An Analysis in Situational Factors on Determining Credit Default Rates

Authors

Spencer Hanson
spha0827@colorado.edu

Benjamin Miller
bemi0682@colorado.edu

Maxim Moghadam
mamo5089@colorado.edu

William Brickowski
wibr5703@colorado.edu

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.

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, and all of this work has been 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

The classic credit score is a system of trust, 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. 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. This system is reactive, and can only predict based on these reactions.

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

[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), 523?547. DOI:http://dx.doi.org/10.1111/1540-6229.00811

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

We will need to convert our data into compatible forms and then merge them together. This will include creating matching null values for unreported or incorrect observations and matching monthly performance data to originated loan data and geographic economic data. We will also need to create an attribute that states whether a person defaulted on their loan. We define this as the event that a loan became 60 days delinquent at some point in its life, this will be used as our standard for the attribute "default".

4.2 Analysis

Using our datasets, we will form logistic regressions with loan default as the dependent variable, 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.

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%.

5 MODELS FOR PREDICTION OF DEFAULT

5.1 Logit and Probit Regression

Regression will be our baseline approach for analyzing default rates. Classic linear regression relates some independent variable X to the behavior of an ostensibly dependent variable Y. In our project, X will represent a matrix of observations while Y will be a categorical or probabilistic vector representing the event of defaulting on a loan, in this event Y_i will take on the value 0 or 1 (categorical) or a value between 0 and 1 (probabilistic).

5.2 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

[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

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.

5.3 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.

5.4 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, Pandas, Numpy, SciKit-Learn, MongoDB, Microsoft Excel, MatPlotlib, MatLab

7 MILESTONES

2/23: Load sample data into python
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/probabilistic regressions on data
4/15: Apply machine learning methods to data
4/28: Document and process results