Want robust explanations? Get smoother predictions first.

Deddy Jobson
deddy@mercari.com
Mercari Inc.

ABSTRACT

Model-agnostic machine learning interpretability methods like LIME which explain the predictions of elaborate machine learning models suffer from a lack of robustness in the explanations they provide. Small targeted changes to the input can result in large changes in explanations even when there are no significant changes in the predictions made by the machine learning model. This is a serious problem as it undermines the trust one has in the explanations made. We propose to solve the problem by smoothening the predictions of the machine learning model as a preprocessing step. We smoothen the predictions by taking multiple samples from the neighbourhood of each input data point and averaging the output predictions. Through our preliminary experiments, we show that the explanations are more robust because of smoothening thus making them more reliable.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → *Data mining*.

KEYWORDS

interpretable machine learning, model agnostic interpretability, LIME, robustness

ACM Reference Format:
Deddy Jobson. 2022. Want robust explanations? Get smoother predictions first.. In *Proceedings of Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI '22).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The sudden improvement in performance of machine learning through deep learning and tree ensemble methods has led to an explosion in the adoption of machine learning in a wide variety of prediction tasks in multiple domains like image, text, tabular data, etc. While the increased performance has made machine learning models much more useful in practice, it has come at the cost of interpretability; one can no longer trivially explain the decisions made by machine learning models the same way one could for statistical models like linear regression in the past. While we can do without interpretability in cases where the consequences of the downstream decisions are little, like in the case of recommending movies, interpretability becomes important in high-stakes situations like predicting whether or not a person has cancer[6]. In such a case, it is not just important to know what the predictions of the model are, but also how the predictions were made.

A number of model-agnostic interpretability methods exist to help explain the predictions made by machine learning models. Partial Dependence Plots[5] show the marginal effect of a feature on the outcome. Individual Conditional Expectation plots[4] do the same by making separate plots for each individual thus allowing one to see the variance (and not just the mean) of the effect of each feature. The above two have a problem wherein we consider the effect of very unlikely counterfactual scenarios in the case where the features in the dataset are strongly correlated.

Shapley values[7] take a game-theoretic approach and assume different features take part in a collaboration to assign a score for an instance. The shapley value for a feature is the average increment in the score obtained by the inclusion of said feature in the collaboration. While using shapley values has a strong mathematical foundation, it has the downside where the computational cost for calculation is exponential to the number of features. While methods like Tree SHAP[8] exist to more efficiently calculate the values, there are issues with the robustness[1] of shapley values which have not yet been resolved.

Local Interpretable Model-Agnostic Explanations (LIME)[9] is a method that estimates a local surrogate model in the vicinity of each data point and uses the coefficients of the local model to interpret the decisions made by the model. It is related to SHAP through Kernel SHAP[2], a way to get approximate SHAP values. One advantage of LIME over shapley values is that LIME can produce sparse explanations which don't rely on too many features resulting in more human-friendly explanations. However, issues regarding the robustness[1] of the explanations provided by LIME have been raised. Our goal in this paper is to find ways to improve the robustness of the interpretations made by LIME to improve the reliability and therefore trustworthiness of the provided explanations.

2 PROBLEM SETUP

The original LIME algorithm works as follows, given a trained model and a target data point:

(1) Sample data around the neighbourhood of the data point.
(2) Get the predicted values for the sampled data points.
(3) Fit a surrogate model to the generated data weighted by distance from the target data point.
(4) Explain the prediction of the main model with the coefficients of the surrogate model.

The explanations generated by the above algorithm can be unstable for a number of reasons.

One source of instability is the sampling of data points[15] that is done randomly, ignoring any correlation between features. Methods

AIMLAI '22, October 21, 2022, Atlanta, GA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

have been developed to estimate the required number of samples to get stable explanations[16] or do away with randomness in the sampling altogether[14].

Another potential cause for instability in explanations, especially pertinent to the case of tabular data, is the discretization of the numerical features. While for the most part this can yield more consistent explanations, target data points near the boundaries can have unstable explanations even when the model predictions (which don't rely on discretization) in the vicinity are relatively stable.

3 RELATED WORK

The measurement of the stability (or lack thereof) of LIME's explanations isn't a new research problem. Alvarez-melis et al.[1] have shown that small pertubations to the input can cause a large change in the output without much of a change in the predictions made by the model. They use the definition of Lipschitz continuity to get the maximum possible difference in explanation within the neighbourhood of the data point to be explained. Their approach is similar to prior work that was done to inspect the lack of robustness of predictions made by neural networks[12].

Visani et al.[13] introduce two novel metrics grounded in statistics to measure the extent to which repeated sampling of the data leads to a variance in the explanations. Their metrics quantify the variance of the selected features and coefficient values, the lower the better.

Much more recently, Garreau et al.[3] performed a very deep analysis into the workings of LIME for tabular data and (among other things) found that when the surrogate model (the one trained for interpretability) uses ordinary least squares, and the number of sampled data points is large, the estimations by LIME are robust to mild perturbations. This suggests that the cause of instability could lie elsewhere.

4 OUR METHOD

For our method, we smoothen the predictions of the model we want to explain with the help of Gaussian noise. We do so because we hypothesize that the lack of robustness in the explanations caused by LIME is not because of LIME itself but rather the jaggedness of the predictions made by the model.

We smoothen the predictions by averaging the predictions made on random perturbations on the data points. We consider the case where all features of the data point are numeric and continuous in this study. We perturb each feature by adding it with gaussian noise of zero mean. We refer to the standard deviation of the gaussian noise to be the "strength" parameter. This is because the greater the "strength" parameter, the larger the perturbations and the smoother the averaged predictions will be (assuming enough samples) and so the "stronger" the smoothening effect. We choose a strength value of 0.1 for our experiments and take 100 random samples for each data point for the smoothening process.

5 EXPERIMENTS AND DISCUSSION

Our hypothesis is that smoothening the predictions will yield explanations that are more robust. To test this hypothesis, we look at the extent to which the variance of LIME's explanations change before and after smoothening the predicting function. We define a metric called Lipschitz Discontinuity Score (LDS) Score which is derived from the expression used in the definition of Lipschitz Continuity. Our approach is similar to the one used in [1]. LDS is defined as follows:

$$LDS = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\frac{||f(x_i) - f(x_j)||_2}{||x_i - x_j||_2}$$

In the above expression, N is the number of records in the dataset, i and j are indices to denote individual records and take values from 1 to N, and $f(x_i)$ is the vector of coefficients we get from the explanations of the LIME algorithm.

Table 1: Preliminary experiments on the Boston dataset (the lower the score the better)

Algorithm	Lipschitz Discontinuity Score
LIME	2.78
LIME smoothed	2.60

We perform preliminary experiments on the publicly available Boston dataset, a dataset with 12 covariates for a regression problem. We parameterize the LIME algorithm to explain with only 3 features. The base model used is the random forest regressor from scikit-learn. We use the default parameters of the random forest since it suffices for the purposes of this study. We estimate the LDS on the Boston dataset using 10-fold cross validation. In table 1, we compare the LDS of the explanations of LIME for two cases: with and without smoothening. We find that there is a substantial improvement in the LDS when smoothening the predictions, in line with our hypothesis.

6 FUTURE WORK

In this paper, we smoothen the predictions of the machine learning model by sampling neighbouring points randomly multiple times and taking the average of the output. We do this to increase the robustness of the explanations by LIME. We chose white noise since the approach is similar to the original LIME algorithm, but since its introduction, various improved sampling strategies have been proposed that result in more robust explanations[10, 11]. Trying those other sampling methods for the purpose of smoothening the predictions is beyond the scope of this extended abstract and can be considered as one avenue for future research.

While we perform preliminary experiments with tabular data, our hypothesis can be potentially true for other forms of data, more so due to the greater dimensionality of data like image, text, etc. In order to extend the idea to other forms of data, the key will be to find how best to perturb the input to get smooth predictions.

Lastly, we test our hypothesis with LIME and found promising results. Since the instability of explanations of other interpretability methods can also be (at least partly) explained by unstable predictions of the machine learning model, we suspect our idea can be applied to improve other model interpretability methods too.

As we can see, there is a lot of scope for future work and we are excited to see how research develops in this direction.

7 CONCLUSION

In this paper, we propose a way to improve the robustness of LIME, a model-agnostic explainer of the predictions of machine learning models. We propose smoothening the predictions made by the model to increase the consistency of the predictions made by the model, thereby making the explanations more trustable. We explain how we smoothen predictions using random noise and perform some preliminary experiments on publicly-available datasets to achieve promising results. We also outline future steps that can be taken to increase the scope of the research.

8 ACKNOWLEDGEMENT

We'd like to thank Mercari Inc. for supporting the research and also the anonymous reviewers who gave very helpful feedback to improve the quality of the paper. Any remaining deficiencies left in the paper belong to the authors.

REFERENCES

[1] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. https://doi.org/10.48550/arXiv.1806.08049 arXiv:1806.08049 [cs, stat].

[2] Ian Covert and Su-In Lee. 2021. Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. PMLR, 3457–3465. https://proceedings.mlr.press/v130/covert21a.html ISSN: 2640-3498.

[3] Damien Garreau and Ulrike von Luxburg. 2022. Looking Deeper into Tabular LIME. http://arxiv.org/abs/2008.11092 arXiv:2008.11092 [cs, stat].

[4] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2014. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. https://doi.org/10.48550/arXiv.1309.6392 arXiv:1309.6392 [stat].

[5] Brandon M. Greenwell, Bradley C. Boehmke, and Andrew J. McCarthy. 2018. A Simple and Effective Model-Based Variable Importance Measure. https://doi.org/10.48550/arXiv.1805.04755 arXiv:1805.04755 [cs, stat].

[6] P. Karatza, K. Dalakleidi, M. Athanasiou, and K.S. Nikita. 2021. Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2310–2313. https://doi.org/10.1109/EMBC46164.2021.9630556 ISSN: 2694-0604.

[7] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/arXiv.1705.07874 arXiv:1705.07874 [cs, stat].

[8] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2019. Consistent Individualized Feature Attribution for Tree Ensembles. https://doi.org/10.48550/arXiv.1802.03888 arXiv:1802.03888 [cs, stat].

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (Feb. 2016). https://doi.org/10.48550/arXiv.1602.04938

[10] Sean Saito, Eugene Chua, Nicholas Capel, and Rocco Hu. 2021. Improving LIME Robustness with Smarter Locality Sampling. https://doi.org/10.48550/arXiv.2006.12302 arXiv:2006.12302 [cs, stat].

[11] Sheng Shi, Xinfeng Zhang, and Wei Fan. 2020. A Modified Perturbed Sampling Method for Local Interpretable Model-agnostic Explanation. https://doi.org/10.48550/arXiv.2002.07434 arXiv:2002.07434 [cs, stat].

[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. https://doi.org/10.48550/arXiv.1312.6199 arXiv:1312.6199 [cs].

[13] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. 2022. Statistical stability indices for LIME: obtaining reliable explanations for Machine Learning models. *Journal of the Operational Research Society* 73, 1 (Jan. 2022), 91–101. https://doi.org/10.1080/01605682.2020.1865846 arXiv:2001.11757 [cs, stat].

[14] Muhammad Rehman Zafar and Naimul Khan. 2021. Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction* 3, 3 (Sept. 2021), 525–541. https://doi.org/10.3390/make3030027 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[15] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. https://doi.org/10.48550/arXiv.1904.12991 arXiv:1904.12991 [cs, stat].

[16] Zhengze Zhou, Giles Hooker, and Fei Wang. 2021. S-LIME: Stabilized-LIME for Model Explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2429–2438. https://doi.org/10.1145/3447548.3467274