Peer Review

1 Introduction

The introduction provides strong motivation for the research topic by describing the decline in affordable housing, stagnating incomes, and the critical role of homeownership for wealth building. It effectively summarizes the evolution from human underwriters to algorithmic decision-making and cites The Markup's 2021 investigation, which found significant racial disparities in mortgage denials. However, to deepen the reader's understanding and local relevance, they could quantify the housing trends with specific statistics (e.g., national or New York State homeownership rates over the past decade) and cite authoritative sources such as HUD or the Census Bureau. Incorporating New York State–specific historical context, for instance, referencing local redlining maps or state housing reports, would motivate the specific selection of the New York housing market. Furthermore, while the research questions are well stated, explicit hypotheses are not formally presented. A brief concluding sentence that states expected directions of bias (e.g., "We hypothesize that, controlling for financial indicators, Black female applicants will face lower approval rates and less favorable loan terms than white applicants") would clarify the study's expectations. Finally, a goal statement, separate from the research questions, was only first mentioned in the second line of the methods section (pg 3). However, adding this to the introduction would clarify the aim of this research much better earlier on in the paper.

2 Data

The Data section correctly identifies the 2023 HMDA Loan Application Register (LAR) dataset, acknowledges the CFPB and FFIEC as the data custodians, and lists the twenty selected variables, blending demographic and financial measures, used in the analysis They describe broadly how they removed records missing race information and coerced fields into numeric types, and they outline categorical coding decisions for variables like action_taken. To enhance transparency and reproducibility, the report should specify the original and final sample sizes (e.g., "Of X million initial records, Y were removed due to missing race values, yielding Z usable observations"). A formal

citation of the exact FFIEC URL, reporting period (e.g., January–December 2023), and any versioning details would strengthen the provenance of the data. A concise table summarizing key variables' means, standard deviations, and missingness rates would demonstrate that the dataset is both clean and tidy. Additionally, a brief codebook or appendix table defining each variable (including coding schemes and units) would aid readers unfamiliar with HMDA conventions. Clarifying why certain action_taken codes (1, 2, 7) were retained and others excluded, as well as any thresholds applied to continuous variables, would help readers' understanding of the data.

3 Methods

The plots on the distribution of demographic information and the distribution of loan application outcomes help us understand the composition of the data. However, additional visualizations for other characteristics of the data, that more closely correspond to the research question (for example, box plots of loan amount or interest rates by race, the overall rate spread distribution), would give readers a better understanding of the data. Overall, the visualizations are well-labeled and follow clear visualization principles.

In the second paragraph of the Methods section, when the authors state that race is recorded as a binary variable between white and black applicants, it's a good idea to mention which race is encoded as 1 and which as 0. This has been done well when discussing the categorization of loan outcomes into binary categories.

The goal of each model is well stated, but more information could be provided about the type of dependent variable (binary/ continuous/ categorical) and the control variable in each model. There is also a mention of the p-value used to assess the statistical significance, but the significance level is not stated. The paper goes on to talk about three classification algorithms, but it is unclear if it is the same three models previously mentioned, or if these classification algorithms are new models. The authors discuss how Figure 3 shows loan success rates by race, but the Figure 3 in the paper is a different plot than the ones they talk about. Likewise, the authors talk about the introduction of synthetic applicant profiles, but there is no discussion about how these profiles were created - was it synthesized based on existing profiles, or was entirely new data created, what were the key assumptions?

Except for the slight tweaks above, the methodology is well-described,

the justification makes sense, and the models are appropriate for answering the research questions.

4 Results

Using only the word significant could invoke confusion regarding whether it is statistical or practical significance. Also, what does it mean to be "the most significant predictor?" Does it imply the lowest p-value or the greatest magnitude of coefficients? Overall, the results section was concise and easy to understand. Maybe having model outputs or a model summary could have made things clearer, especially when drawing out statistical significance. This could also be in the appendix, but something that allows the reader to see how other variables could be affecting the results and outcomes of the model.

5 Discussion

The proposal of other research frameworks to improve the robustness and equity of the models was well thought out. This shows a recognition of the limitations of the methods used in the project to measure fairness and illustrates an awareness of current methods used to resolve these limitations.

The conclusions drawn from the models were also well done and accurately depict the results from the models' analysis. There is still a bit of confusion in the last sentence of the **Discussion** section. It says, " The strongest disparities appear at the pre-approval stage—before rates are finalized—suggesting that the initial decision to approve and the loan terms offered may be more prone to demographic bias than the later stages of the loan process." We wonder how this conclusion was drawn, as there doesn't seem to be any analysis done to distinguish between the pre-approval and post-approval stages in the paper. More insights might be needed to know how this conclusion was arrived at.

Moreover, most of the research questions are well answered in the **Discussion** section. It is clearly shown that there are significant differences in the loan interest rate of applicants based on their race and gender, even though this disparity was contrary to their hypothesis. Methods to improve fairness in the algorithm are also given, but while the importance of these methods is clearly stated, the methods are not explained in detail on how

they actually work. However, it is not clear whether the study quantifies loan approval rates. Concerns are raised about demographic bias in the pre-approval stage, which implies that there could be disparities in the approval rates based on race and gender, but this analysis is not clearly done in the paper. Some advice will be to replace "loan approval rates" in the research question with "loan amount" because that analysis was done and discussed in the paper. This way, the research questions will be fully answered in the paper. More so, it will be much easier to follow if each part of the Discussion section clearly states which research question it's answering. This will make the answers to the research questions more evident. Overall, this section was beautifully written.

The limitations of the analysis are clearly outlined. The assumptions made are clearly stated, and other flexible models that could resolve these assumptions were highlighted. The limitations in the data are also addressed, specifically the high number of white applicants and the high number of approved applicants. In addition, it is also beautiful to see the limitations in the real-life decision-making process addressed. The lack of visibility of the logic and structure of money lenders and the lack of ground truth are important factors to consider when trying to draw conclusions from the results of the paper

The ethics section is superb. It draws well from the limitations in the research to proffer a carefully stated ethical conclusion.

6 Analysis Code

The code used for the analysis is readable and written in proper Python style. Comments are added, and there is a good markdown description of the analysis. The overall methods used are correct. There is good use of categorical variables, and constants are added for the linear regression models, which shows good statistical practice. The predictor variables used were also useful for the research questions being explored. Some good additions might be to explain what FPR and FNR mean in the context of loan approval and why it's important to check these rates (as it may not be immediately obvious why it's needed in the analysis), but overall, it was well done. The diversity of model algorithms also made the analysis richer.

7 General

The report is very well structured. It was easy to read and follow, having been given descriptive variable names and a well-commented code that aligns with the idea of the research. Assuming the data and the setup are provided, it is possible to reproduce the analyses from top to bottom with all the plots made. The only minor suggestions were the mismatch between the figure callouts and the figures themselves, including the importance of having the in-text motivations for FPR/FNR, which may leave readers guessing the necessity of the error rates. The repository is organized perfectly, and our suggestion would just be to work on the minor things that were mentioned.

8 Final

The ethics discussion and being transparent with the methodological limitations stand out as one of the best works, as it tries to explain what is learned while also making sure the audience understands the limitations that the work has. The final question would be, if a model like this were live, would it be possible to regularly check whether it stays fair and that it doesn't treat any group unfairly?