Dataset

The data in this study is collected under the Home Mortgage Disclosure Act (HMDA) from all lenders above a size threshold that are active in metropolitan areas (estimated to cover 90% of first-lien mortgage originations in the United States). It is not within the scope of this paper to perform any empirical analyses on the presence of discrimination in mortgage lending decisions. There is insufficient information in the public data set to draw such conclusions.

The dataset lists different outcomes for a loan application: originated (accepted), withdrawn, closed due to incompleteness, denied, etc. Our main focus is on originated versus denied loans. These decisions are made by financial institutions, and in most we have information on the grounds of why they deny an application: ex. not enough collateral, debt-to-income ratio too high, credit history, etc. However we do not have information on why the loan originated specifically (in this case, why the loan was accepted i.e. because of the credit score, etc.), or the final outcome of the loan, i.e. whether or not the individual defaulted on the loan. While this would limit the internal validity of any analysis assessing the drivers of loan approval, as described above, this should not preclude the demonstration of methodologies.

The approval rate is 75.6% in the full data set; however, a representative sample is not needed, as it is not the aim to draw any empirical conclusions about the drivers of approval. The following steps were taken to facilitate the analysis:

- HMDA data from 2011 are used. Since the borrowers in 2011 experienced an overall increase in house prices, any default risk is likely specific to the borrower rather than reflective of a macroeconomic shock (Fuster et al. 2017);
- Loan records with missing data are removed, given they are from exceptional situations only.
- We subsampled the dataset, and created two datasets:
    - `mortgage_data:` 100,000 originated loans and 100,000 denied loans are randomly sampled without replacement to avoid issues arising from the data imbalance and to facilitate the interpretation of accuracy metrics.
    - `mortgage_data_balanced:` 100,000 originated loans and 100,000 denied loans are randomly sampled without replacement.

  the first subsample preserves as close as possible the original imbalance observed in the full sample; the second one was created to avoid issues arising from the data imbalance and to facilitate the interpretation of accuracy metrics. It also resembles the data used in Lee and Floridi (2020). In both, we tried to preserve as much as possible the original imbalances observed across the protected attributes and the statistical properties of the remaining features.

Both subsets of data can be downloaded from this Github repository: https://github.com/askoshiyama/audit_mortgage. The full data can be obtained here: https://www.consumerfinance.gov/data-research/hmda/historic-data/. Below we present some descriptive statistics on the `mortgage_data_balanced` dataset, highlighting particularly the outcome variable and protected characteristics. To do that, we used the Jupyter

notebook at https://github.com/askoshiyama/audit_mortgage/blob/master/Descriptive%20Stats.ipynb.

Table 1. Descriptive statistics of the protected characteristics in the balanced dataset.

		Balanced	
Features	Group	Originated	Denied
Ethnicity	Hispanic or Latino	3.97%	2.69%
	Not Hispanic or Latino	46.03%	47.31%
Race	Asian	2.58%	3.13%
	Black or African American	5.31%	2.20%
	White	41.41%	44.26%
	American Indian or Alaska Native	0.47%	0.25%
	Native Hawaiian or Other Pacific Islander	0.23%	0.16%
Gender	Female	16.08%	13.33%
	Male	33.92%	36.67%

Modelling

We are interested in estimating the probability of the event that an applicant will have their mortgage denied. The mathematical setup is as follows: given an algorithm $\hat{P}(y = Denied) = f_\theta(x)$ that maps information x from an applicant to a probability, our goal is to learn the optimal parameter set θ for f_θ by minimizing the loss function $L(y, f_\theta(x))$. For example, f_θ could represent a Logistic Regression model; θ the coefficients; x loan amount, application income, etc.; and $L(y, f_\theta(x))$ the cross-entropy risk.

More specifically, below we list the variables used as the input vector x, closely following the study of Lee and Floridi (2020). For more information about the available variables in the full data available by HMDA please refer to: https://files.consumerfinance.gov/hmda-historic-data-dictionaries/lar_record_codes.pdf.

Table 2. Features used in the input vector.

Type	Features
Categorical variables	agency_abbr, owner_occupancy_name, property_type_name, loan_purpose_name, loan_type_name
Non-categorical	loan_amount_000s, applicant_income_000s, population,

variables	minority_population, hud_median_family_income, tract_to_msamd_income, number_of_owner_occupied_units, number_of_1_to_4_family_units

All predictors in the feature set were used, with the exception of race, sex, and minority population of the census tract area (Seng-Lee and Floridi, 2020, Ladd 1998). There are important studies showing that the inclusion of these features may result in both fairer and more accurate outcomes (Kleinberg et al. 2018). The protected characteristics are only used to estimate the fairness metrics. From a model validation perspective, we reported the results in the test set based on a ten-fold cross validation.

As an example, below we present the results of a Logistic Regression model by computing performance and bias (gender) metrics using the balanced dataset. Despite we did not have access to the data used in Lee and Floridi (2020), our results are very close to the ones obtained by these authors. These results can be replicated using the Jupyter notebook at https://github.com/askoshiyama/audit_mortgage/blob/master/Modelling.ipynb.

Table 3. Performance and bias metrics for Logistic Regression in the test set.

Type	Metric	Value
Performance	AUC	0.5753
	Accuracy	0.5561
	Brier	0.2471
	F1-Score	0.5201
	Precision	0.5660
	Recall	0.4810
Bias (Male vs Female)	2-SD Rule	14.26
	Average Odds Difference	0.0018
	Cohen-D	0.2224
	Disparate Impact	1.2786
	Equal Opportunity Difference	0.1040
	Statistical Parity	0.1094

Experiment 1: generating biased datasets