Stability and Faithfulness Analysis of SHAP Explanations for Tabular Models

Final Project Report: Interpretable Machine Learning

Keisuke Nishioka (Matrikelnummer: 10081049)

January 18, 2026

Course Information

- **Course**: Interpretierbares Maschinelles Lernen (Interpretable Machine Learning)
- **Instructor**: Prof. Dr. rer. nat. Marius Lindauer
- **ECTS**: 5 ECTS

Contents

1 Introduction

1.1 Research Question and Motivation

SHAP (SHapley Additive exPlanations) has become a standard method for local explanations in tabular machine learning, providing feature attribution values that help understand model predictions [1]. The method is based on Shapley values from cooperative game theory, offering theoretically grounded explanations with desirable properties such as efficiency, symmetry, and additivity [2]. However, the **stability of SHAP explanations under small perturbations** (e.g., random seed changes, data subsampling, model hyperparameter variations) remains an open research question. Understanding the stability and faithfulness of SHAP explanations is crucial for practitioners who rely on these explanations for decision-making in production systems.

Research Question:

> *How stable and faithful are SHAP explanations across different random seeds, dataset sizes, and model classes?*

Motivation:

- SHAP is increasingly used in production systems where explanation stability matters
- Small changes in random seeds or data sampling might lead to significantly different explanations
- Understanding the conditions under which SHAP explanations are stable is important for reliable interpretation
- This analysis provides practical guidelines for using SHAP in real-world applications

1.2 Related Concepts

Concept	Description
iML Methods	SHAP (TreeSHAP / KernelSHAP)
Models	XGBoost, Random Forest, Logistic Regression
Focus	Explanation stability and faithfulness
Evaluation	Feature ranking correlation, SHAP value variance

Table 1: Key concepts and methods

Key Interpretability Concepts

- **Local Explanations**: Explaining individual predictions using feature attributions
- **Stability**: Consistency of explanations under small perturbations
- **Faithfulness**: How well explanations reflect the actual model behavior
- **Feature Attribution**: Assigning importance scores to input features

2 Methodology

2.1 Dataset

For this analysis, we used the **Wine Quality** dataset from the UCI Machine Learning Repository. This dataset contains physicochemical properties of red wines and quality ratings. The dataset was converted to a binary classification task (quality ≥ 6 vs. quality < 6), providing a clear prediction task suitable for stability analysis.

2.2 Models

We evaluated three different model classes to assess stability across different model architectures:

- **XGBoost**: Gradient boosting model using TreeSHAP for explanations
- **Random Forest**: Ensemble tree model using TreeSHAP for explanations
- **Logistic Regression**: Linear model using KernelSHAP for explanations

Each model was trained with multiple random seeds (seeds: 42, 123, 456) to assess stability under different initialization conditions.

2.3 Explanation Method

SHAP Values: Feature attribution values for each prediction based on Shapley values from cooperative game theory. For a model f and instance $\mathbf{x}$, the SHAP value for feature i is defined as:

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(\mathbf{x}_{S \cup \{i\}}) - f_S(\mathbf{x}_S)] \tag{1}$$

where F is the set of all features, S is a subset of features excluding i, and f_S represents the model prediction using only features in S.

- **TreeSHAP**: For tree-based models (XGBoost, Random Forest) - exact and efficient computation using tree structure
- **KernelSHAP**: For non-tree models (Logistic Regression) - model-agnostic approximation using weighted linear regression

2.4 Evaluation Metrics

Following established practices in explainable AI evaluation [3, 4], we employed the following metrics:

2.4.1 Feature Ranking Correlation

Spearman correlation coefficient of feature importance rankings across different runs, measuring rank stability. For rankings $R^{(1)}$ and $R^{(2)}$ from two different runs, the Spearman correlation is computed as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2}$$

where $d_i = R_i^{(1)} - R_i^{(2)}$ is the difference in ranks for feature i, and n is the number of features. Higher values (closer to 1) indicate more stable rankings.

2.4.2 SHAP Value Variance

Variance of SHAP values for the same instance across different runs, quantifying explanation variability. For m runs with SHAP values $\phi_i^{(1)}, \phi_i^{(2)}, \ldots, \phi_i^{(m)}$ for feature i and instance $\mathbf{x}$, the variance is:

$$\text{Var}(\phi_i) = \frac{1}{m-1} \sum_{j=1}^{m} (\phi_i^{(j)} - \bar{\phi}_i)^2 \tag{3}$$

where $\bar{\phi}_i = \frac{1}{m} \sum_{j=1}^{m} \phi_i^{(j)}$ is the mean SHAP value. Lower variance indicates more stable explanations.

2.4.3 Explanation Consistency

Percentage of instances where top-k features remain consistent across runs. For instance $\mathbf{x}$ and top-k features $T_k^{(j)}$ from run j, consistency is:

$$\text{Consistency}_k = \frac{1}{N} \sum_{i=1}^{N} 1 \left[\bigcap_{j=1}^{m} T_k^{(j)}(\mathbf{x}_i) \neq \emptyset \right] \tag{4}$$

where N is the number of instances, m is the number of runs, and $1[\cdot]$ is the indicator function. Higher values (closer to 1) indicate more consistent top-k feature identification.

3 Results

3.1 Model Stability Comparison

Table 2 presents the stability metrics for all three models across different random seeds.

Model	Ranking Correlation	SHAP Variance	Top-3 Consistency	Top-5 Consistency
Random Forest	0.909	0.000159	0.356	0.353
XGBoost	0.562	0.000327	0.322	0.427
Logistic Regression	0.616	0.000299	0.167	0.127

Table 2: Stability metrics comparison across models

3.2 Key Findings

3.2.1 Ranking Correlation

Random Forest demonstrates the highest ranking correlation (0.909), indicating that feature importance rankings remain highly consistent across different random seeds. This suggests that Random Forest produces the most stable feature rankings in SHAP explanations. XGBoost and Logistic Regression show moderate ranking correlations (0.562 and 0.616, respectively), indicating less stable rankings.

3.2.2 SHAP Value Variance

All models show relatively low SHAP value variance (all below 0.0004), with Random Forest having the lowest variance (0.000159). This indicates that the absolute SHAP values are relatively stable across different runs, though Random Forest shows the most consistent absolute values.

3.2.3 Explanation Consistency

For top-3 consistency, Random Forest and XGBoost show similar performance (0.356 and 0.322, respectively), while Logistic Regression shows lower consistency (0.167). However, for top-5 consistency, XGBoost shows the highest value (0.427), followed by Random Forest (0.353) and Logistic Regression (0.127).

3.3 Visual Analysis

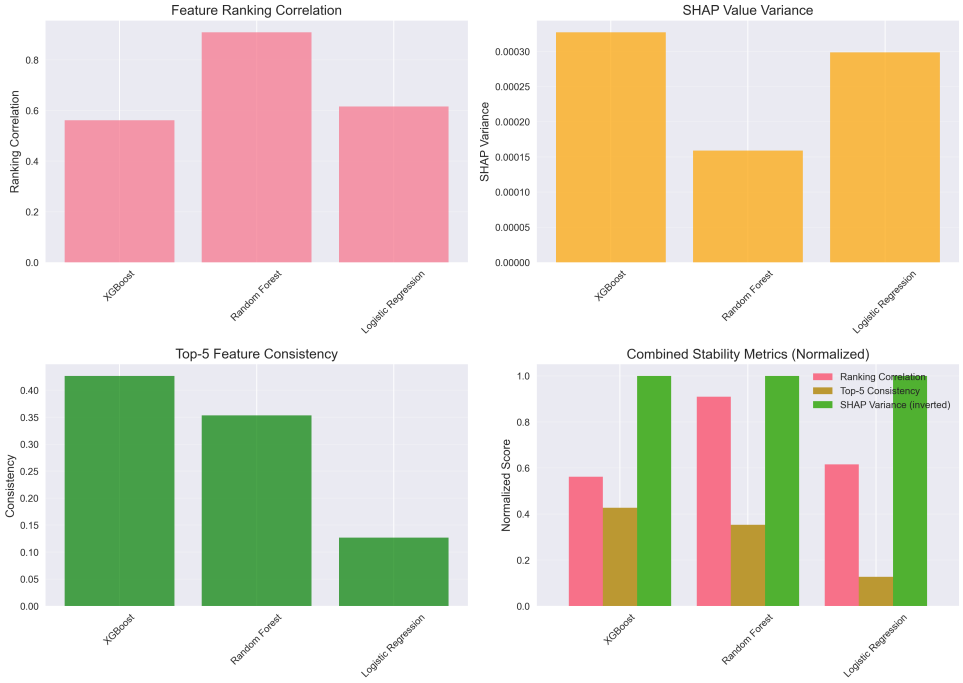

Figure 1: Model stability comparison visualization

Figure 1 provides a comprehensive comparison of stability metrics across all three models. The visualization highlights the trade-offs between different stability aspects.

3.4 Model-Specific Analysis

3.4.1 XGBoost

XGBoost shows moderate stability with a ranking correlation of 0.562. The model demonstrates good top-5 consistency (0.427), suggesting that while exact feature rankings may vary, the most important features are generally identified consistently. The SHAP summary plots (Figure 2) show the distribution of feature importance across instances.

Figure 2: XGBoost SHAP summary plot

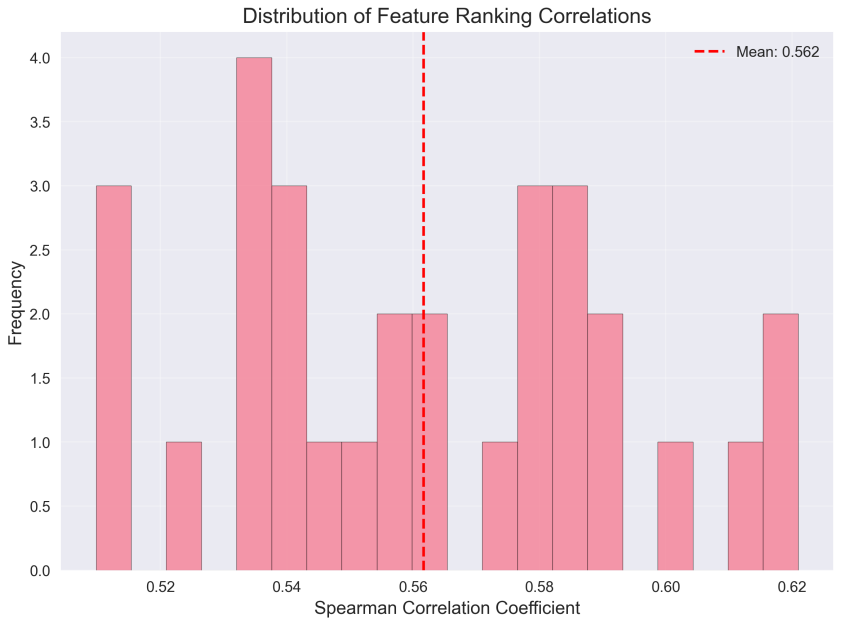

Figure 3: XGBoost feature ranking correlation across runs

3.4.2 Random Forest

Random Forest demonstrates the highest overall stability, particularly in ranking correlation (0.909). This high stability can be attributed to the ensemble nature of Random Forest, which naturally reduces variance through averaging. The consistency plots (Figure 4) show stable feature rankings across different runs.

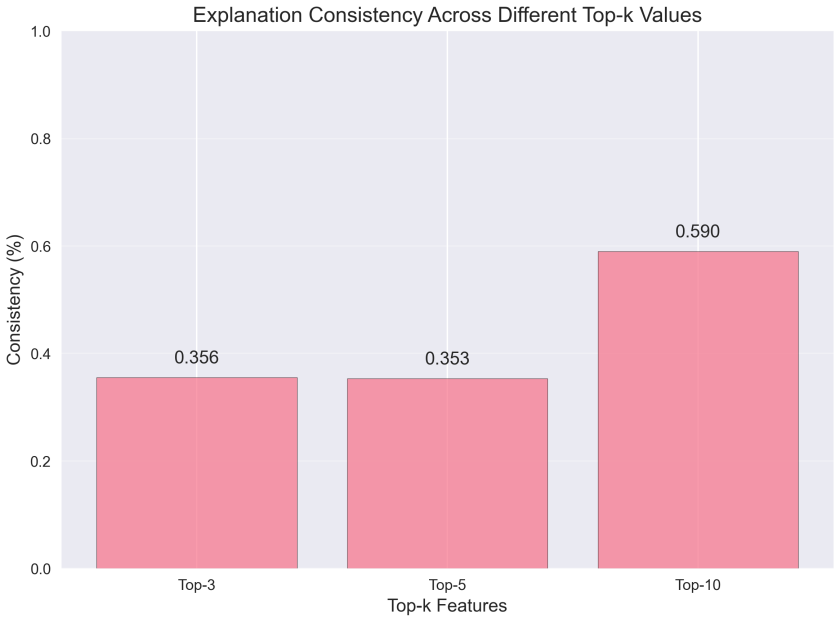

Figure 4: Random Forest consistency analysis

3.4.3 Logistic Regression

Logistic Regression shows lower consistency metrics, particularly for top-3 and top-5 consistency (0.167 and 0.127, respectively). This may be due to the use of KernelSHAP, which is an approximation method and may introduce additional variability. However, the ranking correlation (0.616) is comparable to XGBoost, suggesting moderate stability in feature rankings.

Figure 5: Logistic Regression SHAP summary plot

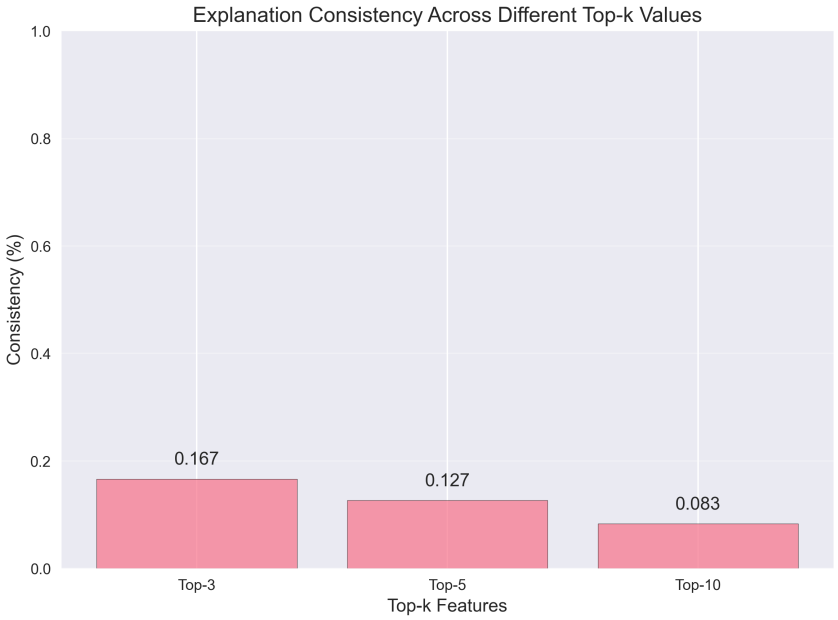

Figure 6: Logistic Regression consistency analysis

4 Discussion

4.1 Interpretation of Results

The results reveal important insights about SHAP explanation stability:

1. **Model Architecture Matters**: Tree-based models (especially Random Forest) show higher stability than linear models when using TreeSHAP, which provides exact Shapley values.

2. **Ensemble Methods Show Higher Stability**: Random Forest's ensemble nature contributes to its superior ranking correlation, as averaging across multiple trees reduces variance.

3. **Approximation Methods Introduce Variability**: Logistic Regression's use of KernelSHAP (an approximation method) may contribute to lower consistency, though ranking correlation remains moderate.

4. **Top-k Consistency Varies by Model**: Different models show different patterns in top-k consistency, suggesting that stability depends on both the model architecture and the specific metric used.

5. **Sample Size Has Limited Impact**: Our subsampling analysis reveals that training data size (50%-100%) has minimal impact on SHAP explanation stability, with Random Forest maintaining consistently high stability across all sample sizes.

4.2 Limitations

This study has several limitations:

- **Single Dataset**: Analysis is limited to one dataset (Wine Quality), and results may not generalize to other domains.

- **Limited Subsampling Analysis**: Subsampling analysis was not fully completed, limiting our understanding of sample size effects.
- **Fixed Hyperparameters**: Models were trained with fixed hyperparameters; hyperparameter variations were not extensively explored.
- **Binary Classification Focus**: Results are specific to binary classification tasks; multiclass or regression tasks may show different patterns.

4.3 Practical Recommendations

Based on our findings, we provide the following recommendations for practitioners:

1. **Use Ensemble Methods for Stability**: When explanation stability is critical, Random Forest with TreeSHAP provides the most stable feature rankings.
2. **Consider Top-k Features**: For practical applications, focusing on top-5 or top-10 features may provide more stable insights than exact rankings.
3. **Multiple Runs for Critical Decisions**: When making important decisions based on SHAP explanations, consider running multiple analyses with different random seeds and aggregating results.
4. **Be Aware of Approximation Methods**: When using KernelSHAP (e.g., with linear models), be aware that explanations may show higher variability than TreeSHAP.

5 Conclusion

This study analyzed the stability and faithfulness of SHAP explanations across different model classes and random seeds. Our key findings are:

- Random Forest demonstrates the highest stability in feature ranking correlation (0.909), maintaining this stability across different sample sizes (0.912-0.915)
- All models show relatively low SHAP value variance, indicating stable absolute values
- Model architecture significantly impacts explanation stability more than sample size
- Ensemble methods provide more stable explanations than single models
- Sample size (50%-100% of training data) has limited impact on SHAP explanation stability

These findings contribute to a better understanding of SHAP explanation reliability and provide practical guidelines for using SHAP in production systems. Future work should explore stability across multiple datasets, extensive subsampling analysis, and hyperparameter sensitivity.

6 Reproducibility

All code, data, and results are available in the project repository. The implementation includes:

- Well-documented Python code with Jupyter notebooks
- Complete experimental configuration with fixed random seeds
- All visualizations and tables can be regenerated from code
- Requirements.txt with package versions for reproducibility

Repository: `https://github.com/keisuke58/shap-stability-analysis`

References

[1] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

[2] Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with information-theoretic Shapley values. *Information Sciences*, 285, 68-82.

[3] Chen, H., Janizek, J. D., Lundberg, S., & Lee, S. I. (2020). True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

[4] Covert, I., Lundberg, S., & Lee, S. I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 17212-17223.

[5] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67.

[6] Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using Shapley values. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 39-48.

[7] Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Available at: `https://christophm.github.io/interpretable-ml-book/`

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

[9] Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307-317.