Adding uncertainty to the estimation of log data using a gradient-boosted tree approach

Marianne Rauch, PhD, TGS, Keyla Gonzalez, TGS

Summary

Machine learning (ML) applications have infiltrated geosciences, especially in wireline log estimations, interpretation of salt bodies, and fault definitions. Other applications such as seismic processing, automated first break picking, velocity analysis, and pre-stack inversions for rock property have been tackled with this technology. Our industry has large datasets that are especially suitable for data-driven methods. However, ML algorithms have been labeled as a "black box" as the actual process is not evident to the end user. Furthermore, the uncertainty of the results is always questioned.

We applied ML to estimate missing log data at a basin scale on millions of wells, and our findings indicate that measuring the uncertainty of the results in a meaningful way is extremely difficult. Instead of spending efforts quantifying uncertainty, we spent more time and work cleaning the input data to ML, which proved to be the most effective in improving the quality of the ML inference results. When estimations are compared to blind data, the data fit is very good if the input data have been cleaned and the uncertainties are small. If the input data are noisy, the comparisons are poor, and the uncertainty is larger.

Method

We used ML algorithms to estimate missing well-log information by training a network to predict one curve from a combination of one or more other curves. We target infill predictions of 5 curves: gamma ray, deep resistivity, neutral porosity, compressional sonic, and density. For each target curve, separate models are trained from different combinations of available curves.

The underlying algorithm for curve prediction is gradient-boosted regression trees. Gradient boosted trees belong to a class of tree-based methods with the capacity for modeling data-driven piece-wise target-feature interactions as opposed to multivariate linear regression, which requires certain assumptions about the correlations between input and target curves. The tree structure is chosen because it is robust - invariant to input scaling, and scalable – additional trees can be used to model higher-order interactions between features. Tree models essentially map a set of feature curves to target curves using iterative cost minimization. Feature curves can be recorded logs or representative attributes extracted from logs. Features can be continuous or categorical.

The inputs for the ML model training are different combinations of feature and target curves. The minimized cost function is the mean squared error between predicted and target curves. During each round of training, a new tree function is built to minimize the mean squared error cost function summed over all samples in the training dataset. At every iteration, the resulting model is used to predict a validation set, and the model which exhibits the lowest validation set errors is deployed over the whole basin on all wells that have at least one of the five curves available, figure 1.

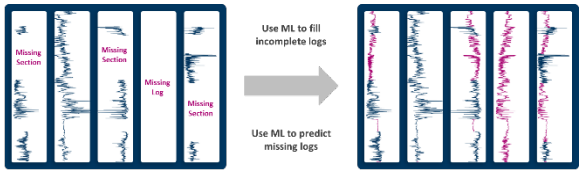

Figure 1: log prediction ML input and estimated logs

Quantile regression uses a typical linear regression model where the mean difference from the ground truth is being tracked, which allows for an optimization of the model. The downside of this approach is that a specific quantile (percentile) is being tracked against the medium of the ground truth. For example, tracking 5% quantile covers 5% of the data, and 95 percent quantile covers 95% of the data. Theoretically, this should provide lower and upper boundaries for the smallest and highest estimates in a regression task.

For regression prediction tasks, rarely, an absolute accurate prediction is performed as those are always inaccurate. Searching for an absolute precision acquired over a prediction interval is better. This is done using a Light Gradient Boosting Machine (LightGBM), which helps to increase the efficiency of a model, reduces memory usage, and is one of the fastest and most accurate libraries for regression tasks. In theory, it adds more utility to the model by implementing prediction intervals that show a range of possible values. Figure 2 shows an example of an estimated density curve. In addition to the actual measured curve (orange), all three predicted curves are represented, P5 (blue), P95 (black) and P50 (red). The ML algorithm effectively estimates density (red curve) compared to the measurement (orange curve). If the measured and predicted curves are similar, the P5 and P95 values shouldn't display a big deviation from the P50 curve. This is not the case in our example, which indicates that the input data are too noisy, and the algorithm has issues dealing with this.

10.1190/image2023-w03-01.1

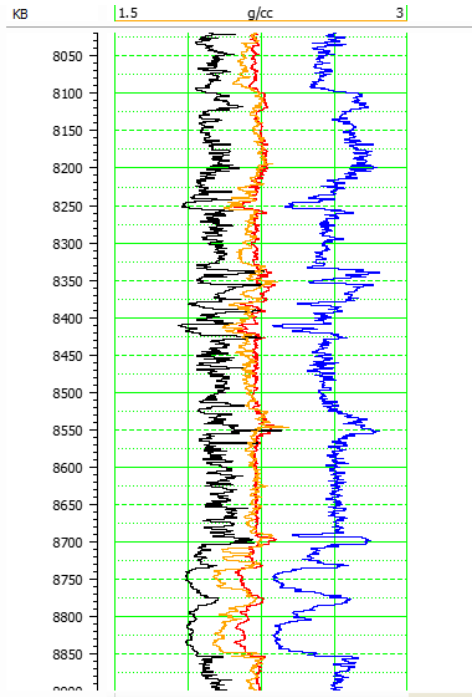

Figure 2: Density curves. Red: P50, black: P95, blue: P5, orange: measured curve.

The input data quality is the problem with using this method or any method to estimate the uncertainty in the prediction. Our research indicates that the quality of the input data plays a very large role in prediction accuracy (Gonzalez et al., 2023).

Conclusions

Artificial Intelligence, AI is playing more and more an important role in how we obtain additional data from existing data. None of the machine learning algorithms can generate a hundred percent result, and estimating these products' uncertainty seems desirable. However, there are caveats to this thinking. First, it isn't straightforward to obtain the uncertainty information. Second, our research has shown that uncertainty is linked to data quality. As such, it is more effective to spend time and effort generating clean input data than calculating uncertainties that are difficult to acquire and problematic to interpret.

Acknowledgement

The authors like to thank TGS to permit publishing these findings.

REFERENCE

Gonzalez, K., O. Brusova, and A. Valenciano, 2023, A machine learning workflow for log data prediction at the basin scale: First Break, **41**, 73–80, doi: https://doi.org/10.3997/1365-2397.fb2023015.