Project Proposal

Graded

Group

Hunter Grimes
Jihan Lee
Vidhi Mittal
View or edit group

Total Points

97 / 100 pts

Question 1

Overview/Motivation/Problem

29 / 30 pts

– **0 pts** Correct

✔ – **1 pt** Unclear problem formulation

what is your task? Classification or regression task on what target?

Question 2

Related Work

10 / 10 pts

✔ – **0 pts** Correct

– **1 pt** Missing reference page

– **1.5 pts** No mention of previous work, if there is any, in the same area as the proposal. Specifically, what machine learning approaches have been taken and what are the results/limitations.

– **2 pts** No discussion on previous work

Question 3

Dataset

19 / 20 pts

– **0 pts** Correct

– **1 pt** Missing details of features

– **1 pt** No explanation on data pre-processing steps (i.e. how are you going to pre-process non-numerical columns such as IDs, timestamps, text, etc.)

✔ – **1 pt** Missing number of samples in dataset

– **10 pts** Missing

Question 4

Methodology

29 / 30 pts

- ✔ **– 0 pts** Correct
- ✔ **– 1 pt** Unclear description of proposed algorithm.
- **– 1 pt** Missing evaluation metrics/methods
- **– 1.5 pts** No discussion on feature pre-processing or feature selection
- **– 1 pt** Missing detail on cross validation technique (train-test split ratio? what's k for k-fold cross validation?)
- **– 2 pts** No discussion on validation/ hyperparameter tuning

how is your logistic regression or neural network model going to solve your problem(evaluate fairness)?

Question 5

Readability

10 / 10 pts

- ✔ **– 0 pts** Correct

Questions assigned to the following page: 1, 3, 4, and 5

Analysis of Biases in Mortgage Lending Using Machine Learning Algorithms

Vidhi Mittal, Hunter Grimes, Jihan Lee

1 Description of Problem

The presence of bias within the mortgage lending system in the United States is an important issue. The Home Mortgage Disclosure Act (HMDA) plays a crucial role in promoting transparency within the lending system by requiring financial institutions to disclose detailed loan information. This kind of transparency plays an important role in ensuring that lenders meet the housing needs of their communities, providing essential information to policymakers and highlighting potential discriminatory lending patterns.

Systemic bias in mortgage lending hinders not only home ownership but also the long-term wealth accumulation of affected communities. By employing various machine learning techniques, including logistic regression and neural networks, we aim to examine relevant data for patterns indicative of discriminatory lending practices. We set our sights beyond the racial disparities, which have been investigated through a number of major studies, to uncover any form of demographic inequities such as socioeconomic status and geographic location. Our approach can potentially contribute to a more comprehensive understanding of fairness in mortgage lending practices.

2 Description of Data

We will be looking at the 2022 dataset from the Federal Financial Institutions Examination Council (FFIEC) under the Home Mortgage Disclosure Act (HMDA). There are up to 110 data attributes per mortgage application which contain a lot of information relevant to our project such as applicant demographics (income, ethnicity, race), loan specifics (amount, type, purpose, outcome), property details (location, type, occupancy), and lender information.

We are interested in specific variables like "denial_reason-1", which include debt-to-income ratio, employment history, credit history, collateral, and more. These variables could shed some light on potential areas of discrimination or bias.

Furthermore, the website allows us to filter data based on geographic location or financial institution, which is beneficial for narrowing our focus to specific regions or types of lenders in order to simplify the project if necessary.

It is interesting that variables like loan amount and income have no upper limit. Hence, they may contain outliers that could skew our findings, which is something to consider while pre-processing.

3 Methodology

We will start off with pre-processing of the data. The data will be cleaned to address missing values, outliers, and duplicate values to ensure model accuracy and reduce potential biases as mentioned above. Categorical features will be transformed using one-hot encoding, while numerical features will be normalized to eliminate scale biases.

Questions assigned to the following page: 2, 3, 4, and 5

Then, the dataset will be split into training, validation, and testing sets to evaluate model performance. We will also try to make feature selection using L1 and L2 regularization methods, commonly known as Lasso and Ridge regression, respectively. Regarding analytical methods, we plan to use logistic regression or deep learning techniques, specifically neural networks, as they can deal with complex non-linear patterns.

Hyperparameter tuning will be conducted through cross-validation and potentially other methods like Grid Search or Random Search to enhance robustness of the model. To improve prediction accuracy and address potential overfitting, ensemble methods like Random Forest and Gradient Boosting will be implemented. Our model will be evaluated on precision, recall, and the F1 Score in predicting mortgage approval. Fairness in model and data would be evaluated using open-source tools like AI Fairness 360 or Fairlearn.

4 Literature Review

Several previous studies have been conducted in this vein on mortgage data. The following are a few relevant examples.

Our first study (Hodges et al., 2024) was the singular machine learning study conducted on our specific dataset. The study found that a deep neural network classifier was the most effective. It also found that the number of features was slightly limiting, and that the algorithm resulted in race and gender biases.

Next, though conducted on a different dataset, the study was able to create an effective deep learning classifier. The researchers remarked on the high levels of interaction between most variables down to the zipcode. (Sadhwani et al., 2021)

Wang, Hong, and Wu focused on predicting loan rates. Specifically, they attempted to test if deep neural networks could be used instead of regression models to alleviate the effects of the gaps and outliers common in mortgage datasets. The study found that deep neural networks were highly effective and recommended their use categorically.(Wang et al., 2023)

Jemai and Zarrad's study was conducted on a wider base of financial data in the service of feature selection for credit risk assessment. They were successfully able to implement a deep learning approach for this feature selection. Additionally, the study found that interest rates on loans were highly predictive of credit risk (which may be relevant to our mortgage default prediction).(Jemai and Zarrad, 2023)

References

Hodges, H., Garrity, C., and Pope, J. (2024). Deep learning, feature selection and model bias with home mortgage loan classification. In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, pages 248–255. SciTePress.

Jemai, J. and Zarrad, A. (2023). Feature selection engineering for credit risk assessment in retail banking. *Information*, 14:200.

Sadhwani, A., Giesecke, K., and Sirignano, J. (2021). Deep learning for mortgage risk. *Journal of Financial Econometrics*, 19(2):313–368.

Wang, D., Hong, D., and Wu, Q. (2023). Prediction of loan rate for mortgage data: Deep learning versus robust regression. *Computational Economics*, 61(3):1137–1150.