Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation

Sarah Tan
Cornell University
ht395@cornell.edu

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Giles Hooker
Cornell University
gjh27@cornell.edu

Yin Lou
Ant Financial
yin.lou@antfin.com

ABSTRACT

Black-box risk scoring models permeate our lives, yet are typically proprietary or opaque. We propose Distill-and-Compare, an approach to audit such models without probing the black-box model API or pre-defining features to audit. To gain insight into black-box models, we treat them as teachers, training transparent student models to mimic the risk scores assigned by the black-box models. We compare the mimic model trained with distillation to a second, un-distilled transparent model trained on ground-truth outcomes, and use differences between the two models to gain insight into the black-box model. We demonstrate the approach on four data sets: COMPAS, Stop-and-Frisk, Chicago Police, and Lending Club. We also propose a statistical test to determine if a data set is missing key features used to train the black-box model. Our test finds that the ProPublica data is likely missing key feature(s) used in COMPAS.

CCS CONCEPTS

• **Computing methodologies** → **Model verification and validation**; • **Mathematics of computing** → *Hypothesis testing and confidence interval computation*;

KEYWORDS

Interpretability; Black-box models; Distillation; Fairness

ACM Reference Format:
Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, 2018, New Orleans, LA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3278721.3278725

1 INTRODUCTION

Risk scoring models have a long history of usage in criminal justice, finance, hiring, and other critical domains [13, 29]. They are designed to predict a future outcome, for example defaulting on a loan. Worryingly, risk scoring models are increasingly used for high-stakes decisions, yet are typically proprietary or opaque.

Several approaches have been proposed [1, 2, 14, 18, 21, 36] to audit black-box risk scoring models: remove, permute, or obscure a protected feature, then see how the the model's predictions change after retraining the model or probing the model API with the transformed data. However, creators of proprietary risk scoring models often do not provide unrestricted access to model APIs, much less release the model form or training data. Moreover, approaches that focus on one or two protected features defined in advance are less likely to detect biases that are not *a priori* known.

In this paper, we study a more realistic setting where we only have a data set labeled with the risk score (as produced by the risk scoring model), the ground-truth outcome, and some or all features; we are not able to probe the model API with new data. We call this data set the *audit data*. We add two potential complications: the audit data may not be the original training data, and the audit data may not have all features used to train the risk scoring model. For example, ProPublica obtained data for their COMPAS study [5] not from the company that created COMPAS, but through a public records request to Broward County (BC), a US jurisdiction that used COMPAS in their criminal justice system [4]. ProPublica may not have had access to all the features BC used for COMPAS.

We propose Distill-and-Compare, an approach to audit black-box risk scoring models using audit data with both black-box risk scores and ground-truth outcomes, without pre-defining feature regions to audit. First, we train a model on the audit data to mimic the black-box model. Then we train another model to predict outcomes (Section 2.1). To gain insight into the black-box model, we uncover feature regions where the two models are significantly different (Section 2.3), and ask "what could be happening in the black-box model, that could explain the differences we are seeing between the mimic and outcome models?". Finally, we use a statistical test (Section 2.2) to determine if the black-box model used additional features we do not have access to (i.e. features not in the audit data).

The contributions of this paper are: 1) We propose an approach to audit black-box risk scoring models under realistic conditions. 2) We show the importance of calibrating risk scores to remove audit data shift or scale post-processing that may been introduced by creators of risk scoring models. 3) We propose a statistical test to determine if the audit data is missing key features used to train the black-box model. 4) We apply the approach to audit four risk scoring models. 5) An ancillary contribution of this paper is a new confidence interval estimate for iGAM [10, 27, 28], a type of transparent model.

2 AUDIT APPROACH

Our goal is to gain insight into a black-box risk scoring model. We draw from model distillation and comparison technique to develop our approach. Section 2.1.1 discusses related work.

2.1 Distill and Compare

Model distillation was first introduced to transfer knowledge from a large, complex model (teacher) to a faster, simpler model (student)

AIES '18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02…$15.00
https://doi.org/10.1145/3278721.3278725

[7, 9, 22]. This was done by running unlabeled samples (either new unlabeled data or training data with labels discarded) through the teacher model to obtain the teacher's outputs, then training the student model to mimic the teacher's outputs. We draw parallels to our setting, taking the risk scoring model to be the teacher and the audit data to be unlabeled samples ran through the teacher (risk scoring model) to obtain the teacher's output (risk scores). We train the mimic model to minimize mean squared error between the teacher and student, i.e.,

$$L(S, \hat{S}) = \frac{1}{T} \sum_{t=1}^{T} \left(S(x^t) - \hat{S}(x^t) \right)^2, \quad (1)$$

where x^t is the t-th sample in the audit data, $S(x^t)$ is the output of the teacher model (risk scores) for sample x^t, $\hat{S}(x^t)$ is the output of the mimic model for sample x^t, and T is the number of samples. Throughout this paper, we will call the teacher model the *black-box model* and the student model the *mimic model*.

Next, we leverage the ground-truth outcome information. We train *our own risk scoring model* on the audit data to predict the ground-truth outcome, i.e.,

$$L(O, \hat{O}) = \frac{1}{T} \sum_{t=1}^{T} \Big\{ O(x^t) \log \left(P(\hat{O}(x^t) = 1) \right) + (1 - O(x^t)) \log \left(P(\hat{O}(x^t) = 0) \right) \Big\}, \quad (2)$$

where $O(x^t) \in \{0, 1\}$ is the ground-truth outcome for sample x^t and $\hat{O}(x^t) \in \{0, 1\}$ is the output of the model for sample x^t. Throughout this paper, we call this model the *outcome model*. Note that the outcome model is not a mimic model.

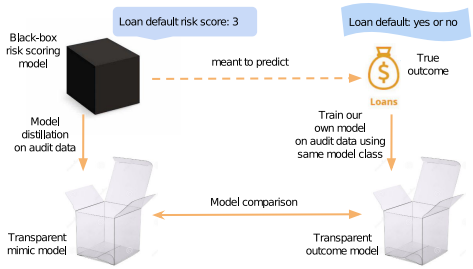

Figure 1: Distill-and-Compare audit approach on a loan risk scoring model.

It is critical that both the mimic model and outcome model are trained using the same model class that allows for interpretation and comparison. Not all model classes have the property that two models of that class can be compared. For example, it is not obvious how to compare two decision trees, random forests or neural nets. We want a model class that is as rich and complex as possible so that the mimic model can be faithful to the black-box model and the outcome model can accurately predict ground-truth outcomes. However, this model class should still be transparent [17] so that we can examine its predictions across different feature regions. In this paper, we use a particular transparent model class, iGAM (Section 2.3.1); other choices are possible.

The risk score and the ground-truth outcome are closely related—the ground-truth outcome is what the black-box risk scoring model was meant to predict. If the black-box model is accurate *and* generalizes to the audit data, it would predict the ground-truth outcomes in the audit data correctly; the converse is true if the black-box model is not accurate *or* does not generalize to the audit data.

Because both the mimic and outcome models are trained with the same model class on the same audit data using the same features, the more faithful the mimic model, and the more accurate the outcome model, the more likely it is that observed differences between the mimic and outcome models stem from differences between the black-box model and ground-truth outcomes. This allows us to ask, "what could be happening in the black-box model, that could explain the differences we are seeing between the mimic and outcome models?". In addition, similarities between the mimic and outcome models (e.g., on COMPAS in Section 3.2, the Number of Priors feature is modeled very similarly by the two models) increases confidence that the mimic model is a faithful representation of the black-box model, and that any differences observed on other features are meaningful.

2.1.1 Related Work. Several auditing approaches also use model distillation techniques to distill black-box models when they cannot be queried or to understand them [1, 2]. Other approaches also train their own outcome models, then uncover feature regions where the model is not accurate [3, 23, 24, 38]. Kim et al.'s iterative procedure [24] not only uncovers such regions but also modifies the model to improve accuracy in these regions. However, they require repeated calls to the model; Agarwal et al. [3] and Kearns et al. [23] similarly require repeated calls or knowledge of the model. Tramer et al. uncovered unexplained associations between black-box outputs and protected features on audit data [35].

Our approach is different from the above, as we avoid repeated calls to the black-box model API (that may not realistically be available), and instead utilize information on both risk scores and outcomes already available in some data sets in this domain (e.g. ProPublica COMPAS data). Some other approaches also compare two models, but not risk scores and outcomes at the same time. Wang et al. trained a model to predict outcomes and another to predict membership in a protected feature region [37]. Chouldechova and G'Sell trained two different outcome models then identified feature regions where the two models differed [12].

2.2 Testing for Missing Features

If the audit data is missing features used by the black-box model, the audit data alone may be insufficient to audit the black-box model. We propose a statistical test to check the likelihood of the audit data missing important features based on the following observation:

> *If the black-box model used features that are missing from the audit data but are useful for predicting the ground-truth outcome, the error between the mimic model (learned on the audit data) and the risk score, $||\hat{S} - S||_E$, should be positively correlated with the error between the outcome model (learned on the audit data) and ground-truth outcome, $||\hat{O} - O||_E$.*

where E is an error metric. Since the test uses predictions from both the mimic and outcome models, the test is performed after both models are trained. In Section 3.4, we perform the test on all risk scoring models we audit in this paper, to check if the audit results

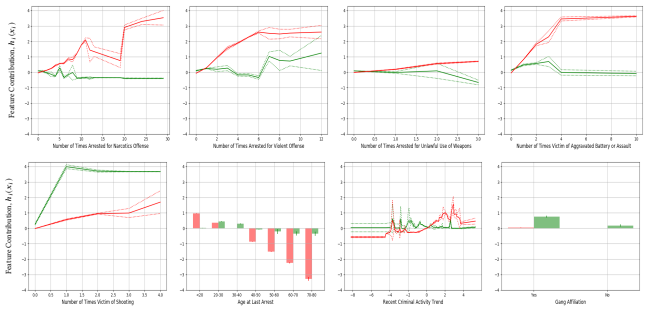

Figure 2: Eight features the Chicago Police says are used in their risk scoring model. Best seen on screen.

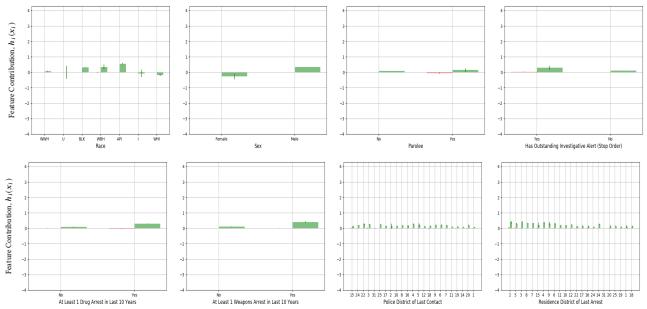

Figure 3: Eight features the Chicago Police says are *not used* in their risk scoring model. Best seen on screen.

are significantly affected by missing features. Note that this test does not require the mimic and outcome models to be transparent.

2.3 Comparing Mimic and Outcome Models

In this section, we provide technical details on how to train the mimic and outcome models so that they are comparable.

2.3.1 Choice of model class. As noted in Section 2.1, we train the mimic model and outcome model using the same transparent model class—in this paper, iGAM [10, 27, 28]. We point the reader to [10, 27, 28] to learn more about iGAMs and to [34] for a distillation example where it was used as a student. Briefly, iGAM has the form

$$E[g(y)] = h_0 + \sum_i h_i(x_i) + \sum_{i \neq j} h_{ij}(x_i, x_j), \tag{3}$$

where g is the logistic function for classification and identity function for regression, h_0 is the intercept, and the contribution of any one feature x_i or pair of features x_i and x_j to the prediction can be visualized in graphs such as Figure 2 (with $h_i(x_i)$ on the y-axis) and Figure 5 (with regions colored by $h_{ij}(x_i, x_j)$). For classical GAMs [20], $h(\cdot)$ are fitted using splines; for iGAM, they are fitted using ensembles of shallow trees and centered for identifiability. Crucially, since iGAM is an additive model, two iGAM models can be compared by simply taking a difference of their feature contributions $h(\cdot)$, which we exploit in Section 2.3.3 to detect differences between the mimic and outcome models.

2.3.2 Calibrating model inputs. Calibration is the process of matching predicted and empirical probabilities [15, 31]. If a risk score is well-calibrated, the relationship between the risk score and empirical probabilities will be linear (e.g., COMPAS and Stop-and-Frisk in the top row of Figure 6 in the Appendix). While developing the method, we discovered that not all risk scores exhibit the desired linear relationship with outcomes in the audit data. For example, the Chicago Police risk score (third column of Figure 6 in the Appendix) is rather flat for risk scores less than 350, then exhibits a sharp kink upwards.

One possible explanation for any nonlinear relationship is that the risk score was well-calibrated on its original training data, but the audit data has a different distribution (data shift) [32]. Another possible explanation is post-processing by model creators to reduce sensitivity in less important parts of the risk score scale and enhance separation in more important parts of the scale [26].

We make the reasonable assumption that risk scores should be monotonic and well-calibrated [26] and use this assumption to undo scale post-processing or audit data shift before training the mimic and outcome models. Specifically, we learn a nonlinear transformation of the risk score (the blue line in Figure 6 in the Appendix), similar to isotonic regression [31], to make the risk scores and outcomes linearly related on a scale of choice. The mimic model is then trained with the transformed risk scores as labels; the outcome model is trained with outcomes, unchanged.

This pre-training calibration step is necessary to compare the mimic and outcome models, as it makes their labels linearly related on a scale that their predicted labels will later be compared on. We select this scale to be logit probability (since the predicted outcomes are already on this scale), and perform this calibration step for Chicago Police and Lending Club but not COMPAS and Stop-and-Frisk, since the latter two already exhibit the desired linear relationships. See Appendix B for details.

2.3.3 Detecting differences. To not mistake random noise for real differences between the mimic and outcome models, we control potential sources of noise during the training process. To avoid data sample-specific effects, we train the mimic and outcome models on the same data sample. Let $sh_i(x_i)$ be feature x_i's contribution to the mimic model, and similarly $oh_i(x_i)$ for the outcome model. We calculate the difference in feature x_i's contribution to the two models, $sh_i(x_i) - oh_i(x_i)$, and construct a confidence interval for this difference to tell if it is statistically significant. One ancillary contribution of this paper is a new method to estimate confidence intervals for the iGAM model class, by employing a *bootstrap-of-little-bags* approach [33] to estimate the variance of $h_i(x_i)$ and $sh_i(x_i) - oh_i(x_i)$. See Appendix A for details. The resulting confidence intervals are the dotted lines in Figures 2–4.

3 RESULTS

3.1 Validating the Audit Approach

In this section, we demonstrate Distill-and-Compare on risk scoring models where we have some information on how they were trained, and check that the approach can recover this information.

3.1.1 Stop-and-Frisk. Using the New York Police Department's Stop-and-Frisk[1] data, Goel et al. [19] proposed a simple risk scoring

[1] http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

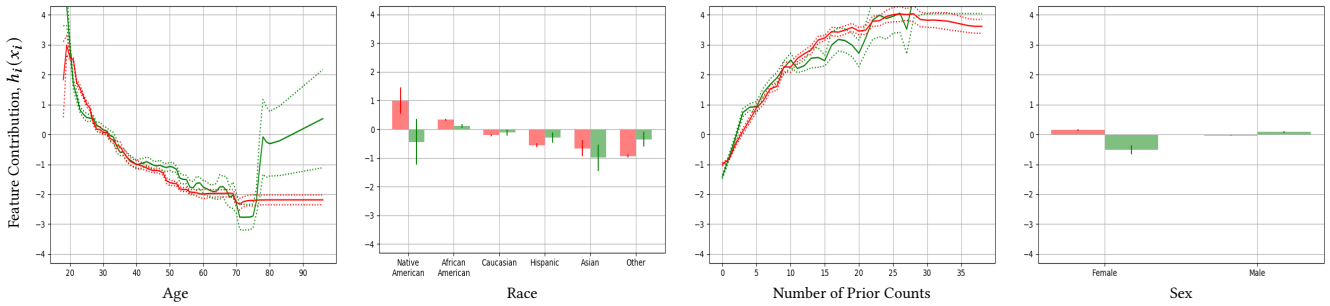

Figure 4: Feature contributions of four features to the COMPAS mimic model (in red) and outcome model (in green).

model for weapon possession: $S = 3 \times \mathbf{1}_{PS} + 1 \times \mathbf{1}_{AS} + 1 \times \mathbf{1}_{Bulge}$, where S is the risk score, *PS* denotes primary stop circumstance being presence of suspicious object, *AS* denotes secondary stop circumstance being sight of criminal activity, and *Bulge* denotes bulge in clothing [19]. Since we know the risk scoring model's functional form, we can verify that the mimic model correctly recovers these coefficients. We apply the risk scoring model to label 2012 data (T=126,457, p=40) after following Goel et al.'s data pre-processing steps [19].

Result. The mimic model recovers the coefficients (3, 1, 1) for the three features used in the risk scoring model (*PS*, *AS*, *Bulge*) and 0 for the remaining features.

3.1.2 Chicago Police "Strategic Subject" List. The Chicago Police Department released arrest data[2] from 2012 to 2016 that was used to create a risk score for an individual being involved in a shooting incident as a victim or offender. This data set contains 16 features, but only 8 are used by the black-box model, which gives us an opportunity to test if Distill-and-Compare can accurately detect which features are and are not used by a black-box model.

We trained a mimic model, intentionally including all 16 features. Figure 2 shows the feature contributions of the mimic model (in red) and outcome model (in green) for the 8 features the Chicago Police says were used by the black-box model; Figure 3 shows the 8 features the Chicago Police says were *not used* in their model.

Result. There is a striking difference between Figures 2 and 3: the mimic model (in red) assigns importance to the features in Figure 2, but does not assign any importance to the features in Figure 3. This agrees with Chicago Police's statement about which features were and were not used in the black-box model. We also note that the outcome model (in green) does assign importance to the unused features (Figure 3), suggesting that there is signal available in the 8 unused features that the Chicago Police risk scoring model could have used, but chose not to use. Race and sex are 2 of these 8 features, which the Chicago Police especially emphasized are not used. These experiments show that mimic models can provide insights into black-box models, and demonstrate the advantages of using outcome information.

3.2 Auditing COMPAS

COMPAS, a proprietary score developed to predict recidivism risk, has been the subject of scrutiny for racial bias [5, 8, 11, 13, 16, 25].

[2]https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np

We do not know what model class, input features or data were used to train COMPAS. As described in Section 1, the COMPAS audit data[3] was collected by ProPublica; it is likely different from the original COMPAS training data. Figure 4 compares the COMPAS mimic model (in red) and outcome model (in green) for four features: Age, Race, Number of Priors, and Gender. The dotted lines are 95% pointwise confidence intervals. We observe the following:

COMPAS agrees with ground-truth outcomes regarding the number of priors. In the 3rd plot in Figure 4, the mimic model and outcome model agree on the impact of Number of Priors on risk; their confidence intervals overlap through most of its range.

COMPAS disagrees with ground-truth outcomes for some age and race groups. The 1st and 2nd plots in Figure 4 show the effect of Age and Race on the mimic and outcome models. The mimic model (red) and the outcome model (green) are very similar between ages 20 to 70, and their confidence intervals overlap. For ages greater than 70, there is evidence that the models disagree as the confidence intervals do not overlap.

The mimic and outcome models are also different for ages 18 and 19: the mimic model predicts low risk for young individuals, but we see no evidence to support this in the outcome model, with risk appearing to be highest for young individuals.

The mimic model predicts that African Americans are even higher risk, and Caucasians lower risk, than the ground-truth outcomes suggest is warranted. Note that the ground-truth outcomes might themselves be biased due to historical discrimination against African Americans.

Gender has opposite effects on COMPAS compared to ground-truth outcome. In the 4th plot in Figure 4, we see a discrepancy between the mimic model and outcome model on Gender. The mimic model predicts higher risk than warranted by ground-truth outcomes for females, and conversely for males.

Using differences to gain insight into COMPAS. We now ask "what could be happening in COMPAS, that could explain the differences we are seeing between the mimic and outcome models?":

(1) Some feature regions may be underrepresented in the black-box model's training data and/or the audit data. In this audit data, only 3% of samples are between 18 and 20 years old, only 0.5% are older than 70 years old, and only 19% are female, which makes learning accurate models in these regions harder.

[3]https://github.com/propublica/COMPAS-analysis

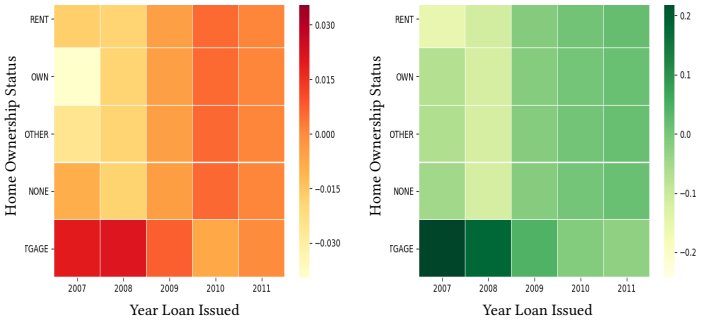

Figure 5: Interaction between loan issue year and home ownership in Lending Club mimic model (in red) and outcome model (in green). Regions colored by $h_{ij}(x_i, x_j)$.

(2) The black-box model may be deliberately simple for some feature regions. For ages greater than 70, the outcome model has much wider confidence intervals than the mimic model. The ground-truth outcomes are potentially high-variance in this region, yet the black-box model's scoring function may have been kept deliberately simple for extreme feature values like this.
(3) The black-box model may have a very different form than the transparent model class. The mimic model predicts low risk for young individuals, but there is no evidence to support this in the outcome model. We trained an iGAM model with interactions between pairs of features, and observed strong interactions between very young age and other variables such as Gender, Charge Degree, and Length of Stay. If COMPAS has a more simple form and does not model interactions well, this may explain why COMPAS needs to predict low risk for very young individuals (because it cannot otherwise predict a reduced risk via interactions of age with other variables).
(4) The black-box model may have used features missing from the audit data, that interact with the non-missing features. We investigate this in Section 3.4.

While we cannot tell (without further investigation) the definitive reason that explains a particular difference between the mimic and outcome models, this has surfaced ideas about the black-box model and uncovered potentially problematic feature regions that we did not *a priori* know, but can now proceed to investigate further.

3.3 Auditing Lending Club

Lending Club, an online peer-to-peer lending company, rates loans it finances on an A1-G5 scale. We use a subset of five years (2007-2011) of loans[4] that have matured, so that we have ground-truth on whether the loan defaulted. We do not know what model class or input features Lending Club used to train their risk scoring model. We believe the data sample we have is similar to the data they would have used to train their models. According to Lending Club, their models are refreshed periodically.

We use this Lending Club example to discuss an insight gained into the black-box model from inspecting feature interactions in the transparent models. Figure 5 shows the interaction of loan issue year and home ownership in the Lending Club mimic model (in red) and ground-truth outcome model (in green). Having a home mortgage in 2007-2008 increases the loan default risk more than having a home mortgage in 2009 and beyond. Recall that 2007-2008 is around the time of the subprime housing crisis. Note the difference in ranges between the two plots—the range goes up to 0.2 for the outcome model (in green) whereas the range is much lower for the mimic model (in red). One possible explanation for this difference is that the Lending Club risk scoring model is updated conservatively (with some lag time), instead of being rapidly updated as economic conditions and behavior change.

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

Table 1: Statistical test for likelihood of audit data missing key features used by black-box model.

Risk Score	Pearson ρ	Spearman ρ	Kendall τ
COMPAS	[0.10, 0.13]	[0.10, 0.14]	[0.08, 0.10]
Lending Club	[0.00, 0.03]	[-0.01, 0.01]	[-0.01, 0.01]
Stop-and-Frisk	[0.00, 0.01]	[-0.03, 0.01]	[-0.02, 0.01]
Chicago Police	[0.00, 0.01]	[0.01, 0.03]	[0.01, 0.02]

3.4 Which Audit Data Are Missing Features?

As black-box models may use additional features we do not have access to, we developed a test (Section 2.2) to assess the impact missing features could have on the audit. Table 1 provides 95% confidence intervals for three correlation measures (linear and nonlinear) used in the test. If zero is in the confidence interval, the error of the mimic model (trained on the audit data) is not correlated with the error of the outcome model (also trained on the audit data). Then, it is unlikely that the audit data is missing key feature(s) that are a) predictive of outcomes (and hence will negatively affect the error of the outcome model if missing); and b) used in the black-box model (and hence will negatively affect the error of the mimic model if missing).

In Lending Club and Stop-and-Frisk we cannot distinguish these correlations from zero, suggesting that no key features are missing from the audit data. For Chicago Police, the confidence intervals contain 0 or are very close to 0 (lower limit 0.01), hence there is little evidence of missing key features. For COMPAS, there is evidence of positive correlation, indicating that the ProPublica data may be missing key features used in the COMPAS model. This is supported by the findings in Section 3.5 that no mimic models trained on the ProPublica data, however powerful (e.g., random forests), could mimic COMPAS well.

3.5 Fidelity and Accuracy

To quantitatively evaluate the audit approach, we report fidelity (how well the mimic model predicts the black-box model's risk scores, measured in RMSE) and accuracy (how well the outcome model predicts the ground-truth outcomes, measured in AUC) for all the risk scoring models we audit in Table 2. For comparison, we also train linear models (a simpler model class than iGAM) and random forests (more complex, but less interpretable).

For COMPAS, all model classes (linear model, iGAM, random forest) have roughly equal fidelity and accuracy. Interestingly, none obtained RMSE lower than 2 on a 1-10 scale. Comparing outcome

Table 2: Fidelity of mimic model and accuracy of outcome model. Lower RMSE is better, higher AUC is better.

	Risk Score	Metric	Linear model	iGAM	iGAM w/ interactions	Random Forest
Fidelity of mimic model	COMPAS	RMSE (1-10)	2.11 ± 0.057	2.01 ± 0.045	**2.00 ± 0.047**	2.02 ± 0.053
	Lending Club	RMSE (2-36)	3.27 ± 0.037	2.60 ± 0.049	2.52 ± 0.051	**2.48 ± 0.033**
	Chicago Police	RMSE (0-500)	17.4 ± 0.102	17.2 ± 0.125	16.5 ± 0.130	**14.0 ± 0.280**
	Stop-and-Frisk	RMSE (0-5)	$\mathbf{0.00 \pm 2 \times 10^{-15}}$	$\mathbf{0.00 \pm 1 \times 10^{-5}}$	$\mathbf{0.00 \pm 2 \times 10^{-5}}$	$0.01 \pm 2 \times 10^{-3}$
Accuracy of outcome model	COMPAS	AUC	0.73 ± 0.029	0.74 ± 0.027	**0.75 ± 0.029**	0.73 ± 0.026
	Lending Club	AUC	0.69 ± 0.006	**0.69 ± 0.016**	**0.69 ± 0.014**	0.68 ± 0.020
	Chicago Police	AUC	**0.95 ± 0.007**	**0.95 ± 0.007**	**0.95 ± 0.007**	0.93 ± 0.009
	Stop-and-Frisk	AUC	0.84 ± 0.020	0.85 ± 0.020	0.85 ± 0.020	**0.87 ± 0.024**

model AUCs across different model classes, iGAM's results are generally comparable to (or slightly better than) more complex random forests (Table 2). For the risk score mimic models, random forests are competitive for Lending Club and Chicago Police. Linear mimic models are not far behind iGAMs for several risk scoring models (COMPAS, Chicago Police, Stop-and-Frisk), suggesting that the black-box model's functional form might be very simple. We know this to be true for Stop-and-Frisk from Section 3.1.1 where the model was a simple linear model.

3.6 Using Additional Data for Distillation

One possible reason why COMPAS is challenging to mimic may be that the ProPublica data is missing key features. This agrees with the results of the statistical test in Section 3.4. Another possible reason is the small sample size (less than 7,000 samples).

One advantage of using a model distillation approach to inspect black-box models is that the approach may be able to benefit from additional unlabeled data if the black-box model can be queried to label the additional data [9]. We found an additional 3,000 individuals in the ProPublica data with COMPAS risk scores *but no ground-truth outcomes*. Adding them to the training (not testing) data for the mimic model and retraining the mimic model, we find marginal improvement in the mimic model's fidelity (from RMSE 2.0 to 1.98). Doing the opposite—removing individuals from the training data in 1,000 increments—decreased the mimic model's fidelity only marginally (to RMSE 2.1, training on only 1,000 individuals). These analyses suggest that for COMPAS, missing key features is a more pressing issue than insufficient data.

4 DISCUSSION

Sometimes we are interested in detecting bias on features intentionally excluded from the black-box model. For example, a credit risk scoring model is probably not allowed to use race as an input. Unfortunately, not using race does not prevent the model from learning to be biased. Racial bias in a data set is likely to be in the outcomes — the labels used for learning; not using race as an input *feature* does not remove the bias from the *labels*.

If race were uncorrelated with all other features (and combinations of features) provided to the model, then removing race would prevent the model from learning to be racially biased because it would not have any input features on which to model this bias. Unfortunately, in any real-world, high-dimensional data set, there is massive correlation among the features, and a model trained to predict credit risk will learn to be biased from correlation of the *excluded* race feature with other features that likely remain in the model (e.g., income or education).

Hence, removing a protected feature like race or gender does not prevent a model from learning to be biased. Instead, removing protected features make it harder to detect how the model is biased, or correct the bias, because the bias is now spread in a complex way among all the correlated features throughout the model instead of being localized to the protected features. The main benefit of excluding protected features like race or gender from the inputs of a machine learning model is that it allows the group deploying the model to claim (incorrectly) that their model is not biased because it did not use these features.

When training a transparent model to mimic a black-box model, we intentionally include all features—even protected features like race and gender—specifically because we are interested in seeing what the mimic model *could* learn from them. If, when the mimic model mimics the black-box model, it does not see any signal on the race or gender features and learns to model them as flat (zero) functions, this suggests whether the black-box model did or did not use these features, but also if the black-box model exhibits race or gender bias even if race or gender were not used as inputs.

5 CONCLUSION

The Distill-and-Compare approach to auditing black-box models was motivated by a realistic setting where access to the black-box model API is not available. Instead, only a data set labeled with the risk score (as produced by the risk scoring model) and the ground-truth outcome is available. The efficacy of Distill-and-Compare increases when a model class that can be highly faithful to the black-box model and highly accurate at predicting the ground-truth outcomes is used, and when the audit data is not missing key features used in the black-box model.

A key advantage of using transparent models to audit black-box models is that we do not need to know in advance what to look for. Many current auditing approaches focus on one or two protected features defined in advance, and thus are less likely to detect biases that are not *a priori* known. The Distill-and-Compare audit approach using transparent models can hence be most useful for real-world, high-dimensional data with multiple, unknown sources of bias.

REFERENCES

[1] Julius Adebayo and Lalana Kagal. 2016. Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models. In *FATML Workshop*.

[2] Philip Adler, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Eduardo Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing Black-Box Models for Indirect Influence. In *ICDM*.

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *ICML*.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm Accessed May 26, 2017.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed May 26, 2017.

[6] Susan Athey, Julie Tibshirani, and Stefan Wager. 2017. Generalized Random Forests. *arXiv preprint arXiv:1610.01271* (2017).

[7] Jimmy Ba and Rich Caruana. 2014. Do Deep Nets Really Need to be Deep?. In *NIPS*.

[8] Thomas Blomberg, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. 2010. *Validation of the COMPAS risk assessment classification instrument.* Technical Report. Florida State University.

[9] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *KDD*.

[10] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*.

[11] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* (2017).

[12] Alexandra Chouldechova and Max G'Sell. 2017. Fairer and more accurate, but for whom?. In *FATML Workshop*.

[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *KDD*.

[14] A. Datta, S. Sen, and Y. Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *IEEE Symposium on Security and Privacy*.

[15] Morris H. DeGroot and Stephen E. Fienberg. 1983. The Comparison and Evaluation of Forecasters. *Journal of the Royal Statistical Society. Series D* (1983).

[16] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. *COMPAS risk scales: Demonstrating accuracy equity and predictive parity.* Technical Report. Northpointe Inc.

[17] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).

[18] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*.

[19] Sharad Goel, Justin M. Rao, and Ravi Shroff. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. (2016).

[20] Trevor J Hastie and Robert J Tibshirani. 1990. *Generalized additive models.* CRC press.

[21] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. *Data Mining and Knowledge Discovery* (2014).

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*.

[23] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *ICML*.

[24] Michael P Kim, Amirata Ghorbani, and James Zou. 2018. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *arXiv preprint arXiv:1805.12317* (2018).

[25] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*.

[26] Manuel Lingo and Gerhard Winkler. 2008. Discriminatory power-an obsolete validation criterion? *Journal of Risk Model Validation* (2008).

[27] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *KDD*.

[28] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *KDD*.

[29] Francisco Louzada, Anderson Ara, and Guilherme B Fernandes. 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science* (2016).

[30] Lucas Mentch and Giles Hooker. 2016. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research* (2016).

[31] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting Good Probabilities with Supervised Learning. In *ICML*.

[32] Richard D Riley, Joie Ensor, Kym IE Snell, Thomas PA Debray, Doug G Altman, Karel GM Moons, and Gary S Collins. 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *The BMJ* (2016).

[33] Joseph Sexton and Petter Laake. 2009. Standard Errors for Bagged and Random Forest Estimators. *Computational Statistics and Data Analysis* (2009).

[34] Sarah Tan, Rich Caruana, Giles Hooker, and Albert Gordo. 2018. Transparent Model Distillation. *arXiv preprint arXiv:1801.08640* (2018).

[35] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering Unwarranted Associations in Data-Driven Applications. In *IEEE European Symposium on Security and Privacy*.

[36] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* (2018).

[37] Hao Wang, Berk Ustun, and Flavio P. Calmon. 2018. On the Direction of Discrimination: An Information-Theoretic Analysis of Disparate Impact in Machine Learning. In *International Symposium on Information Theory*.

[38] Zhe Zhang and Daniel B. Neill. 2017. Identifying Significant Predictive Bias in Classifiers. In *FATML Workshop*.

A A NEW CONFIDENCE INTERVAL ESTIMATE FOR IGAM

It is not trivial to estimate confidence intervals for nonparametric learners such as trees [30]; iGAM's base learners are shallow trees. We employ a *bootstrap-of-little-bags* approach originally developed for bagged models in [33] to estimate the variance of feature x_i's contribution to the model, $h_i(x_i)$, and difference in feature x_i's contribution to the mimic and outcome models, $sh_i(x_i) - oh_i(x_i)$.

Bootstrap-of-little-bags exploits two-level structured cross-validation (e.g. 15% of data points are selected for the test set, with the remaining 85% split into training (70%) and validation (15%) sets). Repeating this inner splitting L times and outer splitting K times gives a total of KL bags on which we train the model. Let $h_i^{lk}(x_i)$ be feature x_i's contribution to the model in the lth inner and kth outer fold. The variance of $h_i(x_i)$ can then be estimated as

$$\widehat{\mathrm{Var}}(h_i(x_i)) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{L}\sum_{l=1}^{L} h_i^{kl}(x_i) - \frac{1}{KL}\sum_{l=1}^{l}\sum_{k=1}^{K} h_i^{kl}(x_i)\right)^2,$$

and its mean $\overline{h_i(x_i)}$ can be estimated by averaging $h_i^{lk}(x_i)$ over KL bags.

We can now construct pointwise confidence intervals (CI) for feature contributions to iGAM models. The 95% CI for feature x_i's contribution to the model, $h_i(x_i)$, is $\overline{h_i(x_i)} \pm 1.96\sqrt{\widehat{\mathrm{Var}}(h_i(x_i))}$ and the 95% CI for the difference in feature x_i's contribution to the mimic and outcome models, $sh_i(x_i) - oh_i(x_i)$, is $\overline{sh_i(x_i)} - \overline{oh_i(x_i)} \pm 1.96\sqrt{\widehat{\mathrm{Var}}(sh_i(x_i)) + \widehat{\mathrm{Var}}(oh_i(x_i)) - 2\widehat{\mathrm{Cov}}(sh_i(x_i), oh_i(x_i))}$, with $\widehat{\mathrm{Cov}}(sh_i(x_i), oh_i(x_i))$ also estimated using bootstrap-of-little-bags.

This variance estimate is conservative (meaning it overestimates true variability), however, given that we are trying to detect differences between the mimic and outcome models, overestimating means we are less likely to mistake random noise for real differences. For large K and L, consistency of this estimate was established in [6].

Note that are pointwise, not uniform, CIs. That is, using the feature Age as an example, these CIs capture the variability of the effect of Age at Age=50, not the entire effect of Age. Uniform CIs can be constructed by adjusting the critical value of the CI.

B CALIBRATION PLOTS

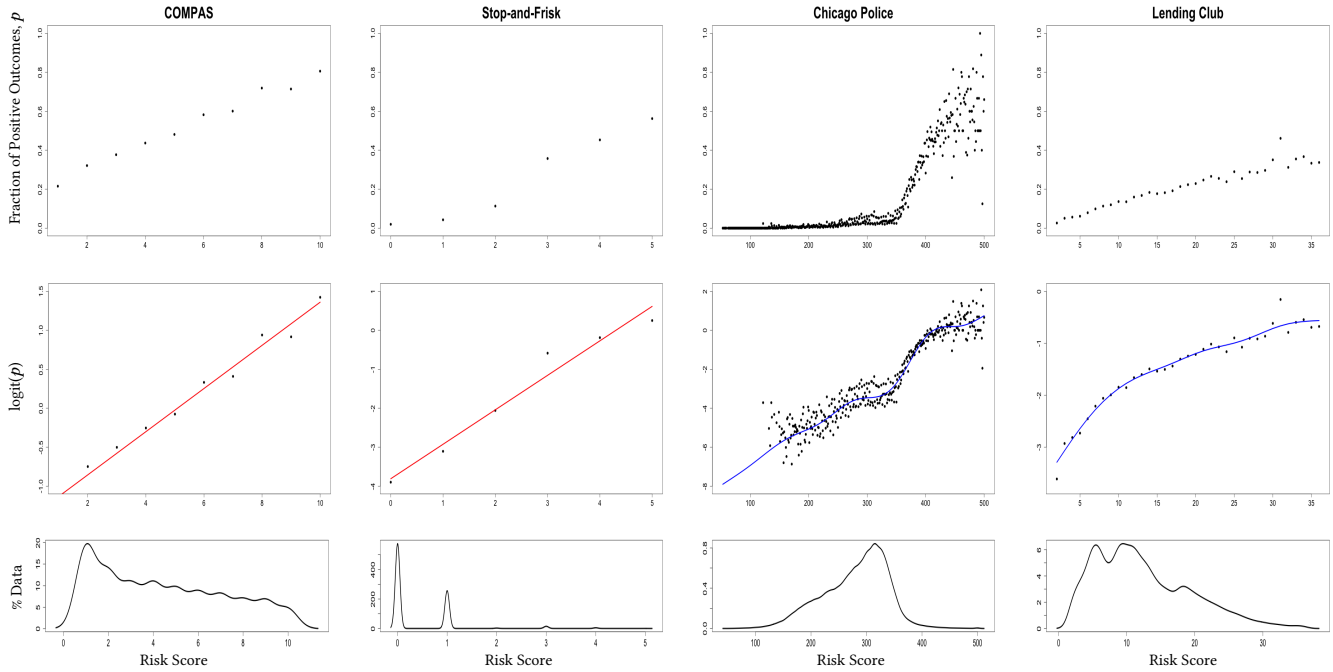

Figure 6: Empirical probability (y-axis) vs. risk score (x-axis) for COMPAS, Stop-and-Frisk, Chicago Police, and Lending Club on probability scale (top row) and logit probability scale (middle row). The risk score distribution is in the bottom row. The red lines on the logit probability scale (middle row) are best-fit straight lines. A good fit (COMPAS and Stop-and-Frisk) suggests that the risk score and logit probability of outcomes (middle row) have a linear relationship. In this case, the mimic model can be trained directly on the raw risk score. When the relationship is not linear (Chicago Police and Lending Club), the risk score must be calibrated (Section 2.3.2). The blue monotonic curves (middle row) are the nonlinear transformations learned during the calibration step. This transformation is applied to the raw risk score to yield the transformed risk score (see Figure 7).

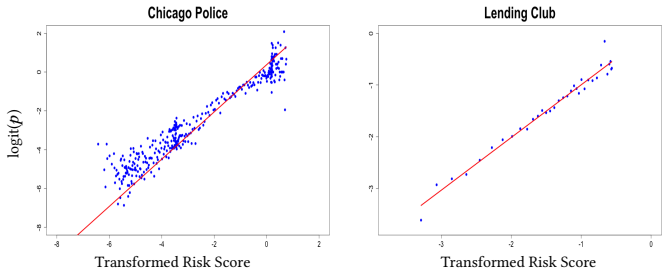

Figure 7: Logit empirical probability (y-axis) vs. transformed risk score (x-axis). The red lines are best-fit straight lines. A good fit suggests that the transformed risk score and logit probability of outcomes now have a linear relationship. The mimic model can now be trained on the transformed risk score. See Section 2.3.2 for more details.