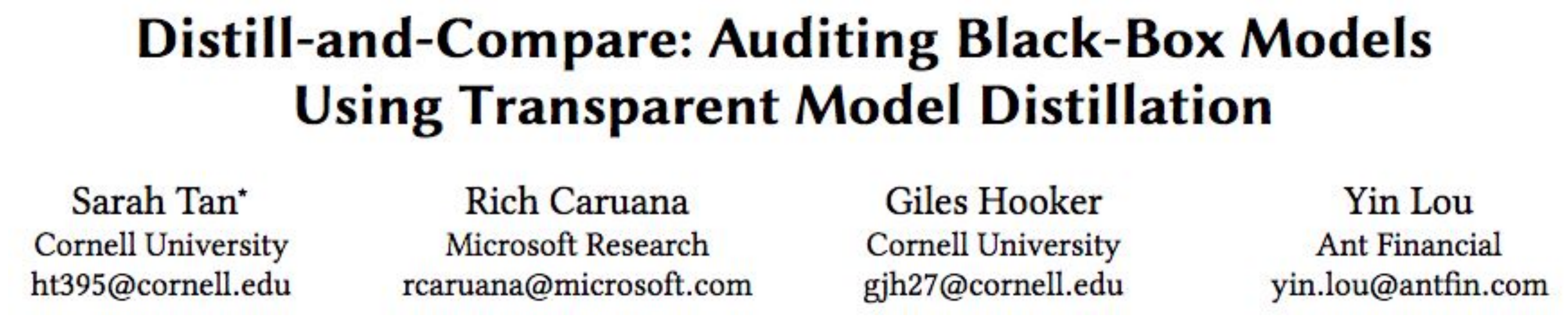


Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation

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

- **Related ML 4 healthcare NeurIPS workshop paper by Authors**
  Learning Global Additive Explanations for Neural Nets Using Model Distillation
  https://arxiv.org/pdf/1801.08640.pdf

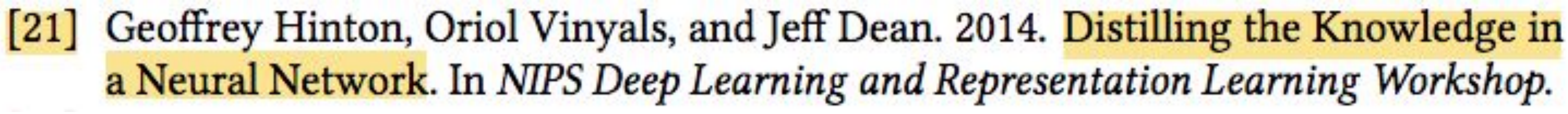
  Slides by Diego Garcia-Olano

Background: Distillation Networks

Teacher Net : “Cumbersome” Deep Network with softmax final layer that produces probability outputs for each target class

Student Network: Smaller Net that can “distill” the knowledge learned by the Teacher.

The Student uses the same data for training as the Teacher, but also uses the predicted soft probability class target outputs from the Teacher as an additional regularizer task.



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

Preliminary experiments on MNIST from Hinton, et al.

Teacher: neural net with **2 hidden layers of 1200** rectified linear hidden units
strongly regularized using **dropout** and weight-constraints.
Achieves 67 test errors

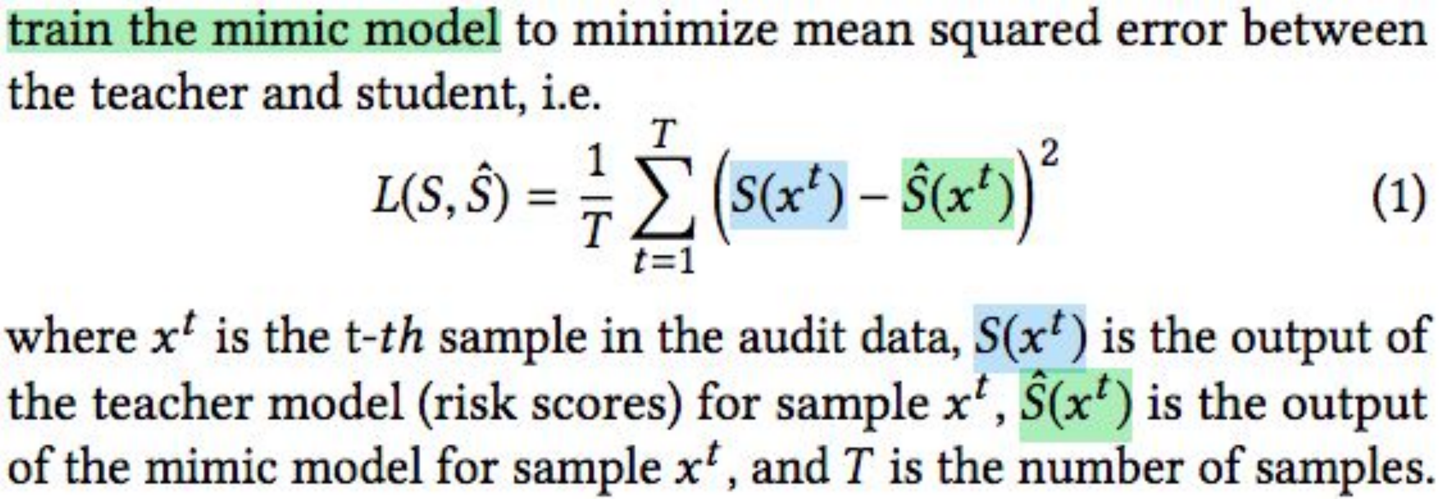
Student: neural net with **2 hidden layers of 30** rectified linear hidden units
Leveraging distilled knowledge from teacher gives similar performance.

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

Logits (zi) can be used for learning the small model
by minimizing the squared difference between the logits
produced by the cumbersome model and the logits produced by the small model.

Now Back to: **Distill and Compare Audit Models**

“Black Box” Teacher Risk Models gives outputs that a **Student** tries to “**Mimic**”



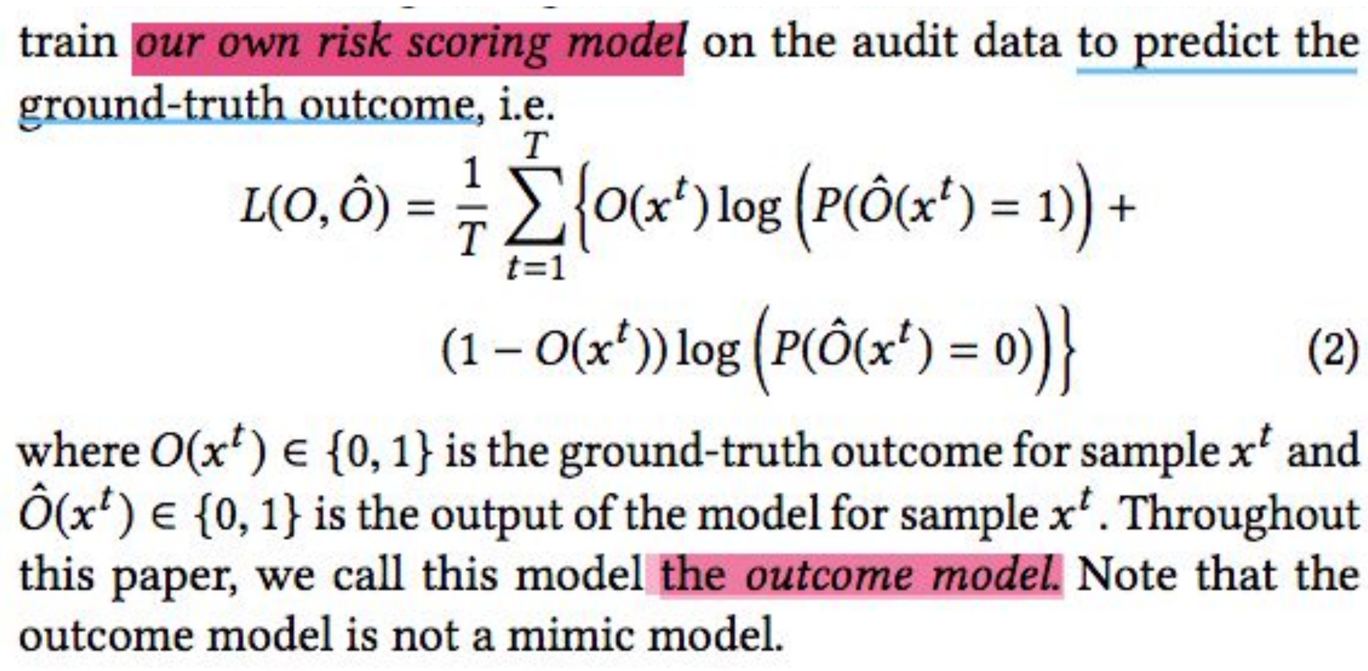
train the mimic model to minimize mean squared error between the teacher and student, i.e.

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

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.

Now Back to: **Distill and Compare Models**

Train separate Outcome Model which can be compared with the Mimic Model



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

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

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.

Paper uses iGAM for both Interpretable Generalized Additive Models (KDD 15)

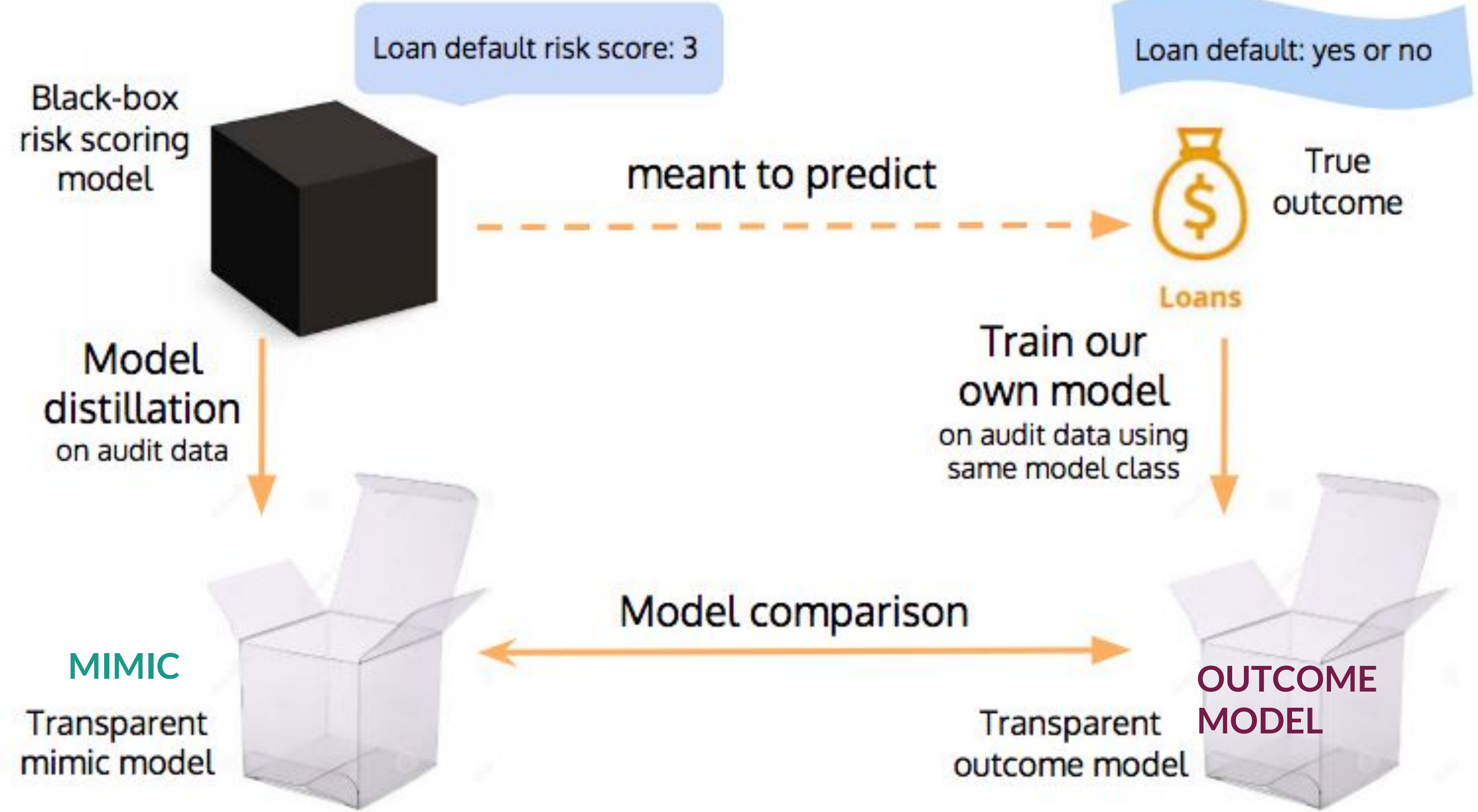


Figure 1: Auditing a loan risk scoring model by training transparent models on data labeled with the risk scores and with ground-truth outcomes for loan defaults.

3 Key Insights:

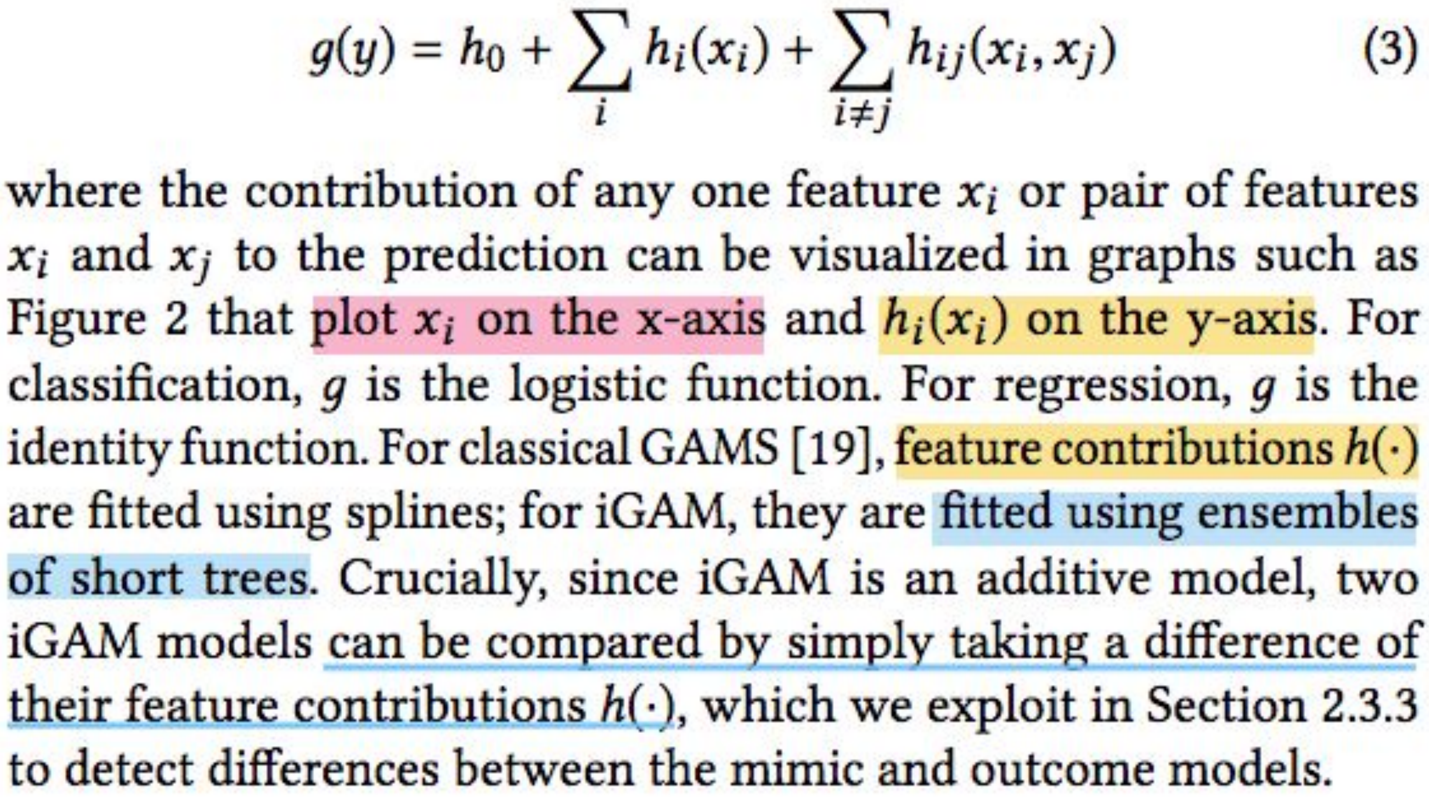
1. The ground-truth outcome is what the black-box risk model was meant to predict. If the **black-box model** is accurate and generalizes to the audit data, it should predict the ground-truth outcomes in the audit data correctly;

2. The **mimic** & **outcome** models are trained with the same model class on the same audit data using the same features. Hence the more faithful the mimic model is to the black box model, and the more accurate the outcome model, the more likely it is that **differences observed between the mimic and outcome models result from differences between** ground-truth outcomes and risk scores given by the **black box** model.

3. Similarities between the mimic and outcome 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.

Training models:

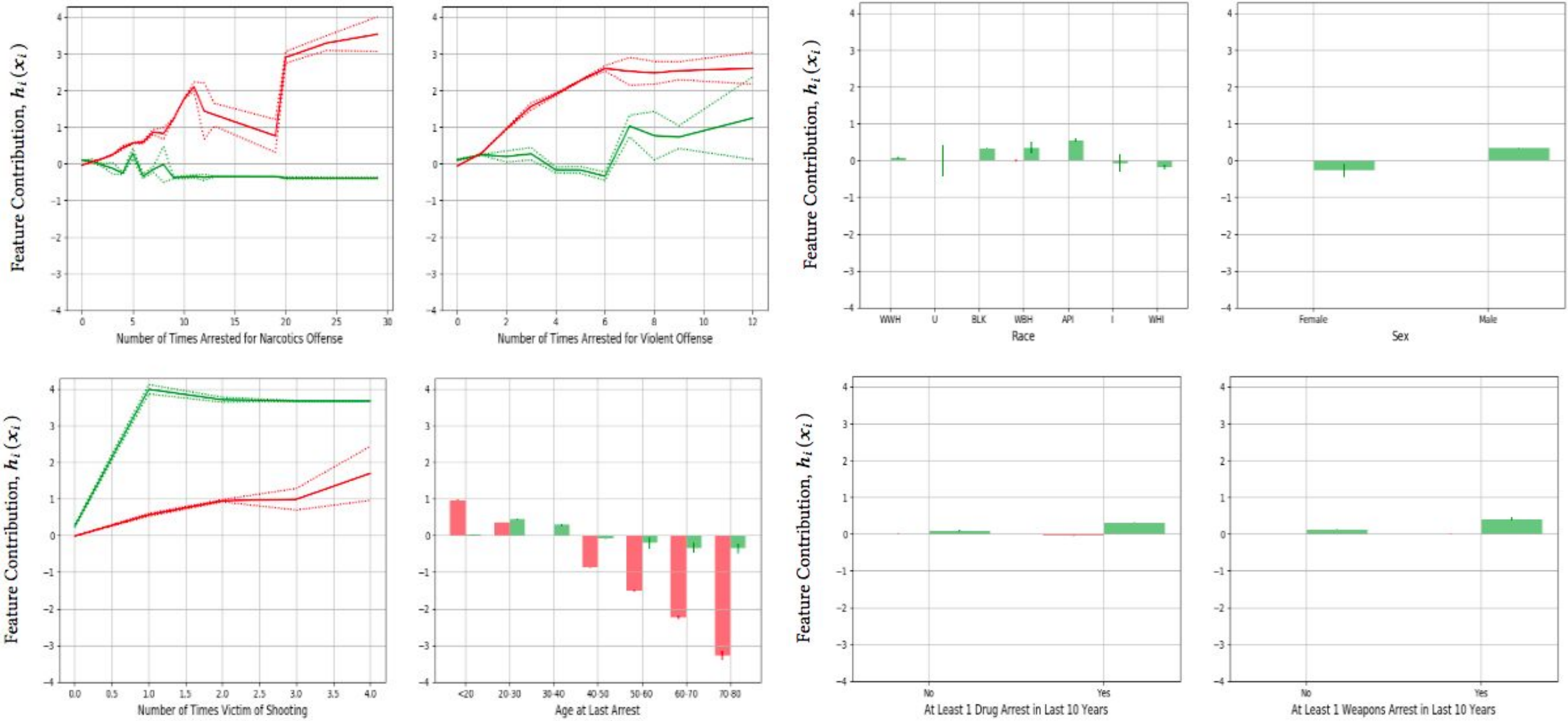
iGAMs : transparent model class for mimic & outcome



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

where 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 that plot x_i on the x-axis and $h_i(x_i)$ on the y-axis. For classification, g is the logistic function. For regression, g is the identity function. For classical GAMS [19], feature contributions $h(\cdot)$ are fitted using splines; for iGAM, they are fitted using ensembles of short trees. 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.

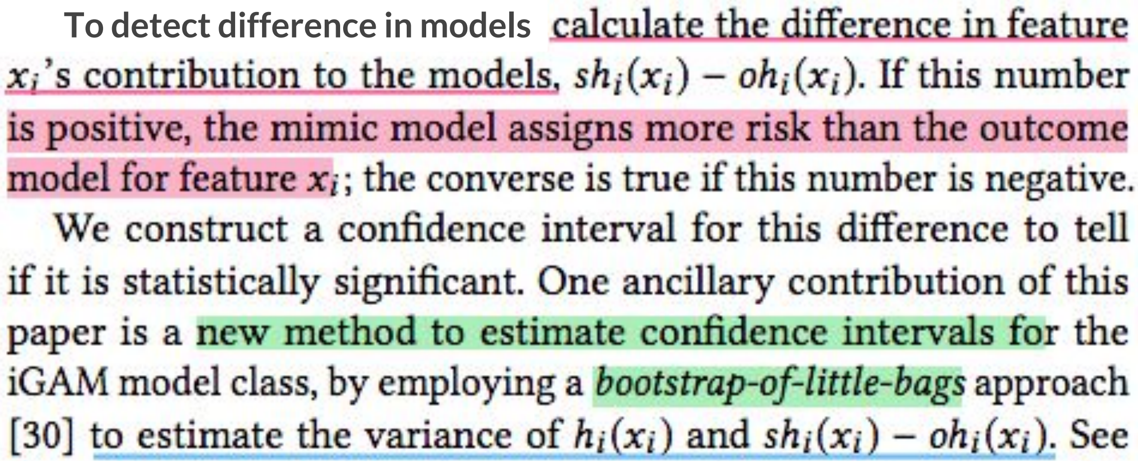
Chicago Police Audit Example. Mimic & Outcome feature contributions



Features Chicago Police used in risk scoring model ...... Features not used in risk scoring model

Comparing/auditing models:

Detecting missing features in audit from mimic & outcome



To detect difference in models calculate the difference in feature x_i's contribution to the models, $sh_i(x_i) - oh_i(x_i)$. If this number is positive, the mimic model assigns more risk than the outcome model for feature x_i; the converse is true if this number is negative.

We 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 [30] to estimate the variance of $h_i(x_i)$ and $sh_i(x_i) - oh_i(x_i)$. See

If **zero is in the confidence interval**,
the error of the mimic model is not correlated to the error of the outcome model.
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).

To conclude:

Read the **Discussion section** on why simply excluding “race”, ”gender” as features to a model does not prevent the model from learning to be biased (because of correlation with other features income, age, etc) and how this teacher/student paradigm allows for checking whether this bias exists even if the protected features are not included in the teacher model.

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; only a data set labeled with the risk score as produced by the risk scoring model and the ground-truth outcome is available. 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.

Thanks!

R code: https://github.com/shftan/auditblackbox