Proposal

VICTOR ZHANG, ELLIE KOGAN, ERIN MUTCHEK, ZHEKA CHYZHYKOVA, and MICHAEL LEE, CMPU 250, Vassar College

1 INTRODUCTION

Over the past several decades, affordable housing has decreased as rent has steadily increased and incomes have stagnated.[4] Home ownership has become increasingly more important as a means to build wealth, especially among Black Americans. Historically, Black homeowners have lower homeownership rates as a byproduct of racism and slavery and homes in black neighborhoods are significantly delayed. Additionally, Black Americans face discriminatory lending practices in the mortgage application process.[3]

When purchasing a house with a mortgage, you first need to get pre-approved for a mortgage loan. During the pre-approval process, the applicants submit simple budget information and lenders use credit reports to determine the maximum amount they are willing to lend. After picking out a property, applicants need to submit more comprehensive employment, income, asset, debt, property, and credit history information.[2] Using this information, lenders choose to approve or deny applications creating a loan estimate in the process. Ultimately determining the final interest rates, hidden algorithms are used throughout the process to evaluate home prices and mortgage applications.[1] During the application process, these algorithms are used to evaluate risk and predict loan repayment. Applicants who are falsely classified on risk or repayment may receive higher interest rates or less money or even get outright rejected, increasing financial burden and lessening financial mobility.

In their August 2021 investigation, *"The Secret Bias Hidden in Mortgage-Approval Algorithms,"* The Markup's Emmanuel Martinez and Lauren Kirchner analyzed over 2 million conventional mortgage applications from 2019. Their findings revealed that even after accounting for factors such as debt-to-income ratio, loan-to-value ratio, and credit score—elements lenders often cite to explain disparities—people of color were denied mortgages at significantly higher rates than White applicants. Nationally, Black applicants were 80% more likely to be denied than their White counterparts with similar financial profiles; Native American applicants faced a 70% higher denial rate, Asian/Pacific Islander applicants 50% higher, and Latino applicants 40% higher. In certain metropolitan areas, these disparities were even more pronounced, exceeding 250%. The investigation highlighted that high-income Black applicants with less debt were denied more often than high-income White applicants with more debt, suggesting that algorithmic underwriting systems may perpetuate existing biases rather than eliminate them.

1.1 Research Questions

In this project, we aim to explore the following research questions:

(RQ1) What are the potential impacts of the credit-scoring system algorithms on the approval rates of applicants of diverse identities and geographical backgrounds?

(RQ2) Do applicants of different race and gender backgrounds experience significantly different loan approval rates?

(RQ3) What disparities, if any, exist in the loan interest rates awarded to applicants of different race and gender backgrounds?

(RQ4) How can we create and assess a more fair credit-scoring model while still prioritizing accuracy?

2 DATA DESCRIPTION

The dataset used in this project outlines the circumstances under which Americans are approved for mortgages from private financial institutions. The collection of this data is mandated by the Home Mortgage Disclosure Act, a federal law passed in 1975 designed to hold banks accountable for declining capital in certain urban areas. In short, financial institutions are now required to report information about every application they process to the Consumer Financial Protection Bureau, which makes it publicly available each year in March. For the purpose of this project we'll be using application data from New York state in 2023. As we continue with preliminary analysis, we'll also choose a few rural, more homogenous counties for the purpose of comparison. Some of the key data points included in this set are the type of property being sought, a demographic description of the applicant(s), their income and credit score, and what loan terms they received. Also of interest is the applicant's debt-to-income ratio, often entered as a range of percentages (e.g. 50%-60%). Financial institutions have defended that this data point, along with a couple others, is often the actual basis for rejections that may appear racially motivated. Therefore, it will be productive to control for this factor in examining decision outcomes.

2.1 Research Hypothesis

Primary Hypothesis:

Black mortgage applicants in New York State in 2023 face significantly higher denial rates compared to White applicants with similar financial profiles, even when controlling for debt-to-income ratio, loan-to-value ratio, and credit score.

Secondary Hypotheses:

- **SP1: Interest Rate Disparities:** Black, Latino, Native American, and Asian/Pacific Islander applicants receive significantly higher mortgage interest rates than White applicants with similar financial qualifications, suggesting bias in algorithmic risk assessment.
- **SP2: Geographic and Racial Lending Disparities:** Mortgage application denial rates are higher for Black applicants in urban areas with historically redlined neighborhoods compared to applicants in more racially homogenous, rural counties, despite similar financial profiles.
- **SP3: Gender and Race Intersectionality:** Black female mortgage applicants might experience disproportionately higher denial rates and less favorable loan terms compared to both Black male applicants and White female applicants, even when controlling for financial metrics.
- **SP4: Algorithmic Bias in Credit Scoring:** Algorithmic credit scoring models used in mortgage underwriting disproportionately classify minority applicants as high risk, leading to higher rejection rates or unfavorable loan terms despite comparable financial standing.

2.2 Plan of Analysis

We will conduct analyses to address the research questions and explore potential biases in the loan approval process. First, we will compute and compare the loan approval rates for different racial groups. A Chi-square test will then be performed to evaluate whether any observed differences are statistically significant, helping us assess if race influences loan approval rates. We will also use a logistic regression model, controlling for financial factors (income, credit score, etc.), to estimate the effect of race on loan approval decisions. We will calculate the disparate impact ratio, comparing minority and White approval rates, and apply the EEOC's 80% rule to determine if potential discrimination exists. Our

analysis will also include investigating whether non-race variables, such as ZIP code or income, act as proxies for race, potentially leading to indirect discrimination. This analysis will be an exploratory last step. We will create an algorithm trained on historical data to examine fairness in machine learning models and feed it synthetic data. We will use fairness metrics (FPR, FNR) across racial groups to evaluate any biases found across racial groups, and we will stimulate changing an applicant's race to assess whether the predicted loan approval would differ. Through this analysis, we will explore racial bias in existing loan-approval algorithms. By creating our algorithm, we will assess if training a fair and accurate model is possible based on the current data available.

3 WORK AGREEMENT

As a group, we all agree that communication, transparency, and accountability are important to making sure we work well together. We'll keep in touch regularly through our group chat, where we can ask for help, share updates, and make sure everyone stays on track. Deadlines will be set and we'll do our best to stick to them, but if anyone has trouble meeting a deadline, we expect a 24-hour heads-up so we can figure it out together. We'll have a weekly check-in on Wednesdays at 3 in the library or a computer lab, and if we need extra time, we'll schedule it as we go. When it comes to getting work done, we'll divide tasks into smaller groups and check in with each other to make sure everyone has a chance to review and edit. We'll also make sure to comment on our code so everyone understands what's going on, and if we change someone else's work, we'll leave a note about what we did. We will practice pair programming when writing our code.

REFERENCES

[1] Mortgage Bankers Association. 2024. *Algorithmic Property Data Collection.* Technical Report. https://www.mba.org/docs/default-source/policy/mba_worksession_propertydatacollection_2.27.24.pdf?sfvrsn=f2caae76_1 Accessed: 2025-02-25.

[2] Investopedia. 2023. Mortgage Process Explained. (2023). https://www.investopedia.com/mortgage-process-explained-5213694 Accessed: 2025-02-25.

[3] Andre M. Perry, Jonathan Rothwell, and David Harshbarger. 2023. Homeownership, racial segregation, and policies for racial wealth equity. *Brookings Institution* (2023). https://www.brookings.edu/articles/homeownership-racial-segregation-and-policies-for-racial-wealth-equity/ Accessed: 2025-02-25.

[4] Local Housing Solutions. 2023. Why Housing Matters. (2023). https://localhousingsolutions.org/bridge/why-housing-matters/ Accessed: 2025-02-25.