"Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations

Yujia Zhang[*1] **Kuangyan Song**[*2] **Yiming Sun**[*1] **Sarah Tan**[1] **Madeleine Udell**[1]

Abstract

Methods for explaining black-box machine learning models aim to increase the transparency of these model and provide insights into the reliability and fairness of such models. However, the explanations themselves could contain significant uncertainty that undermines users' trust in the predictions and raises concern about the model's robustness. Focusing on a particular local explanations method, Local Interpretable Model-Agnostic Explanations (LIME), we demonstrate the presence of three sources of uncertainty, namely randomness in the sampling procedure, variation with sampling proximity, and variation in explained model credibility across different data points. Such uncertainty is present even for black-box models with high test accuracy. We investigate the uncertainty in the LIME method on synthetic data and two public data sets, newsgroups text classification and recidivism risk-scoring.

1. Introduction

While machine learning models have become increasingly important for decision making in many areas (Zeng et al., 2017; Rajkomar et al., 2018), many machine learning models are "black-box" in that the process by which such models make predictions can be hard for humans to understand. Explanations of model predictions can help increase users' trust in the model (Lipton, 2016; Ribeiro et al., 2016), determine if the model achieves desirable properties such as fairness, privacy, etc. (Doshi-Velez & Kim, 2017), and debug possible errors in the model (Ribeiro et al., 2018b).

Indeed, explanation methods aim to help users assess and establish trust in black-box models and their predictions. However, whether the explanations themselves are trustworthy is not obvious. Uncertainty in explanations not only cast doubt on the understanding of a certain prediction, but also raises concerns about the reliability of the black-box model in the first place, hence diminishing the value of the explanation (Ghorbani et al., 2019).

In this paper, we address the question: when can we trust an explanation? In particular, we study the local explanation method Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016). Briefly, LIME explains the prediction of a desired input by sampling its neighboring inputs and learning a sparse linear model based on the predictions of these neighbors; features with large coefficients in the linear model are then considered to be important for that input's prediction. We demonstrate that training LIME explanations involve sources of uncertainty that should not be overlooked. More specifically, generating a local explanation for an input requires sampling around the input to generate an explanation for its prediction. In this paper, we show that this sampling can lead to statistical uncertainty in interpretation.

2. Related Work

The study of interpretable methods can be roughly divided into two fields – designing accurate, yet still inherently interpretable models (Letham et al., 2015; Lakkaraju et al., 2016), and creating post-hoc methods to explain black-box models, either locally around a specific input (Baehrens et al., 2010; Ribeiro et al., 2016) or globally for the entire model (Ribeiro et al., 2018a; Tan et al., 2018a). In this paper, we study one particular local explanation method, LIME (Ribeiro et al., 2016).

Several sensitivity-based explanation methods for neural networks (Shrikumar et al., 2017; Selvaraju et al., 2017; Sundararajan et al., 2017) have been shown to be fragile (Ghorbani et al., 2019; Adebayo et al., 2018). Ghorbani et al. demonstrated that it is possible to generate vastly different explanations for two perceptively indistinguishable inputs with the same predicted labels from the neural network. This paper focuses on the fragility of local post-hoc explanations of models.

[*]Equal contribution [1]Cornell University [2]Zshield Inc. Correspondence to: Yujia Zhang <yz685@cornell.edu>, Kuangyan Song <nachtsky617@163.com>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

Potential issues with LIME's stability and robustness have been pointed out by Alvarez-Melis & Jaakkola, who showed that while LIME explanations can be stable when explaining linear models, for nonlinear models this is not always the case (Alvarez-Melis & Jaakkola, 2018). Testing LIME on images, Lee et al. observed that LIME colored superpixels differently across different iterations and proposed an aggregated visualization to reduce the perception of different explanations over different iterations (Lee et al., 2019). However, they did not study the source of this instability of explanations – the focus of our paper.

3. Approach

3.1. Uncertainty in LIME Explanations

Given a black box model f, and a target point x to be explained, LIME samples neighbors of x and their black-box outcomes and chooses a model g from some interpretable functional space G by solving

$$\text{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

where π_x is some probability distribution around x and $\Omega(g)$ is a penalty for model complexity. Ribeiro et al. (Ribeiro et al., 2016) suggests several methods to achieve sparse solution, including K-LASSO as the interpretable model. For K-LASSO, we let $\Omega = \infty \mathbb{1}[\|w_g\|_0 > K]$, where w denotes the coefficients of the linear model, and sample points near x from π_x to train K-LASSO. We observe that this procedure involves three sources of uncertainty:

- Sampling variance in explaining a single data point;
- Sensitivity to choice of parameters, such as sample size and sampling proximity;
- Variation in explanation on model credibility across different data points.

3.2. Methodology

We use one synthetic data example and two real datasets to demonstrate the three aforementioned sources of uncertainty. To show the sampling variance, we run LIME multiple times for a single data point, record the top few features selected by K-LASSO each time, and observe the cumulative selection probability for each selected feature. Whether features are consistently selected over different trials reflects LIME's instability in explaining the data point. Then, we tune the parameters of LIME to probe the sensitivity of the explanations to sample size and sampling proximity. Finally, we compare LIME explanations of different data points by assessing whether the selected features are informative in the real context. Variation in explanation on model credibility across different data points raises concern about the credibility of LIME as a global explanation for the model.

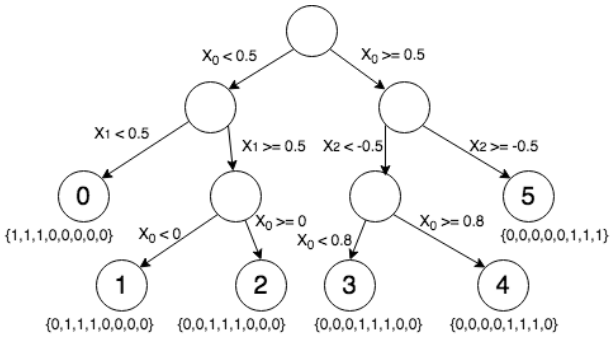


Figure 1: Decision tree partition of eight-feature synthetic data. The local coefficients on each leaf are shown under each end node. Data points are assigned labels based on the linear classifier in Equation 2.

4. Results

We first use synthetic tree-generated data to illustrate the first and second source of uncertainty mentioned above. Then we use examples in text classification to demonstrate the third source of uncertainty. Finally we apply LIME to the COMPAS dataset as a case where LIME explanations are considered trustworthy.

4.1. Synthetic data generated by trees

Data: Given the number of features N, we generate training and test data from local sparse linear models on uniformly distributed input in $[0, 1]^N$. To illustrate LIME's local behavior at different data points, we partition them with a known decision tree. Within each partition, we assign labels on each data point $\mathbf{x}$ based on a linear classifier with known coefficients $\boldsymbol{\beta}$ as shown in Equation 2.

$$y(\mathbf{x}) = \begin{cases} 1 & \mathbf{x}^\top \boldsymbol{\beta} \geq 0 \\ 0 & \mathbf{x}^\top \boldsymbol{\beta} < 0. \end{cases} \tag{2}$$

We consider two cases where the number of features is 4 and 8 respectively. Figure 1 presents a way of splitting the data into six leaves for $N = 8$ with known coefficients, where three out of eight features have coefficients 1 in each leaf. The data splitting and coefficients for $N = 4$ are presented in Figure 4 in Appendix A.

Results: We present results for the case where we apply LIME to interpret black-box models (random forest and gradient boosting tree) trained with eight-feature synthetic data. We run LIME on one data point in each of the six leaves. We first notice that different trials potentially select different features due to sampling variance. Figure 2 shows the cumulative selection frequency of top three fea-

tures in each trial when LIME interprets a random forest model; the case with gradient boosting is shown in Figure 5 in Appendix A. LIME captures the signal of the first three features, which are used globally in the tree splitting of the data. Locally, however, different features are important for each individual leaf, which LIME fails to reflect. Thus, its explanation cannot be considered stable around each input data point in a tree structure. We further notice that LIME by default draws samples from a rescaled standard normal distribution $\mathcal{N}(0, \sigma^2)$ near the test point, where σ^2, the variance of the training data, determines the sampling proximity. The experiments show that LIME tends to capture locally important features better with a smaller sampling proximity and pick up global features with a larger sampling proximity. Since tuning this parameter allows LIME to explore both global and local structure in the data, we suggest users to think consciously about the choice of its value. As an example, we tune LIME's sampling proximity for a data point on leaf 5 in the eight-feature synthetic data, shown in Figure 2f. When a sample is drawn from $\mathcal{N}(0, \sigma^2)$ near the test point, LIME captures the global features used for tree splitting; when a sample is drawn from $\mathcal{N}(0, (0.1\sigma)^2)$, LIME successfully picks up signal from the three local features 5-7. Results for running the same procedure on four-feature synthetic data are presented in Figures 6 and 7 in Appendix A.

4.2. Text Classification

Data: The 20 Newsgroup dataset is a collection of ca. 20,000 news documents across 20 newsgroups. As noted in (Ribeiro et al., 2016), even for text classification models with high test accuracy, some feature words that LIME selects are quite arbitrary and uninformative. To examine this behavior further, we use Multinomial Naive Bayes classifier for two examples of document classification, namely "Atheism vs. Christianity" and "electronics vs. crypt".

Results: Multinomial Naive Bayes classifiers are trained for the aforementioned two classification examples, with test accuracy 0.9066 and 0.9214 respectively. However, as pointed out in (Ribeiro et al., 2016), we need to know the feature importance for each output in order to establish trust in the model. In particular, we find that LIME's local explanations are not always plausible for different test documents. As shown in Figure 3, the selected feature words for the first document ("crypto", "sternlight" and "netcom") display no variation for different trials and are relevant in content, which makes the model seem very credible. However, the selected feature words for the second document are not informative at all. Thus, the model's credibility, as explained by LIME, varies across different input data. We also include results for "Christianity vs. Atheism" in Figure 8 in Appendix A, which also display a difference in model credibility for different documents.

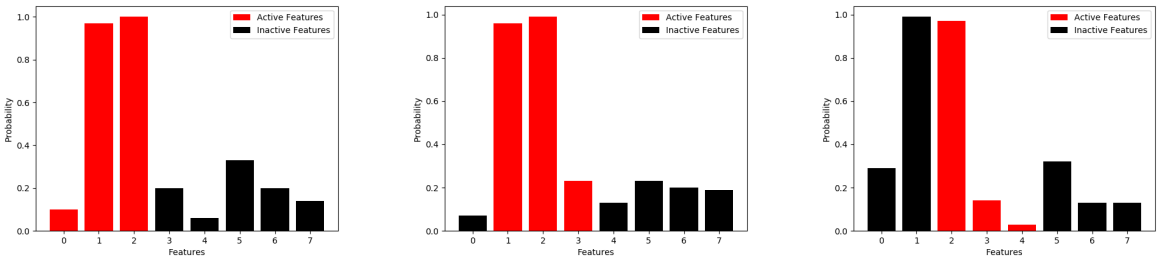
4.3. COMPAS Recidivism Risk Score Dataset

Data: The "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS) is a risk-scoring algorithm developed by Northpointe to assess a criminal defendant's likelihood to recidivate. The risk is classified as "High", "Medium" and "Low" based on crime history and category, jail time, age, demographics, etc. We study a subset of the COMPAS dataset collected and processed by ProPublica (Larson et al., 2016), with the goal of examining the presence of demographic bias in risk-scoring. As we do not have access to the true COMPAS model, we train a random forest classifier as a "mimic model" (Tan et al., 2018b), using selected features and risk assessment text labels from COMPAS. We examine salient features selected by LIME explanations on multiple COMPAS records.

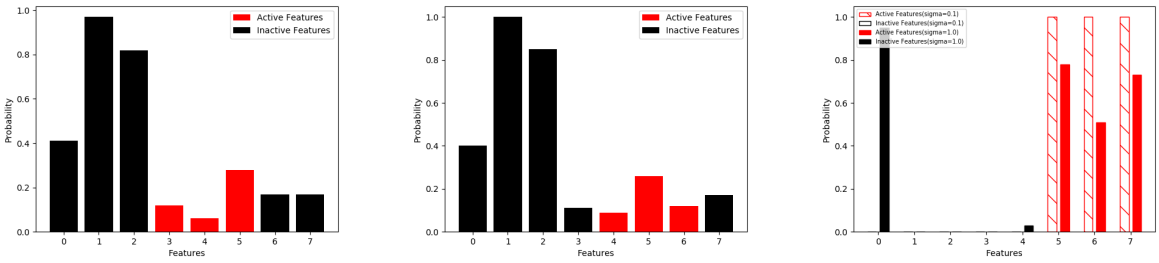
Results: We test and analyze LIME explanation of the random forest classifier with both numerical and categorical features. Unlike the uncertainty we observe in previous experiments on synthetic and 20 Newsgroup data, we see consistent explanation results on different test data points. LIME is applied to two data points that are classified as "high risk" by COMPAS. The results are shown in Figure 9 in Appendix A. We consider these explanations to be trustworthy due to the following two observations: 1) there is little variation in the selection of important features in different trials on the same data point, and 2) explanation is consistent for different data points, since the same features are selected for the two different data points, including race and age. Further analysis using LIME suggests that the mimic model is using demographic properties as important features in predicting a risk score. This in turn shows it is probable that the COMPAS model makes use of demographic features for recidivism risk assessment, so further investigation would be meaningful to gauge the fairness of the algorithm.

5. Conclusion

Explanation methods for black-box models may themselves contain uncertainty that calls into question the reliability of the black-box predictions and the models themselves. We demonstrate the presence of three sources of uncertainty in the explanation method "Local Interpretable Model-agnostic Explanations" (LIME), namely the randomness in its sampling procedure, variation with sampling proximity, and variation in explained model credibility for different data points. The uncertainty in LIME is illustrated by numerical experiments on synthetic data, text classification examples in 20 Newsgroup data and recidivism risk-scoring in COMPAS data.

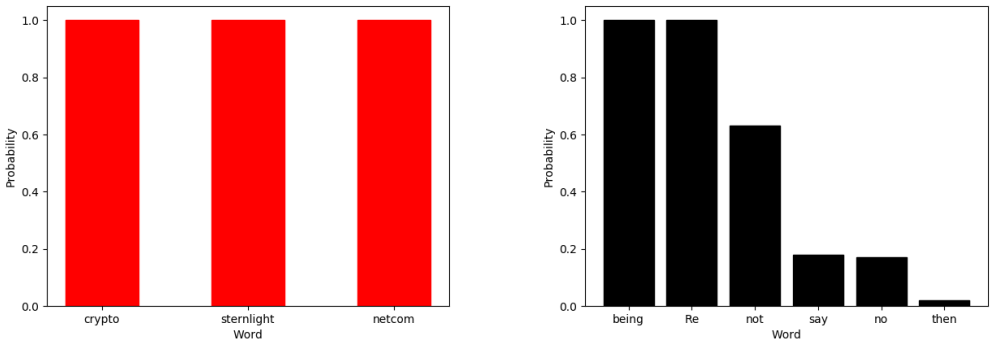


(a) Leaf 0 (b) Leaf 1 (c) Leaf 2



(d) Leaf 3 (e) Leaf 4 (f) Leaf 5

Figure 2: Empirical selection probability in LIME explanations of the random forest model trained by eight-feature synthetic data. A data point is taken from each leaf, and LIME is run 100 times on each point. For each trial of LIME, we record the three features selected by K-Lasso, and calculate the cumulative selection probability for each of the eight features. For each leaf, active features with true coefficients 1 are marked red. Notice that features chosen by LIME are not necessarily locally important features on each leaf. Especially for leaves 3-5, signal from the true features is dominated by signal from the first three features used for tree splitting. For leaf 5, we also tried reducing the sampling proximity by a factor of ten (striped bars), which allows us to recover significant signal from the true local features and rule out the signal of feature 0 used for splitting.



(a) Test document 1 (b) Test document 2

Figure 3: Empirical selection probability for feature words in text classification “electronics vs. crypt”. LIME is run 100 times on the test document; in each LIME trial, we record the three feature words selected by K-Lasso and calculate the empirical selection probability for these words. Words that are informative are marked red. It can be seen that the selected feature words for the first document are consistent and meaningful, while those for the second document are not informative.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *NeurIPS*, 2018.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. In *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Muller, K.-R. How to explain individual classification decisions. *JMLR*, 2010.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *AAAI*, 2019.

Lakkaraju, H., Bach, S. H., and Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, 2016.

Larson, J. L., Mattu, S., Kirchner, L., and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm, 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

Lee, E., Braines, D., Stiffler, M., Hudler, A., and Harborne, D. Developing the sensitivity of lime for better machine learning explanation. In *Proceedings of SPIE: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019.

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 2015.

Lipton, Z. C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *AAAI*, 2018a.

Ribeiro, M. T., Singh, S., and Guestrin, C. Semantically equivalent adversarial rules for debugging nlp models. In *ACL*, 2018b.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *ICML*, 2017.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *ICML*, 2017.

Tan, S., Caruana, R., Hooker, G., Koch, P., and Gordo, A. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640*, 2018a.

Tan, S., Caruana, R., Hooker, G., and Lou, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. In *AIES*, 2018b.

Zeng, J., Ustun, B., and Rudin, C. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society (A)*, 2017.

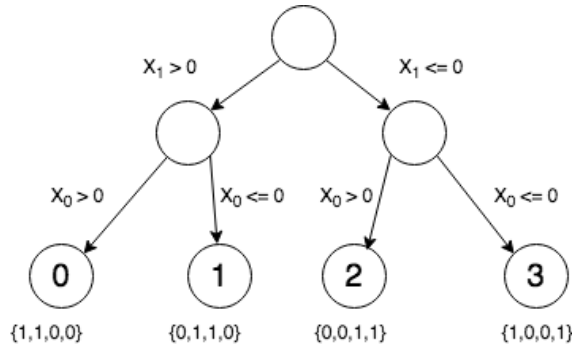


Figure 4: Decision tree partition of four-feature synthetic data.

A. More Simulation Setting

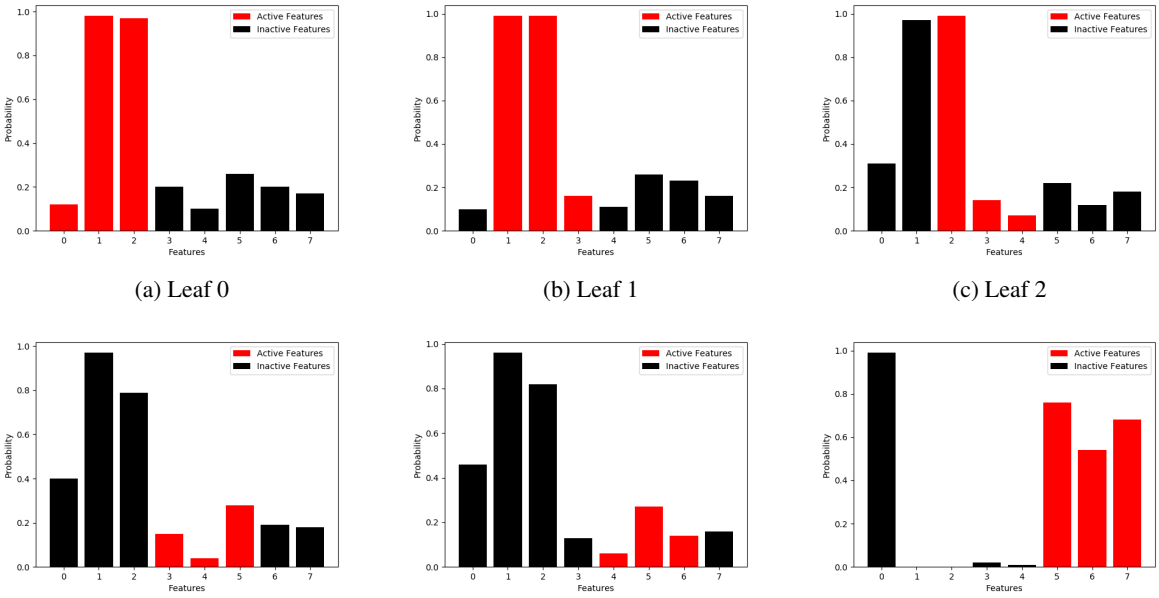
A.1. Setting for four-feature synthetic data

For sample points with four features, we use the first two dimensions of features as their x and y coordinates. We assign each quadrant to different leaf, ignoring the sample points on the x or y axis. For each leaf, we assign different coefficients to their features, as shown in Figure 4. We fit and explain both random forest and gradient boosting classifier within this setting, see Figure 6 and Figure 7.

A.2. More Numerical Results

Text Classification: We select two classes from 20 newsgroup dataset, then apply term frequency-inverse document freque (tf-idf) vectorizer with default settings. Stop words are not removed from resulting tokens as we would like to see if the model is using irrelevant features to predict the results. For the the classification between "Electronics" and "Crypt", we analyze the explanation over two different test data points. We could see from the results that the explanation of test data document one contains several indicative words, such as "crypto", "netcom" and "Sternlight" in this case. However, the explanation results for test data point two contains only one indicative word "information". We also include example for and the result is shown in Figure 8 in Appendix A.

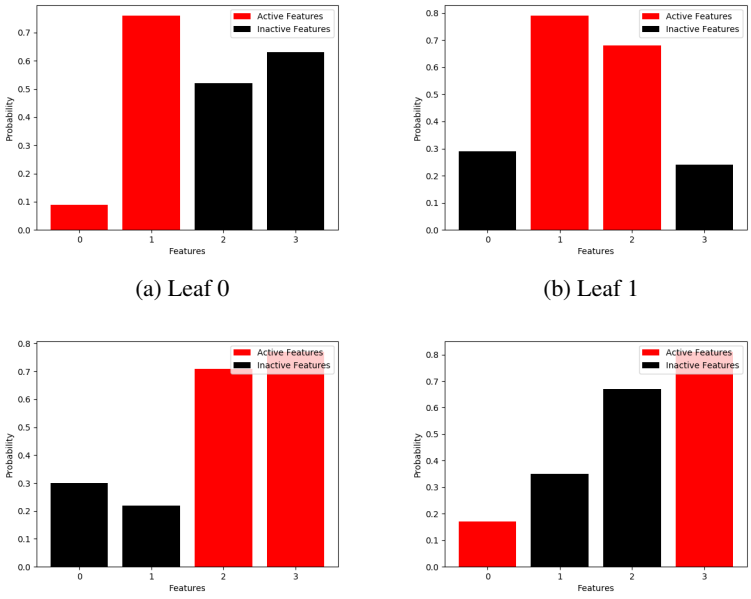
COMPAS Recidivism Risk Score Data: The COMPAS dataset from ProPublica contains a lot irrelevant columns, as well as null values. We selected twelve relevant columns of the dataset, then drop the rows that contain null value. Specifically, we exclude the decile score columns as it is directly related to the text label. We then encode the categorical features, such as "sex" and "race", using one-hot encoder, and encode the label text using label encoder. After the simple data pre-process, we trained a random forest classifier on the processed dataset to mimic the COMPAS black-box model, which we do not have access to.



(a) Leaf 0 (b) Leaf 1 (c) Leaf 2

(d) Leaf 3 (e) Leaf 4 (f) Leaf 5

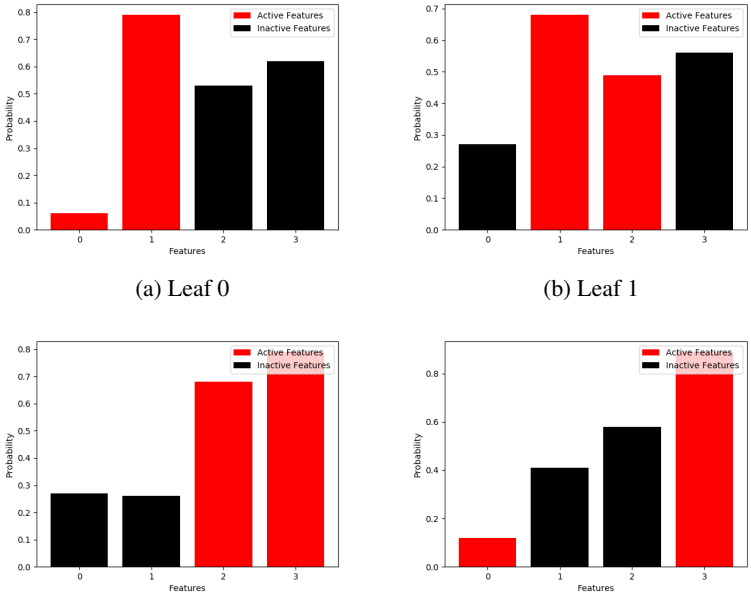
Figure 5: Empirical selection probability in LIME explanations of the gradient boosting tree model trained by eight-feature synthetic data. A data point is taken from each leaf, and LIME is run 100 times on each point. For each trial of LIME, we record the three features selected by K-LASSO, and calculate the cumulative selection probability for each of the eight features. For each leaf, active features with true coefficients 1 are marked red. As in the random forest model, here LIME mainly captures the global features used for tree splitting and fails to reflect the local features important on leaves 3-4. For leaf 5, we reduce LIME's sampling proximity by a factor of 10. This allows us to recover a significant amount of signal from the local features 5-7, although signal from feature 0 is still present.



(a) Leaf 0 (b) Leaf 1

(c) Leaf 2 (d) Leaf 3

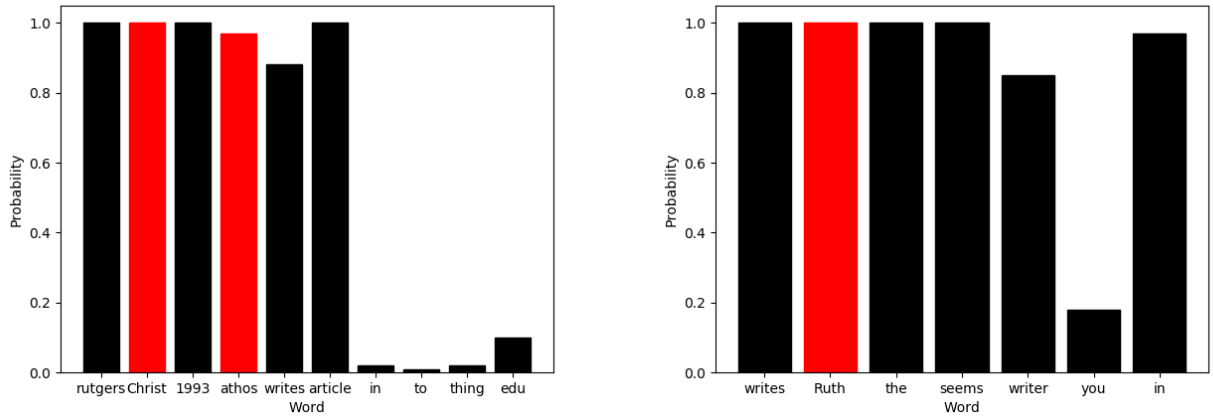
Figure 6: Empirical selection probability in LIME explanations of the random forest model trained by four-feature synthetic data. A data point is taken from each leaf, and LIME is run 100 times on each point. For each trial of LIME, we record the two features selected by K-LASSO, and calculate the cumulative selection probability for each of the four features. For each leaf, active features with true coefficients 1 are marked red.



(a) Leaf 0 (b) Leaf 1

(c) Leaf 2 (d) Leaf 3

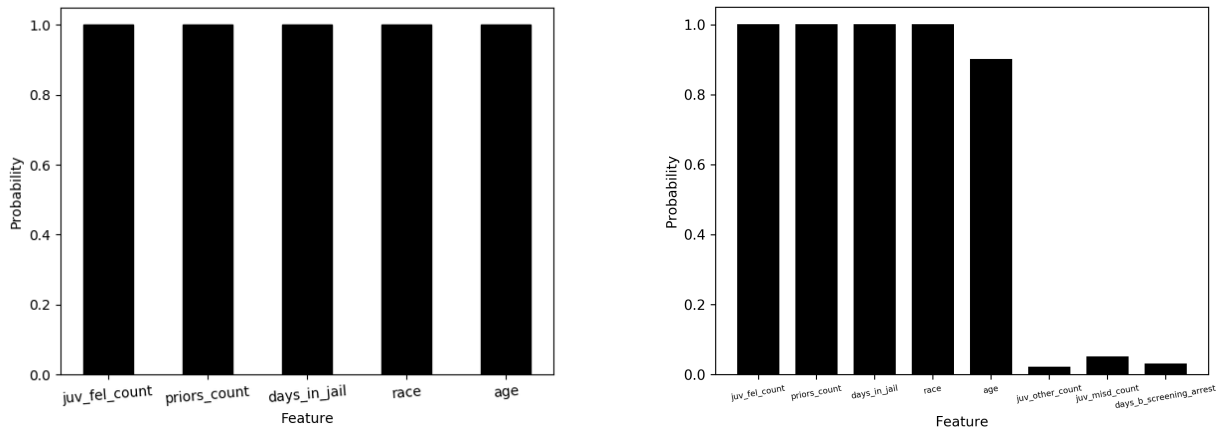
Figure 7: Empirical selection probability in LIME explanations of the gradient boosting tree model trained by four-feature synthetic data. A data point is taken from each leaf, and LIME is run 100 times on each point. For each trial of LIME, we record the two features selected by K-LASSO, and calculate the cumulative selection probability for each of the four features. For each leaf, active features with true coefficients 1 are marked red.



(a) Test document 1

(b) Test document 2

Figure 8: Empirical selection probability for feature words in text classification "Christianity vs. Atheism". LIME is run 100 times on the test document; in each LIME trial, we record the six feature words selected by K-Lasso and calculate the empirical selection probability for these words. Different trials potentially select different feature words. Words that are informative are marked red. It can be seen that many of the frequently selected feature words are not informative.



(a) Sample data 1

(b) Sample data 2

Figure 9: Empirical selection probability in LIME explanations of the COMPAS mimic model. LIME is run 50 times on the test points. We record five top features selected by K-LASSO and calculate the empirical selection probability for these features. The features "juvenile felony count", "priors count", "days in jail", "race", and "age" are consistently selected in different trials on a single data point, as well as for two different data points.