, income, and the terms of default respectively. As people with a higher loan-size to wealth ratio are at a greater risk to default, and people with a smaller ratio

are more able to pay off a loan without bankrputcy, this also affects credit score.
Conditions are the circumstances of the loan, such as interest rate. While the other factors affect credit score, credit score affects conditions; a bank may demand a higher interest rate for a person with a lower credit score.
A high credit score can lead to lower interest rates and larger loans, while a low one can lead to higher interest rates, smaller maximum loans, or even being denied loans.

2.2 Models for Credit Default

The focus of this project will be on creating a model that can accurately predict credit default rates. Much work has been done in forming statistical models to achieve this. These models include Logistic Regression, K-Nearest Neighbors, K-fold Cross Validation, and Random Forest. We look to an archive of statistical papers written on the subject, specifically examples of the use of logit and probit regressions,[1] but also more complicated dynamic models[2] and instances of machine learning models such as random forests.[3]

2.3 Statistical Shortcomings of Credit analysis

Lending institutions employ two measures of scoring credit, namely, bureau scores and application scores. The former measure focuses solely on past credit history, while the latter includes other weighting factors such as age and location in addition to credit history. The issues with bureau credit scores stem primarily from omitted variable bias,[4] where local economic conditions and business cycles are not taken into account in the predictive default model. Additional problems with credit analysis can be attributed to data quality issues.[5]

3 DATA

3.1 Loan-Level Mortgages

Freddie Mac began reporting loan-level credit performance data at the direction of its regulator, the Federal Housing Finance Agency (FHFA) with the stated purpose of increasing transparency, which "helps investors build more accurate credit performance models in support of potential risk-sharing initiatives highlighted in FHFA's Conservatorship Scorecard." We have found a single family loan-level dataset that includes loan-level origination, loan performance, and actual loss data on a proportion of single family mortgages acquired by Freddie Mac. The data contains mortgages from January 1, 1999 to December 31, 2016. [6]

3.2 Macroeconomic Data

The Bureau of Economic Analysis has economic profiles available on every local area in the country. This includes metropolitan and micropolitan areas in every state with measures of income, employment, retirement, and population. [7]

4 PROPOSED WORK

4.1 Preprocessing

The data from an arbitrary quarter in the Freddie Mac dataset include two different data tables. The first of the

[1] Qingfen Zhang. 2015. Modeling the Probability of Mortgage Default via Logistic Regression and Survival Analysis. (2015). http://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1543&context=theses

[2] John Y. Cambell and Joao F. Cocco. 2011. A Model of Mortgage Default. (2011). https://scholar.harvard.edu/files/campbell/files/mortdefault13022014.pdf

[3] Grace Deng. 2011. Analyzing the Risk of Mortgage Default. (2011). https://www.stat.berkeley.edu/ãldous/Research/Ugrad/Grace_Deng_thesis.pdf

[4] Robert B. Avery, Raphael W. Bostic, Paul S. Calem, and Glenn B. Canner. 2000. Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files. Real Estate Economics 28, 3 (2000), 523âĂŞ547. DOI:http://dx.doi.org/10.1111/1540-6229.00811

[5] Robert B. Avery, Paul S. Calem, and Glenn B. Canner. 2011. Credit Report Accuracy and Access to Credit. (2011). https://www.federalreserve.gov/pubs/bulletin/2004/summer04_credit.pdf

[6] https://freddiemac.embs.com/FLoan/Data/download3.php

[7] https://www.bea.gov/regional/downloadzip.cfm

two is the origination data file which contains 27 different attributes, while the other file contains monthly reports of the the respective loans' performances. In the preprocessing stage, we are tasked with combining these two datasets based on a commonly shared attribute, the loan identification number. We then must choose those attributes in the newly merged file that are of interest to us, that is, attributes that we believe are predictors of loan default. Some of these attributes are irrelevant for determine a credit score, others simply have too many missing observations, or cannot be used in regression because they are in a numeric form but represent categorical variables, such as zip code. Additionally, we will need to add an attribute to the merged data frame. This boolean attribute will denote the event that a person defaulted on their loan. We define a default as the event where a loan becomes at least 60 days delinquent at some point in its life, this will be used as our standard for the attribute "default".
After similarly processing the Bureau of Economic Analysis data to trim unwanted columns, we can combine the two datasets by using the zipcode given in the Freddie Mac dataset. We will essentially look up the relevant economic data for each debtor's zipcode and copy the relevant data into a single combined set.

4.2 Analysis

Using our datasets, we will compute a correlation coefficient for each respective attribute compared with the binary data type of whether or not a default has occurred. Will plan to do this prior to performing regression analysis in order to help determine which attributes are most likely to determine default. , deleting those that fall below a certain correlation threshold. This will allow us to minimize the number of dimensionality of our data so that we are more efficiently able to regress on such a large data set. The reason for this is to reduce the number of attributes and in doing so reducing the chance that we over-fit our data and fall prey to the curse of dimensionality. By removing attributes that are not relatively correlated to the default attribute in order to avoid the curse of dimensionality, we are able reduce the risk of over fitting our model. We will then form logistic regressions with loan default as the dependent variable and the attributes that passed the correlation threshold hold as the independent variables. This will form a baseline for our research. We will need to analyze the effect of each attribute on the default rate, for example, perhaps the fact that a person is a first time homeowner should be considered in the regression model considering the fact that they have never owned a house before. After analyzing the relationships between all the variables, we will create a regression model to generate the probability that a person will default on their loan. From this point, we will add in geographic and economic variables that we believe might have a statistical impact on loan default, such as the fact that an individual lives in a rural verses an urban area. From this point, we will explore elements of survival analysis to see if its application can provide a better prediction of loan default. Lastly, we will begin to explore elements of machine learning and, if time permits, form random forests to attempt to form an even more accurate prediction of loan default. From this, we will generate a formula that takes a number of boolean and numeric arguments to generate the probability of default.

4.3 Evaluation

We will evaluate our research by testing to see if its accuracy predicts whether or not a person will default on their loan, we hope to get an accuracy measure of 80-90%. Additionally, we will evaluate the varying models that we used to determine the default rate in order understand why certain models more effective. Models that we are interested in using are logit and probit regression, survival analysis, linear discriminant analysis, and random forest. The discovered knowledge of this project can be applied directly into the credit world and can be used by credit rating agencies to predict credit risk with more accuracy. Additionally, by limiting our scope of research to the time period corresponding to the peak of the mortgage crisis, we will learn about what types of loans and what attributes contributed to the growth and ultimate implosion of the housing bubble. This knowledge can be used to predict future bubbles and characteristic of risky mortgage lending.

5 MODELS FOR PREDICTION OF DEFAULT

5.1 Linear Regression

Regression is a statistical method that uses least squares estimation to formulate a model that expresses the relationship between variables in data. We will be using it as our baseline approach for analyzing default rates. Classic linear regression relates some independent variable X to the behavior of an ostensibly dependent variable Y.

$$Y_i = \beta X_i + \epsilon, \quad i \epsilon n \tag{1}$$

Where Y_i is the i^{th} dependent observation, X_i is the value of the independent variable in the i^{th} observation, and ϵ is a noise factor to account for unobserved variables that follows a normal distribution with mean zero and variance 1. This implies

$$\begin{gathered} Y_1 = \beta X_1 + \epsilon_1 \\ Y_2 = \beta X_2 + \epsilon_2 \\ \vdots \\ Y_n = \beta X_n + \epsilon_n \end{gathered}$$

By defining the following variables:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{2}$$

We can express the linear regressive model as a matrix product

$$Y = X\beta + \epsilon \tag{3}$$

Where Y is a vector of of dependent observations, X is a matrix with columns corresponding to independent variables considered to be of interest in predicting Y and rows corresponding to individual observations, and β is a vector of weights where the ith element denotes the marginal effect of independent variable x_i on the dependent variable Y. To form a predictive model, we estimate the β vector with the least squares estimation method so as to minimize the residual sum of squares. The $\hat{\beta}$ vector is calculated with least squares estimation, and its solution is $\hat{\beta} = (X^T X)^{-1} X^T Y$. We then define

$$\hat{Y} = X^T \hat{\beta}$$

where $\hat{Y}$ represents the predicted dependent outcome given a set of independent variables.

x_i can be either numerical or binary (i.e. 0 or 1) depending on the type of data and the type of regression used. In the latter case, we denote it as a dummy variable. We will generate several of these regression equations using different independent variables, and we will be able to compare their effectiveness, forming a more accurate regression model from their results.

5.2 Logit and Probit Regression

In our project, $\hat{Y}$ will represent the probability of the event of defaulting on a loan, in this event Y_i will take on a value between 0 and 1 (probabilistic). X will represent a matrix of observations with columns corresponding to independent variables considered to be a function of the dependent variable, probability of default. This is known as probit regression, which takes the form:

$$P(Y = 1|X) = \Phi(X^T \hat{\beta}) \tag{4}$$

Where Φ is the cumulative density function of the standard normal distribution.

Additionally, we can look to logistic regression which takes the form

$$\ln \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = X\hat{\beta}_1 \tag{5}$$

To determine which of these to use, we will calculate the likelihood values.

5.3 Survival Analysis

Survival analysis is a branch in statistics that focuses on analyzing the expected time that passes before some event occurs, such as an earthquake or a death from some disease. Our analysis will focus on predicting the probability an individual has defaulted on their loan at time period τ defined as $F(\tau) = P(T < \tau)$. From this we can define a survival function that represents the probability that an individual has not defaulted on their loan at time τ: $S(\tau) = P(T > \tau) = 1 - F(\tau)$. If we set t equal to the maturity date of the loan this is equivalent to the probability an individual will never default on their loan.

5.4 Linear Discriminant Analysis

LDA is a statistical method used in machine learning to find a linear combination of features that characterizes two or more events. We will denote these two events as an individual defaulting ($y = 1$) or not defaulting ($y = 0$) on a loan. LDA approaches this problem by assuming the conditional distributions the functional density ($P(x|y = 0)$ and $P(x|y = 1)$) are normal. Hypothesis testing is then performed on the log of the likelihood ratios to determine the probability of default.

5.5 Bayesian Classifiers

Bayesian Classifiers are a prediction method that uses probability rather than regression, which can be useful for the non-numerical data. Using the formula:

$$P(H|x) = \frac{(x|H)P(H)}{P(x)}$$

We can calculate the probabilities of loan default based on variables or combinations of variables, and will use this to help build decision trees, as detailed in the next sections.

5.6 Decision Tree

Using the regressions and probabilities found from the aforementioned methods, one way of using our findings to create a credit score function will be a decision tree. Using the variables we found most correlated with loan default we can create a function that takes a series of variables and outputs one of a predetermined set of probabilities of default.

5.7 Random Forest

Decision trees are a popular statistical method for in the field of machine learning, but they contain drawbacks in application because of their invariance under scaling, along with other reasons. Random forests applies the technique of bootstrapping to decision trees, selecting random samples with replacement of the training set and fitting trees to each respective sample. This is the most complicated statistical method mentioned and its inclusion will be conditional on the amount of time it takes to form models of the other methods and analyze their results.

6 TOOLS

The tools will be mostly python-based tools, with some database and data manulipation softwares.
Python - To implement statistical and data mining algorithms
Pandas - A package containing some of these algorithms
Numpy - To aid with the large datastructures inherent in large datasets
SciKit-Learn - A package containing basic machine learning algorithms
MongoDB - The database software to store our dataset
MatPlotlib - For visualization of data and generation of graphs
MatLab - For non-statistical calculations and formulas

7 MILESTONES

7.1 Completed

2/23: Load sample data into Mongo
3/5: Download and have all data available
3/12: Clean and standardize all datasets
3/19: Start merge between the datasets
3/26: Perform an analysis of attributes, e.g. correlation
4/1: Perform basic linear regressions on data

7.2 To Do

4/8: Perform logistic/probabilistic regressions on data
4/15: Apply machine learning methods to data
4/28: Document and process results

8 RESULTS

8.1 Correlation

We calculated the correlation matrix of all attributes and created a colored graph of it. In the graph below, lighter hues indicate higher correlations, while white indicates N/A values.

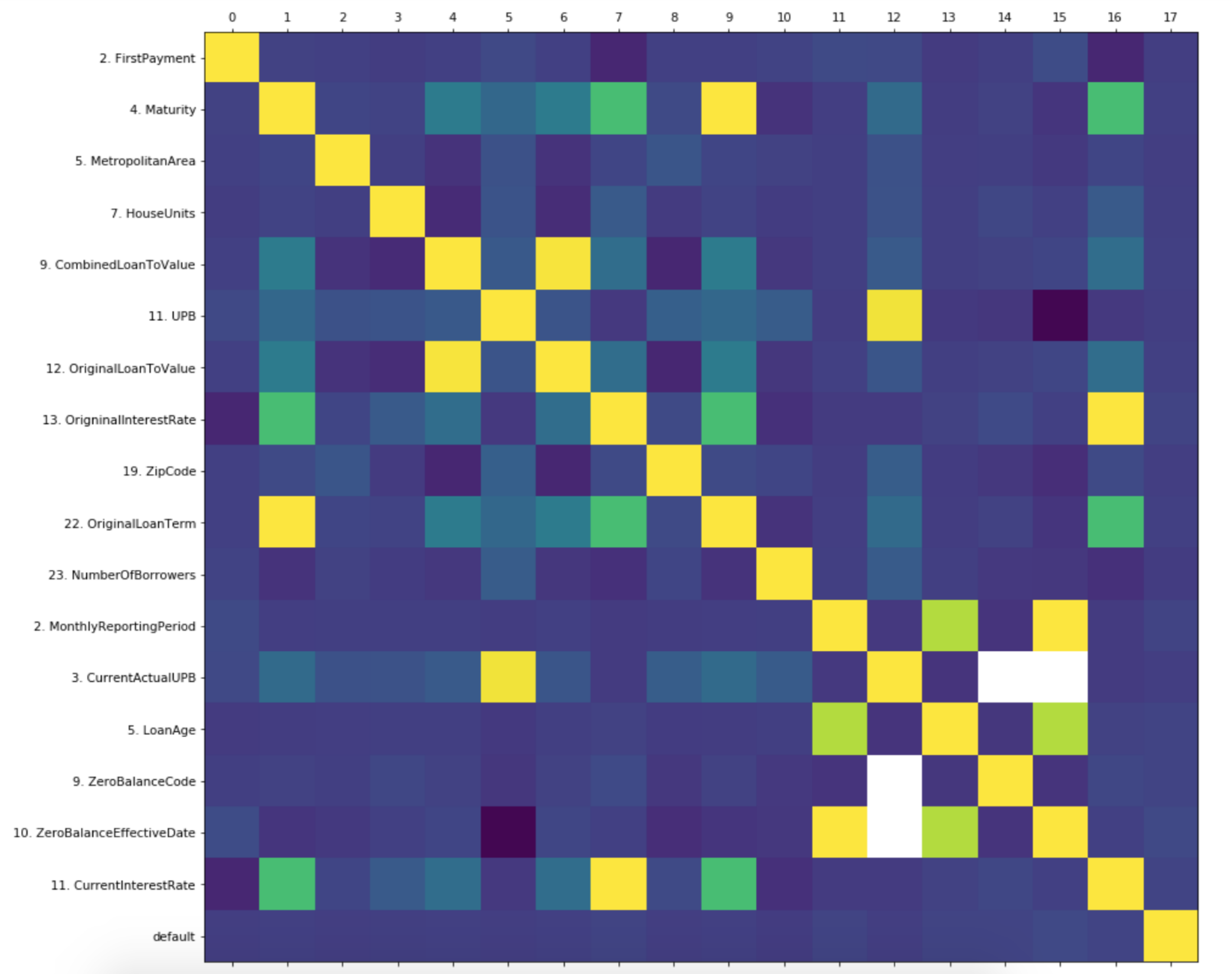
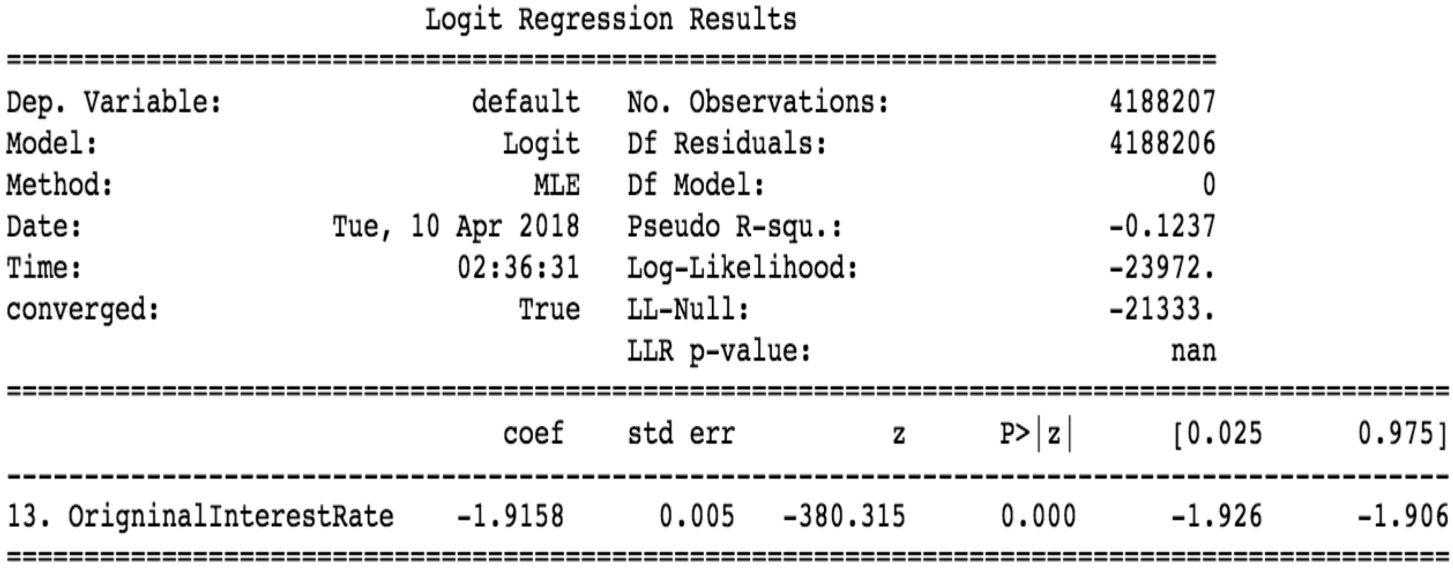


Figure 1: A correlation matrix of all attributes

8.2 Logit Regression

Below is an early regression with one independent variable. The R^2 value is negative, indicating that a horizontal line fits the data better than our model. This is to be expected though, as the model is severely restrictive.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:              4188207
Model:                          Logit   Df Residuals:                  4188206
Method:                           MLE   Df Model:                            0
Date:                Tue, 10 Apr 2018   Pseudo R-squ.:                 -0.1237
Time:                        02:36:31   Log-Likelihood:                -23972.
converged:                       True   LL-Null:                       -21333.
                                        LLR p-value:                       nan
=============================================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
13. OrigninalInterestRate   -1.9158      0.005   -380.315      0.000      -1.926      -1.906
=============================================================================================
```

Figure 2: Logit regression of interest rate on default rate