Final Audit Report

Group Member: Jessica Yang, Vincent Zhu

Introduction

This report presents the findings from an extensive audit of the AI-based hiring system utilized by Bold Bank, developed by Providence Analytica. The primary objective of this audit is to critically assess the decision-making processes within this system, particularly focusing on its adherence to Equal Employment Opportunity Commission (EEOC) guidelines and the bank's standards for fairness and diversity. This involves a detailed examination of the algorithms used for resume scoring and candidate evaluation, and their integration into the broader hiring decisions made by the bank. The necessity for this audit is driven by the growing reliance on automated systems in recruitment and the imperative to ensure these systems are free of bias, transparent, and equitable in decision-making. The audit scrutinizes how decisions are made at each step of the hiring process—from initial applicant screening to final employment decisions—and evaluates the impact of these automated processes on overall hiring outcomes.

Methodology

Data Source

The synthetic dataset we build for querying the API consists of 1000 fake candidate data. The distribution of values of each field of the dataset is shown below:

Applicant ID: Uniform distribution from 1 to 1000 (each ID appears once).
School Name: Providence University: ~167, Providence State University: ~167, State Providence College: ~166, Providence School: ~167, Providence Tech Institute: ~166, Providence City College: ~167
GPA: Distributed between 2.5 and 4.0.
Degree: Bachelor's: ~333, Master's: ~334, PhD: ~333
Location: Providence: ~167, Boston: ~166, Los Angeles: ~167, Miami: ~166, Chicago: ~167, New York: ~167
Gender: Male (M): ~450, Female (F): ~450, Not Applied (N/A): ~100
Veteran status: 0 (No): ~900, 1 (Yes): ~100
Work authorization: 1 (Yes): ~750, 0 (No): ~250
Disability (with -1 representing 'N/A'): 0 (No disability): ~900, 1 (With disability): ~80, -1 ('N/A'): ~20
Ethnicity: 0 (White): ~500, 1 (Black): ~200, 2 (Native American): ~50, 3 (Asian American & Pacific Islander): ~200, 4 (Other): ~50
Role, Start, End: these are randomly generated job names, with some of them as N/A

Evaluation Criteria

We use fairness metrics such as SPD and DI to assess the algorithms, as well as investigating and questioned all parties of the hiring process, including the candidates, the company who provided the model (Providence Analytica), as well as the company who is hiring (Bold Bank).

Based on the replies of the candidates, we suspect that the candidates are being discriminated against work authorization.

Analysis Techniques

We applied the Statistical Parity Difference (SPD) and Disparate Impact (DI) metrics to discover any biases corresponding to "Work Authorization" value. These metrics were selected because they are the parameters on which fairness in such decision-making processes that call for maintaining equal chance and non-discrimination as the law and ethics demand, depends.

The SPD measures the difference between the probability of positive outcomes for the deprived group (the machines that are not allowed to operate - 0) and the probability of positive outcomes for those who are allowed to operate (the machines that have permits to work - 1) therefore, the analysis provides some insight into if any preference exists in any of the subjects.

DI tests the odds of the minority group getting favorable outcomes compared to the majority group, but in a negative manner. A DI value near to 1 signifies fair differences in the treatment of various groups of people. On the other hand, DA values that are far away from 1 are a clue which says that there is an injustice or bias.

Limitations

The limitations and faced us as we conducted a detailed auditing of Bold Bank's AI-driven recruitment system, as a result we had difficulties identifying the magnitude and effect of these constraints in our report. Firstly, we need to use synthetic data to do our audit which is both a useful practical measure to conform with privacy regulations and act responsibly with data. However, the synthetic data is arriving with the disadvantage that it is not representing the full complex, and variation of the real applicant data which in turn results in the gaps in understanding of the AI system with real applicant datasets. In addition, our capacity to execute multiple trials over a desired period of time was somewhat affected by these constraints, thus compromising the thoroughness. Also, the AI models being dynamic features, which improve over time after getting additional training on new facts and observations, made our audit only have captured a cross-section of dynamic conditions only. Our temporal limitations are as such that our finding has been limited to the current state. However, such changes based on adaptability and learning may not have been captured in that analysis. Finally, we should acknowledge that the findings of our audit could be generally attributed. The programmed set-up and the demographic variables we look at in the analysis, together with the model design configuration we audit restrict the finding of our study only to the AI systems in place and therefore will not be broadly applicable to the other AI systems or even different roles in organization. The effect of different AI modal setups or applications is one of the causes of varying results.

Findings

Model predictions:

One thing we noticed about the model, including both candidate scorer (which gives a score based on candidate resume data and evaluator (which gives the final evaluation of whether hiring the candidate) generate very random predictions of the same candidate across different queries. As you can see in the result files: score_1.csv, score_2.csv, score_3.csv, eval_1.csv, eval_2.csv and eval_3.csv, the same candidate get very different scores across different queries, which brings up the question that whether the model is carefully design to achieve consistency.

Fairness metrics results:
We calculate SPD and ID using work authorization as a sensitive attribute, and the result of evaluator. Our analysis yielded the following results across different query results:

Result 1
SPD: -0.030564
DI: 0.8257179397744447

Result 2
SPD: -0.005546
DI: 0.9598286811401565

Result 3
SPD: -0.026935
DI: 0.8281290658339838

These results, supported by quantitative metrics, indicate that while the system generally leans towards fairness, there is a detectable bias favoring candidates with work authorization. These biases are subtle but consistent, emphasizing the need for further refinement of the model to eliminate these disparities fully.

Human Review and Decision-making:
The final decision to grant interviews, despite the initial candidate evaluation by the AI, is made by humans at Bold Bank. This introduces a layer of human judgment that can both mitigate or introduce biases, depending on the reviewers' training, awareness, and adherence to EEOC guidelines and diversity standards.
Human evaluators are responsible for interpreting the results from the AI, which includes making adjustments if a candidate's resume score is low but the candidate is still deemed potentially suitable for the position. This discretion allows for flexibility but also requires stringent oversight to prevent subjective biases.

Feedback Mechanisms and Appeals:
According to Brianna Brown, there was a notable lack of specific feedback on her application rejection, which points to a gap in communication that could affect applicant perceptions of

fairness and transparency. The absence of detailed feedback hinders an applicant's understanding of how decisions are made, which is crucial for trust in AI-driven processes.

The data used to train the AI model is provided by Bold Bank and reflects historical hiring practices. While Providence Analytica claims to preprocess this data to eliminate inequalities, the role of Bold Bank in the initial stages of data preparation is crucial. The quality and integrity of this data, influenced by human decisions and historical biases, directly affect the AI's performance and fairness.

Oversight of AI Decisions:
If the AI makes a decision that is contrary to Bold Bank's core values, there is a mechanism in place to override this decision. This indicates an awareness of the limitations of AI and the necessity for human oversight, but it also raises questions about the frequency, which need to be transparent and consistently applied.

Recommendations

Model Design
In the course of our AI-powered recruiting system audit held by Bold Bank, we have identified some key issues of the system that should be addressed. To enhance the model's reliability, fairness, and transparency, we suggest the following targeted actions for Providence Analytica:

First of all, it is essential to increase robustness of fairness constraints in the AI model so as to better mitigate the biases against protected categories such as gender, ethnicity, and work authorization. It would be advisable to apply Adversarial Debiasing, an advanced machine learning technique, for discovering and reducing the biases by training the model on worst-case scenarios. By tightening these fairness filters, we can mitigate the probability of such discrimination to take place. Instead, candidates are only rated based on the merits and performance that relate to the job.

Additionally, it's necessary to broaden the scoring algorithm's transparency. Our recommendation here would be to devise a means whereby processes in the model become transparent to both candidates and HR managers through clear explanations that will serve as a basis for score reporting. An easy-to-use dashboard can be built that depicts the input of candidate's data into the calculations of their respective scores. With such enhanced transparency, candidates will be able to comprehend how their data is being utilized, moreover, also instilling greater confidence in the AI system which can further help HR professionals to justify their hiring decisions.

Furthermore, we suggest the development of a fixed framework for bias audit and updating the model according to audits outcomes. These audits, especially the ones which are conducted by an independent auditor, should be targeted to recheck the model's performance among certain demographic groups and in different scenarios for ensuring continued accuracy and fairness. Continually updating the model with the latest data and modifying it as necessary will ensure its

performance quality and allow getting rid of emerging biases and issues as a model is processing new data and scenarios. Another vital function is improving the data handling and processing mechanisms. The adaptation of techniques for data augmentation, which simulates a broader range of scenarios will come in handy, in addition to implementing more rigorous data cleaning processes that eliminates errors or inconsistencies that could affect outcomes. Through better data care, the model would be taught on more precise and comprehensive datasets with the aim to have it make fair decisions for a more extensive segment of skillful candidates. Moreover, the conversation of continuous learning can be brought about by establishing feedback mechanisms. With a system that uses inputs from hiring processes outcomes and feedback from candidates, the model can be adapted constantly. This feedback loop would need to include a real-world assessment of the outcomes from application and modifications based on learners' feedback; particularly relating to their impressions of fairness and satisfaction. This continuous refinement will confer the model the ability to adjust and get better, augmenting its accuracy and fairness in a progressive manner.

Company Practices

To have a more meaningful interpretation of the output data of AI models in bailing decisions both ethically and effectively, Bold Bank should think of several strategic enhancements that will help the organization ensure that all concerned stakeholders are efficiently served with the higher interests in mind.

First and foremost, the bank needs to recheck its utilization of both AI and human-generated scores within the hiring framework. While robots can help in the hiring process, having human supervision is of utmost importance and thus, a viable solution is to create a system where the final decision about hiring is taken by humans. It is necessary to reach the stage where the HR personnel are fully aware of AI’s outputs and its limits where the bias and the anomalies in AI systems may require the personnel to override the AI based decisions. This will also help prevent AIs from propagation of biases which might in turn ensure that the recruitment / hiring process is both adaptable and sensitive to nuances that AIs might fail to note. Therefore, avoiding a situation where AI is the only determining factor in the hiring process and ensuring trust among all the candidates is restored as the process stays balanced between both efficiency and fairness.

Additionally, Bold Bank should efficiently inform the applicants how data is being processed and the status of their applications. By giving candidates explanations of their selection progress status and clear information on how their data affects the application outcome they can clarify AI's role in the hiring process and interrogate data privacy or bias concerns. The outcome of all the procedures should be to raise trust in the bank's hiring practices and improve the whole candidate experience. Besides, there is a need for a defined procedure of appeals for candidates who think that their applications have been treated biasedly and unfairly; hence, this is important. This procedure may provide an opportunity for candidates to appeal to a human resources expert for a review particularly where they lose some cases with a strained confidence regarding the accuracy of AI’s evaluation. The AI's decision making process will be safe from any biases as there will be a right to appeal against it.

Moreover, this Application Process will reassure the candidates and the regulatory bodies alike that their hiring practices at Bold Bank follow due processes of fairness and integrity. Furthermore, regular meetings with stakeholders should be scheduled in order to consolidate impressions of all parties affected by the AI hiring system, that is, the candidates, the HR managers, and the senior executives of the company. Interactions in this regard need to highlight the links between the AI system design and operational framework of the bank, and ethical expectations, and its ability to complement instead of clash with the aims of all involved parties. Throughout the course of balancing the interests of all stakeholders, it is essential to establish a suitable algorithm for the AI system, which may allow the bank to be on the right track according to the law and be ethically compliant while achieving the business objectives. The algorithm will undoubtedly steer the hiring process in the desired direction, thus fostering a positive company culture.