Prediction of the wireline log data using machine learning

Soma Roy, GAIL (India) LIMITED, NOIDA-201301, India*
C.S. Singh, GAIL (India) LIMITED, NOIDA-201301, India
soma.roy@gail.co.in

Keywords

Well log, Machine learning, multi-linear regression, ridge regression, Lasso and neural networks

Summary

In petroleum exploration, the acoustic/sonic log (DT) is primarily used as an estimator to calculate formation porosity, to carry out petro-physical studies, or to participate in geological analysis and research. Sometime these logs do not exist in the old wells, drilled twenty years ago, either because of data loss or because of just being not recorded at that time.

The objective of the present study is to predict the missing sonic log by using supervised machine learning (ML) when only common logs such as natural gamma ray, resistivity log, and bulk density are available or no logs are available. In recent years, ML has become very useful approach and a wide range of industries are applying it to their data set for image classifications, sentiment analysis and text prediction. ML approach is successfully use in geosciences for velocity prediction from travel-time curves and geological predictions (de Wit et al., 2013). A more rigorous detail about machine learning and its application in geoscience is discussed by Bergen et al. (2019). Bergen et al., has clearly mentioned that machine learning could play a very important role across various scale of study.

In this paper we used multi-linear regression algorithm and then proceeded to more complicated algorithms such as ridge regressions and neural networks, to predict the missing sonic (DT) logs. The common logs are used as input to the model and the DT log is considered as the target. By using multi-linear regression algorithm, a prediction model is firstly created based on the experimental data and then confirmed and validated by blind-testing the results in wells containing both the predictors and the target (DT) values, used in the supervised training. Finally the optimal model is set up as a predictor. Subsequently, a case study is carried out for the wells in Cambay Basin, located in western part of India. The logs predicted from the model shows a very good agreement with actual log data recorded.

Introduction

Oil and gas exploration in sedimentary basins is very complicated, since all the targets are buried underground and they cannot be viewed or touched directly. So all the properties for the buried targets have to be predicted or estimated by using modern electrical or magnetic tools. The physical properties of the geologic formations include pore-fluid pressure, rock lithology, porosity, permeability, and oil or water saturation. Nowadays, the conventional tool for characterizing these geophysical properties is well logging, and some logs such as gamma ray (GR), dual induction log, formation density (DEN) compensated, deep resistivity (REID), self-potential (SP), and sonic log (DT) are usually recorded. Among them, the sonic log (DT) has largely been used to predict rock porosity, to perform petro-physical analysis, or to carry out well-to-seismic calculus (Bianco, 2013). Owing to the historical operation mistakes, recording or transcription losses, the sonic log may not always be available in well logging suites. The traditional way solving this problem is to transform the DEN or REID log to DT log based on some experimental formula built between these logs. It might be feasible for some area, but sometimes the errors are unacceptable.

Multiple linear regression also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. It attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.

In this paper, multipliple linear regression algorithm is used to predict missing sonic (DT) logs when only common logs (e.g., natural gamma ray—GR, bulk density—DEN, or deep resistivity—REID) are available. By using MLR, we first create and train a supervised network model based on experimental data and then confirm and validate the model by

soma.roy@gail.co.in

blind-testing the results. The optimal model is at last applied to wells containing the predictor data but with lack of DT log. We use this workflow in wells from Cambay Basin, India and predicted the missing sonic logs.

Methods

In this work, four algorithms are optimally used to predict the sonic data, namely; Multiple Linear regression, Ridge regression, Lasso and Neural network.

Multiple Linear regression method

Multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.
Multiple linear regression for $i = 1..n$ obeservations is expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon,$$

y_i= dependent variable
x_i= Explanatory variable
β_0 = y Intercept (constant term)
β_p = Slope co-efficient for each explanatory variable
ϵ = The model's error term

A simple linear regression is a function that allows to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when there are two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

Ridge regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable.

Ridge regression equation is written in matrix form as follows

$$Y = BX + e$$

where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

Lasso regression

Lasso (least absolute shrinkage and selection operator) regression is a type of **linear regression** that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity.

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. Ridge regression) does not result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

The goal of the algorithm is to minimize

$$\sum_{i=1}^{n} \left(y_i - \sum_j x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| ,$$

Some of the βs are shrunk to exactly zero, resulting in a regression model that's easier to interpret.

Neural network

Artificial neural networks (ANN) or connectionist systems are computing systems that are inspired by, but not identical to, biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules.
A simple representation of fully connected neural network along with layers is shown in Figure 1. The input layer to the neural network model, used in this

soma.roy@gail.co.in

study, is fed with predictor log data and output layer will produce the sonic log as output data.

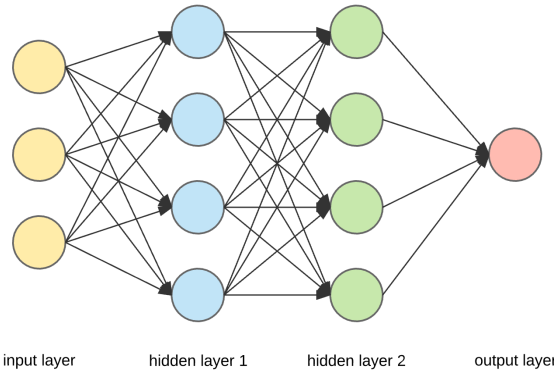

Figure 1: An example of fully connected neural network with input and output layer populated with log

Experimental study

The study uses gamma ray (GR), porosity, formation density, and sonic log (DT) from three wells. Combinations of four different algorithms are used defined by the following workflows:

1. **Problem description and related work**
   First the data has been visualized by cross plotting the DT with all the logs present in the LAS file, as shown in Figure 2. Subsequent to this analysis, for DT prediction in the study area (Cambay Basin), neutron porosity, gamma Ray and density had good correlation with DT, and accordingly used as input features.

2. **Data preparation**

   2.1 The data has been normalized so that all the dataset is in the same range.
   2.2 Then we separate the data into training and validation set. This is to be done to get some metric on how accurate is the model.
   2.3 We shuffle the data to ensure that we get rid of any bias/pattern within the database, before we split the dataset into testing and validation sets.

3. **Training/ test phase**
   3.1 We then split the data into 20% test set and the remaining 80% training set.
   3.2 Next, we will standardize the dataset, which is a way of making the features Gaussian with zero mean and unit variance. We will use the same scalar to standardize the blind dataset that we will test on.
   3.3 Now that we have our normalized training and validation dataset, we will train the data using four algorithms: Linear Regression, Ridge Regression, Lasso and Neural network.
   3.4 We first call the specific regressor using the module and then fit the training data. Then, we can predict the output (DT in our case) on the validation/test set.
   3.5 We can then check the accuracy of prediction using some metrics. We will particularly look at R^2 and Mean Square Error (MSE) statistic.

4. **Model validation and Blind well prediction**

   4.1 Once we have a train model, we will bring in a blind well to do another test on how good the model is. We predict DT in a well that has real DT but was not used for training. We use some metrics to understand the result, mean square error (MSE) or R2 value between the real and predicted curve.
   4.2 Then we have a good model that we are happy with and we predict DT curves in wells that have the input curves (neutron porosity, gamma and density) but are missing the sonic curve.
   4.3 The validation dataset is a way to understand how good the model is doing in the training phase. We want the validation error to be low but also don't want to over fit the data, it is a fine balance and the choice of algorithm is important. Then once we are happy with our model, we would still keep one well with all the curves, including the one we want to predict, aside for another way of checking how accurate the model is. Once we are happy with these two phases of accuracy testing, then we are confident that the model is good and can be applied to other wells where the log to be predicted is missing.

soma.roy@gail.co.in

Results

Testing the model on the validation set is always a good start for any machine learning workflow. However, the real test of robustness is to use the model and predict the blind data. In our case, there is one well that was not used for training, that we will predict the DT on. The blind well contain observed DT log which will be used to test the prediction accuracy. We convert the blind well in a data frame and drop the depth, DT and UWI columns. We will use the same scaler transform that we obtained from the training set.

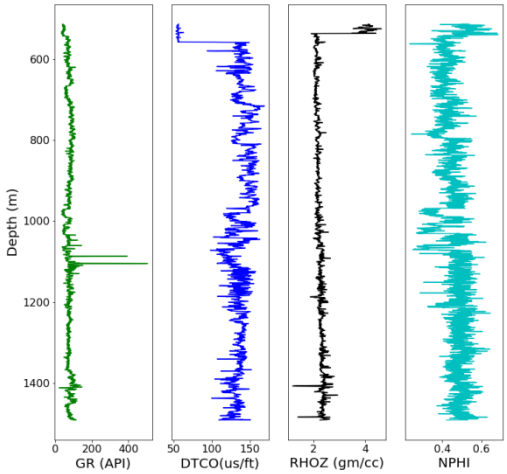

Figure 2: GR, DT, RHOZ and NPHI log plotted with depth

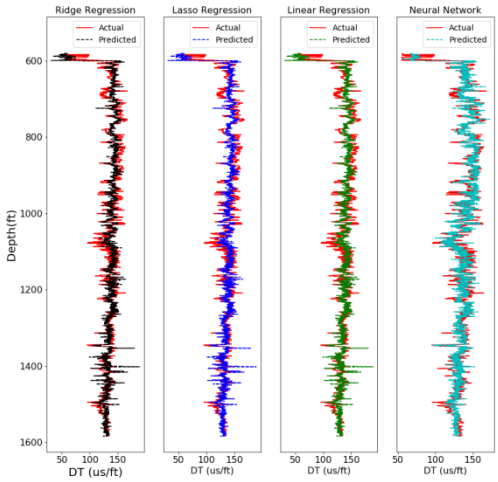

Figure 3: Comparison between predicted DT with actual DT, using all four different algorithms

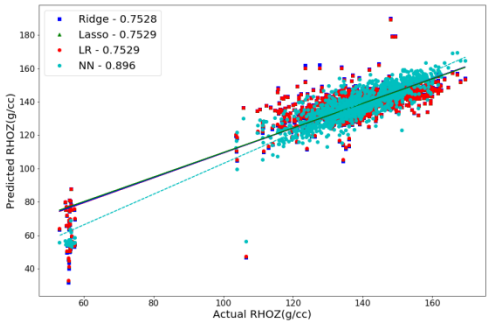

Figure 4: Predicted vs actual RHOZ scatter plot, using all four different algorithms

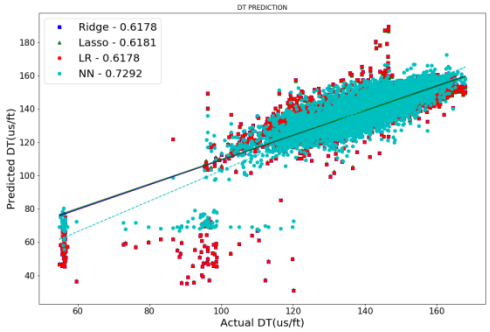

Figure 5: Predicted Vs actual DT scatter plot, using all four different algorithms

Figure 3 shows the plot of actual and predicted sonic log data, which shows a very good agreement. To quantify the misfit, a post prediction multi-variable regression analysis is performed between sets of predicted and actual sonic and density, respectively. The analysis is shown in Figure 4 and 5 for density and sonic, respectively. A detail values of the correlation coefficients and mean square error is shown in Table 1.

Table-1 clearly reflects the fact that the values of R^2 and MSE for the predicted and actual DT are almost same for three regressions namely, linear, ridge and lasso. But for neural network prediction approach, the value of R^2 is higher and value of MSE is lower as compared to the earlier three regressions.

soma.roy@gail.co.in

Table-1: Metric of accuracy: R^2 and MSE

Method	Actual sonic vs Predicted sonic logs		Actual density vs Predicted density logs	
	MSE (%)	R^2	MSE (%)	R^2 coefficients
LR	0.8293	0.6178	0.8221	0.8221
RR	0.8293	0.6178	0.8220	0.8220
LO	0.8286	0.6181	0.8219	0.8219
NN	0.6393	0.7053	0.3220	0.9032
LR: Linear regression, RR: Ridge regression, LO: LASSO, NN: Neural Network				

Conclusions

This study proposes four different regression algorithms and machine learning as a tool for predicting the sonic log in open-hole wells based on other available common logs. Strict steps including data normalization, training set selection, and testing are very important for deciding the prediction power, the generalization capability, and the complexity of the derived regression model. The four different regressors we used were: Linear, Ridge, Lasso and neural network. For training phase, we tested the model on a validation set and saw that all three methods provide very similar results. We then tested the robustness of the model on a blind dataset. Finally we predicted DT using all four different regression algorithms. There are zones within the blind dataset where the prediction is not of very high accuracy. It would be interesting to understand the geological framework and its effect on prediction. We can also try by chopping the data into different lithological zones based on geologic picks before training and predicting. The method presented here is not limited to modeling DT logs only. It can be extended, with appropriate modifications of the algorithm, in any area of well logging studies, where missing log values are needed.

References

1. de Wit, Ralph WL, Andrew P. Valentine, and Jeannot Trampert. "Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks. " Geophysical Journal International 195.1 (2013): 408-422.

2. Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. Science, 363(6433), eaau0323.

3. Bianco, Evan. "Geophysical tutorial: Well-tie calculus." The Leading Edge 33.6 (2014): 674-677.

Acknowledgement

We extend our sincere thanks to GAIL (India) Limited for giving permission to publishing this paper. We would also extend our sincere thanks to Sundeep Sharma, Geophysicist, Devon Energy, Oklahoma City, USA for various technical discussions and his help in creating various modules and frame for machine learning model. We also appreciate Dr. Dharmendra Singh, Sr. Manager (Petrophysics) for his valuable discussion, input and proof reading of the paper.