Well-Log Based Determination of Rock Porosity Through Machine Learning Algorithms (Machine Learning 2021 Course)

Ramilya Sharifullina [1] **Ilya Nachinkin** [1] **Elizaveta Gladchenko** [1]

Abstract

Hydrocarbon (oil and gas) production often involves enormously high costs, including the cost of monitoring many rock's characteristics. One of the ways to reduce it is to implement machine learning to estimate one or several rock properties based on another. In our work, we consider one of the most important characteristics of hydrocarbon-bearing rock - porosity that is usually determined during money- and time-consuming laboratory tests. Based on the well log interpretation, we build and investigate several models for porosity prediction. Our research shows that boosting model not only outperforms all other methods, but also have a great potential to adapt to different oilfield.

Github repo: Well-log ML project
Video presentation: Well-log ML project video

1. Introduction

Porosity is a reservoir characteristic that describes how much oil a particular field contains and, therefore, how economically efficient it is to develop it. Obviously, it is one of the most critical parameters in the Oil and Gas industry. Porosity is used to calculate the amount of oil and gas reserves for reservoir description and reservoir simulation models (Li, 2019).

Rock porosity can be estimated by the interpretation of well-logs. Logging is a time- and money-consuming procedure. Equipment must be lowered and raised into the well, and for this, it is necessary to suspend the hydrocarbon production process. Thus, all process costs a lot of money. Therefore, expenses during field development and logging may be decreased by avoiding at least one logging operation and predicting the change in porosity along the reservoir by three remaining log curves.

Recently, machine learning has been used actively to work with geophysical logs (Belozerov et al., 2018; Antipova et al., 2019; Meshalkin et al., 2020). Modern works show that it is possible to predict various signs of reservoir parameters quite well by using machine learning and deep learning techniques. However, for porosity, there has not been a sufficiently comprehensive study.

In our work, we will fix this and explore both classical machine learning methods and neural network learning methods for predicting porosity.

The main contributions of this report are:

- performance of several standard machine learning models (such as Linear Regression, kNN, Gradient boosting, Random Forest) for the task of well-log interpretation was observed (section 3.2);
- the results of the performance of these methods were compared based on such metrics as RMSE, MAE, R2 (section 3.3);
- peculiar properties of the task such as problems with real data, feature importance (whether they relate to geology or not), adaptation of the model to the data from different wells were investigated (section 4.2);
- deep learning algorithms (neural models such as FNN, RNN) for obtaining porosity values from well-logging data were implemented (section 3.2);
- comparison of the results obtained by neural networks and standard machine learning algorithms was performed (section 4.4).

The rest of this paper is organized as follows. Section 2 presents an overview of old, recent and state-of-the art approaches for well-log data interpretation using machine learning algorithms. Section 3 introduces research object

[1]Skolkovo Institute of Science and Technology, Moscow, Russia. Correspondence to: Ramilya Sharifullina <ramilya.sharifullina@skoltech.ru>.

Final Projects of the Machine Learning 2020 Course, Skoltech, Moscow, Russian Federation, 2020.

and also presents methods and metrics used while solving the stated problem. In Section 4, we introduce all of the performed procedures and obtained experimental results. In Section 5, we discuss achieved objectives and make a conclusion about implementation of machine learning algorithms for the rock porosity determination.

2. Related Work

Traditional and most widely used approaches for estimation of rock porosity involve active use of special equipment, such as porosimeters, computer tomography machines, analytical balances, saturators, etc.However, the field of machine learning also continues to develop proving itself as a fast and accurate method that can compete with traditional ones.

2.1. Estimating porosity in industry

Porosity (ϕ) is a unitless rock property, which is defined as the ratio of the volume of pores (V_p) to the bulk (matrix) volume (V_b) of the sample (Terry et al., 2013):

$$\phi = \frac{V_p}{V_b}.$$

Rock porosity can be estimated by several methods such as:

- Computerized Tomography (CT) scan;
- laboratory test on core samples;
- interpretation of well-logs.

The first two methods are quite expensive and time-consuming because they involve taking samples of the rock and delivering them to the laboratory, where the analysis (CT scan or laboratory test) is performed. The main problem of these methods is difficulty in upscaling obtained results. The reservoir, which stores hydrocarbons (oil and gas), is composed of a large number of various rocks. Thus, even after a thorough examination of several samples, we cannot be absolutely sure that the obtained value of porosity is correct for the entire reservoir.

During field development, in order to determine the intervals of a reservoir containing hydrocarbons, petroleum engineers use the interpretation of various logs. They can also be used to determine the porosity of the rock. Logging refers to the process of recording and analyzing measurements collected discretely or continuously in a wellbore (Aminzadeh & Dasgupta, 2013). Downhole measurements are made by lowering a recording device into the well at the end of an electrical cable.

The group of porosity logs includes:

- density log;
- neutron log;
- acoustic (sonic) log;
- nuclear magnetic resonance (NMR) log.

None of these logs measures porosity directly. Porosity is derived from measuring other parameters related to porosity. The density and neutron are nuclear measurements. The sonic log uses acoustic measurements, and the NMR log records the magnetic resonance of formation nuclei. When used individually, each of the first three logs has a response to a lithological type which it is accounted for, but when used simultaneously, two or three at a time, lithology of an entire reservoir can be estimated, and a more accurate value of porosity derived.

2.2. Well-log interpretation by machine learning methods

The problem of predicting depth profiles from well-logging data and, thus, obtaining values of various rock properties with machine learning algorithms has been addressed since the 1980s. During that time application of machine learning and a so-called 'artificial intelligence' approach was gaining popularity in many fields of study. The Oil and Gas industry was not an exception.

In pioneering works (Kerzner, 1983) first attempts were made towards the determination of formation dip based on the well-log microresistivity measurements. In this work, special attention was given to choosing data for training algorithms and estimation of curves similarities. Thus, two methods were introduced – 'fixed interval' and 'point-to-point'. The first method used intervals of finite length on two curves and compared these intervals using the cross-correlation coefficient as a measure of curve similarity. Two algorithms for point-to-point correlation have been suggested. One algorithm used a pattern classification scheme, while the other algorithm was based on frequency considerations. Before that, manual (optical) correlation, or locating similar events on the curves, was considered to be the most reliable method for the determination of formation characteristics.

Later on, existing solutions were improved by adding extra statistical techniques. In (Lim et al., 1998) it became obvious that applying dimensionality reduction techniques, especially Principal Component Analysis (PCA), is a necessary step to perform due to the fact that, conceptually, all available well-log data should be used for identifying similar characteristics from log traces despite their different scaling in depth or amplitude.

However, this can be a very demanding process – interpreters often use only a selectively reduced data set, such

as resistivity logs. In many situations, data reduction can result in an erroneous or misleading correlation. PCA can establish an effective quantity of data-set information for easier handling by the interpreter.

An interesting approach that included time-series analysis was proposed in (Georgie, 1991). It is a well-known procedure used routinely in many areas of engineering. Although most commonly applied to data sampled with respect to time (e.g., wind records and seismic and electrical data), time-series analysis can be applied equally well to data sampled with respect to depth. The vertical resolution and repeatability of well-log measurements were quantified with time-series analysis. The technique allowed us to choose a filter that maximized the coherent signal and minimized the incoherent noise in data.

A work that was published much later (Belozerov et al., 2018) also used time-series analysis implemented in the recurrent neural network (RNN). RNN represented well-log signals using stochastic Markov processes. It appeared that the model is sensitive to data integrity. Also training dataset was given weights inversely proportional to the class presence due to the fact that the training set was highly unbalanced.

Thus, nowadays we can find many examples of application both supervised and unsupervised machine learning algorithms applied in order to solve the task of either interpreting well-log data or predicting rock properties based on these curves.

Both (Singh et al., 2020) and (Meshalkin et al., 2020) state that although today most studies apply unsupervised learning algorithms for solving mentioned problems, they underperform by a substantial margin compared with nearly all the supervised learning algorithms.

(Singh et al., 2020) showed that unsupervised learning is more stable in terms of its performance irrespective of the type of combination of the data, but the performance of the supervised learning is strongly dependent on the combination of the data that can mimic the output. Among the supervised learning algorithms, the fully connected three-layer neural network performed consistently well, and its prediction performance was better when the training well contained a greater number of data. In (Meshalkin et al., 2020) the Random Forest provided the most accurate predictions of rock properties from well-logging data.

3. Algorithms and Models

Github repo: Well-log ML project

3.1. Data description

The research object is the X oil field. This field is characterized by high energy parameters of the formation, that is why it was called by specialists as one of the most perspective fields.

Since we are considering one field, this suggests that geological parameters will have the same intervals for each well that penetrates this formation. However, this does not mean that we will not encounter any anomalies (for example, thrust fault, which is a break in the Earth crust, across which older rocks are pushed above younger rocks).

It is also worth noting that not all of the above logs are performed for every well.The survey can be carried out not on the entire well but only on a certain depth. It is up to petroleum engineers to decide whether or not conducting a particular survey.

Experimental data from the four wells (named as 169, 170, 171, 183 wells) was involved in our research. We constructed our model in one well and then estimated how well these models is fitted for other wells. The location of the wells relative to each other is shown in Figure 1.

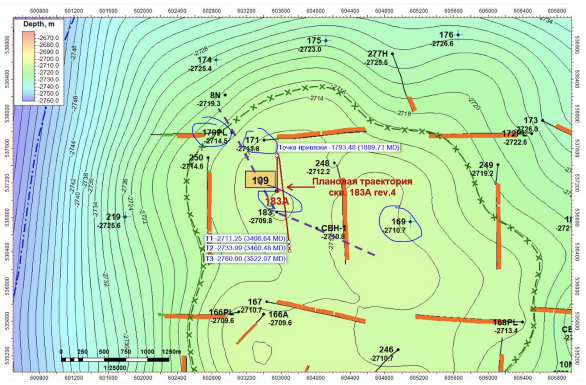

Figure 1. Location of each well in comparison to each other.

Link to the dataset used in the project:
http://bit.ly/3s1QRc2

Each file contains information about each separately taken well. The name of the file corresponds to the number of the well. The data for each well was obtained from the Institute of Geology and Development Combustible Minerals (PJSC "Rosneft"). The data was prepared in software "Techlog" (by Schlumberger Company). Since the wells are horizontal, it was necessary to take into account the inclinometry of the wells. This means that the data for each well was corrected according to the reference to absolute depths (from the sea level).In the data, it can be seen that the measurements were taken not from the absolute zero mark, but from a certain depth (about 2000 meters).

The fact is that there is no need to carry out measurements along the entire length of the well, since the reservoir rock which contains oil and gas under standard conditions lies at a depth of 2000 meters (this value varies depending on the condition of the location of the field, for example, the X field is located in Western Siberia, so this is a standard depth for this region). Then petroleum engineers, based on the initial geological parameters, which are known even before drilling the first well, decide at what point in time or at what depth to launch the device to measure certain geological parameters.

The features of each well are the following: depth; measurements by gamma ray log; measurements by resistivity log; measurements by caliper log; density log measurements; sonic (acoustic) log measurements; neutron log measurements. The target variable is porosity.

Gamma-ray logging is based on the measurement of natural radioactivity of rocks along a borehole. Resistivity logging is a method of well logging that works by characterizing the rock or sediment in a borehole by measuring its electrical resistivity. A caliper is a well logging tool that provides a continuous measurement of the size and shape of a borehole along its depth.

Drilling engineers and geologists use caliper measurements as a qualitative indication of the wellbore condition and measuring the thickness of the mud cake that is needed to maintain the borehole stability. Mud cake is a direct indicator of whether there is the flow of hydrocarbons (oil and gas) from the rock to a well. So, if its thickness is zero, then this means that there is no way to extract oil in this area, and porosity will be interpreted as zero.

As it was mentioned earlier, the following three parameters of logging: density, sonic and neutron logging are principal features for the interpretation of the porosity of the rock. So, density logging tools rely on gamma-gamma scattering or on photoelectric (PE) absorption. Based on the measured density values, we can calculate the formation porosity.

The Sonic (or acoustic) logging allows the determination of the velocity of (ultra)sound in the rocks adjacent to a wellbore.

Neutron logging involves the input of neutrons into the strata and the recording of either the reduction in energy of these particles (neutron—neutron logging) or the gamma-rays emitted from the strata as a result of neutron capture (neutron—gamma logging).

So, we can already assume that when we calculate the importance of model features, we will find out that these three logs will be the most significant when calculating porosity.

3.2. Machine learning algorithms

For the task of predicting porosity from the interpretation of the well-logging data, standard machine learning algorithms were used. Interpretation of the well-logging data was used as input, while experimental data on rock porosity was used as an output. The following set of algorithms was tested for the indirect determination of rock porosity: Linear Regression, Random Forest Regression, XGBoost Regression, k-Nearest Neighbours, Neural Network (Feedforward neural network).

3.2.1. Linear regression

Linear Regression fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ in order to minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation.

We decided to increase the complexity of applied models gradually, so our first choice was a simple Linear Regression model from `sklearn` library. The simplicity of the model allowed us to obtain results rather quickly. However, it was obvious from the very beginning that well-logging data can not be effectively described with such a model due to existing non-linear relations between rock and its properties.

3.2.2. K-Nearest Neighbors

k-Nearest neighbours (k-NN) is another standard machine learning method, which can be used for solving the regression problem. During k-NN method application, the target point is assigned by the average value of the nearest points in the training data based on the similarity metric.

For solving the task of the project `sklearn` implementation `KNeighborsRegressor` with following parameters was used: `weights`: 'distance', `n_neighbors`: 3.

The main benefit of this method is the same as for Linear Regression - simplicity. Unfortunately, k-NN has a very serious drawback that makes it not a good choice for this task - it is extremely sensitive to noisy data, missing values and outliers.

3.2.3. Random Forest

Decision tree-based learning is one of the popular approaches to build predictive models describing non-linear behaviour between input and output data.

A Random Forest is a meta-estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

For solving the task of the project `RandomForestRegressor` from `sklearn` with

following parameters was used: `n_estimators`: 10, `max_features`: 'sqrt', `max_depth`: None, `bootstrap`: False.

On the one hand, Random Forest has a lot of benefits, such as being robust and not prone to overfitting. On the other hand, during solving this task, we face the problem that the number of samples is very small for the large range of values of porosity. Thus, Random Forest might be poorly trained.

3.2.4. XGBOOST

If we deal with regression models, in theory, we can try and use the Gradient Boost model, but it is better to use the XGBoost algorithm since it has some advantages comparing to plain Gradient Boost (XGBoost has built-in L1 (Lasso Regression) and L2 (Ridge Regression) regularization that prevents the model from overfitting).

A big advantage of XGBoost is that it was designed to build models based on large and complicated datasets similar to the one used in this project. However, this method is rather slow comparing to others.

For solving the task of the project `XGBRegressor` with following parameters was used: `max_depth`: 3, `booster`: 'dart'.

3.2.5. NEURAL NETWORK (FNN)

Nowadays, a wide variety of neural network architectures (such as fully connected, convolutional, recurrent, etc.) are successfully implemented and applied for regression problem solving in petroleum engineering (Belozerov et al., 2018). The choice of architecture largely depends on the data structure and problem statement.

In this project, a feedforward three-layer neural network with three hidden layers was used in order to solve the stated problem. We use `Pytorch` for this. The non-linearity was introduced with rectified linear unit activation functions (ReLU). The adaptive moment estimation (Adam) method was used to solve the optimization problem. The initial learning rate was constant and equal to 0.01. The number of epochs was set to 20.

Neural networks show good performance when they are trained on a sufficient amount of data. Because of the fact that it is quite expensive to perform well-logging operations in a big amount of wells, our dataset is limited. Thus, neural networks were not expected to show high-level performance.

3.2.6. NEURAL NETWORK (RNN)

Recurrent Neural Network is a generalization of a feedforward neural network that has internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network.

Because of the fact that well-log curves can be presented as time-series (in our case, a series of depth), RNN seemed to be applicable. We can use the sequence of input samples of the fixed length = 100 in order to predict the porosity on the next step of depth. They can use their internal memory to process sequences of inputs. However, it appeared to be very prone to overfitting and showed bad generalization ability.

In our project we use 1-layer LSTM model (as upgraded version of classic RNN) from `Pytorch` library with `hidden_size` = 30. After the lstm layer there is a Linear layer with an output dimension of 1 to predict porosity.

3.3. Accuracy assessment

The performance of the applied approaches are evaluated on test data using the root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2) :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}, (\phi_i^t - \phi_i^p)^2}{N}},$$

$$MAE = \frac{\sum_{i=1}^{N} |\phi_i^p - \phi_i^t|^2}{N},$$

$$R^2 = 1 - \frac{\sigma_{y|x}^2}{\sigma_y^2},$$

where ϕ stands for porosity, superscripts *t* and *p* stand for true and predicted values, respectively, *N* stands for number of points, σ_y^2 stands for variance of the random variable y, $\sigma_{y|x}^2$ variance of the dependent variable (variance of the model error).

4. Experiments and Results

In this project, we trained several models on well 170 and then tested these models at the same well and on the others (171, 183, 169). All models are tested 1 time with a fixed random state = 42. After we built the models, we assessed their performance using the metrics discussed earlier. Optimal parameters for Linear Regression, XGBoost, Random Forest, KNN were chosen by GridSearch with time-series cross validation. Cross validation scheme is described in the next section.

4.1. Splitting the data

In this project, we split the data using the function TimeSeriesSplit. It provides train/test indices to split time series data samples that are observed at fixed time intervals (in our case, depth intervals), in train/test sets. In each split, test indices must be bigger than before, and, thus, shuffling during cross-validation procedure is inappropriate. The point is that we cannot randomly split the data in this project due to the fact that points are very close to each other.

4.2. Feature importance

The feature importance of the input parameters was assessed based on XGBoost algorithm (giving the best results, as shown in Table 1). Results of relative feature importance assessment are presented on Figure 2. The obtained results prove that the main factor that has an influence on dependencies between well-logging and rock porosity is density, neutron and sonic logging. High contrast in physical properties of the rock matrix and pore-filling fluids significantly changes the density, neutron and sonic log responses. Perhaps there is some confusion about the significance of sonic p-wave and sonic s-wave logs. The fact is that a sound wave can be described not only in general form but also as parallel and shear waves (p-wave and s-wave, respectively). Individually, these waves do not have such a strong importance on porosity prediction.

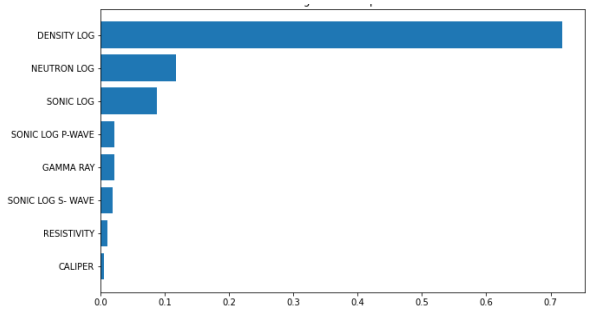

Figure 2. Results of feature importance assessment for well-log based prediction of rock porosity.

4.3. Feature engineering

In order to achieve better performance of our models, feature engineering was applied. At each depth, not only actual well-log values characterize the formation, but also previous values and their relationships among each other bring additional information.

Thus, the new feature was created from the three most important ones (Dense Log, Neutron Log, Gamma Ray), such as the difference between values (1 meter ago) to use in regression models. Other features (window mean, standard deviation, lagged basic feature) were not suitable for our case because those features are applicable to time series models that we will consider further.

As a result, we got the same values for chosen metrics Table 2, so this procedure did not improve the performance of the models.

For RNN separate feature engineering was performed. New features were created (from Dense Log, Neutron Log) and among them were: rolling mean and standard deviation with the window size of 1 meter, and the difference between values on rolling window's borders, lagged basic features (1 meter), standard deviation of original series of depth intervals of 1 meter.

4.4. General results

The performance results of machine learning algorithms application for well-log based prediction of porosity for the well 170, are summarized in Table 1.

Table 1. Results of well-log based determination of rock porosity for well #170 using machine-learning algorithms

ML ALGORITHM	RMSE	MAE	R^2
LINEAR REGRESSION	0.0122	0.0102	0.9024
K-NN	0.0183	0.0128	0.7795
RANDOM FOREST	0.0113	0.0079	0.9152
XGBOOST	0.0094	**0.0061**	0.9413
FNN	0.0207	0.0156	0.7162
RNN	**0.0092**	0.0068	**0.9454**

Table 2. Results of well-log based determination of rock porosity for well #170 using machine-learning algorithms with feature-engineering

ML ALGORITHM	RMSE	MAE	R^2
LINEAR REGRESSION	0.0127	0.0100	0.8933
K-NN	0.0183	0.0128	0.7796
RANDOM FOREST	0.0103	0.0063	0.9303
XGBOOST	**0.0089**	**0.0061**	**0.9477**
FNN	0.0285	0.0255	0.4630
RNN	0.0098	0.0066	0.9372

As it can be seen from Table 2, the XGBoost regression model and RNN shows the best performance for all metrics.

After training on 170 well, these models were tested on other wells (169, 171, 183).

In Figure 3, the actual and predicted values of porosity for all wells obtained with a help of XGBoost regression model are shown. In Table 4, we obtain the metrics of the XGBoost regressor tested on 169, 171, 183 wells.

RNN model will not work if all data is not available (no sonic logs for wells 171 and 183). Moreover, this model

Table 3. Results of implementing RNN algorithm, trained on well 170, on well 169.

WELL	RMSE	MAE	R^2
169	0.1308	0.1117	−9.0226
170	0.0092	0.0068	0.9454

shows poor results if we test it on well 169, which is shown in the Figure 4.

In Table 3, results of implementing RNN model for porosity prediction for wells 169 and 170 can be found.

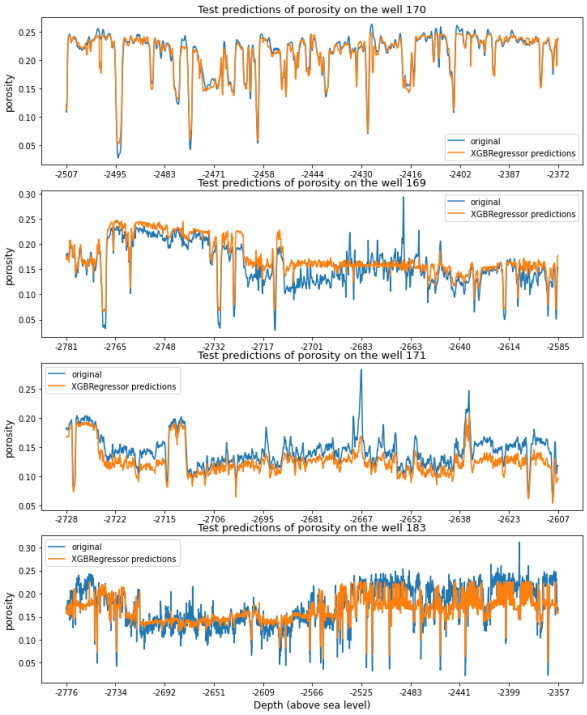

Figure 3. Predicted and actual porosity of the all wells by XGBoost agorithm

All wells are located on the same pad, but well 169 is located farther than others, and it is possible that it is the reason why the models work poorly for this well since the geological conditions of the formation in this well will be significantly different from well 170. Well 171 is the closest to the well 170, and, thus, it is assumed that properties would correlate with properties of the well 170.

Table 4. Results of implementing XGBoost regression algorithm, trained on well 170, on wells 169, 171, and 183.

WELL	RMSE	MAE	R^2
169	0.0211	0.01702	0.7393
171	0.02047	0.0173	0.3923
183	0.0290	0.0223	0.5510

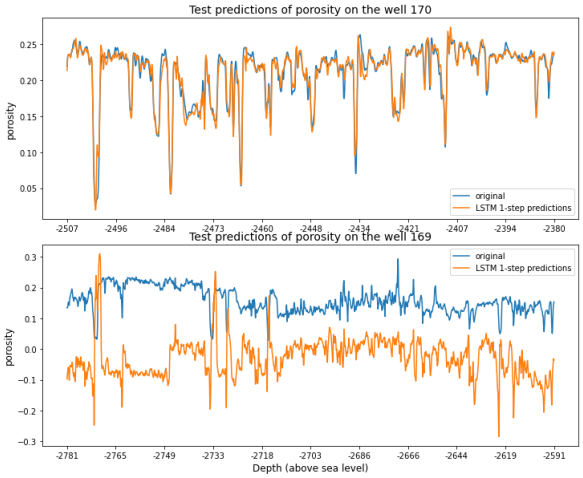

Figure 4. Predicted and actual porosity of the all wells by RNN model

It can be assumed that the model will have a good performance for those wells that are located on the same pad. On the one hand, it is a good way to reduce expenses by avoiding extra logging operations for several wells. However, there is still a chance of heterogeneity in the formation.

In Table 4, results of implementing XGBoost model for porosity prediction for wells 169, 171 and 183 can be found.

So, it can be noticed that the values of the metrics have greatly deteriorated (for example, the standard deviation has become twice as large). But strange enough, coefficient of determination parameter is the highest for well 169 when using this model, most likely due to the fact that for this well we know the values of the sonic logging, sonic logging p-wave and sonic logging s-wave, while the values for the data from wells 171 and 183 are unknown. We can also notice that the results of the metrics are approximately equal for each well, this may indicate that the geological conditions are similar for these wells.

5. Conclusion

We tested widely-used machine-learning algorithms for well-log based determination of rock porosity. We trained five different approaches (Linear Regression, k-Nearest Neighbors, Random Forest, XGBoost, FNN, RNN) and

tested using dataset for the well 170 and we calculated accuracy metrics for the model with the best performance - XGBoost - for wells 169, 170 and 171.

The results of the project allow us to conclude that machine-learning algorithms are a promising framework for fast and effective well-log based predictions of rock porosity. The main advantages of the proposed algorithms are: (1) using machine-learning algorithms is more straightforward comparing to classical approaches, (2) it does not need any extra procedures and equipment, (3) it is more robust compared to laboratory-based approaches, (4) it does not require the upscaling procedure. Implementation of machine-learning techniques for problem solution can radically enhance the results of porosity investigations, basin and petroleum system modelling and optimization of porosity calculation methods.

References

Aminzadeh, F. and Dasgupta, S. N. *Geophysics For Petroleum Engineers*. Elsevier Science, 2013.

Antipova, K., Klyuchnikov, N., Zaytsev, A., Gurina, E., Romanenkova, E., and Koroteev, D. Data-driven model for the drilling accidents prediction. In *Proceedings of the SPE Annual Technical Conference and Exhibition*, Calgary, Alberta, Canada, 2019.

Belozerov, B., Bukhanov, N., Egorov, D., Zakirov, A., Osmonalieva, O., Reshytko, A., and Semenikhin, A. Automatic well log analysis across Priobskoe field using machine learning methods. In *Proceedings of the SPE Russian Petroleum Technology Conference*, Moscow, Russia, 2018.

Georgie, D. Application of time-series analysis to induced gamma ray spectroscopy logs from two cold lake heavy-oil observation wells. In *Proceedings of the 1989 SPE Annual Technical Conference and Exhibition*, pp. 505–512, San Antonio, TX, 1991.

Kerzner, M. G. Formation dip determination - an artificial intelligence approach. *The Log Analyst*, (5):10–22, 1983.

Li, H. Evaluation of oil production potential in fractured porous media. *AIP Physics of Fluids*, (5):2–3, 2019.

Lim, J.-S., Kang, J. M., Seoul, N. U., and Jungwhan, K. Artificial-intelligence approach for well-to-well log correlation. *JPT*, (11):30–32, 1998.

Meshalkin, Y., Shakirov, A., Popov, E., Koroteev, D., and Gurbatova, I. Robust well-log based determination of rock thermal conductivity through machine learning. *Geophysical Journal International*, (05):978–988, 2020.

Singh, H., Seol, Y., and Myshakin, E. M. Automated well-log processing and lithology classification by identifying optimal features through unsupervised and supervised machine-learning algorithms. *SPE Journal*, (10):2778–2800, 2020.

Terry, R. E., Rogers, J. B., and Craft, B. C. *Applied Petroleum Reservoir Engineering*. Prentice Hall, 2nd edition, 2013.

A. Team member's contributions

Explicitly stated contributions of each team member to the final project.

Ramilya Sharifullina (33.3% of work)

- Preparing the Section of Introduction and Research object of this report
- Preparing the Section of Experiment and Results of this report
- Preparing the Section of Conclusion of this report

Ilya Nachinkin (33.3% of work)

- Coding the main algorithm
- Preparing the GitHub Report
- Preparing the Section of Experiment and Results of this report

Elizaveta Gladchenko (33.3% of work)

- Reviewing literature on the topic (8 papers)
- Preparing the Section of Algorithms and Models of this report
- Preparing the Section of Experiment and Results of this report

B. Reproducibility checklist

Answer the questions of following reproducibility checklist. If necessary, you may leave a comment.

1. A ready code was used in this project, e.g. for replication project the code from the corresponding paper was used.

   ☐ Yes.
   ☑ No.
   ☐ Not applicable.

   General comment: If the answer is **yes**, students must explicitly clarify to which extent (e.g. which percentage of your code did you write on your own?) and which code was used.

   Students' comment: None

2. A clear description of the mathematical setting, algorithm, and/or model is included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

3. A link to a downloadable source code, with specification of all dependencies, including external libraries is included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

4. A complete description of the data collection process, including sample size, is included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

5. A link to a downloadable version of the dataset or simulation environment is included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

6. An explanation of any data that were excluded, description of any pre-processing step are included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

7. An explanation of how samples were allocated for training, validation and testing is included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

8. The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results are included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

9. The exact number of evaluation runs is included.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

10. A description of how experiments have been conducted is included.

    ☑ Yes.
    ☐ No.
    ☐ Not applicable.

    Students' comment: None

11. A clear definition of the specific measure or statistics used to report results is included in the report.

    ☑ Yes.
    ☐ No.
    ☐ Not applicable.

    Students' comment: None

12. Clearly defined error bars are included in the report.

    ☐ Yes.
    ☐ No.
    ☑ Not applicable.

    Students' comment: None

13. A description of the computing infrastructure used is included in the report.

- [x] Yes.
- [ ] No.
- [ ] Not applicable.

Students' comment: None