MADS Milestone II Proposal

Michael Schoose
Paul Stotts
John Papazian

Introduction/Overview

The goal of our project is to better understand the characteristics of mortgage borrowers in the United States. We plan to use the National Survey of Mortgage Originations (NSMO). This survey data is a subset of the National Mortgage Database (NMDB), which is managed by the Federal Housing Finance Agency (FHFA) and the Consumer Financial Protection Bureau (CFPB).[1] These government agencies have provided information about the American mortgage market, which is based on a five percent sample of just new residential mortgages. The FHFA completes monthly surveys of the mortgage market, and they collect data on the characteristics of individual mortgages. These include subprime and nontraditional mortgages. In addition, FHFA collects information on the creditworthiness of borrowers. In fact, the National Mortgage Database includes information on Vantage score, which is a proxy of FICO credit scores. The NMDB program supports policymaking and research efforts.

The NMDB is a de-identified loan-level database of closed-end first-lien residential mortgages. The core data in NMDB represent a statistically valid 1-in-20 random sample of all closed-end first-lien mortgages in the files of Experian. It represents the market as a whole and contains detailed loan-level information on the terms and performance of mortgages. It also includes characteristics about the associated borrowers and properties. It has a historical component that dates back before the financial crisis of 2008, and it has been updated over time.

Other scholars have developed Machine Learning models to understand mortgages. Sadhwani, Giesecke, and Sirignano (2021) examined the behavior of mortgage borrowers in the United States between 1995 and 2014.[2] They used a different dataset, and they built a deep learning model of multi-period mortgage delinquency and foreclosures. However, they did not use traditional statistical methods such as logistic regression. Rather, they focused on building just a neural network model. Our team would like to improve their work by building statistical models in addition to various tree ensemble models such as gradient boosting and random forest. This is the first time that any member of our team has analyzed this dataset.

One challenge of this project is that the National Survey of Mortgage Originations is rich with 50,000+ observations and 500+ features. We will first conduct exploratory data analysis. Afterwards, we will subset the data and reduce the number of features before conducting supervised learning and unsupervised learning.

Part A (Supervised learning)

We plan to use the National Survey of Mortgage Originations for supervised learning.[3] There are a number of target variable options to which we could apply classification or regression models. We are still in the process of deciding. For the classification target variable, we might use the performance status “`Perf_Status`” and the associated labels. This target gives us the opportunity to address challenges like an imbalanced class. We can draw valuable conclusions about how borrower education may influence mortgage performance. For the continuous target variable, we will likely use the Loan-to-Value (LTV) ratio. Additionally, we may use the “`Rate_Spread`” as a continuous target variable for regression.

For both, we plan to first develop a statistical model (e.g., logistic regression and linear regression). Then, we will develop various machine learning models (e.g., gradient boosting and random forest). We will compare the statistical model to the various machine learning models. We plan to import the data into Pandas and use Scikit-learn to develop various models. For the binary target models, we plan to evaluate them using metrics such as ROC-AUC. For the continuous target models, we plan to evaluate them using metrics such as RMSE. We plan to conduct visualization for supervised learning using Seaborn, Matplotlib, and/or Altair.

Part B (Unsupervised learning)

We plan to use the National Survey of Mortgage Originations also for unsupervised learning.[4] Due to the feature rich dataset (500+ features), we intend to attempt dimensionality reduction approaches, such as PCA or SVD, prior to fitting the supervised learning models to improve the performance of the supervised learning models.

We would like to cluster survey participants into groups to understand the structure of the data. First, we will need to reformat the survey answers. We plan to use unsupervised learning approaches such as PCA and K-means clustering. We could plot our clusters on two dimensions corresponding to the first and second principal components. This would enable us to evaluate and understand the similarity and differences between clusters of mortgage borrowers. We plan to conduct visualization for unsupervised learning using Seaborn, Matplotlib, and/or Altair.

Team Planning:

- Exploratory Data Analysis completed by May 27th
- Unsupervised Learning completed by June 3rd
- First Standup with Saurabh on June 3rd
- Supervised Learning completed by June 9th
- Draft report completed by June 16th

[1] https://www.fhfa.gov/data/national-survey-mortgage-originations-nsmo-public-use-file/dataset

[2] Apaar Sadhwani, Kay Giesecke , Justin Sirignano. "Deep Learning for Mortgage Risk". Journal of Financial Econometrics, Volume 19, Issue 2, Spring 2021, Pages 313–368, https://doi.org/10.1093/jjfinec/nbaa025

[3] https://www.fhfa.gov/document/NSMO-Technical-Report-v50.pdf

[4] https://www.fhfa.gov/sites/default/files/2024-06/v50-Appendix-C.pdf