Partial dependence plots for inspecting machine learning models of sugarcane yield

Rodrigo Teixeira Polez*, Felipe Ferreira Bocca, Luiz Henrique Antunes Rodrigues

Abstract
Sugarcane yield models are important tools for planning purposes in the sucroenergetic sector. When black-box techniques are used to create such models, methodologies such as partial dependence plots are required for further understanding them. We evaluated partial dependence plots for a few selected important variables. We observed that different techniques learned similar responses. The patterns were consistent across different techniques, feature sets, and the use of feature selection. They also reflected knowledge about the crop.

Key words:
Visual analysis; Data mining; Sugarcane

Introduction
Models are important tools for estimating sugarcane yield, a vital information for the sucroenergetic sector. Machine learning techniques applied to mill data can provide models with good accuracy, but as they are not comprehensible, they are often called black-box models. Partial dependence plots (PDP)¹ may then be used to interpret the patterns learned by the model. Our objective is to create PDP plots to analyze the effect of different variables in the yield according to the models developed.

Results and Discussion
Sugarcane yield models were fitted² in three different datasets (Raw, Ext and Full), using two techniques (Random Forest, RF, and Boosted Regression Trees, BRT), with and without Feature Selection (FS). Five features were chosen based on the FS results². The pattern learned for the effect of number of harvests reflects the knowledge about the crop: yield decreases with successive harvests (Figure 1). The different techniques learned similar patterns in the different datasets, with no difference when FS was applied.

Figure 1. Effect of number of harvests on Yield.

Models were consistent when evaluating the response to the number of harvests. Also consistent, even though not as much, was the response to precipitation during sugarcane tillering. Maximum response is achieved with almost 400 mm of precipitation. The decrease can be related to nutrient leaching and/or the reduction of solar radiation as consequence of rains. The responses found for precipitations over 700 mm are idiosyncratic, for few records in the dataset presented such values. Similar results were observed in the analysis carried for soil pH, Nitrogen application and temperature during sprouting.

The number of harvests was the most important feature during FS². This can be seen in the Yield range when comparing Figure 1 and Figure 2. A small range was also observed in the other features analyzed. As interactions play a significant role in plant growth, we understand that aggregating responses in an average curve is not enough to observe the range of responses of individual features. Other methodologies, where the curves are not aggregated, such as Individual Conditional Expectation³ should address this limitation.

Conclusions
Sugarcane yield models were explored with PDP, increasing the comprehension of the patterns learned by the models. The patterns were consistent across different techniques, feature sets and the use of feature selection. Patterns uncovered by PDPs could be further evaluated by specific research.

Acknowledgement
This project was partially supported by Bioen/Fapesp and Odebrecht Agro-industrial (Process # 12/50049-3).


DOI: 10.19146/pibic-2016-50805