Calculating Mortgage Loan Risk

Georgia Tech College of Computing

Online Master of Science Computer Science (OMS CS)

Team 43

Harris Ashraf hashraf3@gatech.edu
Sharmila Baskaran sbaskaran30@gatech.edu
Bo Chen bchen354@gatech.edu
Travis Jefferies tjefferies3@gatech.edu
Daniel Mower dmower3@gatech.edu
Cody Nguyen cnguyen311@gatech.edu
Ryan Wong rwong33@gatech.edu

Motivation - Introduction

Lenders provide borrowers with a mortgage loan that is paid back monthly within a fixed time-frame. Lenders make profit by imposing interest, which decreases if the borrower introduces risk by defaulting, refinancing, or prepaying their mortgages.

Lenders want to understand influential factors associated with mortgage risk.

Borrowers want to reduce their interest rate when requesting a mortgage loan.

Data Source

Freddie Mac

Freddie Mac provides loan-level origination and performance data to the public for purchased mortgages

26.3 M	75 GB	18 Years	53
US mortgages	large dataset	1999 - 2017	viable features

Project Goal

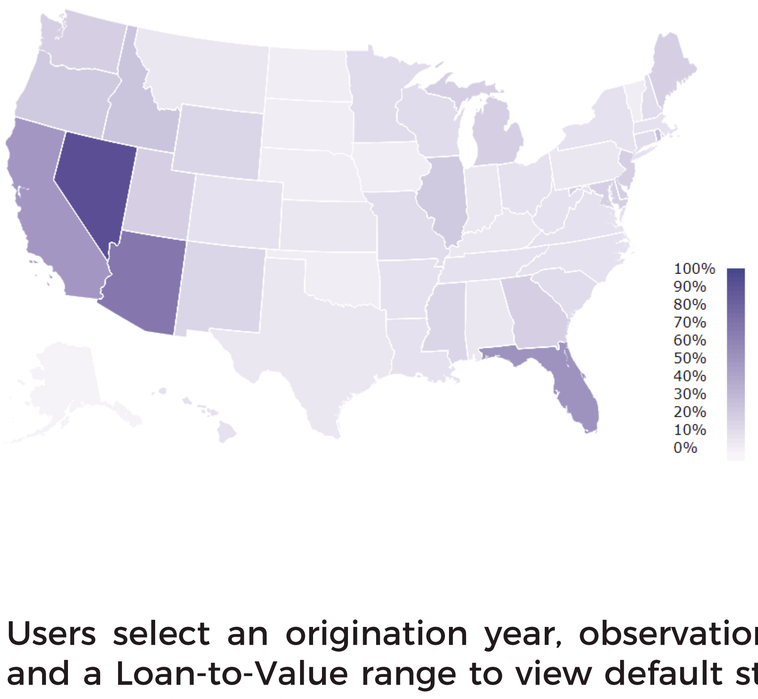
Create a tool that will allow both lenders and borrowers to **easily gauge the financial market and mortgage loan characteristics** to make more informed decisions.

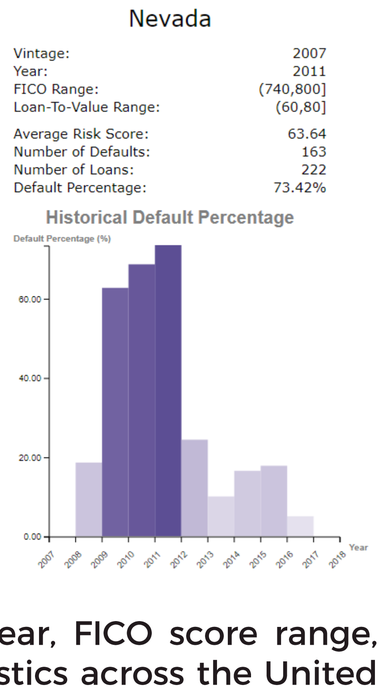
We will do this by allowing users to:

1. Easily access historical mortgage data
2. Efficiently calculate the riskiness of specific loan characteristics

Our method is novel because it is analyzed on an individual-loan level, and conveniently integrates two important financial applications.

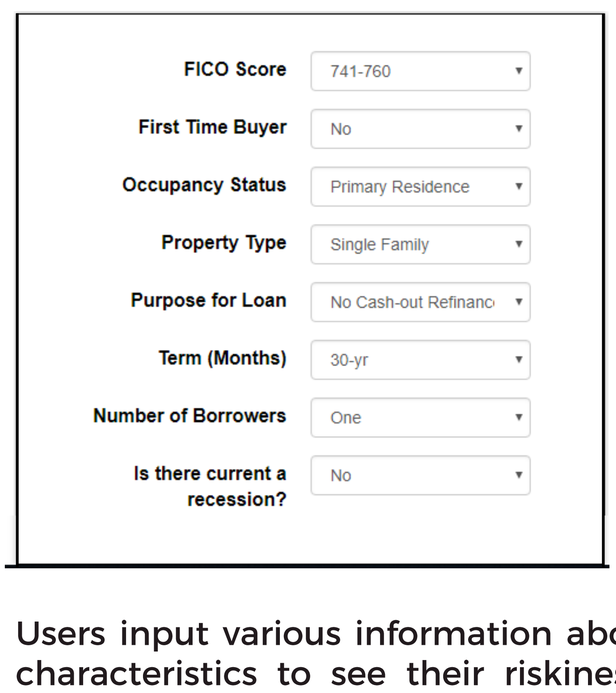
Historical Mortgage Heatmap

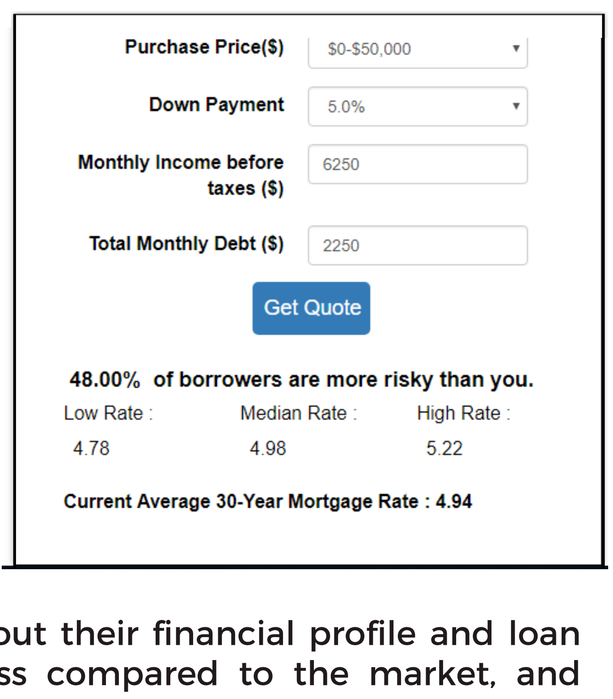




Users select an origination year, observation year, FICO score range, and a Loan-to-Value range to view default statistics across the United States. This provides users with 5,575 queries across 50 states and D.C.

Risk-Based Mortgage Rate Quote





Users input various information about their financial profile and loan characteristics to see their riskiness compared to the market, and rates that resemble the spread appropriate for their situation.

Model Algorithm

Extract the mortgage data from Freddie Mac, and sanitize it to remove missing values, address mixed data columns, and retain relevant features

Use a Distributed Random Forest to perform dimensionality reduction for feature selection to figure out which attributes influence default risk

Use these selected features to provide a Risk Point measure that correlates directly to default risk, better than any individual attribute

Categorize the Risk Point measures into 16 discrete “buckets” and use interquartile spread to provide the user with an interest rate quote range

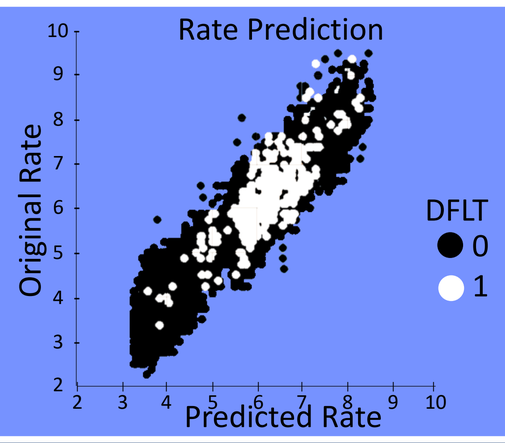
Use the selected features to map the user inputted parameters to a Risk Point bucket, and display the associated range

Scraping / Cleaning | Feature Selection | Risk Score Calculation | Rate Quote Range | Interface Input

Experimentation and Evaluation

We found that traditional methods to classify mortgage default at the time of origination (e.g. K-NN, Random Forest classification) failed because of heavily skewed non-defaulted historical data. Survival Analysis techniques were also negative influenced by the limited natural of mortgage datasets.

By maintaining 15% of the dataset or 4 million historical mortgage data points as our test sample, we were able to consistently predict median interest rate quotes with a ±0.25% difference from historical averages. Our original vs predicted rate has a slope of 0.987 and R-squared value of 0.993, meaning our predictions are close to the market.



Based on our user studies, the most useful capability was the integration between historical data and personalized input analyses via both tools. Our respondents reported this integration provides the right useful information to make requesting a quote easier.

Innovation

Users can quickly understand mortgage risk without knowledge of machine learning

Lenders can easily understand how to maximize expected profits

Borrowers can effortlessly evaluate their mortgage rates

Users can conveniently analyze mortgage trends and access historical data