ShapX Engine: A Demonstration of Shapley Value Approximations

Suchit Gupte
The Ohio State University
Columbus, Ohio, USA
gupte.31@osu.edu

John Paparrizos
The Ohio State University
Columbus, Ohio, USA
paparrizos.1@osu.edu

Abstract

Interpreting decisions made by machine learning models helps build trust in their predictions, ultimately facilitating their practical application. Shapley values have emerged as a popular and theoretically robust method for interpreting models by quantifying the contribution of each feature toward individual predictions. The inherent complexity associated with the computation of Shapley values as an NP-hard problem has driven the development of numerous approximation techniques, leading to a plethora of options in literature. This abundance of choices has created a substantial gap in determining the most appropriate approach for practical applications. To address this gap, we propose ShapX, a web engine that comprehensively evaluates 17 approximation methods across diverse regression and classification tasks. ShapX facilitates an interactive exploration of the strengths and limitations of various Shapley value approximations by guiding users through the suitable selections of replacement and tractable estimation strategies. Ultimately, our study reveals that strategies competent at capturing all the feature interactions leading to accurate estimations of Shapley values. ShapX also allows users to effortlessly upload their own dataset along with the corresponding machine learning model, enabling them to obtain detailed individualized explanations. A detailed walkthrough video of the demonstration is available online[1].

CCS Concepts

• **General and reference** → **Empirical studies**; Surveys and overviews; • **Computing methodologies** → *Feature selection*; • **Information systems** → Data model extensions.

Keywords

Shapley Values; Shapley Value Approximations; Data-Centric AI

ACM Reference Format:
Suchit Gupte and John Paparrizos. 2025. ShapX Engine: A Demonstration of Shapley Value Approximations. In *Companion of the 2025 International Conference on Management of Data (SIGMOD-Companion '25), June 22–27, 2025, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3722212.3725135

1 Introduction

Machine learning (ML) and artificial intelligence (AI) have witnessed significant advances in recent decades. The deployment of ML models to solve real-world problems has increased due to their ability to outperform humans in terms of efficiency. The application of ML models also extends to various life-critical domains, including healthcare [16] and criminal justice [10], where decisions must be accurate, fair, and transparent. A viable strategy for building confidence in ML models is interpretability, which refers to understanding and explaining their decision-making processes. However, as increasingly complex architectures like neural networks [19] gain prominence in pursuit of higher accuracy, the challenge of interpretability grows. Thus, understanding which features contribute most to a model's predictions helps prioritize relevant data, simplifying dataset selection and improving training efficiency and accuracy. To satisfy this need, Shapley values [21, 27] have emerged as a leading feature explanation technique for identifying the impact of individual features in a model's decision-making process.

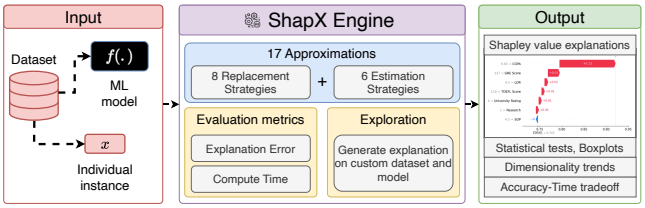

Figure 1: Overview of the Shapley explanation demo system.

The concept of Shapley values [27], originally developed in cooperative game theory, was later adopted to explain machine learning models by modeling the prediction task as a cooperative game, where each feature functions as a player contributing to the prediction. Shapley values estimate a feature's contribution across all possible subsets, capturing nuanced interactions and dependencies to enhance interpretability. Despite its intuitive appeal, Shapley values present two key challenges. The first challenge involves handling missing features when considering subsets of the feature set, as proper treatment is crucial for fair model explanations. Various replacement strategies [12, 17, 20, 21, 29, 31] address this by imputing missing values or using surrogate models to approximate the absent features' behavior. The second challenge stems from the exponential computational complexity of Shapley values, making their exact computation infeasible for high-dimensional data. To mitigate this, numerous estimation strategies [3, 21, 25, 28] have been developed to approximate Shapley values efficiently in polynomial time, balancing computational feasibility with accuracy.

The abundance of approximations highlights the credibility of Shapley values as a reliable technique in model explanations. However, despite significant progress over many years, there is a notable absence of a comprehensive evaluation of these methods in existing literature. To address this gap, we propose ShapX, a novel web engine to facilitate the exploration of model explanations through Shapley values. ShapX is based on the first comprehensive evaluation [14] of the various Shapley value approximations. Specifically,

[1]Video link: https://youtu.be/5uPocjPUAA8

SIGMOD-Companion '25, Berlin, Germany

ACM ISBN 979-8-4007-1564-8/2025/06
https://doi.org/10.1145/3722212.3725135

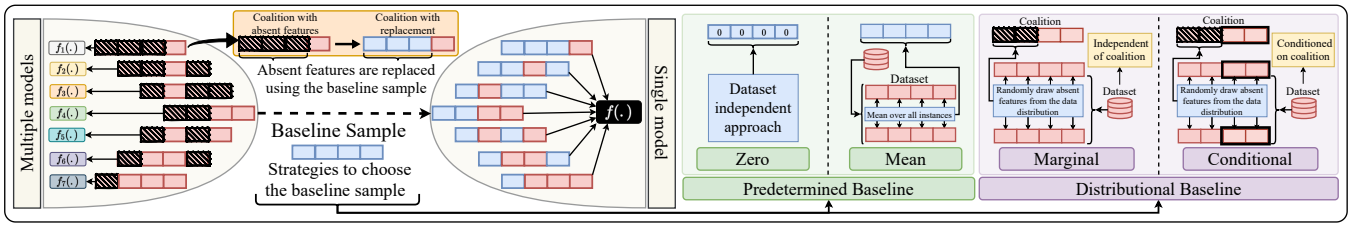

Figure 2: Replacement strategies such as Predetermined Baseline and Distributional Baseline address the absence of features, eliminating the necessity to train an exponential number of models and mitigating computational complexity.

we view Shapley value approximations as a blend of two facets: replacement strategies and tractable estimation strategies, followed by a thorough evaluation of each facet. The ShapX engine enables users to compare these individual facets, facilitating informed strategy selection. Furthermore, ShapX offers the choice to receive personalized explanations for their models based on the most compelling explanation technique identified through statistical analysis. Figure 1 highlights the overview of the ShapX Engine.

2 PRELIMINARIES

Through this section, we lay the foundation for the subsequent content of the paper. We first introduce Shapley value estimation and approximate solutions addressing each facet of Shapley value estimation (Section 2.1), followed by a detailed overview of the evaluation framework (Section 2.2).

2.1 Shapley value estimation

A fundamental supervised machine learning framework involves training a black-box model f on a dataset consisting of features $x_1, \ldots x_d$, where f makes predictions for unknown instances. The most accurate understanding of any black-box ML model is provided by employing the model itself. However, complex models make model interpretation challenging. Shapley values provide a promising solution to this challenge by treating the prediction task as a coalition game and attributing contributions to each input feature towards the prediction. This approach comprehensively explains individual predictions, inducing trust in the model's decision-making process. Formally, given an explicand x^e, feature set D, and a coalition of the feature set $S \subseteq D$, then the Shapley value of input feature i can be expressed as follows:

$$\Phi_i = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} [f(x^e_{S \cup \{i\}}) - f(x^e_S)] \quad (1)$$

The total contribution of feature i is the average marginal contribution of feature i over all possible feature coalitions $S \subseteq D$. However, accurately estimating Shapley values requires knowing the model prediction for every subset $S \subseteq D$, which poses a challenge as the original model is trained on D and does not provide predictions for an arbitrary S. Consequently, there arises a need to train an exponential number of models, which is computationally infeasible for high-dimensional data. To address the identified issue, as depicted in Figure 2, there exist several replacement strategies tailored to manage the absence of features $i \in D \setminus S$.

Despite adopting replacement strategies, Shapley values still necessitate addressing an exponential number of feature subsets.

Table 1: A detailed list of approximations, classified based on estimation and replacement strategies. The approximations form an essential component of our evaluation. "M" denotes replacement via Marginal distribution, while "C" represents Conditional distribution. The "Language" column signifies the implementation language of each approximation.

	Approaches	Estimation	Replacement	Language
Model-agnostic	Exhaustive sampling	Exact	Separate models	*Python*
	IME [28]	RO	Empirical (M)	*Python*
	CES [29]	RO	Empirical (C)	*Python*
	Cohort [23]	RO	Empirical (C)	*Python*
	MLE [25]	MLE	Empirical (M)	*Python*
	Kernel [7]	WLS	Empirical (M)	*Python*
	SGD-Shapley [13]	WLS	Mean	*Python*
	Parametric [12]	WLS	Gaussian/Copula	*Python/R*
	Non-Parametric [12]	WLS	Empirical (C)	*Python/R*
Model-specific	Linear [4]	Linear	Empirical (M)	*Python*
	Correlated Linear [4]	Linear	Gaussian	*Python*
	Tree interventional [20]	Tree	Empirical (M)	*Python*
	Tree path-dependent [20]	Tree	Empirical (C)	*Python/C++*
	DeepLIFT [26]	Deep	All-zeros	*Python*
	DeepSHAP [5]	Deep	Empirical (M)	*Python*
	DASP [1]	Deep	Mean	*Python*

Various tractable estimation strategies offer a pragmatic solution by approximating Shapley values in polynomial time. These tractable estimation strategies, along with the replacement strategies, form a foundation for the various Shapley value approximations. These approximations can be broadly classified into model-agnostic and model-specific approximations. Model-agnostic approximations can be applied to any model regardless of their type. Model-specific approximations are designed to provide an edge by utilizing that specific model's properties. We offer a comprehensive list of the approaches falling under each category in Table 1.

2.2 Evaluation Framework

Datasets: We focus on tabular datasets curated for regression and classification problems. We utilize 200 publicly available datasets from the UCI Machine Learning Repository [22]. Within the datasets, there are as many as 60 input features, and the number of instances ranges from 100 to 1 million.

ML Models: We utilize the supervised machine learning framework used to tackle regression and classification tasks. We use the following model architectures - Linear models [18], Ensemble Learning [2], Gradient Boosting [6], Neural Networks [15], Nearest neighbors [9], Naive Bayes classifiers [30], and Support Vector Machines [8]. To conduct a thorough evaluation, we integrate models

representing each category. Shapley values intend to explain a black box model by leveraging the model itself, negating the significance of its fit quality. Consequently, this allows us to use vanilla versions of each model with default hyperparameters.

Evaluation metrics: To comprehensively assess algorithm performance across diverse dimensions, we establish performance rankings by the Friedman test [11], followed by the posthoc Nemenyi test [24]. We employ the metric Explanation Error [21] to evaluate the accuracy of Shapley value estimates obtained by altering the replacement and estimation strategies. This metric assesses accuracy based on the additive nature of Shapley values.

Since Shapley values are a local feature attribution technique, we compare the instance-wise computational efficiency of different approaches. We anticipate the trend of the compute time results as the dimensionality increases. This comparison is useful in determining which replacement and estimation strategies are most suitable for handling high-dimensional data.

3 System Overview

We introduce the novel ShapX Engine,[2] a modular web engine crafted to enhance the exploration of the multifaceted Shapley value estimation. As demonstrated in Figure 3, the engine comprises five primary frames: **(a) Description, (b) Benchmark Details, (c) Accuracy Evaluation, (d) Compute Time, and (e) Interactive Explanations**. The Description frame provides a compelling rationale for the necessity of a web engine and presents a comprehensive user guide to assist individuals in effectively navigating through the engine. The Benchmark Details provide essential details like the various approximations integrated into the engine, diverse evaluation metrics employed, and the datasets and models utilized to demonstrate the evaluation.

Within the Accuracy Evaluation frame, the focus is on analyzing the performance variability across different dimensions of Shapley value estimation. This frame presents a boxplot and a critical difference diagram for comparing relative accuracy rankings to aid in visualization. Moreover, it supports the investigation of performance variability through diverse replacement and tractable estimation approaches, facilitating a more thorough exploration. In the Compute Time frame, we display the computation time of Shapley values per instance. We offer a selection of replacement and estimation strategies for detailed comparison. The results are presented through visualizations, including bar plot comparison, line plots illustrating the impact of dimensionality, and bubble plots demonstrating the tradeoff between accuracy and compute time.

The Interactive Explanations frame helps users understand their dataset better and provides explanations for any model trained on that dataset. Users are required to upload a CSV data file and a pickle model file, followed by selecting the instance to be explained. The engine then generates Shapley values for the instance using the most effective explanation technique and produces a plot illustrating the Shapley values for each feature.

4 DEMONSTRATION SCENARIOS

This section encompasses four demonstration scenarios designed to aid users in exploring the evaluation framework. The primary objectives of this demo are: (i) summarizing existing research on model explanations using Shapley value estimations (Scenario 1); (ii) comparing the accuracy of different aspects of Shapley value estimations using (Scenario 2); (iii) understanding the computation time of the multifaceted Shapley value approximations, the influence of dimensionality on computation time, and the tradeoff between accuracy and time (Scenario 3); and (iv) enabling users to receive personalized explanations for custom datasets and models, facilitating direct interaction with the framework (Scenario 4).

Scenario 1: An introductory gateway to Shapley value explanations. As shown in Frame 3 (a), this scenario presents the fundamental principles of Shapley values and emphasizes the significance of assessing these explanations from different perspectives. More specifically, it examines the complexities of dealing with missing feature values and investigates the various approaches used to estimate Shapley values effectively. Through establishing this foundational knowledge, we aim to equip users with a comprehensive understanding of Shapley values and the multifaceted criteria used to evaluate their effectiveness.

Scenario 2: A meticulous evaluation of Shapley value approximation methods in terms of accuracy. By employing boxplots and statistical critical diagrams, we seek to facilitate an intricate examination of the performance exhibited by different approximation techniques across a broad spectrum of datasets and model types. Through visual representations like those presented in Frame 3 (b), we empower users to discern trends and make informed decisions regarding adopting the most suitable approximation for their specific application scenario.

Scenario 3: Computational aspects of the multifaceted Shapley value estimation. As presented in Frame 3 (c), we thoroughly assess compute time, offering per-instance comparisons and evaluating the scalability of different approximation methods. We analyze the impact of dimensionality on computational performance. Additionally, we consider the tradeoff between computational efficacy and accuracy. We aim to give users valuable insights into the practical implications of using Shapley value approximation techniques.

Scenario 4: An interactive exploration of custom datasets and models. The Interactive Explorations Frame 3 (d) allows users to upload their datasets and models, thereby facilitating personalized explanations for individual instances. Leveraging the most effective explanation technique identified through rigorous statistical analysis, we enable users to obtain tailored insights pertinent to their specific data and models. The generated explanation plot illustrates the contribution of individual features to the model prediction using Shapley values. Features are arranged based on the magnitude of their Shapley values, with colors indicating the sign of the values: red signifies a positive impact on the model prediction, while blue indicates a negative impact.

5 Conclusion

Through this paper, we present a web-based engine to aid the comprehensive evaluation of Shapley value explanations. The interactive demonstration offers users the opportunity to explore various facets of Shapley value estimation. We provide valuable insights regarding the effectiveness and applicability of different approximation techniques across diverse datasets and model structures.

[2]Available online: https://shapleyexplanations.streamlit.app/

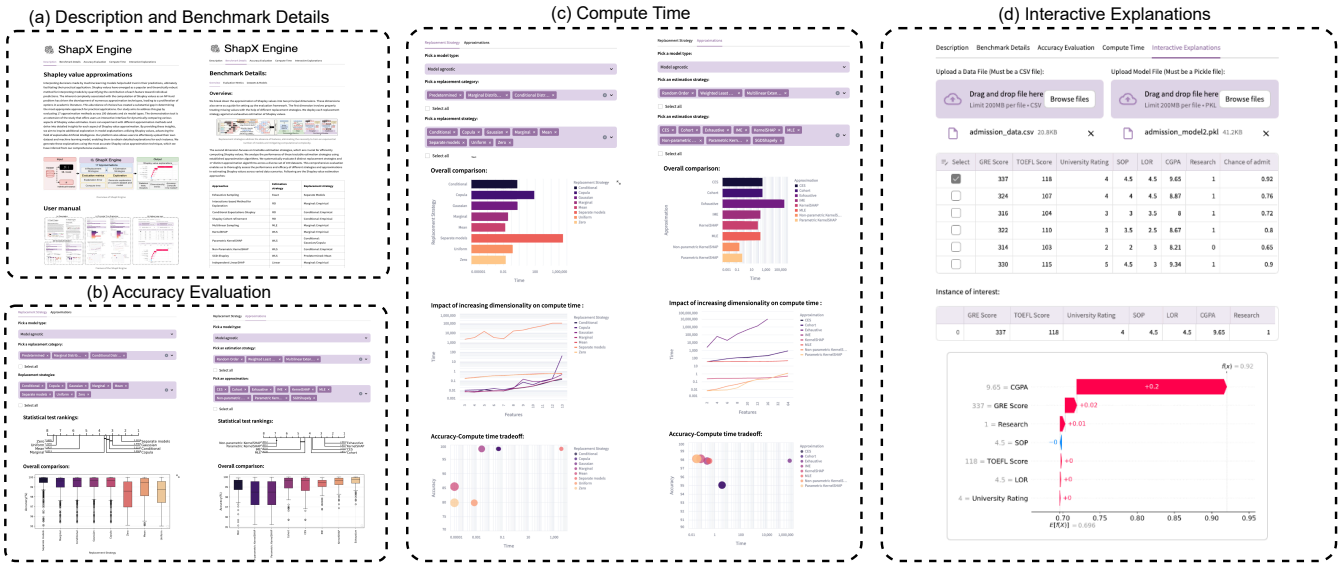

Figure 3: Frames of the ShapX Engine. A detailed walkthrough of the demonstration can be found here: Video Link

Furthermore, users have the opportunity to receive personalized explanations for their models utilizing the most compelling explanation technique determined through our statistical analysis. We hope that this interactive GUI provides users with valuable insights and sparks more progress in the field.

References

[1] Marco Ancona, Cengiz Öztireli, and Markus Gross. 2019. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation. arXiv:1903.10992 [cs.LG]

[2] L. Breiman. 2001. Random Forests. *Machine Learning* 45, 5–32 (2001). doi:10.1023/A:1010933404324

[3] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers& Operations Research* 36, 5 (2009), 1726–1730. doi:10.1016/j.cor.2008.04.004 Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[4] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. 2020. True to the Model or True to the Data? arXiv:2006.16234 [cs.LG]

[5] Hugh Chen, Scott M. Lundberg, and Su-In Lee. 2022. Explaining a series of models by propagating Shapley values. *Nature Communications* 13, 1 (Aug. 2022). doi:10.1038/s41467-022-31384-3

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM. doi:10.1145/2939672.2939785

[7] Ian Covert and Su-In Lee. 2021. Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression. arXiv:2012.01536 [cs.LG]

[8] Nello Cristianini and Elisa Ricci. 2008. *Support Vector Machines.* Springer US, Boston, MA, 928–932. doi:10.1007/978-0-387-30162-4_415

[9] Pádraig Cunningham and Sarah Jane Delany. 2021. k-Nearest Neighbour Classifiers - A Tutorial. *Comput. Surveys* 54, 6 (July 2021), 1–25. doi:10.1145/3459665

[10] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. doi:10.1126/sciadv.aao5580 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aao5580

[11] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* 32 (1937), 675–701.

[12] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. 2021. Shapley explainability on the data manifold. arXiv:2006.01272 [cs.LG]

[13] Simon Grah and Vincent Thouvenot. 2020. *A Projected Stochastic Gradient Algorithm for Estimating Shapley Value Applied in Attribute Importance.* 97–115. doi:10.1007/978-3-030-57321-8_6

[14] Suchit Gupte and John Paparrizos. 2025. Understanding the Black Box: A Deep Empirical Dive into Shapley Value Approximations for Tabular Data. *Proceedings of the ACM on Management of Data* 3, 3, Article 232 (June 2025), 31 pages. doi:10.1145/3725420 SIGMOD.

[15] Simon Haykin. 1994. *Neural networks: a comprehensive foundation.* Prentice Hall PTR.

[16] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. 2019. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery* 9 (2019). Issue 4. doi:10.1002/widm.1312

[17] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. 2021. Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations. arXiv:2103.01890 [stat.ML]

[18] Kamal Kasmaoui. 2019. *Linear Regression.* Springer International Publishing, Cham, 1–11. doi:10.1007/978-3-319-31816-5_478-1

[19] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521, 436–444 (2015). doi:10.1038/nature14539

[20] Lundberg, S.M., G. Erion, and H. et al. Chen. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2, 56–67 (2020). doi:10.1038/s42256-019-0138-9

[21] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[22] Kolby Nottingham Markelle Kelly, Rachel Longjohn. [n. d.]. The UCI Machine Learning Repository. ([n. d.]).

[23] Masayoshi Mase, Art B. Owen, and Benjamin Seiler. 2020. Explaining black box decisions by Shapley cohort refinement. arXiv:1911.00467 [cs.LG]

[24] Peter Nemenyi. 1963. *Distribution-free Multiple Comparisons.* Ph. D. Dissertation. Princeton University.

[25] Ramin Okhrati and Aldo Lipani. 2020. A Multilinear Sampling Algorithm to Estimate Shapley Values. arXiv:2010.12082 [cs.LG]

[26] Jacob Reiter. 2020. Developing an Interpretable Schizophrenia Deep Learning Classifier on fMRI and sMRI using a Patient-Centered DeepSHAP. https://api.semanticscholar.org/CorpusID:220050528

[27] Lloyd S. Shapley. 1988. A Value for n-person Games. https://api.semanticscholar.org/CorpusID:153629957

[28] Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. *J. Mach. Learn. Res.* 11 (mar 2010), 1–18.

[29] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. arXiv:1908.08474 [cs.AI]

[30] Vikramkumar, Vijaykumar B, and Trilochan. 2014. Bayes and Naive Bayes Classifier. arXiv:1404.0933 [cs.LG] https://arxiv.org/abs/1404.0933

[31] Matthew D Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901 [cs.CV]