Auditing Black-Box Models Using Transparent Model Distillation With Side Information

Sarah Tan*
Cornell University
ht395@cornell.edu

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Giles Hooker
Cornell University
gjh27@cornell.edu

Yin Lou
Airbnb, Inc.
yin.lou@airbnb.com

ABSTRACT

Black-box risk scoring models permeate our lives, yet are typically proprietary or opaque. We propose a transparent model distillation approach to audit such models. Model distillation was first introduced to transfer knowledge from a large, complex teacher model to a faster, simpler student model without significant loss in prediction accuracy. To this we add a third criterion - transparency. To gain insight into black-box models, we treat them as teachers, training transparent student models to mimic the risk scores assigned by the teacher. Moreover, we use side information in the form of the actual outcomes the teacher scoring model was intended to predict in the first place. By training a second transparent model on the outcomes, we can compare the two models to each other. When comparing models trained on risk scores to models trained on outcomes, we show that it is necessary to calibrate the risk-scoring model's predictions to remove distortion that may have been added to the black-box risk-scoring model during or after its training process. We also show how to compute confidence intervals for the particular class of transparent student models we use - tree-based additive models with pairwise interactions (GA2Ms) - to support comparison of the two transparent models. We demonstrate the methods on four public datasets: COMPAS, Lending Club, Stop-and-Frisk, and Chicago Police.

KEYWORDS

Transparency, Black-box models, Distillation

1 INTRODUCTION

Risk scoring models have a long history of usage in criminal justice, finance, hiring, and other critical domains that impact people's lives [12, 35]. 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/or opaque[1].

One approach to detecting bias[2] in risk scoring models is to reverse engineer them, i.e., to train a nearly identical model that can be probed, or even re-trained under varying conditions. However, this can be stymied by the lack of access to the data and features used to create the original model, and lack of knowledge of what model type and algorithm was used to train it. Also, if the original model is not available to repeatedly query, it can be difficult to determine how close the reverse engineered model is to the original model. Other approaches have been proposed to audit black-box models for disparate impact [1, 2, 14, 18, 26, 42] by removing, permuting, or obscuring a protected feature and then probing or retraining the black-box outcome model to see how predictions change. Usually these focus on one or two protected features selected in advance, and thus are less likely to detect biases that are not known *a priori*. We do not assume we have access to the black box model, the original training sample, or know which features it used as inputs.

We try to gain insight into the black-box model by treating it as a teacher and distilling its output (the risk scores) into a student model that is transparent or somehow interpretable (*cf.* Section 1.1). Model distillation was first introduced to transfer knowledge from a large, complex teacher model to a faster, simpler student model, by transferring a model function embedded in a data set labeled by that function [5, 8, 27]. To this we add a third requirement: transparency. If the student model is transparent, we can understand how it is making predictions to match its teacher's outputs (the risk scores).

When available, side information can help audit the black-box risk scoring model. Here, side information is the actual outcome for each data point – exactly the true labels that the black-box risk scoring model was intended to predict in the first place[3]. This side information augments distillation in several ways: first, it reveals distortion that may be present in the the risk scores. Consider the risk score reliability diagrams [15, 37] in Figure 1. While the COMPAS and Stop-and-Frisk risk scores (1st two columns) are well-calibrated[4] to the empirical outcome, the Chicago Police risk score (3rd column) is rather flat for risk scores less than 350, then exhibits a sharp kink upwards. Creators of risk-scoring models may distort scores to achieve desired effects such as reduced sensitivity in less important regions and enhanced separation in more important parts of the scale. (Presumably distortions would not violate monotonicity.) We use this side information to detect and then undo distortions in the risk scores before distilling them. Section 2.2 shows how to do this.

A second way to use actual outcome side information is to train a second transparent model to predict the actual outcomes, and compare this transparent model to the transparent student model of the black-box teacher. Because the student and actual outcome models are both trained using the same transparent learning algorithm on the same sample of data using the same features, differences

*This work was performed during an internship at Microsoft Research.

[1] Often they not available to their users (e.g. judges and case workers), individuals being scored, or external parties for assessment and validation.

[2] Legal scholars use two definitions of bias: disparate treatment and disparate impact. An example of the former is explicitly using a protected feature such as race or gender when deciding an outcome. An example of the latter is different categories of a protected feature exhibiting disparate outcomes despite the protected feature not having been used explicitly in the decision. See [32] for a discussion of these concepts in the context of machine learning models. Since discovering disparate treatment in a black-box model requires access to the model or at least knowledge of whether it used a protected feature or not, most research on auditing black-box models has centered on uncovering disparate impact.

[3] Using an analogy from the active learning literature, this side information of actual outcomes is the true label given by an oracle.

[4] Being well-calibrated overall does not imply well-calibrated over subgroups. See [12].

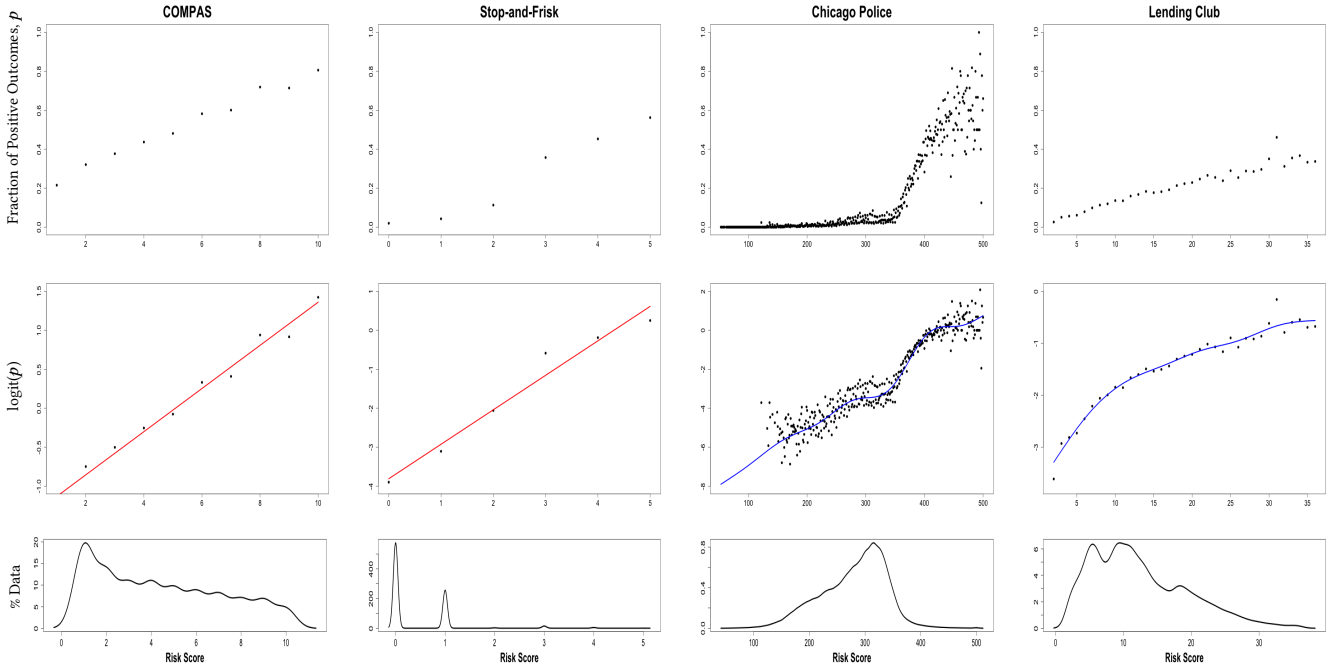

Figure 1: Reliability diagram of *empirical* probability of positive actual outcomes for each risk score bin, on probability scale (top row) and logit probability scale (middle row). The red lines are best-fit straight lines. A good fit suggests that the risk score and the logit probability for the actual outcomes have a linear relationship and can be directly compared (*cf. Section* 2.4). When the relationship is not linear, the risk score must be calibrated prior to comparison. The blue monotonic curves are the learned nonlinear transformations. See Figure 3 for the transformed risk score. In the bottom row are risk score histograms.

between the two models are likely due to differences between the risk scores assigned by the black-box teacher model and the true outcomes. We demonstrate this in Section 6.3 on COMPAS.

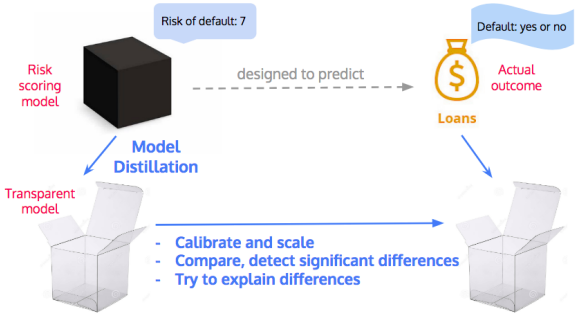

Figure 2: An example of the approach applied to a black-box loan default risk scoring model with side information on whether the loan actually defaulted.

A third way to benefit from side information is to look not for differences between the student and outcome models, but where the models are similar. For example, on COMPAS in Section 6.3, we observe that both the student and outcome models model the most important feature, number of priors, almost identically. When the student and outcome models strongly agree on how to model the effects of some features, it increases confidence that the student model is a faithful representation of the black-box teacher model, and that any differences observed on other features are meaningful.

This approach of comparing a distilled student model with another model of the same type trained on the actual outcomes is very different from just training a single transparent model on actual outcomes and seeing what it learned. The similarities and differences between the transparent student model of the black-box and the transparent model of actual outcomes are more informative than just training a model to mimic the black box, and also more informative than training a model on data that is not the black-box's original training data. We suspect it is better to perform distillation with as strong a student model as possible, as long as the model remains understandable. This allows a single global approximation to be used where weaker models (e.g. linear) would require multiple local approximations[5]. Auditing a black-box by looking at a large number of independent local approximations is not straightforward.

Most approaches to auditing black-box models are complicated by correlations among features, particularly correlation between protected and non-protected features. Our approach is not immune to this. Disentangling correlations between features is difficult and computationally expensive. By studying two models of the same

[5]Simple models sometimes fail to capture the non-linearity of more complex black-box models unless they are *very* local .

model class, trained on the same data sample, using the same features to predict labels that are different yet intimately related, the hope is that the two models learn similar patterns and make similar mistakes *except* when differences between the risk scores (and thus what was learned by the black-box model) and the true outcomes require the models to be different.

The main contributions of this paper are:

(1) Using transparent model distillation to audit four public datasets for potential bias.
(2) Showing how side information (true labels) can be used to aid interpretation of transparent student models.
(3) Providing a calibration procedure that uses side information to remove distortions in risk scores.
(4) Providing a method for calculating confidence intervals for a particular kind of transparent model class we experiment on (Ga2Ms [9, 33, 34]) to support comparing student models to actual outcome models.

1.1 Related Work

We place the proposed approach in the context of existing methods, and overview key concepts used, including model distillation, transparency and interpretability, and calibration.

Auditing black-box models: Several approaches have been proposed to audit black-boxes for disparate impact. Datta et al. [14] propose new feature importance measures that account for correlated features, but require access to the black-box model and knowledge of input data and features used. Others require query access to the black-box [1, 2, 26, 42] or retraining [18].

Some aspects of the proposed approach have been hinted at in previous papers. While not the main focus, in one experiment, Adler et al. [2] trained a model to predict outcomes and then an *overfitted*, interpretable model to predict the first's predictions. This is a different distillation setup from ours, and we use both risk scores and outcomes. Adebayo and Kagal [1] also learn their own risk scoring models when the black-box model cannot be queried.

Like our approach, some papers study two models in tandem, but not risk scores and outcomes at the same time. Wang et al. [43] train a model to predict outcomes and another to predict membership in a protected subgroup, then use the connection between the two models to identify features that proxy for protected features. Chouldechova and G'Sell [11] train two different outcome models then identify subgroups where the two models differ in terms of fairness metric of interest. They do so to compare classifiers (e.g. to determine if a new classifier should be adopted) whereas we want to uncover disparate impact in one risk-scoring model. Other papers work on a single model, discovering subgroups that are miscalibrated when predicting outcomes [45], or exhibiting significant associations between protected features and black-box outputs [41].

Finally, several fairness criteria on the interplay between risk scores and outcomes have ben proposed, such as subgroup well-calibrated, predictive parity and error rate balance [10, 24, 30].

Model distillation transfers a learned mapping of model inputs to outputs. As Hinton et al. point out [27], a conceptual block that may have hindered wider adoption of the approach is the association of the knowledge learned by a model with its learned parameter values, whereas model distillation extracts a model's learned knowledge from a data set labeled by the model [5, 8, 27].

Transparency and interpretability: The definitions of these concepts are still in flux [17, 31]. Transparency and interpretability are highly correlated with model complexity, and simpler models should make it easier for humans to understand the model.

Calibration is the process of matching predicted probabilities with empirical rates in data. Standard calibration techniques include Platt scaling and isotonic regression [37]. Whether a risk score is well calibrated across *subgroups* (not just overall) is one of several fairness metrics. Several papers have shown that it is impossible to simultaneously achieve this and balance for positive and negative classes [10, 24, 30], and Corbett-Davies [12] show an example of classifier that is well-calibrated but intentionally designed to hide racial disparities. We use calibration techniques to undo distortions in the risk score scale.

2 THE APPROACH

We formalize the proposed distillation approach to mimic a black-box model and how side information can help.

2.1 Transparent Model Distillation

Let $r^S : \boldsymbol{x} \rightarrow y^S$ be the (unknown) black-box risk-scoring function. Let $\boldsymbol{x}$ be p-dimensional and x_j denote its jth feature. Data set $\mathcal{D}$ is of size N where data point i has features $\boldsymbol{x}_i$ and two labels: actual outcome y_i^O and risk score y_i^S assigned by the black-box. Let $\mathcal{M}$ be a class of transparent models. We train a student model of this class to predict the outputs of the teacher model (risk scores y^S).

Transparent Model Distillation
Student model of black-box risk score teacher r_S, trained on teacher's outputs y^S:
Input: label y^S, features $\boldsymbol{x}$
Output: prediction $\hat{y}^S$
Model functional form: $\hat{y}^S = f^S(\boldsymbol{x})$

We want f^S to be as rich and complex as possible so it can be a faithful student to r^S, yet still simple enough to remain understandable. We experiment with two types of transparent models in this paper: GA2Ms and linear models.

Generalized Additive Models with Pairwise Interactions (GA2M): A variant of generalized additive models [9, 33, 34] based on shallow trees, GA2Ms' claim to transparency stems from its additive form[6]:

$$g(y) = f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p} h_j(x_j) + \sum_{j=1}^{p} h_{jk}(x_j, x_k)$$

where the contribution of any one feature (or pair of features) to the prediction can be visualized in graphs such as Figure 4. Each term $h_j(x_j)$ is an ensemble of shallow trees restricted to operate on only one feature, and $h_{jk}(x_j, x_k)$ is again an ensemble of shallow trees but operating on pairs of features to capture pairwise interactions. Higher order interactions can be added to further increase accuracy, but they are less easy to visualize and thus more difficult to understand.

[6]For classification, g is the logistic link. For regression, g is the identity.

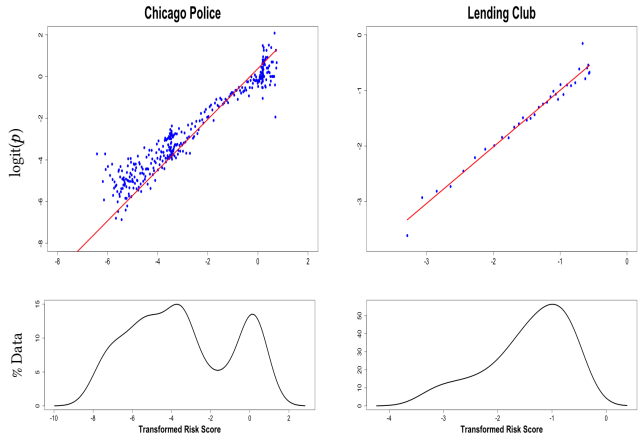

Figure 3: Relationship between logit *empirical* probability and transformed risk score; similar to the middle row of Figure 1, with x-axis being the risk score after applying the non-linear transformation. Red lines are best-fit straight lines. Good fit suggests the *transformed* risk score and logit probability now have a linear relationship and can be compared. The bottom row are *transformed* risk score histograms.

The model is related to additive function decomposition methods [23, 28] used, for example, in hyperparameter optimization to examine the importance and interactions between hyperparameters [29]. Terms are learned together using gradient boosting to obtain an additive formulation. However, unlike classical GAMs [25] where features are shaped using splines, GA2Ms shape features using short trees. A contribution of this paper is devising a new variance estimate for the main feature contributions, $h_j(x_j)$, and pairwise feature contributions, $h_{jk}(x_j, x_k)$, of GA2M models (*cf.* Section 2.5).

Linear and Logistic Regression: linear models of the form:

$$g(y) = f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

the variance of their feature contributions β_j can be estimated according to maximum likelihood theory [36].

Models Not Used: Decision trees have also been suggested to be interpretable. They were used explicitly in [21] and more recently in [11]. However, their transparency depends on the tree structure that can change dramatically from fold to fold and their feature contributions lack confidence intervals. Full complexity models such as random forests, gradient boosted trees, and neural networks are accurate but lack transparency and feature contributions can only be determined post-hoc using methods such as partial dependence [19], feature importance measures [7], etc. that may suffer from bias due to extrapolation off the training manifold [28] or ignored correlations between features [40].

The two transparent models we experiment on in this paper are additive models that:

(1) Do not require the post-hoc methods described above, as their predictions are exactly the sum of individual feature contributions.
(2) Have confidence intervals for their feature contributions, so that statistical significance can be determined.

We leave *post-hoc* transparent models for future work.

2.2 Calibration to Undo Risk Score Distortions

We saw in Section 1 and Figure 1 an example of a risk score (Chicago Police) where distortions may have been applied to some parts of the scale to provide finer resolution. To undo this distortion, we make the reasonable assumption that risk scores should be monotonic and well-calibrated (reflecting empirical probabilities).

Specifically, when the relationship between the risk score and the logit probability (second row of Figure 1) is nonlinear, as is the case for the Chicago Police and Lending Club risk scores, we learn the blue line, a nonlinear monotonic transformation to map the risk scores to a new scale where they are now linearly related to logit probability, as can be seen in Figure 3. This is related to the common calibration method of isotonic regression, except that the transformation is learned using a monotonic spline instead of a monotonic stepped function, giving additional smoothness. We then use the transformed risk scores, instead of the raw risk scores, as the input to the student models. When the risk scores already have a linear relationship with logit probability (COMPAS and Stop-and-Frisk, the 1st two columns of Figure 1), this calibration step is not needed, and their raw risk scores can be used directly as inputs to the student model.

2.3 Transparent Distillation + Side Information

When distilling black-box risk-scoring models, side information on actual outcomes can help. The risk score and the actual outcomes (that the risk score was intended to predict) are intimately related. We train a second transparent model of the same class $\mathcal{M}$ to predict actual outcomes, then compare this second transparent model to the transparent student model trained to mimic the black-box.

Transparent Model Distillation, With Side Information

Student model of black-box risk score teacher r_S, trained on teacher's outputs y^S:
Input: label y^S, features $\boldsymbol{x}$
Output: prediction $\hat{y}^S$
Model functional form: $\hat{y}^S = f^S(\boldsymbol{x})$

Model of actual outcome y_O:
Input: label y^O, features $\boldsymbol{x}$ (same as student model), same data points as student model
Output: prediction $\hat{y}^O$
Model functional form: $\text{logit}(\hat{y}^O) = f^O(\boldsymbol{x})$, since $\hat{y}^O$ is binary. This model is not a student model, as actual outcomes are not labeled by another model.

Now that we have two transparent models, one that assigns risk scores (student model of black-box) and one of actual risk scores (model of actual outcomes), we want to compare how they make their predictions. The two types of transparent models we experiment with both make predictions by summing contributuons from individual features or pairs of features.

2.4 Comparing Two Transparent Models

Comparing feature contributions from two models to each other requires ensuring that: 1) their feature contributions are on the same scale; 2) differences are not due to random noise.

Scaling considerations: Section 2.2 described a nonlinear transformation to transform risk scores to be linearly related to logit probabilities. The logit probability scale is precisely the scale on which the outcome model f^O represents the effect of contributions of individual features x_j, since it uses the logit link.

In addition to transforming the scores, we rescale both distilled models once they are trained to ensure that they provide predictions on the observed data with the same standard deviation. That is, we multiply one to match the standard deviation of the other. This is done to account for student models having an attenuated range of predictions relative to the range of the scores they are trained to predict – a consequence of regression to the mean. In some cases, the score also may have not been trained using the same features as are available for distillation, leading to a greater or smaller range of predictions from the outcome model. Once feature contributions from both models are on the same scale, they can be compared.

Detecting Differences: We want to compare the difference in the contribution of feature x_j to the score student model, $sh_j(x_j)$, compared to the outcome model, $oh_j(x_j)$. To tell if the difference is statistically significant, in Section 2.5 we describe how a confidence interval can be constructed for this difference.

Controlling Variability: To ensure that differences between the two models are not arising from them being trained on different data points, we train the two models only on the same data sample. However, this induces correlation between the feature contributions, which we account for by estimating and including in the confidence interval for the difference (*cf.* Section 2.5).

2.5 Estimating Sample Variance for GA2M

One contribution of this paper is a new variance estimate for GA2M models. We employ a *bootstrap-of-little-bags* approach originally developed for bagged models in [39] to obtain pointwise confidence intervals for GA2M feature contributions, and the difference of two GA2M models' feature contributions.

Bootstrap-of-little-bags is based on two-level structured cross-validation: 15% of data points are selected for the test set. The remaining 85% is split into training (70% of the total data) and validation (15%) sets. We repeat this inner splitting L times, and outer splitting K times, for a total of KL training samples on which we train the student model (on label y^S) and outcome model (on label y^O). Specifically, let $h_j^{lk}(x_j)$ be x_j's feature contribution estimated by the model in the lth inner and kth outer fold. Its mean is the ensemble average of the KL training models, and we estimate its variance as:

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

This variance estimate is conservative (meaning it overestimates true variability), however, given that we are trying to detect differences, overestimating means we are less likely to mistake random noisy differences for real differences. For large K and L, consistency of this estimate was established in [4]. A confidence interval for contribution of feature x_j to a GA2M model is then given by $h_j(x_j) \pm 1.96\sqrt{\widehat{\mathrm{Var}}(h_j(x_j))}$.[7]

3 EMPIRICAL RESULTS

In this section we describe the results of applying transparent model distillation to four black-box models for which we have both risk scores and actual outcomes side information.

3.1 COMPAS: Recidivism Risk and Outcome

COMPAS, a proprietary score developed to predict recidivism risk, has been the subject of scrutiny for racial bias [3, 6, 10, 12, 16, 30, 44]. Because the algorithm does not use race as an input [38], its proponents suggest that it is race-blind. ProPublica collected, analyzed[8], and released data[9] on COMPAS scores and actual recidivism outcomes of defendants in Broward County, Florida. Candidates for protected features in this data set are age, race, and sex. COMPAS is a black-box model because it is protected by IP, not because it is necessarily complex. We do not know what model type, features or data were used to train the original COMPAS model.

Figure 4 shows four shape plots for four of the features available for recidivism prediction: Age, Race, Number of Priors, and Gender. The top row shows what was learned by transparent models trained to predict 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. The transparent model trained on the true outcome (green) shows what can be learned from the data itself. 95% point-wise confidence intervals are shown for both models. The bottom row of Figure 4 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 may be biased for some age and race groups. Examining the plots on the left of Figure 4 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, the variance is large but there is evidence that the models disagree. The difference between the COMPAS mimic model and the true label model is most 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 4 show risk as a function of Race. The COMPAS mimic model predicts that African Americans are higher risk, and that Caucasians are lower risk, than the transparent model trained on the true labels suggests

[7] A confidence interval for the difference in feature contributions of x_j to the score student model compared to the outcome model is $sh_j(x_j) - oh_j(x_j) \pm 1.96\sqrt{\widehat{\mathrm{Var}}(sh_j(x_j) - oh_j(x_j))}$ where $\widehat{\mathrm{Var}}(sh_j(x_j) - oh_j(x_j)) = \widehat{\mathrm{Var}}(sh_j(x_j)) + \widehat{\mathrm{Var}}(oh_j(x_j)) - 2\widehat{\mathrm{Cov}}(sh_j(x_j), oh_j(x_j))$

[8] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

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

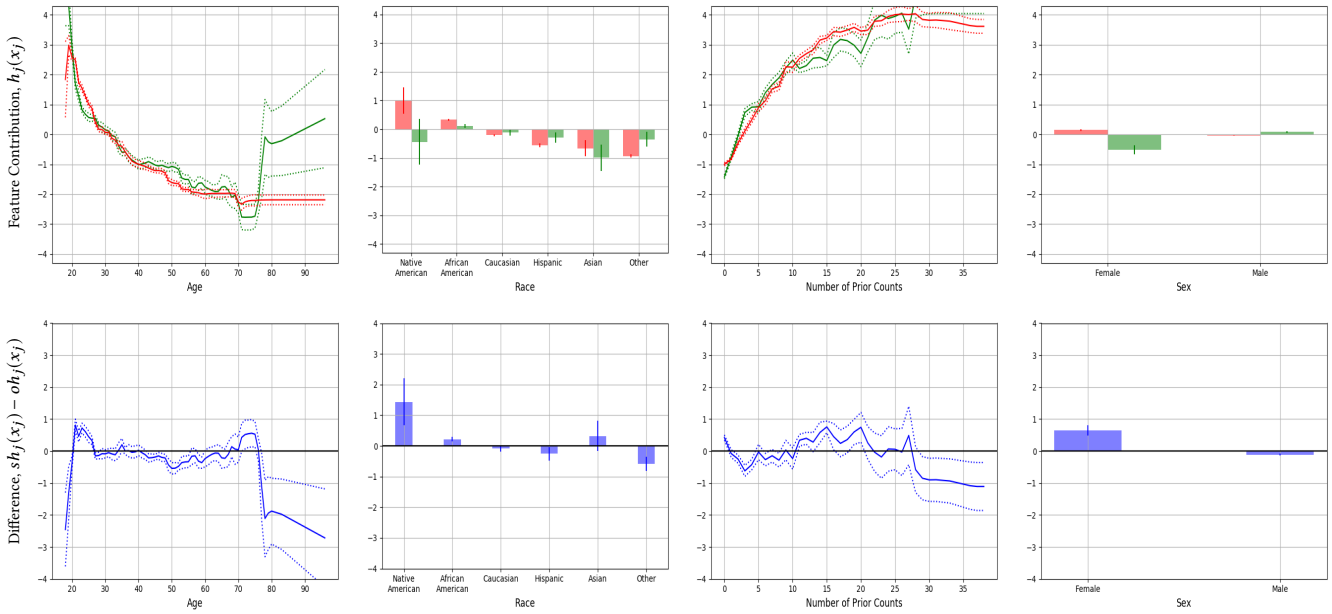

Figure 4: GA2M shaped feature contributions for four features to the COMPAS risk score student model (in red) and the actual recidivism outcome model (in green). For categorical features such as race and sex, categories are ordered in decreasing importance to the score student model. The blue line captures the difference between the two models (score student - outcome). All plots mean-centered on the vertical axes. See Appendix for additional features and interactions between *pairs* of features.

is warranted. The COMPAS model may be 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 are 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. Although we do not have space to include the figures here, the GA2M models identify a number of significant pairwise interactions between gender and other features. This is due in part because the data is predominantly male, so the main effects are more correct for males than females, and the interactions between gender and other features allows the GA2M model to correct the predominately male main effects.

Interactions make GA2M models more accurate by allowing them to model effects that cannot be represented by a sum of main effects on individual features. We observe a number of interesting interactions between gender and other features such as age, length of stay, and number of prior convictions. The pairwise interactions for COMPAS are included in the Appendix. We suspect that the COMPAS black-box model may not be complex enough to properly model interactions, and that this might explain some of the differences we observe between the transparent student model train to mimic COMPAS, and the transparent model trained on the true outcomes. For example, we noted above that the COMPAS model appears to be biased in favor or very young offenders, but we see no evidence to support this in the true outcome model. There are strong interactions between very young age and other variables such as gender, charge degree, and length of stay that we suspect COMPAS is not able to model, and that this may explain why COMPAS needs to predict low risk for very young offenders (because it can't otherwise predict a reduced risk via interactions of age with other variables).

3.2 Lending Club: Loan Risk Score and Defaults

Lending Club, an online peer-to-peer lending company, makes information public on the loans it finances[10]. We use a subset of five years (2007-2011) of loans that have matured, the Lending Club assigned risk score, and the outcome of whether the loan defaulted. 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 and zip code. We do not know what model type Lending Club used for their black-box risk-score model. We believe Lending Club may use additional features that are not available in this public dataset, and their models may not use all of the features

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

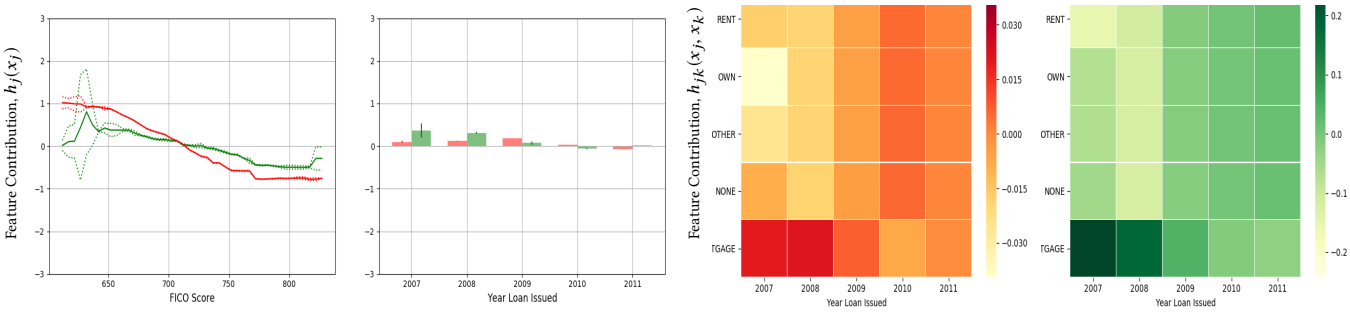

Figure 5: Selected GA2M shaped main and pairwise feature contributions to the Lending Club risk score student model (in red) and the actual loan default outcome model (in green). See Appendix for additional main and pairwise features.

that are in this public dataset. We also believe the sample data we have is similar to the data they would have used to train their models. According to Lending Club, their models are refreshed periodically.

Figure 5 shows selected main and pairwise features for the Lending Club data. Remaining main and pairwise features are in the Appendix. As in Figure 4, red lines show what was learned by the transparent student model trained to mimic the lending model by training on risk scores assigned by that model, and green lines show what a transparent model learned when trained on the true credit default 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.

FICO Score: The models agree qualitatively on FICO score, but not quantitatively — the student model suggests the black-box gives more weight to FICO than the model trained on true outcomes. We suspect this may be because the true outcome model uses more features than the black-box, so the black-box model places greater emphasis on the most important feature it uses: FICO.

Year Issued: The green bars in this graph demonstrate a increase in risk in loan defaults (the actual outcome) for 2007 and 2008, exactly around the time of the subprime crisis. Yet the red bars for the risk-scoring model do not change as dramatically over the years, appearing to lag behind the actual risk. This would make sense if the Lending Club black-box model was updated conservatively, instead of being rapidly updated as economic conditions and behavior change. The pairwise interactions of the year of loan issue and home ownership status (the two graphs on the right) shows that having a home mortgage in 2007-2008 increases the Lending Club loan default risk more than having a home mortgage in 2009 and beyond. Note that difference in ranges between the two pairwise plots - the range goes up to 0.2 for the outcome model (in green) whereas the range is much lower for the student risk scoring model (in red), supporting the hypothesis that the Lending Club black-box model is updated with some lag time.

3.3 Chicago Police Strategic Subject List Risk Score and "Party-to-Violence" Outcome: Detecting Used and Unused Features

The Chicago Police data set is unusual: the data contains 16 features, but the description of the models says only 8 of these are used by the model. This gives us an oppotunity to test if transparent model distillation with side information can accurately detect which features are and are not used by a black-box model. We do this by comparing a transparent GA2M student model trained to mimic the Chicago Police black-box to a second transparent GA2M model trained on the true outcomes *when both GA2M models are trained using all 16 features.* If transparent distillation is providing high-fidelity information about what is in the black-box model, it should be able to detect what features are and are not used by that model.

The Chicago Police Department released arrest data from 2012 to 2016[11] that was used to create a risk assessment score for the probability of an individual being involved in a shooting incident as a victim or offender. Candidates for protected features in this data include race and age. We do not know what model type was used for their black-box, but we do know what 8 features they used in the model of the 16 that were available.

We trained a transparent student model to mimic their model, and intentionally included all 16 features in the student model. Figure 6 shows main effects learned by the student model for the eight features the Chicago Police Department used in their model, and Figure 7 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 a striking difference between the plots in Figure 6 and in Figure 7: there is very little red visible in Figure 7 but a lot of red visible in Figure 6. The transparent model trained to mimic the black-box makes significant use of the features used in the black-box, but little to no use of the features no used by the black-box. Yet the amount of green in the two figures is similar, suggesting that there is signal available in the 8 unused features that the Police model could have used, but chose not to use. This confirms that the transparent student model can provide insight into the inner workings of a black-box model, and demonstrates one of the advantages of using side information.

When comparing what a transparent student learns from the black-box with what a transparent model learns from true labels, it is valuable to train the transparent model on the outcomes two ways: 1) using all available features to see what could have been

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

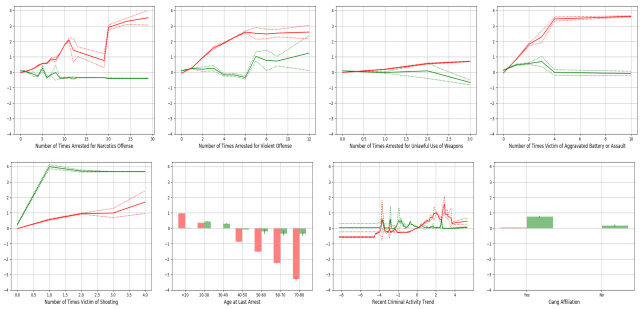

Figure 6: Eight features the Chicago Police says are used in their risk scoring model.

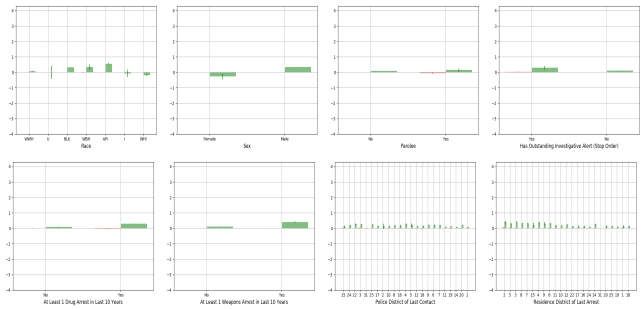

Figure 7: Eight features the Chicago Police says are *not* used in their risk scoring model.

Table 1: 95% confidence intervals for correlation test for likelihood of missing features used by black-box model

Data	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]

learned from the original labels, 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.

3.4 NYPD Stop-and-Frisk Risk Score and Weapon Possession True Outcome

The New York Police Department's stop-and-frisk data[12] has been scrutinized for racial bias [13, 20]. Goel et al. [22] proposed a simple, heuristic risk score model[13], created on 2009-2010 data and tested on 2012 data, for the probability of an individual possessing a weapon:

$$\text{Risk score } y^S = 3 \times \mathbb{1}_{PS} + 1 \times \mathbb{1}_{AS} + 1 \times \mathbb{1}_{Bulge}$$

where *PS* denotes primary stop circumstance is presence of suspicious object, *AS* denotes secondary stop circumstance is sight or sound of criminal activity, and *Bulge* denotes bulge in clothing [22]. We apply the risk scoring model to label 2012 data (n=126,457, p=40) after following Goel et al.'s pre-processing steps. On this problem the risk-scoring model is not a black-box. We know the precise functional form, are using the same features, and can train on a similar sample of data[14]. Both GA2M and linear regression student models recover the coefficients of (3, 1, 1) for the three features used, up to miniscule variance. We note that this is an easy risk-scoring model, with only three features and a simple functional form.

4 DISCUSSION

4.1 Including All Available Features

Sometimes we are interested in detecting potential bias on features that have intentionally been excluded from the black-box model. For example, a credit risk scoring model is probably not allowed to use race as an input. Unfortunately, excluding 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 labels used for learning; not using race as a *feature* does not remove the bias from the *labels*.

When training a transparent student model to mimic a black-box model, we intentionally include all features that may or may not have been originally used to create black-box risk scores, even protected features, specifically because we are interested in examining what the models *could* learn from them.

4.2 Indications of Missing Features

As the black-box may have used additional features which we do not have access to, we developed a test to assess the impact missing features could have on our analysis based on the following:

> *If the black-box risk scoring model has access to hidden features, and these are useful for predicting the outcome, then the error between the student model and risk score should be positively correlated with the error between the student model and the outcome model.*

Table 1 provides confidence intervals for the correlation between score and outcome residuals for f^S – the score student model – based on three correlation statistics. In Lending club and Stop-and-Frisk we cannot distinguish these correlations from zero. In Police and Compas, there appears to be some evidence for correlation, indicating that for these data sets, the student model may not have access to all relevant features. However, the upper end of these intervals is never more than 0.14 – not a very strong effect.

4.3 Fidelity and Accuracy

Table 2 describes the fidelity of the risk score student models, and Table 3 describes the accuracy of the outcome models.

[12] http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
[13] Note that this risk score was proposed by an academic research group, not NYPD.
[14] Since Goel et al. tested their model on 2012 data – the data we train on – and training and testing samples should be similar.

Table 2: Fidelity of student model to teacher outputs (risk score).

Data	Metric	Linear / Logistic Regression	GAM	GA2M	Random Forest
Compas	Risk score (1-10) RMSE	2.11 ± 0.057	2.01 ± 0.045	**2.00 ± 0.047**	2.02 ± 0.053
Lending Club	Transformed risk score (-3.3-0.6) RMSE	0.27 ± 0.006	0.21 ± 0.008	0.20 ± 0.008	**0.19 ± 0.006**
	Risk score (2-36) RMSE	3.27 ± 0.037	2.60 ± 0.049	2.52 ± 0.051	**2.48 ± 0.033**
Chicago Police	Transformed risk score (-7.9-0.8) RMSE	0.32 ± 0.018	0.32 ± 0.018	0.31 ± 0.012	**0.29 ± 0.014**
	Risk score (0-500) RMSE	17.4 ± 0.102	17.2 ± 0.125	16.5 ± 0.130	**14.0 ± 0.280**
Stop-and-Frisk	Risk score (0-5) RMSE	$\mathbf{0.000 \pm 2 \times 10^{-15}}$	$\mathbf{0.000 \pm 1 \times 10^{-5}}$	$\mathbf{0.000 \pm 2 \times 10^{-5}}$	$0.010 \pm 2 \times 10^{-3}$

Table 3: Accuracy of outcome model.

Data	Metric	Linear / Logistic Regression	GAM	GA2M	Random Forest
Compas	Actual outcome AUC	0.73 ± 0.029	0.74 ± 0.027	**0.75 ± 0.029**	0.73 ± 0.026
Lending Club	Actual outcome AUC	0.69 ± 0.006	**0.69 ± 0.016**	**0.69 ± 0.014**	0.68 ± 0.020
Chicago Police	Actual outcome AUC	**0.95 ± 0.007**	**0.95 ± 0.007**	**0.95 ± 0.007**	0.929 ± 0.009
Stop-and-Frisk	Actual outcome AUC	0.84 ± 0.020	0.85 ± 0.020	0.85 ± 0.020	**0.87 ± 0.024**

Fidelity – how close the student is to its teacher – can be assessed as how accurately the student models predict their teachers' outputs (risk scores) on test-sets. For COMPAS, all models have roughly equal fidelity and accuracy. For fidelity, no models had RMSE lower than 2 on a 1-10 scale. Possible reasons why the COMPAS score is challenging to predict include the ProPublica data sample missing essential features. This agrees with the findings of a test for likelihood of missing features we proposed in the previous section. Another possible explanation is the small size of this data set (6,172 points).

One advantage of model distillation is it can benefit from additional unlabeled data if the black-box teacher can be queried to label the data [8]. We found additional data points (3k) from the ProPublica data with assigned risk scores but not actual outcomes. Adding them to the training (not testing) sample for the student model, and retraining the students, we find marginal improvement in the student model's fidelity regardless of (from RMSE 2.0 to 1.98). Doing the opposite – removing points from our training sample in 1,000 increments – the student's fidelity decreases only marginally (to RMSE 2.1 using 1k training points). These analyses suggest that for COMPAS, missing salient features is a more pressing than lack of data.

In terms of cross-model class comparisons, GA2M's AUC results are generally comparable to (or slightly better than) more complex, less intelligible models such as random forests (Table 3). For the risk score student models, random forests are competitive on the Lending Club and Chicago Police data sets. Linear and logistic regression are not as far behind for several of data sets, suggesting that the model functional form might be very simple. For the Stop-and-Frisk data where the model functional form is a simple linear form (*cf.* Section 3.4), unsurprisingly, linear regression performs well whereas the noisiness of forests reduces its accuracy.

4.4 Bias Discovery via Transparency

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. 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.

4.5 Other Transparent Models

Less powerful transparent models may not be sufficiently expressive and flexible to pick up nonlinear relationships. For categorical features, GA2M is equivalent to linear and logistic regression. However, for continuous features, a key advantage of GA2Ms is the ability to model nonlinear relationships.

We compare the GAM estimated effects to that of linear and logistic regression. However, for continuous features, the difference between GA2Ms and logistic regression is significant. Consider Figure 8, which is the equivalent of the bottom row of Figure 4, but using linear and logistic regression as transparent models instead of GA2Ms. Where the GA2M model was able to shape the age feature in a non-linear manner across the entire age range, and detected significant differences between the student risk score model and outcome model for young and old age groups, logistic regression ascribes only one number to the effect of the age feature on the predicted outcome. This resulted in no significant difference detected (because the bar graph for the first column, age, is at at y=0 in Figure 8. This example demonstrates the value of GA2Ms compared to simpler models such as linear and logistic regression. Some of the benefit of GA2Ms arises from their ability to model

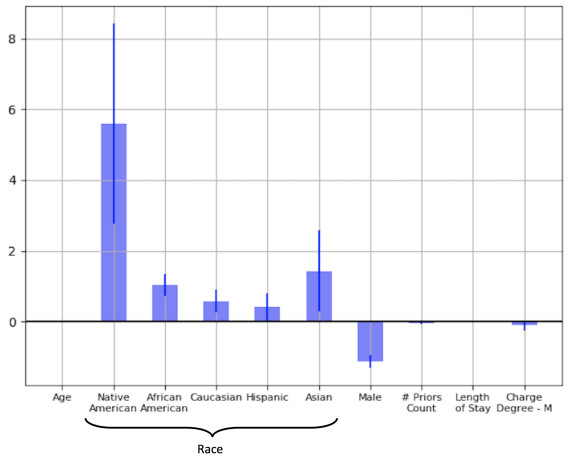

Figure 8: Difference in feature contributions between student risk score model and actual outcome model, when using linear and logistic regression.

pairwise interactions, but much of the benefit occurs when the data has many continuous features that GA2M is able to shape in interesting ways that cannot be represented by linear models.

5 CONCLUSION

We propose a method to audit black-box risk models for potential bias 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 model 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 four public data sets. The key advantages of this approach are that the GA2M transparent models we use are very accurate despite being interpretable, the method generates reliable confidence intervals to aid interpretation, we are able to undo distortions that may be present in the black-box model's predictions, we can often detect which features are and are not used by the black-box model, and one does not need to know in advance what biases to look for.

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] 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. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed May 26, 2017.

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

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

[6] Thomas Blomberg, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. 2010. Validation of the COMPAS risk assessment classification instrument. *College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL* (2010).

[7] Leo Breiman. 2001. Random Forests. *Machine Learning* (2001).

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

[9] 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*.

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

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

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

[13] Decio Coviello and Nicola Persico. 2015. An Economic Analysis of Black-White Disparities in the New York Police Department's Stop-and-Frisk Program. *The Journal of Legal Studies* (2015).

[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] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232.

[20] Andrew Gelman, Jeffrey Fagan, and Alex Kiss. 2007. An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias. *J. Amer. Statist. Assoc.* (2007).

[21] Robert D Gibbons, Giles Hooker, Matthew D Finkelman, David J Weiss, Paul A Pilkonis, Ellen Frank, Tara Moore, and David J Kupfer. 2013. The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): a screening tool for depression. *Journal of Clinical Psychiatry* (2013).

[22] 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).

[23] C. Gu. 2003. *Smoothing Spline ANOVA Models*. Springer, New York.

[24] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.

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

[26] 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).

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

[28] Giles Hooker. 2007. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics* 16, 3 (2007), 709–732.

[29] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. 2014. An Efficient Approach for Assessing Hyperparameter Importance. In *ICML*.

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

[31] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. In *ICML Workshop on Human Interpretability of Machine Learning*.

[32] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. 2017. Does mitigating ML's disparate impact require disparate treatment? *arXiv preprint arXiv:1711.07076* (2017).

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

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

[35] 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).

[36] P. McCullagh and John A. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall/CRC.

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

[38] Avi Feller Sam Corbett-Davies, Emma Pierson and Sharad Goel. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. (2016).

https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas Accessed May 26, 2017.

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

[40] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* (2007).

[41] 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*.

[42] 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).

[43] Hao Wang, Berk Ustun, and Flavio P. Calmon. 2018. On the Direction of Discrimination: An Information-Theoretic Analysis of Disparate Impact in Machine Learning. *arXiv preprint arXiv:1801.05398* (2018).

[44] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *ICWWW*.

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

APPENDIX

Additional Features for COMPAS

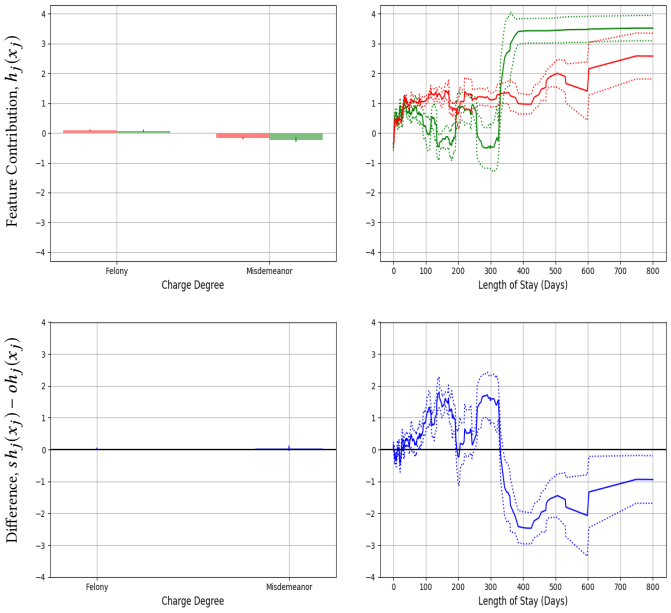

Figure 9: GA2M shaped feature contributions for remaining features to the COMPAS risk score student model (in red) and the actual recidivism outcome model (in green). The blue line captures the difference between the two models (score student - outcome).

Pairwise Interactions for COMPAS

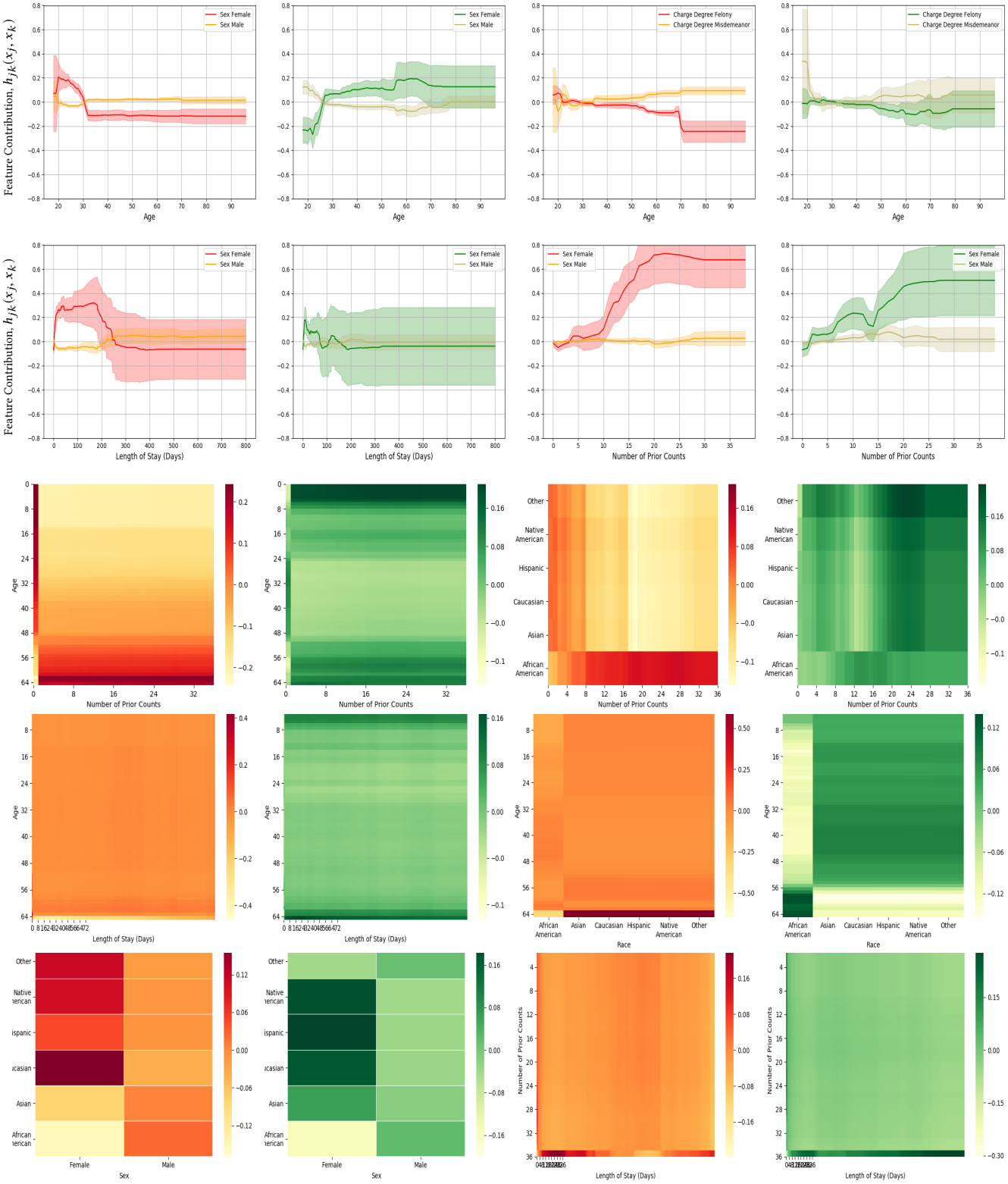

Figure 10: GA2M shaped pairwise feature contributions of top ten pairs to the COMPAS risk score student model (in red) and the actual recidivism outcome model (in green). The pairs in the 1st two rows consist of a continuous feature and a binary categorical feature. The rest of the pairs capture interactions of two continuous features, or categorical features with many levels.

Additional Features for Lending Club

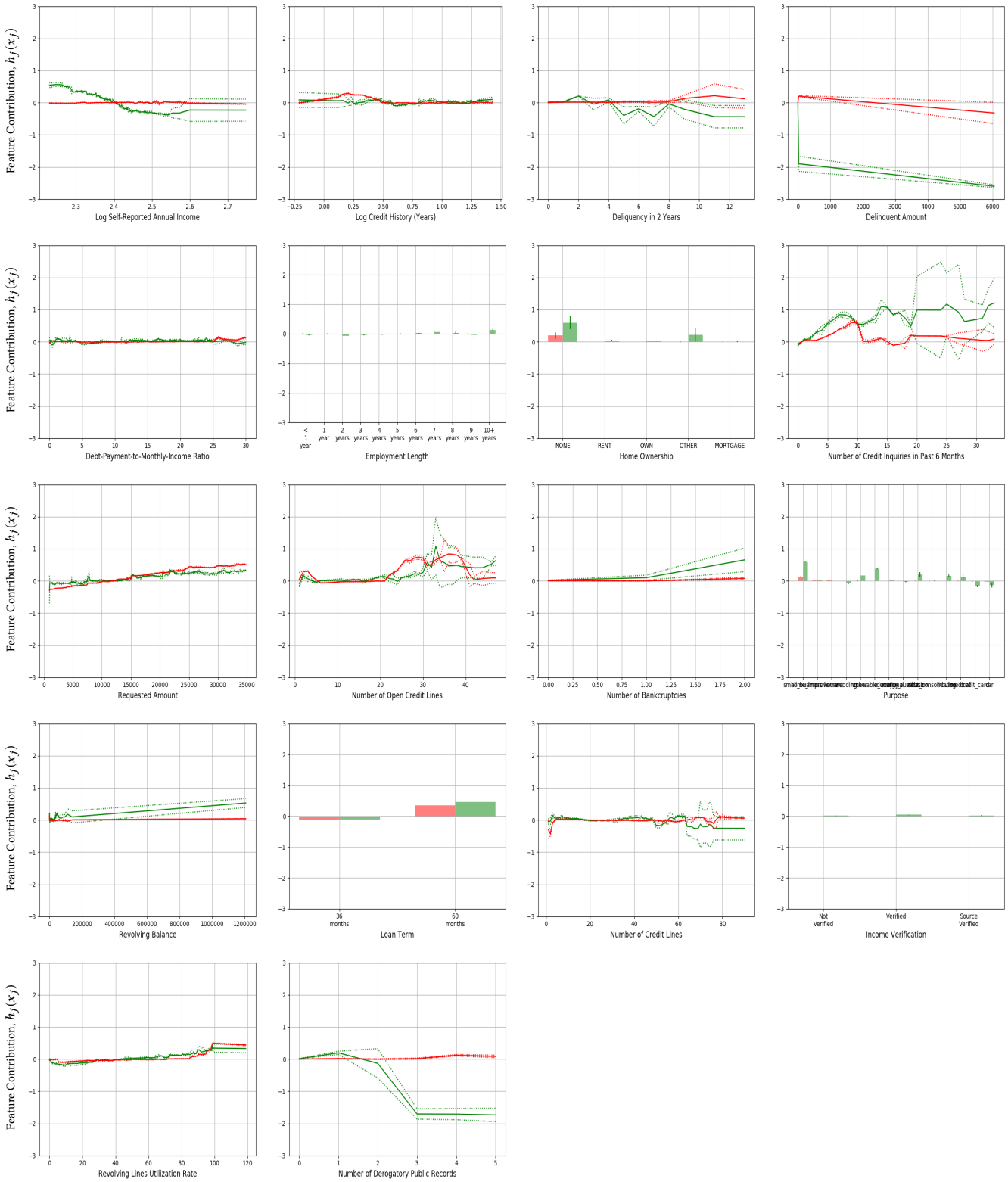

Figure 11: GA2M shaped feature contributions for remaining features to the Lending Club risk score student model (in red) and the actual loan default outcome model (in green).

Additional Pairwise Interactions for Lending Club

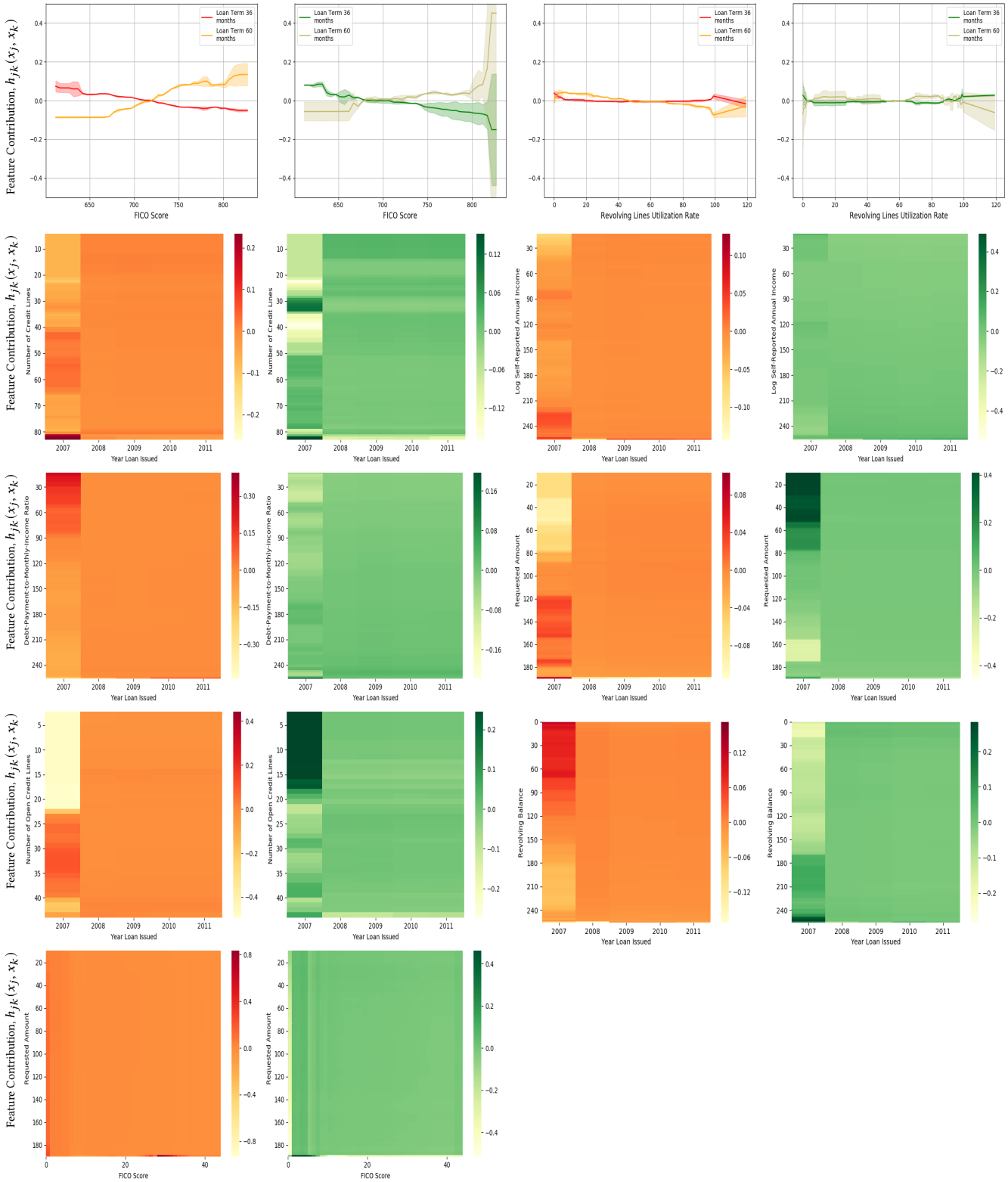

Figure 12: GA2M shaped pairwise feature contributions of remaining pairs to the Lending Club risk score student model (in red) and the actual loan default outcome model (in green). The two pairs in the top row consist of a continuous feature and a binary categorical feature. The rest of the pairs capture interactions of two continuous features, or categorical features with many levels.