Financing the American Dream: Using Logistic Regression and Principal Component Analysis to Identify the Probability of Default in Mortgage Lending

Researcher: Max McDonald [*1], Mentors: Dr. Frances Maloy[†2], and Professor James Morrow [‡2]

[1]University of Washington
[2]Michael G. Foster School of Business
[2]Department of Mathematics

June 8, 2015

Abstract

Financing the American Dream is an investigation of the U.S. mortgage crisis that triggered the Great Recession of December 2009 - June 2012. This paper analyzes the borrower and loan characteristics of a subset of U.S. mortgage loans acquired by Fannie Mae in 2006, whose performance is tracked through 2012. We analyze the data using the statistical technique of **Logistic Regression Analysis** in order to identify the relationship of various independent variables with the probability of mortgage loan default. Additionally, we apply the statistical analysis technique of **Principal Component Analysis** in order to identify which independent variables encode the majority of the variance within the data set. The motivation of the project is to understand the likelihood of defaulting on a mortgage loan given specific borrower and loan characteristics.

Keywords: Securitization, Statistical Analysis, Financial Crisis

[*] `mdjmcdonald@gmail.com`
[†] `thanks@uw.edu`
[‡] `jamorrow@uw.edu`

Acknowledgments

This project would not have been possible without the help from many people within the University of Washington community. I am extremely grateful for the guidance and dedication of my mentor Dr. Frances Maloy at the Michael G. Foster School of Business, who has spent countless hours emailing, editing, finding data, and coaching me on this project from conception to delivery. Also, to my mentor in the Mathematics Department, Professor James Morrow, to the McNair Scholar group for funding and guidance, to my professors Stephan Siegel and Kathryn Dewenter for academic and professional encouragement and guidance at every turn, and to Dr. Therese Mar, my boss at the University of Washington Instructional Center, for genuine support and encouragement. Also, I want to thank my good friend Jammy Phan (Michael G. Foster School of Business, '2014), who spent several days, including his entire birthday, organizing the data with me in Savery 117. I'd also like to thank Laney Routledge (Ph.D. candidate in Economics, UW) for teaching me how to use Stata and to my classmate Zezhou Jing (Advanced Honors Calculus) for teaching me how to use LaTeX. Thank you!!!

Contents

1 Introduction

1.1 Background

In 2008-09 the United States entered into the deepest recession since the Great Depression of 1929-33. Although strikingly similar, the depth of integration between modern global financial centers exacerbated the effects of this recession triggering a global crisis. Although there remains many open questions regarding the blame, most analysts would agree that the runaway residential mortgage lending practices in the United States were the impetus of the shock [see references 2, 3, and 5]. It is clear that the development of a modern mortgage lending infrastructure in the United States has yielded benefits for homeowners and investors alike; in recent years we have witnessed historically low interest rates on home mortgages. Moreover, these rates offer investors higher returns than the would have received on government treasuries. However, many of the financial products created during the period leading up to the Crisis were dubious; the underlying risk was difficult to understand and hence, price. These products set the stage for a financial system collapse.

Several of these products, enabled by the process of securitization, played a central role in the development of the sub-prime mortgage industry. As the financial products continued to develop, wider sectors of the mortgage industry were included. During the years leading up to the Crisis, FI's were encouraged to develop more complex financial instruments to satisfy returned starved investors. One of the products, Adjustable-Rate Mortgages (ARM's), was created in the wake of the OPEC oil shock of the 1970's in order to extend mortgage credit to the market. Before the Crisis, FI's began issuing ARM's at a higher rate than before. There are also interest-only and non-documentation loans. The ARM's are organized such that the interest rates, and hence payments, are very low at the beginning of the contract and they reset later, usually within a period of 1 to 5 years. Interest-only loans are, as the name suggests, interest only for a certain number of years before converting to principal and interest payments. Therefore, the principal mortgage on the loan has no chance of being reduced during the initial contract period of these loans. With non-documentation loans, originating FI's don't require proof of income such as tax statements, bank records, and proof of income; they merely go by what is stated on the application. By pooling mortgages, sub-prime and non sub-prime, securitization obscured the risk inherent in the underlying mortgages.

This paper takes an analytical view on the U.S. mortgage crisis of 2007-2009. Specifically, we will consider the effects of securitization on the extension of mortgage credit to nontraditional sectors of the housing market. We run statistical tests on a data set made available by Fannie Mae, a Government Sponsored Entity dedicated to supporting the residential mortgage industry. From a theoretical perspective, the paper raises the questions of proper governmental oversight, the nature of risk-identification under rapid financial product innovation, and the role of private firms in creating industrial organizational structures that minimize agency problems.

1.2 Definition of Key Terms

Definition 1.2.1 (**Financial Crisis**). An event in which households and firms lose confidence in the financial system; they lose faith in the underlying value of financial assets such as bonds, stock, and money. Financial crises are characterized by "bank runs", which means that depositors go to banks en masse and demand their funds. Because banks lend out the funds that have been deposited, for homes, cars, etc., if enough people line up at the same time to demand their funds, eventually the bank becomes insolvent.

Definition 1.2.2 (**Securitization**). The pooling of cash flows from underlying debt contracts (mortgages, auto loans, credit cards, etc.) to create bonds for which investors pay to receive the rights to the cash flows. Starting in the United States shortly after the Great Depression, securitization has played a central role in bringing liquidity to the market for residential mortgages and has therefore expanded the American Dream of owning a home. Although securitization has contributed greatly to the development of a well-greased mortgage industry, it took on an unprecedented risky nature in the years leading up to the Great Recession.

Definition 1.2.3 (**Government Sponsored Entity, GSE**). Any of the large U.S. Government founded and backed entities which are dedicated to the residential mortgage industry. The largest of these are: Federal Home Mortgage Loan Corporation (Freddie Mac), Federal National Mortgage Association (Fannie Mae), and the Government National Mortgage Association (Ginnie Mae).

Definition 1.2.4 (**Sub-prime Mortgage**). Mortgage credit extended to an applicant with a poor credit history due to any combination of untimely payments on debt obligations, charged-off debt, and high credit risk due to overleverage.

Definition 1.2.5 (**Financial Institution, FI**). Firms dedicated to facilitating the financing needs of households, firms, and governments. These are typically depository institutions (banks), non-bank lenders (the lenders who were originally originating sub-prime mortgages), insurance and finance companies, securities firms, pension and mutual funds. These firms manage everything from deposits and loans to insurance and investment needs.

Definition 1.2.6 (Loan Origination). The process of contracting a financial agreement between a mortgagor and a mortgagee. FI's typically charge a fee for facilitating this service. Fees are typically between 100 to 150 basis points (1-1.5%) of the total amount issued against the property. Before the development of securitization, FI's originated mortgages with funds deposited by their depositors and monitored the loans until term maturity. The Crisis was characterized by the propagation of non-bank FI's. These organization originated the loans and then sold them to larger FI's.

Definition 1.2.7 (Collateralized Mortgage Obligation, CMO). The CMO is a derivative of the Mortgage-Backed Bond. CMO's make the Mortgage-Backed Bond more attractive by reorganizing the risk of the underlying assets. The underlying assets are pooled into different categories, or tranches, according to the perceived risk of the assets. CMO's are used as a way to distribute risk of MBB's according to the investors' risk tolerance levels.

Definition 1.2.8 (Mortgage-Backed Bond, MBB). Created by Financial Institutions, MBB's are investment bonds collateralized by underlying mortgages. These bonds are sold on to investors by the FI's and the funds are used to finance mortgages.

Definition 1.2.9 (Pass-thru Security). A security created by the pooling of mortgages by one or more mortgage holders. Shares, or participation certificates, are sold to investors who receive principal and interest payments commensurate with their investment.

1.3 Literature Review

A significant amount of research has been conducted on this topic. Because of the threat the Crisis posed on the U.S. financial system, the Federal Reserve (”the Fed”) has produced very thorough treatments of many of the problems that made the Crisis possible. In order to gain a deeper understanding of the Crisis, we have chosen to study a few of the papers from the Fed that were written in the immediate wake of the implosion [see references 1, 2, 3, and 5].

Researchers Nadauld and Sherlund find that both the securitization and credit ratings processes provided incentives for securitizing banks to purchase sub-prime mortgage loans in areas of high expected real estate appreciation [5]. In their paper, they document the relationship between the Securities and Exchange Commission's reduction in capital requirements with increased activity in the sub-prime market. They further document the relationship between realized rates of house price appreciation and the credit ratings of the bonds into which the underlying mortgages were pooled; they find that mortgage pools attached to geographic areas with high rates of appreciation were more likely to receive an AAA credit rating.

These findings are very important for the following two reasons. Primarily, they highlight the connection between house price speculation and the extension of mortgage credit to the sub-prime sector. This finding highlights the risky behavior of betting on the unpredictable nature of house price appreciation. Real property values are based on confidence which can change drastically as macroeconomic factors change. The speculative nature of house prices introduced an additional element of risk which proved very challenging to quantify; investors in mortgage-backed securities were making their decisions primarily based on the ratings of the bonds which was not a true reflection of the underlying risk of the assets [see Ashcraft and Schuermann, reference 2].

Secondly, the Nadauld and Sherlund findings identify a close connection between credit ratings on MBB's and the geographic regions for which they were underwritten. Their findings demonstrate that bond ratings were influenced greatly by the underlying property values. Such values can vary greatly between coastal and inland regions in the U.S. As coastal regions are more likely to appreciate in value than inland regions, MBB's constructed primarily of properties from coastal regions concentrate risk to zip codes deemed more likely to to have higher rates of property appreciation. In the event the expectations are inaccurate, this could spell investment disaster due to a lack of diversification. Additionally, Nadauld and Sherlund note that MBB's based on properties in faster appreciating zip codes received lower financing rates often times as much as 150 basis points. This opens the door to price discrimination based on speculative expectations.

Researchers Avery, Bostic, Caleb, and Canner outline, and provide supporting research, for the option model of mortgage loan performance. Within this theoretical framework, loan performance is positively correlated with original home equity, which means that the borrower put a percentage of the price of the house as a down payment. The intuition is that home-buyers who have a significant cash investment in the property are more likely to make mortgage payments even though their finances take a turn for the worse. They are more likely to have thought out the affordability of the home while saving the down payment.

Their research confirms the positive relationship between original equity and loan performance; borrowers with higher investments were more likely to protect their investments. In an analysis of nearly 425,000 mortgage loans spanning the years 1975-83, their data shows a 6.2% default rate on mortgages with a 91 - 95% Loan-to-Value ratio versus 0.24% on those with 10 - 70% ratios [Avery, Bostic, Caleb, and Canner, reference 1]. This represents a multiple of 30 times higher default rate for borrowers with highly levered loans.

The findings of Avery, Bostic, Caleb, and Canner imply that lenders should require having a cash down payment at loan origination. One of the defining characteristics of the Crisis was that mortgage credit was extended to applicants with $0 down. And although many FI's had maximum loan-to-value guidelines, these were thwarted because second mortgages were used as a means to fill in the remaining percentage needed in order to close the loan. Essentially, homebuyers, in these instances, were financing their down payments as a second loan against the property.

Thus, the nature of these loans were, from the start, speculative on the market value of the property. In the event that market confidence would begin to falter, all of a sudden the underlying properties would not be worth as much as before. Therefore, homeowners could easily find themselves in a negative equity situation from one day to the next because the Original Loan-to-Value started out at 100%. At the height of the lending boom borrowers with second mortgages were allowed to borrow up to 125% of the market value of their homes. Once the Crisis took hold, even borrowers who originally had 20% down were forced into negative equity range as average home values fell more than 30% in many areas (Naduald, Sherlund)

2 Mortgage Dynamics

2.1 Securitization and Innovation

Modern day securitization of mortgages began in the U.S. shortly after the second world war in an effort to bring liquidity to the residential mortgage industry. They were issued by the GSE's in the form of pass-through securities which were sold to investors. Because of the resources of the U.S. government, the GSE's' have played a critical role in building the financial infrastructure of the U.S. mortgage industry. During the years just prior to the Crisis, GSE's held nearly $4.5 trillion worth of outstanding mortgage pools (Saunders and Cornett, 810).

In the wake of the events of September 11, 2001, the U.S. economy faced the threat of a deeper economic contraction as it was already in a recession due to the dot-com crash of 2000. As a response, monetary policy loosened and credit, especially mortgage credit, was extended, and even encouraged, through the lowering of interest rates. Subsequently, the number of financial institutions extending mortgage credit began to multiply and this resulted in increased pressure on profit margins as the FI's competed to issue mortgages.Traditionally, mortgage issuing FI's would hold the loans on their books until maturity. But as rates began to fall near record lows, this was business model was no longer feasible and the FI's had to innovate in order to stay in business.

From the perspective of the FI's, securitization is a source of fee income and reduces the transaction costs incurred with holding mortgages on the accounting statements. Hence, income earned by originating mortgages and subsequently selling them became more important to competing FI's. Consequently, securitization accelerated as a market response to the economic survival of the FI's.

The primary source of the costs comes from the risk associated with the duration mismatch of the assets and liabilities. The assets are the issued mortgages that typically amortize over 30 years and the liabilities are the short-term demand deposits used to finance the mortgages. Such a gap increases the transaction costs to the FI's as its investors require a higher risk premium to compensate for the risk inherent in the mismatch. Additionally, FI's are required by the Federal Reserve to maintain sufficient capital reserves for any debt issued, called capital requirements, and to pay insurance premiums to the FDIC which insures short-term demand deposits up to $100,000. The securitization of mortgage debt helps avoid such requirements in the case of the CMO because the issued debt is packaged and quickly moved off the FI's balance sheet. In the case of the MBB the duration gap is mitigated through accounting procedures. With the use of these products the originating FI hedges its risk by sharing it with the intermediaries, investment banks, and the end investors.

2.2 Structured Investment Vehicles: MBB's and CMO's

The rapid product innovation enabled by securitization, which occurred without regulation keeping pace, played a central role in the development of the sub-prime mortgage industry. Although securitization has contributed greatly to the development of a well-greased mortgage industry, the product innovation during the Crisis made it difficult to identify, and hence quantify, the true underlying risk of the securities into which the mortgages were packaged.

The primary products that enabled the financing of the sub-prime market segment are Mortgage Backed Bonds (MBB's) and Collateralized Mortgage Obligations (CMO's). These products effectively wrapped pools of mortgage debt into the form of a bond which could then be sold on the investment market. These bonds offered investors higher than average yields on the existing bonds such as treasuries and corporates and provided the capital necessary to keep the finance wheel of the U.S. real estate market greased and turning.

The CMO works by creating different bond classes, or tranches, based on their perceived risk levels. Typically the originating FI packages the mortgage debt into a pool and moves these packages off their balance sheet, with the assistance of an intermediary firm, to an investment bank who then creates the bonds to be sold on the open market to investors. The investment bank separates the bonds into distinct classes based on their perceived risk levels. The bonds are given a rating and a yield based on their class and investors buy according to their risk tolerance and desired returns.

CMO's are used to mitigate prepayment risk. This is accomplished by applying early principal payments in a hierarchical manner; first the lower rated tranches are retired and then successive higher rated ones as the principal pool is retired. As the principal on the loans is paid off earlier than the original contract stipulates, investors do not receive the interest payments they would have received had the contract followed its original term. Consequently, investors with claims on payments from the lower rated tranches are compensated with a higher interest rate because of the increased prepayment risk.

MBB's differ slightly from pass-throughs and CMO's in that they remain on the balance sheet. The primary motivation of creating an MBB is to concentrate segments of mortgage assets on the FI's balance sheet, and to monitor the value of these assets as home values fluctuate in the market.

Issuing MBB's allows FI's to minimize funding costs by closing the duration gap. The MBB's are issued with a principal balance which is less than the market value of the underlying assets. This over-collateralization allows the issuing FI to reduce the amount of interest paid to its depositors as the bond is considered to be a liquid asset which can be sold on the market. Therefore, depositors don't require compensation for risk inherent in the FI's typical duration mis-match.

2.3 Money Flow and Credit Ratings

At the heart of every mortgage transaction there is a mortgagor (demander of funds) and a mortgagee (investor/supplier of funds). The historically low interest rates made available in the wake of September 11th, 2001, created a glut of deep-pocket investors looking to earn a premium as leaving money in the bank offered no chance of a real positive return (we have recently witnessed negative real rates of interest). In parallel, the possibility of owning a piece of the American Dream began to take root in the hearts and minds of the middle and lower-middle social classes and aggregate demand began to form in the market. In order for this market to realize its potential, risk-seeking investors needed to connect to encouraged home-seekers. This is where market makers emerged to fill the abyss. The following image depicts these market makers, highlighting the roles fulfilled by the specialized functions required in sub-prime mortgage financing.

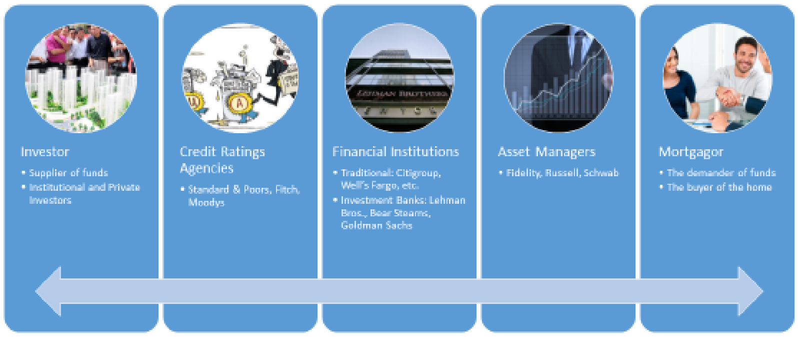
Investor
• Supplier of funds
• Institutional and Private Investors
Credit Ratings Agencies
• Standard & Poors, Fitch, Moodys
Financial Institutions
• Traditional: Citigroup, Well's Fargo, etc.
• Investment Banks: Lehman Bros., Bear Stearns, Goldman Sachs
Asset Managers
• Fidelity, Russell, Schwab
Mortgagor
• The demander of funds
• The buyer of the home

FIGURE 1: PRINCIPAL SUB-PRIME MARKET PARTICIPANTS

As can be seen from the image, on the extreme ends of the transaction are the investor and mortgagor. In order to bring the demand to the capital markets, the FI's have to originate the loans and issue the MBB's. To bring capital to the demanders, the asset managers need to choose which bond(s) to invest in (portfolio selection). Due to investment regulation, most asset firms are required to invest in bonds that have a minimum grade category, typically this is investment-grade. This feature of investment regulation underscores the critical role of the ratings agencies and the significance of their ratings on different bonds available in the capital markets.

The dominant credit agencies, Standard & Poor's and Fitch, use a schematic organization of AAA, AA, A, BBB, BB, B, CCC, CC, C, etc. to categorize bonds offered in capital markets. Moody's uses a slightly modified, but still descending order categorization scheme. The essential idea behind establishing a credit rating is that investors attempt to predict the probability that the bond issuing firm will default on its contractual obligations. The following table based on historical default frequencies, illustrates one way of assigning ratings to different corporate bonds.

Rating	1995	1996	1997	1998	1999	2000
Aaa	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Aa1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Aa2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Aa3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
A1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
A2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
A3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Baa1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Baa2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Baa3	0.00%	0.00%	0.00%	0.00%	0.34%	0.98%
Ba1	0.00%	0.00%	0.00%	0.00%	0.47%	0.91%
Ba2	0.00%	0.00%	0.00%	0.61%	0.00%	0.66%
Ba3	1.72%	0.00%	0.47%	1.09%	2.27%	1.51%
B1	4.35%	1.17%	0.00%	2.13%	3.08%	3.25%
B2	6.36%	0.00%	1.50%	7.57%	6.68%	3.89%
B3	4.10%	3.36%	7.41%	5.61%	9.90%	9.92%

TABLE 1: MOODYS HISTORIC CORPORATE BOND DEFAULT FREQUENCIES

As can be seen from the data, lower rated corporate bonds have a higher frequency of default. We also notice that default rates tended to increase and moved up the schema during the onset of the dot-com crisis. The intuitive idea captured in the data is that as negative macroeconomic events occur, such as interest rate or input shocks (for example, oil), the probability of default expands to include firms with higher credit ratings depending upon their exposure to systematic risk. Several conceptual differences on how to establish the default probabilities exist. These concepts typically range from the empirical (historical) to the more present and forward looking concept based on the financial fitness of a firm's balance sheet. This is an interesting topic which deserves its own analysis which we will not consider for the purpose of this paper.

One of the biggest challenges of the Crisis was that the securitization process pooled sub-prime and prime mortgages together in the same coupon-paying tranche which obfuscated the underlying credit risk. This technique is nontrivial and accomplishes a major financing objective: it raises the rating of the MBB of pooled mortgages which permits investment from a larger set of investment firms. Certainly, the true risk began to realize as the loans began to reset, due to the expiration of the ARM's, in an environment of rising interest rates. The Crisis began to set in as the prime mortgagors were able to refinance, or make the payment on the adjusting terms, while their sub-prime counterparts couldn't. This refinance risk deserves an analysis of its own, which we will not attempt here, but instead we will consider some of the natural frictions that arose as a result of the complex interplay between the principal actors of the sub-prime mortgage industry.

3 Data Analysis

3.1 Data Set

Our data set consists of 899,745 loans which were purchased by Fannie Mae in 2006 whose performance is tracked through December 2012. The loans were purchased from originating lenders and represent mortgages originated throughout all the states of the United States. The loans are all 30-year, fully amortization, full documentation, conventional fixed-rate mortgages. The data set does not include the adjustable-rate mortgages (ARM's), non-full documentation, and interest-only loans which we originally wanted to analyze.

Fannie Mae separates the data into two files, Acquisition and Performance. The Acquisition file contains the variables which encode the borrower and loan characteristics. Examples of these include Borrower Credit Score, Original Interest Rate, Property Type and Original Combined Loan-to-Value. The Performance File tracks loan performance characteristics. Examples of these variables include current loan delinquency status, month-to-month payment history, loan age, mortgage servicer, and maturity date of the loan. Each loan has a unique Loan Identifier which links the Acquisition file to the Performance records.

As can be seen from the data, the FICO scores' distribution is concentrated between the 600 and 800 range. According to Fannie Mae's glossary, the data has been reorganized to reflect current underwriting guidelines. If we had the other interesting loans, we'd see a much broader distribution of FICO scores in Figure 3.

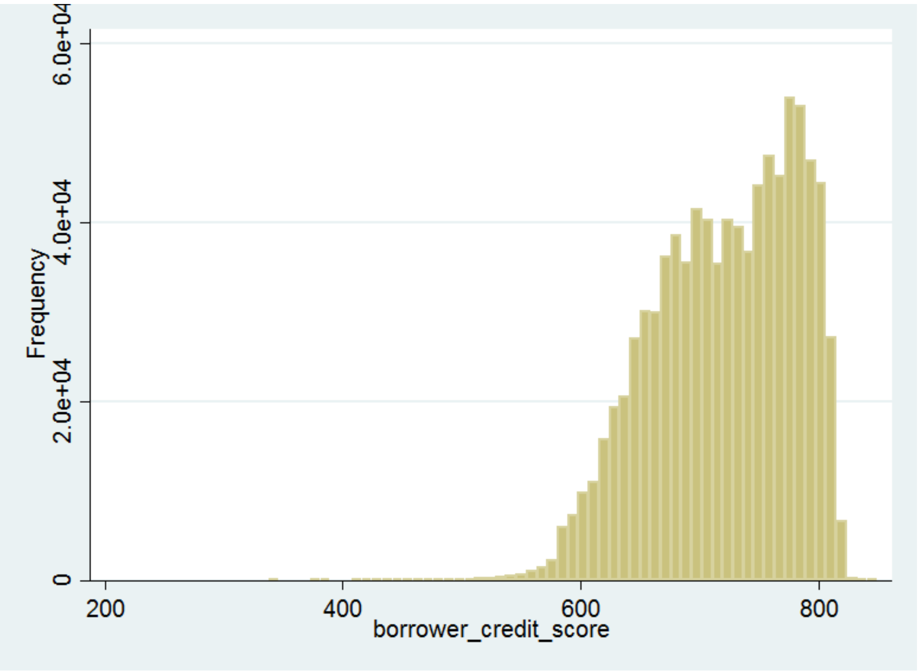

FIGURE 3: DISTRIBUTION OF FICO SCORES IN THE DATA SET

3.2 Variables

In our data preparation, we collapsed the two files, but only kept the Delinquency variable from the performance data set. This is our dependent variable which we coded as binary giving it a 1 if the loan is in default and a 0 if it is not. Of the various independent variables made available in the data set, we chose to focus on the Combined Loan-to-Value, Borrower Credit Score, and the Debt-to-income ratio of the borrower and the relationship of these with the dependent variable.

The foreclosure process can vary in time given differing processes depending on states' laws concerning property rights. Although the data does make record of loans that have are at least 30 days delinquent, we chose to consider only those with 180 days delinquency and greater. The dependent variable is given a 1 if the loan is 180 days or more past due, which means that the borrower hasn't made a payment on the property in at least six months. In either case, six months would reflect the greater of the time delinquent in which a lender would have rights to re-assume the property.

The credit scores range between 300 and 850 and are based on the FICO score, which is a model designed to capture the creditworthiness of individuals. The lower the score, the higher the credit risk. The score is calculated through the effects of several factors including payment history, length of credit experience, and the ratio of outstanding debt against available balances. Given a poor payment performance history and a high ratio of leverage, a particular applicant will result in having a lower score to reflect the higher credit risk.

The Original Combined Loan-to-Value includes the first and second mortgage loans originated against a property's market value. Having a Combined Loan-to-Value of 100% means that the mortgagee didn't put any money down at the time of loan origination.

3.3 Logistic Regression

Logistic Regression is a direct probability model which calculates the probability that a binary event will occur as a response to a single or even several predictor variables. In this model, the log odds of the outcome is modeled as a linear combination of the predictor variables. The independent variables may be continuous or categorical, but the dependent variable has to be binary.

The logistic function is defined as follows:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}, \qquad t \in \mathbb{R} \tag{1}$$

The logistic function is very useful in practice because its domain is the entire real line, but its range is only the positive real numbers so it can be interpreted as a probability. Moreover, t can be viewed as a linear combination of several independent variables in which case we will express it as follows:

$$t = \beta_0 + \beta_1 + \cdots + \beta_n, \qquad t \in \mathbb{R} \tag{2}$$

Where, we can now rewrite (1) in functional notation as:

$$F(t) = \frac{1}{1 + e^{\beta_0 + \beta_1 + \cdots + \beta_n}}, \qquad t \in \mathbb{R} \tag{3}$$

The regression coefficients are estimated using an iterative calculation process, such as Newton's Method. Through such a process, the outcome of each observation updates the value of the coefficients. This implies that with more observations, the value of the coefficients tend towards their limiting (true) value, as is expected from the Central Limit Theorem. The "goodness of fit" is measured with the R squared value; the higher the value, the better the fit between the logistic model and the actual data points.

3.4 Interpretation

For the statistical analysis we ran an logistic regression in the statistical software package Stata. We are analyzing the probability of having a delinquent event (default in our case due to the 180-day delinquency) as a function of the independent variables: Original Combined Loan-to-Value (CLTV), Borrower Credit Score, and Debt-to-Income ratio. In the analysis, the coefficients are relatively small due to the homogeneity of the data set. Each of the independent variables are statistically significant and have the expected sign. In each case the Z scores are well above the absolute value of two standard deviations which indicates statistical significance.

As we can see from Table 2, there is a positive relationship with Delinquency and CLTV and Delinquency and DTI. There is a negative relationship with Delinquency and Borrower Credit Score, which can be noticed from the negative sign in front of the coefficient.

The coefficient on the Borrower Credit Score indicates that a 100 point increase in FICO score will reduce the probability of default by 1%. For the DTI ratio, we notice that a 10% decrease in debt-to-income, reduces the probability of default by 2% and, likewise, a 10% decrease in Original CLTV reduces the probability of default by 2%. These statistics are significant especially when we consider that there are almost 900,000 observations and that the data set does not include the riskier loans that really characterized the Crisis.

TABLE 2: Logistic Regression

Delinquent	**Coeff.**	**Std. Error**	**Z**	**Pr** $> \lvert z \rvert$	**+ 95%**	**- 95%**
Original CLTV	0.0207475	0.0002611	79.48	0	0.0202358	0.0212591
Borrower Credit Score	-0.0119007	0.0000648	-183.75	0	-0.0120276	-0.0117737
DTI Ratio	0.0219776	0.000327	67.21	0	0.0213366	0.0226185
Constant	3.683039	0.0488344	75.42	0	3.587325	3.778752

Table 3 gives our Descriptive Statistics. As the data set has already been reorganized to reflect current lending standards, there is not much deviation. The average FICO score is 721, which is medium to high; sub-prime credit scores are often considered to be 620 and less. There is, however, a range in the Borrower Credit Scores with the lowest being 338 and the highest 850. We also see a mean DTI ratio of approximately 39%, which, we postulate, is significantly lower than what we'd see if we had access to the complete data set. The mean Original CLTV is 73% which means that the average mortgagee in our data set had approximately a 25% cash down payment.

TABLE 3: Descriptive Statistics

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Delinquent (Independent)	0.095816	0.294339	0	1
Original CLTV (Dependent)	73.10547	17.14645	1	155
Borrower FICO (Dependent)	721.8628	59.56459	338	850
Debt-to-Income (Dependent)	38.51873	12.20308	1	64

Source: Fannie Mae Single Family Loan Performance Data
Number of Observations: 899,745

Figure 2 is the Kernel Density of the probability of default as a function of the DTI ratio. Although the data set has been cleaned, we can still notice a positive relationship between DTI ratio and the probability of default. We also notice that the dark blue graph is narrower and taller than the light blue one; this indicates that the defaulted mortgages were the ones with a higher DTI ratio which is concentrated above the upper end of the mean DTI of 38%.

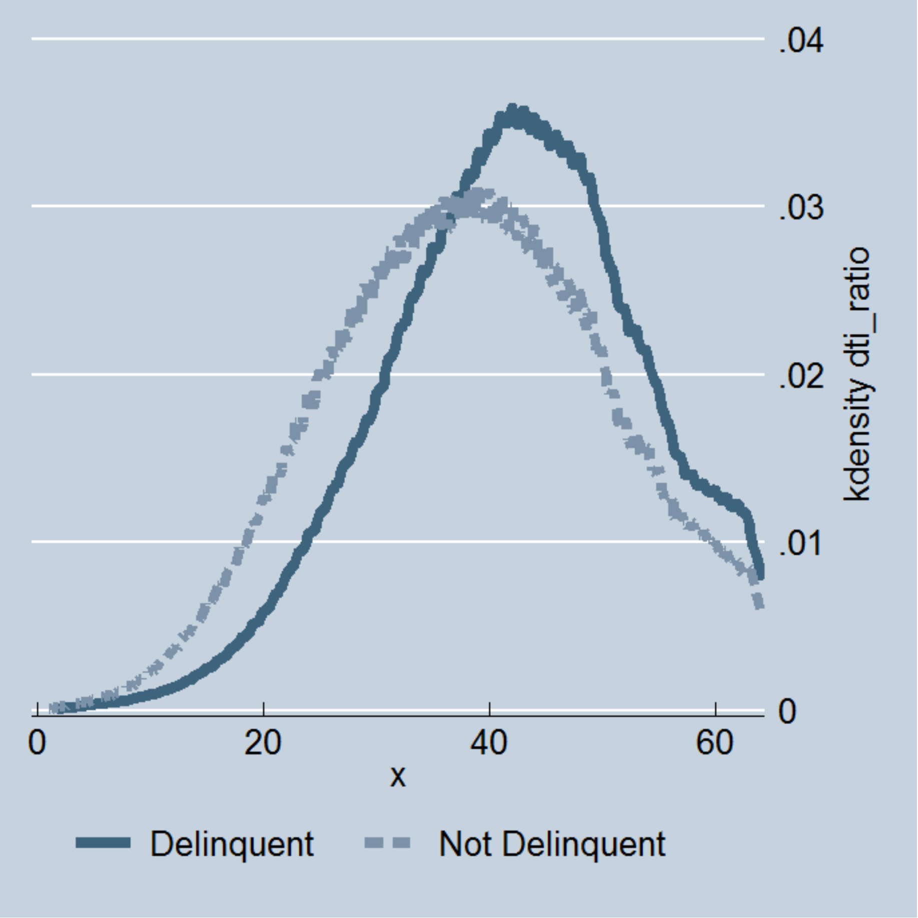

FIGURE 2: Kernel Density of Delinquent vs. Non Delinquent Mortgages as a function of DTI ratio
Number of Observations: 899,745

4 Principal Component Analysis

4.1 Background

Principal Component Analysis (PCA) is a statistical tool used in the analysis of data matrices. PCA takes on different names according to the branch of science to which it is being applied. The most common of these names are singular-value decomposition (SVD), eigenvalue decomposition, factor analysis, or discrete Harhunen-Loeve transform.

Mathematically, PCA is an orthogonal linear transformation of a data matrix into another matrix, which has potentially less dimensionality (i.e. fewer columns). The fundamental motivation of PCA is to reduce the dimensionality of a large data set by transforming the variables of a given set into a matrix of fewer variables, which are those that retain ***most*** of the variation of the original variables. The new set of variables are the Principal Components (PC's). As considered from this context, the PC's of a given data set are the variables which encode the dominant correlations amongst themselves.

This method is an application of linear algebra to statistical analysis in the sense that it models sets of observations as vectors in n-dimensional space such that the principal components are the eigenvalues of the symmetric covariance data matrix. It is also a variant of least squares fitting generalized to n-dimensional space.

PCA is an incredibly powerful technique used in least squares fitting. In the plane, the objective is to create a linear model that minimizes the Euclidean distance between a set of data points. This would be the best fitting line. In the statistical interpretation of this method, Euclidean distance becomes variance from the mean of a data set. The objective then is to identify those variables which have the largest or smallest variance from the mean value of the data.

4.2 Principal Components: Definition and Derivation

Given a vector $\mathbf{x}$ with p random variables, we wish to understand which variables contain the most variance within a given data set. We are also interested in identifying the correlation or covariance between the variables so as to determine which sets of variables co-move the most together. This method is primarily motivated when p is large otherwise we could consider the $\frac{1}{2}p(p-1)$ correlations. Let's consider our vector $\mathbf{x}$ with p random variables and let α be a vector with p random constants. Consider the linear function:

$$\alpha^{\mathbf{T}}{}_{\mathbf{1}}\mathbf{x} = \alpha_{\mathbf{11}}\mathbf{x_1} + \alpha_{\mathbf{12}}\mathbf{x_2} + \cdots + \alpha_{\mathbf{1p}}\mathbf{x_p} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \alpha_{\mathbf{1j}}\mathbf{x_j} \tag{4}$$

We now wish to find a linear function $\alpha_{\mathbf{2}}^{\mathbf{T}}\mathbf{x}$ of the elements of x with maximum variance, that is uncorrelated with (4). This process continues until at the kth stage an uncorrelated linear function $\alpha_{\mathbf{k}}^{\mathbf{T}}\mathbf{x}$ with the maximum variance subject to being uncorrelated with $\alpha_{\mathbf{1}}^{\mathbf{T}}\mathbf{x}, \alpha_{\mathbf{2}}^{\mathbf{T}}\mathbf{x}, \cdots + \alpha_{\mathbf{k-1}}^{\mathbf{T}}\mathbf{x}$ is found. The kth item $\alpha_{\mathbf{k}}^{\mathbf{T}}\mathbf{x}$ of these linear functions is then the kth Principal Component. This definition, from I.T. Jolliffe (see reference 4, page 4), will be our working definition of PC. **Note**: As we began with p random variables, there will be at most p PC's.

Now in order to derive the PC's we first start with the covariance matrix $\sum$ whose *ith*, *jth* element is the covariance between the *ith* and *jth* element of the vector of p random variables. The PC's are then linked to the covariance matrix as the kth vector of constants, α_k, is the eigenvector of the covariance matrix corresponding to the kth largest PC.

For the first PC we consider the vector $\alpha_{\mathbf{1}}^{\mathbf{T}}\mathbf{x}$ whose constant vector α_1 maximizes var$[\alpha_{\mathbf{1}}^{\mathbf{T}}\mathbf{x}] = \alpha_{\mathbf{1}}^{\mathbf{T}}\mathbf{x}\sum\alpha_{\mathbf{1}}$. This relationship is constrained to $\alpha_{\mathbf{1}}^{\mathbf{T}}\alpha_{\mathbf{1}} = 1$ in order that the sum of squares of elements α_1 equals 1. In order to carry out the maximization of the variance subject to our constraint we apply Lagrange's technique with the multiplier λ. Our goal is to maximize the equation:

$$\alpha_{\mathbf{1}}^{\mathbf{T}}\sum\alpha_{\mathbf{1}} - \lambda(\alpha^{\mathbf{T}}\alpha - 1) \tag{5}$$

Differentiating (5) with respect to α_1 gives

$$\sum\alpha_{\mathbf{1}} - \lambda\alpha_{\mathbf{1}} = \mathbf{0} \tag{6}$$

which is equivalent to

$$(\sum -\lambda \mathbf{I}_p)\alpha_{\mathbf{1}} = \mathbf{0} \tag{7}$$

Where $\mathbf{I}$ is the $p\,x\,p$ identity matrix, λ is an eigenvalue of the covariance matrix $\sum$, and $\alpha_{\mathbf{1}}$ is the corresponding eigenvector. The *pth* eigenvector which has maximum variance is then solved by maximizing the quantity

$$\alpha_{\mathbf{1}}^{\mathbf{T}} \sum \alpha_{\mathbf{1}} = \alpha_{\mathbf{1}}^{\mathbf{T}} \lambda \alpha_{\mathbf{1}} = \lambda \alpha_{\mathbf{1}}^{\mathbf{T}} \alpha_{\mathbf{1}} = \lambda \tag{8}$$

so as to yield the largest λ possible. In order to find the second PC, the process is repeated with the added caveat that the this PC has no correlation with the first PC. In other words, cov$[\alpha_{\mathbf{1}}^{\mathbf{T}}\mathbf{x}, \alpha_{\mathbf{2}}^{\mathbf{T}}\mathbf{x}] = \mathbf{0}$ where cov(x,y) denotes the covariance of the random variables x and y. This process is repeated until the $\alpha_{\mathbf{k-1}}$ PC is derived.

4.3 PCA Applied to Fannie Mae Data Set

The data in TABLE 3 is the result of our Principal Component Analysis in Stata. The eigenvalues are the covariances of the *ith* and *jth* random variables in the data vector. We can see from the data that the first six PC's are quite significant in terms of eigenvalues while the last three are relatively small. In the Cumulative column we can see that these first six PC's account for approximately 93% of the variance in the data set.

TABLE 3: Rotation: (unrotated = principal)

Component	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**
Component 1	2.66887	0.923336	0.2965	0.2964
Component 2	1.74554	0.477675	0.1939	0.4905
Component 3	1.26786	0.263545	0.1409	0.6314
Component 4	1.00432	0.0215175	0.1116	0.7430
Component 5	0.982801	0.279588	0.1092	0.8522
Component 6	0.703213	0.394982	0.0781	0.9303
Component 7	0.308232	0.00453517	0.0342	0.9645
Component 8	0.303697	0.288237	0.0337	0.9983
Component 9	0.0154602	0.0000	0.0017	1.0000

The Eigenvalues, which are the PC's, correspond respectively to our following independent variables: 1) Original Interest Rate, 2) Original Unpaid Balance, 3) Debt-to-income Ratio, 4) Loan-to-Value, 5) Combined Loan-to-Value (which includes a first and second mortgage if applicable), 6) Borrower Credit Score, 7) Original Term of Loan, 8) Co-Borrower Credit Score, and 9) Mortgage Insurance. From the table we notice that these first five variables account for approximately 85% of the variation in the data set. **Note**: Please see attached results from Stata analysis for more details.

The following graph is called a Scree Graph and according to Jolliffe (115) was first discussed and so named by Cattell (1966). This two-dimensional plot is used by statisticians to determine how many PC's to use in data analysis. This determination can be subjective, but Jolliffe references taking those PC's that determine the "elbow", which in our case is up through the 5th PC. According to our analysis in TABLE 3 this corresponds to our first 5 PC's accounting for approximately 85% of the variation in the data set.

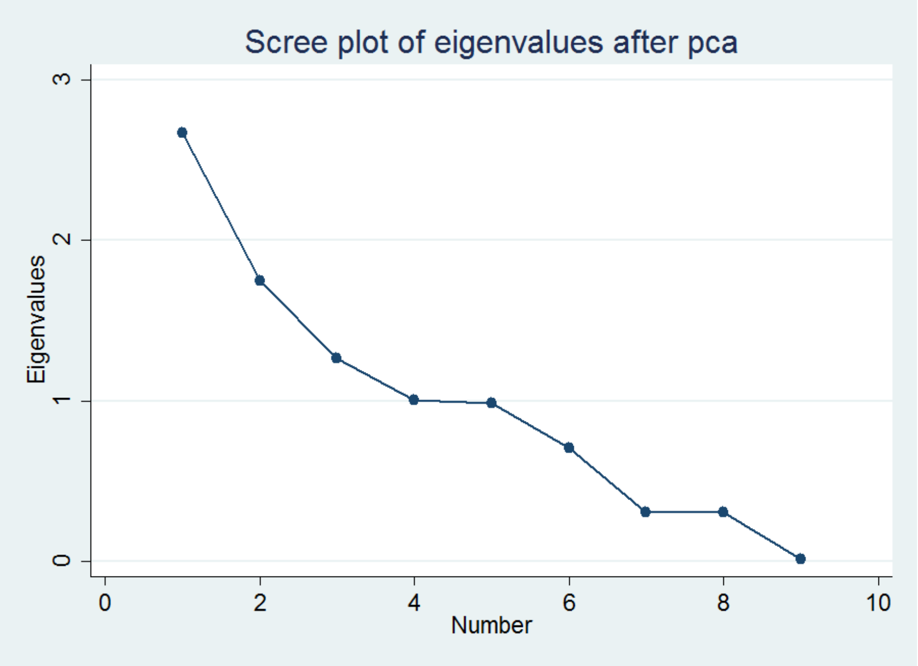

FIGURE 4: Scree Graph for the Correlation Matrix in Fannie Mae Data Set: PC's 1 - 9

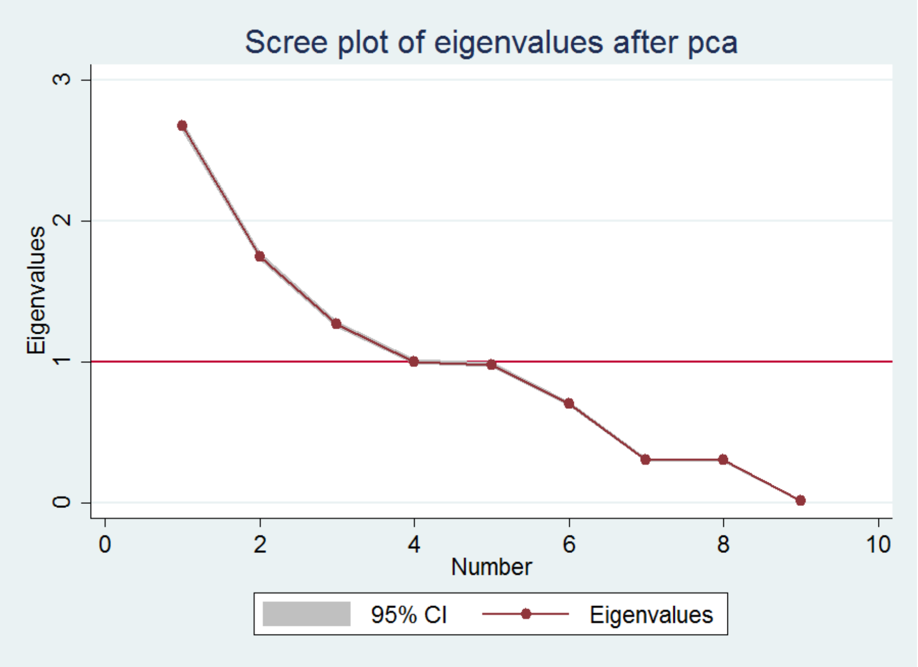

FIGURE 4: Scree Graph Highlighting Recommended PC's to Consider for Analysis

5 Conclusion

Throughout the end of 1999 and early 2000, the United States fell into a recession due to the bursting of the dot-com bubble and the subsequent loss of significant value in financial assets. The events of September 11th, 2001 further threatened an already weakened economy as general demand contracted due to fear. In an effort to end the recession, United States economic policy encouraged the expansion of credit, especially mortgage credit, and long-term interest rates fell to historic lows. Consequently the demand for loanable funds increased dramatically. Supply increased to meet the demand and the possibility of homeownership became a reality for many Americans.

As the middle and lower-middle social strata reacted to the prospect of homeownership, aggregate demand for mortgage loans expanded at a phenomenal rate. Investors seeking a premium began to supply the needed funds and the sub-prime mortgage industry began to gain traction. The financial machinery evolved in tandem and securitization developed as a natural response to market need. The securitization of mortgage debt into bonds became prevalent and the role of ratings agencies became crucial.

As the market evolved rapidly, so did the products. Mortgages with adjustable rates became increasingly popular. These ARM's featured rates initially set lower than a typical conventional loan, but were scheduled to reset at the maturity of their term which was typically three or five years. The expectation of continued low interest rates and high liquidity was near universal and, hence, refinance risk was in the best case mis-priced and in the worst, ignored.

In order to bridge the terrain between the sub-prime mortgage demander and the investor, financial institutions, asset managers, and credit ratings agencies, emerged to fulfill the specialized functions that were needed. Natural frictions arose as a result of moral hazard, due to model error, and pervasive information asymmetry. As the ARM's began to reset at term maturity, refinance risk began to realize as interest rates did begin to increase. Many sub-prime mortgagors were unable to make the higher payments and coupon payments on the MBS's began to default. The subsequent panic became known as the Great Recession of 2008.

The collapse of the sub-prime mortgage market sent shock waves around the world due to the global financial integration. Many of the investors in the CMO's and MBB's were from outside the U.S. and they saw sharp and dramatic declines in their investments in these products. A lack of understanding of the underlying mortgage debt meant that investors relied primarily on the inaccurate bond class ratings to guide their thinking about the risk level of the bonds. Although many investors had experience with some of the products of securitization such as MBB's and CMO's, the combination of these with other innovations such as no documentation, $0 down, and ARM's hid most of the underlying investment risk.

While securitization has provided the necessary capital to enable the American Dream, the correct pricing and recognition of the underlying risk is critical to ensure financial stability in the sub-prime mortgage industry. Obvious areas of interesting research lie in proper levels of governmental oversight and effective private sector compensation policy.

Our analysis has shown that **Logistic Regression** helps us determine relationships between the probability of default on a home mortgage with various independent buyer and/or loan characteristics. **Principal Component Analysis** helps us determine which of these variables contain the largest variance in the data set. This helps us determine those factors most relevant in home mortgage lending.

6 References

References

[1] Arnold, J., Johnson, L. and Riess, R. *Introduction to Linear Algebra, 5th Edition* New York: Pearson Education, Inc., 2002.

[2] Avery, R., Bostic, R., Calem, P., and Canner, G. *Credit Risk, Credit Scoring, and the Performance of Home Mortgages.* Washington D.C.: Federal Reserve, 1996.

[3] Ashcaft B. Adam and Schuermann Til. *Understanding the Securitization of Subprime Mortgage Credit.* New York: Federal Reserve Bank of New York Staff Reports, no. 318, 2008.

[4] Demyanyk, Y. and Van Hemert Otto. *Understanding the Subprime Mortgage Crisis.* St. Louis, MO: Federal Reserve Bank of St. Louis, 2008.

[5] Jolliffe, I.T. *Principal Component Analysis, Second Edition.* New York: Springer-Verlag, 1986.

[6] Nadauld, T. and Sherlund, S. *The Role of the Securitization Process in the Expansion of Subprime Credit.* Washington D.C.: Federal Reserve, 2009.

[7] Saunders, A. and Cornett, M. *Financial Institutions Management: A Risk Management Approach.* New York: McGraw Hill, 2006.

[8] Zandi, Mark. *Financial Shock: A 360 Look at the Subprime Mortgage Implosion and How to Avoid the Next Financial Crisis.* New York: Pearson Education, Inc., 2009.