APPLICATION OF MACHINE LEARNING ALGORITHMS IN PREDICTING PYROLYTIC ANALYSIS RESULT

Le Thi Nhut Suong$^{1,2}$, A V Bondarev$^1$, E V Kozlova$^2$

$^1$Department of Petroleum Prospecting and Exploration, Gubkin Russian State University of Oil and Gas (National Research University), 65 Leninski Avenue, Moscow, Russia, 119991;
$^2$Skoltech Center for Hydrocarbon Recovery, Skolkovo Institute of Science and Technology (Skoltech), Renova Lab 11 Sikorskogo Street, Moscow, Russia, 143005;

Email: lethinhutsuong@gmail.com

Abstract. Geochemical studies of organic matter in source rocks play an important role in predicting the oil and gas accumulation of any territory, especially in oil and gas shale. For deep understanding, pyrolytic analyses are often carried out on samples before and after extraction of hydrocarbon with chloroform. However, extraction is a laborious and time-consuming process and the workload of laboratory equipment and time doubles. In this work, machine learning regression algorithms is applied for forecasting $S_2\text{ex}$ based on the pyrolytic analytic result of non-extracted samples. This study is carried out using more than 300 samples from 3 different wells in Bazhenov formation, Western Siberia. For developing a prediction model, 5 different machine learning regression algorithms including Multiple Linear Regression, Polynomial Regression, Support vector regression, Decision tree and Random forest have been tested and compared. The performance of these algorithms is examined by R-squared coefficient. The data of the X2 well was used for building a model. Simultaneously, this data is divided into 2 parts – 80% for training and 20% for checking. The model also was used for prediction of wells X1 and X3. Then, these predictive results were compared with the real results, which had been obtained from standard experiments. Despite limited amount of data, the result exceeded all expectations. The result of prediction also showcases that the relationship between before and after extraction parameters are complex and non-linear. The proof is R2 value of Multiple Linear Regression and Polynomial Regression is negative, which means the model is broken. However, Random forest and Decision tree give us a good performance. With the same algorithms, we can apply for prediction all geochemical parameters by depth or utilize them for well-logging data.

1. Introduction
Geochemical research of organic matter in source rocks, especially express analysis, Rock-Eval pyrolysis play an important role in determining potential hydrocarbon content of any territory. This is relevant to relatively new regions with high potential of oil and gas, as well as exploiting regions, where additional surveys for non-traditional sources of HC can take place. In sediments, approximately one-third of the matrix is occupied with organic matter, so a separation of generating matter from generated hydrocarbons is required. Separation of kerogen and bitumoid has been proven successful when comparing the results of pyrolytic studies of source rock powder and powder distilled with chloroform. After extraction, we have true kerogen yield ($S_2$), $T_{\text{max}}$ and hydrogen index, which can be used as entry
parameters for basin modeling. However, extraction is a laborious and time consuming process. Moreover, a substantial amount of solvent is required and lab equipment operating loads is doubled. Analysis of kerogen parameters after extraction without performing the extraction itself will reduce costs and time required for analysis. The described problem is typical for “machine learning”. In this work, machine learning regression algorithms is applied for forecasting S2ex based on the pyrolytic analytic result of non-extracted samples. Thank to machine learning algorithms, now we can receive the analysis result for one sample in 40 minutes and just several days for one well, which usually takes 3 up to 4 months.

2. Material and method
The studies were carried out using more than 300 samples from three wells of one field located in the central part of the West Siberian petroleum province within the Frolovskaya basin of Bazhenov formation. For analyzing core samples by geochemical methods, the analysis process goes through several stages and it is indicated by the diagram in Figure 1.

![Figure 1. Scheme of geochemical analysis of samples](image)

After the pyrolytic analysis for non-extracted sample we receive a result table with a set of parameters including Sample ID, Depth, S0, S1, S2, S3, TOC, Tmax, HI, OI, PI, CC, CaCO3, AI, OSI, GOC, NGOC, S3', S3CO, S3'CO, S4CO2, S5 and S4CO. For extracted sample we receive the same parameters, but they were added suffix “ex” for designation (Sample ID, Depth, S0ex, S1ex, S2ex, S3ex, TOCex, Tmaxex, H1ex, OIex, PIex, CCex, CaCO3ex, Axex, OSIex, GOCex, NGOCex, S3'ex, S3'COex, S3'CO2ex, S5ex and S4COex). All the parameters of extracted sample can be forecasted, but in this work only the S2ex has been predicted by machine learning algorithms because it’s the most important parameter, since it’s a real value of kerogen yield and a lot of analysis has been provide just for determining this parameter. In this study, 5 algorithms including: Multiple linear regression, polynomial regression, support vector regression, decision tree and random forest were applied for predicting
2.1. Multiple Linear Regression

Multiple linear regression is a statistical technique that uses multiple explanatory variables to predict the resulting variable. The purpose of this regression is to model a linear relationship between the independent variables and the response (dependent) variable. Formally, the model for multiple linear regression, given $n$ observations, is

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad \text{for } i = 1, 2, \ldots, n \]

where

- $y_i$ – dependent variable
- $x_i$ – independent variable
- $\beta_i$ – bias

2.2. Polynomial Regression

Polynomial regression is a form of linear regression in which the relationship between the independent variable $x$ and the dependent variable $y$ is modelled as an $n$th degree polynomial. As well as multiple linear regression, polynomial regression determines the dependence of the resulting variable on independent variables. The model for polynomial regression, given $n$ observations, is

\[ y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 x_{i2} + \ldots + \beta_n x_{in}^n \quad \text{for } i = 1, 2, \ldots, n \]

where

- $y_i$ – dependent variable
- $x_i$ – independent variable
- $\beta_i$ – bias

2.3. Support vector regression

Support vector regression is a supervised learning type of data analysis developed by Vapnik and colleagues at AT&T Bell Laboratories [8]. The main idea of the method is to translate the initial vectors into a space of higher dimension and search for a separating hyperplane with the largest gap in this space. Two parallel hyperplanes are constructed on both sides of the hyperplane separating the classes. The dividing hyperplane will be the hyperplane that creates the greatest distance to two parallel hyperplanes. The algorithm is based on the assumption that the greater the difference or distance between these parallel hyperplanes, the smaller the average error of the classifier will be. The algorithm is shown in figure 2.

\[
\frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \Rightarrow \min
\]

\[ y_i = (w, x_i) + b + \xi_i \]

\[ y_i = (w, x_i) + b - \xi_i^* \]

- $\xi_i$, $\xi_i^*$ – Slack variables
- $\varepsilon$ – Insensitive tube

*Figure 2. Support vector regression algorithm*
2.4. Decision tree
A decision tree is a supervised machine learning model used to predict a goal by examining decision rules from their characteristics. As the name suggests, we can think of this model as breaking down our data by making a decision based on a series of questions. The decision tree algorithm regression model is illustrated in figure 3.

![Figure 3. Decision tree regression model](image)

2.5. Random forest regression
Random forest regression is a technique that combines predictions from multiple machine learning algorithms together to make more accurate predictions than any single model. The random forest algorithm does an excellent job, it is one of the best algorithms in terms of accuracy, but it does not explain anything at all. It is such a black box that gives good predictions, but does not at all tell about what the addiction is based on, without additional scrupulous analysis. In figure 4 has explained how random forest algorithm works.

![Figure 4. Random forest regression model](image)
2.6. R-squared

All the models were verified by R-squared. This is a coefficient of determination, which ranges from 0 to 1, but it also can be negative. If the coefficient of determination is negative, it means that the selected function \( y = f(i) \) characterizes data worse than \( y = y_{av} \) (horizontal line). Therefore we can’t use this algorithm for predicting.

\[
R^2 = 1 - \frac{RSS}{TSS}
\]

where

- \( R^2 \) - coefficient of determination
- \( RSS \) - sum of squares of residuals
- \( TSS \) - total sum of squares

The sum of squares of residuals, also called the residual sum of squares:

\[
RSS = \sum_i (y_i - f_i)^2
\]

The total sum of squares (proportional to the variance of the data):

\[
TSS = \sum_i (y_i - \bar{y})^2
\]

![Figure 5. The coefficient of determination R²](image)

The data of the X2 well was used for building regression models. Simultaneously, this data is divided into 2 parts – 80% for training and 20% for checking. Those models were applied for prediction of wells X1 and X3. Then, these predictive results were compared with the real results, which had been obtained from standard experiments.

### 3. Result

A summary table of the evaluation of model performance by R-squared is shown in Table 1.

| Method                      | Model Performance |
|-----------------------------|-------------------|
|                             | Model            | X1 well | X3 well  |
| Random forest regression    | 0.95             | 0.88    | 0.96     |
| Support vector regression   | 0.89             | 0.50    | 0.87     |
| Decision tree regression    | 0.90             | 0.77    | 0.91     |
| Polynomial regression       | -0.84            | -15.54  | 0.03     |
| Multiple linear regression  | 0.96             | -0.76   | 0.97     |
To predict the $S_{2ex}$ parameter, first of all, the simplest regression was applied - multiple linear regression. Despite this is the simplest method, it works well with our dataset. A comparison of the model calculation results and the actually measured $S_{2ex}$ is shown in Figure 6. The downside of this method is that it characterizes the data as a function, so it can give us a negative value for the $S_{2ex}$ parameter, which cannot be. With a new dataset that is different from the training dataset, the model can give us a prediction that is very different from the actual values. For example, at well X1, which is far from the original well, so it contains some values that are very different from the actual values (75 to 115 mg difference). Because of these points, the determination estimate is greatly reduced (R2 of well X1 is -0.76). Thanks to this, we can also see how rigorously R2 is evaluating our models. To improve the model, we need more data from different regions. Nevertheless, this method is worthy of predicting the $S_{2ex}$ value of the wells having similar characteristics and facies conditions.

**Figure 6.** Comparison of the results of the model calculation and actually measured $S_{2ex}$ by the method of multiple linear regression.
The next algorithm that was applied is polynomial regression. This method gives us the worst prediction result even in the training model ($R^2 = -0.84$), and obviously the prediction result for wells X1, X3 could not be better. $R^2$ is -15.54 and 0.03, respectively. Different degrees of the polynomial “$n$” were applied to the model, but the larger the degree, the worse the forecast result we got. The result of predicting polynomial regression ($n = 4$) is shown in Figure 7. We can conclude that the relationship between all pyrolytic parameters before extraction and $S_{2ex}$ cannot be modeled as an $n$-th degree polynomial.

Support vector regression is also used to predict the $S_{2ex}$ value. This method does not give the best result, but its result is acceptable. The result is shown in Figure 8. The predicted result of the training well and well X3 is quite good, $R^2$ is 0.89 and 0.87, respectively, but the predicted result of well X1 is not good enough ($R^2 = 0.5$). This algorithm does not give us a negative $S_{2ex}$, which indicates its applicability and satisfactory results. This method characterizes our dataset using a hyperplane or a function, and the relationship between the parameters before and after extraction is complex enough that it cannot be characterized by a function.

Moving on to new data mining and predictive analytics tools - decision tree and random forest algorithms. The decision tree is organized in the form of a hierarchical structure consisting of decision-making nodes for evaluating the values of certain variables to predict the resulting value. The use of classification trees results in a symbolic designation of the class. The geological section consists of different groups of rocks with similar characteristics, i.e. close $S_{2ex}$, so a decision tree algorithm was created to analyze our data. As expected, this method gives very good results (Figure 9). The $R^2$ value of the well test and wells X1, X3 is 0.90, 0.77 and 0.91, respectively.

The decision tree algorithm gives us good predictions, but it's not the best way. The method gives us the best result - random forest (Figure 10). As mentioned above, this method combines predictions from...
multiple machine learning algorithms together to make more accurate predictions than any single model. The R2 values of the well test and wells X1, X3 are 0.95, 0.88 and 0.96, respectively.

**Figure 9.** Comparison of the model calculation results and actually measured S2ex by the decision tree method

The predicted S2ex of all models was also compared with the S2 value, because the S2ex must be lower than S2. The result is shown in Figure 11. As it has been pointed out, the multiple linear regression, support vector regression and decision tree methods have predicted S2ex values greater than S2. Only the random forest method gives us good predictions and less than S2. Based on all the critical examinations, we can conclude that the random forest algorithm is the best regression algorithm for the data.

**Figure 10.** Comparison of the model calculation results and actually measured S2ex by the random forest method

**Figure 11.** Comparison of S2ex by regression model (A. Multiple linear regression, B. Support vector regression, C. Decision tree regression, D. Random forest regression) and S2 of well X3
4. Conclusion
Despite the amount of data for training the model is limited, the result exceeded all expectations. An increase in accuracy should be expected as the input data block for training the model grows. The best results were shown by support vector, decision tree and random forest regressions. Random forest algorithms and decision trees are recommended for predicting the analysis result of any wells, as well as multiple linear regression for all wells close to the training well. The prediction result also demonstrates that the relationship between the parameters before and after extraction is complex and non-linear. The proof is the R2 value of multiple linear regression, and the polynomial regression is negative, which means the model does not work. However, the random forest and decision tree give us good performance. Using the same algorithms, we can apply for predicting all geochemical parameters in different depth within study interval or use them for well-logging data.

References
[1] Arlot S, Genuer R (2014). "Analysis of purely random forests bias”
[2] Ben-Haim Y (2001) Information-gap theory: decisions under severe uncertainty. Academic Press, London.
[3] Biau G. Analysis of a random forests model. Journal of Machine Learning Research. 2012;13:1063–1095.
[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001.
[5] Cortes C. and Vapnik V., "Support-Vector Networks," Machine Learning, vol. 20, pp. 273-297, 1995/09/01 1995.
[6] David M. Lane, Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Rice University.
[7] Draper, N. R.; Smith, H. (1998). Applied Regression Analysis. Wiley-Interscience. ISBN 978-0-471-17082-2.
[8] Glantz, Stanton A.; Slinker, B. K. (1990). Primer of Applied Regression and Analysis of Variance. McGraw-Hill. ISBN 978-0-07-023407-9.
[9] Kamiński, B.; Jakubczyk, M.; Szufel, P. (2017). "A framework for sensitivity analysis of decision trees". Central European Journal of Operations Research. 135–159. doi:10.1007/s10100-017-0479-6. PMC 5767274. PMID 29375266.
[10] Karimi K. and Hamilton H.J. (2011), "Generation and Interpretation of Temporal Decision Rules", International Journal of Computer Information Systems and Industrial Management Applications, Volume 3.
[11] Mathieu Rouaud, 2013: Probability, Statistics and Estimation Propagation of Uncertainties in Experimental Measurement
[12] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.
[13] Liaw and M. Wiener. Classification and regression by random forest. R news, 2:18–22, 2002.
[14] Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies. 221–234p. CiteSeerX 10.1.1.18.4267. doi:10.1016/S0020-7373(87)80053-6.
[15] Scornet, Erwan (2015). "Random forests and kernel methods”
[16] Smola A, Schoelkopf B, A tutorial on support vector regression: Tech. Rep. NeuroCOLT2 NC2-TR-1998-030: 1998.
[17] Tissot B P, Welte D H, Petroleum formation and occurrence 2nd edn. (Springer, Berlin, 1984).
[18] Yan, Xin (2009), Linear Regression Analysis: Theory and Computing, World Scientific, pp. 1–2, ISBN 9789812834119
[19] Yin-Wen Chang; Cho-Jui Hsieh; Kai-Wei Chang; Michael Ringgaard; Chih-Jen Lin (2010). "Training and testing low-degree polynomial data mappings via linear SVM". Journal of Machine Learning Research. 1471–1490p.
[20] Zhu R, Zeng D, Kosorok MR (2015). "Reinforcement Learning Trees". Journal of the American
Statistical Association. 1770–1784p. doi:10.1080/01621459.2015.1036994