Fairly Predicting Happiness

Maxwell Lee

Viterbi School of Engineering, University of Southern California,

Los Angeles, California 90089, USA

(Dated: January 19, 2022)

Abstract

In this report, we attempt to design a model that fairly predicts the happiness of users. In simple approaches, we see that our model tends to over-predict the happiness of areas with high population of members of Group 2, while under-predicting the happiness of areas with high population of members of Group 1. We are able to undo this effect with the use of the helper functions written by the Research and Development team. This resulting model is more in line with the relationships we see in the real world, in which there is some relationship between the ethnic groups and happiness, but not to the extent seen in our initial attempts at modeling.

I. INTRODUCTION

For this task, we are given information about 2110 regions. For each of these regions, we are given the total number of people in ethnic groups 1 and 2, the percentage of the population with at least a bachelor's degree, and the mean household income as independent variables. Our target variable is given as the mean happiness detected in the region. We wish to be able to predict happiness based on this information about the regions in a way that does not unfairly bias members of either ethic groups 1 or 2.

II. DATA PREPROCESSING

The data given is clean, so we are mainly interested in standardizing the data. We use Scikit Learn's StandardScaler [1] function to turn each of our variables into normally distributed variables. This is used to more easily compare different types of models and their fairness.

III. GROUND TRUTH MODEL

We begin by constructing a ground truth model, which we will use to compare the fairness of our predictions. This is done by creating scatter plots to reveal the correlations between the variables in our model. First, we have the percentage of population with at least a bachelor's degree plotted against the happiness. Note that these have be rescaled, so the values themselves are not important, however the relationships are. The horizontal line placed at y=-.25 represents where a happiness score of 5.8 would fall on our chart, the score we use as a baseline for unhappiness. The colorbar provided illustrates the amount of people belonging to ethnic group 1. The result is shown in Figure 1.

Here we can see there is a slightly positive relationship between the bachelor percentage and the happiness, with a correlation of .457. More prominently, we can see members of ethnic group 1 are less likely to have a bachelor's degree.

Next, we can see this same plot but conditioned on ethnic group 2, shown in Figure 2.

Here we see the opposite is the case, as members of ethnic group 2 are more likely to have a bachelor's degree.

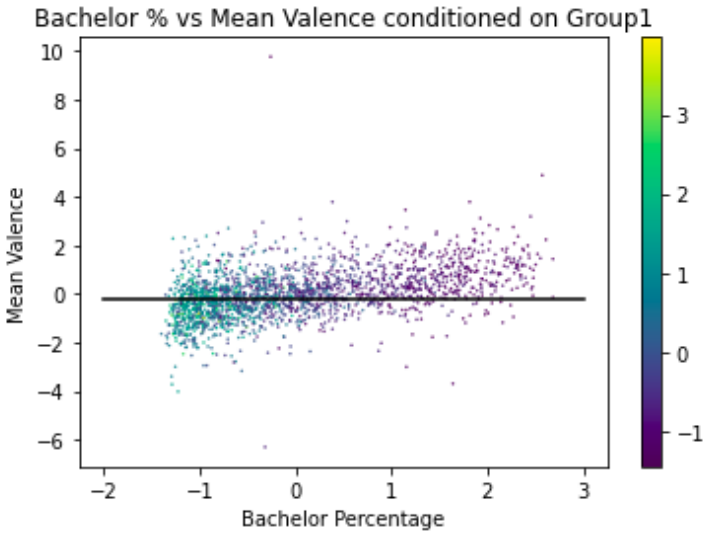

FIG. 1. Bachelor percentage vs Happiness, conditioned on group1

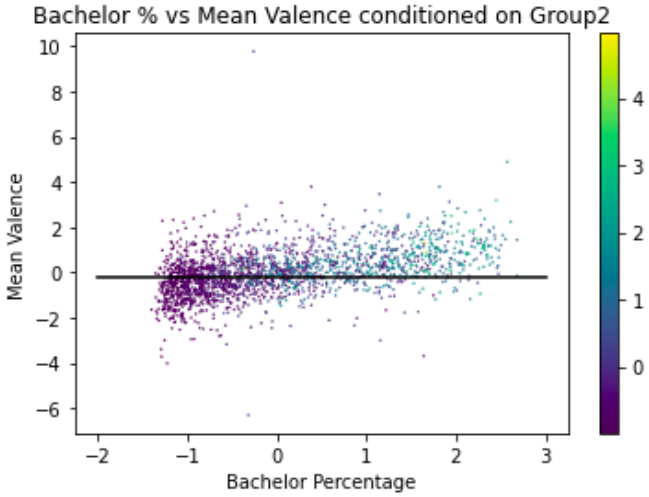

FIG. 2. Bachelor percentage vs Happiness, conditioned on group2

Next, we will show these same plots, but with mean household income as the independent variable. These are shown in Figures 3 and 4

Here we see almost identical relationships. There is a slightly positive relationship between income and happiness at .311. Members of ethnic group 1 are less likely to have a high household income, while members of ethnic group 2 are more likely to have a high household income.

Finally, we wish to see the relationship between the populations of the ethnic groups and happiness. While we obviously do not wish to unfairly bias either group with our

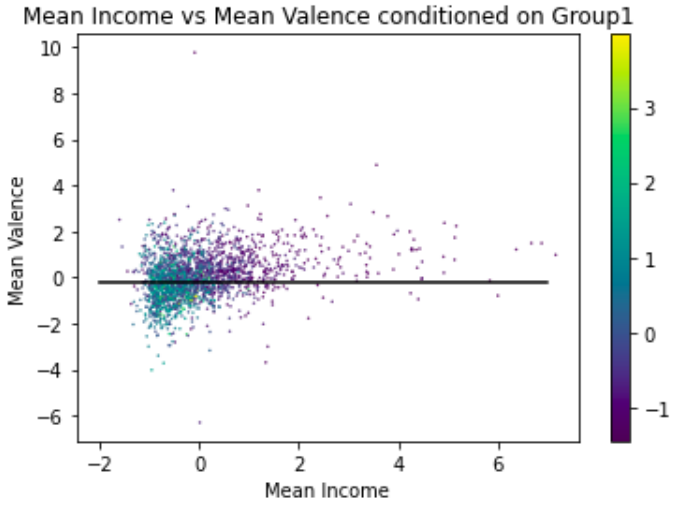

FIG. 3. Household income vs Happiness, conditioned on group1

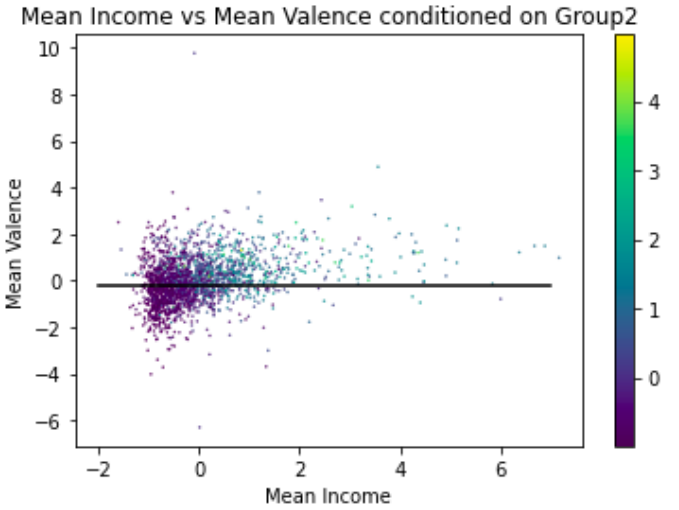

FIG. 4. Household income vs Happiness, conditioned on group2

models, it is important to understand if such a relationship exists. If it does, we should not entirely ignore that relationship, but rather look for our model to similarly replicate that relationship. Figures 5 and 6 show this relationship.

Here we can see that members of group 1 are slightly more likely to be unhappy, with a correlation of -.361, while members of group 2 are slightly more likely to be happy, with a correlation of .335. These relationships, in addition to the previously shown relationships, will be important to the overall fairness of our model.

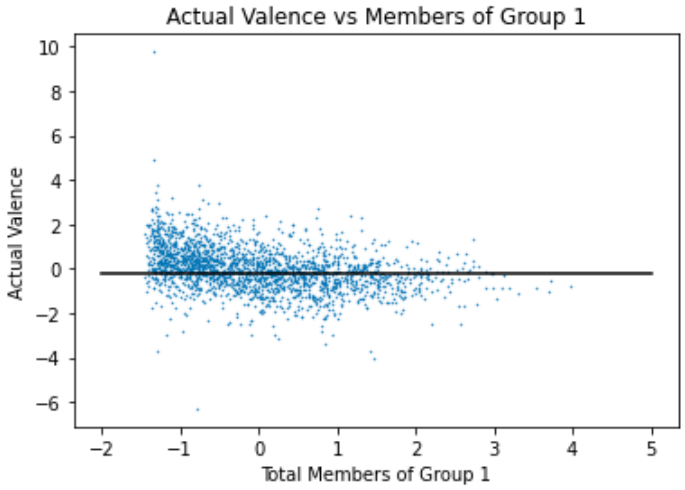

FIG. 5. Group 1 vs Happiness

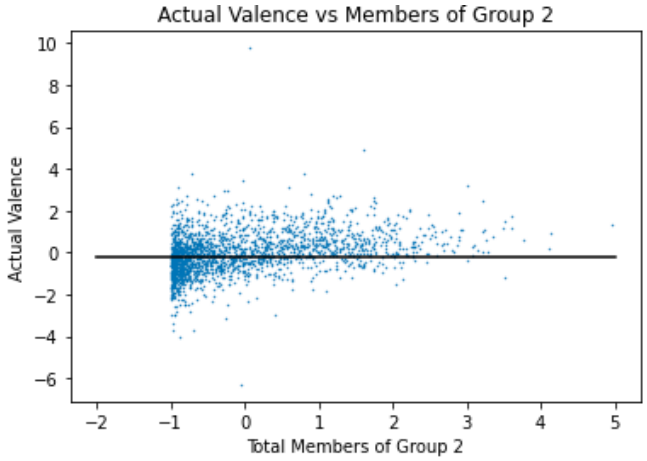

FIG. 6. Group 2 vs Happiness

IV. ETHNIC GROUP AWARE MODEL

In this section, we build a simple linear regression model with all features trying to predict mean happiness. We use 80% of the data to train the model, then make predictions for the entire dataset. We then see how the predictions align with our ground truth observations.The same plots shown above for the ground truth model are shown below in Figures 7 through 12, except with predicted happiness as the y-axis variable.

What we can gather from these figures is a clear issue. Compared to the ground truth representations, our model has biases against the ethnic groups. We have created a model that over-predicts the happiness of members of group 2 while under-predicting the happiness

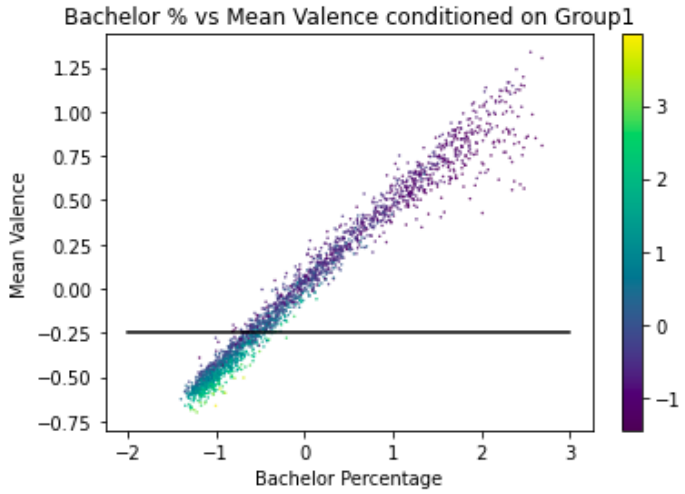

FIG. 7. Bachelor percentage vs Happiness, conditioned on group1

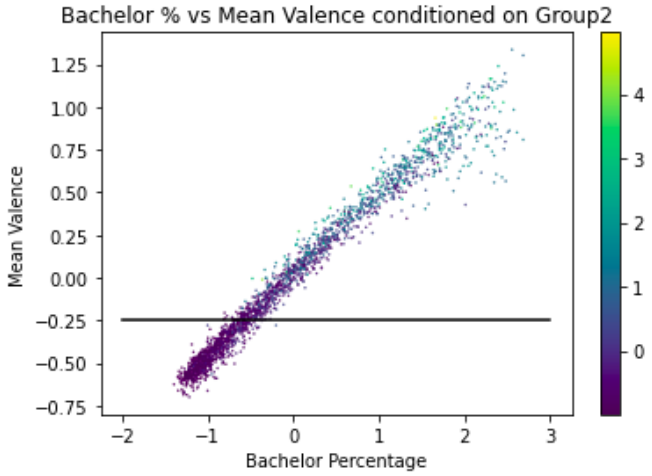

FIG. 8. Bachelor percentage vs Happiness, conditioned on group2

of group 1. We saw in the ground truth models that there is some validity to that relationship, but not nearly to the extent we see in the model. With our model, the correlation between group 1 and happiness is -.771 and the correlation between group 2 and happiness is.749, both more than double the ground truth correlations. All the correlations present are much stronger in the predictions than the ground truth.

This model is behaving in a biased manner towards both the members of group 1 and group 2. Targetting members of group 1 with our notifications due to underpredicting their happiness can certainly be viewed as discriminatory or harrassing. But if we believe in the value of these messages, by over-predicting the happiness of group 2 we are causing them to not receive these notifications. Therefore, we must look to create a model that is less biased

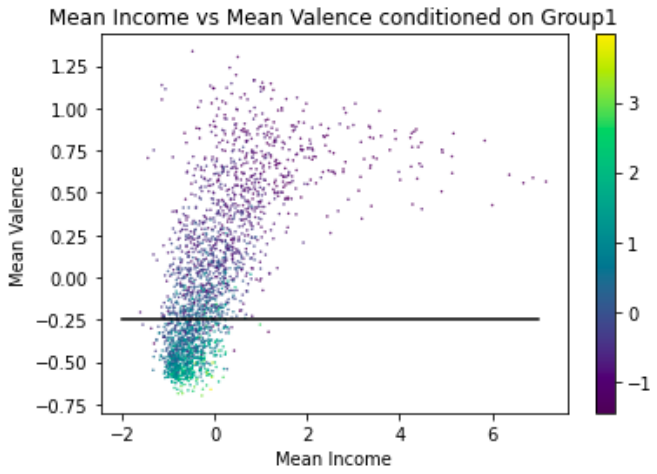

FIG. 9. Household income vs Happiness, conditioned on group1

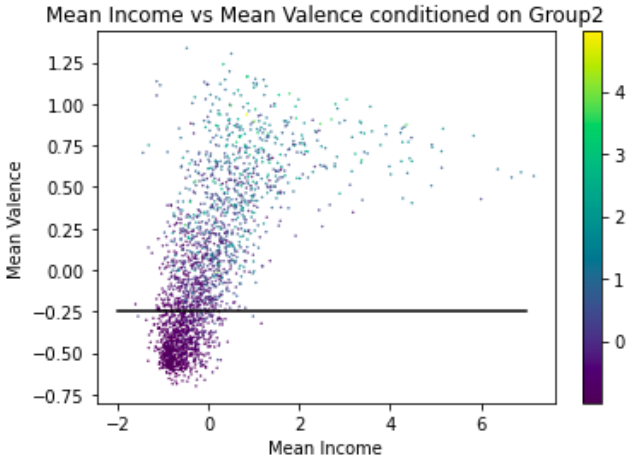

FIG. 10. Household income vs Happiness, conditioned on group2

for both of these groups.

V. ETHNIC GROUP BLIND MODEL

In an attempt to unbias the results, we create a model that does not have any information about the ethnic groups as features. Therefore, the only present features are bachelor percentage and household income. The process for creating the model is the same. We create a linear model with 80% of the data, then predict the entire dataset with that model. The resulting plots are shown in Figures 13 through 18.

These graphs look almost identical to the race aware models. How can this be? Well, as

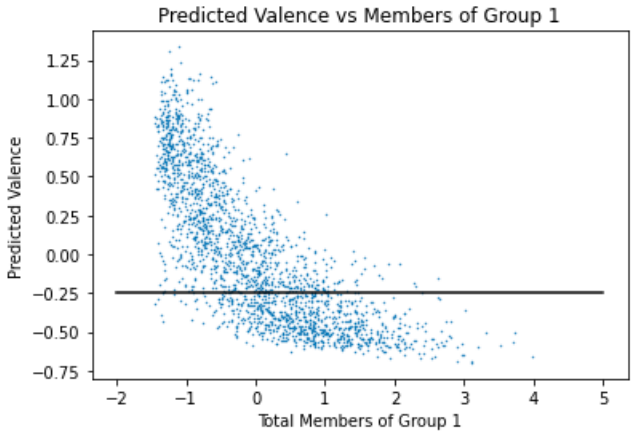

FIG. 11. Group 1 vs Happiness

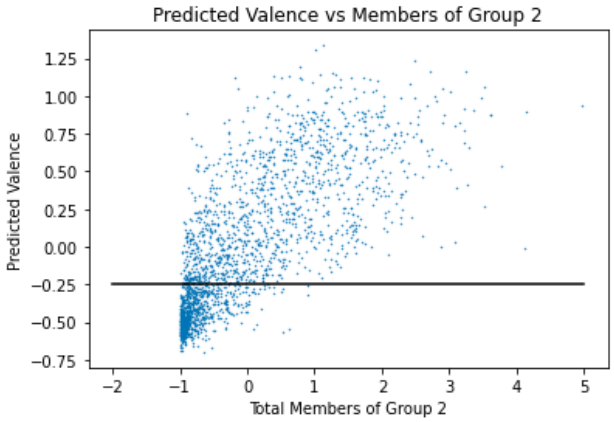

FIG. 12. Group 2 vs Happiness

mentioned in Section III, the racial groups are extremely correlated with having a bachelor's degree and household income. Removing the explicit information about race, but including these features allows us to create a race blind model that is functionally aware of race, exhibiting the exact same biases as the race aware model. Simply removing the race features is not enough to create a fair model. We must work further to create this.

VI. FAIR MODEL

With this model, we look to create a model that preserves the relationships seen in the ground truth model. To do this, we make use of the functions provided by the Research and Development team. We input the protected variables, in this case the two ethnic groups, as well as our target variable and the extent to which we wish to make the target variable

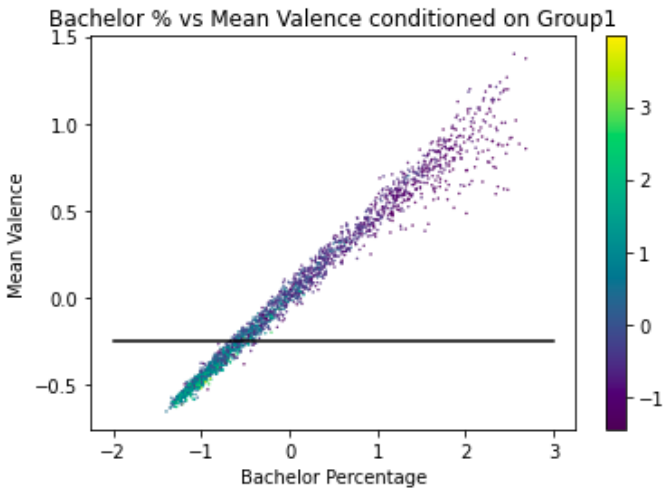

FIG. 13. Bachelor percentage vs Happiness, conditioned on group1

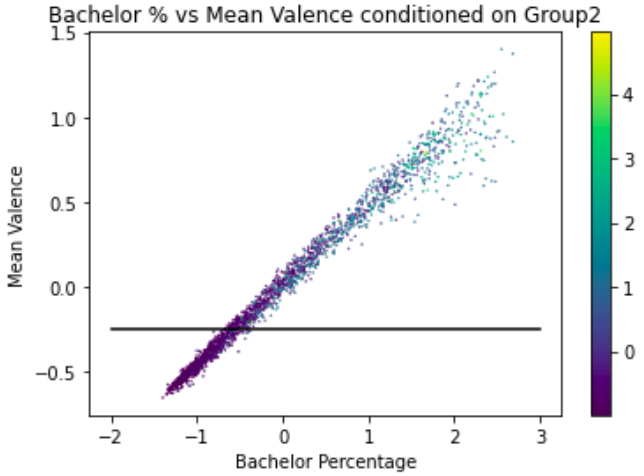

FIG. 14. Bachelor percentage vs Happiness, conditioned on group2

independent from the protected classes. A value of 0 makes an almost perfectly fair model, but this is not necessarily what we want. We know there is some relationship between the ethnic groups and their happiness. The issue with the previous models is that they overstated this importance. After a few attempts, it is found that giving a value of 0.1 accurately preserved the relationships seen in the ground truth models.

Interestingly, even after standardizing the new target variable, the model predicts noticeably fewer regions with low happiness scores. To preserve a similar proportion of regions with low happiness scores, the threshold is raised to -.1 from -.25. This line will be seen on the plots shown below in Figures 19 through 24.

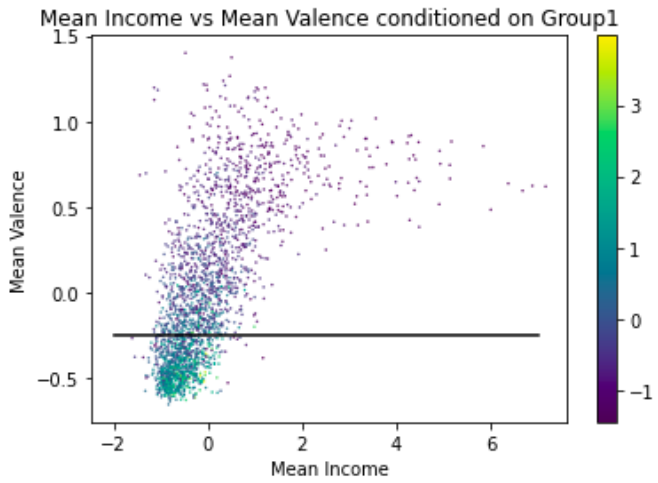

FIG. 15. Household income vs Happiness, conditioned on group1

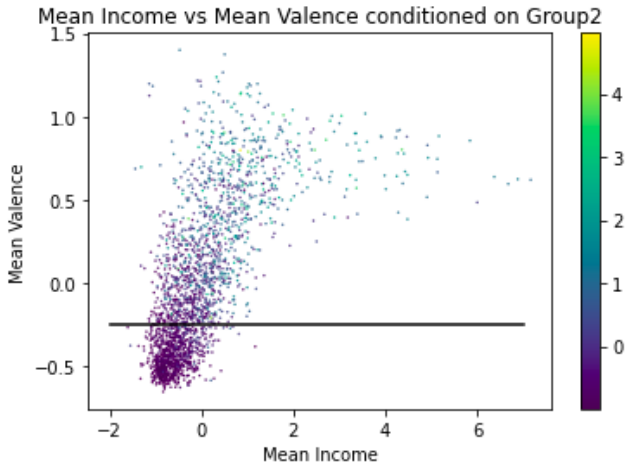

FIG. 16. Household income vs Happiness, conditioned on group2

Note that these relationships aren't perfect. Our model still likely undersells the relationship between racial groups and happiness while overselling the relationship between bachelor percentage and happiness. These relationships can be changed by changing the parameter we pass into the helper functions. But, there is a tradeoff to be had. As we increase the parameter, the relationship of racial groups gets closer to the true correlation, but the relationships of bachelor percentage and household income get further from their true values. If we decrease the parameter, the opposite occurs. This is a somewhat happy medium, where the relationships aren't completely accurate, but is overall a reasonably fair representation.

As things stand with this model, the correlation between bachelor percentage and hap-

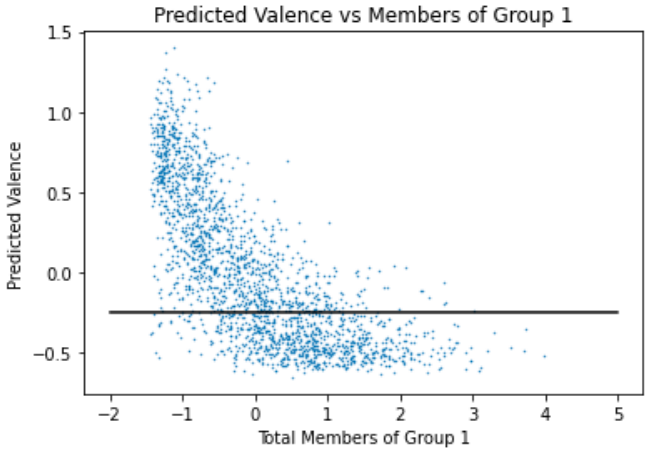

FIG. 17. Group 1 vs Happiness

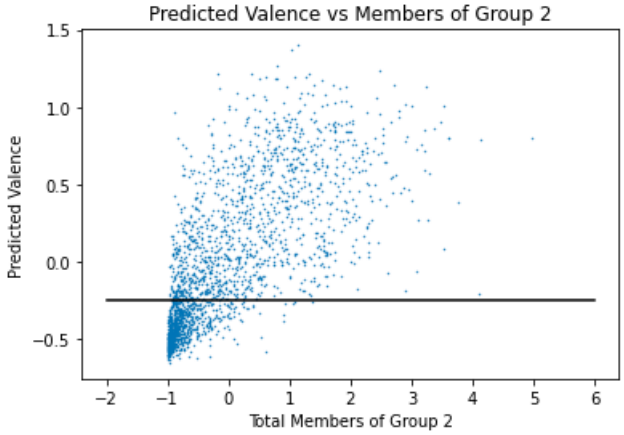

FIG. 18. Group 2 vs Happiness

piness is .67, which is higher than the ground truth value of .31. The correlation between income and happiness is .32, which is very similar to the ground truth happiness of .31. The correlation between ethnic group 1 and happiness is -.137, not too far from the ground truth of -.361, while the correlation between ethnic group 2 and happiness is .19, also not too far from its true value of .335.

VII. CONCLUSION

Overall, we see that we are able to create a reasonably fair model to predict happiness in a region. Through the use of the functions provided by our Research and Development team, our linear model does a decent job of preserving the relationships seen in our ground truth

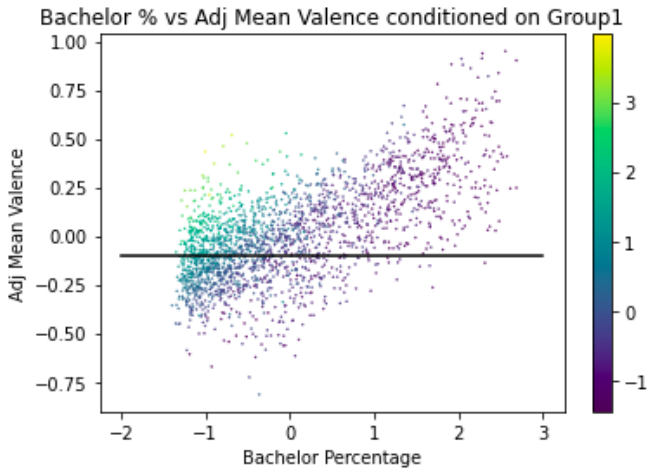

FIG. 19. Bachelor percentage vs Happiness, conditioned on group1

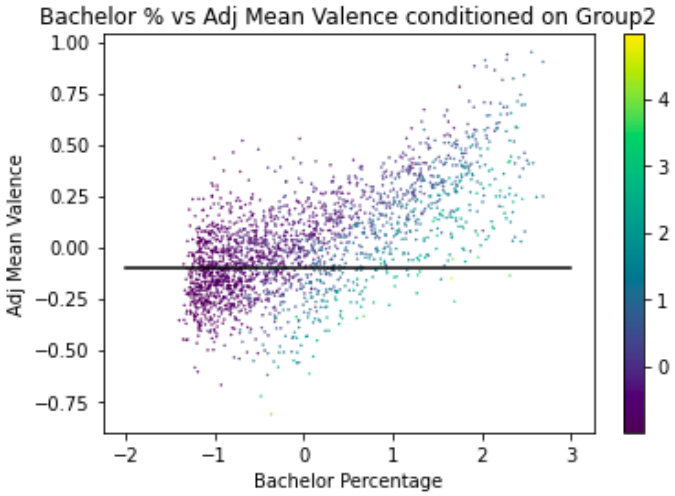

FIG. 20. Bachelor percentage vs Happiness, conditioned on group2

model, and far better than our initial attempts at modeling without these functions. We can be confident that a deployment of these models would not display unfair bias towards either ethnic group.

---

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12**, 2825 (2011).

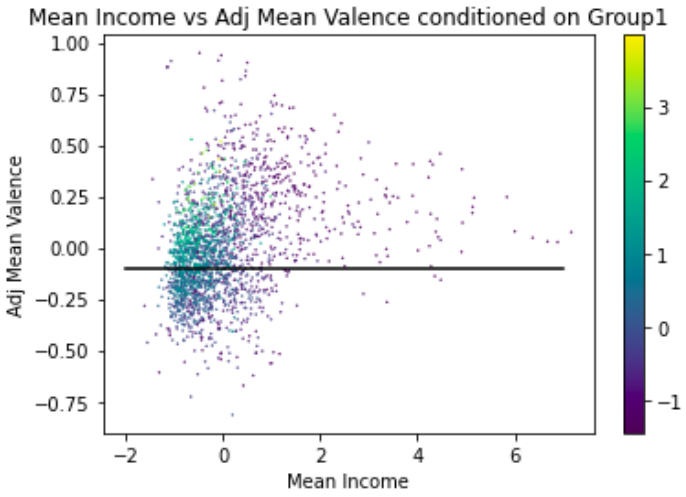

FIG. 21. Household income vs Happiness, conditioned on group1

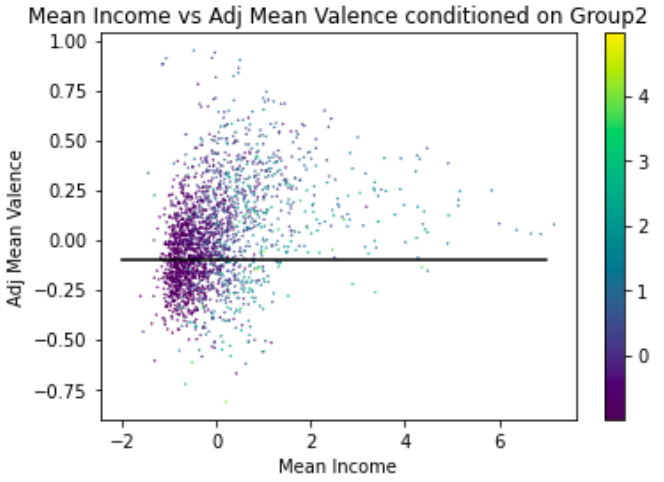

FIG. 22. Household income vs Happiness, conditioned on group2

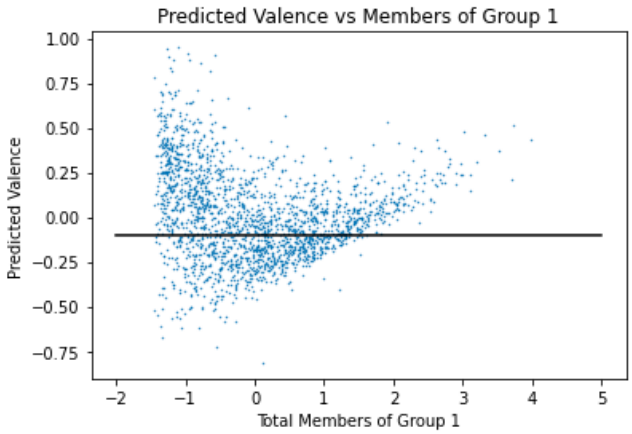

FIG. 23. Group 1 vs Happiness

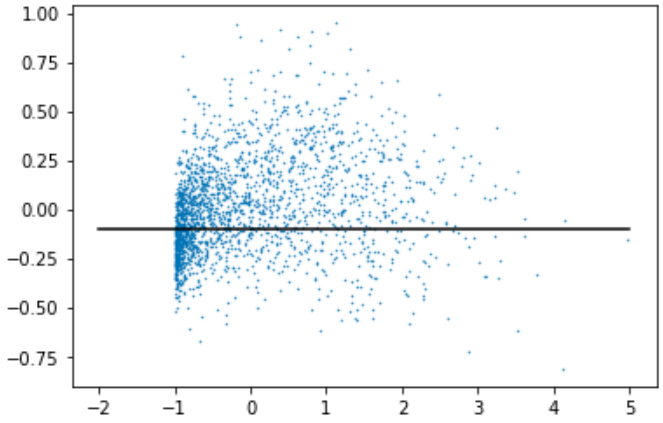

FIG. 24. Group 2 vs Happiness