Audit Report - Github

sarah_prakriti_peters@brown.edu, you_jung_koo@brown.edu

INTRODUCTION

Bold Bank, a financial institution, recently initiated a restructuring of its hiring process. The new recruitment software was implemented by Providence Analytica - which consists of a resume scorer and a candidate evaluator. The resume scorer is responsible for assessing candidates' suitability for specific roles, while the evaluator determines whether a candidate should proceed to the interview stage. This audit was conducted by the Equal Opportunity Commission (EEOC).

The purpose of this audit is to evaluate the efficacy of the new recruitment process and to identify any inherent limitations or biases. Our primary objectives include examining how Providence Analytica's system addresses issues such as missing data or biases caused by sensitive attributes such as race and gender. Additionally, we seek to ascertain how candidate resumes are evaluated, and how candidates are selected for the interview process, focussing on whether certain demographic groups receive disproportionately higher scores.

METHODOLOGY

Data Source

The dataset consists of a mix of categorical and quantitative variables. Categorical variables including locations, job titles, degrees, college names, genders, veteran statuses, work authorization statuses, disability statuses, and ethnicities, were initially constructed as lists. Subsequently, the dataset values were randomly sampled from these lists. Quantitative variables were defined within specific ranges - for example, GPA scores were constrained between 0.0 and 4.0, and then randomly sampled. Throughout the data generation process, our objective was to ensure equitable representation across all subgroups. We ensured this by using a uniform probability distribution across all subgroups, which means that every value has an equal probability of being picked.

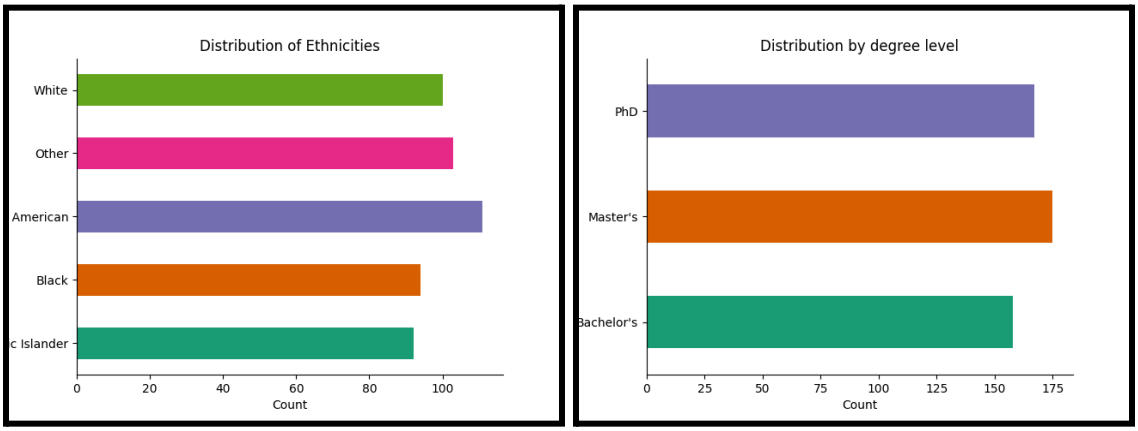

***Fig 1**:* Figure 1 shows us the distribution of candidates by their ethnicities and degree level. As we can see, all groups are mostly equally distributed. This allows us to

compare the results of the resume scorer and candidate evaluator models over various demographic data points.

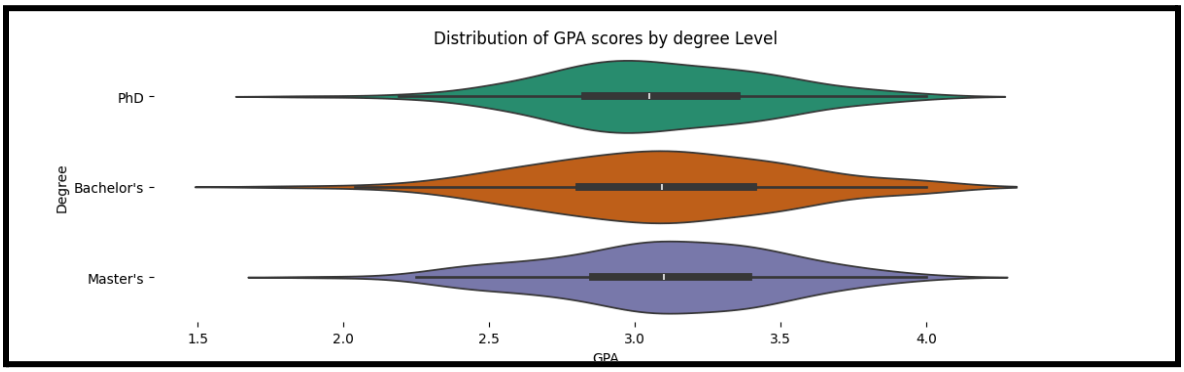

Fig 2: Figure 2 shows us that our distribution of GPA scores remains similar across degree levels. This allows us to compare the results of both models across different degree titles, versus the candidate GPA scores alone.

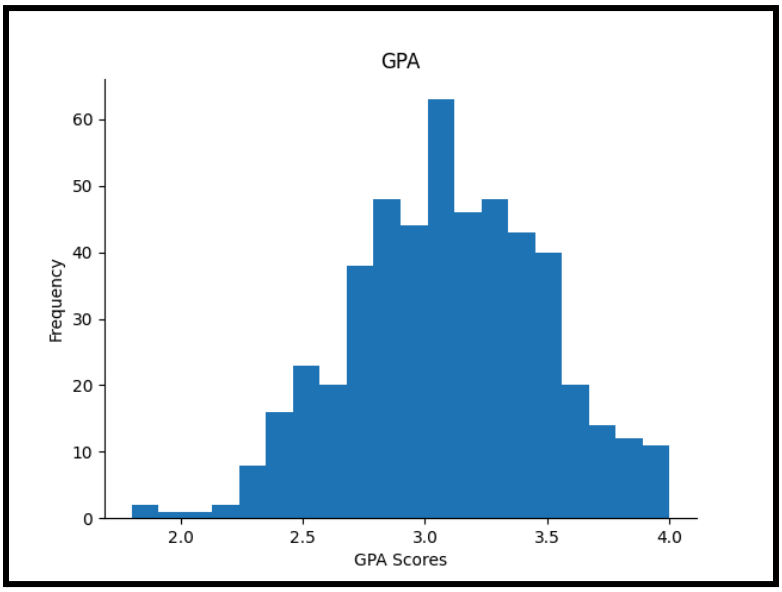

Fig 3*:* Based on the average college GPA in the United States, we created a distribution for GPA scores with a mean of 3.1 and a standard deviation of 0.4. This results in a normal distribution of scores.

We also generated synthetic date data by ensuring the start date is always earlier than the end date. The data generator uses many parameters, one being the number of records. Using this parameter, we generated multiple datasets with 25 to 4,000 instances to test both models.

Evaluation Criteria

While evaluating the results of the two models, we decided to focus on two key metrics - reproducibility and discrimination. We define reproducibility as obtaining consistent results for the same instance over several iterations. While running our generated data through the resume scorer, we came across large fluctuations in the resume scores. This led us to believe that the

probability distribution for resume scores may be randomly sampled. Furthermore, we wanted to evaluate the reproducibility of the candidate evaluator. Essentially, we analyzed if the same instance received interview callbacks regardless of their resume score. While we noticed some trends, we are unable to truly assess the reproducibility of both models.

We define discrimination as favoring a certain group over another. An applicant's dataset consists of several sensitive attributes such as gender, race, disability, veteran status, resume gaps, and GPA scores. Even if a candidate is qualified for the position, they may not receive a call back if they refuse to disclose certain information such as gender. To quantify discrimination, we utilized **Disparate Impact** and **Statistical Parity Difference** scores.

To calculate these scores, we used gender and ethnicity as sensitive attributes. Under gender, male candidates formed the majority group and female candidates formed the minority group. Under ethnicity, we tried various combinations of sub-groups. We only noticed a significant effect in the male versus female experiment. Both scores indicate that females are disadvantaged, and do not receive as many interview calls as males. These scores are illustrated in the table below.

	Majority Group	Minority Group	Disparate Impact	Statistical Parity Difference
0	Male	Female	0.63	-0.23
1	White	Black	0.95	-0.01
2	White	Asian American & Pacific Islander	0.96	-0.01
3	White	Native American	0.94	-0.01
4	White	Other	1.02	0.00
5	Asian American & Pacific Islander	Black	0.99	0.01

Analysis Techniques

With such a diverse dataset, our analysis approach was based on testing group-wise limitations. To precisely determine if specific features influenced any variations in resume scores and candidate evaluations, we maintained consistency across all variables except one. For example, we ensured uniformity in all features while only modifying the candidate's ethnicity. This approach allowed us to evaluate any potential bias directed towards any group. We repeated this process to generate similar instances while changing the values of a single variable such as gender, degree, et cetera.

Another analysis technique was using limited data. By replacing certain data points with missing values, we wanted to estimate if the lack of information would affect the chances of receiving an interview call. Since most demographic information is provided by choice, we aimed to assess whether candidates who refused to share this information would be penalized.

These strategies allowed us to deepen our understanding of how these models work. During each experiment, through the isolation of a single variable, we quantitatively assessed how each feature affected a candidate's opportunities. To elaborate, for one experiment, we looked at the median resume scores for males across different ethnicities over five iterations.

Gender	White	Black	Native American	Pacific Islander/ Asian American	Other
Male 01	7.16	2.00	3.56	6.27	2.41
Male 02	8.10	6.78	6.84	3.79	2.64
Male 03	6.11	0.02	5.38	5.09	6.63
Male 04	5.19	9.44	3.17	4.32	6.78
Male 05	7.00	6.48	5.80	7.26	8.92
Median Resume Score (Male)	7.00	6.48	5.38	5.09	6.63

This table illustrates the experiment above. Since we controlled for all other variables, we quantitatively assessed the impact of race in the resume scorer model. For a small dataset, we notice that certain ethnic groups obtain higher scores. As we increased the number of samples from 25 to 4,000, and as we conducted these experiments over multiple iterations, we found that the resume scores did not follow any particular pattern.

In the table above, we can see that Black males seem to receive scores from 0.02 to 9.44 while controlling for all other variables, and this range is too large to reveal any underlying trends. Based on such experiments, we notice that most scores are extremely randomized, which will be further explained in the findings section.

Limitations

In our methodology for data generation, we aimed to address underrepresentation across groups. Usually, real-world datasets are highly imbalanced and rarely contain information on poorly represented communities. This imbalance poses an undeniable issue in machine learning as models are not usually fed the training data necessary to generate fair results. We assume that Providence Analytica used a dataset that mostly consists of male candidates as the financial industry is highly male-dominated. In the context of this audit, we faced limitations in testing the importance of underrepresented instances. Understanding whether underrepresented instances are weighed differently is crucial in assessing the model's robustness and ability to generalize over diverse populations. Without this information, we remain uncertain about the model's reliability.

We would like to specifically address limitations related to the missing value system, as outlined by Providence Analytica in a previous survey answer. According to the company, null values are either imputed or dropped. In our approach, we generate data based on the original dataset provided by Bold Bank, and we see the presence of certain null values. Sometimes, these values might not need to be dropped but cannot be replaced either - for example, a null value in the gender category could either mean that the candidate did not want to reveal that information, or that the candidate did not find a category that they identified with, and so left that information blank. Imputing the missing value with a different gender could cause a bias, and dropping the instance would penalize the candidate for not answering the question.

We also acknowledge the inherent limitations in interpreting these findings within the context of the actual hiring process. During surveys conducted in this audit, we learned that a human element is associated with resume selection, interview calls, and hiring decisions. While it is rare that a Bold Bank representative will override the results of these models, it is difficult to account for interpersonal dynamics such as recruiters networking with certain candidates or referrals. In most scenarios, candidates who share a personal connection with the recruitment team are more likely to receive interview calls, even if their resumes do not perfectly align with the job description. This limits our understanding of whether the recruitment process has improved or worsened because of the implemented changes.

FINDINGS

Through our analysis, we noticed that resume scores were random, and the resume score distribution was uniform. We could not find any correlation between high resume scores and high callback rates. We generated large datasets containing 4,000 instances each time and ran them through the resume scorer model. We noticed that the distribution of resume scores remained consistent regardless of the changes in the dataset. This distribution is illustrated in the image below.

```
generated_data['score'].describe()

count    4000.000000
mean        5.081960
std         2.896805
min         0.000000
25%         2.560000
50%         5.065000
75%         7.640000
max        10.000000
Name: score, dtype: float64
```

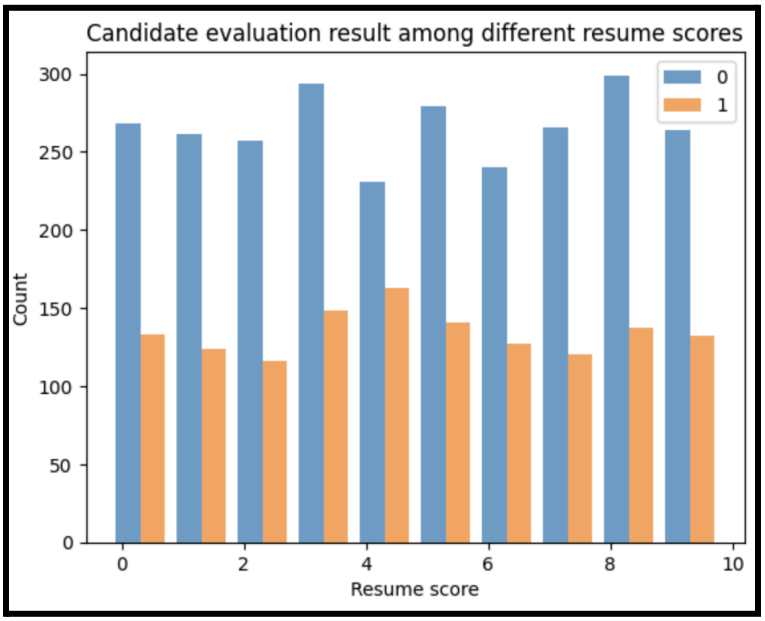

Fig 4**:** This figure illustrates that a high resume score does not result in an interview callback.

We decided to further analyze the results of resume scores and how they relate to callbacks. Ideally, we expected to see a trend where candidates with higher scores received more callbacks. Figure 4 shows us that candidates with a resume score of 4 or slightly above received the most callbacks. This trend was surprising and shows that increasing resume scores and increasing callbacks are not linearly correlated.

Through extensive analysis, we noticed an alarming trend where **males were more likely to receive a callback to attend an interview than females**. We looked at generated data for the same job title and controlled for almost all demographic information except ethnicity and gender. We looked at both males and females who had either a bachelor's degree, a master's degree, or a PhD for all five ethnicities and calculated a callback rate over multiple iterations. Our results showed that in most cases, the male callback rate was higher than the female callback rate. This point is further supported by the data in the table below. We see that for each degree level, males have the highest mean value, taken over the outputs from the candidate evaluator model. We also see that instances with **missing gender or undisclosed gender values never receive an interview call back**.

Degree	Gender	Candidate Evaluator Predictions	
		Mean	Std
Bachelor's	Female	0.41	0.49
	Male	0.62	0.49
	N/A	0.00	0.00
Master's	Female	0.46	0.50
	Male	0.65	0.48
	N/A	0.00	0.00
PhD	Female	0.40	0.49
	Male	0.61	0.49
	N/A	0.00	0.00

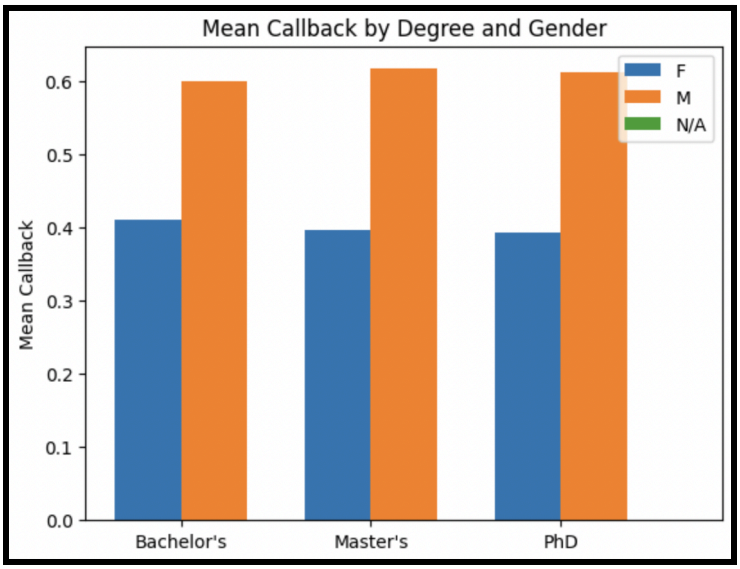

Fig 5: This plot supports our findings that males always receive more interview callbacks than females, and candidates with missing gender never get an interview call.

From our findings, we concluded that candidates were receiving callbacks regardless of their resume scores. While this is a large claim, we believe that either the resume scorer model is incorrectly implemented or the candidate evaluator model does not only consider resume scores in generating outputs. We also ran a dataset that does not contain resume scores through the candidate evaluator and obtained an output. This proves that resume scores are not necessary to receive a call for an interview.

RECOMMENDATIONS

Model Design

Given the observed trend where males receive more interview callbacks than females, it is imperative to investigate any gender-based biases within the model further. One recommendation could be to reevaluate the features used in the model to ensure gender neutrality. Ideally, the model should include features that accurately assess a candidate's ability without accounting for gender, which could be related to GPA scores or past work experience. We also suggest choosing features that rank candidates higher if they share links to project code bases or portfolios. Another suggestion would be to investigate the imbalance between male and female instances to ensure that both groups are equally represented, and if not, to proportionately re-weigh the records.

Since the observed resume score distribution is uniform, we believe these scores are extremely randomized and do not accurately represent a candidate's suitability for the role. An alternative resume-scoring scheme might need to be implemented. We suggest assigning a range of points from 0 to 10 for each feature used in the model. For example, if Bold Bank is looking to hire a senior financial analyst, the resume scorer should assign higher scores to candidates who have more than two years of experience in the financial space. The predicted scores can then be added up and rescaled between 0 and 10. Such a score would be a good starting point to ensure that the right resumes are being selected. We also recommend ensuring the candidate evaluator uses resume scores as a feature and only selects resumes above a certain threshold.

Additionally, we believe that the null value system is not robust enough to handle different cases of missing data. Providence Analytica should develop a more nuanced approach to handling missing values, considering scenarios where candidates opt not to disclose certain information. Imputing or dropping values without context may inadvertently introduce biases into the model.

Lastly, we suggest that these models undergo several tests. We believe that these tests can be quantified using fairness metrics to assess the level of discrimination among various demographics. Disparate Impact and statistical parity difference are useful metrics as they do not require true labels, and give us a comparative value.

Company Practices

While Providence Analytica's software can help speed up the recruitment process, it cannot currently be trusted to provide unbiased outputs. Bold Bank should continue to work with Providence Analytica to develop a fair selection process. However, until that can be achieved, the Bank should empower its recruitment team to exercise their judgment in selecting qualified candidates. At the very least, a recruiter should check all the resumes that pass the candidate evaluator system to ensure that the right candidates receive callbacks.

Bold Bank could also ensure that the recruitment teams are aware of potential biases in the software. Educating stakeholders about the nuances of algorithmic decision-making can help mitigate reliance on the model's outputs as absolute determinants of candidate suitability.

During the hiring process, multiple perspectives need to be considered. For example, suppose Bold Bank tried to hire a data scientist. In that case, the hiring process should include other data scientists, team leads, and HR professionals from different demographic and educational backgrounds. Incorporating a variety of perspectives can help identify any blind spots and avoid any biases in the hiring process.