Building A Fair Machine Learning Model Using Reduction & Threshold Techniques For Loan Acceptance Prediction

By: Aurelio Barrios

1 Introduction

Fairness evaluation in machine learning algorithms has gained traction in recent years due to great informative discoveries of inequities present in real world applications of machine learning. These discoveries show that inequities or biases can arise in different ways. There are instances where the machine learning algorithm is creating an inequity by correlating one sensitive factor with another factor and there are even instances where the data itself is representative of human biases present in the real world. This shows us that even those machine learning algorithms that may appear to be fair can be considered unfair and therefore must be evaluated. Although it is difficult to evaluate the fairness of a machine learning algorithm it is absolutely essential. This must be done in order to ensure that we are not discriminating against minorities, especially when these algorithms are distributing resources or are making life changing decisions such as determining the risk of an inmate reoffending.

It is very important to define what fair means to an algorithm that is producing positive labels. This is not to say that just because distinct protected groups have different rates of positive labels that a model should be considered unfair. There are cases where unequal distribution of positive labels have real world justifications such as in the example of breast cancer by gender. In this example it has been shown that females do in fact have a higher risk of developing breast cancer than males. In our specific case we consider fairness with a machine learning algorithm that predicts whether an individual gets approved for a loan or not. When considering a factor like race it should be considered unfair for a model to have significantly different positive label proportions across different races. This is because race should not be a clear risk factor when determining whether to approve an individual for a loan and to say otherwise is extremely discriminatory.

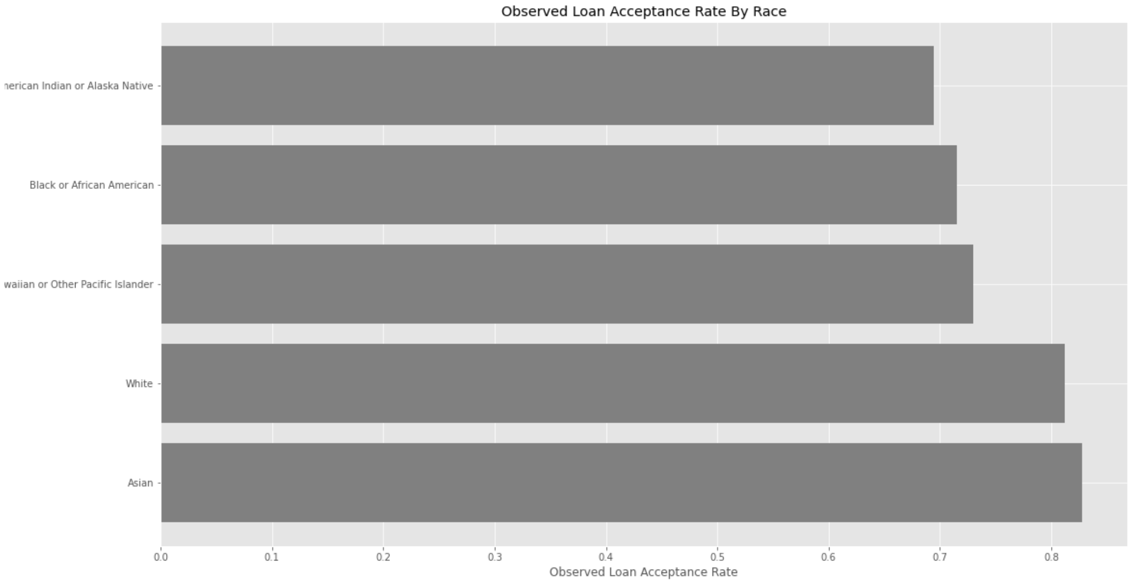
The aim of this paper is to perform accurate binary classification (loan acceptance) while satisfying the constraints required to consider a model fair with respect to sensitive attributes like race and sex. To do this we will rely on in-processing and post-processing techniques like reduction and threshold optimization, under fairness constraints like equalized odds and demographic parity. Furthermore performing intersectional analysis to understand how our model under fairness constraints relates to both gender and race.

2.1 Problem - Data Unfairness

As mentioned before when considering loan acceptance rates between people of different races there should be no significant difference. When we consider a machine learning algorithm that outputs different acceptance rates for different races we are depriving said races with low acceptance rates from a resource and their opportunity to buy a home. This is known as an allocative harm since there is an unjust distribution of a certain resource, in our case loans. A fair model should aim for an equal distribution of resources for those deserving of the resource.

Data may be representative of real world bias. When considering the California Home Mortgage Disclosure Act (HMDA) dataset for the year 2017 we see an example of data capturing real world bias. In the plot below we measure the observed loan acceptance rate by race for our data sample.

Figure 1: Observed Loan Acceptance Rates By Race



As we can see from the plot above there is a clear difference in loan acceptance rates amongst different races. We see that individuals who are White and Asian have an observed loan acceptance rate greater than 80% while individuals in minority groups like American Indian or Alaska Native, Black or African American and Native Hawaiian or Other Pacific Islanders have an observed acceptance rate of around 70%. We expect the observed loan acceptance of individuals across different races to be roughly equal. The differences shown here must be considered if we are to build a fair model, there must be open consideration that this dataset might be capturing real world bias. Now although there are early signs of unfairness it is hard to measure what is fair with just the naked eye.

2.2 Model Building With Unfair Data

In order to determine unfairness we must have quantifiable measurements. These measurements are Demographic Parity which states that acceptance rates of individuals of distinct groups must be equal and Equalized Odds which states that individuals of protected and unprotected groups have equal true positive rates and false positive rates. The calculated measurements used here are demographic parity difference, equalized odds difference and demographic parity ratio. Traditionally for demographic parity difference and equalized odds difference if the absolute value of the calculated difference is smaller than 0.1 then we can consider the model in question to be fair. For demographic parity ratio we traditionally want this value to be in between 0.8 and 1.25 for the model to be considered fair.

In order to begin, the data is split into a training set of 54,381 observations, a testing set of 18,128 observations and a validation set of 18,128 observations. Then a Decision Tree Classifier is built using the training set to train the model, while using the testing set to evaluate the models performance in terms of accuracy. After compiling the machine learning model by gathering the predicted outputs we can compute our unfairness measurements. The table below outlines the fairness measurements for a Decision Tree Classifier trained on the original dataset.

Table 1: Parity Measurements For ML Model On Original Data

Unfairness Measurement	Value	Threshold Result	Fairness Result
Demographic Parity Difference	0.18	> 0.1	Unfair
Equalized Odds Difference	0.2	> 0.1	Unfair
Demographic Parity Ratio	0.78	< 0.8	Unfair

As we can see from the table above according to our fairness measurements our model cannot be considered fair. Our values for Demographic Parity Difference and Equalized Odds Difference are not below the 0.1 threshold needed to be considered fair. Although our

Demographic Parity Ratio is close to the threshold we also cannot consider it to be fair. Overall the initial machine learning model is poor and unfair.

By exploring the outputs of the machine learning model we can get a better sense of the unfairness that is being produced. As mentioned before machine learning models heavily rely on the data and therefore if there are any biases in the data this will certainly influence the machine learning models learning. Although the goal of a model is not to be unfair, the data used may result in an unfair model regardless of initial intention.

Figure 2: Initial ML Model Error Rates For Test Set Predictions

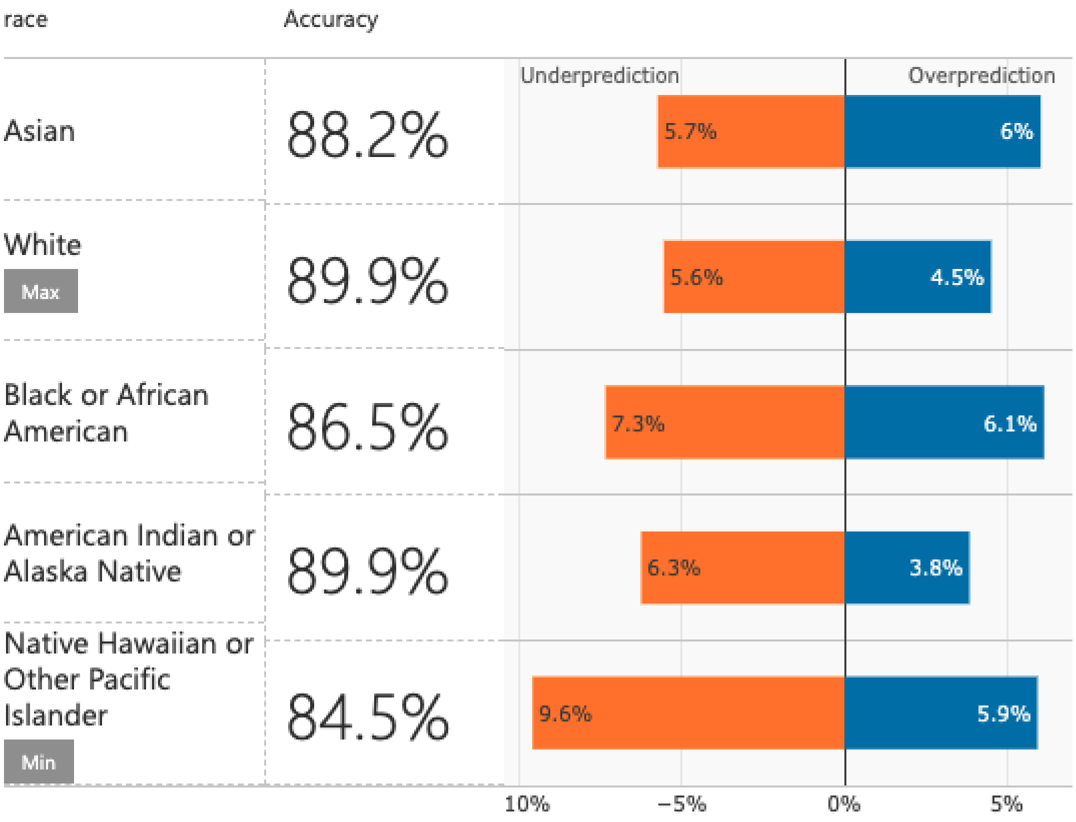
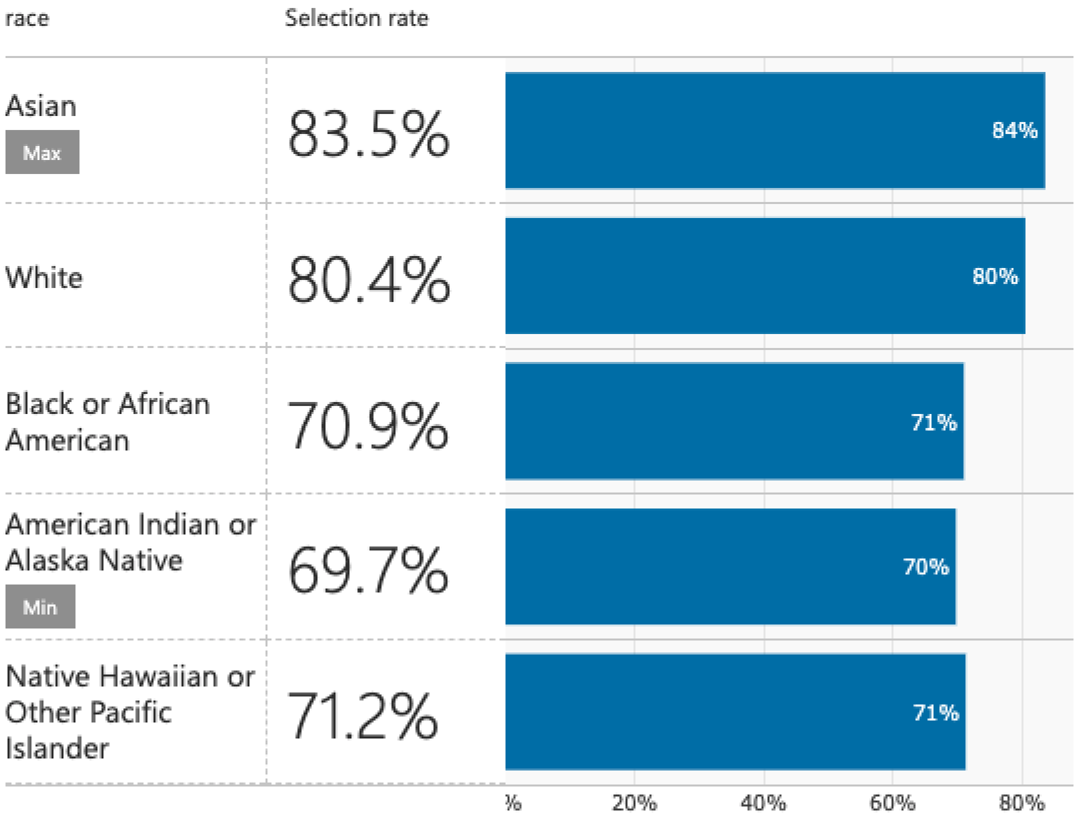


Figure 3: Initial ML Model Selection Rates For Test Set Predictions



After training our model with the training data, it was time to evaluate the model through prediction of the test set. As we can see from the figures above our model learned the biases present in our original data. We begin by viewing the selection rates amongst different race groups and see that the initial biases found before building our model are still present. White and Asian individuals maintain a selection rate of around 80% while the remaining races maintain a selection rate of around 70%. Although this model achieved an accuracy of 89.6% on the test set we see how there are some clear unfair predictions. For example, this model had the highest underprediction rates for the Native Hawaiian or Other Pacific Islander (9.6%) and Black or African American (7.3%) racial groups, which also made up the groups with the lowest accuracies. This shows us how these groups are disadvantaged, around 10% of Pacific Islanders would not have gotten approved for a loan even though they deserved one. This is similar with African Americans, where around 7% did not get approved even if they were deserving of the loan.

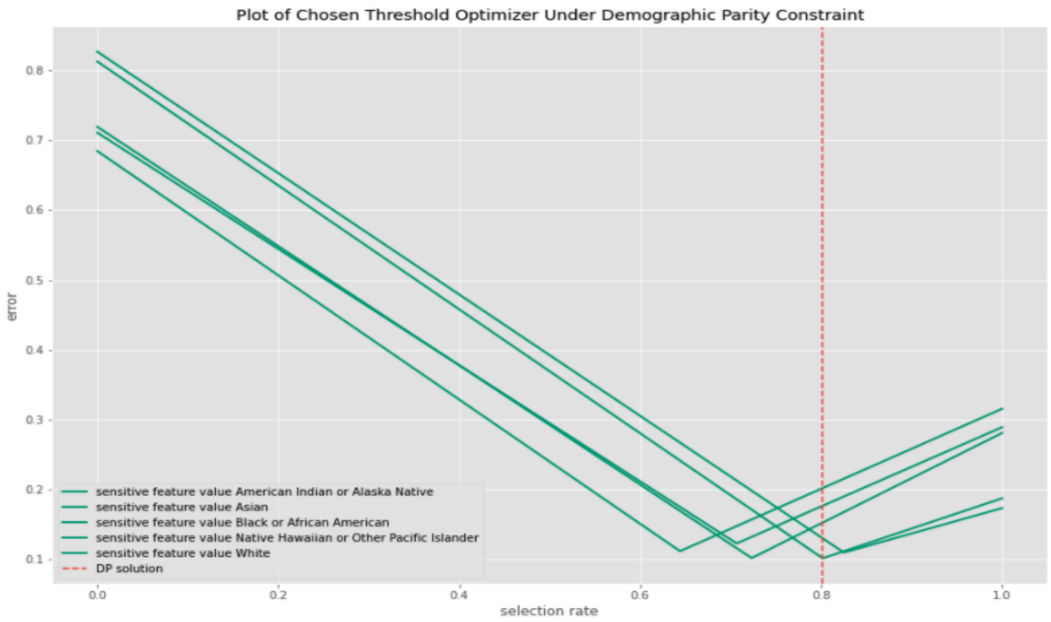
3.1 Model Building Under Fairness Constraints

In order to build a fair model we must build it according to some fairness constraints. Using the Exponentiated Gradient reduction technique with Equalized Odds and Demographic Parity constraints as well as the Threshold Optimizer post processing technique with Demographic Parity constraints we get better, fair estimates for our machine learning algorithm. We outline 4 different models below that each have different fairness constraints and measure their utility and fairness on the testing set.

Table 2: Disparity and Accuracy Outputs For Each Constraint On Test Set Predictions

Constraint Name	Disparity In Predictions	Accuracy
None/Original	13.8	89.4
Threshold Optimizer	6.97	88.8
Equalized Odds	9.03	86.6
Demographic Parity	3.01	84.5

Figure 4: Threshold Optimizer Solution



As we can see from the figure above our threshold optimizer chose a threshold of 0.8 as its solution. This yielded a good balance between a high accuracy and a low disparity in our predictions. The threshold optimizer performed nearly the same accuracy as the original model on the testing set while nearly cutting the disparity by a factor of 2. When we used the exponentiated gradient reduction technique with the equalized odds constraint we got lower accuracy than the optimizer and a higher disparity. The lowest disparity found was reduction with a demographic parity constraint with 3.01 disparity in our predictions but this came at a cost which is reflected on the accuracy of 84.5% on the testing set which is the lowest.

3.2 Model Selection

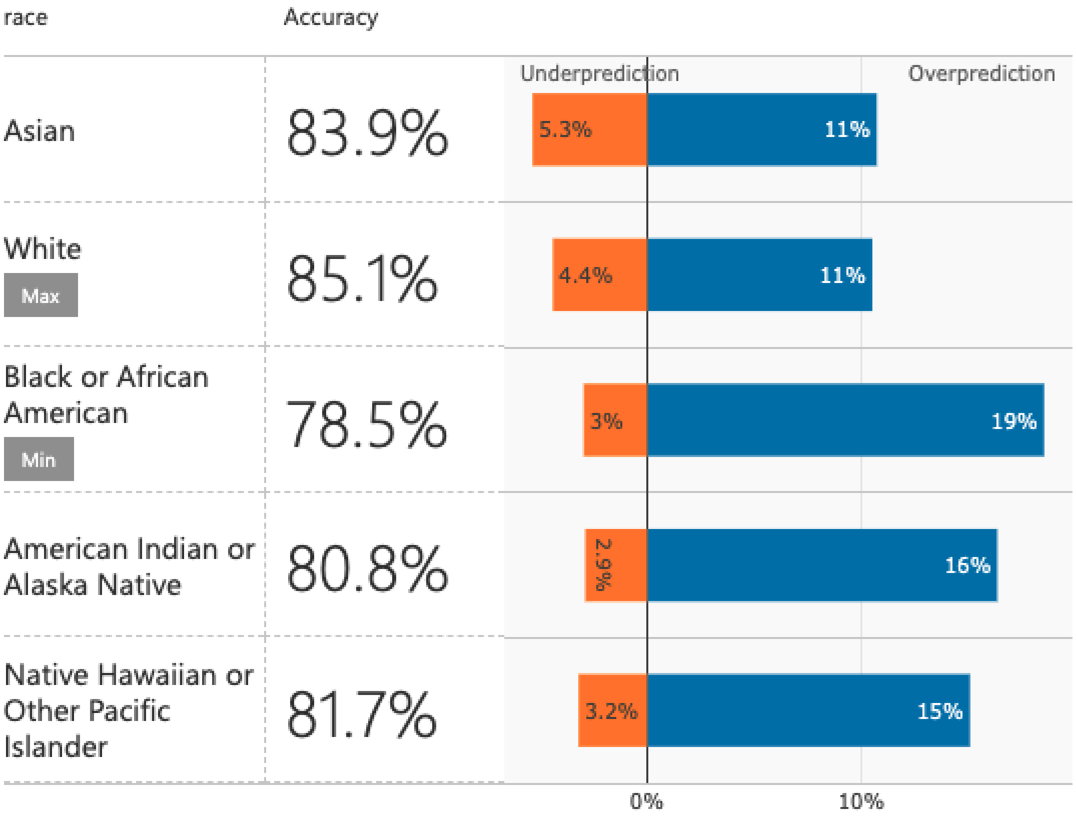
Overall considering that the demographic parity decision tree model had the lowest disparity while also achieving a decent accuracy score this model seemed the best. This model sacrificed accuracy for disparity, but we can evaluate the predictions further to see how this will affect our decisions. As we can see from the table below we meet and match all of the thresholds that were previously not met with the original model with no fairness constraints.

Table 3: Parity Measures For Demographic Parity ML Model

Unfairness Measurement	Value	Threshold Result	Fairness Result
Demographic Parity Difference	0.03	< 0.1	Fair
Equalized Odds Difference	0.1	= 0.1	Borderline Fair
Demographic Parity Ratio	0.97	> 0.8 and < 1.25	Fair

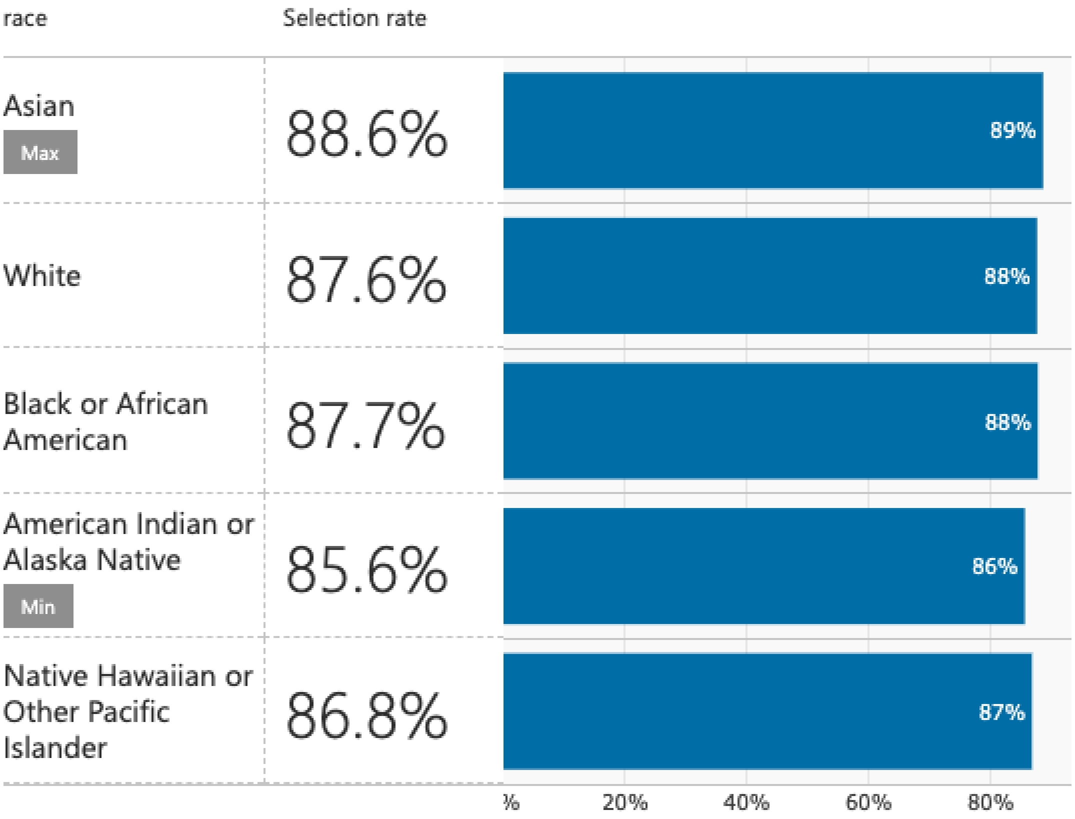
From the figures below we can see how this model achieves its fairness. There is a big difference in the selection rates when compared to the original model. As we can see all groups have roughly the same loan approval rate. Not only this but we see an increase in the approval rate for all different groups of race whilst maintaining fairness. We can also see that underprediction has also decreased for all groups, while overprediction increased for the minority groups. When this model is underpredicting we are denying an individual of an opportunity they deserved, while when we overpredict we are providing an opportunity to an individual who otherwise would have not gotten the opportunity. This model prioritizes making sure that those who deserve a loan are getting a loan. Furthermore it is breaking the cycle of maintaining inequity in the distribution of loans per race by making sure that all races have a similar loan acceptance rate.

Figure 5: Demographic Parity ML Model Error Rates For Test Set Predictions



race	Accuracy	Underprediction	Overprediction
Asian	83.9%	5.3%	11%
White Max	85.1%	4.4%	11%
Black or African American Min	78.5%	3%	19%
American Indian or Alaska Native	80.8%	2.9%	16%
Native Hawaiian or Other Pacific Islander	81.7%	3.2%	15%

Figure 6: Demographic Parity ML Model Selection Rates For Test Set Predictions

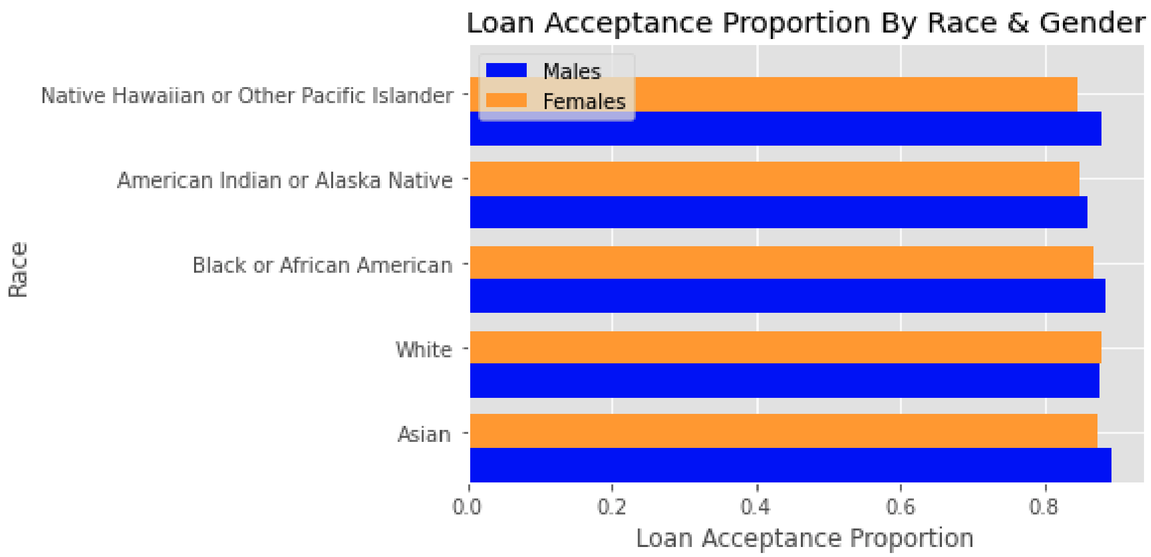


race	Selection rate
Asian Max	88.6%
White	87.6%
Black or African American	87.7%
American Indian or Alaska Native Min	85.6%
Native Hawaiian or Other Pacific Islander	86.8%

3.3 Model Evaluation - Intersectional Analysis

We have seen how the model performs under the protected attribute of race, but we can further explore how this model performs with race and gender. We have achieved an equality for loan acceptance between individuals of different race but it would be interesting to consider the differences between another sensitive attribute in sex for each racial group. As we can see from the figure below, loan acceptance by sex is fairly equal. Loan acceptance for Native Hawaiian or Other Pacific Islander seems to show the greatest difference in loan acceptance by gender. Overall we should expect to see roughly the same loan acceptance predictions for males and females consider that similarly to race, sex should not be considered a key determinant of wether to approve a loan or not.

Figure 7: Loan Acceptance Race & Gender



4 Conclusion

We have seen how data can represent real world bias. This ultimately flows through the machine learning pipeline affecting the predictions that your machine learning algorithm makes. In a real world scenario where these algorithms are being used as decision makers we can see how this would become an issue. The biases found previously are influencing new biases which will then influence the biases further down in the future. This feedback loop would go on forever magnifying the bias present. Through the use of reduction techniques we have managed to not only produce a machine learning algorithm with good accuracy but also an algorithm that is fair. We see equality in loan acceptance for individuals of all different races. We also see a decrease in underprediction for all races, which allows for more allocation of resources to individuals that deserve it. We also see how loan acceptance maintains fairness when considering individuals of different sex when compared between the same race and of different race. Overall it was shown that it is possible to create a model that is fair while maintaining a respectable prediction accuracy on data that represents bias found in the real world.