Issues in Information Systems
Volume 17, Issue I, pp. 206-211, 2016

ANALYZING RISKS FOR HOME LOANS DURING FINANCIAL CRISIS OF 2008 USING DATA MINING

Ben Kim, Seattle University, bkim@seattleu.edu
Nina Tunas, Seattle University, tunasn@seattleu.edu
Adrian Green, Seattle University, greena3@seattleu.edu

ABSTRACT

Freddie Mac's single family loan data provide ample information on historical single family loan originations and performances throughout the loan periods. We analyzed these datasets to answer a critical question for the secondary mortgage market: Were there identifiable characteristics or factors surrounding borrowing and lending leading up to the financial crisis that could have signaled early risks in the resulting mortgage backed securities? For this analysis, the data randomly sampled by Freddie Mac, including the origination and servicing data from year 2002 and 2007, were used. Each year's dataset consists of about a million unique loans and their performance data. After an iterative process of data cleaning and developing several models, we applied the Decision Tree algorithm to identify the major loan risk factors. Initially we thought credit scores might be a major factor for predicting risks. However, we found that MSA (Metropolitan Statistical Area), loan purposes, and original UPB (Unpaid Principal Balance) are the three most significant factors.

Keywords: Data Mining, Decision Trees, Financial Data Analysis, Financial Crisis

INTRODUCTION

The financial crisis of 2008 is widely considered to be one of the worst financial crises since the Great Depression in 1929. The subsequent recession of 2008-2012 impacted financial institutions and economies worldwide, most notably resulting in the European debt crisis of 2009. In the U.S., the Dow Jones industrial average lost 33.8% of its value in 2008, with real GDP declining by 8.9% in the third quarter of 2008. An estimated 8.8 million U.S. jobs were lost, along with $19.2 trillion of household wealth lost (U.S. Department of the Treasury, 2012).

At the heart of the financial crisis was the collapse of the U.S. housing market just prior to 2008. In the previous years, mortgage lenders began offering subprime mortgages to borrowers who were not necessarily qualified for ordinary home loans. These lenders then sold these subprime mortgages to organizations such as Freddie Mac who would then bundle and sell them to investment banks worldwide. Monthly mortgage payments from these bundles provided a continuous revenue stream. As long as the U.S. housing market remained strong, these bundles appeared to be low risk investments. The true risk of these investments became evident when the housing market collapsed and borrowers began to default on their mortgages (Wallison, 2009).

Analysts today continue to contemplate on the crisis in an attempt to understand where things went wrong, and to guard against a repeat of such events in the future. Our analysis seeks to answer the following question: Were there identifiable characteristics or factors surrounding borrowings and lending practices leading up to the financial crisis that could have provided early risk signals in the resulting mortgage-backed securities?

DOMAIN PROBLEMS AND DATASETS

Our analysis is concerned with key measures of risks in home loans and the factors that contribute to, or at least, are associated with these risks. In order to perform an analysis, we are interested in borrower and lender data originating prior to the housing collapse as well as during the collapse itself. In this regard, we analyzed the Freddie Mac Single Family Loan-Level Dataset for years 2002 and 2007. These datasets are available for public at the website of the

Federal Home Loan Mortgage Corporation (FHLMC), also known as Freddie Mac (2016). The dataset covers approximately 17 million 30-year, fixed-rate mortgages originated between January 1, 1999, and September 30, 2013. This dataset is described as a 'living' dataset that may be corrected and updated over time.

Table 1. View Definition for the Final Data Table

```
Create View [dbo].[Comp_View]
As

  Select o.*, s.*
  from [dbo].[orig_2002_CS] o, [dbo].[2002_svcg] s
  where o.[Loan Seq Number O]=s.[Loan Seq Number S]

  Union

  Select o.*, s.*
  from [dbo].[orig_2007_CS] o, [dbo].[2007_svcg] s
  where o.[Loan Seq Number O]=s.[Loan Seq Number S]
```

LITERATURE REVIEW

While the aggressive lending and securitization practices of financial institutions such as Bear Stearns received the brunt of the attention and blame during the peak of the 2008 financial crisis, there are other institutions that may have had a role to play. In fact, Wallison (2009) suggests that government policies and financial regulations played a more significant role in the recent financial collapse than many believe.The Consumer Reinvestment Act (CRA) of 1977 received a major reform in 1993 with a goal of making home loans more accessible to low-and-moderate-income (LMI) borrowers. The more relaxed lending standards resulted in an increase of subprime mortgage loans from 7.2% in 2001 to 18.8% in 2006 (Wallison, 2009). However, Wallison argues that it was the "spreading of these looser standards to the prime loan market" that was the real cause of the housing bubble. Around the same time as the CRA reform, Government Sponsored Enterprises (GSEs) such as Freddie Mac and Fannie Mae were encouraged to modify their lending standards to begin accepting loans that they would have previously denied. The GSEs used this new mission to bypass Congressional regulations that had restricted the size of mortgage portfolios. The new demand for subprime loans created by the GSEs forced private lenders to pursue them as well in an effort to remain competitive. By 2006, approximately a half of all US home loans were non-prime (Wallison, 2009). In addition, Wallison highlights the state laws that allow penalty-free refinancing for homeowners when interest rates fall or home prices rise, enabling them to extract or cash out any equity that had accumulated; these equity loans are tax-deductible, making them even more attractive.

Thompson (2009) explained the political reasons behind the great financial crisis in 2008 involving two government sponsored enterprises, Fannie Mae and Freddie Mac. They borrowed huge amounts of money from Japan and China. Despite the size of their mortgage backed securities purchase, Fannie Mae and Freddie Mac did not meet their performance expectation as they are under-capitalized and overly leveraged. They were also involved in several fraud.

Taylor (2009) addressed three main questions regarding the financial crisis in 2008 using preliminary empirical analysis: a) What caused the financial crisis? b) What prolonged the crisis? c) Why did it worsen so dramatically one year after it began? According to Taylor, the crisis started in mid-2007 after the credit boom period. The first main cause was a loose-fitting monetary policy by the government. Another main cause was that the agencies like Fannie Mae and Freddie Mac were encouraged by the government to prolifically purchase more mortgage backed

securities, including those that are risky and considered bad securities. In terms of factors that prolonged the crisis, the article examined two diagnoses: liquidity problem and counterparty risk.

Ivashina and Scharfstein (2010) discussed the key mechanism of how financial crises affecting real economy through limiting supply of credit for corporation. Victoria and David (2010) used data from Reuters' DealScan database of large bank loans for their analysis. Hormozi and Giles (2004) used data mining for analyzing the banking data. They found data mining as a powerful tool to analyze ever-growing amount of financial data. The banking industry has shown particular interest in the applications of data mining, using clustering/segmentation and

Table 2. Columns used in the Analysis Model using Decision Tree

Credit Score, First Time Buyer, Maturity Date, MSA, MI Percentage, Number of Units, Occupancy Status, Original CLTV, Original DTI Ratio, Original UPB, Original LTV, Original Interest Rate, Channel, PPM, Product Type, Property State, Property Type, Postal Code, Loan Seq Number O, Loan Purpose, Original Loan Term, Number of Borrowers, Seller Name, Servicer Name, First Payment Year, First Payment Month, Loan Seq Number S, Current Actual UPB, Loan Delinquency Status, Loan Age, Months to Legal Maturity, Repurchase Flag, Modification Flag, Zero Balance Code, Zero Balance Date, Current Interest Rate, Current Deferred UPB,DDLPI, MI Recoveries, Net Sales Proceeds, Non MI Recoveries, Expenses, Servicing Year, Servicing Month, Unique ID

link analysis operations in its marketing efforts, but maybe more importantly, using predictive modeling in developing risk profiles.

ETL (EXTRACTION, TRANSFORM, AND LOAD)

The datasets were obtained from the Freddie Mac (2016) website. Loan datasets for year 2002 and 2007 were selected for our analysis. Origination data and servicing data are included in each year's dataset. We used the data randomly sampled by the Freddie Mac – *sample_2002.zip* and *sample_2007.zip*. Each file has two datasets – origination and servicing data. The downloaded datasets were in the text file format. These files were imported to Microsoft SQL Server. From the original datasets, the origination table has a primary key on the *Loan Sequence Number* column and the servicing table has a primary key defined on the two columns - *Loan Sequence Number* and *Monthly Servicing Report*. These two tables are joined using the *Loan Sequence Number* column.

Separating attribute values with the date data type into month and year is a common practice for data mining. This is done to predict the patterns depending on years and months. For this reason, the columns such as *First Payment Date, Maturity Date, Monthly Reporting Period, Due Date Last Paid Installment,* and *Zero Balance Date,* were separated into months and years.

We used the *Loan Delinquency Status* attribute from the servicing table as the predicted attribute to indicate the delinquency in loan payment. The delinquency value of 0, 1, 2, 3, and so on means "Current or less than 30 days past due," "60-89 days delinquent," "90 – 119 days," and so on. We analyzed the delinquent values from 1 (less than 30 days past due) to 12 (less than 360 days past due). Since the size of the data is so huge, we reduced the data size by random sampling and created a *view* by combining the data using *JOIN* and *UNION* operations. After cleaning and transformation of data, the *view* we used for data mining has 497,667 rows. The *view* definition created these operations is shown in Table 1.

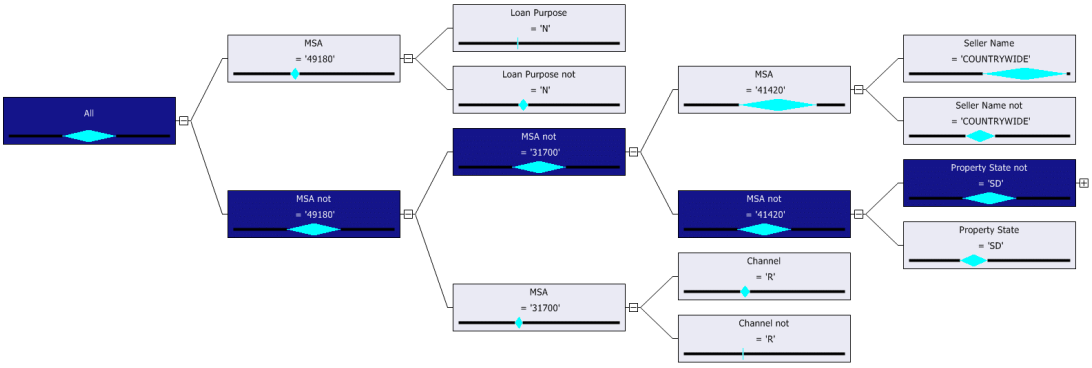

Figure 1. Decision Tree

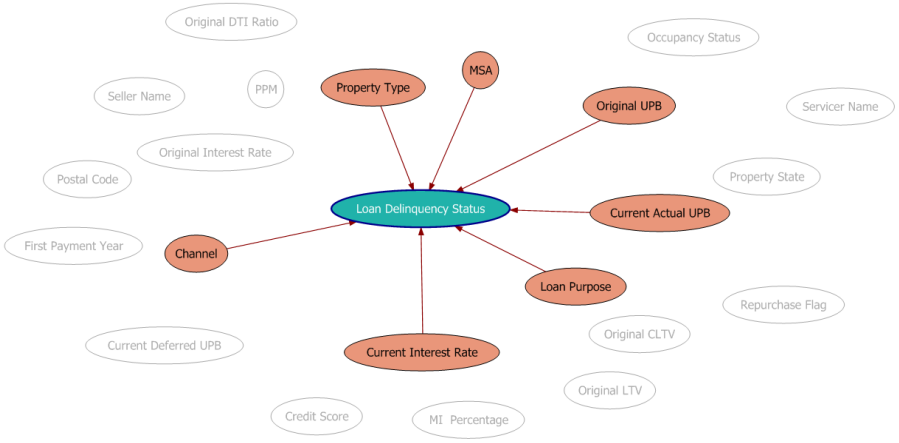

Figure 2. Dependency Network

DATA MINING ANALYSIS AND DISCUSSION

We identified loan payment delinquency as the primary measurement of a risk in home loans. In our Freddie Mac dataset (2016), the column *Loan Delinquency Status* has those measurement values. That column was selected as the predicted column in our mining structure. The factors included in our analysis are shown in Table 2. As a data mining software package, we used *Visual Studio 2013* and the *Analysis Service* of *Microsoft SQL Server 2014*. *Visual Studio* provides a graphical user interface for the *DMX* language used in the *Analysis Service* of *SQL Server*. Once *SQL Server Data Tools (SSDT*) is downloaded and installed, *Visual Studio* can directly access the data mining database on *SQL Server* and create data mining models. We found SQL Server and Visual Studio are relatively easy to learn and use. Also, since the database systems and data mining tools are tightly integrated in the same package, it saves much efforts to clean and manipulate the data iteratively as we refine data mining models.

Using the Decision Tree algorithm, we built a decision tree (Figure 1) and dependency network (Figure 2). The Decision Tree algorithm builds a decision tree by identifying factors providing the most *information gain* at each level of the tree hierarchy. Information gain was calculated using the level of *entropy* for each tree node. Figure 1 shows a decision tree built from the dataset we used. As can be seen there, the factor providing the most

information gain is "MSA." MSA stands for Metropolitan Statistical Area. Metropolitan statistical areas are geographic entities delineated by the Office of Management and Budget (OMB) for use by Federal statistical agencies in collecting, tabulating, and publishing Federal statistics (United States Census, 2016). "Loan Purpose" is the second most significant factor to predict the loan delinquency. It indicates whether the mortgage is a cash-out refinance mortgage, no cash-out refinance mortgage, or a purchase mortgage. (Freddie Mac, 2016). The third one is "Original UPB." UPB stands for Unpaid Principal Balance. The fourth, fifth, sixth, and seventh ones are Channel, Property Type, Loan Purpose, and Current Interest Rate as can be seen in Figure 2. Detailed definitions of these terms can be found at Freddie Mac website (2016). The interesting finding we noted was that credit score was only the ninth significant factor per our decision tree analysis. We believe that this finding appears to be consistent with explanations about the government policies and political reasons for sub-prime loans as major causes for the crisis as discussed earlier in the Literature Review section.

CONCLUSION

When we started analyzing the data, we thought that credit scores of a loan applicant would be the most important factors for delinquency. According to our decision tree analysis, however, it turns out that there are more significant factors than credit scores for determining whether a loan can be delinquent or not. As suggested in the Hormozi and Giles (2004), data mining tools and analysis such as this one could prove invaluable in better gauging the true risks involved in investment activities. In the future, we are interested in identifying the major factors determining the delinquency of loans in recent years and comparing them with the 2007 data.

REFERENCES

Freddie Mac (2016), http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html

Hormozi, A. M., & Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. *Information systems management*, *21*(2), 62-71.

Ivashina, V., & Scharfstein, D. (2010). Bank lending during the financial crisis of 2008. *Journal of Financial economics*, *97*(3), 319-338.

Taylor, J. B. (2009). *The financial crisis and the policy responses: An empirical analysis of what went wrong* (No. w14631). National Bureau of Economic Research.

Thompson, H. (2009). The political origins of the financial crisis: The domestic and international politics of Fannie Mae and Freddie Mac. *The Political Quarterly*, *80*(1), 17-24.

United States Census (2016) http://www.census.gov/population/metro/

U.S. Department of the Treasury. (2012). *The Financial Crisis Response: In Charts.* Retrieved from http://www.treasury.gov/resource-center/data-chart-center/Documents/20120413_FinancialCrisisResponse.pdf

Wallison, P. J. (2009). The true origins of this financial crisis. *American Spectator*, *42*(1), 22-7.