**Draft Final Report**

Michael Schoose

Paul Stotts

John Papazian

June 13, 2025

**Introduction (5 points)**

* *What problem are you trying to solve?*
* *What impact will solving the problem have?*
* *What motivated you to work on it?*
* *Summarize your project’s supervised and unsupervised methods using for the problem, and highlight any novel contributions compared to related projects.*
* *What are your main findings for supervised and unsupervised learning?*

The goal of our project is to better understand the relationship between a borrower’s characteristics and their riskiness, which is expressed as a function of their mortgage interest rate spread at origination. We use the National Survey of Mortgage Originations (NSMO).[[1]](#footnote-1) This dataset is based on a five percent sample of new residential mortgages. The Federal Housing Finance Agency (FHFA) completes monthly surveys of the mortgage market, and they collect data on the characteristics of individual mortgages. We use this dataset for both supervised learning and unsupervised learning.

For our investigation, we create a synthetic target variable of “beta” that represents the riskiness of a borrower. Our synthetic target variable of beta is analogous to the beta (β) used in the Capital Asset Pricing Model (CAPM). This financial formula defines the expected return on an equity investment (Re) as a function of the risk-free rate (Rf) plus the beta of the investment multiplied by the equity risk premium. The equity risk premium itself is defined as the expected rate of return of the market (Rm) minus the risk-free rate (Rf). The expected rate of return of the market (Rm) is akin to the return of a diversified portfolio of stocks that matches the S&P 500 index. Likewise, one could think of the risk-free rate (Rf) as the yield of highly rated government bonds such as the 30-year U.S. Treasury, which has a very low probability of default.

Re = Rf + β(Rm − Rf)

One could rearrange the CAPM formula to define beta (β) as such

Re - Rf = β(Rm − Rf)

(Re - Rf) = β(Rm − Rf)

(Re - Rf) / (Rm − Rf) = β

β = (Re - Rf) / (Rm − Rf)

In the Capital Asset Pricing Model, beta (β) is a measure of the volatility or risk of an investment in relation to the market overall. Investments with a beta of 1 have an expected risk equivalent to the risk of the market overall. Investments with a beta greater than 1 are more risky than the market overall, while investments with a beta less than 1 are less risky than the market overall. There is an underlying connection between the risk of an investment and the investment’s expected rate of return. Investors are willing to tolerate a high-risk investment if the investment is expected to have a rate of return higher than the market overall. Likewise, investors will tolerate an investment with a low rate of return if it is expected to have a lower risk than the market overall. Thus, beta (β) is useful to distinguish a high-risk investment with a high expected rate of return versus a low-risk investment with an low expected rate of return.

In a similar manner, the CAPM formula can be applied to create a beta for specific mortgage. Banks view mortgages similar to how investors view investments. They lend money to homeowners with the expectation of getting a return on their investment. Some homeowners have a higher risk of defaulting on their mortgage. Thus, banks give those homeowners a higher interest rate on their mortgage to compensate the bank for taking on a higher risk. We use the following variables in our data to create the synthetic target variable of beta that represents the riskiness of a borrower.

treasury\_yield = market yield on U.S. Treasury securities at 30-year constant maturity[[2]](#footnote-2)

pmms: Freddie Macs’ primary mortgage market survey rate at origination[[3]](#footnote-3)

rate\_spread: mortgage interest rate spread at origination

We can conceptualize the treasury\_yield as the risk-free rate (Rf). Likewise, we can conceptualize the pmms as the expected risk of the market overall (Rm) because it is an average rate of all mortgages surveyed. Our expected rate of return for any individual mortgage can be defined using the formula below:

Re = rate\_spread + pmms

In other words, our expected rate of return for any individual mortgage is the sum of the mortgage interest rate spread at origination plus Freddie Mac’s average mortgage rate from their market survey. We can substitute these terms into the CAPM formula to compute the beta for each individual mortgage:

β = (Re - Rf) / (Rm − Rf)

β = (rate\_spread + pmms - Rf) / (Rm − Rf)

β = (rate\_spread + pmms - treasury\_yield) / (pmms − treasury\_yield)

Using this formula, we compute beta (β) as the synthetic target variable for our models. Similar to the beta for a stock, it represents the riskiness of an individual mortgage compared to other mortgages. Mortgages with a beta of 1 have an expected risk equivalent to the risk of the mortgage market overall. Mortgages with a beta greater than 1 are riskier than the market overall, while mortgages with a beta less than 1 are less risky than the market overall.

The primary problem that we are trying to solve is to see if we can predict the risk of the mortgages (i.e. the beta) in the National Survey of Mortgage Originations portfolio using mortgage characteristics and macroeconomic factors as features. We are motivated to work on this research to learn how mortgage characteristics influence systematic risk relative to market movements. Moreover, we are interested to see if we can identify high-beta versus low-beta mortgage profiles for portfolio management. Solving this problem will have an impact because it will enable us to learn what are the underlying variables explaining the riskiness of mortgages. To carrying out our research, we are employing both supervised learning and unsupervised learning.  
  
For supervised learning, we are interested to learn which mortgage features (e.g., loan characteristics, borrower profiles, property details) are most predictive of a mortgage beta. We want to learn how accurately different Machine Learning models can predict mortgage beta. Moreover, we are curious which model family (probabilistic, tree-based, instance-based, etc.) performs best for this financial prediction task.  
  
For unsupervised learning, we want to learn if there are any natural groupings/clusters of mortgages based on their characteristics. Moreover, we want to know if these clusters correspond to different risk profiles (i.e., beta levels). Lastly, we are curious to know if we can identify mortgage segments that behave similarly in terms of systematic risk.

**Related work (5 points)**

*Find and reference at least three examples of existing projects or studies that are most similar to your project: include 1-2 sentence descriptions for each, and 1-2 sentences summarizing how what you’re proposing is different or improved compared to this existing work. This is also the place to mention if this project is an extension of a Milestone 1 or other previous course project.*

*For the Data Source and Feature Engineering sections below, if you use two different datasets or different feature engineering for supervised and unsupervised learning, just describe each of those in the appropriate sections of your report and we will merge the grading into the appropriate rubric sections.*

Other scholars have developed Machine Learning models to understand the risk level in mortgages. Sadhwani, Giesecke, and Sirignano (2021) examined the behavior of mortgage borrowers in the United States between 1995 and 2014. They built a deep learning model of multi-period mortgage delinquency and foreclosures. However, they did not use traditional statistical methods such as Ordinary Least Square (OLS) linear regression. Rather, they focused on building just neural network models.[[4]](#footnote-4) We chose to use traditional statistical methods in conjunction with Marching Learning methods. Also, we are not predicting mortgage delinquency nor foreclosure, but rather the riskiness of an individual mortgage as expressed by its beta.

Pillai, Woodbury, Dikshit, Leider, and Tappert (2019) developed Machine Learning models to predict the approval or denial of mortgage applicants. They used a two-tier machine learning model based upon risk exposure from the macro and micro level. They explore a variety of classification models including random forest, gradient boosting, and logistic regression.[[5]](#footnote-5) Our target variable of beta is continuous. Therefore, we explore regression models rather than classification models.

Lastly, Ojha, and Lee (2021) examined the U.S. Residential Mortgage Backed Securities (RMBS) to predict the probability of default. They studied the Fannie Mae loan data that includes around 1.5 million of mortgage loans originating from 2005 to 2009 in the United States. They investigated the probability of default in terms of credit risk based upon loan origination and performance characteristics. Their research focused on different types of neural networks. They developed models using Convolution Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM).[[6]](#footnote-6) In contrast, we would like to focus on traditional Machine Learning approaches such as gradient boosting and random forest.

**Data Source(s) (5 points)**

*Describe the properties of the dataset (or data API service) you used. Be specific. Your information at a minimum should include but not be limited to:*

* *where the datasets or API resource is located,*
* *what formats they returned/used,*
* *what were the important variables contained in them,*
* *how many records you used or retrieved (if using an API), and*
* *what time periods they covered (if there is a time element)*

We use the National Survey of Mortgage Originations (NSMO), which is based upon the time period of 2013 to 2023.[[7]](#footnote-7) 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).[[8]](#footnote-8) These government agencies have provided information about the American mortgage market, which is based on a five percent sample of all residential mortgages. The FHFA completes a monthly survey of the mortgage markets, 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 represents 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.

One challenge of this project is that the National Survey of Mortgage Originations is rich with over 50,000 observations and over 500 features.[[9]](#footnote-9) Many of the survey questions are categorical in nature. We first conduct exploratory data analysis. Afterwards, we subset the data and reduce the number of features before conducting supervised learning and unsupervised learning.

**Feature engineering (8 points)**

* *Describe the steps you used to get from the raw input data to the final features used w/ supervised and unsupervised ML methods.*
* *Explain any initial preprocessing that was required to handle noisy or missing data.*
* *Be sure to include a complete list of all final features (in an appendix if necessary).*

We did a lot of initial preprocessing before doing feature selection. After reading in the raw CSV survey data, we first bifurcate the variables into categorical variables and numeric variables. This is necessary because they have different approaches to imputation. We clean the raw data by converting negative values (representing missing values) into null values. Next, we then join our mortgage survey data with U.S. Treasury yields downloaded from the Federal Reserve. We compute the average 30-year U.S. Treasury yields over each year and month in our survey data, which we use as the risk-free rate (Rf). Using the modified CAPM formula above, we calculate the beta for every mortgage. As discussed above, this beta value serves as our synthetic target variable. There are many outlier values for beta, and therefore we Winsorize these beta values at the 5th and the 95th percentile to reduce the influence of outliers. Lastly, we do one-hot encoding by creating dummy variables representing each category for each categorical variable.

We also engineer a new feature. There are many Vantage scores in the mortgage survey, which is a proxy of FICO credit scores. Banks typically use a “lower middle” credit score. Bank usually pull from the three main credit bureaus for each borrower and then take the middle score for each borrower on the mortgage (e.g., the spouse). We use a similar approach to engineer a ‘low\_score’ feature, which represents the lowest of the borrower’s middle score.

Using subject matter expertise, we decided to exclude certain variables. These variables have been excluded for reasons such as being non-relevant identifiers, components of the target variable, or data from after mortgage origination providing “data leakage”. The excluded variables include the Vantage score after mortgage origination, mortgage performance status over time, forbearance status over time, etc. It is important to remove these variables because they are not useful to predict the beta for each mortgage at origination.

Finally, we do a train\_test\_split() to partition 80% of our observations for model training and keep the remaining 20% as holdout for model validation. We impute missing values for the numeric variables using the mean value imputation. The categorical variables do not need to be imputed because they already have categories representing missing values. We fit a SimpleImputer() on just the training data, which we then apply to both the training partition and the validation partition. We also fit a StandardScaler() on just the training data, which we then apply to both the training partition and the validation partition. These steps are necessary to prevent “data leakage” of information about the validation partition leaking into the training partition.

**Part A. Supervised Learning**

* *Methods description (8 points)*
  + *Briefly describe your supervised learning workflow, the learning methods you used, and the feature representations you chose. You must justify why you chose your methods.*
  + *You should have an adequate number and nature of methods: a minimum of three diverse model families must be described and explored (i.e. with very different underlying mechanisms, e.g. probabilistic, non-probabilistic, tree-based, instance-based, etc).  This might be variable with larger or smaller teams and project specifics, with instructor permission.*
  + *Include a description of how you did hyperparameter tuning or exploration with your models.*
  + *Methods used must be clearly described, each with correct justification.*
* *Supervised Evaluation (22 points)*
  + *In this section you will provide a correct and comprehensive evaluation, analyzing the effectiveness of both your methods, and your feature representations.*
  + *(8 points) Overall results reporting.*
    - *State and justify your choice of evaluation metrics used.*
    - *Provide at least one overall summary of results that compares the best model from each family you used, in a clear, concise table.*
    - *If comparing an evaluation metric between model families (e.g. comparing accuracy of support vector machines vs logistic regression), do not use just the result of a single training/test split: you must report the mean metric across multiple cross-validation folds (typically 5-fold CV), along with the standard deviation of the metric.*
  + *Please see the* ***“Tips for Project Report” video*** *under “Week 6 Project Check-in” for key methods to get insight into your machine learning model and how to report results.*
  + *For the following parts of the evaluation, typically you will do these deeper analysis steps only on your best-performing model (not all of them).*
  + *(6 points) Do a feature importance and ablation analysis on your best model to get insight into which features are or are not contributing to prediction success/failure.*
  + *(4 points) Do at least one sensitivity analysis on your best model: How sensitive are your results to choice of (hyper-)parameters, features, or other varying solution elements?*
  + *(4 points) Given your evaluation results and metrics, what important tradeoffs can you identify?  Some examples:  precision vs recall, training data size vs accuracy, speed vs accuracy, etc.*
* *Failure analysis (5 points)*
  + *Select at least 3 \*specific\* examples (records) where prediction failed, and analyze possible reasons why.*
  + *Ideally you should be able to identify at least three different categories of failure.*
  + *What future improvements might fix the failures?  (You do not need to implement these)*

**Part B. Unsupervised Learning**

* *Methods description (10 points)*
  + *Briefly describe your unsupervised learning workflow, the learning methods you used, and the feature representations you chose. You must justify why you chose your methods.*
  + *You should have an adequate number and nature of methods: a minimum of two unsupervised methods must be described and explored (i.e. with very different underlying mechanisms, e.g. probabilistic, non-probabilistic, tree-based, instance-based, etc).  This might be variable with larger or smaller teams and project specifics, with instructor permission.*
  + *Include a description of how you did hyperparameter tuning or exploration with your models.*
  + *Methods used must be clearly described, each with correct justification.*
* *Unsupervised Evaluation (15 points)*
  + *In this section you will provide a correct and comprehensive evaluation, analyzing the effectiveness of both your methods, and your choice of feature representation.*
  + *(10 points) Overall results reporting.*
    - *State and justify your choice of evaluation metrics used.*
    - *Provide at least one overall summary of results that compares the best model from each family you used, in a clear, concise table.*
    - *To summarize your findings, include at least two visualizations (chart, plot, tag cloud, map or other graphic) for each unsupervised method used that summarize your analysis.*
  + *(5 points) Do at least one sensitivity analysis on your best model: How sensitive are your results to choice of (hyper-)parameters, features, or other varying solution elements?*

**Discussion (8 points)**

* *(4 points) What did you learn from doing Part A?*
  + *What surprised you about your results?*
  + *What challenges did you encounter, and how did you respond to them?*
  + *How could you extend your solution with more time/resources?*
* *(4 points) What did you learn from doing Part B?*
* *What surprised you about your results?*
* *What challenges did you encounter, and how did you respond to them?*
* *How could you extend your solution with more time/resources?*

**Ethical Considerations (4 points)**

* *What ethical issues could arise in providing a solution to Part A, and how could you address them?*
* *What ethical issues could arise in providing a solution to Part B, and how could you address them?*

**Statement of Work (1 point)**

* You must include a statement that describes the contribution that each team member made to the project.

1. <https://www.fhfa.gov/data/national-survey-mortgage-originations-nsmo-public-use-file/dataset> [↑](#footnote-ref-1)
2. <https://fred.stlouisfed.org/series/DGS30> [↑](#footnote-ref-2)
3. <https://www.freddiemac.com/pmms> [↑](#footnote-ref-3)
4. 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> [↑](#footnote-ref-4)
5. Sivakumar Pillai, Jennifer Woodbury, Nikhil Dikshit, Avery Leider, and Charles Tappert. "Machine Learning Analysis of Mortgage Credit Risk". Proceedings of the Future Technologies Conference, October 2019, Pages 107–123, <https://doi.org/10.1007/978-3-030-32520-6\_10> [↑](#footnote-ref-5)
6. Vikram Ojha and JeongHoe Lee. "Default analysis in mortgage risk with conventional and deep machine learning focusing on 2008–2009." Digital Finance 3, 2021, Pages 249–271. <https://doi.org/10.1007/s42521-021-00036-4> [↑](#footnote-ref-6)
7. <https://www.fhfa.gov/data/national-survey-mortgage-originations-nsmo-public-use-file/dataset> [↑](#footnote-ref-7)
8. <https://www.fhfa.gov/document/NSMO-Technical-Report-v50.pdf> [↑](#footnote-ref-8)
9. <https://www.fhfa.gov/sites/default/files/2024-06/v50-Appendix-C.pdf> [↑](#footnote-ref-9)