An Analysis in Situational Factors on Predicting Short Term Credit Default During the Peak of the 2008 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 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 aim to come up with an accurate prediction for default rates and also plan to analyze 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 housing crisis of 2008.

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 bankruptcy, 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 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. We plan to analyze state-level data to see which attributes of a state make its residents likely to default. [7]

[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

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 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 this point, we can collapse the data on each loan sequence number, effectively giving us rows of unique loans, which each row having a column that denotes whether the individual defaulted on their mortgage.
After similarly processing the Bureau of Economic Analysis data to trim unwanted rows, we can combine the two datasets by using the zip code given in the Freddie Mac dataset. We will essentially look up the relevant economic data for each debtor's state 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. We plan to do this prior to performing regression analysis in order to help determine which attributes are most likely to determine default, only considering those that fall above a certain threshold. 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 independent attributes that passed the correlation threshold hold. This will form a baseline for our model. 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. Lastly, we will begin to explore elements of machine learning to attempt to form an even more accurate prediction of loan default. From this, we will generate a model that takes a number of boolean and numeric arguments to estimate the probability of default.

4.3 Evaluation

We will evaluate our research by testing to see if it 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 regression, decision trees, and random forests. 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

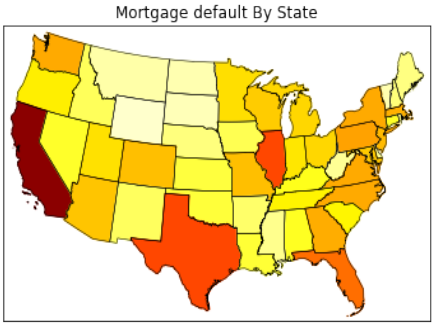
the housing bubble. This knowledge can be used to predict future bubbles and characteristic of risky mortgage lending.

5 ECONOMIC FACTORS IN DEFAULT

5.1 Dataset

For our analysis in local factors of default, we look to the Bureau of Economic Analysis for economic information on each state in the country. We look specifically to personal income, net earnings by residence, employment, wage level, and unemployment compensation. Using the Basemap Python package, we were able to visually map these attributes to states. By doing the same with defaults, we can compare the resulting maps to see which economic factors tend to be most correlated with mortgage default.

5.2 Geographic Analysis of Mortgage Default



Seeing the concentration of defaults in certain states, namely California, Texas, Illinois, and Florida, it may be possible to more accurately determine if a person will default simply on their location, or on the characteristics of that location (i.e. income, crime, etc).

6 MODELS FOR PREDICTION OF DEFAULT

6.1 Regression Analysis

6.1.1 Introduction to Regression. Linear regression is a statistical method that uses least squares estimation to formulate a model that expresses the relationship between variables in data. Classic linear regression relates some independent variable X to the behavior of an ostensibly dependent variable Y through a linear relationship.

$$y_i = \beta x_i + \epsilon, \quad i \, \epsilon \, n \tag{1}$$

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 that follows a normal distribution with mean zero and variance 1 to account for unobserved variables . This implies

$$\begin{aligned} y_1 &= \beta x_1 + \epsilon_1 \\ y_2 &= \beta x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta x_n + \epsilon_n \end{aligned}$$

6.1.2 A Matrix Approach to Regression. By defining the following variables:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 \\ x_1 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{2}$$

We can express the linear regressive model as a matrix product

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

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.

6.1.3 Determining Goodness of Fit. In order to properly assess the goodness of fit (GoF) of a regression model, it is necessary to define some basic measures used in determining the model's efficiency.

$R - Squared$:
R^2 is used to denote the coefficient of determination, which is defined as the proportion of the variance in the dependent variable Y that can be predicted from the independent variables X. It is defined as

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y})}{\sum_{i=0}^{n}(y_i - \bar{y})} = 1 - \frac{RSS}{TSS} \tag{4}$$

where RSS denotes the residual sum of squares and TSS denotes the total sum of squares. If $R^2 = 1$ then the model is said to be perfectly predictive.

In logistic regression, an equivalent statistic to R^2 does not exist. This is because the results of logistic regression are arrived at iteratively with the use of maximum likelihood estimators. The resulting estimates do not have minimized variances, as in the ordinary least-squares regression. As such, in the case of logistic regression we turn to a pseudo R^2. There are various methods for calculating this, in this paper we will be using McFadden's Pseudo R^2, defined as

$$R^2_{pseudo} = 1 - \frac{\ln[\hat{L}(M_{fall})]}{\ln[\hat{L}(M_{intercept})]} \tag{5}$$

where $\ln[\hat{L}]$ denotes the log-likelihood function, M_{fall} denotes models with predictors, and $M_{intercept}$ models without predictors.

Standarderror: To determine the uncertainty in our model, we look to the standard error, defined as

$$\sigma_{x_i} = \sqrt{\sum_{i=0}^{n}(y_i - \hat{y_i})^2} \tag{6}$$

This statistic is used to calculate the confidence interval, where

$$C_{0.025, x_i} = \beta_i - (1.96) * \sigma_{x_i}$$

and

$$C_{0.975, x_i} = \beta_i + (1.96) * \sigma_{x_i}$$

HypothesisTesting: In testing the relevance of variables in a regressive model, we must explore hypothesis testing of variables. Let us define

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

Where H_0 denotes the null hypothesis, that β_i has no effect on Y, and H_a denotes the alternative hypothesis, that X_i is a contributing factor in determining the likelihood of Y. To test these hypotheses, we look to another statistic, the p-value.

$P - value$: We define the probability value p as the probability of rejecting the null hypothesis given that it is true, mathematically as

$$p = P(Reject(H_0)|H_0)$$

When testing the hypothesis, we may quantify our certainty by choosing a significance level α. If p<α, the result is said to be statistically significant, we then reject the null hypothesis and accept the alternative. To find a confidence level, we calculate $C = 1 - \alpha$. If $\alpha = 0.05$ for example, we say that we are 95% certain that the true value of β_i lies between the confidence interval. A confidence level of $\alpha = 0.05$ will be used in this paper for all hypothesis tests.

6.2 Logistic 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:

$$P(Y = 1|X) = \Phi(X^T\beta) + \epsilon \tag{7}$$

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

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

$$\ln \omega = \ln \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = X^T\beta + \epsilon \tag{8}$$

For this paper, we will examine a logistic approach. Let us define omega as the true odds ratio of the regression, and omega-hat as the predicted odds ratio. Then

$$\ln \hat{\omega} = \ln \left[\frac{\hat{Y}}{1 - \hat{Y}} \right] = X^T\hat{\beta} \tag{9}$$

Solving for the probability, $\hat{Y}$, we obtain

$$\hat{Y} = \frac{e^{X^T\hat{\beta}}}{e^{X^T\hat{\beta}} + 1} = \frac{e^{\ln \hat{\omega}}}{e^{\ln \hat{\omega}} + 1} \tag{10}$$

6.3 Bayesian Classifiers

Bayesian Classifiers are a prediction method that uses probability rather than regression, which can be useful for the non-numerical data. Gaussian Naive Bayes makes the assumption that all attributes are independent. Using the formula:

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

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. SciKit Learn has a built in function called GaussianNB() which takes in a target training set and a features training set in order to come up with a classification model. The model then takes in as input a testing features subset of the data and returns the target results. The attributes that we choose to run the classification technique is based on the attributes with the highest correlation coefficient when compared with the binary attribute labeled default. This is done to limit the dimensionality of our analysis with the intention of limiting the negative effects of the curse of dimensionality. The attributes that we deemed most significant and ran with this method were: credit score, debt to income ratio, loan to value ratio, and the original interest rate on the loan.

6.4 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. A decision tree is built by calculating the entropy of a given feature, X.

$$\mathrm{H}(X) = \mathbb{E}[\mathrm{I}(X)] = \mathbb{E}[-\ln(\mathrm{P}(X))]. \tag{11}$$

This is more commonly seen in the summation form,

$$\mathrm{H}(X) = \sum_{i=1}^{n} \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_{i=1}^{n} \mathrm{P}(x_i)\log_b \mathrm{P}(x_i), \tag{12}$$

and this formula is used to pick which feature is the best to split on first. The entropy value is a difference between the current entropy, and what will be "gained" from using this feature.

6.5 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, can cause overfitting, and can be biased towards a more common result in some training sets. The Random Forest technique can improve upon these problems and in some cases can improve accuracy, true positive rate, and false negative rate.

A Random Forest is a large amount of random decision trees. Each of these trees has a random amount of attributes, a, out of the amount of possible attributes, A, where a « A. A good rule for choosing a is to have a = $\sqrt{A}$, but this value can be changed and the best value

for a must be found experimentally. We then generate N decision trees using bootstrap aggregating so that each decision tree has a different subset of the data for training and testing. We then decide on splits and generate each decision tree using the method described in the previous section on the training data for that tree. We then test each tree in turn using its respective testing data. We choose the best n trees, where n < N, using some method of ranking. We can look to maximize accuracy of the tree:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Maximize recall, or true positive rate:

$$\frac{TP}{TP + FN}$$

Or minimize false positve rate:

$$\frac{FP}{TN + FP}$$

or use some combination of these measures to prune our forest. In this case, a positive means default occured while a negative means there was no default. Once we have our new forest of good trees, we can now predict test data with a higher accuracy than any single tree by using all trees. By giving each tree the same datapoint to predict, we get n predctions for the result of the data. Using a majority vote, or some other minimum threshold, we can combine the results from all the trees into one more accurate result.

7 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

8 MILESTONES

8.1 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
4/8: Perform logistic regressions on data

8.2 To Do

4/15: Apply machine learning methods to data
4/28: Document and process results

9 RESULTS

9.1 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.

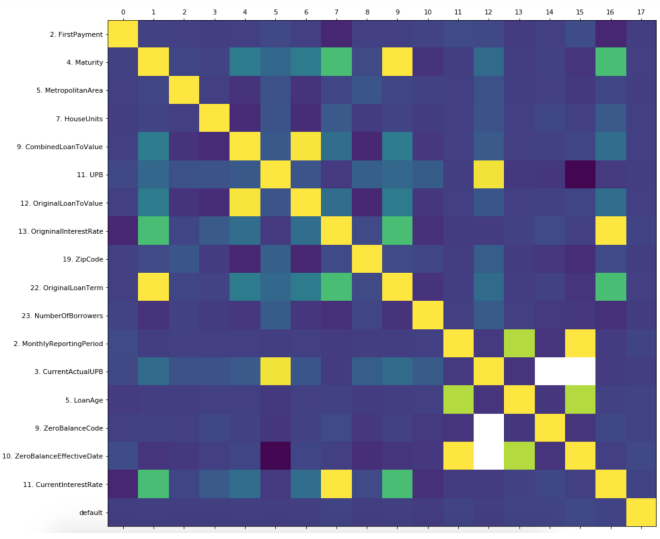


Figure 1: A correlation matrix of all attributes

From these calculations we can grab the attributes with the highest correlation measures with default to form a regression model. The first four variables that are highly correlated with default and relevant to our analysis (i.e. not a date or zip code) are credit score, debt-to-income ratio, original loan-to-value ratio, and original interest rate. For definitions of these terms, please turn to appendix (a).

9.2 Logit Regression

9.2.1 Univariate Logistic Regression. Below is an early regression with one independent variable, credit score. We choose this variable for two reasons. Firstly, we expect credit score to be fairly telling of default rate due to its nature in predicting the ability for its holder to repay a loan. Secondly, according to our initial methodology we choose the attribute that has the highest correlation with default, this is credit score. We form a predictive model

$$\ln \hat{\omega} = \hat{\beta}_0 + X\hat{\beta}_1 \tag{13}$$

where β_0 denotes the intercept of the model and β_1 denotes $\beta_{creditScore}$, the marginal effect of credit score on the log of the odds ratio. Additionally, let us define a hypothesis test for the attribute creditScore, namely

$$H_0 : \beta_1 = 0 \tag{14}$$

$$H_a : \beta_1 \neq 0 \tag{15}$$

Running this, we get the following results

Logit Regression Results

Dep. Variable:	default	No. Observations:	208282
Model:	Logit	Df Residuals:	208280
Method:	MLE	Df Model:	1
Date:	Wed, 18 Apr 2018	Pseudo R-squ.:	0.06268
Time:	18:15:13	Log-Likelihood:	-15839.
converged:	True	LL-Null:	-16898.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
betaNot	6.4457	0.215	29.920	0.000	6.023	6.868
1. creditScore	-0.0145	0.000	-47.845	0.000	-0.015	-0.014

Figure 2: Logit regression of credit score on default rate

Here we have a positive albeit low pseudo R^2, and so our model predicts approximately 6% of the variation in the data. Looking to the attributes in the lower portion of the diagram, we see that $\beta_0 = 6.45$ and $\beta_1 = -0.0145$. Immediately we see that the the marginal effect of credit score on log of the odds ratio is negative, so the model predicts that increasing credit score should decrease the probability of defaulting in some fashion. Using the formula (10) we can calculate the estimated probability of default $\hat{Y}$ as

$$\hat{Y} = \frac{e^{6.45-0.0145*(x)}}{e^{6.45-0.0145*(x)} + 1}$$

where x denotes the credit score of an individual. Note that the standard error for X_i is close to zero, meaning that there is minimal uncertainty in the estimate, and the P-value is less than $\alpha = 0.05$, meaning there is strong evidence that credit score has an effect on default rate. As such, we reject the null hypothesis and accept the alternate hypothesis that credit score does effect default rate.

9.2.2 Multivariate Logistic Regression. Now that we have established a baseline, we can add additional attributes in order of their correlation to the default attribute. Adding the three attributes with the highest correlations after credit score, we form the following multivariate model in a matrix form

$$\ln \hat{\omega} = X^T \hat{\beta} \tag{16}$$

where

$$x = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} \tag{17}$$

Running the regression, we get the following results

Logit Regression Results

Dep. Variable:	default	No. Observations:	205205
Model:	Logit	Df Residuals:	205200
Method:	MLE	Df Model:	4
Date:	Wed, 18 Apr 2018	Pseudo R-squ.:	0.09114
Time:	23:25:40	Log-Likelihood:	-15088.
converged:	True	LL-Null:	-16601.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
betaNot	-1.1695	0.368	-3.179	0.001	-1.890	-0.449
1. creditScore	-0.0117	0.000	-33.771	0.000	-0.012	-0.011
10. DebtToIncome	0.0258	0.002	16.739	0.000	0.023	0.029
12. OriginalLoanToValue	0.0157	0.001	11.986	0.000	0.013	0.018
13. OrigninalInterestRate	0.5541	0.030	18.173	0.000	0.494	0.614

Figure 3: Logit regression of credit score, debt-to-income ratio, loan-to-value ratio, and interest rate on default rate

We see that the value for R^2_{Pseudo} has increased by approximately three percent, hence our new multivariate model is more predictive than our initial univariate model, though not by much. Additionally, every estimated β has a p-value that is less than the significance level $\alpha = 0.05$. Our null hypotheses are

$$H_0 : \quad \beta_i = 0, \ i \epsilon (1, 2, 3, 4)$$

while our alternative hypotheses are

$$H_\alpha : \quad \beta_i \neq 0, \ i \epsilon (1, 2, 3, 4)$$

Since $p < \alpha$ for all X, we reject the null hypothesis in each case and conclude that every independent variable in X has a statistically significant effect on the dependent variable, default.

9.2.3 Logistic Regression with State Data. Knowing that some states carry more risk than others, we consider adding states into our model in an attempt to better represent default. Using the three states with the highest default rate, California, Texas, and Florida, we form a new model. Following the same form as the last two models, the results are as follows:

Logit Regression Results

Dep. Variable:	default	No. Observations:	208432
Model:	Logit	Df Residuals:	208424
Method:	MLE	Df Model:	7
Date:	Wed, 25 Apr 2018	Pseudo R-squ.:	0.1056
Time:	18:42:10	Log-Likelihood:	-25741.
converged:	True	LL-Null:	-28780.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
betaNot	-0.3665	0.265	-1.385	0.166	-0.885	0.152
1. creditScore	-0.0119	0.000	-47.408	0.000	-0.012	-0.011
10. DebtToIncome	0.0254	0.001	22.927	0.000	0.023	0.028
12. OriginalLoanToValue	0.0167	0.001	17.640	0.000	0.015	0.019
13. OrigninalInterestRate	0.5564	0.022	25.269	0.000	0.513	0.600
TX	-0.4200	0.059	-7.174	0.000	-0.535	-0.305
FL	0.3946	0.050	7.908	0.000	0.297	0.492
CA	0.0827	0.039	2.116	0.034	0.006	0.159

We see that including the states has increased the accuracy of the model, as R^2_{Pseudo} has increased to 0.1056. As such this model is the most accurate logistic model in this paper. Though this model is fairly inaccurate, it provides a baseline of attributes to consider, and a model from where to begin our analysis of default rates.

9.3 Bayesian classification

9.4 Decision Trees

9.4.1 Data Cleaning.
Using the decision tree algorithm, we can generate a tree to see if we can predict a given loan is going to be defaulted on, or not. First, we must clean the data of all NaNs values and blanks. This is achieved by iterating through each row in each column, checking if it's valid or not, and then if it's not, we decided to randomly select another valid field to fill in, as to complete that data.

9.4.2 Conversion to scalars.
After cleaning the data, it must be converted into scalar values in order to use the decision tree algorithm. We used a library called sklearn_pandas, which bridges the gap between the pandas library and the sklearn library. It takes data from the dataframe and given a mapping of features with the type of feature (Binary, scalar, etc..) it maps the data from the dataframe into an array suitable for use in sklearn.

9.4.3 Visualization.
Then the data is loaded into the decision tree, and a tree is generated. We then can use the graphviz library to expand that tree into a viewable graphic, as seen in the appendix C.

9.5 Random Forests

10 KEY FINDINGS

10.1 Logistic Regression

Our final model performed rather inaccurately with a pseudo R^2 of approximately 10%. Despite this, from our

hypothesis tests we were able to prove the statistical significance of the inclusion of specific variables, namely credit score, debt to income ratio, loan to value ratio, original interest rate, and residence in Texas, California, Florida, or some other state. Additionally, we were able to weight the respective effect of each attribute on probability of default, with interest trate weighted highest, followed bydebt to income ratio, This provided a baseline for analysis with other methods such as Bayesian Classification, decision trees, and random forests.

10.2 Bayesian Classification

After implementing SciKit Learn's Guassian Naive Bayes function, called GaussianNB(), we were able to build a model will an accuracy of 98 percent. This model can be used by credit rating agency's in conjunction with their current models to enhance accuracy of predicting mortgage defaults.

10.3 Decision Trees

The model that was trained for the decision tree algorithm performed at a 73% accuracy rate on the test set using 4 million data points for the entire set, being split on 3 million training, and 1 million test. This result isn't that great, since I was expecting higher, with such an advanced algorithm, but Freddie Mac must have been very picky in their loans, as most of the data was non-defaulted. This makes sense, since they don't make money on defaulted loans.

10.4 Random Forest

Using the Random Forest method, we were able to generate forests that on average had an accuracy of 98.2%, a recall, or true positive rate of 1.25%, a false positive rate of less than .001%, and a true positive to false positive ratio of 3:1. Again, In a positive means default occured while a negative means there was no default. This means that using this method, Freddie Mac could have denied a few loans to defaulters, and would have only rejected a third as many loans to people who would not have defaulted.

11 APPLICATIONS OF FINDINGS

11.1 Risk Assessment

These methods can be used by credit rating agencies in the financial industry such as Moody's Investors Services and Standard and Poor's in order to more accurately assess credit risk by quantifying the effects of certain attributes to gain a better measure of risk. Such factors can be used in application with mortgage ratings.

11.2 Predicting a market bubble before it happens

This could be used by an economist to look for a bubble in a market. Searching through different attributes in loans, there could be signs that could be useful. Such methods can also be used to analyze the types of loans that are being given out characterized by predicting a number of defaults above a particular threshold.

11.3 Large Scale Default Prediction

Once loans are made, the models we formed could be run to predict the rate of default on the set of loans made. That is, we can answer the question: "How many people out of the number I just gave a mortgage to are going to default?" This is the knowledge gained from the analysis, but this knowledge can be used by analysts to calculate a more accurate measure of risk for the company. Following from this, once a new level of risk has been calculated, assuming it is accurate it can be used to predict profit or loss on the set of loans made. Hence this knowledge can be used to maximize expected profit or minimize expected loss by estimating the proportion of people in a set of loans who will default.

11.4 Marginal Loan Denial

The methods we developed can be combined with preexisting credit-score measurements to marginally increase their accuracy. For example, the random forest method had low recall, or true positive rate, but had even lower false negative rate. Because of this,

we can use methods such as these along with the pre-existing methods and credit score to deny loans to a small amount of would-be defaulters, while denying loans only to an even smaller amount of non-defaulters.

12 APPENDIX

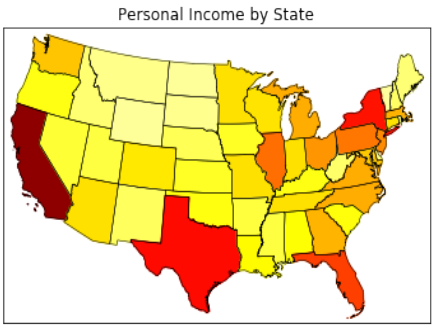
12.1 A: Financial Definitions

12.1.1 Credit Score. The credit score being used in the Freddie Mac dataset is a FICO score. The FICO model is used by the majority of banks as well as credit grantors, and is based on consumer credit files from one or more of the three national credit bureaus: Experian, Equifax, and TransUnion. FICO scores take on values between 300 and 850, with a higher score indicating a lower risk individual. According to FICO, the median score in 2011 was a 711.

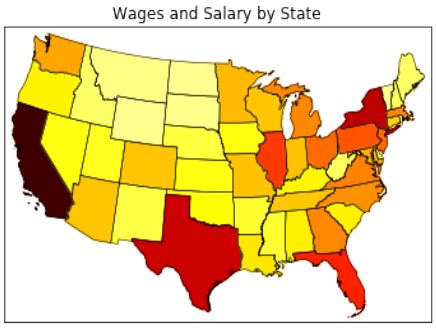
12.1.2 Debt-to-Income Ratio (DTI). DTI is the percentage of a consumers gross income that goes towards paying debt. In our dataset it is calculated by dividing monthly debt payments by monthly income. The higher this value is, the more we would expect the individual to default.

12.1.3 Loan-to-Value Ratio. The loan-to-value ratio is a ratio of the loaned amount of money used to purchase a property and the value of that property. We would expect that the higher this value is, the more likely an individual is to default. That is, the more money an individual has to borrow to pay for a house in terms of the percent value of that house, the more likely they are to default.

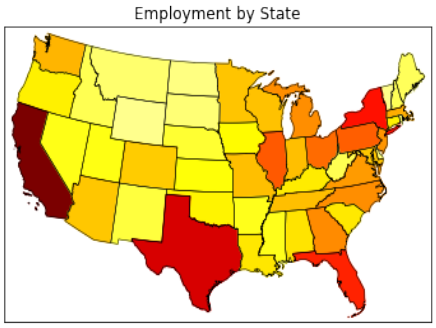
12.2 B: Data Visualizations



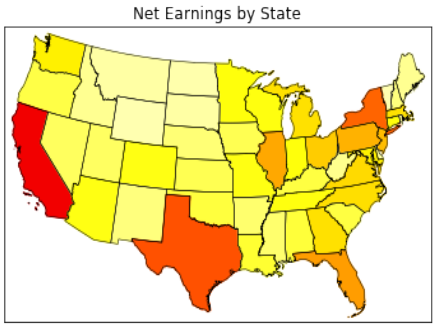
12.2.1 Personal Income.



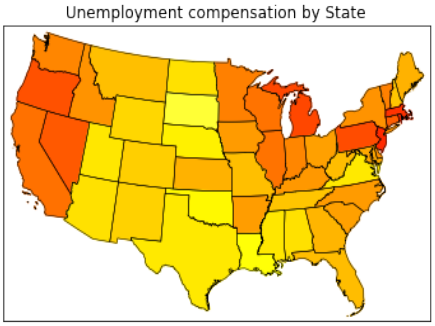
12.2.2 Wages and Salary.



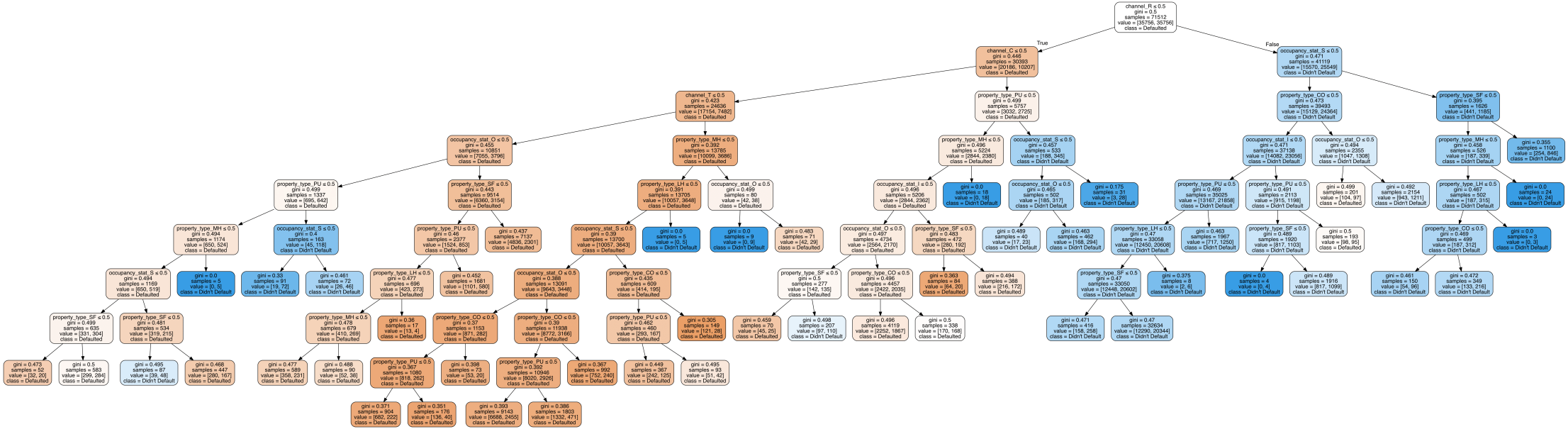
12.2.3 Total Employment.



12.2.4 Net Earnings by Residence.



12.2.5 Unemployment Compensation.


channel_R ≤ 0.5
gini = 0.5
samples = 71512
value = [35756, 35756]
class = Defaulted
True
False
channel_C ≤ 0.5
gini = 0.446
samples = 30393
value = [20186, 10207]
class = Defaulted
occupancy_stat_S ≤ 0.5
gini = 0.471
samples = 41119
value = [15570, 25549]
class = Didn't Default
channel_T ≤ 0.5
gini = 0.423
samples = 24636
value = [17154, 7482]
class = Defaulted
property_type_PU ≤ 0.5
gini = 0.499
samples = 5757
value = [3032, 2725]
class = Defaulted
property_type_CO ≤ 0.5
gini = 0.473
samples = 39493
value = [15129, 24364]
class = Didn't Default
property_type_SF ≤ 0.5
gini = 0.395
samples = 1626
value = [441, 1185]
class = Didn't Default
occupancy_stat_O ≤ 0.5
gini = 0.455
samples = 10851
value = [7055, 3796]
class = Defaulted
property_type_MH ≤ 0.5
gini = 0.392
samples = 13785
value = [10099, 3686]
class = Defaulted
property_type_MH ≤ 0.5
gini = 0.496
samples = 5224
value = [2844, 2380]
class = Defaulted
occupancy_stat_S ≤ 0.5
gini = 0.457
samples = 533
value = [188, 345]
class = Didn't Default
occupancy_stat_I ≤ 0.5
gini = 0.471
samples = 37138
value = [14082, 23056]
class = Didn't Default
occupancy_stat_O ≤ 0.5
gini = 0.494
samples = 2355
value = [1047, 1308]
class = Didn't Default
property_type_MH ≤ 0.5
gini = 0.458
samples = 526
value = [187, 339]
class = Didn't Default
gini = 0.355
samples = 1100
value = [254, 846]
class = Didn't Default
property_type_PU ≤ 0.5
gini = 0.499
samples = 1337
value = [695, 642]
class = Defaulted
property_type_SF ≤ 0.5
gini = 0.443
samples = 9514
value = [6360, 3154]
class = Defaulted
property_type_LH ≤ 0.5
gini = 0.391
samples = 13705
value = [10057, 3648]
class = Defaulted
occupancy_stat_O ≤ 0.5
gini = 0.499
samples = 80
value = [42, 38]
class = Defaulted
occupancy_stat_I ≤ 0.5
gini = 0.496
samples = 5206
value = [2844, 2362]
class = Defaulted
gini = 0.0
samples = 18
value = [0, 18]
class = Didn't Default
occupancy_stat_O ≤ 0.5
gini = 0.465
samples = 502
value = [185, 317]
class = Didn't Default
gini = 0.175
samples = 31
value = [3, 28]
class = Didn't Default
property_type_PU ≤ 0.5
gini = 0.469
samples = 35025
value = [13167, 21858]
class = Didn't Default
property_type_PU ≤ 0.5
gini = 0.491
samples = 2113
value = [815, 1198]
class = Didn't Default
gini = 0.499
samples = 201
value = [104, 97]
class = Defaulted
gini = 0.492
samples = 2154
value = [943, 1211]
class = Didn't Default
property_type_MH ≤ 0.5
gini = 0.467
samples = 502
value = [187, 315]
class = Didn't Default
gini = 0.0
samples = 24
value = [0, 24]
class = Didn't Default
property_type_MH ≤ 0.5
gini = 0.494
samples = 1174
value = [650, 524]
class = Defaulted
occupancy_stat_S ≤ 0.5
gini = 0.4
samples = 163
value = [45, 118]
class = Didn't Default
property_type_PU ≤ 0.5
gini = 0.46
samples = 2377
value = [1524, 853]
class = Defaulted
gini = 0.437
samples = 7137
value = [4836, 2301]
class = Defaulted
occupancy_stat_S ≤ 0.5
gini = 0.39
samples = 13700
value = [10057, 3643]
class = Defaulted
gini = 0.0
samples = 5
value = [0, 5]
class = Didn't Default
gini = 0.489
samples = 40
value = [17, 23]
class = Didn't Default
gini = 0.463
samples = 462
value = [168, 294]
class = Didn't Default
property_type_LH ≤ 0.5
gini = 0.47
samples = 33058
value = [12450, 20608]
class = Didn't Default
gini = 0.463
samples = 1967
value = [717, 1250]
class = Didn't Default
property_type_PU ≤ 0.5
gini = 0.491
samples = 2113
value = [815, 1198]
class = Didn't Default
gini = 0.499
samples = 201
value = [104, 97]
class = Defaulted
gini = 0.492
samples = 2154
value = [943, 1211]
class = Didn't Default
property_type_LH ≤ 0.5
gini = 0.467
samples = 502
value = [187, 315]
class = Didn't Default
gini = 0.0
samples = 24
value = [0, 24]
class = Didn't Default
occupancy_stat_S ≤ 0.5
gini = 0.494
samples = 1169
value = [650, 519]
class = Defaulted
gini = 0.0
samples = 5
value = [0, 5]
class = Didn't Default
gini = 0.33
samples = 91
value = [19, 72]
class = Didn't Default
gini = 0.461
samples = 72
value = [26, 46]
class = Didn't Default
property_type_LH ≤ 0.5
gini = 0.477
samples = 696
value = [423, 273]
class = Defaulted
gini = 0.452
samples = 1681
value = [1101, 580]
class = Defaulted
occupancy_stat_O ≤ 0.5
gini = 0.388
samples = 13091
value = [9643, 3448]
class = Defaulted
property_type_CO ≤ 0.5
gini = 0.435
samples = 609
value = [414, 195]
class = Defaulted
property_type_SF ≤ 0.5
gini = 0.5
samples = 277
value = [142, 135]
class = Defaulted
property_type_CO ≤ 0.5
gini = 0.496
samples = 4457
value = [2422, 2035]
class = Defaulted
gini = 0.363
samples = 84
value = [84, 20]
class = Defaulted
gini = 0.494
samples = 388
value = [216, 172]
class = Defaulted
property_type_CO ≤ 0.5
gini = 0.469
samples = 499
value = [187, 312]
class = Didn't Default
gini = 0.0
samples = 3
value = [0, 3]
class = Didn't Default
property_type_SF ≤ 0.5
gini = 0.499
samples = 635
value = [321, 304]
class = Defaulted
property_type_SF ≤ 0.5
gini = 0.481
samples = 534
value = [319, 215]
class = Defaulted
property_type_MH ≤ 0.5
gini = 0.478
samples = 679
value = [410, 269]
class = Defaulted
gini = 0.36
samples = 17
value = [13, 4]
class = Defaulted
property_type_CO ≤ 0.5
gini = 0.37
samples = 1153
value = [971, 292]
class = Defaulted
property_type_CO ≤ 0.5
gini = 0.39
samples = 11938
value = [8772, 3166]
class = Defaulted
property_type_PU ≤ 0.5
gini = 0.462
samples = 460
value = [293, 167]
class = Defaulted
gini = 0.305
samples = 149
value = [121, 28]
class = Defaulted
gini = 0.459
samples = 70
value = [45, 25]
class = Defaulted
gini = 0.498
samples = 207
value = [97, 110]
class = Didn't Default
gini = 0.496
samples = 4119
value = [2252, 1867]
class = Defaulted
gini = 0.5
samples = 338
value = [170, 168]
class = Defaulted
property_type_SF ≤ 0.5
gini = 0.47
samples = 33050
value = [12448, 20602]
class = Didn't Default
gini = 0.375
samples = 8
value = [2, 6]
class = Didn't Default
property_type_SF ≤ 0.5
gini = 0.469
samples = 1920
value = [817, 1103]
class = Didn't Default
gini = 0.5
samples = 193
value = [96, 95]
class = Defaulted
property_type_CO ≤ 0.5
gini = 0.469
samples = 499
value = [187, 312]
class = Didn't Default
gini = 0.461
samples = 150
value = [54, 96]
class = Didn't Default
gini = 0.473
samples = 52
value = [32, 20]
class = Defaulted
gini = 0.5
samples = 583
value = [299, 284]
class = Defaulted
gini = 0.486
samples = 67
value = [39, 48]
class = Didn't Default
gini = 0.468
samples = 447
value = [280, 167]
class = Defaulted
gini = 0.477
samples = 589
value = [358, 231]
class = Defaulted
gini = 0.488
samples = 90
value = [52, 38]
class = Defaulted
property_type_PU ≤ 0.5
gini = 0.367
samples = 1080
value = [818, 262]
class = Defaulted
gini = 0.398
samples = 73
value = [53, 20]
class = Defaulted
property_type_PU ≤ 0.5
gini = 0.392
samples = 10946
value = [8020, 2926]
class = Defaulted
gini = 0.397
samples = 992
value = [752, 240]
class = Defaulted
gini = 0.449
samples = 367
value = [242, 125]
class = Defaulted
gini = 0.495
samples = 93
value = [51, 42]
class = Defaulted
gini = 0.371
samples = 904
value = [682, 222]
class = Defaulted
gini = 0.351
samples = 176
value = [136, 40]
class = Defaulted
gini = 0.393
samples = 9143
value = [6688, 2455]
class = Defaulted
gini = 0.386
samples = 1803
value = [1332, 471]
class = Defaulted
gini = 0.375
samples = 8
value = [2, 6]
class = Didn't Default
gini = 0.47
samples = 32634
value = [12290, 20344]
class = Didn't Default
gini = 0.489
samples = 1916
value = [817, 1099]
class = Didn't Default