Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230346

Challenges of Estimating Global Feature Importance in Real-World Health Care Data

Aniek F. MARKUS[a,12], Egill A. FRIDGEIRSSON[a], Jan A. KORS[a], Katia M.C. VERHAMME[a] and Peter R. RIJNBEEK[a]
[a] *Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands*

Abstract. Feature importance is often used to explain clinical prediction models. In this work, we examine three challenges using experiments with electronic health record data: computational feasibility, choosing between methods, and interpretation of the resulting explanation. This work aims to create awareness of the disagreement between feature importance methods and underscores the need for guidance to practitioners how to deal with these discrepancies.

Keywords. Prediction modelling, Explainable AI, Variable importance, Shapley values, Evaluating explanations

1. Introduction

Personalized medicine aims to provide treatment and prevention tailored to individual patients. Machine learning (ML) models can help with personalized risk prediction based on individuals' characteristics to pursue this goal. When implementing prediction models in practice, explanations can be useful to validate model behavior and/or to create a shared meaning of the decision-making process [1].

Feature importance is often used to explain ML models and identified as useful explanation by clinicians [2]. A higher score implies a higher importance of the specific feature, i.e. a larger impact on the model predictions ('How does the output rely on a variable?') or model performance ('How much is the loss function reduced?'). In the literature, many methods have been proposed to compute feature importance. In this work we focus on model-agnostic methods (i.e. suitable to explain any kind of ML model) to compute global feature importance (i.e. explaining the model as a whole).

In practice, there are several challenges when aiming to estimate feature importance for prediction models developed using electronic health record (EHR) data: I) computational feasibility, II) choosing between methods, and III) interpretation of the resulting explanation. These challenges are not unique to EHR data, but might be magnified due to the large size, high-dimensionality, and sparsity of the data. In the following paragraphs we discuss the three challenges in more detail:

[1] Corresponding Author: A.F. Markus, E-mail: a.markus@erasmusmc.nl.

[2] This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

I) First, not all methods are computationally feasible for big data. Some feature importance methods require examining all possible combinations of K features or rely on conditional distributions which are unavailable in practice and difficult to estimate. Although there is already work dealing with various ways of approximations (e.g. for Shapley values), formal evaluation of methods (and their approximations) is often lacking and otherwise focused on relatively low-dimensional data.

II) Second, there is a need to choose between methods, but guidance on which method is best to use is lacking. In existing studies, the reason for preferring the chosen feature importance method over alternatives is rarely motivated. However, there is increased awareness that different feature importance methods do not align on the generated explanations (e.g. ranking of top features). This makes the 'interchangeable use' of methods problematic. In recent work, Krishna et al. [3] formalized this as the disagreement problem and analyzed how users deal with this problem in practice.

III) Finally, feature importance explanations might not be in line with (user) expectations. Hase et al. [4] state that explanations are socially misaligned when they convey a different kind of information than what users expect. As an example, they mention some unexpected factors (e.g. model seed, hyperparameters) that might influence the resulting explanations more than expected factors (e.g. the data). However, there are also large differences in how feature importance methods work (e.g. different definitions and/or assumptions), which can formally explain variation in explanations. Resulting explanations miss this nuance and are interpreted similarly by end-users. The general lack of consensus on a definition for feature importance makes it impossible to systematically evaluate whether the selected subset of features is truly important.

In this work, we examine these challenges using experiments with real-world health data when estimating feature importance for a model predicting hospital readmission within 30 days.

2. Methods

2.1. Real-world data

We developed a prediction model on the Dutch Integrated Primary Care Information (IPCI) database [5] to answer the following question: "Among adult patients discharged from the hospital (target population), which patients will be readmitted (outcome) within 30 days (time-at-risk) after the visit?". The IPCI database has been mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), which enables standardized extraction and analysis of health care data. This study was approved by the IPCI Governance Board (number 09/2020).

2.2. Model development

We selected a random sample of N = 100,000 patients of which 75% was used for model development ('training set'). The remaining 25% of patients ('test set') was used for validation. We trained prediction models using the python library `sklearn` using logistic regression with L2-regularization (C=0.1) on the training set. Candidate covariates included sex, age (in 5-year groups), and binary variables indicating the presence or absence of recorded conditions and drugs (measured 30 days, 1 year, and any time prior

to index). To obtain a dataset with dimensionality K, we selected the K features with the strongest relation to the outcome based on the Pearson correlation coefficient.

2.3. Feature importance methods

We investigated the following feature importance methods:

- **Permutation feature importance (FI)**: measures the decrease in model predictive performance after random shuffling the values of a certain feature. We measured model performance using the area under the receiver operator curve (AUC) and balanced accuracy (BA). This method has the advantage that it is fast because it does not require retraining of the model, but has the disadvantage that it might extrapolate to unlikely data points.
- **Shapley Additive exPlanation (SHAP) value**: measures the marginal contribution of features, averaged over all orderings in which the subset of features can be constructed. The resulting explanations are considered to result in a 'fair' allocation because they satisfy five desirable properties (efficiency, symmetry, dummy, monotonicity, and linearity). However, depending on the estimation strategy the method can still extrapolate to unlikely data points in case of correlated features. The computation time of exact Shapley values is an NP-hard problem, therefore we studied two approximations:
  - **KernelSHAP**: uses a weighted linear regression to estimate local SHAP values [6], we implemented this using the python library `shap` (with nsamples = 10*num_features + 2048 and l1_reg = 'num_features(10)').
  - **SAGE**: a sampling-based approximation to compute global SHAP values [7], we implemented this using the python library `sage-importance` (with MarginalImputer() and n_permutations = 1000*num_features).

2.4. Experiments

I. Computational feasibility: we investigated which feature importance methods are able to deal with the high dimensionality of EHR data. We measured the computation time when calculating feature importance using each method across different data dimensionalities K = [20, 50, 100]. For Shapley values we evaluated using 500 and 1000 background samples. The experiments were run using 16 cores of an Intel® Xeon® CPU E5 v4.

II. Choosing between methods: we investigated to what extent different feature importance methods result in different rankings of features for EHR data. We investigated the top-10 ranked features as users typically focus only on the most important features.

III. Interpretation of the resulting explanation: we investigated alignment with user expectations. We argue that users expect feature importance for additive models to be in line with the size of the model coefficients. We measured alignment of the feature importance methods with model coefficients, comparing both the ranking (using top-5 overlap, top-5 sign agreement, and Kendall's tau) and normalized values (using mean absolute error).

3. Results

3.1. Experiment I: computational feasibility

Table 1 indicates the required computation times for each of the methods. This shows that the computation times to explain a prediction model quickly explode. Permutation methods are very fast, but the approximations of Shapley values take up to multiple hours. These results suggest that it is critical to further improve the speed of Shapley methods as the explanation times might be a hurdle for model explanation in practice.

Table 1. Computation time (in minutes) for various FI methods.

	Permutation FI		**KernelSHAP**		**SAGE**	
	AUC	BA	500 samples	1000 samples	500 samples	1000 samples
K=20	0,0	0,0	23,7	93,4	1,2	2,5
K=50	0,1	0,1	34,6	129,8	16,6	33,6
K=100	0,2	0,3	52,5	193,8	129,2	260,5

3.2. Experiment II: choosing between methods

The prediction models for K = [20, 50, 100] resulted in performance (AUC) of 0.66, 0.66, and 0.67, respectively, on the test set. Figure 1 presents the most important features identified by each method for the prediction model for K=50. This shows significant differences in the explanations across methods. Not only is there a mismatch in the top features, but also the ordering of features and the direction of effect differ (e.g. X28). Model coefficients, permutation FI, and SAGE roughly identify the same set of important features. However, KernelSHAP leads to a very different set of features. SAGE can additionally capture the direction of effect as this does not capture the change in model error (as permutation FI), but the difference between the actual and average prediction.

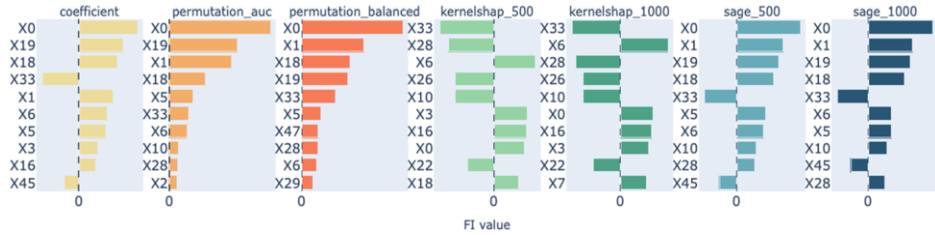

Figure 1. Top-10 ranked features in the prediction model according to different FI methods (K = 50).

3.3. Experiment III: interpretation of the resulting explanation

Figure 2 shows the alignment of feature importance methods with model coefficients measured using three rank-based metrics and one value-based metric. The agreement between model coefficients is highest with SAGE across all metrics, as can be seen from the figure because Top-5/Sign agreement are both 1 (indicating perfect agreement) and the MAE is low (indicating the normalized values are close). These metrics make the discrepancies between model coefficients and KernelSHAP very clear for the top features (Top-5/Sign agreement are both 0.2), but also show that the overall produced ranking (Kendall's tau) and values (MAE) are better than for permutation FI.

Figure 2. Agreement of FI methods with model coefficients. The metrics are scaled between 0-1, with values closer to 1 indicating more agreement.

4. Discussion and Conclusion

It is often important to understand which features are most important for a given prediction model. We have shown there are challenges in different phases of creating such explanations for EHR data, even for a simple classifier. First, the computation times for state-of-the-art feature importance methods are significant, which may be a hurdle to implementation in practice. Second, different feature importance methods often result in different explanations, hence it is important to make an informed choice between methods. Third, even though these observed differences are not always unexpected, e.g. permutation FI/SAGE explains model performance and KernelSHAP explains model predictions, it remains a challenge to communicate these differences to end users.

We only investigated one type of classifier and one prediction task. Results may vary depending on the studied example, but the main findings (e.g. disagreement between methods) have been found in other studies as well (e.g. [3]). Moreover, when investigating other classifiers such as tree-based and/or deep learning methods we expect the problems will only be larger due to the non-linearity of these methods.

This work aims to create awareness of the disagreement between feature importance methods and underscores the need for guidance to practitioners how to deal with the discrepancies between feature importance methods. For this, we argue it is important to make explicit what we mean with feature importance (also for non-linear models) and which goal we aim to fulfill (e.g. to understand model decisions or to improve the model), as this can guide how methods should be formally evaluated and selected.

References

[1] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655. doi: 10.1016/j.jbi.2020.103655.

[2] Tonekaboni S, et al. What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of Machine Learning research. 2019:1-21.

[3] Krishna S, Han T, Gu A, Pombra J, Jabbari S, Wu S, et al. The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:220201602. 2022.

[4] Hase P, Xie H, Bansal M. The out-of-distribution problem in explainability and search methods for feature importance explanations. Adv Neural Inf Process Syst. 2021;34.

[5] de Ridder MAJ, de Wilde M, de Ben C, Leyba AR, Mosseveld BMT, Verhamme KMC, et al. Data resource profile: The integrated primary care information (IPCI) database, the Netherlands. Int J Epidemiol. 2022. doi: 10.1093/ije/dyac026

[6] Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30; 2017.

[7] Covert I, Lundberg S, Lee S-I. Understanding global feature contributions with additive importance measures. Adv Neural Inf Process Syst. 2020;33.