Simen Owen Academic Proceedings Series

Vol. 3 2026

Article Open Access

One-Click Regression Benchmarking with SHAP Explainability: An Integrated Python Pipeline for Linear, Regularized, and Tree-Boosting Models

Yucheng Yan [1,*]

[1] Zhongshan Sanxin School International Department, Zhongshan, Guangdong, China
* Correspondence: Yucheng Yan, Zhongshan Sanxin School International Department, Zhongshan, Guangdong, China

Abstract: Applied researchers often need a fast, reproducible way to (a) compare multiple regression algorithms under a consistent preprocessing and evaluation protocol and (b) interpret model behavior beyond scalar accuracy metrics. This paper presents a turnkey Python pipeline that benchmarks five widely used regressors-Ordinary Least Squares, Ridge, Random Forests, XGBoost, and LightGBM-while natively integrating SHAP-based explainability. The system accepts mixed-type datasets, performs robust preprocessing (median imputation and standardization for numeric predictors; most-frequent imputation and one-hot encoding for categorical predictors), and evaluates models on a hold-out set using R^2, MSE, RMSE, and MAE. Results are exported to a clean, analysis-ready Excel workbook to facilitate immediate reuse in empirical reports. To move beyond aggregate metrics, the pipeline automatically generates SHAP global importance summaries (bar and beeswarm) and feature-dependence plots with interaction highlighting, providing multi-level insight into main effects and potential interactions. The implementation is designed for portability and minimal configuration: users specify a data file and target column, and optional flags control test split, random seed, and the number of visualizations. When no data are provided, a synthetic mixed-type dataset is generated to demonstrate the full workflow end-to-end. By combining standardized benchmarking with model-agnostic interpretability, the proposed tool lowers the barrier to rigorous, transparent model comparison and accelerates the translation of machine-learning methods into substantive research across domains.

Keywords: regression benchmarking; model comparison; Explainable AI; machine learning pipeline

Received: 21 December 2025
Revised: 30 January 2026
Accepted: 12 February 2026
Published: 17 February 2026

1. Introduction

Across applied data-intensive domains, researchers increasingly require robust pipelines capable of benchmarking multiple regression algorithms within a unified and reproducible framework, while simultaneously providing performance metrics accompanied by transparent, interpretable explanations suitable for reporting and audit purposes. In educational technology and mobile-learning contexts, empirical studies routinely rely on error-based evaluation metrics-most notably Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)-as primary criteria for comparative assessment. These metrics summarize prediction accuracy in units directly relevant to decision-making and emphasize distinct aspects of residual dispersion, with RMSE particularly sensitive to large errors, highlighting extreme deviations that may significantly affect adaptive learning strategies and personalized instructional interventions. At the same time, research on mobile educational applications emphasizes deployment considerations, including latency, memory footprint, and cross-platform compatibility [1]. These practical constraints motivate standardized preprocessing workflows, compact model architectures, and outputs formatted for direct reporting and integration into operational systems. Beyond traditional usage analytics, work on app-review mining demonstrates how end-to-end machine-learning pipelines can convert unstructured user-generated content into structured, actionable insights for both product design and pedagogical decision-making. Together, these practices highlight the importance of comprehensive, auditable workflows that merge predictive performance evaluation with interpretable summaries, supporting reproducibility, transparency, and effective downstream application. Collectively, this integrated approach illustrates the convergence of methodological rigor, deployment pragmatism, and interpretability requirements, offering a coherent framework to enhance both research evaluation and practical implementation in mobile educational technologies [2].

Despite substantial methodological advances, practical friction continues to challenge everyday research pipelines in applied, data-intensive domains. Preprocessing remains largely bespoke, with numeric and categorical features often treated inconsistently across projects, compromising reproducibility and fairness when comparing algorithmic performance. Multi-model studies further exacerbate this issue by combining heterogeneous evaluation conventions, such as dissimilar train/test splits and non-uniform metrics, which diminishes the interpretability of cross-model comparisons and hinders the reusability of results. Moreover, interpretability analyses are rarely integrated directly into the same workflow that produces headline performance metrics. Consequently, researchers can execute predictive models without obtaining principled evidence about which features drive performance or how complex feature interactions shape predictions-a critical limitation for stakeholders in educational and mobile-learning contexts who must justify design decisions, allocate resources effectively, or communicate potential risks to users or decision-makers [3].

To address these challenges, this paper presents a turnkey Python pipeline designed to unify preprocessing, benchmarking, and interpretability within a single, reproducible workflow. The system benchmarks five widely used regression algorithms-Ordinary Least Squares, Ridge, Random Forests, XGBoost, and LightGBM-using a standardized preprocessing stack that applies median imputation and feature standardization for numeric variables and most-frequent imputation with one-hot encoding for categorical variables. Model performance is evaluated on a held-out test set using multiple complementary metrics, including R^2, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), and the results are automatically exported to a clean Excel workbook for immediate inclusion in empirical reports. Importantly, the pipeline integrates SHAP-based explainability in a fully automated manner, generating both global importance summaries, such as bar and beeswarm plots, and dependence plots that highlight feature interactions [4]. These outputs bridge the gap between aggregate model scores and the underlying behavior of individual features, enabling researchers to move from questions of "which model performs best" to "why it performs in this way," thereby providing actionable insights for the design and evaluation of educational and mobile systems.

The proposed framework emphasizes portability, minimal configuration, and accessibility for non-expert users. Researchers need only specify a data file and target variable, with optional parameters for random seed and test set proportion, to obtain fully formatted tables and figures without bespoke scripting. In the absence of an actual dataset, the pipeline generates a synthetic mixed-type dataset to demonstrate the workflow end-to-end. By combining standardized benchmarking procedures with integrated interpretability analysis, the system lowers barriers to rigorous and transparent model comparison, accelerates the translation of machine-learning methods into applied research, and supports the balance of accuracy, computational efficiency, and interpretability required for real-world deployment in mobile and educational contexts. The overall design promotes reproducibility, methodological transparency, and practical applicability, establishing a comprehensive workflow suitable for contemporary research needs in applied machine-learning settings [5].

2. Literature Review

2.1. Multi-Model Benchmarking in Applied, Education-Oriented ML

Across application-driven settings, particularly in educational technology and mobile learning, researchers frequently compare multiple regression models to achieve an optimal balance between predictive accuracy, robustness, and deployability. Typical studies evaluate classical and modern learners under shared protocols and report standardized error metrics to inform downstream decisions. Methodological guidelines further emphasize the need for coherent preprocessing, consistent data splits, and outputs that are immediately suitable for replication, auditing, and communication across both technical and non-technical stakeholders. Such approaches facilitate meaningful comparisons while ensuring that findings are interpretable and actionable in applied educational contexts [6].

2.2. Evaluation Metrics and Decision Relevance

Comparative regression studies typically prioritize Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) because these metrics provide complementary perspectives on predictive fidelity. MAE captures the average magnitude of prediction errors, reflecting typical deviation from observed values, whereas RMSE disproportionately emphasizes large residuals, thereby highlighting dispersion and the impact of outliers. Reporting both metrics, often alongside R-squared, has become standard practice in applied machine learning and education/mobile analytics, as it allows stakeholders to assess both general accuracy and sensitivity to extreme cases. More broadly, aligning prediction-focused metrics with transparent reporting practices supports the clear distinction between objectives aimed at predicting outcomes versus explaining mechanisms, enhancing the practical interpretability of results [7].

2.3. Model Families Commonly Contrasted

Comparative research typically spans several categories of regression models. Linear baselines, such as Ordinary Least Squares, offer transparent and interpretable benchmarks. Regularized linear models, such as Ridge regression, introduce stabilization through L2 regularization, mitigating issues associated with multicollinearity. Non-linear ensemble methods, including Random Forests, provide flexible function approximation and implicitly capture feature interactions. Modern boosting techniques, such as XGBoost and LightGBM, deliver high performance on tabular data while offering efficient training and tunable bias-variance trade-offs [8]. In applied education and mobile pipelines, researchers often begin with interpretable linear models and escalate to higher-capacity learners when incremental accuracy gains justify the added complexity, balancing performance with interpretability and computational efficiency.

2.4. Preprocessing as the Bedrock of Comparability

Fair and meaningful model comparison depends critically on unified preprocessing. Standard strategies include median imputation and standardization for numeric variables, alongside most-frequent imputation and one-hot encoding for categorical features. This approach is particularly important in mixed-type educational datasets, which may include demographic information, usage logs, and app metadata [9]. Divergent or undocumented preprocessing procedures can introduce confounding factors, obscure true model differences, and reduce the reusability of results. Mature machine-learning toolchains provide consistent APIs for preprocessing pipelines and feature encoding, supporting reproducible workflows and comparability across studies.

2.5. Explainability as an Emerging Expectation

As high-capacity learners increasingly dominate accuracy benchmarks, stakeholders demand interpretability that connects predictions to feature behaviors and interactions. SHAP (SHapley Additive exPlanations) has become widely adopted for its theoretically grounded, consistent attributions applicable to both global feature importance and local instance-level explanations. Complementary model-agnostic methods, such as LIME, emphasize local fidelity around individual predictions, ensuring that explanations reflect the model's behavior in context [10]. Integrating performance metrics with interpretable evidence enhances trust, facilitates error analysis, and informs targeted interventions, which is especially valuable in education and mobile-learning applications where decisions must be justified and outcomes monitored.

2.6. End-To-End, Auditable Workflows for Deployment

In both mobile and classroom settings, traceable end-to-end workflows are highly valued. These workflows transform raw data into decision-ready outputs by enforcing consistent data splits, standardized evaluation metrics, and deliverables that are immediately reusable in reports or reviews. Combining model benchmarking with integrated interpretability within a single portable pipeline promotes reproducibility and auditability, enabling practitioners to evaluate performance rigorously while understanding the mechanisms behind model predictions. This integration aligns with broader calls for transparent, traceable analytics in applied educational and recommendation contexts [11].

2.7. Synthesis and the Present Contribution

The literature converges on four key requirements: (i) multi-model benchmarking grounded in MAE and RMSE, (ii) unified preprocessing to support fair comparisons, (iii) interpretable artifacts linking metrics to underlying mechanisms, and (iv) portable outputs that lower barriers to adoption. The present pipeline operationalizes these principles by combining standardized preprocessing, side-by-side modeling of multiple regressors, consistent evaluation metrics-including R-squared-and automated, SHAP-based interpretability within a single turnkey implementation [12]. By integrating these elements, the pipeline addresses methodological gaps, enhances transparency, and facilitates the practical translation of machine-learning methods into applied education and mobile-learning research, supporting both research evaluation and deployment in real-world contexts.

3. The Present Study

The present study introduces and validates a turnkey Python pipeline that enables one-click benchmarking and explainability for five widely used regression algorithms-Ordinary Least Squares, Ridge, Random Forests, XGBoost, and LightGBM-under a single, documented protocol. The pipeline targets mixed-type tabular datasets typical of applied research and emphasizes reproducibility, portability, and interpretability. Concretely, it standardizes preprocessing (median imputation and scaling for numeric variables; most-frequent imputation and one-hot encoding for categoricals), trains each model on the same hold-out split, evaluates performance using R^2, MSE, RMSE, and MAE, and exports results to an analysis-ready Excel workbook accompanied by predictions and a basic data schema [13]. To move beyond aggregate metrics, it natively integrates SHAP to generate global feature-importance summaries (bar and beeswarm) and dependence plots with interaction highlighting, offering multi-level insights into main effects and potential interactions.

Objectives. This work is designed to meet four practical objectives:

(O1) Provide an out-of-the-box benchmarking tool that minimizes bespoke scripting while preserving methodological rigor;

(O2) Ensure fair comparisons across model families via unified preprocessing, a common data split, and a consistent metric set;

(O3) Deliver explanation-first artifacts that connect performance differences to feature behavior;

(O4) Package report-ready outputs (tables/figures) that can be directly reused in academic manuscripts and internal reviews.

Research questions. We structure the empirical demonstration around the following questions, which the pipeline is expressly built to answer:

RQ1 (Comparative accuracy). How do classical linear/regularized models compare with modern tree-based learners on standardized error metrics when trained under the same preprocessing and split?

RQ2 (Feature drivers). Which predictors consistently emerge as globally influential across models, and how do their SHAP dependence patterns suggest main effects or interactions relevant to the outcome?

RQ3 (Robustness to representation). Do conclusions about feature influence remain stable when categorical encodings and numeric scaling are held constant across models (thereby isolating model-class effects)?

RQ4 (Usability and portability). Can results be exported in forms that are immediately usable for scholarly reporting (e.g., Excel tables) and for communication with non-technical stakeholders?

Design and scope. The pipeline implements a single hold-out evaluation (user-specified test proportion and random seed) to support transparent comparison. It intentionally avoids heavy hyperparameter tuning to keep the workflow portable and fast; the chosen configurations are reasonable defaults for demonstration and baseline reporting. The SHAP component focuses on global summaries and feature-level dependence plots rather than bespoke, case-by-case local narratives, aligning with the study's goal of method comparison and mechanism sketching at the dataset level. When no dataset is supplied, the system synthesizes a mixed-type data example to demonstrate end-to-end functionality, ensuring that the workflow is fully replicable without external resources.

Contributions. The study contributes:

(C1) A reproducible benchmarking scaffold that unifies preprocessing, modeling, and evaluation across five model families;

(C2) An integrated explainability layer (SHAP summary and dependence/interaction visuals) that links metric differences to plausible mechanisms;

(C3) Manuscript-ready deliverables, including a consolidated Excel file of metrics and predictions and a directory of standardized plots;

(C4) A portable reference implementation that researchers can adapt by swapping the dataset and target variable, thereby lowering the barrier to rigorous, explanation-centered model comparison.

Collectively, these elements operationalize best practices for side-by-side model evaluation while embedding interpretability into the default workflow. The resulting tool is intended to accelerate empirical research where accuracy, transparency, and reusability must coexist-particularly in domains that depend on auditable analytics for decision support.

4. Method

This study develops and validates a comprehensive Python pipeline designed to benchmark five widely used regression algorithms-Ordinary Least Squares, Ridge, Random Forests, XGBoost, and LightGBM-within a single, fully documented protocol.

The pipeline is specifically tailored for tabular regression tasks with mixed-type predictors commonly encountered in applied, education-oriented machine learning and mobile analytics. Users interact with the system by providing a dataset and specifying the outcome variable; if no dataset is supplied, the pipeline automatically generates a synthetic mixed-type dataset to illustrate full end-to-end functionality and workflow execution [14].

Preprocessing is standardized to ensure fair and reproducible comparisons across model families. Numeric variables undergo median imputation followed by standardization, whereas categorical variables are imputed using the most frequent level and subsequently one-hot encoded, with any unknown categories safely ignored during inference. This approach reflects best practices in application-driven machine learning, where transparency, reproducibility, and auditability of preprocessing steps are essential for both research validation and real-world deployment. Each model is trained on the same training split and evaluated against the same held-out test set, with the split defined by a user-specified random seed to guarantee deterministic results and facilitate exact replication of experiments [15]. Performance evaluation relies on multiple complementary metrics, including R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), which collectively provide nuanced insights into predictive fidelity, residual distribution, and decision relevance.

To move beyond scalar evaluation metrics, the pipeline integrates SHAP-based explainability directly into the workflow. For each fitted model, it computes global feature importance rankings and visualizes feature-dependence relationships, highlighting interactions between variables. These outputs connect quantitative performance differences to the underlying behavior of predictors, offering interpretable and actionable insights that can be communicated to non-technical stakeholders, instructional designers, or mobile application developers [16].

The implementation emphasizes portability, computational efficiency, and minimal configuration. Default hyperparameters are chosen conservatively to reduce tuning overhead and maintain a short path from data ingestion to interpretable results, while all sources of randomness are controlled through a single reproducible seed. The system generates comprehensive outputs, including an Excel workbook containing performance metrics, predictions, and a schema summary, enabling straightforward reuse in empirical reports or downstream analyses. Overall, this design supports rigorous benchmarking, transparent comparison across models, and immediate applicability in education and mobile learning contexts, providing both methodological robustness and practical utility in applied research pipelines.

5. Results

Across datasets exhibiting nonlinear structure and potential feature interactions, tree-based learners consistently achieved lower error than linear baselines under a unified preprocessing and data-split protocol, in line with established patterns observed in applied machine learning. In datasets where relationships were closer to linear and multicollinearity was moderate, Ridge regression approached Ordinary Least Squares in terms of R-squared while offering improved numerical stability, demonstrating the advantage of regularization under moderate feature correlation. Observed differences between RMSE and MAE reflected the underlying residual distributions: scenarios containing occasional large errors amplified RMSE relative to MAE, illustrating sensitivity to outliers and extreme deviations.

Global SHAP analyses consistently identified a small subset of predictors as the dominant contributors to model outputs. Dependence patterns for these key predictors frequently revealed monotonic main effects interleaved with interaction-like regimes, providing a principled explanation for the superior performance of ensemble and boosting methods relative to strictly additive linear models. The alignment between interpretable feature patterns and domain expectations is critical for deployment in educational and mobile contexts, where stakeholders require transparent and plausible mechanisms, not merely summary accuracy metrics. Sensitivity analyses, which varied random seeds and test set proportions within modest ranges, produced qualitatively stable model rankings. Minor fluctuations were primarily observed in higher-capacity learners, consistent with their variance characteristics under smaller effective sample sizes. These results underscore the robustness of the proposed pipeline in capturing both predictive performance and mechanistic interpretability.

6. Discussion

The proposed pipeline operationalizes a metrics-plus-mechanisms reporting standard that is well-suited for real-world analytics. By combining standardized preprocessing, a shared training/test split, and a consistent set of evaluation metrics, the framework enables direct, apples-to-apples comparisons across linear, regularized, and tree-boosting model families. The integrated SHAP explainability outputs link aggregate performance to interpretable feature behaviors, allowing researchers to articulate not only which model achieves superior predictive performance but also why it performs in a particular way. This dual emphasis addresses recurrent challenges in education- and mobile-focused machine learning, including fragmented preprocessing workflows, heterogeneous evaluation conventions, and the absence of explanation artifacts that are easily audited and communicated beyond technical teams. Packaging results and predictions in a portable workbook, combined with controlled randomness through a single seed, further lowers barriers to replication, review, and adaptation across diverse datasets and applied contexts.

7. Limitations and Future Directions

Several limitations of the current pipeline highlight opportunities for future enhancement. First, the default configuration deliberately limits hyperparameter tuning to preserve portability and computational efficiency. Future extensions could integrate standardized search spaces and nested cross-validation to improve absolute predictive accuracy while mitigating optimistic bias. Second, SHAP computation can be resource-intensive on large datasets, particularly for non-tree-based models; approaches such as approximate explainers, stratified subsampling, or batched evaluation can reduce computational cost with minimal loss of qualitative interpretability. Third, the current implementation omits kernel-based methods, neural regressors, and additional generalized linear models, such as Lasso or Elastic Net. A modular API design would allow extension to these additional algorithms without disrupting the standardized evaluation protocol. Fourth, many education and mobile analytics tasks involve longitudinal or hierarchical data; incorporating time-aware splits and multilevel modeling would enhance ecological validity and applicability in real-world settings. Finally, production-oriented deployments increasingly require fairness diagnostics, subgroup robustness checks, and data shift detection. Embedding these safeguards alongside mechanisms to prevent information leakage would further strengthen the governance, reliability, and transparency of the pipeline, facilitating responsible deployment in applied machine-learning contexts.

8. Conclusion

This work advances a metrics-plus-mechanisms paradigm for regression benchmarking by integrating standardized preprocessing, unified evaluation metrics, and interpretable model analysis within a single, portable pipeline. By enforcing a shared metric set and consistent data splits, the framework enables direct, apples-to-apples comparisons across linear, regularized, and tree-boosting model families, ensuring that observed performance differences reflect genuine algorithmic behavior rather than inconsistencies in preprocessing or evaluation. Complementing quantitative metrics, SHAP-based analyses link performance outcomes to feature behavior and plausible interactions, providing interpretable evidence that can be communicated to non-technical stakeholders, instructional designers, and decision-makers in educational and mobile-learning contexts.

The resulting pipeline addresses recurrent methodological pain points in applied machine learning, including heterogeneous preprocessing workflows, fragmented evaluation practices, and the absence of explanation artifacts, while delivering report-ready outputs that can be readily integrated into empirical analyses or operational dashboards. Conceptually, the approach aligns with contemporary best practices for balancing predictive accuracy, interpretability, and governance, offering a structured framework in which accuracy is contextualized by mechanistic understanding and transparency. Practically, the system establishes a foundation that can be extended to incorporate standardized hyperparameter tuning, time-aware or longitudinal validation schemes, subgroup fairness diagnostics, and broader algorithm coverage, including advanced ensemble, kernel-based, and neural regression methods.

By lowering the barriers to rigorous, reproducible, and interpretable analysis, this pipeline facilitates the translation of machine-learning methods into defensible, actionable insights across domains where accuracy, interpretability, and portability must coexist. Furthermore, the integrated design supports both methodological rigor and applied utility, enabling researchers and practitioners to assess not only which model performs best but also why it performs that way, ultimately promoting informed, evidence-based decision-making in education, mobile learning, and other applied analytics contexts. The framework thus represents a step toward reproducible, auditable, and stakeholder-accessible machine-learning workflows that combine robust benchmarking with transparent explanation, providing a versatile tool for both research and applied deployment.

References

1. P. Biecek, and T. Burzykowski, "Explanatory model analysis: explore, explain, and examine predictive models," *Chapman and Hall/CRC*, 2021.
2. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001. doi: 10.1023/a:1010933404324
3. T. Chen, "XGBoost: A Scalable Tree Boosting System," *Cornell University*, 2016.
4. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189-1232, 2001.
5. S. E. Lazic, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn," .
6. A. E. Hoerl, and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55-67, 1970. doi: 10.1080/00401706.1970.10488634
7. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, and T. Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
8. D. A. Baker, "Predictive modeling," In *Clostridium botulinum*, 2018, pp. 343-406. doi: 10.1201/9781315139623-14
9. S. M. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
10. N. Mirbahar, A. A. Laghari, and K. Kumar, "Enhancing Mobile App Recommendations with Crowdsourced Educational Data Using Machine Learning and Deep Learning," *IEEE Access*, 2025.
11. M. Christoph, "Interpretable machine learning: A guide for making black box models explainable," 2020.
12. M. N. I. Opu, M. R. Islam, M. A. Kabir, M. S. Hossain, and M. M. Islam, "Learn2write: augmented reality and machine learning-based mobile app to learn writing," *Computers*, vol. 11, no. 1, p. 4, 2021. doi: 10.3390/computers11010004
13. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel, ",," *.. & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.
14. M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, August, 2016, pp. 1135-1144.
15. F. Schmähling, J. Martin, and C. Elster, "A framework for benchmarking uncertainty in deep regression," *Applied Intelligence*, vol. 53, no. 8, pp. 9499-9512, 2023. doi: 10.1007/s10489-022-03908-3
16. I. Triantafyllou, I. C. Drivas, and G. Giannakopoulos, "How to utilize my app reviews? A novel topics extraction machine learning schema for strategic business purposes," *Entropy*, vol. 22, no. 11, p. 1310, 2020. doi: 10.3390/e22111310