Dominique Barnes
Jingyi Lu

Final Audit Report

Introduction

We have been hired as an auditor by the Equal Employment Opportunity Commission (EEOC) to investigate the new hiring system developed by Providence Analytica to streamline Bold Bank's recruiting process. Due to the sensitive nature of inviting applicants for interviews, we have taken this audit very seriously to ensure that fairness is maintained and the top priority of the system developed by Providence Analytica.

The hiring system developed by Providence Analytica consists of a resume scorer and a candidate evaluator. It is important to note that the candidate evaluator is dependent on the resume scorer. The resume scorer processes resumes in a CSV format and assigns a score from 0 to 10 (0–worst, 10–best). The candidate evaluator processes the same resumes in a CSV format however, it also includes the resume score produced by the resume scorer. Utilizing the resume score, the candidate evaluator generates a binary outcome (0 or 1) for whether the candidates should receive an interview.

The EEOC has called on us to conduct an external audit on the basis of possible discrimination the system is alleged to have from previous complaints. By federal laws it is illegal to discriminate against a job applicant or employee because of the person's race, color, religion, sex, gender identity, sexual orientation, national origin, age, disability or genetic information. Our role is important because we will fairly and accurately assess the allegations in the charge and make a finding by conducting a very thorough audit to ultimately determine if it complies with the laws of discrimination.

The goal of this audit will be to determine whether or not the resume scorer and candidate evaluator demonstrate bias behavior. We will check the fairness-accuracy tradeoffs in the model to ensure there is a balance between the trade-off of accurately selecting good candidates that align with the company's policies and practices and expectations in a fair manner that does not discriminate based on the contents of the resume. A source of bias may come from the data provided by the bank in order to train the resume scorer and candidate evaluator. The distributions of applicants from different demographic groups might be influenced by advertisement techniques utilized by the bank and the career fairs attended. From our interviews with Providence Analytica we have also learned that historical hiring data was involved in the construction of the models which can also introduce biases, therefore, disproportionately affecting one or more groups of applicants in the decision making process.

Due to the confidential nature of the historical data, we will create our own datasets that span a range of different demographics and variables that should directly influence hiring decisions. We will also create datasets to test whether similar or same demographics across applicants will result in similar resume scores. Hypothetically, the scores should all be the same if the contents of each applicant's resume is the same or very similar. If this is not the case then we will have reason to believe that there is bias and we will determine how this bias was introduced to the model.

Methodology

Data Source

The dataset was created with a Jupyter Notebook in Python using Google Colabs. We have created a dataframe that was used as inputs for the resume scorer and candidate evaluator. Our algorithm used variables for school name, grade point average (GPA), degree (Bachelor's, Master's, PhD), location, gender, veteran status, work authorization, disability, ethnicity, roles in companies, and start/stop dates for the roles. The output of the resume scorer was attached to the dataset to be fed into the candidate evaluator for analyzing fairness. Using the Data Dictionary provided by our employer, we were able to ensure the values of each variable was formatted correctly for the resume scorer and candidate evaluator. We created a dataset of 3000 applicants to ensure the robustness of the models during the audit and randomly generated values for each variable of each applicant to prevent computational bias in our dataset. The distributions of the sensitive attributes can be visualized in **Figure *1***.

Evaluation Criteria

To assess whether the two algorithms were biased or not, we used some fairness metrics including Disparate Impact (DI) and Equal Opportunity Difference (EOD). We believed that one possible presentation of bias was the difference between the proportions of people who received an interview in the majority group of a sensitive attribute and those in the minority group of the attribute, so we used disparate impact to investigate this difference. DI measures the ratio of the rate of favorable outcomes for the unprivileged group to the rate of favorable outcomes for the privileged group. A value of 1 indicates no disparity, values less than 1 indicate potential adverse impact on the unprivileged group, while values greater than 1 indicate a potential adverse impact on the privileged group. In addition to the predicted probabilities of receiving an interview, we would like to examine the predictive performance of the models, so we incorporated true positive rates and true negative rates in our analysis. Furthermore, employed EOD to measure the difference in true positive rates between the privileged and unprivileged groups. A value of 0 indicates no difference between the true positive rates between the privileged and unprivileged groups.

5 sensitive attributes were taken into consideration from the resume: 1. Gender, 2. Veteran Status, 3. Work Authorization, 4. Disability, and 5. Ethnicity. Since these attributes described the demographic information of the candidates and were unrelated to their ability of working, they were ideal variables to investigate the fairness of the algorithms. As the definition of bias in Artificial Intelligence mentioned, the most common systemic biases in AI are racism and sexism[1], so it is crucial to include gender and ethnicity in the sensitive attributes. A research study has been conducted by Stone et al. that has demonstrated evidence of potential bias that may be a contributing factor to high veteran unemployment rates, which is why veteran status was also included as a sensitive attribute[2]. Disability is another important focus of sociological

[1] Schwartz, R., Apostol, A., Vassilev, V., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *NIST Special Publication 1270*. https://doi.org/10.6028/NIST.SP.1270

[2] Stone, C. B., Lengnick-Hall, M., & Muldoon, J. (2018). Do stereotypes of veterans affect chances of employment? *The Psychologist-Manager Journal, 21*(1), 1–33. https://doi.org/10.1037/mgr0000068

and psychological research in employment[3], indicating the significance of fairness analysis using this attribute. In 1960 immigrant workers only accounted for 1 in 17 workers in the United States, however, as of 2015 immigrants have represented nearly 17% of the workforce[4,5]. In order to keep the growing trend of hiring diverse candidates and creating an inclusive work culture by including applicants from all over the world in the application process, work authorization was included as a sensitive attribute. The privileged group was determined by identifying the majority value for the specific sensitive attribute, while the unprivileged group was any other class/value listed for the sensitive attribute.

Analysis Techniques

In order to investigate the model thoroughly it was necessary to create an output of true predictions. These true predictions were determined in a two-step process. The first step was to use important factors that were selected during our interview process by Bold Bank. We inquired about the expected level of education and years of experience for applicants for both entry level and senior positions for the position that Bold Bank was hiring for which was Financial Analyst.

During the interview, Bold Bank had informed us that they are generally looking for a candidate that possesses a minimum of a bachelor's degree, while prior experience may be beneficial, it is not always required for entry level positions. However, for senior level positions they seek candidates with higher levels of education (e.g. Master's or PhD) and significant professional experience. We also added another consideration of GPA, which was evaluated with three tiers. The first tier was a GPA less than 3.0, the second was a GPA between 3.0 and 3.5, and the highest awarded tier was a GPA greater than 3.5. This was used to reward applicants with a higher GPA which is often considered during application processes.

We evaluated each applicant and provided a resume score by summing the positive attributes that were present in their application. The score begins at zero, the applicant's GPA would add an additional 1, 1.5, or 2 points to their overall score depending on the tier they were placed in for the GPA score described above. The second attribute considered was whether the applicant had Bachelor's, Master's, or PhD, where a Bachelors would add an additional 1 point to their overall score and Masters or PhD would add 1.5 to the overall score. The applicants with a Bachelor's degree were awarded highly for having multiple roles. This was quantified by identifying students that had a role description under the categories of Role 1,2, and 3 with the score increasing from 1.5, 2, and 2.5 respectively depending on the amount of roles the applicant reported in their resume. However, because this is a mandatory requirement for

---

[3] Vornholt, K., Villotti, P., Muschalla, B., Bauer, J., Colella, A., Zijlstra, F., … Corbière, M. (2018). Disability and employment – overview and highlights. *European Journal of Work and Organizational Psychology*, 27(1), 40–55. https://doi.org/10.1080/1359432X.2017.1387536

[4] Center for Immigration Studies. (2023, Month Day). Employment Situation of Immigrants and US-born: Fourth Quarter 2023. Retrieved from https://cis.org/Report/Employment-Situation-Immigrants-and-USborn-Fourth-Quarter-2023

[5] Lambert, J. R., Basuil, D. A., Bell, M. P., & Marquardt, D. J. (2019). Coming to America: work visas, international diversity, and organizational attractiveness among highly skilled Asian immigrants. *The International Journal of Human Resource Management*, *30*(15), 2293–2319. https://doi.org/10.1080/09585192.2017.1322116

Masters and PhD applicants, the scores awarded to having multiple roles were less than that of a Bachelor's candidate with an increase in score from 0.5, 1, and1.5 respectively for having reported roles. The final resume scores varied between 5.5 and 9.0 for the applicants in our test data (**Figure *2***).

The second step in determining the true prediction score was to add the resume scores into the candidate evaluator DataFrame. Once these values were added to each candidate we decided on a threshold on whether or not to give the candidate an interview. The reason why only the resume score was considered for interview prediction, was because the same variables considered in the resume scorer would be considered in the candidate evaluator. In order to avoid redundancy, we used the resume scorer as the only variable to determine the prediction of the interview. We concluded that a resume score greater than or equal to 7 would provide a prediction of 1, which would signify the applicant of receiving an interview. The threshold of 7 was determined because the most competitive candidates received this score or higher. The frequency of interview offers was plotted below, and we can see that our model reflects the competitive nature of interviews offered to candidates in this industry (**Figure *3***). We were able to create a list of true positives that we could then use to conduct further analysis of fairness metrics.

Limitations

One major limitation of our dataset is that it cannot give a realistic representation of when some of the variables are systematically biased because it is completely randomly generated. For example, gender bias appears extensively in most of the datasets in the real world, so the scores of metrics using true positive rates may be slightly worsened by higher theoretical proportion of females in our dataset. The ratio of different ethnic groups are also distorted by the randomization method as the groups are distributed according to geographical areas, which makes it impossible to have similar group sizes in any population in a specific area.

A limitation in the analysis technique used in our auditing process is that we are creating what we believe would be positive attributes to the company. Further analysis would be needed in order to accurately determine the weights given to each variable used in the resume scorer algorithm (GPA, education, and work experience). Although we actively collaborated with Bold Bank in inquiring about specific qualification requirements needed for the position they are hiring for, the resume score was calculated based on weights determined by us, the auditors. It is also very important to incorporate more variables in the resume score as the score is used to determine whether or not the candidate is provided an interview. This is another limitation of our analysis technique, as the candidate evaluator solely used the resume score to provide an interview outcome. Ideally more variables would be used in combination with the resume score, however, for this evaluation of the model it provided sufficient information. Our analysis technique can be improved upon, however, for the task of creating true predictions in order to complete our analysis, we believe it has created a true representation of the financial job market.

There were also limitations inherent to the circumstances of the audit. This was due to limited resources being available to us as auditors. Providence Analytica was not able to share any information with regards to how the models they used for Bold Bank were created or trained. We did not have access to their training or testing data. We also did not have access to

the fairness metrics they used to validate the models. Due to the limited amount of resources provided by Providence Analytica, we were not able to directly compare how our fairness metrics may have matched with the fairness metrics used by Providence Analytica.

Findings

Disparate Impact

The disparate impact for the sensitive attribute of gender is 0.315, which suggests the rate of positive predictions for the unprivileged gender group (female) is lower compared to the privileged gender group (male). The disparate impact value for veteran status is 1.068, indicating that the rate of positive predictions for the privileged group (veterans) is slightly higher when compared to the unprivileged group. The disparate impact value for work authorization is 1.014 which is very close to 1, this indicates minimal disparity of positive predictions between the privileged (those not requiring work authorization) and unprivileged groups. The disparate impact value for disability status is less than 1 with a value of 0.874, this indicates a potential adverse impact on the rate of positive prepositions for the unprivileged group when compared to the privileged group (those without disabilities). However, according to the four-fifth rule generally adopted by the US government, this value for disability status is acceptable. The last reported sensitive attribute is ethnicity with a disparate impact value of 1.047 which is slightly higher than 1, suggesting that the rate of positive predictions for the privileged group (3 for Asian American & Pacific Islander) is slightly higher compared to the unprivileged groups. These interpretations provide insight into potential disparities in the outcomes predicted by the model across different sensitive attributes listed in **Table *I***.

Sensitivity and Specificity Analysis

Sensitivity (**Equation *1***) and specificity (**Equation *2***) were calculated with the equations reported in **Table *II***. The true positives and negatives and false positives and negatives determined by the predicted outcomes which are the results from the Providence Analytica model, and the true predictions which were calculated and determined through the method described in the “Analysis Technique" section. Sensitivity, also known as the True Positive Rate (TPR) measures the proportion of actual positive cases that are correctly identified by the classifier. In this case the sensitivity was measured as 0.243. This means that out of all the actual positive cases, only 24.3% were identified correctly by the classifier. Specificity, also known as True Negative Rate (TNR) measures the proportion of actual negative cases that are correctly identified as negative by the classifier. In our audit case, specificity is 0.773. Out of all the actual negative cases, 77.3% were correctly identified as negative by the classifier.

Equal Opportunity Differences

The results of the Equal Opportunity Difference calculations are presented in **Table *III***. The EOD value for gender is 0.274 suggesting that the model is more likely to correctly predict positive outcomes for males than females. The EOD for veteran status is -0.021 which is a negative value, indicating that the true positive rate for the unprivileged group is slightly higher than the privileged group (veterans) on a small scale of difference. For the work authorization attribute, the EOD is -0.030, again indicating that the true positive rate for the unprivileged group

is higher compared to the privileged group (those not requiring work authorization). The EOD for disability is 0.016, suggesting that the model is more likely to correctly predict positive outcomes for the privileged group (those without disabilities) than the unprivileged group. The final sensitive attribute is ethnicity and the EOD is -0.063, suggesting that the model is more likely to predict positive outcomes for individuals with the majority ethnicity (3 for Asian American & Pacific Islander) when compared to the rates for all other ethnicities.

According to the results above, we were able to observe a significantly lower disparate impact value and higher EOD value in gender than in other sensitive attributes. More specifically, these values showed that females were less likely to be predicted qualified to receive an interview compared to males and the predictive power of the algorithm was better for males than for females. Therefore, we can conclude that the evaluating system constructed by Providence Analytica was significantly biased with respect to gender. However, as we mentioned in the limitations, we do not have enough access to the datasets, algorithms, and evaluation metrics used in constructing the models, so we are not able to specify the exact source of bias in the whole process.

Recommendations

Model Design

To reduce the impact of bias from data, Providence Analytica may need to do exploratory data analysis and preprocessing of training data more carefully. All the sensitive attributes shall be covered in EDA, and the results shall be utilized in data transformation and model construction, with which extra emphasis should be put on the sensitive attributes that are unequally distributed. Useful techniques include the disparate impact remover in preprocessing and regularization in loss functions.

A recommendation to the model developers at Providence Analytica would be to ensure they are incorporating DI and EOD into their fairness metrics. If they had calculated the DI and EOD values for the sensitive attributes in the model, they might have been able to find disparities in their model before deployment. DI is important because it highlights systems that are meant to be inherently neutral but actually present an unfair bias to unprivileged groups. Similarly with EOD, it is important to measure because it provides assurance that each group is being treated equally. It is crucial to know where sources of bias may occur and DI and EOD are good metrics from identifying it. By addressing these biases early on, Providence Analytica will be able to ensure their clients that their models are fair with respect to interview offers and they will be able to avoid any future issues that may arise with candidates feeling a sense of bias from their models.

Company Practices

From our audit we have determined that gender is a sensitive attribute that may be creating bias in Providence Analytica's models for determining interview offers. Due to this, we believe it is in the best practice of Bold Bank to increase collaboration with Providence Analytica prior to deployment of the models. Early involvement of the model development process would have allowed Providence Analytica to understand the expectations and requirements for candidates applying through this system. If biases were to persist after early preventive steps were taken, it

would be important for Bold Bank to consider alternative options for their hiring process. For instance, they could find a third-party to participate in the investigating process to make sure the models and results from Providence Analytica were unbiased. Fair and unbiased evaluation of resumes would be in the best interest of each stakeholder involved in the decision-making process.

Appendix

1. Figures

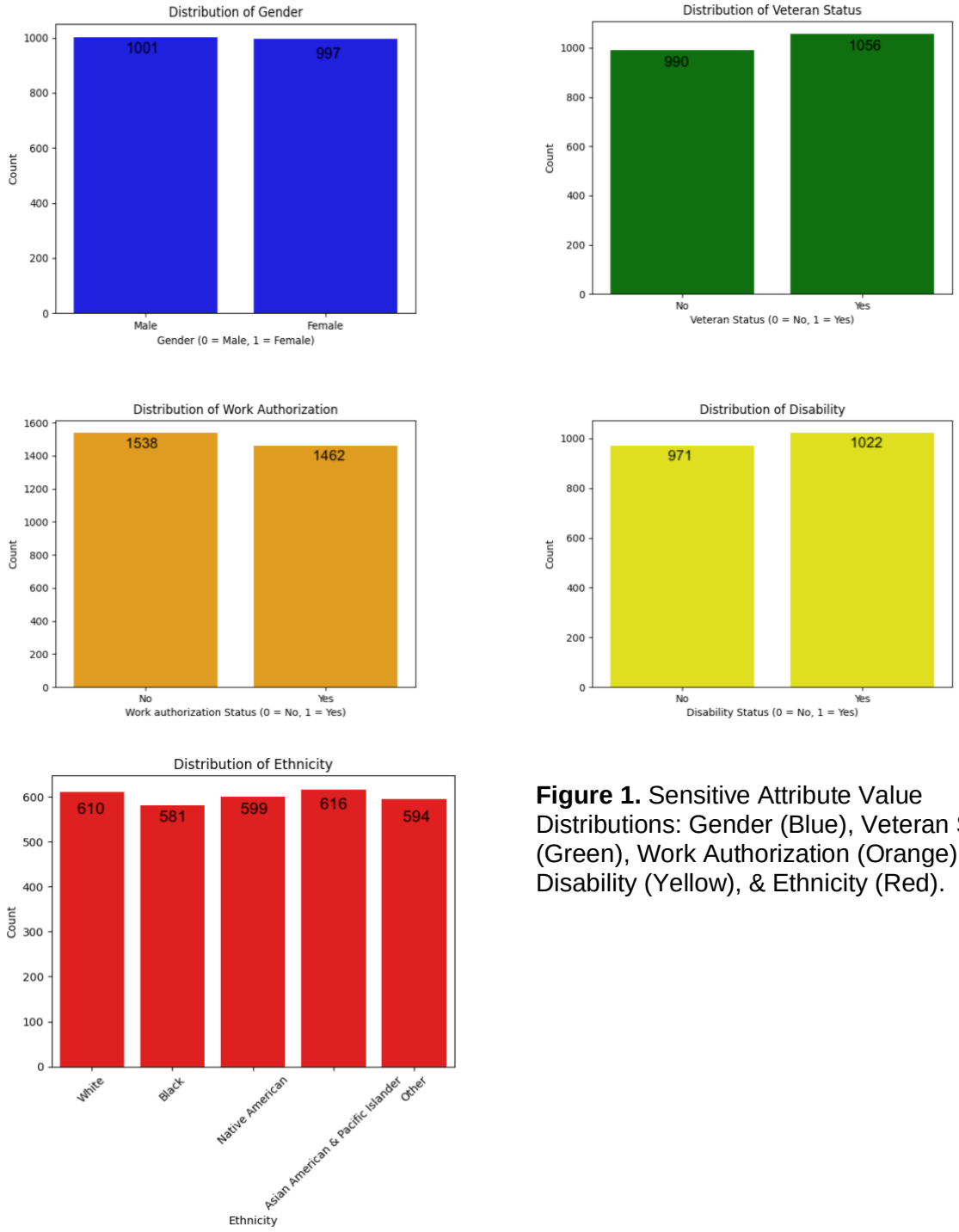

Figure 1. Sensitive Attribute Value Distributions: Gender (Blue), Veteran Status (Green), Work Authorization (Orange), Disability (Yellow), & Ethnicity (Red).

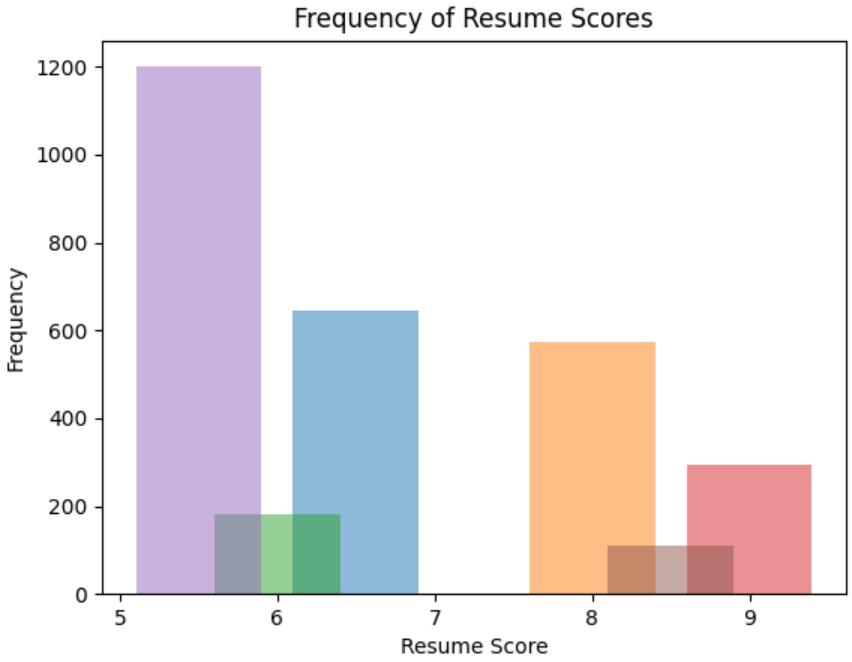

Figure 2. Frequency of resume scores calculated from the analysis technique

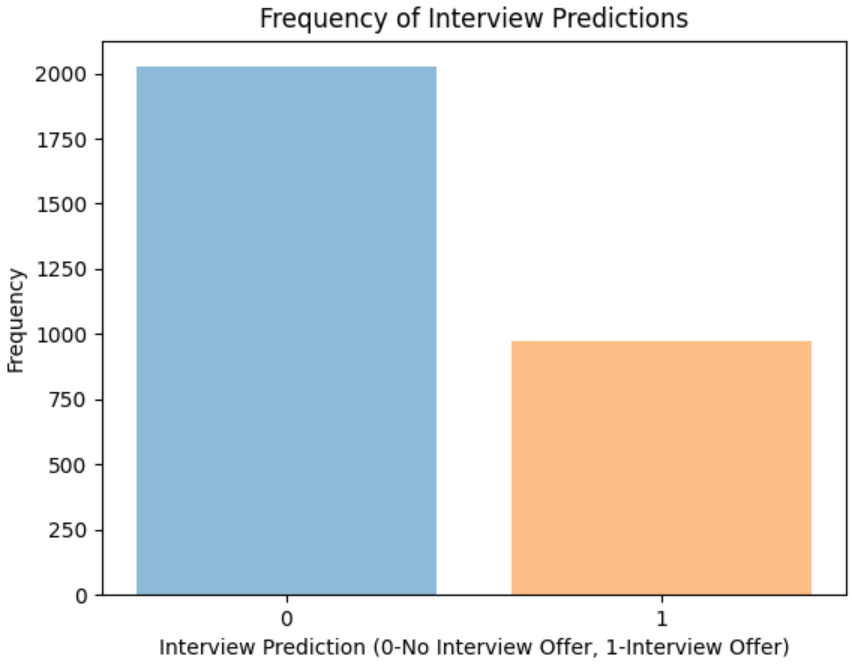

Figure 3. Interview predictions scores from true resume scores

2. Tables

Table I. Disparate Impact Results

Sensitive Attribute	Gender	Veteran Status	Work Authorization	Disability	Ethnicity
DI Value	0.315	1.068	1.014	0.874	1.047

Table II. Sensitivity and Specificity Results

Sensitivity	**Specificity**
0.243	0.773

Table III. Equal Opportunity Differences Results

Sensitive Attribute	Gender	Veteran Status	Work Authorization	Disability	Ethnicity
EOD Value	0.274	-0.021	-0.030	0.016	-0.063

3. Equations

Equation 1: $Sensitivity = True\ Positives\ /\ (True\ Positives + False\ Negatives)$

Equation 2: $Specificity = True\ Negatives\ /\ (True\ Negatives + False\ Positives)$