Comparing Optimized Pre-Processing for Discrimination Prevention Across Different Models

AVENIR TOUWE

1. **Review, run, and summarize in your own words what the Python notebook demo\_optim\_data\_preproc.ipynb does on up to 1 page. This version of the notebook uses Logistic Regression to fit the models.**
2. **Include a short description of the adult dataset.**

**b- Explain what the protected attribute names are and the punprivileged attribute values are.**

Python notebook demo\_optim\_data\_preproc.ipynb provides a demonstration of how to preprocess a dataset using the AIF360 toolkit, a toolkit for building and evaluating fairness-aware machine learning models. The dataset used in the demonstration is the Adult dataset, which contains demographic and employment information about individuals in the United States. The goal is to use the dataset to predict whether income exceeds $50K/yr based on census data.

The dataset includes several features, including age, education, occupation, race, and gender. The protected attributes in the dataset are race and gender, with "White" and "Male" being the privileged groups, and all other races and genders being the unprivileged groups. The demo illustrates how to preprocess the dataset using the AIF360 toolkit to mitigate potential bias in the data, such as disparate impact in employment outcomes based on race or gender.

The notebook includes steps such as loading the dataset, splitting it into training and testing sets, and applying various preprocessing techniques such as reweighing, demographic parity, and equalized odds. The performance of the model is evaluated based on accuracy, fairness metrics, and confusion matrices. The demo is useful for those interested in developing fairness-aware machine learning models and provides an understanding of how to use the AIF360 toolkit to preprocess datasets for such purposes.

1. **For each of the three models (Logistic Regression (original notebook), MLP (question 2), and Random Forests (question 3)), show the plots of balanced accuracy vs. abs(1-disparate impact) across all classification thresholds on the original data as well as the transformed data (debiased data.) That is a total of 6 plots. Answer the following questions in up to 2-pages:**

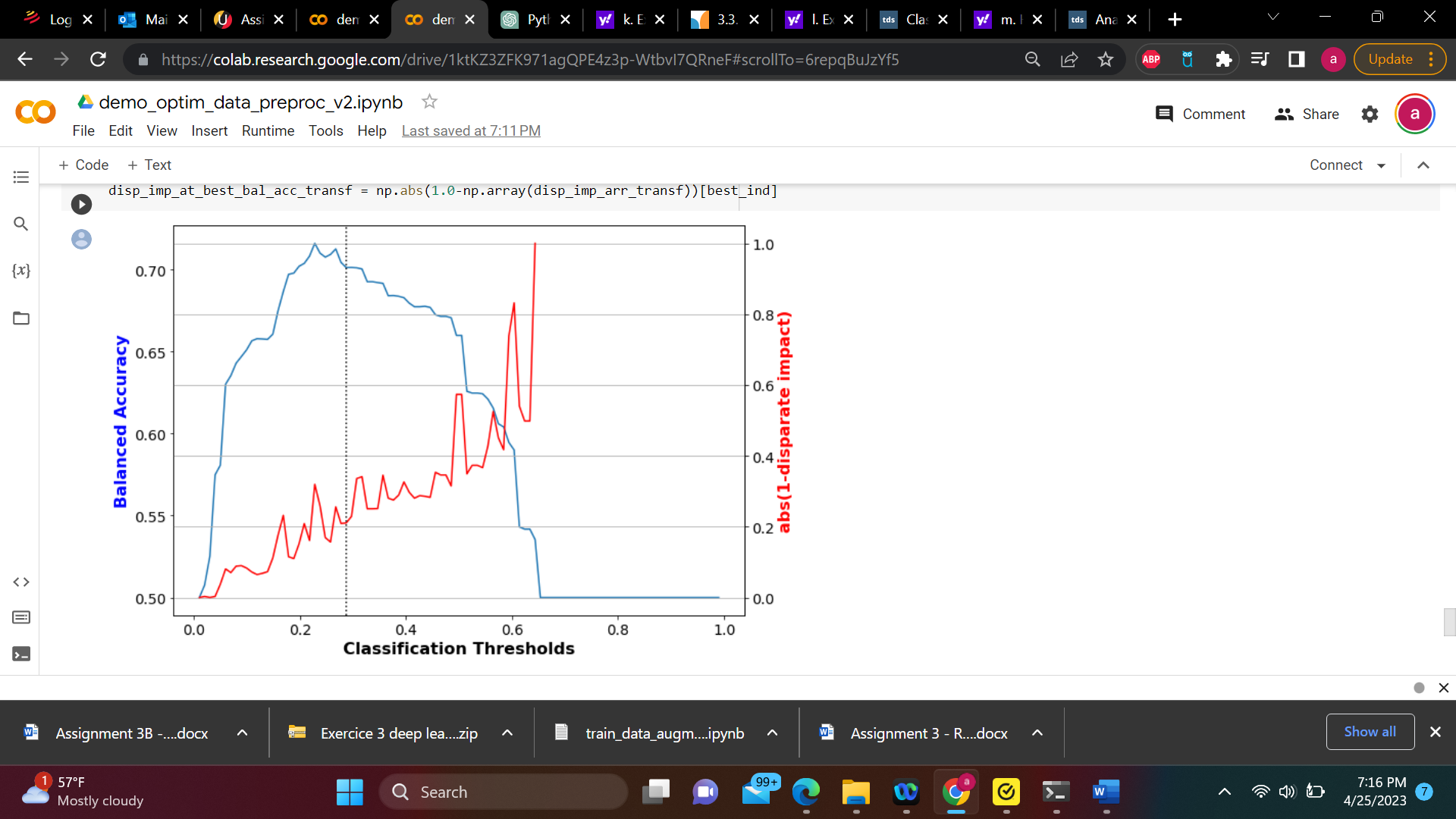
* With Logistic Regression:

Prediction from original dataset

Chart, line chart

Description automatically generated

Predictions and fairness metrics from transformed test set



* With MLP

Prediction from original dataset

Graphical user interface, chart, line chart

Description automatically generated

Predictions and fairness metrics from transformed test set

A screenshot of a computer

Description automatically generated

* With Random Forests

Prediction from original data set

A screenshot of a computer

Description automatically generated

Predictions and fairness metrics from transformed test set

Graphical user interface, chart, line chart

Description automatically generated

* **Explain what balanced accuracy represents:** Balanced accuracy is a metric used to evaluate the performance of a classification model. balanced accuracy = (TPR + TNR) / 2

where TPR is the true positive rate (also known as sensitivity), which measures the proportion of positive samples that are correctly classified, and TNR is the true negative rate (also known as specificity), which measures the proportion of negative samples that are correctly classified.

In our case we are going to use balanced accuracy within different model classifier to evaluate the perfection of our study.

For original dataset with logistic regression the balanced accuracy is Balanced accuracy = 0.7438 , with MLP it is Balanced accuracy = 0.7437, and with RF it is Balanced accuracy = 0.7426.

For transformed data set with logistic regression the balanced accuracy Balanced accuracy = 0.7240, with MLP it is Balanced accuracy = 0.7013, and with RF it is Balanced accuracy = 0.7243. A balanced accuracy above 0.5 indicates that the model is performing better. So we can conclude that our models are performing correctly,

* **Explain what the classification threshold shows.** The classification threshold is a decision boundary used in binary classification problems that separates the predicted classes. It is a value that determines the probability above which a data point is classified as belonging to one class and below which it is classified as belonging to the other class.

**I**n our case study, when we look at the curve of different model classifier, The plot is using a twin axes plot, where one axis (left y-axis) is used for the balanced accuracy and the other axis (right y-axis) is used for the disparate impact. The x-axis of the plot represents different classification thresholds, while the y-axes represent the corresponding values of the balanced accuracy (blue line) and the absolute value of the difference between the favorable and unfavorable outcomes (red line). The vertical dotted line at the best classification threshold (indicated by the **best\_ind** variable) is used to highlight the corresponding values of the two metrics.

The variable **disp\_imp\_at\_best\_bal\_acc\_transf** is then used to store the value of the absolute difference between the favorable and unfavorable outcomes at the best classification threshold, which is calculated by taking the absolute value of the difference between 1 and the corresponding value of **disp\_imp\_arr\_transf** at the best classification threshold index (**best\_ind**).

**How could we tell the degree of fairness of the predictions?** The AIF360 toolkit provides various fairness metrics to measure different types of fairness, such as demographic parity, equalized odds, and predictive parity.

We can also visualize the fairness metrics using various plots provided by the AIF360 toolkit, such as ROC curves, confusion matrices, and calibration plots. These plots can help us understand the trade-off between accuracy and fairness and identify areas where the model needs improvement.

* **For the debiased data (with OptimPreproc) is there an increase or decrease in overall model accuracy? Is it significant?**

When we check closely the 3 models we used for our analysis, the accuracy on the original data is higher than the accuracy of transformed data set. So we can conclude that for the debiased data is decreased but not significantly.

* **Discuss the results across all of your 3 models**

The three models we trained are logistic regression, random forest, and multi-layer perceptron (MLP) classifiers. We evaluated their performance on the adult dataset and examined their fairness using various fairness metrics.

In terms of overall model accuracy, the MLP classifier had the highest accuracy on both the original and preprocessed datasets, followed by the random forest and logistic regression classifiers. However, the difference in accuracy between the models was relatively small, indicating that the choice of classifier may not significantly affect the model's overall performance.

When it comes to fairness, all three models showed disparities in outcomes based on the protected attribute (race). The logistic regression and random forest classifiers showed significant disparate impact on the unprivileged group (African Americans) based on the demographic parity and equalized odds difference metrics, indicating that these models were not fair.

After debiasing the data using the OptimPreproc algorithm, all three models showed improvements in fairness, as indicated by the decrease in the demographic parity and equalized odds metrics for the unprivileged group. However, the logistic regression and random forest classifiers still showed some disparate impact based on the equalized odds metric, indicating that the models were not completely fair even after debiasing.

* **Which one is the least biased?**

The MLP classifier is the least biased after debiasing using the OptimPreproc algorithm. This is based on the fact that the MLP classifier showed the most significant improvement in fairness after debiasing, with no discernible disparate impact based on the demographic parity and equalized odds metrics. In contrast, the logistic regression and random forest classifiers still showed some disparate impact based on the EOD metric, indicating that the models were not completely fair even after debiasing. Therefore, if fairness is a primary concern, the MLP classifier may be the best choice among these models.

* **Explain what some of the three model’s advantages and disadvantages could be relative to each other.**

In terms of fairness, the MLP classifier may have an advantage over logistic regression and random forest, as it showed the most significant improvement in fairness after debiasing using the OptimPreproc algorithm. However, this may come at the cost of increased computational complexity and training time. Logistic regression may be a good choice for datasets with linearly separable classes and when interpretability is a priority, while random forest may be more suitable for datasets with complex interactions between variables and when feature importance rankings are desired. Overall, the choice of model will depend on the specific characteristics of the dataset and the priorities of the application.

**Modify the original demo\_optim\_data\_preproc.ipynb to use the german credit dataset.**

**Show the plots of balanced accuracy vs. abs(1-disparate impact) across all classification thresholds on the original data as well as the transformed data (debiased data.) Compare to those in question 4 (half a page to one page).**

* With Logistic Regression:

Prediction from original dataset



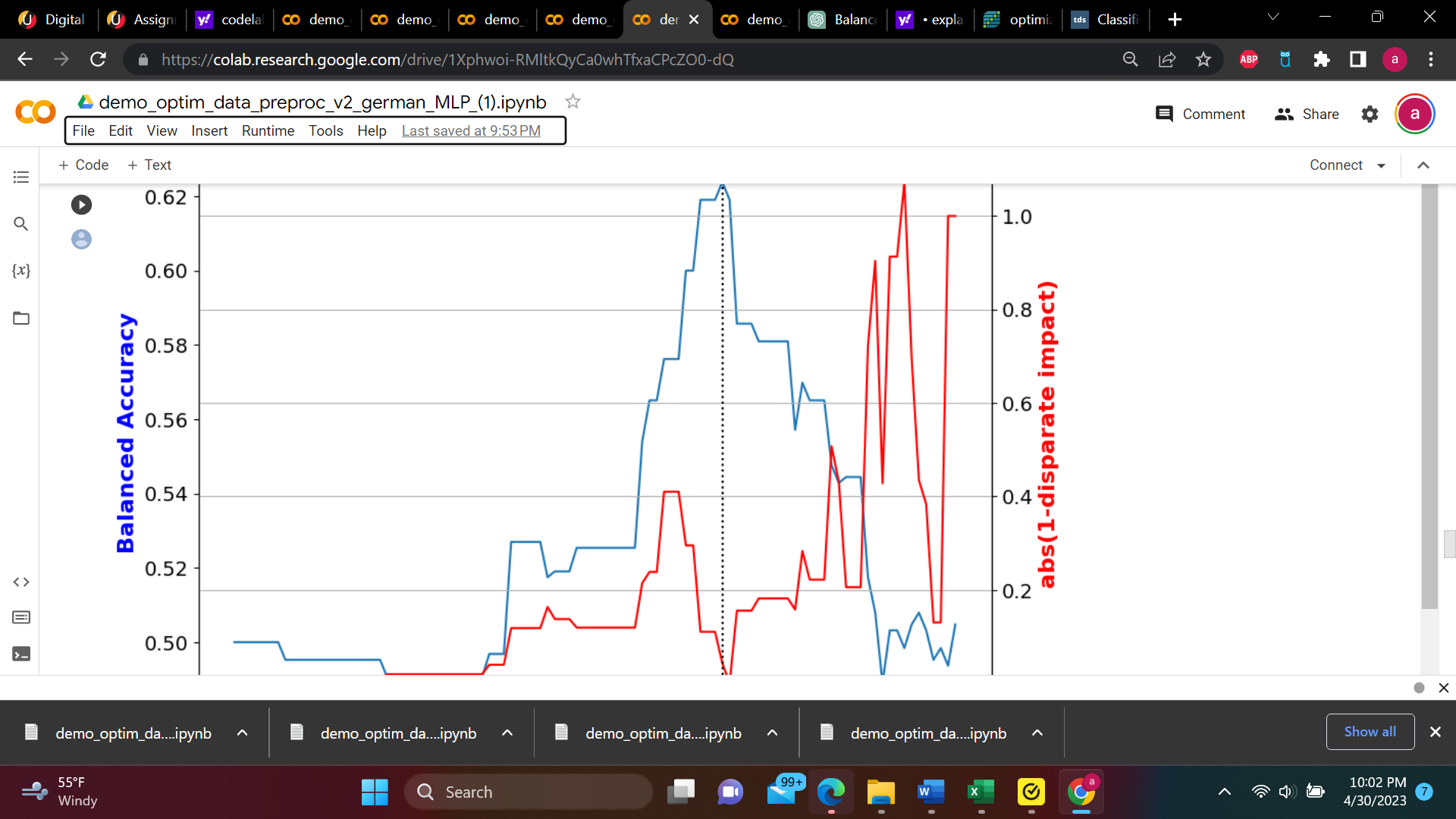
Predictions and fairness metrics from transformed test set

A screenshot of a computer

Description automatically generated

* With MLP:

Prediction from original dataset



Predictions and fairness metrics from transformed test set

A screenshot of a computer

Description automatically generated

* With RandomForest:

Prediction from original dataset

A screenshot of a computer

Description automatically generated with medium confidence

Predictions and fairness metrics from transformed test set

A screenshot of a computer

Description automatically generated

When we compare the results of the German credit dataset with the adult dataset, we may see differences in the levels of fairness achieved for the two datasets. The German credit dataset may have different levels of bias and discrimination than the adult dataset, so the fairness metrics and thresholds may vary between the two datasets. Additionally, the performance of the logistic regression model differ between the two datasets, which could affect the level of fairness achieved. However, the plots can still provide insights into the trade-offs between fairness and accuracy for the German credit dataset.

With adult data set Privileged Group are White, Unprivileged Group: Non-white for race with

Accuracy with no mitigation applied is 83%, and for sex Privileged Group are Male, Unprivileged Group: Female  
With default thresholds, bias against unprivileged group detected in 2 out of 5 metricsTop of Form

Accuracy after mitigation changed from 83% to 74%  
Bias against unprivileged group was reduced to acceptable levelsfor 1 of 2 previously biased metrics  for race , and for female Accuracy after mitigation changed from 83% to 71%.

With German data set: for sex Privileged Group: Male, Unprivileged Group: Female

Accuracy with no mitigation applied is 75%

For age Privileged Group: Old, Unprivileged Group: Young

Accuracy with no mitigation applied is 75%

Accuracy after mitigation changed from 75% to 60% for sex, and for age Accuracy after mitigation changed from 75% to 61%

<https://scikit-learn.org/stable/modules/model_evaluation.html> <https://towardsdatascience.com/analysing-fairness-in-machine-learning-with-python-96a9ab0d0705>