1. INTRODUCTION

Machine learning systems and algorithms have become ubiquitous in almost every industry and business that we rely on. This includes the food, finance, marketing, banking, and housing industries. Although getting the models to be incorporated properly is a task, but also having the model to perform as expected is a challenge that must be solved properly. Machine learning models rely heavily on the data that it is provided with, without considering various factors such as equal representation in data, loss of data, bias etc. This is where fairness comes in and is used to balance out the data by adding weights to the unprivileged data and preventing the ML model from teaching itself bias. This report will cover two methods that can be used to not only perform fairness analysis, but also to try and increase the fairness to support the unprivileged groups. All the datasets and models run will use the AIF360 toolkit which is an extensible open-source toolkit that can help examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle.

2. DATASETS ANALYSED

The report analyses two different datasets. Namely, the Adult Census dataset and the German Credit dataset. The Adult dataset is a very commonly used dataset for performing and testing various prediction and fairness algorithms. The adult dataset consists of information regarding the earnings of the people living in the USA and the German dataset consists of credit data from banks.

3. ADULT DATASET

The adult dataset has 15 columns and 32561 rows, and contains attributes such as age, sex, work class, education, income, relationship etc. In this dataset, the ‘sex’ class is the sensitive class, with ‘Male’ being considered as privileged and ‘Female’ being considered as unprivileged. The data is directly imported from the aif360 library, to take direct advantage of the library’s features. Once the data is imported and the protected, privileged, and unprivileged classes are provided, the numerical and categorical features are separated. The numerical features are age, education-num, capital-gain, capital-loss, and hours worked per week. The categorical features are work class, education, marital-status, occupation, relationship, race, and native country. The target feature is income-per-year.

3.1 STANDARD MODEL

3.1.1 EDA

A Fairness based exploratory data analysis is performed on the dataset using the binary label dataset metric as we are checking for group fairness on a single binary dataset. Information such as base rate of the entire population, privileged group, unprivileged group, statistical parity difference and disparate impact is calculated.

Base rate on entire population	0.24
Base rate on privileged group (males)	0.30
Base rate on unprivileged group (females)	0.11
Statistical parity difference	-0.19
Disparate impact	0.36

Figure 1: Binary Label Dataset Metric table with the values

The annual income is also calculated, and it is grouped by sex. The probability of having income greater than $50k is also calculated for the to sex categories. The difference of income distribution can be clearly observed in this bar graph.

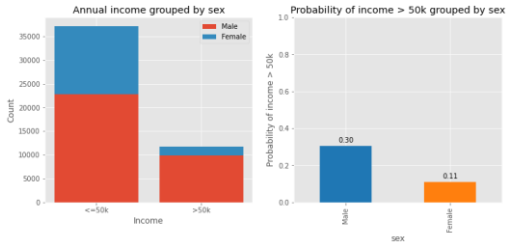

Figure 2: a)Annual income grouped by sex, b)Probability of income greater than $50k

From the above graphs, the following observations can be made:

1. The 'sex' attribute is dominated by 'Male' with 67% compared to 'Female' values with 33%.
2. The dominance of 'Male' value does not by itself indicate bias/unfair treatment.
3. The percentage of males with income > 50k w.r.t. total sample is 20%.
4. The respective percentage of females is 4%. The within-class percentage of females with income > 50k is 11%.
5. The respective percentage for males is 30%. Several fairness metrics were calculated for the dataset at hand. More specifically, statistical parity difference and disparate impact indicate the existence of bias in the dataset.

The base rate is defined as P(Y=1) = P/(P+N), optionally conditioned on protected attribute.

3.1.2 TRAIN TEST SPLIT AND PRE-PROCESSING

After performing the exploratory data analysis, and recording the initial fairness metrics, the data is split into two, namely train and test. The split is done as per the specified percentage of 70% for the training data. Once splitting is done, the data is scaled using the MinMaxScalar. The MinMaxScalar scales the minimum and maximum value between the 0 and 1. Statistical Parity Difference measures the difference of the above values instead of ratios; hence we would like it to be close to 0.

For the y datasets, ravel is used to flatten the array, so that it can be used later.

3.1.3 CROSS-VALIDATION

To perform the required analysis, the classification algorithm used is Logistic Regression. Since the relationship between the features and the target aren't too complex, and tuning is easy, Logistic Regression is used. Five-fold classification is performed on the model to find the best parameters using GridSearchCV. GridSearchCV runs through all the provided different parameters that is fed into the parameter grid and produces the best combination of parameters based on the scoring metric.

The solvers that were used to tune the hyperparameters are liblinear, and newton-cg. Library for Large Linear Classification, or short for liblinear uses a coordinate descent algorithm which is based on minimizing a multivariate function by solving univariate optimization problems in a loop. In other words, it moves toward the minimum in one direction at a time. Newton.cg which is short for newton method use an exact Hessian matrix. It's slow for large datasets because it computes the second derivatives. The cross-validation is also run for random-state, to find if there are any state's that provides better accuracy. The model is run for 5 folds and has 8 candidates and 40 fits. The whole cross validation model is run twice, once for finding the parameters with the best accuracy and another to find the parameters that will find the best roc_auc model. The best parameter model is rerun for finding the accuracy and the confusion matrix. Our confusion matrix thus represents four possible states:

- **True positive**: Model predicts >50K, and that is the ground truth.
  True negative: Model predicts <50K, and that is the ground truth.
- **False positive**: Model predicts >50K, and that contradicts reality.
- **False negative**: Model predicts <50K, and that contradicts reality.

Along with the confusion matrix, different metrics are also calculated, such as statistical parity difference, equal opportunity difference, average of difference, balanced accuracy, and disparate impact for both the models.

Equal opportunity difference measures the ability of the classifier to accurately classify a datapoint as positive regardless of the presence of the unprivileged feature. We would like it to be close to 0. A negative value signals bias towards privileged.

Average of difference in FPR and TPR for unprivileged and privileged groups. A value of 0 indicates equality of odds.

Balanced accuracy is a general metric, not dependent on bias. We would like to have it close to 1, meaning that the classifier can equally detect positive and negative classes.

We would like disparate impact to be close to 1. It measures the ratio between the likelihood of the class being predicted as positive if we have the unprivileged feature and the same likelihood with the privileged feature. Values close to 0 indicate strong bias.

Statistical Parity Difference	-0.1661
Equal opportunity difference	-0.2731
Average of difference	-0.1700
Balanced accuracy	0.6856
Disparate impact	0.1516

Figure 3: Metric info of standard model

3.2 FAIRNESS MODEL

Fairness models are used to help balance out the data. There are multiple methods to do this, including reweighing, mitigation with pre-processing and disparate impact. In this project, Disparate Impact was used as a metric to evaluate fairness. DI model works by comparing proportions of individuals that receive a positive output among the two groups: privileged and unprivileged.

$$\frac{Pr(Y=1|D=\text{unprivileged})}{Pr(Y=1|D=\text{privileged})}$$

Figure 4: Formula to calculate Disparate Impact

The Disparate Impact Remover is a pre-processing technique that edits the values that can be used as features to increase the fairness between group. The algorithm makes use of the repair level value that help is changing this feature. Classification Metric is also used to classify the data. The false positive rate difference, false negative rate difference, error rate difference, true positive rate difference and the mean difference is also re-calculated towards the end. The fairness model is run using the logistic regression with the liblinear solver and balanced class weight.

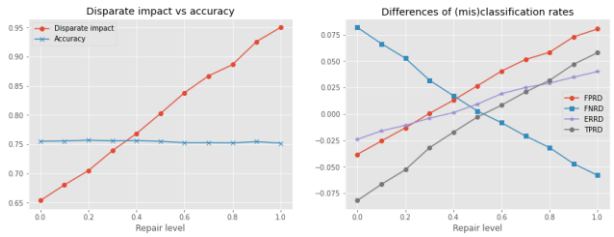

Figure 5: a) Disparate impact vs accuracy, b) Difference of misclassification rates

FPRD(False Positive Rate Difference): $FPR_{D=\text{unprivileged}} - FPR_{D=\text{privileged}}$
FNRD (False Negative Rate Difference): $FNR_{D=\text{unprivileged}} - FNR_{D=\text{privileged}}$
ERRD(Error Rate Difference): Difference in error rates for unprivileged and privileged groups
$ERR_{D=\text{unprivileged}} - ERR_{D=\text{privileged}}$, and ERR=(FP+FN)/(P+N)

The fairness-unaware performance of the Logistic Regression model is obtained with repair_level=0, for a specified set of model hyperparameters e.g., solver. For example, the newton-cg solver achieves better accuracy

than liblinear solver, however the bias removal task is more challenging.
By increasing the repair_level, disparate impact score increases resulting to a fairer classifier; at the same time, the accuracy is not affected. For repair_level=1 we achieve the best disparate impact ratio equal to 0.95.
As far as (mis)classification rate differences are concerned, by removing disparate impact the respective FPRD and FNRD metrics are also reduced, up until a point (repair_level~0.4) where they reach a value close to 0. By increasing repair_level more than this threshold, the differences begin to increase again, in the opposite direction.
After the fairness model is run, the following values are given out by the model. It can be observed that the Disparate Impact has increased drastically.

Statistical Parity Difference	0.34
Base rate on privileged group (males)	0.34
Base rate on unprivileged group (females)	0.33
Statistical parity difference	-0.02
Disparate impact	0.95

Figure 6: Fairness model metrics

From the below graph it can be observed that there is not as much variation is represented data between the classes.

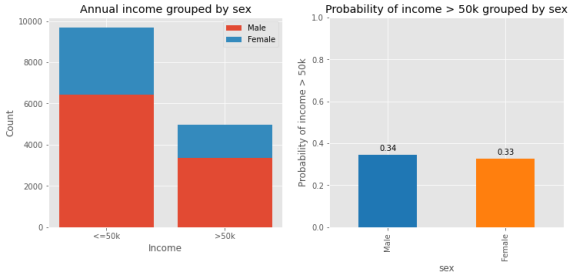

Figure 7: Fairness visualized data

Extra work was also done by taking 'race' to be the sensitive group, in which white is the privileged class and black is the unprivileged class. The results are discussed in the appendix.

4. GERMAN DATASET

One method by which a bank makes money is by collecting interest from loans. Which mean that the most important factor here is repayment. A bank receives multiple applications for loans, but only sanctions the loan after going through the applicant's profile. The application process has to be well thought of and must be able to provide loans to right applicants. There are generally two types of risks involved in this process:

- If an applicant with a good credit risk, who can repay the loan applied for and not getting an approval could result in a loss for the bank.
- If the applicant has a bad credit risk and can most likely not pay the loan and getting it approved from the bank could result in a loss to the bank.

4.1 STANDARD MODEL

4.1.1 EDA

A similar type of fairness EDA is performed on the dataset to get a reading of the data, and to understand the bias. The bias between the sexes is not that much, but when the data is observed for credit provided the bias is clearer and more apparent. The data is directly imported from the AIF360 library. Binary Label Dataset Metric is used to get the base rate for the privileged and unprivileged data, the mean and the disparate impact.

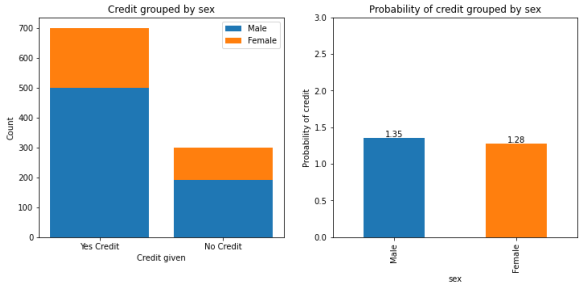

Figure 8: Credit Fairness metric graph

The base rate of the entire population is found to be 0.70, and the disparate impact is at 0.90, which isn't much. But the base rate of the privileged group (Male) is found at 0.72 and the unprivileged group (Female) is at 0.65, which is a lot of difference.

Base rate on entire population	0.70
Base rate on privileged group (males)	0.72
Base rate on unprivileged group (females)	0.65
Statistical parity difference	-0.07
Disparate impact	0.90

Figure 9: Fairness visualized data

4.1.2TRAIN TEST SPLIT

The data is split into 70% train and 30% test data as per the report requirements. Then the standard scalar is applied on the train data, to scale the data before it is to be processed in cross validation.

4.1.3 CROSS VALIDATION

The model used here is Logistic Regression. Since the data is small, the hyperparameters that are to be tuned here are newton-cg, lbfgs and liblinear in the solver, and the number of maximum iterations is also tuned between 200 to 2000. The tuning is done to find the parameters that provide the best accuracy for 5 folds. After the 5-fold cross validation is performed, the best model is the one

with a maximum of 200 iterations and the one that uses the newton-cg solver. On running this model, and accuracy of 71.33% was achieved.

Statistical Parity Difference	-0.1079
Equal opportunity difference	0.0269
Average of difference	-0.1195
Balanced accuracy	0.5813
Disparate Impact	0.8832

Figure 10: Metric info of credit standard model

The disparate impact, as checked above is at 0.88, and the statistical parity difference is -0.107. The equal opportunity difference is 0.0269, which means that there is not much bias in the data. The -0.11 average of difference states that there is an equality of the odds, and the 0.58 of the balanced accuracy means that the model is not very good at detecting the negative and positive classes.

4.2 FAIRNESS MODEL

The method to help remove the bias was disparate impact remover as it aims to remove this ability to distinguish between group membership. The algorithm requires the user to specify a repair_level, this indicates how much you wish for the distributions of the groups to overlap.

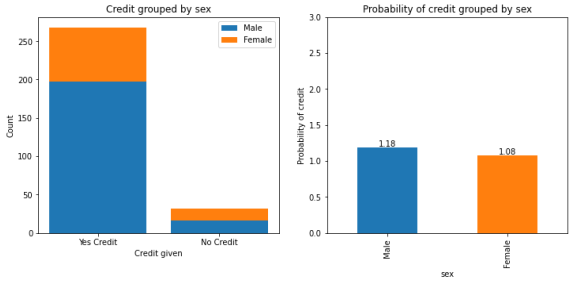

Figure 11: Fairness metric post biasing

From the above graph, biasing has been done and data has been added to support the unprivileged class. But this came at the cost of the model losing fairness and providing more loans. This is not an ideal model as it would mean that the bank might lose money by providing loans to bad credit people. Some further work and tuning is required in this matter.

5. TASK 3

The method that can be used to compare the models above, can be brought from the FairMLHealth library. This library has many metrics and comparison feature that will be useful to compare two models and evaluate which one is better. Unfortunately, the model couldn't be executed in the code as a deeper understanding was required on how to use the tool. The FairMLHealth library has a tool such as report, measure, and compare. The compare tool can be used to compare two regression models with the following code:

```
report.compare(test_data = X_test,  targets = y_test,
                protected_attr = X_test['gender'],
              models = {'Any Name 1':model_1, 'Model
                2':model_2},
             pred_type="regression")
```

This code will compare the models and the best model can be analyzed and chosen.

6. COMPARING REWEIGHING AND DISPARATE IMPACT REMOVER

Reweighing is a pre-processing algorithm that assigns weights to dataset points in such a way that their weighted discrimination is 0 when compared to the designated group. Disparate Impact compares the proportion of individual group outputs, the proportion of the unprivileged group that received the positive outcome divided by the proportion of the privileged group that received the positive outcome. The industry standard is a four-fifths rule: if the unprivileged group receives a positive outcome less than 80% of their proportion of the privileged group, this is a disparate impact violation. However, this maybe increased for the business. Both the models are used to improve the fairness metric.

7. CONCLUSION

Bias in the machine world is very complex to solve. There is a plethora of tools that can be used to manipulate the data, or the results in such a way that the bias is mitigated, and equal opportunity is provided. Due to the integration of AI and ML in the daily lives of us humans, having an unbiased system is vital to have equal opportunities for everybody. Making a balanced system that is not only unbiased, but also maintains accuracy is very difficult and requires a lot of tuning. This project has made these concepts clear and has also brought to light the challenges that have to be overcome to be able to make a proper unbiased and accurate system.

8. APPENDIX

Base rate on entire population	0.24
Base rate on privileged group (white)	0.25
Base rate on unprivileged group (black)	0.15
Statistical parity difference	-0.10
Disparate impact	0.60

Figure 12: Pre-Fairness metric of adult dataset, with race as sensitive data(Adult-race)

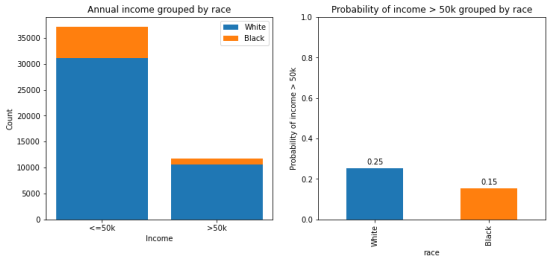

Figure 13: Pre-Fairness metric graph of adult dataset, with race as sensitive data where a clear bias is visible(Adult-race)

LOGISTIC REGRESSION-1

```
lr1=LogisticRegression(solver='liblinear', random_state=0)
lr1.fit(X_train,y_train)
predictions1 = lr1.predict(X_test)
print("Logistic Regression 1 Accuracy: %0.2f" %(100 * accuracy_score(y_test,predictions1)) + "%")
print("\nConfusion Matrix: \n", confusion_matrix(y_test, predictions1))
```

```
Logistic Regression 1 Accuracy: 81.44%

Confusion Matrix:
 [[10578   517]
 [ 2202  1356]]
```

Figure 14: The best parameter after logistic regression is liblinear and random state 0, the accuracy is 81.44% (Adult-race)

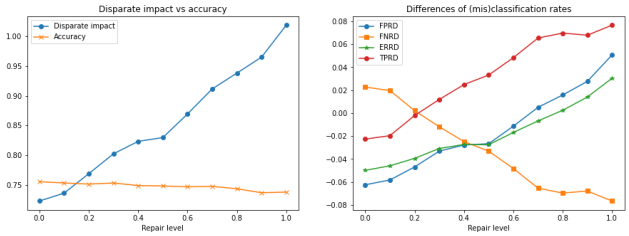

Figure 15: The graphs in which the repair value is changed to observe the variation of accuracy and disparate impact, and other metric values. Increase in repair value is causing the accuracy to drop and the disparate impact to improve (Adult-race)

Base rate on entire population	0.38
Base rate on privileged group (white)	0.38
Base rate on unprivileged group (black)	0.39
Statistical parity difference	0.01
Disparate impact	1.02

Figure 16: The metric table after Disparate Impact is removed. A clear change in the bias is observed, and the results are really satisfactory (Adult-race)

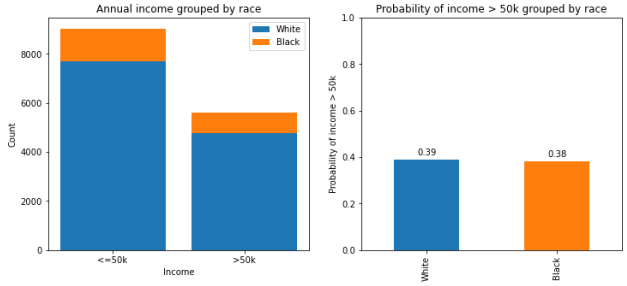

Figure 17: Graphical representation of the data post disparate impact removal (Adult-race)

9. REFERENCES

[1] *AIF360* - https://github.com/Trusted-AI/AIF360

[2] *AIF 360 official website* - https://aif360.readthedocs.io/en/latest/index.html

[3] *What is Logistic Regression* - https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

[4] Stacey Ronaghan, '*AI Fairness — Explanation of Disparate Impact Remover*' - https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f1

[5] *FairMLHealth Documentation on Github* - https://github.cm/KenSciResearch/fairMLHealth