Debugging the Black-Box COMPAS Risk Assessment Instrument to Diagnose and Remediate Bias

Patrick Hall [1] **Navdeep Gill** [1]

Abstract

The black-box Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) criminal risk assessment instrument (RAI) is analyzed for confounding racial bias and a novel procedure is proposed for remediating bias from individual criminal risk predictions. A repeatable *global versus local* analysis motif is introduced in which global and local model behavior are compared to debug and diagnose unwanted bias in a black-box prediction system using tools such as surrogate models, gradient boosting machine feature importance, leave-one-covariate-out (LOCO) feature importance, partial dependence plots, and individual conditional expectation (ICE) plots. LOCO-derived feature importance is also used to remove prediction contributions from bias-inducing input features. The proposed *global versus local* approach and remediation strategy can be applied to many black-box and machine learning (ML) decision-making systems.

1. Introduction

Many criminologists and criminal justice reformers agree that entirely too many people are imprisoned in the U.S. (Travis et al., 2014). While RAIs, somewhat like the COMPAS instrument analyzed here, have been used to lower pretrial incarceration rates for some populations (Fratello et al., 2011), they are applied across the U.S. to determine the risk of pretrial flight, the level of service an inmate or other individual should receive while under penal supervision (Christin et al., 2015), and are now being used or considered for use in sentencing decisions (Hyatt & Chanenson, 2017).

Proponents of these instruments believe they can be used to make data-driven, objective decisions about an individual's future criminal risk and reduce jail and prison populations by diverting individuals that score low or medium on risk scales to alternative supervision programs (Arnold & Arnold, 2016). Detractors have raised pointed concerns about racial bias in these instruments; in particular, investigative journalists at ProPublica recently uncovered potentially serious problems in COMPAS (Angwin et al., 2016). However, a subsequent rejoinder from a group of respected criminologists pointed out noteworthy flaws in the ProPublica analysis (Flores et al., 2016). Is COMPAS biased? This paper will present evidence that the instrument behaves as expected globally, but locally it appears to produce biased risk scores for certain individuals.

In the following sections, decision tree and gradient boosting machine (GBM) surrogate models (Craven & Shavlik, 1996) are trained to simulate COMPAS risk scores using demographic features collected by ProPublica. Then several additional model debugging techniques including:

- GBM feature importance (Friedman, 2001)
- Leave-one-covariate-out (LOCO) feature importance (Lei et al., 2017)
- Partial dependence plots (Hastie et al., 2008)
- Individual conditional expectation (ICE) plots (Goldstein et al., 2015)

are employed to compare global and local behavior in COMPAS predictions and diagnose confounding bias stemming from a latent race feature. Finally, using COMPAS as a representative example, a novel, prototype idea is introduced for numerically remediating the impact of unwanted bias in black-box ML and artificial intelligence (AI) prediction systems.

1.1. The Multiplicity of Good Models

It is well understood that for the same set of input features and prediction targets, complex ML algorithms can produce multiple accurate models with very similar, but not exactly the same, internal mechanisms (Breiman, 2001). This alone is an obstacle to interpretation, but when using

[1]H2O.ai, Mountain View, California, USA. Correspondence to: Patrick Hall <phall@h2o.ai>.

these types of algorithms as interpretation tools, or with interpretation tools, details of the explanation can change across multiple accurate models. In the reported results, conclusions are drawn only from trends and results seen across multiple accurate surrogate models or across repeated applications of interpretability tools.

2. Surrogate Models

A surrogate model is typically trained to predict the output of another intractably complex or opaque black-box model using some set of interesting training features as inputs and the predictions of the more complex model as a target. A trained surrogate model enables a heuristic understanding of the complex model's internal mechanisms. In this case, both the original inputs and internal mechanisms of the proprietary COMPAS instrument are unknown. By using the data set constructed by ProPublica[1], in which COMPAS risk scores are manually associated with an individual's demographic attributes, surrogate models can be trained in which age, race, and criminal history are used as input features to predict COMPAS risk scores.[2]

2.1. Single Decision Tree

By training a single decision tree, the resulting surrogate model displayed in Figure 1 is a global, approximate flow chart for the COMPAS instrument's decisions. According to the flow chart, the most important features in determining an individual's COMPAS risk scores are an individual's number of prior convictions and an individual's age. The path to the leaf node with the highest normalized risk score value (0.40993544) indicates that individuals younger than 34.5 years of age with more than 10.5 prior convictions are generally the most likely to receive the highest risk score from COMPAS. Conversely the path leading to the leaf node with the lowest normalized risk score value (-0.23085095) shows that those individuals with less than 4.5 prior convictions and an age of greater than 37.5 years are typically expected to receive the lowest COMPAS risk scores.

Race is notably missing from this simple surrogate model, potentially indicating that race is not a first order consideration in the COMPAS instrument. However, the RMSE of this surrogate model is approximately 2.7, out of the ten-point COMPAS risk score scale. Hence, this surrogate model can provide only an overview of the COMPAS instrument's decision making processes. Next a highly accurate surrogate model is trained to provide more refined insights into COMPAS.

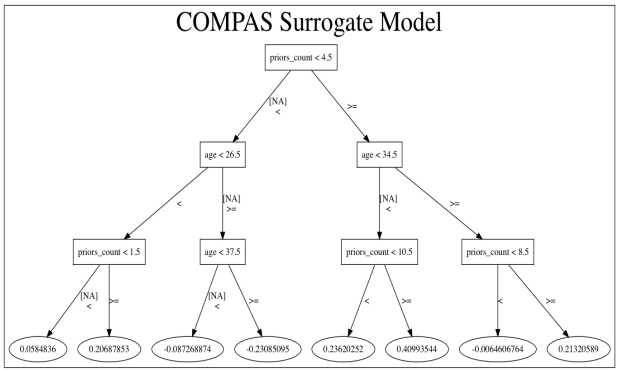

Figure 1. A global, graphical depiction of a decision tree surrogate model for the COMPAS instrument.

Table 1. Hyperparameter values for the GBM surrogate model.

HYPERPARAMETER	VALUE
COL_SAMPLE_RATE	0.6
COL_SAMPLE_RATE_CHANGE_PER_LEVEL	1.1
COL_SAMPLE_RATE_PER_TREE	0.68
HISTOGRAM_TYPE	ROUNDROBIN
MAX_DEPTH	15
MIN_ROWS	1
MIN_SPLIT_IMPROVEMENT	1E-8
NBINS	112
NBINS_CATS	3344
NTREES	201
SAMPLE_RATE	0.4

2.2. Gradient Boosting Machine

A random grid search strategy (Bergstra & Bengio, 2012) is used to train a gradient boosting machine surrogate model for the COMPAS model on the ProPublica data set. 250 candidate models are considered in the random search. The hyperparameters of the most accurate model found in the search are presented in Table 1 for reproducibility purposes. This model is used for all further analysis.

The RMSE of the GBM surrogate model is approximately 0.2. Since COMPAS risk scores are on a scale of 1 to 10, this indicates that the GBM surrogate model simulates the risk score predictions generated by COMPAS with only 2% error. Although COMPAS reportedly uses over 100 inputs, none of which are an individual's race, this surrogate model can very accurately simulate COMPAS in the data provided using only 11 features, including an individual's race.

[1]Data and other information from the ProPublica analysis is publicly available: https://github.com/propublica/compas-analysis.

[2]According to Flores (2016), the data on pretrial individuals collected by ProPublica may not be well-suited to analyze their associated COMPAS scores.

3. Global and Local Feature Importance Comparison

Global feature importance is calculated following Friedman (2001). Local feature importance is calculated using a variant of the LOCO technique (Lei et al., 2017) as defined in Equation 1.

$$I_{i,j} = \hat{y}(X_i) - \hat{y}(X_{i,(-j)}) \tag{1}$$

Each local feature importance $I_{i,j}$ is found by re-scoring the trained surrogate model for each row *i* while setting the feature of interest *j* to missing giving $\hat{y}(X_{i,(-j)})$. This modified prediction is then subtracted from the original prediction $\hat{y}(X_i)$ to find the raw importance for feature *j* in row *i*. Local feature importance values are then scaled for direct comparison with global values. As these local feature importance values can be misleading in the presence of highly correlated input features (Adebayo, 2016), Pearson correlations were calculated and found to be acceptable before evaluating 1 for the input features and GBM surrogate model.

Local feature importance values are compared to global values for two test individuals to debug any unexpected contributions from race in COMPAS. One test individual is a relatively older Caucasian male facing felony theft charges with two prior convictions. The other is a younger, African-American male arrested on possession of cannabis with no prior convictions. The COMPAS model gives these two individuals drastically different risk scores. The Caucasian male is predicted to be low risk with a numeric score of 3, whereas the African-American male is rated as a very high risk with a numeric score of 10. Both global and local feature importance values provide insight into how COMPAS risk scores are influenced by age, criminal history, and the latent race feature in the ProPublica data. Figure 2 provides a visual comparison of the feature importance values for the two test individuals to each other and to the global feature importance values.

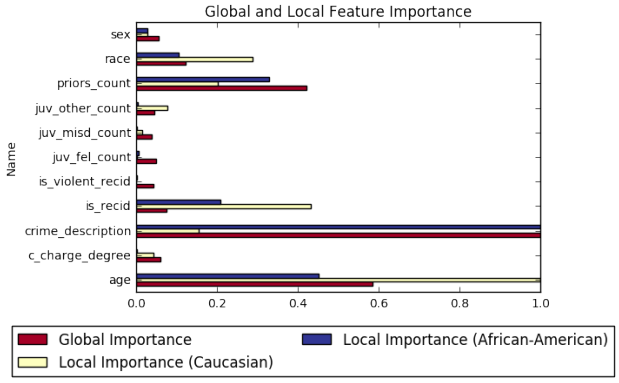

Figure 2. Global feature importance compared to local feature importance for the test individuals.

Between the global and local feature importance values, age and criminal history information dominate the COMPAS instrument decision process, as is expected. It is also clear that race plays a role in COMPAS predictions. Some feature or features must have pushed the African-American male's very high score away from the lower score assigned to the Caucasian male. Given that the African-American male received the highest possible score, it is not unreasonable to assume that each important feature, including race, likely pushed him up toward this highest possible score. (Numerical evidence for this assertion is presented in Section 6 using raw local feature importance values.) Given his feature importance values, it is logical for his younger age and drug arrest to have increased his risk score, but it seems highly questionable that an individual's very low prior conviction count and race should also contribute to a higher risk score.

4. Global Partial Dependence and Local ICE Comparison

Partial dependence plots are constructed following Hastie et al (2008). ICE plots, a newer and less well-known adaptation of partial dependence plots, can be used to create localized explanations by performing a sensitivity analysis of model predictions for a given individual and input feature. ICE plots are constructed following Goldstein et al (2015).

$$ICE_{i,j,k} = \hat{y}(X_i, x_{k,j}), x_{k,j} \in x_j \tag{2}$$

ICE values are simply disaggregated partial dependence. When displayed for a single individual $ICE_{i,j,k}$ they represent the model predictions for a single row *i*, where a feature of interest *j* is varied over its domain *k*. Overlaying ICE plots onto partial dependence plots enables the debugging of a model's treatment of certain individuals by comparing an individual's local predictions to the model's global predictions across the domain of an input feature.

4.1. Partial Dependence and ICE for Age

As expected, global, average COMPAS risk scores decrease with increasing age as displayed in Figure 3. However, an African-American male with no prior criminal history and a Caucasian male with a short criminal history appear to be treated differently for almost all values of age. (For this specific surrogate model, the African-American male is nearly always assessed with a noticeably higher risk score.) It seems logical that predictions would converge at high age as risk diminishes, but that does not occur here. Unless there is some non-obvious criminological justification for treating such individuals differently at all ages, this finding could point to a potential racial bias in the COMPAS risk scores.

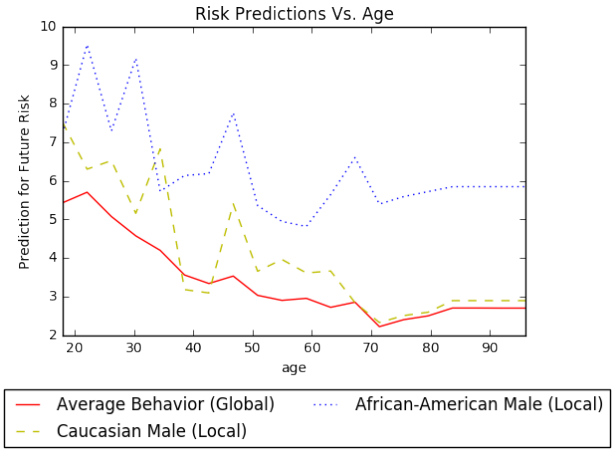

Figure 3. Partial dependence compared to ICE for the test individuals across simulated ages from 18 to 96.

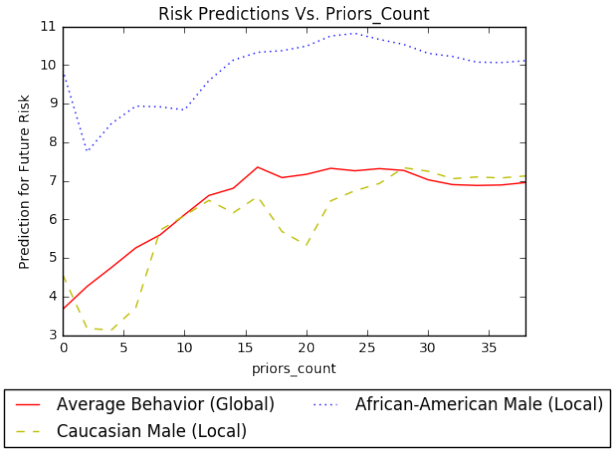

Figure 4. Partial dependence compared to ICE for the test individuals across simulated prior counts from 0 to 38.

4.2. Partial Dependence and ICE for Prior Convictions

Global, average COMPAS risk scores increase as prior convictions increase. Again, this prediction behavior agrees with generally accepted standards. Yet, comparing partial dependence and ICE for the number of prior convictions for the two test individuals exposes an unnerving pattern in the COMPAS predictions as presented in Figure 4. No matter the number of prior convictions, a young African-American male facing drug charges would always be assessed a much higher risk score than an older Caucasian male facing felony theft charges and always be assessed a much higher than average score. Moreover, this difference is starkest for individuals with no or short criminal histories. This behavior is observed across multiple accurate surrogate models.

4.3. Partial Dependence and ICE for Race

Figure 5 compares the surrogate model's average treatment of race with its treatment of race in the test individuals. The partial dependence of the model on each race is roughly equivalent, although the partial dependence for African-American scores is slightly higher than for other groups. Also, using ICE to treat the African-American male as every other race still results in the highest risk scores for each race group, indicating that something besides race is driving the test individual's very high risk score.

However, comparing the two test individuals again points to the same pattern of potential racial bias:

- Using ICE to treat the African-American male as a Caucasian male with all other traits held constant would result in his risk score being cut by roughly one-half to two-thirds across several tested surrogate models.
- Using ICE to the treat the Caucasian male as an African-American male holding his other traits constant would nearly double his score across several tested surrogate models.

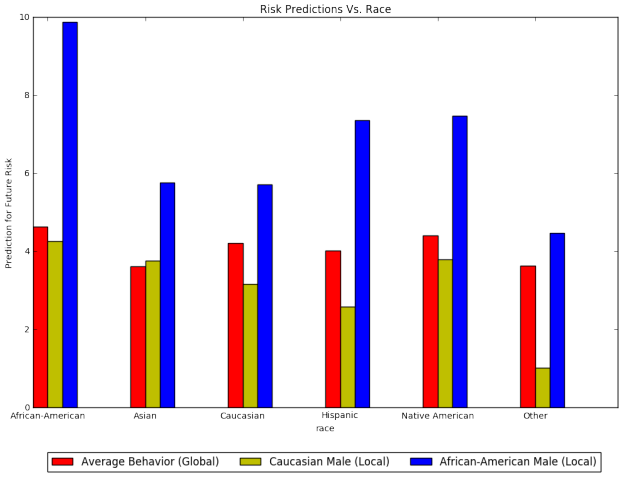

Figure 5. Partial dependence compared to ICE for the test individuals across simulated races.

5. Conclusions

This analysis highlights a set of ideas and tools that can be used to examine the many and growing number of ML and AI systems that are impacting human lives in important, and sometimes undesirable, ways. It is the authors' opinion that the use of these systems will continue to grow in number and societal impact. If so, more tools to understand

their decisions and debug and remediate their mistakes will be needed in coming years.

By applying the proposed *global versus local* analysis strategy to COMPAS, expected global behavior is confirmed, but compelling local patterns of racial bias are exposed for the test individuals. According to ProPublica, COMPAS does not consider race explicitly (Angwin et al., 2016). However, race likely cannot be excluded from a model where any available training data could be tainted by institutional racial bias (Travis et al., 2014). More likely race is a latent, confounding feature in COMPAS that interacts with all other important input features resulting in biased predictions for certain individuals.

While the analyzed COMPAS instrument is not recommended for sentencing decisions, it is possible that it could be misused for serious incarceration decisions, that a previous risk score could linger in a judge's mind during sentencing, or that other risk tools could be used for serious incarceration decisions (Hyatt & Chanenson, 2017). Even in the best case, if racial bias in COMPAS model is limited to corner cases, it's unlikely that any amount of racial bias is acceptable to individuals who would be negatively affected by COMPAS model risk scores or other RAI risk scores. It seems obvious that even a model with minimal bias should not be used for decisions involving an individual's freedom. Unfortunately, this may not have been the case in the past and may continue to be a terribly unfair problem into the future.

6. Suggestions for Future Work

Future work entails applying differentially private learning techniques to build new risk models and testing a novel, prototype procedure to remediate any unwanted bias from COMPAS and other ML and AI decision-making systems.

6.1. Differential Privacy

If racial bias is present in training data, racial bias will appear in any model trained on that data. Differential privacy approaches offer a theoretical and practical framework for altering data in ways that could potentially reduce unwanted bias in models based on an individual's attributes while preserving predictive utility (Zemel et al., 2013). Differentially private learning techniques could be applied to build new risk models or COMPAS surrogates and the *global versus local* analysis strategy could be applied to test for successful bias remediation.

6.2. Prototype for Bias Remediation

When making decisions that have a large impact on people's live, no level of unwanted bias is acceptable. This necessity clashes with the realities of both statistical models, and frankly, the realities of human decision making. In the future, it may be better to accept that unwanted bias exists in most data and models and to use tools to identify and remove unwanted bias before scores are presented to human decision makers. A prototype procedure using local feature importance values for removing unwanted bias from predictions is presented in this section.

Table 2. Unscaled local feature importance values for the two test individuals.

FEATURE	AF.-AM. MALE	CAUC. MALE
SEX	-0.15	-0.08
AGE	2.35	-2.87
RACE	0.55	-0.80
JUV_FEL_COUNT	0.03	0.00
JUV_MISD_COUNT	-0.02	-0.04
JUV_OTHER_COUNT	0.02	0.22
C_CHARGE_DEGREE	-0.02	-0.12
IS_RECID	1.08	-1.20
IS_VIOLENT_RECID	-0.01	0.00
PRIORS_COUNT	1.72	-0.56
CRIME_DESCRIPTION	5.20	-0.43

The simulated risk score produced by the GBM surrogate model for the African-American test individual is 9.87, very close the actual COMPAS-assigned risk score of 10. Table 2 presents the unscaled local feature importance values for the two test individuals. These values can be used to decompose this score into signed, individual feature contributions.

Since the *global versus local* analysis calls the COMPAS instrument's treatment of race and priors count into question, unscaled local feature importance values are used to remediate the contributions of these features from the COMPAS risk scores. Also, in the case of the African-American test individual, his current crime *not* being classified as recidivism contributed positively to his simulated risk score, and the contribution of the `is_recid` feature is also be remediated. The new remediated score can be defined as

$$\hat{y}_{rem_i} = \hat{y}_i - \sum_r I_{(i,j)_r} \tag{3}$$

where $\hat{y}_{rem_i}$ is the remediated risk score for row *i*, $\hat{y}_i$ is the original risk score for row *i*, and each $I_{(i,j)_r}$ represents the local feature importance values for the *r* remediated features. For the African-American test individual, the remediated numeric score is

$$\hat{y}_i - (I_{i,race} + I_{i,priors\_count} + I_{i,is\_recid}) = 6.53 \tag{4}$$

The simulated COMPAS score for the Caucasian test individual is 3.17. Carrying out the same remediation procedure yields the remediated risk score in 5.

$$\hat{y}_i - (I_{i,race} + I_{i,priors\_count} + I_{i,is\_recid}) = 5.74 \quad (5)$$

Given that this Caucasian male individual went on to commit a serious theft and the African-American male went on to commit no additional crimes, these remediated risk scores may be presenting a more accurate overall evaluation of future criminal risk. Of course, broader testing is required.

Additionally, the unscaled values add more seriousness to claims asserted in Section 3 that these two individuals are treated differently by COMPAS based on their race. The unscaled local value for race for the African-American is positive, whereas the unscaled race feature importance value for the Caucasian male is the negative. While race does not contribute strongly to either prediction, it does contribute to a difference of 1.35 risk score points between the two test individuals.

References

Adebayo, Julius A. Fairml: Toolbox for diagnosing bias in predictive modeling. Master's thesis, MIT, 2016.

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

Arnold, Laura and Arnold, John. Reforming systems to improve lives: 2016 annual report, 2016.

Bergstra, James and Bengio, Yoshua. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

Breiman, Leo. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

Christin, Angele, Rosenblat, Alex, and Boyd, Danah. Courts and predictive algorithms, 2015.

Craven, Mark W. and Shavlik, Jude W. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, pp. 24–30, 1996.

Flores, Anthony W., Bechtel, Kristin, and Lowenkamp, Christopher T. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". *Federal probation*, 80 (2), 2016.

Fratello, Jennifer, Salsich, Annie, and Mogulescu, Sara. Juvenile detention reform in new york city: Measuring risk through research, 2011.

Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pp. 1189–1232, 2001.

Goldstein, Alex, Kapelner, Adam, Bleich, Justin, and Pitkin, Emil. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer, 2008.

Hyatt, Jordan M. and Chanenson, Steven L. The use of risk assessment at sentencing: Implications for research and policy. Public Policy Research Paper 2017-1040, Villanova Law, 2017.

Lei, Jing, G'Sell, Max, Rinaldo, Alessandro, Tibshirani, Ryan J., and Wasserman, Larry. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 2017.

Travis, Jeremy, Western, Bruce, and Redburn, Steve (eds.). *The Growth of Incarceration in the United States: Exploring Causes and Consequences*. The National Academies Press, 2014.

Zemel, Rich, Wu, Yu, Swersky, Kevin, Pitassi, Toni, and Dwork, Cynthia. Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning*, ICML-13:325–333, 2013.