CSCI 1951Z Final Report

Li-Heng Pan, Sagar Raichandani

Introduction

In this project, we play the role of an auditing team placed by EEOC to investigate Bold Bank's hiring practices. Bold Bank is a financial institution that has recently implemented a new hiring system developed by Providence Analytica to streamline its recruitment process. The purpose of this audit is to investigate complaints that allege discriminatory outcomes in the new hiring process. Through our analysis, we hope to ascertain that the candidate selection process is not plagued by discriminatory outcomes because of a person's race, color, religion, sex, national origin, age, and disability.

To conduct our review, we will rely on a combination of stakeholder interviews– job applicant, Providence Analytica, and Bold Bank representatives– and automated flows to probe the resume scorer and candidate evaluator models. At a high level, the hiring system consists of two phases:

1. A resume scoring model that assigns a score of 0 to 10 to a CSV-formatted resume, indicating the candidate's suitability for a specific role
2. A candidate evaluator model customized to the company's preferences, utilizes the resume scores to generate a binary outcome for whether the candidate should receive an interview.

While these two systems are intricately linked, in our analysis we will be examining their outputs independently and their relation to each other.

Methodology

Bearing the objectives discussed in the previous section, our approach was structured as:

1. We conducted thorough interviews with 3 stakeholders, gathering relevant information as follows:
   (a) From Brianna Brown, a recent graduate from Central University, we learned about the job application process, ensuring that candidates of a specific gender, race, nationality, or disability were not disadvantaged in the application process.
   (b) We probed Providence Analytica for information on their data collection strategy, treatment of personal information, fairness considerations in each of the models, and evaluations of fairness trade-offs with accuracy and Bold Bank's requirements.
   (c) Finally, with Bold Bank, we delved into their hiring practices, information shared with Providence Analytica, fair hiring checks and balances employed with the Bank, and mechanisms for redressal.
2. We set up a data generation pipeline that picked attributes of each applicant from a random distribution– usually, a uniform random distribution from a fixed set for categorical features and a normal distribution for numeric features. While we experimented with different distributions, uniform distributions made the dataset easy to work with and more interpretable. More details are in the subsection that follows.
3. For evaluation and analysis, we analyzed the distributions of the resume scores and interview selection rates and stratified these results by gender, race, and disability status. Our findings along with fairness metrics and interpretability measures will be presented in the subsections that follow.
4. Finally, we noted some of the limitation with our approach, challenges working with Providence Analytica and Bold Bank teams, and recommendations for improvement.

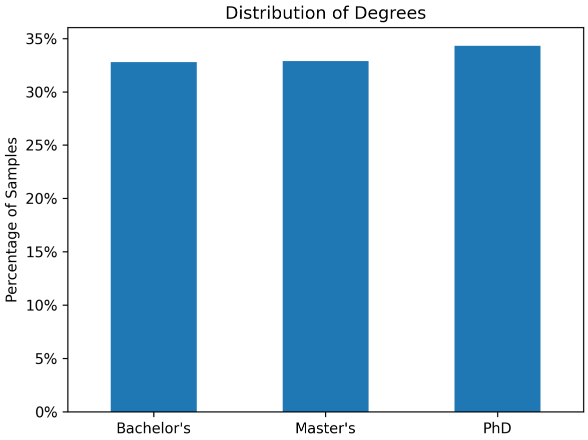

Figure 1: Distribution of Degree

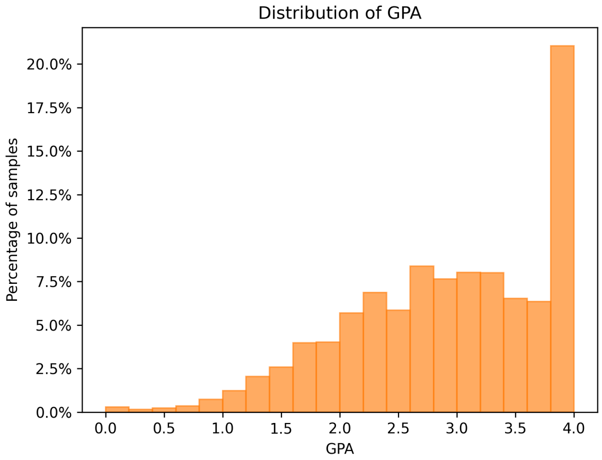

Figure 2: Distribution of GPA

Data Source

We set up a data generation pipeline to create a synthetic dataset where each feature was drawn independently from a random distribution. We also found it advantageous to generate identical samples multiple times to gauge the variability in the models' outputs. With such a strategy we repeated each sample 10 times, for a set of 4000 distinct applicants, totaling 40000 rows of data. Each feature is sampled using standard Numpy functions in a reproducible manner. More details can be found here[1].

Our choice to rely on uniform distribution for each feature– as opposed to a more realistic sampling strategy– ensured that we had a sufficient balance to make observations for different groups with the same confidence. As depicted in the bar graph above (Figure 1), we created a dataset that had a comparable number of Bachelor's, Master's, and PhD applicants. Likewise, for GPA (Figure 2) we sampled from a normal distribution with mean $(\mu) = 3$ and standard deviation $(\sigma) = 1$. Table 1 provides a view into the number of applications we queried for each gender-ethnicity sub-group. Likewise, Table 2 demonstrates that we created query samples where individuals had a range of experience and prior job history.

	Ethnicity				
Gender	**Asian**	**Black**	**Native American**	**Pacific Islander**	**White**
Male	2440	2740	2730	2730	2750
Female	2460	2710	2640	2610	2760
N/A	3000	2340	2730	2480	2880

Table 1: Distribution of applicants' ethinicity and gender

[1] Check project repo file: src/dataprep-flow.ipynb

	Years of Experience					
Number of Jobs	Less than 1	1-3	3-5	5-10	10-15	15+
0	3630	0	0	0	0	0
1	0	6750	3240	0	0	0
2	0	3300	0	0	0	0
3	0	0	3270	13360	3140	3310

Table 2: Distribution of applicants' length of experience

Evaluation Criteria

We started off by considering some baseline statistical measures based on the mean, median, and standard deviation of the scores returned by the resume scorer model and predictions returned by the candidate evaluator. Making 10 calls to for each sample allowed us to estimate confidence intervals and test for statistical significance of each of our findings. In addition to statistical analysis, we identified sensitive features, and sensitive feature combinations to measure the changes in mean responses across groups. We quantified the same through two fairness metrics:

1. Statistical parity difference:

$$\text{SPD} = \mathbb{P}(\hat{Y} = 1 | A = \text{minority}) - \mathbb{P}(\hat{Y} = 1 | A = \text{majority})$$

where, $\hat{Y}$ is the predicted outcome, and A is the sensitive attribute

2. Disparate impact:

$$\text{DI} = \frac{\mathbb{P}(\hat{Y} = 1 | A = \text{minority})}{\mathbb{P}(\hat{Y} = 1 | A = \text{majority})}$$

where, $\hat{Y}$ is the predicted outcome, and A is the sensitive attribute

For the aforementioned statistical summary measures and the fairness metrics, we thorougly investigated patterns of bias due to gender, age (estimated based on prior work experience), gaps in job history, ethnicity, two-way combination of ethnicity and gender, and disability. The reasons for picking these as sensitive attributes was mostly based on our intuition of factors that have a history of societal/institutional bias– typically against people of color, women and non-binary individuals, the differently-abled, and experienced professionals who may be discriminated with due to their age.

Finally, we considered looking at the global intepretability SHAP values to better understand the features the resume scorer and candidate evaluator models depend upon, and identify sensitive features that are modeled for incorrectly. This was also in an effort to gauge if the model has any sources of computational or human bias– especially due to interaction effects– which may not be immediately apparent through the more rudimentary statistical techniques listed above.

Analysis Techniques

During the early phases of the audit pipeline, we recognized that the model's responses are not always stable– with resume scores showing an approximate population standard deviation (σ) of 2.85, and mean (μ) of 5– and closely resembled a uniform random distribution. Likewise with the predictions, it was difficult to spot a reproducible pattern. Bearing these observations, we found it advantageous to query the same candidate 10 times, and take a mean of each instances response as the output. This helped us gain more insight into larger trends in the dataset, and estimate confidence intervals for each of our findings and observations.

Next, we directed our efforts towards quantifying the differences in the model's response for several subgroups based on a single feature or multiple. We looked at various fairness metrics, as detailed in the previous section, recorded our observations through graphs and tables, and finally considered Shapley values to gauge model intepretability. All of these will be explained in more detail in the following sections.

Limitations

While our audit strategy has helped us identify deep insights into the model's workings, we do find that the following aspects could have been improved:

1. Without access to training data and ground-truth values for resume scores or candidate evaluations, we found it difficult to explore fairness metrics like separation, sufficiency, false positive rate (FPR), and average absolute odds difference (AAOD) which are conditioned on the true value by definition.
2. While we did investigate two-feature interactions for certain sensitive attributes– for example, race and gender, number of jobs and employment gaps, and degree levels and GPA– we were unable to look at many more combinations, and especially failed to investigate interaction effects involving more than 2 features.
3. During the investigation process, we did not receive a sufficiently detailed response from the representatives of Bold Bank and Providence Analytica. This hampered our progress, and despite providing sufficient evidence of the randomness in the models' outputs, their teams failed to acknowledge our assumptions.
4. We were slowed by the compute limitations placed by the API. Model calls were slow, and required a lot of manual effort to get right. While we eventually set up an automation pipeline, the responses were limited to 4000 rows per call, reducing the scope of sending large samples to verify statistical integrity.

Findings

Our analysis revealed several crucial findings regarding the hiring practices at Bold Bank. These findings highlight potential concerns in the fairness and functionality of the recruitment process:

1. **Unfair treatment towards individuals identifying as female or non-binary:**
   Our data indicates a glaring issue with the selection of female candidates and individuals who refused to divulge gender information either because they are non-binary or other personal reasons. What stands out is that the resume scores are similar across gender categories, but the selection rates are vastly higher for male candidates. This is also beyond any reasonable bounds such as the $\frac{4}{5}^{\text{th}}$ rule.

   Table 3 highlights the same and shows that controlling for a similar sample size, while the mean resume scores, their standard deviations– and by extension the standard errors, since group size is fixed at 10– are rather similar, the candidate evaluator model's selection rate is vastly different. The mean selection rate highlights this difference– where individuals identifying as female have a 38% lower selection rate when compared to men. For those who deny sharing gender information or identify as non-binary, the selection rate is 0%.

   The above analysis can also be viewed through the lens of fairness metrics covered in class.

Gender	# applicants	Mean resume score	std error resume score	Mean selection rate
Male	1,339	5.02	2.85	0.49
Female	1,318	4.99	2.83	0.31
Non-binary	1,342	4.98	2.85	0

Table 3: Gender-based discrepancies

Let's consider the statistical parity difference (SPD)– defined as the difference between the proportion of positive outcomes for the majority and minority groups. Mathematically,

$$\text{SPD} = \mathbb{P}(\hat{Y} = 1|A = \text{minority}) - \mathbb{P}(\hat{Y} = 1|A = \text{majority})$$

Setting the predicted outcome $(\hat{Y})$, in the above equation, to be the output of the candidate evaluator model, we can see that $\mathbb{P}(\hat{Y} = 1|A = \text{minority}) = \mathbb{E}(\hat{Y}|A = \text{minority})$, since our outcome is binary. Thus, substituting the mean selection rate in the above equation, we find that:

$$\begin{aligned}\text{SPD}_{\text{f-m}} &= \mathbb{E}(\hat{Y}|\text{gender} = \text{female}) - \mathbb{E}(\hat{Y}|\text{gender} = \text{male}) \\ &= 0.31 - 0.49 = -0.18\end{aligned}$$

Likewise, for non-binary candidates, we see:

$$\begin{aligned}\text{SPD}_{\text{na-m}} &= \mathbb{E}(\hat{Y}|\text{gender} = \text{N/A}) - \mathbb{E}(\hat{Y}|\text{gender} = \text{male}) \\ &= 0.00 - 0.49 = -0.49\end{aligned}$$

Disparate impact is very similar to SPD, except that it is a ratio of the above proportions as opposed to a difference. Mathematically,

$$\text{DI} = \frac{\mathbb{P}(\hat{Y} = 1|A = \text{minority})}{\mathbb{P}(\hat{Y} = 1|A = \text{majority})}$$

Substituting the values as we did above:

$$\begin{aligned}\text{DI}_{\text{f/m}} &= \frac{\mathbb{E}(\hat{Y}|\text{gender} = \text{female})}{\mathbb{E}(\hat{Y}|\text{gender} = \text{male})} \\ &= \frac{0.31}{0.49} = 0.63\end{aligned}$$

Similarly, for non-binary individuals:

$$\begin{aligned}\text{DI}_{\text{na/m}} &= \frac{\mathbb{E}(\hat{Y}|\text{gender} = \text{N/A})}{\mathbb{E}(\hat{Y}|\text{gender} = \text{male})} \\ &= \frac{0.00}{0.49} = 0.0\end{aligned}$$

2. **Penalization for career gaps over 2 months:**
   Another critical finding is the model's response to candidates with career gaps exceeding two months. Despite similar resume scores, the selection rates for candidates who have a gap more than 2 months is 0%.

   Table 4 highlights the same and shows that even with a sizable sample, while the mean resume scores, the standard deviations– and by extension the standard errors– are rather similar, the candidate evaluator model's average selection rate is approximately 30%, while the same for individuals with as little as a 2-month gap in their work history it is 0%.

Total job gap	# applicants	Mean resume score	std error resume score	Mean selection rate
< 2 months	3,334	4.98	2.83	0.32
>= 2 months	665	5.04	2.87	0.00

Table 4: Career-gap-based discrepancies

Expressing the same using some of the fairness metrics we have seen above, we can say that:

$$\begin{aligned}\text{SPD}_{\text{gap}} &= \mathbb{E}(\hat{Y}|\text{gap >= 2 months}) - \mathbb{E}(\hat{Y}|\text{gap < 2 months}) \\ &= 0.00 - 0.32 = -0.32\end{aligned}$$

Similarly with disparate impact,

$$\begin{aligned}\text{DI}_{\text{gap}} &= \frac{\mathbb{E}(\hat{Y}|\text{gap >= 2 months})}{\mathbb{E}(\hat{Y}|\text{gap < 2 months})} \\ &= \frac{0.00}{0.32} = 0.0\end{aligned}$$

3. **Lack of correlation between resume scores and selection rates:**
   In Figure 3, we observe that there is no clear relationship between resume scores and the likelihood of selection for an interview. This lack of correlation indicates that the resume scoring model might not effectively influence the final hiring decisions and that the candidate evaluator model holds a more significant position in the hiring process.

4. **Randomness in resume scores:**
   The resume scoring model exhibited substantial variability in scores assigned to identical profiles. This inconsistency suggests potential flaws in the model's stability or its underlying algorithms, leading to questions about its reproducibility and the fairness of the candidate assessment process.

   We further investigated this trend by employing Shapley values to study the resume scorer model's dependence on its input features. The SHAP values presented in Figure 4 indicate that the model does not significantly utilize any of the features to determine its output. All the features display a minimal impact close to zero on the model's decisions. This could imply potential issues with how the model is structured or its overall functionality in assessing candidate data.

,

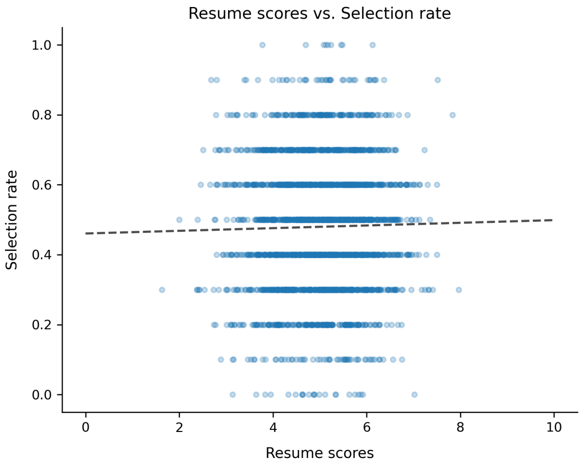

Figure 3: Scatterplot between resume scores and selection rates

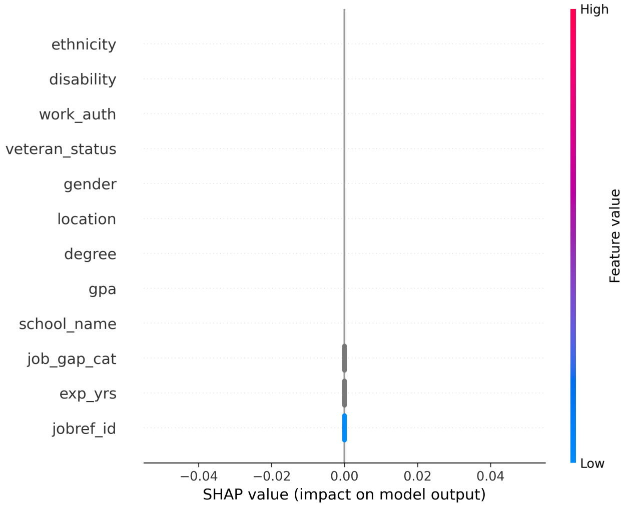

Figure 4: SHAP

Recommendations

Having identified some consistent trends demonstrating bias, we recommend the following actions to Providence Analytica and Bold Bank.

1. **Model design**

   - The candidate evaluator model is extremely unfair to female and non-binary individuals, with a 38% and 100% lower selection rate respectively. To solve for this, and reduce disparate impact, we recommend that Providence Analytica experiments with different classification thresholds; tuning for fairness as opposed to accuracy. Techniques like Uniform Sampling– a resampling procedure that balances the dataset through a learned weighting scheme or Cost-Sensitive Learning may be used.
   - We also observed that the candidate evaluator model's prediction bears no correlation to the resume scorer model's output. We'd recommend another training cycle where feature importances are thoroughly examined through techniques like Permutation Feature Importance, LIME, and Shapley values.
   - Through our review and stakeholder interviews, it seemed like representatives at Providence Analytica were unaware of the bias against applicants with gaps in their work history. Employ better checks and balances to make sure that sub-groups formed from the interaction of two or more features are examined, and in addition to accuracy fairness measures are prioritized.
   - Finally, we recommend stricter guidelines that must be shared with clientele on fairness thresholds and modeling procedures that do not solely prioritize accuracy.

2. **Company practices**

   - We found the fairness checks employed by the Bank to be very poor, with little justification provided for any evaluation their teams have made of the gender-based discrepancies we observed. We recommend building internal teams consisting of technical stakeholders and recruiters who can evaluate the responses shared by Providence Analytica for accuracy and fairness.
   - From our stakeholder interviews, it came to light that the condition to not hire individuals with a gap in their work history likely emerged from the preferences shared by Bold Bank with Providence Analytica. We do not approve of this and believe that resume gaps are a reasonable part of motivated individuals who may have found themselves unable to work due to health reasons, maternity leaves, and educational pursuits. These instances must be thoroughly investigated and any instances of deliberate exclusion due to correlation with age or gender must be stopped.
   - We found no mechanism through which candidates can seek redressal. Candidates must have the right to know that their applications are being evaluated through a model and must be able to opt for human evaluation if they believe their individual circumstances warrant the same. Not only does Bold Bank fail to provide for a redressal mechanism, by not providing alternate mechanisms to send in applications they're likely discriminatory against individuals with certain disabilities.