**Visualising Accuracy vs. Fairness in ML models**

**Introduction**

In the pursuit of building highly accurate machine learning models, it's often easy to overlook the potential biases hidden within these mathematical constructs. Our primary focus becomes achieving the best predictive performance, sometimes at the cost of introducing unfair or discriminatory outcomes. AI Fairness 360 is a pioneering initiative dedicated to unearthing and addressing these biases. It conducts rigorous research and offers powerful methods to mitigate disparities in machine learning models.

But wait, how can a mathematical model be biased against humans? Well, the reasons are as follows:

1. **Biased Training Data**: Machine learning models learn from historical data, and if this data contains biases or discrimination, the model is likely to inherit those biases. For example, if historical hiring data favours certain demographic groups over others, a model trained on this data may perpetuate hiring biases.
2. **Data Collection Methods**: Bias can be introduced during data collection. For example, if data collection processes favour certain groups or areas over others, the resulting data will reflect these biases.
3. **Historical Inequities**: Biases may also stem from systemic inequalities that have existed in society for a long time. These inequities can be reflected in the data used to train models, leading to biased outcomes.
4. **Labelling Bias**: Biases can arise from the way data is labelled. Human annotators or data labellers may inadvertently introduce their own biases when labelling data.
5. **Sample Size**: If certain groups are underrepresented in the training data, the model may not learn to make accurate predictions for those groups, leading to bias against them.
6. **Feature Engineering**: The selection of features or variables used in the model can introduce bias if certain features are chosen based on subjective criteria or if they carry underlying biases.

In this exploration, we delve into one of AI Fairness 360's datasets, known as the "Adult Income" dataset. Our aim is to visualize and understand modelling bias concerning an individual's sex, shedding light on the importance of fairness and equity in the world of artificial intelligence.

**About the dataset**

The AI Fairness 360 (AIF360) toolkit provides a collection of datasets for use in fairness, bias, and discrimination research in machine learning. One of the datasets included in AIF360 is the "Adult Income" dataset. This dataset is used for income prediction, typically for tasks like income classification.

**Task:** The primary task associated with the Adult Income dataset is income prediction. The goal is to predict whether an individual's income exceeds $50,000 per year based on various demographic and employment-related features.

**Features:** The dataset contains a set of features that describe individuals, including age, education, occupation, marital status, race, gender, hours worked per week, etc.

**Label:** The target variable or label is a binary variable that indicates whether an individual's income exceeds $50,000 per year (1 for income > $50K, 0 for income <= $50K).

**Protected class:** There are two protected classes in the dataset, namely race and sex. For this analysis, we have chosen to focus on sex as a factor that may introduce bias. In this context, the privileged class is "Male" (labelled as 1), while the unprivileged class is "Female" (labelled as 0).

**Python specifications**

Environment – Jupyter

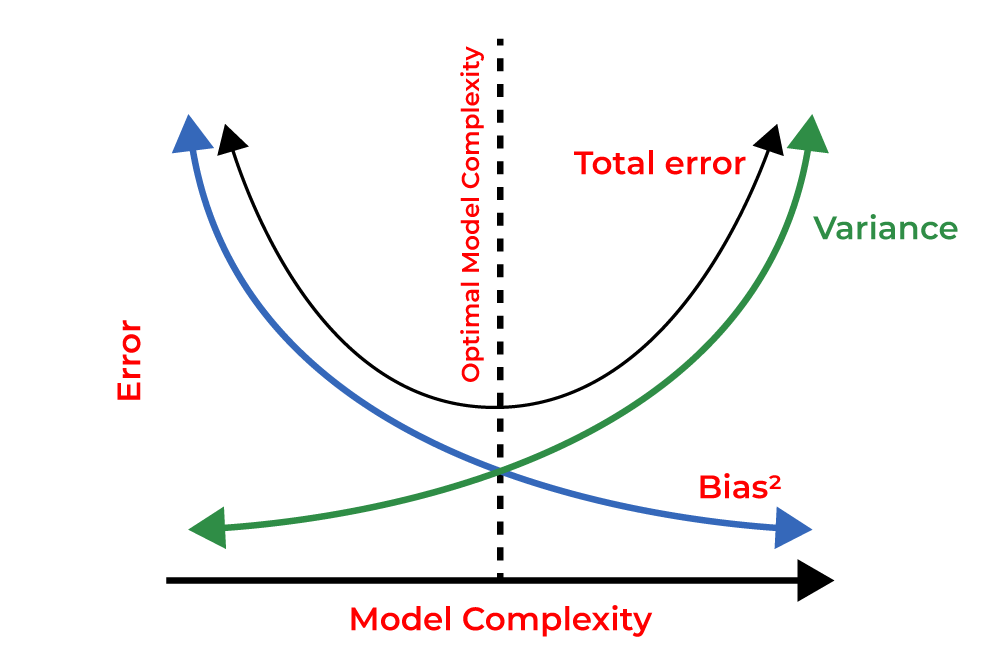
Classifier – Random Forest

Key modules – AIF360 (similar to sklearn but for fairness analysis)

**Bias-variance**

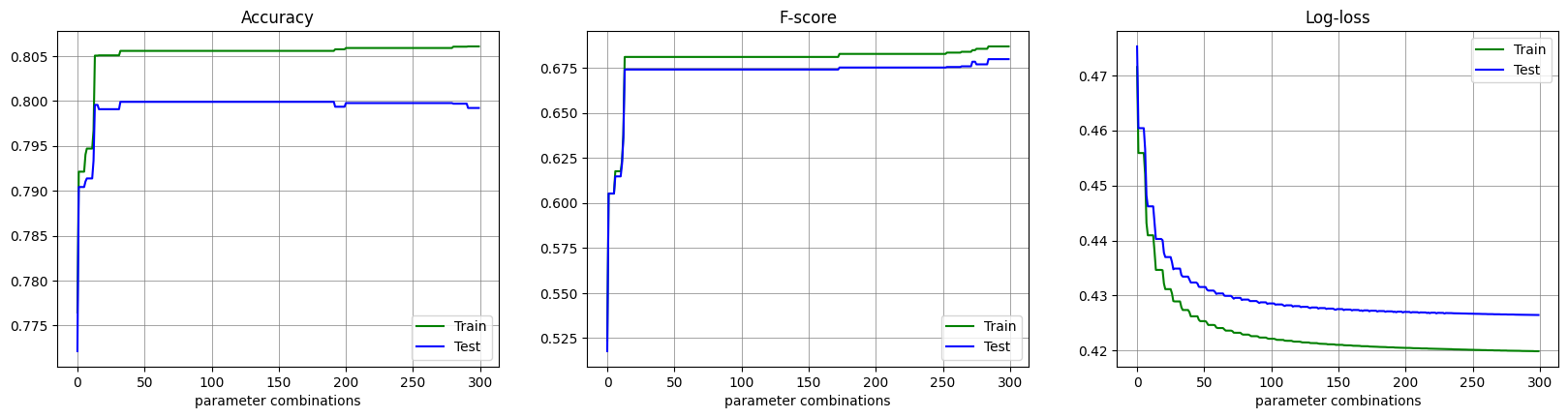
Before delving into assessing the fairness of a predictive model, it's crucial to grasp the trade-off between bias and variance. These two concepts are fundamental in the context of model performance and have a significant impact on a model's ability to make accurate predictions.

Bias in a machine learning model represents the error that occurs when the model simplifies complex real-world data and fails to capture the underlying patterns and trends. High bias leads to underfitting, where the model is overly simplistic and performs poorly on both the data it was trained on (the training data) and unseen data (the testing data). To mitigate bias, one can explore strategies such as using more complex models, increasing the model's capacity, or incorporating more informative features.

Variance, on the other hand, describes how sensitive a model is to small fluctuations or noise in the training data. High variance results in overfitting, a scenario where the model fits the training data extremely closely but struggles to generalize to new, unseen data. To reduce variance, it's advisable to simplify the model, apply regularization techniques, increase the amount of training data, or carefully select relevant features.

The bias-variance trade-off is a central concept in machine learning. Striking the right balance between bias and variance is essential for a model to generalize effectively to new, unseen data. Techniques like cross-validation, hyperparameter tuning, and feature engineering play a pivotal role in finding this equilibrium and constructing models that excel in making predictions on previously unobserved data.

In cases of highly imbalanced datasets, the importance of proper data handling becomes evident. Unlike balanced datasets where accuracy is a commonly chosen evaluation metric, imbalanced datasets like this one benefit from metrics like the F-score or log loss for more precise model assessment. These metrics account for the imbalance and provide a more accurate measure of a model's performance, particularly in situations where minority classes are of significant importance.



The provided figure illustrates a common scenario where a model's performance diverges at a certain point, indicating a shift from better performance on the training data to less favourable results on the test data. Ideally, the chosen model should strike a balance between these two data sets, delivering optimal results within acceptable error margins. The three graphs, from left to right, portray the progression from underfitting to overfitting, with the desired model residing somewhere in between.

Apart from the techniques mentioned for addressing overfitting, model selection plays a pivotal role. Ensemble algorithms like Random Forest offer a robust solution to overfitting because they consider results from multiple trees with varying performance levels. This diversity often results in a more balanced and reliable model that generalizes well to new, unseen data.

Having discussed bias and variance in predictions, let's delve into the core of our analysis. The analysis is structured into two main sections:

In the first section, we conducted hyperparameter tuning for a Random Forest model, resulting in two models with distinct focuses. One model prioritizes accurate predictions, while the other emphasizes fairness in the selection process, particularly for privileged and unprivileged classes. In this section, all samples are assigned uniform weight.

In the second section, we repeated the same steps, but this time we introduced sample weights that consider not only the label class but also the protected class. This approach categorizes samples into four distinct groups: positive-privileged, positive-unprivileged, negative-privileged, and negative-unprivileged.

Before proceeding, a mild overview of two of the fairness metrics used:

**Equal opportunity difference**

EOD measures the disparity in the model's true positive rates (sensitivity or recall) between the privileged and unprivileged groups in a protected attribute (e.g., gender or race). It quantifies the difference in the probability that a true positive prediction is made for each group, indicating whether one group has an advantage in terms of positive predictions. EOD assesses whether the model provides equal opportunities for positive outcomes across different groups. A value of 0 indicates fairness, while a positive value suggests a favourable bias towards the privileged group, and a negative value indicates bias towards the unprivileged group.

**Disparate Impact**

Disparate Impact, also known as Adverse Impact, focuses on the impact of a model's decisions on different protected groups, often with respect to adverse or unfavourable outcomes such as being denied a loan or job. It is typically measured as the ratio of the predicted positive outcomes for the privileged group to the predicted positive outcomes for the unprivileged group. Disparate Impact analysis aims to identify and rectify situations where one group is disproportionately affected or favoured by the model's decisions. A value of 1 indicates no disparate impact, while values significantly above or below 1 suggest potential discrimination.

**Accuracy and fairness with uniform weight**

**Accurate model:**

The selection criteria used for the accuracy focused model was log-loss or cross-entropy loss whose formula is

The metric used for evaluating model fairness was disparate impact. The function to calculate DI is available from the aif360 library.

|  |  |
| --- | --- |
| **Accurate model** | **Fair model** |
| precision recall f1-score support  0.0 0.91 0.70 0.80 11132  1.0 0.46 0.79 0.58 3521  accuracy 0.73 14653  macro avg 0.69 0.75 0.69 14653  weighted avg 0.81 0.73 0.74 14653 | precision recall f1-score support  0.0 0.84 0.83 0.83 11132  1.0 0.48 0.51 0.50 3521  accuracy 0.75 14653  macro avg 0.66 0.67 0.66 14653  weighted avg 0.76 0.75 0.75 14653 |
|  |  |
| Fairness metrics of accurate model on unweighted samples: -  Equal opportunity difference: -0.3934  Disparate Impact: 0.2722 | Fairness metrics of fair model on unweighted samples: -  Equal opportunity difference: 0.0155  Disparate Impact: 0.8343 |

**Discussion:**

Upon initial observation, it appears that the fair model outperforms the accurate model in classification, boasting a higher accuracy of 75% compared to the accurate model's 73%. However, it's important to consider the dataset's high imbalance, primarily favoring one label class (0). Consequently, accuracy alone is an inadequate metric for model evaluation.

A closer examination of recall and F1-score reveals that the accurate model excels as a classifier when compared to the fair model. This is further bolstered by the confusion matrix where we can clearly see which model was better able to accommodate for the class label imbalance and predict the minority class better. It's worth noting that this improvement in predicting the minority class often comes at the cost of misclassifying some of the majority class labels. However, this trade-off is an essential aspect of achieving a more generalized model that accommodates both class extremes.

In the third row, the fair model lives up to its name, delivering values that approach absolute fairness for the evaluation metrics, as seen in the Equal Opportunity Difference (EOD) close to 0 and Disparate Impact (DI) close to 1. The accompanying bar plots provide further evidence that the fairness model narrows the gap in predicting true positives between privileged and unprivileged classes, while the accurate model predominantly favours the privileged class in this regard.



But, can we enhance fairness while preserving optimal classification performance? AI Fairness 360 offers a practical solution for this challenge: assigning weights to samples through a technique known as "Reweighing."

Reweighing is a technique designed to mitigate bias and promote fairness in machine learning models. It can be found as a preprocessing algorithm in the aif360 library. The technique works by adjusting the weights of the samples in the dataset. It aims to give different weights to different samples, particularly focusing on the privileged and unprivileged groups defined by a protected attribute (in this case sex). The goal is to create a balanced dataset, where the model's training process is less influenced by the privileged group.

Here's how reweighing works:

* Calculate Initial Weights: Initially, the dataset is assigned uniform weights for all samples. This means that each sample has the same weight in the training process.
* Identify Protected Groups: The protected attribute is used to categorize the dataset into privileged and unprivileged groups. For example, in gender fairness, male and female groups may be identified.
* Compute Disparities: The disparities in prediction outcomes between the privileged and unprivileged groups are measured. These disparities include metrics like false positive rates, false negative rates, or other relevant fairness metrics.
* Adjust Sample Weights: The sample weights are then adjusted based on the disparities identified. The goal is to assign more weight to underprivileged or disadvantaged groups and reduce the weight of privileged groups. This helps to balance the influence of these groups during model training.
* Iterative Process: The reweighing process is often iterative. The sample weights are recalculated, disparities are reevaluated, and the weights are adjusted again. This cycle continues until the model's predictions become fairer and more unbiased.

By reweighing the dataset, it ensures that the model's training process considers the needs and characteristics of both privileged and unprivileged groups more equally. It's a valuable tool in promoting fairness and reducing bias in machine learning models, particularly when dealing with imbalanced datasets or scenarios where fairness concerns are paramount, for example while developing a loan approval system or student admission systems.

**Accuracy and fairness with reweighed samples**

|  |  |
| --- | --- |
| **Accurate model** | **Fair model** |
| precision recall f1-score support  0.0 0.89 0.71 0.79 11132  1.0 0.44 0.72 0.55 3521  accuracy 0.71 14653  macro avg 0.67 0.72 0.67 14653  weighted avg 0.78 0.71 0.73 14653 | precision recall f1-score support  0.0 0.89 0.72 0.79 11132  1.0 0.44 0.71 0.55 3521  accuracy 0.72 14653  macro avg 0.67 0.71 0.67 14653  weighted avg 0.78 0.72 0.73 14653 |
|  |  |
| Fairness metrics of accurate model on reweighed samples:-  Equal opportunity difference: 0.0329  Disparate Impact: 0.9351 | Fairness metrics of fair model on reweighed samples:-  Equal opportunity difference: 0.0461  Disparate Impact: 0.961 |

**Discussion:**

Following the application of sample reweighing, the accuracy-focused and fairness-focused models yield more balanced results, with similar outputs. These models not only excel in classification performance but also strike a balance between privileged and unprivileged classes.

****

**Conclusion**

In this task, we analysed the effects of bias and the requirements of fairness in machine learning models trough a simple classification problem. We began by discussing about the bias-variance trade-off in machine learning that almost always takes the front seat obscuring the effects of unfairness and bias towards specific classes. To demonstrate this, we obtained two similar models fine-tuned to produce i) the most accurate results and ii) the fairest result. After observing their differences and understanding the impact, a method called reweighing was discussed which aims to mitigate the effects of bias while training a machine learning model. At the end, we visualised the effects of reweighing and confirmed the minimization of bias in the training process.

The complete code in python can be found here.