Interpretability in machine learning predictions: case of Random Forest regression using Partial Dependence Plots

Danesh Tina, Rachid Ouaret, Pascal Floquet

Interpretability in machine learning predictions: case of Random Forest regression using Partial Dependence Plots

Tina Danesh[1], Rachid Ouaret[1], Pascal Floquet[1]

Affiliation 1 : Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France

Introduction

Deep Neural Networks, Random Forests, and Support Vector Machines are some of the Machine Learning (ML) methods. Due to their accuracy, they have been increasingly used in the prediction domain in various engineering applications. They could be applied when dealing with large databases that do not meet the assumptions of traditional statistical techniques. However, it may be difficult for experts to understand the results of these models and the reason behind them. Also, it is hard to explain them to one plan to act based on these predictions, and most of the time, they remain as a black box.

In this work, the procedure of gaining related information from a model about the relationships between inputs and model outputs is defined as the interpretability in prediction problems. Enhancing interpretability comprises distinguishing the effect of each input uncertainty on the model output variance. Therefore, we aim to enhance ML methods' interpretability to better understand a chosen problem (Murdoch et al., 2019). For this aim, there are two choices, model-specific methods and model-agnostic methods. There are some tools to interpret the results in model-specific methods. These tools are specific to the model and cannot be applied to other ML models (for example, in linear regression, coefficient significance is considered a useful tool to interpret the model). The model-agnostic method could overcome this drawback and be used on any machine learning prediction model (Molnar, 2019).

We applied Random Forest (RF) models as the ML methods to predict a Combined Cycle Power Plant (CCPP) power output as an illustrative example. The data set of the CCPP used in this study is taken from Tüfekci's paper (2014). After that, the Partial Dependence Plot (PDP) and Individual Conditional Expectation (ICE) are used as the model-agnostic method. The PDP features the change in the average predicted value as the specified input(s) change over their marginal distribution. The plots are considered ICE for each data sample (Goldstein et al., 2015). A PDP averages the individual lines of an ICE plot. They are low-dimensional graphical explanations of the prediction function to easily understand the relationship between the output and predictors.

The contribution of this work lies precisely in the interpretability methods of random forest predictions. The aim is to describe the relationships between different input and output variables from model-agnostic points of view.

Results

The parameters that affect the CCPP are the ambient conditions such as ambient temperature (AT), atmospheric pressure (AP), relative humidity (RH), and the exhaust steam pressure (or vacuum, V) effect on the steam turbine. These parameters are the input variables of the system, and the electrical power from both gas and steam turbines is the output variable.

Figure 1 shows the PDPs and ICE simultaneously of a random forest for the power output predictions for each input parameter and reveals the main effect of input variables. The RH and AP main effects

behave quasilinearly and have a low effect on PE. In comparison, AT and V main effects have inverse sigmoidal behavior, and changing each has more effect on PE than AP and RH. For example, we can divide the AT plot into three parts. It is partly linear in the first and third parts. In the middle, we have a complex variation. Increasing the AT makes PE decrease. The AP plot has a very slow slope, and increasing the AP makes PE increase, but for a minimal range.

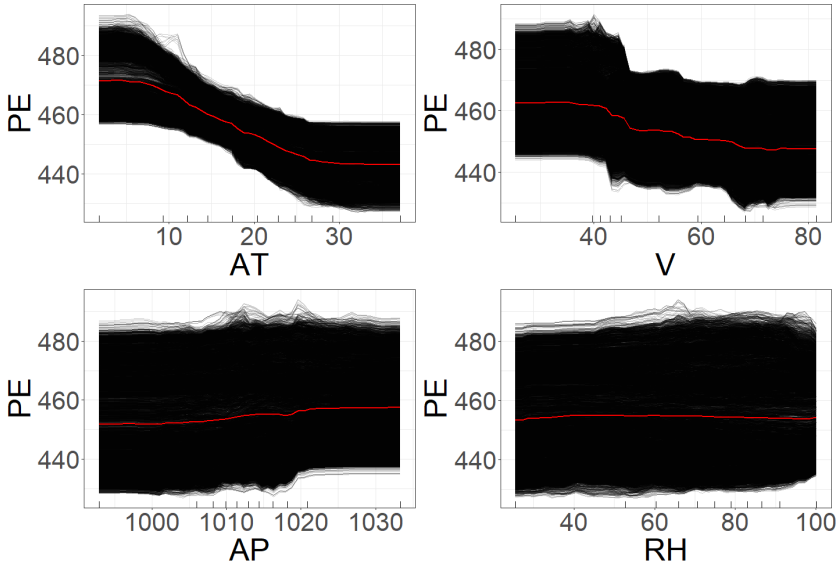

Figure 1 : PDP (red) and ICE plots (black) of a random forest for power output predictions. The PDP is computed after learning the ML model for PE predictions. It reveals the main effect of each input variable.

Conclusion

The main purpose of this study is to enhance the interpretability of the random forest prediction model with the help of a model-agnostic method. To this end, we perform PDP as a model-agnostic method after the RF prediction model. The PD plots make it easier to grasp the effect of each input on the model's output, and based on this information, it is much easier to decide for the decision-maker. For example, in our case, we can conclude AT and V have more influence on PE, and the variety of PE according to AT and V is in a wider range compared to RH and AP based on PDP.

A further research objective could include applying and comparing different model-agnostic methods, such as accumulated local effects, global surrogate models, and permutation feature importance. Moreover, the interpretability of other machine learning prediction methods such as support vector machines, recurrent neural nets, and long short-term memory could be examined.

Références bibliographiques

Molnar, Christoph. *Interpretable machine learning*., 2020.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. "Definitions, methods, and applications in interpretable machine learning." *Proceedings of the National Academy of Sciences* 116, no. 44 (2019): 22071-22080.

Tüfekci, Pınar. "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods." *International Journal of Electrical Power & Energy Systems* 60 (2014): 126-140.