Zoey Katzive and Huda Saeed
Professor Suresh Venkatasubramanian
CSCI 1951Z: Algorithmic Fairness
8 May 2024

Bold Bank Hiring Audit

Introduction

Bold Bank is a financial institution that has recently implemented a new hiring system developed by Providence Analytica. The hiring system consists of a resume scorer, which provides a resume score from 0 to 10, and this is used as a feature in the candidate evaluator along with other features that incorporate the company's priorities. The candidate evaluator then outputs a binary decision of whether they receive an interview.

The Equal Employment Opportunity Commission received complaints that this system may be discriminatory. Hence, this audit aims to investigate whether these allegations of discrimination are true concerning race, color, religion, sex (including pregnancy and related conditions, gender identity, and sexual orientation), national origin, age (40 or older), disability, or genetic information.

Methodology

1. Data Source

   Once the APIs for the resume scorer and candidate evaluator were released along with the data format that they accepted, random data was generated on Python 3.11.5. 3,995 candidates were generated due to computational expense and their attributes were determined through the following method:

   The schools, locations, and roles were chosen randomly from the schools, locations, and roles in the sample dataset provided with the APIs; since these features are not sensitive attributes this was a simple method to fill these features given their infinite possibilities.

   The GPA was randomly chosen from a Normal distribution with a mean of 2.8 and a standard deviation of .4 to reflect rare instances of extremely high and low GPAs [1].

   Degree, gender, veteran status, work authorization, disability, and ethnicity were all randomly chosen from their fixed choices which are (Bachelor's, Master's, PhD), (Male, Female, N/A), (Yes, No, N/A), (Yes, No), (Yes, No, N/A), and (White, Black, Native American, Asian American & Pacific Islander, Other) respectively.

   The start date for each first role was randomly chosen from 2020 to 2023 as this represents the candidate's most recent job, and the end date was randomly selected from N/A and anywhere between 1 and 12 months after the start date. The second role was randomly selected from N/A and the roles in the sample dataset; if it was not N/A the start date was randomly chosen from 6 to 12 months before the start date of their first role and the end date was randomly chosen from 1 to 5 months before the start date of their first role. The third role was randomly selected from N/A and the roles in the sample dataset; if it was not N/A the start date was randomly chosen from 12 to 18 months before the start date of their first role and the end date was randomly chosen from 6 to 12 months before the start date of their first role. This maintained the

chronological order of the candidates' work history. If any of the roles were N/A, so were their dates and the following roles and dates.

All the features are roughly uniformly distributed since they were randomly chosen except for GPA which followed a Normal distribution and Role 3, the end dates, and the start dates for the second and third roles all of which had a mode of N/A. This is because if Role 2 was N/A that automatically resulted in an N/A for Role 3 so there was a greater likelihood of N/A in Role 3. Furthermore, N/A was also the most likely in all of the dates except for the Start 1 date as it was always considered an option in addition to dates relative to each candidate's Start 1 date which varied.

2. Evaluation Criteria

Correlations were used to identify potential proxies for sensitive attributes as they are dropped during training. Additionally, independence, Statistical Parity Difference, and Disparate Impact were calculated as these fairness metrics do not require knowledge of the "true" outcome as only the model's predictions are available. The attributes of gender, disability, work authorization, veteran status, and ethnicity were investigated as these may have legal implications if discriminated against in hiring.

3. Analysis Techniques

After the resume scores and candidate evaluations were obtained from the generated data, visualizations were first created to investigate if there seemed to be a disparity between interview attainment and the five sensitive attributes defined earlier. Then, correlations were computed between features to identify potential proxies of the sensitive attributes. Finally, independence, Statistical Parity Difference, and Disparate Impact were calculated for these sensitive attributes (pair-wise with each ethnicity) with the four-fifths rule as a cutoff to determine the presence of bias in independence and Disparate Impact [2]. To facilitate these metric calculations, female and male and no work authorization and work authorization were recoded as 1 and 0 which represent minority and majority groups respectively. Lastly, N/A responses were also investigated against the majority and minority groups in the attributes that allow an N/A response (disability, work authorization, and gender) through independence metrics, which is the same as Disparate Impact and a symbol manipulation of Statistical Parity Difference.

4. Limitations

The limitations of the methodology include the fact that the dataset was synthetically produced, thus the assumption that GPA is Normally distributed and many of the other features are uniformly distributed may not reflect reality which may have a bearing on the results below. Furthermore, since it is synthetically produced, proxies of sensitive attributes may be missed since all the features are randomly generated. Additionally, the time constraints within the dates of each candidate's roles also do not reflect the reality of all candidates (ie, perhaps some applied after unemployment for an extended period of time) which may also have a bearing on the results. Furthermore, other fairness metrics could not be computed due to the nature of the problem where the "true" outcome of whether a candidate is interviewed is the same as the predicted outcome since Bold Bank uses this algorithm in their hiring process; these metrics could have provided greater insight into the algorithm's behavior. Another limitation is that the algorithm itself was not accessible which would have provided a greater understanding

of where the bias manifested itself, i.e. in the cost function, the missing values handling, etc.

Findings

The central finding of this audit is that candidates who apply as female or N/A in the gender field and Native Americans are discriminated against in this hiring system.

The distribution of applicants who received an interview and those who did not is roughly uniform across ethnicity, work authorization, veterans status, and disability (see Appendix). However, about 12.9% of male applicants received an interview compared to 8.1% of female applicants, almost a 60% higher rate for males which warrants additional scrutiny into the algorithm's treatment of the gender attribute.

prediction	Not Interviewed	Interviewed
Gender		
F	0.919044	0.080956
M	0.871157	0.128843
N/A	1.000000	NaN

Table 1: Proportions of each gender given an interview versus not

Our analysis of variable correlations reveals that there are no potential proxy variables for the sensitive attributes in the data. The greatest correlation in the dataset is -0.051 between GPA and gender which is extremely low. Earlier in this audit process, Providence Analytica reported that they omit sensitive attributes during training; this is not an effective way to account for bias, as it does not account for potential proxies. Although proxies were not identified in this investigation since it relied on synthetic data where each feature was randomly and independently generated, there still may be potential proxies for these sensitive attributes which facilitate bias.

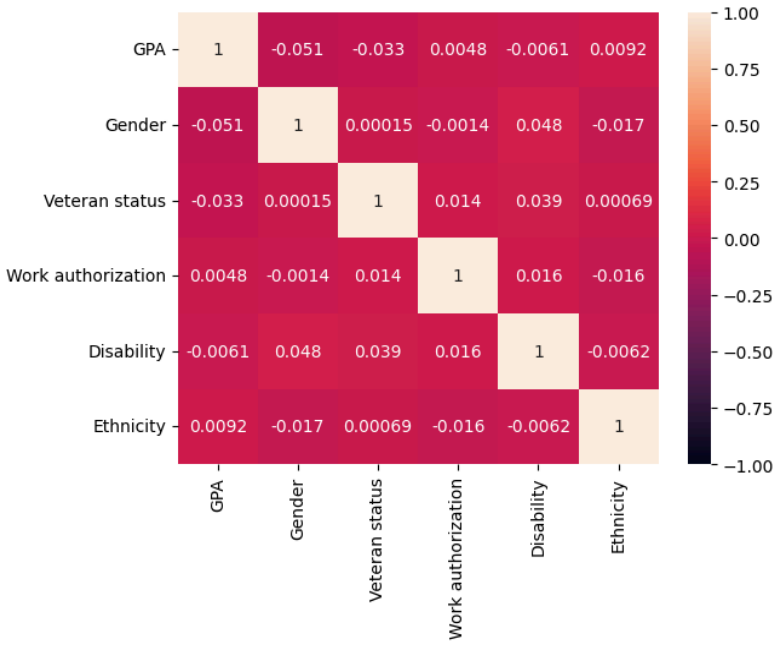

Figure 1: The dataset contains no significant proxies.

After no potential proxies were identified, independence, Statistical Parity Difference, and Disparate Impact score analyses yielded potential biases against women, unknown genders, and Native Americans.

The independence scores (which are equivalent to the Disparate Impact scores as minority hiring rates were calculated on the numerators) for disability, veteran status, and work authorization all pass the four-fifths rule (they fall between 0.8 and 1.25). Gender has an independence of 0.63 across both genders, which falls outside of the four-fifths range and suggests that male applicants are more likely to receive an interview than female applicants.

The independence scores across each different ethnicity mostly pass the four-fifths rule, with the exception of the scores between Black and Native American applicants (1.26), Native American and Asian American and Pacific Islander applicants (0.78), and Native American and Other ethnicity applicants (0.68). This implies that Native Americans are discriminated against in hiring compared to all other ethnic groups except Whites at a concerningly higher rate.

Variable	Independence
Gender	0.628329
Veteran status	1.034860
Work authorization	1.064096
Disability	1.041002

Table 2: Independence scores for gender, veteran status, work authorization, and disability

Group 1	Group 2	Independence
White	Black	0.954545
White	Native American	1.204489
White	Asian American & Pacific Islander	0.943442
White	Other	0.822500
Black	Native American	1.261845
Black	Asian American & Pacific Islander	0.988367
Black	Other	0.861667
Native American	Asian American & Pacific Islander	0.783271
Native American	Other	0.682862
Asian American & Pacific Islander	Other	0.871808

Table 3: Independence scores between ethnicities

Statistical Parity Difference (SPD) scores for each sensitive attribute may also indicate that the model exhibits a gender and Native American bias. Veteran status, work authorization, and disability received SPD scores extremely close to the perfectly fair score of 0 (0.002, 0.004, and 0.003, respectively), while gender's SPD score had a magnitude more than 10 times higher (-0.048).

The SPD scores across each ethnicity were all of lower magnitude than the SPD score for gender, but note that the SPD scores with a magnitude of at least .01 were all the pairs with Native Americans (all of which imply Native Americans were hired at a lower rate than the other ethnicity) as well as between White and Other which had an SPD of -.01 which implies that Other ethnicity's hire rate was .01 greater than the White hire rate.

Variable	SPD
Gender	-0.047887
Disability	0.002845
Veteran status	0.002417
Work authorization	0.004370

Table 4: SPD scores for gender, veteran status, work authorization, and disability

Group 1	Group 2	SPD
White	Black	-0.003247
White	Native American	0.014606
White	Asian American & Pacific Islander	-0.004040
White	Other	-0.012679
Black	Native American	0.017853
Black	Asian American & Pacific Islander	-0.000793
Black	Other	-0.009432
Native American	Asian American & Pacific Islander	-0.018646
Native American	Other	-0.027285
Asian American & Pacific Islander	Other	-0.008639

Table 5: SPD scores across different ethnicities

Lastly, a potential flaw was discovered in the model that causes it to predict 0 (no interview) whenever the gender field is N/A. However, this was not the case for the disability or veteran status variables, both of which passed the four-fifths rule in comparison to both the majority and minority classes in their feature.

Variable	Independence (minority versus NA class)
Gender	0.000000
Disability	0.960613
Veteran status	0.975141

Table 6: Independence scores between minority and NA groups for gender, disability, and veteran status

Variable	Independence (majority versus NA class)
Gender	0.000000
Disability	1.000000
Veteran status	1.009134

Table 7: Independence scores between majority and NA groups for gender, disability, and veteran status

Recommendations

1. Model Design

   First, it is crucial that Providence Analytica fix the bug that denies an interview automatically when an applicant's gender is not provided. This flaw unfairly discriminates against applicants who choose not to provide their gender or whose identities do not fit into the male/female binary. If this is not a bug and instead a legitimate output from the model, it needs to be immediately addressed as an extreme case of discrimination.

   The incorporation of fairness measures for gender and ethnicity (particularly Native Americans) are also recommended. The apparent discrimination against female and Native American applicants can be mitigated through preprocessing techniques such as a disparate impact remover, which would limit the model's ability to discriminate by creating a common distribution between the gender and ethnicity groups. Furthermore, a prejudice remover regularizer could be incorporated in the algorithm's cost function to optimize fair performance. Additionally, a post-processing strategy like Reject Option based Classification (ROC), which defines a threshold and swaps the classifications of male and female applicants who score near that threshold, could prove beneficial here as well. This would allow Bold Bank to continue interviewing high quality applicants while simultaneously removing some of the model's inherent bias. We recommend that Providence Analytica test the implementation of fairness measures to protect non-male and Native American applicants and use the four-fifths rule to comply with fairness regulations.

2. Company Practices

   Earlier in this audit process, Bold Bank reported that they have "implemented rigorous oversight mechanisms and conducted comprehensive evaluations" to affirm that their procedures are in full compliance with the Equal Opportunity Employment Commission. Given that this audit reveals significant discrimination against women, Native Americans, and applicants with unknown genders, it is evident that these mechanisms are not rigorous enough. Thus, we recommend that Bold Bank implement stronger oversight procedures that specifically calculate fairness metrics across each group for every sensitive attribute. This will ensure that Bold Bank is in compliance with federal regulations and has access to the best talent, that applicants no matter their demographic background have a fair chance at employment, and that Providence Analytica is held accountable and made aware of any concerns in their algorithm so they can develop better-performing technologies.

Citations

[1] Talbott, Tyler. "Average College GPA by Major - 2024." *College Transitions*, 29 Jan. 2024, www.collegetransitions.com/blog/average-college-gpa-by-major/.

[2] "Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures under Title VII of the Civil Rights Act of 1964." *US EEOC*, www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial. Accessed 8 May 2024.

Appendix

prediction Disability	Not Interviewed	Interviewed
No	0.930618	0.069382
Yes	0.927774	0.072226
N/A	0.930618	0.069382

Appendix A: Proportions of each disability status interviewed versus not

prediction Ethnicity	Not Interviewed	Interviewed
White	0.928571	0.071429
Black	0.931818	0.068182
Native American	0.913965	0.086035
Asian American & Pacific Islander	0.932611	0.067389
Other	0.941250	0.058750

Appendix B: Proportions of each ethnicity interviewed versus not

prediction Work authorization	Not Interviewed	Interviewed
No	0.927448	0.072552
Yes	0.931818	0.068182

Appendix C: Proportions of each work authorization status interviewed versus not

prediction	Not Interviewed	Interviewed
Veteran status		
No	0.930665	0.069335
Yes	0.928248	0.071752
N/A	0.930031	0.069969

Appendix D: Proportions of each veteran status interviewed versus not