UNCONVENTIONAL RESOURCES TECHNOLOGY CONFERENCE
FUELED BY SPE • AAPG • SEG

URTeC: 2670157

Application of Data Analytics for Production Optimization in Unconventional Reservoirs: A Critical Review

Srikanta Mishra* and Luan Lin, Battelle Memorial Institute.

Abstract

This paper provides an overview of data analytics concepts and applications in the context of production data analysis for unconventional reservoirs. Topics discussed under key concepts in data analytics include: predictive modeling methods, handling missing variables, model evaluation and validation, automatic tuning of model parameters and variable importance. Key features of a number of representative features are summarized, and observations are made regarding the current state of practice with respect to: limited model evaluation, restricted number of alternative models, ignoring data imputation, and skipping variable importance. Finally, some comments are presented about how the past may not be prologue for predictive model applications.

Introduction and Scope

Big data analytics and data-driven modeling have become quite the buzzwords in recent years - especially in the context of analyzing the production behavior of oil and gas reservoirs [1]. Their growing application has been predicated on the potential to usher in exciting new developments related to: (1) acquiring and managing data in large volumes, of different varieties, and at high velocities, and (2) using statistical techniques to "mine" the data and discover hidden patterns of association and relationships in large, complex, multivariate datasets [2]. The ultimate goal is to develop data-driven insights for understanding and optimizing the performance of unconventional reservoirs, where robust mechanistic models of flow (from nano pores through a network of natural and induced fractures into a multi-stage hydraulically fractured horizontal well) are still under development. However, despite its success in fields ranging from consumer marketing to cyber security to health care [3], the subject of data analytics remains a mystery to most petroleum engineers and geoscientists because of the statistics-heavy jargon and the use of complex "black-box" algorithms.

The process of extracting important patterns and trends from the data to understand "what the data says", and building a predictive model or "learner" using a training data set to predict the value of an outcome based on a number of inputs, is generally referred to as "supervised learning" [4]. Such problems can be further subdivided into: (a) regression problems, where the response variable is continuous, or (b) classification problems, where the response variable is categorical. In both cases, the predictor variables can be continuous and/or categorical. For example, building a predictive model for the cumulative annual production in the first 12 months is a regression problem, whereas determining the factors responsible for separating the top 25% of the wells ("good") from the bottom 25% ("bad") in terms of cumulative production is a classification problem.

This paper will provide a critical review of the state of the art regarding applications of data analytics (*aka* machine learning or data mining) for production optimization in unconventional reservoirs by focusing on: (a) ***premises***, i.e., easy-to-understand descriptions of the commonly-used concepts and techniques, (b) ***promises***, i.e., case studies demonstrating successful practical applications, and (c) ***perils***, i.e., honest appraisal of challenges and potential pitfalls. Key concepts such as machine learning algorithms, handling of missing variables, predictive model building using automatically tuned parameters, cross-validation, and variable importance will be first introduced. Next, examples from the literature will be used to showcase how data analytics has been applied to study unconventional well performance - with an emphasis on the applicability of various techniques, and their limitations.

Key Concepts in Data Analytics

Predictive Modeling Methods

Some of the common predictive modeling techniques for regression and classification problems in the context of production optimization problems (i.e., analysis of production, well completion, and/or formation evaluation data) can be identified as follows:

- Classification and Regression Trees (CART) – Binary decision trees where the predictor space is split into nested rectangular regions, each with a constant value or categorical label for the response [5]
- Random Forest Regression (RF) - Ensemble of simple regression trees, each of which is trained using a random subset of observations and predictors [6]
- Gradient Boosting Machine (GBM) - Ensemble of regression trees which are trained sequentially, with each new tree designed to address shortcomings in predictions made by earlier trees [7]
- Support Vector Machine (SVM) - Transforms the data into another space in which a linear regression or linear classification-style approach can be used to model them [8]
- Artificial Neural Networks (ANN) - Computing system made up of simple, highly interconnected processing elements, which process input information to predict output in an iterative manner [9]
- Kriging Model or Gaussian Emulation (KM) - Models the response as a trend term with an autocorrelation structure, where neighboring observations have similar responses [10]
- Self-Organizing Map (SOM) – Classifies data in such a way so the multi-dimensional input dataset is converted to a lower-dimensional grid space without losing geometrical relationship among data points [11]

Handling Missing Variables

A common occurrence in data-driven modeling is the situation where some of the observations have missing values for one or more predictors. Therefore, it is necessary to impute (i.e., fill-in) these missing values in some manner under the typical assumption that the missing data mechanism has not distorted the observed data [4]. The simplest strategy is to impute the missing value with the mean or median of the non-missing values for that input. A more involved approach is based on the assumption that the inputs have some moderate degree of dependence. Subsequently, a predictive model can be built for each input given the other inputs, and the missing values imputed based on their predictions from the model.

Another imputation strategy, used specifically for the RF algorithm, uses the concept of proximity [6]. The proximity statistic is a measure of similarity of different data points to one another in the form of a symmetric matrix with 1 on the diagonal and values between 0 and 1 off the diagonal. The imputed value is the average of the non-missing observations weighted by the corresponding proximities. The GBM algorithm handles the missing data issue using surrogate splits during the tree building process [7]. On the other hand, the SVM and ANN algorithms do not have any built-in model-specific data imputation methodology.

Model Evaluation and Validation

Three common metrics for evaluating the goodness-of-fit are: (1) average absolute error or AAE, (2) mean squared error or MSE, and (3) pseudo-R^2 [12]. The AAE is defined as the average magnitude of the difference between the true response and predicted response (i.e., the average size of the residuals. MSE is similar to AAE, but measures the average squared difference between observations and their corresponding predictions, rather than the absolute value. Note that AAE has units matching those of the response, while MSE is measured in squared units of the response. A common variant of MSE is the root mean square error or RMSE, which is simply the square root of MSE. The third metric, pseudo-R^2 compares the sum of squared differences between the true responses y_i and predicted responses to the overall sum of squares, which is proportional to the variance of the responses. That is, it measures how much of the variability in the response is explained by the model.

For model validation, a common approach is to use an independent test set in the form of completely new data, or a "held out" portion of the training dataset. In both cases, one can fit the model using the training portion of the dataset (typically 70-90% of the data), and then evaluate the fit on the independent test observations (i.e., remaining

10-30% of the data) to gauge the predictive ability of the model for new data. A more robust alternative is *k*-fold cross-validation [4]. Here, the training dataset is randomly split into *k* different groups or "folds". Next, each of the *k* groups is held out one at a time, the model is trained on the remaining *k*-1 groups, and used to make predictions on the group that was held out. After cycling through all *k* groups, there will be a single cross-validated prediction for every observation in the dataset, where the predictions were made using a model for which that observation was not included in the training set.

Automatic Tuning of Model Parameters

Selecting the values of the tuning parameters in various "black box" algorithms often becomes a manual time sink, with the added potential for significant subjective bias. Examples of such tuning parameters include: (a) number of variables randomly sampled as candidates at each split and number of trees for the RF algorithm, (b) number of trees for the GBM algorithm, (c) cost parameter for the SVM algorithm, and (d) number of hidden layers and hidden units for the ANN algorithm.

To his end, an automated process that relies on cross-validation has been suggested in the literature [13]. The basic steps are as follows: (a) define a set of candidate values for the tuning parameters(s), (b) for each candidate set, resample data, fit model and predict hold outs for a k-fold cross-validation strategy, (c) aggregate the resampling into a performance profile, (d) determine the final tuning parameters using cross-validated RMSE as the metric of choice, and (e) using the final tuning parameters, refit the model with the entire training set.

Variable Importance

For complex data-driven models, where there is no transparent functional relationship between the input and output variables, it is difficult to infer key input-output dependencies based on a simple evaluation of model results. In general, the identification of variable importance tends to be model specific, reflecting the unique nature of the model building process in each algorithm. For example, the relative importance computed for an RF model measures the prediction strength of each variable by calculating the increase of RMSE when that variable is permuted while all others are left unchanged [6]. On the other hand, relative importance for GBM models is based on the number of times a predictor variable was selected for splitting, weighted by the squared improvement to the model as a result of each split, averaged over all trees, and rescaled with a total sum of 100 [7].

One straightforward approach for variable importance that is not tied to any particular model is based on the concept of R^2-loss [14]. This method works for any regression model, and the reasoning is that if an influential predictor is removed from a model, the accuracy of that model will be significantly reduced. Alternatively, if a superfluous predictor is removed from the model, there should be little to no impact on the accuracy. To measure variable importance, one can compute pseudo-R^2 as defined using all the predictors, and then for a reduced model that uses all of the predictors *except* the predictor of interest. The "R^2-loss" metric is simply the difference between the pseudo-R^2 for the full model and the reduced model. The larger the loss in pseudo-R^2 for any given predictor, the greater its influence on the model response.

Example Applications

Table 1 summarizes a representative set of publications dealing with both formation evaluation and production optimization in unconventional reservoirs. These studies [15-27] are fairly recent, most within the past 5 years. They have primarily been applied to the shale basins in the USA, with a few applications in overseas assets. The methods that have been applied cover a broad range of methodologies. However, most of the studies appear to use just one or two preferred techniques for all of their predictive modeling needs, as opposed to comparing the performance of a larger suite of alternative approaches.

A second point to note that the methods represented in **Table 1** cover the standard repertoire of statistical learning techniques that are common in the word of data science [4]. There is, however, an alternative paradigm that has also been used for modeling the behavior of shale formations. This approach combines fuzzy pattern recognition with neural network type learning methods to develop predictive models with good accuracy [28 and references therein].

Table 1. Summary of representative data analytics applications for unconventional reservoirs.

Author	Type	Application	Methods
[15] Zhong et al.	Regression	Modeling of cumulative production in the Wolfcamp dataset	OLS, SVM, RF, GBM
[16] Schuetter et al.	Regression and Classification	Modeling of cumulative production and extreme well production performance in the Wolfcamp dataset	Regression – OLS, RF, GBM, SVM, KM; Classification – CART
[17] Bhatta-charya et al.	Classification	Mudstone lithofacies classification for Bakken and Marcellus wells	SVM, ANN, SOM and Multi-Resolution Graph-based Clustering (MRGC).
[18] Yu et al.	Regression	TOC estimation in Zhangjiatan shale, south-eastern Ordos Basin (China)	Gaussian Process Regression (same as KM)
[19] Lolon et al.	Regression	Predictive statistical model for evaluating the impact of various fracture treatment and well completion designs on production	Ordinary regression (OLS), GBM, RF
[20] Alim-khanov et al.	Regression	Selection of candidate wells for hydraulic fracturing and prediction of post-frac well performance Povkh Field	ANN, CART
[21] Zhao et al.	Classification	Classification of different lithofacies and petrotypes in the Barnett shale play	proximal SVM
[22] Zhao et al.	Regression	Estimation of TOC from triple combo logs in a Barnett shale play	Proximal SVM
[23] Gupta et al.	Regression	Production forecasting in unconventional resources	ANN, SOM
[24] Bhatta-charya et al.	Classification and Regression	Causal analysis and data mining of well stimulation data	CART
[25] LaFollette et al.	Regression	Predictive modeling of oil production in the Bakken light oil play	OLS, GBM
[26] Shelley et al.	Regression	Understanding multifractured horizontal Marcellus completions	ANN
[27] Nejad et al.	Regression	Predictive modeling of Eagle Ford production based on well completions	ANN

Observations

- ***Limited model evaluation –*** In general, most of the studies tend to base their model evaluations using the full training data, or at best, one hold-out test dataset. Multi-fold cross-validation does not appear to be a routine element of the predictive modeling workflow, even though the robustness of cross-validated goodness-of-fit measures as a proxy for a model's predictive accuracy for new data is well established [13].

- ***Restricted number of alternative models*** – Most analysts appear to prefer one or two model building techniques. The default choices appear to be ANN and SVM. Our experience suggests that multiple models can provide very similar goodness-of-fit measures on training or test data. However, their performance with respect to future predictions or identification of variable importance can be quite different [15]. It is therefore advisable to use multiple modeling strategies to build the predictive model, especially as the incremental workload for fitting additional models is quite trivial.

- ***Ignoring data imputation*** – The literature is somewhat reticent about the application of imputation for filling missing values for unconventional oil and gas applications. It appears that most studies just eliminate the data records that have missing values for one or more predictors. In reality, several modeling techniques such as RF and GBF have built-in methods for data imputation which can be applied for approximating the missing values, even when the final predictive model is something different such as an ANN.

- ***Skipping variable importance*** – Another aspect of model building that seems to garner limited attention is the subject of variable importance. The inner workings of a black-box type model are not easy to understand as compared to a traditional closed-form model. Hence, the need for a formal variable importance analysis to extract insights regarding the relative importance of different predictors on the forecasts for the response variable. In particular, the model-independent R^2-loss approach could be a simple but useful practical tool in this regard [16].

- ***Past may not be prologue*** – It is quite likely for unconventional reservoirs that the space-time continuum corresponding to a given data set (and the resulting data-driven model) may not exactly match the conditions for the next application. For example, the shale facies may be slightly different, or the flow regime may be more transient than boundary dominated. Under such conditions, it is important to keep in mind that data-driven models have a limited ability to project the unseen.

Acknowledgments

We thank our colleagues Jared Schuetter and Shuvajit Bhattacharya for several helpful discussions. This study was supported by a Battelle internal R&D project.

References

1. L. Saputelli, 2016, "Technology Focus: Petroleum Data Analytics", *Society of Petroleum Engineers*, doi:10.2118/1016-0066-JPT
2. K. Holdaway, 2014, *Harnessing Oil and Gas Big Data with Analytics*, Wiley.
3. A. Bahga and V. Madisetti, 2016, *Big Data Science and Analytics: A Hands-On Approach*, VPT.
4. Hastie, T., R. Tibshirani, and J.H. Friedman, 2008. *The elements of statistical learning: data mining, inference, and prediction*, Springer.
5. Breiman L., Friedman J.H., Olshen R.A. and Stone CJ, 1984. *Classification and Regression trees.* Monterey, CA: Wadsworth and Brooks/Cole; 1984.
6. Breiman, L. 2001. Random forests, *Machine Learning*, 45(1): 5-32
7. Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine, *Annals of Statistics* (2001): 1189-1232.
8. Vapnik, V., 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

9. Rumelhart, D. E. and McClelland, J. L., 1986. *Parallel Distributed Processing, 1: Foundations*, MIT Press, Cambridge.

10. N. Cressie, 1993, *Statistics for Spatial Data*, New York: Wiley.

11. Kordon, A.K., 2010. *Applying Computational Intelligence: How to Create Value*, Springer-Verlag, Berlin.

12. Navidi, W., 2008, Statistics for Engineers and Scientists, McGraw Hill, New York.

13. Kuhn, M. and K. Johnson, 2013, *Applied Predictive Modeling*, 2013. Springer.

14. Mishra, S., N.E. Deeds and G.J. Ruskauff, 2009. Global sensitivity analysis techniques for groundwater models, *Ground Water*, 47(5): 730-747.

15. Zhong, M., J. Schuetter, S. Mishra and R. Lafollete, 2015. Do data mining methods matter? A "Wolfcamp" shale case study. *SPE Hydraulic Fracturing Technology Conference*.

16. Schuetter, J., S. Mishra, M. Zhong and R. LaFollette, 2015. Data Analytics for Production Optimization in Unconventional Reservoirs. *SPE/AAPG/SEG Unconventional Resources Technology Conference*.

17. Bhattacharya, S., T.R. Carr and M. Pal, 2016, Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA, *Journal of Natural Gas Science and Engineering*, 33, http://dx.doi.org/10.1016/j.jngse.2016.04.055

18. H. Yu, Z. Wang, R. Rezaee, Y. Zhang, L. Xiao, X. Wang and L. Zhang, 2016, The Gaussian Process Regression for TOC Estimation Using Wireline Logs in Shale Gas Reservoirs, *International Petroleum Technology Conference*, 10.2523/IPTC-18636-MS, IPTC-18636-MS.

19. E. Lolon, K. Hamidieh, L. Weijers, M. Mayerhofer, H. Melcher and O. Oduba, 2016, Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models: A Middle Bakken and Three Forks Case History, *Society of Petroleum Engineers*, 10.2118/179171-MS, SPE-179171-MS.

20. R. Alimkhanov and I. Samoylova, 2014, Application of Data Mining Tools for Analysis and Prediction of Hydraulic Fracturing Efficiency for the BV8 Reservoir of the Povkh Oil Field, *Society of Petroleum Engineers*, 10.2118/171332-MS, SPE-171332-MS.

21. T. Zhao, V. Jayaram, K. J. Marfurt and H. Zhou, 2014, Lithofacies Classification in Barnett Shale Using Proximal Support Vector Machines, *Society of Exploration Geophysicists*, SEG-2014-1210.

22. T. Zhao, S. Verma, D. Devegowda and V. Jayaram, 2015, TOC Estimation in the Barnett Shale From Triple Combo Logs Using Support Vector Machine, *Society of Exploration Geophysicists*, SEG-2015-5922788.

23. Gupta, S., Fuehrer, F., & Jeyachandra, B. C., 2014, Production Forecasting in Unconventional Resources using Data Mining and Time Series Analysis. *Society of Petroleum Engineers*. doi:10.2118/171588-MS.

24. Bhattacharya, S., Maucec, M., Yarus, J. M., Fulton, D. D., Orth, J. M., & Singh, A. P., 2013, Causal Analysis and Data Mining of Well Stimulation Data Using Classification and Regression Tree with Enhancements. *Society of Petroleum Engineers*. doi:10.2118/166472-MS

25. Izadi, G., Zhong, M., & LaFollette, R. F., 2013, Application of Multivariate Analysis and Geographic Information Systems Pattern-Recognition Analysis to Production Results in the Bakken Light Tight Oil Play. *Society of Petroleum Engineers*. doi:10.2118/163852-MS

26. Shelley, R., Nejad, A., Guliyev, N., Raleigh, M., & Matz, D., 2014, Understanding Multi-Fractured Horizontal Marcellus Completions. *Society of Petroleum Engineers*. doi:10.2118/171003-MS

27. Nejad, A. M., Sheludko, S., Shelley, R. F., Hodgson, T., & Mcfall, P. R., 2015, February. A Case History: Evaluating Well Completions in Eagle Ford Shale Using a Data-Driven Approach. *Society of Petroleum Engineers*. doi:10.2118/173336-MS

28. S. Mohaghegh, 2017, *Shale Analytics: Data-driven Analytics in Unconventional Reservoirs*, Springer.