




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

T. Cech¹ , E. Kohlros², W. Scheibel²  and J. Döllner¹ 

¹Digital Engineering Faculty, University of Potsdam, Germany

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

Abstract

Machine Learning models underlie a trade-off between accuracy and explainability. Given a trained, complex model, we contribute a dashboard that supports the process to derive more explainable models, here: Fast-and-Frugal Trees, with further introspection using feature importances and spurious correlation analyses. The dashboard further allows to iterate over the feature selection and assess the trees' performance in comparison to the complex model.

CCS Concepts

• **Human-centered computing** → **Visualization techniques**; • **Information systems** → **Users and interactive retrieval**;

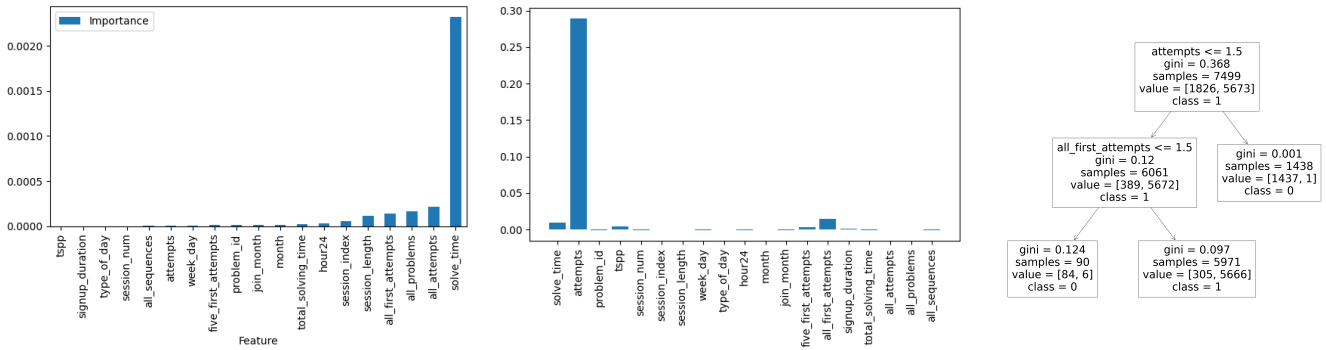
1. Introduction

In recent years, *Artificial Intelligence (AI)* models become more prevalent in public discourse. Often, AI models are complex and can be considered a *Black Box*, as they suffer from a lack of explainability; especially when used in high-stake decision contexts. The field of *Explainable AI (XAI)* wants to provide techniques that target to explain such 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 lead 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 to state-of-the-art models [CFKM18]. They argued that for an FFT model to show high quality, one must select few but high-quality features. 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. For it, we show how spurious correlations could help identify important features for training an FFT on them to obtain a simple yet good enough model. This model is benchmarked side-

by-side with the complex model and other FFT variants created by the user.

Related Work. Several techniques were proposed for obtaining Feature Importance Scores [LRBB*23]. One example of such a technique is *FeatureExplorer* by Zhao et al. [ZKM*19]. *FeatureExplorer* trains consecutive small regression models using selected features for determining an importance score. Additionally, we consider permute-and-predict methods [YSOL09]. For permute-and-predict, the values of several features are permuted, and then consequently the trained model predicts the target based on the permuted feature values. If the model changes its prediction, a feature is considered of high importance since it can not be changed significantly without changing the prediction. Hooker et al. have argued that the permute-and-predict technique can be misleading if the features are highly correlated [HMZ21]. We mitigate this risk by detecting Spurious Correlations especially Simpson's Paradox [AFL18b] and, additionally, considering a second feature importance score provided by *FeatureExplorer*. Simpson's Paradox describes the phenomenon that an overall trend that is present in an aggregated dataset might be missing entirely when disseminating the dataset according to the categories of the dataset [GBK17, TAL24]. For finding Spurious Correlations, we use the method of Alipourfard et al. who repeatedly disseminate trends according to the classes and detects whether Simpson's Paradox occurred [AFL18a]. The task of using domain expert knowledge to refine a model is usually referred to as *model steering* [DCCE19]. We follow Chen et al. [CFKM18] in focusing on finding an easy-to-interpret good enough model instead of supporting domain experts to refine an optimal model [DCCE19, CMKK22].



(a) The Feature Importance Score calculated by Feature Explorer [ZKM*19]. (b) The Feature Importance Score as calculated by permute-and-predict [YSOL09]. (c) The FFT generated in our example.

Figure 1: The Feature Importance Scores and the resulting FFT when choosing the top 2 features in our example.

2. Approach

Our dashboard provides an interface for obtaining a simple FFT model trained on a server by letting a user consider Feature Importance Scores and analyzing Spurious Correlations.

Overall Process. First, the user uploads a numerical dataset in the CSV format. In addition, the user has to upload CSV files for the training data, training target, test data, and test target. After uploading the dataset, the user can upload a complex pickled model. Then, the model is unpickled and used for prediction on the test dataset. Consequently, a confusion matrix, a correlation matrix, and our two Feature Importance Scores are shown. By clicking the button “Calculate Spurious Correlations” the algorithm of Alipoufard et al. [AFL18a] is performed resulting in one of two outcomes: Either no instance of Simpson’s Paradox is found and this status is returned or several line plots are shown which shows an overall trend in the aggregated dataset and its disseminated counterpart.

FFT Training. After reassuring oneself that either no instance of Simpson’s Paradox is present in the data or getting the knowledge which features are spoiled by Simpson’s Paradox one can judge based on the Feature Importance Scores which features are reliable and important for prediction. Especially, when both of our Feature Importance Scores agree or are at least as similar one should consider this feature for Training an FFT. After selecting the features one can train the FFT on the training data by clicking “Train FFT with Selected Features”. Afterwards, the FFT is evaluated on the test dataset. The results as well as the trained FFT is shown in a separate view. One can repeat the process until a good-enough model is obtained. For an overview over our Dashboard and further details about the components, we refer to our supplemental material.

3. Preliminary Case-Study

We exemplify our approach on the Khan student dataset [AFL18a]. The authors have shown that some instances of Simpson’s Paradox are present in this dataset. As our complex model, we use a Ran-

dom Forest with standard parameters from the scikit-learn library¹. We use a randomized stratified train-test split with 75% of data used for training and 25% for testing. As we already know from the study of Alipoufard et al. that some Spurious Correlations are present in the dataset and we replicate the method from the authors, we expected to find such instances. After reviewing all Feature Importance Scores and Spurious Correlations as shown in Figure 1, we conclude that “all_first_attempts” and “attempts” are an important indicator for “Performance”. Indeed, by training an FFT on only those two features, we obtain a model with similar performance as shown in Figure 1c. Therefore, we obtained a model with over 96% accuracy by only using 2 of the available 19 features while also increasing the trust in our model because the FFT can be easily interpreted. This model only performs 1% worse in terms of Accuracy, Precision, and Recall than our original complex Random Forest classifier. Additionally, by only considering two features, we avoided all Spurious Correlations which may mislead us [HMZ21].

4. 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. Furthermore, they gain control over training and understanding of the Machine Learning model [Rud19]. For future work, we want to extend on 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.

¹ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

Acknowledgements

This work was partially funded by the German Federal Ministry for Education and Research (BMBF) through 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.

References

- [AFL18a] ALIPOURFARD N., FENNEL 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, 2
- [AFL18b] ALIPOURFARD N., FENNEL P. G., LERMAN K.: Can you trust the trend?: Discovering Simpson’s paradoxes in social data. In *Proc. 11th International Conference on Web Search and Data Mining* (2018), WSDM ’18, ACM, pp. 19–27. doi:10.1145/3159652.3159684. 1
- [CFKM18] 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. doi:10.1145/3236024.3236050. 1
- [CMKK22] CHATZIMPARMPAS A., MARTINS R. M., KUCHER K., KERREN A.: FeatureEnVi: Visual analytics for feature engineering using stepwise selection and semi-automatic extraction approaches. *IEEE Transactions on Visualization and Computer Graphics* 28, 4 (2022), 1773–1791. doi:10.1109/TVCG.2022.3141040. 1
- [DCCE19] DAS S., CASHMAN D., CHANG R., ENDERT A.: BEAMES: Interactive multimodel steering, selection, and inspection for regression tasks. *IEEE Computer Graphics and Applications* 39, 5 (2019), 20–32. doi:10.1109/MCG.2019.2922592. 1
- [DHA*21] 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. 2
- [GBK17] GUO Y., BINNIG C., KRASKA T.: What you see is not what you get!: Detecting simpson’s paradoxes during data exploration. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* (2017), ACM, pp. 1–5. URL: <https://dl.acm.org/doi/10.1145/3077257.3077266>, doi:10.1145/3077257.3077266. 1
- [HMZ21] HOOKER G., MENTCH L., ZHOU S.: Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing* 31, 6 (2021), 82:1–16. doi:10.1007/s11222-021-10057-z. 1, 2
- [LRBB*23] 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. doi:10.1111/cgf.14733. 1
- [Rud19] 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. doi:10.1038/s42256-019-0048-x. 1, 2
- [TAL24] 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. doi:10.1109/TVCG.2023.3326587. 1
- [Vig15] VIGEN T.: *Spurious correlations*. Hachette UK, 2015. 2
- [YSOL09] YANG J.-B., SHEN K.-Q., ONG C.-J., LI X.-P.: Feature selection for MLP neural network: the use of random permutation of probabilistic outputs. *IEEE Transactions on Neural Networks* 20, 12 (2009), 1911–1922. doi:10.1109/TNN.2009.2032543. 1, 2
- [ZKM*19] ZHAO J., KARIMZADEH M., MASJEDI A., WANG T., ZHANG X., CRAWFORD M. M., EBERT D. S.: FeatureExplorer: Interactive feature selection and exploration of regression models for hyperspectral images. In *Proc. Visualization Conference* (2019), VIS ’19, IEEE. doi:10.1109/VISUAL.2019.8933619. 1, 2