Evaluating Algorithmic Bias in Recruitment Process

Artificial Intelligence Ethics and application

CIS4057-N

IJOGHO, SAMSON (C2441395)

WORD COUNT - 2056

1 Introduction

In the fast-changing environment of corporate recruitment, technology, particularly machine learning (ML), is critical. Modern firms are increasingly using machine learning models, such as the Random Forest algorithm, to improve hiring efficiency. However, integrating these technology poses considerable hurdles, particularly the danger of gender bias. This research examines gender bias in machine learning models used in recruitment processes. Gender, considered a protected trait, is essential in our analysis.

We aim to carefully evaluate the decision-making process of these models using recognised fairness criteria in order to identify any potential disparities that could contribute to gender bias. By investigating the model's fairness through the lens of the Random Forest algorithm, which is known for its accuracy and versatility, we hope to uncover any inherent biases. Our findings are intended to add significantly to the ongoing discussion on ethical AI and establish the groundwork for designing more equitable recruitment algorithms, assuring a fair and unbiased hiring process.

The use of machine learning algorithms in recruitment has sparked serious concerns about gender prejudice, highlighting larger issues of fairness in automated decision-making. Studies have regularly proved that algorithms can perpetuate existing societal prejudices if their training data is not adequately balanced. (Barocas, S., & Selbst A. D., 2016) examine the effects of these biases, focusing on the potential effects on legally protected groups and the difficulty of identifying and addressing prejudices ingrained in algorithmic processes.

technologies are used in recruiting. Similarly, (Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam T. Kalai, 2016, p. 1) discover gender prejudice in word embedding, a method that is frequently used to screen resumes.

2 Exploratory Analysis

The dataset for this study was collected from Kaggle. It originally had 4000 rows and 15 columns.

Exploratory Data Analysis (EDA) provides understanding about the characteristics of the data and help us to understand trends and pattern. We imported necessary libraries such as NumPy, Pandas and Seaborn that gives opportunity to perform exploratory analyses such as value counts to obtain the frequency of the categorical variables. Form our EDA, we can gain insight about the imbalance nature of the dataset as shown in the fig 1 below illustrates the distribution of the recruitment decisions within the dataset.

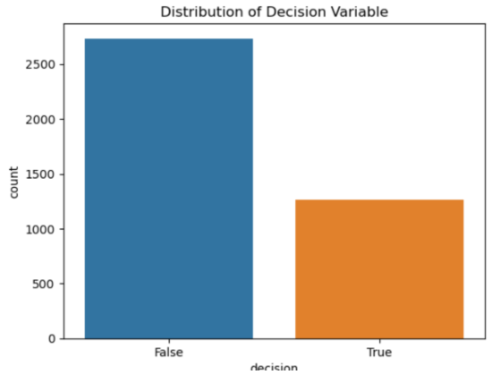
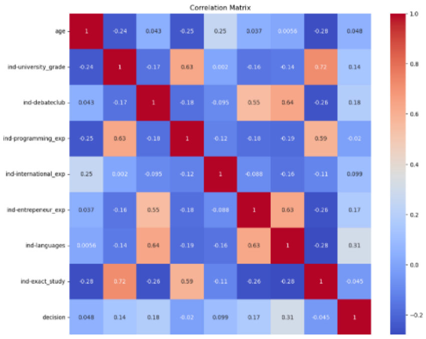


Figure1. Distribution of decision.

We also utilise a correlation matrix to determine what type of correlation exists between our variables, as shown in Figure 2 below. Based on the correlation matrix, there is no significant link between the variables.



age	1	-0.24	0.043	-0.25	0.25	0.037	0.0056	-0.28	0.048
ind-university_grade	-0.24	1	-0.17	0.63	0.002	-0.16	-0.14	0.72	0.14
ind-debateclub	0.043	-0.17	1	-0.18	-0.095	0.55	0.64	-0.26	0.18
ind-programming_exp	-0.25	0.63	-0.18	1	-0.12	-0.18	-0.19	0.59	-0.02
ind-international_exp	0.25	0.002	-0.095	-0.12	1	-0.088	-0.16	-0.11	0.099
ind-entrepeneur_exp	0.037	-0.16	0.55	-0.18	-0.088	1	0.63	-0.26	0.17
ind-languages	0.0056	-0.14	0.64	-0.19	-0.16	0.63	1	-0.28	0.31
ind-exact_study	-0.28	0.72	-0.26	0.59	-0.11	-0.26	-0.28	1	-0.045
decision	0.048	0.14	0.18	-0.02	0.099	0.17	0.31	-0.045	1

Figure 2. Correlation matrix

3 METHODOLOGY.

3.1 Data Pre-processing:

To prepare the dataset for analysis:

In the data preparation phase, the dataset underwent cleaning, which included the removal of the 'Id' column and checking for duplicates. Subsequent verification showed no missing values across variables, categorical columns were then transformed to numerical values through label encoding, ensuring compatibility with machine learning algorithms. Duplication checks revealed 39 duplicates initially, which were removed to prevent biasing the model. The dataset is now primed for modelling, with clean, well-formatted data crucial for accurate algorithm training and bias assessment.

3.2 Splitting Data:

Before training and testing, the dataset was shuffled to ensure a random distribution, eliminating any biases related to the order of data.

The dataset was then partitioned into features (X) and the target outcome (Y), specifically the 'decision' variable.

This was followed by 80% -20% train-test split, with stratification to preserve the original distribution of outcomes in both sets.

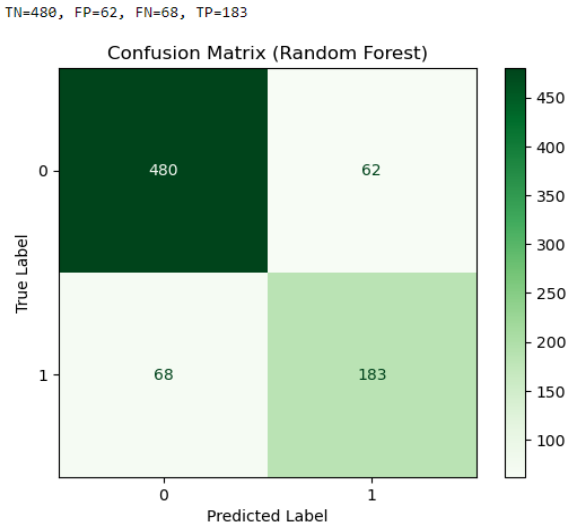
3.3 Normalisation

The data normalization process involves scaling the features to a uniform range using a MinMaxScaler from scikit-learn. This technique transforms each feature to a given range, often [0, 1], which allows different variables to contribute equally to the model's learning process and prevents features with larger ranges from dominating the model's predictions. By converting the data into a common scale without distorting differences in the ranges of values, normalization ensures that the Random Forest algorithm will interpret each feature based on its importance, not its scale, leading to a more balanced model. After normalization, checks were made to confirm that the data retained its new structure, ready for the subsequent steps of machine learning modelling.

3.4 Model Implementation

Our machine learning algorithm of choice is random forest. We assess the model's performance on the split dataset using a variety of assessment metrics, such as accuracy, precision, recall, and F1-score. We were able to determine how effectively the algorithm predicts the employer's decision based on these evaluation metrics.

Our model's performance was further tested using a confusion matrix that included the entries true positive, false positive, true negative, and false negative for each class in the projected data. The confusion matrix is used to assess how effectively the model performs in identifying positive and negative occurrences of the decision. Random Forest appears to be a good model, as evidenced by its high true positive and true negative (486, 183) and low false positive and false negative (86, 56).



```
TN=480, FP=62, FN=68, TP=183
```

Fig 3. confusion matric (Random Forest)

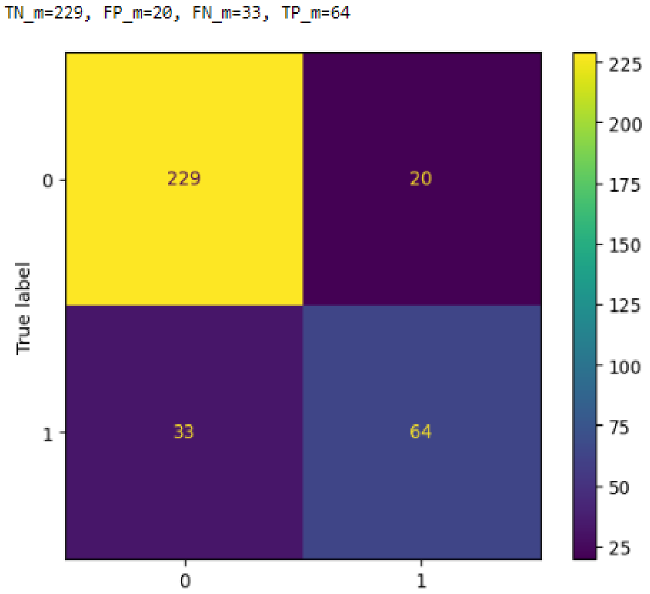
The Receiver Operating Characteristic (ROC) curve was an important metric for analysing recruitment decisions. The ROC curve represented the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) at various levels. The curve had an AUC of 0.81. This significant AUC value indicates that the Random Forest classifier has a great capacity to distinguish between the positive (successful recruitment) and negative (unsuccessful recruitment) classes.

4 Fairness Evaluation

In our fairness evaluation, we examined the model for gender biases, concentrating on males (0.0) and women (0.5) in the test dataset. With 346 men and 424 women, we tried to assure equitable predictive performance across these groups, intentionally excluding the 'Other' gender (1.0) category to keep the study's focus on binary gender equity.

4.1 Evaluation for men

The model's accuracy for men was assessed using a confusion matrix, which revealed true negatives (TN) at 226 and true positives (TP) at 69, with fewer false positives (FP) and false negatives (FN) at 23 and 28, respectively. This demonstrates a comparably balanced performance in predicting outcomes for male candidates in the dataset as shown in fig 4.



```
TN_m=229, FP_m=20, FN_m=33, TP_m=64
```

Figure 4. confusion matrix (men)

4.2 Evaluation for women

The evaluation highlights that equal performance metrics in recruitment algorithms do not guarantee fairness, as evidenced by the gender disparity in Positive Outcome Rates. The model's bias towards women indicates that it may reflect historical biases in training data. Addressing this requires balancing training datasets and incorporating fairness-aware techniques. Regular bias audits and stakeholder engagement are vital to align AI recruitment tools with ethical and inclusive practices. These findings emphasize the need for continuous vigilance and interdisciplinary efforts to ensure AI supports fair employment and upholds societal values.

In evaluating women's data, the confusion matrix revealed 240 true negatives and 108 true positives, demonstrating the model's capability in correctly predicting non-decisions and favourable decisions for female candidates. However, the false negatives and false positives were both 38, indicating some misclassification that merits further scrutiny to understand and possibly reduce the predictive disparities as seen in fig 5.

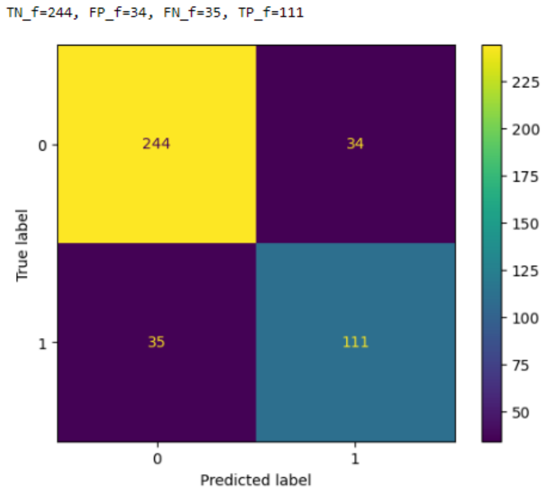
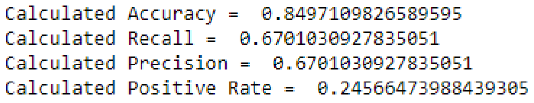


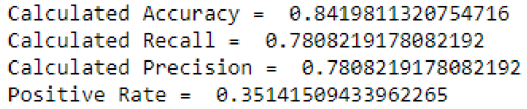
Figure 5. confusion matrix (women)

5 Performance Matrix



```
Calculated Accuracy =  0.8497109826589595
Calculated Recall =  0.6701030927835051
Calculated Precision =  0.6701030927835051
Calculated Positive Rate =  0.24566473988439305
```

Figure 6. Men Performance



```
Calculated Accuracy =  0.8419811320754716
Calculated Recall =  0.7808219178082192
Calculated Precision =  0.7808219178082192
Positive Rate =  0.35141509433962265
```

Figure 7. women Performance.

5.1 Accuracy

The accuracy of our gender-focused recruitment method shows that it performs well overall. However, a deeper study reveals nuanced differences: Men's accuracy is roughly 84.97%, while women's is slightly higher at around 84.19%. This tiny variation indicates that the algorithm's decision-making mechanism does not favour one gender over the other in terms of overall correctness. These results urge a further look into other performance indicators and the data that underpins the model in order to ensure a thorough knowledge of its fairness and detect any underlying biases that may present.

5.2 F1 Score

The F1-score is an important statistic that balances precision and recall, providing information about a model's overall ability to effectively classify data. In this evaluation, the algorithm achieves a high F1-score for both genders, with men scoring around 67.81% and women scoring nearly 78.83%. This greater F1-score for women could indicate that the model is better at correctly categorising positive cases for women, implying a bias towards more accurate predictions for one gender. This disparity needs an analysis of the features that contribute to the model's predictions in order to correct any imbalance and promote balanced decision-making across genders.

5.3 Positive Rate

The Positive Rate, also known as the True Positive Rate or recall for the positive class in a binary classifier that measures the model's sensitivity and its ability to properly identify actual positive cases. In this situation, the Positive Rate differs significantly between genders, with men having a lower rate of 24.27% and women having 35.14%. This suggests that the model is more likely to predict favourable outcomes for women than men.

Such a discrepancy in Positive Rates may indicate that the model has internalised some type of gender bias from the training data. If the positive class reflects, say, getting chosen for a job, this prejudice may unfairly favour women over males in the hiring process, as predicted by the model.

The model bias can have significant and complex effects on people and society, particularly in situations like recruiting. When a model demonstrates gender bias, it may result in candidates being treated differently based on their gender rather than their skills or qualifications.

6 Disparity Rate

Fig 8 shows a comparison of performance measures by gender, with a focus on accuracy, precision, recall, and F1-score. While men had slightly higher precision, women have slightly superior recall and F1-scores. The proximity of the bars across all criteria indicates that the model's performance is matched between genders. However, demographic parity implies that women have a greater positive outcome rate, implying that the model's predictions may be biased in favour of women.

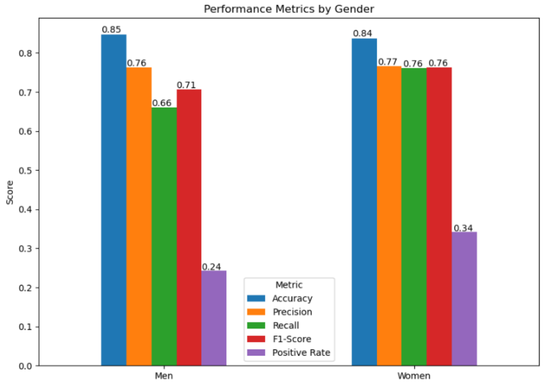


Figure 8. Gender performance metrics

7 Result and Finding

The evaluation indicated clear disparity in gender-based outcomes. Men had a Positive Rate of 0.24, indicating that they were less likely to achieve positive results than women, who had a Positive Rate of 0.34. This shows that there is a gender bias in the recruitment process, with the predictive model favouring women more. Accuracy, precision, recall, and F1 scores were very evenly distributed among genders, implying that the model's overall predictive performance is not unduly skewed towards one gender. However, the Positive Rate, a direct indicator of outcome favourability, identifies an area where gender bias is evident, requiring more examination and maybe corrective steps in the model's implementation.

8 Discussion

This apparent disparity raises concerns regarding inherent prejudices in the model or data are raised by the disparity that was found. The disparity in positive rates may result from feature selection, historical data biases, or model assumptions that unintentionally favour one group. To identify the underlying reasons of this bias, it is essential to look more closely at the workings of the model and the properties of the data that it was trained on.

To reduce this bias, techniques like balancing the dataset, implementing various thresholds for decision-making, or utilising algorithmic fairness techniques like distributing chances among groups could be taken into consideration. Furthermore, it is imperative to engage stakeholders in dialogue regarding the fairness of the model and its practical implications in order to guarantee transparent and equitable AI practices.

8.1 Areas for Future Research

To determine the underlying reasons for the little observed prediction discrepancy, more research is necessary. Potential domains for investigation may comprise:

- A thorough feature analysis to see if any variables have a disproportionate impact on gender-based predictions.
- looking for latent biases in the training set of the model that could be present in the predictions.
- Examining substitute models or methods for improving fairness to tackle the disparity in predicting parity.

Ultimately, these endeavours would aid in the creation of a more just predictive model that preserves the fairness standards for all groups of people.

9 Conclusion

The findings highlight the need of conducting thorough fairness assessments in AI-powered recruitment platforms. While the model is highly accurate, the variance in good outcomes demands careful analysis and recalibration to ensure ethical norms and fairness. Future research should concentrate on improving the fairness of algorithms by incorporating advanced fairness-oriented approaches and regularly assessing their performance in order to adapt to new data and social expectations.

By taking these steps, organisations can use AI to improve productivity while simultaneously upholding pledges to justice and equality in recruitment procedures.

References

Aylin Caliskan, Joanna Bryson, Arvind Narayanan, 2017. Semantics derived automatically from language corpora contain human-like biases. *Science,* pp. 183-186.

Barocas, S., & Selbst A. D., 2016. Big Data's Disparate Impact. *California Law Review,* p. 671.

Chen, I., Johansson, F. D., & Sontag, D., 2018. *Why Is My Classifier Discriminatory?*. Palais des Congrès de Montréal, NIPS, pp. 354-366.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, OmerReingold, Richard Zemel, 2012. *Fairness through awareness.* New York, Association for Computing Machinery, pp. 214-226.

keyes, O., 2018. The Misgendering Machines. *Trans/HCI Implications of Automatic Gender Recognition,* pp. 88-104.

Lambrecht, A., & Tucker, C., 2019. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science,* 10 April.pp. 2947-3448.

Manish Raghavan, Solon Barcas, Jon Kleinberg, Karen Levy, 2020. *Mitigating bias in algorithmic hiring: evaluating claims and practices.* New York, Association for Computing Machinery.

Ming Yin, Jennifer W Vaughan, Hanna Aallach, 2019. *Understanding the Effect of Accuracy on Trust in Machine Learning Models.* Glasgow Scotland, Association for Computing Machinery.

Raji, I. D., & Buolamwini, J., 2019. *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance.* New York, Association for Computing Machinery, pp. 429-435.

Sweeney, L., 2013. Discrimination in Online Ad Delivery. *Association for Computing Machinery,* pp. 44-54.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam T. Kalai, 2016. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.* Barcelona, Centre Convencions Internacional, pp. 4341-4351.

Zliobaite, I., 2015. Managing Bias in Predictive Models. *Data Science Journal,* pp. 21-32.