"Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations
Seminar: Uncertainty quantification in machine learning

Rahul Poovassery
Matriculation Number: 229295
Summer Semester 2021/22

Supervisor: Chiara Balestra

Abstract

Machine learning models have been widely used, but they still primarily exist as "black boxes". As a result, methods for understanding these black-box models are crucial because they attempt to explain the models' fairness and dependability by making them more transparent. Nevertheless, if these explanations are accompanied by uncertainties, users will be less inclined to trust the predictions and will become more concerned about the robustness of the model. Here, the emphasis will be on one such method, Local Interpretable Model-Agnostic Explanations (LIME) and to illustrate three sources of uncertainty in this method: randomness in the sampling process, variance with sampling proximity, and variation in explained model credibility across various data points. The existence of these sources of uncertainty is proven by analysing the uncertainty in the LIME method with the aid of synthetic data and two public data sets, a newsgroup data set and a data set for recidivism risk-scoring.

1. Introduction

Machine learning is at the heart of many recent scientific and technological breakthroughs, and it is playing an increasingly significant role in decision-making across a wide range of fields. Modern machine learning models are sometimes effectively "black boxes", as it is practically hard to understand how they work. For instance, a doctor would never operate on a patient merely because "the model said so" as the stakes are quite high if a machine assumes the place of the doctor. Therefore, knowing the reasoning behind the model's predictions will enable us to assess whether the model achieves desired features such as fairness, privacy, etc. (Doshi-Velez and Kim (2017)), as well as to identify and fix any model flaws (Ribeiro et al. (2018b)) and this would help users to decide whether or not to trust the model.

As a result, methods for comprehending and illuminating these black-box models were subsequently created with the goal of assisting users in evaluating and establishing trust in black-box models and their predictions. But what if these justifications themselves aren't reliable and have flaws? Uncertainty in explanations raises concern about the understanding of a specific prediction which in turn leads to questions about the reliability of the black-box model, reducing the significance of the explanation (Ghorbani et al. (2019)).

Therefore, the topic we are addressing here is when can we trust an explanation, and to answer this question, we study a particular explanation method called Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al. (2016)). It is a model-agnostic method to generate local explanations for the model. The method's fundamental notion is simple and straightforward: Assume we have a complex classifier with a very nonlinear decision boundary. However, if we focus on a single prediction and examine it, we can use a simple interpretable model to explain how the model behaved in that particular locale. LIME explains the prediction of the desired input by using a local surrogate model trained on perturbations of that data point, which is achieved by sampling its neighbouring inputs and learning a sparse linear model utilizing the complex model's predictions for these neighbours as labels. Then the features in the linear model with the highest coefficients are then assumed to be significant for predicting that input. We illustrate that there are uncertainties in the explanations of LIME after investigating it with the aid of a few trials. These

uncertainties should not be disregarded. In order to explain the predictions, LIME requires sampling close to the desired input, which introduces uncertainty because the sampling process introduces concerns like randomness, variance with the sampling proximity, and variation in the credibility of the explained model.
Additionally, two recent methods that are useful and will be utilized in the future to help users interpret models and their predictions are discussed.

2. Methods

The local explanation method investigated in this experiment is LIME. The explanation produced by LIME is obtained by the following:

$$\xi(x) = argmin_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (1)$$

An explanation $\xi(x)$ is defined as a model $g \in G$, where G is a family of interpretable models, for instance, linear models, decision trees etc. It is an optimization function with two loss terms, to put it succinctly. The first term $\mathcal{L}(f, g, \pi_x)$, is a measure of how unfaithful the approximation of the black-box model$f$, by using simple interpretable model g is in the neighbourhood of the data point x. π_x is a proximity measure between an instance z to x, so as to define locality around x. And the second term $\Omega(g)$ is a measure of complexity of the explanation $g \in G$ that is it is to regularize the complexity of our simple surrogate model. For decision trees, for instance, (g) might refer to the depth of the tree, whereas for linear models, (g) might refer to the total number of non-zero weights. We must minimize $\mathcal{L}(f, g, \pi_x)$ while keeping $\Omega(g)$ at a level that is human interpretable (Ribeiro et al., 2016).
This approach can be extended to different fidelity functions $\mathcal{L}$, explanation families G, and complexity measures Ω. Here, the search is conducted using perturbations, and K-LASSO is employed as the interpretable model. We set $\Omega = \infty 1\left[\|w_g\|_0 > K\right]$ for K-LASSO, where w denotes for the linear model's coefficients and π_x samples points nearby x to train K-LASSO (Zhang et al. (2019)).

3. Experiment Design

We run LIME on three different types of data sets, one synthetic data example and two real data sets, to show and establish the existence of uncertainties in the LIME method and to identify their sources.
In the first experiment, we will run LIME on the synthetic data produced by trees for a single data point in multiple iterations and then use K-LASSO to select the top features from it each time. Then, by looking at the cumulative selection probability of the chosen features, we will see if the top features that LIME chose throughout various trials are consistent or not. This will demonstrate whether LIME can explain a given data point across various trials. The experiment then tweaks different LIME parameters to determine how sensitive LIME's explanations are to sampling proximity and different sample sizes. Finally, we'll run LIME on the two real data sets. We will first use the Newsgroup data set as an example of text classification, and then we will use the "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS) Recidivism Risk Score Data set as an example of risk classification. In these tests, we will apply LIME to different data points, compare the explanations of LIME on those data points, observe the cumulative selection probability of the top features we've chosen, and see if the top features are actually informative in the real context.

4. Data Summary

Three data sets, one synthetic in a simulation setting and two actual data sets are being utilized in the experiment to demonstrate the existence of uncertainties in LIME.

4.1. Simulation setting: Synthetic data generated by trees

Using local sparse linear models on uniformly distributed input in $[0, 1]^N$, we create a training and test data set with N number of features. Out of the two scenarios with 8 and 4 features that are taken into consideration, we will only use the case with 8 features in this example. To observe LIME's local behaviour at different data points, the data is then partitioned using a known decision tree.

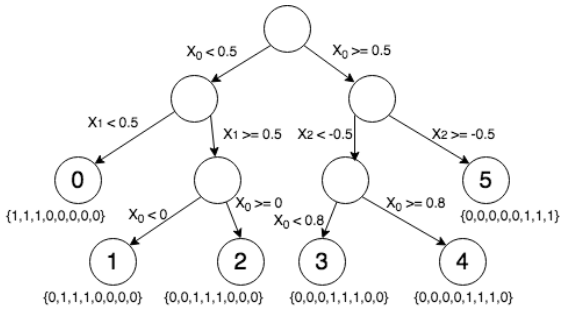


Figure 1. Decision tree partition of eight-feature synthetic data (Zhang et al. (2019)).

In this instance, data is split into six leaves with known coefficients β and labels are assigned to each data point x using the given linear classifier.

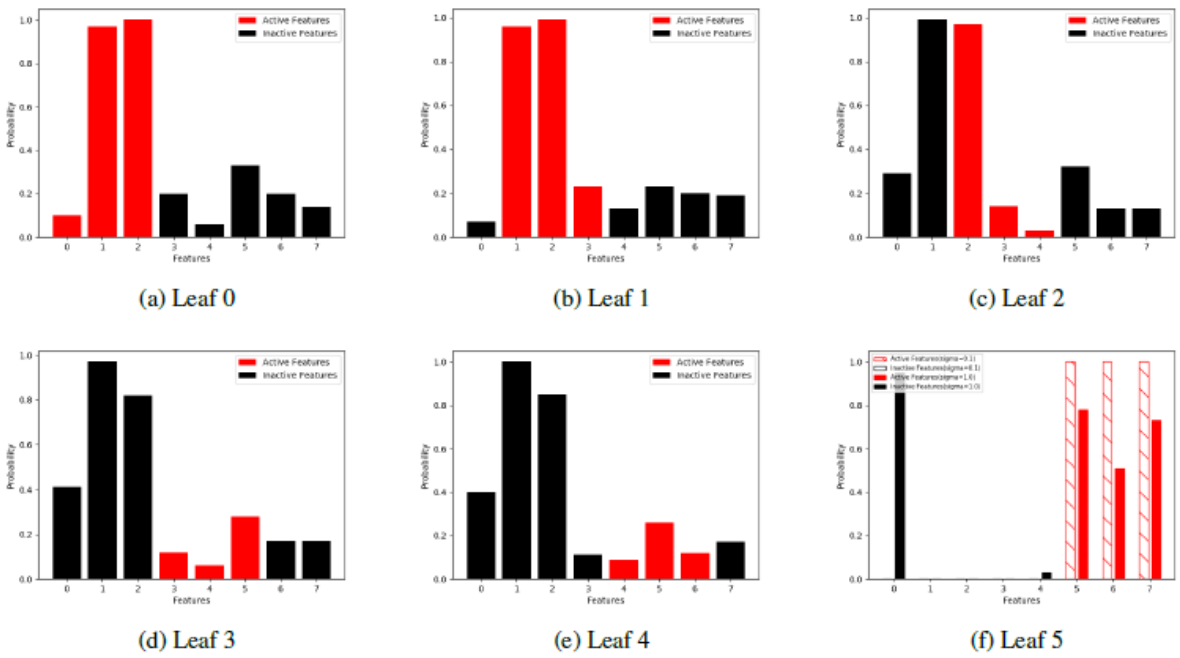


Figure 2. Empirical selection probability in LIME explanations of the random forest model trained by eight-feature synthetic data.

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

In each leaf, three of the eight characteristics are given coefficients 1, as seen in Figure 1. Under each end node in the figure, each leaf's local coefficients are indicated.

4.2. Text Classification Data set

Here, The 20 Newsgroups data set is utilized as a text classification example. It consists of about 20,000 newsgroup documents distributed evenly across 20 different newsgroups. The data set is typically used for machine learning experiments on text applications, like text classification and text clustering. The two document classification examples "Atheism vs. Christianity" and "electronics vs. crypt" are taken from the data set in order to observe LIME's behaviour on text classification models with high accuracy. LIME is then applied to these examples to determine whether the features it chooses are informative or not (Ribeiro et al. (2016)).

4.3. COMPAS Recidivism Risk Score Data set

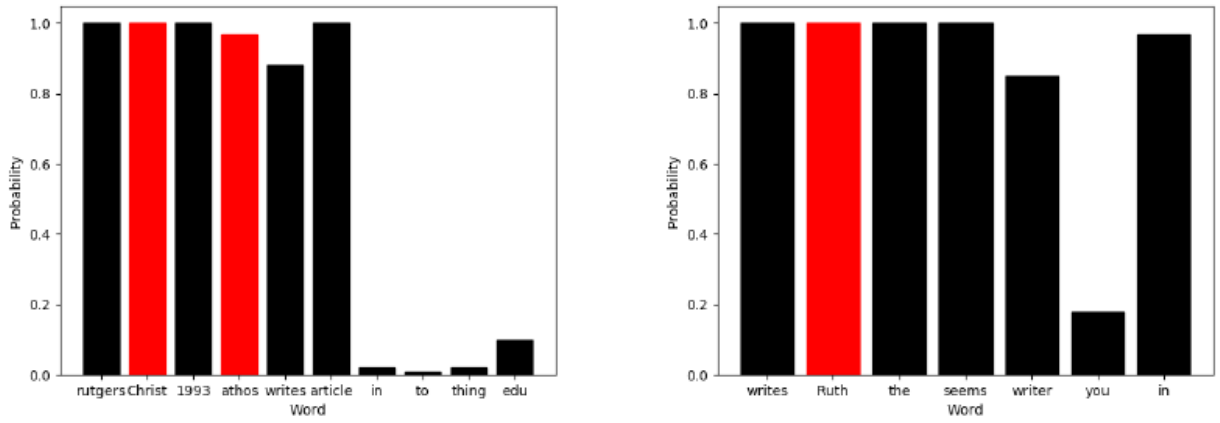
The "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS) dataset is utilized in the final experiment. COMPAS is a risk scoring algorithm developed by Northpointe Inc. It is a commercial algorithm that judges and parole authorities use to determine whether a criminal defendant is likely to commit another crime (recidivism). Criminal history, jail and prison time, demographics, and COMPAS risk scores for defendants from Broward County are used to determine the risk scores, which are then categorized as "High," "Medium," and "Low." We use a portion of the COMPAS dataset compiled and processed by ProPublica to apply LIME to two data points that COMPAS has categorized as "high risk" in order to observe how LIME behaves when applied to a risk classification model.

5. Analysis

On the three data sets, various experiments are run to analyze the LIME explanation method and determine whether there are any sources of uncertainty. The following section is an overview of the key points and observations from these experiments:

5.1. Simulation Setting

In this instance, we employ Random Forest Classifier as the black-box model to train on synthetic data with eight features. Following this, we run LIME 100 times on one data point from each of the six leaves, and the



(a) Test document 1

(b) Test document 2

Figure 3. Empirical selection probability for feature words in text classification "Christianity vs. Atheism".

top features are selected using K-LASSO for each trial by computing the cumulative selection probability for each of the eight features.

Results are illustrated in Figure 2, where active features for each leaf that have true coefficients of 1 are highlighted in red. We can see from the figure that the top features chosen by LIME are not always the locally significant features on each leaf. We can see that the features chosen by LIME with the highest cumulative selection probability are identical to their true features in leaves 0 and 1. One of the features chosen in leaf 2 also coincides with the true features. However, signals of the first three features are chosen for leaves 3–5, rather than the actual features.

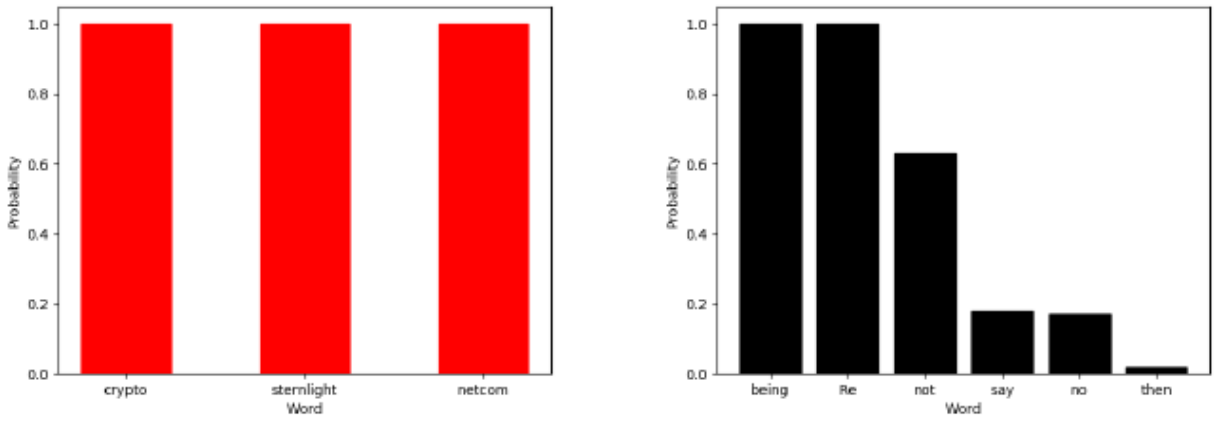
The first three attributes are employed for data tree splitting on a global scale. LIME records this global information rather than local information for each leaf, although each leaf has different features that are important. LIME's explanations are therefore contradictory locally for each data point from each leaf in this instance because it is unable to collect the relevant information accurately. It has been observed that LIME by default selects samples from a standard normal distribution $\mathcal{N}\left(0, \sigma^2\right)$ near the test point, where σ^2 is the variance of the training data, and that different trials frequently choose different features as a result of sampling variance because the variance of the training data determines the sampling proximity.

So, in order to verify this observation, the sampling proximity of LIME is altered by pulling a sample from $\mathcal{N}\left(0, (0.1\sigma)^2\right)$ near the test point instead of $\mathcal{N}\left(0, \sigma^2\right)$ for a data point on leaf 5. Figure 2f illustrates this, showing a tenfold decrease in the sampling proximity for leaf 5. LIME thus picks locally significant features in leaf 5 with smaller sample proximity. We can see from this trial that LIME typically picks up global features with larger sampling proximity and locally important features with smaller sampling proximity.

5.2. Text Classification

Two classification examples—"Atheism vs. Christianity" and "electronics vs. crypt"—are chosen from a data set of 20 newsgroups, and a term frequency-inverse document frequency (tf-idf) vectorizer is used on the data with the default settings. Following that, Multinomial Naive Bayes classifiers—which in this case serve as our black-box model—are trained on each of these classification cases. For each of the classes, the model provides test accuracy values of 0.9066 and 0.9214, respectively (Ribeiro et al. (2016)). In order to determine the feature importance for each output and to examine the explanation over two different test data points, we now run LIME on the two different text documents for each of these cases. On the test documents, LIME is run 100 times for the examples and top features are chosen using K-LASSO utilizing empirical selection probability. Figure 3 and Figure 4 provide empirical selection probabilities for the feature words in the text classification "Atheism vs. Christianity" and "Electronics vs. Crypt", respectively. The informative words in the figures are highlighted in red.

Figure 3 demonstrates that different trials select different words as their top features. We can observe that just two words from Test Document one, "Christ" and "Athos" and one word from Test Document 2, "Ruth" are considered informative in relation to the context out of the six features chosen by LIME. It is clear that none of the other frequently used feature



(a) Test document 1

(b) Test document 2

Figure 4. Empirical selection probability for feature words in text classification "electronics vs. crypt".

words are informative. In the second example, Figure 4 shows that the top features in Test Document 1 are consistent and informative, as shown by the explanations chosen by LIME. In other words, the top features are all informative and they don't vary with different trials, in this case, "crypto," "netcom," and "Sternlight." However, none of the features that were chosen for test document 2 is informative. Therefore, the model in this instance appears to be quite credible with Test Document 1, but ineffective with Test Document 2. Thus, the model's credibility fluctuates across different input data and is not always reasonable for different test documents, according to LIME's local explanations.

5.3. Risk Classification

In the last experiment, two data points with both numerical and categorical variables that COMPAS has classified to be "high risk" are chosen. and is trained using a random forest classifier as the "mimic black-box model" since we do not have access to the original COMPAS black-box model (Tan, Caruana, Hooker, and Lou (2018)). On these two data points, LIME is then run 50 times, and the top features are chosen using K-LASSO utilizing the empirical selection probabilities for these features.

The results are shown in Figure 5. The figure shows that the features "juvenile felony count", "priors count", "days in jail", "race", and "age" are consistently selected in different trials on both a single data point and for two different data points.

As there are very few deviations in the top features that were selected on the same data point between different trials, while at the same time the selected features are consistent across data points, we can state that LIME's explanations in this situation are reliable and consistent.

6. Related Works

In many situations, knowing why a model produces a particular prediction can be just as important as knowing if the prediction is accurate. In response, a number of approaches have been put out to assist users in interpreting the predictions of complex models. Designing accurate models that are still inherently interpretable (Lakkaraju et al. (2016) and Letham et al. (2015)) and developing post-hoc ways to explain black-box models are the two main areas of study for interpretable approaches. Post-hoc techniques can either be applied globally for the entire model (Ribeiro et al. (2018a) and Tan, Caruana, Hooker, Koch, et al. (2018)) or locally for a particular input (Baehrens et al. (2010) and Ribeiro et al. (2016)) . In this seminar, we looked at one specific local explanation method, LIME, and discovered the sources of potential uncertainties with LIME's stability and robustness.

The development of techniques that aid users in interpreting predictions have been driven by the growing tension between the accuracy and interpretability of model predictions.

Deep Learning Important Features (DeepLIFT) is one such contemporary technique with potential for the future. By backpropagating the contributions of each neuron in the network to each feature of the input, it is a method for decomposing the output prediction of a neural network on a given input. Each neuron's activation is compared to its "reference activation" by

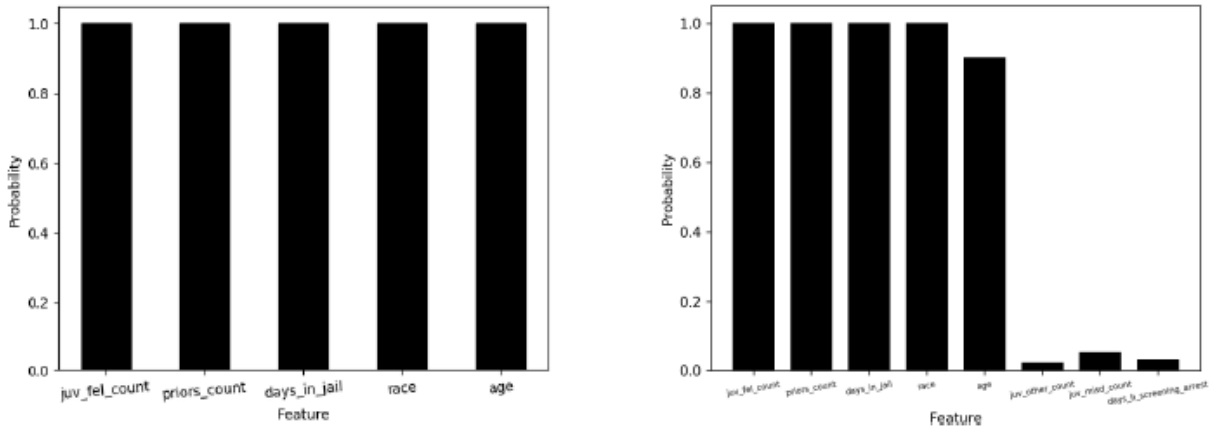


Figure 5. Empirical selection probability in LIME explanations of the COMPAS mimic model.

DeepLIFT, which then calculates contribution scores based on the difference (Shrikumar et al. (2017)).

And last, there is a relatively new method called Shapley Additive explanations (SHAP), which uses game theory to explain the output of any machine learning model. Using the standard Shapley values and their related expansions from game theory, it links optimal credit allocation with local explanations. It proposes a unified method of understanding model predictions (Lundberg and Lee (2017)). It proposes a system that integrates six pre-existing feature attribution techniques, including LIME and DeepLIFT, and they offer their framework as an additive feature attribution model. They demonstrate the ease with which the Shapely Values can be determined and the ability to structure each of these methods as an equation. Once these strategies are applied to this framework of estimating the Shapely values, they have a solid theoretical basis on which to build, such as LIME. The authors of the research have developed a new, model-neutral method for approximating Shapely Values called Kernel SHAP (LIME + Shapely Values), as well as certain model-specific methods like DeepSHAP, which is an adaption of DeepLIFT.

Since LIME is an attribute feature addition method, the author showed that the Shapely Values provide the desired properties for an additive feature explanation method. We are aware that LIME uses a heuristic to select its kernel function and kernel distance as hyperparameters, which may result in uncertainties, and would lead to other consequences. Kernel SHAP reduces this uncertainty and guarantees that the solution to the problem will create Shapely values while also benefiting from the associated mathematical guarantees by presenting a Shapely Kernel and a corresponding loss function (Lundberg and Lee (2017)).

7. Conclusion

Three experiments were used to examine LIME's explanations, and the results allowed us to demonstrate that LIME does contain significant uncertainties. We discovered that the main sources of uncertainty in LIME were three.

The numerical experiments on synthetic data revealed that there is sampling variance, or randomness in LIME's sampling technique when describing a single point. Additionally, it was discovered that it is sensitive to the parameters used, such as sample size and sampling proximity, meaning that its explanation changes with sampling distance. It demonstrates that LIME tends to pick up local features with smaller sample proximity and global features with higher sampling proximity, indicating that users should tune it and select the value effectively to allow LIME to explore both local and global structures in the data. The experiment on text classification examples in 20 Newsgroup data revealed that LIME's explanations for the models' credibility aren't always accurate and vary depending on the type of input data used, that is, it varies across different input data points. Last but not least, the experiment on COMPAS data demonstrated that LIME does function in certain instances where LIME explanations are really trusted and believed to be reliable.

Explanation methods were developed to support users in evaluating and establishing trust in black-box models and their predictions, but if they themselves are questionable, it casts doubt on the validity of the black-box model and its predictions.

The increasing friction between the accuracy and interpretability of model predictions has motivated the development of different methods to assist users in interpreting predictions. Recent methods, such as DeepLIFT, which aids in predicting feature importance for neural networks, and SHAP, which even uses LIME, DeepLIFT, and other feature attribution techniques, will be employed in the future to help users in understanding these black-box models.

References

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, *11*, 1803–1831.

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

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 3681–3688.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1675–1684.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350–1371.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018a). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI conference on artificial intelligence*, *32*(1).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018b). Semantically equivalent adversarial rules for debugging nlp models. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *International conference on machine learning*, 3145–3153.

Tan, S., Caruana, R., Hooker, G., Koch, P., & Gordo, A. (2018). Learning global additive explanations for neural nets using model distillation.

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-compare: Auditing black-box models using transparent model distillation. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 303–310.

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*.