DSC 167 Paper 2

Daniel Son
dson@ucsd.edu

Derek Leung
djleung@ucsd.edu

1 Introduction

In our previous work, we examined the potential for undeserved biases against certain racial groups in the United States that may stem from discriminatory attitudes and/or profit driven motives within past mortgage decisions. The mortgage data analyzed in the report is a compilation of public mortgage records that were filed in accordance with the Home Mortgage Disclosure Act (HMDA), enacted in 1975 to "monitor minority access to the mortgage market" (Munnell et al., 25) with the intention of eliminating barriers for those of lower income to "escape or improve or improve poor neighborhoods" (Munnell et al., 25). The results of this work revealed that White and Asian groups had much higher recall than other minority groups. This means there were proportionally less people rejected for a loan in these majority groups that should be qualified to get one. Conversely, this means that other, minority groups exhibit a higher likelihood to be rejected for a loan that should qualify to get one. Moving forward, this paper seeks to address the loan approval disparity by implementing the preprocessing technique of label flipping. Then, the results of our model will be evaluated on the same criteria of recall, as well as on the selective labeling approach of contraction.

2 Context

In our previous work, we speculate that a utilitarianism distribution scheme regarding the allocation of loans, seeks to provide loans based on the perceived utility, or the overall benefit, to the lender. In an effort to judge whether or not they will make money from the transaction, lenders may draw conclusions about a racial group based on generalizations as an oversimplified and biased way to judge their utility. We found that some conceptual evidence may include "that minorities typically are less able to rely on friends or relatives to help them through tough economic times" (Ladd, 46) or that specific racial groups have lower income on average which will impact their ability to pay back the loans. In our findings, while we would need further investigation to identify a specific cause or causes, it was clear that the proportion of loans rejected varied significantly by race.

Regardless of cause, this finding violates the Formal Equality of Opportunity which protects against unneeded use of traits, such as race, when judging people for an opportunity. While loans are advertised to be open to all, it has been discovered in many cases that race is being used to screen applicants since it is simple to judge whether an applicant will pay back a loan based on generalizations about the group that they are a part of. This is the motivation for our work using label flipping with a new model. By training a model after flipping loan approval and rejection between the most qualified people from minority groups who were rejected and the least qualified people from majority groups

Preprint. Under review.

who were approved, we hope to artificially tone down the impact of race in the decision making process.

One potential argument against this method is that there may be people who a lender truly believes to be unqualified for a loan, for which this new model will recommend loan approval. We justify this choice with Rawls' difference principle. Favoring an egalitarian distribution of goods, the Difference Principle states that inequalities within society are permissible only when it "make[s] the least advantaged in society materially better off" (Stanford Encyclopedia of Philosophy). Rawls' principles are protected in American society by several laws including: the Equal Credit Opportunity Act which prohibits discrimination in credit transactions based on "race, color, religion, With these factors in mind, the HMDA data sets should not display any demographic disparity that can be directly attributed to discrimination based on race.

3 Data and Model

Model Details:

This model is a Decision Tree Classifier developed by Daniel Son and Derek Leung. The current version is the second of the model and it incorporates fairness balancing techniques, specifically label flipping pre-processing in order to provide a more ethical loan granting algorithm that has a comparatively more balanced approval and denial rates between racial groups.

Intended Use:

The primary intended use for this model is to predict loan approvals based on the 2017 HMDA mortgage data set along with incorporation of fairness balancing techniques, such as label flipping. The intention behind this model is to provide a more fair framework for a mortgage prediction model for comparison purposes between what is used in reality, not to make official loan granting decisions. The target users are those who aim to do research regarding inequality within the HMDA mortgage data sets. This model is not intended to be used in place of the current system as the data set was limited in that it did not contain information important to loan decisions, such as credit score, debt history, and information if loans have been paid back or not. Due to this several inferences were made in order to analyze the data, including using income as a proxy for the ability of a borrower to pay back a loan.

Factors:

Significant group information within this data set includes racial, ethnic, and gender groups. The primary relevant factor that heavily affects the performance of the model—particularly recall—are the different racial groups. The evaluation factors are racial group and ethnic group. The labels as given from the HMDA data set were modified slightly, for our purposes, to take into account the minority ethnic group for the "White" race. This group was split into "White" and "Hispanic/Latino White". The other groups are as follows: Black or African American, Asian, American Indian or Alaska Native, and Native Hawaiian or Other Pacific Islander.

Metrics:

The evaluation metrics for this model includes: true positive rate, acceptance rate, and failure rate to measure unequal model performance between racial groups, as well to compare the performance of the models with and without fairness balancing techniques. In this specific data set, the true positive rate measures the number of people who got a loan (positives) out of all the people who should have qualified for a loan (all positives). The measure of true positive rate can be calculated from

the confusion matrix generated by sklearn based on the model's binary predictions of "Approved" and "Not Approved" with respect to loans. The acceptance rate and failure rate are used for the contraction fairness evaluation technique to compare the model's before and after label flipping performance as shown in Figure 4. The performance is judged by comparing similar acceptance rate groups' failure rate. Maximizing acceptance rate and minimizing failure rate is the ideal goal, so for each racial group, the maximal equal acceptance rate of the ground truth and label flipped model was found and the failure rate was reported for each.

Training Data:

The breakdown of racial groups in the data set can be seen in Figure 1. It can be seen that there is significantly more entries in the "White" category, so to equalize the representation within the training data, ten thousand entries from each race was randomly sampled. The number of selected observations was limited by the size of the smallest group, which was "Native Hawaiian or Other Pacific Islander". The data from the 2017 HMDA mortgage data set was then split with an 80:2 ratio for the training and test sizes for each race. The training set without the random sampling would have contained 6945201 observations while the testing set without sampling would have contained 1736301 observations. With sampling to equalize the representation of racial groups, there were 7591 training observations and 1898 testing observations for each race. The columns used for the model were: "tract_to_msamd_income", "loan_amount_000s", and "applicant_income_000s".

Evaluation Data:

The test data to check the models performance was generated as described above in the training data section. There were 7591 randomly selected observations for each race. Each race was evaluated separately due to differing distributions of income within each group, as illustrated in Figure 2. It can be seen that Asian, White, and Hispanic/Latino White have the smallest spread of income between the approved and not approved groups. All the other minority races have a significant larger spread between the means of the approved and not approved groups. Having a model for each race would allow the model to find trends unique to each race. The random sampling to equalize the representation of each racial group is important due to the extremely different representations in the original data set. Since our primary focus is to evaluate the demographic parity, we want all groups to be equally represented in a fair model.

The data was preprocessed by standard scaling the numerical values. The outcome column which was originally a binary column with "Approved" and "Not approved" as its unique values was converted to a binary column with 1's and 0's. For other categorical columns, those values can be one-hot encoded before being passed in for evaluation.

Ethical Considerations:

As mentioned in the Intended Use section, this model is to be used for research purposes and should not have any real life mortgage decisions based off its predictions. Although this model takes steps to encourage fairness between demographic groups through label flipping and representative sampling, there are too many factors considered in the application process that are not present in the given data to make an accurate prediction. Misuse of this model could potentially sentence applicants to a non-representative decision of whether they should receive a loan or not, which in turn can affect the course of their life if they are not able to receive a life-changing loan that they deserve. Additionally, due to a lack of data and context, we do not consider this work enough to classify any particular lender or lending decision individually as unfair. Results are visualized and interpreted based on observed trends and should be treated generally as such.

Caveats and Recommendations:

In order to make this model more accurate for potential real-life implementation, it would require much more data related to the applicant screening process for mortgages, including if the loan was paid back or not, debt history, credit score, etc.

Quantitative Analysis:

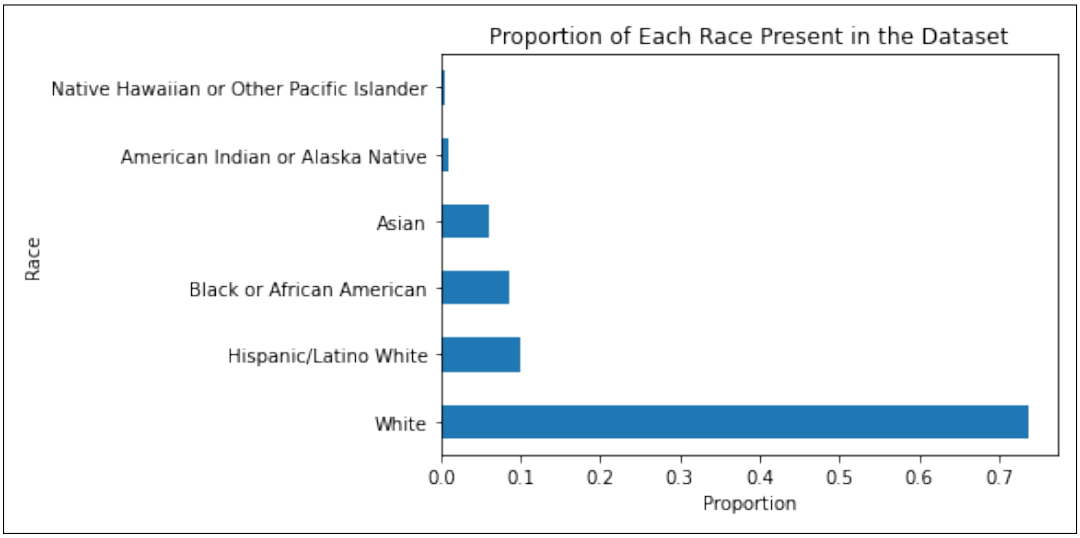

Figure 1: Proportion of Racial Groups Within the 2017 HMDA Dataset

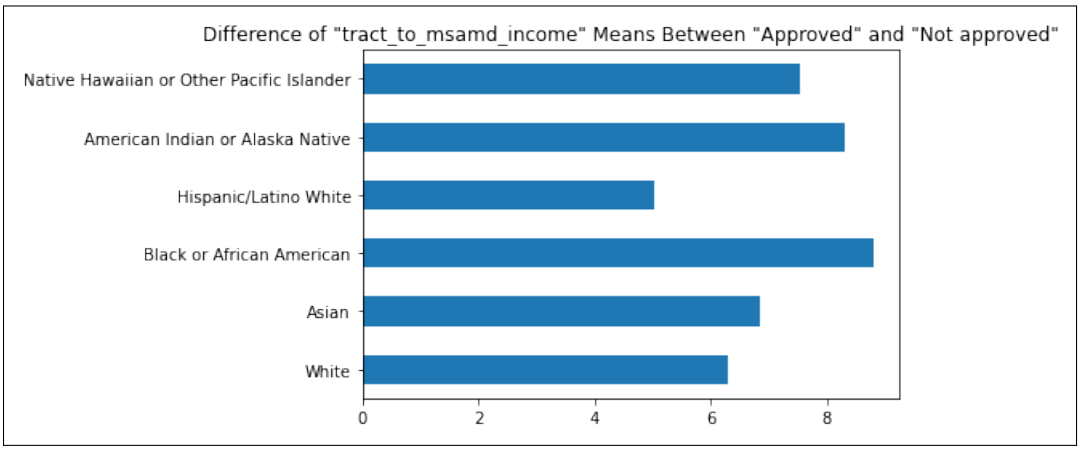

Figure 2: Proportion of Racial Groups Within the 2017 HMDA Dataset

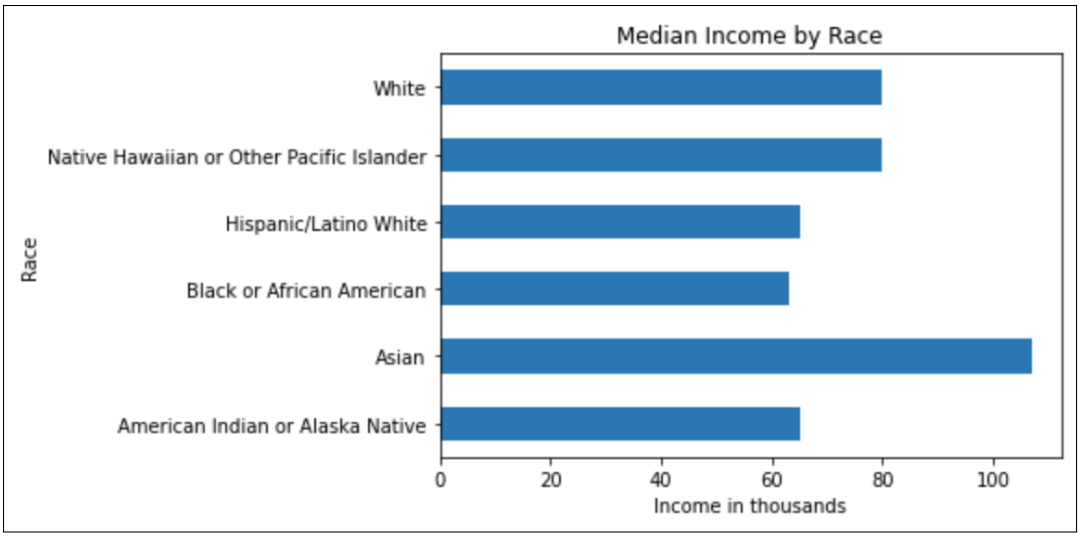

Figure 3: Median Income for Each Racial Group

4 Fairness Metric

Following up with our fairness evaluation from our previous work, we measure true positive parity, or recall, across racial groups. To reiterate our reasoning, the examination of demographic disparity was chosen to determine if the HMDA data set would uphold the ideals put forth by Rawls' Difference Principle which are also reflected in the American laws passed, such as the Equal Credit Opportunity Act, with the purpose of allowing minorities access to equal opportunities of resources including mortgage loans. The metric of true positive parity was chosen because this investigation focuses on a primary harm of false negatives, those who are justified to be given a loan being denied, which may suggest the presence of racial discrimination between similarly qualified groups. We should also note here that for this specific investigation, we used the true outcomes of loan approval and denial from the HMDA data set as a ground truth. In practice, however, both true negatives and false negatives are invisible. More specifically, if we do not grant someone a loan, we are unable to see if they pay it back. This fact is notable as justification for our use of the selective labeling approach of contraction later in this section.

Here, we will evaluate both the original decision tree model from our previous work and our new model which incorporates label flipping to examine the effect the fairness-based preprocessing techniques had. These are both decision tree classifiers which uses income percentile by tract, loan amount, and flat income in order to predict whether a loan was approved or denied. As per the fairness metric described earlier, recall, we evaluate our results from the model.

In the original, pre-label flip, model we found that White and Asian applicants are treated most generously, with the best recall at .8518 and .8568 respectively. Native Hawaiian or Other Pacific Islander and Hispanic/Latino White were worse off than the previous group at .8004 and .8059 respectively. Lastly, Black/African American and American Indian or Alaska Native applicants are worst off at .7226 and .7264 respectively. It is notable that the Hispanic/Latino White group had a lower recall score than the White applicant group.

After label flipping, we see an overall improvement, but with some clear trade off. Two of the minority groups, namely Black/African American and Native Hawaiian, see a clear increase of 0.0156 and 0.0238 to 0.7383 and 0.8243 respectively. Two other minority groups, namely American Indian or

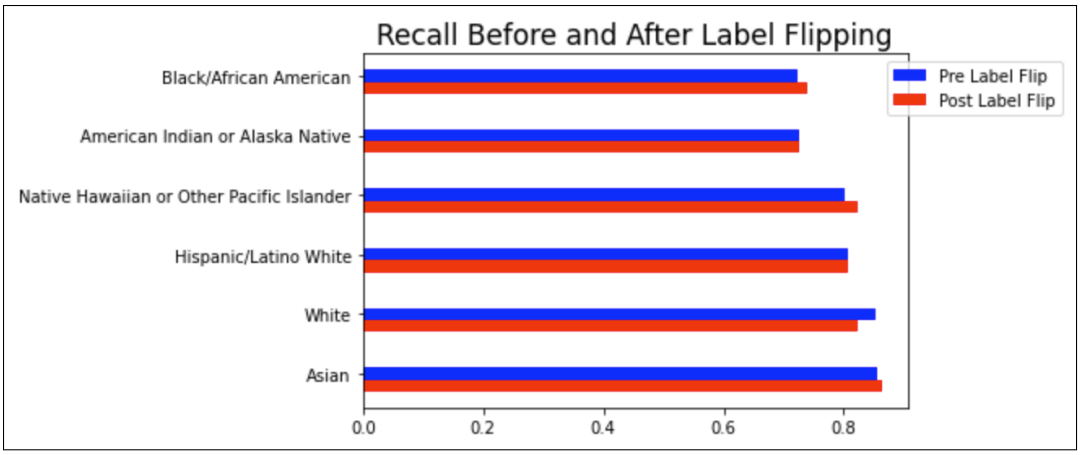

Figure 4: Recall by Racial Group Before and After Label Flipping

Alaska Native and Hispanic/Latino White, see almost no difference. A 0.0025 decrease to 0.7239 and a 0.0004 increase to 0.8063 respectively. Lastly, our majority groups exhibit an interesting adjustment. White applicants experience a 0.0282 decrease in recall to 0.8236 while Asian applicants experience a 0.0062 increase to 0.8630. A further look into the model suggests that this may occur because White applicants average a lower income than Asian applicants, which is illustrated in Figure 3. This difference may indicate that among those approved for a loan, the lower income White applicants were evaluated to be more risky, so were demoted, or their label was changed from approved to not approved, at a higher rate during the label flipping process. From this analysis, we see that the metric of recall may still leave much to be desired, but there is an overall improvement to the demographic parity of the model regardless.

Next, we want to evaluate the the performance of the model further, beyond simply examining recall amongst racial groups. This is done using the selective labeling approach of contraction previously mentioned. In this section specifically, we visualize how well the model performs by plotting acceptance rates and failure rates. Here, we see the acceptance rate, the proportion of each group that the model decision is to approve a loan, and the failure rate, the proportion of incorrect decisions to approve a loan.

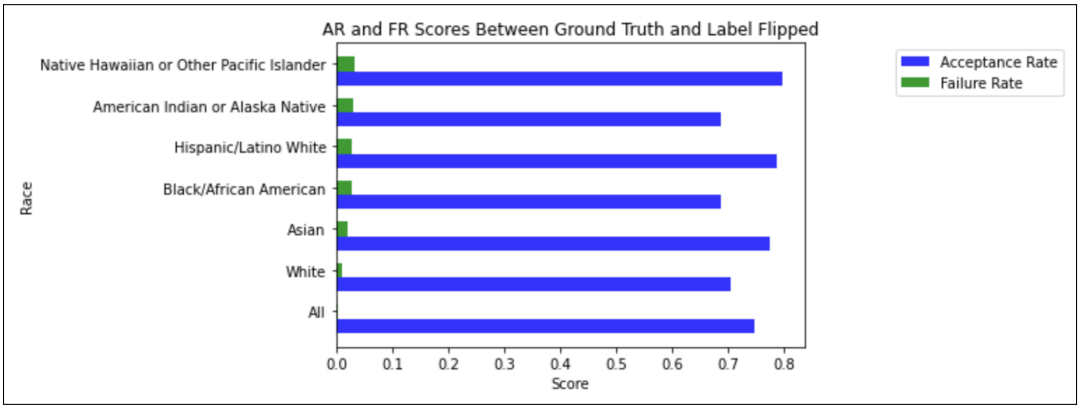

Figure 5: Acceptance Rate and Failure Rate Scores Between Label Flipped Model and HDMA Ground Truth Data

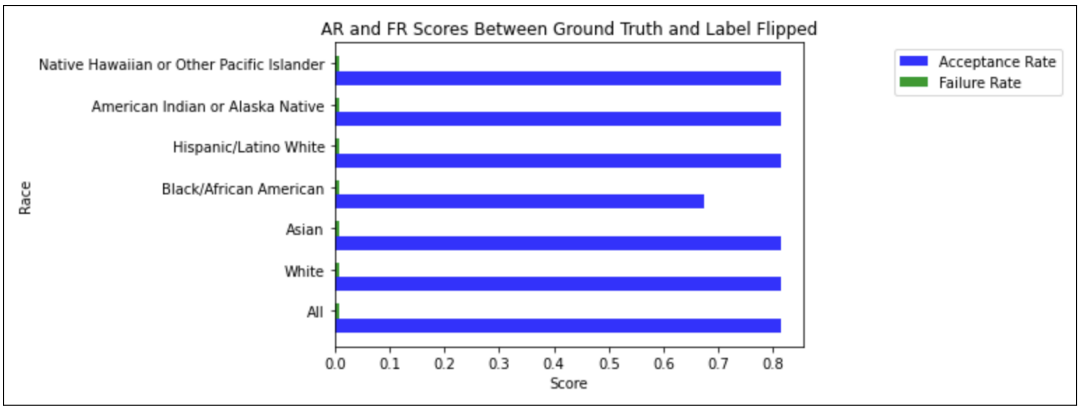

Figure 6: Acceptance Rate and Failure Rate Scores Between Label Flipped Model and HDMA Ground Truth Data

Comparing the results of the label flipped and non-label flipped acceptance and failure rates, we can conclude that there is a trade-off between failure rate and demographic parity. In initial appearance at least, it seems like we sacrifice some accuracy in correct loan decision to achieve fairness across groups, making more incorrect labeled decisions with less represented groups. However, it is very important to note in this scenario that "correct" labeled decisions are purely based on the HDMA data set decision. In practice, we still do not know the actual result of whether or not those were were approved for a loan actually paid it back. Overall, we believe that it may be worth considering that this small accuracy trade off could be worth the increase in proportional equality between racial groups.

5 Results and Interpretation

The goal of the classification model generated for this report is to generate a fair prediction in terms of demographic parity given the 2017 HMDA housing mortgage data set. Comparing the proportion of approved and not approved outcomes, as illustrated in Figure 7 and Figure 8, it is clear that the preprocessing techniques, particularly label flipping, did in fact have a positive effect in promoting equality between demographic groups in loan decisions. It can be seen that the bars representing the proportions of each decisions after label flipping (red bars) have a much tighter distribution compared to the bars for the outcomes without label flipping, which proves the point that the preprocessing techniques aided in promoting demographic parity, or equality of positive outcomes between demographic groups.

Although this equalization came with a trade off: the approval rates for Asian and White races dropped significantly in exchange for all the other minority races' rates increasing, it can be seen from Figure 7 that both the races whose approval rates decreased, White and Asian, were better off in terms in terms of loan approval before label flipping, both boasting more than 80% of all their applications being approved for a loan. This trade off, is justified by Rawls' Difference Principle, which promotes an egalitarian distributions of goods by benefiting the worst off groups in society. In this case, the worst off groups—Black or African American, Hispanic/Latino White, American Indian or Alaska Native, and Native Hawaiian or Other Pacific Islander—had increased access to the good the model is distributing (loans) while the best off—White and Asian—suffered an inequality through their reduction of loan approval rates. Such inequality is exactly what Rawls' theory states is ethically permissible since the best off sacrificed resources to benefit the worst off groups. Through

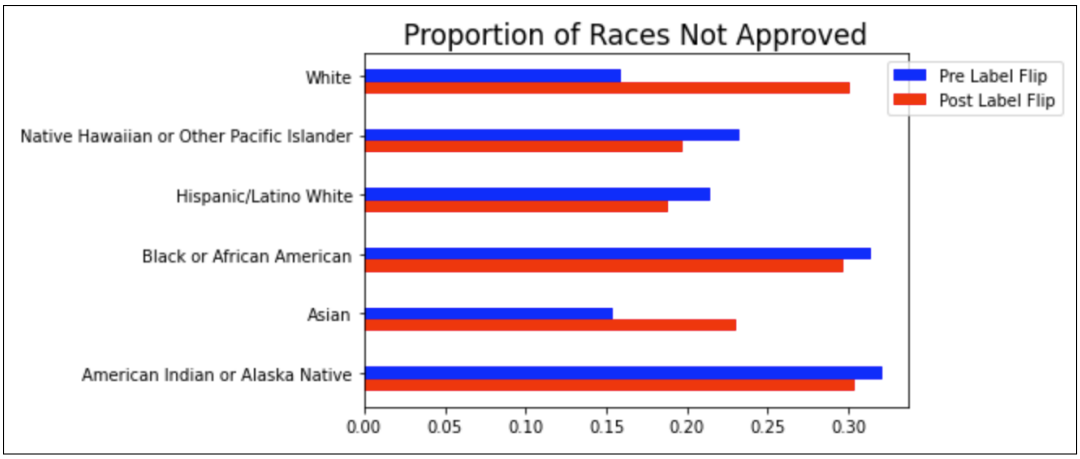

Figure 7: Proportion of Denied Loans Before and After Label Flipping

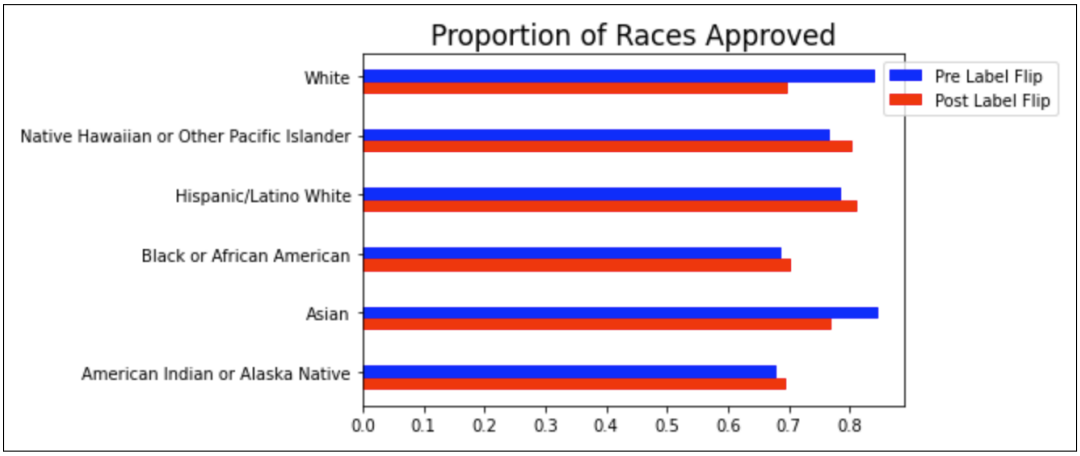

Figure 8: Proportion of Approved Loans Before and After Label Flipping

the redistribution of approved loans by the fairness tuned decision making model, it can be seen from the results that this model does indeed promote demographic parity through the equalization of loan approval and denial rates between all racial groups present in the dataset.

References

Ladd, Helen F. “Evidence on Discrimination in Mortgage Lending.” Journal of Economic Perspectives, vol. 12, no. 2, 1998, pp. 41–62., https://doi.org/10.1257/jep.12.2.41.

Lamont, Julian, and Christi Favor. “Distributive Justice.” Stanford Encyclopedia of Philosophy, Stanford University, 26 Sept. 2017, https://plato.stanford.edu/entries/justice-distributive/.

Munnell, Alicia H. Mortgage Lending in Boston: Interpreting HMDA Data - JSTOR HOME. https://www.jstor.org/stable/2118254.