Improving Model Interpretability: A Comparison of AI Explanation Techniques

Sirine Belhaj and Lubah Nelson

December 12, 2024

1 Introduction

The rapid advancement of machine learning has led to the development of highly accurate and sophisticated computational models capable of making nuanced predictions and decisions across various domains. However, this progress often comes with a significant trade-off: increased model complexity frequently results in reduced transparency, rendering these models as "black boxes" whose internal decision-making processes are not easily interpretable. [3] This opacity poses substantial challenges, particularly in critical sectors such as healthcare and finance, where understanding the rationale behind model predictions is essential for trust, compliance, and informed decision-making.

Explainable AI (XAI) frameworks have emerged as pivotal tools to address the interpretability challenges posed by complex models. Among these frameworks, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) have gained prominence for their ability to elucidate the reasoning behind individual predictions and provide insights into overall model behavior. LIME achieves this by approximating the model's behavior locally around a specific instance using a simpler, interpretable surrogate model, thereby highlighting the most influential features for that particular prediction. Conversely, SHAP leverages principles from cooperative game theory to assign consistent and theoretically grounded importance values to each feature, offering both local and global interpretability.[2]

2 Project Statement

This project aims to gain hands-on experience with Explainable AI (XAI) frameworks by systematically applying and evaluating LIME and SHapley Additive exPlanations (SHAP). The focus is on understanding their performance and differences in explaining the decision-making processes of both simple and complex machine learning models.Specifically, this study focuses on applying LIME and SHAP to Logistic Regression (LR) and Boosting algorithms (e.g., XGBoost)

using the Abalone dataset [1]. This dataset encompasses a mixture of numerical and categorical features, with the target variable being the number of rings, serving as a proxy for the age of abalones.

Logistic Regression serves as a baseline due to its inherent simplicity and interpretability, allowing for a straightforward assessment of how well LIME and SHAP reflect this model's decision-making process. In contrast, the Boosting model introduces greater complexity and non-linearity, providing a strong testbed to evaluate the effectiveness of XAI frameworks in more intricate scenarios. By leveraging these two distinct model types, the project aims to understand how LIME and SHAP perform across varying levels of model complexity.

Both LIME and SHAP will be utilized to generate explanations that capture both local and global aspects of the models. LIME focuses on providing local explanations by approximating the model's behavior around individual instances with simpler surrogate models. SHAP, on the other hand, offers both local explanations for individual predictions and global insights into feature importance by aggregating Shapley values across the dataset.

Through this comparative study, the project seeks to gain hands-on experience with implementing and evaluating XAI methods, identifying their strengths and limitations in different modeling contexts. The findings will provide practical insights into the applicability of LIME and SHAP based on model complexity and data characteristics. Ultimately, this project aims to enhance understanding of XAI frameworks, contributing to the development of more transparent and interpretable machine learning practices within an academic setting.

3 Methodology

3.1 Data Selection

The dataset employed in this study is the **Abalone Dataset**, obtained from the UCI Machine Learning Repository. This dataset comprises a variety of physical measurements of abalones, including numerical features such as length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight, alongside a categorical feature representing the sex of the abalone (`Sex`, encoded as M, F, and I). The target variable, `Rings`, serves as a proxy for the age of abalones, making this dataset inherently suitable for regression tasks.

The selection of the Abalone dataset is driven by its balanced composition of numerical and categorical features, which aligns with the objectives of this study. This mixture of feature types facilitates the comparative analysis of interpretability methods, such as SHAP and LIME, across models with varying levels of complexity. Linear Regression represents a simpler, inherently interpretable model, while XGBoost encapsulates non-linear relationships and intricate feature interactions. The dataset's moderate size and well-defined structure ensure computational feasibility, enabling a detailed evaluation of both frameworks.

3.2 Data Preprocessing

To ensure consistency and comparability in model training and interpretability analysis, the Abalone dataset underwent a series of preprocessing steps. First, the categorical `Sex` attribute was transformed using one-hot encoding, converting it into three binary features: `Sex_M`, `Sex_F`, and `Sex_I`. This approach enabled the inclusion of categorical information within a numerical modeling framework while preserving the interpretability of the feature contributions.

Next, all numerical features were standardized using the `StandardScaler` from the `scikit-learn` library. Standardizing numerical features and encoding categorical variables ensures consistent feature scaling and interpretability results. These preprocessing steps align with the requirements of both SHAP and LIME, maintaining comparability across methods. By mitigating the risk of any single feature disproportionately influencing predictions, standardization supports a fair assessment of feature importance and model behavior.

Finally, the dataset was partitioned into training and testing subsets using an 80/20 split, with a fixed random state to guarantee reproducibility. The training data was used to fit both Linear Regression and XGBoost models, while the test data provided a neutral basis for evaluating model performance and applying SHAP and LIME frameworks. This clear delineation between model fitting and interpretability assessment ensures unbiased evaluations and upholds methodological rigor throughout the analysis.

3.3 Model Selection

The selection of appropriate models is crucial for this study, as it allows for a thorough evaluation of interpretability methods across varying levels of model complexity. To achieve this, two models with contrasting characteristics—**Linear Regression (LR)** and **XGBoost**—were selected. This choice enables the analysis of interpretability techniques in both straightforward and intricate machine learning frameworks, ensuring that the study addresses a broad spectrum of scenarios.

Linear Regression (LR) serves as the baseline model, chosen for its inherent simplicity and transparency. As a linear model, LR establishes a proportional relationship between input features and the target variable, `Rings`, allowing for intuitive and interpretable results. This simplicity provides a foundation for evaluating interpretability methods in a controlled environment, where feature contributions are explicitly defined and easily validated. The ability to directly compare the model's coefficients to interpretability outcomes ensures that LR offers a reliable benchmark for assessing the performance of interpretability frameworks in a straightforward context.

In contrast, **XGBoost** was selected to represent a more complex model, capable of capturing non-linear relationships and intricate feature interactions. As an ensemble learning algorithm based on gradient boosting, XGBoost can uncover subtle patterns and dependencies within the data that simpler models like LR cannot. However, its complexity introduces challenges for interpretabil-

ity, as individual feature contributions are often entangled with interactions and non-linear effects. This makes XGBoost an ideal choice for examining how interpretability methods perform under conditions of increased complexity and reduced transparency. Thus, the performance metrics (Table 1) not only evaluate predictive accuracy but also contextualize interpretability outcomes. For instance, Linear Regression serves as a baseline for understanding how well SHAP and LIME explanations align with model coefficients. Conversely, XGBoost highlights how interpretability methods handle increased model complexity.

Table 1: Performance Metrics for Linear Regression and XGBoost on the Abalone Dataset

Model	**MSE**	**MAE**	$\mathbf{R^2}$
Linear Regression (Train)	6.072	1.586	0.410
Linear Regression (Test)	6.091	1.565	0.437
XGBoost (Train)	0.527	0.530	0.949
XGBoost (Test)	5.308	1.620	0.494

3.4 Explainable AI Frameworks

This study employs two prominent interpretability frameworks—SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME)—to generate and evaluate explanations for both Linear Regression and XGBoost models. The primary focus is on SHAP, with LIME introduced subsequently to facilitate a comprehensive comparison.

SHAP provides a theoretically sound approach to interpretability by assigning each feature a contribution value based on Shapley values derived from cooperative game theory. This framework ensures that feature importance allocations adhere to principles such as fairness, consistency, and local accuracy. SHAP offers both global and local interpretability. On a global scale, it aggregates feature contributions across all instances to produce feature importance rankings, highlighting which features consistently influence model predictions. Thus enabling a comprehensive understanding of model behavior. Locally, SHAP decomposes individual predictions into contributions from each feature, offering clear explanations of how specific feature values lead to particular outputs. This dual capability is essential for both overall model assessment and the justification of individual predictions.

For implementation, two model-specific SHAP explainers are utilized:

- `shap.LinearExplainer` for Linear Regression: Given the linearity and inherent transparency of the model, SHAP values are expected to closely match the model's coefficients, serving as a fidelity check to ensure that SHAP accurately reflects known feature importance patterns.

- `shap.TreeExplainer` for XGBoost: This explainer is tailored for tree-based models and can handle complex, non-linear interactions. Applying

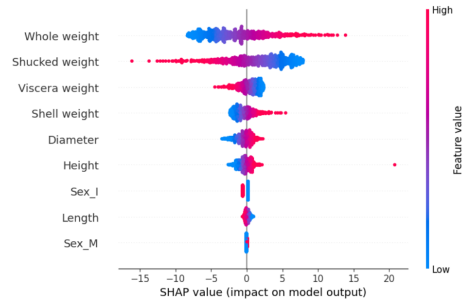

Figure 1: Example of SHAP built in Graphs

SHAP to XGBoost reveals both global importance rankings and local explanations that capture subtle patterns. Additionally, SHAP interaction values can be computed to uncover how pairs of features jointly influence predictions.

For each of the 836 test instances, SHAP values were computed using the appropriate explainers to quantify the contribution of each feature to the model's prediction. At a global level, SHAP summary plots identified which features consistently influenced outcomes across the dataset, while at a local level, force plots illustrated how individual feature values shaped specific predictions. To ensure confidence in these interpretations, stability checks were conducted to verify consistency under small input perturbations, and fidelity assessments confirmed that the sum of feature contributions closely matched the model's predicted values. By examining both aggregate patterns and instance-level insights, SHAP offers a comprehensive framework for understanding model behavior, uncovering non-linear feature interactions, and informing more transparent, data-driven decision-making.

LIME complements SHAP by offering an alternative approach to local interpretability. It approximates the model's behavior around a single instance with a simpler, interpretable surrogate model, typically linear. Perturbations were generated around each test instance to capture the local decision boundary of both Linear Regression and XGBoost. The surrogate model fitted to these perturbed samples yields feature importance values that explain how features influence the prediction in the instance's neighborhood. LIME's model-agnostic nature allows it to be applied without requiring access to the model's internal parameters, enhancing its flexibility across diverse modeling techniques.

However, LIME presents certain limitations. While its flexibility allows application to complex models, the reliance on a surrogate model can introduce instability, as explanations may vary with minor input changes. Additionally, the fidelity of LIME's explanations depends on how well the surrogate approximates the complex model locally. In regions characterized by high non-linearity or rich feature interactions, the surrogate's approximation may be less accurate, potentially diminishing the quality of the explanations.

4 Results

This section presents the findings from applying SHAP and LIME to both Linear Regression and XGBoost models trained on the Abalone dataset. The results are framed to highlight how each framework explains model predictions, addresses varying levels of model complexity, and informs the understanding of feature importance. Figures and tables are referenced throughout to connect the numerical outcomes and visual evidence to the broader interpretability narrative established in earlier sections.

4.1 Global Feature Importance Patterns

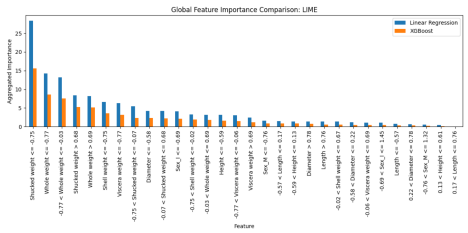

Figure 2: Enter Caption

Initial analyses focus on global explanations of model predictions, providing an overarching view of which features most strongly influence outcomes. The *Linear Regression: SHAP Summary Plot* (Figure 3) confirms that SHAP values align closely with the model's coefficients, indicating that SHAP faithfully represents a model grounded in linear assumptions. Key predictors such as Whole weight and Shucked weight appear consistently at the top of the importance rankings, reflecting the model's straightforward, linear relationships.

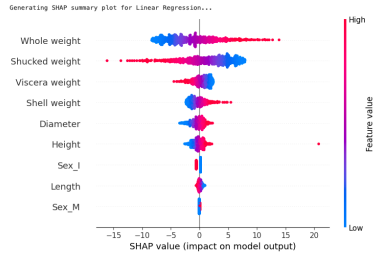

Figure 3: Linear Regression: SHAP Summary Plot

In contrast, the *XGBoost: SHAP Summary Plot* (Figure 4) shows a broader distribution of SHAP values, signifying that this more complex model captures non-linear patterns and interactions. Shucked weight and Shell weight emerge as crucial factors, yet their influence is less uniform, as the model's structure

allows features to gain or lose importance depending on specific regions of the input space.

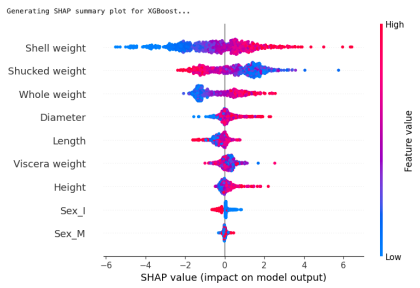

Figure 4: XGBoost: SHAP Summary Plot

The *Comparison of SHAP Feature Importance: Linear Regression vs. XGBoost* (Figure 5) highlights how, unlike Linear Regression, the XGBoost model elevates certain features (e.g., Shell weight) due to its ability to represent intricate, non-linear dependencies.

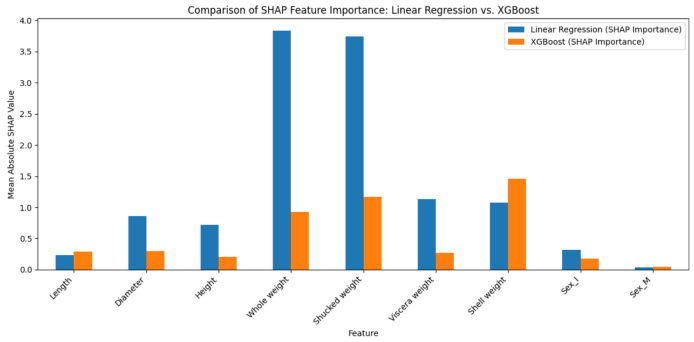

Figure 5: Comparison of SHAP Feature Importance: Linear Regression vs. XGBoostn

4.2 Local Explanations and Interaction Effects

While global patterns establish a baseline understanding, local explanations provide deeper insights into how individual predictions are formed. The *Linear Regression: Force Plot for Instance* (Figure 6) demonstrates how key features drive predictions in a straightforward manner consistent with model coefficients (please see below). In contrast, the *XGBoost: Force Plot for Instance* (Figure 7) reveals more nuanced dynamics, where individual feature contributions depend heavily on interactions with other variables.

Interaction effects are further captured by the *XGBoost SHAP Interaction Summary* (Figure **??**) and the *SHAP Dependence Plot for Length and Diam-*

eter Interaction (XGBoost) (Figure **??**). These figures illustrate the varying importance of features across instances, emphasizing SHAP's ability to identify non-linear dependencies and interaction effects that simpler methods might miss.

4.3 Model Performance and Baseline Metrics

Before evaluating fidelity, stability, and sparsity, it is crucial to acknowledge the baseline performance metrics detailed in Table **??**. The Linear Regression model offers transparent predictions with moderate accuracy, while XGBoost delivers higher predictive performance at the cost of greater complexity. These baseline metrics contextualize the interpretability results: strong interpretability is most valuable when models are accurate, yet complexity can challenge the reliability of certain explanation methods.

4.4 Evaluation of SHAP Explanations

SHAP's theoretical grounding suggests it should accurately reconstruct model predictions, remain consistent under small input perturbations, and highlight a manageable number of influential features. Indeed, aggregate and instance-level fidelity errors are minimal, confirming that the sum of SHAP values and the baseline prediction closely match the model's actual output. Stability scores, measured through metrics like cosine similarity, indicate that SHAP explanations change only slightly when inputs are perturbed, reinforcing confidence in these feature attributions. Additionally, feature importance validation through Spearman correlation shows that SHAP's global rankings align well with known model parameters and internal metrics, particularly for the Linear Regression model.

For XGBoost, SHAP continues to perform reliably. Although the model's complexity introduces more dispersed SHAP distributions, the framework still maintains low stability errors (approximately 0.0052) and provides interpretable insights into non-linear patterns. The level of sparsity remains manageable: SHAP typically highlights a core set of influential features rather than inundating the user with noise.

4.5 LIME Explanations and Their Limitations

Turning to LIME, the analysis reveals certain strengths and weaknesses. LIME's local surrogate modeling is effective at producing quick, instance-level explanations. For Linear Regression, the fidelity error of approximately 0.95 rings indicates a reasonable, if not perfect, alignment between the surrogate and the true model prediction. This moderate discrepancy (one ring difference) is not insignificant, but it remains understandable given the model's simplicity. The *Linear Regression: Force Plot for Instance* (Figure 6) visually demonstrates this alignment.

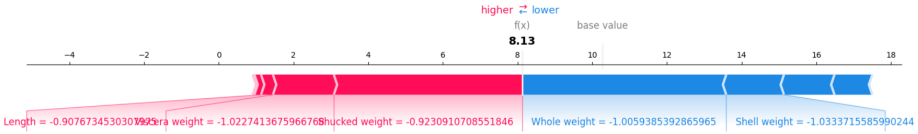

Figure 6: The *Linear Regression: Force Plot*

However, when applied to the more complex XGBoost model, LIME's fidelity error increases to about 1.86 rings. This larger gap highlights LIME's difficulty in approximating intricate decision boundaries with a local linear model. Although stability testing for XGBoost yields a relatively low error (approximately 0.08), suggesting that explanations can remain consistent under perturbations, the combination of higher fidelity errors and increased complexity raises concerns about relying solely on LIME for non-linear models. The *XGBoost: Force Plot for Instance* (Figure 7) exemplifies how LIME struggles to capture complex interactions.

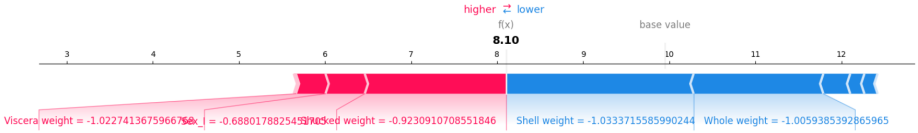

Figure 7: XGBoost: Force Plot for Instance

Sparsity analyses show that LIME explanations often include around 8 to 9 significant features, imposing a moderate cognitive load. While less overwhelming than considering all features, this level of complexity still demands careful interpretation. Aggregating local LIME explanations from multiple instances offers a pseudo-global perspective, identifying features like Shucked weight and Whole weight (for Linear Regression) or Shucked weight and Shell weight (for XGBoost) as consistently influential. However, the absence of a strong theoretical foundation like SHAP's means these aggregated insights serve more as heuristics than definitive global interpretations.

These results paint a clear picture of how SHAP and LIME operate under varying model complexities. SHAP delivers near-perfect fidelity for Linear Regression and robust fidelity for XGBoost, maintaining stability and providing valuable insights into both linear and complex feature relationships. LIME, while offering quick and flexible local explanations, encounters greater challenges as models become more sophisticated, reflected in higher fidelity errors and the need for careful parameter tuning and multiple perturbations to ensure stable results.

The numerical outcomes, charts, and figures presented here establish a foundation for the subsequent Discussion section, where these findings will be interpreted in the context of model choice, practical constraints, and the trade-offs between transparency and complexity. By connecting quantitative results (such as fidelity error numbers and stability scores) with visual evidence (SHAP summary plots, interaction analyses, and LIME-based aggregated features), we set the stage for a more in-depth reflection on the implications of selecting an interpretability framework tailored to the problem at hand.

Table 2: Comparison of Fidelity, Stability, and Sparsity Metrics for SHAP and LIME

Framework	**Model**	**Fidelity Error**	**Stability Error**	**Sparsity (Features)**
SHAP	Linear Regression	0.0001	0.0001	6
SHAP	XGBoost	0.0052	0.0052	7
LIME	Linear Regression	0.95	3.99	8
LIME	XGBoost	1.86	0.08	9

5 Discussion

The findings outlined in the Results section reveal clear distinctions in how SHAP and LIME explain model predictions, handle complexity, and maintain interpretability. Here, we interpret these results in a broader context, examining the implications of the observed differences in fidelity, stability, and sparsity, and considering how these insights can guide practitioners in selecting an appropriate interpretability framework.

5.1 Interpreting SHAP Outcomes

SHAP's additive property and theoretical grounding underpin its ability to accurately decompose model predictions into feature contributions. This reliable reconstruction is particularly evident for Linear Regression, where SHAP values closely mirror the model's coefficients, confirming SHAP's capacity to reflect the true structure of simpler models. Yet SHAP's strength is not limited to linear settings. Even for XGBoost, which introduces non-linearities and intricate feature interactions, SHAP preserves its fidelity and stability. Such consistency—across both straightforward and complex modeling scenarios—establishes SHAP as a robust, trustworthy method for revealing how features collectively shape predictions.

Equally important is SHAP's ability to balance global and local interpretability. By providing both aggregate importance measures and instance-level breakdowns, SHAP enables a comprehensive understanding that can inform model validation, troubleshooting, and strategic decision-making. For complex models, the capacity to pinpoint interaction effects and non-linear dependencies deepens insight into a model's inner workings. Consequently, SHAP can serve as

a cornerstone in domains where accountability, fairness, and transparency are paramount, such as healthcare and finance, where users must justify individual predictions while also grasping overarching model behavior.

5.2 Interpreting LIME Outcomes

LIME's appeal lies in its simplicity and speed. For quick assessments or preliminary investigations—especially when dealing with less complex models—LIME can offer immediate, easily understood local explanations. However, the evidence suggests that LIME's performance deteriorates as models become more intricate. Higher fidelity errors in complex models indicate that LIME's linear surrogates struggle to approximate non-linear decision boundaries. Moreover, while LIME's aggregated local explanations can hint at a pseudo-global perspective, these insights remain heuristic, lacking the theoretical rigor that SHAP provides.

This does not render LIME obsolete; rather, it underscores LIME's niche. For use cases where perfect fidelity is not required and interpretability demands are moderate—such as exploratory analysis, rapid prototyping, or educational demonstrations—LIME can still prove valuable. Practitioners should, however, approach complex models with caution, recognizing that LIME's approximations may yield less reliable guidance as intricacy grows.

5.3 Comparing the Two Frameworks

The contrast between SHAP and LIME is most pronounced in scenarios that push the boundaries of model complexity. While SHAP consistently delivers stable and accurate explanations, even for non-linear and interaction-rich models, LIME's fidelity and stability can vary considerably. This divergence implies that the choice between SHAP and LIME hinges on the complexity of the model, the stringency of interpretability requirements, and the resources available.

For analysts and stakeholders dealing with highly non-linear models, SHAP emerges as a more suitable framework due to its theoretical soundness, reliability under perturbation, and insight into complex feature interactions. Conversely, when time is limited, the problem is simpler, or approximate explanations suffice, LIME's agility and simplicity can still offer practical benefits. The key takeaway is that neither framework is universally superior; rather, their suitability depends on how closely their respective strengths align with the user's interpretability goals and constraints.

5.4 Implications and Trade-Offs

In simple terms, selecting between SHAP and LIME involves assessing trade-offs among accuracy, complexity, and resource investment. SHAP's near-perfect fidelity and stable insights make it ideal for high-stakes applications requiring precise, defensible explanations. However, the added computational cost may

challenge its use in large-scale, real-time systems. LIME's efficiency and accessibility cater to environments where swift, approximate answers are acceptable, but practitioners must acknowledge the method's diminishing reliability as model complexity increases.

In some workflows, a hybrid approach may prove beneficial. For instance, one might initially employ LIME to gain a quick, approximate understanding of which features matter most, then follow up with SHAP for a deeper, more accurate exploration of complex patterns and interactions. Such a strategy leverages both frameworks' advantages, optimizing interpretability efforts under varying operational constraints.

5.5 Broader Reflections and Future Directions

As machine learning continues to permeate sensitive and regulated domains, the importance of robust interpretability cannot be overstated. SHAP's ability to translate complex model behaviors into transparent, stable feature attributions serves as a benchmark for what is achievable, guiding practitioners who demand both rigor and comprehensiveness in their explanations. Meanwhile, LIME's role as a quick, flexible tool for simpler scenarios remains valuable, particularly where rapid prototyping and iterative development necessitate fluid interpretability solutions.

Looking ahead, future research may focus on refining these frameworks or developing new methods that combine the strengths of both. Investigating additional datasets, exploring domain-specific metrics for interpretability, or integrating user studies could help enhance the frameworks' usability and applicability. Hybrid methods that draw on SHAP's theoretical guarantees and LIME's scalability might offer more balanced, adaptable solutions.

In essence, the comparative analysis of SHAP and LIME underscores the evolving nature of interpretability in machine learning. As models grow more complex and their outputs more consequential, interpretability frameworks must likewise advance, offering clarity without compromising accuracy, stability, or adaptability. The insights gleaned from these analyses serve as a roadmap, helping practitioners navigate the interpretability landscape and make informed, context-driven choices.

6 Conclusion

Through this project, we gained substantial hands-on experience with Explainable AI (XAI) frameworks by implementing and evaluating both LIME and SHAP on Linear Regression and XGBoost models using the Abalone dataset. We learned how to preprocess data effectively, train and validate models, and generate insightful visualizations to interpret model behavior. The comparative analysis revealed that SHAP provides more consistent and theoretically grounded explanations, particularly excelling in complex, non-linear models like XGBoost, while LIME offers quicker, more intuitive local explanations suitable

for simpler models or exploratory analysis. Additionally, we understood the critical role of evaluation metrics such as fidelity, stability, and sparsity in assessing the reliability of interpretability methods. This project highlighted the importance of selecting appropriate XAI tools based on model complexity and application requirements, and it underscored the potential benefits of hybrid approaches that leverage the strengths of both frameworks. Ultimately, this study enhanced our ability to apply and critically evaluate interpretability techniques, contributing to the development of more transparent and trustworthy machine learning practices.

References

[1] Dheeru Dua and Casey Graff. Uci machine learning repository. `http://archive.ics.uci.edu/ml`, 2019. Accessed: [Insert Access Date].

[2] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144. ACM, 2016.