Improve Fairness in Credit Card Approvals

Chinmay Wadnerkar[*]
cwadne2@uic.edu
University of Illinois at Chicago
Chicago, Illinois, USA

ABSTRACT

Over the last couple of decades our dependency on machine learning and AI systems has increased exponentially increased. Every aspect of life is touched by these systems We rely on these machines to provide us with accurate and objective answers, thinking that they are faithful. But sometimes because of reasons discussed further in this report we get models which are not fair or faithful.

This course project for *CS594* is all about finding and trying different solutions to introduce or improve fairness in our machine learning models.

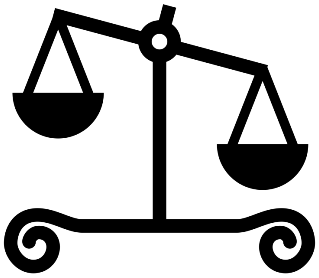


Figure 1: Fairness **(Source - Georgia Institute of Technology)**

1 INTRODUCTION

As mentioned earlier every aspect of our life is touched by machine learning models and AI systems that have been employed in almost every field. Banks are no exception. They use machine learning models to predict if a person has probability to default on their card payments and thus should they be given a credit card or not. It can be a formulated as a classification problem.

Sometimes the data used for training these models can be biased thus causing the models to give out biased output reducing their fairness and trustworthiness. For this course project I want to check if the Credit Card approvals based on defaulting payments and GENDER are related. If they are related, then the model is not fair and we need to find ways to make them fair.

Sensitive attributes should be given special consideration and, in most cases, should not be linked to a specific entity. Some examples are race, age, education, sex, marital statues etc. As all data contains some bias, the machine learning models are usually trained on these skewed samples and can learn to link these sensitive attributes to the target label. The sensitive attribute column in my dataset is GENDER and the target label is the DEFAULT PAYMENT COLUMN.

The main goal of my project is to induce fairness in my models to reduce the disparity between the genders while still giving out a relatively accurate model.

The dataset contains 30000 samples that has 24 attributes, but I am currently focusing on the GENDER column and DEFAULT PAYMENT column. Default payment has binary values (1 meaning that the user will default the next payment and 0 meaning that the user will make the next payment).

2 METHOD

In this section I'll discuss about the various methods that I used to train the models, methods to show the fairness of the models and then finally the method used to improve the fairness of the model.

2.1 Methods to train the model

The methods I used to train this classification model are listed below:

- RandomForestClassifier - It fits a number of decision tree classifiers on various samller parts of the dataset and uses averaging to improve the predictive accuracy.
- AdaBoostClassifier - It fits a classifier on the original dataset and then fits additional copies of the classifier on the same dataset. It then adjusts the weights of the incorrectly classified instances making sure that the next classifiers focus more on the difficult samples.
- DecisionTreeClassifier - It simply builds a tree where each node is a test on the given attribute.
- SVC - It maps data points to a high-dimensional space and then finds the best hyperplane that divides the data into two classes.

2.2 Methods to show the fairness of the model

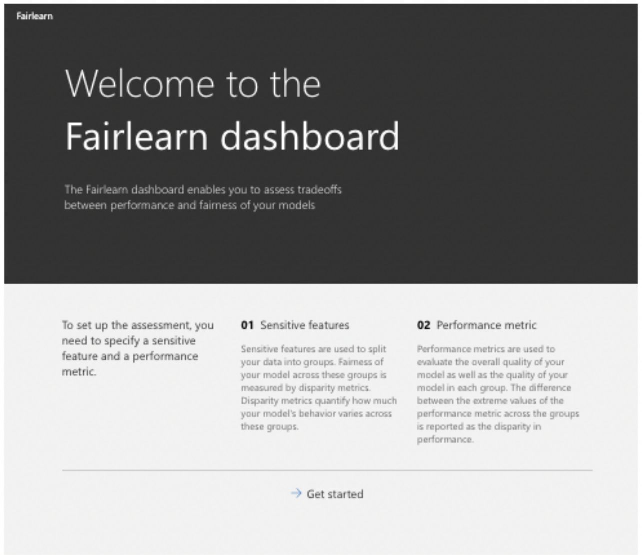
I used a package called Fairlearn in order to visualize and improve the fairness of the model. Fairlearn has a module called widgets and this module contains the Fairlearndashboard. We need to pass the sensitive attributes, the name for the sensitive attributes, the

[*] Worked on this course project independently

ground truth label and the predicted outcome as input parameter to the dashboard method. The output is an interactive dashboard that lets us select multiple sensitive groups and the metric that we want to visualize. I decided to go with accuracy for this project.

I used the Fairlearn DashBoard to give out the current accuracies and disparity between the selected sensitive attributes for both the algorithm.

The dashboard looks like in the images attached below:



Welcome to the Fairlearn dashboard

The Fairlearn dashboard enables you to assess tradeoffs between performance and fairness of your models

To set up the assessment, you need to specify a sensitive feature and a performance metric.

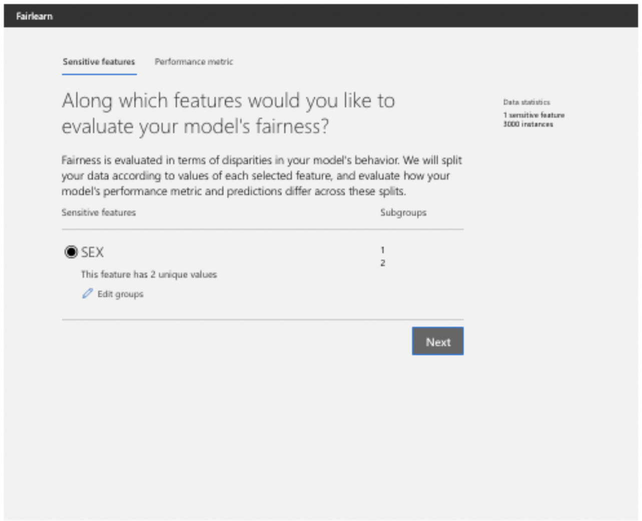
01 Sensitive features

Sensitive features are used to split your data into groups. Fairness of your model across these groups is measured by disparity metrics. Disparity metrics quantify how much your model's behavior varies across these groups.

02 Performance metric

Performance metrics are used to evaluate the overall quality of your model as well as the quality of your model in each group. The difference between the extreme values of the performance metric across the groups is reported as the disparity in performance.

Figure 2: Fairness Dashboard Home screen



Along which features would you like to evaluate your model's fairness?

Data statistics
1 sensitive feature
3000 instances

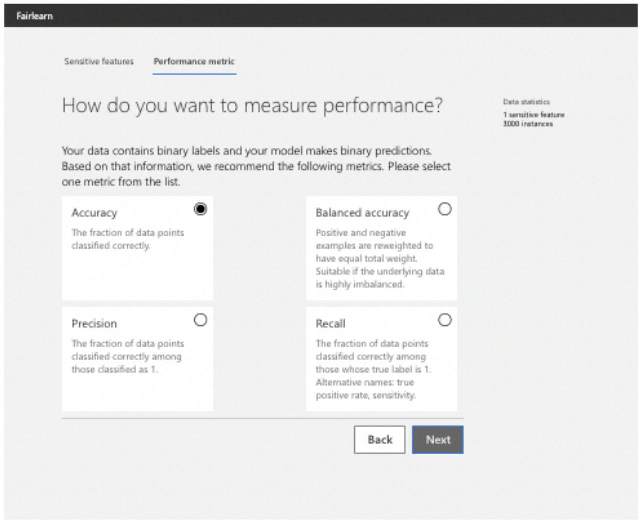
Fairness is evaluated in terms of disparities in your model's behavior. We will split your data according to values of each selected feature, and evaluate how your model's performance metric and predictions differ across these splits.

Sensitive features	Subgroups
● SEX This feature has 2 unique values Edit groups	1 2

Figure 3: Sensitive attribute selector

2.3 Methods to improve Fairness

To reduce the disparity in accuracy, I used a post processing algorithm which takes as input an existing classifier and the sensitive features and derives a monotone transformation to preserve the order of these outcomes of the classifier's prediction to enforce the specified parity constraints.



How do you want to measure performance?

Data statistics
1 sensitive feature
3000 instances

Your data contains binary labels and your model makes binary predictions. Based on that information, we recommend the following metrics. Please select one metric from the list.

Accuracy ●
The fraction of data points classified correctly.

Balanced accuracy ○
Positive and negative examples are reweighted to have equal total weight. Suitable if the underlying data is highly imbalanced.

Precision ○
The fraction of data points classified correctly among those classified as 1.

Recall ○
The fraction of data points classified correctly among those whose true label is 1. Alternative names: true positive rate, sensitivity.

Figure 4: Metric Selector

The fairness algorithm called as ThresholdOptimizer. Here we take the original model and add this optimization layer on top of it to equalize the performance between the selected demographic group.

I have used Fairlearn package to implement the fairness algorithms and improvements, fairness metrics and for the results dashboard. It works seamlessly with matplotlib and sklearn for easier visualizations.

3 RESULTS

In this section I'll discuss about the various results that I got after training the models and after applying the fairness algorithm to increase the fairness.

3.1 Results for the Training Algorithms

I have used accuracy and the ROC curves for picking out the best algorithms. I will use the best performing and an intermediate algorithm to show a comparison between them after we use the fairness algorithm. The accuracies are given below:

- RandomForestClassifier - 0.826
- AdaBoostClassifie - 0.820
- DecisionTreeClassifier - 0.751
- SVC - 0.776

The ROC curves are given below from Figure 5 till Figure 8:

3.2 Results for Visualizing fairness

The following results from Figure 9 to Figure 10 show the fairness of the Random Forest and Decision Tree classifier:

The accuracy for Random Forest model can be seen in the figure 9. We can also see how accurately the model predicts the output for the 2 genders. The orange graph shows that how much % of both

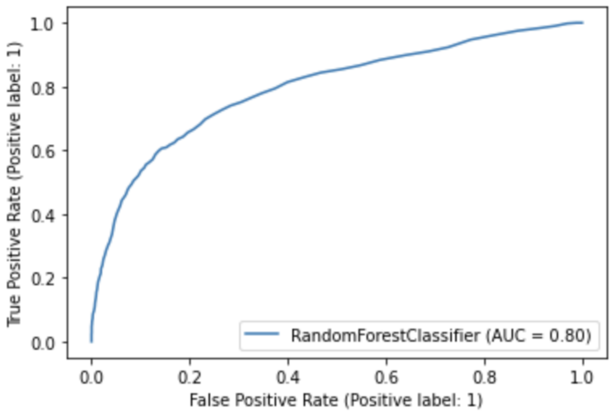


Figure 5: Random Forest Classifier

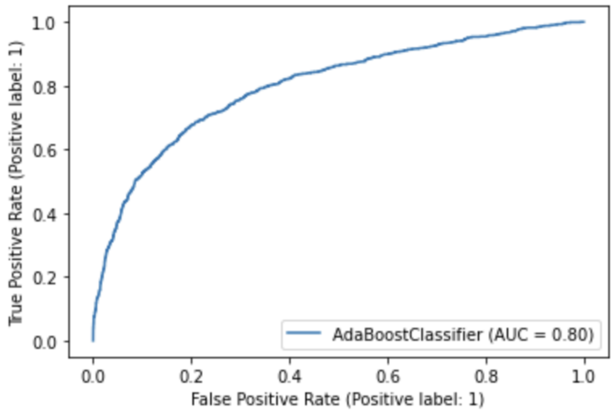


Figure 6: Ada Boost Classifier

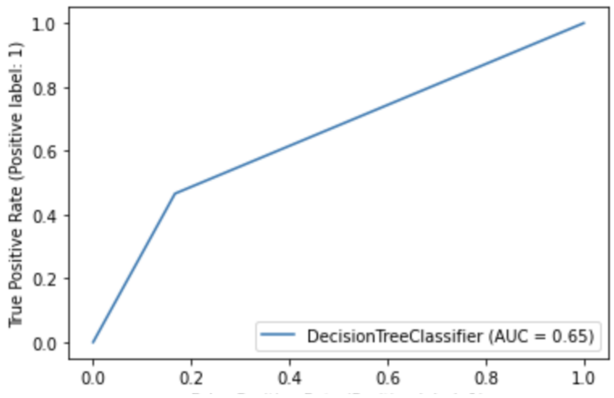


Figure 7: Decision Tree Classifier

genders is incorrectly underpredicted. We can clearly see that for Random forest on the left only 13% of the males are underpredicted i.e these should be predicted as defaulter for the next payment but

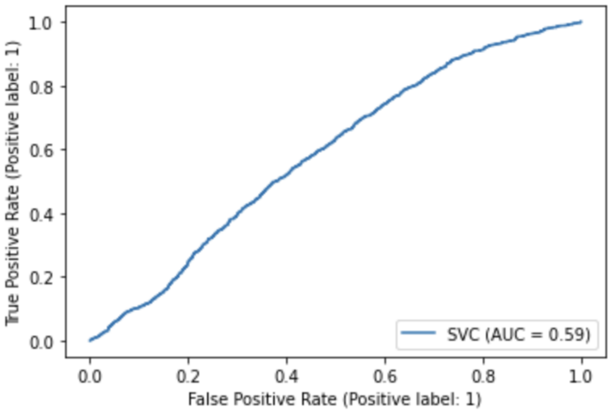
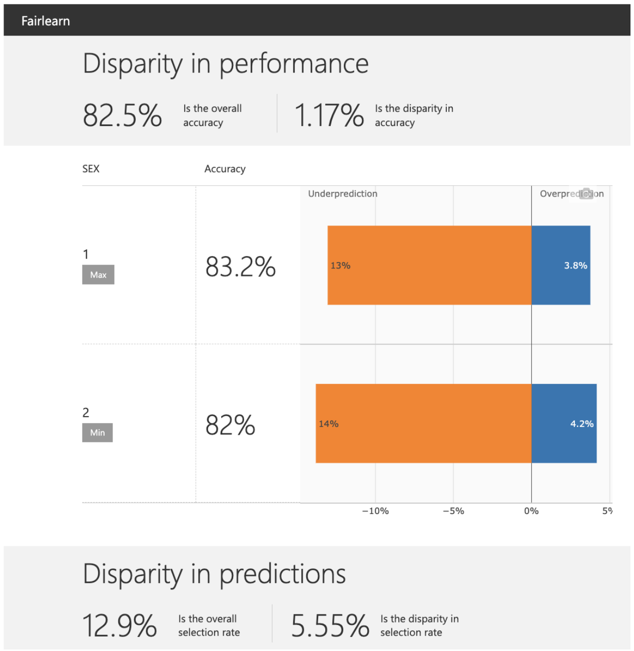


Figure 8: SVC


Fairlearn

Disparity in performance

82.5% Is the overall accuracy

1.17% Is the disparity in accuracy

SEX	Accuracy
1 Max	83.2%
2 Min	82%

Disparity in predictions

12.9% Is the overall selection rate

5.55% Is the disparity in selection rate

Figure 9: Fairness in Random Forest

are not, while 14% of females are underpredicted i.e these should be predicted as defaulter for the next payment but are not.

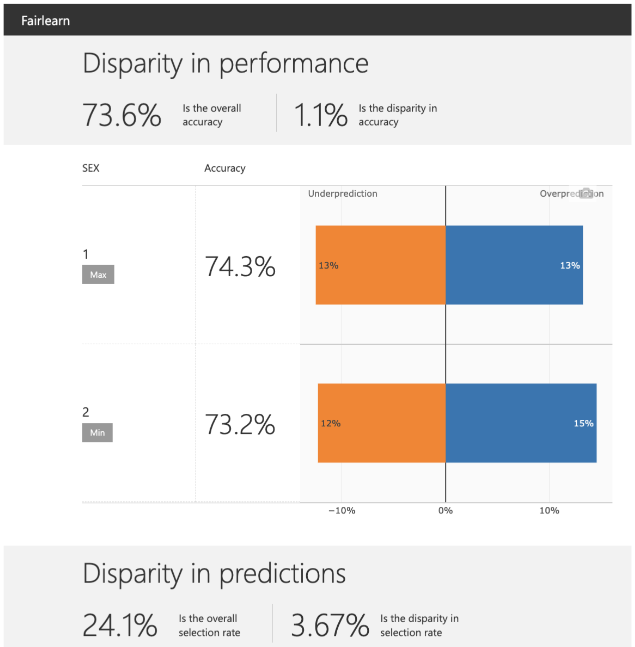
This clearly shows that the data and model are biased towards males.

Similar pattern can be found for decision tree classifier and other classification models, although there is a difference between the actual values.

3.3 Results after improving the Fairness

These results show the models after the fairness improvement algorithm has been applied.

The fairness of the new models is shown in Figure 11 and Figure 12:



Fairlearn

Disparity in performance

73.6% Is the overall accuracy

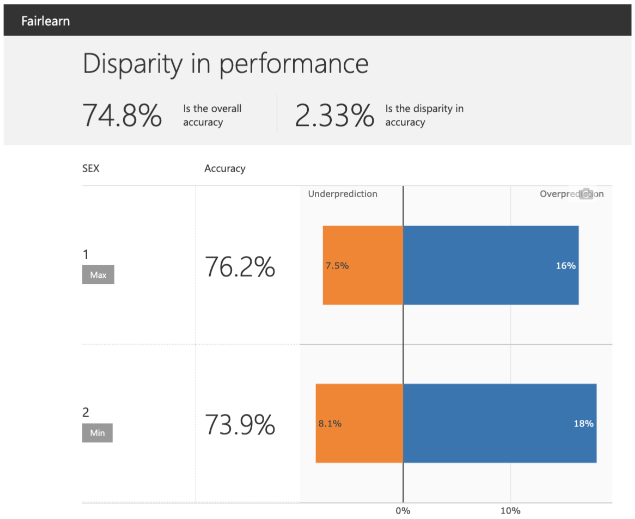
1.1% Is the disparity in accuracy

Disparity in predictions

24.1% Is the overall selection rate

3.67% Is the disparity in selection rate

Figure 10: Fairness in Decision Tree



Fairlearn

Disparity in performance

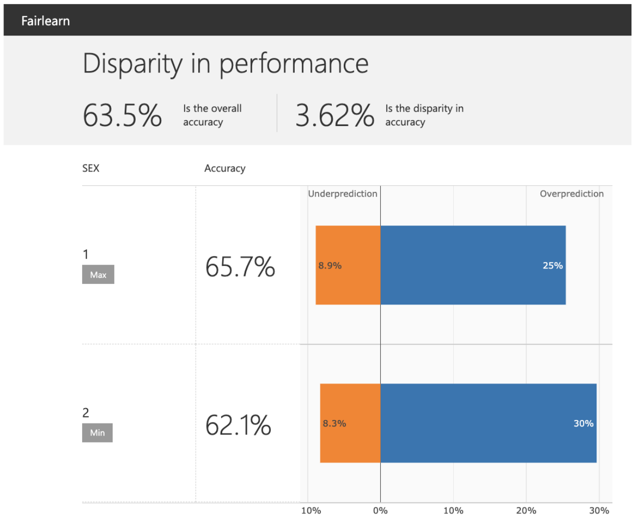
74.8% Is the overall accuracy

2.33% Is the disparity in accuracy

Figure 11: Improved Fairness in Random Forest

After implementing the Threshold optimizer function which is a post processing algorithm, we see that the underpredictions for men and women have been reduced significantly for both the model thus significantly reducing the disparity. We also see a reduction in the overall accuracy of our models, which is a trade-off that one has to make to make the model less biased and fairer.

The underpredictions for females in random forest have reduced by more than 5.9%, making the model less biased for males in Random Forest.



Fairlearn

Disparity in performance

63.5% Is the overall accuracy

3.62% Is the disparity in accuracy

Figure 12: Improved Fairness in Decision Tree

3.4 Results after comparing the Fairness

These results show the comparison between the old and the new models after the fairness improvement algorithm has been applied.

The fairness is shown in Figure 13 and Figure 14:

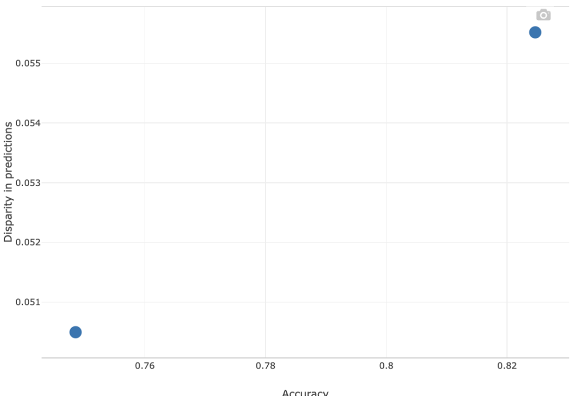


Figure 13: Comparison between old and new Random Forest Model

The figure 13 shows the accuracies and disparity of the older model compared with the model that has the fair algorithm applied to it. It should be read as:

- Accuracy ranges from 74.9% to 82.5%. The disparity ranges from 5.05% to 5.55%.
- The most accurate model achieves accuracy of 82.5% and a disparity of 5.55%.
- The lowest-disparity model achieves accuracy of 74.9% and a disparity of 5.05%

This pattern can be seen for all the classification models mentioned earlier during this presentation where we see a drop in accuracy but a significant rise in the fairness.

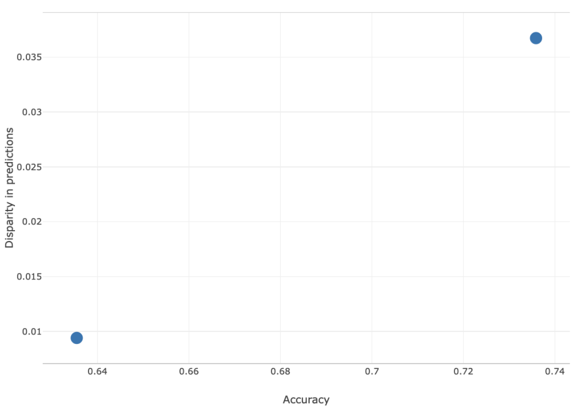


Figure 14: Comparison between old and new Decision Tree Model

4 LIMITATIONS

Although we see a increase in the fairness of our models, there are limitations that I have faced while working on the project. They are mentioned below.

- The dataset contains only 30000 samples which is a relatively small size. If I had access to a similar but larger dataset, the model could have gotten more insights and thus the accuracy would have been higher.
- While working with Fairlearn, I faced multiple versioning issues, where some versions did not work with Google Chrome and the documentation does not make it clear for us.
  So, it took a lot training and retraining with different versions of Fairlearn to get the dashboard to give outputs as the visualizations.
- As seen in the results, there is a clear trade-off between the fairness and accuracy of the models. We see that increasing the fairness hurts accuracy because it diverts the objective from only accuracy to both accuracy and fairness.
- I retrained the models' multiple times, but I was not able to get the same improvement in models thus making it difficult to reproduce the best results.

5 FUTURE WORK

Observing the results and after going through the limitations we can see that there is a scope of improvements in the future. The future works include but are not limited to the list given below:

- I am working on implementing other preprocessing fairness algorithms like CorrelationRemover to see if the results change. I also want to implement a wrapper approach called GridSearch to see if it can serve as an alternative to the current solution.
- Trying out other fairness tools like AI fairness 360 to check whether it works in a similar fashion as Fairlearn . I am also working on using other fairness dashboard called 'raiwidgets' to see if it is easier to use than Fairlearn dashboard.
- I have currently been working with only GENDER attribute, but I want to include multiple sensitive attributes like AGE and EDUCATION to see their impact on the final results.

6 DIVISION OF WORK

I have worked independently on this project and hence all the tasks have been performed by me. Since this was a solo project, it is impossible for me to conduct a peer evaluation. I have listed the tasks below that I had to perform in this project.

- Read papers on Fairness.
- Searched and collected the dataset.
- Preprocessed the data.
- Trained the models.
- Improved and Visualized the fairness.
- Writing the report.
- Making the presentation and presenting it in class.

7 CONCLUSION

In conclusion, we see that there were bias against males in the given models.

After implementing the fairness we were able to successfully reduce this disparity in the model and hence able to induce and improve the fairness of our models

8 REFERENCES

- Project Code
- Fairlearn from https://fairlearn.org/
- Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." ACM Computing Surveys (CSUR) 54.6 (2021): 1-35.