DBS
Live more, Bank less

CBG Analytics

Model Performance, Fairness and Explainability Report

Model ID:
Model Name:
Country:
Model Developer (Project Lead):
Date:

I. Model Description

This is a supervised classification task for credit default risk model. The objective is to use historical loan application data to predict whether or not an applicant will be able to repay a loan. The target is a 0 for the loan was repaid on time, or a 1 indicating the client had payment difficulties. There are over 750 features/input variables that includes CODE_GENDER, FLAG_OWN_CAR, AMT_INCOME_TOTAL, AMT_CREDIT, NAME_EDUCATION_TYPE, OCCUPATION_TYPE and NAME_HOUSING_TYPE.

II. List of Prohibited Features

religion, nationality, birth place, gender, race

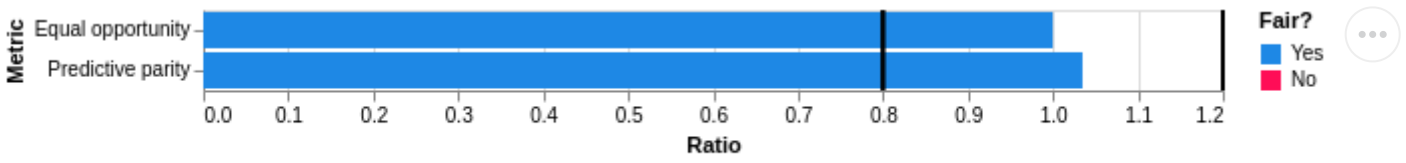
III. Algorithmic Fairness

Algorithmic fairness assesses the models based on technical definitions of fairness. If all are met, the model is deemed to be fair.

Fairness deviation threshold is set at **0.2**. Absolute fairness is 1, so a model is considered fair for the metric when the **metric is between 0.80 and 1.20**.

Prohibited Feature: `CODE_GENDER`

	Metric	Ratio	Fair?
0	Equal opportunity	0.999884	Yes
1	Predictive parity	1.034783	Yes



Overall: **Fair**

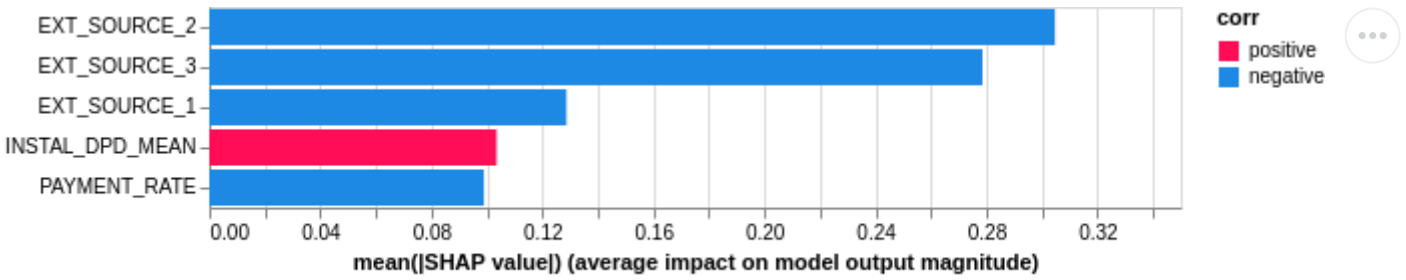
Prohibited Feature: `NAME_EDUCATION_TYPE_Higher_education`

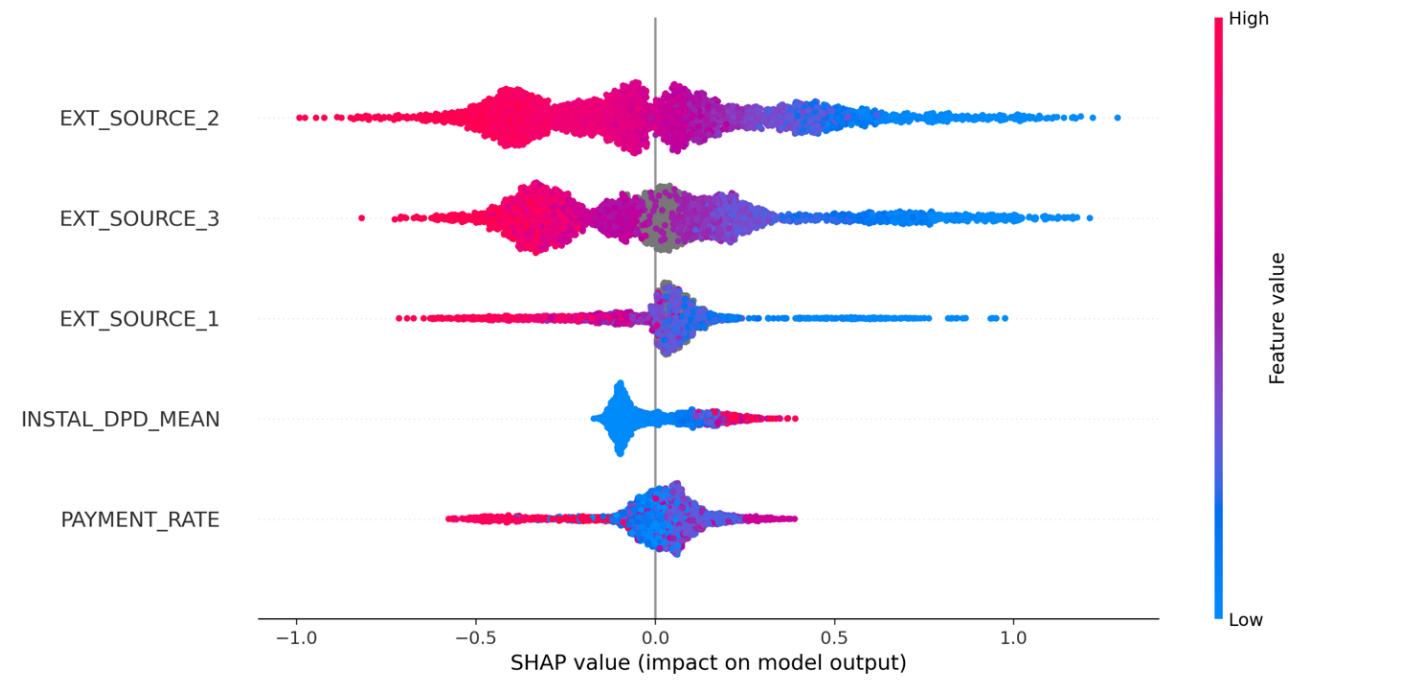
	Metric	Ratio	Fair?
0	Equal opportunity	0.925512	Yes
1	Predictive parity	inf	No

Overall: **Not Fair**

IV. Model Explainability

SHAP Summary Plots of Top Features





The top features are `EXT_SOURCE_2` , `EXT_SOURCE_3` , `EXT_SOURCE_1` , `INSTAL_DPD_MEAN` , `PAYMENT_RATE` .

EXT_SOURCE_2, EXT_SOURCE_3, EXT_SOURCE_1 are scores derived from past records of the client transactions with our bank. Thus, they are important predictors, which are correctly reflected in the feature importance plots.

V. Model Performance

```
Model accuracy = 0.9239
```

	precision	recall	f1-score	support
0	0.9271	0.9958	0.9602	7701
1	0.5556	0.0622	0.1119	643
accuracy			0.9239	8344
macro avg	0.7413	0.5290	0.5361	8344
weighted avg	0.8985	0.9239	0.8949	8344

VI. Conclusion

Model performance: As the precision and recall values are above 70%, the model is considered performing well.

Explainability: Having reviewed the plots in Section 4, we have assessed that the prohibited features do not contribute significantly to the model performance. The direction of the feature impact is also as expected. We are confident that if asked to explain the key factors in the model, a clear explanation can be given to stakeholders and customers.

The top features that have positive correlation with their model output are `INSTAL_DPD_MEAN` .

The top features that have negative correlation with their model output are `EXT_SOURCE_2` , `EXT_SOURCE_3` , `EXT_SOURCE_1` , `PAYMENT_RATE` .

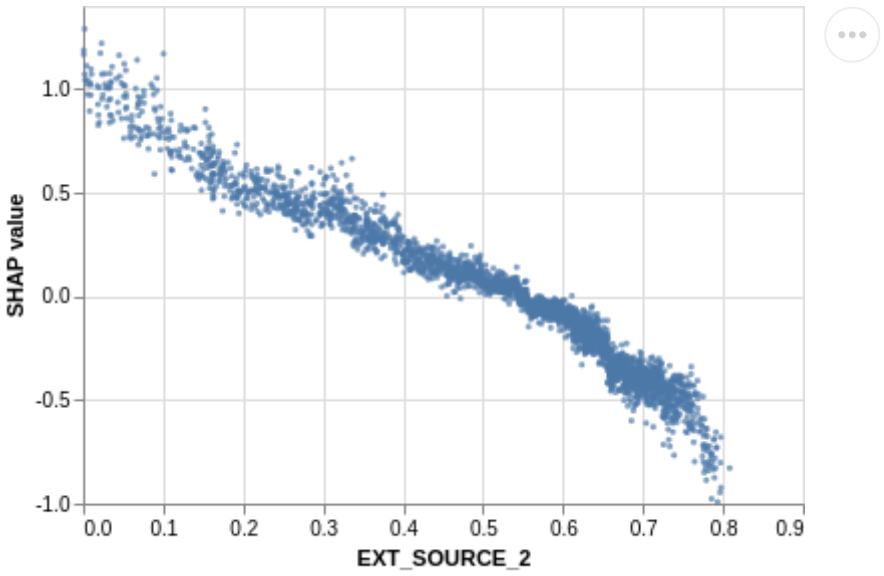
Fairness: We consider the model to be fair if it is deemed to be fair for all metrics. From the table below, overall the model is considered **not fair**.

	Prohibited Variable	Fair?
0	CODE_GENDER-class1	Yes
1	NAME_EDUCATION_TYPE_Higher_education-class1	No

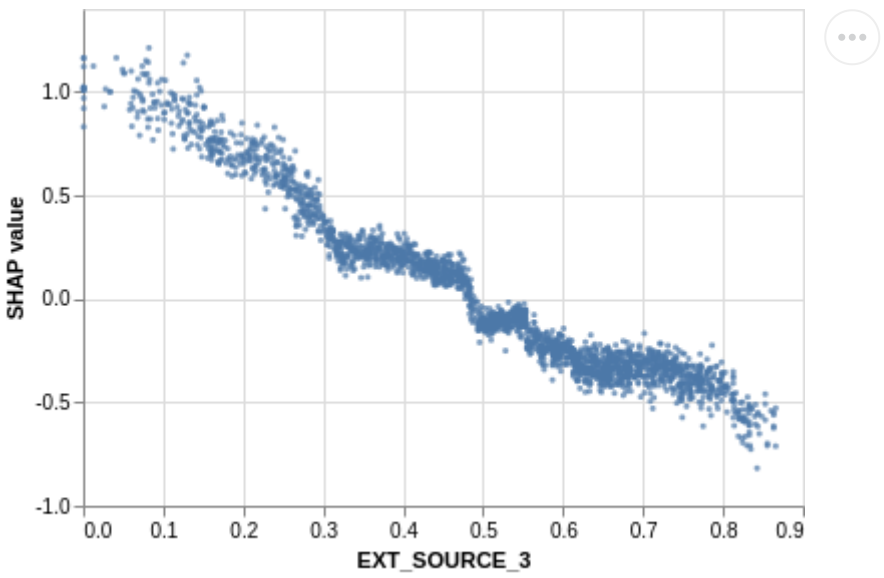
Appendix

Dependence Plots of Top Features

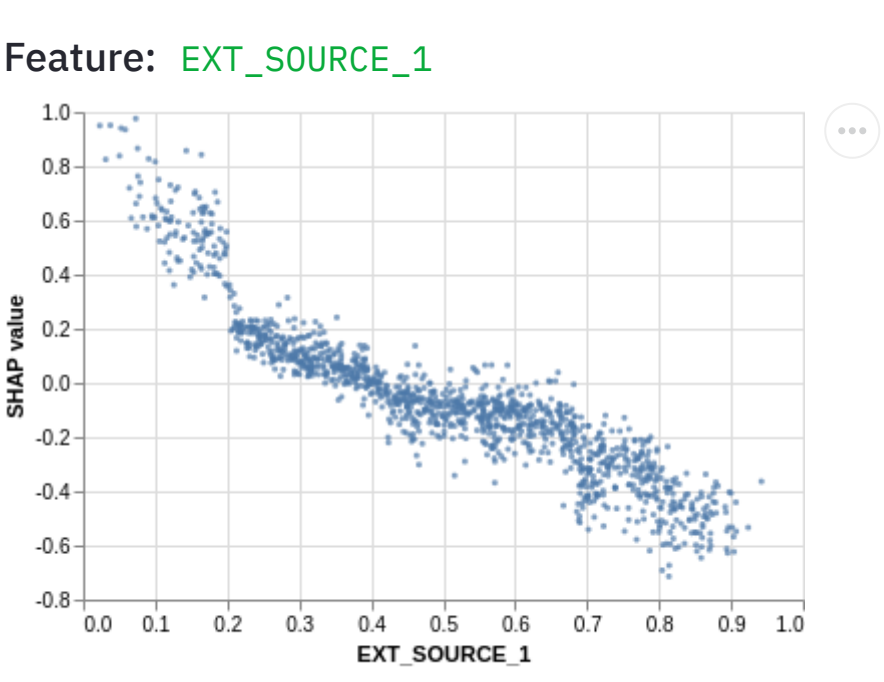
Feature: `EXT_SOURCE_2`



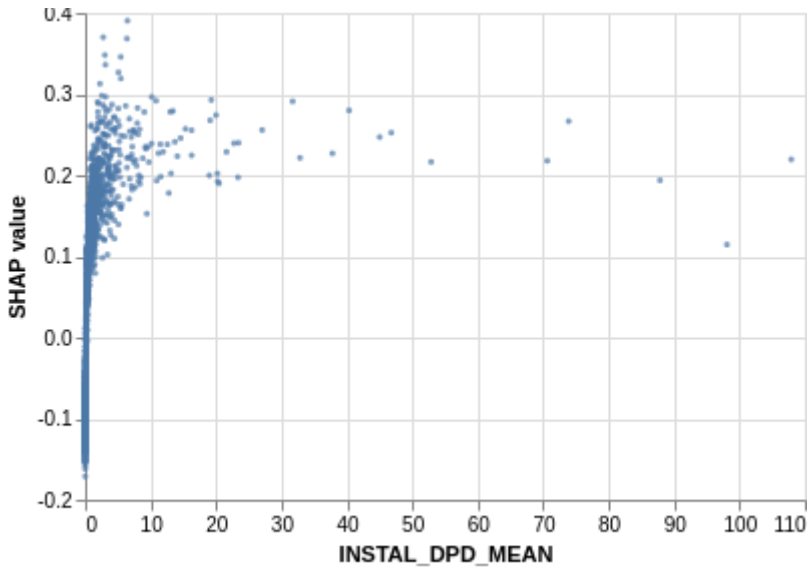
Feature: `EXT_SOURCE_3`



Feature: `EXT_SOURCE_1`



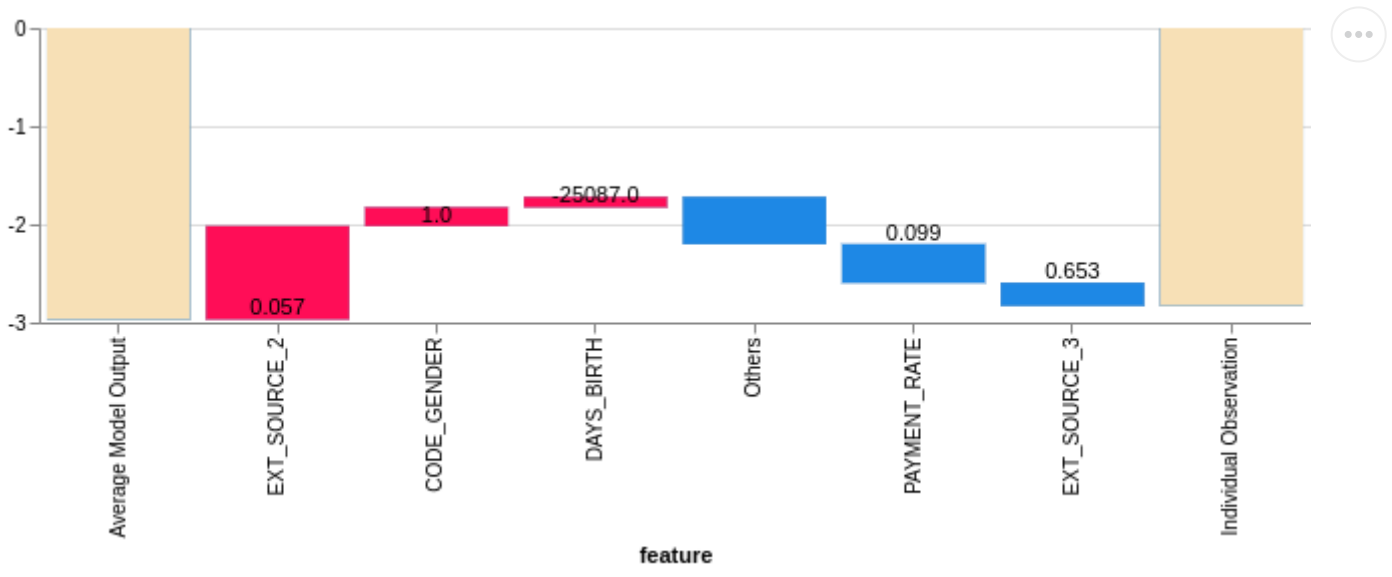
Feature: `INSTAL_DPD_MEAN`



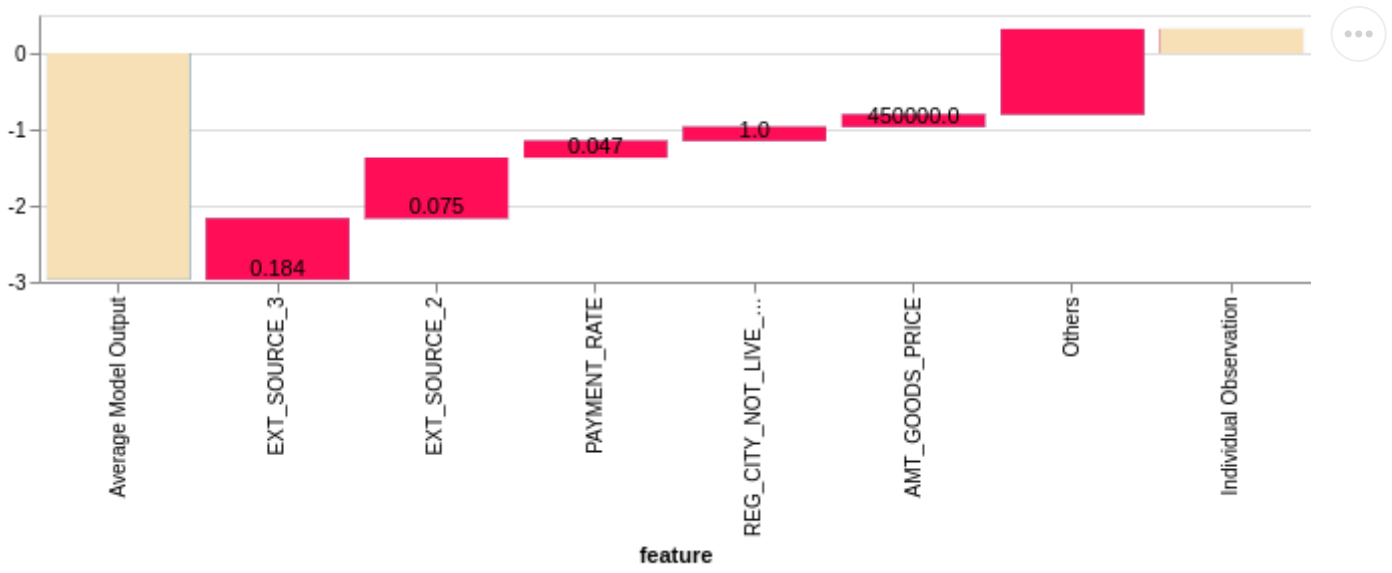
Feature: PAYMENT_RATE

Sample Individual Explainability

Sample from Class=0: SHAP Contribution to Model Prediction



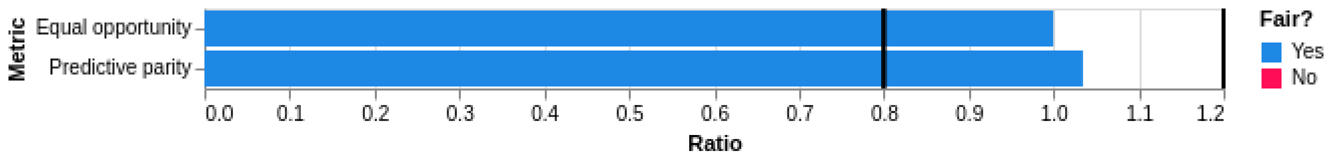
Sample from Class=1: SHAP Contribution to Model Prediction



Algorithmic Fairness

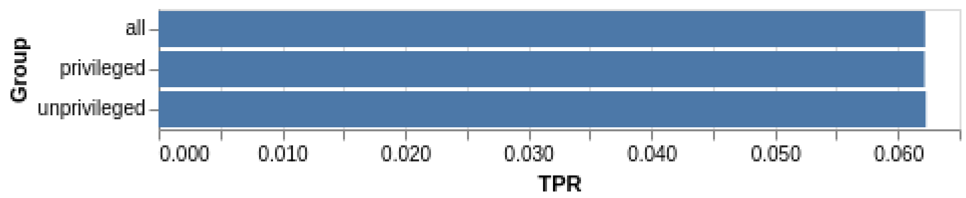
Prohibited Feature: `CODE_GENDER`

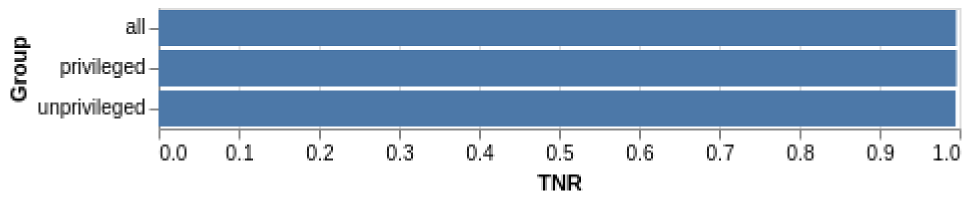
Fairness is when **ratio is between 0.80 and 1.20**.

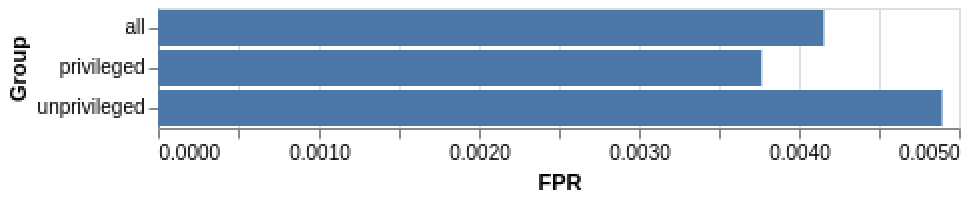


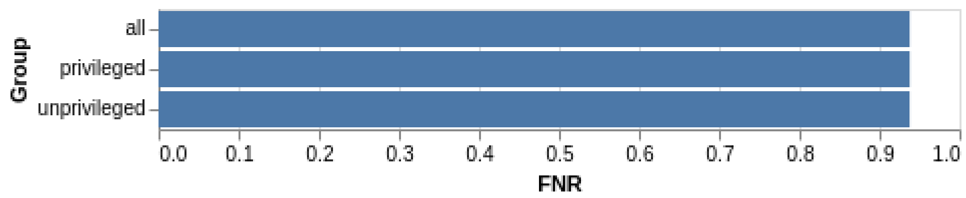
	Metric	Unprivileged	Privileged	Ratio	Fair?
0	Equal opportunity	0.937729	0.937838	0.999884	Yes
1	Predictive parity	0.566667	0.547619	1.034783	Yes

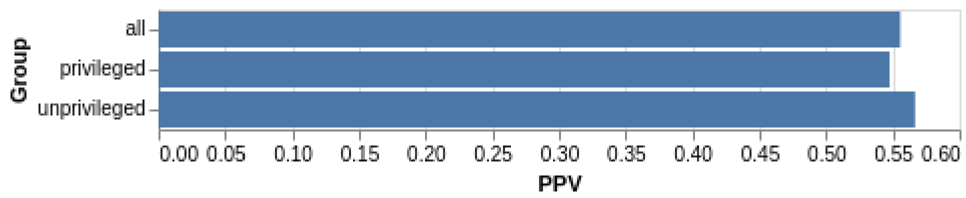
Performance Metrics

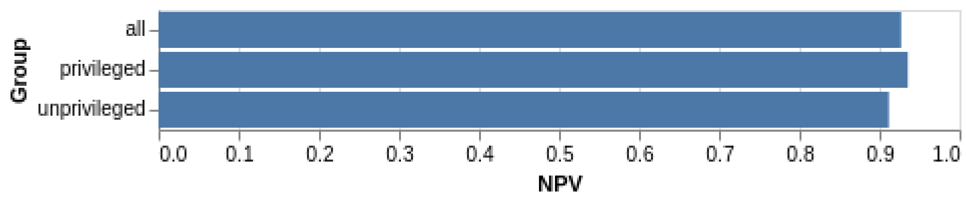


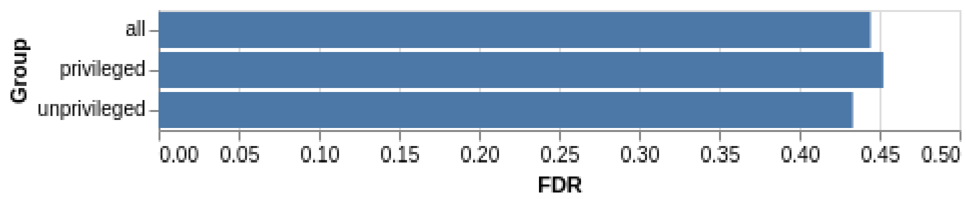


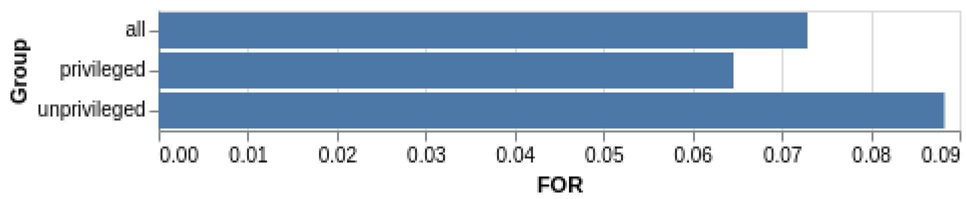


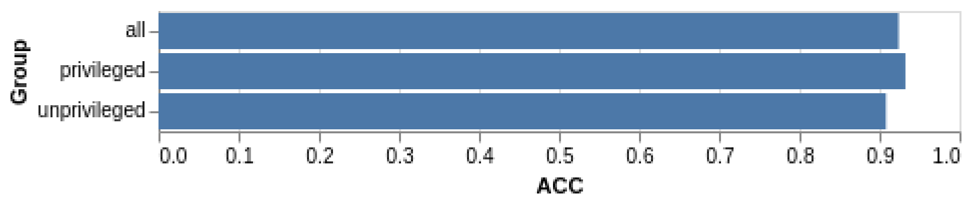


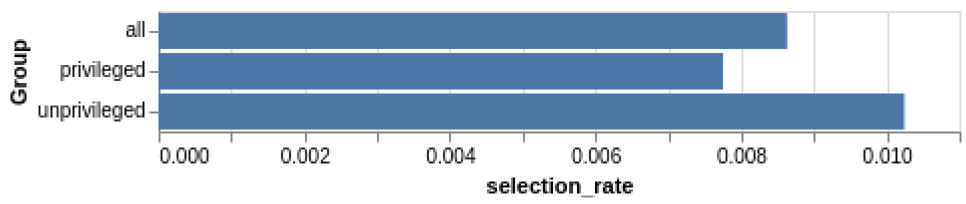


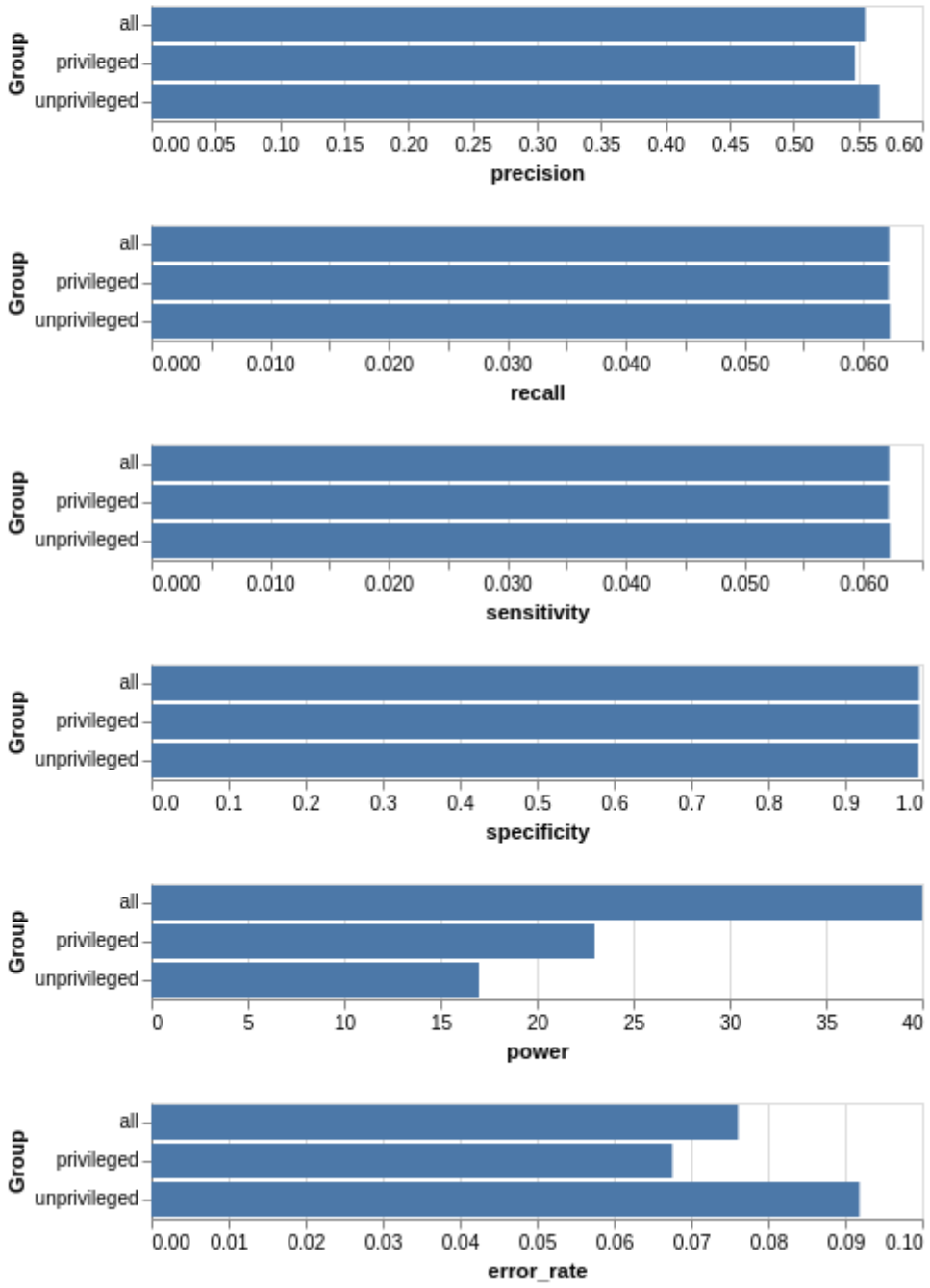




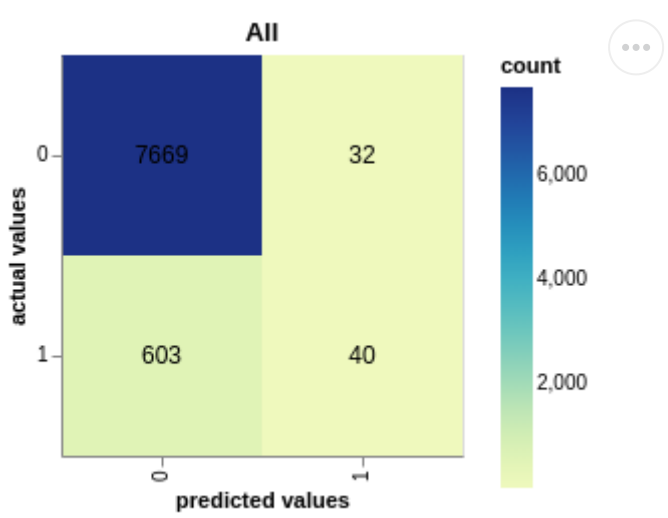


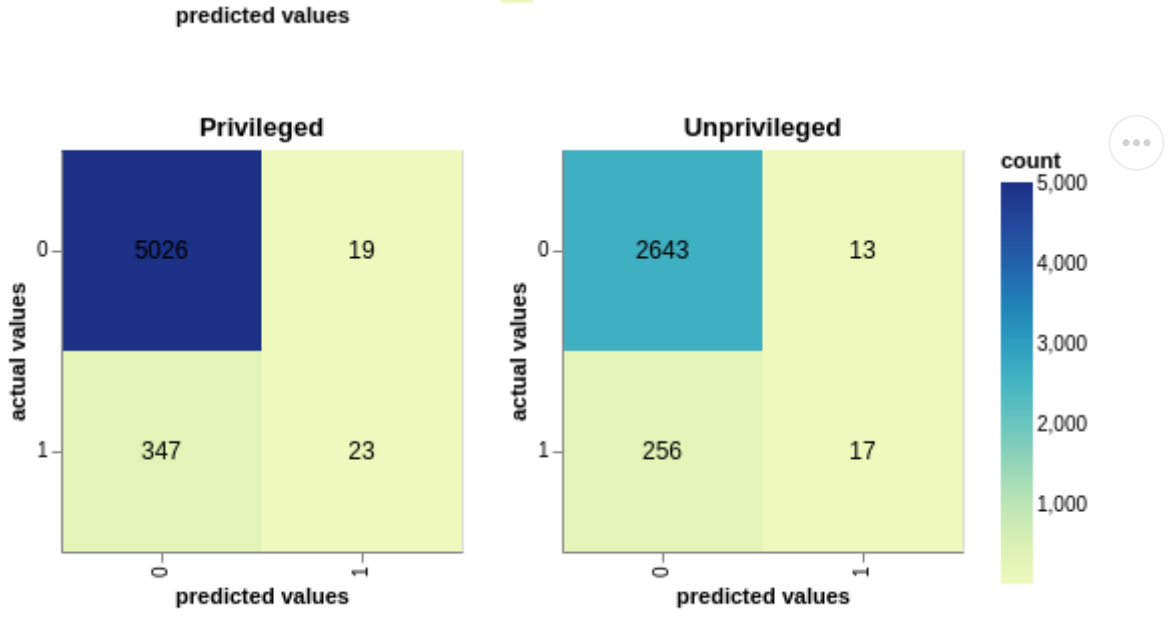






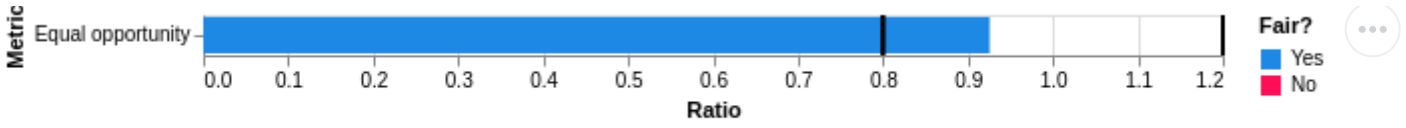
Confusion Matrices





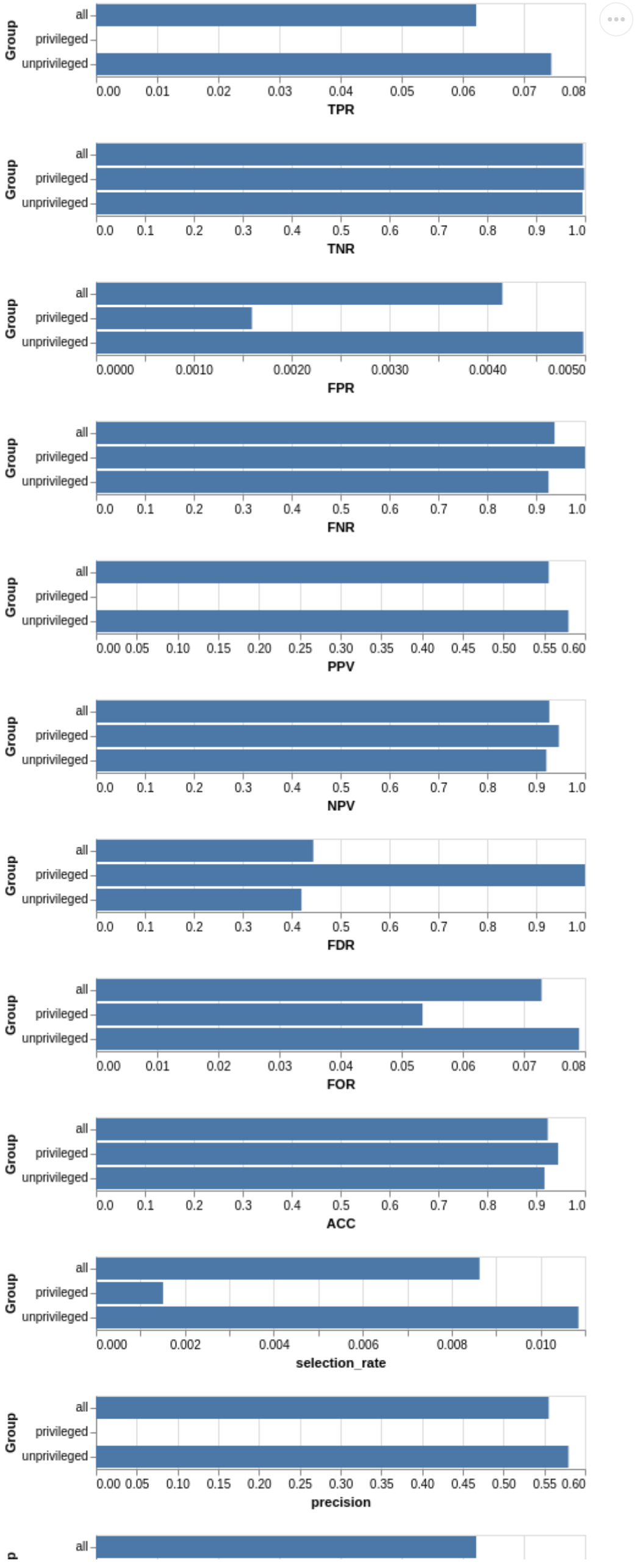
Prohibited Feature: `NAME_EDUCATION_TYPE_Higher_education`

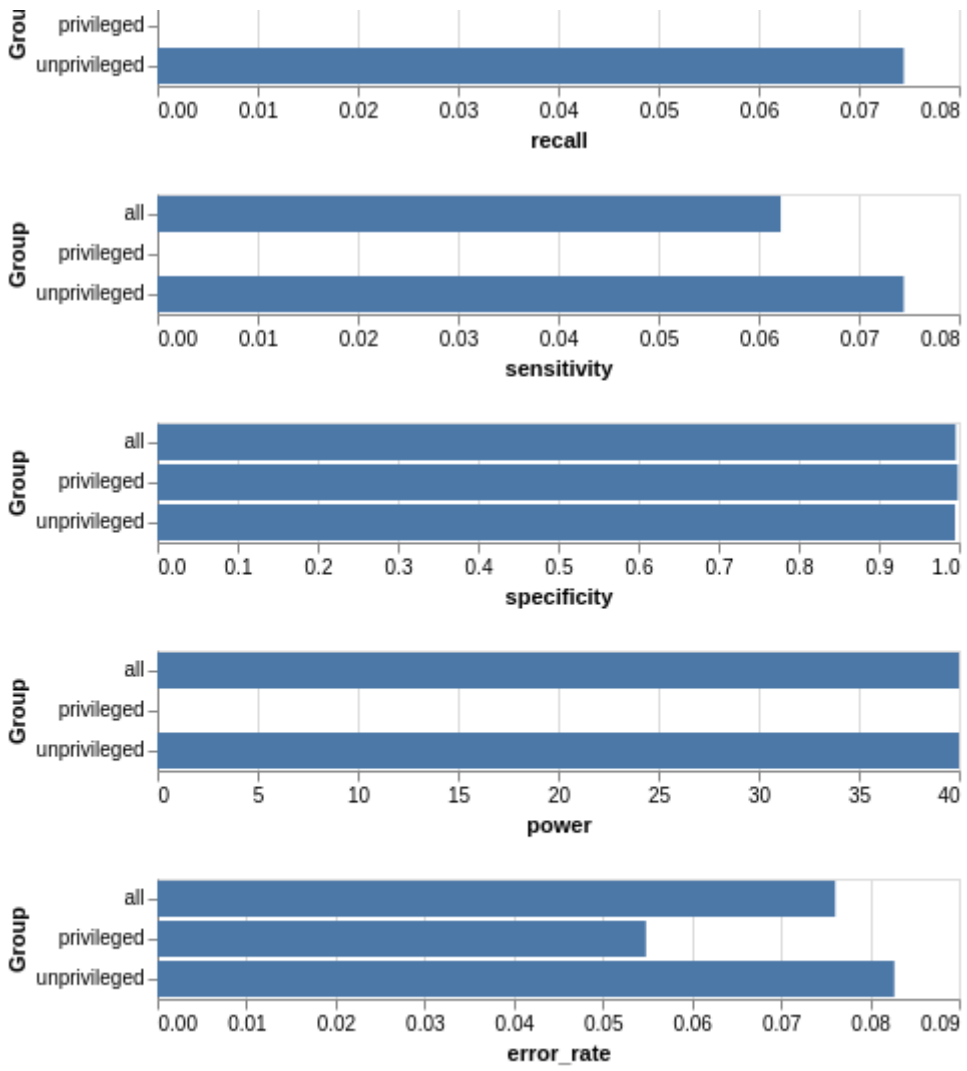
Fairness is when **ratio is between 0.80 and 1.20**.



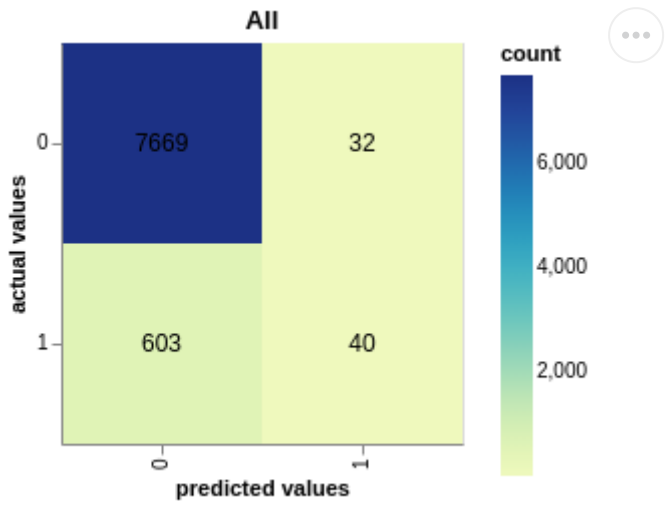
	Metric	Unprivileged	Privileged	Ratio	Fair?
0	Equal opportunity	0.925512	1.000000	0.925512	Yes
1	Predictive parity	0.579710	0.000000	inf	No

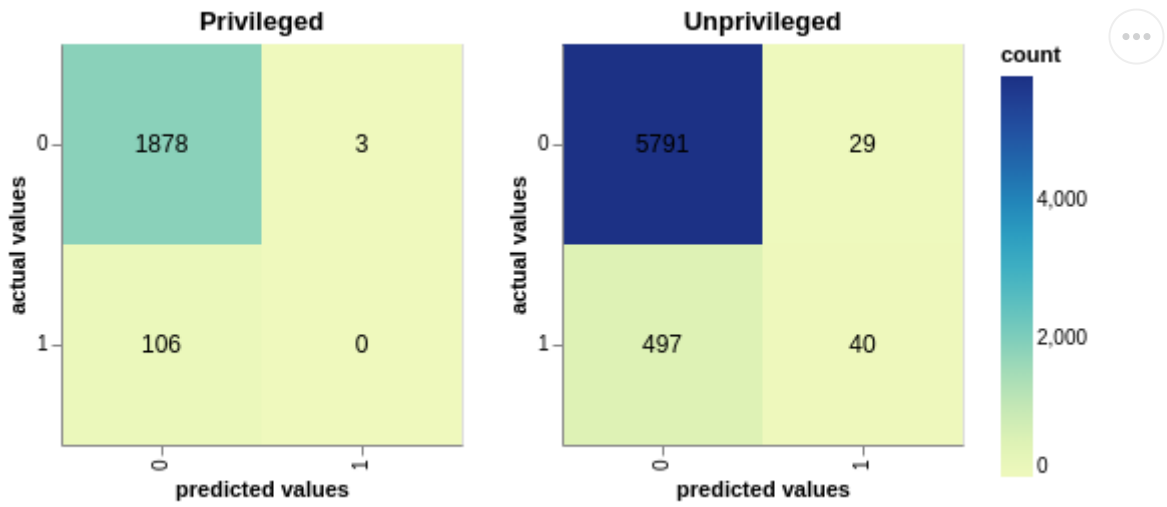
Performance Metrics





Confusion Matrices





Notes

Equal opportunity:

$$\frac{\mathrm{FNR}(D = \mathrm{unprivileged})}{\mathrm{FNR}(D = \mathrm{privileged})}$$

Predictive parity:

$$\frac{\mathrm{PPV}(D = \mathrm{unprivileged})}{\mathrm{PPV}(D = \mathrm{privileged})}$$

Made with Streamlit