Interpretability for the Baseline Classification Model

Daniel Wang
Computer Science & Data Science
College of LSA
University of Michigan

Maria Han Veiga
Department of Mathematics &
Michigan Institute for Data Science
University of Michigan

1 Abstract

Interpretability refers to the ability to understand and interpret how a machine learning model or algorithm makes decisions based on input data. In reality, there has been an increasing demand for model interpretability as more and more organizations reply on machine learning models to make business decisions. One of the main reasons for the importance of model interpretability is that it helps to build trust and confidence in the decision-making process. By understanding how a model makes its predictions or recommendations, stakeholders can better evaluate the accuracy and reliability of the model's outputs. This is especially important in industries, for example healthcare, where incorrect decisions can have serious consequences.

Interpretability can also help to identify biases or errors in a model's decision-making process. If the data feed to the model or model itself is biased or flawed, it can lead to inaccurate or unfair results. By understanding how a model is making its decisions, it is possible to identify and correct these biases.

In this research, a baseline classification model has been trained using the Brazilian Peritoneal Dialysis Multicenter Study BRAZPD dataset (Fernandes et al., 2008). The goal is to predict patient mortality outcomes with the model and explore the interpretability of the model using SHAP values (Lundberg and Lee, 2017).

2 Introduction

The Brazilian Peritoneal Dialysis Multicenter Study (BRAZPD) (Fernandes et al., 2008) is a long-term study that monitors the peritoneal dialysis (PD) reality at a monthly frequency in Brazil.

Renal failure along with several comorbidities are critical conditions that could result in patients' death. BRAZPD dataset (Fernandes et al., 2008) collected data associated with renal failure and the comorbidities from patients who participated in this PD study with a follow-up from December 2004 to February 2007.

However, the dataset has some artifacts, such as the incorrect handling of drop out records and time-dependent features, which made the data cleaning process necessary before. In the initial phase, data cleaning and relabeling the features were applied to the dataset. Later on, a baseline classification model has been setup targeting on patients' mortality (binary classification) and patient cause of death (multiclass classification). The multiclass classification was then further broken down into binary classification tasks, such that model will only predict the top 2 cause of death: Cardiovascular (CVD) and Sepsis not-related to treatment.

The model constructed in this research shows that "Age" was the most contributing factor for the patients' mortality within the follow-up window. Further investigations of the the model's output will be analyzed later in this paper.

3 Method

3.1 Data Processing

Due to the existence of artifacts, such as outliers and incorrect handling of the data, data cleaning were applied to the BRAZPD dataset (Fernandes et al., 2008).

Parts of the sample did not last through the duration of follow-up window, namely 75 months. Reasons can be attribute to patients' death during the follow-up window, patients switching to other types of dialysis and opting out of the PD study, etc.Therefore, data cleaning was done by determining the end-point for each patient record in the sample based on the the individual record duration. Any data that contradicts with its corresponding following up endpoint was set to NaN.

Outliers was also filtered out with standard deviations for a later use in visualizing the data. As the

visualization of the patients' glucose level shown in Figure 1, The visualization showed the distribution of each feature associated with the PD study. As an example, Figure 1 showed the patients' glucose level during the observation window.

As previously mentioned, the number of participants in the PD study decreases over time mainly due to patient mortality and switching to alternative dialysis methods, as shown in Figure 2.

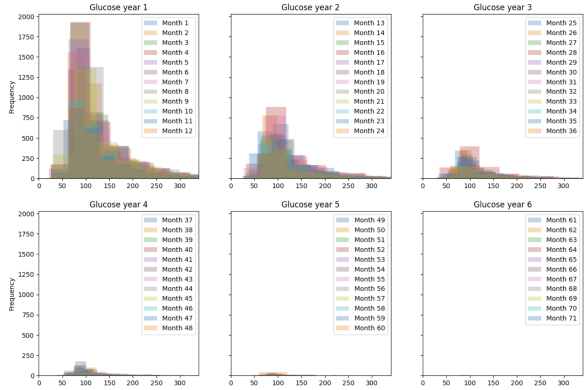

Figure 1: Distribution of patients' glucose over time

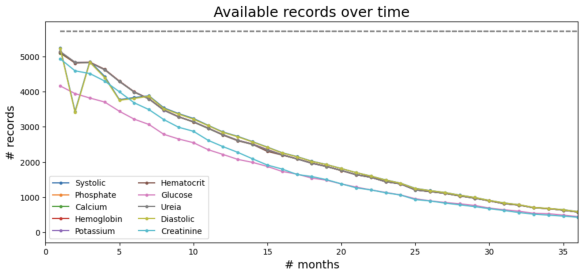

Figure 2: Available records overtime

3.2 Predictive models

A baseline classification model was set up on patient mortality (binary classification):

$$f : X \rightarrow \{0, 1\}$$

The function f takes on a value of 1 to indicate that a patient has died and a value of 0 to indicate that the patient is alive

Another binary classification model was set up on patient cause of death. There were many targets in patients' cause of death. Therefore, to break down the problem of classifying different cause of death into a binary classification problem, the top 2 causes were manually picked among all other causes based on the distribution histogram of the causes. As Figure 3 indicates, top 2 causes are 1 and 5. They correspond to CVD (Cardiovascular) and Sepsis not-related to treatment respectively. Each cause is then determined by a single binary classification model.

Both models used the same set of features, which included various features related to the patient's body condition. However, the models for the cause of death further restrained the dataset where they only consider the patients who has died during the study.

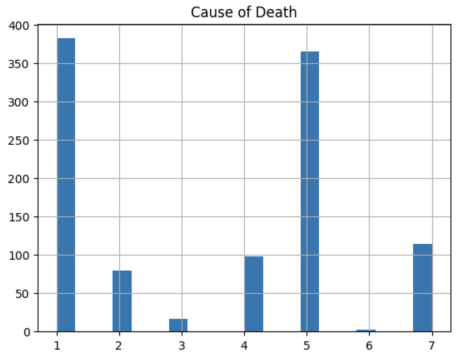

Figure 3: Distribution of cause of death

The classification model was implemented using XGBoost (Chen and Guestrin, 2016), a popular machine learning algorithm for supervised learning tasks. The classification model is trained on a given dataset that contains 22 features that contribute to renal failure. The model also applied cross-validation: stratified k-fold procedure where k = 10 in this study. This strategy preserves the class distribution of the data in each fold.

During training, the model was fit to the data to learn the relationship between the feature set and the target variable. After training, the model's predictions are evaluated using various performance metrics, such as the normalized accuracy, F1 score, recall score, and precision score.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} * 100$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1\ score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

The model implemented in this research also provides an option to use SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) to deal with class imbalance in the data, which creates synthetic examples of the minority class to balance the class distribution.

3.3 Interpretability

Additionally, this research focused on the interpretability of the model's prediction with SHAP values (Shapley Additive exPlanations) (Lundberg and Lee, 2017). SHAP values are used to measure the contribution of each feature to the prediction for a given data point. The SHAP values can help understand which features are most important in making predictions and can be used to identify potential biases of the model or to perform feature selection for improvement.

4 Result

4.1 Interpret Features

As Figure 4 shows, the result of this research indicates that within the follow-up window, "Age" and "Davies Score" were the most contributing features to the patients' mortality. The higher the patient's age the more likely the patient's mortality within the follow-up window. The Davies comorbidity score was used to assess the severity of comorbidity conditions (Davies et al., 2002). The higher the Davies comorbidity score, the higher the patients' mortality.

In addition to the features related to the patients, the years of experience and the size of the medical centers also showed major impact on patient's mortality. Centers with a smaller number of patients may provide better care for each patient, resulting in lower patient mortality rates compared to centers with a larger number of patients. Centers with more years of experience may be better equipped to handle the medical needs of patients, resulting in lower patient mortality rates compared to centers with less experience.

Though the model used in this research makes suggestions correspond to common truth in a medical setting, its performance was also measured by a series of metrics, such as: mean accuracy, mean f1-score, etc. The mean accuracy for this binary classification model is around 0.827, which is relatively acceptable. On the contrary, the mean f1-score was only around 0.2, which indicates the model's poor performance on classifying positive and negatives examples in the dataset.

Similar to the classification model for patients' mortality, the models for the mortality caused by CVD and non-treatment-related Sepsis were also measured by mean accuracy and mean f1-score across all folds in the stratified k-fold procedure. The mean accuracy and mean f1-score for the mortality caused by CVD are 0.6 and 0.34 respectively. And those for mortality caused by non-treatment-related Sepsis are 0.62 and 0.33 respectively.

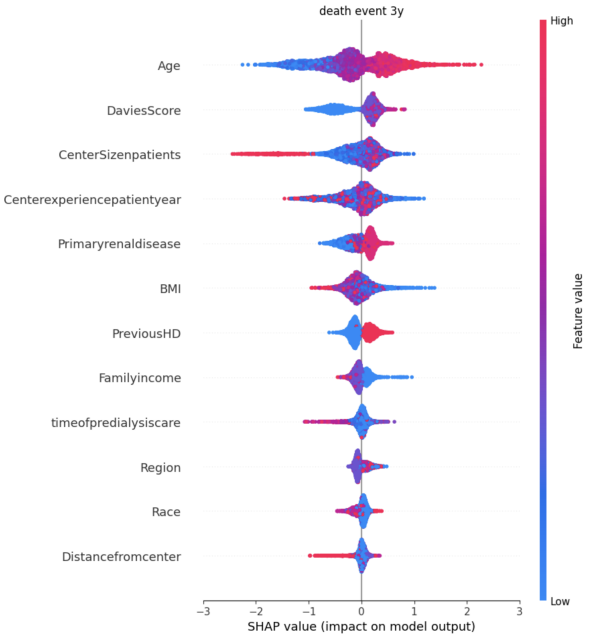

Figure 4: Interpretability for patients' mortality model

According to Figure 5 and Figure 6, the mortality caused by CVD and non-treatment-related Sepsis were mainly affected by "Age" and "BMI". Though "Age" didn't show a clear separation between the value and its relation with the aforementioned two causes, a higher "BMI" would contribute more toward the morality caused by CVD. The model also showed that "Gender" was a major feature that affect mortality caused by CVD, with female more prone to have CVD related issues compared to male.

The model for mortality caused by non-treatment-related sepsis showed a clear relation between patients' BMI and the target variable. A lower BMI would contribute more towards the mortality caused by non-treatment-related sepsis. The years of experience and the size of the medical center were found to have a similar effect on the target variable, similar to the model for patients' mortality as shown in Figure 6.

4.2 Validate Model's Performance

The performance of our classification models are not promising with the BRAZPD dataset (Fernandes et al., 2008), as indicated by the evaluation metrics used in this study. The accuracy ranged from 60% to 90%, while the f1-score ranged from 0.5% to 30%. Neither of these metrics is an indica-

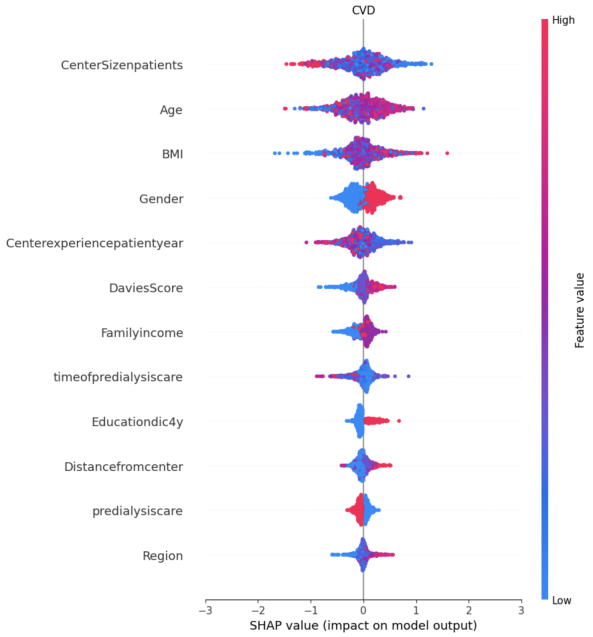

Figure 5: Interpretability for mortality caused by CVD

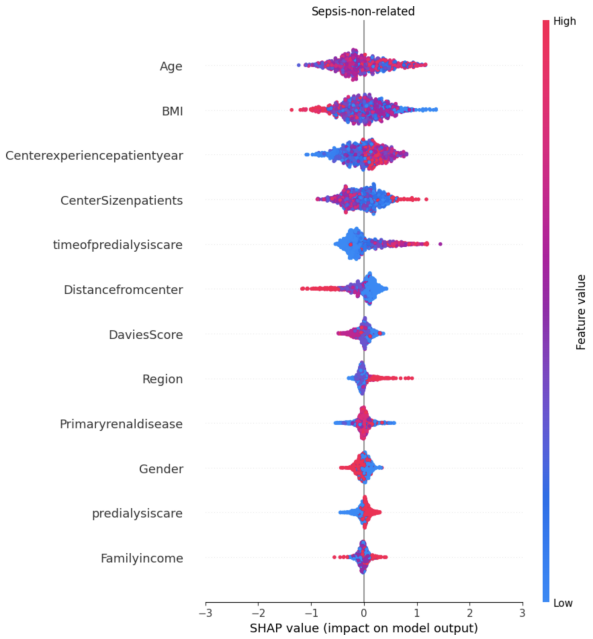

Figure 6: interpretability for mortality caused by non-related Sepsis

tion of good performance.

To further break down the problem of poor performance, the Iris dataset (Fisher, 1936) was also applied to the classification model, and results were generated. With the clean and organized data provided by the Iris dataset, the model achieved above 90% for both accuracy and f1-score. As shown in Figure 7, our model shows that a smaller petal length and petal width indicate a clear distinction between Setosa and the other two classes: Versicolour and Virginica. This aligns with the fact that one of the classes in the Iris data (Fisher, 1936), Setosa, is linearly separable from the other two classes. Versicolour and Virginica, on the other hand, do not have a clear distinction on petal length. Rather, as Figures 8 and 9 indicate, the trend indicated by our model shows that the smaller the petal width, the more it aligns with Virginica, and the higher the petal width, the more it aligns with Versicolour. Therefore, it is safe to say that the model generated in this research achieved promising performance with less noisy data.

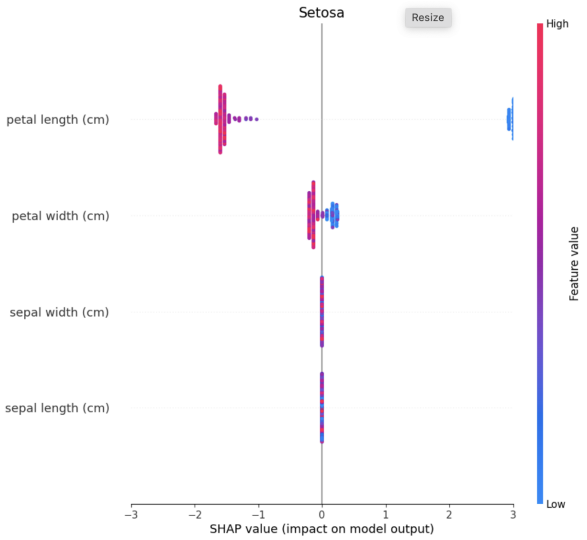

Figure 7: Binary classification model for Setosa

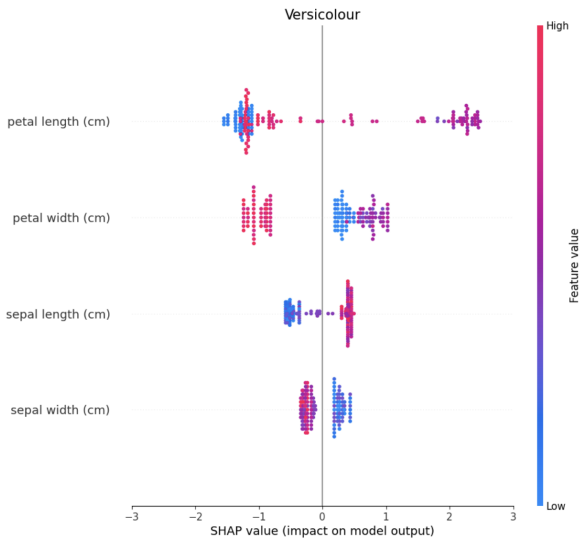

Figure 8: Binary classification model for Versicolour

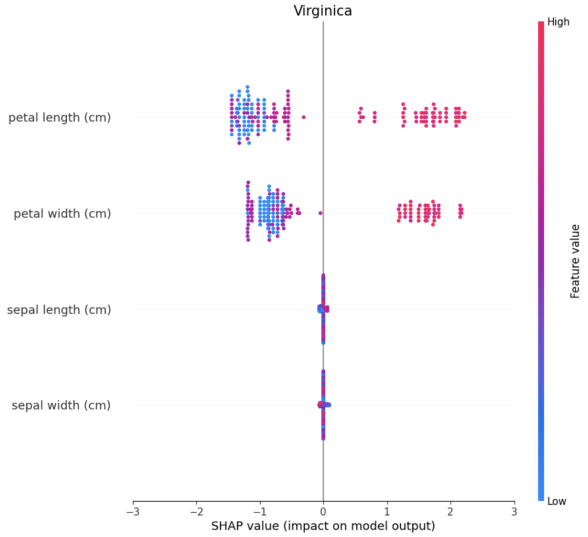

Figure 9: Binary classification model for Virginica

References

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Simon J Davies, Louise Phillips, Patrick F Naish, and Gavin I Russell. 2002. Quantifying comorbidity in peritoneal dialysis patients and its relationship to other predictors of survival. *Nephrol. Dial. Transplant*, 17(6):1085–1092.

N Fernandes, M G Bastos, H V Cassi, N L Machado, J A Ribeiro, G Martins, O Mourão, K Bastos, S R Ferreira Filho, V M Lemos, M Abdo, M T I Vannuchi, A Mocelin, S L Bettoni, R V Valenzuela, M M Lima, S W Pinto, M C Riella, A R Qureshi, J C Divino Filho, R Pecoits-Filho, and Brazilian Peritoneal Dialysis Multicenter Study. 2008. The brazilian peritoneal dialysis multicenter study (BRAZPD) : characterization of the cohort. *Kidney Int. Suppl.*, 73(108):S145–51.

R. A. Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.