Calculating Mortgage Loan Risk
Team 43 - Progress Report

Harris Ashraf[1], Sharmila Baskaran[2], Bo Chen[3], Travis Jefferies[4], Daniel Mower[5], Cody Nguyen[6], and Ryan Wong[7]

[1]Georgia Institute of Technology, hashraf3@gatech.edu
[2]Georgia Institute of Technology, sbaskaran30@gatech.edu
[3]Georgia Institute of Technology, bchen354@gatech.edu
[4]Georgia Institute of Technology, tjefferies3@gatech.edu
[5]Georgia Institute of Technology, dmower3@gatech.edu
[6]Georgia Institute of Technology, cnguyen311@gatech.edu
[7]Georgia Institute of Technology, rwong33@gatech.edu

1 Introduction

Lenders such as banks provide borrowers with a mortgage loan that is paid back monthly for a fixed time-frame. Lenders make profit by imposing interest; however, the lender assumes risk since a borrower can default, refinance, or prepay their loan, thus decreasing profitability for the lender. Lenders want to understand the influential factors associated within mortgage loans to reduce their loss from potentially-risky loans, while borrowers want to minimize their rate.

2 Problem Definition

We calculated the appropriate mortgage rate to charge a potential borrower. This rate is determined by creating a default risk measure based on borrower characteristics and macroeconomic conditions while accounting for new customers. Based on our default risk measure, we estimated an interest rate range that represents high, average, and low profitability for lenders. Ultimately, our tool aims to allow users to easily explore historical mortgage default data, and borrowers and lenders to quickly obtain a range of acceptable interest rate quotes before engaging in the rate setting process.

3 Literature Survey

In [2][4][5][9][12], logistic regression and regression trees were used to make non-parametric predictions of default probability. In [2][8][18], K-Nearest Neighbors resulted in poor accuracy. In [2][8][14][17][21], Random Forest was found to have accuracy in the mid-90%. In [1][5], deep learning models were used to test the importance of feature extraction on accuracy, which is further stressed in [10]. In [6][12][20], Survival Analysis techniques were used to predict the time a mortgage will default, which is useful because our data contains loans that haven't defaulted within the recorded time frame. We will evaluate whether all these models will satisfy the constraints of our problem. We can evaluate our model using statistical techniques mentioned in [11][15] such as R-squared value.

In [3][7], geovisualization tools were developed to analyze navigability effectiveness of interaction features such as sliders and drop-down menus. In [3], the longitude, latitude, and duration of the mouse cursor over the tool was recorded to determine ease of navigability. [16] analyzes various formats of choropleth maps to demonstration how they provide different information. This alludes to spatial treemaps, which [19] found to be effective in displaying geographically-defined information. [13] explains how to use Human-Computer Interaction (HCI) concepts to overcome age and color-blindness impediments. We will consider all of these suggestions when designing our front-end interface and evaluating the usefulness of our interface.

4 Our Method

4.1 Tool Goal

Our aim is to help users understand mortgage risk. Our first tool shown in Figure 1 provides a convenient interface to display historical mortgage data that originated between 1999 and 2017. A choropleth segmented by states is dynamically adjusted based on four user-selected parameters: 1) "Vintage" or origination year, 2) "Year" or observational year, 3) "FICO" or range of credit scores, and 4) "LTV" (Loan-to-Value) or mortgage loan amount relative to value of the property. This helps users analyze past market conditions.

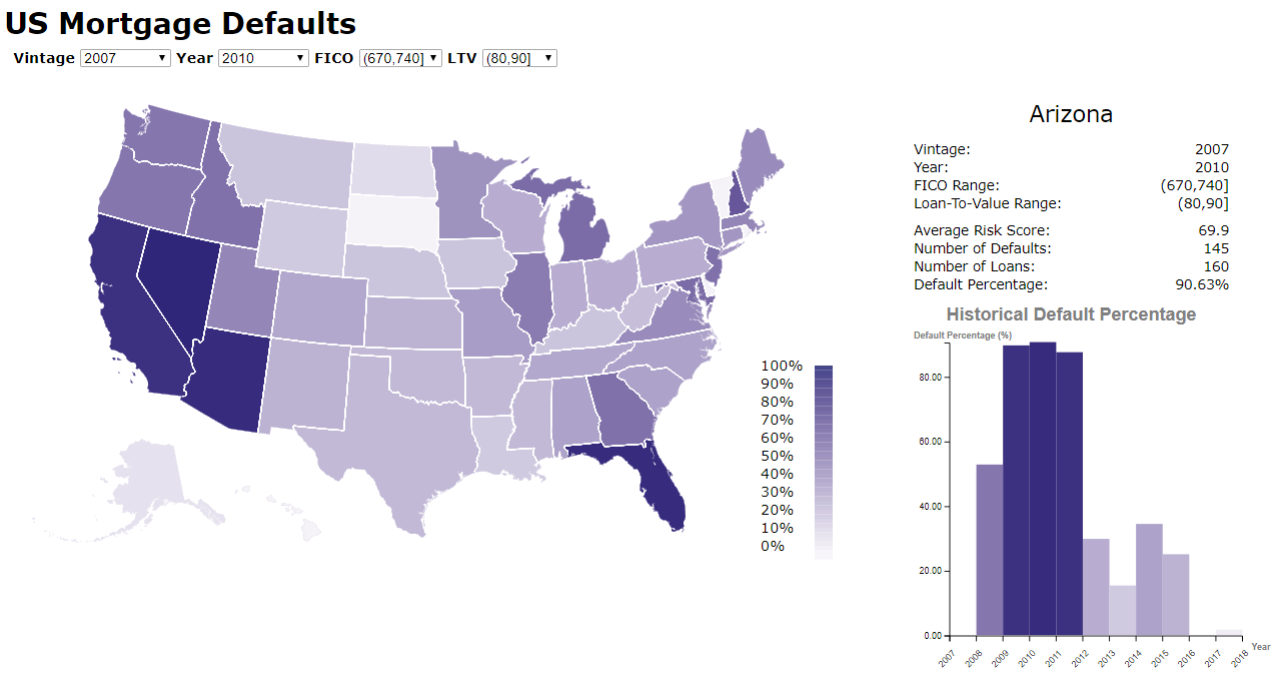

Figure 1. Interface for the choropleth tool.

The second tool provides an easy-to-use mortgage rate quote tool. On the front end, the tool allows the user to input relevant loan characteristics. On the back-end the tool also accounts for current macroeconomic market conditions. The tool is shown in Figure 2, provides rate quote estimates that users can use to determine their best interest rate.

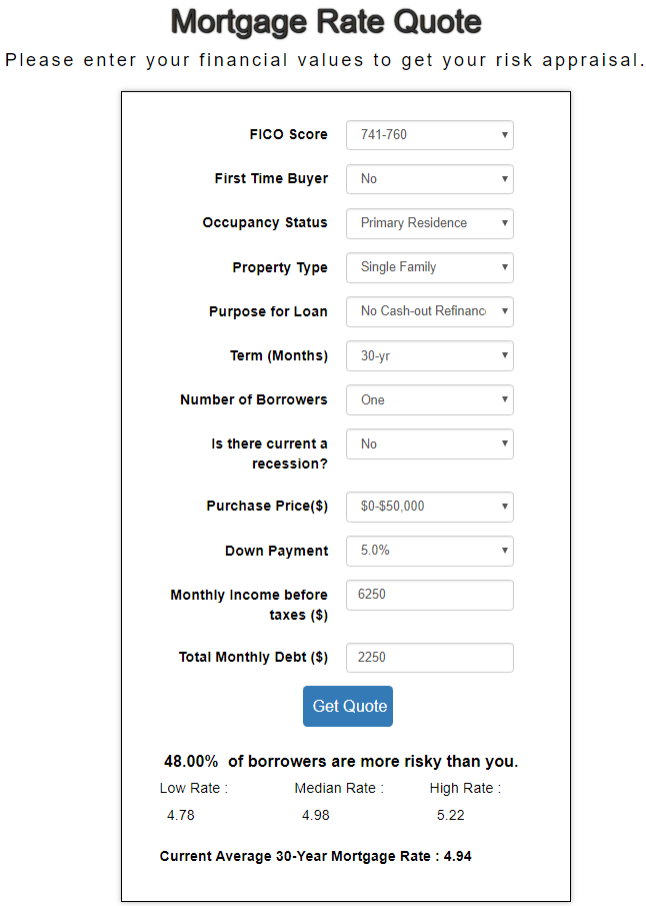

Figure 2. Interface for the risk score calculator tool.

4.2 Data Source and Tools

Freddie Mac (FM) operates in the secondary mortgage market by purchasing mortgages from lenders. FM provides origination and performance data to the public for the purchased mortgages. **The 75 GB dataset used in our project contains information for more than 25 million US mortgages originating between 1999 - 2017**. In addition, we use macroeconomic data obtained from the Federal Reserve Bank of St. Louis Economic Data (FRED).

To help us process the data, we use the H2O.ai machine learning and scikit-learn Python libraries to efficiently utilize various machine learning models, including Distributed Random Forest and Linear Regression. Our interface was built upon HTML, PHP, CSS, and JavaScript, using libraries such as Data-Driven-Documents (D3), boostrap, Jquery, and Selenium.

4.3 Model Algorithm

4.3.1 Data Cleaning and Sampling

All data is first cleaned for missing values, converting mixed data columns to the same data type, and retaining important information. The data was pre-processed as in Figure 3 in order to present the data easily in the choropleth.

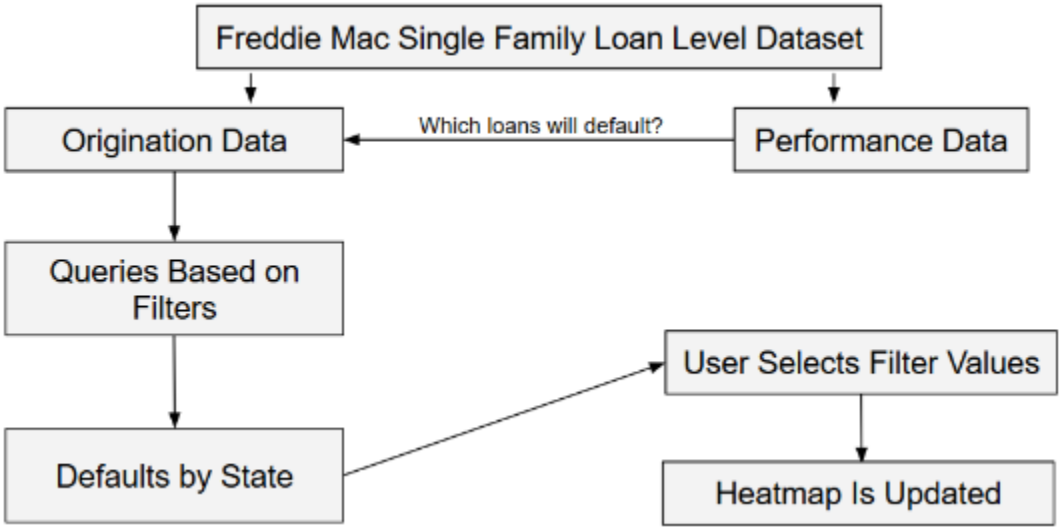

Figure 3. Data flow for the choropleth tool data processing.

Figure 4 shows the steps associated with data sampling and model handling for the risk calculator, which will be explained in more detail throughout later sections. The performance data is used to label the origination data as eventually defaulting if payments aren't made within 120 days.

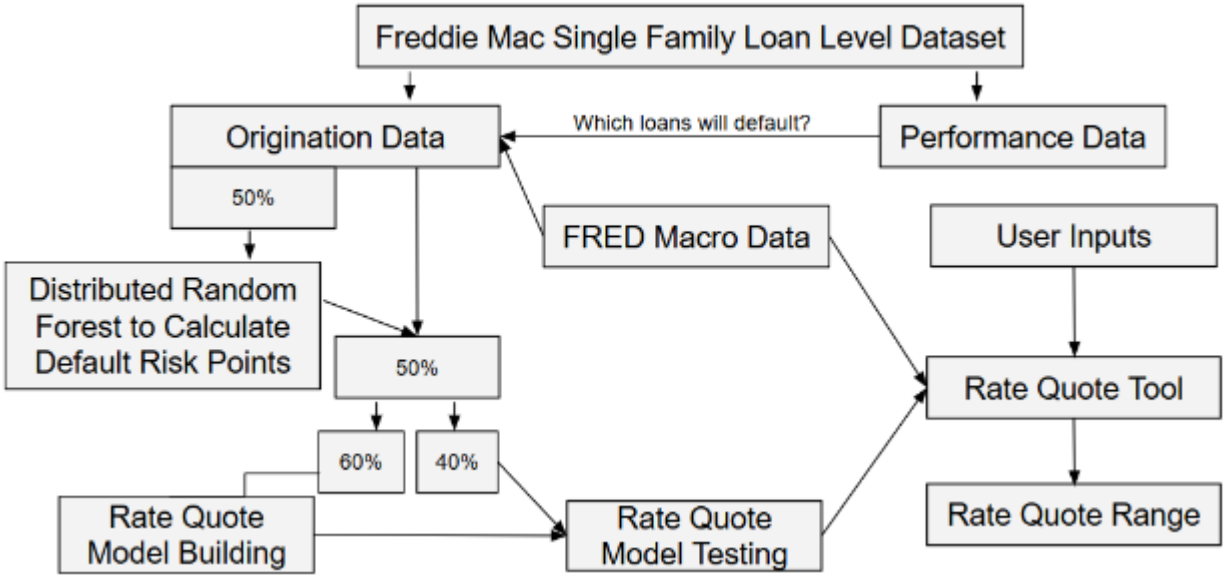

Figure 4. Data flow for the risk score calculator tool data processing.

4.3.2 Risk Score Calculation

An inherent goal of our tool is to understand which features of the dataset highly influence mortgage's default probability. We employ a Distributed Random Forest (DRF) model dimensionality reduction from the relevant origination attributes and macroeconomic variables used in the Federal Reserve to stress test financial institutions. Determination for feature splitting and thresholds are randomized as defined by H2O's library. We set the DRF parameters as trees of 150, maximum depth of 35, and minimum data points per leaf of 100; these were determined through grid search testing and experience with the data.

We used two key features of the DRF framework: 1) Variable Importance (VI) which assigns an importance weight to each feature, and 2) Partial Dependence Plots (PDP) which determines the relationship between each feature and their mean default

rate. VI and PDPs were used to extract information from the data. Figure 5 shows each of the features and their VI, where not surprisingly FICO score is the highest weighted feature at about 30%.

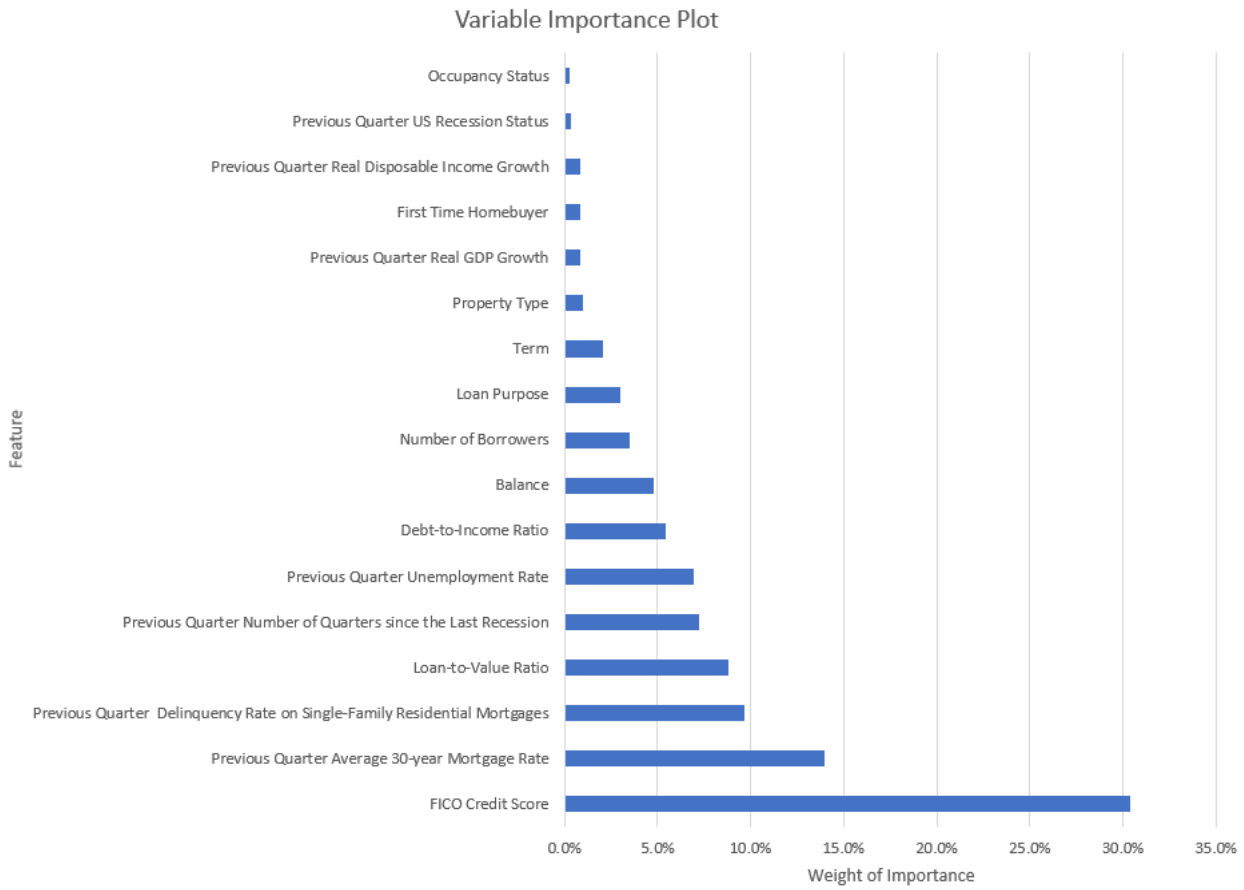

Figure 5. Variable Importance (VI) plot for the ten highest weighted features.

Figure 6 shows the PDP for OrigLTV (Original Loan-to-Value), where the red line corresponds to the mean response (i.e. default probability). This PDP analysis was done on all the features.

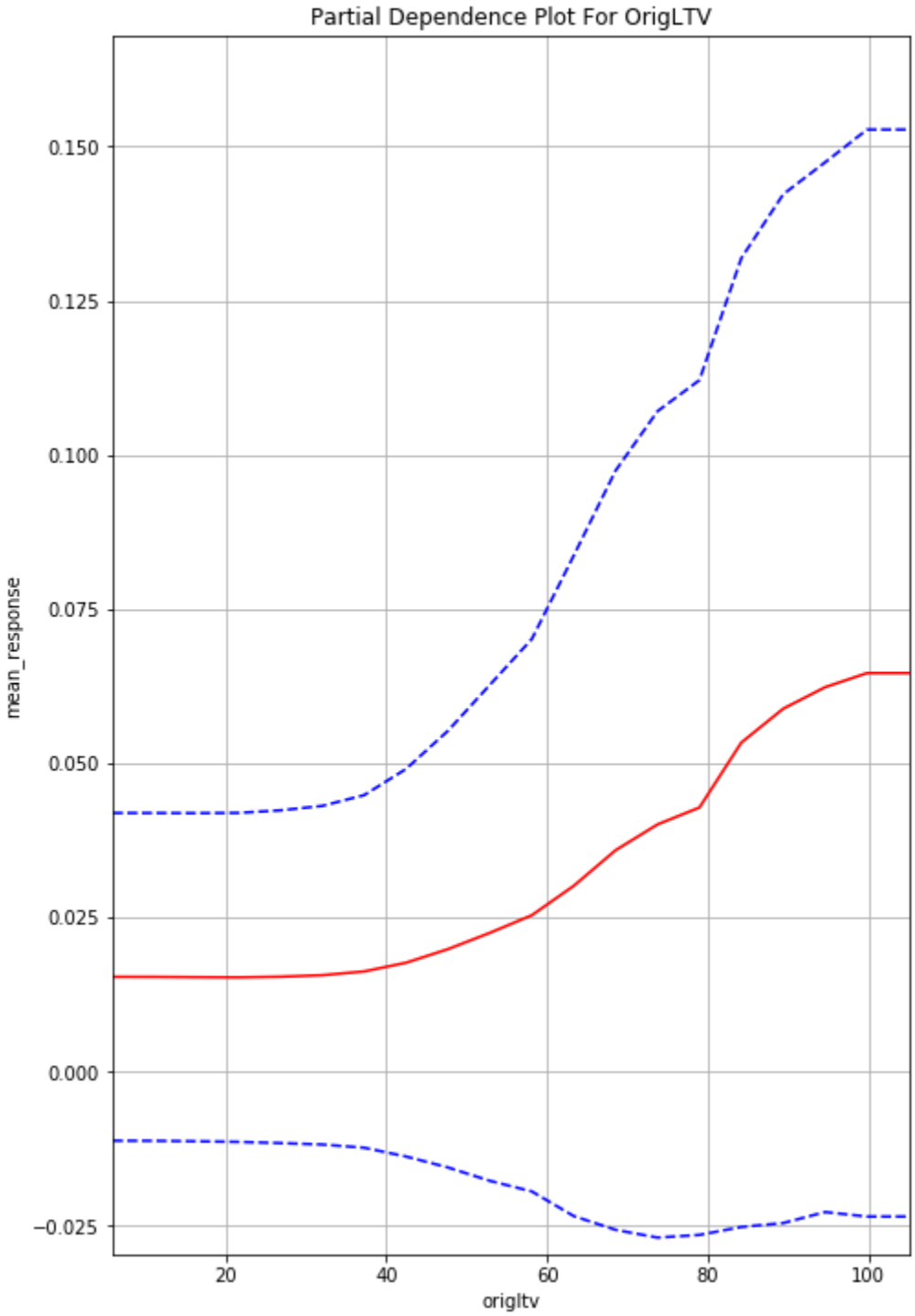

Figure 6. Partial Depedence Plot (PDP) for OrigLTV.

To illustrate the process of deriving the risk points (RP) measure, consider the example with the loan-to-value feature which has a VI weight of 8.8%. Suppose the loan in question has an LTV of 100. Note that in the PDP plot above, over 6% of loans in the 100 LTV range default. The formula for assigning RP based on on any feature is shown below.

Risk Points = VI * PDP / Max(PDP) * 100

For this loan we would assign 0.088 * 0.06 / 0.06 * 100 = 8.8 RP.

This process is repeated for each feature and the RP are summed together. Thus the RP measure uses information from every feature based on it's weight of importance. The RP measure is bound between 0 and 100, where 0 corresponds to the least risk and 100 to the most risk.

4.3.3 Rate Quote Model

Pricing a mortgage rate is dependent on several factors including a borrower's default risk and the average rate all lenders offers. This is an imprecise science as rate quotes may vary from lender to lender. Each party has its own model and unique portfolio of other assets which affects pricing decisions. To make the tool useful to lenders, we create a rate quote model based on data segmented according to RP ranges shown in Figure 7 below.

Risk Point Range	Corresponding Percentile Range Among the Dataset	% of Loans More Risky thatn the given Risk Point Range
(0,48.63595]	(0.0% - 6.25%]	93.75
(48.63595,50.5222]	(6.25% - 12.5%]	87.50
(50.5222,51.97334]	(12.5% - 18.75%]	81.25
(51.97334,53.18575]	(18.75% - 25.0%]	75.00
(53.18575,54.31676]	(25.0% - 31.25%]	68.75
(54.31676,55.47339]	(31.25% - 37.5%]	62.50
(55.47339,56.71045]	(37.5% - 43.75%]	56.25
(56.71045,58.04532]	(43.75% - 50.0%]	50.00
(58.04532,59.41849]	(50.0% - 56.25%]	43.75
(59.41849,60.8515]	(56.25% - 62.5%]	37.50
(60.8515,62.40068]	(62.5% - 68.75%]	31.25
(62.40068,64.19618]	(68.75% - 75.0%]	25.00
(64.19618,66.34118]	(75.0% - 81.25%]	18.75
(66.34118,69.37149]	(81.25% - 87.5%]	12.50
(69.37149,74.51278]	(87.5% - 93.75%]	6.25
(74.51278,100]	(93.75% - 100.0%]	0.00

Figure 7. Table of the 16 risk score buckets and their corresponding percentage relationship to the rest of the data.

By grouping the data as shown above we are able to measure the 25th, 50th, and 75th percentiles of the spread between the rates offered and the average rate offered by all lenders for each RP range.

We center the bucketed data at the 50th percentile RP range to ensure average-risked borrowers receive a rate quote close to the current average rate. We force the distance between the 25th and 75th percentile lines at the 50th percentile RP range to be equal to the historical average. This ensures that out model is in line with current market forces. We then estimate the linear relationship between the RP ranges based on all combinations of the following scenarios.

1. Product Type (30yr mortgages if the Term is greater than 180 months, and 15yr for Term below 180 months)
2. Increasing Rate Environments if the last federal funds rate change was positive and Decreasing if the last change was negative
3. High or Low Rate environment (current average 30-yr mortgage rates > 5%, or <= 5 for High and Low respectively)

Figure 8 below shows the model for the spread between current average rates for a scenario when pricing a 30 year mortgage in an increasing yet low rate environment. Notice that for the (56.9,58.3] RP range, the median spread is calibrated to be zero. The slopes of each of the three lines are then driven by the data. The positive slope indicates that the RP measure rank-orders default risk and that in general the market participants price loans based primarily on default risk.

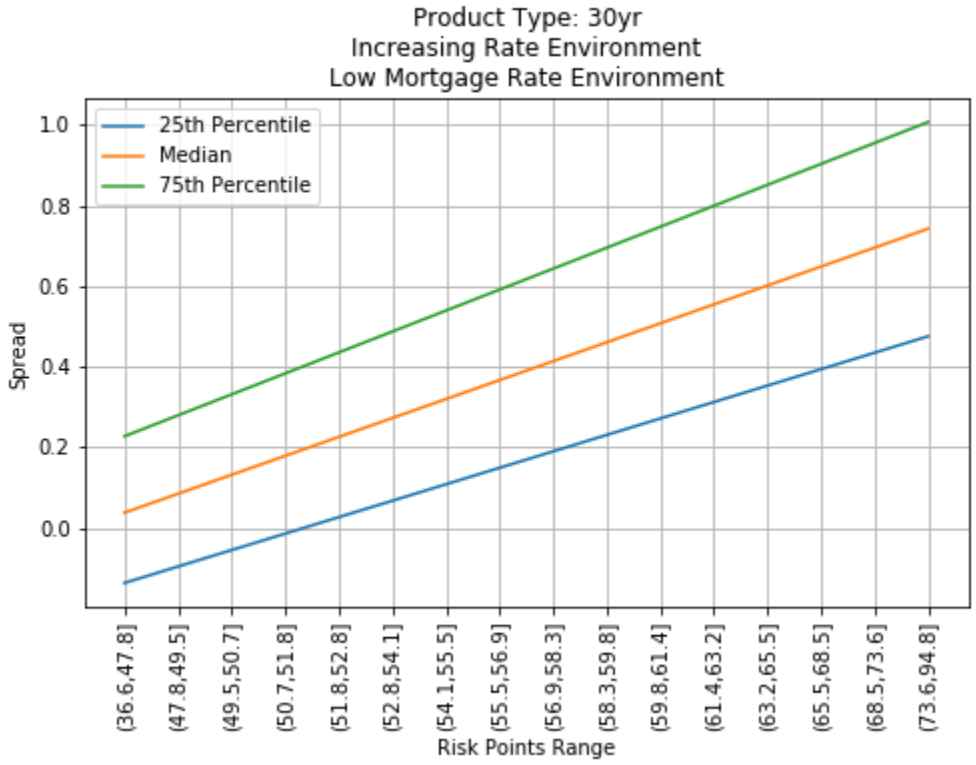

Figure 8. Interest rate thresholds across the risk score buckets for a 30-year fixed rate term, increasing rate environment, and low mortgage rate environment.

4.3.4 Tool Data Handling

The pre-processed risk point buckets and associated interest rate quotes were provided to our tool. When the user inputs a specific loan characteristic, the tool will query the pre-processed data to calculate and display the associated risk point, find an appropriate risk point bucket, and display the interest rate thresholds.

For the choropleth, we segment the data based on the two most dominant features used by banks. FICO score is segmented on poor (300-670], average (670,740] , good (740,800], and excellent credit (800,850]. LTV is segmented on low (0,60], medium,-low (60,80], medium-high (80,90], and high exposure (90,105] to the bank. Additionally for informational purposes, we segment the data based on state, origination/vintage year, and observational year. For each combination of the aforementioned filters we run queries on the entire dataset. This allows us to provide the user with access to the output of 5,575 queries across 50 states and D.C., providing a vast range of analysis capabilities against historical mortgage data.

4.4 Distribution of Teamwork

Team member contribution was equally distributed; however, each member was assigned responsibilities according to sub-teams in Table 1. Data Scraping researched, extracted, and consolidated the data for modeling. Data Modeling designed and tested machine learning models to predict loan risk. Data Visualization developed the front-end interface. Project Management facilitated project progress to adhere to deliverables.

Table 1. Distribution of work through sub-teams

Team member	Data Scraping	Data Modeling	Data Visualization	Project Management
Harris Ashraf			x	
Sharmila Baskaran	x	x		
Bo Chen			x	
Travis Jefferies	x	x		
Daniel Mower	x	x		
Cody Nguyen			x	
Ryan Wong			x	x

5 Experimentation / Evaluation

5.1 Model Accuracy

In order to validate the effectiveness of our modeling technique, we looked at current methods of analyzing mortgage loan defaults. Originally, we wanted to understand the probability of default at the time of origination; however, historical mortgage datasets are heavily skewed with more data points having not defaulted. This ruled out the usability of classification techniques such as K-Nearest Neighbors, Random Forest classification, etc. that were used previously in our literature survey. We then considered using Survival Analysis techniques, which could account for skewed datasets, but we realized that we were dealing with a limited dataset as most of the mortgage are right censored.

We found that it is most useful for users to be able to identify significant predictors that influence default risk. This allows us to gain a more in-depth understanding of risk factors to consider at origination. We then can calculate an RP score associated with loan's characteristics macroeconomic market conditions. This score can then be used to calculate the aforementioned interest rate thresholds that a loan officer or a borrower can use to make informed decisions when a loan is being requested.

We found success using this risk score calculation method. To evaluate the accuracy of our model in predicting interest rate quotes, we sample our data to create a test dataset as shown in Figure 4, comprising of 15% of the Freddie Mac origination data, or approximately 4 million historical loan data points. Figure 9 shows the predicted versus actual interest rate from our test dataset. The slope of a line fit to the predicted vs. actual rates is 0.987, while the R-squared value is 0.993. This means that our model accurately predicts mortgage rates close to actual interest rates provided by previous lenders for all levels of default risk.

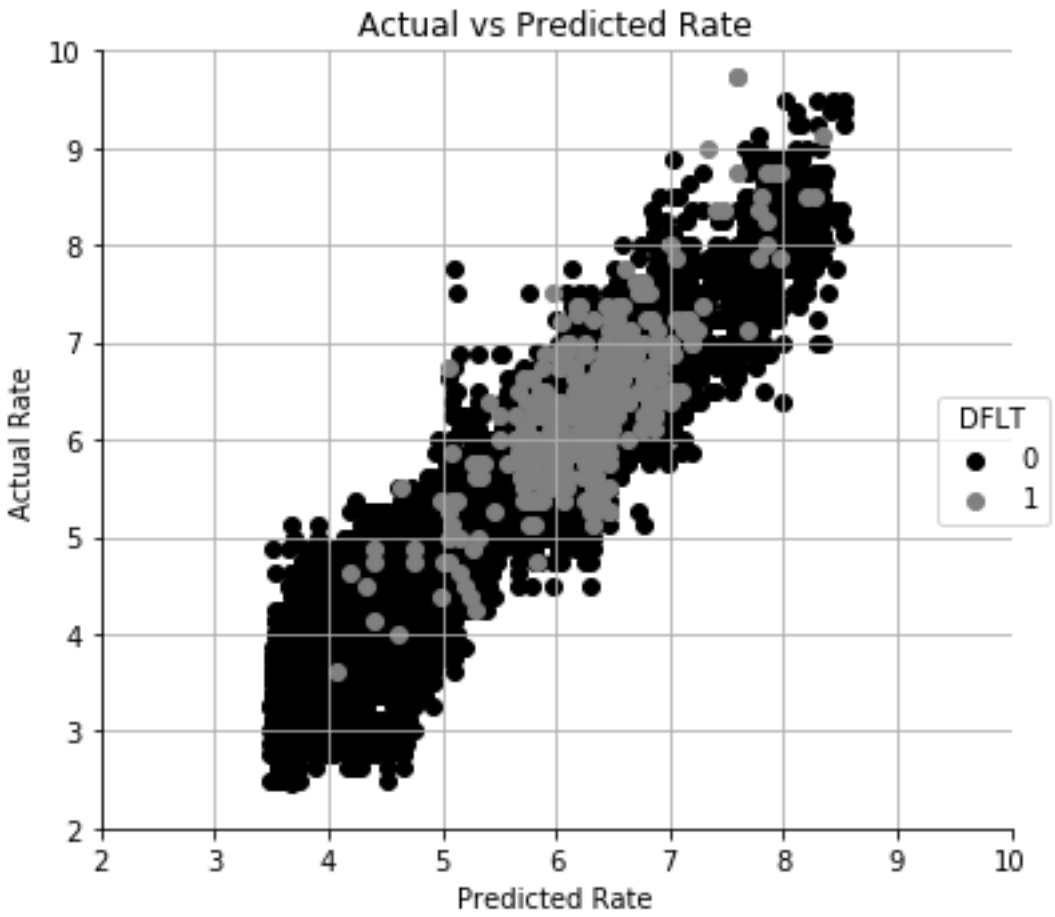

Figure 9. Predicted vs actually receive interest rate based on our test sample.

Figure 10 shows the default class proportion associated with each risk point bucket/range described previously. As predicted, more risky risk point buckets/range correlate to higher proportions of defaulting mortgages. This shows that we were able to effectively assign a risk score depending on specific loan characteristics.

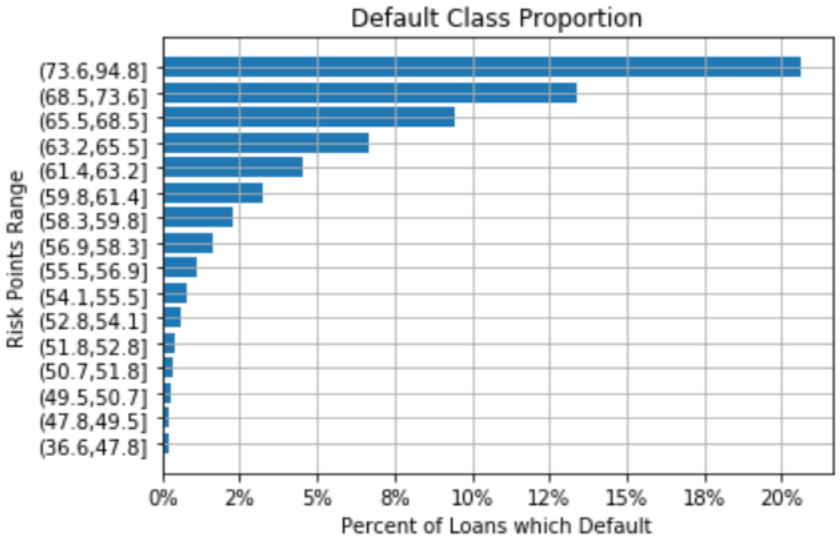

Figure 10. Percentage of defaulted mortgages for each risk score bucket.

If we look at the histogram in Figure 11 of origination default risk point distributions, we see that defaulted mortgages have higher risk points, unsurprisingly.

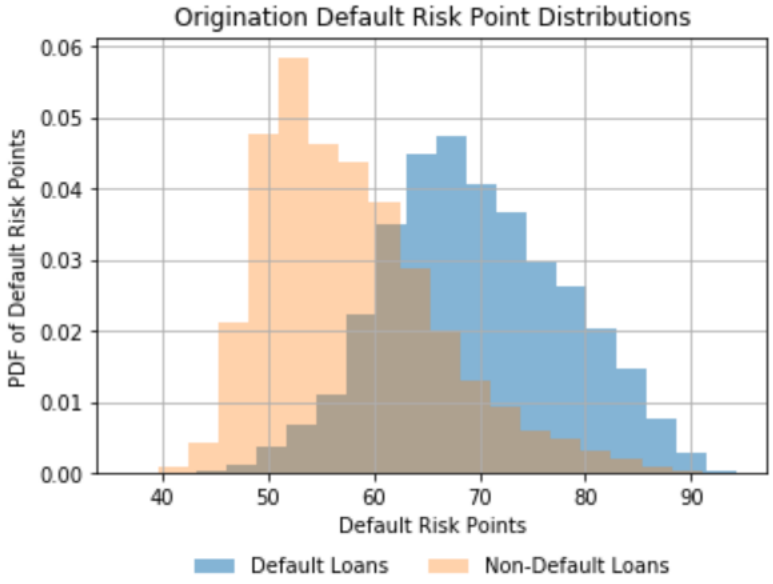

Figure 11. Risk point distribution for defaulted and non-defaulted loans.

We also evaluated our output based on similar interest rate calculators such as Bankrate.com's, and performed a Linear Shift of our results to better align with market forces. We've observed that our average rates were $\pm 0.25\%$ different than average market rates, which demonstrates how accurate are predictions were.

5.1.1 Tool Usability

From our risk point assignment, we calculate a range of spread for predicted interest rate quotes. To evaluate the usability of these predictions, we looked at the trends associated with this spread across the risk point buckets. Figure 12 shows this trend for a 15 year term, decreasing interest rate environment, and a low interest rate environment. Notice that there is an increasing slope, indicating that there are higher spreads for higher risk loans. Our pricing model based on risk points provides a risk score and financial market-based methodology for determining a mortgage rate. We performed this trend analysis for all the other factors and found similar results.

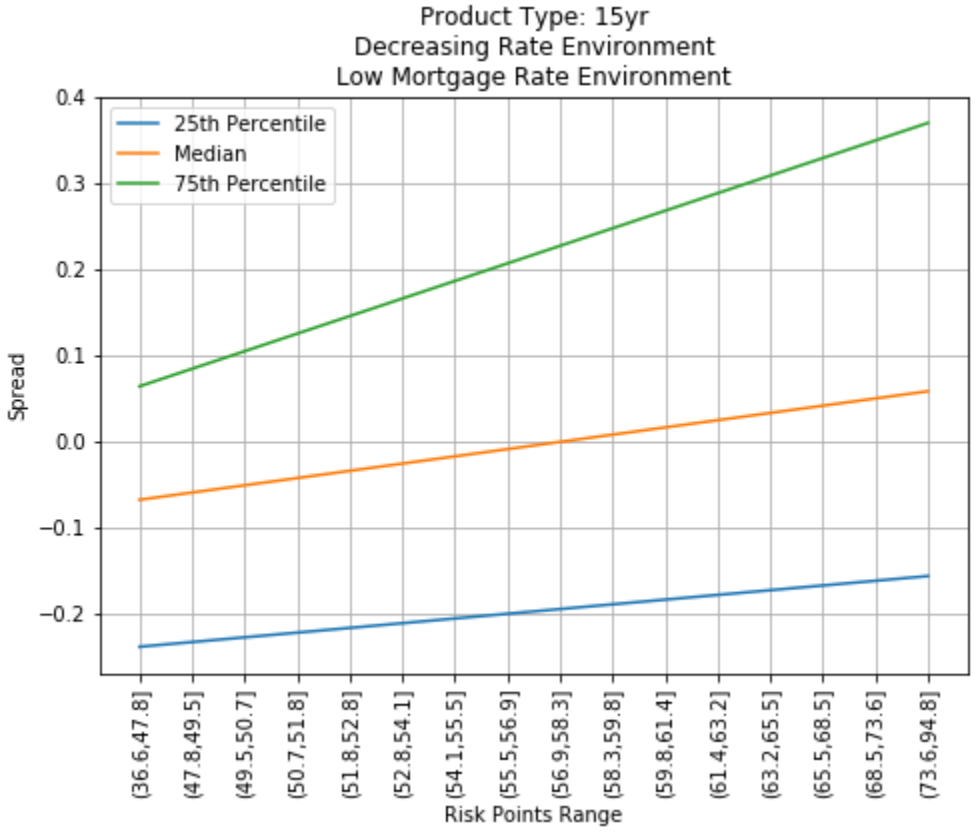

Figure 12. Interest rate thresholds across the risk score buckets for a 15-year fixed rate term, decreasing rate environment, and low mortgage rate environment.

To gauge the usability of our tool as a whole, we perform two user studies. The first is understand whether our tool easily allows the user to accomplish a set of intended use cases. This allows us to perform quantitative analysis of our tool's usability.

1. Understand historical default rates by state
2. Understand default trends that occurred by state
3. Understand the riskiness of a loan characteristic
4. Obtain interest rates that would increase profitability for a lender
5. Obtain interest rates that would reduce costs for a borrower

For each of the use cases, we create and provide a set of inputs, and establish indicators such as expected output that allow us to gauge tool effectiveness. We measure the duration a user takes to perform each use case, which allows us to understand bottlenecks in our tool. The uses cases are provided in Appendix A.

Our second user study involves a list of post-evaluation questions to understand user opinions of our tool. This allow us to perform qualitative analysis of our tool's usability.

1. What do you think were the intended purposes of our tool?
2. If you were interested in requesting a mortgage from the bank, how useful would this tool be in helping you understand the best interest rate for you?
3. If you were a loan officer providing a borrower with an interest rate, how useful would this tool be in helping you determine the best interest rate for your bank?

4. Based on the information provided in the tool, how easy was it for you to understand the various financial terms?
5. How visually appealing is the tool (e.g. with respects to color, layout)?
6. How easy was it to navigate through the tool?

By combining the results of both our quantitative and qualitative analyses, we can understand which components of our tool attributes to confusion or reduced usability. We modified our color scheme to avoid huge variations and the red-yellow spectrum to adjust for color blindness, and rephrased our definitions of financial terms to simplify terminology for average users. Our user studies indicated that the most useful capability of our tool was the integration between historical data analysis, and personalized input analysis via both tools. Our respondents also reported that this integration provides the right useful information to make requesting a quote much easier. Interesting analyses are in **Appendix B**.

6 Conclusion and Discussion

Lenders and borrowers are interested in understanding mortgage default risk. Our tool provides historical data analysis through state-level choropleth mapping, and risk score and rate quote calculations to help users make more informed decisions when requesting a quote. Instead of predicting default probability at origination or the time when a loan defaults, we weigh features based on how they influence default risk, map a loan's characteristic to risk score buckets, and calculate interest rate thresholds that indicate above-, at-, and below-average interest rates. This methodology and tool results in four innovations:

1. Created a easily interpreted model deployable at a bank. A loan officer doesn't have to understand machine learning to provide a customer a rate quote.
2. A loan officer can know the upper and lower bound of rates as in Figure 12 to retain the customer while compensating the bank for risk and maximizing expected profits.
3. Borrowers have a convenient approach to evaluate their rates through the calculator.
4. The historical data choropleth allows easy comparison to past mortgage loans.

7 Appendix A

7.1 Use Case 1

You are interested in understanding whether the loan you received from the bank is average compared to the mortgage market. Pretend you have a FICO score of 721, make $8,500/month pre-tax, have a monthly debt of $1,000, and have $38,000 for a downpayment. You want to buy your first property, which will be a single family home priced at $500,000.

The bank gave you an interest rate of 5.51% for your 30 year loan, one borrower, and during a recession.

Questions:

1. What percentage of people are you more risky than?
2. How did your interest rate compare to the national average?
3. How long did the use case take you?

7.2 Use Case 2

You are interested in understanding historical mortgage data for your state. Pretend you want to figure out how loans that originated in 2008 are doing in 2010 for the state Nevada. You are interested in people with FICO scores similar to yours (721), but don't care about the LTV ratio.

Questions:

1. How many loans were recorded for that specific input?
2. What percentage of mortgages defaulted?
3. What was the average risk score?
4. How long did the use case take you?

8 Appendix B

8.1 House Price Declines

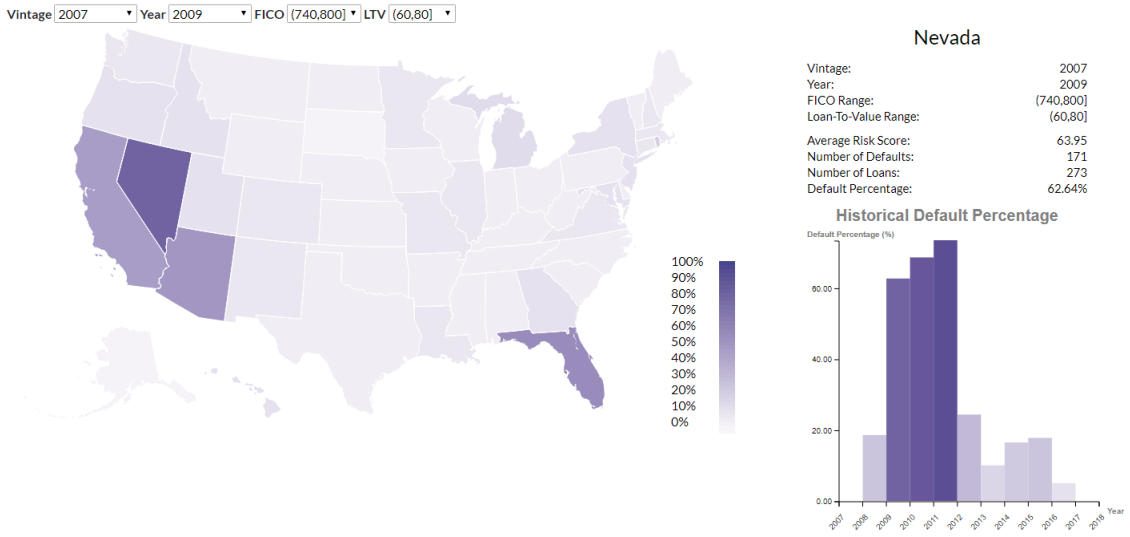

Figure 13. Using our choropleth tool, we show House Price Declines.

An interesting event happened during the financial crisis. US home prices fell dramatically. From Jan 1992 to April 2007, the US housing market had not observed a decrease in home prices. However, as the housing crisis began to emerge, some states experienced dramatic declines. For example, from August 2006 to June 2009 home prices in Nevada fell by roughly 54%! In general banks can protect themselves by requiring larger down-payments (high LTVs). For example, if I had a mortgage loan with an LTV of 80 and home prices dropped by 15%. The bank would still have a 5% buffer to protect them in the event of me defaulting. If you look at the map below you can see that the low LTV's and high credit scores in Nevada weren't enough to protect banks from default risk. In Nevada, 62.64% of loans with between 40 and 20% down payments (LTV (60-80]) that originated just after the 2006 housing price peak defaulted in 2009!

8.2 Extreme Weather Events

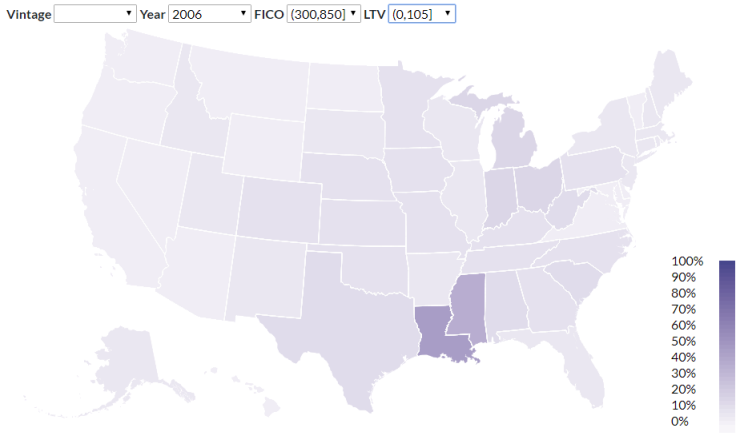

Figure 14. Using our choropleth tool, we show Extreme Weather Events.

Do extreme weather events affect defaults? Look at the figure below. Post Katrina (2005) Louisiana and Mississippi had 36% and 28% default rates respectively!

8.3 Financial Armageddon

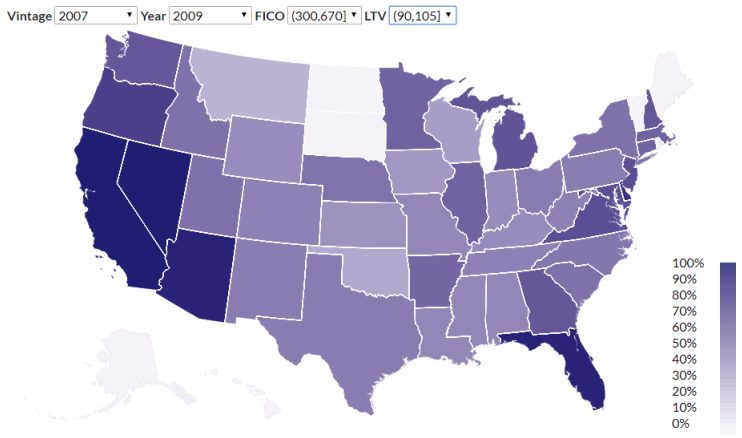

Figure 15. Using our choropleth tool, we show Financial Armageddon.

Loans originating in 2007 are widely considered the worst vintage. Particularly based on their performance in 2009. The chart below shows a financial Armageddon for the riskiest borrowers (low credit scores and high LTVs) in this category. Nearly 100% of these loans defaulted in California, Nevada, Arizona, and Florida during 2009.

9 Bibliography

[1] Addo, P. M., Guegan, D., Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risk, 6(38). doi:doi:10.3390/risks6020038. Retrieved from https://www.mdpi.com/2227-9091/6/2/38/pdf.

[2] Akindaini, B. (2017). Machine Learning Applications in Mortgage Default Prediction (Unpublished master's thesis). University of Tampere. Retrieved from http://tampub.uta.fi/bitstream/handle/10024/102533/1513083673.pdf.

[3] Aoidh, E. M., Bertolotto, M., Wilson, D. C. (2008). Understanding geospatial interests by visualizing map interaction behavior. Information Visualization, 7(3-4). doi:10.1057/palgrave.ivs.200824. Retrieved from http://journals.sagepub.com/doi/pdf/10.1057/IVS.2008.24.

[4] Bagherpour, A. (2017). Predicting Mortgage Loan Default with Machine Learning Methods. Retrieved from http://economics.ucr.edu/job_candidates/Bagherpour-Paper.pdf.

[5] Baldominos, A., Jose Moreno, A., Iturrarte, R., Bernardez, O., Alfonso, C. (2018). Identifying Real Estate Opportunities using Machine Learning. Retrieved from http://adsabs.harvard.edu/cgi-bin/bib_query?arXiv:1809.04933.

[6] Bhattacharya, Arnab P. Wilson, Simon Soyer, Refik. (2017). A Bayesian approach to modeling mortgage default and prepayment. https://arxiv.org/pdf/1706.07677.pdf

[7] Dang, G., North, C., Schneiderman, B. (2001). Dynamic queries and brushing on choropleth maps. Retrieved from https://ieeexplore.ieee.org/document/942141/authorsauthors.

[8] Deng, G. (2016). Analyzing the Risk of Mortgage Default (Unpublished master's thesis). Retrieved from https://www.stat.berkeley.edu/ aldous/Research/Ugrad/Grace_Deng_thesis.pdf

[9] Hoaglin, David C., Frederick Mosteller, and John W. Tukey. Exploring data tables, trends, and shapes. New York: Wiley-Interscience, 1985. Print. Chapters 10-11

[10] Hwang, S., Park, M., Lee, H. (2013). Dynamic analysis of the effects of mortgage-lending policies in a real estate market. Mathematical and Computer Modelling, 57(9-10), 2106-2120. doi:10.1016/j.mcm.2011.06.023

[11] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. Loss models : from data to decisions. Hoboken, N.J: John Wiley Sons, 2008. Print. Chapters 16-17

[12] Li, M. (2014, October). Residential Mortgage Probability of Default Models and Methods (Rep.). Retrieved from https://www.fic.gov.bc.ca/pdf/fid/14-0877-sup.pdf

[13] Marsh, S. L. (2007). Using and Evaluating HCI Techniques in Geovisualization: Applying Standard and Adapted Methods in Research and Education (Unpublished master's thesis). Retrieved from https://pdfs.semanticscholar.org/c096/5d4427673401320ff927e860600f65d6e89a.pdf

[14] Moosavi, V. (2017). Urban Data Streams and Machine Learning: A Case of Swiss Real Estate Market. Retrieved from https://arxiv.org/abs/1704.04979.

[15] Servigny, and Olivier Renault. Measuring and managing credit risk. New York: McGraw-Hill, 2004. Print. Chapter 3: Credit Scoring

[16] Skowronnek, Alsino. "Beyond Choropleth Maps: A Review of Techniques to Visualize Quantitative Areal Geodata." Alsino.io, INFOVIS READING GROUP WS 2015/16, 2015 alsino.io/static/papers/BeyondChoropleths_AlsinoSkowronnek.pdf. https://alsino.io/static/papers/BeyondChoropleths_AlsinoSkowronnek.pdf

[17] Sonderby, S. (2014). Non-parametric survival analysis in breast cancer using clinical and genomic markers (Unpublished master's thesis). Technical University of Denmark. Retrieved from http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6779/pdf/imm6779.pdf

[18] Tan, P., Steinbach, M., Karpatne, A., Kumar, V. (2019). Introduction to Data Mining, 2nd Edition. Retrieved from http://www.mypearsonstore.com/bookstore/introduction-to-data-mining-9780133128901

[19] Wood, J. Dykes, J. (2008). Spatially Ordered Treemaps. IEEE Transactions on Visualization and Computer Graphics, 14(6), pp. 1348-1355. doi: 10.1109/TVCG.2008.165 http://openaccess.city.ac.uk/536/1/wood_spatially_2008.pdf

[20] Zhang, Q. (2015). Modeling the Probability of Mortgage Default via Logistic Regression and Survival Analysis (Unpublished master's thesis). University of Rhode Island. Retrieved from https://digitalcommons.uri.edu/cgi/viewcontent.cgi?referer=https://www.google.com/httpsredir=1article=1543context=theses

[21] Zhou, L., Wang, H. (2012). Loan Default Prediction on Large Imbalanced Data Using Random Forests. Indonesian Journal of Electrical Engineering, 10(6), 1519-1525. Retrieved from https://www.researchgate.net/publication/267864165_Loan_Default_Prediction_on_Large_Imbalanced_Data_Using_Random_Forests.