Coded Bias: An Investigation of Bias in Lending Algorithms

Arul Nigam
TJHSST Computer Systems Lab

Abstract

Machine learning algorithms have seen widespread adoption in recent years. However, as the use of these algorithms becomes ubiquitous, the associated problems do as well, making it crucial to identify, measure, and respond to these threats. The issue of algorithmic bias is foremost among the challenges posed by machine learning that scientists, policymakers, and others must swiftly address in order to ensure fair and efficient outcomes. Research has identified substantial racial algorithmic bias in mortgage-lending decisions, most prominently against Black and Latino borrowers. In this paper, we explore various techniques to measure this bias, uncover "proxy" sources, and identify potential solutions.

Keywords: algorithmic bias, algorithmic fairness, artificial intelligence ethics, credit scoring, machine learning bias

Introduction

Machine learning can be used to improve people's lives dramatically, from enabling early detection of cancer to enhancing military intelligence to automating self-driving cars, and it has the potential to do even more. Still, the field must overcome multiple fundamental problems, of which one of the most prominent is implicit bias in algorithms, resulting in discriminatory outcomes. As the use of machine learning becomes pervasive, it is increasingly important to understand, identify, quantify, and mitigate the degree of implicit bias in those algorithms. This is particularly important because of their role in automating important decisions, from medical diagnoses to fraud detection (Khetan 2019). Already, there are numerous instances in which these algorithms yield results that inadvertently discriminate against groups even when explicitly designed to avoid considering protected attributes such as sex or race. For instance, Amazon was forced to cancel a project that used machine learning to review resumes and select the best candidates, because the algorithm downgraded female candidates (Dastin 2018).

The most common culprit is the presence of data proxies, seemingly harmless data points which may be used to infer protected attributes. One example is ZIP Code; homophily results in members of similar groups clustering together geographically, causing ZIP Codes to be correlated with the racial or ethnic groups that dominate that area and in turn serve as a data proxy for race.

An example of an impacted field is criminal justice. Algorithmic bias was found to affect a risk-assessment tool used by many judges to determine whether to grant bail or parole. Research has shown that the results of the tool exhibited significant bias. resulting in discriminatory treatment of people of color. Even healthcare is riddled with this problem. A recent study showed that a triaging system, used in many hospitals, systematically referred Black patients for additional care at much lower rates than White patients, even when key medical information was

the same. This bias was again largely due to training data influenced by historical racism, abuse, and exploitation in healthcare.

In this research, we focused on examining algorithmic bias in another major area: the mortgage lending industry.

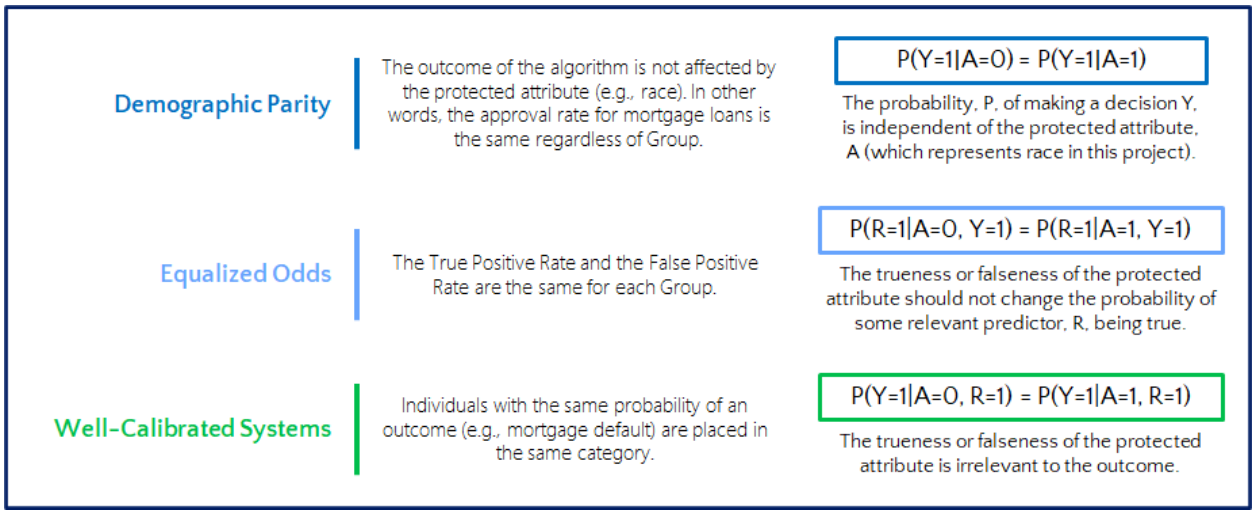
A common application of machine learning in the financial services industry in recent years has been automated credit scoring and lending algorithms. However, these algorithms have been shown to exhibit unintended racially-biased outcomes, even when the algorithms appear to be race-blind. Even though AI-based algorithms have reduced racial bias in mortgage lending by up to 40%, impactful disparities are still perpetuated by these algorithms. Research conducted at the University of California Berkeley found that Black and Latino borrowers pay up to 8.6 basis points higher interest on mortgage loans than White or Asian borrowers do, costing them between $250 million to $500 million every single year (Bartlett et al., 2019). Even though protected attributes like race or sex are excluded from an algorithm's training data, the resulting models can still reflect bias. This research seeks to measure and identify the sources of this bias and develop potential solutions.

Machine Learning-based Lending Algorithm Model Development

Machine learning techniques were employed to first create an ensemble of lending algorithms that model the decision-making process in banks. These algorithms were applied to understand data proxies and the underlying factors that lead to discriminatory outcomes in mortgage lending.

The modeled lending algorithms were used to make recommendations for limiting discrimination while also preserving the efficacy of the algorithm.

The Models of Fairness



This project relies on a combination of three of the many possible models for measuring and testing fairness. The basis of our understanding of fairness is set in Demographic Parity as it controls for protected attributes and Group.

Methods

Given that lending algorithms are a core part of a bank's business, the design of these algorithms are proprietary and confidential. In order to conduct this research, we developed a representative lending algorithm to support our analysis. The lending algorithm used in this project was developed from scratch, using training data to create a model to predict whether a loan application would be approved or denied.

The primary training dataset that was used to develop the lending algorithm was the substantial loan information reported by major banks in compliance with the Home Mortgage Disclosure Act (HMDA) and published by the Consumer Financial Protection Bureau.

An ensemble of different algorithms including Support Vector Machines, ElasticNets, Random Forests, Gradient Boosting, and Multivariate Adaptive Regression Splines (MARS) were employed to develop the algorithm.

A variety of representations of data produced by the mortgage lending models were analyzed to identify the underlying factors that contribute to racial bias in lending algorithms.

Results and Discussion

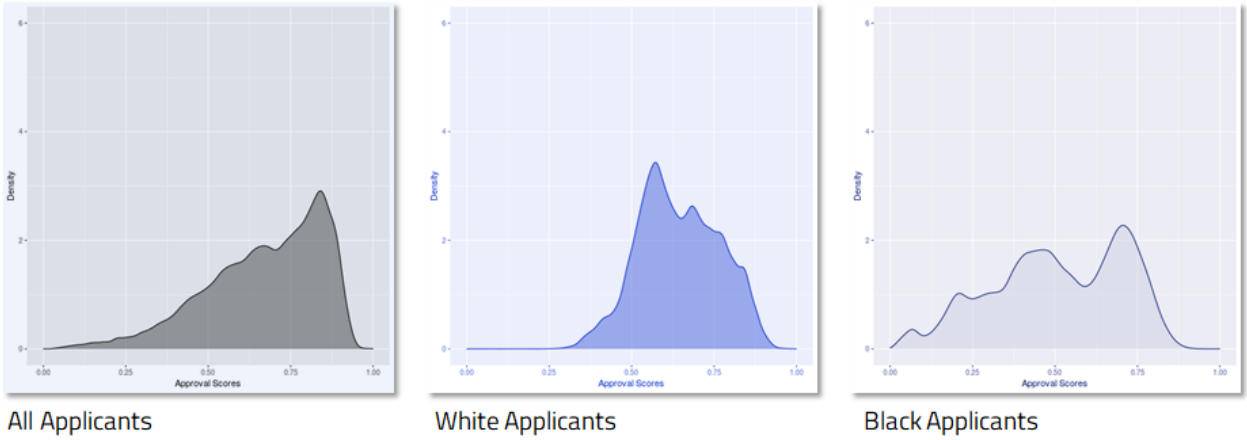


Figure 1. Distribution of Probability Score Nationwide

Figure 1 illustrates the probability distribution for a loan application acceptance based on race. These probability distributions were derived from the ElasticNet model trained on nationwide

data. Very few White applicants have a probability of less than 30% of being accepted for a loan and the probability of acceptance is skewed towards the positive end of the spectrum. This indicates that, on average, applicants who are White have a higher probability of having their applications approved. On the other hand, very few Black applicants have a probability greater than 85% of being accepted for a loan and the distribution of probabilities is much more spread out. The skew is much more severe, and the data may be bimodal with a peak below 0.5, meaning those profiles are more likely than not to get rejected.

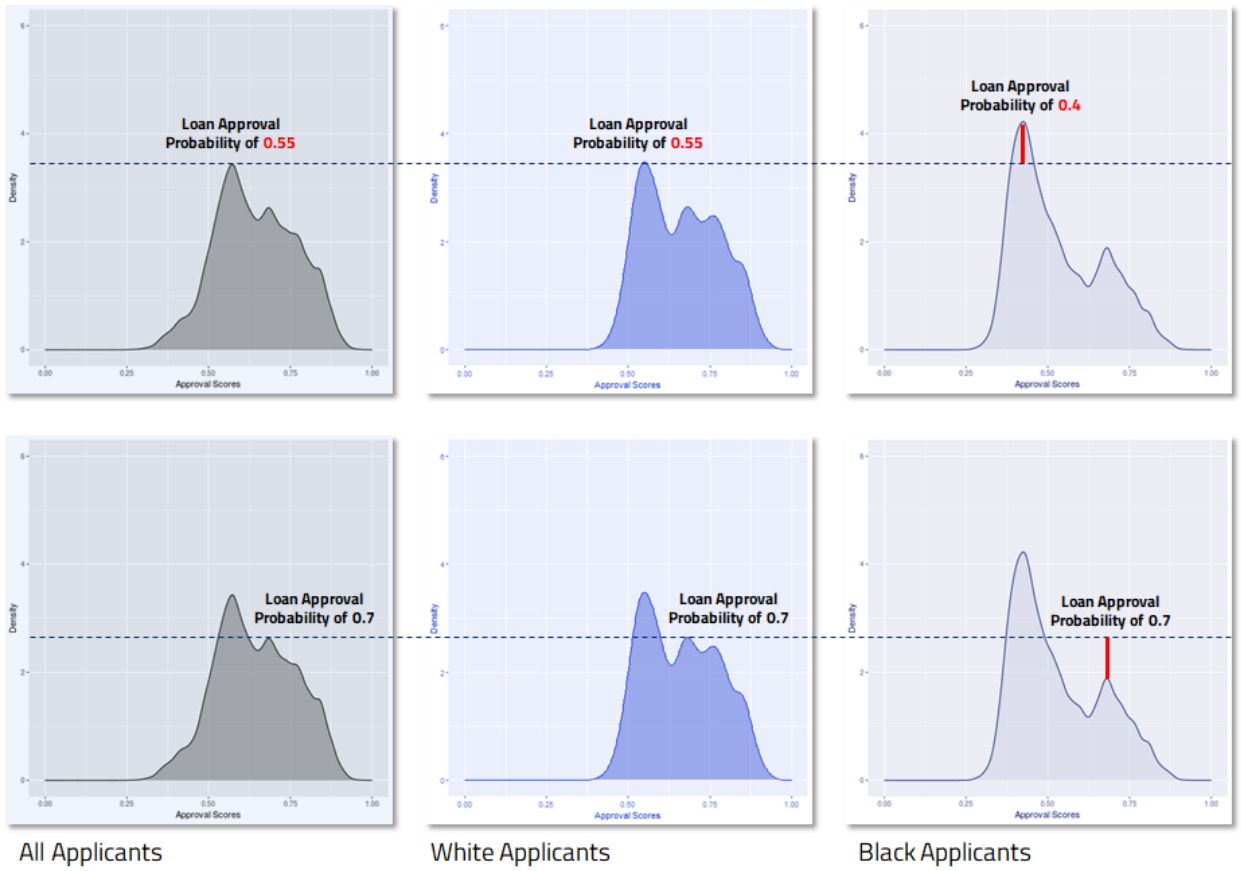


Figure 2. Probability Distribution of Loan Approvals in California

Figure 2, visualizations from a model that also used ElasticNet, uses data from the state of California as a case study. Although some features are exaggerated in different profiles, the shapes of the distributions are quite similar. However, a close examination at the two peaks makes the disparities evident. For White applicants, the main peak is on par with the peak for all applicants indicating that they have a higher probability of their loan application being accepted compared to Black applicants. Furthermore, a greater proportion of White applicants have a high probability of their loan application being accepted compared to a smaller proportion of Black applicants at a lower probability.

For Black applicants, we can see that the main peak is much higher and significantly further to the left at approximately 0.4 or about 0.15 points below the peak for other applicants, meaning that on average, Black applicants have a lower probability of having a mortgage loan being accepted. Furthermore, the smaller second peak indicates that a smaller proportion of Black applicants have high probabilities of their loans being approved.

These discrepancies indicate that models based on the statewide distribution are ineffective in fairly evaluating Black loan applications.

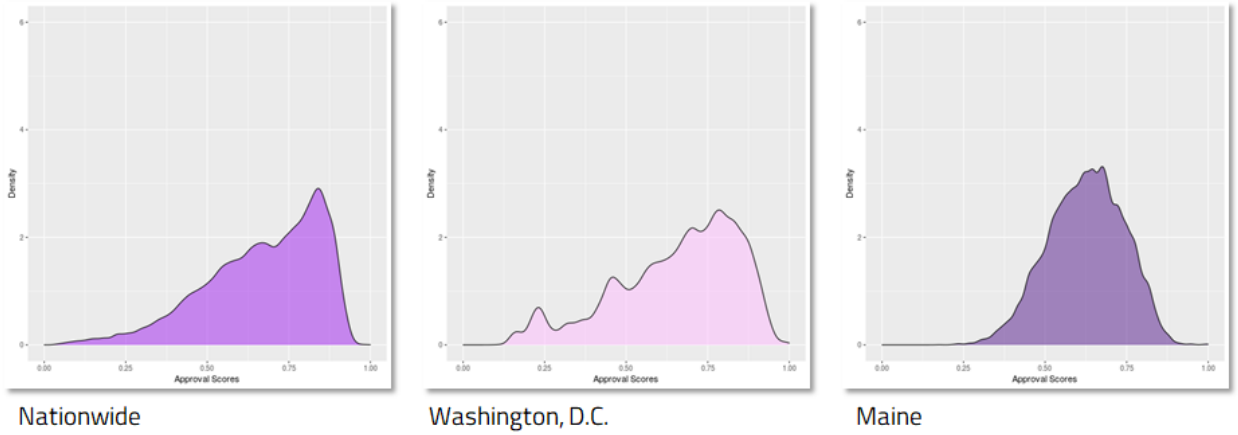


Figure 3. Distribution of Probability Scores by State

The St. Louis Federal Reserve established a standard that compared economic racial inequality by state by looking at the median income of a Black household per $1 of income in a White household. This number ranges from just $0.32 in Washington, D.C. to $0.87 in Maine.

Compared to the nation as a whole, *Figure 3* shows that Washington, D.C.'s distribution is similar in shape but leans much more towards the bottom half, with multiple smaller peaks below 0.50 and even below 0.25, indicating that there are more applicants with lower probabilities of their loan applications being accepted. Maine, however, is representative of the ideal: a relatively normal distribution centered well above 0.50.

This comparison shows us that the issue is not just the algorithm acting in a vacuum: rather, it amplifies underlying inequality in the places where it is applied.

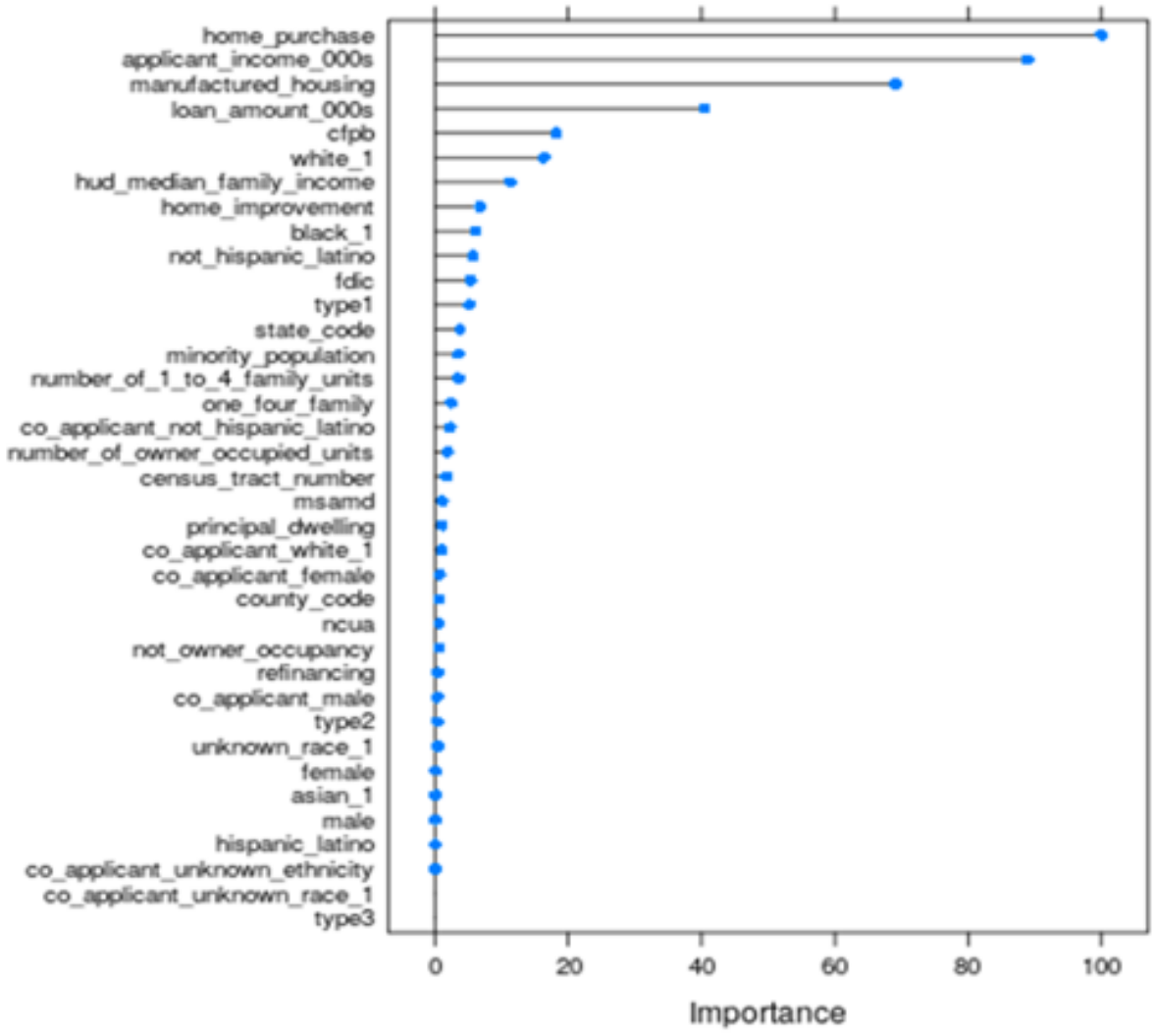


Figure 4. Relative Variable Importance

We considered the relative importance of several variables in a GradientBoosting model. There are some other variables included in the HMDA dataset that are excluded here as they have zero- or near-zero variance, or because they are linearly dependent on a different variable. As illustrated in *Figure 4*, three factors—whether the loan purpose is for a home purchase, applicant income, and whether the mortgage is for manufactured housing—are dominant, closely followed by loan amount. This indicates that even moderate levels of "proxy" behavior, where some explicit factor that is strongly correlated with some non-explicit factor begins to implicitly represent that factor in the algorithm, may be particularly consequential if observed in any of these major variables.

	White	Black	Absolute Difference	Ratio
Predicted Approval Probability	29.2%	-27.9%	57.0%	1.0
Minority Population	-29.1%	26.4%	55.5%	1.1
Male Applicant	23.7%	-4.6%	28.2%	5.2
Female Co-Applicant	16.1%	-7.3%	23.4%	2.2
Female Applicant	4.5%	10.4%	5.9%	2.3
Applicant Income	0.1%	-1.1%	1.2%	8.3

Figure 5. Race Correlation with Key Variables

Predicted Approval Probability, the first factor in *Figure 5* is the response variable: the probability of an application being approved as predicted by the machine learning model. The correlation between the prediction of approval and being White or being Black is roughly equal in magnitude, as indicated by the absolute ratio which rounds to 1. However, being White is positively correlated with approval, meaning being White makes an applicant more likely to be approved for the loan. A Black applicant, on the other hand, is equally less likely to be approved, due to the negative correlation, which results in a significant difference of 57%.

Minority Population & *Applicant Income* which represents the percentage of the population of the applicant's ZIP Code are racial minorities, has the opposite relationship with White and Black applicants as the predicted probability did. Black applicants are much more likely to be in a ZIP Code with a high minority population, while White applicants are just as likely not to be. This results in another substantial absolute difference. The absolute difference between races for Applicant Income, however, is rather low. While the Minority Population variable represents a significant racial disparity despite a relatively low variable importance, as seen in *Figure 4*, the applicant income represents a small racial disparity despite being of high relative importance.

Finally, reviewing the correlation of applicant and co-applicant sex with race reveals that Black applicants are significantly more likely to be female, and much less likely to be applying with a female co-applicant. White applicants however are more likely to be male and applying with a female co-applicant. We can generally infer from this combination that the applicants are a married couple. The differences in correlation between these factors and each race speak to a substantial difference in marriage rates between these two groups. This information makes it possible to infer familial status, which the Fair Housing Act prohibits, amongst other protected attributes, from being considered in the mortgage underwriting process. Taken together, these factors are another data proxy for familial status rather than race.

	Loan originated	Application approved but not accepted	Application denied by financial institution	Application withdrawn by applicant	File closed for incompleteness	Total number of applicants
White	64.41%	3.34%	14.38%	13.36%	4.51%	7,687,800
Black or African American	49.51%	3.50%	25.71%	15.02%	6.25%	817,630
Asian	64.46%	3.43%	12.90%	14.59%	4.62%	570,056
American Indian or Alaska Native	47.88%	3.33%	26.11%	15.55%	7.13%	90,340
Native Hawaiian or Other Pacific Islander	56.54%	3.17%	19.22%	15.25%	5.81%	48,122

Figure 6. Race vs. Loan Application Outcome

The application outcomes for each racial group listed in *Figure 6* indicate that White and Asian applicants were the most successful by far, with almost ⅔ of applications resulting in loans being originated, meaning that the loans were accepted and disbursed. Other minority groups had approximately 15% fewer loans originated. Asian and White borrowers also had the lowest proportion of denied applications, with Black and American Indian / Alaska Native borrowers being denied at twice the rate of Asian applicants. Notably, applications were withdrawn or closed for incompleteness at similar but alarmingly high rates across all races, suggesting that improving completion rates may improve outcomes.

Next Steps

The federal government requires factors like credit scores to be considered in lending decisions. This requirement is established through Freddie Mac and Fannie Mae. However, the use of these government directed factors can perpetuate racial disparities, and are not subject to the discretion of a financial institution. For example, there is a more than 50 point difference in the mean credit score of African Americans and the country as a whole. So, using credit scores is virtually guaranteed to result in disparate outcomes based on race (Choi et al., 2019).

Future studies should evaluate whether using race-conscious algorithms can reduce biased outcomes, by explicitly adjusting scores to correct measured racial bias. If so, we will need to partner with policymakers to understand the public policy constraints and revise the law to enable the explicit use of protected attributes like race, without losing sight of the spirit of legislation such as the Fair Housing Act. Furthermore, by assessing other models of fairness, we can continue to pinpoint areas of concern. Overcoming algorithmic bias to ensure that the mortgage lending process is fair for everyone requires the insight of both policymakers and

developers. A comprehensive model considering the complex interplay of technical and public policy decisions will enable us to make adjustments and achieve our societal goals.

As machine learning becomes more prominent it's our responsibility to mitigate the challenges of algorithmic bias before they are brought to bear at too great a cost.

Sources

Bartlett, Robert, et al. “Consumer-Lending Discrimination in the FinTech Era.” *NBER*, 17 June 2019, www.nber.org/papers/w25943.

Choi, Jung Hyun, et al. *Explaining the Black-White Homeownership Gap*. www.urban.org/sites/default/files/publication/101160/explaining_the_black-white_homeownership_gap_a_closer_look_at_disparities_across_local_markets_0.pdf.

Dastin, Jeffrey. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.” *Reuters*, Thomson Reuters, 10 Oct. 2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Federal Reserve Bank of Boston. “Geoffrey Tootell .” *Federal Reserve Bank of Boston*, 1 Jan. 2017, www.bostonfed.org/people/bank/geoffrey-tootell.aspx.

Fuster, Andreas, et al. “Predictably Unequal? The Effects of Machine Learning on Credit Markets.” *SSRN*, 17 Nov. 2017, papers.ssrn.com/sol3/papers.cfm?abstract_id=3072038.

Gillis, Talia B. “The Input Fallacy.” *SSRN*, 5 May 2020, papers.ssrn.com/sol3/papers.cfm?abstract_id=3571266.

Kent, Ana Hernández. “Examining U.S. Economic Racial Inequality by State.” *St. Louis Fed*, Federal Reserve Bank of St. Louis, 20 Nov. 2020, www.stlouisfed.org/publications/bridges/volume-3-2020/examining-us-economic-racial-inequality-by-state.

Khetan, Vivek. “Bias in Machine Learning Algorithms.” *Medium*, Towards Data Science, 30 Apr. 2019, towardsdatascience.com/bias-in-machine-learning-algorithms-f36ddc2514c0.

Klein, Aaron. “Reducing Bias in AI-Based Financial Services .” *Brookings*, Brookings, 17 July 2020, www.brookings.edu/research/reducing-bias-in-ai-based-financial-services.

Matthew Stewart, PhD Researcher. “Handling Discriminatory Biases in Data for Machine Learning.” *Medium*, Towards Data Science, 29 July 2020, towardsdatascience.com/machine-learning-and-discrimination-2ed1a8b01038.

Obermeyer, Ziad, et al. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science*, American Association for the Advancement of Science, 25 Oct. 2019, science.sciencemag.org/content/366/6464/447.full.

Pierson, Emma, et al. “Fast Threshold Tests for Detecting Discrimination.” ArXiv.org, 10 Mar. 2018, arxiv.org/abs/1702.08536.

"The Home Mortgage Disclosure Act." *Consumer Financial Protection Bureau*, www.consumerfinance.gov/data-research/hmda.

Vartan, Starre. "Racial Bias Found in a Major Health Care Risk Algorithm." *Scientific American*, Scientific American, 24 Oct. 2019, www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm.