Analysis of Bias in Mortgage Lending

Hunter Grimes, Jihan Lee, Vidhi Mittal

May 8, 2024

Abstract

This project investigates biases in U.S. mortgage lending using the Home Mortgage Disclosure Act (HMDA) dataset, with a focus on demographic disparities affecting mortgage approvals. Building machine learning models such as Random Forest and XGBoost, and utilizing the AI Fairness 360 toolkit, we analyze potential biases to mitigate those related to race, ethnicity, sex, and age. Our study quantifies these biases and proposes mitigation strategies through fairness assessments. Findings indicate moderate to significant biases, especially against racial minorities, highlighting the need for measures to ensure fairness in mortgage lending practices. This project contributes to efforts aimed at enhancing financial equity and regulatory compliance in the housing market.

1 Introduction

1.1 Motivation and Significance

The mortgage lending system in the United States plays a crucial role in enabling homeownership, a key component of financial stability and wealth creation for many Americans. Despite regulatory efforts such as the Home Mortgage Disclosure Act (HMDA), which mandates transparency by requiring the disclosure of loan information, demographic disparities persist. These biases not only hinder fair access to homeownership but also impact long-term wealth accumulation among disadvantaged communities. Addressing discriminatory practices in mortgage lending is essential not only for ensuring equity in housing opportunities but also for helping financial institutions adhere to fairness laws such as the Fair Housing Act and the Equal Credit Opportunity Act.

Our project aims to employ machine learning techniques to train models on datasets provided under HMDA to predict mortgage approvals. We then plan to use fairness techniques to detect and address biases, expanding the scope beyond the commonly studied racial disparities to include other demographic factors like socioeconomic status and geographic distribution.

The novelty of our project lies in this comprehensive approach to understanding biases within mortgage lending. By moving beyond traditional focus areas like race, and including other variables, our project offers a more holistic view of the issues at hand. This helps in developing nuanced models that reflect real-world complexities.

The significance of our project is that it has practical implications for enhancing transparency and accountability within the financial sector. Our project aims to generate actionable insights that can help financial institutions in complying with legal standards and promote trust among consumers, thereby fostering a more just financial ecosystem.

1.2 Background

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

First study we examined (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 zip code. (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)

2 Data

For our project, we utilized the 2022 Home Mortgage Disclosure Act (HMDA) Public Loan/Application (LAR) Data, which encompasses a comprehensive compilation of mortgage application filed for the year. This dataset is publicly available and features a substantial volume of data, including 16,080,210 observations across 99 variables. It incorporates demographic attributes such as race, ethnicity, age, and sex, alongside loan-specific details like loan purpose, type, interest rate, and loan-to-value ratio. This rich dataset's high dimensionality requires careful preprocessing and feature selection before proceeding with training and evaluating machine learning models.

2.1 Protected Attributes

For the purposes of this project, we identified protected attributes to evaluate potential biases in mortgage lending processes. We selected ethnicity, race, sex, and age as the critical attributes. To facilitate our fairness assessments, we categorized these attributes into historically privileged and unprivileged groups as following:

- Ethnicity
  - Privileged: Not Hispanic or Latino
  - Unprivileged: Hispanic or Latino
- Race
  - Privileged: White
  - Unprivileged: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, 2 or more minority races
- Sex
  - Privileged: Male
  - Unprivileged: Female
- Age
  - Privileged: 25-74
  - Unprivileged: <25, >74

This categorization will be instrumental in applying various fairness metrics to determine if there are any statistical biases or instances of unfairness against the unprivileged groups in subsequent analyses.

2.2 Data Preprocessing

In the data preprocessing phase of our project, meticulous attention was required due to the dataset's large size and high dimensionality. This phase was the most time-consuming aspect of the project, primarily because of the numerous challenges presented by missing values, data inconsistencies, undesirable formats, and the sheer volume of data, along with the need to address potential data leakage issues.

We initiated our preprocessing by eliminating features that were deemed unnecessary, redundant, or had excessive missing data (over two-thirds missing). For example, geographical information was available at multiple levels (state, metropolitan area, county, census tract), but we chose to retain only the state code to avoid issues of high cardinality that exacerbate the dimensionality problem. Additionally, we removed observations that provided no useful information, such as those with entries like 'Not Available' or 'Free Form Text Only' for race, ethnicity, and sex.

Further, the target variable 'action_taken' was recoded into a binary variable named 'loan_approved,' which simplified the output to just 'loan approved' or 'loan denied' statuses. Categorical features denoted by integers were recoded to reflect their actual categories to enhance interpretability. To manage the dataset size and ensure computational efficiency, we randomly sampled 100,000 observations for model building.

The data was then split into a training set and a test set in a 70:30 ratio. Missing values were imputed using the median for floating-point features and the mode (or the most frequent value) for other types of features. Numerical features were standardized using z-score standardization, and categorical features were one-hot encoded.

2.2.1 Feature Selection

Feature selection was rigorously performed using Pearson correlation coefficients. Features highly correlated with the target (above 0.8) or with each other (above 0.9) were evaluated, and the less relevant ones (based on their correlation with the target) were dropped. This process resulted in a refined set of 130 features and one target variable, with the training data consisting of 70,000 entries and the test data comprising 30,000 entries. Despite reducing the number of original features, due to one-hot encoding, the remaining high dimensionality necessitate careful consideration in subsequent modeling stages.

2.3 Data Exploration

Before diving into machine learning modeling, we conducted a brief exploratory data analysis and visualization to gain insights from our dataset. Initially, we examined the distribution of our target variable - loan approval - via a pie chart (see Figure 1), which revealed a 75:25 split between loans approved and denied. Given the binary nature of our classification task, this class imbalance poses a potential challenge. To address this issue, we plan to employ multiple evaluation metrics such as F1-score, precision, recall, and area under receiver operating characteristic curve, and consider using techniques like the Synthetic Minority Oversampling Technique (SMOTE) to balance the classes before modeling.

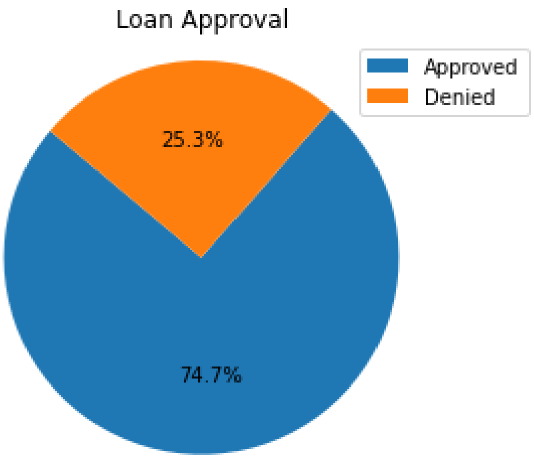

Figure 1: Pie Chart of Loan Approval

Further, Tables 1 to 4 present the distributions of loan approval across protected attributes, highlighting potential disparities.

	Total	Approved	Denied
Non-Hispanic	87.2%	88.0%	84.7%
Hispanic	12.8%	12.0%	15.3%

Table 1: Distribution of Loan Approval by Ethnicity

For ethnicity, despite the overall data comprising 87.2% non-Hispanic and 12.8% Hispanic individuals, non-Hispanic applicants constituted a higher proportion of the approved loans (88.0% vs. 12.0% for Non-Hispanic) and a lower proportion of the denied loans (84.7% vs. 15.3%).

	Total	Approved	Denied
White	77.4%	79.3%	71.8%
Black	12.9%	11.0%	18.7%
Asian	8.0%	8.3%	7.0%
Native	1.0%	0.8%	1.5%
Pacific Islander	0.3%	0.3%	0.5%
2+ Minority Races	0.3%	0.3%	0.5%

Table 2: Distribution of Loan Approval by Race

A similar pattern emerged with race, where white individuals were overrepresented in the approved category relative to their general population proportion (79.3% vs. 77.4%), and black individuals were underrepresented in approvals (11.0% vs. 12.9%) but overrepresented in denials (18.7% vs. 12.9%).

	Total	Approved	Denied
Male	59.5%	60.0%	58.0%
Female	40.5%	40.0%	42.0%

Table 3: Distribution of Loan Approval by Sex

	Total	Approved	Denied
<25	3.6%	3.9%	3.0%
25-34	20.3%	21.6%	16.6%
35-44	24.9%	25.2%	24.0%
45-54	21.9%	21.6%	23.1%
55-64	16.7%	16.1%	18.3%
65-74	9.0%	8.5%	10.4%
>74	3.5%	3.2%	4.5%

Table 4: Distribution of Loan Approval by Age

3 Methods

Our approach was structured to ensure the development of robust predictive models before evaluating their fairness. Initially, using our preprocessed data, we aimed to establish a baseline model that performed reasonably well to set a standard for subsequent more complex models. To this end, we chose logistic regression without any tuning or regularization as our baseline model.

To explore the potential of achieving higher performance and better handling of the high-dimensional data, we incorporated more sophisticated ensemble models such as Random Forest and XGBoost. These models are known for their robust performance across various types of data scenarios and their ability to model non-linear relationships more effectively than simpler models.

3.1 Random Forest

We chose Random Forest because it is known for performing well at handling high dimensionality and reducing the risk of overfitting through its ensemble approach. Also, it offers feature importance metrics, which are helpful in identifying key factors that influence lending decisions.

3.2 XGBoost

We chose XGBoost for similar reasons as Random Forest, i.e., due to its robust performance with high-dimensional datasets and its capability to handle class imbalance effectively, especially when combined with techniques like SMOTE. XGBoost provides feature importance information as well.

3.3 AI Fairness 360

To address the critical aspect of fairness in our model predictions, we utilized the AI Fairness 360 toolkit provided by IBM. This toolkit is designed to help detect bias and potentially mitigate it in machine learning models and datasets. It offers a comprehensive suite of metrics and algorithms that enabled us to assess fairness in both our dataset and the models we developed. We examined fairness in the dataset using two metrics (Disparate Impact and Statistical Parity Difference) and the models using four metrics (Disparate Impact, Statistical Parity Difference, Average Odds Difference, and Equal Opportunity Difference).

4 Experiments and Results

4.1 Hyperparameter Tuning

We performed hyperparameter tuning using random search cross-validation instead of exhaustive grid search due to large amount of data. The parameter grids we chose for random forest and XGBoost look as follows:

- Random Forest
  - 'n_estimators': [10, 50, 100]
  - 'max_depth': [None, 5, 10, 20]
  - 'min_samples_split': [2, 5, 10]
  - 'min_samples_leaf': [1, 2, 4]
  - 'max_features': [None, 'sqrt', 'log2']
- XGBoost
  - 'booster': ['gbtree', 'dart']
  - 'learning_rate': [0.1, 0.2, 0.3, 0.4, 0.5]

- 'min_split_loss': [0, 10, 100, 1000]
- 'max_depth': [4, 6, 8, 10]
- 'scale_pos_weight': [0.2, 0.25, 0.5, 0.8, 1]

4.2 Model Evaluation

After tuning our models, we trained and evaluated them. Both of our models were relatively close in performance, with XGBoost outperforming by only approximately 1.5% point in accuracy. As in our initial evaluation of the data, feature imbalance negatively affected our final results. We can see significantly lower accuracy and precision for the denied class as compared to approved due to its lower sample size.

Model	Accuracy	Precision(approved)	Precision(denied)	Accuracy(approved)	Accuracy(denied)
Random Forest	0.8195	0.8814	0.6424	0.8758	0.6543
XGBoost	0.8328	0.8696	0.7003	0.9127	0.5984

Table 5: Model Performance

Our ROC curve as shown in Figure 2 indicates that both our models were similarly performing, with Random forest having a slight edge. Once again, our improved pre-processing resulted in a much healthier ROC Curve. Both models performed to our expectations with this dataset, and given our extensive hyperparameter tuning and pre-processing which took the majority of our time for this project, we acknowledge that these aspects cannot be further optimized. For further research, we would suggest either acquiring additional data from rejected loans to to mitigate the class imbalance, or exploring different set of models.

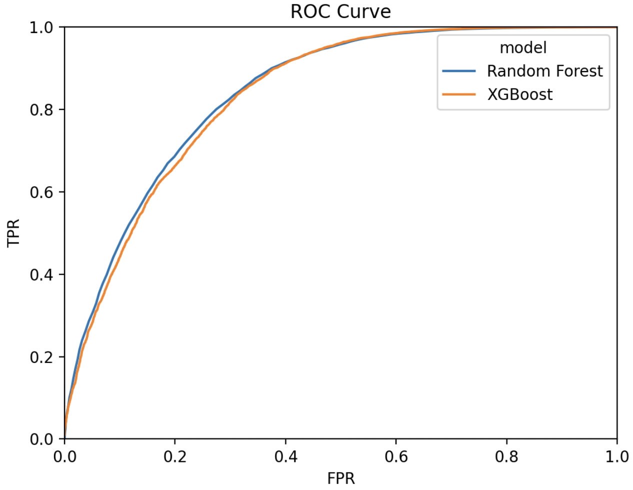

Figure 2: Receiver Operating Characteristic (ROC) Curve

4.3 Feature Importance

After evaluating our models, we decided to examine the top 20 most important features in predicting loan approval, which is shown in Figures 3 and 4.

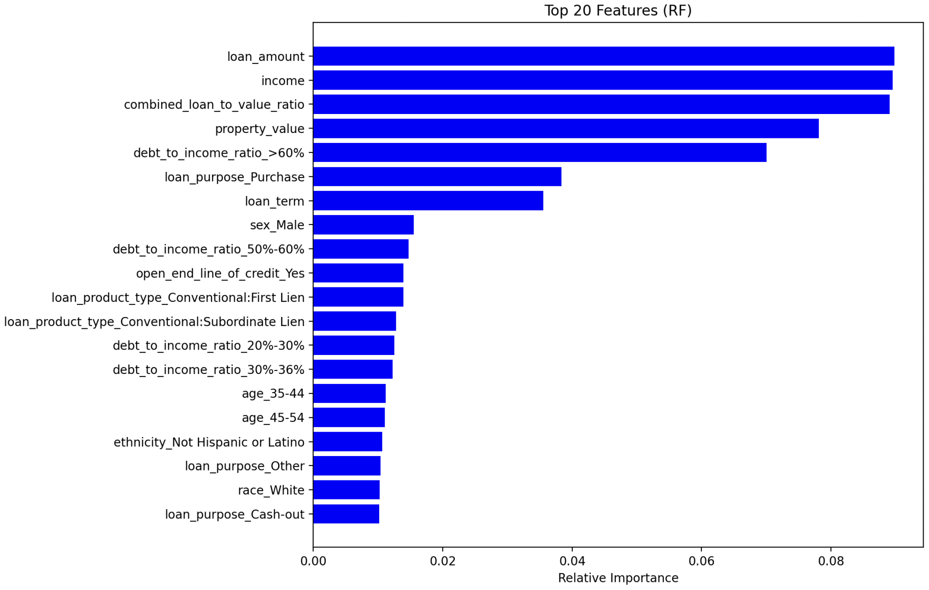

Figure 3: Top 20 Feature Importance in Random Forest

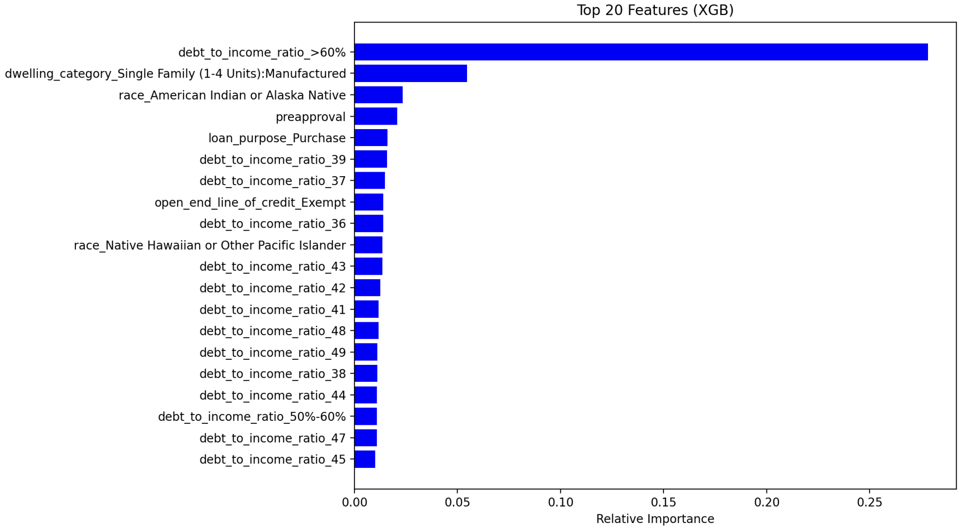

Figure 4: Top 20 Feature Importance in XGBoost

Notably, demographic attributes such as ‘race_American Indian or Alaska Native’, ’sex_Male’, and ‘ethnicity_Not Hispanic or Latino’ appear in the top 20 features. Both models heavily weigh factors related to mortgage like loan amount, income, debt-to-income ratio, and property value, which is logical and appropriate predictors for loan approval as they are directly related to a borrower's ability to repay the loan.

4.4 Fairness Evaluation

We proceeded to evaluation of fairness in our dataset and models, which was our primary goal in this project. We aimed to utilize a variety of statistical metrics assessing potential bias and fairness, including Disparate Impact, Statistical Parity Difference, Average Odds Difference, and Equal Opportunity Difference. Mathematical definition of these metrics and their interpretations are established as following.

4.4.1 Disparate Impact

$$\frac{Pr(\hat{Y}=1|D=unprivileged)}{Pr(\hat{Y}=1|D=privileged)} \tag{1}$$

Disparate Impact measures the ratio of positive outcomes between the unprivileged and privileged groups. A value of 1 indicates perfect fairness, implying that both groups have equal chances of receiving a positive outcome. Values less than 1 indicate fewer positive outcomes for the unprivileged group, suggesting potential discrimination.

4.4.2 Statistical Parity Difference

$$Pr(\hat{Y}=1|D=unprivileged) - Pr(\hat{Y}=1|D=privileged) \tag{2}$$

This metric measures the difference in the probability of a positive outcome between the unprivileged and privileged groups. A value of 0 indicates no disparity - both groups have equal probabilities of a positive outcome. Negative values indicate a higher probability of positive outcomes for the privileged group.

4.4.3 Average Odds Difference

$$\frac{1}{2}[(FPR_{D=unprivileged} - FPR_{D=privileged}) + (TPR_{D=unprivileged} - TPR_{D=privileged})] \tag{3}$$

This metric averages the difference in the true positive rates (sensitivity) and false positive rates between the unprivileged and privileged groups. A value of 0 indicates equal true positive and false positive rates for both groups. Negative values suggest better or more favorable outcomes for the privileged group.

4.4.4 Equal Opportunity Difference

$$TPR_{D=unprivileged} - TPR_{D=privileged} \tag{4}$$

Equal Opportunity Difference specifically looks at the difference in true positive rates between the unprivileged and privileged groups. It focuses solely on those instances that should receive a positive outcome. A value of 0 indicates that both groups have the same true positive rate. Negative values indicate a bias favoring the privileged group.

4.4.5 Dataset Fairness Metrics

Before evaluating fairness in our models, we proceeded to assess potential biases present in the dataset. This step was done by binarizing different demographic groups into privileged and unprivileged groups as we mentioned in the Protected Attributes section. The metrics used were disparate impact and statistical parity difference, and the result can be seen in Table 5.

Metrics	Disparate Impact	Statistical Parity Difference
Ethnicity	0.94	-0.05
Race	0.84	-0.12
Sex	0.98	-0.02
Age	0.95	-0.04

Table 6: Dataset Fairness Metrics

We set our benchmark for significant biases against unprivileged groups as 0.9 for disparate impact, which means that unprivileged group had 10% point lower probability of receiving a positive outcome (loan approval, in this case), and -0.1 for statistical parity difference. Based on the result in Table 5, with regard to ethnicity, there was minimal bias against unprivileged group (Hispanic or Latino) with a disparate impact value close to 1 (e.g., 0.94) and the statistical parity difference of -0.05. When it comes to race, things were a lot different since the metrics suggested significant bias against the unprivileged group with a disparate impact value of 0.84 and the statistical parity difference of -0.12. The metrics for sex show very minimal bias. Indeed, they showed the most equality among privileged and unprivileged groups across all protected attributes. Bias with regard to age was also fairly minimal. Its metrics did not exceed our benchmark level for significant biases.

4.4.6 Model Fairness Metrics

It is crucial to examine whether machine learning models, which are getting used more and more by mortgage lenders nowadays, would perpetuate or amplify biases present in the dataset. We decided to use the best performing model above, which was XGBoost, to assess fairness in machine learning classification. As mentioned above, we employed four statistical metrics, including Disparate Impact, Statistical Parity Difference, Average Odds Difference, and Equal Opportunity Difference. The result is shown in Table 7.

Metrics	Disparate Impact	Statistical Parity Difference	Average Odds Difference	Equal Opportunity Difference
Ethnicity	0.93	-0.05	-0.03	-0.01
Race	0.77	-0.19	-0.15	-0.10
Sex	0.96	-0.03	-0.03	-0.00
Age	0.94	-0.04	-0.03	-0.01

Table 7: XGBoost Fairness Metrics

As in the dataset fairness metrics, we decided to set benchmark for significant biases against unprivileged group as 0.85 for Disparate Impact, -0.1 for Statistical Parity Difference and Average Odds Difference, and -0.05 Equal Opportunity Difference.

The evaluation of the XGBoost loan approval classification model based on various fairness metrics indicated varied performance across different protected attributes. For ethnicity, sex, and age, the model performed relatively well, maintaining a disparate impact close to 1 (0.93, 0.96, and 0.94, respectively) and minimal negative values in statistical parity difference, average odds difference, and equal opportunity difference, all within acceptable thresholds. This result suggests a relatively fair treatment of these groups or a minimal biases against these groups in terms of mortgage loan approval rates compared to the overall population.

However, the assessment for race revealed significant biases. The disparate impact was notably below the threshold of 0.85 at 0.77, which was indeed lower than that in the dataset. This metric not only indicates potential unfairness towards certain racial minority groups, but also highlights the fact that the machine learning classification model can perpetuate or worsen the biases against those groups. Furthermore, the statistical parity difference, average odds difference, and equal opportunity difference all exceeded negative thresholds, highlighting a pronounced disparity in loan approval chances across racial lines.

While the mortgage loan approval classification model demonstrates adequate fairness concerning ethnicity, sex, and age, it exhibits concerning biases against certain racial groups. Addressing these disparities will be crucial for ensuring equity in loan approvals across all demographic categories.

5 Discussion and Conclusion

Overall, we observed moderate to significant biases against certain demographic groups in mortgage lending. Even a moderate level of bias can lead to significant real-world issues, particularly for individuals who depend on fair lending practices. It is critical to explore methods to mitigate these biases effectively. Our suggestion for the mitigation consists of three parts: Preprocessing, Inprocessing, and Postprocessing. Preprocessing techniques such as reweighing and optimized preprocessing (OP) play a crucial role by adjusting the weights of training instances and modifying labels and features, respectively, to ensure fairness prior to model training. Additionally, inprocessing techniques like adversarial debiasing, which simultaneously aims to maximize prediction accuracy and minimize adversarial loss, contribute to reducing bias. Postprocessing methods, such as equalized odds postprocessing, further refine the model by adjusting the decision boundary to balance false positive and negative rates among groups. Alongside these techniques, conducting thorough exploratory data analysis and visualization remains essential to understand and address the underlying biases in the data comprehensively.

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.

A Contributions

- Hunter: Model Training and Evaluation, Hyperparameter Tuning, Editing of Report
- Jihan: Data Preprocessing, Fairness Evaluation, Compilation and Editing of Report

- Vidhi: Correlation Analysis, SMOTE, Model Training and Evaluation, Editing of Report

B Code and Dataset

B.1 Code

- GitHub Repository: cs-334-project

B.2 Dataset

- HMDA Data Website
- HMDA Data Publication
- 2022 Public Loan/Application Records(LAR)
- Data Publications (Public HMDA - LAR Data Fields)

B.3 Additional Toolkit

- AI Fairness 360