EVALUATING EXPLAINABILITY IN MACHINE LEARNING PREDICTIONS THROUGH EXPLAINER-AGNOSTIC METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence (AI) continues to transform industries and research at an accelerated pace, bringing forth numerous challenges related to transparency and accountability in AI-driven decision-making. Decision-makers and stakeholders require not only a clear understanding of how these systems generate predictions but also assurance that these processes are conducted ethically and responsibly. These challenges highlight a critical need for effective tools to evaluate and enhance the interpretability of AI models. To address this gap, we propose a new set of explainer-agnostic metrics aimed at evaluating the interpretability of AI models in the context of specific explainers. By focusing on global and local feature importance, as well as surrogate models, our metrics capture key elements such as feature stability, fluctuations in prediction behavior, and contrasts in feature relevance across conditional subsets. By quantifying these complex dynamics as clear scalar measures, we offer a structured framework for assessing model transparency, fairness, and robustness. We demonstrate the practical utility of our approach through case studies on a set of benchmark datasets, revealing deeper insights into model interpretability that facilitate more informed decision-making among AI developers and stakeholders. Ultimately, our work aims to foster AI systems that are not only technically reliable but also transparent, fair, and accountable, thereby advancing the development of ethical AI practices.

1 INTRODUCTION

Despite the remarkable recent evolution in prediction performance by artificial intelligence (AI) models, they are often deemed as "black boxes", i.e. models whose prediction mechanisms cannot be understood simply from their parameters. An explainable or interpretable algorithm is one for which the rules guiding its prediction decisions can be questioned and explained in a way that is intelligible to humans. Specifically, *interpretability* regards the ability to extract causal knowledge about the world from a model, and *explainability* pertains to the capability to articulate precisely how a complex model arrived at specific predictions, detailing its mechanics. Understanding AI models' behavior is essential for explaining predictions to support decision-making, debugging unexpected behaviors (contributing to improving model accuracy), refining modeling and data mining processes, verifying that model behavior is reasonable and fair, and effectively presenting predictions to stakeholders.

The main goal of explainable artificial intelligence (XAI) encompasses several critical objectives Ali et al. (2023). Firstly, XAI aims to empower individuals by enabling them to make informed decisions, mitigating the potential harms of fully autonomous decision-making systems. Secondly, it seeks to enhance decision-making by providing transparent information about the outputs of AI models, facilitating well-informed choices. Thirdly, XAI identifies and addresses vulnerabilities that could compromise machine learning-based systems, bolstering their resilience. Lastly, Lastly, it endeavors to boost user confidence in AI systems by promoting transparency and fostering a clear understanding of the decisions made by these models.

The literature offers various approaches to assessing explainability methods. Quantitative metrics often evaluate whether these methods meet specific quality and reliability criteria (Bodria et al., 2023). Common metrics include fidelity—how well the explanation aligns with the underlying model

(Guidotti et al., 2018), stability—whether similar inputs yield consistent explanations (Alvarez Melis & Jaakkola, 2018), faithfulness—how accurately the explanation reflects the true behavior of the model (Alvarez Melis & Jaakkola, 2018), monotonicity—whether more of a certain feature leads to a stronger explanation (Luss et al., 2021), and complexity—how easily the explanation can be understood. Qualitative assessments are similarly varied and are categorized into functionally-founded, application-grounded, and human-grounded, each offering different perspectives on the utility and interpretability of the explainability method.

However, most existing approaches are tied to specific explainability frameworks or model architectures, which limits their generalizability and usefulness in varied deployment scenarios. To address these challenges, this study introduces a novel set of explainer-agnostic metrics that can evaluate the outputs of any XAI method used for classification or regression tasks. These metrics encapsulate the behavior of explanations into a singular, concise representation, making them highly applicable in automated systems that monitor the trade-off between model accuracy and explainability. By quantifying this balance, the proposed metrics help assess both the transparency and risk associated with AI models in deployed environments.

Our main contributions are:

- Explainer-Agnostic Metrics: We propose a set of explainer-agnostic metrics designed to evaluate the outputs of any explainer, making them applicable across various model types and settings.
- Consistent Feature Evaluation: Our metrics capture fundamental behaviors of the explanations, providing a consistent framework for evaluating the feature importance of any explainer in deployment scenarios, ensuring uniformity across models and explainers.
- Publicly Available Tools: All metrics and benchmark datasets are made publicly available via a Python library, providing the community with tools to implement and further develop these methods.

The paper is organized as follows: Section 2 presents our proposed methods for explainable AI at local and global levels; Section 3 discusses a set of applications for these methods; and finally, Section 4 outlines our main findings and potential directions for future research.

2 Proposed explainer-agnostic metrics

The proposed metrics are model-agnostic and explainer-agnostic, meaning they can be applied to any type of model without requiring access to the model's internal structure or parameter estimates. These metrics only require the predictions $\hat{y}$ from a trained model f and the explanations of any explainer $\mathcal{E}$ applied in f. Additionally, our methods extend existing concepts in explainable AI (XAI), including (i) permutation feature importance, (ii) partial dependence plots, and (iii) surrogate models. By summarizing explainer outputs into a single value, proposed metrics provide a concise measure that captures essential behaviors of the explanations, enabling a consistent evaluation of model's explainations across different settings.

Figure 1 illustrates a simplified framework for explainer-agnostic metrics. Given a set of features X and a target vector Y, both the black-box model f and its surrogate $\hat{f}$ serve as inputs to the explainer $\mathcal{E}$. The nature of $\mathcal{E}$ can vary, but we assume three possible types of outputs: (1) local feature importance focuses on understanding the contribution of features in the predictions of each instance; (2) global importance provides insights into the overall model by indicating how much each feature contributes to the model's predictions across the entire dataset. The (3) surrogate importance involves approximating the complex model with a simpler, interpretable model (the surrogate model $\hat{f}$).

Following, we describe the explainer-agnostic metrics proposed.

2.1 Metrics Based on Global Feature Importance

2.1.1 Feature Importance Spread

This metric evaluates the feature importance distribution using a divergence measure. Consider a uniform distribution of feature importance $U = \{\bar{f}, \bar{f}, \ldots, \bar{f}\}$, where $\bar{f} = 1/|F|$, indicating that

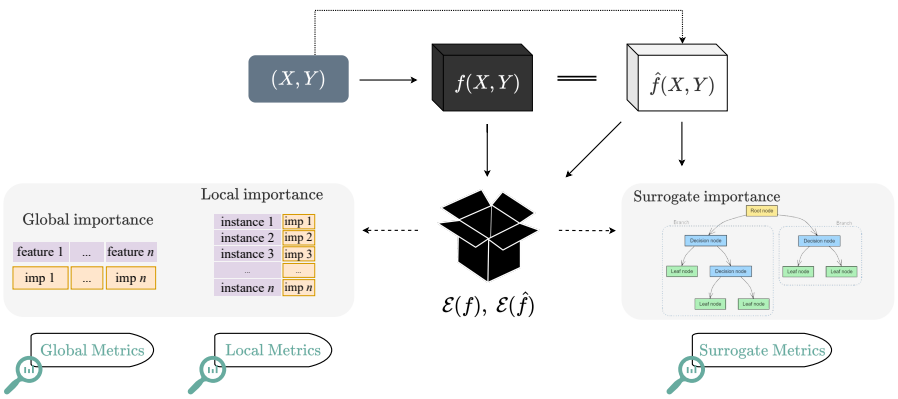

Figure 1: A simplified representation of explainer-agnostic metrics framework

all features are equally relevant. Such a distribution often suggests that understanding the model's decision-making process may be more complex, as it implies equal contributions from many features.

To quantify the deviation from this uniformity, we apply the Jensen-Shannon Divergence (JS), denoted as S_D. Unlike the Kullback-Leibler Divergence (D_{KL}), JS is symmetric and bounded between 0 and 1, making it a more interpretable measure of dissimilarity. A S_D value close to 0 indicates that the feature importance distribution is near-uniform, implying that all features contribute similarly, making it harder to discern the key driving factors in the model. On the other hand, a higher S_D value (closer to 1) indicates that feature importance is concentrated on a few features, which simplifies model interpretation by highlighting the most influential features.

Definition 2.1 (feature importance divergence)**.** Let P a normalized feature importance distribution, U a uniform distribution, and $M = \frac{1}{2}(P + U)$. The feature importance spread is defined by

$$S_D(F) = \sum_{j=1}^{F} \frac{1}{2} D_{KL}(P_j \| M_j) + \frac{1}{2} D_{KL}(U \| M_j) \tag{1}$$

2.1.2 α-FEATURE IMPORTANCE

The α-Feature Importance metric measures the smallest subset of features needed to represent at least α of the model's total feature importance.

Definition 2.2 (α-Feature Importance)**.** Let f_j represent the importance of the j-th feature, and F be the total number of features. The α-Feature Importance metric, $FI_\alpha(F)$, is the proportion of features required to capture at least α of the total importance:

$$FI_\alpha(F) = \frac{\min\left\{k \mid \sum_{j=1}^{k} f_{(j)} \geq \alpha \cdot \sum_{j=1}^{F} f_{(j)}\right\}}{F} \tag{2}$$

where $f_{(j)}$ are the ordered feature importances (from highest to lowest). α is the fraction of total importance you want to capture (e.g., $\alpha = 0.8$ for 80%). $FI_\alpha(F)$ ranges from 0 to 1, indicating the proportion of features needed. A low $FI_\alpha(F)$ means a small number of features explain most of the model's behavior. Conversely, a high $FI_\alpha(F)$ means that many features are necessary to explain the model.

2.1.3 FLUCTUATION RATIO

The Fluctuation Ratio (F_R) quantifies the oscillatory behavior present in Partial Dependence Plots (PDPs), providing a measure of the stability and interpretability of the relationship between individual features and a model's predictions. PDPs show how the predicted outcome changes with a feature while keeping all other features averaged out. However, fluctuations in PDPs can indicate unstable or complex relationships between the feature and the target variable, which may hinder interpretability.

Definition 2.3 (fluctuation ratio)**.** Given a trained model f and feature x, the partial dependence function $PD(x) = \mathbb{E}_{\mathbf{X}_{-x}}[f(x, \mathbf{X}_{-x})]$ is the expected prediction $f(x, X_{-x})$, keeping the feature x fixed and averaging over all possible combinations of the other features X_{-x}. To compute the fluctuation ratio (F_R):

$$\Delta PD_i = PD(x_{i+1}) - PD(x_i) \textbf{ and } D_i = \text{sign}(\Delta PD_i) \tag{3}$$

$$F_R = \frac{1}{n} \sum_{i=1}^{n-2} \mathbb{I}(D_i \neq D_{i+1}) \tag{4}$$

where ΔPD_i represents the discrete derivative of the partial dependence function at point i, and D_i is the sign of this derivative, indicating whether the curve is rising ($D_i > 0$) or falling ($D_i < 0$) at point i. The fluctuation ratio (F_R) measures how often the curve changes direction by counting how many times consecutive signs of the slope (D_i) differ. $\mathbb{I}$ is a boolean operator that returns 1 if there is a change in direction ($D_i \neq D_{i+1}$), and 0 otherwise. A higher F_R indicates more frequent oscillations in the PDP, which suggests a less stable and potentially less interpretable relationship between the feature and the prediction.

2.1.4 Rank Alignment

The Rank Alignment metric assesses the consistency of feature importance rankings between the overall dataset and specific subsets. It measures how well the top-ranked features from the entire dataset align with those from conditional subsets, such as different predicted classes in classification tasks or output quartiles in regression tasks.

Definition 2.4 (rank alignment)**.** Let $\mathbf{F}_\alpha$ be the top α proportion of features based on their importance in the overall dataset, $\mathbf{F}_\alpha^g$ be the set of top α proportion of features for group g, and G be the total number of groups. The Rank Alignment score is the average Jaccard similarities across all groups:

$$R_A = \frac{1}{G} \sum_{g=1}^{G} \frac{|\mathbf{F}_\alpha \cap \mathbf{F}_\alpha^g|}{|\mathbf{F}_\alpha \cup \mathbf{F}_\alpha^g|} \tag{5}$$

A Rank Alignment score close to 1 indicates high consistency in feature importance rankings across groups, while a lower score suggests variations that may signal bias, instability, or group-specific behaviors.

2.2 Metrics Based on Local Feature Importance

2.2.1 Position Consistency

The Position Consistency metric evaluates how consistent the feature importance rankings are across different instances in the dataset. It measures the degree to which each feature maintains a consistent rank of importance when assessed at the local (instance-specific) level.

Definition 2.5 (position consistency)**.** For each instance i, the vector of local feature importances $f_i \in \mathbb{R}^d$ is converted into a ranking $r_i \in \mathbb{N}^d$, where $r_{ij} = \text{rank}(f_{ij})$. In this case, r_{ij} represents the rank of feature j in sample i, with lower ranks corresponding to higher importance. Next, for each feature j, the rank stability is calculated based on the ranks it holds across all iterations. First, the most frequent rank r_{freq} among iterations is determined. Then the actual deviation (D) of the feature's ranks from this most frequent rank and the maximum possible deviation ($\max D$) are computed as follows:

$$D_j = \sum_{i=1}^{M} |r_{ij} - r_{\text{freq},j}| \textbf{ and } \max D_j = M \times (\max_i r_{ij} - \min_i r_{ij}) \tag{6}$$

The position consistency of feature j is then calculated as: The position consistency C_j for feature j and the overall stability is for all features I_V are calculated as:

$$P_C = \frac{1}{d}\sum_{j=1}^{d} C_j \quad \textbf{where} \quad C_j = 1 - \frac{D_j}{\max D_j} \tag{7}$$

2.3 Importance Variability

The Importance Variability (I_V) metric measures the extent to which the importance values of each feature vary across different instances in the dataset. It provides insight into how stable the importance of each feature is when assessed locally.

Definition 2.6 (importance variability)**.** For each feature j, the mean importance across the samples is calculated $\mu_j = \frac{1}{M}\sum_{i=1}^{M} f_{ij}$ where f_{ij} is the importance of feature j in sample i, and M is the number of samples. The sample variance of the feature's importances is given by V_j .

$$V_j = \frac{1}{M}\sum_{i=1}^{M}(f_{ij} - \mu_j)^2 \quad \textbf{and} \quad V_{\text{max},j} = \frac{(f_{\text{max},j} - f_{\text{min},j})^2}{4}, \tag{8}$$

Otherwise, the stability S_j for feature j and the overall stability is for all features I_V are calculated as:

$$I_V = \frac{1}{d}\sum_{j=1}^{d} S_j \quad \textbf{where} \quad S_j = 1 - \frac{V_j}{V_{\text{max},j}} \tag{9}$$

2.4 Metrics Based on Surrogate Models

Surrogate models are interpretable models that approximate the behavior of complex black-box models, allowing for interpretability in tasks such as feature importance analysis. These models are especially valuable in Explainable AI (XAI) because they enable the decomposition of predictions into understandable components. To ensure reliability in model interpretation, it is essential to evaluate the *stability* of the features and their importances across different samples.

Feature and importance stability assess how consistently surrogate models rely on specific features or feature rankings when subjected to data variations (e.g., bootstrapping). High stability implies that the surrogate model provides robust explanations that do not fluctuate significantly with changes in data, which is critical for trustworthiness in model interpretability.

2.4.1 Surrogate Performance

Surrogate performance metrics are based on a simple surrogate model, typically a decision tree with a maximum depth of 3. This model is chosen for its interpretability and ease of visualization. The objective of this analysis is to quantify the discrepancy between the predictions of the original complex model and the surrogate model, using two key metrics: *surrogate accuracy* and *accuracy difference*.

Definition 2.7 (Surrogate Performance)**.** Let y represent the true target values, y_{pred} the predicted values from the original model, and $y_{\text{surrogate}}$ the predicted values from the surrogate model. Let $\mathcal{L}$ denote an accuracy measure, such as classification accuracy or another appropriate metric. The surrogate accuracy (SG_A) and accuracy difference (SG_D) are defined as follows:

- **Surrogate Accuracy** (SG_A): The accuracy of the surrogate model compared to the original model's predictions:
$$SG_A = \mathcal{L}(y_{\text{pred}}, y_{\text{surrogate}}).$$
This metric reflects how well the surrogate model replicates the predictions of the original model.
- **Accuracy Difference** (SG_D): The difference in accuracy between the original model and the surrogate model, defined as:
$$SG_D = \mathcal{L}(y, y_{\text{pred}}) - \mathcal{L}(y, y_{\text{surrogate}}).$$

This metric measures the performance loss when using the surrogate model instead of the original model.

A lower value of SG_D indicates that the surrogate model closely approximates the original model, whereas a higher value indicates a larger discrepancy between the two models. SG_A quantifies how well the surrogate captures the behavior of the original model, while SG_D reveals the trade-off between interpretability (using the surrogate) and accuracy (compared to the original model).

2.4.2 SURROGATE FEATURE STABILITY

Surrogate feature stability refers to the consistency of the features *used to construct the decision tree* across multiple bootstrap samples. This metric evaluates whether the surrogate model repeatedly selects the same features when exposed to slightly different versions of the data. Inconsistent feature selection may indicate that the surrogate model's explanations are not reliable.

Definition 2.8 (feature stability score)**.** The *feature stability score*, SG_f, is defined as the average Jaccard similarity between the sets of features selected in the original dataset and in bootstrap samples:

$$SG_f = \frac{1}{B} \sum_{i=1}^{B} \frac{|F_0 \cap F_i|}{|F_0 \cup F_i|} \tag{10}$$

where F_0 is the set of features selected by the decision tree on the original dataset, F_i is the set of features selected in the i-th bootstrap sample, and B is the number of bootstrap samples. This score quantifies how consistently the decision tree uses the same set of features for construction, with values closer to 1 indicating high stability.

3 EXPERIMENTS

This section provides a detailed account of the experimental setup and presents the results derived from the application of explainability metrics to various machine learning models across standard benchmark datasets. The analysis is structured around two case studies: the first investigates the Adult dataset, while the second focuses on the US-Crime dataset. These case studies are designed to illustrate the effectiveness of the proposed explainability metrics in both classification and regression tasks. A comprehensive description of the datasets, model configurations, and explainability techniques employed is provided in the Appendix B.

3.1 CASE STUDY 1: ADULT DATASET

For the classification task, we employed the Adult Dataset considering training four classification models: Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), and a Multi-Layer Perceptron (MLP).

3.1.1 GLOBAL FEATURE IMPORTANCE

The analysis of Global Feature Importance metrics *alpha score* reveals that the majority of the models concentrate feature importance on approximately 13% of the total features, as measured by the alpha score. An exception is the MLP model, which distributes its importance across roughly 30% of the features. Moreover, XGB exhibits the highest spread divergence, indicating a stronger concentration of importance in a few key features compared to other models.

Additionally, the fluctuation ratio, a metric that quantifies the complexity of the relationship between features and predictions, is more pronounced in models with higher non-linear behavior, such as RF, XGB, and MLP. Conversely, LR, being a linear model, has the lowest fluctuation ratio close to 0, signifying that its predictions are more straightforward and interpretable (see Appendix **??**). Among non-linear models, XGB shows the highest degree of fluctuation, followed by RF, reflecting the increased complexity of these models. Figure 4 visualizes the top feature importances (in blue) along with their fluctuation ratios. This visualization allows us to assess not only the importance of a feature for decision-making but also the complexity of its relationship with the target prediction. Another

noteworthy observation is the alignment of feature importance across predicted labels, particularly within the alpha range ($\alpha = 0.8$). In this regard, Random Forest (RF) shows the highest alignment of features that contribute to both label 0 and label 1 predictions, demonstrating consistent feature usage. As shown in Appendix 6, Random Forest maintains the top three most important features consistently across all subsets, providing better alignment compared to other models.

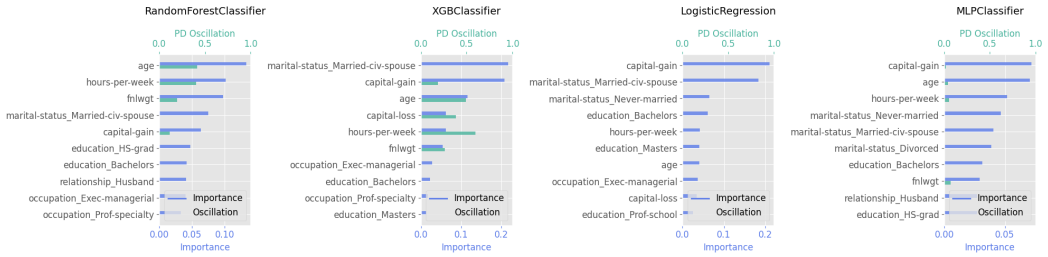

Figure 2: Feature Importance and Fluctuation Ratio for the ML model trained on the Adult Dataset.

3.1.2 Local Feature Importance

For Local Feature Importance, extracted from SHAP explainer, the ranking of its features across all samples are showed in Figure 3. The x-axis represents feature rankings, while the y-axis corresponds to the samples used to compute the local feature importances. The upper half of the graph displays samples with label 0, and the lower half represents samples with label 1. Some models exhibit changes in feature importance rankings between labels, and the Position Consistency metric quantifies how stable these rankings are across different subsets. For instance, the MLP model demonstrated the highest consistency in feature rankings, while XGBoost showed the lowest. The upper part of the figure illustrates this Position Consistency metric.

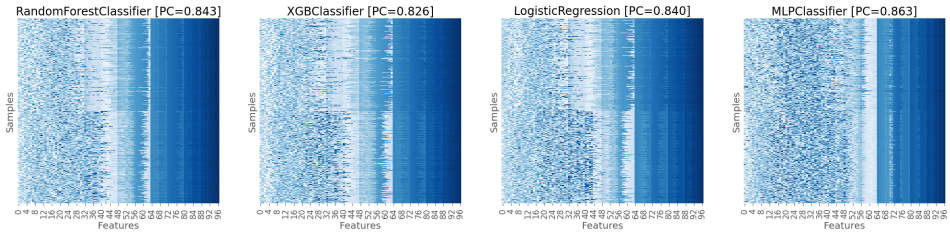

Figure 3: Feature Importance Contrast between samples grouped by labels for the ML model trained on the Adult Dataset. The upper half of the graph displays samples with label 0, and the lower half represents samples with label 1

3.1.3 Surrogate Metrics

The surrogate model analysis reveals that the Random Forest (RF) model shows the largest deviation from its surrogate, with an accuracy difference of 0.1876. This indicates that the surrogate struggles to accurately mimic the original model's predictions. When using decision trees, several metrics can help interpret the model's complexity, such as tree depth, number of features, and the number of rules. The reliability of these metrics, along with surrogate performance metrics, can be evaluated by assessing the stability of feature selection during the surrogate's creation.

The surrogate feature stability metric provides insight into this, indicating, for example, that in our results, the surrogate used for MLP showed the highest feature stability (0.3543), while Logistic Regression (LR) had the lowest (0.1939). Based on the Jaccard index used to calculate this metric, it can be interpreted that, on average, 35% of the features were consistently used across all surrogate models for MLP, while only 19% were used consistently for LR.

Table 1: Results XAI metrics for models trained on Adult Dataset

Metrics	RF	XGB	LR	MLP	REF
Efficacy					
Accuracy	0.852	**0.874**	0.850	0.839	1
F1-Score	0.671	**0.718**	0.659	0.636	1
Global Feature Importance					
Spread Divergence	0.6689	**0.7189**	0.6902	0.5231	1
Alpha Score	0.1649	**0.1134**	0.1443	0.3093	0
Fluctation Ratio	0.0855	0.1743	**0.0002**	0.0070	0
Rank Alignment	**0.8889**	0.5000	0.3913	0.5000	1
Local Feature Importance					
Position Consistency	0.843	0.826	0.840	**0.863**	1
Importance Stability	0.942	**0.944**	0.934	0.893	1
Surrogate					
Accuracy Difference	0.1876	0.0836	**0.0410**	0.0792	0
Surrogate Accuracy	0.8123	0.8734	**0.8872**	0.8573	1
Surrogate Feature Stability	0.3030	0.2444	0.1939	**0.3543**	1

3.2 Case Study 2: US-Crime Dataset

For the regression task, we employed the US-Crime Dataset to train and evaluate four models: Random Forest (RF), Bayesian Ridge Regression (BR), Logistic Regression (LR), and Multi-Layer Perceptron (MLP). In terms of efficacy, the Bayesian Ridge model achieved the lowest mean squared error (MSE) at 0.0186, closely followed by Logistic Regression and Random Forest with MSEs of 0.0190 and 0.0198 respectively. The MLP model showed relatively poorer performance with the highest MSE at 0.0248.

3.3 Global Feature Importance

The analysis of global feature importance reveals interesting patterns in how different models utilize features. MLP shows the highest spread divergence at 0.7628, significantly higher than other models, indicating it relies heavily on a small subset of features. In contrast, Logistic Regression has the lowest spread divergence at 0.3292, suggesting more balanced feature utilization across the dataset. When examining the complexity of feature relationships, Random Forest exhibits the highest fluctuation ratio at 0.2625, indicating complex, non-linear relationships between features and predictions. Both Bayesian Ridge and Logistic Regression show zero fluctuation, reflecting their linear nature, while MLP has a minimal ratio of 0.0047.

For consistency across prediction ranges, Logistic Regression demonstrates the highest rank alignment at 0.7778, suggesting consistent feature importance rankings regardless of the predicted value. Bayesian Ridge shows the lowest alignment at 0.3256, indicating that it uses different features depending on the prediction value. This variance in feature utilization across models highlights the different approaches each algorithm takes to the regression task.

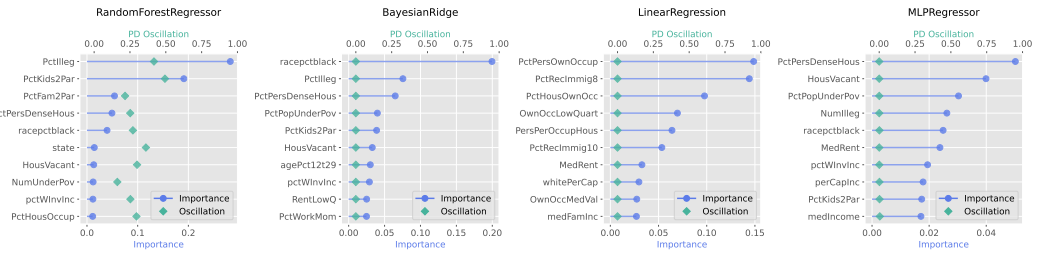

Figure 4: Feature Importance and Fluctuation Ratio for the ML model trained on the US-Crime Dataset.

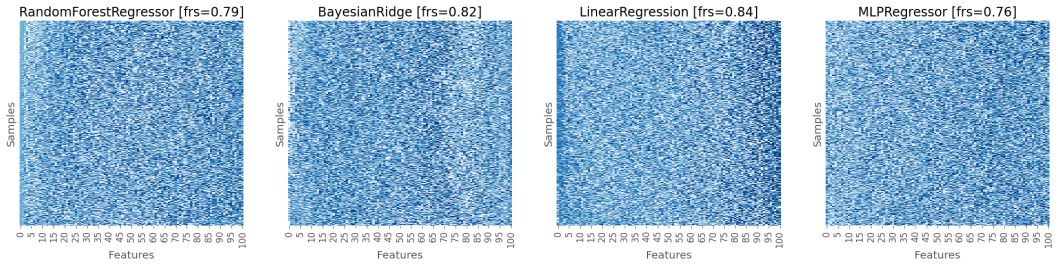

Figure 5: Feature Importance Contrast between labels for the ML model trained on the US-Crime Dataset.

Table 2: Results XAI metrics for models trained on US-Crime Dataset

Metrics	RF	BR	LR	MLP	REF
Efficacy					
MSE	0.0198	**0.0186**	0.0190	0.0248	0
Global Feature Importance					
Spread Divergence	0.4041	0.4087	0.3292	**0.7628**	1
Fluctuation Ratio	0.2625	**0.0000**	**0.0000**	0.0047	0
Rank Alignment	0.6757	0.3256	**0.7778**	0.5769	1
Local Feature Importance					
Position Consistency	0.7941	0.8188	0.8353	**0.7586**	0
Importance Stability	0.9972	0.9944	**0.9825**	0.9847	0
Surrogate					
Surrogate MSE	0.0047	**0.0016**	0.0024	0.0083	0
MSE Difference	-0.0075	**-0.0020**	-0.0024	-0.0092	0
Features Stability	**1.0000**	0.9980	0.9960	0.9980	1
Feature Importance Stability	0.2117	**0.9956**	0.9854	0.9772	1

3.3.1 Local Feature Importance

The local feature importance analysis provides insights into how models behave at the individual sample level. Logistic Regression leads with the highest position consistency at 0.8353, indicating stable feature rankings across different samples. MLP shows the lowest consistency at 0.7586, suggesting more variable feature utilization across samples. Notably, all models demonstrate remarkably high importance stability (above 0.98), with Random Forest showing the highest at 0.9972. This suggests that while the ranking of features might change, the relative importance of features remains highly consistent across different samples for all models.

3.3.2 Surrogate Metrics

The surrogate model analysis offers additional insights into model complexity and interpretability. Bayesian Ridge achieves the lowest surrogate MSE at 0.0016, indicating that its behavior is the easiest to approximate with a simpler model. MLP has the highest surrogate MSE at 0.0083, suggesting more complex decision-making patterns. Intriguingly, all models show negative MSE differences, indicating that surrogate models perform better than the original models - an unusual result that might warrant further investigation. In terms of surrogate feature importance stability, Bayesian Ridge, Logistic Regression, and MLP all show very high values above 0.97, while Random Forest has a notably low metric of 0.2117. This suggests that Random Forest's decision-making process is significantly more complex and harder to approximate with an interpretable model.

4 Conclusion

The increasing focus on transparency in AI systems has highlighted the trade-off between accuracy and explainability in machine learning models. This paper introduces a new set of explainability metrics aimed at capturing global and local feature importance. These metrics offer a structured way to understand model complexity without the need for extensive graphical analysis of feature

importance, providing a more systematic and interpretable framework for evaluating AI models. It is crucial to note that these metrics are explainer-agnostic, being applied for any explainability method.

As for future research, we plan on exploring the statistical properties of the metrics proposed, potentially creating inference tools for model selection and model auditing based on these tools. While the scope of the article focused on constructing model-agnostic metrics to evaluate model predictions, future studies may explore the development of metrics for methods focused on inner interpretability.

REFERENCES

Rasheed Omobolaji Alabi, Mohammed Elmusrati, Ilmo Leivo, Alhadi Almangush, and Antti A Mäkitie. Machine learning explainability in nasopharyngeal cancer survival using lime and shap. *Scientific Reports*, 13(1):8984, 2023.

S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99: 101805, 2023. ISSN 1566-2535. doi:10.1016/j.inffus.2023.101805. URL https://www.sciencedirect.com/science/article/pii/S1566253523001148.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023.

M. Buckmann, A. Joseph, and H. Robertson. An interpretable machine learning workflow with an application to economic forecasting. Technical report, Bank of England, 2022. URL https://www.bankofengland.co.uk/working-paper/2022/an-interpretable-machine-learning-workflow-with-an-application-to-economic-forecas

R. Elshawi, M. H. Al-Mallah, and S. Sakr. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, 19(1):146, Jul 2019. ISSN 1472-6947. doi:10.1186/s12911-019-0874-0. URL https://doi.org/10.1186/s12911-019-0874-0.

R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, 2020. doi:10.1111/coin.12410. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12410.

T. Fel and D. Vigouroux. Representativity and consistency measures for deep neural network explanations. 2020.

T. Fel, D. Vigouroux, R. Cadène, and T. Serre. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 720–730, 2022.

De-Cheng Feng, Wen-Jie Wang, Sujith Mangalathu, and Ertugrul Taciroglu. Interpretable xgboost-shap machine-learning model for shear strength prediction of squat rc walls. *Journal of Structural Engineering*, 147(11):04021173, 2021.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. 2019.

Freddy Gabbay, Shirly Bar-Lev, Ofer Montano, and Noam Hadad. A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. *Applied Sciences*, 11(21):10417, 2021.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects. 2018.

U. M. Khaire and R. Dhanalakshmi. Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1060–1073, 2022.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

Ziqi Li. Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. *Computers, Environment and Urban Systems*, 96:101845, 2022.

S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. 2017. URL https://arxiv.org/abs/1705.07874.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. Leveraging latent features for local explanations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1139–1149, 2021.

Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine*, 126:104041, 2020.

Xin Man and Ernest P Chan. The best way to select features? comparing mda, lime, and shap. *The Journal of Financial Data Science*, 3(1):127–139, 2021.

Yuan Meng, Nianhua Yang, Zhilin Qian, and Gaoyu Zhang. What makes an online review more helpful: an interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(3):466–490, 2020.

R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 652–663, 2021. ISBN 9781450384735. doi:10.1145/3461702.3462597. URL https://doi.org/10.1145/3461702.3462597.

Menaka Narayanan et al. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. 2018.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. 2020.

S. Nogueira, K. Sechidis, and G. Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018.

Yasunobu Nohara, Koutarou Matsumoto, Hidehisa Soejima, and Naoki Nakashima. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214:106584, 2022.

Amir Bahador Parsa, Ali Movahedi, Homa Taghipour, Sybil Derrible, and Abolfazl Kouros Mohammadian. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, 136:105405, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002. ISSN 0377-2217. doi:https://doi.org/10.1016/S0377-2217(01)00264-8. URL https://www.sciencedirect.com/science/article/pii/S0377221701002648.

M. T. Ribeiro, S. Singh, and C. Guestrin. 'why should i trust you?': Explaining the predictions of any classifier. 2016. URL https://arxiv.org/abs/1602.04938.

Raquel Rodríguez-Pérez and Jürgen Bajorath. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of computer-aided molecular design*, 34:1013–1026, 2020.

Yao Rong et al. A consistent and efficient evaluation strategy for attribution methods. 2022.

P. Schmidt and F. Biessmann. Quantifying interpretability and trust in machine learning systems. 2019.

Lloyd S Shapley et al. A value for n-person games. 1953.

L. M. Thimoteo, M. M. Vellasco, J. Amaral, K. Figueiredo, C. L. Yokoyama, and E. Marques. Explainable artificial intelligence for covid-19 diagnosis through blood test variables. *Journal of Control, Automation and Electrical Systems*, 33(2):625–644, 2022. URL https://link.springer.com/article/10.1007/s40313-021-00858-y.

J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021. ISSN 2079-9292. URL https://www.mdpi.com/2079-9292/10/5/593.

A Literature and Technical Background

The literature in explainable methods for AI models has been gaining prominence as AI models proliferate in virtually all areas of society (such as predicting hypertension (Elshawi et al., 2019), healthcare ElShawi et al. (2020), COVID-19 diagnosis Buckmann et al. (2022), economics and finance Thimoteo et al. (2022)). In this section, we discuss some of the XAI methods and forms to categorize them based on their proprieties. The discussion is far from being exhaustive, given how effervescent this field is, but the goal of this section is to help the reader understand our proposed XAI metrics and how they relate to the rest of the literature.

In this way, we can attribute a prominent role to metrics in achieving these objectives. Unlike other strategies that may lead to misleading interpretations of their outcomes, well-defined metrics tend to efficiently elucidate issues and enhance risk management in sensitive processes. Moreover, the use of metrics facilitates comparisons between different sets of results, making it effective, for instance, in hypothesis validation. Several factors, however, complicate the definition of metrics. In the case of XAI, the absence of a ground truth for explainability adds a layer of complexity to the comparison of different strategies, as pointed out by Zhou et al. (2021). In this context, potential strategies involve considering aspects such as fidelity, unambiguity, and overlap, as discussed by Lakkaraju et al. (2017). The existing literature introduces various methods for evaluating explainability techniques Mothilal et al. (2021). Mothilal et al. (2021) presents a framework that unifies strategies centred around feature attribution and counterfactual generation. In contrast, other studies propose explainability metrics grounded in algorithmic stability Fel et al. (2022); Khaire & Dhanalakshmi (2022); Nogueira et al. (2018). However, it is essential to clarify that our objective is not to evaluate the explainability methods themselves. Instead, our focus is on providing insights based on the importance attributed to features by different models. In doing so, we aim to facilitate comparisons of explanations across models.

We can define at least three categories of metrics with an emphasis on the explainability of results generated for AI models: (i) subjective, (ii) objective, and (iii) computational. The *subjective metrics* are employed when evaluating aspects that elicit subjective responses from users. This type of metric may be based on trust, understanding, and satisfaction, as proposed in Hoffman et al. (2018). The *objective metrics* are those related to observed aspects, for example, in users performing a particular task. They can be measures of time for the execution of the task or accuracy. Schmidt & Biessmann

(2019) seeks to objectively measure the quality of explainability methods and shows that quick and highly accurate decisions represent a good understanding of explainability. Narayanan et al. (2018) evaluate explainability results based on subjective (satisfaction) and objective (response time and accuracy) metrics. Furthermore, *computational metrics* are derived from mathematical indicators that assess the quality of explanations generated by an XAI method. Since these metrics are based on specific equations, user intervention is not necessary for obtaining them, making this type of metric suitable for automated systems.

A.1 Computational Metrics for Evaluating XAI Outputs

With the growing use of explainability methods for machine learning models, there is also an increase in studies that seek to evaluate the results based on feature importance or feature-attribution. Works investigating *computational metrics* may aim to construct metrics that assess methods based on their desired properties, such as fidelity, stability, comprehensibility, representativity, and consistency. Fel & Vigouroux (2020) focuses on constructing metrics that address representativity and consistency. The work proposed by Nguyen & Martínez (2020) suggests a set of metrics that also rely on certain properties, and in the case of evaluating feature-attribution methods, the metrics (monotonicity and non-sensitivity) seek to assess how faithful the methods are. On the other hand, other strategies evaluate feature-attribution methods based on different factors. Rong et al. (2022) proposes an information-theoretic strategy to evaluate feature-attribution methods.

A.2 Relevant approaches in XAI methods

The permutation feature importance was first introduced by Fisher et al. (2019). The work proposed by Fisher et al. (2019) constructs a model-agnostic method for machine learning predictions. Intuitively, if altering the values of certain features results in a considerable change in the AI model error, this feature is considered to be important. Alternatively, features are deemed unimportant if the AI model error remains unchanged after altering its values. Permutation feature importance is a powerful tool because its interpretation is intuitive and it can be applied to any model – i.e. it is an easy-to-understand global XAI method. In addition, it does not require retraining the model, nor knowing its estimates and nor its *modus operandi*. On the other hand, there is no consensus in the literature about whether a training or test set should be used to compute the feature's importance.

Another explainability method is the partial dependence curve. A machine learning model f is often a function of a multitude of features x, which makes it infeasible to plot the estimated model in a high-dimensional space. Instead, the *partial dependence curve*, can be used to assess how the predicted outcome of a model f behaves as a function of values of a particular *selected* feature x_s, after averaging f over the values of *all* other features x_A. Intuitively, the partial dependence curve can be interpreted as the expected/average model response as a function of the input feature of interest Pedregosa et al. (2011). This helps with model explainability since it makes it possible to assess whether the relationship between the outcome and a feature is linear, and/or monotonic, or more complex. For example, if applied to a linear regression, the partial dependence plot always shows a linear relationship. The estimation assumes that features are not correlated. The violation of this assumption indicates that the averages calculated for the partial dependence may include data points that are implausible.

The relationship between features and the outcome of an AI model is often too complex to be easily summarized in *black-box* models. On the other hand, there are models where explainability is straightforward, such as linear regression or relatively small decision trees. It turns out that one could use the latter simpler models – also referred to as surrogate models – to approximate a complex *black-box* model, often locally (i.e. for a subset of observations), and reap the explainability properties of the approximating model. In other words, for a subset of observations, it is possible that a simple and explainable model can approximate reasonably well complex predictions made by a complex model, and, in turn, offer a straightforward connection between features and output in the prediction process. This surrogate model approximation can be implemented by fitting a simple explainable model to a dataset corresponding to the same features used to train the complex AI model, while using the outcome generated by the AI model, as opposed to using the observed outcome.

The additive feature attribution methods are also well-defined in literature, with many applications Li (2022); Man & Chan (2021); Feng et al. (2021); Meng et al. (2020); Parsa et al. (2020); Nohara

et al. (2022); Rodríguez-Pérez & Bajorath (2020); Alabi et al. (2023); Gabbay et al. (2021); Magesh et al. (2020). As described by Lundberg & Lee (2017), this class of methods have an explanation model there is a linear function of binary variables. Several models follow this additive feature attribution definition. The local interpretable model-agnostic explanations (LIME) proposed by Ribeiro et al. (2016) focuses on providing explanations for any classifier (or regressor) at the local level. As an additive feature attribution method, it aims to explain why a particular prediction was made for a specific instance. The purpose of the Shapley additive explanations (SHAP) method introduced by Lundberg & Lee (2017), uses Shapley values Shapley et al. (1953) to compute feature attribution. Taking as input a set function $v : 2^n \to \mathbb{R}$, we can define the Shapley value $\phi_i(v)$ for a specific variable *i* as your contribution to the payoff through the weighted average of all possible combinations.

B EXPERIMENTS SETUP

B.1 DATASETS

Classification For the classification task, we employ the Adult dataset Becker & Kohavi (1996), which is designed to predict whether an individual's income exceeds $50,000 per year based on demographic and work-related attributes. This dataset includes 14 features, such as 'age', 'workclass', 'education', 'marital-status', 'occupation', 'relationship', 'race', 'sex', 'capital-gain', 'capital-loss', 'hours-per-week', and 'native-country'. It contains 48,842 records, each representing an individual with associated feature values.

Certain variables in the dataset require clarification. For instance, "fnlwgt" (final weight) is a continuous variable that indicates the number of people in the population represented by each record. Similarly, "education-num" quantifies the total years of education as a continuous counterpart to the categorical "education" variable. The "relationship" variable defines the individual's role within their household (e.g., 'husband', 'not-in-family'), while "capital-gain" and "capital-loss" capture additional income derived from investments, separate from wages.

The target variable is binary, where 0 represents individuals with income less than $50,000 and 1 represents those earning more. The objective of the model is to identify patterns in the features that reliably predict whether an individual's income surpasses this threshold. Before modeling, we handle common data preprocessing steps, such as encoding categorical variables and addressing missing values, to ensure robust model performance.

Regression For the regression task, we employ the United States Crime dataset (Redmond & Baveja, 2002). The dataset is an extensive and multifaceted collection of data that provides critical insights into the patterns and prevalence of crime across various communities within the United States. This comprehensive dataset combines socio-economic information from the 1990 US Census, law enforcement data from the 1990 US Law Enforcement Management and Administrative Statistics (LEMAS) survey, and detailed crime reports from the 1995 FBI Uniform Crime Reporting (UCR) program. By integrating these diverse data sources, the dataset offers a holistic view of the factors influencing crime and the efficacy of law enforcement responses.

The dataset includes many variables to allow for testing algorithms that select or learn weights for attributes. In total, the dataset contains 1993 instances and 101 attributes. However, attributes unrelated to crime were excluded. Variables were selected if there was any plausible connection to crime (N=122), along with the target attribute, Violent Crimes Per Population (total number of violent crimes per 100K population). The dataset comprises community-related variables, such as the percentage of the population considered urban and median family income, as well as law enforcement-related variables, such as the per capita number of police officers and the percentage of officers assigned to drug units.

B.2 MODELS AND PARAMETERS

Classification models The models used in Case Study 1 were Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), and a Multi-Layer Perceptron (MLP). The MLP architecture

consisted of two hidden layers with 200 and 100 neurons using the hyperbolic tangent (tanh) activation function.

Regression models The models used in Case Study 2 were Random Forest (RF), Bayesian Ridge (BR), Linear Regression (LR), and a Multi-Layer Perceptron (MLP). MLP architecture is consisted of two hidden layers with 300 and 200 neurons using the hyperbolic tangent (tanh) activation function.

C CASE STUDY 1: ADDITIONAL RESULTS

C.1 GLOBAL FEATURE IMPORTANCE

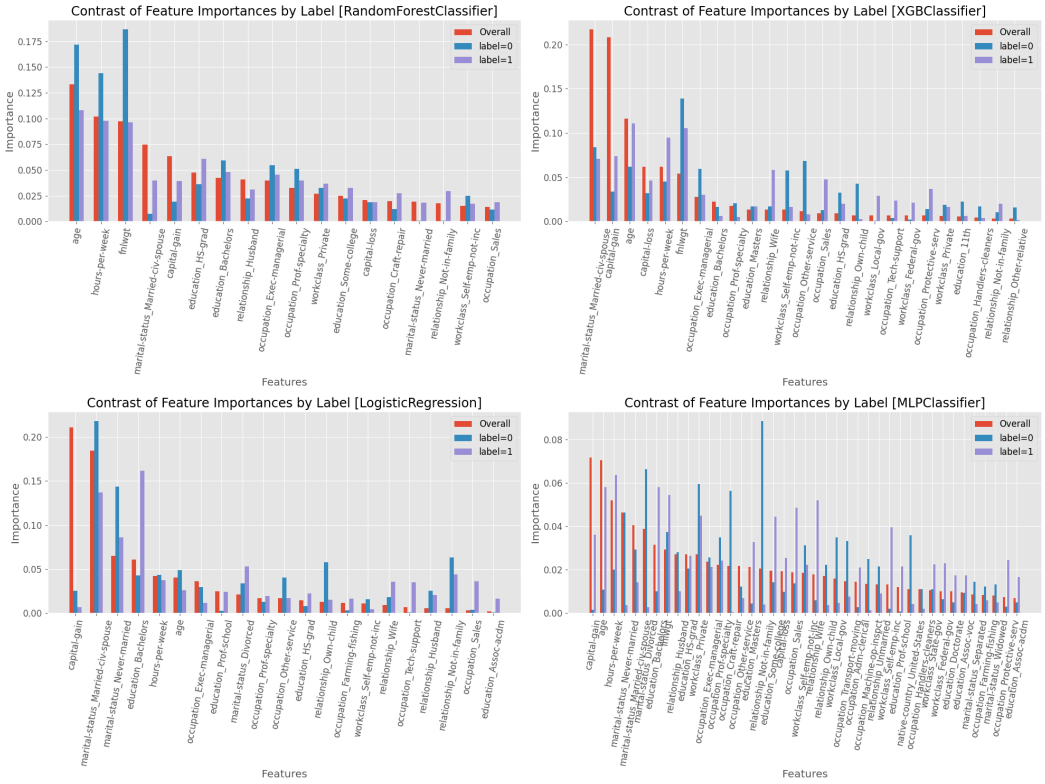

Figure 6: Feature Importance Comparison: Global importance versus class-specific importances for labels 0 and 1 in a machine learning model trained on the Adult Dataset.

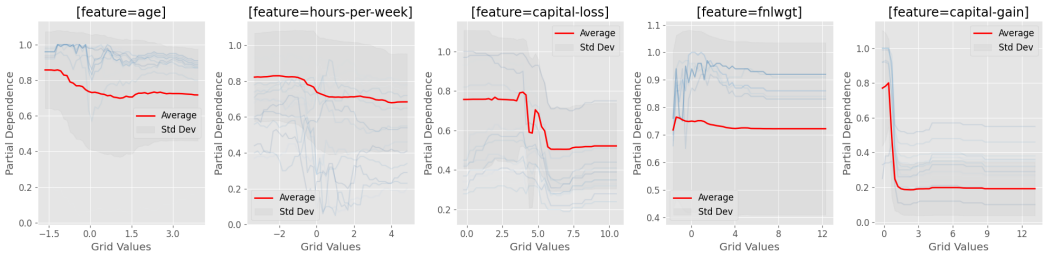

Figure 7: Partial Dependence Plot for Random Forest Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation.

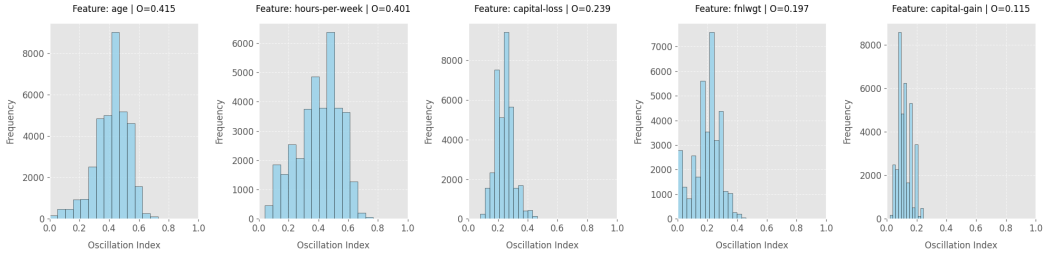

Figure 8: Fluctuation Ratio Distribution for the Most Important Features in the Random Forest Model: Analyzing the variation in feature importances.

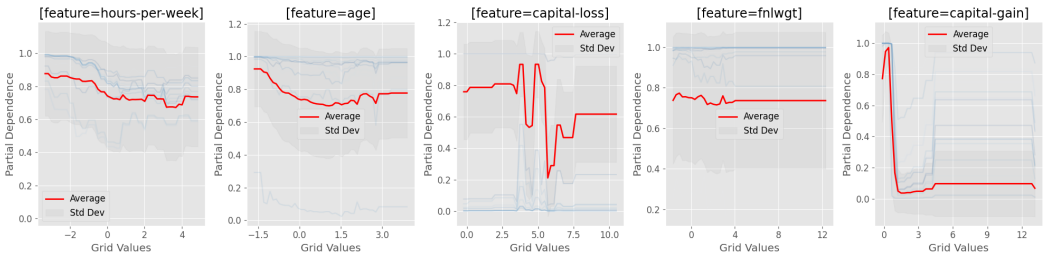

Figure 9: Partial Dependence Plot for XGBoost Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation.

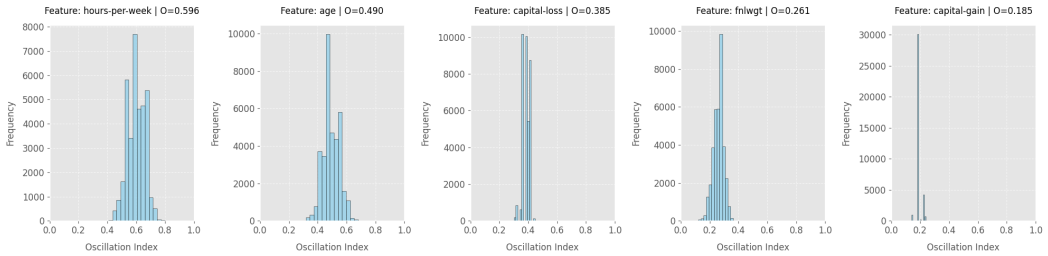

Figure 10: Fluctuation Ratio Distribution for the Most Important Features in the XGBoost Model: Analyzing the variation in feature importances.

D Case Study 2: Additional Results

D.1 Global Feature Importance

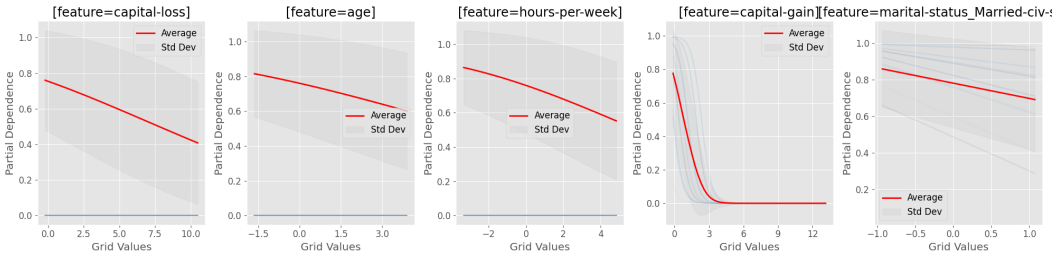

Figure 11: Partial Dependence Plot for Logistic Regression Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation.

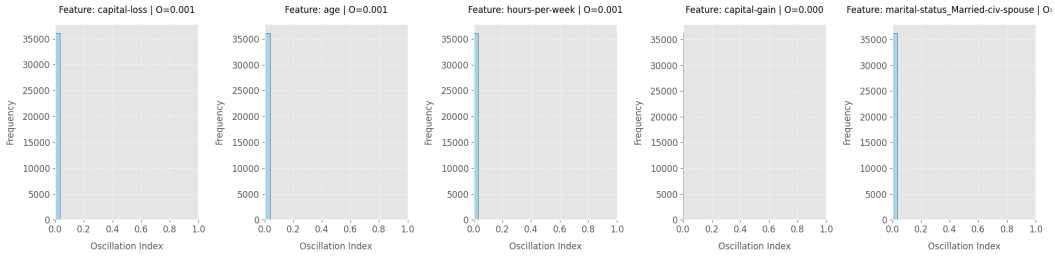

Figure 12: Fluctuation Ratio Distribution for the Most Important Features in the Logistic Regression Model: Analyzing the variation in feature importances.

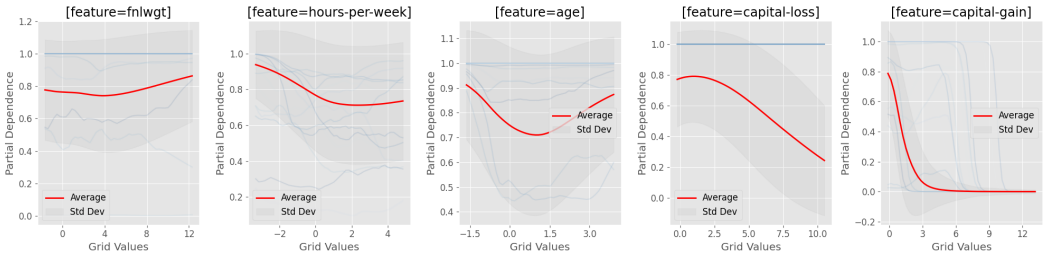

Figure 13: Partial Dependence Plot for Multi Layer Perceptron Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation.

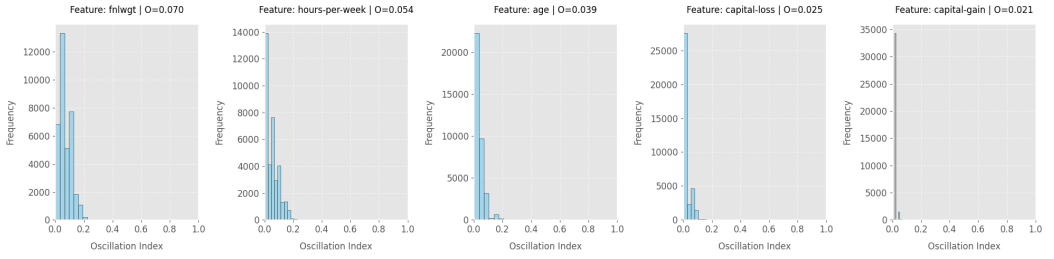

Figure 14: Fluctuation Ratio Distribution for the Most Important Features in the Multi Layer Perceptron Model: Analyzing the variation in feature importances.

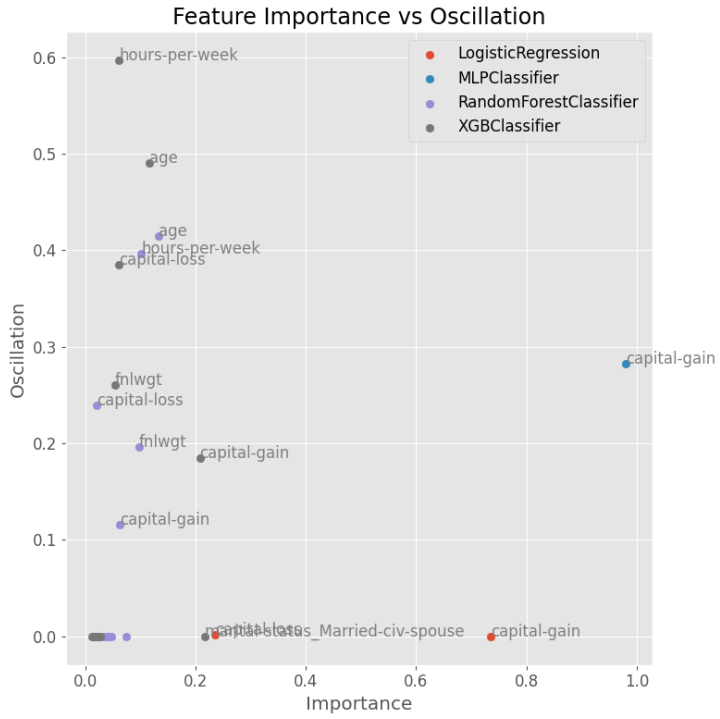

Figure 15: Feature Importance Contrast between labels for ML model trained on adult dataset

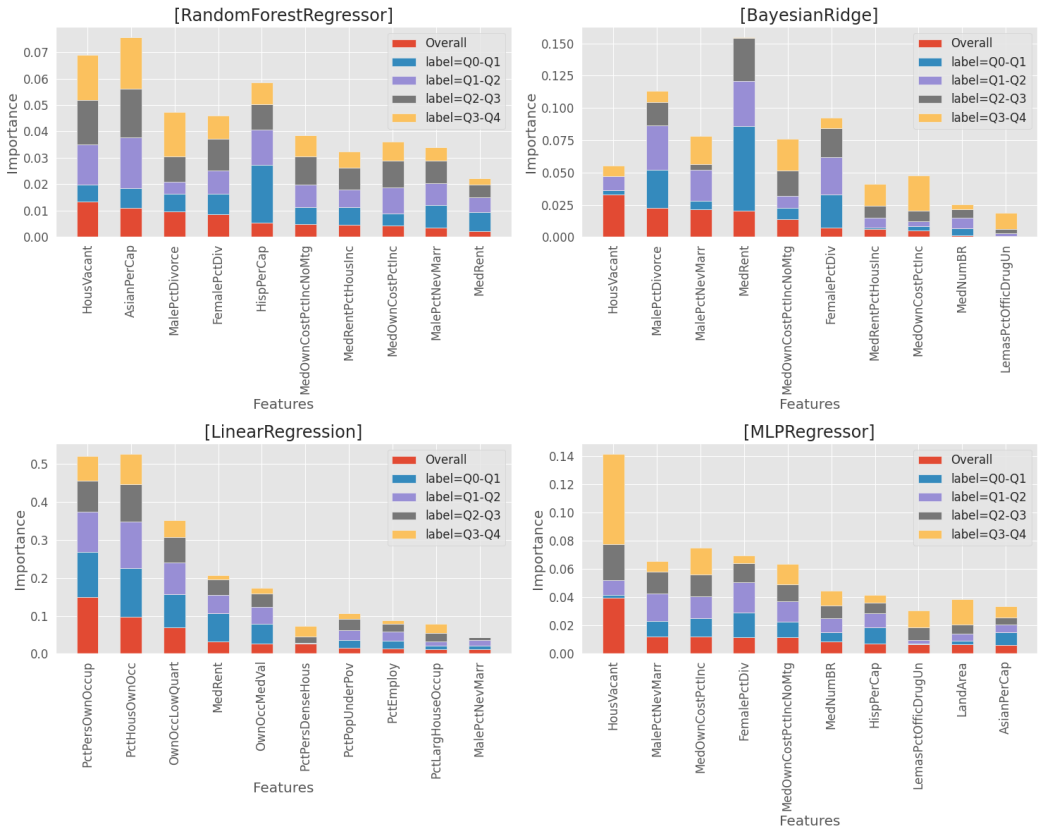

Figure 16: Feature Importance Comparison (top-10): Global importance versus class-specific importances for labels 0 and 1 in a machine learning model trained on the US-Crime Dataset.

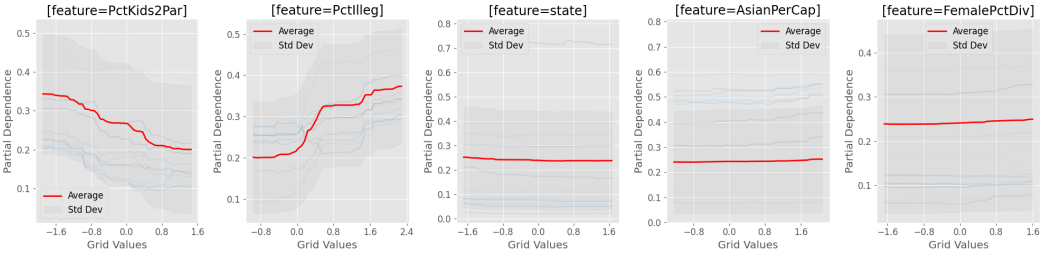

Figure 17: Partial Dependence Plot for Random Forest Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation for US-Crime Dataset.

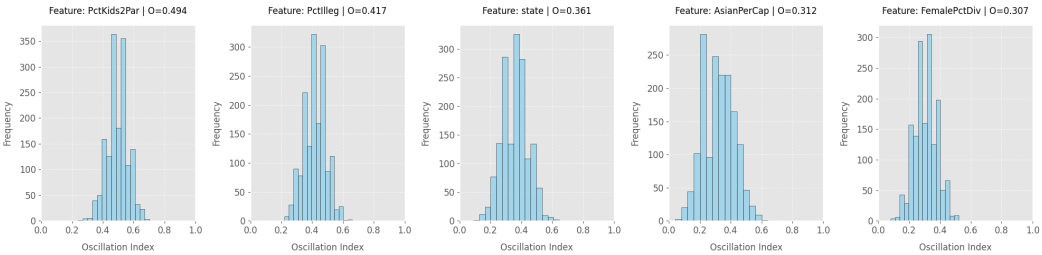

Figure 18: Fluctuation Ratio Distribution for the Most Important Features in the Random Forest Model: Analyzing the variation in feature importance for US-Crime Dataset.

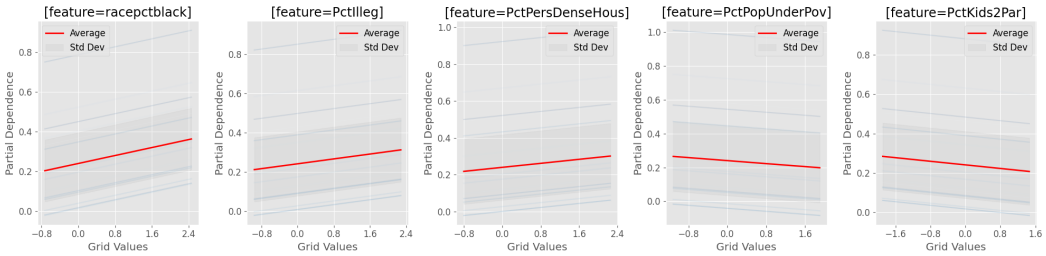

Figure 19: Partial Dependence Plot for Bayesian Ridge Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation for US-Crime Dataset.

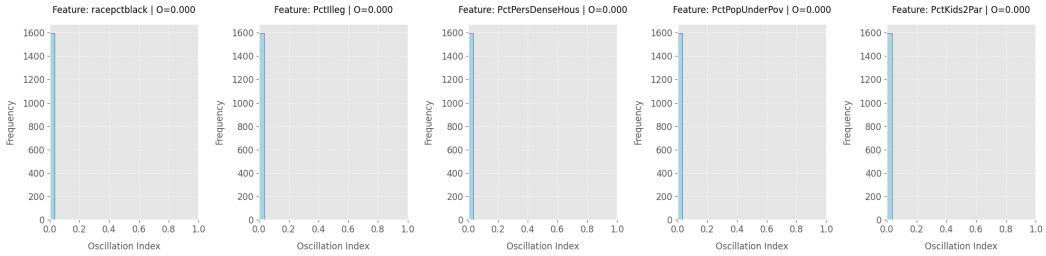

Figure 20: Fluctuation Ratio Distribution for the Most Important Features in the Bayesian Ridge: Analyzing the variation in feature importance for US-Crime Dataset.

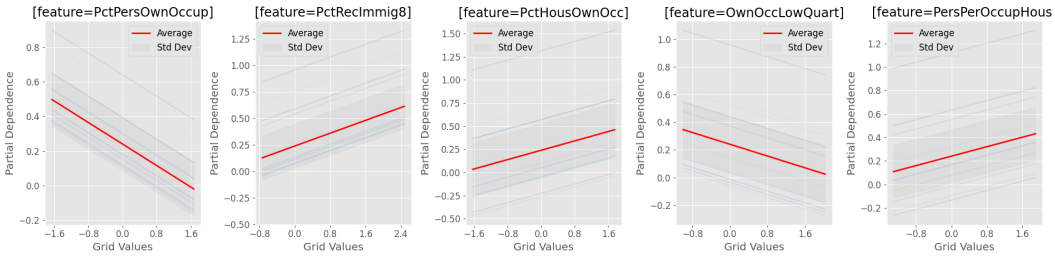

Figure 21: Partial Dependence Plot for Linear Regression Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation for US-Crime Dataset.

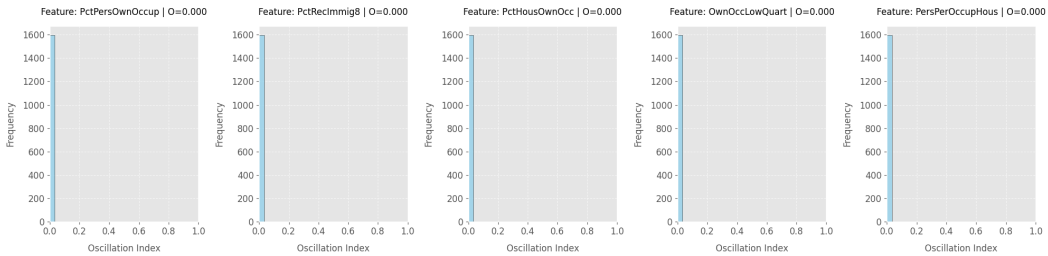

Figure 22: Fluctuation Ratio Distribution for the Most Important Features in the Linear Regression Model: Analyzing the variation in feature importance for US-Crime Dataset.

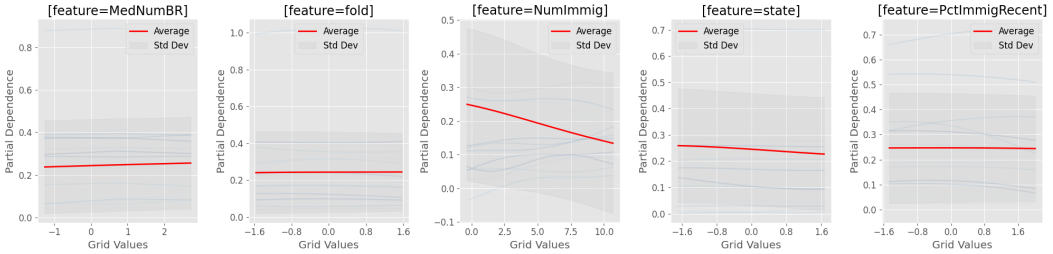

Figure 23: Partial Dependence Plot for Multi Layer Perceptron Model: Displaying the mean, standard deviation, and the 15 curves with the highest fluctuation for US-Crime Dataset.

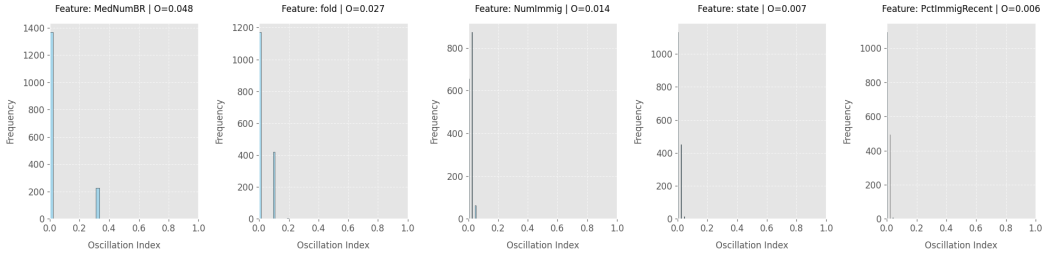

Figure 24: Fluctuation Ratio Distribution for the Most Important Features in the Multi-Layer Perceptron Model: Analyzing the variation in feature importance for US-Crime Dataset.

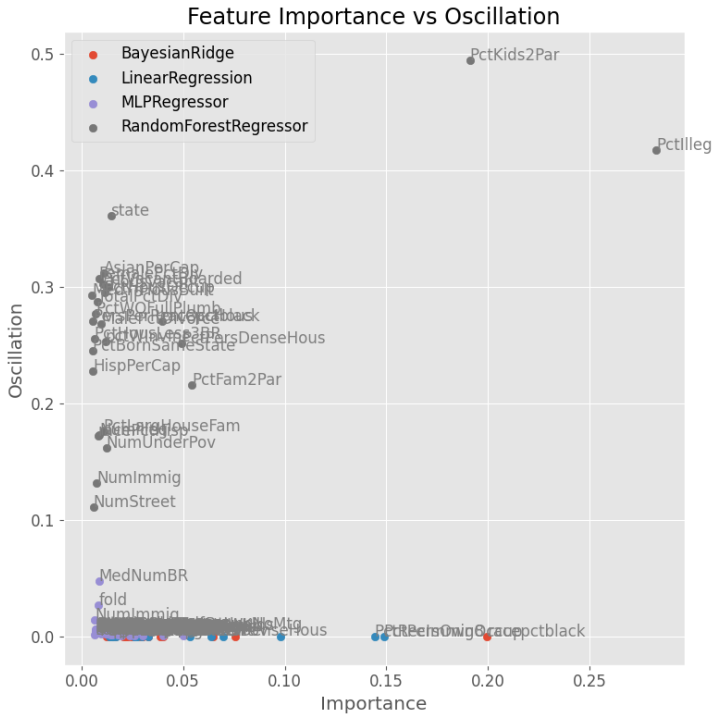

Figure 25: Feature Importance Contrast between labels for ML model trained on US-Crime dataset