Fairness in Machine Learning for 2nd Order Solutions

Contributors: Yiyu Lin, Yinting Zhong, Genesis Qu, Pragya Raghuvanshi

Table of Contents

1.Introduction

2nd Order Solutions, our client in this capstone project, is a financial consulting firm that works on providing analytical solutions to their financial partners – mainly banks domestically and internationally. The company uses most of its time to build statistical models to help clients craft valuation and credit lending policies, fraud detection, and due diligence.

As an institution that provides financial services to the public, 2OS, and its clients operate under a strict network of regulatory frameworks and oversight bodies. A key aspect of such regulation is the requirement – under the Equal Credit Opportunity Act – that the models that decide what consumers receive financial products may not discriminate on protected characteristics of the clients such as gender, race, disability status, and ethnicity. Such requirements are fundamental to the service that 2OS provides because current regulations render any model that introduces biases unusable. How machine learning algorithms perpetuate bias is keenly researched in academia and the tech media world. A frequent way bias shows up is through biased training data. For example, if most women were denied opportunities in a company while few men were, then an algorithm trained on this data to screen resumes would doubtlessly recommend men disproportionately. In sensitive fields such as healthcare and finance, such bias needs to be carefully guarded against. Our goal is to provide 2OS with tools to assess fairness and mitigate the biases before model handover, enhancing their business processes and value proposition.

2.Project Objective & Goals

The purpose of our capstone team and this project is to research the evaluation of fairness in financial machine-learning products and evaluate current packages that quantify algorithmic bias in models. We define fairness as an equal opportunity to obtain a positive outcome for both the underprivileged and the privileged groups. The goal is to make recommendations to 2OS on which statistical package(s) best fulfills its need to remain compliant with financial regulations. The purpose of our capstone team and this project is to research the evaluation of fairness in financial machine-learning products and evaluate current packages that quantify algorithmic bias in models.

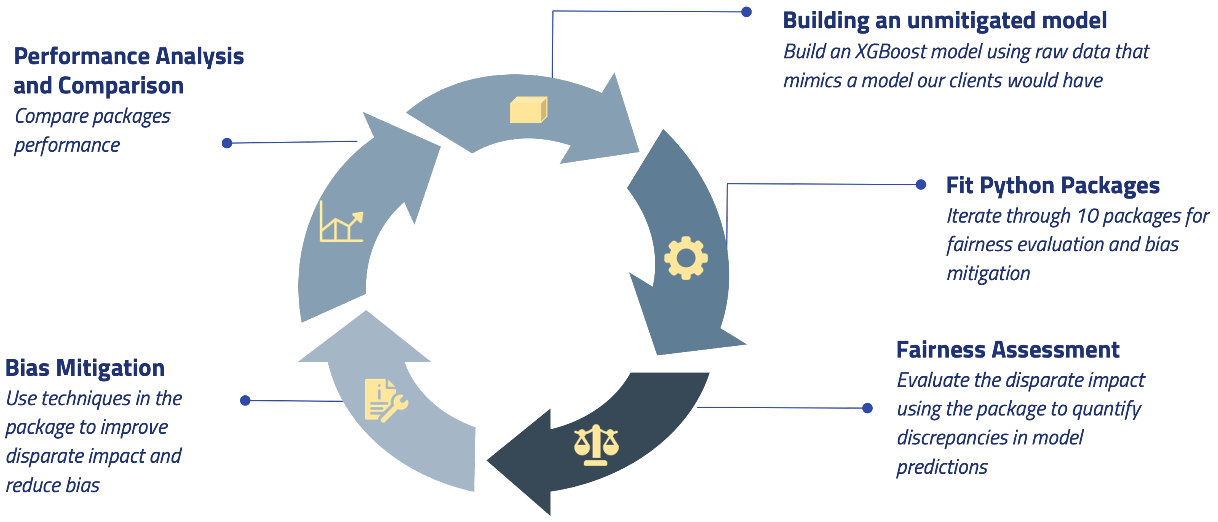
3.Datasets

Evaluation and mitigation of biases are applied to two datasets:

1. Taiwanese Credit Card Dataset: This dataset comprises customers' default payments in Taiwan in 2005.
   Some features of the data:
   Target Variable: *Default/Non-Default*
   Features: *23*
   Instances: *30000*

2. Adult(Census) Data set: This dataset comprises an individual’s annual income results from various factors. Also known as the "Census Income" dataset.
   Some features of the data:
   Target Variable: *Income >50K, <=50K*
   Features: *14*
   Instances: *48842*

4.Overall Methodology



Building an unmitigated model
Build an XGBoost model using raw data that mimics a model our clients would have
Fit Python Packages
Iterate through 10 packages for fairness evaluation and bias mitigation
Fairness Assessment
Evaluate the disparate impact using the package to quantify discrepancies in model predictions
Bias Mitigation
Use techniques in the package to improve disparate impact and reduce bias
Performance Analysis and Comparison
Compare packages performance

4.1 Building an unmitigated model

We opted for a baseline model using XGBoost, selected to align with client stipulations and preferences. XGBoost is preferred by clients due to its scalability and its enhanced performance capabilities.

Additionally, feature importance plots from XGBoost models help make the model explainable.

Models trained in folder "02_model/model" includes 3 models. Model1 is trained by all variables of the dataset, Model2 is trained by all variables excluding four Protected Variables (age, education, sex, marriage), and Model3 is trained excluding three Protected Variables (age, education, sex).

4.2 Fit Python Packages

Using various packages, we tried to mitigate the bias in the baseline model. We rigorously assess these fairness tools to ensure their reliability and effectiveness, allowing us to offer informed recommendations on their practical use. We try various methods and metrics for each package to evaluate their effectiveness and compatibility with the client’s needs.

The 10 Python packages we explored are:

Package Name, GitHub Repository

1. AI Fairness 360, https://github.com/Trusted-AI/AIF360
2. DALEX, https://github.com/ModelOriented/DALEX
3. Deon, https://github.com/drivendataorg/deon
4. Fairlearn, https://github.com/fairlearn/fairlearn
5. fairness-in-ml, https://github.com/equialgo/fairness-in-ml
6. FairSight, https://github.com/ayong8/FairSight
7. Responsible AI toolbox, https://github.com/microsoft/responsible-ai-toolbox
8. Smclarify, https://github.com/aws/amazon-sagemaker-clarify
9. Themis-ML, https://github.com/cosmicBboy/themis-ml
10. PiML, https://github.com/SelfExplainML/PiML-Toolbox

Out of the ten packages, AI Fairness 360 outperforms the rest in terms of bias mitigation, usability, and generalizability.

AIFairness 360

AI Fairness 360(AIF 360) is an open-source library designed to help researchers detect, evaluate, and mitigate biases in machine learning algorithms. Depending on how they reduce bias, it offers a comprehensive suite of algorithms categorized as preprocessing, in-processing, and post-processing. Pre-processing techniques are applied before the training process on the dataset and produce fair representations which can be subsequently used by the machine learning model. In-processing is a bias mitigation algorithm applied to a model during its training. Post-processors are algorithms that edit the model outputs/predictions to maximize an objective function and prioritize fairness. We have employed the following algorithms to evaluate and mitigate bias across the groups.

1. Disparate Impact Remover: Disparate Impact Remover is a preprocessing tool that modifies feature values to increase fairness across groups. This technique works by adjusting data points to reduce bias while maintaining the relative data ordering within each group. Essentially, it seeks to obscure any strong indicators that might reveal group affiliation without disrupting the inherent structure of the data.
2. Learning Fair Representations: LFR addresses the group and individual fairness by finding latent representations that encode data and obfuscate information about protected attributes. It works by identifying protected attributes and encoding them thereby creating a new set of features, removing any association between the attributes and the output variable.
3. Reweighing: Reweighing is a technique that assigns differentiated weights to the training instances based on their group categories to promote fairness prior to the classification process. This approach doesn't alter any actual feature or label values but adjusts the significance of each sample during model training. Higher

weights are assigned to instances that are underrepresented and lower weights are assigned to instances that are overrepresented.

4. Calibrated equalized odds postprocessing: This method specifically focuses on the principle of equalized odds. This technique operates on the output scores of an already trained classifier and adjusts the final decision labels to achieve fairness. The idea is to optimize the model's output probabilities to ensure that the odds of a positive classification are balanced across different groups for individuals with the same true outcome.
5. Reject option classification: This method works by identifying a 'confidence band' around the decision boundary of the model. Within this band, the algorithm preferentially alters outcomes to benefit the unprivileged groups and assigns unfavorable outcomes to the privileged groups. The rationale is to correct potential biases that might have influenced the model's training, particularly in those cases where the model is most uncertain about its decisions.

Below, we include a brief description of the other 9 packages that we tested for bias mitigation and the reasons behind their exclusion from the experimental analysis.

DALEX

DALEX is designed to work with any predictive model, regardless of its internal structure. It offers various tools for exploring different aspects of a model, including model performance, conditional effects of variables, and variable importance. This enables a deeper understanding of how models make their predictions. However, the key limitation of this package is that it serves only as a starting point for understanding a model but does not include any tools to pinpoint model fairness deficiencies and mitigate the unfairness. Because of this limitation, we consider DALEX unfit for the purpose of bias mitigation proposed by our clients and removed it from consideration.

Deon

Deon is a command-line tool that appends an ethics checklist to the project for data scientists to assess biases manually. Even though the checklist has a comprehensive list of criteria to evaluate bias, the package so solely based on data scientists' subjective opinions on evaluating biases. There is no algorithm to calculate bias and perform mitigation. Thus, the Deon package was removed from our consideration.

Fairlearn

Fairlearn is an open-source library designed to help researchers assess and mitigate unfairness in machine learning models. It provides tools for evaluating and visualizing fairness metrics and helps users understand and address potential biases in their models. In addition to assessment, Fairlearn provides algorithms to mitigate unfairness in machine learning models. These algorithms aim to adjust model predictions to reduce disparities while maintaining overall predictive performance. There are two different mitigation strategies in the Fairlearn package, Postprocessing and Reductions. We have decided to remove Fairlearn for consideration because one of the only two mitigation methods, “Reductions”, failed to reduce fairness for our dataset. We speculated that this may be due to the reason that the “Reductions” approach aims to reduce biases from multiple features holistically. Thus, in our analysis when we only focus on one feature at a time, the mitigation was not prioritizing the feature we selected. Since the inner workings of which feature the “Reductions” approach prioritizes was missing from the documentation, we also find this package relatively difficult to work with.

Themis-ML

Themis-ML is designed to promote fairness-aware machine learning. It builds upon pandas and scikit-learn, implementing various algorithms that address discrimination and bias in machine learning models. However, many mitigation methods are missing documentation and explanations, making the package difficult to use. Thus, we have decided to move on with other packages.

PiML

PiML is a toolbox for interpretable machine learning model development and validation. It helps users to use simple codes to build ML models and interpret their performance results. Within the results section, fairness was evaluated, under the integration of another fairness package called solas-ai. The PiML package itself was not a package aimed at bias evaluation and mitigation. Thus, we don’t think PiML is useful for our clients’ purposes. As for the fairness package solas-ai, it is useful for disparity testing and bias evaluation, but it has no fairness mitigation methods.

Responsible AI toolbox

The Responsible AI Toolbox encompasses a comprehensive array of tools, including libraries and user interfaces, aimed at enhancing the scrutiny and assessment of data and models. This toolkit is instrumental in enriching the understanding of AI systems, providing those involved in AI development and oversight with the necessary resources to foster ethical and responsible AI practices. It supports informed decision-making based on data, ensuring AI technologies are developed and managed with greater responsibility. However, it falls short of offering a tool for mitigation.

Smclarify

The smclarify package, provided by Amazon Web Services (AWS), is a powerful tool within the AWS SageMaker suite designed for bias detection and explainability in machine learning models. It aids in identifying and reporting various types of biases in datasets and models—both pre and post-training—to promote fairness. Additionally, it offers explainability features to elucidate the impact of input features on model predictions, which is vital for model transparency, debugging, and improvement. However, It focuses on identifying and explaining bias rather than providing direct solutions or strategies to mitigate these biases within machine learning models.

fairness-in-ml

Fairness in ML mimics the Generative Adversarial Network logic of a zero-sum game, where the generative model is replaced by the predictive classifier model and the task of the adversarial model is to predict the sensitive attribute value from the output of the classifier. The adversarial training of the classifier is done through the extension of the original network architecture with an adversarial component. This technique ranks low in terms of Generalizability, Usability, and Interpretation as it involves the architecture of neural networks which is harder to implement and interpret when compared to the XGBoost model. Therefore, we also decided against moving this package forward in our recommendation.

4.3 Performance Analysis and Comparison

4.3.1 Fairness Evaluation Metrics

There has been a wealth of studies on fairness in machine learning and algorithmic biases in recent years. Specifically, scholars have proposed several definitions of fairness and different metrics that quantify bias – such as statistical parity, equalized odds, and disparate impact. Our analysis will cover a suite of fairness metrics but will focus on Disparate Impact. We zero in on the disparate impact metric due to its salience in the consumer lending space. Disparate Impact measures the ratio between the proportion of each group receiving the positive outcome. This is a commonly cited metric measuring fairness in financial decisions. In fact, the Consumer Compliance handbook published by the Board of Governors at the Federal Reserve highlights disparate impact as a textbook example of a violation of the ECOA.

The Disparate Impact (DI) is calculated using the formula:

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

4.3.2 Balanced Accuracy

Balanced Accuracy calculates the mean between the True Positive Rate and the True Negative Rate in the model predictions. We want our clients to be able to grant financial products to people who would not default and deny them to people who would. Balanced Accuracy is also ideal for Unbalanced Data, in cases where there are very few defaults in the data, this metric can capture model performance accurately.

5.Results

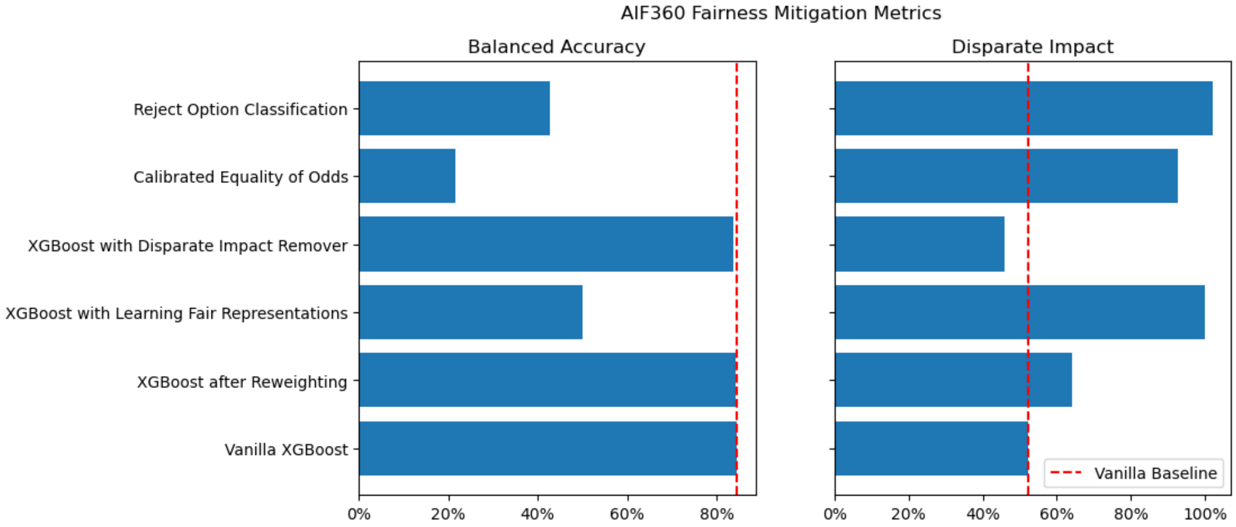
In actual assessment of fairness in financial machine learning models, we'll have to control for the effects of variables that are not protected under the ECOA – such as income and educational background. The effects of these unprotected variables can make our estimate of bias in the model completely inaccurate. For example, if a higher proportion of female applicants in the data hold higher income and have more years of education, it would make sense if they receive a loan at a higher rate than their male counterparts. A simplistic assessment of disparate impact would gloss over these nuances.

To address the confounding effect of unprotected variables, we applied the Matching technique often used in Causal Inference, where we matched pairs of observations in the data between the unprivileged group and the privileged group such that a subset of columns share similar support. In other words, and picking up from our example earlier, we sample a smaller subset of the whole dataset where the female group and the male group both have similar distribution in income and years of education, allowing us an apples-to-apples comparison.

We apply this principle to our Adult dataset. Since the target variable in the data is income, the possible confounders are years of education and hours worked each week. Using race as the protected attribute here (note that this process can and should be repeated for all protected attributes of interest), we applied the DAME algorithm in Python and picked the subset of data after an appropriate number of iterations. This matching process filtered out about 27% of White applicants and 14% of Non-White applicants. Consequently, we identified that the bias in the data due to race was slightly less than we thought: the new disparate impact was 0.62 instead of the original 0.60.

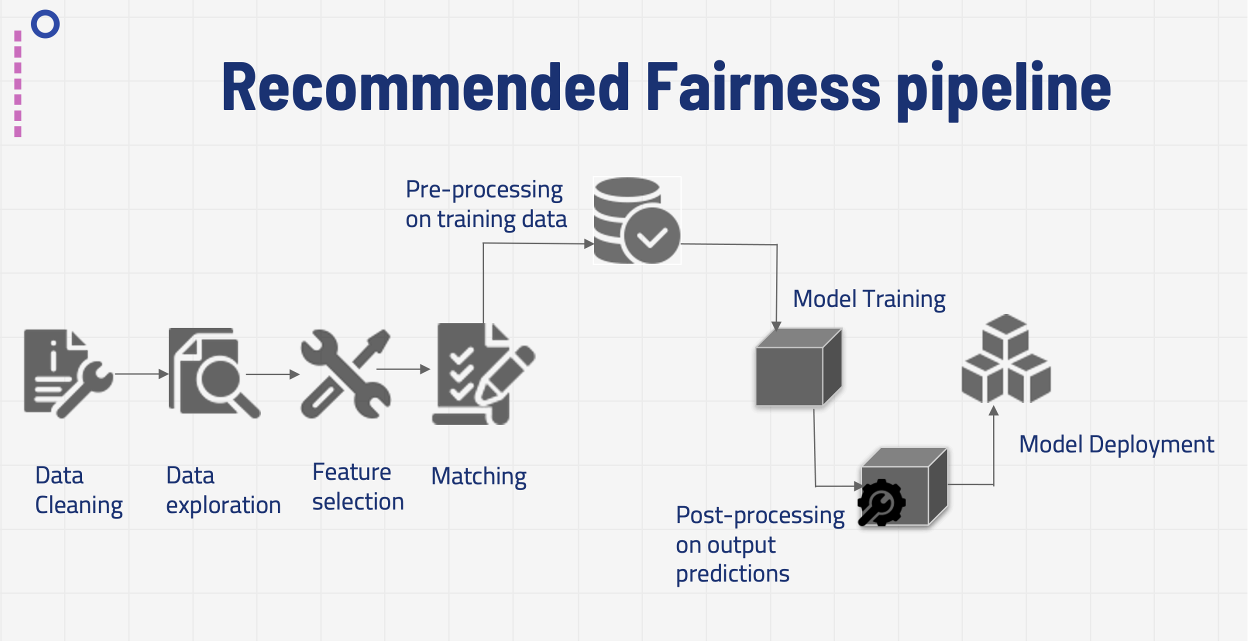
After this initial matching process, we can move on to applying the many bias mitigation techniques in the model and testing their efficacies.

The pre and post-processing techniques of AIF 360 on mitigation of biases across race as the sensitive attribute on the adult dataset yield the following results. We can significant reduction in disparate impacts in many of these techniques while maintaining manageable degradation in performance.



6.Conclusion

In conclusion of the analysis performed, we recommend to follow a fairness pipeline as shown:


Recommended Fairness pipeline
Pre-processing on training data
Model Training
Model Deployment
Data Cleaning
Data exploration
Feature selection
Matching
Post-processing on output predictions

7.Usage

```
# Create a virtual env to install the required packages
python -m venv /path/to/new/virtual/environment
source /path/to/new/virtual/environment/bin/activate

# Clone the repository and install the required packages
git clone https://github.com/YYLinn/2nd-order-solution-ML-Fairness-.git
cd 2nd-order-solution-ML-Fairness
pip install -r requirements.txt

# Run the analysis script to perform the required experimentation
python 03_analysis/Experimentation/main.py --technique <<Fairness_technique>> --
sensitive_attr <<sensitive attribute>>
```

The code will give json files with various performance metics and disparate impact of the new model, according to the technique specified.

Documentation related to the project can be found at: Documentation