Unbiased Kernel Shap in Dalex

Michał Kucharczyk mk439959@students.mimuw.edu.pl
Edward Sucharda es446519@students.mimuw.edu.pl
University of Warsaw, Poland

Abstract

In this paper we describe work we did in implementing and testing a modified version of Kernel Shap called by us Unbiased Kernel Shap. We successfully extended Dalex python library to calculate and plot this new form of shapley value estimation. We run tests to prove that Unbiased Kernel Shap is really unbiased and check at what cost it really is. We compered methods in terms of relative error to exact shapley values. Our experiments does not show any advantages of using our implemetnation instead of original Kernel Shap.

1. Introduction

One of the state of art in explainable machine learning is research on SHAP(Lundberg and Lee, 2017). This idea allows explaining local outputs for any model presenting impact of every feature based on cooperative game theory - Shapley value(Shapley, 1951). One of the main problems of this approach is that for explaining machine learning models number of the operations for single output is 2^N, where N is number of input features. This cause that this method is too time consuming for most problems. One of the solutions is using combination of another state of art work LIME(Ribeiro et al., 2016) and aforementioned method, known as Kernel Shap(Lundberg and Lee, 2017). This method calculates very good estimation of shapley values in much fewer operations. But as every estimation it has some drawbacks. One of them is the problem that there is no proof that this method is unbiased. The solution to this issue was introduced in paper *Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression*(Covert and Lee, 2020). The goal of this research is to implement and test this method in Python version of library Dalex(Baniecki et al., 2021).

2. Methodology

2.1 Implementation

Every explanation made by Dalex is implemented within methods assigned tIn this paper we describe work we did in implementing and testing a modified version of Kernel Shap call by us Unbiased Kernel Shap. We successfully extended Dalex python library to calculate and plot this new form of shapley value estimations. We run tests to prove that Unbiased Kernel Shap is really unbiased and check at what cost it really is. We compered methods in terms of relative error to exact shapley values. Our experiments does not show any advantages of using our implemetnation instead of original Kernel Shap.o class *Explainer*. We imple-

mented a new possible input parameter for method *predict_parts* keeping the structure of the library. Our implementation allows similar functionality to Dalex's shap or break down. We kept the structure of classes and preserve the graphical style of ploted explanations. We present our work in fork of original repository(Kucharczyk and Sucharda, 2023).

2.2 Tests

We tested our algorithm on two datasets. The first one is a house price estimation(House). We have chosen a subset of nine features: number of bedrooms, number of bathrooms, square footage of the home, square footage of the lot, total floors, overall grade given to the housing unit, square footage of house apart from basement and year house was built. We only left 9 features from original 20 firstly because some were categorical and more importantly because we wanted a dataset which explanations can be compared with exact shapley values. The second dataset was death rate regression (Death). From 32 original columns we dropped two (binnedinc and geography) that did not contain numeric data. For each of the datasets we trained two models: SVM with linear kernel and Xgboost on 300 training samples.

For both datasets we explained 10 observation with Kernel Shap from Shap(Lundberg and Lee, 2017) python library and our implemented version of Unbiased Kernel Shap. For each model and data point we tested 50, 200 and 500 number of samples (parameter of the estimation) and for each of these made 5000 measurements. Finally we calculated with Exact Explainer from aforementioned Shap package precise shapley values for Housing dataset. For Death dataset exact shapley values could not be obtained because of the number of features in this dataset.

3. Experimental results

First analysis confirmed fact that Unbiased Kernel Shap has higher variance than Kernel Shap. In the Figure 1 are presented results of average value estimations over 5000 measurements and their 0.95 confidence interval. Fact that Unbiased Kernel Shap is very susceptible to number of samples makes first concerns about it's unbiased property. The predicted average should be close to constant which is not fulfilled in any of the 9 plots.

On the second experiment we calculated relative shap value estimator error thanks to Exact Shap calculated on Housing dataset. Results for one of the input features are presented on Figure 2. Original Kernel Shap is here more stable while changing number of samples (n_sample). This leads to conclusion that it is much faster that Unbiased Kernel Shap because even small number of samples is enough to achieve results similar to Unbiased Kernel Shap with much more samples.

The final experiment was calculating average error and average standard deviation over all tested observations. The results on the Figure 3 show that Unbiased Kernel Shap gives worse results than Kernel Shap. We suspect that this is mainly caused by much bigger standard deviation. The standard deviations comparison for Death dataset is presented in Figure 4 and the results for House dataset similarly confirm high variance of Unbiased Kernel Shap.

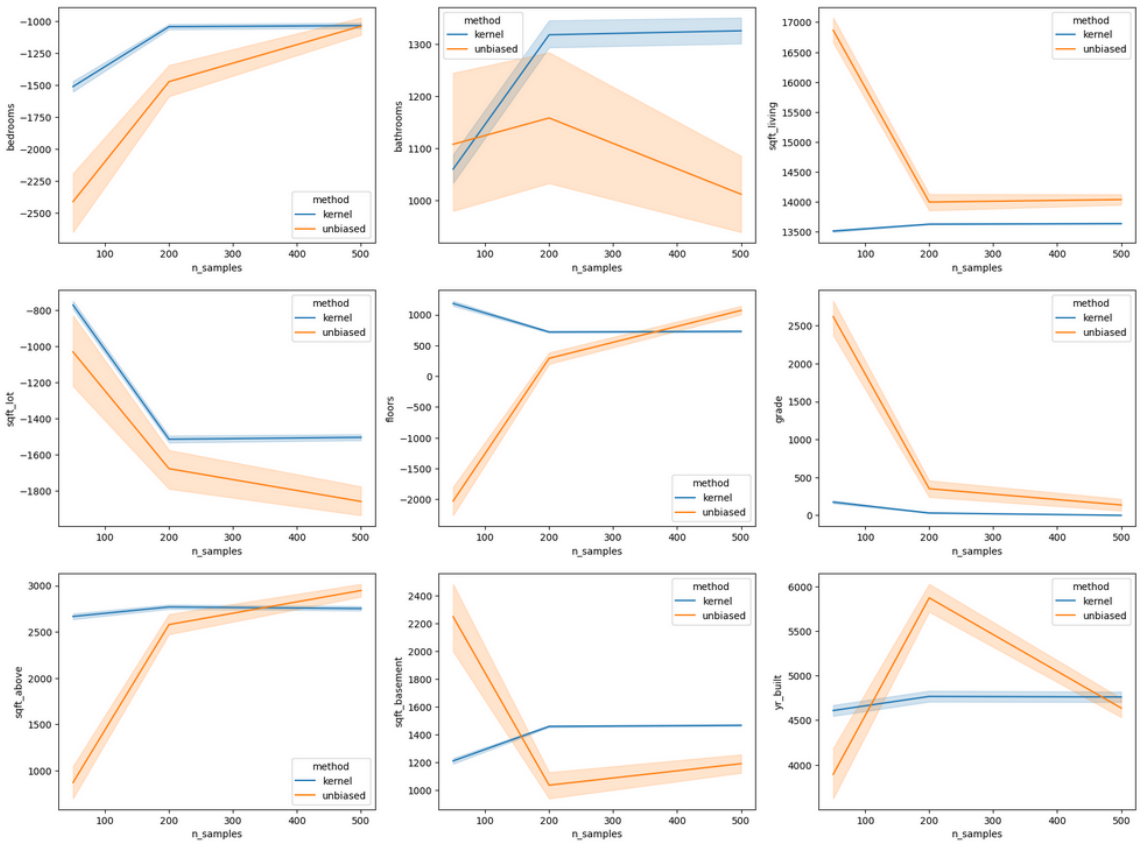

Figure 1: Shapley value estimation on SVM trained on House dataset for single observation.

4. Conclusion

Our research was to check if in some cases Unbiased Kernel Shap can be a good alternative to original Kernel Shap. Despite fact that Unbiased Kernel Shap can be proven mathematically that it is unbiased, our research show that it's high variance in total makes it worse than original method. Moreover we did not obtained results that would confirm results in article (Covert and Lee, 2020) because for single observation average over 5000 measurements was changing with number of samples. This last fact leads to conclusion that the main positive feature - fact that estimator is unbiased were not confirmed empirically by our experiments.

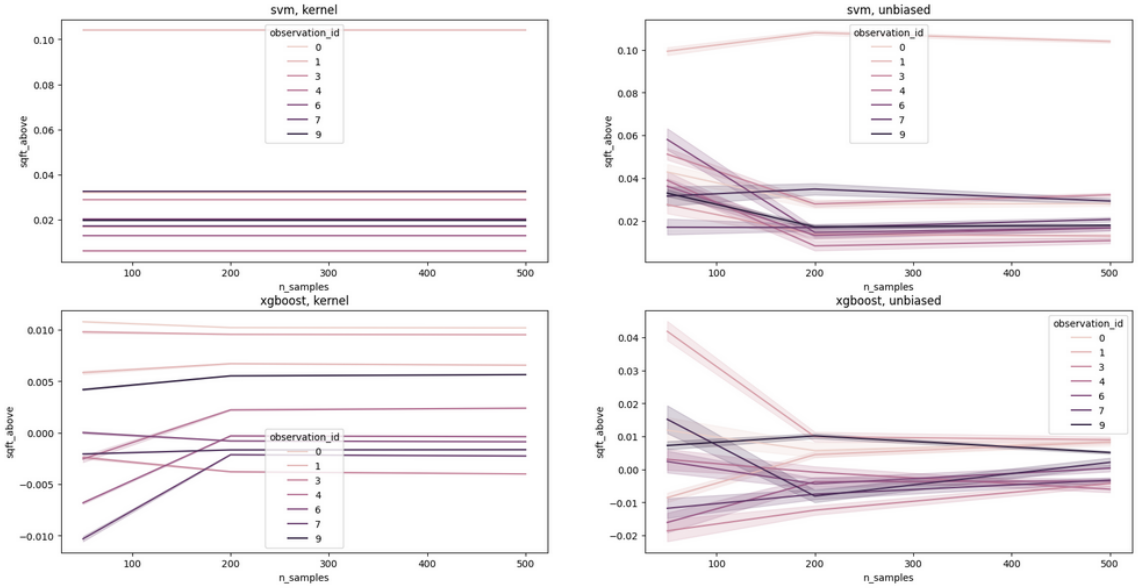

Figure 2: Raltive error of shapley value estimation for square footage of house apart from basement.

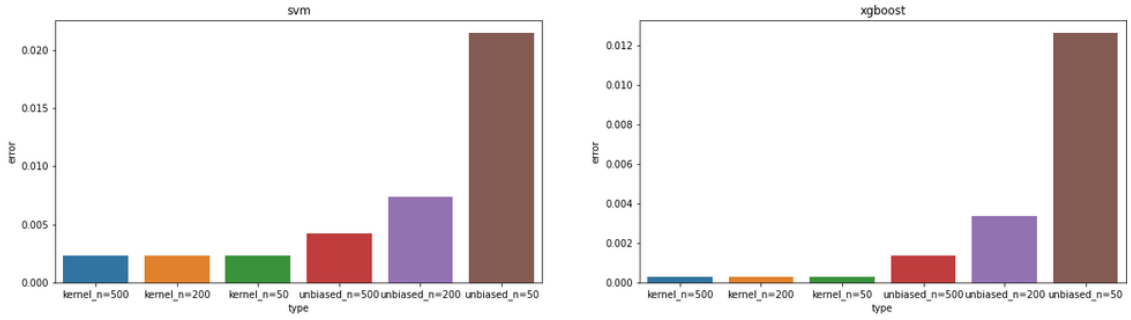

Figure 3: Relative error averaged over all features and whole House test dataset.

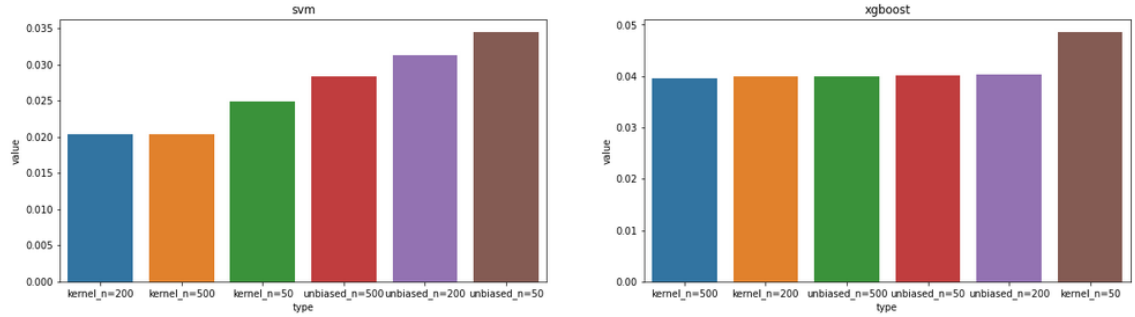

Figure 4: Standard deviation error averaged over all features and whole House test dataset.

References

Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, and Przemyslaw Biecek. dalex: Responsible machine learning with interactive explainability and fairness in python. *Journal of Machine Learning Research*, 22(214):1–7, 2021. URL http://jmlr.org/papers/v22/20-1473.html.

Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation via linear regression. *CoRR*, abs/2012.01536, 2020. URL https://arxiv.org/abs/2012.01536.

Death. Ols regression challenge - death rate. https://data.world/nrippner/ols-regression-challenge. Accessed: 2023-01-31.

House. Kaggle - house sales in king county, usa. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction. Accessed: 2023-01-31.

Michał Kucharczyk and Edward Sucharda. Unbiased kernel shap. https://github.com/quczer/DALEX, 2023. Accessed: 2023-01-31.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

Lloyd S. Shapley. *Notes on the N-Person Game mdash; II: The Value of an N-Person Game*. RAND Corporation, Santa Monica, CA, 1951. doi: 10.7249/RM0670.