Algorithmic Bias in Machine Learning Credit Scoring Models

Shashwat Sharma
Parkland High School
Allentown PA

Abstract

This study evaluates potential biases in machine learning models used for predicting mortgage loan approvals. Mortgage data was pulled from HMDA (Home Mortgage Disclosure Act) and was filtered to contain mortgage approvals, denials, and applications completely filled out for homes in Pennsylvania. The dataset includes anonymized information about the house being purchased, personal and financial information, community statistics, and more. Sensitive attributes of race, ethnicity, and sex were omitted from training to make sure the prediction models were not trained to discriminate based on those attributes. The predictions are then modeled in a confusion matrix to test possible bias based on the sensitive attributes. There was no significant bias found between the different machine learning algorithms of Linear Regression, Random Forest, XGBoost, and K-Nearest Neighbors. There was a statistically significant higher rate of false negatives for black applicants and a lower rate of false negatives for Asian applicants compared to the dataset's total proportion. We are unable to make conclusions on Native Americans or Alaskan Native and Native Hawaiian or Pacific Islanders due to a very small data size. We are also unable to find any statistically significant bias based on ethnicity or sex. Although race was not a factor used for training, attributes correlated with it may influence the prediction model’s decisions. AI poses a new threat to fairness in credit scoring and lending, a necessary factor in building wealth and home ownership. This research provides broader implications for identifying bias in other important sectors where AI automates decisions.

1 Introduction

The emergence of machine learning (ML) models has revolutionized the credit scoring industry by enabling the processing of large volumes of data to assess creditworthiness with unprecedented speed and efficiency. However, integrating algorithmic systems in financial services raises concerns about fairness and bias, with the potential for algorithms to perpetuate or even exacerbate existing inequalities. This is particularly concerning in mortgage lending, a critical area of financial stability and equality for many individuals and families. The Home Mortgage Disclosure Act (HMDA) dataset provides a comprehensive overview of mortgage lending activities in the United States. It has become a cornerstone for analysis and regulation in this domain (Consumer Financial Protection Bureau, 2021).

Studies have found significant disparities in loan denial rates and credit costs among African-American and Hispanic borrowers compared to their white counterparts (Rugh & Massey, 2010). Such disparities have tangible economic consequences and impede efforts toward achieving equitable access to credit (Quillian et al., 2020).

This research paper looks for potential bias in machine learning (ML) credit scoring models using Pennsylvania's HMDA (Home Mortgage Disclosure Data) dataset. We consciously exclude race, ethnicity, and sex from the training data to prevent embedding historical prejudices into our models. Post-training, our analysis incorporates these protected attributes to detect any biases in loan predictions. This dual approach allows us to assess whether ML algorithms, trained without sensitive demographic data, nonetheless perpetuate discrimination against certain groups. This paper advances that without careful consideration and adjustment, ML models may not only replicate but potentially intensify the discrimination faced by minority groups in mortgage lending.

The importance of this investigation cannot be overstated. Homeownership is a principal driver of wealth accumulation in the United States; thus, any bias in mortgage lending has long-lasting repercussions on wealth disparity (Shapiro et al., 2013). As of 2021, the homeownership rate among white Americans stood at 72.1%, compared to 49.3% among African Americans and 58.1% among Hispanic Americans, illustrating significant racial disparities (U.S. Census Bureau,

2021). Moreover, using machine learning in credit scoring impacts a substantial portion of the population, with approximately 63% of Americans relying on mortgages to finance their homes (National Association of Realtors, 2022).

This research aligns with the increasing demand for algorithmic accountability in financial services (European Union, 2019). It resonates with broader societal calls for justice and equity in critical decision-making processes.

2 Fairness of Machine Learning

The concept of fairness in machine learning (ML) is complicated and technically challenging. It involves many mathematical definitions and statistical properties, often reflecting broader social values and legal principles. Fairness cannot be condensed into a single metric or formula but is instead an ensemble of criteria that ML models might be required to satisfy, depending on the context of their application (Kleinberg et al., 2016). We navigate through some of these statistical fairness criteria and their mathematical structure, focusing on the equal odds property and exploring the nuances of true positives, true negatives, false positives, and false negatives in the context of credit scoring models.

2.1 Definition of Fairness and Equal Odds

Fairness in algorithmic decision-making can be understood as the absence of prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics. In mathematical terms, fairness is often translated into a set of constraints that an algorithm must satisfy. One of the most discussed fairness constraints in the literature is the equal odds property, which requires that a classifier's predictions are independent of the protected attributes (e.g., race, gender) given the actual outcome (Hardt et al., 2016). Mathematically, this can be expressed as:

$$P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = a')$$

for all groups a, a' in the protected attribute A, and for each outcome A in the actual label A, where A represents the predicted label

2.2 Disparate Impact and Error Rates

To understand fairness through error rates, it's crucial to define the terms true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In credit scoring, a TP is a correctly predicted approval for someone who should be approved, whereas an FP is an incorrect approval for someone who should not be. Conversely, a TN is a correct rejection for someone who should be rejected, and an FN is an incorrect rejection for someone who should be approved. The fairness of an ML model in this setting can be evaluated by analyzing the rates of these outcomes across different groups defined by protected attributes.

Disparate impact can be quantified by comparing error rates: the false positive rate (FPR) and the false negative rate (FNR). For a fair model, these rates should be similar across all groups:

$$FPR(a) = P(\hat{Y} = 1|Y = 0, A = a)$$

$$FNR(a) = P(\hat{Y} = 0|Y = 1, A = a)$$

The goal is to minimize the difference in these rates between groups, which can be mathematically represented as:

$$|FPR(a) - FPR(a')| < \epsilon$$

$$|FNR(a) - FNR(a')| < \epsilon$$

where ϵ is a small positive value representing an acceptable margin of error

2.3 Trade-offs in Fairness Criteria

Fairness in ML models often involves managing trade-offs between different fairness criteria (Chouldechova, 2017). These trade-offs exemplify the inherent tensions between error rates, such as the trade-off between equalizing FPRs and FNRs across groups. The tension arises because achieving equal FPRs across groups may lead to unequal FNRs, and vice versa, due to differing base rates and distributions of the predictor variables in the data (Kleinberg et al.,

2016). This can be mathematically formalized using conditional probability and Bayes' Theorem, which relates the predictive values and base rates:

$$P(Y = 1|\hat{Y} = 1, A = a) = \frac{P(\hat{Y}=1|Y=1,A=a)\cdot P(Y=1|A=a)}{P(\hat{Y}=1|A=a)}$$

This equation illustrates how the probability of an event (e.g., being creditworthy) conditional on a prediction and a protected attribute depends on both the accuracy of the prediction and the base rate of the outcome in the group. The complexity of these relationships highlights why a nuanced approach to fairness is necessary when developing and evaluating ML models.

2.4 Chi-squared Test for Statistical Independence

In the fairness evaluation of machine learning credit scoring algorithms, the Chi-squared test for independence helps determine if there is a significant bias in loan approvals related to protected attributes. This statistical test is essential for examining the false negative rate (FNR)—cases where credit-worthy individuals are wrongly denied loans—which are of particular concern in fairness assessments due to their potential to reinforce economic inequality.

To conduct the test, false negatives are cross-tabulated with protected attributes, and the Chi-squared statistic is calculated to compare observed and expected frequencies of loan denials. A high Chi-squared value indicates a significant difference between the groups, suggesting potential bias:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Oi represents observed counts of false negatives and Ei represents expected counts if the decision process were independent of the protected attribute. A non-significant Chi-squared value would imply that the distribution of false negatives does not statistically differ from what would be expected in a fair system, indicating no evidence of bias in the algorithm concerning the attribute tested. The Chi-squared test thus serves as a quantitative method to flag significant disparities in the false negative rate.

A crucial outcome of the Chi-squared test is the p-value, which quantifies the probability of observing the calculated statistic, or one more extreme, by chance if there were no true association between the protected attributes and the false negatives. A low p-value (< 0.05 for this experiment) is considered evidence against the null hypothesis of independence, indicating that the differences in false negative rates are statistically significant. Conversely, a p-value higher than the threshold suggests that any observed disparities in loan denials could reasonably occur under random conditions, and thus the evidence does not support a claim of bias with respect to the protected attribute. This p-value serves as a benchmark for determining whether the detected patterns of errors are likely due to inherent biases in the credit scoring algorithm or random variations.

3 Predictions

3.1 Data

The publicly available Home Mortgage Disclosure Data (HMDA) dataset for 2021 provided by the Consumer Financial Protection Bureau (CFPB) is used to train and test the machine learning models. It contains over 90% of mortgage records and anonymized sensitive and non-sensitive features (Consumer Financial Protection Bureau, 2021). I narrowed down the dataset to Pennsylvania mortgages that were either originated (1), approved but not accepted (2), or denied (3), applications that completely filled out our sensitive features of race, ethnicity, and sex, and important financial information including the debt-to-income ratios, amount of the loan, income, and others. In the dataset's 'action_taken' target variable, loans that were approved but not accepted (previously labeled '2') have been reclassified and combined with originated loans (labeled '1') to represent all approved loans. After cleansing the dataset, it is left with 314096 entries.

Fig. 1: Feature Overview

#	Feature	Feature Description
0	action_taken	Outcome of the loan application (target variable)
1	purchaser_type	Type of entity purchasing the loan
2	preapproval	Indicates if loan preapproval was requested
3	loan_type	Classification of the loan type
4	loan_purpose	The purpose for which the loan was applied
5	lien_status	Status of the lien on the property
6	reverse_mortgage	Indicates if the mortgage is a reverse mortgage
7	open-end_line_of_credit	Indicates if credit is extended under an open-end credit plan
8	loan_amount	The amount of the loan
9	hoepa_status	Status under the Home Ownership and Equity Protection Act
10	loan_term	Term of the loan in months
11	negative_amortization	Indicates if the loan can have negative amortization
12	interest_only_payment	Indicates if the loan can have interest-only payments
13	balloon_payment	Indicates if the loan includes a balloon payment
14	other_nonamortizing_features	Indicates other non-amortizing loan features
15	property_value	Value of the property being mortgaged
16	construction_method	Method of construction
17	occupancy_type	Type of property occupancy
18	manufactured_home_secured_property_type	Type of secured property for a manufactured home

#	Feature	Feature Description
19	manufactured_home_land_property_interest	Interest details for land where a manufactured home is placed
20	total_units	Number of units on the property
21	income	Income of the applicant in thousands
22	debt_to_income_ratio	Applicant's debt-to-income ratio
23	applicant_credit_score_type	Type of credit score used to evaluate the applicant
24	applicant_ethnicity-1	Ethnicity of the loan applicant
25	applicant_race-1	Race of the loan applicant
26	applicant_sex	Sex of the loan applicant
27	submission_of_application	Indicates how the application was submitted
28	initially_payable_to_institution	Indicates if the loan is payable to the institution initially
29	aus-1	Automated underwriting system used
30	tract_population	Population in the tract where the property is located
31	ffiec_msa_md_median_family_income	Median family income for the MSA/MD
32	tract_to_msa_income_percentage	Income percentage of the tract relative to MSA/MD
33	tract_owner_occupied_units	Number of owner-occupied units in the tract
34	tract_one_to_four_family_homes	Number of one-to-four family homes in the tract
35	tract_median_age_of_housing_units	Median age of housing units in the tract

Fig. 2: Dataset overview by ethnicity

	Entries
Hispanic or Latino	15274
Not Hispanic or Latino	298821

Fig. 3: Dataset overview by race

	Entries
Native American or Alaskan Native	969
Asian	14459
Black or African American	26018
Native Hawaiian or Pacific Islander	344
White	272305

Fig. 4: Dataset overview by sex

	Entries
Male	199378
Female	114717

A potential flaw of this dataset is the exclusion of credit scores, which most lenders use to assess creditworthiness. This could change the accuracy score (higher or lower); however, the primary objective of our research is not to assert the absolute accuracy of predictions but to investigate whether the predictions made by machine learning models exhibit bias against certain groups. By focusing on bias rather than accuracy, our study contributes to the understanding of how machine learning algorithms perform across diverse populations and highlights the importance of designing algorithms that are fair and equitable.

3.2 Machine Learning Algorithms

Four machine learning methods are used to predict the mortgage outcomes using the data described above: Linear Regression, Random Forest, XGBoost, and K-Nearest Neighbors

(KNN). The dataset was split into 80% for training and 20% for testing. Sensitive features relating to race, ethnicity, and sex were excluded from the training set so that the algorithm doesn't use it as a factor in whether they are approved or not. This adheres to anti-discrimination laws such as the Fair Housing Act (FHA) and the Equal Credit Opportunity Act (Consumer Financial Protection Bureau [CFPB], 2021). After training, the models are evaluated to determine if they are biased toward these sensitive attributes in their predictions despite not being explicitly provided with such information.

Fig. 5, Predicted Approval Metrics

	Linear Regression	Random Forest	XGBoost	KNN
Precision	0.87	0.87	0.88	0.87
Recall	0.91	0.91	0.96	0.91
F1-Score	0.89	0.89	0.92	0.89

Fig. 6, Predicted Denial Metrics

	Linear Regression	Random Forest	XGBoost	KNN
Precision	0.98	0.98	0.99	0.98
Recall	0.97	0.97	0.97	0.97
F1-Score	0.97	0.97	0.98	0.97

Fig. 7, Overall Accuracy

	Linear Regression	Random Forest	XGBoost	KNN
Overall Accuracy	0.96	0.96	0.97	0.96

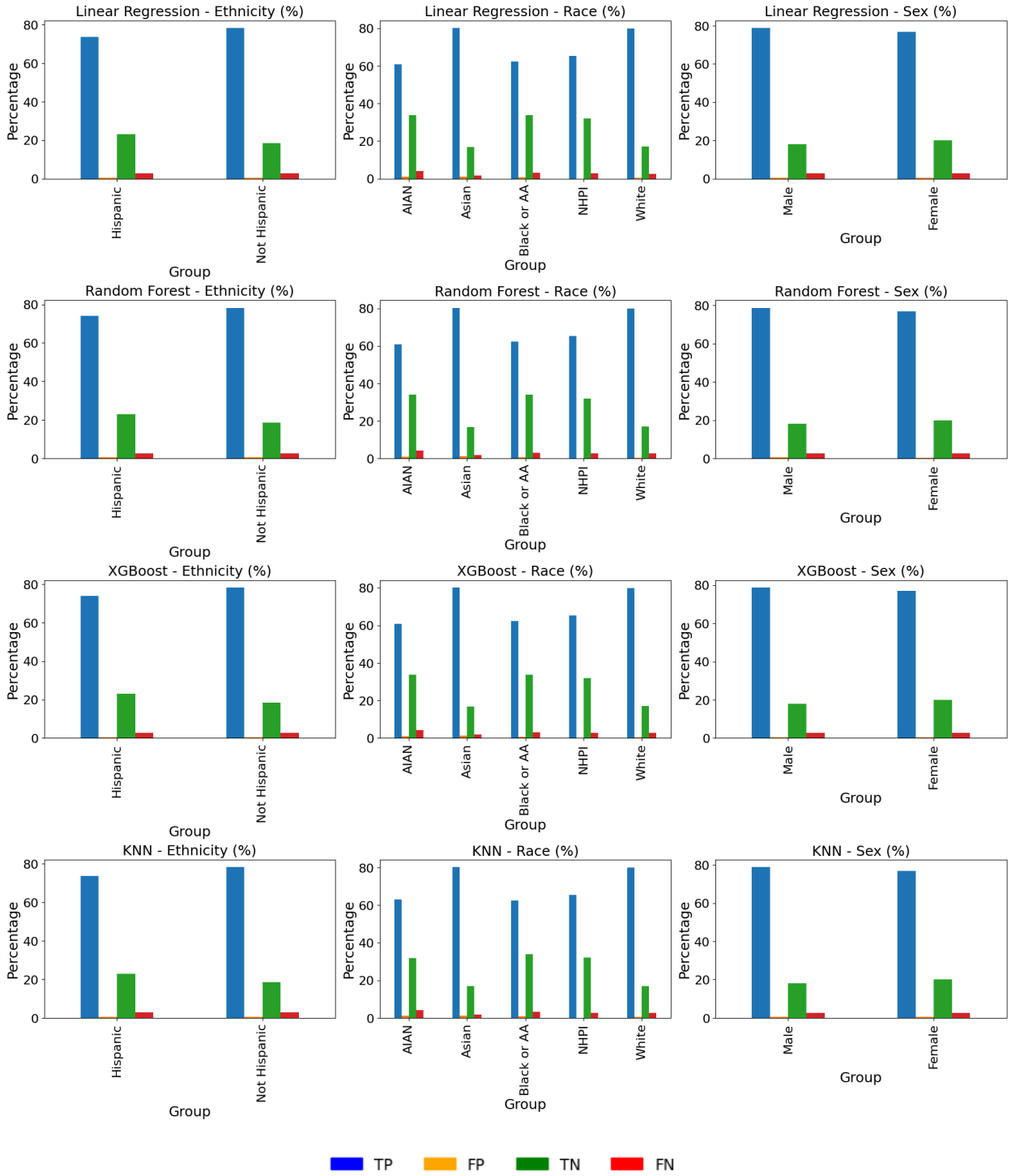
The overall accuracy between the different models is very closely aligned, with XGBoost marginally leading the other three models at 0.97 vs. 0.96. These high accuracy rates indicate a robust but not perfect performance. The machine learning algorithms are much more accurate at predicted denials rather than approvals. The generally higher accuracy in predicting denials could be attributed to several factors, including more consistent or defining features among the denied loans within the dataset, which makes them easier for models to learn and recognize. This could reflect a real-world propensity for loan applications with specific negative financial indicators to be denied, creating clear patterns for the algorithms to detect.

4 Bias Analysis

While sensitive attributes were excluded during training, evaluating confusion matrices for each model by protected features (race, ethnicity, and sex) provides insight into potential biases in the model's predictions.

4.1 Bias between Different Machine Learning Models

Fig. 8, Table of Confusion by Protected Features and Different Machine Learning Models



AIAN: American Indian or Alaskan Native, NHPI: Native Hawaiian and Pacific Islander

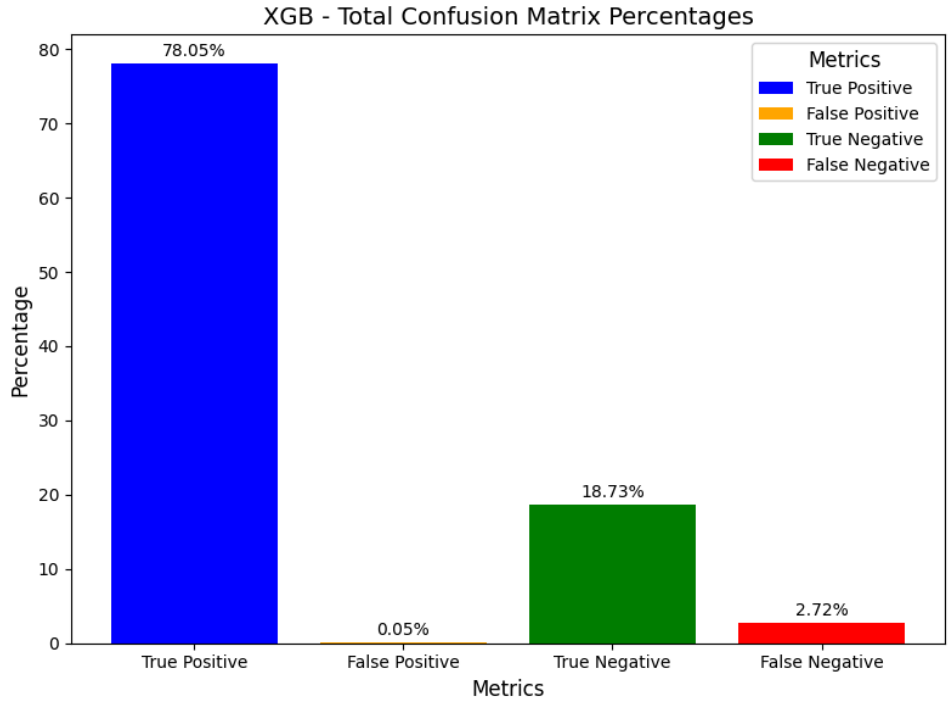
There is negligible variance in the performance between the different machine learning models tested in 3.2. All models showed similar true and false positive/negative rates and they all had a p-value of 1.0.

Since there is no statistically significant bias between the different machine learning models, the bias analysis between the different protected features will focus on XGBoost, because it is slightly more accurate.

Fig. 9, Table of true and false positive/negatives

	True Positive %	False Positive %	True Negative %	False Negative %
Total	78.05%	0.05%	18.73%	2.72%

Fig. 10, Table of true and false positive/negatives



4.2 Ethnicity

Fig. 11, Table of true and false positive/negatives by ethnicity

	True Positive %	False Positive %	True Negative %	False Negative %
Hispanic or Latino	73.80%	0.05%	23.00%	2.75%
Not Hispanic or Latino	78.26%	0.05%	18.52%	2.72%

Fig. 12, Graph of true and false positive/negatives by ethnicity

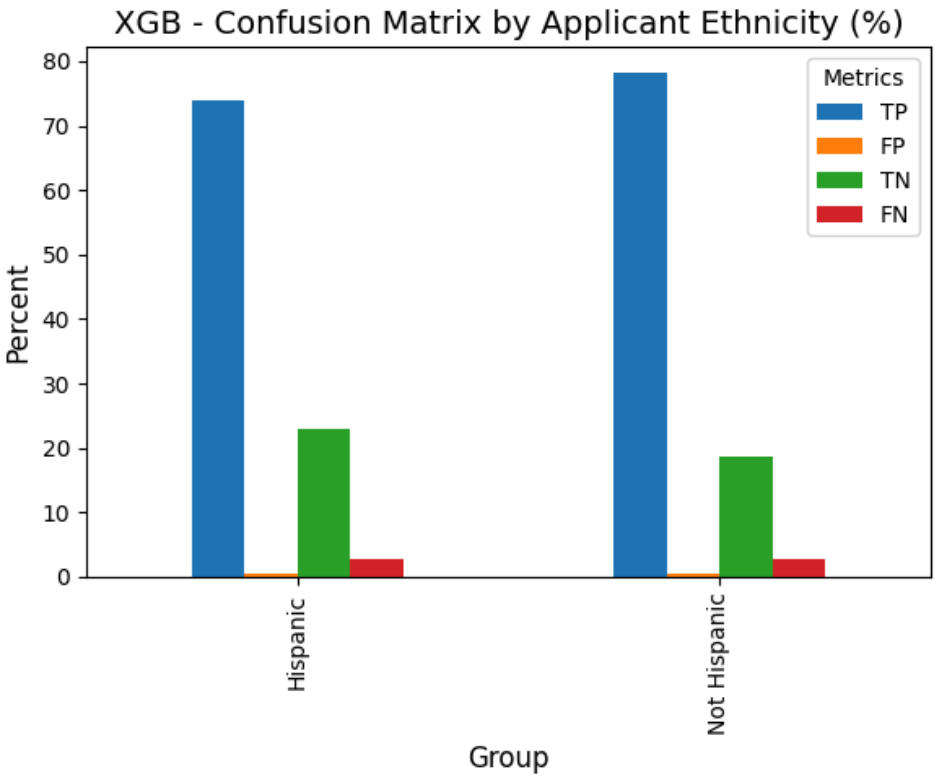


Fig. 13, Significance test of false negative rate by race

	Difference of False Negative from Total	p-value
Hispanic or Latino	+0.03%	0.943
Not Hispanic or Latino	-0.00%	1.0

Both p-values exceed the usual 0.05 significance level, which suggests there is no statistical evidence to conclude that the false negative rates for these two ethnic groups are different from the overall false negative rate of 2.72%.

4.3 Race

Fig. 14, Table of true and false positive/negatives by race

	True Positive %	False Positive %	True Negative %	False Negative %
American Indian and Alaskan Native	60.93%	1.04%	33.85%	4.17%
Asian	80.14%	1.20%	16.86%	1.81%
Black or African American	61.74%	0.72%	33.86%	3.69%
Native Hawaiian or Pacific Islander	65.28%	0%	31.95%	2.78%
White	79.50%	0.45%	17.32%	2.73%

Fig. 15, Graph of true and false positive/negatives by race

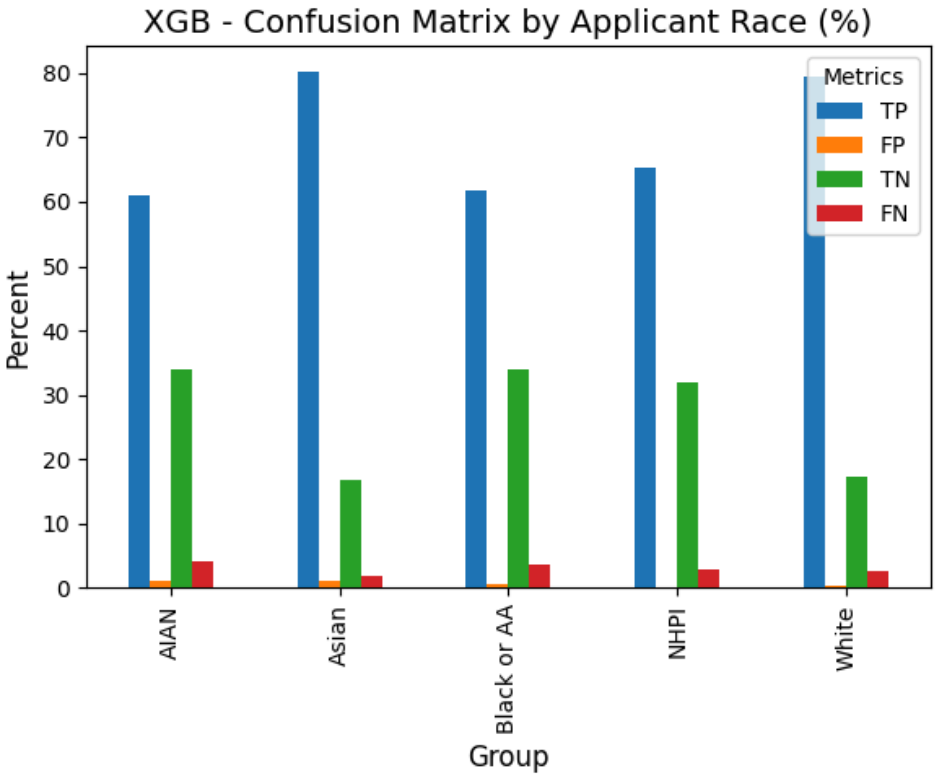


Fig. 16, Significance test of false negative rate by race

	Difference of False Negative from Total	p-value
American Indian and Alaskan Native	+1.45%	0.436
Asian	-0.91%	0.021
Black or African American	+0.97%	0.0063
Native Hawaiian or Pacific Islander	+0.06%	0.982
White	+0.01%	0.919

The p-value is a measure of the probability that an observed difference could have occurred just by random chance. Typically, a p-value < 0.05 is considered statistically significant.

According to these results, the differences in false negative rates are statistically significant for the Black and Asian populations compared to the overall population rates. Black applicants is at a significant disadvantage due to bias in the machine learning algorithm, while bias against Asian applicants are much less than the average. For White, Native Hawaiian, and Native American populations, the differences are not statistically significant.

Despite Native Americans experiencing a substantially higher rate of false negatives in credit decisions compared to the overall population, this group's limited number of data entries results in a higher p-value. Consequently, the current data does not provide sufficient statistical power to draw conclusions about bias.

4.4 Sex

Fig. 17, Table of true and false positive/negatives by sex

	True Positive %	False Positive %	True Negative %	False Negative %
Male	78.74%	0.56%	18.00%	2.72%
Female	76.85%	0.41%	20.00%	2.73%

Fig. 18, Graph of true and false positive/negatives by sex

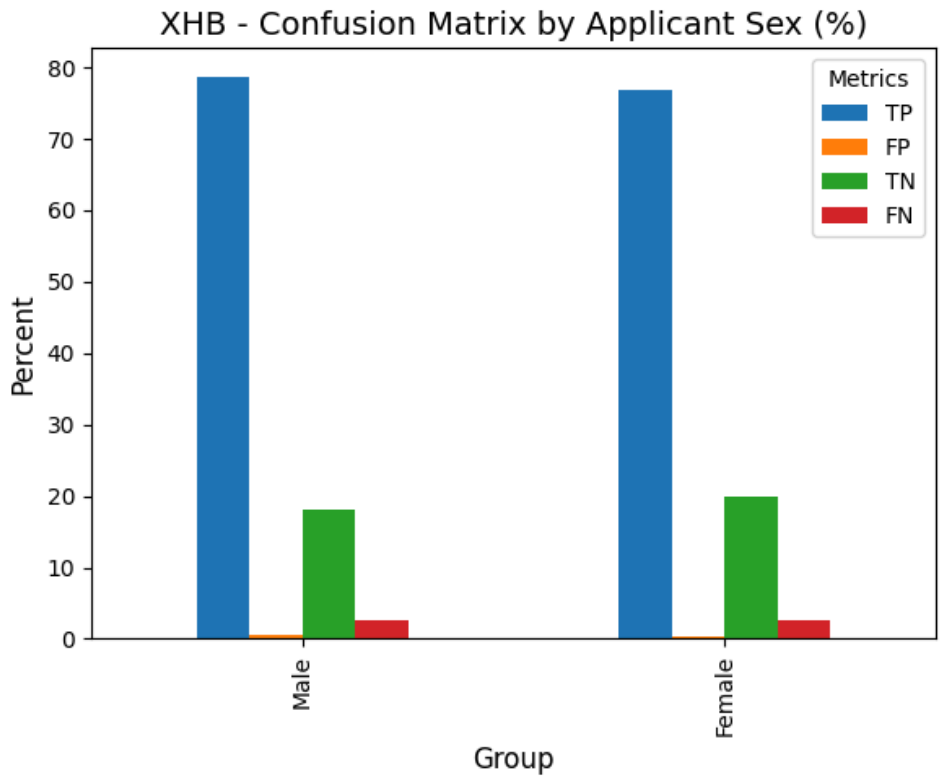


Fig. 19, Significance test of false Negative rate by sex

	Difference of False Negative from Total	p-value
Male	-0.00%	1.0
Female	+0.01%	0.947

Since both p-values are well above the standard threshold of 0.05 for statistical significance, we conclude that there are no significant differences in the false negative rates between males and females when compared to the overall rate.

5 Conclusions

5.1 Feature Importance

Feature importance measures how much each feature contributes to a model's predictions. In tree-based models like XGBoost, it's often gauged by how much a feature decreases uncertainty when creating decision trees. Even without race as an input, a model can still be biased if important features correlate with race. For instance, a loan model may use location as a key feature. If some locations have historically been inhabited by certain racial groups, the model might deny loans to those groups more often.

Columns of tract_to_msa_income_percentage and ffiec_msa_md_median_family_income which track income and median family income of the town or area of the applicant had a strong negative linear correlation with mortgage approvals, while Black applicants had a moderate positive linear correlation with those features, and Asian applicants had a moderate negative linear correlation with those features.

Removing those features resulted in the model's accuracy dropping to 94.8%, while Black applicants' false negative rate dropped to 3.42% and Asian applicants' false negative rate increased to 2.33%, while the model's false negative rate increased to 2.89%. Further testing and lending data is required to find the optimal accuracy rate between equity, profit, and accuracy.

5.2 Summary

In this paper, we created a machine learning algorithm that correctly predicted whether the applicant should be approved or denied for a mortgage loan with a 97% accuracy rate, then analyzed its inherent bias against certain groups of people. However, our analysis revealed that the model, despite not being directly informed by race, ethnicity, or gender data, showed significant discriminatory bias. We are able to support the hypothesis that Black applicants are discriminated against more heavily in machine learning credit scoring algorithms. Black applicants were incorrectly denied loans 35.6% more frequently than the average rate of incorrect denials, highlighting a critical bias issue. Conversely, the model was more accurate for Asian applicants, who experienced a lower rate of false negatives. While the low false negative

rate for Asian applicants is good, the goal is to reduce the false negative rates across all groups to ensure equity in mortgage lending. We were unable to find any statistically significant bias based on ethnicity or sex.

I am also unable to find any statistically significant bias between the different machine learning algorithms I tested: Linear Regression, Random Forest, XGBoost, and K-Nearest Neighbors. They all resulted in a very similar rate of false negatives between different groups and XGBoost was very slightly more accurate.

The Equal Credit Opportunity Act was a significant milestone in preventing credit discrimination based on race, ethnicity, and sex, yet algorithmic bias in AI and machine learning still poses a threat to fair credit scoring. Banks should be particularly attentive to these biases, not only because fairness is imperative, but also because eliminating such biases can unlock new revenue opportunities. By ensuring that credit scoring algorithms are bias-free, banks can extend credit to deserving applicants who might otherwise be unfairly assessed, leading to increased customer bases and market expansion. Furthermore, the correction of algorithmic bias aligns with the socially responsible banking models that attract modern customers (Federal Reserve, 2021).

This research into machine learning algorithms in credit scoring extends beyond the financial sector, providing a crucial framework for identifying and correcting bias. It has broader implications for other sectors, like housing and employment, where machine learning algorithms are used for tenant selection and job applications (FTC, 2021).

References

[1] Brandeis University, Waltham, MA: Institute on Assets and Social Policy. Shapiro, T. M., Meschede, T., & Osoro, S. (2013). The Roots of the Widening Racial Wealth Gap: Explaining the Black-White Economic Divide. Research and Policy Brief.

[2] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104, 671.

[3] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163.

[4] Consumer Financial Protection Bureau. (2021). About HMDA. Retrieved from https://www.consumerfinance.gov/policy-compliance/rulemaking/final-rules/regulation-c-home-mortgage-disclosure-act/

[5] Consumer Financial Protection Bureau. (2021). The Home Mortgage Disclosure Act.

[6] European Union. (2019). Ethics guidelines for trustworthy AI.

[7] Federal Trade Commission (FTC). (2021). "Big Data: A Tool for Inclusion or Exclusion?"

[8] Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. Advances in Neural Information Processing Systems, 29.

[9] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer Credit-Risk Models via Machine-Learning Algorithms. Journal of Banking & Finance, 34(11), 2767-2787.

[10] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv preprint arXiv:1609.05807.

[11] National Association of Realtors. (2022). Quick Real Estate Statistics.

[12] Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2020). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. Proceedings of the National Academy of Sciences, 117(42), 26223-26229.

[13] Rugh, J. S., & Massey, D. S. (2010). Racial Segregation and the American Foreclosure

Crisis. American Sociological Review, 75(5), 629-651.

[14] U.S. Census Bureau. (2021). Homeownership Rate by Race and Ethnicity of Householder.