Project Report – Observing and Eradicating Implicit Bias found in Recruitment Models

Aryan Prakash Rao

I. Project Proposal

A. Introduction

In this project, I will be attempting to remove implicit Bias found in recruitment/hiring machine learning models. An implicit bias is an unconscious association, belief, or attitude toward any social group [1]. This sort of Bias is usually more disadvantageous to minority groups. This poses some bad implications in recruitment processes in companies. Companies have started actively looking to attract a more diverse repertoire of candidates through various methods as this creates a more positive work environment. These include blind resume screenings or female-only career days [2].

A similar sort of Bias can be found in machine leaning models. This sort of Bias could be referred to as a 'Prejudice Bias'. This sort of Bias occurs when a dataset reflects common prejudices/stereotypes found in the real world [3]. For example, using data about Computer Science recruitment where the ratio of male to female is 9:1 could perpetuate a real-world gender stereotype, creating a positive correlation between being a male and aptitude for Computer Science.

The way I plan to mitigate this, is by applying the Rooney-Rule as a mathematical constraint to the model during Post-Processing. The Rooney Rule is an NFL (National Football League) policy where all teams are required to interview (not hire) a certain number of ethnic-minority candidates for specific managerial positions. This rule was first established in 2003 and variations of it are now present in various unrelated industries. Through the application of this rule, the overall percentage of African American coaches in the NFL was up by 16% in 2006 [4].

I will be using a ranking model using regression combined with a subset- selection algorithm. A ranking model is one in which each item is given a rank based on their perceived utility, higher the rank, higher the utility (Eg. Deciding amount of Bonus given to employees). A subset-selection algorithm involves choosing a set of items with the highest perceived utility from a dataset (Eg. University Recruitment based on exam scores). This model is very similar to a binary classification model as there are always only 2 potential outcomes for each item in a dataset.

This model is susceptible to 'Prejudice Bias' and is an apt candidate for the application of the Rooney Rule. It is to be noted that my goal is to create a shortlist of potential candidates to be chosen for a particular role; I do not plan on explicitly choosing candidates for a particular role. This follows the Rooney Rule very closely, as the rule requires disadvantaged candidates to be interviewed, not selected.

B. Motivation

This problem is a very relevant one in this day and age, it is something a lot of us have encountered. I personally may have even faced this sort of Bias in some of my university applications 2 years prior. You can never truly know if this Bias has affected you. This sort of uncertainty undermines the fairness of recruitment policies. To uphold a certain sense of fairness, it is essential to have measures in place to uphold the integrity of these processes.

A machine learning model is a great choice for shortlisting candidates to be considered for a job or, a seat at a university. This would nullify possible implicit biases of recruitment officers. The only task left is to mitigate possible biases in the dataset and machine learning model itself. This is the problem I have set out to solve.

C. About the Dataset

The Dataset [5] can be downloaded at [6]. The dataset contains results of the IIT JEE 2009 Entrance Exam. It contains the Physics, Chemistry and Math mark of each candidate along with their Full Name, Parent's Name, Gender, Social/Financial Category along with some registration information.

Sadly, recruitment statistics are not available for the 2009 Exam. I will instead use the 2011 recruitment statistics to make comparisons[7]. Refer to the Data Analysis part of the project progress section for all data analysis.

D. Tasks to Complete

Firstly, I would have to prepare the dataset so it can be used as an input for the algorithm. There is a lot of unnecessary data within the dataset which can safely be removed. Some data points need will need adjusting as well. I will then implement a conventional machine learning model to rank students, allot IIT seats and visualize these results.

Finally, I will then implement a fair machine learning solution loosely based on the fairness article I have chosen [8]. It is to be noted that the paper's[8] goal it to remove bias algorithmically and there is no mention of machine learning. So, I cannot really follow the algorithm's constraints.

I will be trying out a regression based ranking model and use a subset-selection algorithm for the shortlisting part. The main benefit of a ranking model, in this case, is that there are multiple IIT institutions, some better than others; the rankings could be used to decide who gets into which IIT. I feel the ranking model proposed by the article does not suit my goals as it blindly ranks underprivileged groups to the highest positions. Therefore, I will use a simple regression model instead. I will be factoring in the Rooney Rule in the subset-selection part of the solution.

E. Technologies to be used

All implementation aspects will be done through python, all code will be written. in .py files. I will be using the pandas, numpy, matplotlib and scipy(sklearn) packages for the implementation.

II. Data Analysis

A. First Look

384,977 candidates attended the 2009 rendition of the IIT JEE exam. The dataset[5][6] contains the Physics, Chemistry and Math mark of each candidate along with their Full Name, Parent's Name, Gender, Birth Category along with some registration information for each candidate.

The 'Gender' and 'Category' sections are areas of potential Bias. 'Gender has 2 potential values; M: Male; F: Female. 'Category' has 5 potential values; GE: General; ON: Open; SC: Scheduled Castes: ST: Scheduled Tribes; OC: Other Backward Castes. The ON and GE category can be classified together as they are roughly the same category. Refer to Fig 1 and 2 for gender and category demographics. I have catalogued more detailed stats in the '*Key Statistics'* section. This is because, I have made some required adjustments to numerical data within the *'Data Cleaning and Transformation'* sub-heading.

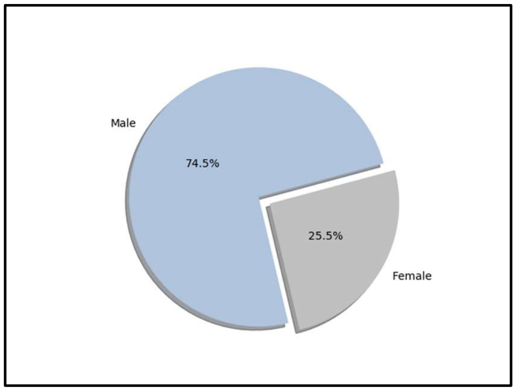

Fig 1: Gender Ratio of Candidates

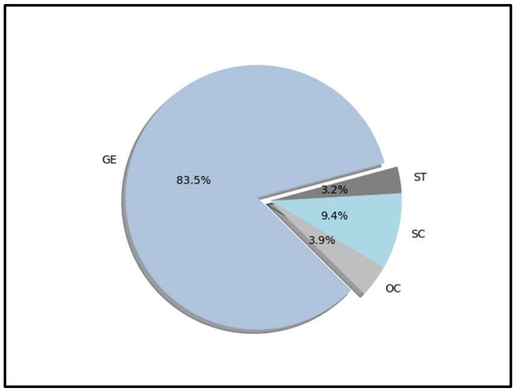

Fig 2: Category Ratio of candidates

B. 2011 Recruitment Statistics

As 2009 recruitment data is not available, I will be referring to the recruitment statistics of the 2011 exam[7]. This data can be used to get a good idea of what the recruitment process looks like. In 2011, 468,280 students attended the exam and only 13,196(roughly 2.8%) were admitted. 23.2% of all candidates were female but only 10.9% of the students admitted were female. This already indicates there might be some bias as you would expect the % of females to be closer to 20%. The stats for the OC category also look like this; OC made up for 29% of the candidates but only 19% of the admitted students belonged to the OC category. This may indicate potential bias based on Gender and Category.

I will attempt to create a shortlist consisting of 7% of all candidates to be considered for selection.

C. Data Cleaning and Transformation

The raw dataset contains eleven columns, when reading the dataset, I only read 6 columns. The 7 columns included are, REGST_NO, GENDER, CATEGORY, MATH, PHYSICS and CHEMISTRY. The columns left out are the SUB_CATEGORY, PIN_RES (residence pin), PARENT NAME , TOTAL_MARKS and NAME. The sub-category was left out due to it containing around 99% of the same value and the lack of recruitment information available for the underprivileged group. The parent's name, student name and the residence pin aren't really required as I already have a unique identifier(REGST_NO).The total marks were left as I will be calculating my own total marks using the individual MATH, PHYSICS and CHEMISTRY scores.

The GENDER and CATEGORY values are converted to integers, {"M":0, "F":1}, {"GE":0, "SC":1, "OC":2, "ST":3}. As you can see, I have given larger values to smaller demographics, this will benefit minority groups in the fair solution. I have converted the range of individual subject scores from -35:165 to 0:200. The PHYSICS mark on average is lower than MATH and CHEMISTRY marks, I have manipulated these MARKS based on the average of averages (average of the 3 subject averages). This caused an increase in average PHYSICS marks and slight decreases in the MATH and CHEMISTRY averages. These scores and then used to calculate the TOTAL MARKS, signified by the equation:

$$T = (M + P + C)^{x} \tag{1}$$

The value x is the scoring coefficient, this mediates the effect of the TOTAL MARKS on the machine learning model. This will vary for the conventional and fair solutions. After the TOTAL MARKS is calculated, the MATH, PHYSICS and CHEMISTRY columns are removed as they are no longer required. I have observed no bias within the dataset itself other than the Physics score being law.

D. Key Statistics from the Dataset

Stat	Gender Statistics		
	Overall	**Male**	**Female**
Frequency	384,977	286,942	98,028
Frequency(%)	100.0%	74.5%	25.5%
Mean Score	133.7	136.1	126.5
Variance of Score	2424.8	2702.3	1543.9

Stat	Category Statistics			
	GE	**OC**	**SC**	**ST**
Frequency	321,266	15,109	36,177	12,484
Frequency(%)	83.5%	3.9%	9.4%	3.2%
Mean Score	137.4	118.5	114.9	112.0
Variance of Score	2607.1	1531.2	1002.7	813.4

III. Conventional Implementation

For the conventional implementation, I have chosen a linear regression algorithm to assign a ranking to each candidate based on TOTAL MARKS. The Y Column used to train and test data is the percentile (as an integer) of each candidate. The highest scorer has 100, while the lowest scorer has 0. After the pre-processing techniques are applied the bottom 60% (Score wise) of the dataset is completely removed as this will only increase computation time and lower the accuracy. As it is, there is no chance a low scoring candidate would pass the entrance exam.

This parsed dataset is then split into train and test using sklearns train_test_split method. The ratio of train:test is 75:25. The trained model generalized very well to the testing dataset, obtaining an accuracy of 93.3% accuracy. The scoring coefficient (x, refer to equation 1) was set to 0.85 to maximize the accuracy. The accuracy was calculated using the following equation:

$$A(\%) = 100 * \frac{R(\sum|p(i) - t(i)| i \in l)}{R} \tag{2}$$

l: length of testing dataset, i: each member of the testing dataset, p(i): predicted Y column, t(i): actual Y column. R is the difference between smallest and largest predicted value

After obtaining the rankings through the model, I then decide who gets shortlisted. The top 7% of all candidates according to the predicted model are shortlisted for further consideration. For the testing dataset, 6949 candidates were shortlisted. 6018 (86.6%) were male and 931 were female (13.4%). 6667 were GE(95.9%); 103 were OC(1.4%); 148 were SC(2.1%) and 31 were ST(0.4%). These were the results obtained without any diversifying measures.

I then applied diversifying measures to the dataset, the ratio of Male:Female was made 51:49. The ratio of GE:OC:SC:ST was made 32:33:28:6. Refer to Fig 1 and 2 for previous ratios. To diversify the dataset, I had to remove a lot of datapoints, so the resulting testing dataset is much smaller (around 90% smaller). For the diversified testing dataset 649 candidates were shortlisted. 522 were male(80.4%) and 174 were Female(19.6%),. 507 were GE(78.1%), 144 were OC(22.2%), 42 were SC(22.2%) and only 3 were ST(0.5%). The accuracy of the model with a diverse dataset is still high at 92.4%.

Interestingly, the majority groups still tended to dominate the shortlist even when the dataset was diversified to include similar amounts of each group. Although, this also indicates there is some bias as shortlisted demographics did shift towards the minority groups. In India, more emphasis on education (especially in technical fields) is given to Males than to Females. This is probably why males tended to outscore females. The GE category dominating can be attributed to a better education and a greater access to resources. This is because the SC, ST and OC categories are generally financially deprived. I believe the lower averages of minority groups are mostly due to societal issues in India. A minority candidate with a lower score might have higher utility than a majority candidate with a higher score as the majority candidate would have more access to resources. In the fair solution, I will attempt to account for these situations. I will try to solve this problem by applying a Rooney-Rule-Like algorithm loosely based on my research paper.

IV. Fair Implementation

The fair implementation is somewhat similar to the conventional solution but there are some key differences. For the fair solution I will instead be using a logistic regression model. This is to counteract societal advantages of majority groups, this benefits minorities as the integer values to signify a minority group is higher than the one given to a majority group (Refer to Data Analysis part C). This causes minority candidates to achieve higher ranks. I have used the same Y values as the conventional implementation. I have also sub-sampled the dataset to include a diverse set of candidates. The ratios are GE:OC:SC:ST -> 32:33:28:6 and M:F -> 51:49 for category and gender. The scoring coefficient (x, refer to equation 1) was set to 2.9 to increase accuracy. This sub-sampled dataset is used to generate training and testing data. This data is then fed directly to the Logistic Regression model. The predicted rank from this model is parsed onto the testing dataset and then fed to my subset-selection algorithm.

This algorithm ensures at least α(0.07×S) candidates from each category are shortlisted. 'S' is the size of the category and α (0< α<1) is a constant used in calculations. I have used an α of 0.6 for calculations, but this can be adjusted as seen fit. Increasing the α will make the shortlisted subset resemble the original demographics more closely. This factor of the algorithm enforces the Rooney Rule. After the Rooney's Rule constraint is satisfied of, subset spaces are assigned on a merit basis. Refer to Fig 3 for a comparison of the results of a conventional solution versus a fair solution. Both algorithms were applied to the same diverse sub-sample of the dataset.

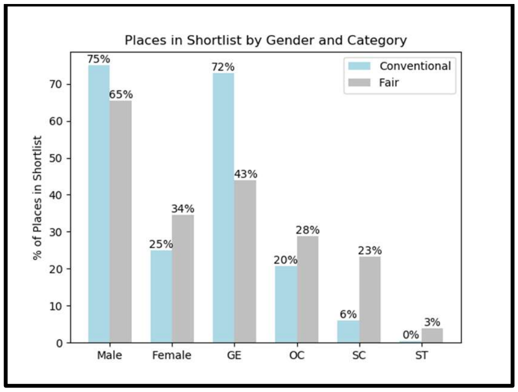

Fig 3: Comparison of Demographics, Conventional vs Fair

As you can see, the number of places allotted for all minority groups have increased. This is a clear indication of reduction in bias. I have clearly got different results from the paper. The paper didn't even consider the CATEGORY of candidates. I have used a completely different algorithm to my paper's algorithm and my goal was completely different. The paper also uses different metrics/ visualizations to showcase results. The results of my algorithm will vary for different α values. My algorithm manages to create a diverse subset of shortlisted candidates with decent accuracy. This was my goal and I have achieved it.

The accuracy of the algorithm is 84.2% (calculated the same way as the conventional algorithm). This is 10.2% decrease from the conventional model's accuracy (92.4%). This is mainly due to the fair solution using a logistic regression model rather than linear regression one. The logistic regression model considers integer values assigned to the GENDER and CATEGORY features.

References

[1] K.Cherry, How Does Implicit Bias Influence Behaviour?, https://www.verywellmind.com/implicit-bias-overview-4178401

[2] C.Pavlou, Unconscious bias in recruitment: How can you remove it?, https://resources.workable.com/stories-and-insights/unconscious-bias-in-recruitment

[3] M.K.Pratt, machine learning bias (AI bias), https://searchenterpriseai.techtarget.com/definition/machine-learning-bias-algorithm-bias-or-AI-bias

[4] B.W.Collins, Tackling Unconscious Bias in Hiring Practice: The Plight of the Rooney Rule, New York University Law Review. 82 (3): 870–912

[5] H.Alderman and E.M.King, Gender differences in parental investment in education. Structural Change and Economic Dynamics, 9(4):453–468, 1998

[6] A. Rana, IIT JEE 2009 Results, https://captnemo.in/projects/iitjee/ https://jumpshare.com/v/R3YeZko2gkeZ4P2Xkk4N

[7] JEE Team, JEE 2011 Results, https://iitk.ac.in/new/data/jee-report/Report%20JEE%202011_EDITED_Aug14.pdf

[8] Celis, L. Elisa, Anay Mehrotra, and Nisheeth K. Vishnoi. Interventions for ranking in the presence of implicit bias, FAT 2020