A Dashboard for Simplifying Machine Learning Models using Feature Importances and Spurious Correlation Analysis

Tim Cech *, Erik Kohlros, Willy Scheibel, and Jürgen Döllner

* University of Potsdam, Digital Engineering Faculty, Germany

1 Motivation

The field of *Explainable AI (XAI)* wants to provide techniques that target to explain Black Box Models, e.g., by analyzing certain properties which are considered decisive for the prediction of the model [LRBB*23]. As Rudin argues, this basic property of many XAI methods leads to some sort of obfuscation since the complex model is not directly explained [Rud19]. She argues that the XAI community should focus on obtaining interpretable models instead. One such model is the *Fast-and-Frugal Tree (FFT)*, a basic Decision Tree with the additional property that there are at max two nodes per level. In the past, Chen et al. have used FFTs in the context of Software Defect Prediction and exemplified, that an FFT model can be competitive with state-of-the-art models[CFKM18]. One way to determine which feature could qualify for such a selection is *Feature Importance Scores* [LRBB*23]. However, Teng et al. have shown that such feature-based analysis can lead to misleading conclusions when not considering potential *Spurious Correlations* as exemplified in their VISPUR system [TAL24]. In this work, we present a proof-of-concept for a dashboard that combines Feature Importance Scores with the analysis of Spurious Correlations.

2 Dashboard

An example overview of our dashboard for the case discussed in the paper can be found in Figure . 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).

Preliminary Case Study

We exemplify our approach on the Khan student dataset [AFL18]. The authors have shown that some instants of Simpson's Paradox are present in this dataset. As our complex model, we use a Random Forest with standard parameters from the scikit-learn library. We use a randomized stratified train-test split with 75% of data used for training and 25% for testing. One example of a Spurious Correlation is shown in Figure 2. Here, one could assume that the "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". After reviewing all Feature Importance Scores and Spurious Correlations as shown in Figure 3, we conclude that "all_first_attempts" and "attempts" are an important indicator of "Performance". Indeed, by training an FFT on only those two features, we obtain a model with similar performance. Therefore, we obtained a model with over 96 % accuracy by only using 2 of the available 19 features. This model only performs 1 % worse in all metrics compared to our original complex Random Forest classifier.

Acknowledgements

This work was partially funded by the German Federal Ministry for Education and Research (BMBF, grant 01IS22062, "AI research group FFS-AI"). This work is part of project 16KN086467 ("DecodingFood") funded by the Federal Ministry for Economic Affairs and Climate Action of Germany.

Dashboard Examples

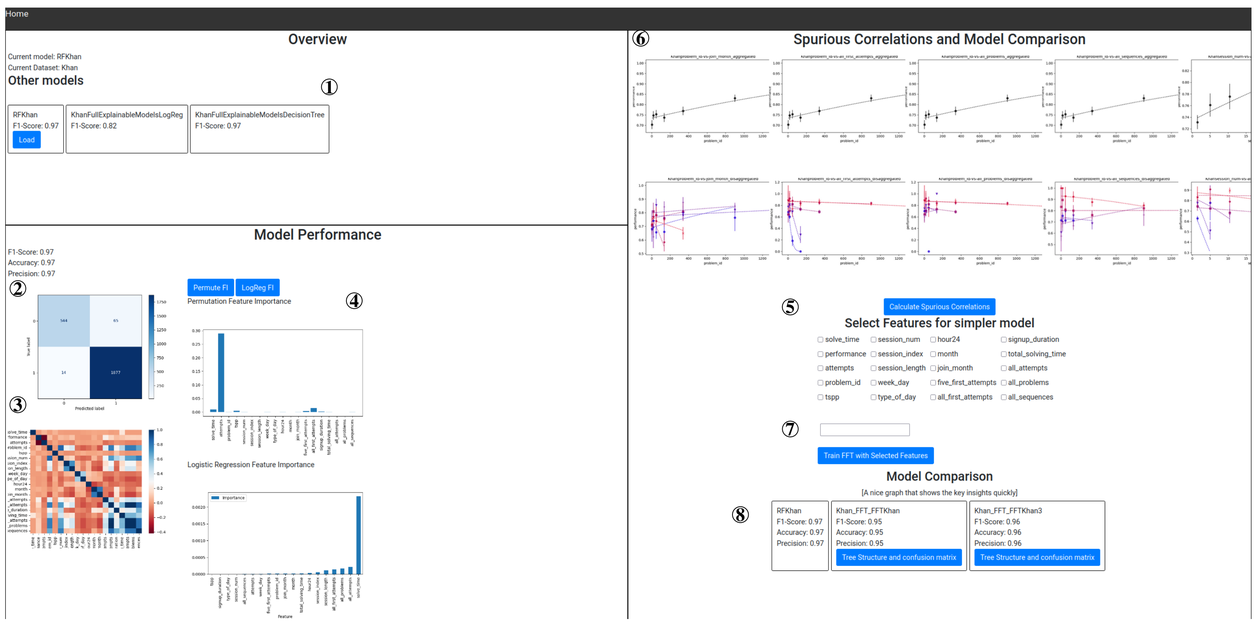

Figure 1: *An overview of the dashboard for the Khan student dataset. In this case, spurious correlations are detected (top right), which informs the Feature Importance scores (bottom left). Based on this information, a user can train an FFT.*

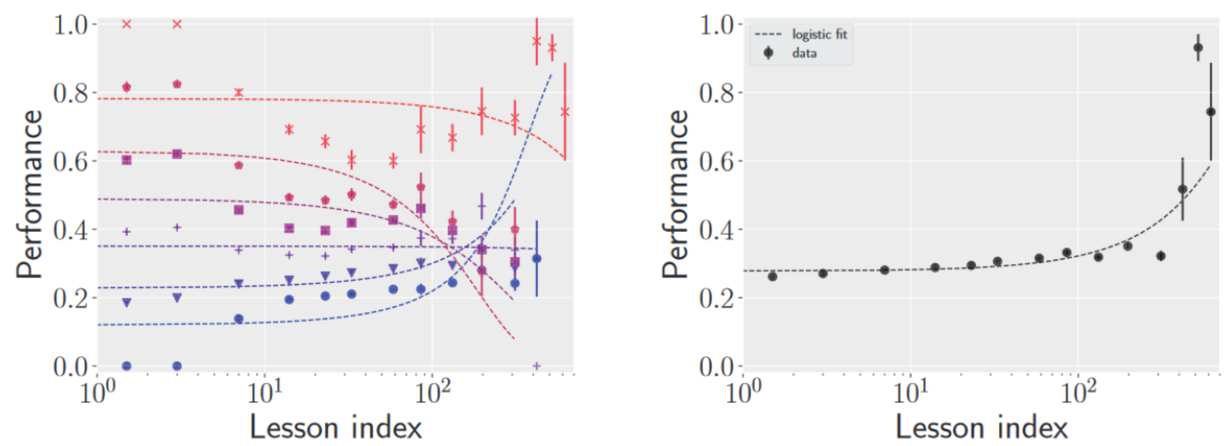

Figure 2: *Shown is an example of Spurious Correlation from the overview mentioned in Figure 1. Here, one can see that the Lesson Index (the number of Lessons taken by a student) seems to correlate well with the student's performance. The trend is created by students with bad performance getting better and better students getting worse.*

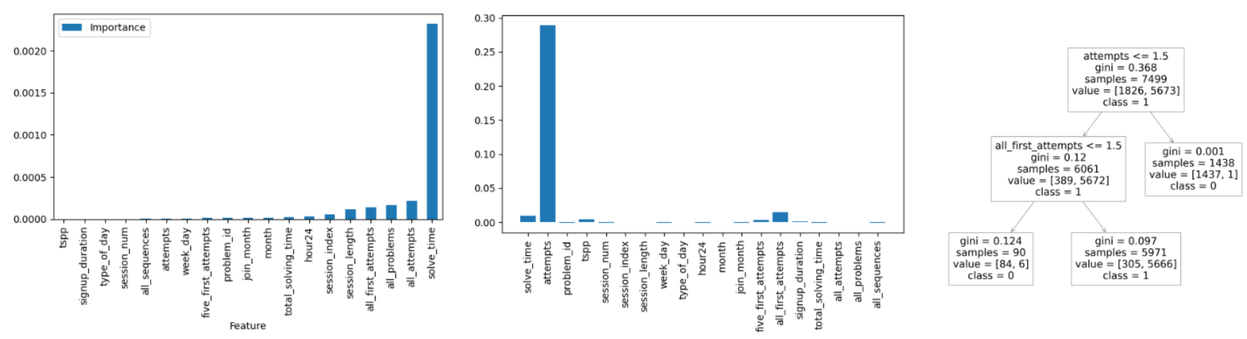

Figure 3: *Based on the information from Figure 1 and Figure 2, the user decides that attempts and all_first_attempts are the best predictors for final performance. Indeed, the FFT trained on these two features achieves a competitive score compared to the initial Random Forest.*

3 Conclusions

The current state of the dashboard allows for an overview of multiple trained FFTs that are derived using a user's domain knowledge. Although these models usually have a reduced accuracy, they are more explainable and use only lightweight abstractions of the data. Such an interface enables users to explore Spurious Correlations and Feature Importances to increase their trust in the used features. For future work, we want to extend the evaluation and document users' decisions on their final FFT. Further, the dashboard can be extended to further support the identification of harmful correlations [DHA*21] or of other types of Spurious Correlations [Vig15]. Furthermore, our dashboard design is clearly in an early stage that should be improved upon in the future.

References

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.

CHEN D., FU W., KRISHNA R., MENZIES T.: Applications of psychological science for actionable analytics. In *Proc. Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2018), ESEC/FSE '18, ACM, pp. 456–467.

DENTON E., HANNA A., AMIRONESEI R., SMART A., NICOLE H.: On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society 8*, 2 (2021), 1–14.

LA ROSA B., BLASILLI G., BOURQUI R., AUBER D., SANTUCCI G., CAPOBIANCO R., BERTINI E., GIOT R., ANGELINI M.: State of the art of visual analytics for explainable deep learning. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, pp. 319–355.

RUDIN C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*, 5 (2019), 206–215.

TENG X., AHN Y., LIN Y.-R.: VISPUR: Visual aids for identifying and interpreting spurious associations in data-driven decisions. *IEEE Transactions on Visualization and Computer Graphics 30*, 1 (2024), 219–229.

VIGEN T.: *Spurious correlations*. Hachette UK, 2015.

Tim Cech, M.Sc.
tim.cech@uni-potsdam.de
+49(0)176 540 760 48
hpi3d.de
www.timcech.de/

Prof. Dr. Jürgen Döllner
office-doellner@hpi.de
www.hpi3d.de

Computer Graphics Systems Group
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2–3
D-14482 Potsdam, Germany

www.hpi3d.de

SPONSORED BY THE
Federal Ministry of Education and Research

Universität Potsdam