Fairness in Mortgage Approvals: An Optimization Approach

Kiran Gite, John Lazenby
15.095 Final Project, Fall 2020

Abstract

As a result of the Fair Housing Act of 1968, financial institutions must take great care to avoid considering protected characteristics like race, ethnicity, age, and gender when determining to whom to underwrite a mortgage. However, hundreds of years of economic discrimination against minorities, particularly Black households, has resulted in racial wealth gaps. We would like to investigate the tradeoffs between fairness and profit that banks face when giving mortgage loans.

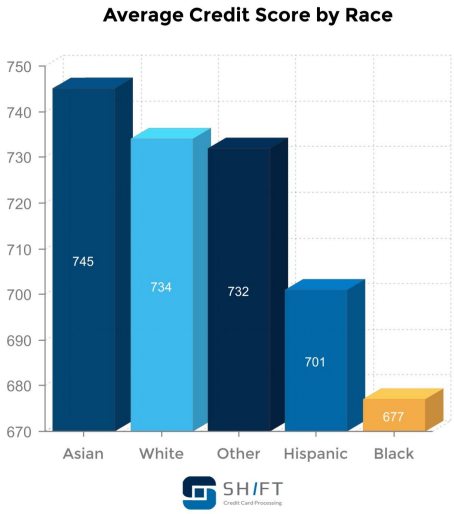
We obtain and clean HMDA data containing demographic and financial information about mortgage applicants for Flagstar Bank in 2019, generating a feature indicating whether an applicant is Black or not. We take an optimization approach to experiment with variations on classification methods to predict denials for Black and non-Black mortgage applicants. We add different types of linear and mixed-integer constraints to support vector machine formulations in order to obtain “fair” models. We compare performance metrics (AUC, negative predictive value) and fairness metrics (denial rate disparity, false positive disparity, false negative disparity) between fair models and baseline models. We see that fair models can approve loans for qualified minority applicants who would unfairly be excluded from standard models, increasing a bank’s profit from minority applicants by 46% over the baseline. We conclude that fairness in models can be achieved without significantly impacting model performance. Future work includes faster solution methods for our mixed-integer formulations, and considering the interactions of different types of bias in consumer finance.

Problem Statement

The Fair Housing Act of 1968 states that it is illegal to discriminate in the sale, rental, and financing of residential dwellings. As a result of this law, financial institutions must take great care to avoid considering protected characteristics like race, ethnicity, age, and gender when determining to whom to underwrite a mortgage. We are interested in investigating how these fair lending restrictions impact a financial institution's ability to make accurate predictions for which applicants to deny and which to approve. To do so we take an optimization approach to experiment with variations on classification methods to predict denials for Black and non-Black mortgage applicants. We observe how different constraints impact performance and fairness relative to baseline methods.

Fairness in Consumer Finance

The United States has a history of economic discrimination against minorities. For example, the practice of "redlining" involved rejecting mortgage applications not because of the financial status of applicants, but solely because the mortgages were for homes in "high-risk" Latino and Black neighborhoods [1]. This specific form of systemic discrimination like this was outlawed by the Fair Housing Act, but its legacy is very much present today, in the form of some minority households continuing to have reduced wealth and assets compared to non-minority households as a result of hundreds of years of discriminatory laws and practices. The figure on the right shows a simple analysis done by Shift Credit Card processing that shows the average FICO score (popular credit score that is meant to be a general measure of a borrower's ability to repay) by different race categories [2]. Asian and White borrowers have dramatically higher scores than Hispanic and Black borrowers.



When determining to whom to extend credit, financial institutions need to make a judgement about a borrower's ability to repay. As mentioned above, FICO is one way to do this. Additional factors that may indicate a borrower's ability to repay are their income relative to the loan payment they will be making (debt to income ratio) and the loan size relative to the value of collateral (loan to value ratio) among others. An ill-intentioned banker eager to determine whether a borrower is likely to repay may base their decision solely on the borrower's race, and sadly because of the correlation between race and financial circumstance this may be a good decision from a purely profit seeking point of view. A more common scenario could arise when a well-intentioned banker finds themselves treating racial minorities unfairly even if they do not consider an applicant's race as an input to their decision. This is because many characteristics that indicate ability to repay are heavily correlated with race. We seek to investigate and better understand this push and pull that exists between fairness and profit seeking in consumer finance.

HMDA Data

In our attempt to mimic the scenario a banker might face when making a loan approval or denial decision, we use HMDA data. Ideally, we would have the exact factors a bank receives when an applicant applies for a loan and historical information on delinquency. However, this kind of data is kept secret by financial institutions, so HMDA data is a solid alternative. The HMDA data for 2019 contains mortgage applications for 5,508 financial institutions in the US for a total of 15.1 million applications [3]. The data contains features related to the proposed loan itself, the borrower's financial situation, and demographic characteristics of the borrower(s), as well as whether a loan was approved or denied (our response variable). The full list of features can be found in reference [4].

One limitation of this data is the lack of information on applicant credit characteristics. Credit score is an essential measure any banker would have available, so our models may have less accuracy than financial institutions' real models. Another limitation is that we do not have any data on loan defaults, so we do not know whether the bank correctly approved or denied loans. In the absence of this information, we must make a significant assumption that the financial institution made the correct approval decisions in our dataset. We acknowledge the possibility that these historical approval decisions are faulty or contain bias.

The HMDA data is complex, so we decided to focus on Flagstar Bank's 2019 applications. Flagstar is one of the largest mortgage lenders in the USA and received 105,354 applications in 2019. Feature engineering involved extracting races and ethnicities of borrowers and co-borrowers and removing all variables containing data filled in after loan approval or denial. Since Black applicants have the highest rates of loan application denial in our dataset (Appendix A, Figure A.2), we want to focus on the discrepancies between Black and non-Black applicants, so we created a new feature indicating whether a borrower or co-borrower is Black or not. More details on data cleaning can be found in Appendix A.

Methods and Results

We prepared our data by randomly selecting 80% of points for a training set and 20% for a testing set. We performed undersampling to balance our training set, since our models performed poorly on the imbalanced data. Even after undersampling, the final training set contained over 32,000 rows. After one-hot-encoding categorical variables, we obtained 126 features that we used to classify whether a loan application was approved (negative outcome) or denied (positive outcome). Our prediction models are intended to mimic a model that a financial institution would use, taking the HMDA data as the truth about who should receive a loan and who should not. For more details on final features used, see Appendix B.

We added another tweak to the process of obtaining model predictions based on the assumption that Flagstar has a fixed number of loans they must give out annually and cannot go above or below that number. After a model gives predictions for approvals and denials, only a certain number of applications can actually be approved because of this constraint. For each method, we adjusted the threshold that determined whether a loan was approved or denied so that the overall denial rate matches Flagstar's true loan denial rate in 2019.

We will train different prediction models to try to achieve multiple interpretations of "fairness" and evaluate performance in model results with the following metrics:

- AUC: A higher AUC indicates the model is better able to rank borrowers in terms of their credit risk. A high value is good for the profitability of a bank.
- Overall Negative Predictive Value (NPV; True Negatives/(True Negatives + False Negatives)): Because the bank extends the same number of loans each time, this rate is the percentage of those loans they extend that do not default. The higher it is, the more profit the bank receives.
- False Positive Rate of Minority Borrowers vs. Majority Borrowers: This rate represents the percentage of borrowers that were incorrectly rejected. If this rate is higher for minorities borrowers relative to majority borrowers it could indicate unfairness.
- False Negative Rate of Minority Borrowers vs. Majority Borrowers: This rate represents the percentage of unqualified borrowers that were approved. If this rate is higher for majority borrowers relative to minority borrowers it could indicate unfairness.
- Denial rate among minorities: The lower this rate the better the outcomes are for minority applicants. If rates are different between groups this could indicate unfairness.

We use logistic regression and support vector machines as classifiers. We use SVM as a basis for our fair models, as SVM can be formulated as a linear program, allowing us to easily add additional fairness constraints. We chose not to include regularization in our SVM models for the sake of simplicity, although that is an avenue for future work. All logistic regression models were generated with the glm function in R; all SVM models were formulated as optimization models in Julia with JuMP and solved using Gurobi 9.0.

Baseline Method 1 (Explicitly Discriminatory): Unconstrained SVM and Logistic Regression with Minority Status as a Feature

Below are the results for our initial baseline methods: standard SVM and logistic regression. For our first two baseline methods we include an indicator of whether the applicant is Black in the set of features (referred to as the minority feature from here onwards). IS means in-sample and OOS means out-of-sample in all below tables.

Method	AUC	Overall NPV (%)	FPR (%)			FNR (%)			Minority Denial Rate (%)
			Min.	Maj.	Diff.	Min.	Maj.	Diff.	
SVM w/ Minority Feature, IS	0.710	58.2	38.0	8.9	29.1	41.2	66.7	-25.5	51.3
SVM w/ Minority Feature, OOS	0.713	86.0	84.9	13.9	71.0	7.3	59.0	-51.7	87.4
Logistic w/ Minority Feature, IS	0.715	58.3	35.9	8.9	27.0	43.0	66.4	-23.4	49.4
Logistic w/ Minority Feature, OOS	0.717	86.0	54.2	15.4	38.8	29.1	56.6	-27.5	59.6

These results clearly demonstrate unfairness for the minority applicants. Both models deny >50% of minorities out-of-sample. False positive rates are significantly higher for minority applicants, showing that this potential denial model is biased against such applicants. SVM and logistic regression models perform similarly in sample and SVM is much more biased out of sample. By including minority status as a feature in the model, the model seems to be incorrectly learning that minority status is what determines loan worthiness, as opposed to applicants' financial characteristics, and thus treating minority candidates unfavorably. This is the kind of model an ill-willed discriminating banker might use and is definitely illegal.

Baseline Method 2: Unconstrained SVM, Logistic Regression without Minority Feature

For our next set of baseline models we once again use standard logistic regression and support vector machines but this time do not include minority status as a feature.

Method	AUC	Overall NPV (%)	FPR (%)			FNR (%)			Minority Denial Rate (%)
			Min.	Maj.	Diff.	Min.	Maj.	Diff.	
SVM No Minority Feature - IS	0.709	58.1	21.7	10.0	11.7	60.7	65.0	-4.2	32.9
SVM No Minority Feature - OOS	0.710	85.7	27.7	17.1	10.6	52.7	55.5	-2.8	34.0
Logistic No Minority Feature - IS	0.714	58.2	22.1	9.7	12.4	57.1	65.1	-8.0	35.4
Logistic No Minority Feature - OOS	0.714	85.9	33.1	16.6	16.5	50.3	54.9	-4.6	38.4

These results appear to be more fair than the previous baseline results. The minority denial rate decreases and the gaps in false positive rate and false negative rate are more favorable for minority borrowers. Surprisingly, the performance measured in both AUC and overall net predictive value stays about the same. Seeing these results, a banker has no justification to be explicitly discriminatory because they can increase fairness without sacrificing any performance. However, this does not make the results fair. Minority borrowers still face much higher false positive rates and lower false negative rates relative to majority candidates.

Optimization-based Fairness Method 1: Holistic Constrained SVM

The first fairness method we implemented is a restriction on the percentage of total loss that can come from non-minority (majority) training observations. Our theory is that the previous baselines are approving majority borrowers by default. If we force models to place more scrutiny on majority borrowers, the results will be more favorable for minority borrowers. We add a constraint requiring that the percentage of total loss coming from majority borrowers multiplied by some constant hyperparameter has to be less than or equal to the percentage of majority applicants in the training set. We tried values of 1.25, 15, 2, and 5 for the hyperparameter without meaningfully different results.

Formulation with hyperparameter = 1.5:

Let the training data set have n observations and p features

$majority$ is an indicator that comes from our data and is 1 when the i^{th} training observation is a majority and 0 otherwise

$MajorityCount$ is the total number of majority applicants in the training data set.

Minimize $\sum_{i=1}^{n} t_i$

Subject to:

$$0 \leq t_i \ \forall \ i \ \in \{1, ..., n\}$$

$$1 - y_i * (\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \leq t_i \ \forall \ i \ \in \{1, ..., n\}$$

$$n * 1.5 * \sum_{i=1}^{n} majority_i * t_i \leq MajorityCount * \sum_{i=1}^{n} t_i$$

Method	AUC	Overall NPV (%)	FPR (%)			FNR (%)			Minority Denial Rate (%)
			Min.	Maj.	Diff.	Min.	Maj.	Diff.	
Constrained SVM - IS	0.705	57.8	21.8	10.5	11.3	62.2	65.3	-3.1	32.0
Constrained SVM - OOS	0.705	85.4	25.5	17.5	8.0	57.1	56.5	0.6	31.1

This method does not perform as well performance-wise as the previous model. The AUC and overall negative predictive value both decrease. This training set and test results for this method show improved fairness. In the test set the false positive rate difference decreases from 11.3 to 8.0, the false negative rate difference is actually favorable for minorities and the minority denial rate decreases from 32.0 to 31.1.

Optimization-based Fairness Method 2: Two Step Holistic Constrained SVM

The second fairness method comes at a similar angle to the first and acknowledges the likely trade off that occurs between accuracy and emphasis on a particular class. This method takes as an input the objective value from the unconstrained original SVM formulation. The objective function now only considers loss from majority applicants but there is an additional constraint stipulating that the combined loss from both minority and majority applicants cannot exceed the unconstrained SVM by a constant factor (a hyperparameter). We tried values of 1.25, 2, and 5 without meaningfully different results.

Formulation with a hyperparameter of 1.5:

Let the training data set have n observations and p features

$majority_i$ comes from our data and is 1 when the i^{th} training observation is a majority and 0 otherwise

$PreviousLoss$ is the objective value of the unconstrained SVM using the same training data.

Minimize $\sum_{i=1}^{n} majority_i * t_i$

Subject to:

$0 \leq t_i \ \forall \ i \in \{1, ..., n\}$

$1 - y_i * (\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \leq t_i \ \forall \ i \in \{1, ..., n\}$

$\sum_{i=1}^{n} t_i \leq 1.5 * PreviousLoss$

Method	AUC	Overall TNR (%)	FPR (%)			FNR (%)			Minority Denial Rate (%)
			Min.	Maj.	Diff.	Min.	Maj.	Diff.	
Two Stage SVM - IS	0.705	57.8	21.7	10.5	11.2	62.2	65.3	-3.1	32.0
Two Stage SVM - OOS	0.705	85.4	25.3	17.5	7.7	57.3	56.5	0.9	30.8

The results here are very similar to the first fairness method. This is likely because their motivations are so similar. The minority denial rate decreases slightly more in this method relative to this previous one but they are virtually identical. Minority denial rate and the gaps in false positive rate and false negative rate actually improve for out-of-sample testing.

Optimization-based Fairness Method 3: Constraining Approval Rates

This method interprets "fairness" as "having similar outcomes across groups". We would like to ensure that mortgage approval rates are similar for minorities and non-minorities. To incorporate this into an SVM formulation, we use binary variables z_i equaling 1 if applicant i is approved and 0 otherwise. The full formulation is below:

Let K be the set of indices of non-minorities: $K = \{i \mid minority_i = 0, \forall\ i \in 1...n\}$

Let J be the set of indices of minorities: $J = \{i \mid minority_i = 1, \forall\ i \in 1...n\}$

Let M be a big-M constant that is bigger than all possible loss values

Let ε be the largest acceptable gap between minority and nonminority approval rate

Minimize $\sum_{i=1}^{n} t_i$

Subject to:

$1 - y_i * (\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \leq t_i \quad \forall\ i \in \{1, ..., n\}$

$t_i \geq 0 \quad \forall\ i \in \{1, ..., n\}$

$(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \leq M * (1 - z_i) \quad \forall\ i \in \{1, ..., n\}$ (pred. > 0 → denial → $z_i = 0$)

$(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \geq -M * z_i \quad \forall\ i \in \{1, ..., n\}$ (pred. < 0 → approval → $z_i = 1$)

$\frac{1}{|K|} \sum_{i \in K} z_i - \varepsilon \leq \frac{1}{|J|} \sum_{i \in J} z_i \leq \frac{1}{|K|} \sum_{i \in K} z_i + \varepsilon$ (minority approval rate within ε of non-minority rate)

$z_i \in \{0, 1\} \quad \forall\ i \in \{1, ..., n\}$

Due to the addition of over 30,000 binary variables and big-M constraints, the runtime of this model is extremely large and the model cannot reach optimality in 24 hours. By setting the MIP gap in our solver to be 2%, we obtained model results in 7 hours that are within 2% of the true optimal solution. Below are results with $\varepsilon = 0.4,\ M = 2,000$:

Method	AUC	Overall NPV (%)	FPR (%)			FNR (%)			Minority Denial Rate (%)
			Min.	Maj.	Diff.	Min.	Maj.	Diff.	
SVM Approval Constraint - IS	0.708	58.2	24.5	9.7	14.8	56.5	65.2	-8.7	36.6
SVM Approval Constraint - OOS	0.710	85.8	38.1	16.4	21.6	44.0	55.9	-11.9	43.8

This model appears to perform slightly better in terms of AUC and overall net predictive value, but its fairness is not as good; gaps in false positive rate and false negative rate are higher, as is the minority denial rate. It is possible that the allowable gap of $\varepsilon = 0.4$ was too large to put any real fairness constraints on the problem. It is also possible that trying to equalize approval rates does not directly impact how well the model is able to separate qualified and nonqualified applicants based on financial characteristics.

Optimization-based Fairness Method 4: Constraining Gap in False Positive Rate

This method interprets "fairness" as "having similar error rates across groups". We specifically ensure similar false positive rates for minorities and non-minorities. Let the binary variable f_i equal 1 if applicant i is a false positive ($y_i = -1$ and loss $(\beta_0 + \beta_1 x_2 + ... + \beta_p x_p)$ is positive), and 0 otherwise. The formulation is below:

Let $K = \{i \mid minority_i = 0 \quad \forall\ i \in 1...n\}$; Let $J = \{i \mid minority_i = 1 \quad \forall\ i \in 1...n\}$

Let I be the set of "negative" minorities: $I = \{i \mid minority_i = 1, y_i = -1 \quad \forall\ i \in 1...n\}$

Let $H = \{i \mid minority_i = 0, y_i = -1 \quad \forall\ i \in 1...n\}$

Let M be a big-M constant larger than all loss values; Let ε be the largest acceptable gap between minority and nonminority false positive rate

Minimize $\sum_{i=1}^{n} t_i$

Subject to:

$1 - y_i * (\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \leq t_i \quad \forall\ i \in \{1, ..., n\}$

$t_i \geq 0 \quad \forall\ i \in \{1, ..., n\}$

$f_i \geq \frac{-y_i + (1/M) * (\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) - 1}{2} \quad \forall\ i \in \{1, ..., n\}$ (y < 0, pred. > 0 → f_i = 1)

$f_i \leq y_i + 2 + (1/M) * (\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \quad \forall\ i \in \{1, ..., n\}$ (y < 0, pred. < 0 → f_i = 0)

$f_i \leq M * (1 - y_i) \quad \forall\ i \in \{1, ..., n\}$ (y > 0 → f_i = 0)

$\frac{1}{|H|} \sum_{i \in K} f_i - \varepsilon \leq \frac{1}{|I|} \sum_{i \in J} f_i \leq \frac{1}{|H|} \sum_{i \in K} f_i + \varepsilon$ (minority FPR within ε of non-minority FPR)

$f_i \in \{0, 1\} \quad \forall\ i \in \{1, ..., n\}$

We choose $\varepsilon = 0.1,\ M = 2,000$ to obtain the results below. This formulation also required us to set a MIP optimality gap of 2% so that the solver would finish in less than 24 hours.

Method	AUC	Overall TNR (%)	FPR (%)			FNR (%)			Minority Denial Rate (%)
			Min.	Maj.	Diff.	Min.	Maj.	Diff.	
SVM FPR Constraint - IS	0.697	58.0	11.2	10.6	0.6	71.5	64.0	7.5	22.3
SVM FPR Constraint - OOS	0.697	85.8	21.3	17.4	3.9	62.5	54.5	8	26.5

This model gives interesting results, with the highest parity between false positive rates for minorities and nonminorities, and the lowest minority denial rates across all models. Qualified minority applicants are no longer being unfairly flagged as nonqualified. However, we see that the false negative rates are now biased towards minorities for the first time, meaning that a slightly higher proportion of nonqualified minority applicants are receiving loans. AUC is slightly lower than but close to AUC for other methods. The next step for this model would be to add constraints that correct the false negative rate gap.

While this false positive rate-based formulation is not efficient, it does attack error disparities more directly than our other methods. Future work in this area could be identifying tractable formulations as well as effective algorithms for solving such problems.

Results

Considering model performance measures, the "unfair" baseline models including the minority feature outperformed all fair models. However, these gaps are extremely small: the difference between the best and worst out-of-sample AUC is only 0.02, and the difference in

negative predictive value is 0.6%. We now compare our optimization-based fairness models to a "naive fairness baseline" of an SVM model trained without the minority feature.

Model (out-of-sample)	AUC	Overall NPV (%)	FPR Gap (%) (min. - maj.)	FNR Gap (%) (min. - maj.)	Minority denial rate (%)
SVM Approval Constraint	0.710	85.8	21.6	-11.9	43.8
Constrained SVM	0.705	85.4	8.0	0.6	31.1
Two Stage SVM	0.705	85.4	7.7	0.9	30.8
SVM FPR Constraint	0.697	85.8	3.9	8.0	26.5
Fair baseline	0.710	85.7	10.6	-2.8	34.0

Each of our methods beats the baseline in some metric (cells highlighted in green). Among our methods, Constrained SVM gives the best false negative rate gap, which means the model does not admit more unqualified borrowers from any particular group. SVM with the false positive rate constraint gives the smallest false positive rate gap and minority denial rate. We conclude that different types of fairness constraints do improve various model fairness metrics without significantly impacting model performance. For a full results table, see Appendix C.

Quantifying Model Impact

To investigate the financial impact of each model, we calculated a rough profit measure in our test set as the total monetary value of loans given to qualified candidates (true negative applicants) subtracted by the monetary value loaned to unqualified candidates (false negative applicants). Comparing profit between the baseline, the fair baseline, and Two-Stage SVM:

Model	Bank Profit = Good loan $ - Bad loan $ (USD)		
	Minorities	Non-minorities	Total
Baseline ("unfair")	0.065 billion	3.006 billion	3.071 billion
Fair baseline, % change over baseline	33.1%	-1.0%	-0.2%
Two Stage SVM, % change over baseline	46.6%	-1.1%	-0.1%

Considering profit from minorities, the Two-Stage SVM model beat the unfair baseline by 46.6% and the fair baseline by 13.5%. This means that our model is able to give loans to qualified minority applicants who would be unfairly rejected by other models. The downsides are that the profit from non-minorities drops by 1.1% and total profit drops by 0.1%. However, it would be up to bank leaders to decide if they would rather risk a loss or negative publicity and a lawsuit for violating the Fair Housing Act. Full table in Appendix D.

Conclusions and Future Work

Fairness in machine learning has many different definitions, and different constraints on ML problems can achieve various types of fairness without significantly impacting performance. Future work on the technical side of this problem could involve making models sparse for interpretability. Future work on the fairness side could involve investigating how gender or age discrimination can interact with race/ethnicity based bias. Finally, we must consider that no matter how fair our model is for a set of data, the data itself contains bias, as we may not be

able to remove the effects of hundreds of years of discrimination. It is possible that a truly fair model for predicting loan-worthiness may need to incorporate both quantitative and qualitative features, making this a fascinating and challenging machine learning problem.

Team Contributions

- John was responsible for data gathering, cleaning and preprocessing, formulating and coding linear models, running SVM baseline models, calculating model results and metrics
- Kiran was responsible for data visualization, formulating and coding mixed-integer models, calculating profit numbers, running logistic regression baseline models

References

1. https://www.cbsnews.com/news/redlining-what-is-history-mike-bloomberg-comments/
2. https://shiftprocessing.com/credit-score/
3. https://www.consumerfinance.gov/about-us/newsroom/ffiec-announces-availability-2019-data- mortgage-lending/
4. https://github.com/cfpb/hmda-platform/blob/master/docs/v2/spec/markdown/modified_lar/2019_Modified_LAR_Data_Dictionary.md
5. https://orwh.od.nih.gov/toolkit/other-relevant-federal-policies/OMB-standards

Appendices

Appendix A: Data acquisition and visualization

We downloaded our dataset from the HMDA data browser (link in reference [3]).

Our HMDA data contains:

- Information about the proposed loan: type, purpose, amount, term property location, lien status
- Information about the borrower and co-borrower's financial situation: income, cumulative loan to value ratio, and debt-to-income ratio
- Information about the demographic characteristics of the borrower and co-borrower: race, ethnicity, gender, and age
- Whether a loan was approved or denied by the financial institution

With a full list of fields in reference [4].

We performed standard data cleaning like dropping columns that had mostly null values and dropping rows with null values. We dropped county and tract information as these columns would have produced far too many dummy columns after one-hot encoding. We downloaded data for three banks (Bank of America, NFCU, and Flagstar) for 2018-2019 but ended up only using Flagstar 2019 data.

The HMDA data allows both borrower and co-borrower to report up to 5 different races and 5 different ethnicities. We combined these field into a race/ethnicity characterizations of Hispanic, American Indian, Asian, African American, Native Hawaiian, and Non-Hispanic White according to the office of management and budget standards by considering an applicant to belong to the category if either borrower or co-borrower reported membership to a race in any of the 5 fields [5]. We decided to focus on the fairness impact on African American borrowers because they have the highest denial rate in our data (Figure A.2 below).

Before beginning the modeling process we took care to remove target leaking variables. These variables are those that are determined after the decision on the loan has already been made. They can not be considered as inputs to our models because they would not have been available to a loan officer at the time the decision was being made. The most obvious example of a variable like this is "Reason for Denial" where a value of "Not Applicable" corresponds perfectly with approval. Other examples include the fields preapproval, rate spread, HOEPA, total cost, origination charges, and others.

The following visualizations are all based on our initial HMDA pull (years 2018-2019, 3 banks).

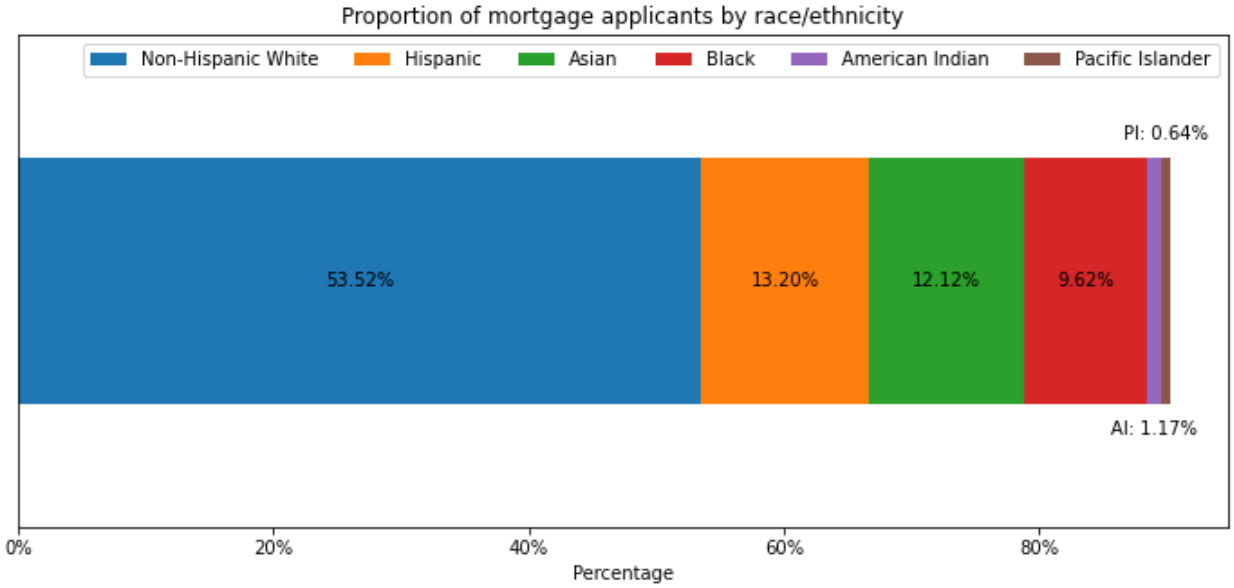


Figure A.1: Proportion of total applicants by race/ethnicity

There are significant overlaps in the data's racial/ethnic categories (they are not mutually exclusive). We will consider any applicants that indicate themselves as Black as minorities, and any applicants that do not indicate themselves as Black as non-minorities.

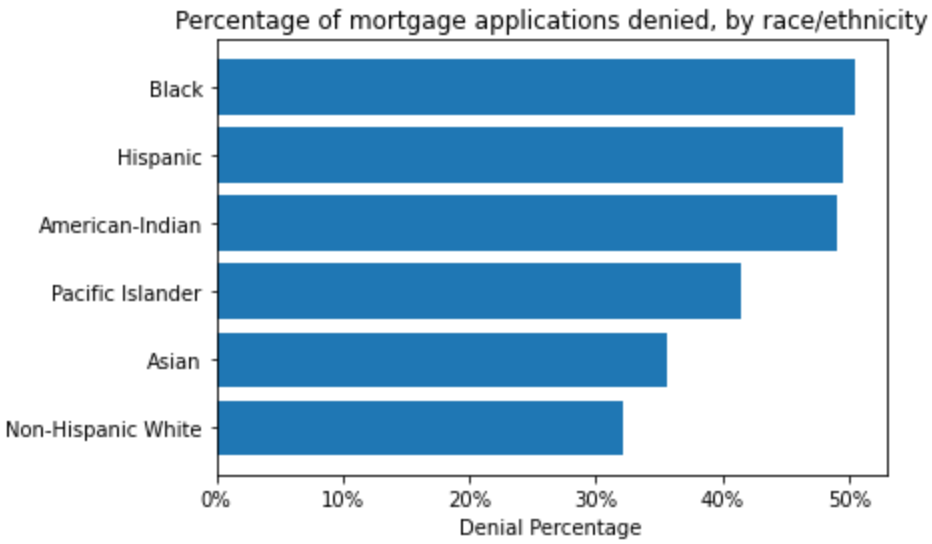


Figure A.2: Percentage of applications denied by race/ethnicity

Black applicants have the highest mortgage denial rates, almost 20% higher than non-Hispanic white applicants.

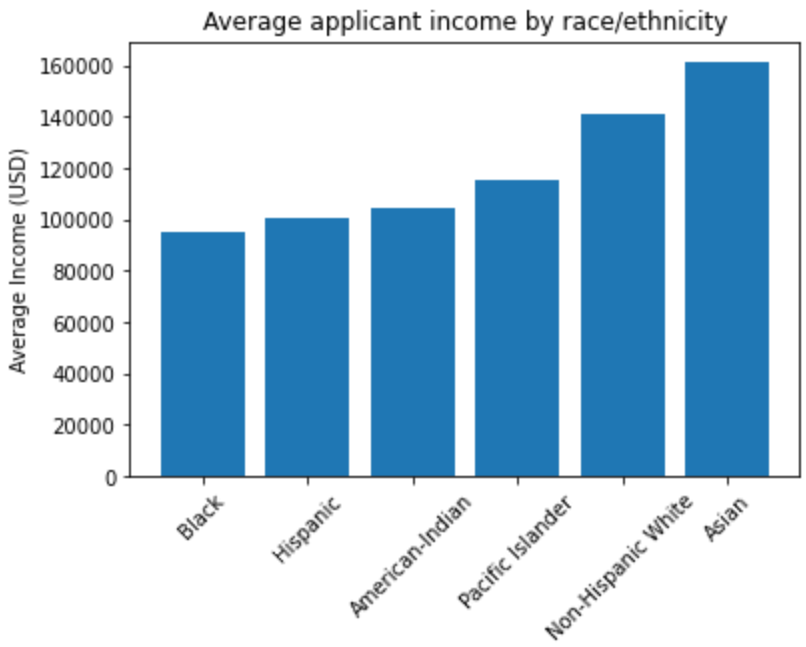


Figure A.3: Income by race/ethnicity

Here, we see strong correlations between race and average income, which are somewhat similar to the denial rate patterns in Figure A.2. This is an example of a variable that is correlated with race/ethnicity, so simply removing a race/ethnicity feature from training data may not be enough to ensure fairness.

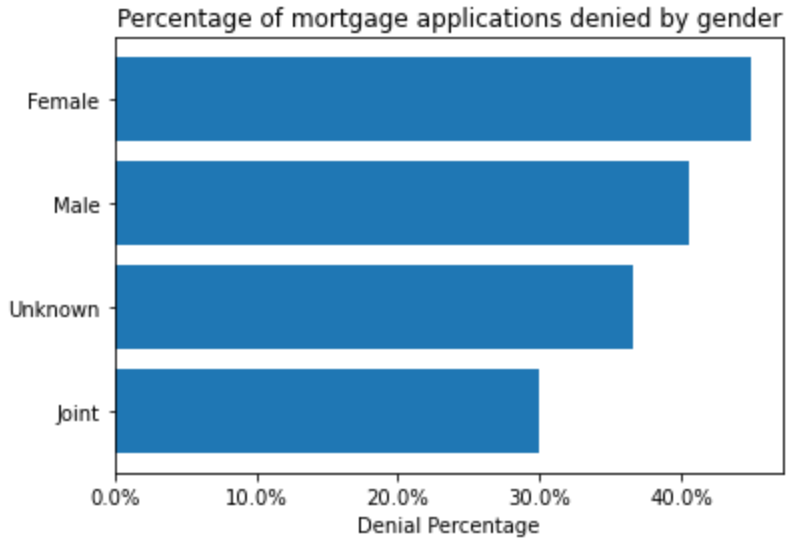


Figure A.4: Denial rates by gender on application

While we are only focusing on race/ethnicity in this project, other forms of bias are also present in our data. We can see that applicants who designate themselves as female have the highest mortgage denial rates. The fact that joint applications have the lowest denial rate could also be its own kind of bias against same-sex couples who are applying for home loans.

Appendix B: Preparation of data for modeling

We performed a random split on our cleaned Flagstar HMDA data, assigning 80% of observations to a training data set and 20% to a test data set. The overall dataset is imbalanced, with 20% of applications denied and 80% approved. We found that this imbalance was seriously hindering the model's performance. To remedy this we performed undersampling, randomly excluding approvals until there were even amounts of approvals and denials. After undersampling, there were 32,028 observations in the training set. We did not undersample for the testing set to test on realistic data. The testing set contained 20,072 observations. We one-hot encoded each categorical column to obtain one dummy column for each category. The full list of columns used for modeling is below.

Loan amount (numeric)	State (50 categories)	Manufactured Home Secured Property Type (2 categories)
Income (numeric)	Lien status (2 categories)	Manufactured Home Land Property Interest (2 categories)
Combined loan-to-value ratio (numeric)	Borrower credit score type (6 categories)	Total units (4 categories)
Property value (numeric)	Coborrower credit score type (7 categories)	Submission type (2 categories)
Loan type (4 categories)	Debt-to-income ratio (19 buckets)	Open-End Line of Credit (2 categories)
Loan purpose (6 categories)	Loan term (3 buckets)	Business purpose (2 categories)
Construction method (2 categories)	Interest-only payments (2 categories)	Minority (binary, generated from race/ethnicity features)
Occupancy (3 categories)	Other non-amortizing features (2 categories)	Denied (binary, the response variable we are trying to predict)

Prior to modeling, the data was normalized, as that is the best practice for SVM and logistic regression.

Appendix C: Full results

Table C.1: Relevant in-sample and out-of-sample metrics for all models. “Unfair” baseline models are highlighted in gray, fair baseline models are in yellow, and optimization-based fairness models are in white.

Method	AUC	Overall TNR (%)	FPR (%)			FNR (%)			Minority Denial Rate (%)
			Min.	Maj.	Diff.	Min.	Maj.	Diff.	
SVM Includes Minority Feature - Training Set	0.710	58.2	38.0	8.9	29.1	41.2	66.7	-25.5	51.3
SVM Includes Minority Feature - Test Set	0.713	86.0	84.9	13.9	71.0	7.3	59.0	-51.7	87.4
Logistic Includes Minority Feature - Training Set	0.715	58.3	35.9	8.9	27.0	43.0	66.4	-23.4	49.4
Logistic Includes Minority Feature - Test Set	0.717	86.0	54.2	15.4	38.8	29.1	56.6	-27.5	59.6
SVM Includes Minority Feature - Test Set as if All White	0.710	85.8	29.1	16.9	12.2	51.1	55.3	-4.2	35.5
Logistic Includes Minority Feature - Test Set as if All White	0.714	85.9	32.6	16.7	15.9	51.1	54.9	-3.9	37.8
SVM No Minority Feature - Training Set	0.709	58.1	21.7	10.0	11.7	60.7	65.0	-4.2	32.9
SVM No Minority Feature - Test Set	0.710	85.7	27.7	17.1	10.6	52.7	55.5	-2.8	34.0
Logistic No Minority Feature - Training Set	0.714	58.2	22.1	9.7	12.4	57.1	65.1	-8.0	35.4
Logistic No Minority Feature - Test Set	0.714	85.9	33.1	16.6	16.5	50.3	54.9	-4.6	38.4
Constrained SVM - Training Set	0.705	57.8	21.7	10.5	11.2	62.2	65.3	-3.1	32.0
Constrained SVM - Test Set	0.705	85.4	25.3	17.5	7.7	57.3	56.5	0.9	30.8
Two Stage SVM - Training Set	0.705	57.8	21.8	10.5	11.3	62.2	65.3	-3.1	32.0
Two Stage SVM - Test Set	0.705	85.4	25.5	17.5	8.0	57.1	56.5	0.6	31.1
SVM Approval Constraint - Training Set	0.708	58.2	24.5	9.7	14.8	56.5	65.2	-8.7	36.6
SVM Approval Constraint - Test Set	0.710	85.8	38.1	16.4	21.6	44.0	55.9	-11.9	43.8
SVM FPR Constraint - Training Set	0.697	58.0	11.2	10.6	0.6	71.5	64.0	7.5	22.3
SVM FPR Constraint - Test Set	0.697	85.8	21.3	17.4	3.9	62.5	54.5	8	26.5

The “Includes Minority Feature - Test Set as if All White” models are a combination of baseline methods 1 and 2 (including the minority feature in the training set, and artificially setting all observations as non-minorities in the testing set),

Appendix D: Profit calculations

Table D.1: Loan comparisons between models, testing dataset

	Bad loans (false negatives)			Bad loan amount (billion $)			Good loans (true negatives)			Good loan amount (billion $)			Profit (good - bad) (billion $)		
Model	Min.	Maj.	Total	Min.	Maj.	Total	Min.	Maj.	Total	Min.	Maj.	Total	Min.	Maj.	Total
Baseline	107	2051	2158	0.024	0.572	0.596	357	12940	13297	0.088	3.578	3.667	0.065	3.006	3.071
Fair baseline	185	1990	2175	0.041	0.560	0.601	522	12758	13280	0.127	3.538	3.665	0.086	2.978	3.064
Two Stage SVM	211	2046	2257	0.046	0.593	0.639	583	12615	13198	0.141	3.567	3.709	0.095	2.974	3.069

We see that while the two-stage SVM model gives slightly more bad loans to minorities, that is counteracted by the many more good loans given to minorities compared to the fair baseline. The two-stage SVM model also gives out a higher monetary amount of good loans (almost $42M more) than both the fair and unfair baselines.