Improving Feature Selection in Predictive Model Analytics Using Explainable AI (XAI)

Avraham Sikirov

March 20, 2025

Abstract

High-dimensional tabular datasets often include irrelevant or redundant features that can degrade model performance and interpretability. Traditional feature selection methods (e.g., correlation-based or mutual information) may effectively reduce dimensionality but often provide limited insight into *why* certain features matter. In this project, we propose integrating Explainable AI (XAI) methods such as SHAP and LIME into the feature selection pipeline to enhance both interpretability and efficiency. We evaluate our approach on four publicly available datasets: the UCI Adult Income, Titanic Survival, Diabetes Prediction, and Loan Default datasets. Our findings show that, while correlation-based feature selection sometimes outperforms the baseline in terms of accuracy and F1-score, SHAP-based selection yields mixed results and often underperforms in our specific experimental setup. Potential reasons include fewer available features and dropping crucial but seemingly low-importance features. Nonetheless, XAI-based methods provide richer explanations of feature importance, which can be critical for transparency and trust in high-stakes applications. We conclude that combining XAI insights with careful feature engineering may improve both performance and interpretability, though computational cost and dataset characteristics must be carefully considered.

1 Problem Description

In the data science (DS) pipeline for tabular data, feature selection is critical to improving model performance and interpretability. However, traditional methods frequently ignore the underlying *reasons* why features are deemed valuable or irrelevant. This black-box approach can be problematic

in high-stakes environments, where domain experts must understand model behavior. Hence, this research seeks to enhance feature selection by integrating Explainable AI (XAI) methods, aiming to:

- Improve model performance by removing non-informative or redundant features.
- Increase interpretability by explaining the significance of selected features.
- Reduce training time and complexity while preserving accuracy.

Despite existing feature selection strategies such as recursive feature elimination or mutual information-based filters, these methods offer limited transparency. Our premise is that methods like SHAP and LIME can supply additional insights into individual feature contributions, ultimately leading to a more interpretable and efficient pipeline.

2 Solution Overview

Our solution integrates traditional selection techniques (e.g., correlation analysis) with XAI-based scoring methods, specifically:

1. **Baseline**: Train and evaluate a model (Random Forest) with *all* features to measure initial performance.
2. **Traditional FS**: Apply correlation-based (or similar) feature selection to filter out features having low correlation with the target.
3. **SHAP-based FS**: Train a model on the full feature set and compute SHAP values. Rank features by their average SHAP importance and keep the top k.
4. **Evaluation**: Compare each approach (baseline, correlation-based, and SHAP-based) using Accuracy and F1.

We implemented this workflow on four datasets that reflect different domains and data characteristics:

- **UCI Adult Income**: Predicting whether annual income exceeds 50K.
- **Titanic**: Predicting passenger survival.
- **Diabetes**: Predicting the onset of diabetes.
- **Loan Default**: Predicting likelihood of loan repayment default.

3 Experimental Evaluation

To evaluate each feature selection method, we measured both *accuracy* and *F1-score* on a held-out test set. Table 1 summarizes the results.

Dataset	**All Features**		**Corr-Based FS**		**SHAP-Based FS**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Adult	0.840	0.605	0.848	0.630	0.745	0.292
Titanic	0.804	0.750	0.748	0.684	0.559	0.432
Diabetes	0.960	0.778	0.960	0.778	0.890	0.083
Loan	0.880	0.111	0.858	0.240	0.858	0.174

Table 1: Comparison of Accuracy and F1 across the four datasets for three feature selection approaches.

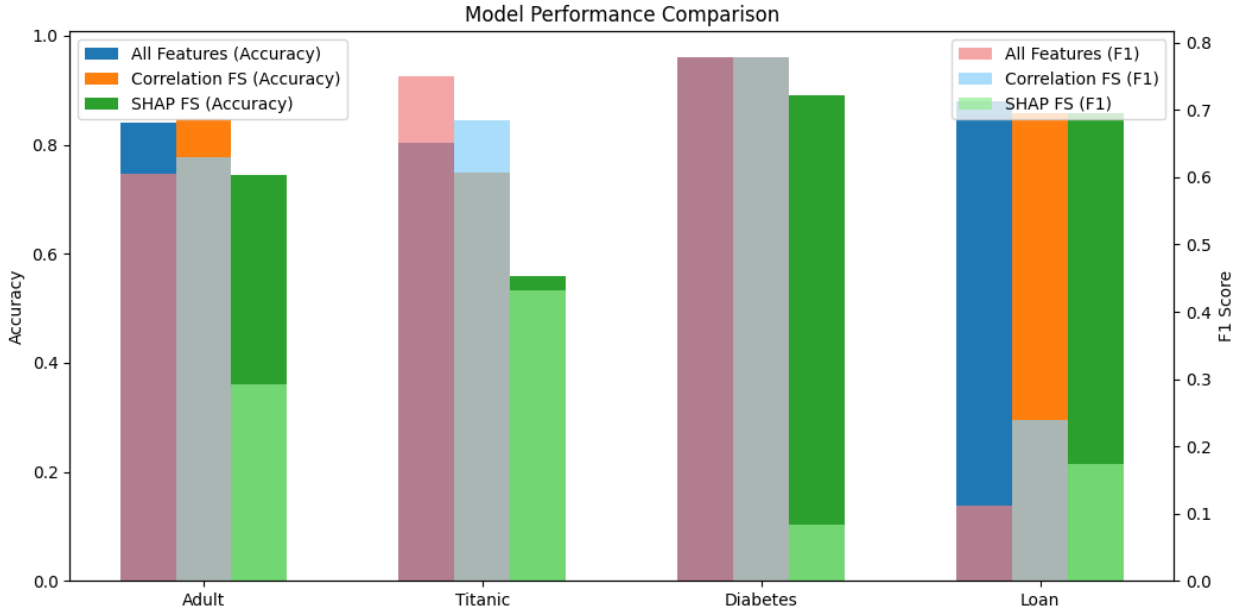

Figure 1: Model Performance Comparison

3.1 Observations and Analysis

1. Correlation-Based FS vs. Baseline: For the Adult dataset, correlation-based FS improved both accuracy (from 0.840 to 0.848) and F1-score (from 0.605 to 0.630). A similar trend is ob-

served for the Titanic dataset, though the improvement is smaller. This method seems to remove predominantly non-informative features without discarding crucial predictors.

2. SHAP-Based FS Results: Our SHAP-based feature selection *underperformed* in several cases. Notably, on the Adult dataset, the F1-score dropped from 0.605 to 0.292. On the Titanic dataset, accuracy decreased from 0.804 to 0.559. A plausible explanation is that, in relatively small or low-dimensional datasets, limiting to only the top k SHAP features can inadvertently remove features that are moderately—but collectively—important. Moreover, the small number of retained features might lead to insufficient coverage of relevant information (e.g., certain categorical or numeric attributes that become pivotal after initial splits).

3. Computational Cost: SHAP computations (particularly for tree-based models with large datasets) can be very expensive, scaling poorly with many rows and features. In our experiments, SHAP analysis sometimes took significantly longer than correlation-based methods, underscoring the importance of balancing interpretability with available computation resources.

4. Potential Improvements: Rather than selecting a fixed top-k, an iterative or adaptive selection approach might more effectively combine XAI-based importance with validation feedback. Future work can explore alternative interpretability methods like LIME or integrated gradient techniques, or combine SHAP with a wrapper-based method to systematically test performance increments.

4 Related Work

Feature Selection. The seminal work by [1] laid the groundwork for analyzing feature relevance using heuristic and statistical methods. Since then, filter-based and wrapper-based techniques have proliferated, focusing mainly on reducing dimensionality or computational burden rather than explainability.

Explainable AI (XAI). The LIME framework introduced by [2] provides local interpretability by approximating complex models around a single prediction. More recently, [3] proposed SHAP, unifying various measures of feature importance into a cohesive theoretical framework. While SHAP offers powerful insights into model predictions, it can be computationally expensive for large-scale data.

In this work, we combined the above lines of research, using correlation-based feature selection as a baseline and employing SHAP for more nuanced explanations. However, as our experiments

show, caution must be exercised when discarding features solely by rank; moderate-importance features can be collectively predictive and thus remain essential.

5 Conclusion

We set out to integrate XAI methods into the feature selection pipeline for tabular data, aiming to enhance both model performance and interpretability. Our results indicate that while correlation-based methods can improve upon the baseline in certain scenarios, a straightforward top-k SHAP-based selection may sometimes degrade accuracy and F1. Possible reasons include the removal of multiple moderately important features, the small effective dimensionality in certain datasets, and the computational costs associated with SHAP.

Nevertheless, these findings do not negate the potential of XAI in feature selection. Instead, they highlight the importance of balancing interpretability, performance, and computational feasibility. Hybrid strategies that incorporate user/domain knowledge, adaptive cutoff points, or iterative wrapper-based methods could yield better results in practice. Future work might also explore additional interpretability frameworks (e.g., integrated gradients, counterfactual explanations) for more robust and flexible feature selection strategies.

References

[1] Kira, K. and Rendell, L. (1992). *The Feature Selection Problem: Traditional Methods and a New Algorithm.* In AAAI Proceedings of the Tenth National Conference on Artificial Intelligence (Vol. 2).

[2] Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). *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 (KDD).

[3] Lundberg, S. M. and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions.* In Advances in Neural Information Processing Systems, 30.