CSCI 1951Z Final Audit Report

Introduction

Bold Bank, an ambitious mid-sized bank, recently adopted a new hiring system designed by Providence Analytica, a software company working on designing state-of-the-art AI products for companies in various industries, to enhance its hiring process.

Recently, the Equal Employment Opportunity Commission (EEOC) has received complaints alleging that this new recruitment system may be discriminatory. According to the EEOC's mandate, it is crucial to enforce the federal laws that make discrimination against a job applicant because of the person's sensitive information such as gender, ethnicity, and disability illegal. As a group of special assistants for the EEOC, our task is thus to conduct a thorough audit of Bold Bank's new hiring system in order to determine whether it complies with the guidelines and regulations of EEOC.

This hiring system consists of a resume scoring model and a candidate evaluation model. The resume scoring model takes the candidates' resumes in CSV format as input, and assigns a score from 0 to 10 to each resume, indicating whether the candidate is suitable for the role. The candidate evaluation model then takes the resumes, along with the resume scores (output of resume scoring model) as input, and returns a binary outcome (0 or 1) to indicate whether the candidate should receive an interview.

EEOC has provided us the opportunity to conduct interviews with one representative from each stakeholder group, namely, job applicants, Providence Analytica, and Bold Bank, to learn more about the context. In particular, we interviewed Brianna Brown, a new graduate from Central University applying to be a financial analyst at Bold Bank, a representative from Providence Analytica, and a representative from Bold Bank.

According to Brianna's answers to our questions, she complained that she felt discriminated against by the part of her application that asked her to disclose whether she needed work authorization, especially after knowing that an AI model will be used to screen resumes. In addition, we also got the information from her that she was not only allowed, but even "encouraged", to disclose sensitive information such as ethnicity, gender, and disability status. Such encouragement seems dubious to us although Bold Bank claimed in the application that such information was only for statistical purposes.

We have found other sources from which potential discrimination could arise in the answers of both the representatives from Providence Analytica and Bold Bank to our questions. For example, the representative of Providence Analytica claimed that they have incorporated the values and priorities of Bold Bank into the models they built. However, we noticed that the representative said that if a company values diversity and inclusion, the algorithm may be designed to give higher scores to candidates from diverse backgrounds. This seemed to be discriminatory to us in the first place because the definition of "diverse" can vary significantly, and giving different scores to candidates based on their backgrounds does sound like discrimination. Moreover, if this is indeed what Bold Bank values, then this could be an internal problematic bias present inside the company that does not comply with what the EEOC, and the federal laws, require.

Nevertheless, our investigation would be largely untenable and unjustifiable if we relied solely on these answers, representative of the true situation or not, from our interviews. It is essential for us to actually look into the models deployed to substantiate the discriminations suspected or exonerate the models from potential violations of federal requirements.

Methodology

1. Original Data Source

The original dataset was collected randomly in a simulated environment for querying the API. It consists of 3,000 entries, each representing an individual job applicant.

The original dataset was generated with an intention to maintain a roughly even distribution across various attributes, including gender, work authorization status, veteran status, disability status, etc., thereby ensuring no inherent biases in attribute distribution. The balanced nature of the dataset was intentional to eliminate any pre-existing biases and focus on assessing the model's fairness in handling diverse applicant profiles.

Thus, all the features were generated from a list of values randomly with equal probability of appearing, with possible values shown below:

Each individual's school is randomly chosen from this list: ["Brown University", "Providence University", "Providence College", "University of Rhode Island", "Providence State University", "State Providence College", "Providence School", "Bryant University", "Boston College", "Boston University", "Northeastern University", "Columbia University", "Harvard University", "MIT", "UC Berkeley", "UCLA", "UCSB"]

Each individual's GPA is a random number (rounded to two decimal places) from the list of 100 evenly spaced numbers between 1 and 4.

Each individual's degree is randomly assigned as one of "Bachelors", "Masters", and "PhD".

Each individual's location is randomly chosen from this list: ["Boston", "Providence", "Los Angeles", "Philadelphia", "Miami", "Chicago", "New York", "San Francisco", "Las Vegas", "Washington"].

Each individual's gender is randomly assigned as "M" or "F" or "N/A", where "M" indicates male, "F" indicates female, and "N/A" means the information was not provided.

Each individual's veteran status is randomly assigned as 0 or 1 or "N/A", where 0 indicates not a veteran and 1 indicates a veteran. "N/A" means the information was not provided.

Each individual's work authorization status is randomly assigned as either 0 or 1, where 1 means the individual has work authorization and 0 means the individual does not.

Each individual's disability status is randomly assigned as 0 or 1 or "N/A", where 1 means the individual has a disability and 0 means the individual does not. "N/A" means the information was not provided.

Each individual's ethnicity is randomly chosen from [0, 1, 2, 3, 4].
The following correspondence exists between these numbers and actual ethnicity groups:

0: White	1: Black	2: Native Americans
3: Asian Americans & Pacific Islanders	4: Other	

To make the individuals' applications not easily distinguishable by their role history, we decided not to leave any of the three most recent jobs blank.

The list of all possible jobs is: ["Software Engineer", "Machine Learning Engineer", "Data Scientist", "Hardware Engineer", "Data Analyst", "Research Scientist", "Quantitative Analyst", "Software Tester"].

The three most recent job roles were chosen randomly with replacement from the above list. The start and end dates of each role were randomly generated but we have made sure that the end dates are always at least one month after the corresponding start dates.

1.1 Evaluation Criteria for Original Data

The primary criterion for assessing the algorithm was its fairness across different demographic groups. The fairness metrics chosen for this analysis included the Disparate Impact (DI) measure and ANOVA, which is pivotal for evaluating whether the outcomes of a predictive model disproportionately favor or disadvantage any specific group. Attributes (i.e., sensitive information) investigated for DI analysis were: Gender, Work Authorization, Veteran Status, Disability, and Ethnicity. We have added School Name and Degree to this list, for ANOVA analysis. These attributes were selected based on their common consideration in anti-discrimination law and policies related to employment. We also looked at the correlation coefficients between the predictions and the features.

1.2 Analyses on Original Data

1.2.1 Disparate Impact Analysis

Disparate Impact analysis was conducted to evaluate whether the predicted outcomes (i.e., binary outcomes from the candidate evaluation model indicating whether the candidate should receive an interview) disproportionately favored or disadvantaged any specific group based on sensitive attributes.

DI is calculated as the ratio of the probability of a positive outcome (i.e., receiving an interview) for the protected group to that of the non-protected group. A DI value less than 1 indicates potential discrimination against the protected group, while a value greater than 1 means a bias in favor of the protected group. Values close to 1 indicate a fairer system.

In the case that the attribute is not binary, such as Gender and Ethnicity, one group is selected as the protected group and all the other groups are combined to form the unprotected group. Below are the results:

Gender: If the non-male individuals are set as the protected group, then the DI is 0.322. This indicates that non-male individuals are significantly less likely than males to receive a positive prediction (i.e., an interview).

Work Authorization: For work authorization, if the group of individuals that do not have work authorization is set as the protected group, then the DI is 0.94. This suggests a bias favoring individuals with work authorization.

Veteran Status: The veteran status shows a DI of 1.023 if veterans are the protected group. This indicates a very small bias towards veterans.

Disability: If the individuals with disabilities are the protected group, then the DI is 0.945. This suggests a disadvantage in receiving positive predictions for individuals with disability compared to those without disabilities.

Ethnicity: The DI for ethnicity stands at 0.757 if letting the group of non-white individuals (ethnicity≠0) be the protected group. This indicates that White people are slightly favored towards the positive outcome (i.e., receiving an interview) than people with other ethnicities.

The conclusions are:
(1) The predictions on the original dataset did have significant biases for males in terms of receiving interviews. This is also shown by the fact that 296 out of 669 males in the dataset received an interview but only 201 out of 1834 non-male individuals received an interview.

(2) For sensitive features Work Authorization, Veteran Status, and Disability, no conspicuous biases were found in terms of different groups' receiving interviews.
(3) For Ethnicity, the predictions did show a slight bias towards white people because the DI is less than 0.8, which means some biases favoring white people. We also examined the DI values when other groups are chosen to be the protected group, as shown below:

Protected Group	Eth_1	Eth_2	Eth_3	Eth_4
DI	0.913	1.024	0.904	0.875

The results show that the scores are all above the threshold of 0.8 and thus, no significant biases towards these groups are identified.

1.2.2 ANOVA Analysis

ANOVA (Analysis of Variance) tests were carried out to determine if there are statistically significant differences in the predicted outcomes across different groups defined by sensitive attributes. The attributes and their respective groups are described below along with the ANOVA results:

Attributes	Degree	Veteran Status	School Name	Work Authorization	Disability Status	Ethnicity
Statistic	2.854	0.327	1.547	0.562	0.185	2.355
P-Value	0.058	0.567	0.081	0.454	0.667	0.052

For attributes Degree and Ethnicity, the p-values are slightly above the conventional significance indicator of 0.05, suggesting no strong evidence of statistically significant differences in predicted outcomes among different degree levels. However, the p-values being very close to 0.05 might warrant a closer look or a study with a larger sample size to confirm these findings. For other features, the p-values indicate that there are no significant differences between the predicted outcome across groups, suggesting that they may not be influencing the predictive outcomes a lot.

1.2.3 Correlations (Original Data)

In this section, we looked at the correlations between the binary predictions and the selected features. The correlation values are shown as Plot 1 in Appendix.

The plot shows that apart from the males group, the gender-not-provided (N/A) group, and the White ethnicity group, all the other groups, and thereby the features, have low correlation with the outcome. Thus, this may indicate that the biases towards/against these groups are not existing or minimal.

1.2.4 Limitations of Original Data

- Lack of Ground Truth

This original dataset only contains the predicted outcomes (i.e., predictions) so there are no ground truth labels for whether the individuals actually received an interview or not. This absence of true labels prevented us from conducting comprehensive fairness evaluations using metrics that require the knowledge of ground truth labels for computation such as Equal Opportunity Difference (EOD) and Average (Absolute) Odds Difference (A(A)OD). This limitation significantly restricted the analyses that can be conducted, resulting in only a partial view of the models' fairness.

- Data Coverage

This original dataset only contains simulated (i.e., not real) data. This synthetic nature made this dataset not able to represent the full complexity of real-world demographics. As a result, this dataset may lack nuances and variations present in real-world applicant pools, limiting its applicability to practical auditing scenarios.

2. Biased Data Source:

Since in the original dataset, the different values of all the features are distributed roughly evenly, it is impossible to know the effect of uneven distributions on the models' biases using this original dataset. However, uneven distributions can frequently occur in the real world. Thus, we decided to modify the original dataset to make some distributions uneven and compare the results. The evaluation criteria used here is Disparate Impact (DI).

2.1 Making Gender Distribution Uneven

In this section, we only modified the original dataset's Gender column to make one gender be the majority/unprotected group and the other two combined be the minority/protected group. The Male-Biased dataset was created by randomly setting originally non-male individuals' gender to "M" to make the total number of male individuals in the dataset 2700, which is 90% the total number of individuals. The Female-Biased dataset and the NA-Biased dataset were created in a similar way.

The DI's for the male-biased, female-biased, and NA-biased datasets are:

Bias Added	Male 90%	Female 90%	N/A 90%
Disparate Impact	0.400	0.732	0

This means:
(1) When 90% applicants are males, the males are favored a lot for getting an interview.
(2) When 90% applicants are females, the females are not favored as the males are in (1).
(3) When 90% applicants do not report their gender, none of them get an interview.

These show that the models were significantly biased towards males for giving interviews when the majority of applicants are males. In contrast, when the majority of applicants are females, the models were not favoring them as much as males were favored in the former male-dominant case. Also, the models were significantly disadvantageous to applicants who did not report the gender when the majority of applicants do not report it.

2.2 Making Work Authorization Distribution Uneven

The work-authorization-biased datasets were created in a similar way by randomly setting individuals' work authorization status to 0 or 1 to ultimately make 90% of individuals have or do not have work authorization respectively. All other features were kept intact.

The results are:

Bias Added	Have 90%	Not Have 90%
Disparate Impact	0.776	0.814

This shows that models were slightly favoring applicants with work authorization when the majority of applicants have a work authorization. If the majority of applicants do not have a work authorization, then the models tend to treat the two groups roughly equally with no or minimal bias towards/against any group.

2.3 Making Veteran Status Distribution Uneven

The veteran-status-biased datasets were created in a similar way by randomly setting individuals' veteran status to 0, 1, or N/A to ultimately make 90% of individuals non-veteran, veteran or no-information-provided respectively. All other features were kept intact.

The results are:

Bias Added	Veteran 90%	Non-Veteran 90%	N/A 90%
Disparate Impact	1.098	0.931	1.002

These DI values do not show conspicuous biases towards/against veterans, non-veterans, and those who did not report their veteran status. This could suggest that the models might have treated the individuals with different veteran statuses equally or with minimal biases.

2.4 Making Disability Status Distribution Uneven

The disability-status-biased datasets were created in a similar way by randomly setting individuals' disability status to 0, 1, or N/A to ultimately make 90% of individuals do not have disabilities, have disabilities, or no-information-provided respectively. All other features were kept intact.

The results are:

Bias Added	Disability 90%	No-Disability 90%	N/A 90%
Disparate Impact	1.027	1.032	0.958

This also shows that there do not exist significant biases in Disability Status and the models might have treated applicants with different disability statuses roughly equally.

2.5 Making Ethnicity Distribution Uneven

The ethnicity-biased datasets were created in a similar way by randomly setting individuals' ethnicity to the ethnicity we are trying to add bias to. The ultimate goal was to make 90% of individuals in each dataset have the same ethnicity. All other features were kept intact.

The results are:

Bias Added	Eth_0 90%	Eth_1 90%	Eth_2 90%	Eth_3 90%	Eth_4 90%
Disparate Impact	1.010	1.140	0.912	0.920	1.092

This is showing that when Ethnicity Group 1 (Black people) is the majority group, then the models tend to favor them a bit. For all other ethnicity groups, the DI values do not show significant biases towards/against them.

2.6 Limitations of Biased Data

- The threshold set as 90% to represent a majority group may not be representative in all cases and in the real world, the actual situation might be much more complex.
- These biased datasets were still built on the original one so might inherit internal biases from the original dataset that can influence predictions.

Recomendations

1. Model Design

- Review Gender Bias

The models showed significant bias towards males, particularly when the gender distribution was highly skewed. It is crucial to reevaluate how gender information is being used in the scoring model and consider anonymizing this attribute or employing techniques like gender-neutral resume processing.

- Address Ethnic Bias

While the ANOVA analysis indicated only minor biases, the Disparate Impact (DI) analysis revealed potential biases favoring specific ethnicities. Further investigation and targeted improvements in data processing, anonymization, or reweighting of certain ethnic groups can help mitigate these biases.

- Incorporate Feature Engineering for Sensitive Attributes

Implement feature engineering to handle sensitive attributes more effectively. For instance, consider grouping similar job titles or normalizing school names to prevent direct or indirect biases associated with them.

2. Company Practices

- Refine Usage of Model Outputs

Make sure that model outputs like resume scores and interview recommendations are used as one part of a holistic decision-making process, rather than as the sole determining factor.

- Monitor Regularly for Bias

Establish regular audits and analyses of the model's outcomes to identify and correct biases in practice. This may involve setting up a dedicated team to track disparities across sensitive attributes and respond promptly.

- Maintain Transparency

Ensure transparency around how the models are used in decision-making, and clearly communicate the purpose and limitations of model usage to applicants, stakeholders, and internal teams.

Appendix

Plot 1