# Final Audit Report

## Introduction

This audit was initiated to evaluate the hiring system employed by Bold Bank, which uses decision-making models developed in collaboration with Providence Analytica. The system is designed to streamline the recruitment process through the use of a resume scorer and a candidate evaluator. These components work together to analyze data pertaining to a candidate to determine whether this candidate should be invited for interviews. The primary purpose of this audit is to ensure that the implementation of these models follows the guidelines and regulations enforced by the Equal Employment Opportunity Commission (EEOC).

This audit tries to evaluate whether the decision-making processes might discriminate against applicants based on sensitive attributes like gender, disability, work authorization, ethnicity and veteran status. These attributes serve as inputs for the Resume Scorer and Candidate Evaluator models developed by Providence Analytica. Importantly, various federal laws safeguard protected classes such as ethnicity and gender. Additional legislation, which we will discuss later in the audit, also forbids hiring discrimination on the basis of characteristics such as disability, veteran status, and citizenship status (considered here in terms of work authorization), which are collected during the employment process.

This examination was done after recent complaints that suggested that the system might systematically disadvantage some candidates based on their demographic information. Consequently, the focus of this audit will revolve around the integrity and fairness of the decision-making algorithms used in Bold Bank's hiring process.

## Methodology

### Data Source

When testing our resume scorer, we initially suspected that the model might be generating random values. To investigate this, we created a dataset (resume_tester_large.csv) replicating each participant in the given dataset resume_scorer.csv 800 times, given the API's 4000-row limit. We submitted this data to the API to analyze the distribution of scores for each participant and determine if it followed a random pattern. As such, the distribution of the csv used to query the API followed exactly the same distribution as the one originally given by resume_scorer.csv, it just had more rows.

To detect potential biases in the API we generated specific datasets for each sensitive attribute i.e gender, veteran status, work authorization, disability and ethnicity, including an "N/A" option where the candidate was allowed to select that option. Each dataset consisted of entries with the same value for one sensitive attribute while the other sensitive attributes were randomly generated based on predefined probabilities. This approach allowed us to isolate the impact of each sensitive attribute in a dataset on the final output of the model. The distribution of the other sensitive attributes (not the one held constant by choice) remained consistent across different datasets due to predefined probabilities, ensuring that any changes in the results could be attributed to the selected sensitive attribute that was kept fixed. Specifically, the probability of being a part of the gender categories were

set at 45% for both male and female and 10% unspecified, veteran status at 70% non-veteran, 25% veteran, and 5% unspecified, work authorization at 20% no and 80% yes, and disability status at 85% no, 10% yes, and 5% unspecified. Ethnicity was uniformly distributed, with each category having a 20% probability of being picked for any candidate.

After creating the datasets, an analysis showed that the numerical means for the non-focused sensitive attributes aligned closely with their predefined probabilities for all datasets that we created to query the API. For instance, in datasets that were not specifically testing the veteran_status attribute, approximately 25% of the datapoints were veterans and 75% were not, reflecting our intended probability setup. Ultimately, this means that all the datasets we created and used to query the candidate evaluator had the same distributions for the sensitive attributes we were not testing, as desired.

Additionally, we constructed a control dataset named `control_data` comprising 4000 rows where no specific sensitive attribute was kept fixed. Each attribute was assigned randomly using the same predefined probabilities as above. This dataset was then placed through the API to get its corresponding result. After this, both the dataset and its results were employed to train an XGBoost model that was used as a tool to interpret the API (which to us was a black box model as we could not see the code but just the results it outputted).

Furthermore, to assess if the model displayed biases towards certain sensitive feature combinations, we examined the combined effects of different attributes. To do so, we created other datasets of 4000 rows, keeping some combinations of features constant (say, everyone was a black female) while varying the others randomly, through our predefined probabilities. This helped us identify if the model showed preferential treatment towards combinations like black male, black female, white male, female without work authorization, and/or male with work authorization. This allowed us to evaluate the model's fairness and accuracy not only for one sensitive attribute but also for a combination of sensitive attributes.

Moreover, we conducted interviews with the main stakeholders involved in the admissions process which were: a job applicant, Bold Bank, and Providence Analytica. By analyzing their responses to our questions we tried to identify any potential issues in the hiring process that could show the existence of bias in the hiring process.

**Evaluation Criteria**

From the start, we suspected that the resume scoring model might be generating random scores as its output. To verify whether this was true, we replicated each of the five candidates in the provided resume_scorer dataset 800 times into a dataset (resume_tester_large.csv), plugged this dataset into the API to get the results (stored in resume_results_large.csv) and examined the score distribution for each candidate. This method was used to assess whether our suspicions about the model generating random numbers for the resume score were correct.

In evaluating potential biases within the candidate evaluator model, we used the Disparate Impact metric, and then used the “Four-Fifths Rule” to assess the outcome. This is because this rule gives us

a widely used legal standard that can be used to determine if a model's outputs could be seen as discriminatory. The Four-Fifths Rule states that any value below 0.8 for Disparate Impact indicates a discriminatory output from the algorithm.

Additionally, once our XGBoost model was developed, we wanted to examine the importance of sensitive attributes in influencing the model's predictions. While not conclusive, we believed that if the sensitive attributes emerged as the most significant factors in the XGBoost model's decisions, this could suggest potential biases of the candidate evaluator itself. Essentially, the most relevant features identified by our interpretability tool (XGBoost model) could reveal potential discriminatory patterns by indicating which potentially sensitive attributes influence the decision-making of the candidate evaluator.

The sensitive attributes examined were veteran status, gender, ethnicity, disability, and work authorization. These attributes are protected under various anti-discrimination laws, which include the Civil Rights Act of 1964, prohibiting discrimination based on ethnicity or sex; the Immigration Reform and Control Act of 1986, which forbids discrimination based on citizenship or work authorization status; the Americans with Disabilities Act of 1990, preventing discrimination against individuals with disabilities; and the Vietnam Era Veterans' Readjustment Assistance Act of 1974, which prohibits discrimination in hiring against veterans. Focusing on these attributes aimed to ensure that the model's outputs did not cause any unfair bias against these legally protected groups in the hiring process.

**Analysis Techniques**

When assessing our resume scorer, we initially thought it might be outputting random scores. To investigate this, we created a dataset in which each participant was replicated 800 times, respecting the API's 4000-row limit (resume_tester_large.csv). This dataset was subsequently submitted to the API, and we analyzed the distribution of scores for each participant to determine if it showed a random pattern. Once the dataset was processed by the API, we collected the scores and constructed a histogram of the scores for all participants. If the histograms for the different candidates all displayed a uniform distribution, it would suggest that the resume scorer was essentially generating random numbers, confirming that the resume scorer was just outputting random numbers.

To calculate Disparate Impact, we created multiple datasets where one sensitive attribute was consistently applied while the others were randomly generated. This approach allowed us to isolate and evaluate the impact of a specific sensitive attribute while maintaining a consistent distribution for the others. For example, in the female dataset, all 4000 rows featured women but the values for the other sensitive attributes that were not gender were randomly generated. These datasets were then sent through the API, and we calculated the mean of the results. Since the results were binary (0 or 1), the mean value represented the proportion of candidates of a certain sensitive attribute that the model recommended for interviews. We compared the outcomes across different sensitive attributes from these datasets to see the variations in positive outcomes associated with each sensitive attribute. We then ranked the datasets (each of which contained all points who had the same score for a certain sensitive attribute) based on which had the highest and lowest average predictions of positive outcomes. This analysis helped identify which classes were privileged and which were not.

By dividing the averages (which represent the probability of success for each sensitive attribute), we could then find the Disparate Impact. Consequently, we were able to check if the result complies with the Four-Fifths Rule.

$$\text{DI} = \frac{P(\hat{Y}=1|A=\text{minority})}{P(\hat{Y}=1|A=\text{majority})}$$

After assessing Disparate Impact for individual sensitive attributes, we further explored whether the model exhibited bias against specific combinations of sensitive attributes. To do this, we generated datasets where two sensitive attributes were held constant instead of just one. The specific combinations we tested included black male, black female, white male, female without work authorization, and male with work authorization. These combinations were chosen because preliminary results suggested that gender might be a factor in bias, and we hypothesized that combining gender with another sensitive attribute could reveal more discriminatory patterns. We then submitted these datasets to the API and compared the average results to the highest averages obtained by the dataset where just the gender attribute was kept fixed. This comparison helped us determine if the combination of two attributes could result in an increase in the discriminatory effect of the algorithm.

Since we didn't have access to the underlying code of the model, we developed our own XGBoost model to provide some level of interpretability regarding the decision-making process. We trained this model using `control_data.csv`, which included randomly generated values for all sensitive attributes, along with the outcomes from the candidate_eval website, used as our target variable. This approach was designed to replicate the functionality of the candidate_eval model, enabling us to identify which features were most influential in its decision-making process. Utilizing the feature importance functionality of XGBoost, we could determine which attributes the model deemed significant. If our XGBoost model accurately predicted the outcomes generated by the candidate_eval model, and particularly highlighted a sensitive attribute as a key feature, it would suggest that the candidate evaluator model might be using this sensitive attribute, which it should not do. Using feature importance in our XGBoost model for interpretation helped us further explore whether the candidate_eval model might be using sensitive attributes in its decisions. Additionally, we used SHAP values to evaluate the importance and impact of these sensitive features on the decision made by the model on a single candidate.

**Limitations**

One limitation we had in this audit was that we did not have direct access to the code used by Providence Analytica. As such, we could only make conclusions on the dataset based on its outputs and not the code itself. This means that although we can try and identify the existence of bias, it would be hard to understand why this bias arises. Therefore, our recommendations must remain somewhat general, as we are unable to specify exact changes needed in the code.

Furthermore, while the XGBoost model used to interpret the results from the candidate evaluator achieved a solid 67% accuracy, it does not consistently correctly predict the output of the candidate evaluator. While it can help us interpret the results of the API, it is not a highly reliable predictor, and

the most significant features it identifies may not always necessarily align with those considered most crucial by the candidate evaluator model created by Providence Analytica.

Another limitation we encountered was the lack of information regarding actual outcomes, such as whether candidates truly received interviews. This prevented us from using other fairness metrics like EOD or AAOD, which depend on the true outcome, to assess the fairness of the hiring process. Consequently, while we later determined that the model exhibits bias through the disparate impact analysis, we were unable to determine if the bias extended in multiple ways or if it disproportionately affected other sensitive attributes through different fairness metrics.

Time constraints also prevented us from testing every possible combination of sensitive attributes. While it's unlikely, there's a chance that we may have overlooked specific interactions between variables that could demonstrate discrimination. With additional time and resources, we would investigate all combinations of sensitive features and examine their impacts.

## Findings

When we analyzed the resume scores produced by the Resume Scorer model, our goal was to assess the consistency and pattern of the scores to determine if the model was generating random numbers as resume scores. To do this, we replicated each of the five individuals in resume_scorer 800 times into a dataset (resume_tester_large.csv), placed this dataset in the evaluator model, and plotted the distribution of the results (resume_results_large.csv). The histograms for the outputs of the resume scorer for the five participants are shown in *Figures 1-5* in the Appendix.

The histograms for all candidates reveal that the distribution of the resume scores is uniformly distributed between 0 and 10, exhibiting a similar count for both the minimum and maximum scores. Consequently, it appears that the resume scorer does not exhibit bias but instead generates random scores for all candidates evaluated. Ultimately, this shows that our initial suspicion was correct and the resume scorer was just outputting random numbers as resume scores.

As mentioned in our analysis, we used the Disparate Impact Ratio to check for the existence of bias. The observed disparate impact ratios among the sensitive attributes that were given by the candidate evaluator model can be seen in *Table 1* in the Appendix. From the table, it is evident that the algorithm does not meet the “Four-Fifths Rule” for all sensitive attributes, as it recorded a value below 0.8 for the gender category. This disparity highlights a bias in the algorithm's outputs, favoring the privileged class (males) over the non-privileged classes (females and candidates marked as 'N/A'). The results suggest that the model discriminates not only against females but also against individuals who opt for 'N/A' as their gender. This issue is particularly concerning as it may affect those who are non-binary and do not see themselves represented in the binary gender options, leading to their potential candidate rejection based on this. This is particularly concerning as interviews revealed that applicants were assured by Bold Bank that choosing "N/A" for any demographic information would not influence their application outcomes. However, our audit findings contradict this, demonstrating that selecting "N/A" for gender consistently results in rejection by the algorithm. This indicates not only the presence of bias within the algorithm but also suggests that the bank, which commissioned the algorithm, was unaware of this discriminatory effect.

According to our disparity impact analysis, the model did not appear to discriminate on any sensitive attributes other than gender. Additionally, as shown in *Table 2* in the Appendix, no combination of sensitive attributes resulted in a significant increase in the discriminatory effect of the model.

From our use of the XGBoost model as an interpretability tool, we discovered that the 'N/A' option in the gender field was the most significant predictor for the outcome of the candidate evaluator model, with an importance score of 0.79. 'Male' and 'Female' followed as the second and third most important features, with scores of 0.09 and 0.08, respectively. In contrast, all other metrics were marginal, each scoring 0.004 or less in feature importance. Given that the top three features identified by our interpreter are all related to gender, this suggests that the candidate evaluator may also be using gender to make its decisions, which is not legally allowed. Using the different feature importance interpreter we got similar results for global feature importance, where gender seemed to play a big role in the XGBoost model's decision making. We then also used SHAP to check local feature importame for a single point, which showed that gender was the main feature that was used by the model to make the decision for that point. These results can be seen in *Figure 6* and *Figure 7* in the Appendix.

## Recommendations

### Model Design

To prevent the model from resulting in discriminatory hirings we provide recommendations for both Providence Analytica, who created the model and Bold Bank who commissioned and uses the model.

Providence Analytica has been provided with a dataset spanning thirteen years by Bold Bank, which could contribute to the observed outcomes and potential biases in the hiring model. Although new data is being fed into the system every hiring cycle, the limited thirteen-year timeframe may still be a problem. The dataset's historical nature might be perpetuating biases in hiring or not reflect the current diversity of the applicants, impacting the model's accuracy and fairness due to the scarcity of data for training from underrepresented groups such as women or people who are non-binary (and would place "N/A" on the gender). This could lead to less accurate predictions for these groups, which can potentially be disadvantageous to women and non-binary individuals in hiring decisions made by the model. To address these issues, it is crucial for Providence Analytica to not only continuously update and review the dataset but also apply techniques like data re-balancing or re-weighting the data used to train the model. These methods can help ensure equitable representation in the training set, encouraging a model whose outputs will display a fairer treatment across all genders options.

Furthermore, as indicated by the interpreter model, despite Providence Analytica denying it, the model seems to use gender as a factor in its decision-making process. To prevent this, we recommend that Providence Analytica verifies whether the model utilizes gender or any proxies of it in its decision-making. If so, removing the feature or proxy could promote fairness through unawareness as the model could not use the gender attribute or its proxies to make a decision.

Moreover, we found that Providence Analytica's statement "We use a range of metrics to validate the model's performance, including accuracy and F1 score" indicates Providence Analytica's reliance on traditional performance metrics to tune its model in training. However, these alone don't reveal biases or the distribution of sensitive attributes in decisions. Incorporating fairness metrics like the Disparate Impact (DI) ratio during model tuning could help identify biases and ensure different demographic groups are treated equitably by the model. This approach would help Providence Analytica train a model that will follow the "Four-Fifths Rule", which would follow the legally accepted hiring fairness standards.

Finally, we suggest that Providence Analytica revisits the resume scoring model, as it appears to generate random scores. While these scores may not exhibit bias based on sensitive attributes, they could disadvantage individuals who deserve an interview but then receive a low score randomly.

**Company Practices**

Additionally, Bold Bank's absence of protocols for overriding algorithms suggests that the situation becomes particularly concerning when biases are detected. This gap could allow biased outputs to influence hiring decisions for a long time before they are noticed and then addressed. To improve this, Bold Bank should create clear guidelines and checks for human reviewers. This could involve more training programs focused on recognizing and mitigating algorithmic bias, or clear criteria for hiring managers as to when to ignore the algorithm's recommendations.

Furthermore, Bold Bank has not specified to Providence Analytica any instructions regarding the fairness metrics to be used in training and assessing the algorithm. We recommend that in the future, Bold Bank employees responsible for sourcing this algorithm receive training on identifying bias and utilizing numerical methods, such as the Disparate Impact, to detect it.

**Appendix**

*Figure 1: Histogram of Resume Scores for Applicant ID 1 in the given 'resume_scorer.csv'*

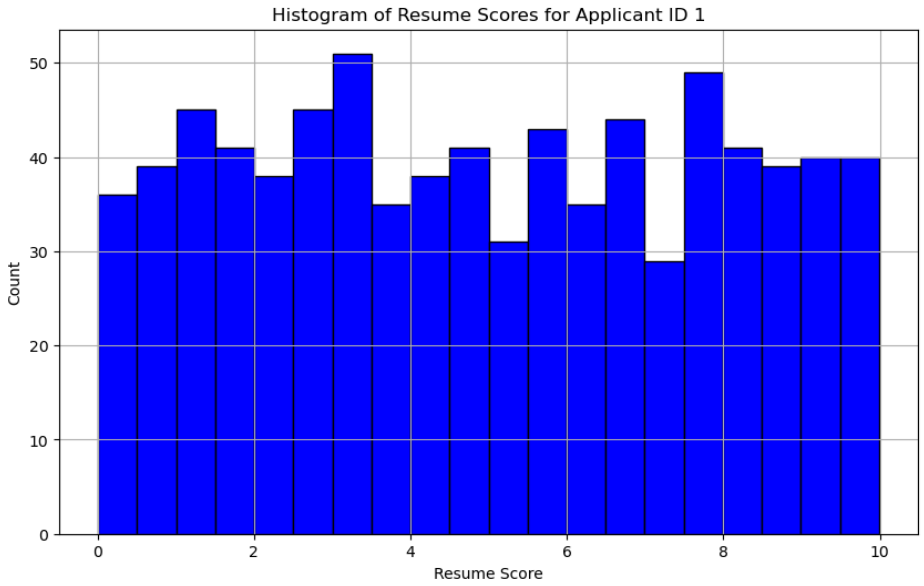


*Figure 2: Histogram of Resume Scores for Applicant ID 2 in the given 'resume_scorer.csv'*

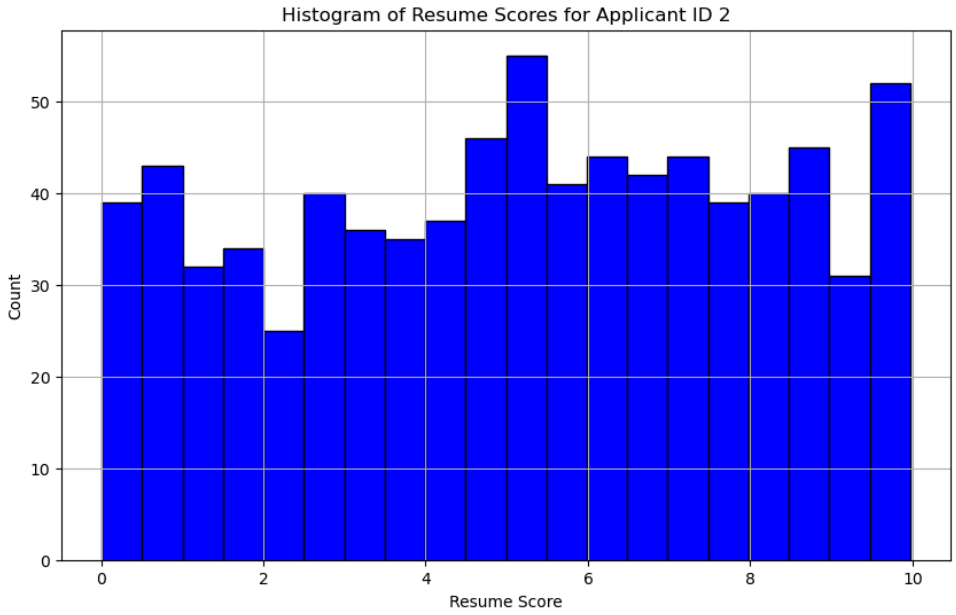

*Figure 3: Histogram of Resume Scores for Applicant ID 3 in the given 'resume_scorer.csv'*

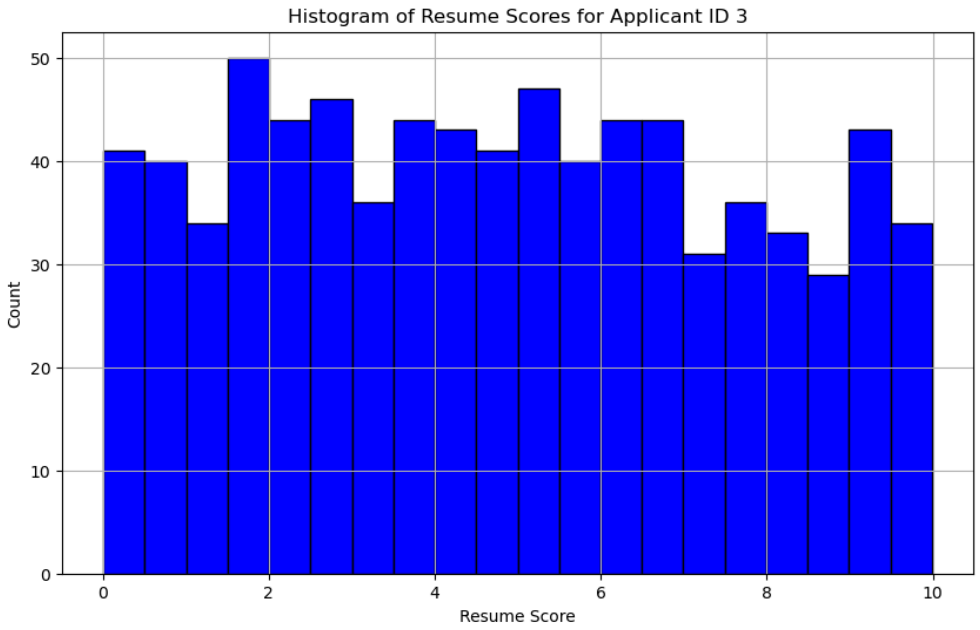


*Figure 4: Histogram of Resume Scores for Applicant ID 4 in the given 'resume_scorer.csv'*

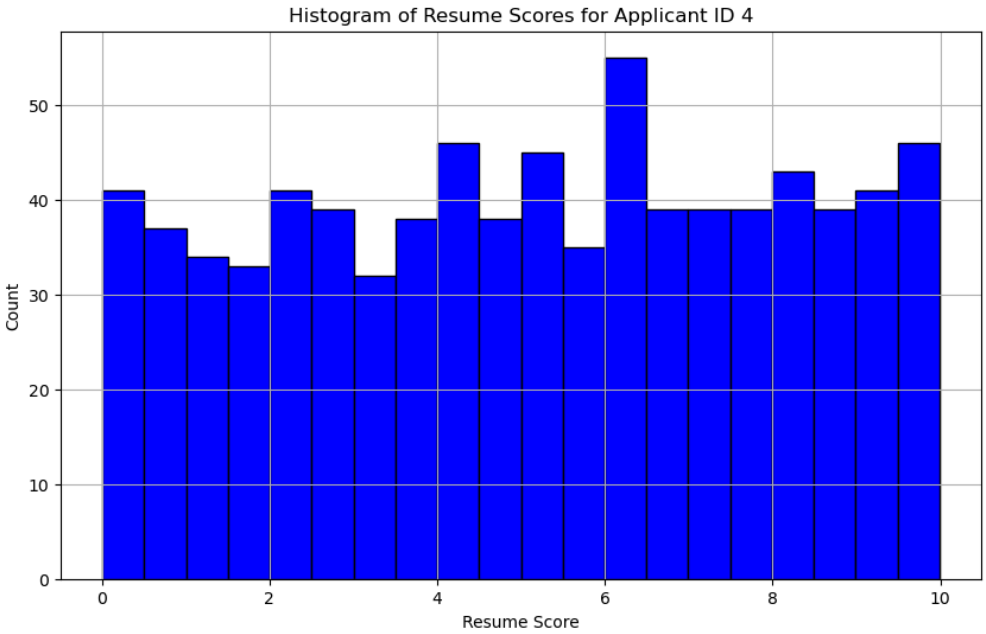


*Figure 5: Histogram of Resume Scores for Applicant ID 5 in the given 'resume_scorer.csv'*

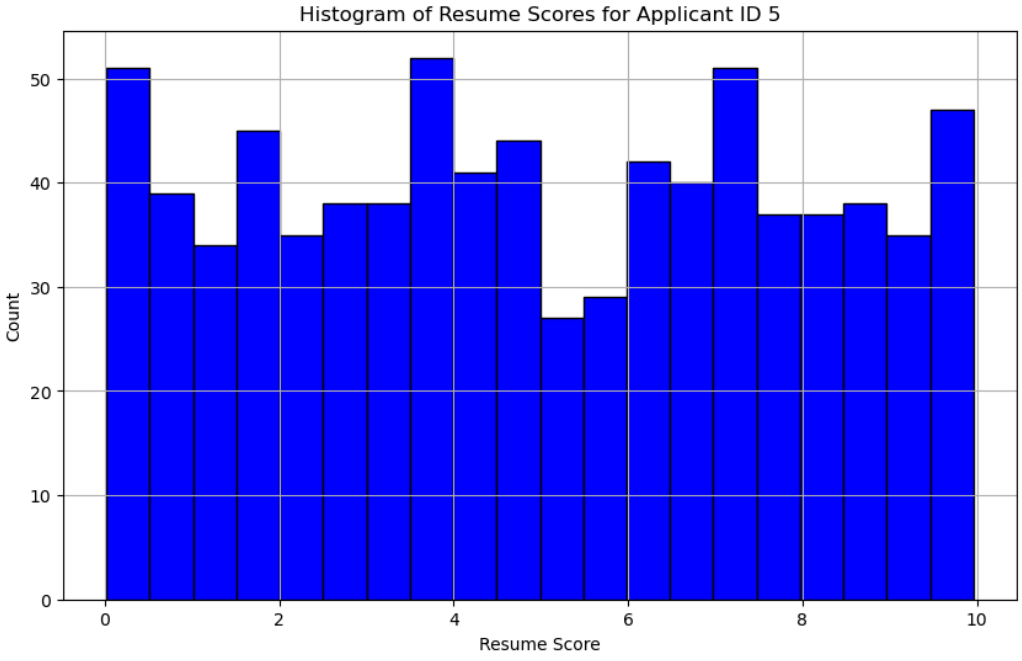

*Figure 6: XGBoost Model Most Important Global Features*

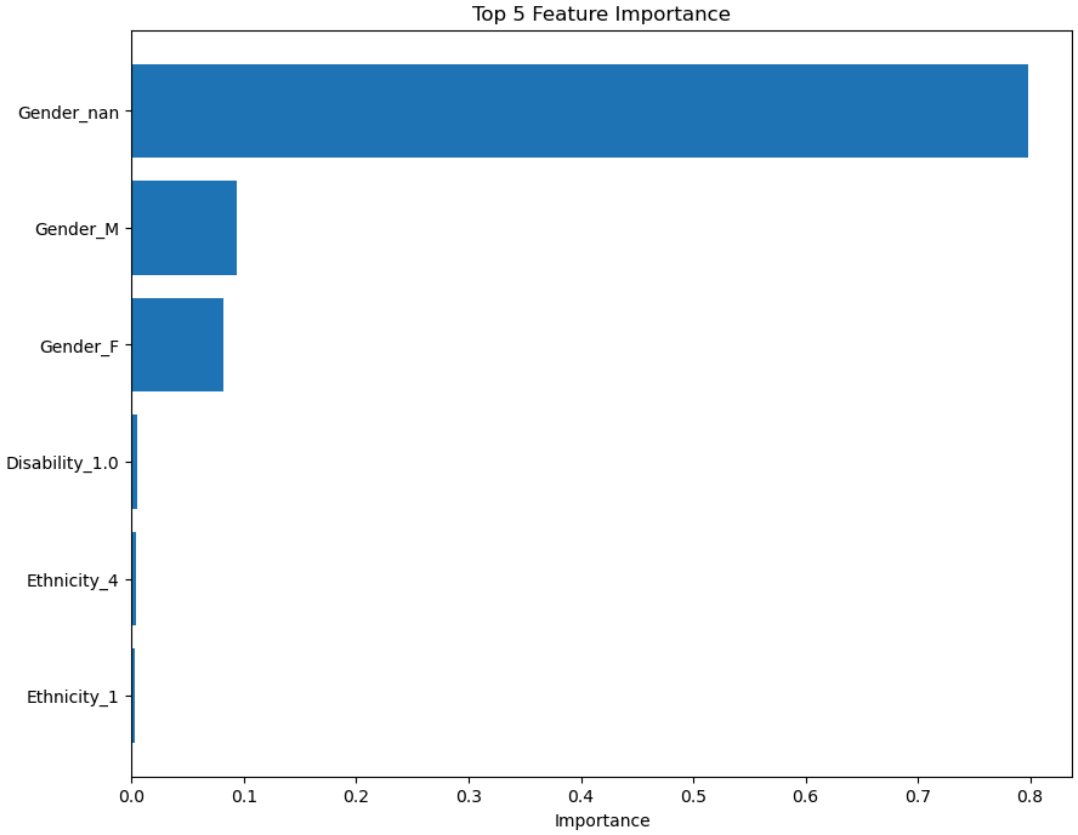

*Figure 7: SHAP Waterfall Plot With Most Important Features Used by XGBoost Model for a Single Candidate*

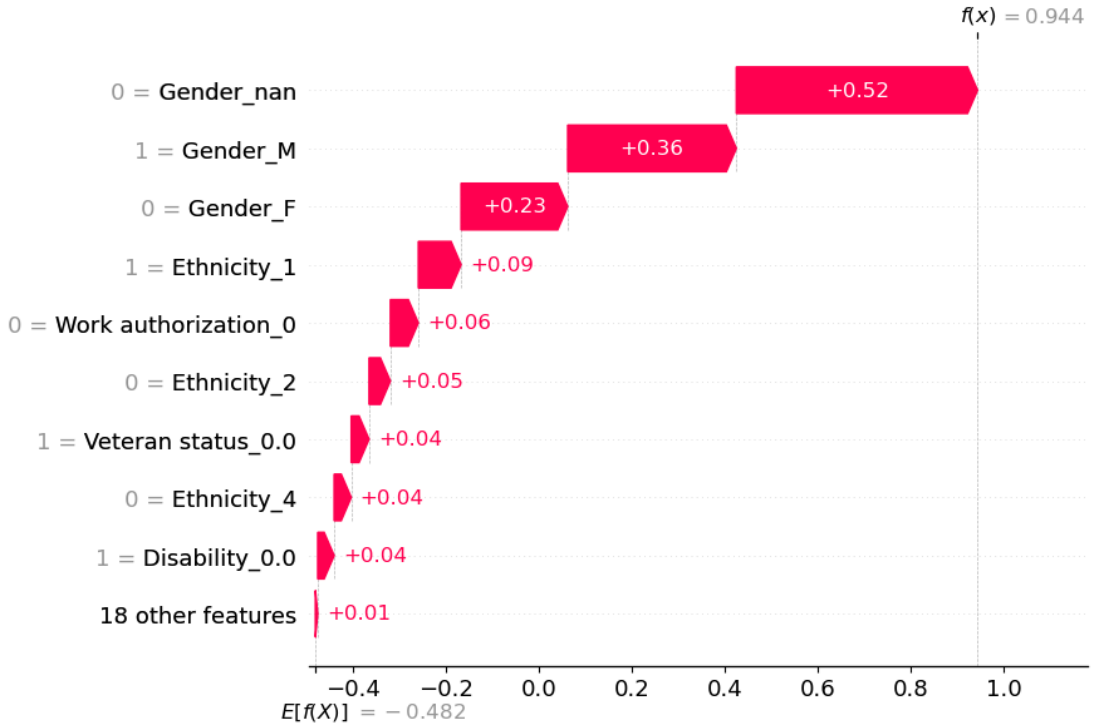

*Table 1: Disparate Impact Analysis Results, Single Sensitive Attribute*

| Sensitive Attribute | Average Acceptance Rate | Majority/Privileged Group Acceptance Rate | Disparate Impact Ratio |
|---|---|---|---|
| ***Gender*** | | | |
| Female | 0.419 | 0.651 (Male) | 0.643 (<0.8) |
| Male | 0.651 | 0.651 (Male) | 1.000 |
| N/A | 0 | 0.651 (Male) | 0 (<0.8) |
| ***Ethnicity*** | | | |
| White | 0.479 | 0.485 (Black) | 0.987 |
| Black | 0.485 | 0.485 (Black) | 1.000 |
| Native American | 0.469 | 0.485 (Black) | 0.967 |
| Asian American & Pacific Islander | 0.462 | 0.485 (Black) | 0.952 |
| Other | 0.476 | 0.485 (Black) | 0.981 |
| ***Veteran Status*** | | | |
| No | 0.467 | 0.478 (Yes) | 0.976 |
| Yes | 0.478 | 0.478 (Yes) | 1.000 |
| N/A | 0.467 | 0.478 (Yes) | 0.976 |
| ***Disability Status*** | | | |

| | | | |
|---|---|---|---|
| Yes | 0.468 | 0.488 (No) | 0.959 |
| No | 0.488 | 0.488 (No) | 1.000 |
| N/A | 0.472 | 0.488 (No) | 0.967 |
| ***Work Authorization*** | | | |
| No | 0.460 | 0.479 (Yes) | 0.960 |
| Yes | 0.479 | 0.479 (Yes) | 1.000 |

*Table 2: Disparate Impact Analysis Results, Combination of Sensitive Attributes*

| Combination of Sensitive Attributes | Average Acceptance Rate | Average Acceptance Rate for the Respective Gender | Disparate Impact Ratio |
|---|---|---|---|
| Male, Black | 0.649 | 0.651 (Male) | 0.997 |
| Male, White | 0.649 | 0.651 (Male) | 0.997 |
| Female, Black | 0.414 | 0.419 (Female) | 0.988 |
| Male, Work Authorization | 0.660 | 0.651 (Male) | 1.014 |
| Female, No Work Authorization | 0.397 | 0.419 (Female) | 0.947 |