Discover Discriminatory Bias in High Accuracy Models Embedded in Machine Learning Algorithms

Jianing Zhang and Vibhu Verma

Department of Decision Sciences
The George Washington University, Washington, DC 20052, USA
avecsally@gwu.edu; vibhuverma@gwmail.gwu.edu

Abstract. For all the excitement about Machine Learning Algorithms, there are serious impediments to its widespread adoption. Accuracy is not the only criteria in measuring the model performance, in real life, current model assessment techniques, like cross-validation or receiver operator characteristic (ROC) and lift curves, simply don't tell us about all the nasty things that can happen such as Opaqueness, Social discrimination etc. within the models. And that's why model debugging, the art and science of understanding and fixing problems in Machine Learning models, is so critical to the future of Machine Learning. Without being able to troubleshoot models when they under perform or misbehave, organizations simply won't be able to adopt and deploy the algorithm for good and at scale. Inspired by this, we want to challenge the models that have very high accuracy, conduct model debugging and discrimination testing to discover the hidden inaccuracy. The results bring up the concern of how a seemingly reliable model can present bias that would be damaging if actually deployed in the future ...

Keywords: XGBoost, Model Debugging, Discrimination Testing

1 Introduction

Pursuing the best performing model that eliminates any human intervention is undoubtedly the optimal solution for any industry. Most of the time, data scientists train the model regardless of sensitive social or ethical issues. Nonetheless, all models are made by humans and reflect human biases.

The unwanted bias can creep into the models no matter how accurate our model performs. When we deploy the model in a professional setting, it will be affecting real lives and data scientists should have urged caution and skepticism in using the algorithm. While we cannot do much in fighting to change injustices in the world and discrimination that happens outside of the machine learning systems, there is a huge potential to use algorithms to make a more equal society. Only when we can make algorithms as unbiased as possible so that they do not perpetuate social injustices that are embedded in the data so that we can trust

the algorithms. In this paper, we will be focused to discover the hidden discrimination fact from those seemingly accurately performed models and bring up the discussion upon the challenges we have in finding the optimal best performing model.

2 Related Works

Bias exist in many stages of the machine learning process: from how we frame the problem to collect and process the data, and it is hard to fix because the computers are not designed to detect it. [2] The Gender Shades Study done by MIT [3] showed that machine learning algorithms can discriminate based on classes like race and gender in the facial recognition field. The darker-skinned females are the most misclassified group (with error rates of up to 34.7%), while the maximum error rate for lighter-skinned males is only 0.8%. Bias also exists in the health system, Trishan et al.[11] defined it for the first time that algorithm bias exists "when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities in health systems."

With more studies shedding light on the existence of discrimination in AI-based products, Navdeep et al.[8] have provided an optimized workflows in 2019 for machine learning applications that require high accuracy and interpretability that mitigate risks of discrimination by introducing interpretable models, post-hoc explanation, disparate impact and discrimination testing into the workflow.

3 Methodology and Experiments

3.1 Methodology

The approaches we have taken to identify the discrimination in a model are as follows: first, we picked a popular data set from UCI - Census Income Data set to predict whether the person's income exceeds 50K dollars per year. Then, train the data set with different machine learning algorithms to optimize the accuracy by picking out the best AUC number, which provides an aggregate measure of performance across all possible classification thresholds. After carefully compare each models' performance regarding to their accuracy, we select the best model and dig into each social groups to test whether the discrimination exists. Discrimination test will then be done measured with the Youden Criteria threshold to find the accepting range and compares the predicted results with the actual values to get a clear map of how the model is presenting bias for different social groups. By checking the Adverse Impact Ratio (AIR), Standard Mean Difference (SMD) and Marginal Effect (ME) would show us how the best performing model

can actually predict the results successfully. Our paper will not be focusing on how to improve the models' performance on those biased groups but to conduct the residual analysis on them and provide possible explanations regarding the existence of the bias.

3.2 Model Selection

The UCI Adult data set contains 15 attributes (age, gender, educational level, capital gain and loss, mirital status, hours worked per week and etc.) with more than 32000 records to determine the person's income level (greater or less than $50K).

Generalized Linear Models (GLM) has been heavily relied upon by data scientists and statisticians for its flexible performance for ordinary linear regression that allows response variables. [13] While Gradient Boosting Machines (GBM) are more sophisticated by using decision tree algorithms to capture the model, they are more accurate in predicting non-linear relationships. [9] Both XGbBoost and GBM follow the principal gradient boosting algorithm but XGBoost uses a more regularized model formalization to control over-fitting and gives a better performance.

While these three algorithms are popular for their efficiency and accuracy, we tested the UCI Adult data and trained models using all three algorithms. Clearly XGBoost outperformed the other two with the highest accuracy rate of 92.81%, which means the model can predict weather income level larger than $50K or not with more than 90% chances of success.

Table 1. GLM, GBM and XGBoost Model Performance Comparison

	Model Performance Comparison					
	MSE	**RMSE**	**Logloss**	**AUC**	**AUCPR**	**Gini**
GLM	0.10106	0.3179	0.3179	0.90888	0.77125	0.81775
GBM	0.09034	0.30056	0.28856	0.92344	0.81766	0.84688
XGBoost	0.08753	0.29585	0.27546	0.92806	0.82646	0.85611

3.3 Model Interpretation

Now that we have found the most accurate model, we want to take the step further to discover the real performance of the model. SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. Shapley values correspond to the mean contribution of a feature and allows us to compare how the model weighs features compared to the variable importance and Pearson Correlation.

Picking up a random person, the prediction starts from the baseline. In the force plot, it is the average of all predictions. Each feature value is a force plot

that either increases or decreases the prediction. And each Shapley value is an arrow that pushes to increase (positive value) or decrease (negative value) the prediction. We can see that "capital-gain" and "education-num" are the two most positively affecting variables for the model output pushing it away from the base value 0.026 up to 0.43, while on the other side, 'hours-per-week", "workclass" are dragging the probability down. These forces balance each other out at the actual prediction of the data instance.

From the feature importance plot, we can observe how each variable is affecting the model output. Age is the most important feature, changing the predicted absolute probability of income class by 70 percentage points. Followed along is education level and capital gain by 50 percent and 45 percent. Clearly the income level has its nature highly correlated with educational level, hours work per week, capital gain and loss and etc., these attributes will not be considered as discriminated factors in the model, instead, we will mainly focus on the existence of discrimination for Gender and Race social groups.

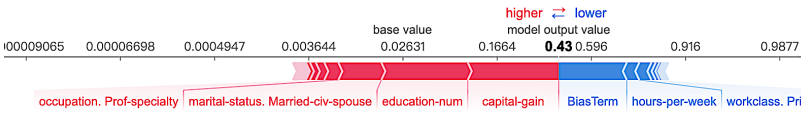

Fig. 1. SHAP explanation Force Plot for a Random Person in UCI Adult data set

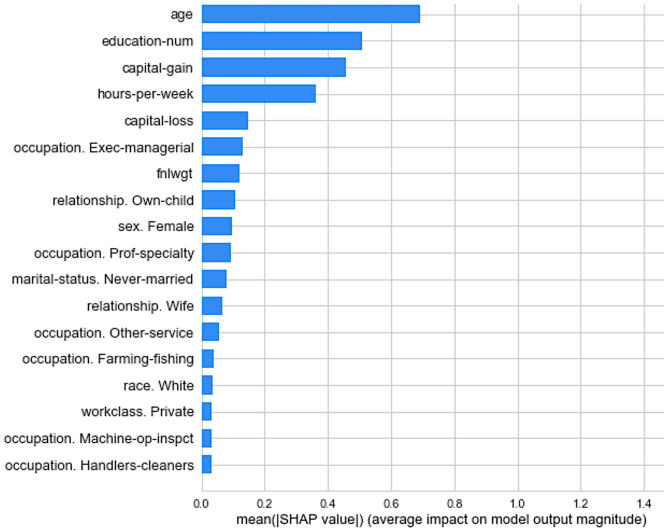

Fig. 2. SHAP Feature Importance Histogram UCI Adult data set

3.4 Discrimination Testing

Youden's J Statistic[14] is a single statistic that captures the performance of a dichotomous diagnostic test. A value of 1 indicates that there are no false positives or false negatives, which means the test is perfect. Choosing a threshold range would help to identify the discriminated groups. Based on this best cut-off point criteria of [0.8,1.3] chosen for the highest F1 score, [12] we can test the existence of discrimination based on different Gender and Race groups. By looking into the disparity table will help us gain the insight of the disparity between different groups. Any statistics that are not within the range used bold font in Table 2.

Table 2. Disparity Table for Gender (Male as Reference Group) and Race (While as Reference Group)

Disparity Ratio Table for Gender and Race

		Prevalence	Accuracy	Precision	False Omission Rate
Gender	Female	**0.358023**	**1.251710**	**0.428902**	**0.351136**
	Male	1.000000	1.000000	1.000000	1.000000
Race	Amer_Indian_ Eskimo	**0.452418**	1.152264	**0.480303**	**0.449047**
	Asian_Pac_Islander	1.038224	0.989363	0.949305	1.046497
	Black	**0.484170**	1.157155	**0.467897**	**0.485789**
	Others	**0.360552**	1.185365	**0.167062**	**0.378884**
	White	1.000000	1.000000	1.000000	1.000000

Precision Ratio shows how well the model classifies the groups. Males have a higher precision ratio for the higher salary subgroup compared to Females. For every 100 Males correctly classified into higher salary subgroups, we have only 43 Females. While 100 Whites are correctly classified as a higher salary subgroup, only 16 Others racial identities successfully be considered as in higher income levels.

Prevalence Ratio returns the true condition of the data set, for every 100 Males getting a salary more than $50K, we have only 35 Females. For every 100 Whites getting more than $50K, no more than 50 Blacks, American Indian Eskimo or Other racial groups. This disparity might have existed as a part of our data set, and even within our Machine Learning algorithm. This may also imply that the model would wrongly classify those groups that are not in favor - Female, Black, American Indian Eskimo or Other racial groups - into the lower salary subgroup.

Now we want to know by what proportion the model misclassified the groups. The False Omissions Rate measures the proportion of false negatives which are incorrectly rejected. With that of Females lower than Males, it is the fact that more of the less deserving Males is getting favored with their salaries as compared to less deserving Females. Similar in the Race table, the False Omission Rate is high for both White and Asian Pacific Islanders, which means these groups

who should be classified as negative class (earnings less than $50K) were wrongly classified to positive class (earnings greater than $50K). These can be interpreted as our model is more in favor of Male to Female, and more in favor of Whites and Asian Pacific Islander to other racial groups.

Adverse Impact Ratio (AIR) is a well-known discrimination measure. Adverse impact refers to practices that appear neutral but have a discriminatory effect on a protected group. Any values below 0.8 can be considered evidence of illegal discrimination in many lending or employment scenarios. In Table 3, we can see an almost ideal result where the protected and reference groups have very similar acceptance rates and AIR is near 1.

Standard Mean Difference (SMD) is used as a summary statistic in analysis when the studies assess the same outcome for different groups. SMD has prescribed thresholds: 0.2, 0.5, and 0.8 for small, medium, and large differences, respectively. Running a SMD for Gender and Race we can see that there's no difference in the model prediction for both the labels in both for Gender and Race. This means that the mean difference in all the classes is small and the model predictions are doing fairly well for each class.

Marginal Effect (ME) describes the difference between the percent of the reference group awarded a salary greater than $50K and the percent of the protected group awarded a salary more than $50K under our model. The marginal effect numbers are negative in Table 4, indicating that a higher percentage of individuals in the protected group were getting higher salary than in the reference group, this value would likely not indicate a discrimination problem in most scenarios.

Table 3. Adverse Impact Ratio (AIR), Standard Mean Difference (SMD) and Marginal Effect (ME) for Gender and Race

	Reference Group	Protected Group	AIR	SMD	ME(%)
Gender	Male	Female	1.001	0	-0.70%
Race	White	Black	1.007	0.09	-0.44%
	White	Others	1.005	0.03	-0.63%

3.5 Residual Analysis

Figure 3 shows the results of model residuals including global value and different residuals in Gender and Race subgroups. Residuals are a numeric measurement of model errors, essentially the difference between the model's prediction and the known true outcome. Residual plots helps us to get where influential outliers, data-quality problems, and other types of bugs often become plainly visible. Here, the pink line corresponds to the class of people who had salaries greater than $50K and blue - who did not. There appears to be several cases in the validation data where we have very high residual values as the outliers are presented on the global log loss residual plot. To understand how the values of this

input variable affects prediction error, we can see that our model is struggling to accurately predict cases where race is "Amer-Indian-Eskimo" and "Other". Due to the fact that we have fewer values in the data set and the model predicts most of them inaccurately, so we have to be extremely careful when testing such data sets.

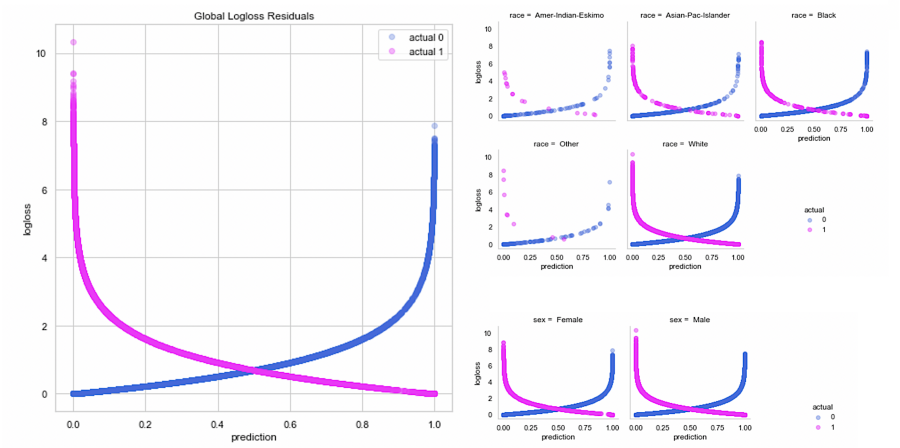

Fig. 3. Residual Plots for the while UCI Adult data set (left), Race (right top), and Gender groups (right bottom)

While it may seem that the model performs similarly for both the genders, but the model is not performing well to predict classes accurately. There are a lot of Male who should have been classified to lower income level class but actually been predicted as higher level, while Females are facing the opposite. We should throw some light on these incorrectly observed values and take more efforts in understanding the data quality issues to debug the model and lead to a more accurate model. Possible solution to overcome this is by human intervention with more targeted algorithms regarding specific groups to get an estimate.

4 Discussion

From the above experiments, we have conducted Discrimination Testing and Residual Analysis to test the fairness of the best performing model which gives us the highest AUC number (the aggregate measure of performance across all possible classification thresholds). The results confirm that even with our best performing XGBoost model, which gives 92.8% accuracy, the discrimination still exists. Regardless of direct indicators of income level such as educational level, hours work per week, capital gain and etc., the focus of the discrimination testing was around the Gender and Race identity groups.

Extensive evidence showed that Males are favored by our model than females by having higher precision ratio of 1.0 while Female only 0.43 and false omission rate of 1.0 while Female 0.35. This disparity may have existed before the conduction of Machine Learning Algorithm, by checking the prevalence ratio would give us the true condition of the data set. The answer was not surprising holding Males' prevalence ratio as 1.0, Female only 0.36. As for different racial groups, Whites clearly are favored by the model. Simply compare the number of precision ratio and false omission rate ratio, except for Asian Pacific Islander group, all other racial groups fell below our threshold of 0.8. Not surprisingly, the prevalence ratio for those groups are also low compared to that of Whites. These indicate the bias are there even before the implement of the Machine Learning Algorithm. All these differences can be accounted as data quality issues or discrimination due to under training of a few classes compared to others. Algorithm itself would not create discrimination. By checking the Adverse Impact Ratio (AIR), Standard Mean Difference (SMD) and Marginal Effect (ME) gives us the faith that the best performing model can actually predict successfully in most scenarios with no discrimination problems. However, if the data set contains the unfair information, algorithm would learn from it and the discrimination would become robust in the workflow since algorithms are not designed to eliminate the discrimination. This result is consistent with what has been found in previous Gender Shades Study by MIT and can be treated as a motivating example to show the need for increased transparency in the performance of any AI products and any services that focused on human subjects. Ongoing oversight and context limitations are needed to provide appealable outcomes that exhibit minimal social discrimination.

5 Conclusion

Machine Learning models learn from data to become accurate, and Machine Learning models require data that's truly representative of the entire problem space being modeled. If a model is failing, adding representative data into its training set can work wonders. Data augmentation can be a remediation strategy for discrimination in Machine Learning models, too. Through our study, we implemented different model debugging techniques to do the analysis and demonstrated that simple analysis can catch many errors and that they can be used as a form of model supervision as well as model debugging technique to significantly improve model accuracy.

While everyone wants reliable Machine Learning models, we must make sure we have addressed all the data quality issues in retrieving the data and collaborate with the domain experts to impute the missing data rather than statistics, which would give us a more realistic data set. Machine Learning is a powerful tool. We still believe there is a huge potential in using algorithms to make a more fair and equal society so that people can trust algorithm and data to perpetuate unknowing social injustices that are embedded in the history.

References

1. Dua, D., & Graff, C. (2017). UCI machine learning repositor. http://archive.ics.uci.edu/ml/
2. Hao, K. (2020, June 12). This is how ai bias really happens-and why it's so hard to fix. MIT Technology Review. https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/
3. Joy, B., & Timnit, G. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15.
4. Kenneth, M. (2002). Disparate results in adverse impact tests: The 4/5ths rule and the chi square test. Public Personnel Management, 31, 2.
5. Lundberg, S. (2019, Dec 12). Shap(shapley additive explanations). GitHub.https://shap.readthedocs.io/en/latest/
6. McKenna, M. (2019, October 14). Three notable examples of ai bias. AI Business.https://aibusiness.com/document.asp?doc_id=761095&site=aibusiness
7. Mishel, L. (2013, January 30). Vast majority of wage earners are working harder, and for not much more. Economic Policy Institute, 348. https://www.epi.org/publication/ib348-trends-us-work-hours-wages-1979-2007/
8. Navdeep, G., Patrick, H., Kim, M.,& Nicholas, S. (2020, February 29). A responsible machine learn- ing workflow with focus on interpretable models, post-hoc explanation, and discrimination testing. information, 11, 137.
9. Proksch, M. (2020, June 09). From glm to gbm. towards data science. https://towardsdatascience.com/from-glm-to-gbm-5ff7dbdd7e2f
10. Stewart, M. (2019, March 30). Handling discriminatory biases in data for machine learning. towards data science. https://towardsdatascience.com/machine-learning-and-discrimination-2ed1a8b01038
11. Trishan, P., Heather, M., & Rifat, A. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. J Glob Health, 9.
12. Unal I. (2017). Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. Computational and mathematical methods in medicine, 2017, 3762651. https://doi.org/10.1155/2017/3762651
13. Wikipedia. (2020). Generalized linear model — Wikipedia, the free encyclopedia [Online; accessed 15-July-2020].
14. Wikipedia. (2020). Youden's J statistic — Wikipedia, the free encyclopedia [Online; accessed 15- July-2020].