Kerr Tan (st4153), Joyce Zhang (kz2085)

Course Project: Technical Audit of an Automated Decision System

Background:

Due to the growing need for educated job candidates, a number of companies begin to utilize algorithms to optimize their recruitment methods and selection processes. The "Job Placement Dataset" on the Kaggle platform provides an algorithmic tool for the company to assess candidates during the fresh graduates' recruiting process. The dataset includes 12 attributes such as gender, educational background, work experience, and test scores to evaluate the competitiveness of the candidates and make a prediction concerning whether the candidate gets placed or not, indicated with "status" in the dataset. It collects information from 215 fresh graduates and labeled each graduate depending on its job placement prediction. This ADS' main goal is to develop a machine-learning model that helps the recruiter make a better judgment to select the most experienced and qualified candidates.

Input and Output:

Our data includes the following 7 categorical features and 6 numeric features (Kaggle Dataset):

1. gender : Gender of the candidate
2. ssc_percentage : Senior secondary exams percentage (10th Grade)
3. ssc_board : Board of education for ssc exams
4. hsc_percentage : Higher secondary exams percentage (12th Grade)
5. hsc_borad : Board of education for hsc exams
6. hsc_subject : Subject of study for hsc
7. degree_percentage : Percentage of marks in undergrad degree
8. undergrad_degree : Undergrad degree majors
9. Work_experience : Past work experience
10. emp_test_percentage : Aptitude test percentage
11. specialization : Postgrad degree majors - (MBA specialization)
12. mba_percent : Percentage of marks in MBA degree
13. status (TARGET) : Status of placement. Placed / Not Placed

Below is the graph that shows the data type of each variable:

#	Column	Non-Null Count	Dtype
0	gender	215 non-null	object
1	ssc_percentage	215 non-null	float64
2	ssc_board	215 non-null	object
3	hsc_percentage	215 non-null	float64
4	hsc_board	215 non-null	object
5	hsc_subject	215 non-null	object
6	degree_percentage	215 non-null	float64
7	undergrad_degree	215 non-null	object
8	work_experience	215 non-null	object
9	emp_test_percentage	215 non-null	float64
10	specialisation	215 non-null	object
11	mba_percent	215 non-null	float64
12	status	215 non-null	object

Figure 1: Datatype

The Kaggle platform does not clearly state how was the data collected. We assume that the company collected data through the questionnaire in the job application. It collects 215 fresh graduates' information, and each candidate has provided all answers to the 12 attributes, so the dataset contains no missing values and can be directly used for analysis (Figure 2).

```
data.isnull().sum()
```

```
gender                  0
ssc_percentage          0
ssc_board               0
hsc_percentage          0
hsc_board               0
hsc_subject             0
degree_percentage       0
undergrad_degree        0
work_experience         0
emp_test_percentage     0
specialisation          0
mba_percent             0
status                  0
dtype: int64
```

Figure 2: Missing Value

Our target variables focused on "status", which is a binary labeled 'Placed' and 'Not Placed' that predicts whether the candidate gets the job. For a candidate labeled "Placed", the model predicts that he/she is more likely to be selected for this job position. For a candidate labeled "Not Placed", the model predicts that he/she may not be selected for this job position.

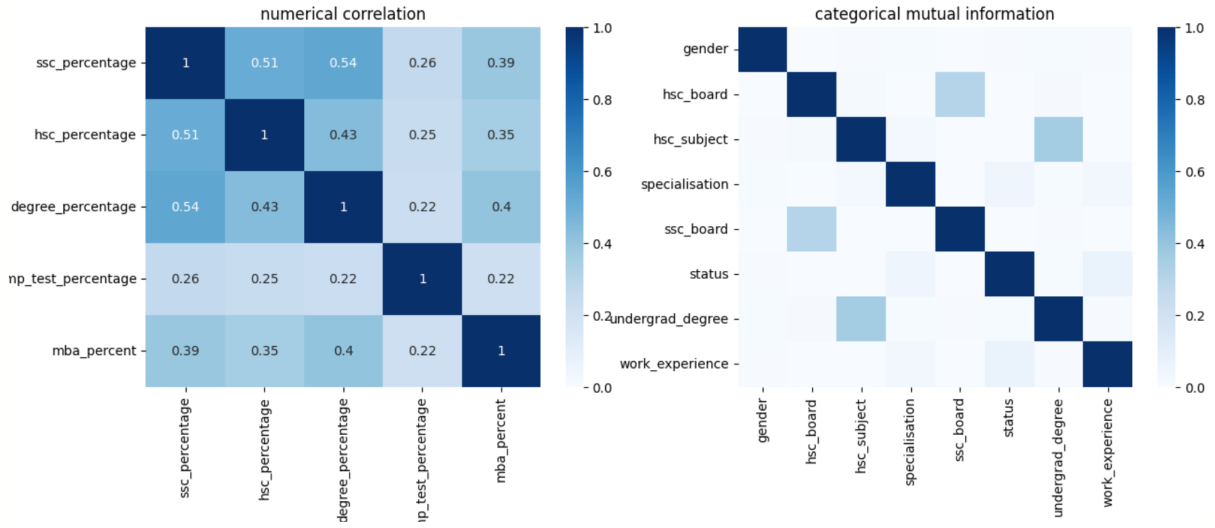


Figure 3: Correlation Heatmaps

According to the first graph in Figure 3, the "ssc_percentage", "hsc_percentage", "degree_percentage", and "mba_percentage" are correlated. "Ssc_percentage" and "hsc_percentage" refer to the senior secondary exam and higher secondary exam, which are both taken during high school. It makes sense that those who gain higher scores in secondary exams perform well in college, so it is reasonable for those variables to be correlated. The second graph of Figure 3 indicates that the protected variable "gender" does not show a strong correlation with other attributes. "Hsc_board" and "ssc_board" are correlated, which is reasonable for the two exams may be held by the same board. Additionally, "undergrad_degree" and "hsc_subject" has a strong correlation. Students may choose the subject they wish to pursue as a future career path to take the exam. As a result, all the correlations stated above will not impact the prediction model.

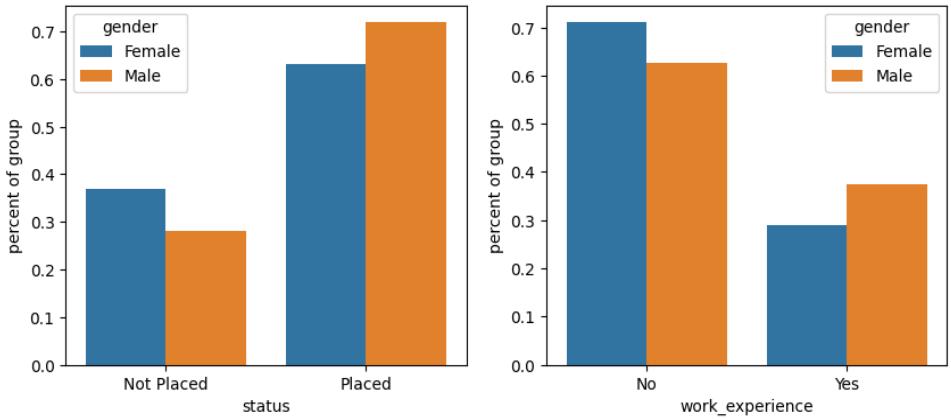


Figure 4: Gender Percentage in Placement Status and Work Experience

Before implementing the algorithm, it's necessary to understand the distribution of the target variable "status" in different gender. As shown above, females tend to have a higher rate of "Not Placed"

compared to males and tend to have a lower rate of “Placed” compared to males. On the other hand, females tend to have a higher rate for “no working experience” and a lower rate for “had working experience”.

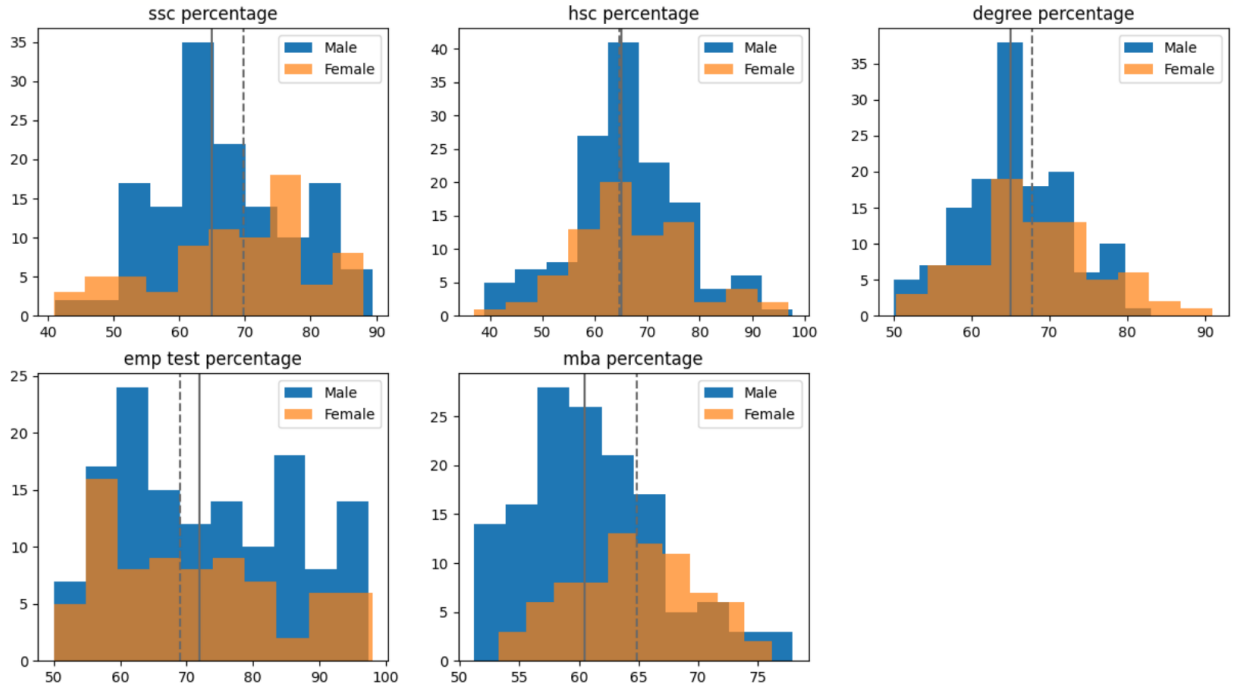


Figure 5: Scores Distribution in Different Gender Groups

The above attributes including “ssc_percentage”, “hsc_percentage”, “degree_percentage”, “emp_test_percentage” and “mba_percentage” are used to train the algorithm, representing different score results. Knowing the distribution between males and females associated with placement status shown as above, it’s also important to understand the distribution of score tests, interpreting working abilities for each group. The solid line in the graphs indicates the median of males for this specific test, while the dashed line indicates the median of females for this specific test. Besides “emp_test_percentage”, all test scores distribution graphs show that females tend to have higher median scores compared to males. However, with the results that females have a lower rate in job placement status from graphs of the distribution between males and females in job placement status (Figure 4), it can be seen that the dataset has a pre-existing bias against females. If implementing a machine learning model without addressing, it might also evoke technical bias against females by labeling them as “Not Placed”.

The output of the ADS will be a label for each candidate provided with training attributes, either “Placed” or “Not Placed”, respectively representing most likely to be selected or not likely to be selected.

Implementation and Validation:

The data does not contain missing values, so there is no data-cleaning process in the code. Since the dataset is quite small with only 215 respondents and all the strong correlations demonstrated in Figure 3 between variables are reasonable, the author did not include preprocessing algorithms in his code. Then, the author implemented 6 different machine learning models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Multinomial Naive Bayes. For each model, the code generated four metrics: accuracy, F1 score, recall, and precision. Our focus is the Logistic Regression Model, and its metrics output is shown below. Other models implemented by the author will have a similar layout in outputs.

```
Logisitic Regression
Model performance for Training set
- Accuracy: 0.8895
- F1 score: 0.888060
- Precision: 0.888324
- Recall: 0.889535
----------------------------------
Model performance for Test set
- Accuracy: 0.8837
- Fl score: 0.8821
- Precision: 0.8817
- Recall: 0.8837
==================================
```

Figure 6: Metric for Logistic Regression

For the logistic regression model, its accuracy for both the training and testing set is 0.89 and the F1 score is 0.88, which indicates that the system is quite accurate. The precision is 0.88, and the recall rate is also 0.88, which means that the ADS correctly labeled 88 percent of candidates' placement results. All of these metrics suggest that the model has relatively high accuracy in predicting the job placement of candidates. Meanwhile, among other models, the Decision Tree and Random Forest have the highest accuracy, which is shown in the below figures.

```
Decision Tree
Model performance for Training set
- Accuracy: 1.0000
- F1 score: 1.000000
- Precision: 1.000000
- Recall: 1.000000
----------------------------------
Model performance for Test set
- Accuracy: 0.8372
- Fl score: 0.8391
- Precision: 0.8420
- Recall: 0.8372
==================================
```

Figure 7: Metric for Decision Tree

```
Random Forest
Model performance for Training set
- Accuracy: 1.0000
- F1 score: 1.000000
- Precision: 1.000000
- Recall: 1.000000
----------------------------------
Model performance for Test set
- Accuracy: 0.7907
- Fl score: 0.7710
- Precision: 0.7801
- Recall: 0.7907
==================================
```

Figure 8: Metric for Random Forest

For the Decision Tree and the Random Forest model, the results of the four metrics all equal 1, which shows that the model has perfectly labeled all candidates and can make predictions with 100 percent accuracy. However, these results are suspicious as indicating a tendency of overfitting for the training set. On the other hand, considering the dataset only has a small number of entries with only 215, both Decision Tree and Random Forest models are overly complicated for the training set, capturing all the noise within it. As a result, there is a pattern of lower accuracy in the testing set for both models, which also indicates low robustness for these models, and the next section will explain more about it.

Outcomes:

Before evaluating the fairness metrics for each model implemented by the author, we made a change for 'Gender' and 'y_train' / 'y_pred' to become binary numerical encoding for later interpretation. For 'Gender', males are represented as 0 while females are represented as 1. For 'y_train' and 'y_pred', the target status of 'Placed' are represented as 1 while 'Not Placed' are represented as 0. For 6 models that were implemented by the author, we all did a fairness metrics evaluation and outputting overall performance as well as by-group fairness metrics performance for each model. The sensitive feature, in this case, will be 'Gender'. Then we generalized them into a visual presentation.

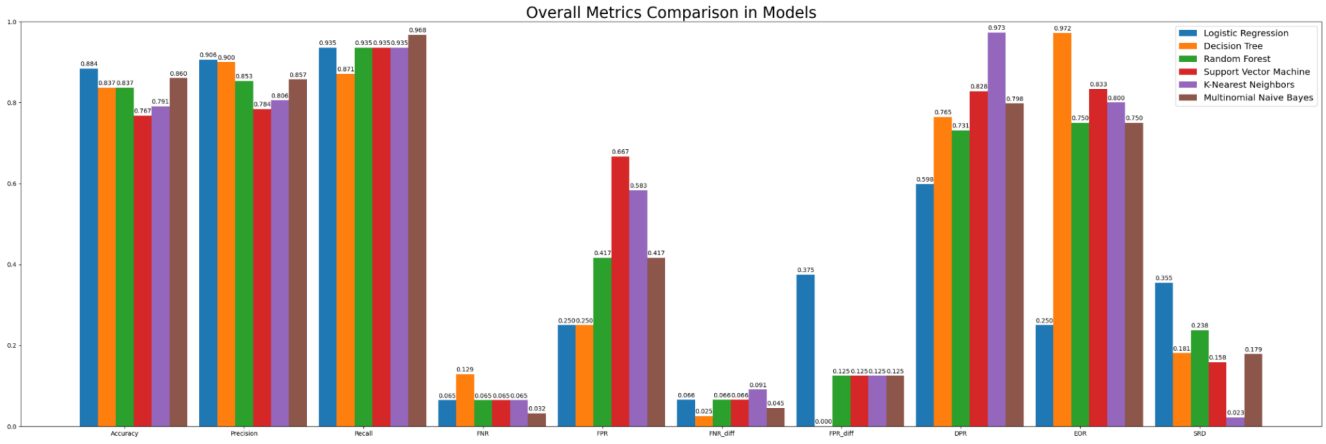


Figure 9: Overall Fairness Performance Comparison for Models

The figure above shows the overall fairness metric performance for models that have been implemented by the author. In the evaluation, 10 metrics are presented in the graph including Accuracy, Precision, Recall, False Negative Rate, False Positive Rate, False Negative Rate Difference, False Positive Rate Difference, Demographic Parity Ratio, Equalized Odds Ratio, and Selection Rate Difference. For metrics of accuracy, precision, and recall, all the models maintain a high score, which means that they are making accurate predictions. But again, since the dataset is small, machine learning models especially those that are more complicated are easily overfitting the training set, leading to high

accuracy. Another pattern that we observed from the graph is that FPR in models tends to be higher than FNR, which means that the models are more likely to make false-positive errors. This could also be explainable since our dataset consists of more positive labels than negative labels, and the models will learn to predict the majority classification more often.

```
false negative rate difference    0.065657
false positive rate difference    0.375000
demographic parity ratio          0.598465
equalized odds ratio              0.250000
selection rate difference         0.355204

false negative rate difference    0.025253
false positive rate difference    0.000000
demographic parity ratio          0.764706
equalized odds ratio              0.971591
selection rate difference         0.180995

false negative rate difference    0.065657
false positive rate difference    0.250000
demographic parity ratio          0.764706
equalized odds ratio              0.666667
selection rate difference         0.217195
```

Figure 9: Fairness Metrices for Logistic Regression, Decision Tree, and Random Forest

For the fairness metrics, it's noticeable that the Logistic Regression model among other models has the lowest Demographic Partiy Ratio and Equalized Odd Ratio. Low DPR and EOR for the Logistic Regression Model indicate that there is a difference between males and females in terms of predictions, respectively indicating that low positive predictions and low false positive predictions across groups in this model. The demographic parity ratio of 0.6 indicates that the proportion of positive cases in male and female groups is very different, leading to gender bias. Overall, the logistic regression model fails to enhance the fairness. On the other hand, both the Decision Tree model and the Random Forest model have a higher demographic parity ratio and equalized odds ratio than the logistics regression model, which indicates that the prediction for males and females is not evenly distributed.

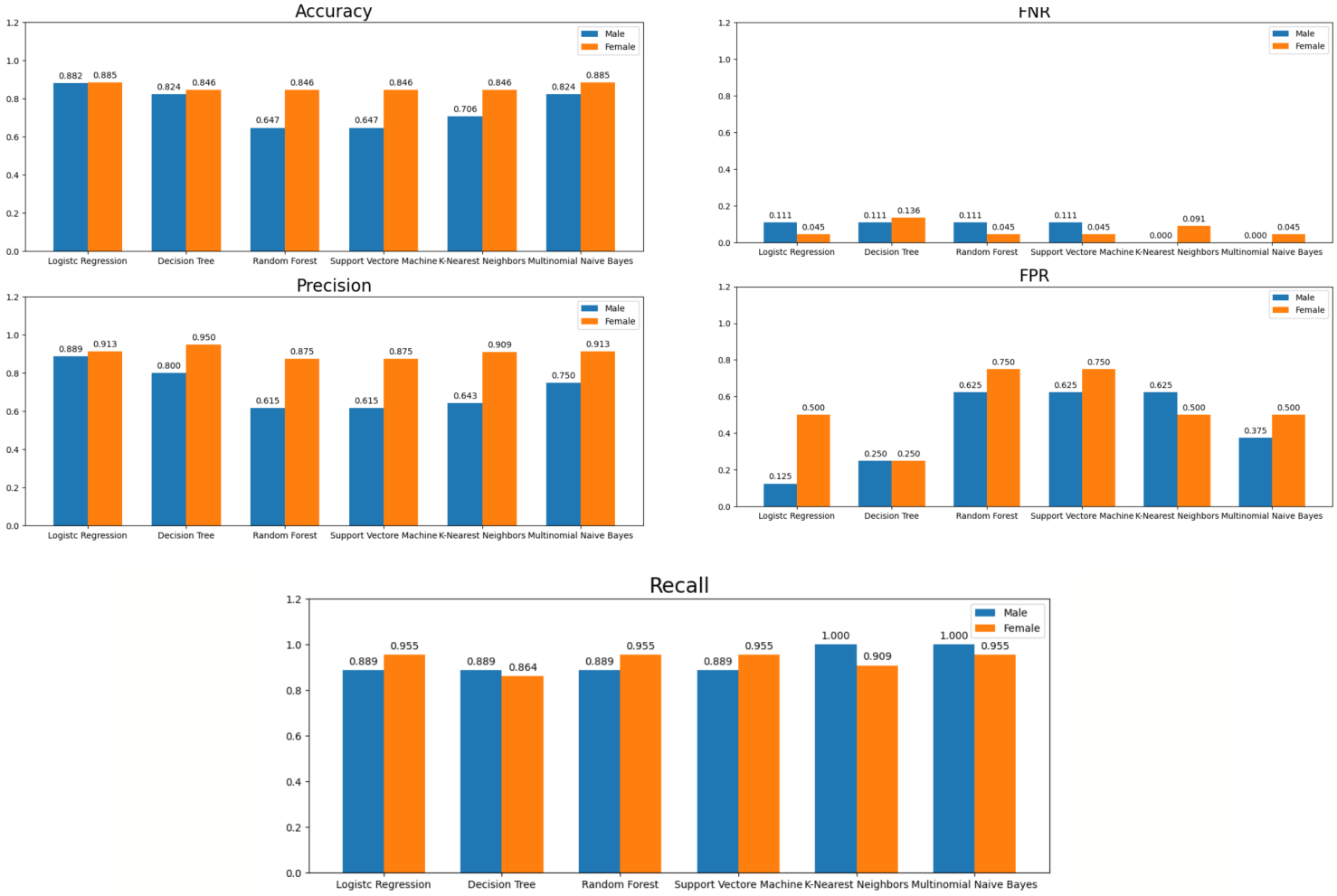


Figure 10: Accuracy, Precision, Recall, FNR, and FPR in Models between Females and Males

Therefore, to visualize the difference between males and females within the models, our group constructed graphs to present Accuracy, Precision, Recall, FNR, and FPR across groups. Intuitively, females tend to have higher accuracy and precision than males across models, as shown in the above figure. The difference between females and males in terms of precision is greater than that in terms of accuracy, which means that females are more accurate in predicting true positive labels. Meanwhile, the FNR is lower than FPR across all models, but males have higher FNR compared to females, while females have higher FPR compared to males. This indicates that females are more likely to be predicted as positive labels compared to males, and a small proportion of males are more likely to be expected as negative labels. Considering the original dataset that more females are labeled as ‘Not Placed’ than males with more females having no working experience, the models developed by the author seem to favor females in this case.

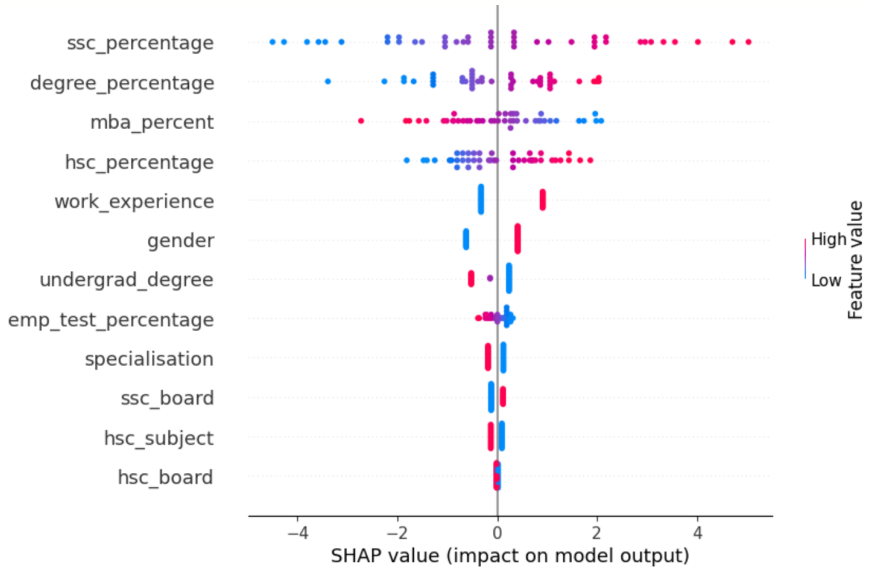


Figure 11: SHAP Value Summary Plot

Furthermore, SHAP was employed to test the robustness of ADS. First, we initialize the SHAP explainer and applying on the test data. The figure above plot shows each feature's impact on the prediction result. We observe that for categorical variables such as "work_experience", the graph shows that all red points have relatively high positive SHAP values, while all the blue points have the same negative SHAP value, which means that work experience is a crucial factor and a significant influence on the prediction. The same SHAP values also show consistency in feature impact. Also, the same phenomenon happened to gender as well. The red bar for "gender" is a bit longer than the blue bar so the red bar represents male candidates. Since all red points have the same positive SHAP value, it means that male candidates tend to be labeled as "Placed". For numerical variables, we notice that a higher value of features like "ssc_percentage", "degree_percentage" and "hsc_percentage" leads to a positive prediction, while a higher value for "mba_percentage" and "emp_test_percentage" pushes the prediction to a negative result.

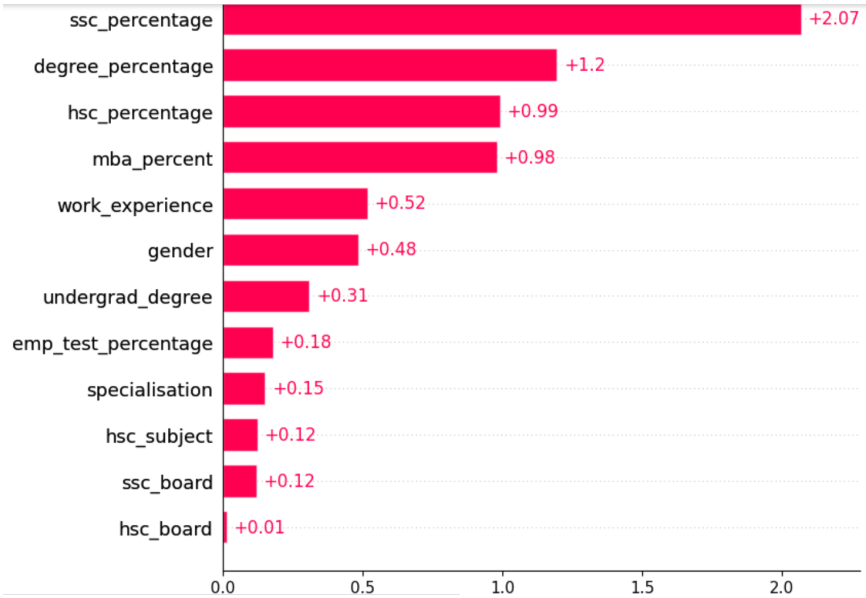
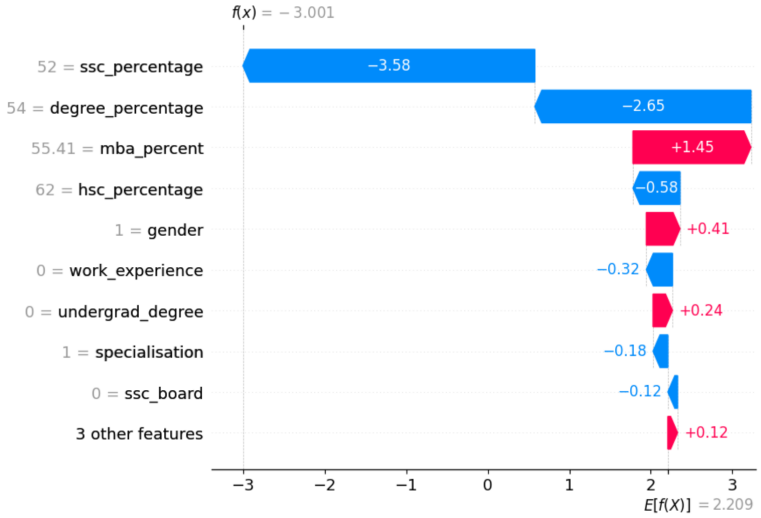
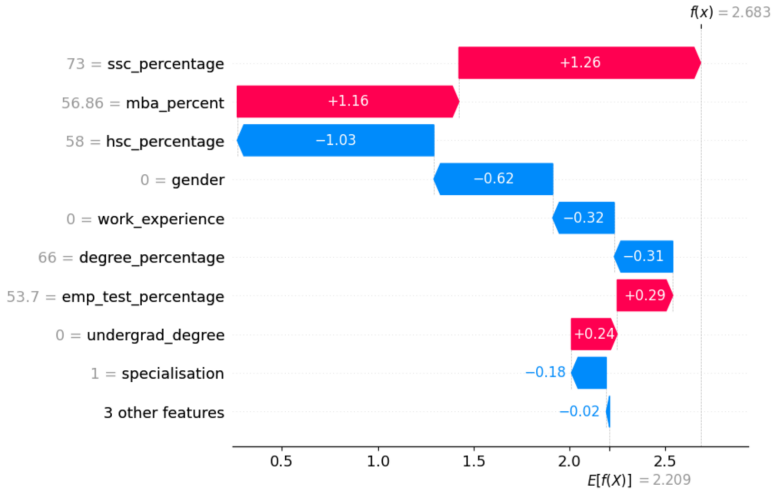


Figure 12: SHAP Value Bar Chart

Next, we plot a bar chart to examine the overall influential magnitude of each feature on the prediction result. Based on the above graph, we notice that the most influential variable is "ssc_percentage", which on average contributes 2.07 to each prediction. Other education performance features such as "degree_percentage" and "hsc_percentage" also have a high contributing magnitude to the prediction. On the other hand, features like "hsc_board" only affect the prediction with 0.01 magnitude on average, which can be ignored. For the features that we are concerned with, "gender" contribute 0.48 to the prediction, and the "work_experience" contribute 0.52, which are relatively low and did not match our expectations.





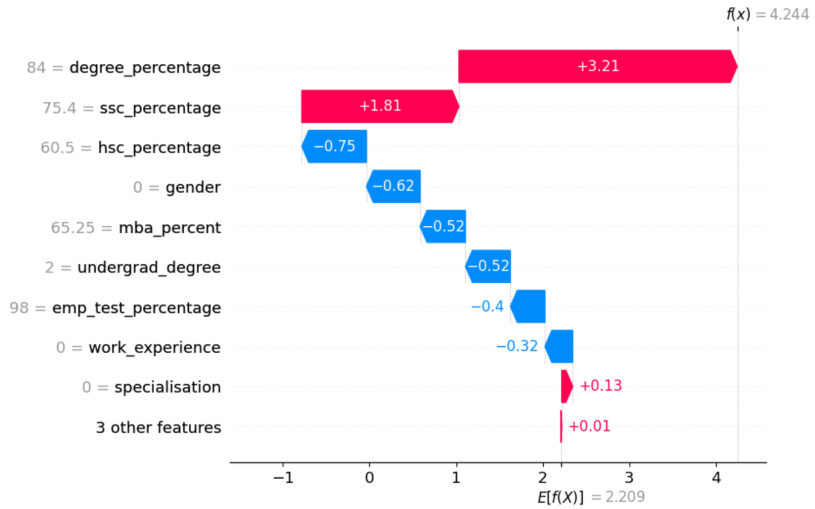


Figure 12: SHAP Value Waterfall Plot for Three Instances

Finally, we employ three waterfall plots to test three different instances respectively to provide a detailed breakdown of the contribution of variables to the prediction. We observe that after implementing the model, females are more likely to be labeled as "Placed" with a feature contribution of 0.41, while male candidates lead to a negative prediction result with a contribution of -0.62. With that being said, the model effectively impairs gender bias and takes care of female candidates. We also notice that the model seems to put great importance on education performance features including "ssc_percentage" and "degree_percentage". Based on the plots, a "degree_percentage" of 80 leads to a positive prediction result with a magnitude of 3.21, which is the highest among other features in the third instance. For the first instance, a "degree_percentage" of 54 contributes to a negative result with a magnitude of -2.65. This means that the model favored the candidate with a good undergraduate academic performance. On the other hand, so far, "ssc_percentage" is the most influential feature, which is somehow unreasonable and may overestimate those candidates with a high senior secondary test grade. For example, in the second instance, the candidate with a 73 "ssc_percentage" is pushed to a positive prediction result with the contribution of 1.26, but this candidate actually has no working experience and has a relatively bad academic performance in college, so it is unreasonable to expect that this candidate will be placed to the job. As a result, the model's robustness might be harmed due to the large magnitude of the contribution of some irrelevant features.

Summary:

The dataset we are dealing with in this project has an intrinsic shortcoming, which consists of a small number of entries. Furthermore, males are more than females with more males having the target status of 'Placed' than females do. Naturally, this will lead to concerns regarding the implementation of

models, which might reinforce biases against females by labeling them as ‘Not Placed’. However, with the evaluation of fairness and performance of the ADS model with SHAP, we found out that the models help to take care of the disadvantaged group, females, in terms of predictions.

Our research is mainly concerned with the Logistic Regression Model developed by the author, and we believe that the Logistic Regression Model is appropriate for the dataset we are dealing with. The target variable that we would like to predict is a binary label, which fits Logistic Regression for this scenario. We found out that the accuracy of the Logistic Regression Model is high, as well as precision and recall. Also, it helps to distribute fairness to the disadvantaged group of females by being more likely to predict them as positive labels, but at the same time, it drags down the equalized odd ratio with only 0.25. Meanwhile, the SHAP values for some features are pretty consistent in distribution as well as impact, for example, ‘specialization’ and ‘work_experience’ maintain almost a single SHAP value for prediction. However, using SHAP to explain the Logistic Regression Model, we also found that the model values the importance of some relevant features like ‘ssc_percentage’, which it’s a qualification for high school students.

Overall, the ADS is accurate, but not quite fair or robust. Considering this ADS is served for the condition of job placement, employers might find these measures like accuracy appropriate to help them select potential candidates. On the other hand, candidates might value measures like fairness especially demographic parity ratio, selection rate difference, and equalized odd ratio to ensure the algorithm will be fair in selecting candidates. Particularly, female candidates who might concern about their selection due to zero work experience would care about how the algorithm weighs different features and make predictions. Furthermore, we think this might not be a good ADS to implement in the public sector or industry. The original input data is small, and it might result in low generalizability. Also, the features that this ADS values are irrelevant in choosing candidates for a job, which will lead to incorrect labeling when there is more data in the public sector or industry.

This dataset needs to improve its data collection by collecting more data and providing more information for females with work experience and being placed into a job. This measure is to allow more data from an underrepresented group for better generalization in predictions. The features inside the dataset might need to include more relevant features and disregard those that are less relevant, for example, this dataset consists of a lot of qualifications including those from high school period and college period. Also, we notice that the author doesn’t do any data processing before implementing the models, and within the dataset, there are a lot of categorical variables. We would recommend performing one-hot encoding to convert categorical variables to numerical variables that can be used in machine learning models, and it might be more accurate compared to before.

Work Cited

Kaggle Dataset:

https://www.kaggle.com/datasets/ahsan81/job-placement-dataset

Shngare, Prasad:

https://www.kaggle.com/code/prasadshingare/job-placement-dataset#Preparing-Dataset-for-ML-model