Interpretable Active Learning

Richard L. Phillips[1] **Kyu Hyun Chang**[2] **Sorelle Friedler**[1]

Abstract

Active learning has long been a topic of study in machine learning. However, as increasingly complex and opaque models have become standard practice, the process of active learning, too, has become more opaque. There has been little investigation into interpreting what specific trends and patterns an active learning strategy may be exploring. This work expands on the Local Interpretable Model-agnostic Explanations framework (LIME) to provide explanations for active learning recommendations. We demonstrate how LIME can be used to generate locally faithful explanations for an active learning strategy, and how these explanations can be used to understand how different models and datasets explore a problem space over time. We propose a measure for uncertainty bias based on disparate impact that allows further exploration of the relative exploitation of different data subgroups. We combine the LIME framework with the uncertainty bias metric to demonstrate how clusters of unlabeled points can be made automatically based on common sources of uncertainty. We show that this allows for an interpretable explanation of what an active learning algorithm is learning as points with similar sources of uncertainty have their uncertainty bias resolved.

1. Introduction

The importance of interpretability and explainability of machine-learned decisions has recently been an area of active interest, with the EU even declaring what has been called a “right to an explanation” (Goodman & Flaxman, 2016). Recent work on interpretability has included both local explanations about an individual’s decision (Ribeiro et al., 2016) and global explanations about the model’s actions overall, and has included interpretable techniques in clustering (Chen et al., 2016), integer programming (Zeng et al., 2016), rule lists (Wang & Rudin, 2015), and methods for understanding deep nets (Zeiler & Fergus, 2014; Le et al., 2011) in addition to historical work on decision trees (Quinlan, 1993) and random forests (Breiman, 2001). In these traditional machine learning contexts, the focus of interpretability has been two-fold, first on the receiver of the decision (“why was I rejected for this job?”) and second on the model creator (“why is my model giving these answers?”).

Here, we extend this interest in interpretability to active learning, a domain in which the explanation is additionally of interest to the labeler (”why am I being asked these questions and why is it worth it to answer?”). Since active learning is generally applied in scenarios such as drug discovery where it is expensive (whether in terms of time or money) to label a query, the labeler in these contexts is often a domain expert in their own right (e.g., a chemist). Given this, a query explanation can serve as a way to both justify an expensive request and allow the domain expert to give feedback to the model. We demonstrate how active learning choices can be made more interpretable to non-experts and show that expert-driven learning performs at least as well as traditional active learning strategies on several simulated and real datasets.

1.1. Results

We demonstrate how active learning choices can be made more interpretable to non-experts. Using per-query explanations of uncertainty, we develop a system that allows experts to choose whether to label a query. This allows experts to incorporate domain knowledge and their own interests into the labeling process. For example, in the case of a chemist’s knowledge of a chemical system, this might allow a model to focus on the reactions of interest to the chemist, the ones for which reagents are already purchased, or even take advantage of the chemist’s existing knowl-

[1]Haverford College, Haverford, PA [2]Google Inc., Mountain View, CA. Correspondence to: Sorelle Friedler <sorelle@cs.haverford.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

edge to learn targeted information faster. Indeed, we demonstrate the potential for such expert-driven active learning systems to outperform traditional active learning strategies.

In addition, we introduce a quantified notion of *uncertainty bias*, the idea that an algorithm may be less certain about its decisions on some data clusters than others. In the context of decision-making about people, this may mean that some protected groups (e.g., races or genders) may receive less favorable decisions due to risk aversion (Goodman & Flaxman, 2016). In the context of active learning, this means that these groups are more likely to be targeted for exploratory queries in order to improve the model. We combine this idea with the explanations generated per query to describe the groups most targeted by uncertainty bias.

2. Related Work

Active Learning. Active learning has a long history that is detailed in this comprehensive survey (Settles, 2009). Our work will focus on explaining query uncertainty. Uncertainty querying for active learning was first proposed in 1994 by Lewis and Gale (Lewis & Gale, 1994). Since then, it has become perhaps the most common strategy for active learning and several strategies for quantifying uncertainty have been developed (Settles, 2009). Strategies used to quantify uncertainty for actively learning multi-class classification problems include selecting the sample with the minimum maximum-class probability, selecting the sample with the minimum difference in probabilities between the two most probable classes, and choosing the sample with maximal label entropy. All three of the above strategies are equivalent for binary classification tasks, such as the tasks we focus on in this paper (Settles, 2009).

3. Local Interpretable Model-Agnostic Explanations

We will build specifically on a method for creating local explanations introduced in (Ribeiro et al., 2016). Local Interpretable Model-Agnostic Explanations (LIME) is an algorithm for offering prediction explanations for individual predictions. This works well on even very complex models by training an interpretable model on the local space around a prediction. This local approximation is useful for creating annotations of factors that are influential in a model's prediction. For a given predicted instance, LIME generates a perturbed sample set in the neighborhood of the instance. Then, based on the sample and the model predictions, LIME searches for the most interpretable model and derives an explanation.

4. Explaining Active Learning Queries

The goal of this work is to explain, beyond the attributes and specific data points queried, a strategy to understand what uncertainty an active learning algorithm is attempting to resolve and to determine whether any subgroups need to be monitored during an active learning run.

Toy example. An example multi-class classification problem is used to explore explanations on uncertainty. Four Gaussian distributions with unit variance are centered at $(-3, -3)$, $(3, -3)$, $(3, 3)$, and $(-3, 3)$. The Gaussians are assigned labels such that the first two represent one class and the second another. Initially, 50 points are randomly selected from the Gaussian at $(-3, -3)$ and $(3, 3)$ to be labeled. The points have been purposefully drawn in such a way as to label none of the points Gaussians centered in the second and fourth quadrants. An initial logistic regression model, W, is trained on the initial 50 labeled points. Based on the resulting model of the probability distribution, the certainty scores across the problem space are mapped. The labeled points, decision boundary, and certainty scores can be seen in Figure 1.

Using LIME (Ribeiro et al., 2016), we can ask for locally-faithful weighted explanations of the certainty values provided by W. We will refer to combinations of LIME explanations of uncertainty as "uncertainty regions." These regions can be useful for grouping points together based on identical sources of uncertainty. In these cases, we would expect points explored in a given uncertainty region to increase the certainty we have about points with the same sources of uncertainty.

5. Identifying Uncertainty Bias

In situations where some instance populations are smaller (minority groups) or where the initial training data distribution is skewed, the active learner may prefer queries that are disproportionally drawn from a single region (or population group). For example, in our toy example above, we saw that upper left quadrant is underrepresented in the labeled dataset. The points in this region have higher uncertainty and were more likely to be targeted for active learning queries. In order to understand both what and how an active learning method is learning and whether there is a disparate impact among groups targeted to be labeled, it

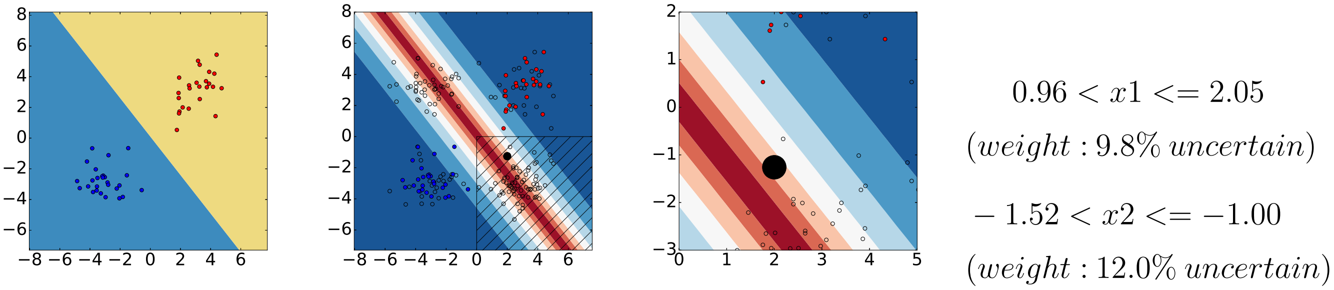

Figure 1. Left: Labeled starting pool. Center: Certainty over the problem space. The black points represent unlabeled points in the pool. Blue represents regions of certainty and red represents regions that are uncertain. Pictured in the hatched quadrant is the 'explanatory region' that contains the target point $(1.88, -1.26)$. Right: A local view of the target point and the LIME explanation for the model's prediction of 80% uncertainty for the black point at $(1.88, -1.26)$.

is important to identify this *uncertainty bias*.

Definition 5.1 (Uncertainty bias)**.** *Given a dataset $D = (\mathbb{X}, U)$ with d-dimensional feature vector $\mathbb{X}$ and corresponding (discrete) uncertainty labels U and its disjoint set R of uncertainty regions (groups), let $X_r = \{x \in \mathbb{X} | x \in r, r \in R\}$ be the items of the data set within R. U takes values $+$ (certain) and $-$ (uncertain). The* uncertainty bias *with respect to region $r \in R$ is defined to be:*

$$1 - \frac{Pr(U = + | x \in r)}{Pr(U = + | x \in R \setminus r)}$$

Note that this is the same as $1 - DI$ where DI is the disparate impact value (Feldman et al., 2015) applied where the region of focus is the protected class and the positive value is a label of $+$. For the purposes of this work, we consider any point with certainty greater than or equal to the median over our pool to be certain $(U = +)$.

5.1. Clustering Over Uncertainty Labels to Increase Interpretability

It will not always be practical to manually define explanatory regions to observe when doing active learning. To automatically create groups for tracking the principle patterns explored during active learning, k-means clustering is used to cluster the samples' explanations and weights. We use each label as its own dimension where the value for each point is the weight of that label on the point's uncertainty, or 0 if the point falls outside of the constraint. This means that for a set of possible uncertainty labels U, all of our original data points have an equivalent point in our uncertainty space $\mathbb{R}^{|U|}$. The objective of k-means (using Lloyd's algorithm) is thus to minimize the pairwise squared deviations for all of the points in each cluster: $\sum_i^k \sum_{d \in U} \sum_{x,y \in C_i} \|x_d - y_d\|^2$

Each cluster centroid is then used to keep track of the principle sources of uncertainty for that cluster. The number of clusters, k is chosen by trying a wide range of potential values and finding the value that maximizes the proportion of points that share their top uncertainty constraints with their respective cluster centroids. As this ratio will likely continue to trend upwards as k grows, k is simply increased until adding another k will not improve this proportion over some small threshold. It is possible to largely capture all of the uncertainty labels for a pool within a relatively small number of clusters, greatly simplifying the task of tracking what regions of uncertainty are explored.

5.2. Experiments: Identifying Uncertainty Bias

In addition to our toy data set outlined above, we consider uncertainty bias under the explanatory active learning framework described above on the ProPublica dataset for recidivism prediction (Angwin et al., 2016) as well as the Dark Reaction Project dataset of chemical reactions for synthesis prediction (Raccuglia et al., 2016).

LIME Setup. All of the experiments in the this work use LIME as described in (Ribeiro et al., 2016) to explain continuous (regressor) predictions.[1,2] We apply this to our active learning selection criterion. Ridge regression is the local 'interpretable' model to estimate feature importance. For the experiments in this work, our active learning criterion is the max class probability. Continuous features were split into at most 8 bins by greedily maximizing information gain to make the splits. To track our active learning strate-

[1] https://github.com/marcotcr/lime/

[2] https://github.com/datascienceinc/lime

gies on the various datasets, we simply recorded every time a point was sampled for each cluster. Similarly, we track the uncertainty bias across every cluster each time a point was labeled and the model trained on the now-expanded data.

Toy Data Set. The results of the toy data set are consistent with our understanding of active learning. The underexplored regions, Quadrant 2 and Quadrant 4, begin with a high uncertainty bias and are quickly emphasized by the active learning algorithm. The general classifier boundary quickly approaches the correct one, and the uncertainty biases begin to even out. The learning progress can be seen in Figure 2.

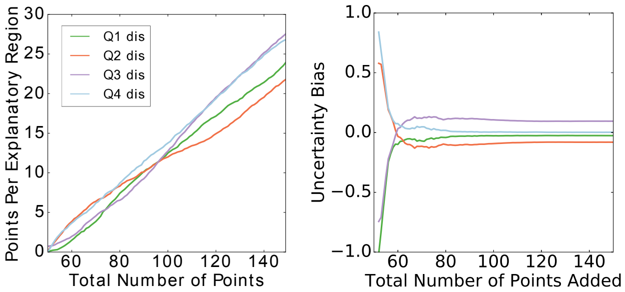

Figure 2. Left: Average counts of points taken from each quadrant during 50 active learning runs on the toy data set. Right: Uncertainty bias per quadrant over 50 active learning runs. As we would expect, the under-explored quadrants, Quadrant 2 and Quadrant 4, start off with high uncertainty bias that is gradually resolved. This corresponds to the the exploration of Quadrant 2 represented by the slope of the orange curve between 50 and 90 training points in the left graph.

ProPublica Recidivism and Race. The ProPublica dataset includes attributes describing the sex, age, race, juvenile felony and misdemeanor counts, number of adult priors, charge degree (felony or misdemeanor), and charge description for 6172 people arrested in Broward County, Florida, along with a boolean value indicating whether they were rearrested within two years of the original arrest date. For this experiment, the goal was to see if sensible explanations could be made about what the active learning algorithm is exploring. We used a logistic regression model trained on an initial pool of 400 randomly selected points as our starting point. Each data point was given two uncertainty labels and the uncertainty labels and weights were clustered with $k = 40$ clusters. Eighty-three percent of the points in the pool were in clusters with centroids that matched their own uncertainty constraints and 100% shared at least one uncertainty label with their cluster's centroid.

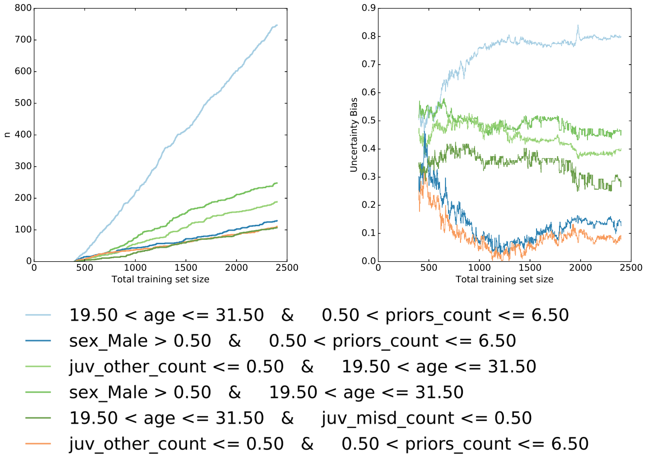

Figure 3. The most frequently queried uncertainty clusters for the ProPublica 2-year recidivism dataset. The first two clusters, covering middle aged people and men with a few priors seem like reasonably uncertain classes. It seems reasonable that a high number of priors or no priors might make the chance of recidivism more certain and that the age range and gender are likely very common in the dataset and naturally have high variance. Five of the six clusters all display overall downward trends.

We can see clearly that, of the top most explored clusters, uncertainty bias trended downwards on five of them. This demonstrates that our active learning strategy is successfully exploring the regions of greatest uncertainty. It also provides support for the validity of our explanatory labels, as the uncertainty labels for points that are frequently queried truly correlate with the resolved uncertainty.

Moving forward, we wanted to see how uncertainty bias might be dependent on race and how this might be affected by the active learning process. To test this we tested the new model for uncertainty bias based on race with each point added to our pool. The results can be found in Figure 4. It is evident that, from the very beginning, there is a notable disparity in our model's ability to make confident predictions between the different racial categories. While active learning seems to resolve most of the difference in uncertainty bias between the people labeled 'White' and 'Hispanic,' Black people arrested in Broward County were subject to considerable uncertainty bias by our logistic regression model even after 2000 more points were actively queried.

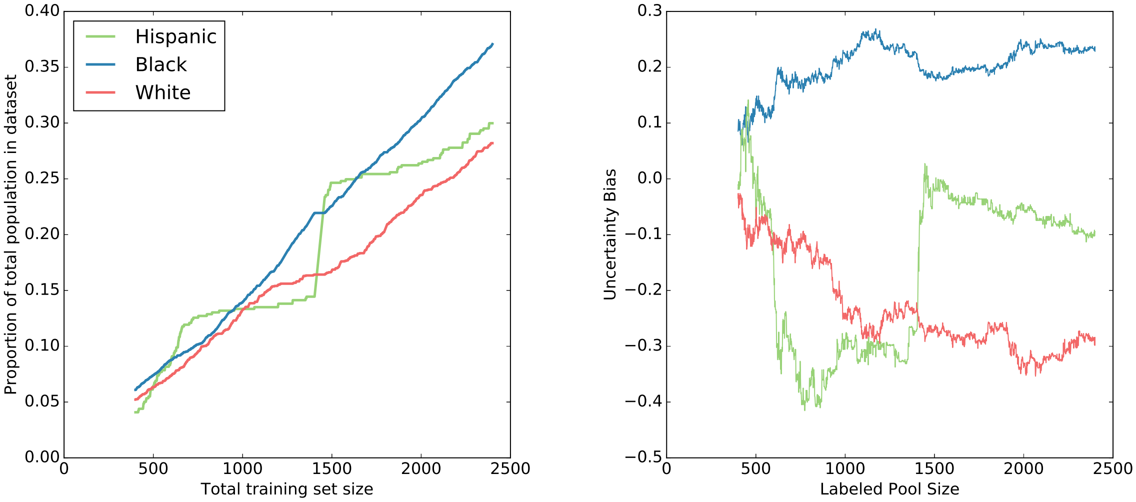

Figure 4. Uncertainty bias based on racial labels found in the dataset. It is useful when active learning resolves uncertainty bias, but this run demonstrates why this cannot be assumed. Near-margin active learning does not reduce the uncertainty bias conditioned by racial labels for the recidivism dataset. In this situation, further steps ought to be taken to resolve this bias.

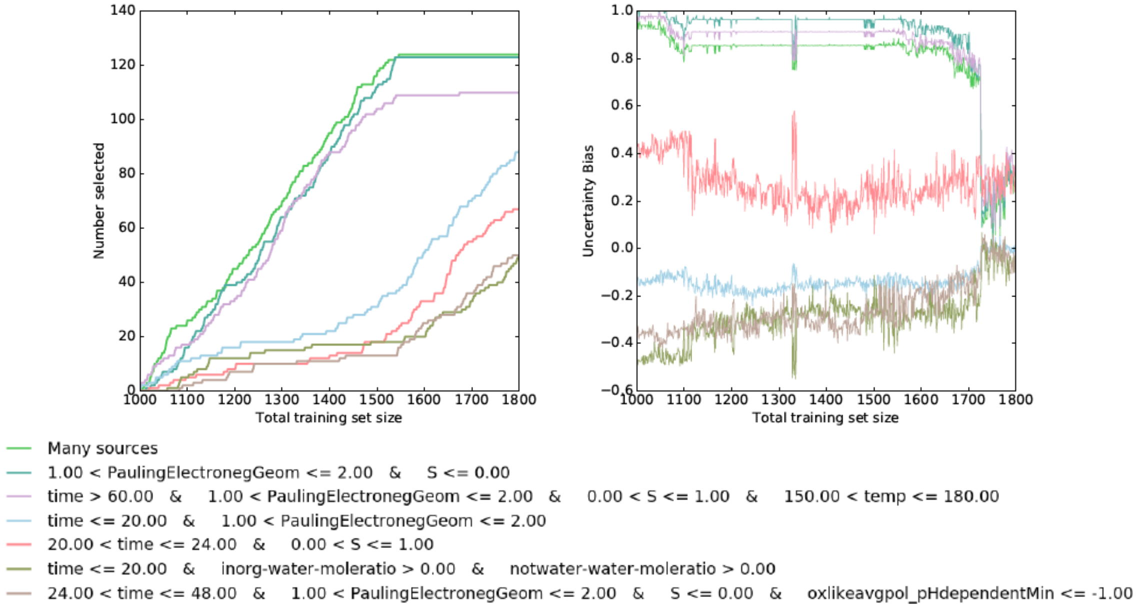

Figure 5. Counts and uncertainty biases for the most frequently sampled clusters from the Dark Reactions dataset. It is interesting to observe how the first three clusters are explored thoroughly together, and then the other four clusters begin to dominate.

Dark Reactions Dataset. The Dark Reaction Project dataset includes 6114 hydrothermal synthesis reactions and 274 attributes describing chemical properties that might predict the associated boolean classification indicating whether the experiment successfully created a crystalline product or not. To predict this outcome, we used AdaBoost with 200 decision stump weak learners. The certainty function thus estimates the probability of each class through a weighted average of the fraction of training samples within each leaf of the decision stumps.

The Dark Reaction dataset was a challenge because there are many more features than the the ProPublica dataset. We started by considering $k = 7$ clusters and the associated uncertainty bias values (see Figure 5). We find a large gap in uncertainty between the top three clusters and the rest, and the most uncertain cluster also has an explanation indicating this increased uncertainty. To produce more specific class labels, 10 weighted explanatory labels were generated to explain the certainty of each point. As there are many more attributes used for each explanation in the DRP dataset, the labels listed for each cluster indicate that a given attribute implies more than 2% more or less certainty for points in that cluster, on average. 'Many sources' is the leading cluster, which refers to a cluster with no primary source of uncertainty above this threshold. Given that there are 274 attributes in the DRP dataset, it is notable that most of the curves do have prominent sources of uncertainty. By allowing a domain expert to control the cutoff parameter and the number of explanations to use, we can adjust towards more precise explanations of uncertainty.

6. Conclusion

This work has demonstrated a straightforward application of LIME to explain single active learning queries. We also define a quantitative measure of uncertainty bias. With these tools in hand we first demonstrate how we can track the exploration of groups of points with common uncertainty and confirm that that uncertainty is being resolved with the uncertainty bias measure. We then demonstrate on more complex, real-world datasets how regions of uncertainty can be generated automatically to create meaningful groups to track during learning. We hope to draw attention to the problem of uncertainty bias, to highlight the lack of transparency research in active learning, and to encourage active learning usage to be interpretable and accountable.

References

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. Machine bias. *ProPublica*, May 23, 2016.

Breiman, Leo. Random forests. *Machine learning*, 45 (1):5–32, 2001.

Chen, Junxiang, Chang, Yale, Hobbs, Brian, Castaldi, Peter, Cho, Michael, Silverman, Edwin, and Dy, Jennifer. Interpretable clustering via discriminative rectangle mixture model. In *IEEE 16th International Conference on Data Mining (ICDM)*, 2016.

Feldman, Michael, Friedler, Sorelle A., Moeller, John, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Certifying and removing disparate impact. *Proc. 21st ACM KDD*, pp. 259–268, 2015.

Goodman, Bryce and Flaxman, Seth. European union regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

Le, Quoc V., Ranzato, Marc'Aurelio, Monga, Rajat, Devin, Matthieu, Chen, Kai, Corrado, Greg S., Dean, Jeff, and Ng, Andrew Y. Building high-level features using large scale unsupervised learning. In *Proc. ICML*, 2011.

Lewis, David D. and Gale, William A. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pp. 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. URL `http://dl.acm.org/citation.cfm?id=188490.188495`.

Quinlan, Ross. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Raccuglia, Paul, Elbert, Katherine C., Adler, Philip D. F., Falk, Casey, Wenny, Malia B., Mollo, Aurelio, Zeller, Matthias, Friedler, Sorelle A., Schrier, Joshua, and Norquist, Alexander J. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533:73–76, May 5, 2016.

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. ACM KDD*, 2016.

Settles, Burr. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Wang, Fulton and Rudin, Cynthia. Falling rule lists. In *AISTATS*, 2015.

Zeiler, Matthew D. and Fergus, Rob. Visualizing and understanding convolutional networks. In *Computer Vision — ECCV 2014*, pp. 818–833. Springer, 2014.

Zeng, Jiaming, Ustun, Berk, and Rudin., Cynthia. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society*, Sept. 2016.