Bias Check in Hiring

Team members:
Donghuai Li, Yue Wang, Yichen Zhang
Kehan Wang, Guanting Gai

Advisor: Jason Kuruzovich

2020.04

Contents

1 Overview

1.1 Introduction

Nowadays, machine learning technology is used by more and more companies to help make correct business decisions. There is no doubt such new technology significantly works pretty well when it is used to predict future sales, the performance of employees, and so on. But some problems also come with such an advantage, one of which is algorithms bias. When algorithms and models are applied to real-world problems, they are not always behaving fairly. It has become a crucial problem for those companies using predictive algorithms, and OutMatch faces this kind of problem as well. And our goal is to try to find out whether or not there is a bias in OutMatch prediction and try to solve it.

1.2 Company Background

OutMatch is a human resource company whose business consists of job-fit assessments, behavioral interviewing, and online reference. Founded in Dallas 2015, OutMatch now has developed into an experienced human resource assessment company. Their Predictive Talent Platform provides its clients with a suite of technology to assist their recruitment, talent acquisition, and leadership development.

To be specific, OutMatch helps clients put a strategy around recruiting decisions, such as who will be hired, how to develop applicants, and how they care for company culture. OutMatch can fully assist its clients to make the best possible decisions about candidates from hiring and development to leadership and culture. By doing this, clients expect to see an ordinary workforce transformed into a high-growth, high-performance company.

Up to now, OutMatch delivers nearly 20 million scientifically-proven employment assessments each year at over 200,000 client locations worldwide. The wonderful services empower companies to reach their full potential and strong alignment with companies' unique Culture DNA™. The clients of OutMatch include HCA Healthcare, American Airlines, 7-Eleven, etc.

1.3 Problem Statement

Recruiting companies provide applicants' information to OutMatch. Then OutMatch gives everyone an assessment score based on the personality online questionnaire. After that, OutMatch sets a threshold of about 20% of applicants that are excluded from consideration in the hiring process. The rest 80% of applicants move forward to the next step of the hiring companies.

Now recruiting companies want to know the process of screening and whether it would create bias. Thus, OutMatch needs to ensure the bottom 20% of applicants are not being discriminated against.

The team OutMatch will assess machine learning predictive models and develop approaches that can assess as well as address algorithmic bias. The specific tasks are as follows:

a. Develop a definition of algorithmic fairness in the context of OutMatch's modeling.

b. Develop a measure of algorithmic fairness for machine learning models.

c. Develop a method to address potential features causing algorithmic bias.

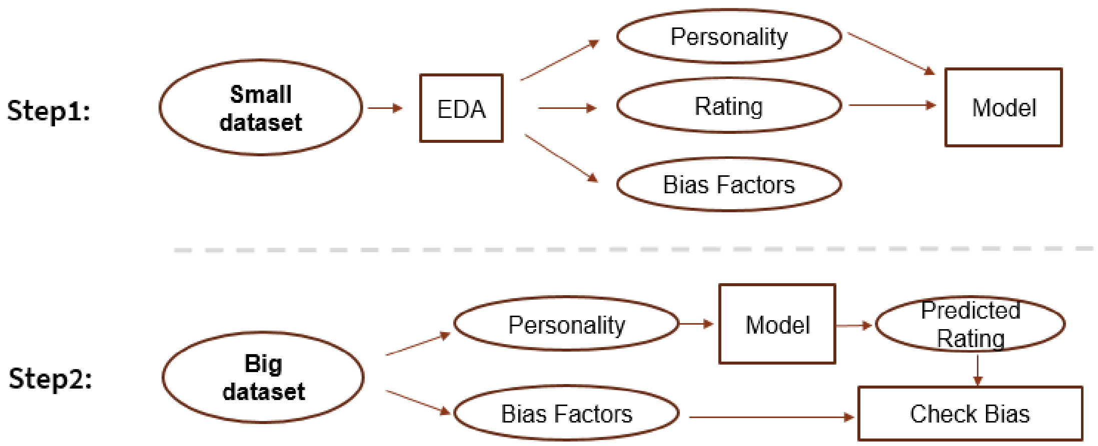
1.4 Evaluate metrics

Equalized Odds: Equalize the outcomes across the protected and non-protected groups, in other words, to make sure every group has the same percentage.

Four-Fifth rule: The selection rate for any non-protected group is not less than four-fifths (80%) of that for the protected group.

1.5 Process

We have two datasets now, and we plan to use the small dataset to train a ml model and fit it into the big dataset to predict candidates' performance. After filtering out the bottom 20% candidates, we check out the remaining 80% and try to find out whether or not there is a significant bias between different group:



2 Exploratory Data Analysis

2.1 Brief description of the dataset

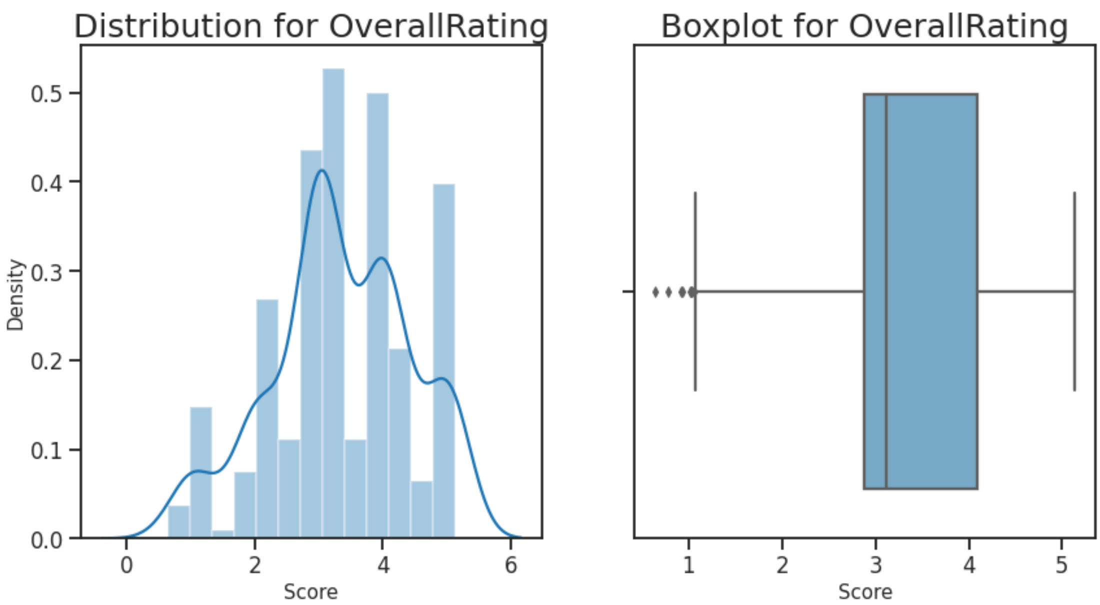
The dataset we are working on describes the three parts: 1) Bias factors: candidate's Gender, Race and AgeBand; 2) Rating: the score given by supervisors to certain candidates based on their actual performance, which is also our dependent variables; 3) scores of personality derived from an online questionnaire. The dataset contains 484 records and 26 features.

We need to check whether there is any missing value in the dataset before data visualizations. As shown below, there is 141 missing information about the ID and personality score. Without the useful information, the record of data is useless. So, we decided to remove them. Now we have 313 rows and 26 variables.

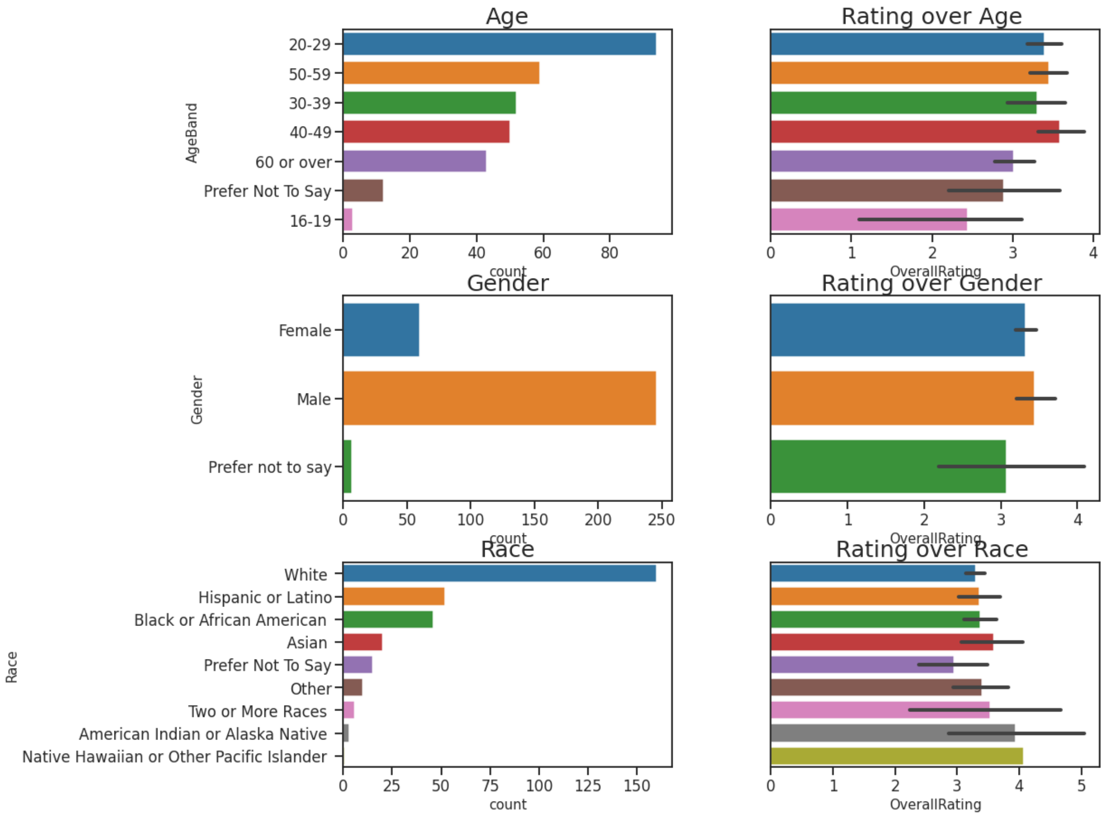
Features	Missing Values
CandidateID	141
AgeBand	0
Race	0
Gender	0
QualityofHirePoint	0
OverallRating	66
Accommodation_tile	141
Assertiveness_tile	141
CautiousThinking_tile	141
Competitiveness_tile	141
CriticismTolerance_tile	141
DetailInterest_tile	141
FollowThrough_tile	141
InterpersonalInsight_tile	141
Multitasking_tile	141
ObjectiveThinking_tile	141
Optimism_tile	141
PositiveViewofPeople_tile	141
PreferenceforStructure_tile	141
ProcessFocused_tile	141
WorkIntensity_tile	141
RealisticThinking_tile	141
ReflectiveThinking_tile	141
Sociability_tile	141
SocialRestraint_tile	141
WorkIndependence_tile	141

2.2 OverallRating

OverallRating is the dependent variable in the dataset. The median of OverallRating is 3.34, with a standard deviation of 1.08, and most of the applicants have scores between 3 - 4. It indicates that there might be many candidates who have the same score. Apart from that, the minimum rating is 0.64 and the maximum rating is 5.13. There are those outliers below 1, which means these employees leave a bad image to their managers.



2.3 OverallRating & Age, Gender, Race



Most of the candidates are young people whose ages between 20 - 29, which takes up 30% of all applicants, followed by people from 30 – 59, which is easy to understand since they are the main labor in society.

After that, we analyzed the relationship between OverallRating and AgeBand. It seems the performance of each age group is very close. The people aged 40 - 59 perform a little better, people between 20-29 perform well maybe because they just step into society and they are motivated and diligent to make some achievements. However, people between 30 – 39 have a huge range and the reason maybe they get a little tired and bored and may focus more on life events, family or pursuing pleasure. People either older or younger have barely satisfactory performance.

Then we connected gender with OverallRating. We can see that there is no significant difference between males and females, which is in line with the reality since in practical terms, both men and women have a large number of them to be the salesmen. The only thing worth mentioning is that females have lower scores and bottom outliers than men. Additionally, those people who prefer not to say their gender has lower performance than other genders.

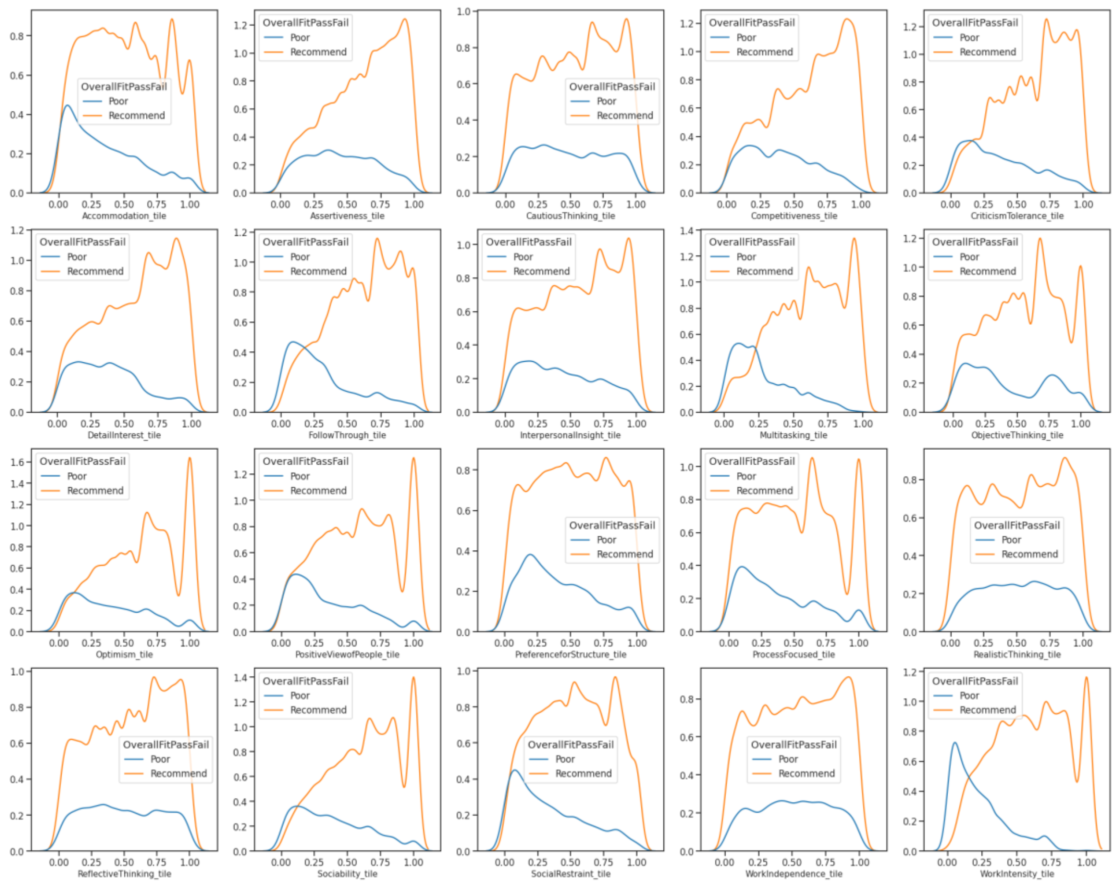
In our dataset, about half of them are white people (51.1%), with 16.6% Hispanic or Latino and 14.7% Black or African American. Since the database has huge differences among different races, we should examine the ratio of each statistics and consider the base number effect. Also, some people have two or more races, which makes our analysis more complicated.

We also would like to see the race influence on overall rating. As the above bar chart shows, it manifests that for the position of salesmen, American Indian Alaska Native has the best performance, followed by White people and Asian. The rest of them are basically at the same level while “prefer not to say” has the lowest performance. On the whole, there is no huge difference between various races. Moreover, the sample size for some races is too small (there is only 1 Native Hawaiian or Other Pacific Islander) to draw conclusions based on it.

2.6 Kernal density estimation plot

We made a Kernal density estimation plot of all assessment attributes. Our main customer is Carmax and they are trying to hire wonderful salesmen, so as an outstanding salesperson, he or she has to be very outgoing or say, sociability, to persuade clients to buy their cars. Also, he or she should be very careful and cautious about the sales data, properties and contracts to make a good deal. The superior salesman should have strong abilities of multitasking and competitiveness as well, to work for more clients at the same time. Most personalities are highly positively correlated with the overall rating, while only 4 personalities are negatively correlated with OverallRating.

We display 20 percentile distributions we can see that in some attributes, recommend
 candidates are more likely to have a higher rank while poor candidates are not. For i
nstance, the Assertiveness distribution, the second graph in the first row, implies reco
mmend candidates have a peak near 100 percentile, while the last graph of row 3, the
realistic thinking distribution, recommend candidates have no much difference in distr
ibution comparing to poor candidates.



3 Model

3.1 Predicting the target variable

3.1.1 Detecting discrimination in the OverallRating

The first step is determining whether there are already some discriminations in the existing OverallRating given by the supervisor based on applicants' real performances, because the OverallRating is the target value in the model training process. If there are some discriminations when the supervisors give the overall score,

the model trained by this target will be untrusted. We used a simple linear regression to detect the impact of categories on OverallRating. OverallRating is the dependent variable. and Age, Gender and Race are independent variables.

Since the dataset only has over 300 records after cleaning, some of the attributes are relatively small, only including less than 10 observations. Thus we integrated some of the choices. For instance, in race, we categorized “Two or More Races (not Hispanic or Latino", "American Indian or Alaska Native (not Hispanic or Latino", "Native Hawaiian or Other Pacific Islander (not Hispanic or Latino) " all as "others".

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          OverallRating   R-squared:                       0.048
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     1.161
Date:                Wed, 18 Mar 2020   Prob (F-statistic):              0.307
Time:                        13:46:32   Log-Likelihood:                -460.00
No. Observations:                 313   AIC:                             948.0
Df Residuals:                     299   BIC:                             1000.
Df Model:                          13
Covariance Type:            nonrobust
==================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                      2.7151      0.659      4.121      0.000       1.419       4.012
Gender[T.Male]                -0.0756      0.159     -0.475      0.635      -0.389       0.238
Gender[T.Prefer not to say]    0.4348      0.676      0.644      0.520      -0.895       1.764
race[0]                       -0.3522      0.293     -1.200      0.231      -0.930       0.225
race[1]                       -0.3231      0.288     -1.124      0.262      -0.889       0.243
race[2]                       -0.2209      0.351     -0.630      0.529      -0.911       0.470
race[3]                       -0.5698      0.497     -1.147      0.252      -1.547       0.408
race[4]                       -0.3675      0.260     -1.412      0.159      -0.880       0.145
Ageband[0]                     1.0622      0.638      1.666      0.097      -0.193       2.317
Ageband[1]                     0.9670      0.644      1.501      0.134      -0.301       2.234
Ageband[2]                     1.2651      0.647      1.956      0.051      -0.008       2.538
Ageband[3]                     1.1425      0.647      1.765      0.079      -0.131       2.416
Ageband[4]                     0.7031      0.653      1.076      0.283      -0.582       1.989
Ageband[5]                     0.4917      0.837      0.587      0.558      -1.156       2.140
==============================================================================
Omnibus:                        6.306   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.043   Jarque-Bera (JB):                6.062
Skew:                          -0.296   Prob(JB):                       0.0483
Kurtosis:                       2.660   Cond. No.                         41.1
==============================================================================
```

The result of regression for existing OverallRating with Age, Gender, Ethics in train data set.

We could see from the output that all independent variables are not significantly correlated with OverallRating. Hence, we can conclude that there's **<u>no bias</u>** in the current score system and the research can go on.

3.1.2 Model training by using training data set

Next, we fit a baseline model. The dependent variable is OverallRating and independent variables are 20 characteristics scores derived from questionnaire that Outmatch designed, which includes: accommodation, assertiveness, cautious thinking and etc. This linear regression model is the way we predict OverallRating in the larger test data set.

	Coefficient		
Accommodation_tile	-0.253933	Optimism_tile	-0.501215
Assertiveness_tile	-0.379527	PositiveViewofPeople_tile	0.136935
CautiousThinking_tile	-0.346315	PreferenceforStructure_tile	0.111477
Competitiveness_tile	0.770784	ProcessFocused_tile	0.257311
CriticismTolerance_tile	-0.370401	WorkIntensity_tile	0.853233
DetailInterest_tile	-0.851393	RealisticThinking_tile	0.454297
FollowThrough_tile	-0.663007	ReflectiveThinking_tile	0.258378
InterpersonalInsight_tile	-0.280138	Sociability_tile	0.425462
Multitasking_tile	0.229177	SocialRestraint_tile	1.361354
ObjectiveThinking_tile	0.057681	WorkIndependence_tile	-0.282271

Coefficient of baseline model

From the coefficient we could see that SocialRestraint has the highest positive number of 1.36, followed by WorkIntensity and Sociability, which means these are the main features that needed to be a good car salesperson. The R-squared is 0.2, not a wonderful model, perhaps due to the small number of the data set.

3.1.3 Detecting bias in the baseline model

After using the baseline model to predict OverallRating in the train data set, we tried to identify whether there's bias existing in the predicted OverallRating. We used the same method as fitting a linear regression model with predicted OverallRating and Age, Gender, Ethics in test data to see whether there's bias in prediction rating in our own model result.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Prediction   R-squared:                       0.040
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                    0.9493
Date:                Wed, 18 Mar 2020   Prob (F-statistic):              0.502
Time:                        14:12:33   Log-Likelihood:                -194.31
No. Observations:                 313   AIC:                             416.6
Df Residuals:                     299   BIC:                             469.1
Df Model:                          13
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                     3.1249      0.282     11.085      0.000       2.570       3.680
Gender[T.Male]               -0.1101      0.068     -1.617      0.107      -0.244       0.024
Gender[T.Prefer not to say]  -0.2431      0.289     -0.841      0.401      -0.812       0.326
race[0]                      -0.0081      0.126     -0.064      0.949      -0.255       0.239
race[1]                      -0.1626      0.123     -1.321      0.187      -0.405       0.080
race[2]                      -0.1290      0.150     -0.859      0.391      -0.424       0.166
race[3]                      -0.1445      0.213     -0.680      0.497      -0.563       0.274
race[4]                      -0.1644      0.111     -1.476      0.141      -0.384       0.055
Ageband[0]                    0.3814      0.273      1.398      0.163      -0.156       0.918
Ageband[1]                    0.4306      0.276      1.563      0.119      -0.112       0.973
Ageband[2]                    0.3707      0.277      1.339      0.182      -0.174       0.915
Ageband[3]                    0.3334      0.277      1.204      0.230      -0.212       0.878
Ageband[4]                    0.3570      0.280      1.277      0.202      -0.193       0.907
Ageband[5]                    0.4249      0.358      1.186      0.237      -0.280       1.130
==============================================================================
Omnibus:                        6.647   Durbin-Watson:                   1.836
Prob(Omnibus):                  0.036   Jarque-Bera (JB):                7.175
Skew:                          -0.247   Prob(JB):                       0.0277
Kurtosis:                       3.553   Cond. No.                         41.1
==============================================================================
```

The result of regression for Predictive OverallRating with Age, Gender, Ethics in train data set.

3.1.4 Generate recommendation criteria

In the third step, we created recommendation criteria based on Outmarch's method.

Outmatch would let the top 80% of people pass the test and we decided to do the same. We sorted the prediction OverallRating from high to low, and selected the top 80% as recommended, the bottom 20% as not recommended. Thus, we created this dummy variable - recommendation, which is either 1 or 0.

e.g. prediction 5.54 >- recommended >- assign to 1

prediction 0.82 >- not recommended >- assign to 0

pred	recommend
0.824642	0
0.824642	0
0.851670	0
0.880179	0
0.886711	0

3.1.5 Detecting bias in the recommendation result

Our next step is to fit a linear regression model with predicted recommendation, Age, Gender, Ethics in test dataset to test whether there's bias in "recommend" in our own model result.

OLS Regression Results

Dep. Variable:	recommend	R-squared:	0.034
Model:	OLS	Adj. R-squared:	-0.008
Method:	Least Squares	F-statistic:	0.8104
Date:	Wed, 18 Mar 2020	Prob (F-statistic):	0.649
Time:	20:48:48	Log-Likelihood:	-154.49
No. Observations:	313	AIC:	337.0
Df Residuals:	299	BIC:	389.4
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6480	0.248	2.611	0.009	0.160	1.136
Race[T.Black or African American (not Hispanic or Latino)]	0.0216	0.111	0.195	0.845	-0.196	0.239
Race[T.Hispanic or Latino]	0.0305	0.108	0.281	0.779	-0.183	0.244
Race[T.Other]	-0.0979	0.132	-0.740	0.460	-0.358	0.162
Race[T.Prefer Not To Say]	0.0794	0.187	0.424	0.672	-0.289	0.448
Race[T.White (not Hispanic or Latino)]	0.0019	0.098	0.019	0.985	-0.191	0.195
AgeBand[T.20-29]	0.1547	0.240	0.644	0.520	-0.318	0.628
AgeBand[T.30-39]	0.2429	0.243	1.001	0.318	-0.235	0.720
AgeBand[T.40-49]	0.1671	0.244	0.685	0.494	-0.313	0.647
AgeBand[T.50-59]	0.0991	0.244	0.406	0.685	-0.381	0.579
AgeBand[T.60 or over]	0.0649	0.246	0.264	0.792	-0.419	0.549
AgeBand[T.Prefer Not To Say]	0.1016	0.316	0.322	0.748	-0.519	0.723
Gender[T.Male]	0.0020	0.060	0.033	0.974	-0.116	0.120
Gender[T.Prefer not to say]	-0.2576	0.255	-1.012	0.312	-0.759	0.243

Omnibus:	63.567	Durbin-Watson:	0.087
Prob(Omnibus):	0.000	Jarque-Bera (JB):	100.990
Skew:	-1.390	Prob(JB):	1.18e-22
Kurtosis:	3.139	Cond. No.	41.1

The result of regression for Recommendation with Age, Gender, Ethics in test data set

Still, there's **<u>no bias</u>**.

3.2 Bias in Single Group

3.2.1 Gender

After we get the predicted result, we try to see the different pass rates for different groups: Gender, Age and Race. First, we look at Gender, the table below is the confusion matrix for male and female.

	RECOMMEND	NOT RECOMMEND	**TOTAL**
MALE	15225	3585	18810
FEMALE	7431	2084	9515
TOTAL	22656	5669	28325

In order to see clearly, we convert the count number into the pass rate, which is shown below:

	RECOMMEND	NOT RECOMMEND
MALE	0.81	0.19
FEMALE	0.78	0.22

This is the pass rate for male and female, we can see from the chart that the pass rate for the male is 81% and female is 78%.

Now we would like to know whether there's conscious bias in our model.

A. Four-Fifth rules

The selection rate of any disadvantage group is no less than 80% percent of the advantage group.

The selection rate of male is 0.81. The selection rate of females is 0.78 and is higher than 80% of the selection rate of male.

Since the selection rate of male group is not more than 80% of the female group, we believe that there is **no bias; perhaps** the difference may be random.

B. Z-test /Chi Square test

H0: There is no difference between the percentage of female and male

H1: There is difference between the percentage of female and male

Because chisq= 31.909, p value= 0.000, we reject H0.

We have the reason to believe that there's **indeed some bias** in the screening process. However, since this is such a big data set, it is easy to reject the null hypothesis.

3.2.2 Age

Also, the same process for age, almost all age group get a pass rate of about 80%, the distribution is shown below:

	RECOMMEND	NOT RECOMMEND
16-19	0.803	0.197
20-29	0.791	0.209
30-39	0.792	0.208
40-49	0.802	0.198
50-59	0.824	0.176
60 or over	0.833	0.167
Prefer Not To Say	0.822	0.178

We did the same two tests, and the results are same:

No bias using Four-Fifth rule;

Bias exists using Chi Square test.

3.2.3 Race

Again, the pass rate for different races also doesn't show a significant difference. Although the difference is a bit bigger than that in different age groups, we still get a fairly stable pass rate of about 80%.

	RECOMMEND	NOT RECOMMEND
white	0.827	0.173
black	0.773	0.227
hispanic or latino	0.779	0.221
asian	0.779	0.221
other	0.806	0.194
prefer not to say	0.790	0.210

We did the same two tests, and the results are same:

No bias using Four-Fifth rule;

Bias exists using Chi Square test.

3.3 Bias in Sub-Group

Now we have a distribution of pass rates for each group. However, since three factors cause discrimination, the absence of discrimination against a single factor does not mean that there is no discrimination against all sub-groups. For example, while from the point of gender and age, there is no discrimination between the young and the old or male and female, but for the young men and old men these two sub-groups, there are likely some differences in the pass rate. What we need to do next is to examine whether the pass rate of different sub-group will vary.

We already have the final prediction, to make it easy, we export the result to excel and produce a pivot table to see pass rate for all sub-groups.

3.3.1 Bias of two bias factors

If we only consider the combination of two bias factors. The result is as below:

Row Labels	Passrate
Asian (not Hispanic or Latino)	**77.96%**
Female	81.00%
Male	77.26%
Black or African American (not Hispanic or Latino)	**77.30%**
Female	75.29%
Male	79.01%
Hispanic or Latino	**77.91%**
Female	77.36%
Male	78.18%
Prefer Not To Say	**78.83%**
Female	72.73%
Male	82.65%
Two or More Races (not Hispanic or Latino)	**80.37%**
Female	79.45%
Male	81.01%
White (not Hispanic or Latino)	**82.71%**
Female	81.40%
Male	83.17%
(blank)	
(blank)	
Grand Total	**79.96%**

Gender and Race

For different gender and race groups, there is some difference but not quite significant. The highest pass rate is 83.17% and the lowest is72.73%.

Row Labels	Passrate
Female	**78.07%**
16-19	76.44%
20-29	76.85%
30-39	77.09%
40-49	80.52%
50-59	82.48%
60 or over	85.79%
Prefer Not To Say	84.17%
Male	**80.92%**
16-19	82.44%
20-29	80.31%
30-39	80.19%
40-49	80.09%
50-59	82.53%
60 or over	82.83%
Prefer Not To Say	83.41%
(blank)	
(blank)	
Grand Total	**79.96%**

Gender and Age

For gender and age, the highest pass rate is 85.79% and the lowest is 76.44%.

16-19	**80.33%**
Asian (not Hispanic or Latino)	83.87%
Black or African American (not Hispanic or Latino)	77.47%
Hispanic or Latino	79.17%
Prefer Not To Say	81.82%
Two or More Races (not Hispanic or Latino)	79.47%
White (not Hispanic or Latino)	82.86%
20-29	**79.02%**
Asian (not Hispanic or Latino)	74.43%
Black or African American (not Hispanic or Latino)	76.11%
Hispanic or Latino	78.66%
Prefer Not To Say	79.27%
Two or More Races (not Hispanic or Latino)	80.05%
White (not Hispanic or Latino)	82.40%
30-39	**79.13%**
Asian (not Hispanic or Latino)	82.77%
Black or African American (not Hispanic or Latino)	77.04%
Hispanic or Latino	76.03%
Prefer Not To Say	78.00%
Two or More Races (not Hispanic or Latino)	80.00%
White (not Hispanic or Latino)	81.33%

50-59	**82.52%**
Asian (not Hispanic or Latino)	73.56%
Black or African American (not Hispanic or Latino)	83.48%
Hispanic or Latino	74.07%
Prefer Not To Say	74.29%
Two or More Races (not Hispanic or Latino)	75.32%
White (not Hispanic or Latino)	84.60%
60 or over	**83.22%**
Asian (not Hispanic or Latino)	67.35%
Black or African American (not Hispanic or Latino)	85.31%
Hispanic or Latino	76.27%
Prefer Not To Say	57.14%
Two or More Races (not Hispanic or Latino)	86.67%
White (not Hispanic or Latino)	84.46%
Prefer Not To Say	**83.70%**
Asian (not Hispanic or Latino)	85.19%
Black or African American (not Hispanic or Latino)	82.76%
Hispanic or Latino	81.58%
Prefer Not To Say	78.81%
Two or More Races (not Hispanic or Latino)	83.87%
White (not Hispanic or Latino)	85.56%

Race and Age

For Race and Age, the highest pass rate is 86.67% and the lowest is 57.14%.

3.3.2 Bias of three bias factors

Row Labels	Passrate
16-19	**80.33%**
Asian (not Hispanic or Latino)	**83.87%**
Female	80.95%
Male	84.47%
Black or African American (not Hispanic or Latino)	**77.47%**
Female	72.40%
Male	82.08%
Hispanic or Latino	**79.17%**
Female	76.62%
Male	80.66%
Prefer Not To Say	**81.82%**
Female	50.00%
Male	88.89%
Two or More Races (not Hispanic or Latino)	**79.47%**
Female	76.67%
Male	81.32%
White (not Hispanic or Latino)	**82.86%**
Female	81.36%
Male	83.42%

Sub-group age 16-19

20-29	**79.02%**
Asian (not Hispanic or Latino)	**74.43%**
Female	77.00%
Male	73.77%
Black or African American (not Hispanic or Latino)	**76.11%**
Female	73.98%
Male	78.16%
Hispanic or Latino	**78.66%**
Female	78.03%
Male	78.99%
Prefer Not To Say	**79.27%**
Female	69.88%
Male	86.36%
Two or More Races (not Hispanic or Latino)	**80.05%**
Female	78.90%
Male	80.87%
White (not Hispanic or Latino)	**82.40%**
Female	80.76%
Male	83.06%

Sub-group age 20-29

30-39	**79.13%**
Asian (not Hispanic or Latino)	**82.77%**
Female	86.67%
Male	81.98%
Black or African American (not Hispanic or Latino)	**77.04%**
Female	75.27%
Male	78.41%
Hispanic or Latino	**76.03%**
Female	74.91%
Male	76.62%
Prefer Not To Say	**78.00%**
Female	77.78%
Male	78.18%
Two or More Races (not Hispanic or Latino)	**80.00%**
Female	78.81%
Male	80.86%
White (not Hispanic or Latino)	**81.33%**
Female	79.06%
Male	82.19%

Sub-group age 30-39

40-49	**80.23%**
Asian (not Hispanic or Latino)	**80.69%**
Female	86.67%
Male	79.13%
Black or African American (not Hispanic or Latino)	**78.45%**
Female	76.79%
Male	79.53%
Hispanic or Latino	**77.80%**
Female	80.14%
Male	76.86%
Prefer Not To Say	**85.00%**
Female	80.00%
Male	86.67%
Two or More Races (not Hispanic or Latino)	**84.35%**
Female	85.71%
Male	83.52%
White (not Hispanic or Latino)	**81.22%**
Female	81.84%
Male	80.96%

Sub-group age 40-49

50-59	**82.52%**
Asian (not Hispanic or Latino)	**73.56%**
Female	83.33%
Male	72.00%
Black or African American (not Hispanic or Latino)	**83.48%**
Female	85.55%
Male	82.21%
Hispanic or Latino	**74.07%**
Female	80.56%
Male	72.00%
Prefer Not To Say	**74.29%**
Female	57.14%
Male	78.57%
Two or More Races (not Hispanic or Latino)	**75.32%**
Female	75.86%
Male	75.00%
White (not Hispanic or Latino)	**84.60%**
Female	82.39%
Male	85.30%

Sub-group age 50-59

60 or over	**83.22%**
Asian (not Hispanic or Latino)	**67.35%**
Female	66.67%
Male	67.39%
Black or African American (not Hispanic or Latino)	**85.31%**
Female	91.18%
Male	83.49%
Hispanic or Latino	**76.27%**
Female	62.50%
Male	78.43%
Prefer Not To Say	**57.14%**
Female	66.67%
Male	50.00%
Two or More Races (not Hispanic or Latino)	**86.67%**
Female	75.00%
Male	90.91%
White (not Hispanic or Latino)	**84.46%**
Female	88.89%
Male	83.89%

Sub-group age over 60

Although ground pass rates for three bias factors are not significantly different, there are, however, some quite significant differences in pass rates of different sub-groups. For example, the highest pass rate exists in females with two or more races and without clarifying her age, which is 93.33%, and the lowest pass rate is female whose age is from 16 to 19 and prefer not to say his race, which is only 50%. It's a considerable difference. However, the results may be caused by a small data set(only 313 observations in total).

4 Result

Based on our analysis and outcomes of models, we conclude as follow:

For a single group such as Gender, Age and Race: There is a small difference between different groups, so we don't think there is a bias against different ages, gender or races.

For the two-factor sub-group: Our recommendation is fairly even on Gender-Age and Gender-Race. But when it comes to Race-Age, we get the highest 86.67% and the lowest 57.14%, which means there is bias to some degree. The discrimination is against the elders who don't want to clarify their race.

For the three-factor sub-group: The highest pass rate exists in the female with two or more races and without clarifying her age, which is 93.33%, and the lowest pass rate is female whose age is from 16 to 19 and prefer not to say his race, which is only 50%. Also this is quite a huge bias.

5 Packages- detect and address the bias

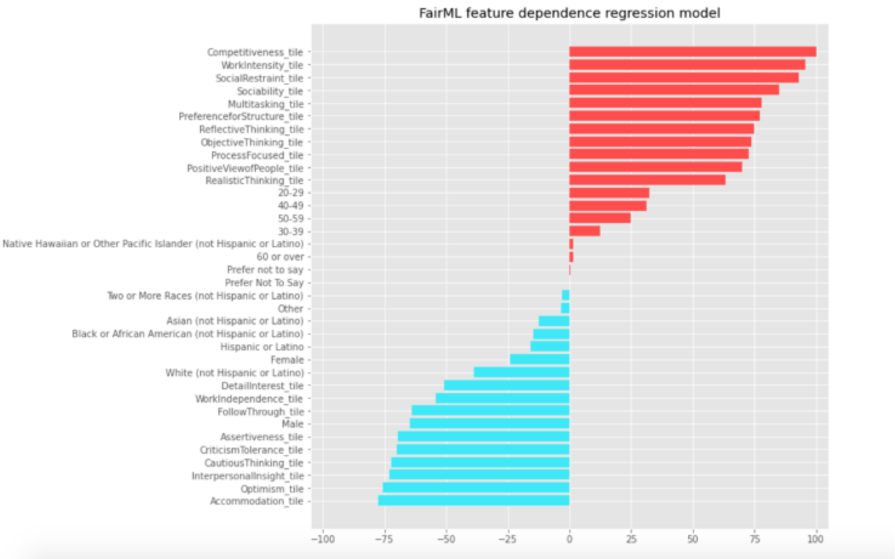
Now we have already found out the existence in our model manually, then we would like to see whether there are already python packages to do it automatically. There two packages we find are FairML, Fairlearn, and IBM360.

5.1 FairML

The basic idea behind FairML (and many other attempts to audit or interpret model behavior) is to measure a model's dependence on its inputs by changing them. If a small change to an input feature dramatically changes the output, the model is sensitive to the feature.

Orthogonal projection of vectors is important for FairML because it allows us to completely remove the linear dependence between attributes. If two vectors are orthogonal to one another, then no linear transformation of one vector can produce the other. This intuition underlies the feature dependence measure in FairML.

After implementing our train model, we can get a bar plot to detect bias.



Feature importance for Age, Gender, Ethics

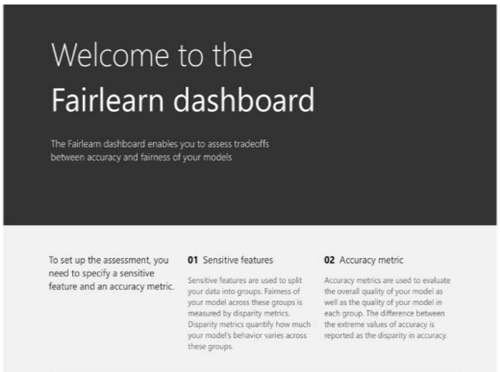
From the plot we could see that aside from those characteristic tiles derived from the questionnaire, the model indeed depends on age, gender and ethics but in a relatively small proportion. The biggest positive one is age 20-29 with 30 and the most negative one is race white with -35. Those numbers are small compared to those characteristics', thus, we can conclude that **<u>the chances of bias existing are not very high</u>** in the model.

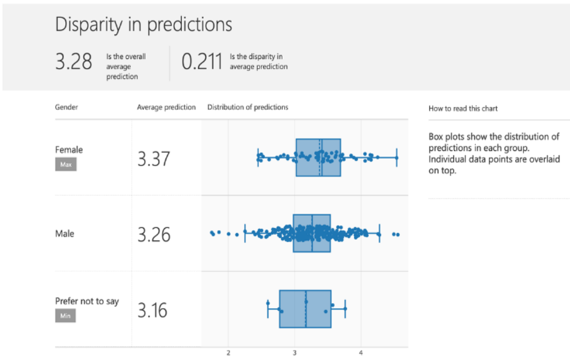
5.2 Fairlearn

In order to better visualize the bias result, another package called fairlearn can create a dashboard in Jupyter Notebook to ease the analysis process.

Fairlearn dashboard is a Jupyter notebook widget to assess how a model's predictions impact different groups (e.g., different races), and also for comparing multiple models along different fairness and accuracy metrics.

There is an illustration of the dashboard. We can choose the sensitive features and the accuracy metrics by our own preference, and thus see the disparity(difference) in predictions and accuracy.





The logic behind the package is the same as the model part we did before.

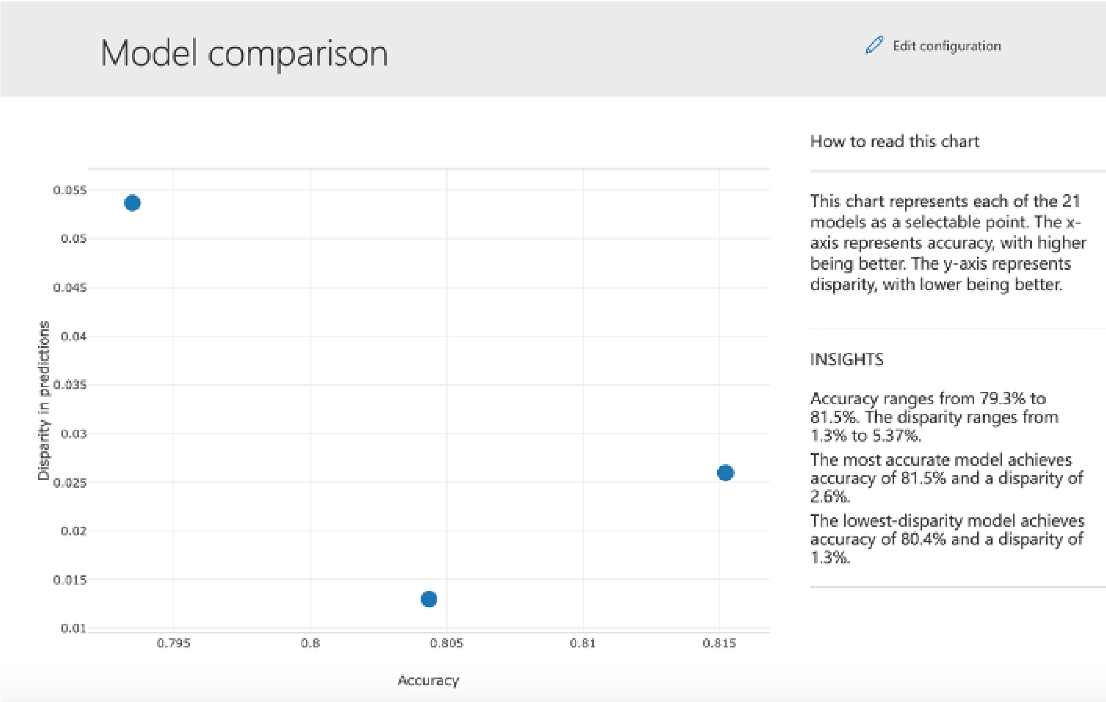
More impressively, the package contains algorithms for mitigating unfairness. In our part, we tried the GridSearch algorithm of Fairlearn to reduce bias. The basic idea of the algorithm is to generate several models that achieve various trade-offs between accuracy (measured by AUC) and disparity.

This algorithm comes from the paper "A Reduction Approach to Fair Classification" (Agarwal et al. 2018). (https://arxiv.org/abs/1803.02453)

The key idea is to reduce fair classification to a sequence of cost-sensitive classification problems, whose solutions yield a randomized classifier with the lowest (empirical) error subject to the desired constraints.

GridSearch generates models corresponding to various Lagrange multiplier vectors of the underlying constraint optimization problem. This tries a series of different models, parameterized by a Lagrange multiplier. For each value lambda, the algorithm reweights and relabels the input data, and trains a fresh model (lambda=0 corresponds to the unaltered case).

For example, there is an example of trade-off between accuracy and disparity in prediction. We prefer the model with the lowest disparity and highest accuracy. And it can provide a choice for companies to decide which model to choose. This is a really powerful tool to use.



5.3 AI Fairness 360

5.3.1 Detecting the bias by comparing pass rate with the same score level

This thought of detecting bias is from a demo in the AI fairness 360. In that case, the recidivism risk categories predicted by the COMPAS tool is compared to the actual recidivism rates of defendants in the two years after they were scored. COMPAS scores for each defendant ranged from 1 to 10, with ten being the highest risk. Scores 1 to 4 were labeled by COMPAS as “Low”; 5 to 7 were labeled “Medium”; and 8 to 10 were labeled “High.” And by comparing to the actual result, they found that High-risk white defendants are 3.61 times as likely to recidivate as low-risk white defendants, while high-risk black defendants are only 2.99 times as likely to recidivate as low-risk black defendants. Thus, the bias could be confirmed.

Then we tried to find bias in our model using this method, the steps are as follow:

- Using the actual OverallRating that provided supervisors to get a result of recommend or not recommend based on the ratio of 80%, which means top 80% are recommended.

- Fit the model using train data set through linear regression and get predicted OverallRating.

- Separate score as low, high medium based on the top, middle and bottom 33.33%.

- Calculate proportion and check differences between recommendation (recommend/ not recommend) and score level (low, middle, high).

White_high_recommend	93.18%	White_mid_recommend	77.19%
Black_high_recommend	95.83%	Black_mid_recommend	80.00%

Pass rates of race groups with high and mid score level

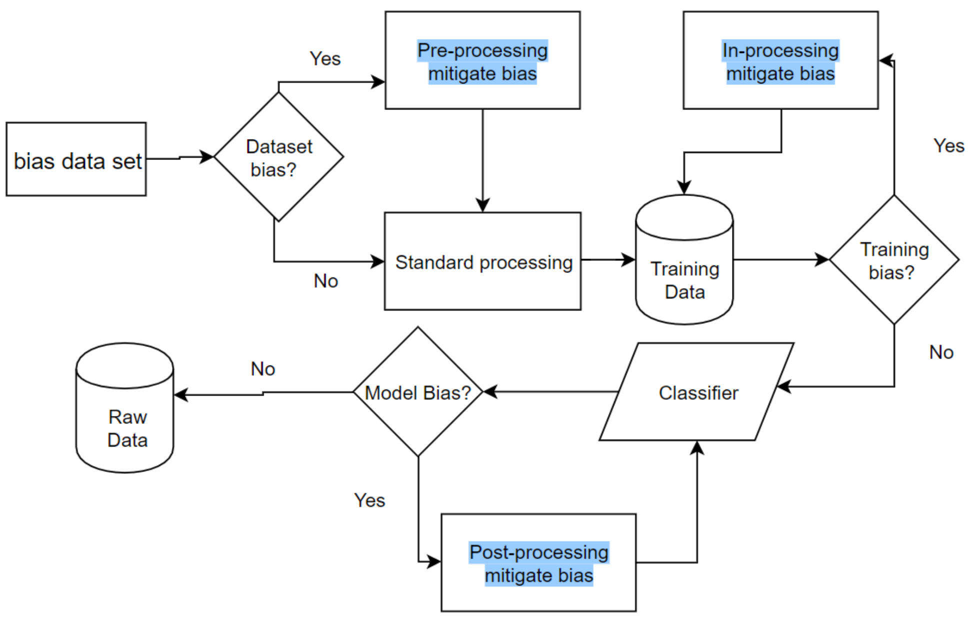
Male_high_recommend	95.89%	Male_mid_recommend	79.54%
Female_high_recommend	92.86%	Female_mid_recommend	75.00%

Pass rates of gender groups with high and mid score level

From the table we could see that there was no obvious difference between different age or races. Hence, we believe that there's **no bias** in age and gender in the model.

5.3.2 Using AI Fairness 360 Open Source Toolkit for Outmatch dataset

5.3.2.1 AI Fairness 360 detect bias pipeline & fair worldviews



The above graph shows that there are 3 steps bias may exist: Pre-processing, In-Processing, Post-processing.

In the process of learning AIF360, we find two different worldviews on group fairness:

we're all equal (WAE) and what you see is what you get (WYSIWYG) The WAE worldview holds that all groups have similar abilities with respect to the task (even if we cannot observe this properly), whereas the WYSIWYG worldview holds that the observations reflect ability with respect to the task.

For example in predicting an applicant's future performance, using OverallRating Score as a feature for predicting success in the company, the WYSIWYG worldview says that the score correlates well with future success and that there is a way to use the score to correctly compare the abilities of applicants. In contrast, the WAE worldview says that the OverallRating score may contain structural biases so its distribution being different across groups should not be mistaken for a difference in the distribution in ability.

In our data sets, we want to check whether there exist structural biases in our both dataset, so we pick the WYSIWYG world view in the following steps. Under WYSIWYG we use demographic parity metrics that should be used: disparate_impact and statistical_parity_difference. The definition of metrics are followed:

Metrics Name	Disparate Impact	Statistical Parity Difference
Definition	Computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.	Computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.
Fairness Range	(0.8,1.2)	(-0.1,0.1)

5.3.2.2 Apply AI Fairness 360 to our data set

We defined three privileged groups: 'White' in race, 'Male' in gender, 'Young' (age band between 16-39) in age. Using the transformed binary label data set we calculate the fairness metrics for the training data set and test data set, respectively.

Metrics Name		Disparate Impact	Statistical Parity Difference
Fairness Range		(0.8,1.2)	(-0.1,0.1)
Small data set	Gender	0.964	-0.029

(311 rows) Demographic Variables	Age	1.067	0.051
	Race	1.013	0.01

Fairness metric result for training data set

Metrics Name		Disparate Impact	Statistical Parity Difference
Fairness Range		(0.8,1.2)	(-0.1,0.1)
Big data set (28729 rows) Demographic Variables	Gender	0.99	-0.007
	Age	1.064	0.049
	Race	0.973	-0.021

Fairness metric result for test data set

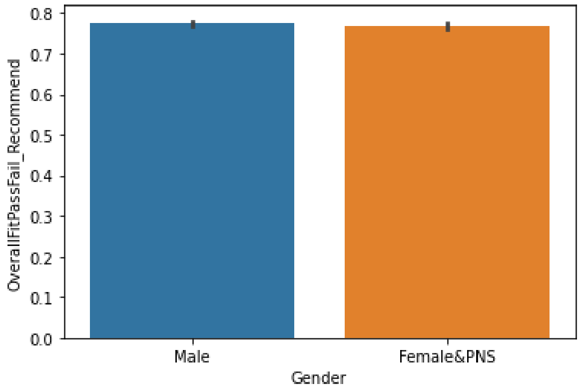
As the results suggest, both data sets exist tolerable bias under the fairness range. We asked a question in further discussion: What if our data set does have bias out of fairness range? Can we mitigate it? So we made a fake data and tried to figure out how to apply bias mitigation algorithms into our data.

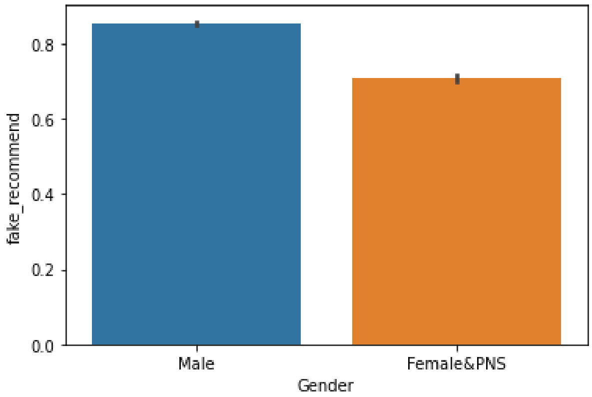
5.3.2.3 Fake data and bias mitigation algorithm to our data set

First, we choose the Gender as the target variable. We increase the recommended rate of males and decrease the female & PNS recommend rate. The new rate difference still doesn't break the Four-Fifth rule. The results are followed.

	Before FAKE recommend rate	After Fake recommend rate
Male	0.774	0.853
Female & Prefer	0.767	0.708

Not to Say		





Now we check the fairness metrics as before

Metrics Name	Disparate Impact	Statistical Parity Difference
Fairness Range	(0.8,1.2)	(-0.1,0.1)
Demographic Variables: Gender	0.83	-0.145

The Statistical Parity Difference metric is below the fairness range, the bias can not be neglected. Because this bias shows up before we use it as training data, we need to use the pre-processing bias mitigation algorithm. Now we introduce the pre-processing algorithm “Reweighing”, it weighs the examples in each group combination differently to ensure fairness before classification.

The fairness metric after we apply reweihing algorithm is followed:

we check the fairness metrics as before

Metrics Name	Disparate Impact	Statistical Parity Difference
Fairness Range	(0.8,1.2)	(-0.1,0.1)
Demographic	0	0

Variables: Gender		

Now we can say there is no bias in the data set.

5.3.3 Summary of AI Fairness 360

AI Fairness 360 Open Source Toolkit contains over 70 fairness metrics and 10 bias mitigation algorithms. There are some examples of other widely used metrics and bias mitigation algorithms in this package.

a. Example of metrics

- Average odds difference

Computed as average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups.

Fairness for this metric is between -0.1 and 0.1

- Equal opportunity difference

This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.

Fairness for this metric is between -0.1 and 0.1

- Theil index

Computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1. It measures the inequality in benefit allocation for individuals.

A value of 0 implies perfect fairness. Fairness is indicated by lower scores, higher scores are problematic.

b. Bias Mitigation Algorithms For Each Phase of the Pipeline

AIF 360 currently contains 10 bias mitigation algorithms that span three categories, including pre-processing, in-processing, and post-processing algorithms.

Pre-Processing Algorithms Mitigating bias in **Training Data**	In-Processing Algorithms Mitigating bias in **Classifiers**	Post-Processing Algorithms Mitigating bias in **Predictions**
Reweighing Modifies the weights of different training examples	**Adversarial Debiasing** Uses adversarial techniques to maximize accuracy & reduce evidence of protected attributes in predictions	**Reject Option Classification** Changes predictions from a classifier to make them fairer
Disparate Impact Remover Edits feature values to improve group fairness	**Prejudice Remover** Adds a discrimination-aware regularization term to the learning objective	**Calibrated Equalized Odds** Optimizes over calibrated classifier score outputs that lead to fair output labels
Optimized Preprocessing Modifies training data features & labels	**Meta Fair Classifier** Takes the fairness metric as part of the input & returns a classifier optimized for the metric	**Equalized Odds** Modifies the predicted label using an optimization scheme to make predictions fairer
Learning Fair Representations Learns fair representations by obfuscating information about protected attributes		

6 General ways of addressing bias

1. Choose the right learning model for the problem. Instead of using a simple linear regression model to generate scores, use some more complicated matrices, algorithms or models. Also, Increase the sophistication of the model and make a conscious decision to progress at every stage.

2. Choose the right representative train dataset. Make sure that the parameters or variables are reliable and persuasive. Polish the online questionnaire and generate various percentile questions. Additionally, expand the training dataset like social media data or something else. Go through an empirical test first.

3. Monitor performance using real data: Use new and realistic data set to retrain the model periodically. Check the fairness of the model in fixed time.

4. Remove variables that may generate problems: If there's truly bias in the model, just delete those data to prevent problems.

7 Conclusions

1. Our predictions of Outmatch data for different groups vary to some degree. According to different metrics, we get different answers about the existence of bias. In this case, there is bias when measured by Equalized Odds but not by Four-Fifth rule.

2. We find some python packages, such as AI Fairness 360, FairML, and Fairlearn, to detect and address the bias.

3. We also find some common practices for eliminating discrimination, which provides a good reference when we encounter other related fairness problems.

Reference

https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/

https://towardsdatascience.com/is-your-machine-learning-model-biased-

94f9ee176b67

https://arxiv.org/pdf/1412.3756.pdf

https://techcrunch.com/2018/11/06/3-ways-to-avoid-bias-in-machine-learning/

https://www.logikk.com/articles/prevent-machine-bias-in-ai/

https://towardsdatascience.com/evaluating-machine-learning-models-fairness-and-bias-4ec82512f7c3

http://erichorvitz.com/biases_classifier_emotion_study.pdf

https://becominghuman.ai/how-to-prevent-bias-in-machine-learning-fbd9adf1198

https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases

https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

https://www.kdnuggets.com/2018/05/machine-learning-breaking-bad-bias-fairness.html

https://blog.insightdatascience.com/tackling-discrimination-in-machine-learning-5c95fde95e95

https://insidebigdata.com/2018/08/23/report-explores-machine-learning-ai-bias/

https://towardsdatascience.com/understanding-and-reducing-bias-in-machine-learning-6565e23900ac

https://ieeexplore.ieee.org/document/8025197

https://cdn.oreillystatic.com/en/assets/1/event/295/Removing%20unfair%20bias%20in%20machine%20learning%20using%20open%20source%20_sponsored%20by%20IBM_%20Presentation.pdf