Detecting Bias in Black-Box Models Using Transparent Model Distillation

Sarah Tan[*]
Cornell University
Ithaca, NY 14850
ht395@cornell.edu

Rich Caruana
Microsoft Research
Redmond, WA 98052
rcaruana@microsoft.com

Giles Hooker
Cornell University
Ithaca, NY 14850
gjh27@cornell.edu

Yin Lou
Airbnb Incorporation
San Francisco, CA 94103
yin.lou@airbnb.com

Abstract

Black-box risk scoring models permeate our lives, yet are typically proprietary and opaque. We propose a transparent model distillation approach to detect bias in such models. Model distillation was originally designed to distill knowledge from a large, complex teacher model to a faster, simpler student model without significant loss in prediction accuracy. We add a third restriction - transparency. In this paper we use data sets that contain *two labels* to train on: the risk score predicted by a black-box model, as well as the actual outcome the risk score was intended to predict. This allows us to compare models that predict each label. For a particular class of student models - interpretable generalized additive models with pariwise interactions (GA2Ms) - we provide confidence intervals for the difference between the risk score and actual outcome models. This presents a new method for detecting bias in black-box risk scores by assessing if contributions of protected features to the risk score are *statistically different* from their contributions to the actual outcome.

Introduction

Risk scoring models have a long history of usage in criminal justice, finance, hiring, and other critical domains that impact people's lives (Corbett-Davies et al.; Louzada, Ara, and Fernandes). They are designed to predict a future outcome, for example defaulting on a loan or re-offending. Worryingly, risk scoring models are increasingly used for high-stakes decisions, yet are typically proprietary and opaque.

One attempt to detect bias in risk scoring models could be to reverse engineer them. However, this can be stymied by the lack of access to all features and the same data sample used to create the model, or a means of determining how close the result is to the true, unknown model. On the other hand, one could study the actual outcome, testing for disparate impact through methods such as training a model to predict the outcome, removing, permuting, or obscuring a protected feature (Adler et al.; Feldman et al.), and then retraining the model to see if it changes. One challenge with this approach is that advance knowledge of which feature to act on is needed, and there may be biases in the data that are not *a priori* known.

We propose a third approach - transparent model distillation for bias detection - that examines both the risk score as well as the actual outcome, leveraging the difference between them to detect potential bias. The approach involves training two separate models: a transparent student model to mimic a black-box risk score teacher, and another transparent model to predict the actual outcome that the risk score was intended to predict. We then ask the question:

> *Are there systematic differences in the risk scoring model compared to the actual outcome?*

Of particular concern is when such systematic differences occur on protected features such as race. Hence, when we train both models, we intentionally include all features that may or may not be originally used to create the black-box risk score, even protected features, specifically because we are interested in examining what the model *could* learn from these variables.

The student models we use are interpretable generalized additive models with pairwise interactions (GA2Ms) (Lou, Caruana, and Gehrke; Lou et al.; Caruana et al.) based on a variant of short, boosted trees. We provide confidence intervals for the difference between two models of this class, which we then use to detect systematic differences between the risk score and actual outcome. We study this class of models in terms of accuracy, fidelity, and transparency, and conduct comparisons to other student model candidates.

Related Work

Chouldechova and G'Sell proposed an approach to identify subgroups where two classifiers differ in terms of fairness, recursively partitioning covariate space by testing for homogeneity of a fairness metric of interest (e.g. false positive rate) between the two classifiers (Chouldechova and G'Sell). Our approaches differ in that they compare two classifiers of the outcome whereas we compare a black-box risk score and the outcome. Zhang and Neill work on a single model, to identify subgroups where estimated outcome probabilities differ significantly from observed probabilities (Zhang and Neill). Like these approaches, we formulate bias detection as a statistical hypothesis testing problem.

Method

Let $\mathcal{D} = \{(y_i^S, y_i^O, \boldsymbol{x}_i)\}_{i=1}^N$ be a data set of size N, where y^S is a risk score and y^O is the actual outcome the risk score was intended to predict. Let $y^S = r^S(\boldsymbol{x})$ be the true risk

[*] Work performed during an internship at Microsoft Research.

scoring model. $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ is a vector of p features for person i, and x_j is the jth variable in feature space. Our goal is to automatically detect regions of feature space where the risk scoring model significantly differs from the actual outcome.

Transparent Model Distillation for Bias Detection

Let $\mathcal{M}$ denote a class of transparent models. We first describe the proposed distillation setup before delving into the choice of transparent model in the next section.

We train two models of class $\mathcal{M}$:

Transparent Model Distillation

(Model S - regression): **Student model of true black-box risk score teacher** r_S:

$$y^S = f^S(\boldsymbol{x}) \tag{1}$$

(Model O - classification): **Model of actual outcome**. g is the logistic link function:

$$g(y^O) = f^O(\boldsymbol{x}) \tag{2}$$

As in classic model distillation, if the student model (Model S) is a high fidelity approximation of its teacher, we can then attempt to understand r^S by looking at f^S. In other words, whenever the unknown r^S is needed, f^S is used instead. The model of the actual outcome (Model O) comes into the picture because r^S, now approximated by f^S, was designed to predict the actual outcome y^O - precisely what we trained f^O to predict.

Our goal is to detect regions of feature space with systematic differences between y^S and y^O. Examining the contribution of a protected feature to f^S compared to f^O will invariably yield differences. Hypothesis testing tells us if these differences are real and systematic, or due to random noise. Denoting feature x_j's contribution to the risk score student model and outcome model as $C^S(x_j)$ and $C^O(x_j)$ respectively, the null hypothesis of no bias - that a protected feature does not contribute to the risk score any higher (or lower) than it contributes to the actual outcome - is:

Detecting Bias Using Hypothesis Testing

Null hypothesis (no bias):

$$C^S(x_j) = C^O(x_j)$$

Two-sided alternative hypothesis:

$$C^S(x_j) \neq C^O(x_j)$$

The null hypothesis is rejected when the p-value, a function of $P(C^S(x_j) - C^O(x_j))$, is small. Hence we require a model class $\mathcal{M}$ where the probability distribution and uncertainty of the *difference* between a feature's contribution to two different models of the class is known or estimable.

Choice of Transparent Model Class $\mathcal{M}$

We use interpretable generalized additive models (Lou, Caruana, and Gehrke; Lou et al.; Caruana et al.), a class of transparent models based on a variant of bagged, short trees learned using gradient boosting. Its transparency stems from its additive form[1]:

$$g(y) = \beta_0 + \sum h_j(x_j) + \sum h_{jk}(x_j, x_k)$$

where each term $h_j(x_j)$ is a shallow tree restricted to only operate on one feature, and $h_{jk}(x_j, x_k)$ is again a shallow tree but operating on two features. h_j is called the shape function of feature x_j and can be plotted against x_j in graphs such as the red or green lines in Figure 1. h_{jk} is the pairwise interaction of x_j and x_k and can be visualized in a heat map such as in Figure 2. This allows the contribution of any one feature to the prediction to be directly examined, making the model transparent. Multiple terms are learned together using gradient boosting to obtain an additive formulation, hence the name generalized additive models (GAMs). However, unlike classical GAMs where features are shaped using splines, tree-based GAMs shape features using short trees.

We do not use decision trees, a popular class of transparent models, as their instability (Gibbons et al.) may lead to spurious inferences of bias. Full complexity models such as random forests and neural networks are accurate but lack transparency and the contribution of individual features to the prediction, while visualizable using partial dependence plots, does not have estimable uncertainty, hence we do not consider them. Linear models (including logistic regression) are transparent and the difference in feature contributions between two models of the same class is estimable like for GA2Ms. We experiment on them as alternative student models below.

Characterizing the Distribution of Differences

Comparing the feature contributions across two models presents some challenges; the Appendix details how we solved them. Briefly, the first challenge is that one model (Model S) performs regression and the other (Model O) performs classification. We describe a way to make these feature contributions comparable. The second challenge is the several possible sources of variation can affect the difference between these feature contributions, leading to spurious judgments of bias. We use a structured bagging setup to remove avoidable sources of noise. Finally, we use a bootstrap-of-little-bags approach to estimate variance and covariances for feature contributions of this class of student models. These covariance estimates account for the fact that the same training observations with the same feature values were used to train the two different models.

Experimental Setup

Data

We use publicly available data sets containing both a black-box risk score and the actual outcome. Table 1 summarizes them.

(1) **COMPAS Risk Score and True Recidivism Outcome**: COMPAS, a proprietary score developed to predict recidivism risk, has been the subject of scrutiny for racial bias (Angwin et al., Kleinberg, Mullainathan, and Raghavan;

[1] g is the logistic link function for classification. For regression, g is the identity function.

Table 1: Description of data sets

Data	Number of	Number of	Label 1 Risk Score		Label 2 Outcome "Yes" or "No"
	Observations	Features	Scale*	Mean ± SD	Baseline Rate for "Yes" Class
COMPAS recidivism	6172	6	1 - 10	4.4 ± 2.8	0.46
Lending Club loans	42,506	28	A - G converted to 0 - 6	1.7 ± 1.4	0.15

* We use the convention that the higher the risk score, the more likely the outcome.

Chouldechova; Corbett-Davies et al.; Zafar et al.; Blomberg et al.; Dieterich, Mendoza, and Brennan). Because the algorithm does not use race as an input (Sam Corbett-Davies and Goel), its proponents suggest that it is race-blind. ProPublica collected, analyzed (Angwin et al.), and released data[2] on COMPAS scores and true recidivism outcomes of defendants in Broward County, Florida. Potentially protected features in this data set are age, race, and sex.

(2) **Lending Club Loan Risk Score and True Default Outcome**: Lending Club, an online peer-to-peer lending company, makes public information on the loans it finances[3]. We use a subset of five years (2007-2011) of loans. This time period was chosen because all loans have matured, and the outcome of whether the loan defaulted has been observed. We use only individual, not joint loans, and remove non-baseline features such as loan payment information that could leak information into the label. Candidates for protected features in this data include state.

Training Procedure

We train a GA2M regression model on risk score and a GA2M classification model on actual outcome. Each GAM model is trained using 5000 gradient boosting iterations, with the optimal number of iterations ($\leq$ 5,000) selected based on minimum validation set loss. For comparison, we also train random forest models and linear and logistic regression.

Evaluation Metrics

We assess *fidelity*, the notion that the student model should match the teacher model (Craven and Shavlik 1995), using the closeness of the student model to the teacher's predictions. In effect, this is the accuracy of the student model. We also assess the *accuracy* of the outcome model.

Detecting Bias Using Confidence Intervals

To determine if the difference between the risk score label and outcome label models is statistically significant, we could report p-values. However, the duality between confidence intervals and hypothesis testing p-values affords a quicker method - a visual inspection of whether the horizontal line at $y = 0$ representing zero difference was within the confidence interval. If this is not the case, the null hypothesis that there is no difference between the two models is rejected.

Results

In this section we describe insights from comparing transparent student models trained to predict black-box risk scores to transparent models trained to predict actual outcomes. We also validate that GAM models are good student models. Due to space constraints, we describe our findings on the Lending Club data in the Appendix.

Detecting Bias in COMPAS

Figure 1 shows shape plots for four features used by COMPAS for recidivism prediction: Age, Race, Number of Priors, and Gender. The top row shows what was learned by the transparent models trained to predict either the COMPAS risk score (red), or true recidivism outcome (green). The transparent model trained to mimic the COMPAS model (red) gives insight into how the COMPAS model works, where as the transparent model trained on the true outcome shows what can be learned from the data itself. 95% confidence intervals are shown for both models. The bottom row of Figure 1 shows the difference between the red and green terms in the top row, along with 95% confidence intervals for this difference that takes into account the covariace between the red and green terms.

COMPAS is biased for some age and race groups. Examining the plots on the left of Figure 1 for Age, we see that the red mimic model and the green true outcome model are very similar for ages 20 to 70: the confidence intervals in the top plot overlap significantly, and the confidence intervals in the difference plot (bottom row) usually include zero. For Age greater than 70 where the number of samples is low (bottom plot), the variance of the true-labels model is large and there is not enough evidence to conclude that the models disagree. However, the difference between the COMPAS mimic model and the true label model is significant for ages 18 and 19: the COMPAS model apparently predicts low risk for very young offenders, but we see no evidence to support this in the model trained on the true labels where risk appears to be highest for young offenders. This suggests an interesting bias favoring young offenders in the COMPAS model that does not appear to be explained by the data.

The next set of graphs in Figure 1 show risk as a function of Race. COMPAS apparently predicts Native Americans

[2] https://github.com/propublica/compas-analysis

[3] https://www.lendingclub.com/info/download-data.action

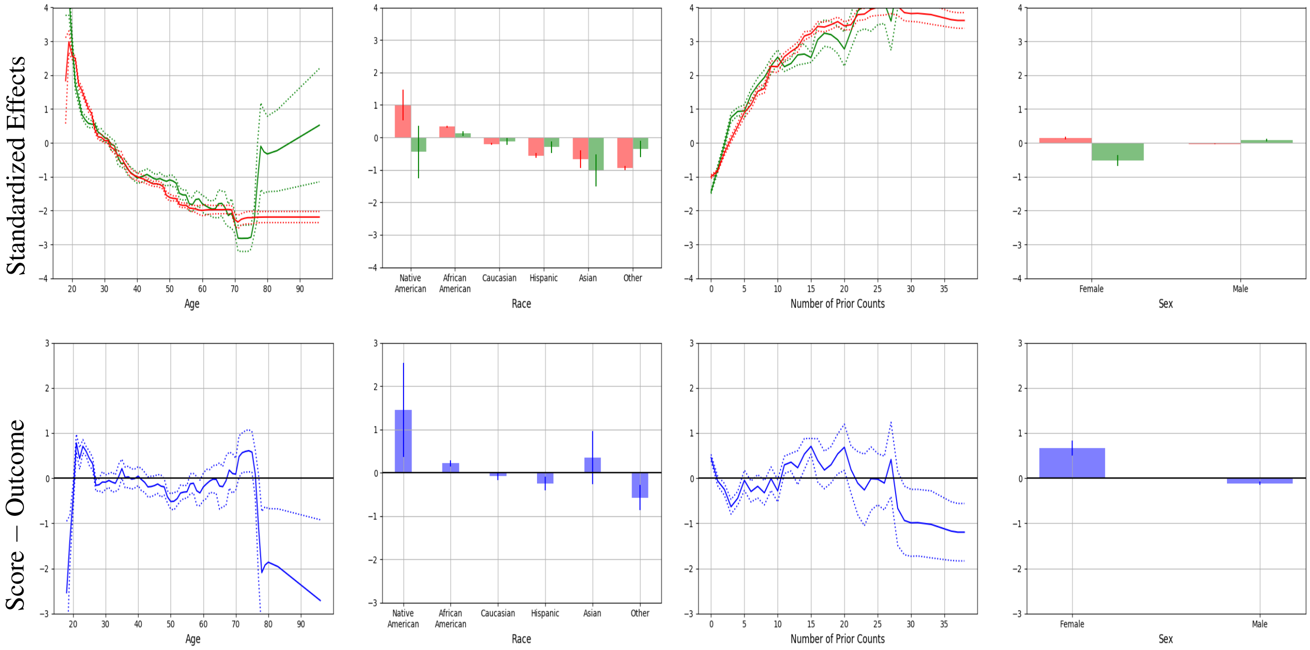

Figure 1: Shape plots for four of six features for recidivism prediction. The remaining features are in the Appendix. Top row: Red lines: effect of feature on COMPAS risk score. Green lines: standardized effect of feature on actual recidivism outcome. Categorical terms ordered in decreasing predicted risk of the score. All plots mean-centered on the vertical axes to allows individual terms to be easily added or subtracted from the model.
Bottom row: Blue lines: difference between score and outcome models (score - outcome).

are a high risk group, despite the fact that the model trained on the true outcomes predicts this group is relatively low risk. The COMPAS mimic model also predicts that African Americans are higher risk, and that Caucasians are lower risk, than the transparent model trained on the true labels suggests is warranted. Apparently the COMPAS model is even more biased against African Americans and towards Caucasians than the (potentially biased) training data warrants.

COMPAS agrees with data on number of priors. In the 3rd column, the COMPAS mimic model and the true-labels model agree on the impact of Number of Priors on risk — the error bars overlap through most of the range and become very wide when the largest difference is observed for more than 30 priors.

Gender has opposite effects on COMPAS compared to true outcome. In the 4th column, we see a discrepancy between what the COMPAS mimic model and the true-labels model learned for Gender. The COMPAS model predicts that Females are higher risk than the data suggests is correct for women, and that males are lower risk than the data suggests is correct for men. We suspect this difference arises because most of the training data is for males (bottom graph), and that this COMPAS model is not as good at distinguishing between male and female as it could be.

Pairwise interactions. Figure 2 shows one of the more significant pairwise interactions the GA2M model for recidivism. Interactions make GA2M models more accurate by allowing them to model effects that can not be represented by a sum of main effects on individual features. The interaction in Figure 2 is between gender and number of prior convictions. The graph on the left is the interaction for the student model trained to mimic COMPAS, and the one on the right is for the model trained on the true outcomes. The two graphs are qualitatively similar, suggesting that COMPAS models this interaction similarly to how the model trained on the raw outcomes models it. Both models are essentially flat for males (the top of each graph). Both graphs, however, increase risk significantly for females with more than 10-20 priors, and reduce risk for females with few prior convictions. We suspect that this interaction would not be necessary if the data consisted only of males or only of females because these effects could then be absorbed into the priors main effect. Because this data is predominantly male, the main effect for number of priors is more correct for males than for females, and the interaction between gender and priors is used to correct the model for females because there apparently is a significant difference between the impact of number of priors on risk for the two genders.

Are GA2Ms Successful Mimic Models: Fidelity?

We compared the accuracy of interpretable GA2Ms in predicting its teacher black-box risk score to that of another class of transparent student models - logistic regression - and a non-transparent student model - random forests. The results are in Table 2. For both risk scores, GA2M has RMSE similar to random forests and lower than logistic regression. None of the methods could go lower than RMSE of 2 on

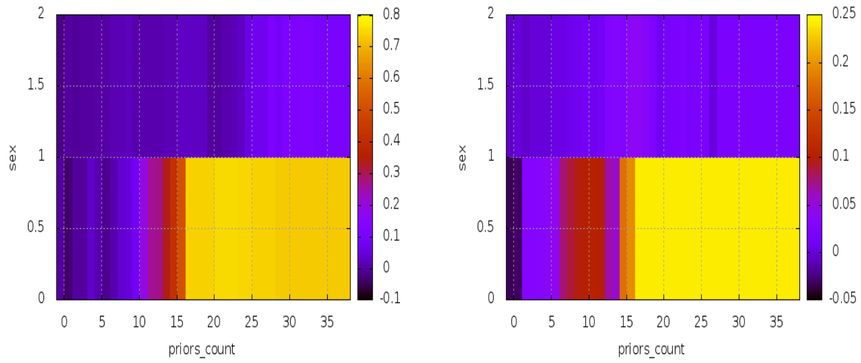

Figure 2: Pairwise interaction between gender and number of priors for recidivism prediction. The heatmap on the left is for the student trained to mimic the COMPAS scores. The heatmap on the right is trained on the actual recidivism outcomes.

a 1-10 scale, likely reasons why the COMPAS risk score is challenging to predict include the lack of essential features from the public data set released, and its smaller number of observations compared to modern data sets.

Are GA2Ms Successful Mimic Models: Transparency?

We compare the GAM estimated effects to that of linear and logistic regression. For categorical features, GA2Ms are equivalent to logistic regression. However, for continuous features, the difference between GA2Ms and logistic regression is significant. Consider Figure in the Appendix, which is the equivalent of the bottom row of Figure 1. Where the GAM model was able to shape the age feature in a non-linear manner across the age range, and detected systematic bias in the young and old ages, logistic regression ascribes only one number as the effect of the age variable on the predicted outcome. This resulted in no systematic difference detected (because the bar graph for the first column, age, is at at y=0 in Figure . This example demonstrates that the use of GA2Ms for bias detection is especially valuable when the data has many continuous features that the GA2M is able to shape in interesting ways.

Discussion

Bias Discovery via Hypothesis Testing

One of the key advantages of using transparent models to understand bias in data, and bias in black-box models trained on that data, is that you do not need to know in advance what biases to look for. Examining the black-box model often shows bias that would not have been anticipated in advance. For example, in addition to the expected bias on race, the COMPAS recidivism model appears also to be biased in favor of very young offenders age 18-19, and against women. Once unexpected biases like these are discovered by examining the transparent model, further testing can be done to study the bias and determine its source. Not having to know what to look for in advance is very useful because there are many kinds of bias in real data, some of which we would not have known to look for or design statistical tests to test for.

Detecting Used and Unused Variables

One way to evaluate if a transparent student model trained to mimic a black-box model can provide insight into the black-box model is to test if the student model can distinguish between variables that are used or not used in the black-box model. The Chicago Police Department dataset contains 16 variables that could be used for prediction, but the description of their model states that only eight of these variables are used in the model.

We trained a transparent student model to mimic their model, and intentionally included all 16 variables in the student model. Figure 4 shows main effects learned by the student model for the eight features the Chicago Police Department used in their model, and Figure 5 in the Appendix shows the main effects learned by the student model for features the Department says were *not* used in their model. As in other figures, red is what the transparent student model learns when trained to mimic the Chicago Police Department model, and green is what the transparent model learns when trained on true outcomes. There is very little red visible in Figure 5: the transparent model trained to mimimc the Police Department model correctly learns to put almost zero weight on the features not used in the Police Department model. But there is significant green in most of the graphs in Figure 4 in the Appendix: transparent models trained on the true outcome find these variables useful to predict risk even if the decision was made not to use them in the Chicago model.

This confirms that the transparent student mimic model is able to properly detect when variables are not used in the black-box teacher model if there is little or no bias on these variables. This is important not only because it shows that transparent student models can be used to tell what features are used in a black-box model, but also because it increases the fidelity of the student model to the black-box model — if the student puts little or no weight on the same features not used in the black-box model, then it is more likely to model

Table 2: Fidelity of student model of teacher risk score (first row of each data set) and accuracy of outcome model (second row of each data set).

Data	Metric	Linear / Logistic Regression	GAM	GA2M	Random Forest
Compas	Risk score (1-10) RMSE	2.09 ± 0.014	2.01 ± 0.031	$\mathbf{1.99 \pm 0.032}$	2.02 ± 0.019
	Actual outcome AUC	0.74 ± 0.007	0.74 ± 0.016	$\mathbf{0.75 \pm 0.015}$	0.73 ± 0.013
Lending Club	Risk score (0-6) RMSE	0.46 ± 0.002	0.25 ± 0.002	0.23 ± 0.003	$\mathbf{0.21 \pm 0.005}$
	Actual outcome AUC	0.68 ± 0.010	0.70 ± 0.015	$\mathbf{0.71 \pm 0.015}$	0.69 ± 0.010

the effects of features that were used in the black-box model correctly.

When comparing what a transparent student learns from the black-box with what a transparent model learns from the raw data, it is valuable to train the transparent model on the raw data two different ways: 1) using all available features to see what could have been learned from the original targets, and 2) using only those features used in the black-box model for direct comparison with a student model trained using only those features to mimic the black-box. Because of space restrictions we do not include both of these in this paper; to avoid confusion, the model presented in green in both Figures 4 and 5 are from the same transparent model trained on all 16 features. In Figure 4 there is strong evidence that both the Chicago black-box model, and a transparent model trained on the true outcome, find the eight features used in the Chicago model useful, but there are significant differences between how the two models use these features because the transparent model trained on the true outcomes (green) uses all 16 features and places significant mass on the 8 unused features, thus altering what is learned by the true outcome model on the other eight features that are used by the Chicago model.

Using Excluded Variables to Detect Bias

Sometimes we are interested in detecting bias on variables that have intentionally been excluded from the black-box model. For example, a model trained for recidivism prediction or credit scoring is probably not allowed to use race as an input to prevent the model from learning to be racially biased. Unfortunately, excluding a variable like race from the inputs does not prevent the model from learning to be biased. Racial bias in a data set is likely to be in the outcomes — the targets used for learning; removing the race *input race variable* does not remove the bias from the *targets*. If race was uncorrelated with all other variables (and combinations of variables) provided to the model as inputs, then removing the race variable would prevent the model from learning to be biased because it would not have any input variables on which to model the bias. Unfortunately, in any large, real-world data set, there is massive correlation among the high-dimensional input variables, and a model trained to predict recidivism or credit risk will learn to be biased from the correlation between other input variables that must remain in the model (e.g., income, education, employment) and the excluded race variable because these other correlated variables enable the model to more accurately predict the (biased) outcome, recidivism or credit risk. Unfortunately, removing a variable like race or gender does not prevent a model from learning to be biased. Instead, removing protected variables like race or gender make it harder to detect how the model is biased because the bias is now spread in a complex way among all of the correlated variables, and also makes correcting the bias more difficult because the bias is now spread in a complex way through the model instead of being localized to the protected race or gender variables. The main benefit of removing a protected variable like race or gender from the input of a machine learning model is that it allows the group deploying the model to claim (incorrectly) that they model is not biased because it did not use the protected variable.

When training a transparent student model to mimic a black-box model, we intentionally include as an input variable any protected variables that the model excluded, but for which we are interested in detecting bias. We are careful to *include* both race and gender as input variables on the transparent student mimic model, and on the transparent model trained to predict the true outcomes, specifically because we are interested in examining what the model learns from these variables. If, when the student model mimics the black-box model, it does not see any signal on the race or gender variable and learns to model them as flat (zero) functions, this indicates whether the teacher model (the black-box model) probably did or did not use these variables, but also if the teacher model exhibits race or gender bias even if the model did not use race or gender as inputs.

Conclusion

We propose a method to detect bias in black-box risk models by using model distillation to train a transparent student model to mimic the black-box model, and then comparing the transparent mimic model to a transparent model trained using the same features on true outcomes instead of the labels predicted by the black-box model. Differences between the transparent mimic and true-labels model indicate differences between how the black-box model makes predictions, and how a model trained on the true outcomes makes predictions, highlighting potential biases in the black-box model. We demonstrate this method on two data sets and uncover a number of interesting differences (potential biases). The key advantages of this approach are that the transparent models are very accurate despite being intelligible, the method generates reliable confidence intervals to aid interpretation, and one does not to know in advance what biases to look for.

References

Adler, P.; Falk, C.; Friedler, S. A.; Rybeck, G.; Scheidegger, C. E.; Smith, B.; and Venkatasubramanian, S. 2016. Auditing black-box models for indirect influence. In *IEEE 16th International Conference on Data Mining*, 1–10.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016a. How we analyzed the compas recidivism algorithm. Accessed May 26, 2017.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016b. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Accessed May 26, 2017.

Athey, S.; Tibshirani, J.; and Wager, S. 2017. Generalized random forests. *arXiv preprint arXiv:1610.01271*.

Blomberg, T.; Bales, W.; Mann, K.; Meldrum, R.; and Nedelec, J. 2010. Validation of the compas risk assessment classification instrument. *College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL*.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. ACM.

Chouldechova, A., and G'Sell, M. 2017. Fairer and more accurate, but for whom? In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230*.

Craven, M. W., and Shavlik, J. W. 1995. Extracting tree-structured representations of trained networks. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, 24–30. MIT Press.

Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. Compas risk scales: Demonstrating accuracy equity and predictive parity. Technical report.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Gibbons, R. D.; Hooker, G.; Finkelman, M. D.; Weiss, D. J.; Pilkonis, P. A.; Frank, E.; Moore, T.; and Kupfer, D. J. 2013. The computerized adaptive diagnostic test for major depressive disorder (cad-mdd): a screening tool for depression. *The Journal of clinical psychiatry* 74(7):1–478.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*.

Lou, Y.; Caruana, R.; Gehrke, J.; and Hooker, G. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 623–631. ACM.

Lou, Y.; Caruana, R.; and Gehrke, J. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150–158. ACM.

Louzada, F.; Ara, A.; and Fernandes, G. B. 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*.

Neukrug, E., and Fawcett, R. 2014. *Essentials of testing and assessment: A practical guide for counselors, social workers, and psychologists*. Nelson Education.

Sam Corbett-Davies, Emma Pierson, A. F., and Goel, S. 2016. Accessed May 26, 2017.

Sexton, J., and Laake, P. 2009. Standard errors for bagged and random forest estimators. *Computational Statistics and Data Analysis* 53(3):801–811.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. ACM.

Zhang, Z., and Neill, D. B. 2017. Identifying significant predictive bias in classifiers. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Appendix

Characterizing the Distribution of Differences

Making Outcomes and Scores Comparable. Additive models built on different labels are measured in different units. In our approach, binary outcomes result in additive effects measured in units of inverse logit probabilities while those based on scores are given by the units of the score. In order to make these comparable, we construct standardized effects by subtracting the average from each effect and scaling by its standard deviation. Specifically, for each effect we compare terms of the form

$$h_j^*(x_j) = \frac{h_j(x_j) - \int h_j(s)ds}{\sqrt{\int \left(h_j(s)\right)^2 ds - \left(\int h_j(s)ds\right)^2}}. \quad (3)$$

between models. These terms are now dimensionless (Neukrug and Fawcett 2014) and centered on zero and can be compared.

Structured Bagging Setup. Given the dataset, we randomly select 15% of the samples to be the test set. The remaining 85% of the data is further split into training (70% of the data) and validation (15%). The random split of this 85% of the data into train and validation is repeated L times (while keeping the test set constant), and the model predictions and feature contributions are averaged. This whole process (selecting a test set and then performing L folds of the remaining data) is repeated K times. Figure illustrates this.

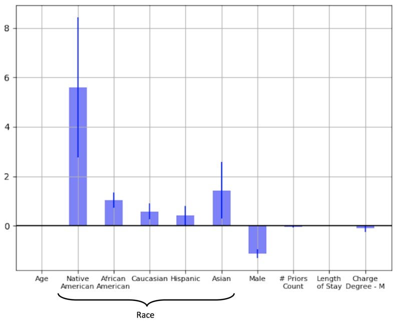

Figure 3: Logistic regression difference between score and outcome models.

We use the same K outer folds and L inner folds to train both models, and compare the models within each inner fold. This removes variation due to training the two models on different splits of the data.

Variance Estimation. Sexton and Laake (Sexton and Laake 2009) proposed a bootstrap-of-little-bags approach to estimate the variance of functions learned from bagged ensembles, and Athey et al. proved the consistency of this variance estimate (Athey, Tibshirani, and Wager 2017). While bagging was originally introduced to this short boosted tree model to reduce the variance of the learned shape functions, in this paper, bagging has additional importance, as we can use the bootstrap-of-little-bags variance estimate to estimate the variance of $h_j(x_j)$: feature x_j's contribution to the prediction.

Using Sexton and Laake's bootstrap-of-little bags approach, given $k = 1, \ldots, K$ outer folds, each of which has $l = 1, \ldots, L$ inner folds, we first take the average of $h_j(x_j)$ across inner folds of the same outer folds, yielding K averages. Then we take the variance of these K averages as our variance estimate for one model. The variance estimate for the difference between two models is then a sum of the individual variances and covariance, also estimated using bootstrap-of-little-bags.

It is important to note that this gives us pointwise confidence intervals, i.e. confidence intervals at specific values of x_j e.g. race=African-American, age=50, etc., which is sufficient for our goal of detecting specific feature values that exhibit systematic differences between the risk score and actual outcome. We leave the construction of uniform confidence intervals for future work.

Investigating Lending Club Loan Risk Scores

Figure 7 shows shape plots for four of the features used in the loan default risk prediction model: Annual Income, FICO Score, Interest Rate, and Loan Purpose. As in Figure 1, red lines show what was learned by the transparent student model trained to mimic the lending model by training on scores predicted by that model, and green lines show what a transparent model learned when training on the true credit fault labels. Comparing the red and green lines helps us understand what the black-box lending model learned, and how it differs from what a model could have learned from the true labels.

The black-box lending model probably ignores income and loan purpose. Interestingly, the black-box lending model appears to ignore Self-Reported Income (red), even though a model trained on the true labels shows a strong effect for income on risk (green). We suspect the lending model may ignore income because it is self-reported and thus easy to fake.

In the last plot for the Loan Purpose term, we see that the lending model probably also ignores the loan purpose (red), but a model trained on the true labels (green) suggests that purpose is a useful feature and that risk is highest for small business loans, and least for loans taken out for weddings. For the lending model we see a number of graphs like this where the mimic model is essentially flat on a feature that the model trained on the true labels finds useful, and suspect that this is one way of determining which features are or are not used by a black-box model.

The models agree and disagree on FICO. As for FICO score, the models agree for low to intermediate FICO scores, but disagree for scores around 800. A transparent model trained on the true labels predicts that FICO score near 800 indicate very low risk, but the mimic model suggests that there is little difference in risk between scores of 750 and 800. The black-box lending model appears to be a smooth, simple function of FICO scores above 675 (and possibly below 675, but the error bars are too large to be conclusive).

The models are qualitatively similar but quantitatively different on interest rate. On interest rate, the mimic model (red) suggests that the black-box lending model places significantly more emphasis on interest rate than a model trained on the true labels. Both models show a strong, linear increase in risk with interest rate, but the slope is twice as high on the mimic model.

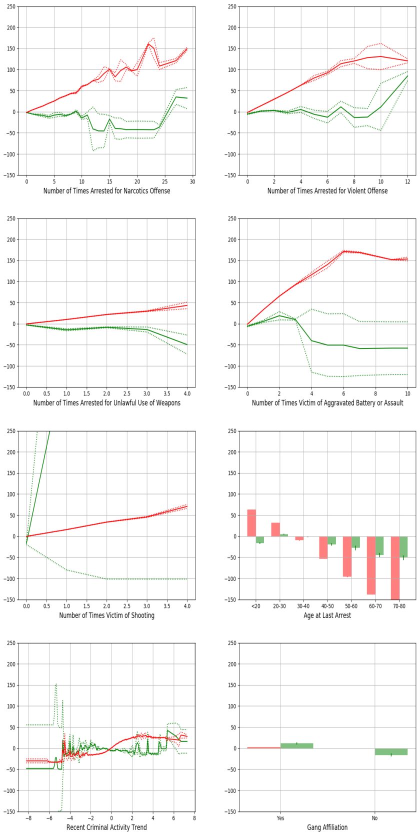

Figure 4: Shape plots for eight features the Chicago Police Department claims are used to predict the risk score

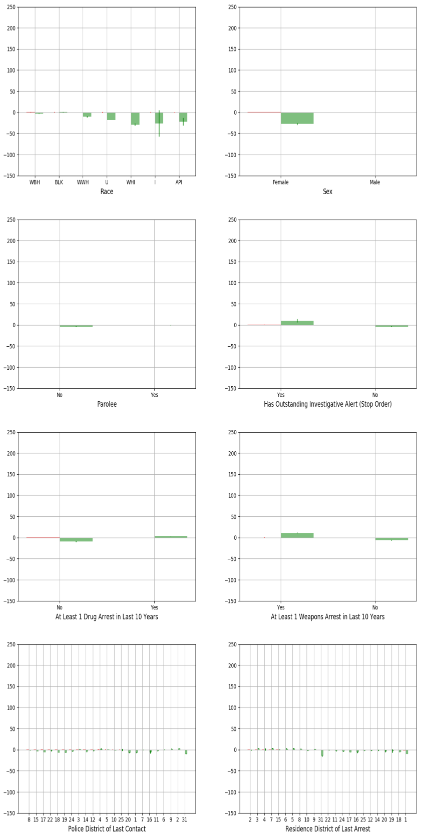

Figure 5: Shape plots for eight features the Chicago PoliceDepartment claims are *not* used to predict the risk score

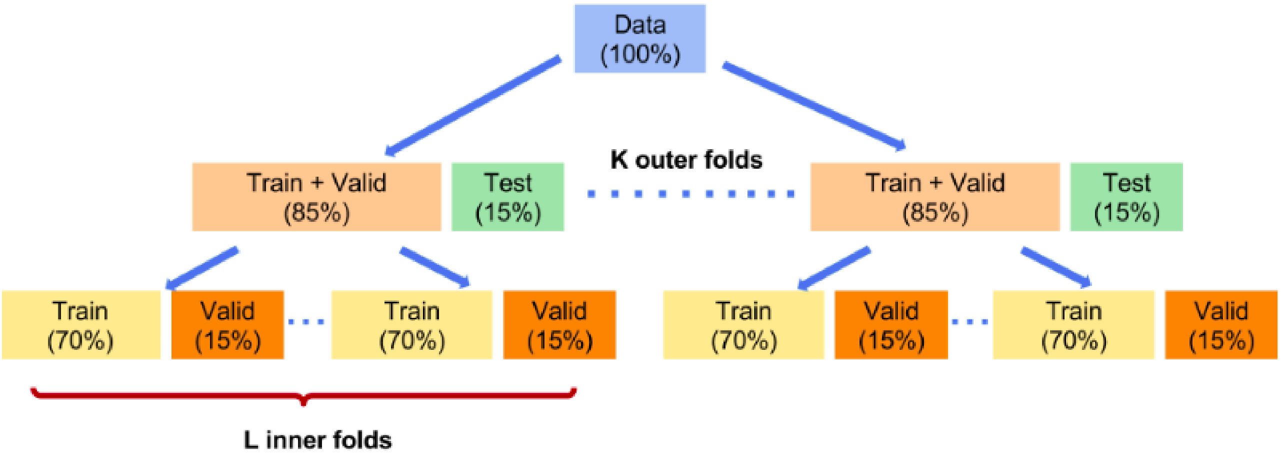

Figure 6: Structured Bagging Setup

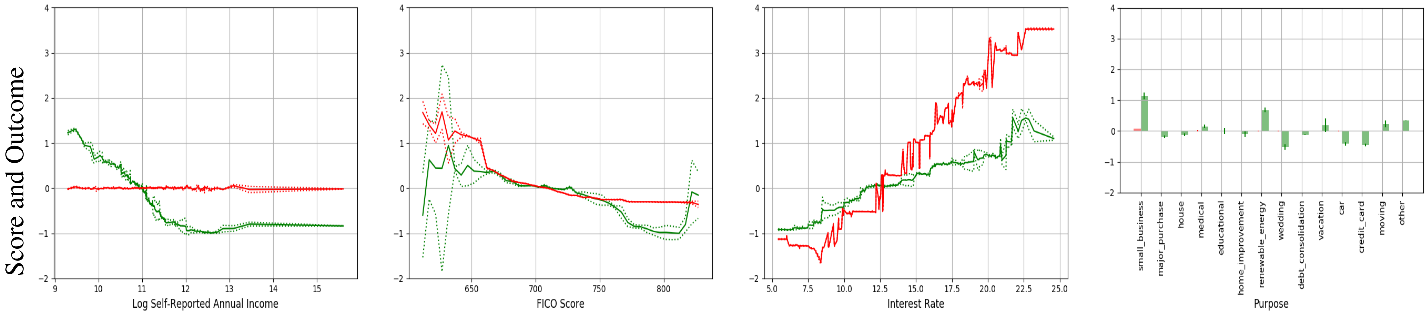

Figure 7: Shape plots for four features for loan default prediction

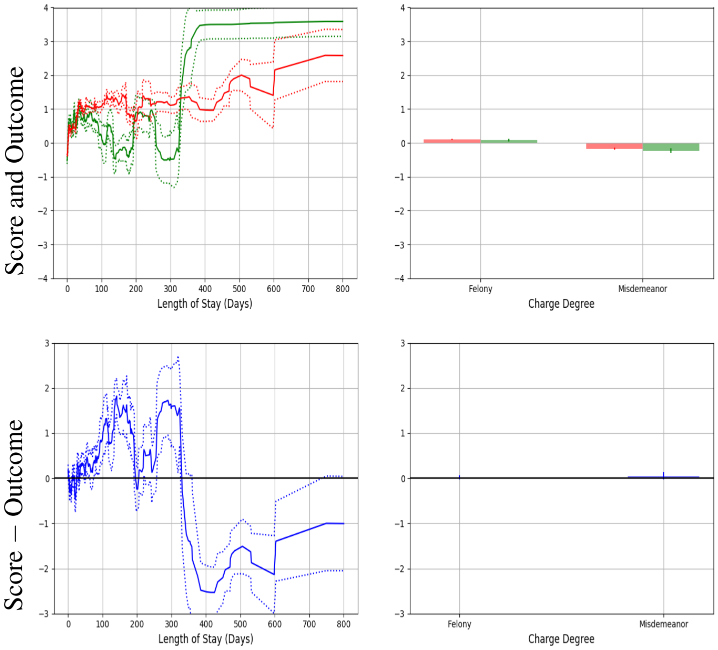

Figure 8: Additional features in recidivism risk data. See caption in Figure 1.