Consumer Data Research Centre

An ESRC Data Investment

experian

Assessing fairness/bias in binary classification machine learning models on consumers

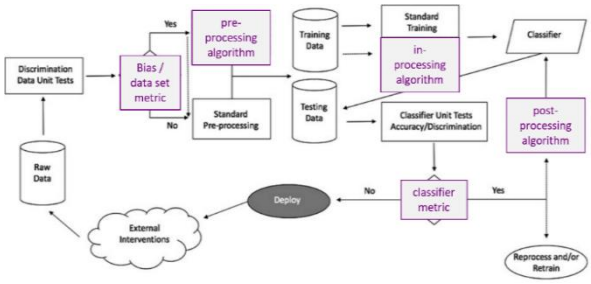
Benedict Faria[1], Roman Kontchakov[1] and Ernest Chow[2]
[1]Birkbeck University of London, [2]Experian Ltd

Project Background

Measuring and mitigating bias in classification machine learning algorithms is a new field in data science. As more cases of bias in machine learning algorithms emerge, it is imperative that it is addressed and mitigated against. Left untreated, bias in classification datasets can expose businesses to legal risk, and limit opportunity. Training datasets used to train classification models are the main source of bias. They comprise historic data and feature selections that reflects social and economic disparities. Bias mitigation techniques aim to minimise the bias without compromising model performance. This assessment reviewed the various approaches in mitigating bias in datasets. It assessed the metrics used to test a dataset for bias, and the various open-source algorithms available to mitigate this bias.

Data and Methods

The This assessment used selected metrics and mitigation algorithms from the AI Fairness 360 (AIF360) toolkit on three public datasets. Bias was measured in terms of achieving algorithmic fairness and used statistical and similarity-based metrics. Statistical metrics use the 'Sensitive' attribute and classification labels. These measure group fairness and use label counts to measure fairness between privileged and unprivileged groups. Similarity based measures use non-Sensitive attributes to measure for individual fairness in classification outcomes. Mitigation algorithms either transform and debias a training dataset, tweak a classifier to perform in a non-discriminatory way, or re-label the classification outcomes to achieve fairness. Three bias mitigation approaches were used. Pre-processing algorithms transform the training data to obfuscate any discriminatory patterns. In-processing algorithms use adversarial methods to tweak the classifier into becoming bias-aware. Post-processing algorithms modify the post classification labels to achieve fairness. Baseline measures of bias were compared with post-mitigation measures to assess the effectiveness of each mitigation algorithm in terms of maximising bias mitigation and classifier performance.



Bias mitigation algorithms in a typical machine learning pipeline.

Baseline measures of bias were compared with post-mitigation measures to assess the effectiveness of each mitigation algorithm in terms of maximising bias mitigation and classifier performance.

Key Findings

Baseline measurements for fairness showed low levels of bias in two of the three public datasets. In this context, the best performing mitigation algorithms were Learning Fair Representations LFR) in pre-processing and Adversarial Debiasing in in-processing. LFR creates a set of intermediate fair representations of the training data that obfuscates the Sensitive variable and can be generalised for any classification problem. Adversarial Debiasing uses Generative Adversarial Networks (GANs) to achieve fairness. GANs require substantial volumes of training data.

Bias measurement and mitigation is a non-trivial exercise. It requires considerable domain expertise for data cleansing, feature engineering, fairness metrics selection and result interpretation. Fairness metric and mitigation algorithm selection is based on what task the classification seeks to do. Numerous formulae for defining algorithmic fairness continue to emerge, reflecting the complex and social nature of determining fairness. Correspondingly, the number of bias mitigation algorithms continues to evolve. There are also proposals for public datasets to demonstrate their provenance and publish metrics for any inherent bias. To minimise commercial risk, businesses need to establish an acceptable trade-off between optimising model accuracy and minimising bias. Setting a high classification threshold means that an unbiased classifier produces fewer favourable labels, e.g. offers lower true positive rates (TPR) and higher true negative rates (TNR). One way to measure the business impacts of the trade-off is to assign unit costs for TPR and TNR, so that businesses can vary thresholds for acceptable risk. However, determining such unit costs may not be trivial for most businesses.

Value of the Research

This assessment provides a starting point for businesses to understand, measure and mitigate bias in training datasets for binary classifiers. It highlights the sources of potential bias, discusses the various fairness metrics to quantify the bias, and implements mitigation algorithms at three points in a machine learning pipeline. Following this assessment's approach, businesses can test their training datasets for bias, and assess the effectiveness of the mitigation algorithms in terms of achieving fairness and maintaining classification model performance.