Final Project: Fairness in Bold Bank's Hiring Practices

CSCI1951Z: Fairness in Automated Decision Making

Lyatte Liu, Ziao Zhang
Email: yutang_liu@brown.edu or ziao_zhang@brown.edu
Taught by Prof. Suresh Venkatasubramanian

Spring 2024

Contents

0 Introduction

As auditors appointed by the Equal Employment Opportunity Commission (EEOC), our team investigated Bold Bank's hiring practices of a newly implemented hiring system developed by Providence Analytica. In our investigation process, we conducted two rounds of interviews with Bold Bank, Providence Analytica, and job applicant Brianna Brown, and thoroughly tested and analyzed the resume scorer and evaluator models. We have concluded that the current model should raise fairness and performance concerns for all stakeholders. For detailed implementation, please refer to https://github.com/JackeyLove36/CSCI1951Z-Fairness.

1 Methodology

1.1 Data Source

Since the auditing team does not have access to the original training or testing data, we have designed a resume generator to test the given scorer and evaluator model. Given a number of samples, the generator will randomly generate resume data with the following attributes:

Explanation 1.1

The attributes are as following:

- **School name and location**: randomly selected from databases of university names and city names
- **GPA**: a random floating point number between 2.0 and 4.0
- **Degree**: randomly chosen from "Bachelors", "Masters", "Phd"
- **Gender**: a string randomly selected from "M", "F", and "N/A"
- **Veteran status and Disability**: randomly selected from 0, 1, and "N/A"
- **Work authorization**: randomly chosen from 0 and 1
- **Ethnicity**: randomly selected from 0, 1, 2, 3, 4 (white, black, native American, Asian American and Pacific Islanders, and others)
- **Role 1-3, Start 1-3, and End 1-3**: the previous work experiences of the candidate. The generator first randomly determines the number of experiences of the candidate (1, 2, or 3), and fills the role name(s) randomly from a list of role names. The start date is selected randomly from a date between 2014-1-1 and 2022-1-1. The end date is between 60 days and 740 days after the start date.

We generated a dataset of 4000 thousand samples **(Dataset One)**. To explore further the decision-making process of the models, we have also tested two modified datasets in addition to the random dataset. We identified two types of variables, quantifiable (GPA, degree, gender, experience length, etc) and non-quantifiable (location, school name, role name, etc). The former is either a categorical or numerical value; the latter is a string that we had been told would be evaluated by NLP techniques. To single out the effect of the non-quantifiable variables, we took the original random dataset, and defaulted the school name, location, and role 1 to the same value (”Providence College”, ”Providence”, and ”Data Scientist”, respectively), and experience 2 and 3 to ”N/A” **(Dataset Two)**.

The purpose of the other dataset **(Dataset Three)** was to explore the relationship between the type of experience the candidates have **(”Role”)** and the resume scores and final prediction. We wrote four sets of roles, categorized by their relevance to the job description of a financial analyst Bold Bank has provided:

Explanation 1.2

Four Sets of Roles:

- Highly relevant: ”Investment Banker”, ”Financial Analyst”, · · ·
- Somewhat relevant: ”Software Engineer”, ”Researcher”, · · ·
- Somewhat irrelevant: ”Sales Representative”, ”Social Media Coordinator”,· · ·
- Highly irrelevant: ”Pharmacist”, ”Multimedia Artist”, ”Dietitian”, · · ·

For this last dataset, we took the original random dataset and substituted **”Role 1”** to one of the roles from the above sets. Based on a randomly selected relevance category, the resume generator randomly chooses a role name. It keeps track of the category for future analysis.

1.2 Evaluation Criteria

Question 1.3

Our evaluation process mainly asks two questions:

1. Do the resume scorer and evaluator discriminate upon any attributes, including gender, ethnicity, nationality, disability status, etc?
2. Do the resume scorer and evaluator make sensible recruiting decisions that correspond to Bold Bank’s hiring priorities to assess the candidates’ skills, experiences, and qualifications?

If both questions have positive answers, we consider Bold Bank’s hiring process and Providence Analytica’s models to be fair and accurate.

1.3 Analysis Techniques

Explanation 1.4

We have used a variety of analysis techniques to determine whether the above criteria are met:

- We calculated and visualized the distribution of the resume scores and predictions of **Dataset One** by each demographic attribute (gender, ethnicity, veteran status, work authorization, disability)

- We used logistic regression to explore the effect of each demographic attribute on the resume scores and predictions.
- We derived the fairness metric through Independence in evaluating the effect of classification attributes in resume scores and predictions.
- Controlling all non-quantifiable attributes, we calculated and visualized the distribution of the resume scores and predictions of **Dataset Two** by each categorical attribute.
- Controlling all non-quantifiable attributes, we used logistic regression to explore the effect of each demographic attribute on the resume scores and predictions **(Dataset Two)**.
- Controlling all non-quantifiable attributes, we used logistic regression and visualizations to explore the effect of a candidate's GPA, level of degree, and the length of their experience on the resume score and final prediction **(Dataset Two)**.
- We used logistic regression to explore the effect of a role's relevance on the resume score and the final prediction, controlling the number of experiences **(Dataset Three)**. We also calculated and visualized the distribution of the four relevance categories.

1.4 Limitations

The auditing team did not have access to the resume scorer and evaluator model, the training data, or the testing data, and thus had limited resources and information as to the architecture and training process. We were also unaware of the model's previous performance in terms of accuracy and fairness. Furthermore, the team at first experienced some difficulties with the provided API and was only able to test with 4000 samples. Therefore, **our conclusion is subject to further investigation if more resources or information become available in the future.**

2 Findings

2.1 Conclusion to Discriminatory behavior of the resume scorer

Conclusion 2.1

Claim One: **The Resume Scorer does not exhibit Discriminatory Behavior against any Demographic Group.**

2.1.1 Evidence from Distribution of resume scores

We visualized the distribution of resume scores by the attributes using violin plots and box plots. Comparing the median and quartile values, we do not see a significant discrepancy among different groups for all attributes, including gender, ethnicity, veteran status, disability, and work authorization. Take **Gender** as an example. The medium values for male, female, and non-disclosed-gendered candidates are all around 5.0, the 25th quartile around 2.8, and the 75th quartile around 7.7 **(Figure 1)**.

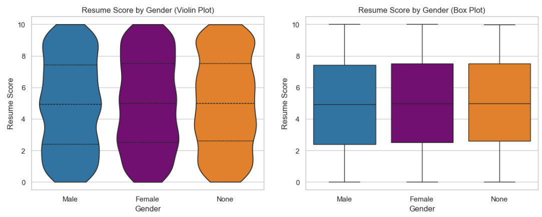

Figure 1: Distribution of Resume Scores by Gender through Violin Figure and Box Plot

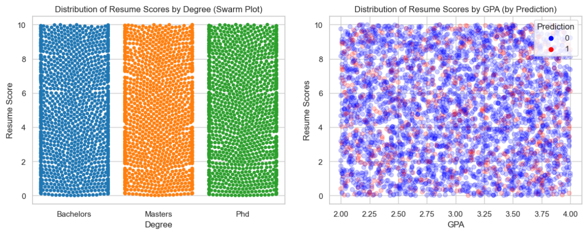

Figure 2: Distribution of Resume Scores to Degree (Left) and GPA by predictions (Right)

The same applies when we control all non-quantifiable attributes (location, school name, experience name, and number) **(Figure 1)**. Moreover, in both datasets, the distributions of resume scores by ethnicity, veteran status, work authorization, and disability are similar, meaning that the scorer is not biased toward any group. For example, **Figure 2** illustrate the relationship between Resume Scores to Degree and GPA respectively (The patterns are relatively the same, which corresponding to all Claim One)

2.1.2 Evidence from Statistical Test

We used Statistical Test (OLS) to evaluate further in quantitative whether the unfairness exists in resume score. After successfully implemented OLS Regression Results, we find that p-value is 0.939 > 0.05, which reflects that **Gender** is not statistical important (Corresponding to Claim One).

2.2 Conclusion to Discriminatory behavior of the evaluator

Conclusion 2.2

Claim Two: **The Evaluator Model discriminates against Female Candidates and Candidates of non-disclosed gender**.

2.2.1 Evidence from Distribution of predictions

We calculated the selection rate of each gender in **Dataset One** and **Dataset Two (Table 1)**, and plotted the distribution of predictions in both datasets **(Figure 3)**. Most tellingly, the selection rate of non-disclosed gender for both datasets is 0, meaning that all candidates of non-disclosed gender are defaulted to a negative label by the evaluator. In addition, the distribution is significantly discrepant between male and female candidates. In **Dataset One**, the selection rate of male candidates is 0.334, over 10 percent higher than that of females (0.218). The gap is even larger (0.620 and 0.370) in dataset 2 when the non-quantifiable attributes are controlled.

Dataset	F	M	N/A
1	0.218	0.334	0.000
2	0.370	0.620	0.000

Table 1: Selection rate by gender

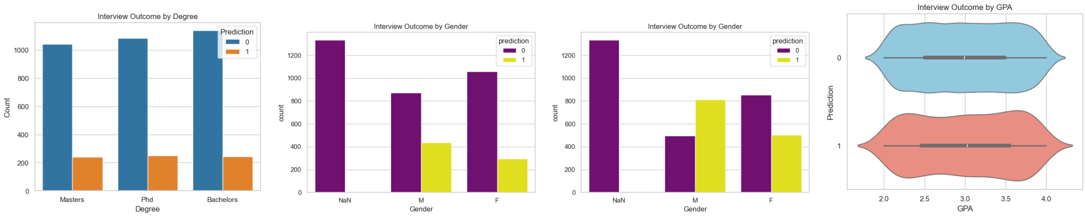

Figure 3: Relationships between Predictions and Degree, GPA and Gender **Dataset One** (Left), **Dataset Two** (Right)

Except for gender, the evaluator does not exhibit discriminatory behavior towards other demographic groups based on distribution.

2.2.2 Evidence from Logistic Regression Results

We fitted logistic regression models to explore the combined effect of the attributes and resume scores on the final prediction, the result from which also signifies discrimination against female candidates. In the completely random dataset **(Dataset One)**, the coefficient of the male gender (C(Gender)[T.M]) is 0.5678, with a statistically significant p-value of 0.001, meaning that being male increases the log odds of receiving an interview by 0.5678. Similarly, in Dataset 2, the coefficient of the male gender is 1.1001 with a statistically significant p-value of 0. Other than gender, we do not see other forms of discriminatory behavior toward other groups based on the logistic regression models.

2.2.3 Evidence from Independence

Our team also implemented the Independence Fairness Metric in Evaluator Models. We treated Males in the denominator and Females in the numerator. With the successful implementation in Independence, we derived 0.32780869547587915 $>$ 0, suggesting potential biases between two groups (Corresponds to our Claim Two).

2.3 The Resume Scorer and the Evaluator's Correspondence to Bold Bank's Hiring Priorities

Our investigation suggests that **the resume scorer and the evaluator model fail to carry out Bold Bank's hiring priorities.** During our interview with Bold Bank, they expressed that they value qualified and experienced candidates suitable for a financial analyst role. We used visualizations, linear and logistic regression models to explore the effect of GPA, degree, experience length, and experience type on the model's decisions, but none showed a distinct relationship between these attributes and the model outcome.

2.3.1 Effect of Degree and Role Type

In Dataset 2, controlling all non-quantifiable attributes, we visualized the score and prediction distribution of candidates with Bachelor's, Master's, and PhD degrees **(Figure 3)**. All degree levels have similar score distributions and selection rates (around 0.33). The regression models also show minimal influence of the degree levels on the resume scores and final predictions. Though the Bold Bank representative has indicated that the role prefers qualifications such as a master's degree in a finance-related field than a bachelor's, the model does not seem to incorporate this consideration (see Appendix 4.1 for details). Similarly, we do not see a clear relation between the relevance of experience and the model outcomes. All categories of roles are equally distributed with regard to score and prediction **(Figure 4)**. Although the company would prefer finance-related and data-related experiences, both models fail to reflect this priority in their decision-making process.

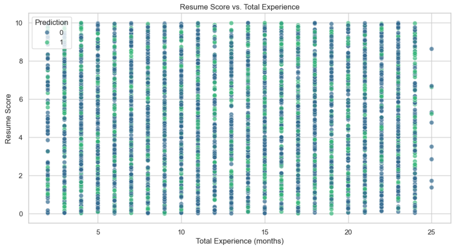

Figure 4: Relationship between Resume Scores and Total Experience

2.3.2 Effect of GPA and Length of Experience

We have also used scatter plots to visualize the effect of GPA and length of experience on the scores and evaluations. Similarly, we do not notice any patterns that show how the models process GPA and experience length, which are traditionally used by companies to determine candidates' qualifications **(Figure 2, 3)**.

2.3.3 Candidate Complaints

We interviewed a recent job candidate Brianna Brown, who felt discriminated against during the screening process. She specifically stated that she is a female international student requiring visa sponsorship and that she might be rejected because of her work authorization status even when being more qualified than others (see Appendix 4.2 for details). As we did not see any biases in terms of work authorization, we suspect that the model's failure to identify candidates' qualifications could have led to the job applicant's complaint.

3 Recommendations

3.1 Model Design

Based on the above findings, we come to two major conclusions: first, the evaluator model is biased against marginalized genders; second, both the scorer and the evaluator models fail to assess candidates fairly based on their qualifications.

Recommendations 3.1. We recommend the following action items for Providence Analytica:

1. Include candidates of non-disclosed gender or non-binary genders in the training and testing data to produce fair and accurate results for these historically marginalized groups. If they are already included, make sure there are enough

2. Identify and mitigate biases against female candidates in the evaluator's predictions. The bias might stem from an imbalanced distribution of training data and the lack of female candidates with positive labels.

3. Investigate the scorer's and the evaluator's decision-making process, especially regarding attributes that relate to a candidate's qualifications, such as GPA, degree level, experience name, experience number, and experience length. Make sure that a more experienced and qualified candidate translates into a higher resume score and a better chance of a positive label.

3.2 Company Practices

Currently, the screening decisions generated by the evaluator model are used directly without human intervention. Given the findings that the current automated decision-making process involves significant bias and inaccuracy, we recommend that the company establish a designated team of specialists who will monitor the fairness and accuracy of the automated hiring decisions. For instance, regularly perform a check on a randomly selected set of candidates. From the auditing team's perspective, it is in the company's best interest that the new hires are diverse and best qualified for the positions, as it strengthens the company in their future development and expansion. In addition, more just and transparent hiring criteria can motivate the candidates and match them with their best fit. Similarly, any findings and issues on the machine learning models reported will help Providence Analytica perfect its product. In conclusion, we suggest that Bold Bank adjust its hiring practices during the automated resume screening process.

4 Appendix

4.1 Financial Analyst's job description provided by Bold Bank

Job Description: Bold Bank is seeking a highly motivated and detail-oriented individual to join our Finance team as a Financial Analyst. In this role, you will be responsible for analyzing financial data, preparing reports, and providing insights to support strategic decision-making within the organization. The ideal candidate will have strong analytical skills, a solid understanding of financial principles, and the ability to thrive in a fast-paced environment.

Qualifications:

- Bachelor's degree in Finance, Accounting, Economics, or a related field.
- Strong analytical and quantitative skills.
- Proficiency in Microsoft Excel.
- Good communication skills.
- Ability to work effectively in a team environment.
- Internship or relevant work experience in finance or banking is a plus.

Preferred Qualifications:

- Master's degree in Finance, Accounting, or Business Administration.
- Certification such as CFA (Chartered Financial Analyst) or CPA (Certified Public Accountant).
- Familiarity with financial analysis software (e.g., Bloomberg, SAP).

4.2 Job Applicant Brianna Brown's complaint

Q (Auditing team): Is there any particular piece of information that you feel uncomfortable submitting? Is there any particular piece of information that the recruiter might have used to discriminate against you?
A (Job Applicant): I felt uncomfortable submitting information about my nationality and work authorization status, as I believe it might have been used against me in the application process due to the company's preference for candidates with existing work authorization.

Q: Have you heard back from the recruiters about your application status? (rejected/proceed to the interview/etc.) If so, how do you feel about this result?
A: I have heard back, and I was rejected. I believe that I was discriminated against based on my work authorization status. I'm an international student and require visa sponsorship. There are other students with fewer qualifications who were hired. I also was made aware that an AI model is applied to screen my resume as part of their hiring process, but I'm unsure what the general criteria and skills the algorithm is designed to assess.