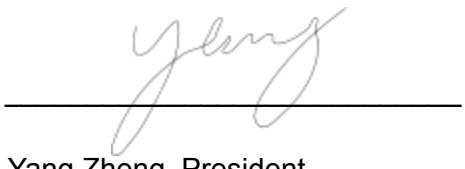
Bold Bank Algorithm Audit: Findings and Recommendations

Introduction

On March 11, 2024, Zheng, Wecht, and Associates (the Auditor) was contracted by the Equal Employment Opportunity Commission (EEOC) to conduct an audit of the hiring algorithm currently being deployed by Bold Bank (the Bank). The Auditor was contracted to perform an investigative audit of the Bank's algorithm and underlying hiring practices after a discrimination complaint was issued against the Bank. The complaint alleged that Bold Bank's hiring system, which relies primarily on a third-party proprietary algorithm, may be in violation of federal laws that make it illegal to discriminate against job applicants based on specified sensitive attributes. The algorithm in question was developed by Providence Analytica (the Developer). The EEOC entrusted the Auditor to perform a comprehensive audit to determine whether Bold Bank's hiring practices meet the guidelines and regulations instituted by the EEOC to protect applicants from discriminatory practices.

In order to fulfill its obligations, the Auditor developed a rigorous auditing procedure that incorporated both qualitative and quantitative assessments of the Bank's hiring practices. The procedure involved comprehensive interviews with relevant stakeholders, including representatives from both the Bank and the Developer. These interviews were used to inform the Auditor's investigations into the hiring algorithm. This algorithm is specifically comprised of two main modules: a Resume Scorer and a Candidate Evaluator. The Resume Scorer is an algorithm that processes a candidate's resume, which includes demographic information, and produces a numeric score ranging from 0 to 10. This score ostensibly indicates how qualified the candidate is for a given position. The Candidate Evaluator is a separate model that considers both the resume score and the Bank's hiring preferences to determine whether or not the candidate should be selected for an interview.

The Auditor assumes that either or both of the models outlined above may potentially demonstrate discriminatory behaviors that are in violation of federal laws and EEOC regulations. As such, the audit procedure and findings detailed in this report investigate the two models independently, as well as the hiring system holistically. This audit was conducted by the Auditor in an impartial and thorough manner. The findings and recommendations outlined in the report were reached after careful consideration and comprehensive analysis.



Yang Zheng, President

Isaac Wecht, Chief Auditing Officer

Methodology

Data Source

In order to investigate the Developer's algorithm, the Auditor created a bespoke dataset that captures specific demographic trends related to sensitivity attributes. The generated dataset contains 4,000 rows and 19 columns, each representing job candidates and their respective attributes. The data generation process entailed randomly selecting attributes within key feature groupings. The feature types were selected based on the information that the Bank claimed was used in its hiring algorithm. Additionally, the potential attributes for each feature group were selected from a predefined list. This structured method allowed for a systematic approach for controlled data generation. The below table outlines the information types, as well as the sensitivity level of the attributes used in the generated dataset. The sensitivity level denotes the relative risk each feature type poses for potential discriminatory behavior:

Information Type	Description	Features	Sensitivity Level
Education	Information related to a candidate's educational achievements	School Name GPA Degree	**MEDIUM**
Demographic	Information pertaining to the candidate's diversity background	Gender Veteran Status Disability Ethnicity	**HIGH**
Geographic	Information regarding the candidate's current location	Location Work Auth.	**MEDIUM**
Work Experience	Information regarding the candidate's previous work experience	Role Start Date End Date	**LOW**

The generated dataset demonstrates uniform distribution across most attributes. The distributions for Gender and Ethnicity are visualized in the below figures. In addition to the 4000 datapoint dataset, the Auditor manually created a smaller set of data points that are controlled for specific sensitive attributes. Within this dataset, the same datapoint is repeated multiple times, with only one attribute being changed. This systematic approach allows us to identify whether certain features are more highly discriminated against by the algorithm. Details of the smaller dataset are below:

Datapoints	Features of Focus	Sensitivity Level
52	Gender, Veteran Status, Work Auth, Disability, Ethnicity	**HIGH**

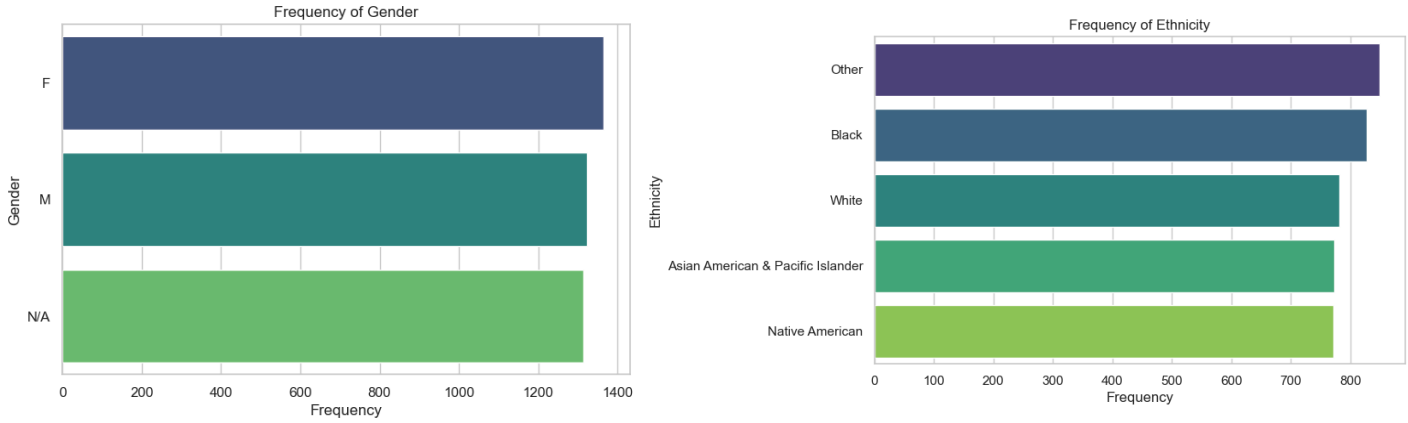


figure one: *The gender attribute is evenly distributed across all 4000 datapoints. Potential gender types include Female, Male, and NA.* ***figure two:*** *The ethnicity attribute is also evenly distributed across all datapoints.*

Evaluation Criteria

In assessing the fairness of Bold Bank's hiring algorithm, the Auditor utilized two established fairness metrics: Statistical Parity Difference (SPD) and Disparate Impact (DI). These metrics were chosen because Bold Bank did not provide true results from their hiring practices and these metrics do not rely on ground truth labels. In addition to the fairness metrics, the Auditor analyzed the confidence intervals of the resume scores produced by the algorithm. Lastly, the Auditor created its own metric, the Reproducibility Metric (RM), which captures how deterministic the algorithm is when producing results for similarly qualified candidates. The below section outlines these various metrics in more detail:

1. **Statistical Parity Difference (SPD)**
    a. Metric Description: SPD measures the difference in the probability of receiving a favorable outcome between a specified majority and minority group.
    b. Metric Range: [-1, 1]
        i. 0 - perfect fairness
        ii. < 0 - bias against the minority group
        iii. > 0 - bias against the majority group
    c. Attributes of Focus: Gender and Ethnicity. These attributes were selected because they represent historic avenues for discrimination by employers. Additionally, Federal law explicitly prohibits employers from evaluating candidates based on these attributes.
2. **Disparate Impact (DI)**
    a. Metric Description: DI assesses the ratio of probabilities for receiving favorable outcomes between minority and majority groups.
    b. Metric Range: [0, ∞)
        i. 1 - perfect fairness
        ii. A disparate impact may exist when this metric is lower than 80%

c. Attributes of Focus: Gender and Ethnicity. These attributes were selected because they represent historic avenues for discrimination by employers. Additionally, Federal law explicitly prohibits employers from evaluating candidates based on these attributes.

3. **Confidence Interval**
    a. Metric Description: Statistical measurement that gives a range of numbers in which the standard mean of a variable should appear.
    b. Attributes of Focus: Resume scores produced by the Resume Scorer. By studying the CI for these scores, we can assess the distribution of model results, and the degree of confidence the model has in its predictions.
4. **Reproducibility Metric (RM)**
    a. Metric Description: Statistical measurement that quantifies how similar results are for the same candidate, who is passed through the algorithm multiple times. The metric is calculated as the average difference between the candidate's results.
    b. Metric Range: [0, ∞)
        i. 0 - perfectly reproducible
        ii. The higher the RM, the less reproducible the results are
    c. Attributes of Focus: Resume scores produced by the Resume Scorer. By studying the RM for these scores, we can assess how deterministic the model is in its predictions for similarly qualified candidates.

Analysis Techniques

The Auditor implemented three distinct strategies in order to investigate the model's decision process. These strategies included (1) large scale data generation and analysis, (2) smaller, controlled data generation and analysis, and (3) True Positive and False Positive analysis with accuracy assumptions. Details of these approaches are outlined below:

1. **Large Scale Data Analysis**
    a. Technique: As described in the Data Source section, the Auditor generated a synthetic dataset with 4000 data points, each representing a unique candidate. This same dataset was twice passed through the Resume Scorer and Candidate Evaluator. Results from these passes were analyzed using the above criteria.
    b. Justification: Large scale data allows the Auditor to gain a holistic understanding of how the model operates. Additionally, multiple passes through the model with the same data can be used to estimate the model's reproducibility.
2. **Smaller, Targeted Data Analysis**
    a. Technique: As described in the Data Source section, the Auditor generated a synthetic dataset with 52 data points, each representing a unique candidate. This dataset is controlled for specific sensitive attributes. Results from two passes through the model were analyzed using the above criteria.
    b. Justification: A controlled dataset allows the Auditor to study the model's behavior regarding specific attributes. Additionally, multiple passes through the model with the same data can be used to estimate the model's reproducibility.

3. **True Positive and False Positive Analysis with Derived True Labels**
    a. Technique: Due to the unavailability of actual hiring outcomes, we devised an approach to simulate 'true labels' based on an assumed interview reception rate. The rate of receiving an interview at the Bank was calculated to be approximately 20%. This threshold was derived by analyzing the frequency of interview recommendations across two evaluations of the dataset. Specifically, candidates with a resume score above 8 were considered to have a positive outcome (label = 1), mirroring the bank's operational benchmark for selecting candidates.
    b. Justification: With simulated true labels the Auditor is able to conduct additional evaluations that test the algorithm's ability to correctly predict whether a candidate should receive an interview.

Limitations

Despite the Auditor's best efforts, there are several limitations that hindered our ability to efficiently audit the hiring system. The below table outlines a non-exhaustive list of limitations and their level of hindrance to our auditing goals:

Limitation	Description	Impact on Audit	Level
Lack of Ground Truth Labels	Generated data points were not labeled with their true outcomes from hiring	Unable to derive fairness metrics that rely on true labels	**HIGH**
Results provided in binary terms	Model produced results in binary terms, without providing probabilities for its predictions	Key analysis techniques, such as LIME, rely on probabilities from a model's output	**Medium**
Insufficient information from key stakeholders	Representatives from the Bank and the Developers refused to provide answers to several interview questions	Limited the scope of qualitative analysis available from insider descriptions	**Medium**

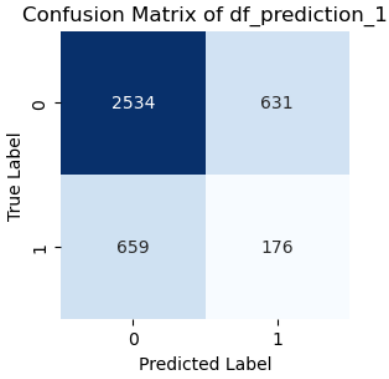
In addition to the methodology limitations denoted above, the Auditor faced inherent limitations from the circumstances in which the audit was conducted. Most crucially, the Auditor was not granted access to the code base from the Developer. As a result, the entire audit was conducted as a black box analysis, where the model was probed for results. It is important to note that this dynamic will inherently limit the effectiveness of the audit.

Findings

As outlined above, our audit utilized several key fairness metrics and statistical standards to evaluate how the Bank's algorithm performs across different demographic groups. The below table details the overall summary of our findings with respect to these metrics:

Evaluation Metric	Large Dataset Average	Small Dataset Average	Summary
Statistical Parity Diff. (Gender)	-0.157	-0.365	Findings suggest a pattern of disadvantage for the minority gender group
Disparate Impact (Gender)	0.583	0.441	Findings substantially below the threshold of 0.80; suggest a significant disparate impact against the minority gender group
Statistical Parity Diff. (Ethnicity)	0.006	0.006	Findings suggest there is no pattern of disadvantage for the minority ethnicity group
Disparate Impact (Ethnicity)	1.035	1.030	Findings above the threshold of 0.80; suggest there is no significant disparate impact against the minority ethnicity group
Confidence Interval	[4.97, 5.16]	[4.12, 5.76]	Mean score: 5.15. The narrow CIs show the model's stability in scores across runs
Reproducibility Metric	3.35	3.47	Significant variation in the scores when the same candidates are assessed multiple times

We see that the DI metric, on average 0.512, is significantly below the Federal 0.8 guideline. This indicates that the algorithm is biased against female and black candidates. Additionally, we measured the Resume Scorer's RM to be 3.41 on average, which demonstrates that the algorithm is not deterministic in its evaluation of similar candidates, raising concerns for maintaining consistent hiring standards. Lastly, we conducted an analysis of the TP and FP rates of the algorithm as mentioned above.



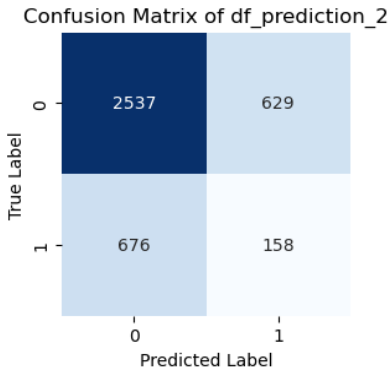


figure three: *Confusion Matrix of the first Evaluation.* ***figure four****: Confusion Matrix of the Second Evaluation.*

The above figures show that the model has a 17% false positive rate, meaning that the algorithm will fail to properly identify 17% of candidates that should be selected for interviews. This finding calls into question the accuracy and reliability of the algorithm for all candidates.

The audit of the hiring algorithm, using robust statistical and fairness metrics, has revealed substantial biases that may impact fairness and compliance with anti-discrimination laws.

Recommendations

The critical findings outlined above illustrate the need for immediate and targeted revisions to the Bank's hiring algorithm. We provide the following recommendations for both the Developer and the Bank:

The Developer - Model Design

1. **Bias Mitigation Algorithms:** Implement machine learning algorithms specifically designed to reduce bias, such as fairness-aware algorithms that adjust weights and training methods to level the playing field for all demographic groups.
2. **Enhanced Data Collection:** Expand the dataset to include more diverse demographic information to better understand and mitigate biases across genders and ethnicities. Ensuring a balanced representation can help reinforce fairness in model outcomes.
3. **Regular Audits and Updates:** Establish a routine for periodic audits of the model to identify and address any emerging biases or discrepancies. Updates should be made based on audit findings to continually improve model fairness.
4. **Transparency in Algorithm Design:** Increase transparency around how decisions are made within the algorithm, especially decisions that involve critical outcomes like interview recommendations. Involve all relevant stakeholders in conversations.

The Bank - Company Practices

1. **Stakeholder Collaboration and Transparency:** The Bank should establish a more collaborative relationship with Providence Analytica to ensure that both parties regularly review and understand how the AI model influences hiring decisions. Additionally, the Bank should provide clear explanations on its career portal about how AI is used in the recruitment process, both holistically and for individual applicants.
2. **Ethical Oversight and Compliance:** Implement an ethics board that includes representatives from Providence Analytica, Bold Bank, and an independent third party to oversee the application of AI in hiring processes. This board should ensure compliance with EEOC guidelines and address any ethical concerns that arise.
3. **Bias Monitoring and Mitigation:** Develop and implement ongoing training programs for HR teams to recognize and mitigate potential biases in AI-assisted decisions. The training should be informed by regular bias audits of the AI model and should empower HR personnel to make adjustments to the AI recommendations when necessary.
4. **Documentation and Reporting:** Keep detailed records of all AI model updates and key decisions. These documents should be accessible to all stakeholders to ensure a unified understanding and approach to addressing biases.