# A Dashboard for Simplifying Machine Learning Models using Feature Importances and Spurious Correlation Analysis – Supplemental Material

T. Cech[1], E. Kohlros[2], W. Scheibel[2] and J. Döllner[1]

[1]Digital Engineering Faculty, University of Potsdam, Germany
[2]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Germany

## 1. An example Overview

An example overview of our dashboard for the case discussed in the extended abstract can be found in Figure 1. The standard workflow with our dashboard is as follows: First, the user determines the train-test split of their data offline, trains a machine learning (ML) model, and uploads the dataset, the train-test split, and the final model. Then, the user can assess the quality of all uploaded ML models for the dataset and load one (1). This will result in the model performance view below. There, the user can further assess the quality of the model by seeing further metrics and the confusion matrix (2). Then, the user can interrogate the feature importance scores by looking at the correlation heat map (3) and request the feature importance scores according to our two methods (4). Optionally, the user can test for the occurrence of spurious correlations by clicking the button at (5). If spurious correlations can be found, all of them will be displayed at (6). Otherwise, the dashboard will report that no spurious correlations could be found. After assessing the features according to the feature importance scores and the spurious correlations, the user can select a subset of features to train a Fast-and-Frugal Tree (FFT) and evaluate it on the test dataset (7). This results in an overview of the achieved scores of all FFTs (8). By clicking on the button "Tree Structure and Confusion Matrix", the user gets the additional visualizations as exemplified in Figure 14 and Figure 16 for assessing a single FFT.

## 2. Example Analysis of a Concrete Spurious Correlation

One example of an instance of Simpson's Paradox is shown in Figure 2 for the Khan student dataset [AFL18]. Here, one could assume that "Lesson index" describing the number of already taken lessons is a good indicator of our target "Performance" since they appear positively correlated in the aggregated view. But, by disseminating the dataset per student one can see that this trend is deceptive as students who have shown a strong performance in the beginning get worse and only students with lower performance get stronger. Therefore, it can be concluded that the feature "Lesson index" is no good indicator of "Performance".



## 3. Examples on Additional Datasets

We tested two additional datasets. (1) The Apple Quality dataset (https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality) and (2) the Credit Card fraud dataset (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). It is unknown if either of those datasets include an instance of Spurious Correlation. We could not identify any such instance with our approach for both datasets.

Figure 3–Figure 10 showcases the results for the Apple Quality dataset. The Feature Importance Scores are inconclusive hinting that we can not find a simple heuristic that can be encoded by a FFT. Our tests with the Top-3 features and Top-5 features FFT support this claim. Both models suffer high quality losses of over 15% Precision, Recall and Accuracy. Still, we obtain a simple interpretable model which may be acceptable in certain circumstances.

Figure 11–Figure 16 showcase the results for the Credit Card Fraud dataset. The Feature Importance Scores identify two features that are important for prediction, namely "V12" and "V17". The FFT trained on those two features even identifies that one feature is enough for prediction. The evaluation reveals that this is enough for reaching a competitive result to the more complex model (compared to Figure 13). We want to emphasize that this results mainly from not taking the special properties of the dataset into account. We naively trained on 75 % of the data of a very imbalanced dataset. Therefore, all classifiers might tend to simply predict that a transaction is not fraudulent to achieve a high score. This is exemplified by our Dashboard which suspiciously points out that such a model is trainable by using only one feature.

## References

[AFL18] ALIPOURFARD N., FENNELL P., LERMAN K.: Using Simpson's paradox to discover interesting patterns in behavioral data. In *Proc. 12th International Conference on Web and Social Media* (2018), ICWSM '18, AAAI, pp. 2–11. doi:10.1609/icwsm.v12i1.15017. 1

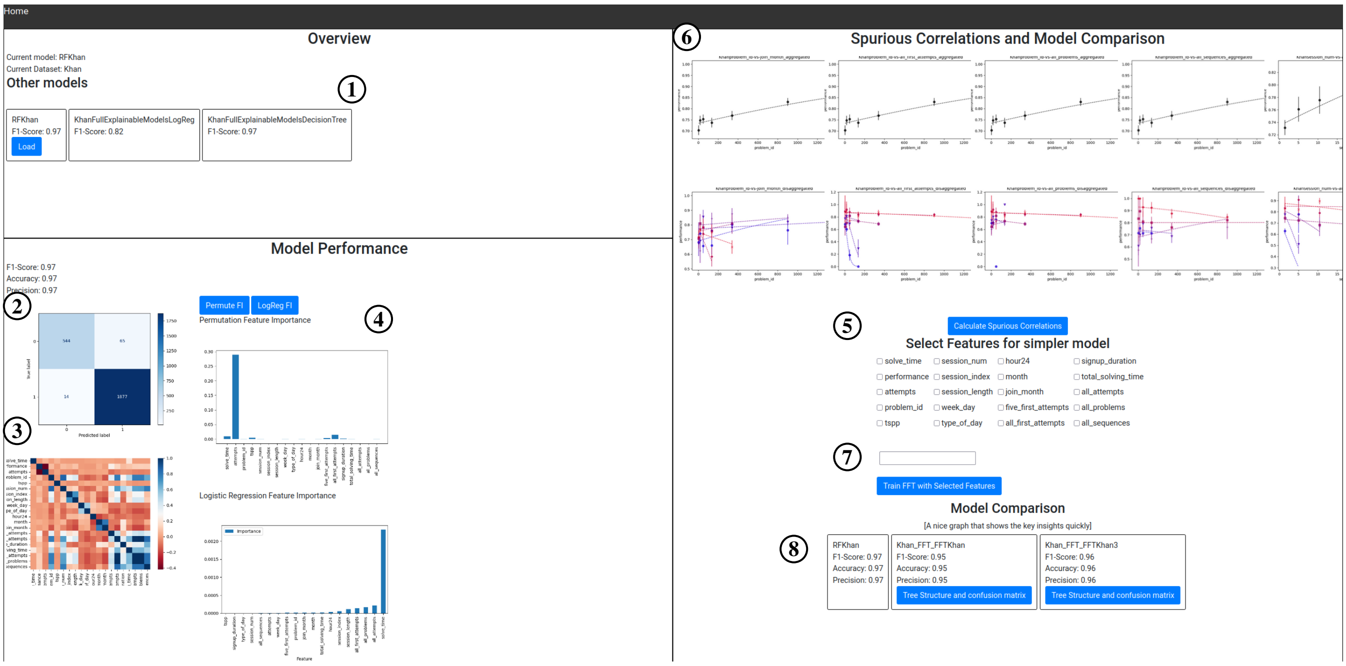


**Figure 1:** *Overview of the Dashboard.*

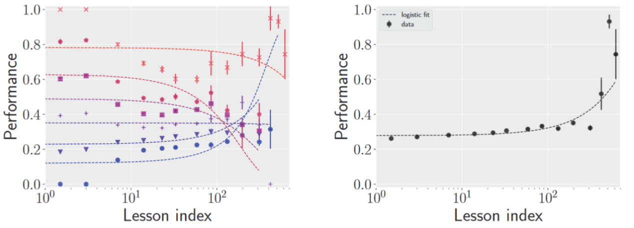


**Figure 2:** *Spurious Correlations and Instances of the Simpson's Paradoxon.*

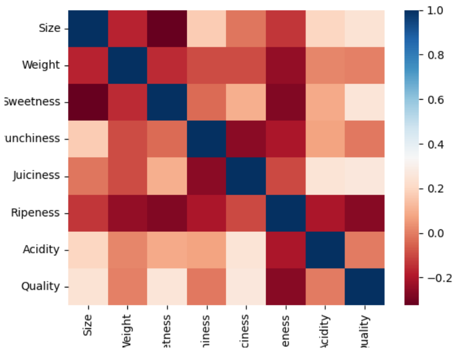


**Figure 3:** *The correlation matrix for the Apple Quality dataset.*

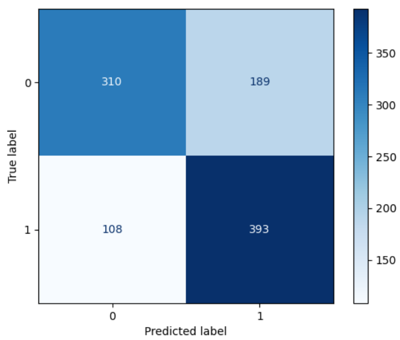


**Figure 4:** *The confusion matrix for the Apple Quality dataset for the original Random Forest model.*

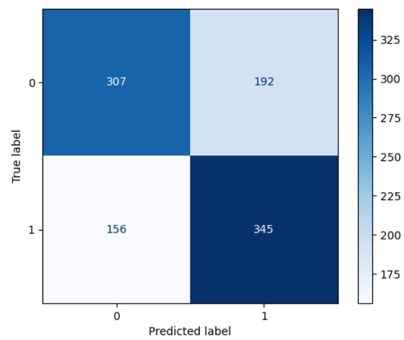


**Figure 5:** *The confusion matrix for the Apple Quality dataset for the FFT using the Top-3 features according to the Feature Importance Scores.*

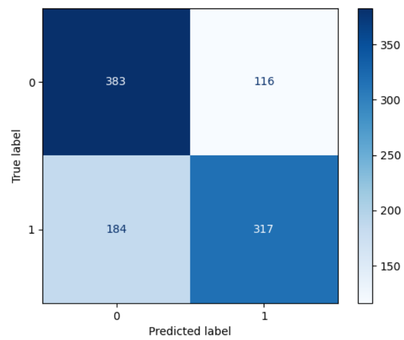


**Figure 6:** *The confusion matrix for the Apple Quality Dataset for the FFT using the Top-5 features according to our Feature Importance Scores.*

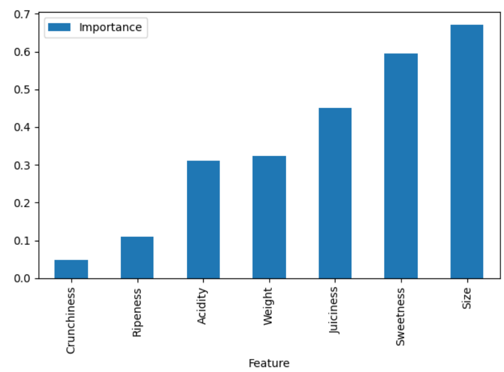


**Figure 7:** *The Feature Importance Score according to FeatureExplorer for the Apple Quality dataset.*

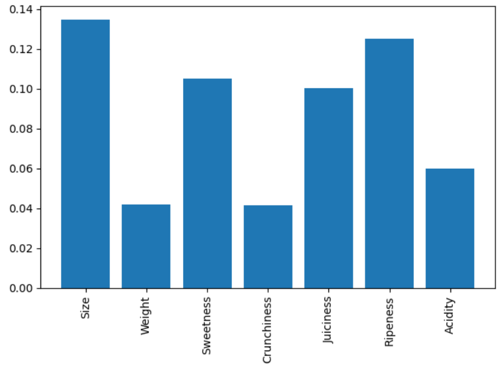


**Figure 8:** *The Feature Importance Score according to the permute-and-predict technique for the Apple Quality dataset.*

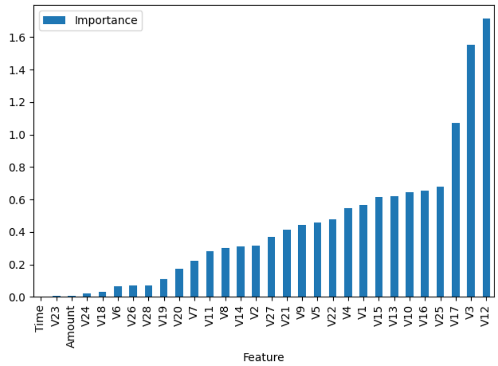


**Figure 11:** *The Feature Importance Score according to FeatureExplorer for the Credit Card Fraud dataset.*

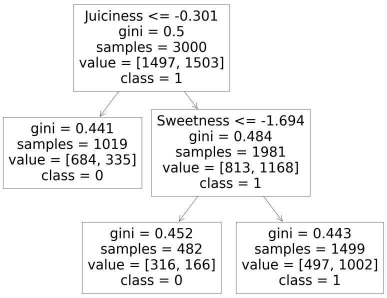


**Figure 9:** *The FFT Tree with the Top-3 features for the Apple quality dataset according to the Feature Importance Scores.*

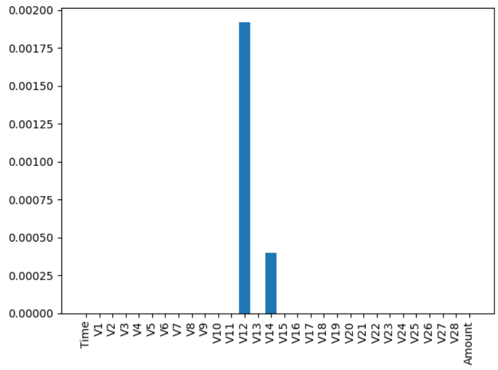

**Figure 12:** *The Feature Importance Score according to permute-and-predict for the Credit Card Fraud dataset.*

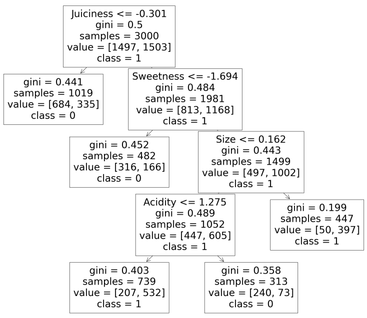


**Figure 10:** *The FFT Tree with the Top-5 features for the Apple quality dataset according to the Feature Importance Scores.*

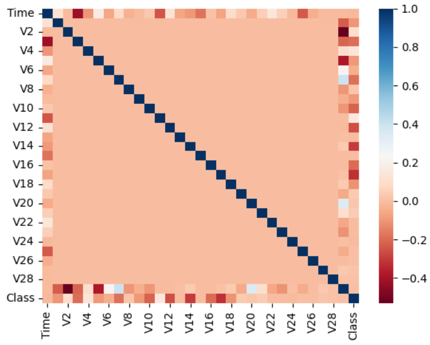

**Figure 13:** *The correlation matrix for the Credit Card fraud dataset.*

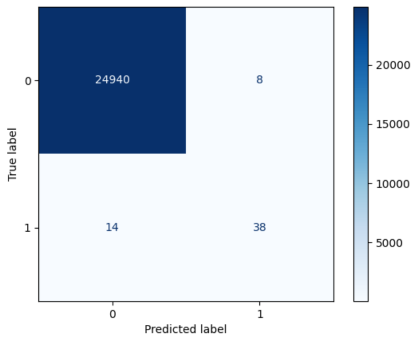


**Figure 14:** *The confusion matrix of the FFT matrix using only the Top-2 features of the Credit Card Fraud dataset according to the Feature Importance Scores.*

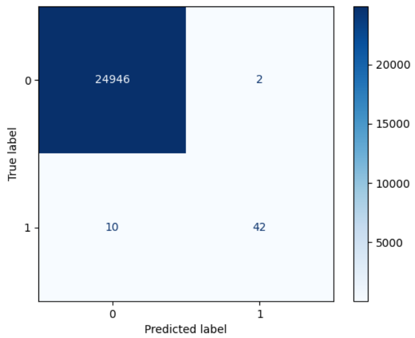


**Figure 15:** *The confusion matrix of the original Random Forest model for the Credit Card Fraud dataset.*

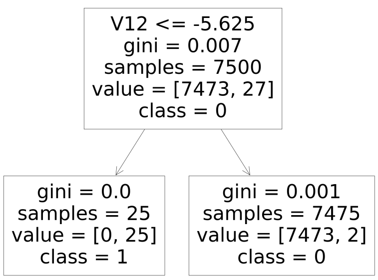


**Figure 16:** *The FFT with the Top-2 features for the Credit Card fraud dataset according to the Feature Importance Scores.*