Final Audit Report

Group Member: Chujun Chen, Yunfei (Cynthia) Xing
Repository link: https://github.com/Christina-Chen01/CSCI-1951Z-FairnessAduitProject

Introduction

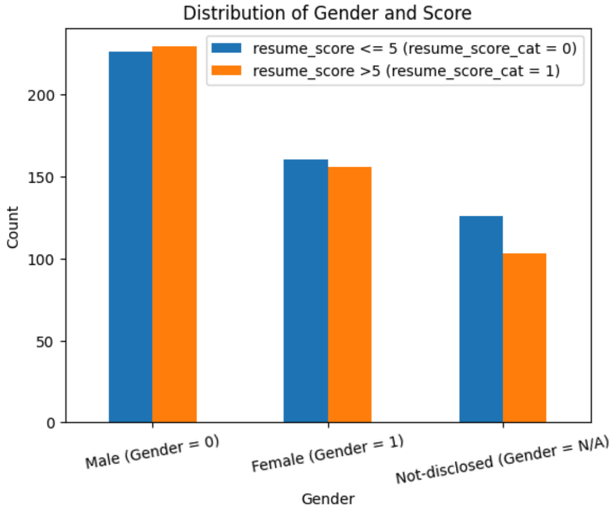
This report outlines the findings of an audit conducted on two AI-driven models developed by Providence Analytica and employed by Bold Bank to facilitate HR decision-making. Although such automated models are designed to automate and optimize the candidate selection process, concerns arise about how these models process and interpret sensitive attributes like gender, ethnicity, work authorization status, etc.

To mitigate such concerns, this audit report aims to assess the performance of the models by pinpointing any potential biases in decision-making. Focusing on these aspects allows us to deliver actionable insights and recommendations that will assist Providence Analytica and Bold Bank in improving the transparency, accountability, and fairness of their automated decision systems.

Methodology

Data Source

We created a synthetic dataset containing 1,000 entries designed to represent applicants with diverse academic and professional backgrounds. Each entry features randomly assigned attributes including veteran status, work authorization, school names, GPAs ranging from 2.8 to 4.0, and degrees spanning six academic levels. To accurately simulate the demographic composition of real-world applicants, we include a demographic distribution where ethnicity is weighted, with approximately 70% of individuals assigned to either white or asian. Applicants' past role experiences encompass a range of technology-related professions, from semi-related fields like Teaching Assistants to strongly related positions such as Software Engineers and Full-stack Developers. This diversity in roles, alongside variable employment dates allows for a dynamic assessment of technical expertise and work history simulation.



During EDA, it is noted that applicants marked with a gender of 'N/A' receive a default prediction of zero indicates a significant oversight in handling missing or non-standard gender data.

Analysis Techniques

To investigate the decision-making processes of our models, we employed a combination of pre-processing, in-processing, and sensitive attribute filtering techniques. The model utilized in the audit is a customized logistic regression which includes multiple layers with regularization to prevent overfitting and uses a cross-entropy loss function to optimize binary classification tasks effectively. This setup helps to assess how well the model performs under varying conditions and identifies potential biases in its predictions.

In the pre-processing stage, we utilize the *Disparate Impact (DI) Remover* technique. This method is designed to adjust the distribution of data to minimize disparate impact where the fairness metric DI gauges the ratio of positive outcomes between disfavored and favored groups. By modifying data distributions before they are input into the model, this technique helps in reducing the potential for the model to perpetuate existing biases.

In the in-processing phase, we integrated algorithmic fairness by incorporating a *Prejudice Remover Regularizer* into our training algorithms. This approach introduces a penalty term to the learning objective, which actively discourages reliance on sensitive attributes like gender or ethnicity, promoting a decision-making process that is fairer and less biased. Sensitive attribute filtering ensures that explicitly sensitive attributes are not used directly as features in our models. By confirming the consistent disregard of attributes like disability status, we maintain the integrity of our fairness efforts and align our methodology with best practices in ethical AI development.

The introduction of *Reject Option Classification (ROC)* in the post-processing phase is pivotal. After initial model decisions are processed, ROC intervenes by reassessing decisions within a defined probability interval near the decision boundary, which we refer to as the grey zone. For decisions that are not made with high confidence (probabilities between 0.5 and a threshold theta), ROC adjusts classifications to correct potential biases. This method is essential for cases where the decision is tentative and more likely to be influenced by underlying biases. By recalibrating these borderline decisions, ROC enhances the fairness of the model, ensuring that individuals are neither unjustly favored nor unfairly penalized based on sensitive attributes.

Evaluation Criteria

In this audit, our evaluation strategy encompasses a comprehensive application of fairness and accuracy metrics across pre-processing(Feldman et al.) , in-processing (Kamishima et al, 2012), and post-processing techniques (Karmiran et al, 2012). These metrics, grounded in recognized research such as Feldman et al.'s work on certifying and removing disparate impacts, are deployed at each stage of the algorithmic processing pipeline to ensure our assessments align with both legal and ethical standards.

- Pre-processing and In-processing Techniques:

During both the pre-processing and in-processing phases, we employ the following metrics to systematically evaluate and address potential biases. Disparate Impact (DI) measures the ratio of positive outcomes between disfavored and favored groups, with a threshold of 0.8 for fairness evaluation, while Statistical Parity Difference (SPD) quantifies the discrepancy in favorable outcomes across groups, aiming for 0 disparity to signify perfect fairness. Similarly, the Average Absolute Odds Difference (AAOD) and Equal Opportunity Difference (EOD) assess disparities in true positive and false positive rates, and true positive rates alone, respectively, both aiming for a score of 0 to indicate no bias and equal treatment across all demographic groups.

- Post-processing Techniques:

In the post-processing phase, we further refine our evaluation using the Reject Option Classification (ROC) with repair levels represented on the x-axis and the above metrics on the y-axis. This approach allows us to visually and quantitatively assess how adjustments in the model's decision thresholds affect fairness metrics as repair level changes.

- Gender as Attributes:

Including gender as an attribute in our analysis is critical due to findings from EDA that raised significant concerns. During the EDA, it was observed that applicants marked with a gender of 'N/A' consistently received a default prediction of zero. This pattern suggests a substantial oversight in the model's handling of missing or non-standard gender data, potentially leading to discrimination against applicants who do not specify their gender. This explains why we incorporate the DI metric, as it will help us quantify and correct any biases that arise from how gender data is processed, ensuring the model treats all applicants fairly regardless of their gender identification.

Limitations

One significant limitation of our methodology is the use of a black box model, where the internal workings and the training data are not accessible. This limits our ability to audit or directly modify the model for bias mitigation. Our reliance on synthetic data also complicates the evaluation, as it may not accurately reflect real-world data complexities, potentially leading to inaccuracies in assessing the model's fairness and performance.

Furthermore, the tools and techniques we employ, which focus on pre-processing inputs and in-processing constraints, can mitigate some forms of bias but do not address inherent biases within the model's architecture or learning algorithms. Additionally, while filtering out sensitive attributes helps prevent direct discrimination, it doesn't tackle indirect biases stemming from correlated non-sensitive attributes, leaving potential for systemic discriminatory behaviors.

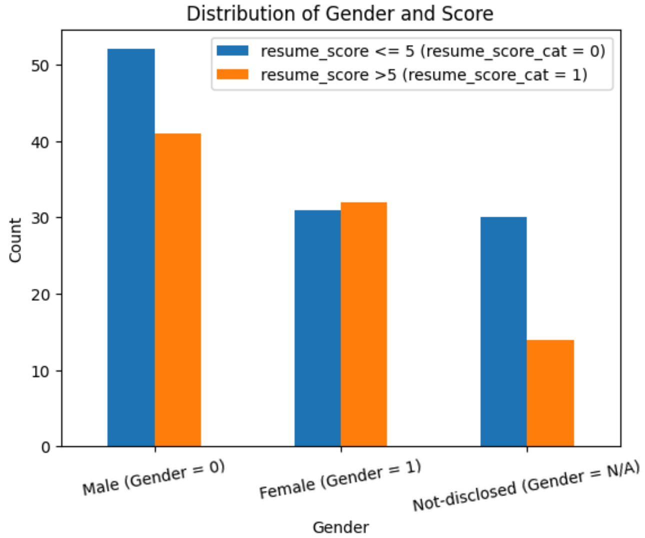
Findings

Resume Scoring System

- Model Metrics

	Before DI Remover	**After DI Remover**
Balanced accuracy	0.5170	0.5441
Disparate impact	0.8692	0.9305
Statistical parity difference	0.0760	0.0635
Average odds difference	0.0793	0.0687
Equal opportunity difference	0.1096	0.0894

- DI Remover

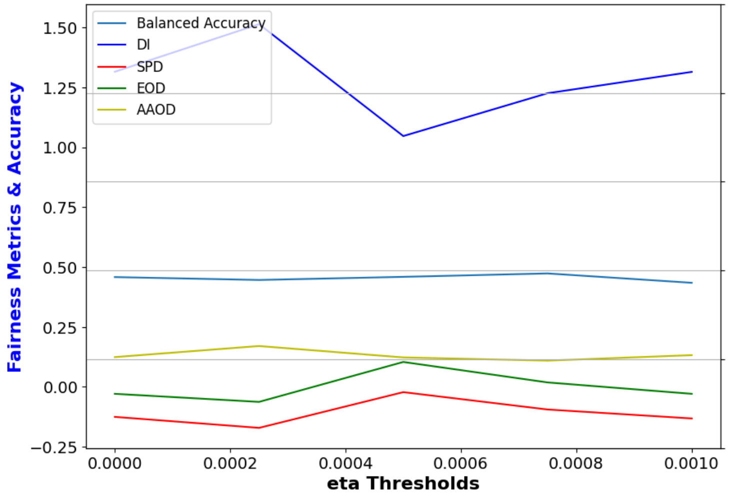


The results above show that the resume scorer model is somewhat biased in the first place with a DI score of 0.8692. It has moderate ability to predict accurate resume scores given a balanced accuracy of 0.5170. SPD, AAOD and EOD have near zero values, implying that the model is slightly biased in favor of the privileged group but can be considered fair in general.

With DI remover, balanced accuracy increased to 0.5441, indicating an improvement in the model's overall prediction accuracy. Disparate impact was further reduced closer to 1 (0.8692 to 0.9305), suggesting a decrease in bias between different groups. SPD, AAOD and EOD all show marginal improvements further closer to zero, indicating more equitable treatment across groups in terms of positive outcome rates and true positive rates.

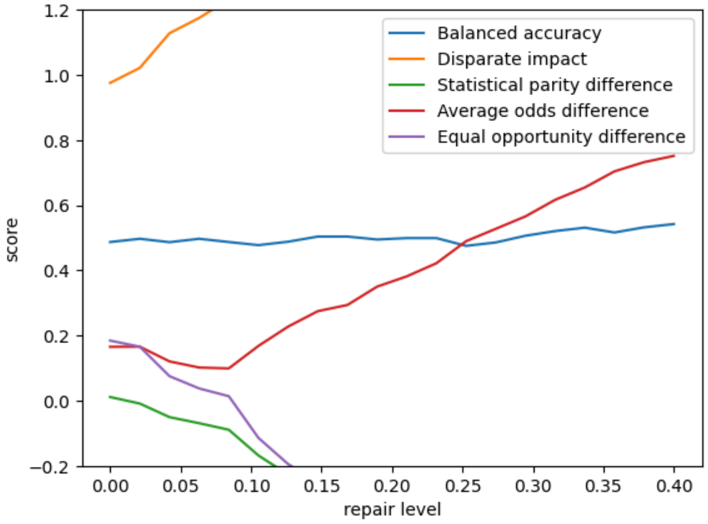
- Prejudice Remover Regularizer

eta	Balanced Accuracy	Disparate Impact	Statistical Parity Difference	Average Odds Difference	Equal Opportunity Difference
0.0	0.4581	1.3155	-0.1255	0.1241	-0.0292
0.00025	0.4461	1.5140	-0.1713	0.1703	-0.0629
0.0005	0.4591	1.0469	-0.0222	0.1227	0.1038
0.00075	0.4734	1.2257	-0.0947	0.1094	0.0184
0.001	0.4343	1.3149	-0.1320	0.1322	-0.0292



The use of PRR shows a complex impact on fairness metrics. While it improves fairness in terms of reducing disparities (as seen in the diminishing absolute values of SPD, AAOD, and improvements in DI and EOD), it also causes fluctuations in model accuracy. This indicates a trade-off that typically accompanies the use of fairness-enhancing techniques, where adjustments in one area can lead to temporary setbacks in others. It is also noteworthy that the performance of these impacts are fairly unstable, as we see fluctuations in metrics across different eta values.

- ROC



The repair algor ut significantly
compromising balancec ecially EOD,

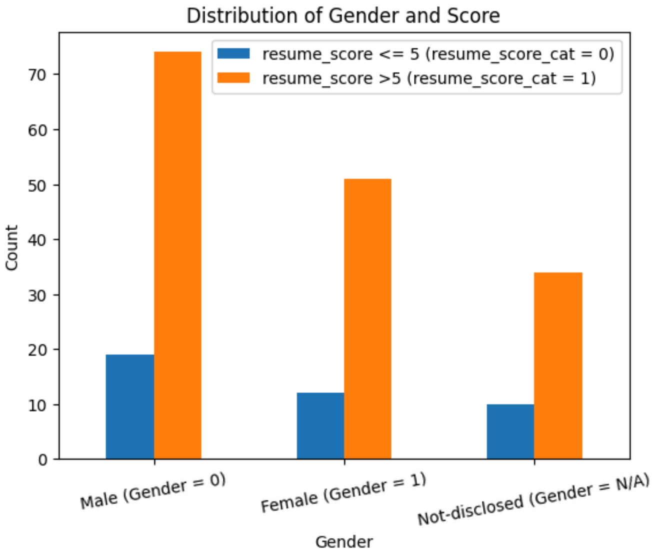
alongside the decreasing trend in SPD, highlight a successful adjustment towards fairer model outcomes as the repair level is increased. These results are promising, especially since the original resume scorer model is not heavily biased and leaves little room for fairness adjustments.

Candidate Evaluation Model

- Model Metrics

	Before DI Remover	**After DI Remover**
Balanced accuracy	0.5809	0.5472
Disparate impact	0.0828	0.9984
Statistical parity difference	0.4142	0.0013
Average odds difference	0.4013	0.0786
Equal opportunity difference	0.3752	0.0837

- DI Remover



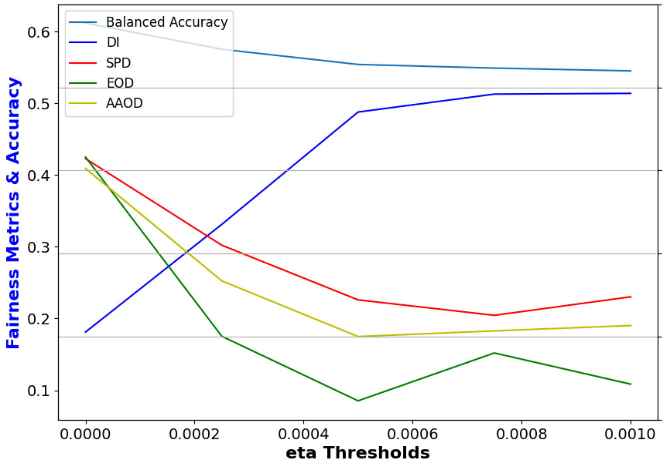
The results above show that the resume scorer model is extremely unbalanced with a low DI value at 0.0828. Such a low DI typically signifies that one group is highly favored over another. It is reasonably effective in accurately evaluating candidates given a balanced accuracy of 0.5809. SPD, AAOD and EOD are fairly high, suggesting notable disparities in both the rates of false and true positives across groups.

With DI remover, balanced accuracy decreased to 0.5472, indicating that the model sacrificed accuracy for fairness, which is fairly reasonable. Disparate impact drastically improved to 0.9984, showing nearly equal probabilities of positive outcomes across different groups. This is a significant enhancement towards fairness. SPD, AAOD and EOD all reduced to near zero values, suggesting notable improvement in fairness in the model. The bar chart reflects the distribution after the DI Remover, where a more balanced representation across groups in terms of score categories can be expected.

- Prejudice Remover Regularizer

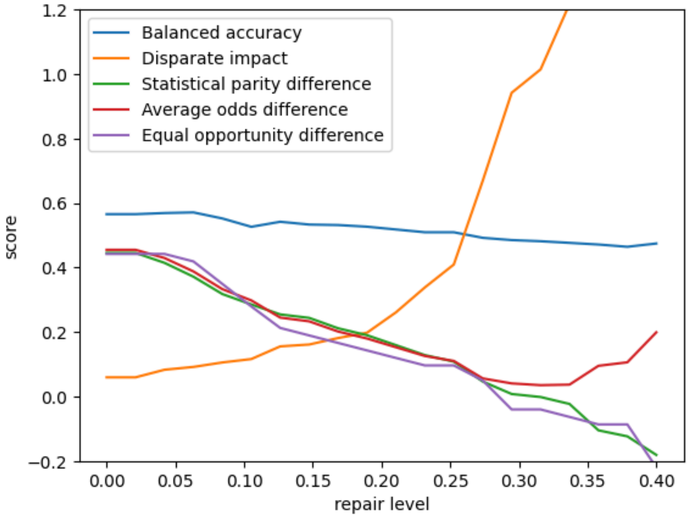
eta	**Balanced Accuracy**	**Disparate Impact**	**Statistical Parity Difference**	**Average Odds Difference**	**Equal Opportunity Difference**
0.0	0.6115	0.1811	0.4227	0.4089	0.4248

0.00025	0.5750	0.3311	0.3021	0.2524	0.1752
0.0005	0.5539	0.4876	0.2259	0.1748	0.0853
0.00075	0.5488	0.5126	0.2044	0.1827	0.1519
0.001	0.5449	0.5136	0.2301	0.1901	0.1085



The PRR is very effective in improving fairness metrics across the board as the regularization strength increases. While it does so, there is a trade-off with a decline in balanced accuracy, which is common when enforcing fairness constraints. The regularizer significantly corrects for disparate impact, statistical parity, and error rate disparities between groups, making the model substantially fairer.

- <u>ROC</u>



The repair algorithm also effectively enhances fairness in model predictions, particularly evident in the higher repair levels. It maintains balanced accuracy reasonably well, which is commendable given the substantial adjustments in other metrics. However, the potential for overcompensation in DI at higher levels suggests that finding an optimal repair level that maximizes fairness without introducing new biases is crucial. A repair level near 0.3 seems to be a good choice to enhance the candidate evaluation model.

Following the implementation of techniques, our analysis revealed an improvement in DI metrics and a reduction in other fairness metrics, suggesting that the original model did indeed incur biases. Our findings underscore the necessity of integrating fairness-enhancing methodologies throughout the model development process to prevent and mitigate discriminatory effects.

Recommendations

Model Design

Based on the findings from our audit, we recommend several enhancements to the model developed by Providence Analytica. Firstly, the imbalance gender distribution in the dataset suggests an underlying bias in the model's decision-making process, potentially due to the way missing data is treated or encoded. We recommend implementing comprehensive data preprocessing techniques that better manage missing values and ensure they do not systematically disadvantage any group. Moreover, integrating bias detection metrics directly into the model development and testing phases can provide ongoing oversight to identify and mitigate such biases before the model is deployed.

Additionally, Providence Analytica should consider enhancing transparency regarding how their models handle sensitive attributes. While the omission of direct sensitive attributes is a step towards fairness, the unintended side effects observed indicate that indirect biases might still be influencing the outcomes. Introducing explainability tools and methods can help both developers and clients understand how decisions are made, particularly in cases where model outputs significantly impact individuals' opportunities.

Company Practices

The use of the model's outputs in Bold Bank's HR decisions, such as offering first-round interviews, warrants a reassessment due to Providence Analytica's concerns about how it handles 'N/A' gender data. It is recommended that HR personnel be trained to critically evaluate model outputs and incorporate a manual review process for cases that may affect diversity and inclusion efforts negatively. Given the heavy reliance on the model's decisions, it is crucial to establish robust oversight mechanisms that combine model-based decisions with human judgment. This dual approach will help mitigate biases and enhance the fairness and effectiveness of the hiring process.

In summary, while the model developed by Providence Analytica aims to streamline and objectify the recruitment process, the identified biases necessitate a balanced approach that combines technological solutions with human oversight. By adopting these recommendations, Providence Analytica and Bold Bank can better align their operations with ethical standards and enhance their decision-making frameworks to be more inclusive and just.

Reference

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2783258.2783311

Kamiran, F., Karim A., , X. Zhang, "Decision Theory for Discrimination-Aware Classification," 2012. IEEE 12th International Conference on Data Mining, Brussels, Belgium, 2012, pp. 924-929, doi: 10.1109/ICDM.2012.45.

Kamishima, Toshihiro & Akaho, Shotaro & Asoh, Hideki & Sakuma, Jun. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. 35-50. 10.1007/978-3-642-33486-3_3.