Final Audit Report

https://github.com/9r0x/Fairness_audit

Introduction

We conducted a review of Bold Bank's automated hiring system, which uses Providence Analytica AI technology. Our primary focus was on the resume scoring and candidate evaluation models. These are the tools that determine who is in for an interview.

The goal of this review was to ensure that these AI tools are fair and adhere to the rules regulating equal employment opportunities. It is important that these systems do not discriminate against anyone based on their race, gender, religion, age, or other personal characteristics. Discrimination is not only illegal, but also unfair. It can prevent talented and qualified people from getting the jobs they deserve, and it can also harm the bank by excluding diverse perspectives that could help the business thrive.

In addition to testing for bias, we wanted to see how well these AI tools are performing—are they selecting the best candidates? Are they trustworthy and fair in their decisions? Ensuring that the hiring process is both effective and fair allows us to maintain a level playing field in which everyone has a fair chance based on their skills and potential.

Methodology

Data Source

The dataset we used contained anonymized resumes in CSV format, which captured a variety of attributes such as educational background, work experience, skills, and personal demographics. Our resume scorer assigned each resume a suitability score ranging from 0 to 10. This scoring and additional data were then used to communicate with our API, allowing the candidate evaluator model to determine whether a candidate should be invited for an interview.

Here's a glimpse into how we processed the data:

1. Loading the Data: We started by loading the sample resume data from a CSV file.
2. Handling Sensitive Information: We defined a dictionary to manage sensitive demographic data, such as gender, veteran status, and ethnicity, to define what is fairness in our analysis.
3. Creating Duplicates for Analysis: To test fairness, we generated multiple versions of each row by varying the sensitive attributes while keeping other data points consistent. This helped us analyze if different groups received different suitability scores under the same conditions.
4. Merging and Saving the New Dataset: We merged the altered sensitive attributes back into our dataset and prepared it for further analysis.
5. In an ideal scenario, with other columns unchanged, simply changing the sensitive attributes should not impact the score or the candidate decision.

Evaluation Criteria

In assessing the fairness of Bold Bank's hiring system, our evaluation criteria centered on the use of fairness metrics specifically tailored to the model's output type—regression for resume scores and classification for hiring decisions. These metrics were chosen based on their relevance to the types of decisions being automated and their ability to reveal potential biases in treatment and outcome across protected groups defined by sensitive attributes including `'Gender'`, `'Veteran status'`, `'Work authorization'`, `'Disability'`, `'Ethnicity'`.

Metrics Used:

- **Resume Scores:** We applied metrics such as Disparate Impact, Statistical Parity Difference, Balanced Accuracy, and Average Odds Difference. These metrics can help us assess whether the scores assigned by the resume scoring algorithm were distributed equitably across different demographic groups, thus affecting the fairness in the initial phase.
- **Hiring Decisions:** We employed classification-based fairness metrics, including Disparate Impact, Statistical Parity Difference. These metrics allowed us to evaluate whether the final hiring decisions exhibited any bias towards or against certain groups, thereby affecting the fairness in actual hiring outcomes.

Attributes Investigated:

- We primarily focus on sensitive attributes such as `'Gender'`, `'Veteran status'`, `'Work authorization'`, `'Disability'`, `'Ethnicity'`. We consider all other attributes legal in making a difference in the company's decision making process such as `'GPA'`, `'School Name'`, and `'Degree'`.

Analysis Techniques

Quantitative Analysis:

- Statistical Analysis: We calculated fairness metrics for both resume scoring and candidate evaluation. This involved statistical tests to compare outputs across demographic groups and identify significant differences.
- Regression Analysis: For the resume scorer, regression analysis is significant in understanding how different variables influence the scoring process, especially the sensitive attribute which should not contribute at all in the first place.

Data Validation Techniques:

- Data Quality Review: We ensure the data used for analysis was complete and accurately reflected the candidate population, validating the integrity and representativeness of the dataset.
- Model Audit: We examine the algorithms for transparency and consistency, checking for model stability across different subsets of data and ensuring that the model logic aligned with legal and ethical standards.

Limitations

Data Limitations: Our analysis depends heavily on the availability of and quality of the data provided. Incomplete data on applicant demographics or outcomes could have led to an incomplete assessment of the model's fairness. It is noticeable that our data contains a considerable amount of null values, which could be a potential issue for the fairness analysis.

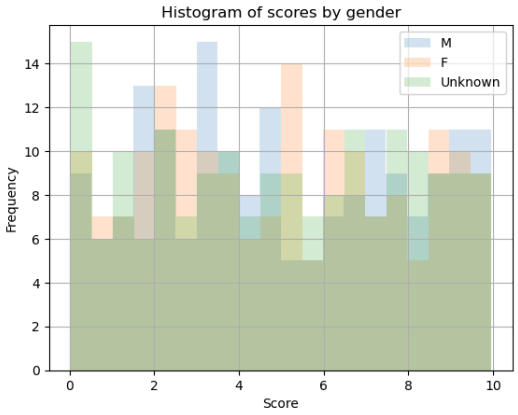
Metrics Limitations: The fairness metrics used, though widely accepted, can sometimes oversimplify the complexities of sensitive attributes and societal biases. Metrics like Disparate Impact and Equal Opportunity Difference provide useful insights but do not capture all nuances of fairness or ethical considerations.

Model Complexity: It is decided by the model's nature and certain aspects of the decision processes are not fully transparent, limiting our ability to inspect the complete logic and all underlying assumptions of the models.

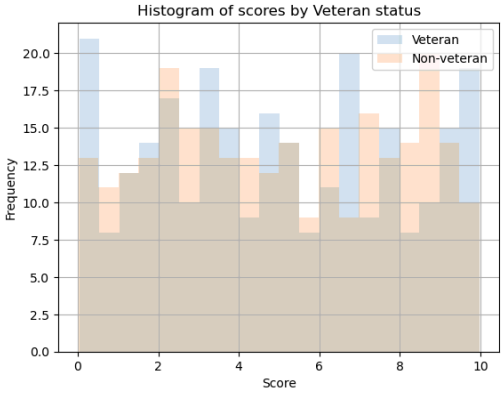
Findings

Exploratory data analysis

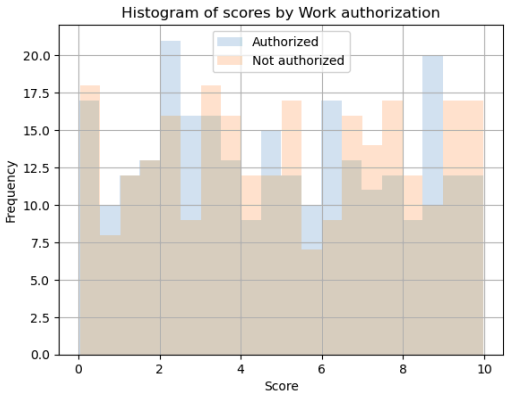
For scores by gender, we noticed that scores are relatively evenly distributed across the range. The overall frequency of scores for females tends to be lower compared to males, which could suggest fewer female entries or potential scoring biases.



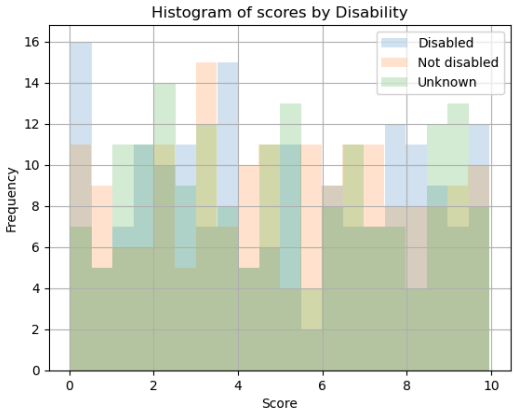
The scores for veterans are spread relatively evenly across the range. The frequency of scores for veterans is notably higher at the lowest end (0-2), suggesting either a pattern in the scoring model or a characteristic of the veteran resumes that results in lower scores.



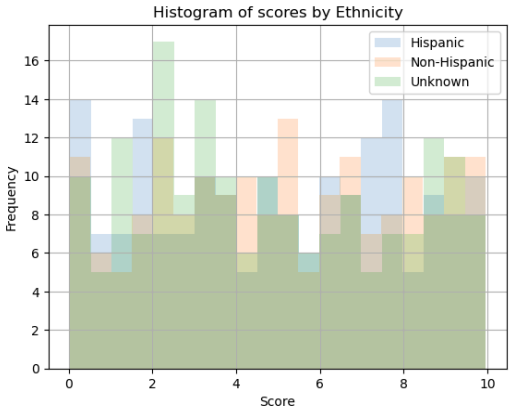
Both groups show a relatively uniform distribution across all scores with a gradual increase towards the middle ranges, but the frequency of high scores (8-10) is significantly lower for the Not Authorized group. The distribution suggests a potential bias where individuals without work authorization might be scoring lower.



The distribution for individuals identified as disabled shows a higher frequency at lower scores. Scores for the not disabled group are more evenly distributed across all ranges with a slight increase around mid-scores. The frequency of lower scores is notably higher for the Disabled group, which may suggest potential biases or differences in how their resumes are evaluated.



The distribution for Hispanics shows a higher frequency at lower scores while scores for Non-Hispanics are spread more evenly across all ranges. Lower scores appear to be more prevalent among Hispanics, suggesting potential issues in scoring fairness or differences in the quality or presentation of their resumes, especially if their qualifications are comparable to those of Non-Hispanics.



Audit findings

As was presented in the previous section, we have systematically computed the metrics for each of the sensitive attributes, and the results were presented below. We could see that there are slight but noticeable biases in all of the following attributes against the unprivileged group, but the bias is not at a high degree.

Metrics for Gender
Disparate Impact based on Gender: 0.9848074875798438
Statistical Parity Difference based on Gender: -0.07611111111111146
Mean Absolute Error (Privileged): 2.572015020576132
Mean Absolute Error (Unprivileged): 2.490090946502058
R-squared (Privileged): -0.00013386002675597197
R-squared (Unprivileged): -0.00021342060485673997
Average Odds Difference (Regression) based on Gender: 0.0761111111111111

Metrics for Veteran status
Disparate Impact based on Veteran status: 1.0194429413179413
Statistical Parity Difference based on Veteran status: 0.09581481481481458
Mean Absolute Error (Privileged): 2.578026886145404
Mean Absolute Error (Unprivileged): 2.512111111111111
R-squared (Privileged): -0.0002594992913607097
R-squared (Unprivileged): -0.00027743276929159677
Average Odds Difference (Regression) based on Veteran status: 0.09581481481481482

Metrics for Work authorization
Disparate Impact based on Work authorization: 1.047854584254249
Statistical Parity Difference based on Work authorization: 0.23255555555555585
Mean Absolute Error (Privileged): 2.5355958847736626
Mean Absolute Error (Unprivileged): 2.5545421124828533
R-squared (Privileged): -0.001598048136636887
R-squared (Unprivileged): -0.0015659583534435306
Average Odds Difference (Regression) based on Work authorization: 0.23255555555555557

Metrics for Disability
Disparate Impact based on Disability: 1.0203532853285329
Statistical Parity Difference based on Disability: 0.10050000000000026
Mean Absolute Error (Privileged): 2.6672878600823045
Mean Absolute Error (Unprivileged): 2.4816767489711937
R-squared (Privileged): -0.00015876828811411947
R-squared (Unprivileged): -0.0004685867753499995
Average Odds Difference (Regression) based on Disability: 0.10049999999999999

Metrics for Ethnicity
Disparate Impact based on Ethnicity: 1.0275380738956896
Statistical Parity Difference based on Ethnicity: 0.13722222222222147
Mean Absolute Error (Privileged): 2.57715658436214
Mean Absolute Error (Unprivileged): 2.4962475308641974
R-squared (Privileged): -5.76721622413956e-06
R-squared (Unprivileged): -0.0024738503239893905
Average Odds Difference (Regression) based on Ethnicity: 0.13722222222222227

In addition to the score analysis, we also performed a special analysis against the decision making process. From the results pasted below, we noticed that this process adds more unfairness than the scoring itself. We found that the process shows clear differences for

different groups. Assuming that the scoring model was relatively fair, we found that this decision process has clear preferences for one group over the other.

Metrics for Gender
Disparate Impact based on Gender: 0.6507936507936508
Statistical Parity Difference based on Gender: -0.1222222222222222

Metrics for Veteran status
Disparate Impact based on Veteran status: 0.8571428571428572
Statistical Parity Difference based on Veteran status: -0.029629629629629617

Metrics for Work authorization
Disparate Impact based on Work authorization: 0.9622641509433961
Statistical Parity Difference based on Work authorization: -0.007407407407407418

Metrics for Disability
Disparate Impact based on Disability: 0.9411764705882354
Statistical Parity Difference based on Disability: -0.011111111111111

Metrics for Ethnicity
Disparate Impact based on Ethnicity: 1.090909090909091
Statistical Parity Difference based on Ethnicity: 0.01666666666666669

Regression Analysis Findings:

Gender: 0.2902313299911564
Veteran status: -0.23038445093969992
Work authorization: -0.17185752868660079
Disability: -0.17041185503105022
Ethnicity: -0.007419267085687381

From our regression analysis, we found biases linked to specific attributes such as gender, veteran states, work authorization, disability, and ethnicity. While the positive coefficient for gender suggests a preference towards males, negative coefficients for veteran status, work authorization, and disability indicate that these groups are likely to have lower scores. Even the minor negative influence with ethnicity can show at some degree of bias. Although these biases appear differently in magnitude, they are crucial and require adjustments in the model to ensure fairness and prevent discriminatory outcomes.

Recommendations

Model Design

After reviewing the audit and feedback, we have a few suggestions to help Providence Analytica make their model better and more equitable:

Improve Transparency: Providence Analytica should make its model more transparent. They should clearly explain which factors the model considers and how these factors influence the scores. This will help everyone understand how decisions are made, resulting in increased trust in the AI system.

Conduct Regular Fairness Checks: Just like our audit, it would be beneficial to make fairness checks a regular practice as according to the interview, the model uses the data from the bank.

This way, any biases that arise can be addressed quickly. Additionally, Providence Analytica should use the feedback from these checks to update the model on a regular basis.

Adjust Factor Importance: Rather than treating all factors equally, Providence Analytica should consider adjusting the importance of each factor based on its relevance to the job and its impact on fairness.

Add More Data Varieties: To avoid biases caused by uneven data, Providence Analytica should use methods such as creating synthetic data or enhancing existing data. This will allow the model to train on a more balanced set of data.

Company Practices
Here's how Bold Bank can better use the AI system in their hiring and align it with their goals:

Add a Review Step: Bold Bank should take a second look at candidates who are on the edge or may have been unfairly judged by the AI. This step, taken by human HR professionals, will ensure that everyone has a fair chance.

Check Strategic Alignment: Bold Bank should check on a regular basis to see if the AI system is helping them meet their goals, such as making more money and growing the business, while also ensuring that it attracts and retains the best people without violating any job laws.

Keep Things Clear and Open: It's important to keep candidates informed about how AI is used in the hiring process. Bold Bank could include detailed info on their website and in job applications about the use of AI.