COMPASS FAIRNESS PRIVACY TRADEOFF ANALYSIS

Ayush Oturkar, Keshvi Gupta, Atharva Sherekar
Department of Statistics - Rutgers University

ABSTRACT

The concerns of fairness, and privacy, in machine learning and AI systems have received a lot of attention worldwide recently. However, achieving privacy and fairness along with good performance of these algorithms is not theoretically possible. In this work, we observe this tradeoff using differential privacy and reject option classification to achieve fairness and model accuracy on interpretable models. We observe that, depending on the situation, a certain level of performance of any one metric of interest, the other two metrics must be traded off. We also observe the positive impact of applying bias mitigation techniques on the model explanations - a welcome cost at the expense of reduced model performance.

1 INTRODUCTION

When using Machine Learning techniques, one should always be careful about the implications of the predictions. This is especially true when the outcomes of such models directly affect humans. Developers of these models must ensure that such models are fair and do not harm the privacy of the users. However, achieving ethical standards is often at odds with developing well-performing models in terms of traditional performance measurements. Therefore, in this project, the tradeoff between model performance, user privacy, and model fairness is observed and analyzed. Various models, methodologies, and techniques to come to the most optimal solution that finds the most optimal balance between all three, without compromising too heavily on one particular aspect is explored.

2 METHODOLOGY

2.1 FINALIZING DATASET

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism dataset is a public dataset that contains information about criminal defendants in Broward County, Florida. The dataset includes information such as the defendant's age, race, gender, criminal history, and risk of recidivism. The dataset's relevance to this project stems from its inherent ethical considerations

2.1.1 TRADEOFF BETWEEN FAIRNESS AND ACCURACY

The COMPAS algorithm, employed to assess recidivism risk, has been criticized for its potential bias against minority groups. This raises concerns about the fairness of the model, as it may unfairly disadvantage certain individuals based on their race or ethnicity. The tradeoff between fairness and accuracy is particularly relevant in the context of recidivism risk assessment, as the consequences of an inaccurate prediction can be severe. For instance, a false positive prediction could lead to an individual being denied bail or parole, while a false negative prediction could result in a dangerous individual being released into the community. The COMPAS dataset provides a valuable opportunity to examine this tradeoff empirically and develop strategies for mitigating bias in risk assessment algorithms.

2.1.2 Implications of Data Privacy

The use of sensitive personal data, such as criminal history, raises concerns about data privacy. While utilizing this data may enhance the model's accuracy, it is essential to balance the benefits of predictive power with the protection of individual privacy rights.

2.2 Evaluation Metrics

Evaluation metrics were decided based on the context of the problem to ensure that they uphold the values of the judiciary system. The values we focused on were 'the guilty should not be punished' and that 'individual's identity should get disclosed by model parameters'.

To ensure this judgment criterion Precision, FPR (False Positivity Rate) and epsilon (Differential Privacy Budget) were decided as the evaluation metrics for Model Performance, Fairness Evaluation and Privacy Quantification respectively.

Maximizing the Precision over Accuracy or Recall ensures that we are minimizing the False Positives, i.e. not punishing the innocent.

False Positivity Rate indicates the number of False Positives given the Number of Actual Negatives. In this context, FPR indicates the chances that the model will predict an individual reoffending even when in reality, he is innocent. A biased model is expected to have significantly higher FPR for discriminated groups than non-discriminated groups.

ϵ typically involves adding noise to the model's coefficients or model's outputs. This noise helps to ensure that small changes in the input data do not significantly affect the model's predictions, thereby protecting the privacy of individual data points.

Therefore, it acts as a privacy constraint, limiting the amount of information that can be inferred about individual data points.

2.3 Data PreProcessing

Data preprocessing is a crucial step in the data analysis and machine learning pipeline, aimed at enhancing the quality of raw data to make it suitable for further analysis or model training. The need for data preprocessing arises due to several challenges and imperfections associated with real-world datasets. Here are some key reasons why data preprocessing is essential:

2.3.1 Analysing Target Distribution

From the data it was observed close to 45% of the offenders reoffend in two years.

2.3.2 Data Cleaning

The subsequent phase involved meticulous data cleaning, where missing values were systematically assessed across the entire data frame, and necessary treatment procedures were applied. Features with an absence exceeding 50% of the total data were deliberately excluded, as their inclusion might not significantly contribute to the overall model performance. Additionally, certain numeric features exhibiting skewness underwent imputation using the median to address potential bias and ensure the integrity of the data.

2.3.3 Exploratory Data Analysis

Some key observations from the dataset include:

1. **Crime Distribution:** Approximately 65% of offenders were involved in felonies, while misdemeanor crimes accounted for about 35% of the cases.
2. **Age and Recidivism:** A significant decline in the recidivism rate is noticeable among individuals below the age of 30, suggesting a higher susceptibility to reoffense in the younger age group.

3. **Impact of Variables:** Variables such as felony count and time spent in jail or custody do not exhibit discernible patterns related to criminal reoffense.
4. **Handling Sparse Data:** Due to the dense concentration of data near zero, flagging any non-zero value as 1 may reveal meaningful patterns.
5. **Prior Criminal Offenses:** The total count of prior criminal offenses, combining juvenile and adult offenses, displays a discernible pattern in relation to criminal reoffense.
6. **Gender Disparity:** Males show a higher tendency to reoffend compared to females, as indicated by the data.
7. **Racial Disparities:** The African American community demonstrates a higher recurrence of re-offenses compared to other racial groups. However, it is crucial to address this factor to ensure model fairness and mitigate bias, emphasizing the need for careful consideration in subsequent modeling steps.

2.3.4 Feature Engineering

Feature engineering plays a pivotal role in the Data Science lifecycle as it involves crafting features that can enhance model performance by uncovering meaningful patterns in the data. Some of features that were engineered were:

- Jail in time was calculated using jail_in_date and jail_out_date
- Time in custody was calculated using the in_custody and out_custody date

2.3.5 Modelling Data Preparation

Under Modelling data preparation, a pivotal phase is undertaken to facilitate linear machine learning, comprising two fundamental stages:

1. **Categorical Encoding** : Within this phase, features characterized by lower cardinality, such as 'race' and 'c_charge_degree' (indicating the degree of the criminal charge), undergo one-hot encoding—a technique involving the creation of dummy variables. Simultaneously, the 'c_charge_desc' feature, providing details about the criminal charge, undergoes Target encoding. This method employs the mean of the target variable, accompanied by slight random perturbations, effectively averting data leakage.
2. **Addressing Multicollinearity** : The management of multicollinearity emerges as a critical consideration for securing precise coefficient elasticities. Detection of significant multicollinearity is accomplished through the calculation of the Variance Inflation Factor (VIF). In response to this identification, a strategic decision is made to alleviate the issue by selectively dropping highly collinear features. This action includes features such as 'age_cat,' 'two_year_recid,' 'Misdemeanor_c_charge_degree,' 'Medium_v_score_text,' 'Caucasian_race,' 'Medium_score_text,' and 'Low_v_score_text.' The purpose is to curtail the VIF below a predefined threshold, set at 10 for optimal model performance. This meticulous process ensures the integrity of the model coefficients and enhances the overall robustness of the linear machine learning framework.

2.4 Explored Techniques

2.4.1 Baseline Modeling

In this research approach, general modeling techniques commonly utilized by Data Scientists are employed. However, given the presence of bias in the data, particularly towards variables such as race, sex, or other protected classes, there is a risk of introducing bias into the models. The primary objective is to establish a baseline model using preprocessed and cleaned data to obtain foundational results for this study. For testing model performance the data was split in train (80%) and test (20%) respectively using stratified sampling towards using two years recidivism (target variable)

For this purpose, two essential models were implemented: Logistic Regression and a Multi-Layer Perceptron (MLP). Logistic Regression was fitted with default hyperparameters, serving as a fundamental reference point for our analysis. Recognizing the complexity of underlying patterns and aiming to capture intricate relationships within the data, a Multi-Layer Perceptron was employed. The MLP is configured as a 3-layered neural network with 8x8x4 hidden neurons, providing a more sophisticated representation of the data.

In the comparative analysis between MLP (Multilayer Perceptron) and Logistic Regression models, it was determined that Logistic Regression exhibited comparable recall accuracy for the positive class (1) in comparison to MLP. To streamline the study and mitigate potential overfitting, as well as to maintain simplicity, Logistic Regression was chosen as the baseline model.

Upon subjecting the Logistic Regression model to testing using a dedicated test set, an examination of biases towards the protected class was conducted. Visual inspection revealed an inherent bias towards African Americans, with a tendency to predict more false positives in this demographic. Additionally, the model exhibited a leaning towards males in terms of false positive predictions.

The evaluation of **Statistical Parity** indicated positive biases towards both gender and the African American race. The Statistical Parity Difference metric, which gauges the difference in the ratio of favorable outcomes between monitored groups and reference groups, yielded values deviating from the ideal of zero. The assessment of statistical parity for the variable representing the sex of the accused yielded a value of 0.25 and yielded close to 0.2 for African American race. Ideally, this metric should approach 0 to signify minimal bias with regard to gender. The obtained value of 0.25 & 0.2 respectively indicates a notable deviation from the desired neutrality, underscoring the presence of bias in the model's predictions concerning the accused person's gender and race. Addressing and mitigating this bias is imperative for fostering fairness and equity in the model's outcomes.

Analyzing the **False Positive Rate (FPR)**, a measure of the proportion of positive cases incorrectly identified as positive in a test, revealed significantly higher FPRs for both gender and African Americans when compared to privileged groups. This disparity underscored a potential bias in the model's predictions. African Americans are 16% more likely to be wrongly accused of reoffense through model predictions on test data.

In the pursuit of addressing these observed biases, an adjustment was made with the anticipation of a potential decrease in the model's recall. This modification aimed to alleviate the occurrence of False Positives, particularly concerning gender and African American groups. These efforts align with the overarching objective of enhancing fairness and equity in the outcomes generated by the model.

2.4.2 Differential Privacy

Differential privacy offers a rigorous mathematical framework for quantifying and bounding the privacy risks associated with releasing statistical information about a dataset. By applying differential privacy techniques, it is possible to protect the privacy of individual defendants while still allowing for the release of aggregate statistics about recidivism rates. In our exploration of differential privacy techniques on the COMPAS dataset, we implemented the following privacy-preserving methods:

- **Laplace Mechanism for Differential Privacy:** To enhance the privacy of our logistic regression model, we employed the Laplace Mechanism. This technique introduces carefully calibrated Laplace noise to the model parameters, providing a level of privacy determined by the privacy parameter epsilon (ϵ). Setting ϵ to 1.00, we observed a trade-off in model accuracy, emphasizing the impact of privacy preservation on predictive performance.

- **DP-SGD (Differentially Private Stochastic Gradient Descent):** For our multi-layer perceptron (MLP), we adopted DP-SGD, a widely-used algorithm for training differentially private models. DP-SGD modifies the standard stochastic gradient descent by incorporating privacy-preserving mechanisms. This includes gradient clipping and the addition of noise to gradients during the training process. By adjusting privacy parameters, such as setting epsilon (ϵ) initially to 8.00 and later to 4.00, we explored the influence of varying privacy levels on the model's accuracy.

2.4.3 Fairness

The exploratory data analysis (EDA) uncovered bias against African Americans and males in recidivism risk prediction. To ensure fairness, we identified African American race and gender as sensitive classes, removing variables contributing to bias in the COMPAS model. We consciously avoided suppression of sensitive attributes, aiming to develop a discrimination-aware algorithm that effectively addresses bias.

Exploring Preprocessing Techniques: In our quest to mitigate bias, we explored preprocessing techniques, including reweighing and Learning Fair Representations (LFR). Reweighing adjusted weights for underrepresented groups, but concerns arose about data integrity and interpretability. LFR transformed data for reduced bias but presented challenges in interpretability. Despite the limitations, alternative fairness techniques were considered more effective for recidivism risk prediction.

In-Processing with Prejudice Remover: Prejudice Remover (PR), an in-processing technique, targeted bias by incorporating a regularization term into the learning objective. PR effectively reduced false positive rate (FPR) disparity, achieving a precision of 72%. However, implementation limitations in the Holisticai package necessitated the construction of a new base model, impacting the overall workflow.

Post-Processing to Address Bias: Post-processing techniques were explored to address bias after model predictions. Equality of Opportunity had minimal impact on FPR rates, while Calibrated Equalized Odds (CEO) compromised accuracy and recall. Reject Option Classification (ROC) emerged as a practical choice, allowing the classifier to abstain from predictions in uncertain situations. ROC demonstrated exceptional performance with a precision of 70% and an FRPD of 0.03, aligning well with real-world scenarios and legal considerations.

Selecting Reject Option Classification (ROC): Given ROC's ability to abstain from predictions in uncertain scenarios, its exceptional performance, and its practicality for real-world situations, we have chosen it as our post-processing technique. This approach aligns with our focus on individual fairness, model explainability, and legal justifiability in the context of recidivism risk prediction. ROC provides a robust solution without necessitating data manipulation or process rebuilding.

Considerations for Real-World Applicability: In our pursuit of fairness in recidivism risk prediction, we emphasize the importance of selecting techniques that balance performance, context, implementation feasibility, and real-world applicability. The chosen approach, Reject Option Classification (ROC), stands out as a practical and effective solution, particularly in high-stakes situations where incorrect decisions have significant consequences.

2.5 Finalized Approach

The finalized approach for addressing bias in the recidivism risk prediction model consisted of three main steps:

1. **Considered Race as a Sensitive Attribute:** Race was identified as a sensitive attribute due to the observed bias in the baseline logistic regression model. This model exhibited a precision of 69%, a recall of 60%, and an FPRD of 0.14. Additionally, the model's beta estimates indicated a bias towards race, with estimates of 0.01 for Sex and 0.02 for African American race.
2. **Implemented Differential Privacy:** Differential privacy was employed to enhance privacy while preserving model performance. Experimenting with various epsilon values led to the selection of epsilon = 10, which maintained model performance while improving fairness by reducing FPRD from 0.14 to 0.11. This reduction in FPRD is attributed to the obfuscation of the relationship between sensitive attributes and the outcome by differential privacy.
3. **Implemented Fairness using Reject Option Classification (ROC):** ROC was implemented to further mitigate bias. ROC has the advantage of being fairness definition agnostic and allows for optimizing for various fairness metrics. In this case, optimizing for 'Statistical Parity Difference' was chosen. Statistical Parity Difference measures the difference in the positive prediction rate/Precision between the privileged and unprivileged groups. By setting low and high thresholds smartly, ROC could be optimized for FRPD by

maximizing precision. The chosen thresholds were 0.3 for the low threshold and 0.7 for the high threshold, considering the ambiguity in the model's predictions.

The implementation of ROC resulted in an increase in precision to 75% and a reduction in FPRD to 0.06. Recall decreased to 30%, but this is acceptable given the high-stakes nature of the recidivism risk prediction task. Additionally, the model's beta estimates moved closer to zero, indicating a reduction in bias after applying the fairness technique.

3 Results

The tradeoff of model performance, fairness and privacy. As the bias mitigation measures (FPRD and Epsilon) get better, the model performance (Accuracy) takes the hit. Note that in this tradeoff, Model Precision is getting better - the metric we wanted to maximize to ensure fairer judicial predictions.

Model	Accuracy	Precision	Bias: FPRD	Privacy: Epsilon
Baseline model A	70%	69%	0.14	-
Baseline + Privacy implementation	69%	68%	0.12	10
Baseline + Privacy + Fairness	63%	75%	0.06	10

References

1. https://www.iguazio.com/glossary/false-positive-rate/

2. https://www.statsmodels.org/stable/index.html