Cole Harris, Chevron

"Potential pitfalls in exploration and production applications of machine learning" (SPE 169523)

Abstract

Exploration and production applications of machine learning algorithms are varied and numerous. However less attention has been given to the underlying assumptions critical to the application of such techniques. As the breadth of applications increases, it is critical to understand those characteristics of the problem and data that may impact the results.

Independent of the particulars of any specific algorithm, the standard model development and evaluation process may be described as follows. From a set of data points consisting of features and a response, a machine learning algorithm produces a model that may then be used to compute predicted responses on new data. This model can be applied to additional data for which the response is available, and performance estimated by comparing the predicted and actual response. However the reliability of this estimate may be strongly dependent on the statistical characteristics of the data. If the observations are not independent, then the results may not reflect performance in application.

To explore the impact of the violation of the assumption of independence on predictive model development and evaluation, standard machine learning algorithms were used to develop models from synthetic time series data and real monthly oil production data from the Wolfcamp play in the Midland basin. For both, standard approaches to model evaluation fail. For the oil production data, an alternative approach to model development and evaluation is shown to produce both more reliable estimates of model performance and improved model performance.