Why Do I Trust Your Model? Building and Explaining Predictive Models for Peritoneal Dialysis Eligibility

George Michalopoulos	University of Waterloo, ON, Canada
Helen H. Chen	University of Waterloo, ON, Canada
Yang Yang	University of Waterloo, ON, Canada
Sujan Subendran	University of Waterloo, ON, Canada
Robert R. Quinn	University of Calgary, AB, Canada
Matthew J. Oliver	Sunnybrook Health Sciences Centre, ON, Canada
Zahid Butt	University of Waterloo, ON, Canada
Alexander Wong	University of Waterloo, ON, Canada

Email: {gmichalo,helen.chen,y24yang,sujan.subendran,alexander.wong,zahid.butt}@uwaterloo.ca
rquinn@ucalgary.ca, matthew.oliver@sunnybrook.ca

Abstract

Achieving fairness, accountability and transparency is of vital importance when using machine learning (ML) techniques in the healthcare realm. Yet, the myths of the "black box" of ML algorithms still exist among healthcare professionals. In this research, we developed a ML model for the eligibility of patients for peritoneal dialysis and employed various interpretability techniques to explain the models to nephrologists to gain their trust in the model. We compared different model-specific and model-agnostic ML interpretability strategies with traditional statistical analysis methods and we analyzed their applicability in healthcare domain.

1 Introduction

The "black-box" nature of complex machine learning (ML) models hinders the acceptance of predictive models in healthcare domain. Appropriate techniques for interpretable ML models can be an effective solution to achieve fairness, accountability and transparency (FAT). There are two main groups of interpretable ML, model-specific and model-agnostic. Model-specific or intrinsic interpretability is achieved by the adoption of explainable models and model-agnostic methods are created by using a second model to provide explanations for an existing model.

2 Methods

2.1 Predicting PD Eligibility

Peritoneal Dialysis (PD) is an effective home-based therapy with comparable outcomes to in-center hemodialysis, with potentials to maintain a better quality of life for dialysis patients. We used the Dialysis Measurement, Analysis and Reporting System (DMAR®), collected from renal programs at multiple hospitals in Alberta. The dataset was randomly divided into training (80%) and testing (20%) to examine the accuracy, sensitivity, specificity, and balanced accuracy (average of sensitivity and specificity) for each model. The goal of this study was to investigate the crucial question of how efficiently we could explain different ML models that predict a patient being a viable candidate for PD modality.

2.2 Model-specific Explanation

Each model-specific explanation was designed for a unique ML model, and their functionality is usually based on examining the internal model structures and parameters. We considered two main groups of model-specific methods: (i) tree-based models, including the feature selection mechanism of Random Forest (RF) or XGBoost where the average impurity decrease of each feature is calculated; and (ii) generalised linear models (GLMs) such as linear regression and logistic regression where the weight of each feature reflect their importance. In addition, we considered different types of regularization like LASSO (L1) and Ridge (L2).

2.3 Model-agnostic Explanation

These methods treat a model as a black-box and do not inspect internal model parameters and thus a model-agnostic analysis can be broadly applicable to various ML models. We analyzed the results of three popular model-agnostic methods on our case study: (i) Permutation Feature Importance (PFI) where the importance of a specific feature can be calculated by measuring the increase in the prediction error of the model after feature's values are permuted [1] (ii) local interpretable model-agnostic explanations (LIME) which is a method for fitting local, interpretable models for explaining a single prediction in any black-box ML model [2], and (iii) the SHAP framework that calculates the contribution of each feature to the prediction of a specific instance (patient), by calculating the Shapley values [3] (the average marginal contribution of a feature value over all possible coalitions)

2.4 Traditional Statistical Modelling

We performed a univariate analysis using the Student's t-test for determining the feature importance. Furthermore, we experimented with three popular (in the statistical field) methods for selecting the appropriate features for our model: (i) Forward Selection where the model starts with no features and in each iteration, the best feature is added to the model (ii) Backward Elimination which starts with a set of features and in each iteration the least significant feature is removed (iii) Stepwise selection which is a method that is a combination of the forward and the backward technique.

3 Results and Discussion

All methods agreed on the importance of some features like Albumin and BMI as significant features in the prediction of PD eligibility. However, between the methods that are using multivariate analysis (e.g., Forward, Backward, Stepwise selection) and the tree-based interpretability methods, there are many features that are chosen in one method but discarded from the other methods. This is why the ML developer should always consider the strong and weak points of each method and carefully consider the method that is most suited to their needs.

4 Conclusions and Future Work

In this research, we presented different machine learning interpretability methods and we analyzed their effectiveness and applicability with a real-world health dataset. We also provided a comparison of these methods to statistical methods that are traditionally used in health research. We plan to investigate how the different interpretation artifacts generated from each interpretability method could impact the users' trust in the model.

References

[1] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, pp. 1340–1347, 04 2010.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 1135–1144, ACM, 2016.

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (USA), pp. 4768–4777, Curran Associates Inc., 2017.