Explaining patient-level prediction models using permutation feature importance and SHAP

Aniek F. Markus, Egill A. Fridgeirsson, Jan A. Kors, Katia M.C. Verhamme, Peter R. Rijnbeek

Background

Feature (or variable) importance is often used to explain prediction models to end-users. These methods rank or measure the predictive power of features. A higher score implies a higher importance of the specific feature, i.e. a larger impact on the model predictions (sensitivity based perspective: "how does the output rely on a variable?") or model performance (predictive power approach: "how much is the loss function reduced?"). Many methods to compute (global) feature importance exist (1), but clear guidance on which method is best to use is lacking. Permutation feature importance (2, 3) is a popular, straightforward approach that is currently supported in the OHDSI PLP package (4). More recently, Shapley values have been advocated for in the literature (5). In this work, we compare which features are important according to permutation feature importance and Shapley Additive exPlanations (SHAP) for a given patient-level prediction model. We give insight into both methods and explore whether it is desired to add SHAP (or other feature importance methods) to the OHDSI PLP package.

Methods

Model development

We developed a prediction model on the Dutch Integrated Primary Care Information (IPCI) database to answer the following question: "Among patients 60 years or older presenting at the general practitioner for an outpatient visit (target population), which patients will die (outcome) within 90 days (time-at-risk) after the visit?". We use a random sample containing 75% of patients ('training set') for model development. The remaining 25% of patients ('test set') was used for validation. To obtain a small model for illustrative purposes, we first selected the top 50 features indicated by LASSO logistic regression and then obtained the final model by retraining using only these 50 features.

Feature importance methods

We investigated the following two model-agnostic feature importance methods:

1. **Permutation feature importance** (2, 3): measures the decrease in model performance after random shuffling the values of a certain feature. We measured model performance using the default scoring option: accuracy.

2. **Shapley Additive exPlanations (SHAP)** (5): inspired from game theory, this method views the prediction task as a game with players ('features') that need to fairly distribute a total payout ('output'). The Shapley value is a solution concept that assigns each feature their marginal contribution to the output averaged over all orderings in which the subset of features can be constructed. As exact computation of SHAP values is not feasible for large data, hence we use the approximation method KernelSHAP (5). To obtain global feature importance we averaged the Shapley values per feature for a (random) sample of 1000 patients. We used a subsample of data to decrease the computational burden of the method.

We compare both methods against the model coefficients as these coefficients are a model-specific alternative to assess feature importance for models where the prediction is the weighted sum of the input values. For more details on the methods used see Table 1.

Table 1: Comparison of feature importance methods.

	Model coefficient	Permutation feature importance	SHAP
Interpretation	- The change in the mean output associated with a change in that term, while the other terms in the model are held constant.	- Decrease in model performance - Importance of all features together does not add up to the model performance (sum can be larger)	- Contribution to predicted output - Importance of all features together adds up to the difference between the actual and average prediction
Advantages	- Readily available for linear models.	- Does not require retraining of the model - Can be compared across problems because of using error ratio	- 'Fair' distribution of importance by satisfying theoretical properties (efficiency, symmetry, dummy and linearity) - Allows contrastive explanation by comparing against subset of data
Disadvantages	- Not model-agnostic, only available for linear models. - Assuming features have the same scale or have been scaled prior to model development.	- Linked to error of the model, which might not be of interest (e.g. to investigate model robustness) - Extrapolates to unlikely unrealistic points (especially if features are correlated)	- Computationally expensive for large data and we have to rely on approximations - Can give importance to features that are not used by the model and have no influence on spread of importance between correlated features

Feature importance evaluation

We performed 3-fold cross validation for a robust estimate of the feature importances. We repeatedly computed permutation feature importance and SHAP on 2/3 of the training set and averaged the resulting importance values. We investigated the top 5 ranked features (using absolute values) and visualized the normalized feature importances.

Results

We identified 289,082 patients 60 years or older presenting at the general practitioner for an outpatient visit in IPCI, of whom 5,623 died during the time at risk (1.9%). The final prediction model showed good internal discrimination (AUC = 0.78). Note it might be possible to improve the performance with a larger, more complex model, but this was not the goal of this work. The observed performance is common for clinical prediction models and was therefore considered appropriate for illustrative purposes.

We first investigated the top 5 ranked features, see Table 2. We found some overlapping features between methods as indicated by the numbers in brackets, e.g. 'infective corneal ulcer previous 30 days' and 'vascular disorder of pelvis previous 30 days' were included in the top 5 by all methods. However, there are also differences; some features included in the top 5 by one method are ranked much lower by another (e.g. 'drug use temozolomide previous 365 days' is on place 13 – 4 – 11 and 'drug use goserelin previous 30 days' on place 45 – 27 – 4). In total 9 different features were identified by the three methods. Note these features should not be interpreted as 'risk factors', but only as features that are important for the given prediction model.

Table 2: Top 5 ranked features in the model according to different feature importance methods (in brackets the number of times each feature was included in the top 5).

	Model coefficients	Permutation feature importance	SHAP
1.	Infective corneal ulcer previous 30 days (3x)	Infective corneal ulcer previous 30 days (3x)	Vascular disorder of pelvis previous 30 days (3x)
2.	Vascular disorder of pelvis previous 30 days (3x)	Vascular disorder of pelvis previous 30 days (3x)	Infective corneal ulcer previous 30 days (3x)
3.	Affective psychosis previous 30 days (2x)	Antibiotics and chemotherapeutics for dermatological use at day of visit	Affective psychosis previous 30 days (2x)
4.	Hypertensive disorder previous 30 days	Drug use temozolomide previous 365 days	Drug use goserelin previous 30 days
5.	Neoplasm of intrathoracic organs previous 30 days (2x)	Neoplasm of intrathoracic organs previous 30 days (2x)	Selective calcium channel blockers with direct cardiac effects use previous 365 days

Next, Figure 1 shows the feature importance of all features included in the model. We normalized values as we are interested in the relative importance of features indicated by each method. This also shows the large variation between methods, especially also for features not included in the top 5.

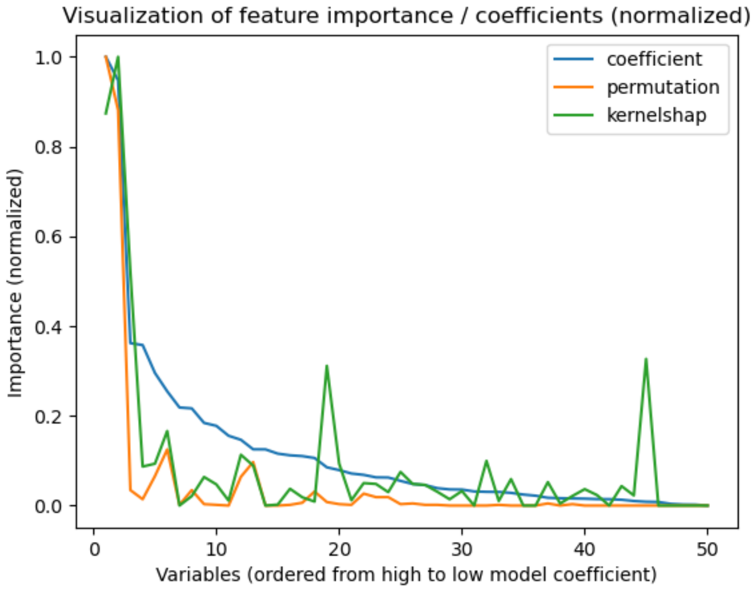

Figure 1: Visualization of (normalized) feature importance.

Conclusion

The results show some agreement between the methods on the relative importance of features, but also large variation. This is not surprising, given that methods vary in their intended behavior (6). For example, permutation feature importance explains *model performance* and model coefficients/SHAP explain *model predictions*. However, the effect of these differences (i.e. a different feature importance ranking) is not always clear to users of these methods. Knowing which feature importance method is best to use is important for reliable interpretation and presentation of prediction models developed within OHDSI. This is a first step in that direction and we plan to do a more in-depth analysis of feature importance methods to understand the influence of e.g. correlated features and model misspecification.

Funding

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.

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. https://doi.org/https://doi.org/10.1016/j.jbi.2020.103655.

2. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J Mach Learn Res. 2019;20(177):1-81.

3. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

4. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-75. https://doi.org/10.1093/jamia/ocy032.

5. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30; 2017 22-11-2019.

6. Covert I, Lundberg S, Lee S-I. Explaining by removing: A unified framework for model explanation. J Mach Learn Res. 2021;22(209):1-90.