Stability and Faithfulness Analysis of SHAP Explanations

for Tabular Machine Learning Models

Keisuke Nishioka (Matrikelnummer: 10081049)
Interpretierbares Maschinelles Lernen | Prof. Dr. rer. nat. Marius Lindauer

Introduction

Research Question:
How stable and faithful are SHAP explanations across different random seeds, dataset sizes, and model classes?

Motivation:
- SHAP is widely used in production systems
- Explanation stability is crucial for reliable interpretation
- Understanding stability conditions helps practitioners

Key Concepts:
- **SHAP**: SHapley Additive exPlanations
- **TreeSHAP**: For tree-based models (exact)
- **KernelSHAP**: For any model (approximation)
- **Stability**: Consistency under perturbations

Key Results

Stability Metrics Comparison:

Model	Ranking Corr.	SHAP Var.	Top-5 Consist.
Random Forest	**0.909**	**0.00016**	0.353
XGBoost	0.562	0.00033	**0.427**
Logistic Reg.	0.616	0.00030	0.127

Key Findings:
1. **Random Forest** shows highest ranking correlation (0.909)
2. All models show low SHAP variance (< 0.0004)
3. **XGBoost** shows best top-5 consistency (0.427)
4. Ensemble methods provide more stable explanations

Methodology

Dataset:
Wine Quality (UCI Repository)
Binary classification task

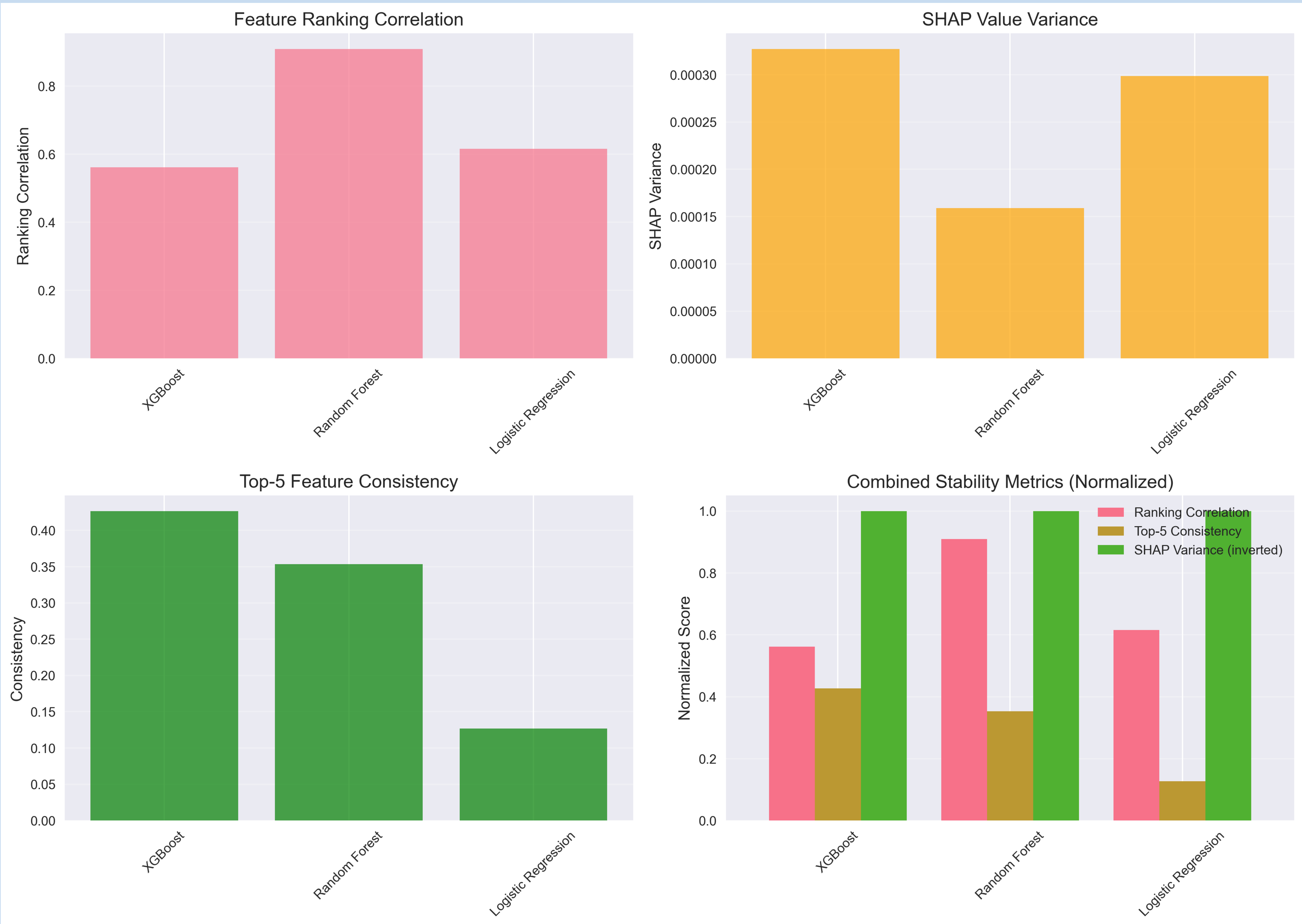
Models:
- XGBoost (TreeSHAP)
- Random Forest (TreeSHAP)
- Logistic Regression (KernelSHAP)

SHAP Value:
- $\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S \cup \{i\}} - f_S]$

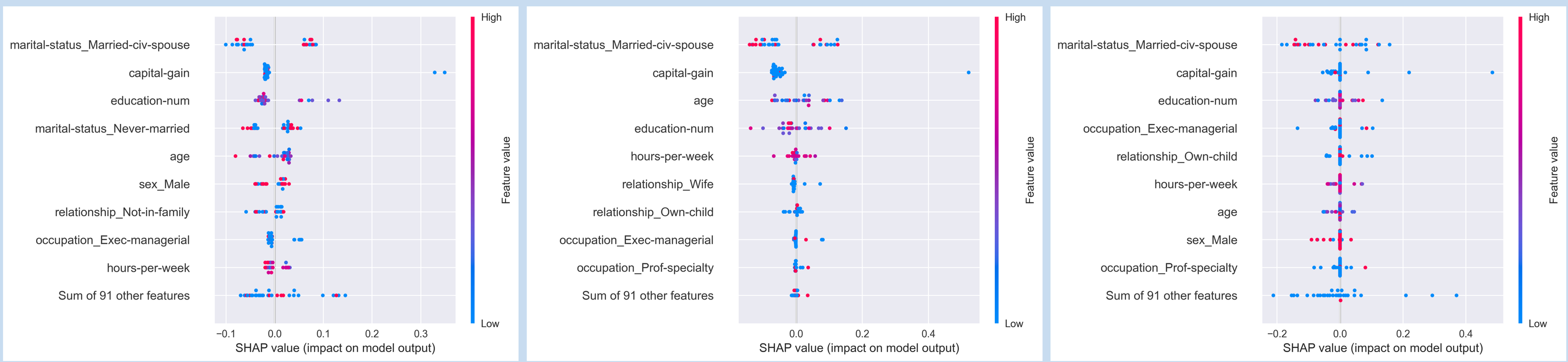
Evaluation Metrics:
- **Ranking Correlation**: Spearman ρ (0-1)
- **SHAP Variance**: $\mathrm{Var}(\phi_i)$
- **Top-k Consistency**: % consistent features

Experimental Setup:
- Random seeds: 42, 123, 456
- 100 test instances analyzed
- Top-k: 3, 5, 10 features

Visualizations

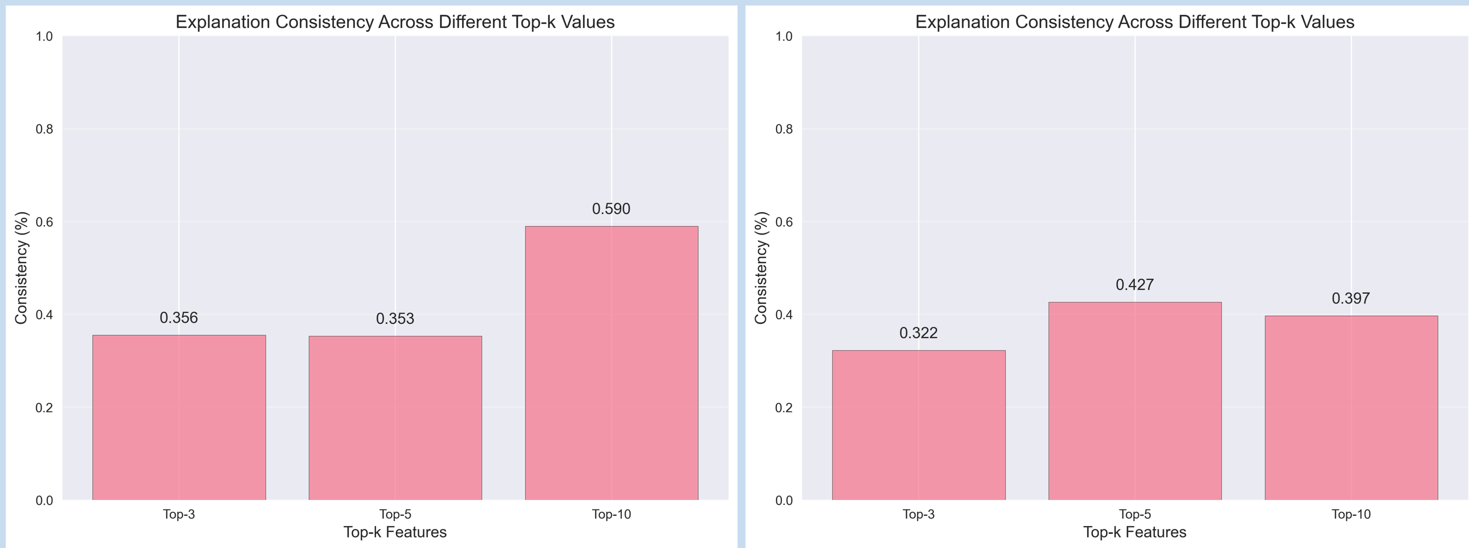


Model Stability Comparison



SHAP Summary: RF, XGBoost, Logistic Regression

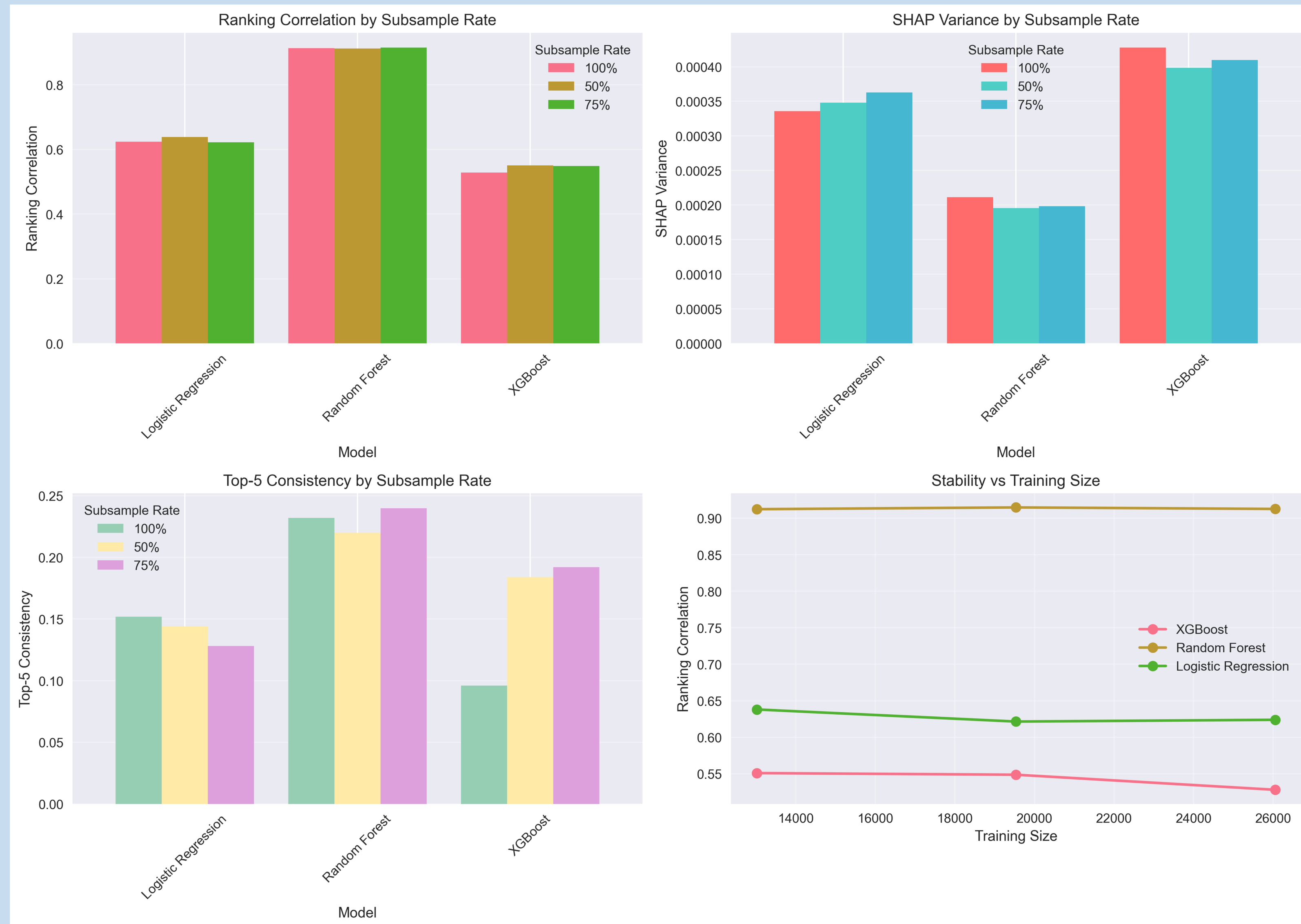
Detailed Analysis



Consistency: RF (left), XGBoost (right)

Model Insights:
- **RF**: Highest stability (0.909), ensemble reduces variance
- **XGBoost**: Moderate (0.562), best top-5 consistency (0.427)
- **LR**: Lower consistency, KernelSHAP variability

Subsampling Analysis



Stability vs. Sample Size

Ranking Correlation by Sample Size:

Model	50%	75%	100%
RF	0.912	0.915	0.913
XGB	0.551	0.549	0.528
LR	0.638	0.621	0.624

Key Findings:
- RF: High stability (0.91+) across all sizes
- Sample size: **Limited impact** on stability
- Architecture: Primary factor

Conclusions & Recommendations

Main Conclusions:
1. Random Forest provides the most stable SHAP explanations
2. Sample size has limited impact on explanation stability
3. Model architecture significantly impacts stability
4. Ensemble methods offer superior stability

Recommendations:
- Use **Random Forest** with TreeSHAP for critical applications
- Focus on **top-5 or top-10** features for stable insights
- Sample size (50%-100%) has **minimal impact** on stability
- Run **multiple analyses** with different seeds