Second look Insights

Fairgen

Background knowledge: Shapley Values

- Invented by Lloyd Shapley to provide a fair solution to this question: if a coalition C of individuals collaborate to get a final value V, how much did each individual contribute to the final value
- Shapley values are also used for **Explainability** of machine learning prediction models
  - ❖ We use them to determine how individual feature contribute to the model output and rank their importance in the model output
  - ❖ We also measure the correlation of all attribute of the feature to explain how the feature contribute to the output prediction
    - ➢ A positive Shapley value for a feature means that the feature has a positive contribution to the prediction of the target variable. This means that when the feature is present in the data, it increases the accuracy of the model's predictions for the target variable.
    - ➢ A negative Shapley value for a feature would indicate that the feature has a negative impact on the prediction of the target variable.

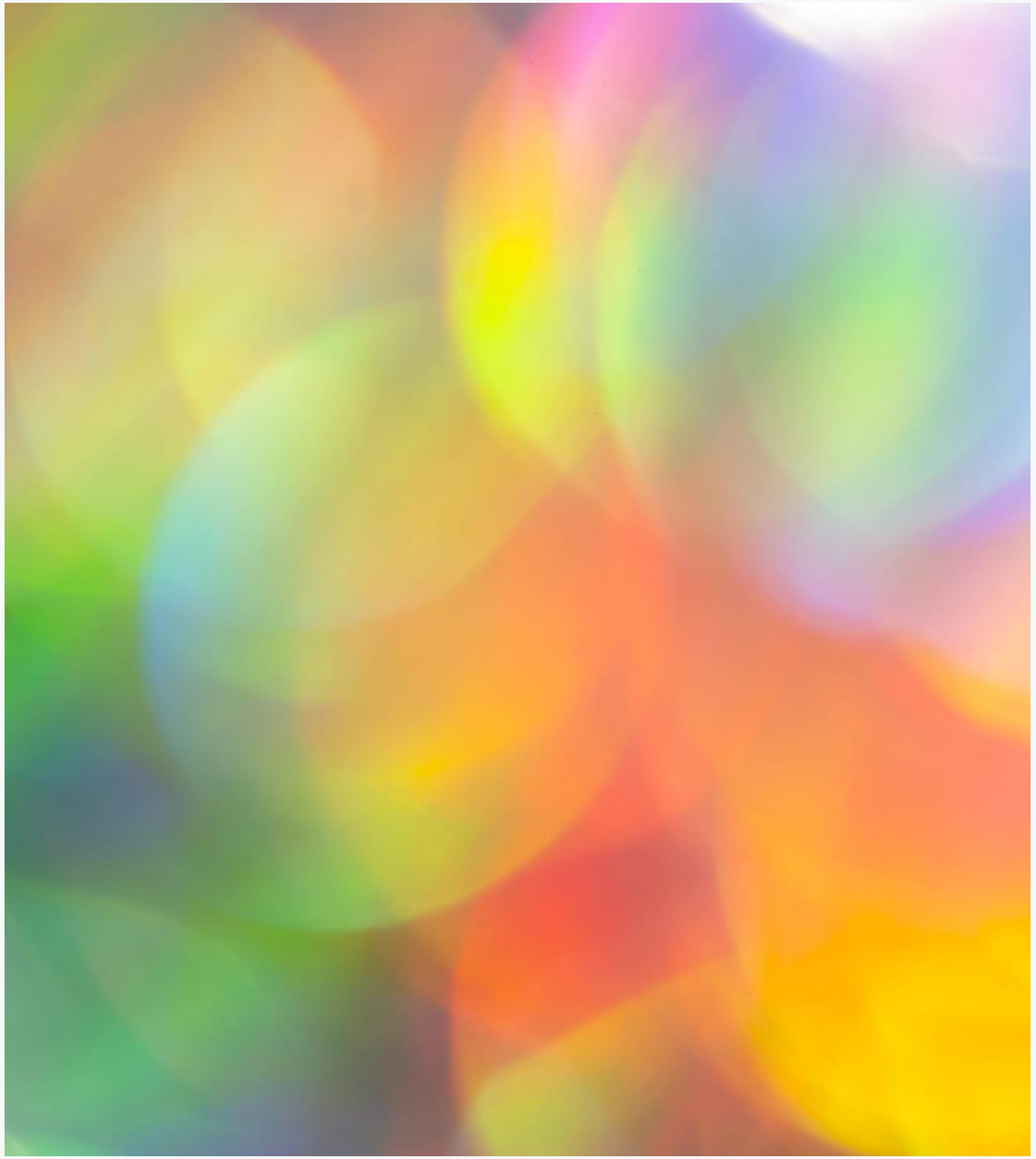
Experiment

We are going to use the three wrappers to try to solve the sex bias in Adult Census dataset.

- ❖ Adding more women in the generated dataset with proportion wrapper
- ❖ Adding more women with high income with allocation wrapper
- ❖ Ensuring the model achieve equal opportunity for high income between men and women with the performance wrapper

GOALS:

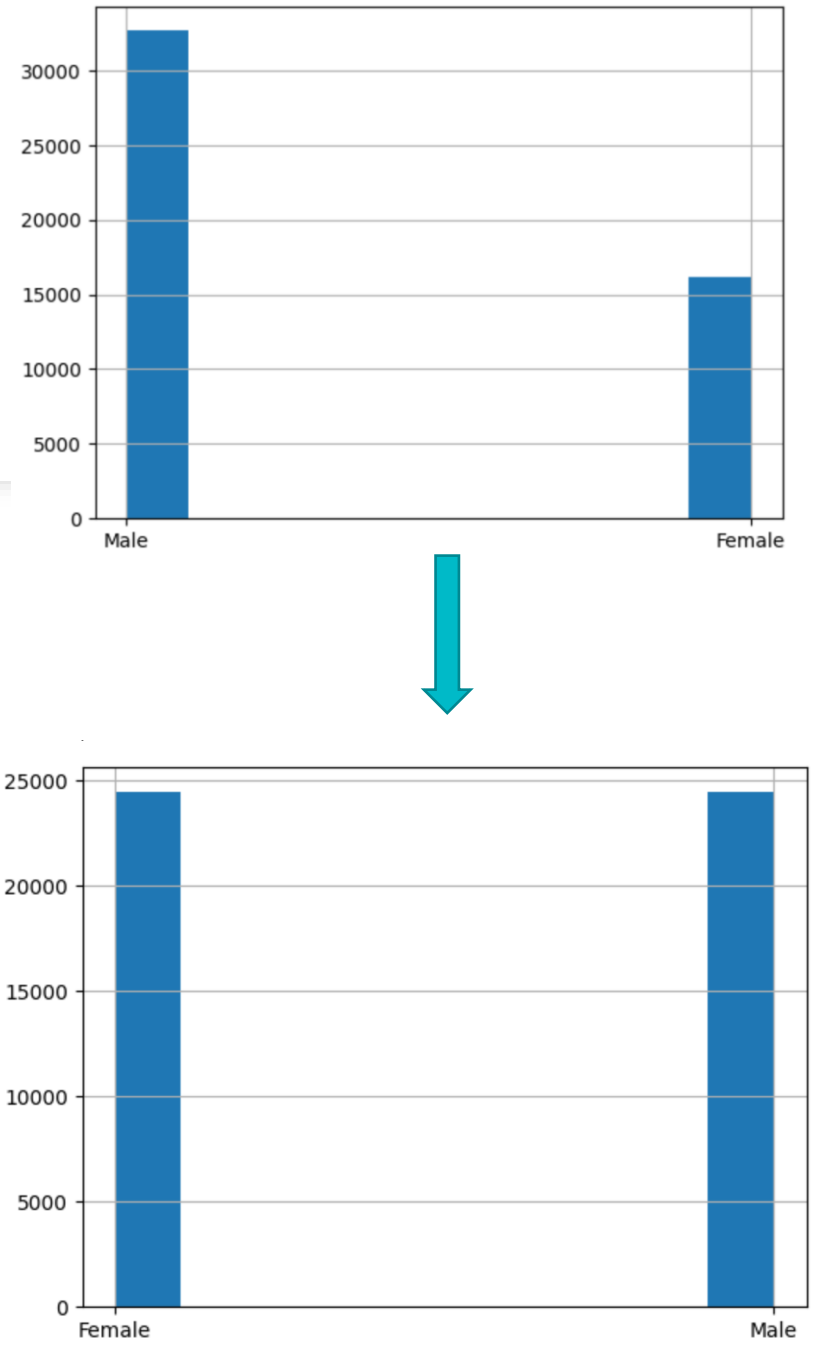
- **Get insights on how the generated dataset differs from the real one to meet the constraints**
- **Explain how the models trained on these new generated datasets differ from the model trained on the real dataset**
- **Conclude on the way each wrapper solve the bias**



Proportion GAN

Proportion Wrapper

- Generate a new dataset with the proportion requested for a selection of attributes
- New examples are therefore generated to fit the proportions
- No big impact on the positive target



80% Women 20% Men in AdultCensus

- Summary of differences between the two datasets for high income under majority proportion of women
- No major differences in every attributes between the two dataset
- We see that the proportion of women with high income didn't increase, but has decreased instead

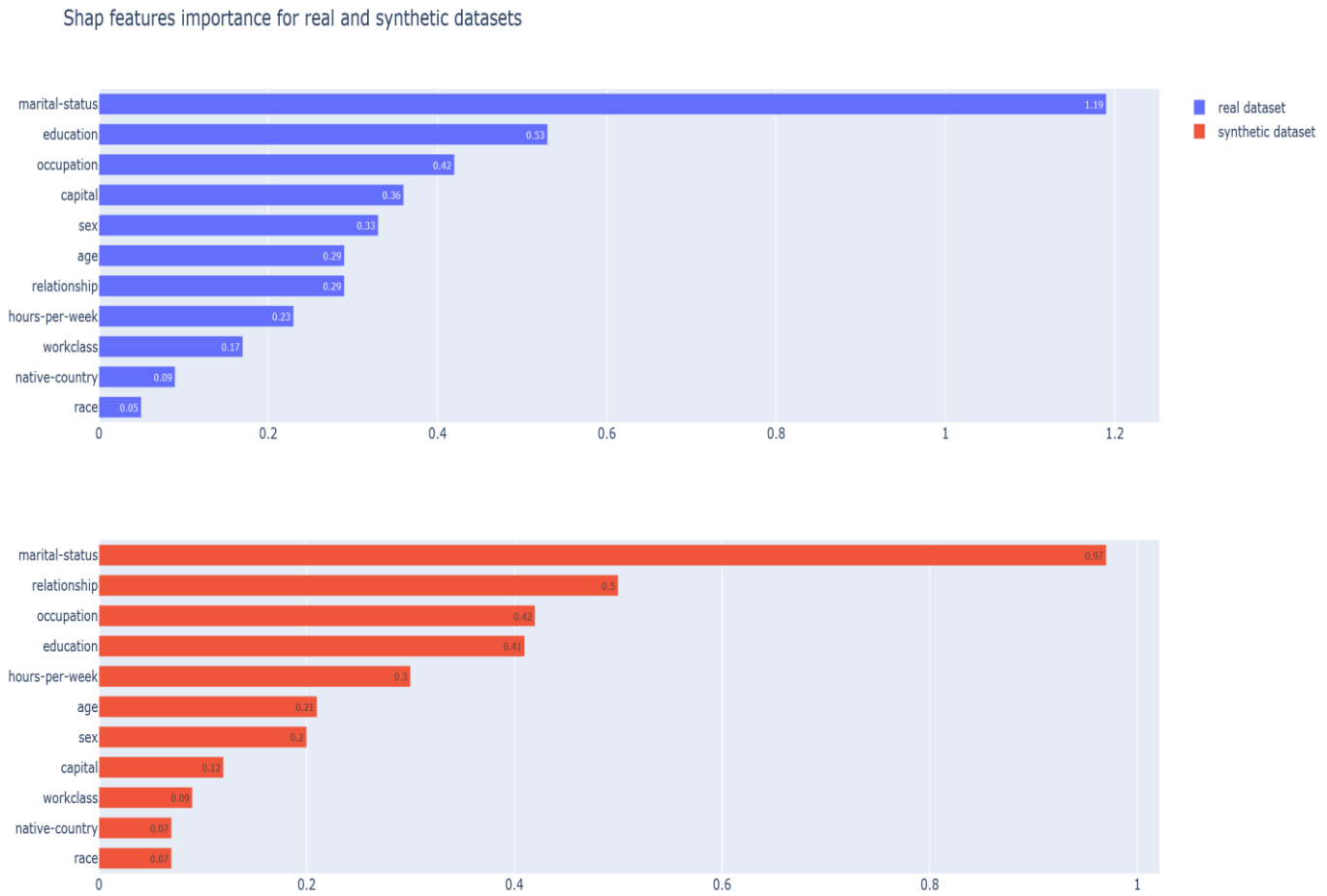
Conclusion: in the generated dataset, there is more women with low income

	demographic	differences_with_real_dataset	proportion in dataset
0	Male,White	4%	58.13%
0	Male	3.5%	66.45%
1	White	2.5%	85.28%
1	Male,Asian-Pac-Islander	1%	2.09%
3	Female,Black	1%	4.53%
3	Asian-Pac-Islander	0.4%	3.37%
4	Black	0.4%	9.75%
2	Female	-0.4%	33.55%
2	Female,White	-1%	27.16%

80% Women 20% Men in AdultCensus

- This plot shows the importance of each feature for the model prediction for the two datasets
- Here, the rankings are very similar except for capital that decrease in importance and relationship that increase in importance

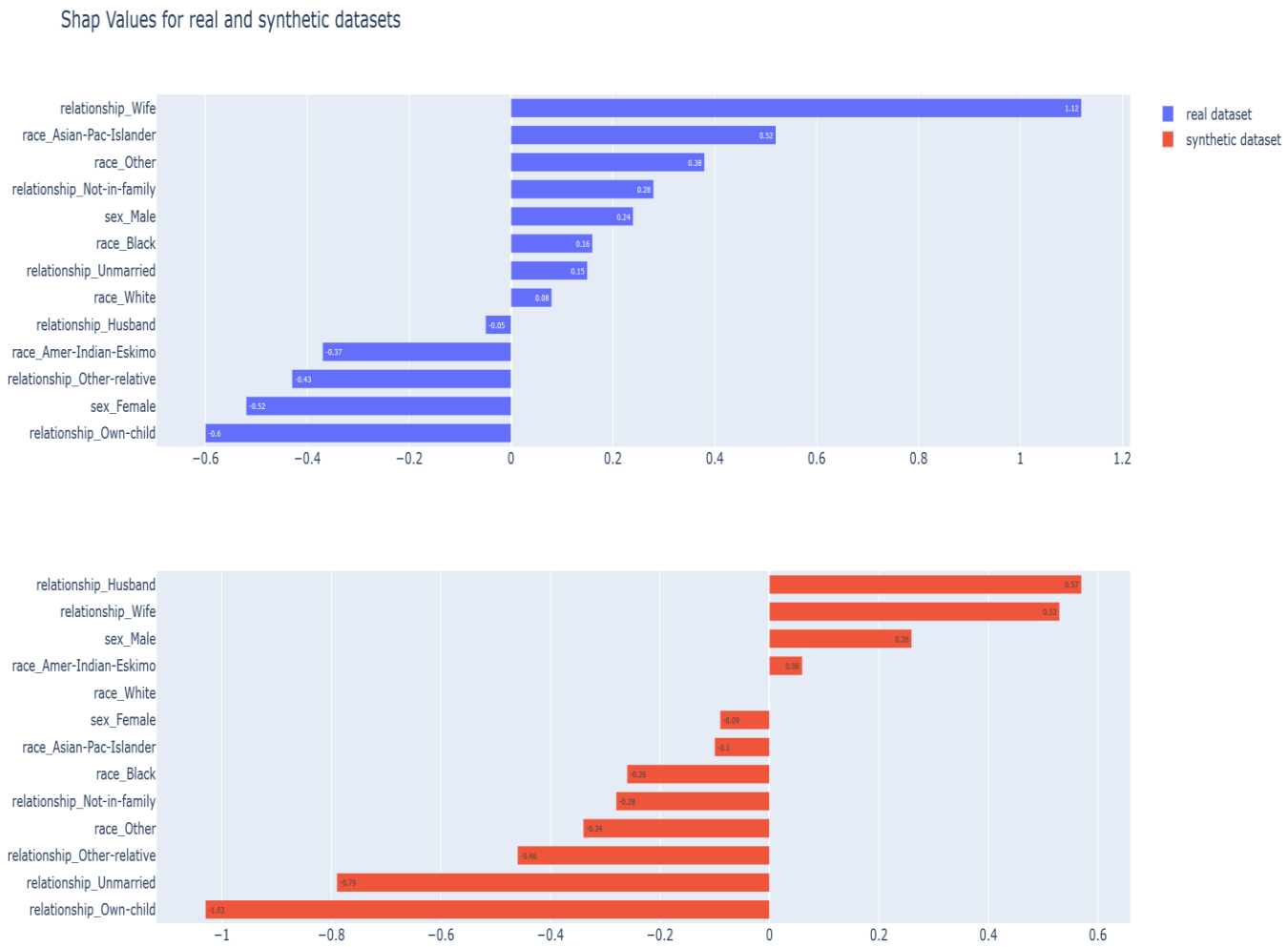
Conclusion: interestingly, this means that the model become less fair when just modifying the proportions. Capital that should be very linked to high income is less relevant and relationship which is sort of proxy for sex is now more important for high income

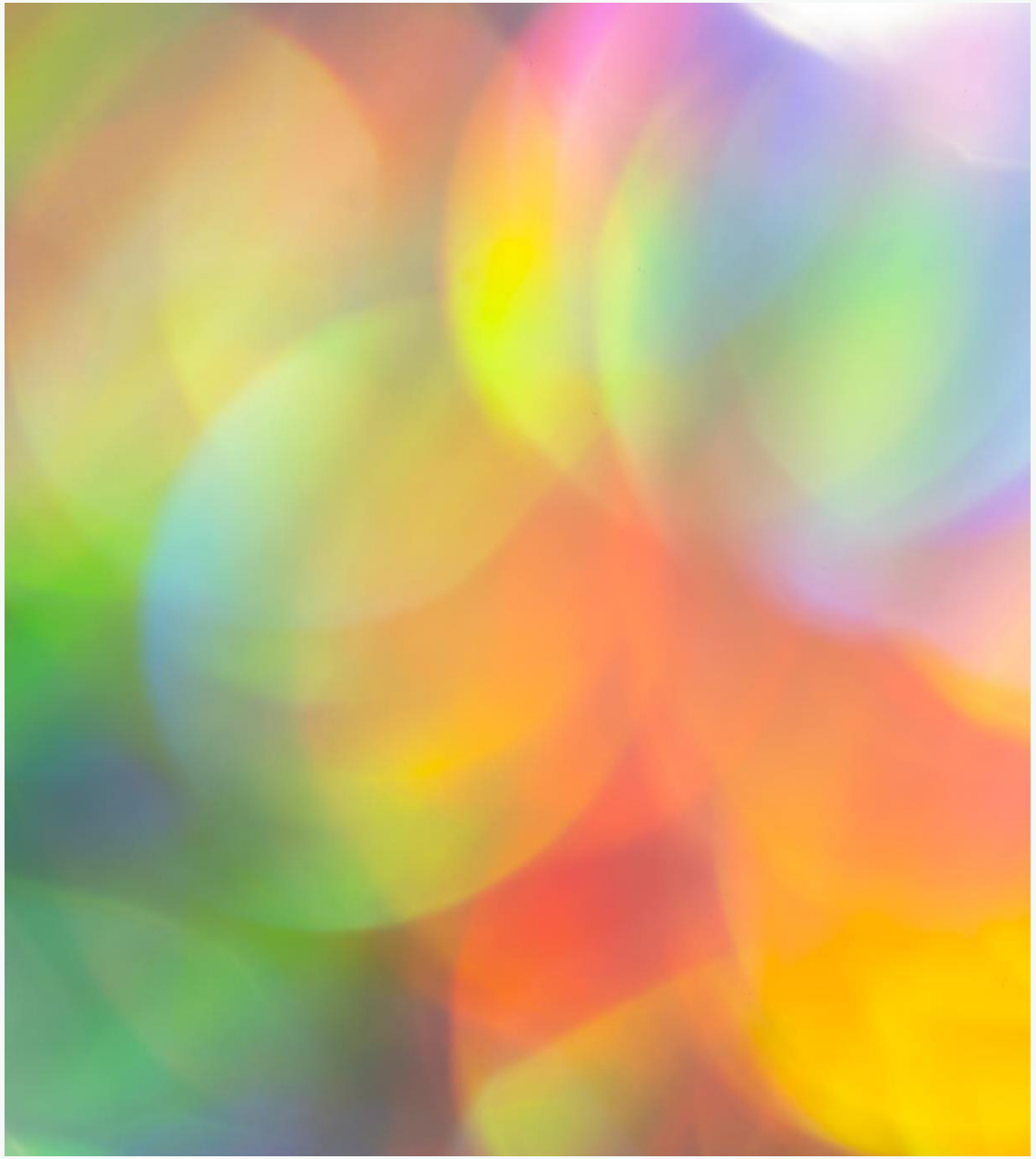


80% Women 20% Men in AdultCensus

- Plots of Shapley Values from the two datasets
- We have seen before that we have more women with low income and more men with high income
- Sex_female is still negatively correlated with high income in the synthetic dataset
- Interestingly, no major change for sex_male but relationship_husband becomes very correlated with high income
- Relationship_husband is a proxy for male as 94% of individuals with this attribute are male in the synthetic dataset

Conclusion: Based on this plot, we see that the model predicting high income is more biased and based on gender as it was before

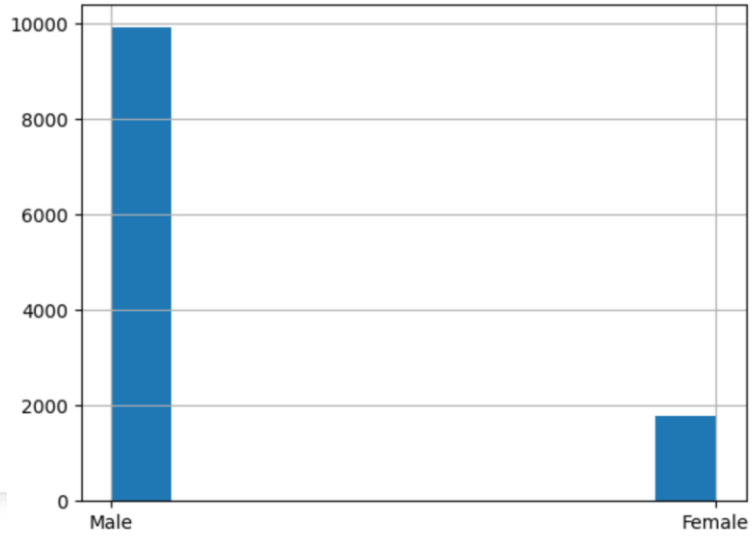




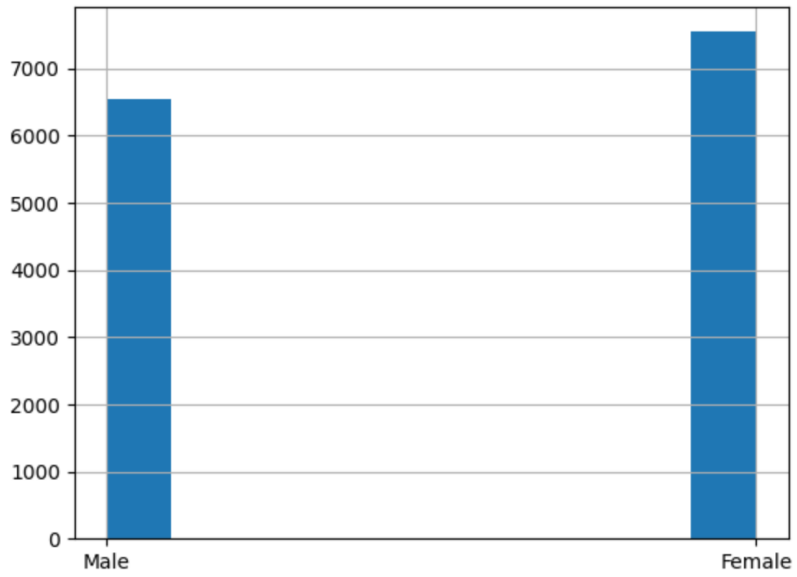
Allocation GAN

Allocation Wrapper

- Generate a new dataset where constraints linking attributes and target are met
- **Example:** we want to increase the proportion of women with high income



Histogram of gender proportion with high income



70% Women with high income 30% Men with high income in AdultCensus

- Summary of differences between the two datasets for high income under majority proportion of women with high income
- As expected, the differences with the real dataset are huge.
- However, we noticed that mostly white female with high income were generated

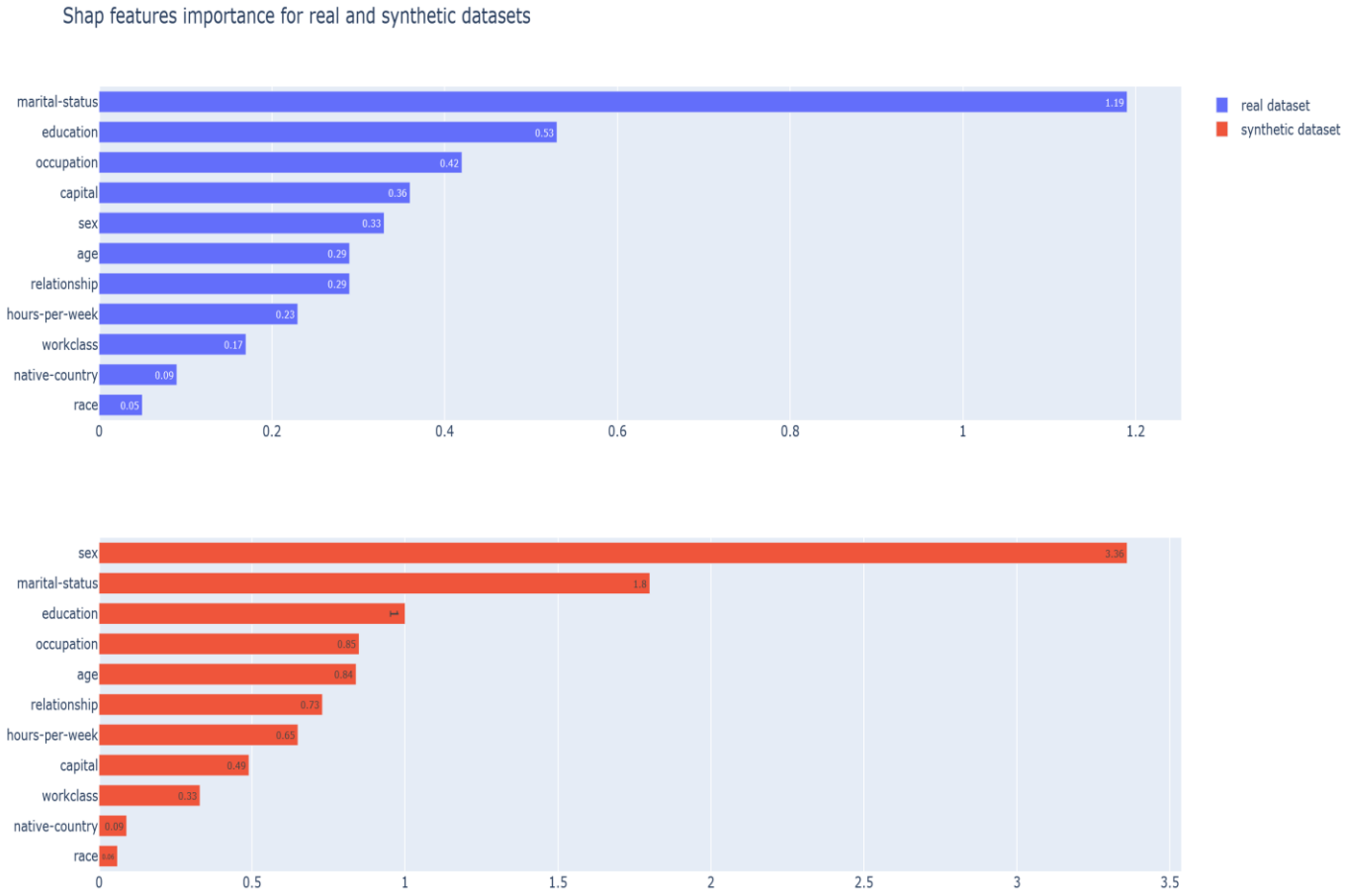
Conclusion: in the generated dataset, there is more women with high income but we need to check if it increased another bias like race for example

	demographic	differences_with_real_dataset	proportion in dataset
0	Female	64.8%	33.55%
0	Female,White	64%	27.16%
1	White	22.9%	85.28%
2	Never-married	21.3%	33.04%
1	White,Never-married	21%	27.02%
3	Married-civ-spouse	10.8%	45.66%
2	White,Married-civ-spouse	10%	40.72%
3	Male,Married-civ-spouse	6%	40.81%

70% Women with high income 30% Men with high income in AdultCensus

- This plot shows the importance of each feature for the model prediction for the two datasets
- Here, the rankings is similar for the feature importance except for the feature sex that has become the most important feature to contribute to the prediction of the model
- However, race still has a low contribution in the prediction for high income which is good

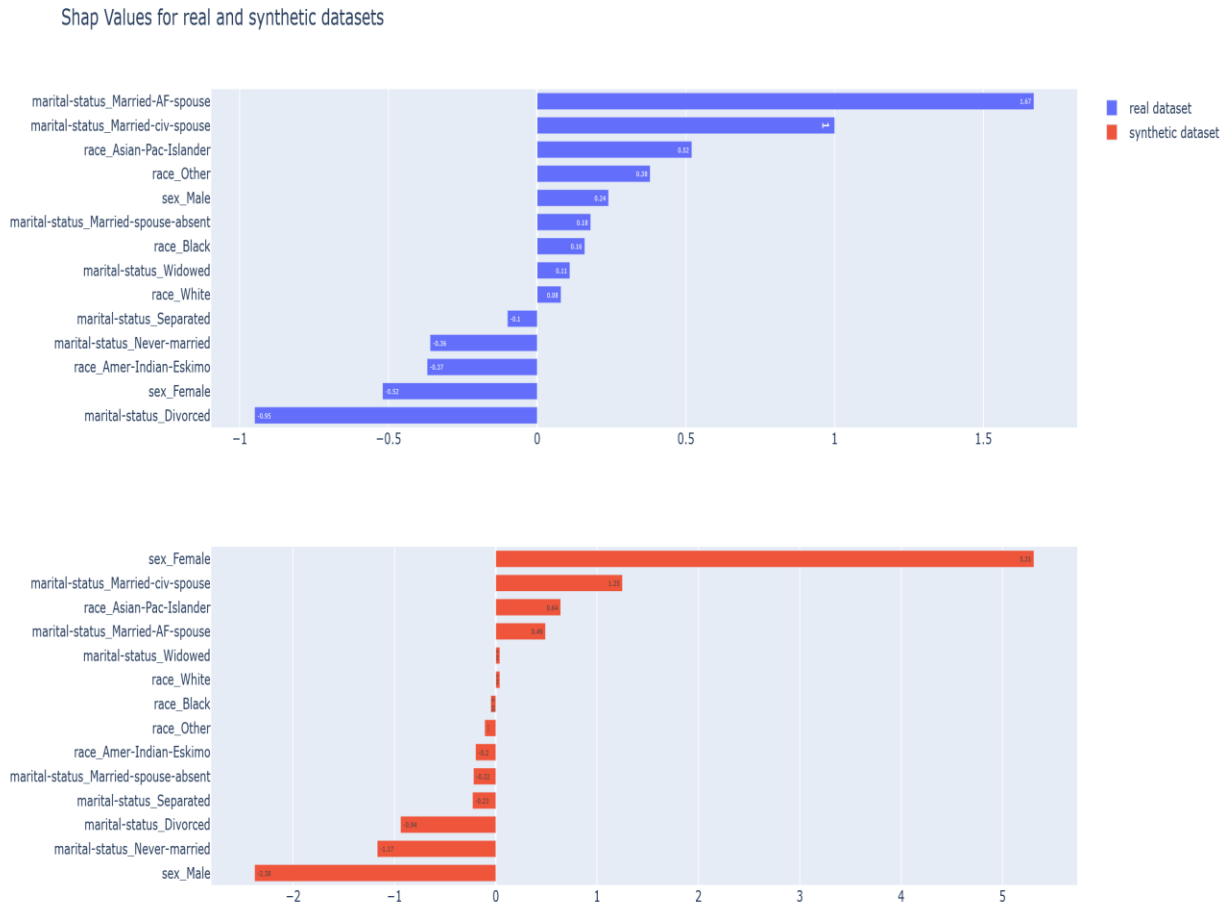
Conclusion: It seems that when increasing the number of women with high income so that much more women has high income than men, the model learns that gender becomes a great indicator of high income. This was not the case in the real dataset.

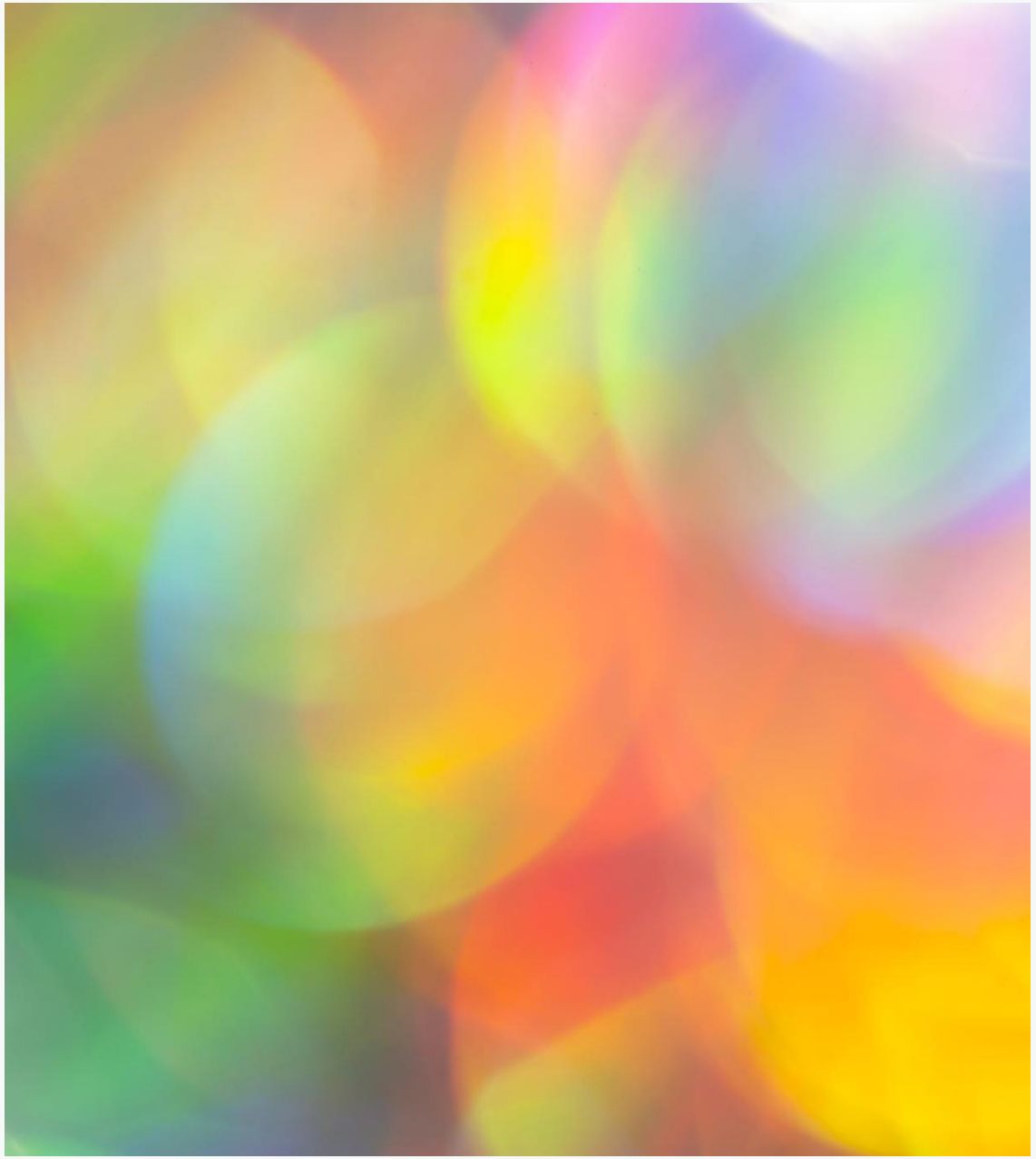


70% Women with high income 30% Men with high income in AdultCensus

- Plots of Shapley Values from the two datasets. We'll look at attributes of features
- Very predictively, we see here an inversion of correlation between sex_male and sex_female in the two shap values plots
- Sex female is highly correlated with high income while in the model trained on real dataset, sex female is correlated with low income
- Sex male is highly correlated with low income while in the model trained on real dataset, sex male is correlated with high income
- White race isn't more correlated with high income than previously which means that the model is not using it as a factor to predict high income

Conclusion: This shows proof of a positive discrimination happening when predicting high income. If the subject is a woman, then she'll more likely have a high income no matter her other attributes like capital, education..





Performance GAN

Performance Wrapper

- **Recall: measure how many relevant elements detected**

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

- In the context of Adult Census, it means for example: out of all women with high income, how many were correctly detected
- Performance wrapper generate a new dataset where the recall is equal for each attribute of the selected features
- We are allowing a gap for the recall between the attributes of a feature
- This type of fairness is called Equal Opportunity

Same Recall for feature Sex and Race, 1% gap

- We notice in the cross differences that proportion of minority attributes (female, black) from the selected features (sex, race) increases
- We also notice that in the synthetic data, there is an augmentation of high degree (masters, bachelors) for these minorities

Conclusion: At this point, it seems that education will have a high impact on the prediction and therefore we want the minorities to have equivalent opportunity to have high income prediction by increasing their proportion with high education

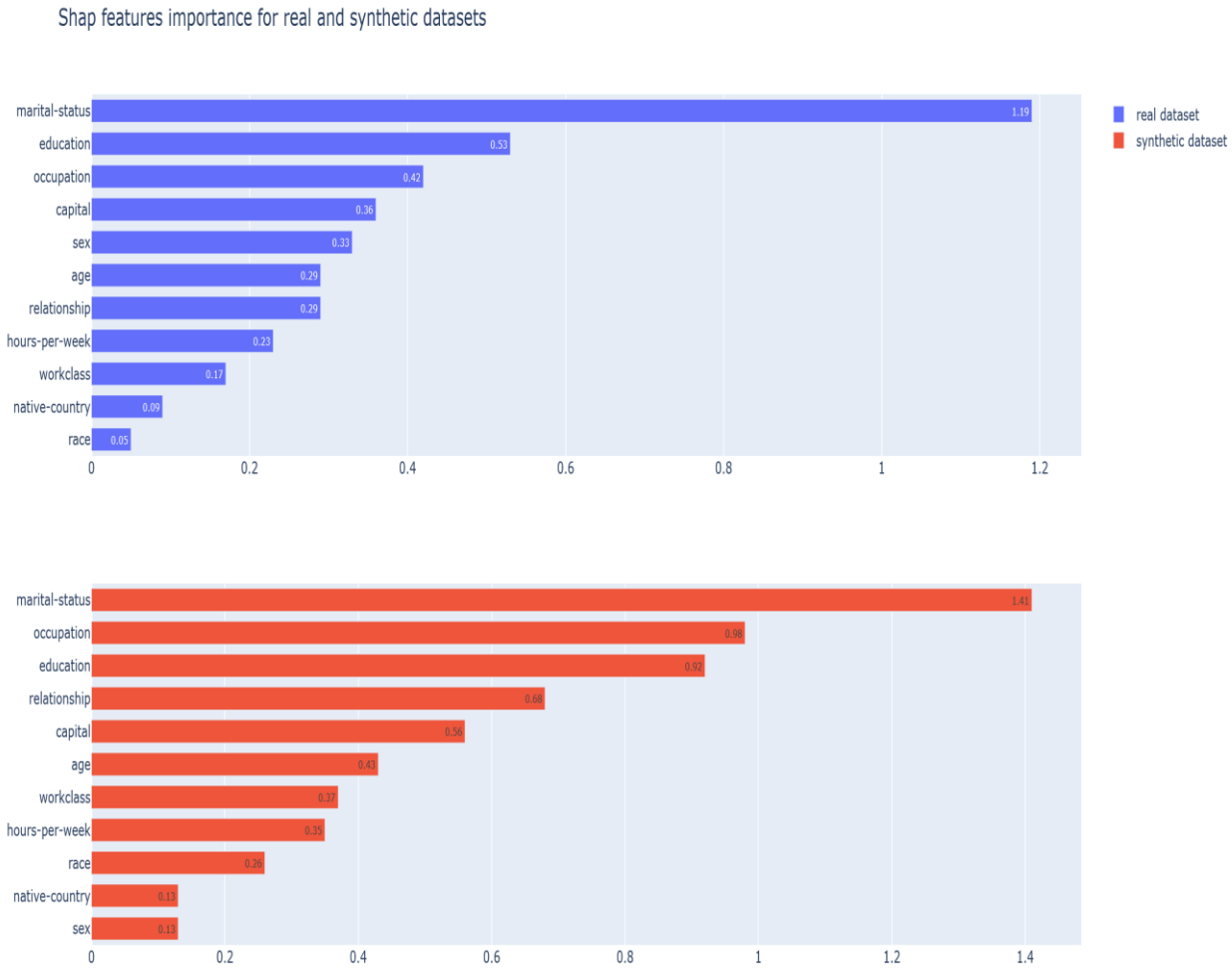
	demographic	differences_with_real_dataset	proportion in dataset
1	Black,Bachelors	7%	1.12%
2	Black,Some-college	7%	2.35%
4	Male,Black	6%	5.22%
0	Doctorate	5.1%	1.22%
2	Black	4.0%	9.75%
8	Female,Masters	4%	1.61%
13	Female,Bachelors	3%	5.28%
11	Female,Some-college	3%	8.77%
10	White,Doctorate	3%	1.03%
4	Female	2.2%	33.55%

Same Recall for feature Sex and Race, 1% gap

- When looking at non selected features, we observe that their feature importance value has increased.
- Education has indeed a more important role in the prediction, as well as occupation
- Sex feature who had a rather important role decrease in ranking and value
- Race who had a very low importance increase in ranking and value

Conclusion: It seems that in order to meet the criteria of recall for all subgroup, the model had to reduce the importance of sex and increase the importance of race. This means that sex is a bad indicator if we want the same recall for men and women. This makes sense because when using sex as indicator, the model tends to discriminate one above the other as seen previously. And since race was almost not acknowledged previously, it needs to have more importance for the model to predict correctly and equivalently the subjects with high income for each subgroup of race.

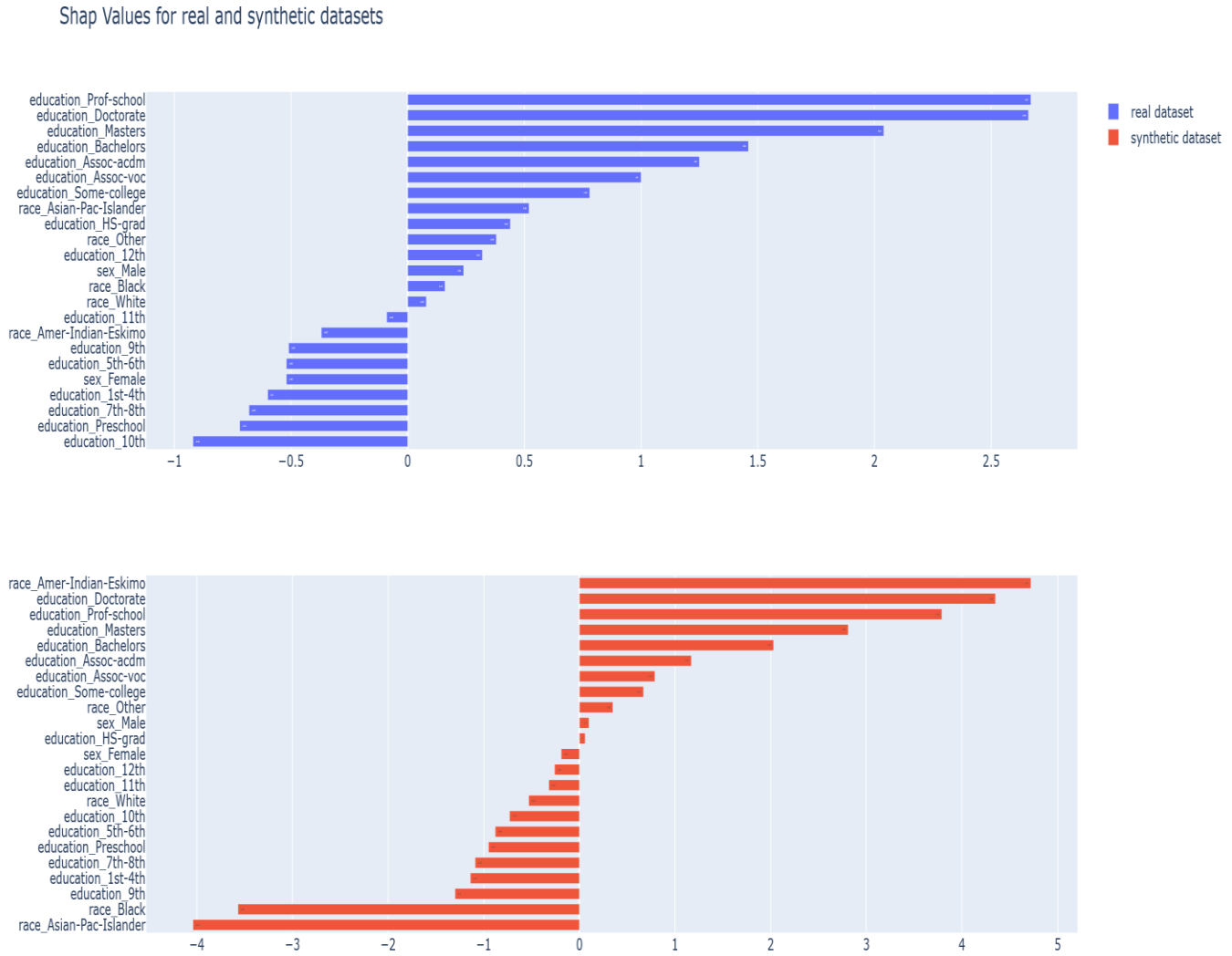
Also, the importance of other features that are better indicators for high income like education and occupation increased.



Same Recall for feature Sex and Race, 1% gap

- We see here that sex is much less correlated with the income no matter the direction of the correlation
- Race_white and race_black becomes negatively correlated with high income or correlated with low income especially race_black
- High education like bachelor, doctorate have their positive correlation increased and low education have their negative correlation decreased even more

Conclusion: This plot validate that sex feature male or female have only a small correlation with high income and therefore a small contribution to the prediction. However, the model will seek high education as indicator to predict high income with more certainty than before in order to respect and optimize the recall for all subgroups of race and sex.



Summary

- Proportion wrapper by adding more women with low income had increased the bias and attribute relationship_husband who is a proxy for sex_male becomes much more correlated with high income
- Allocation wrapper by adding more women with high income inverse the existing correlation between sex and income and perform positive discrimination against women
- Performance wrapper diminish the importance of feature sex, diminish the correlation of women and men with income and increase the importance of features like education and occupation to predict high income with the same recall for men and women