Consulting Report: Addressing Ethical Issues in a Hiring Algorithm

Group Members:

Sofia Torres
Mia Marte
James Yackanich
Alex Major

Organization:

Florida Atlantic University

Date Created:

July 30th, 2024

1. Introduction

In this project, we're working as consultants for a tech company to take a closer look at their hiring algorithm and see if there are any ethical issues. Our job is to analyze how the algorithm works, check out the data it was trained on, and figure out if it's fair, transparent, and respects privacy. We will examine aspects such as bias and fairness ensuring the algorithm does not favor or disadvantage specific groups and privacy concerns, to verify that personal information is handled appropriately. Additionally, we will assess the transparency and explainability of the algorithm's decisions, ensuring that the rationale behind hiring choices is clear and understandable. To achieve this, we will utilize tools such as ChatX, a conversational AI to assist with brainstorming and analysis, and Facets Dive for data visualization. Our goal is to identify any ethical concerns and provide recommendations to enhance the fairness, transparency, and trustworthiness of the hiring process.

2. Methodology

2.1 Subjective Analysis

Based on the provided description of the hiring algorithm, "NIST Trustworthy and Responsible AI Framework (NIST AI 600-1)" document, and co-ideation with ChatX, these questions focus on fairness, bias, transparency, accountability, and privacy:

- Fairness and Bias
  - Is the training dataset representative of all demographic groups (e.g., gender, ethnicity, age)?
  - Does the algorithm disproportionately favor or disadvantage specific demographic groups (e.g., based on gender, race, or age)?
  - How is bias detection performed during the training and evaluation phases?
  - Are certain skills or attributes from underrepresented groups weighted less favorably compared to others?
  - How are fairness adjustments applied when ranking candidates? Are these adjustments transparent and justified?
- Transparency and Explainability
  - Can the algorithm's decisions (e.g., candidate ranking) be easily explained to both hiring managers and candidates?
  - What steps are taken to ensure that hiring managers can understand and trust the algorithm's decisions?

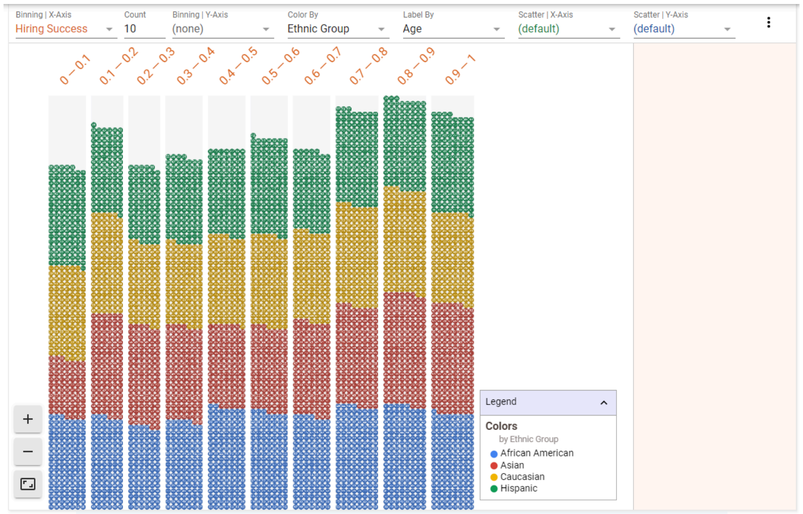
- How are candidates scored and ranked? Can the specific factors influencing a candidate's score be communicated clearly?
- Does the algorithm offer insights into why a particular candidate was shortlisted or rejected?

- Privacy and Data Security
    - Does the algorithm comply with data privacy laws (e.g., GDPR, CCPA)? How is personal information protected?
    - Are candidates' personal data anonymized or minimized? Are data protection measures in place to safeguard personal data against misuse or breaches?

2.2 Data Analysis

The three datasets share the same classification inputs/features, which are considered by the algorithm when analyzing job applicants. Key features in the datasets are as follows: gender, marital status, ethnic group, age, education, experience in various programming languages (Java, SQL, Python, ML) and job requirements match score. The job requirements match score is a value that is used to see how well the job applicant's experience and qualifications match with the predefined requirements of the job role. The prediction output of the dataset is the hiring success value, which takes into account a wider range of features such as ethnicity, gender and marital status to represent how successful a candidate is expected to be during the hiring process.

While using Facet Dive to analyze the data, testing for potential bias in the algorithm was conducted. Features such as gender, marital status, ethnicity and age were compared with the hiring success value to see which type of candidate was expected to be successful during the hiring process, as these features are some that do not correlate with the job requirements match score but can still be taken into account by the hiring success value. Pictured below is dataset one, categorized by hiring success value based on various ethnic groups, which can be used to showcase how potential biases were looked for and found in the datasets.



Binning | X-Axis: Hiring Success
Count: 10
Binning | Y-Axis: (none)
Color By: Ethnic Group
Label By: Age
Scatter | X-Axis: (default)
Scatter | Y-Axis: (default)

After conducting more data analysis on the three datasets and testing more features, potential biases in the algorithm include gender, marital status, ethnic group, and age.

2.3 Tools

- Google PAIR's Facet Dive for data visualization.
- ChatX for ideation and analysis.

---

3. Findings

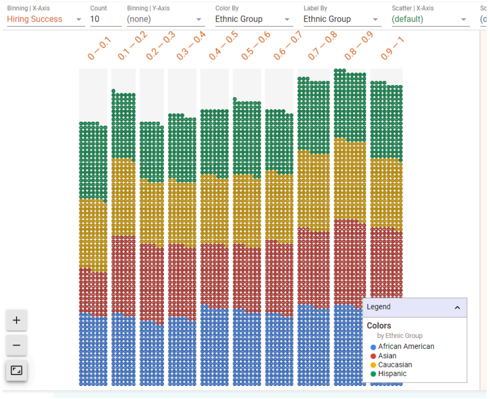
3.1 Bias and Fairness

Potential biases discovered in the algorithm (e.g., gender, race, indirect biases).

According to the algorithm, this demographic information is collected and used so that bias is able to be mitigated, so below the datasets will be the analysis to see if the algorithm would be trained on datasets that could lead to bias.

DATASET 1

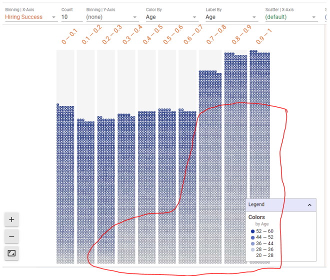
For dataset 1, the hiring success distribution based on ethnic group, age, gender, marital status, and education all had pretty much the same even distribution throughout(Below is an image of this distribution that repeats throughout).

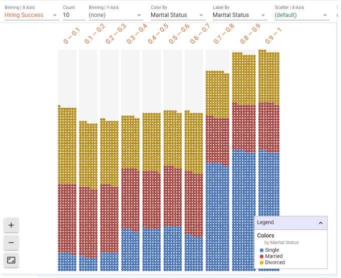


Binning | X-Axis Hiring Success
Count 10
Binning | Y-Axis (none)
Color By Ethnic Group
Label By Ethnic Group
Scatter | X-Axis (default)

DATASET 2

This dataset seems to have hired more younger people, and more single people. Other than that, the distributions are the same as the first dataset.

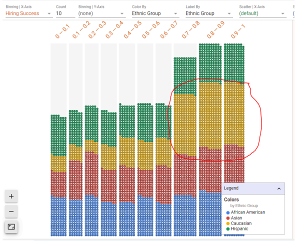

Binning | X-Axis: Hiring Success
Count: 10
Binning | Y-Axis: (none)
Color By: Age
Label By: Age
Scatter | X-Axis: (default)


Binning | X-Axis: Hiring Success
Count: 10
Binning | Y-Axis: (none)
Color By: Marital Status
Label By: Marital Status
Scatter | X-Axis: (default)

(Blue is single, light colored is young)

DATASET 3

This dataset seems to have more hired caucasians(image below), and again has more single young people hired, like dataset 2.


Binning | X-Axis: Hiring Success
Count: 10
Binning | Y-Axis: (none)
Color By: Ethnic Group
Label By: Ethnic Group
Scatter | X-Axis: (default)

Analysis of how these biases could affect hiring outcomes.

The algorithm description, it states "The design relies on a comprehensive dataset containing candidate profiles and historical hiring outcomes, which has significantly impacted the algorithm's development and functionality.". If these datasets are not balanced, I could see machine learning leaning more towards what it is familiar with and was trained on potentially giving certain candidates an advantage.

Datasets 2 and 3 have more hiring for single and young people. This could cause problems with the algorithm potentially being biased towards younger or single candidates being desired. Also for dataset 3, there was a higher distribution of caucasians being hired, so again, if this demographic is not handled correctly, the algorithm could opt to hiring more caucasian people due to being trained off of seeing more of them being hired.

In the data normalization portion of the algorithm, I could see potential issues of fairness with missing or incomplete data. It states that incomplete profiles missing data will be intelligently filled or flagged for manual review. If the algorithm handles the missing information itself, by filling something in, then this could cause fairness issues. These people with incomplete resumes could have false information added to their profile by the algorithm which could in turn affect their ranking in the algorithm depending on what the algorithm decides to handle this missing data. This is a problem, because missing data should just be flagged and not auto filled for a more fair option.

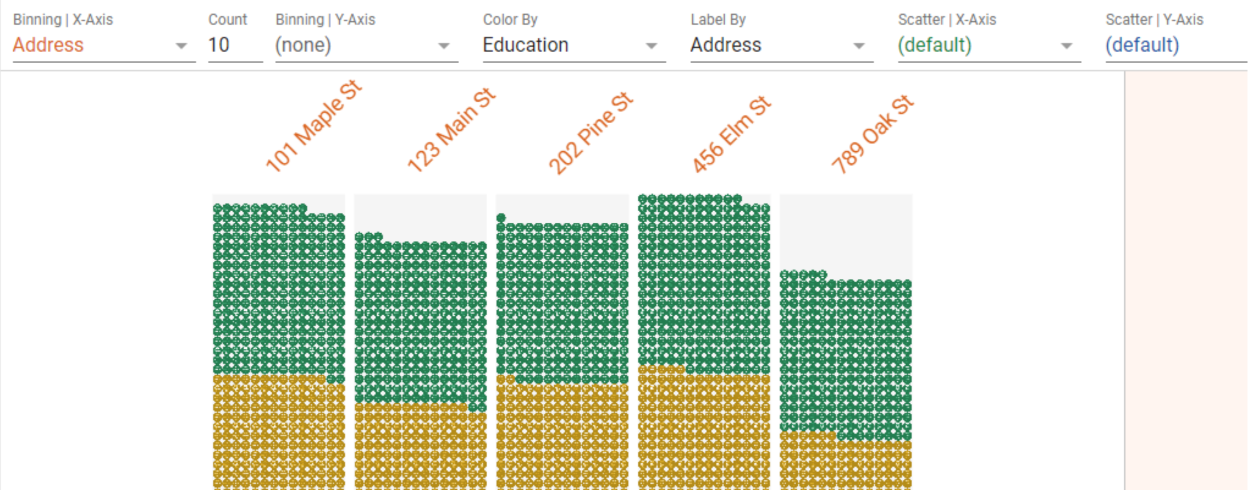
3.2 Privacy Concerns

Examination of privacy risks based on the data used.

Phone Numbers,email,addresses and Names are all areas with privacy risks. Each of these is collected in these data sets and used for making profiles for the candidates and helping with the interview scheduling assistance part of the algorithm that helps suggest interview schedules. While this sensitive information is important to be able to contact the applicants back, handling of this sensitive information should be done with extreme care so that privacy of the candidates is not breached. While the algorithm says that it anonymizes or excludes demographic details from the dataset unless relevant and consented to, the following analysis will check the datasets to see if this handling is being done.

DATASET 1

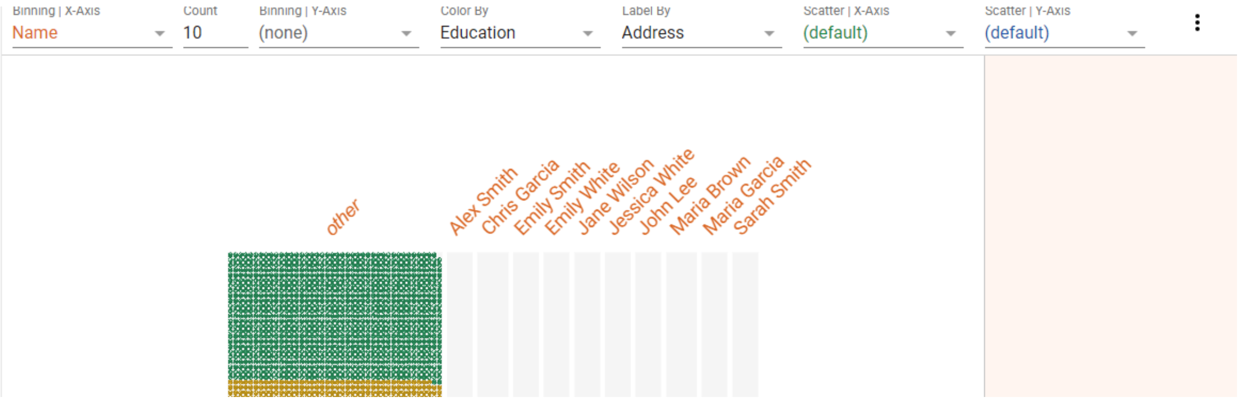
Data set 1 seems to correctly handle showing the sensitive information for im assuming consented peoples, and then having a separate category for others for the people who wish to hide the information. The only possible issue I could see is with the address category. I'm not sure if they gave everyone random addresses from these 5 categories, or if these are the actual addresses and none are being omitted for privacy (since this is training data), but this would cause privacy concerns. If none of this information was omitted, even though other categories clearly had things like their name or phone number filled with garbage information, then this would be an example of a privacy concern in the dataset. However, if it just autofilled everyone's address, then the algorithm is working as intended and correctly concealing the address and private information of the candidates.(Below is an image of the possible privacy issue with the address. There is no other category like "other" unlike the other sensitive information)



Binning | X-Axis Address
Count 10
Binning | Y-Axis (none)
Color By Education
Label By Address
Scatter | X-Axis (default)
Scatter | Y-Axis (default)

DATASET 2

As for dataset 2, it seems to be the same in terms of omitting personal information when needed. As seen below, it has another category for people wishing to hide the names in the dataset.



Binning | X-Axis Name
Count 10
Binning | Y-Axis (none)
Color By Education
Label By Address
Scatter | X-Axis (default)
Scatter | Y-Axis (default)

And, Like dataset 1, also has the same 5 addresses, so its either working correctly giving these random 5 addresses to each person, hiding their addresses, or there is an issue and the addresses are being displayed without protection of privacy.

DATASET 3

Dataset 3 also seems to be the same as the other two, so it is safe to say that the datasets are consistent in the handling of the sensitive information like name, email, address, etc. Again, the only possible issue could be with the addresses since it was the only one without another category, but again, maybe this was implemented differently to give everyone one of these 5 random addresses to protect privacy.

Identification of sensitive information (e.g., age, gender, race) and whether its use aligns with privacy standards.

The identification of this sensitive demographic information is valid since it is a hiring algorithm and this data is used to prevent bias. However, after analyzing the datasets, age, marital status, and ethnicity, should be carefully handled, because if not, it could lead to bias and unfairness in the algorithm.

3.3 Transparency and Explainability (Alex)

The algorithm takes into consideration the key features being: gender, marital status, ethnic group, age, education, experience in various programming languages (Java, SQL, Python, ML). The education and coding experience is then used to give each job candidate their own individual job requirements match score value, which shows how well the candidate fits the job description requirements. The algorithm also creates a hiring success value, which calculates the chance of the job candidate succeeding in the hiring process by taking into account the age, gender, marital status and ethnicity of the person. Overall, dataset hiring can be explained as the most qualified candidate in terms of experience being hired for the most part. There are some potential instances for bias towards individuals who are either caucasian, young or single, which cannot be explained as these things should not be considered in terms of hiring job applicants. The dataset is overall transparent in terms of what leads to candidates having a high hiring success value.

---

4. Recommendations

- **4.1 Bias Mitigation**
  - To reduce bias in the hiring algorithm, one way is to remove details like age, race, gender, or marital status if they aren't needed for the job. This can sometimes create unfairness, especially if the training data includes more people from certain groups. If removing them completely isn't possible, their impact should be lessened so that the focus stays on skills and qualifications. The algorithm should make sure it's fair for everyone, and it should be regularly checked for bias. We could also have humans review the decisions made by the algorithm to prevent unfairness amongst applicants.
- **4.2 Privacy Improvements**
  - To better protect privacy, personal details like addresses, names, and phone numbers should be hidden or replaced with fake information when possible. For data anonymization, we can remove or alter personal details so that individuals cannot be identified. For example, replacing names with unique codes or removing specific identifying information like addresses. Adding random data that doesn't affect the overall trends can make it harder to identify specific individuals. For example, if an exact age is known, you might add a random value to it, so it's no longer clear if someone is exactly 25 or 30 years old. We should also limit the amount of personal information being collected, only using what's necessary for the hiring process.
- **4.3 Enhancing Transparency**
  - To enhance transparency and explainability of the algorithm's decisions, rationale behind the hiring decisions should be elaborated upon. Key features such as experience in languages and education should be taken into account the most, so auditing/monitoring the algorithm to mitigate bias will be the best approach. This will allow job candidates to be fairly critiqued on their skills pertaining to the job description requirements instead of unimportant qualities such as gender or ethnicity. Creating a feedback loop for candidates will also be beneficial in

enhancing transparency, as a summary of how their resume was evaluated and ranked amongst the other candidates can help explain to each candidate why they were considered or not for the position.

---

5. Conclusion

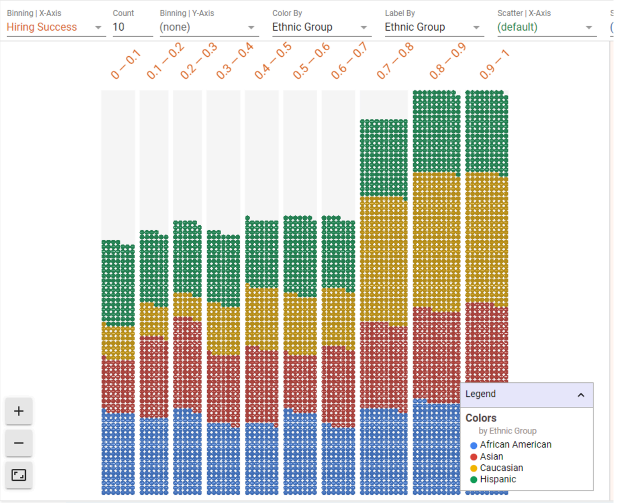
In conclusion, several ethical issues have been identified with the hiring algorithm, particularly concerning bias, privacy, and transparency. The algorithm tends to favor younger, single, or Caucasian candidates, which could result in unfair hiring outcomes. Privacy concerns also arise from how sensitive information, like addresses and personal details, is handled, indicating a need for better data protection. Additionally, improving transparency in the algorithm's decision-making process is essential to ensure that hiring managers and candidates understand why certain applicants are chosen. To address these issues, it's recommended to reduce the impact of demographic factors, enhance privacy protections, and make the algorithm's decisions clearer and more transparent for everyone involved.

---

6. Appendices

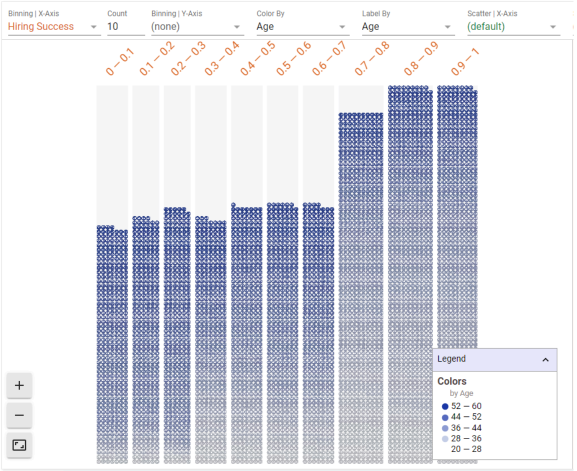
6.1 Data Analysis Screenshots

- ○ Screenshots from Facets Dive showing any identified biases.
- ○ Explanations of each screenshot and its relevance to the ethical concerns.



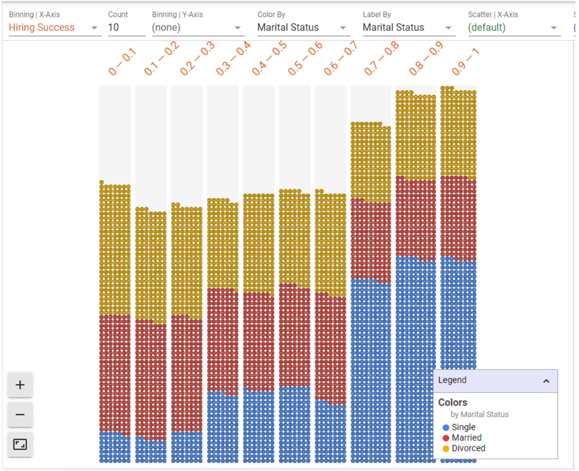
Binning | X-Axis
Hiring Success
Count
10
Binning | Y-Axis
(none)
Color By
Ethnic Group
Label By
Ethnic Group
Scatter | X-Axis
(default)

Here is bias on dataset 3 for hiring caucasian people (yellow). This could cause diversity issues and fairness issues in the hiring algorithm.



Binning | X-Axis
Hiring Success
Count
10
Binning | Y-Axis
(none)
Color By
Age
Label By
Age
Scatter | X-Axis
(default)

Here is bias that is present on dataset 3(same with dataset 2). The lighter colors indicate younger people, and you can see more younger people are being hired. This could lead to unfairness in the hiring algorithm having younger people be more likely to get hired.



Binning | X-Axis
Hiring Success
Count
10
Binning | Y-Axis
(none)
Color By
Marital Status
Label By
Marital Status
Scatter | X-Axis
(default)

Here is the bias that was found in marital status on dataset 3(same with dataset 2). You can see the single (blue) people are getting hired more, and this could cause issues with singles being preferred. This bias also was directly affected by the age bias since younger people have a higher percentage of single people.

References

1. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024, July 26). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence
2. *Facets - know your data*. Facets - Visualizations for ML datasets. (n.d.). https://pair-code.github.io/facets/index.html#facets-dive