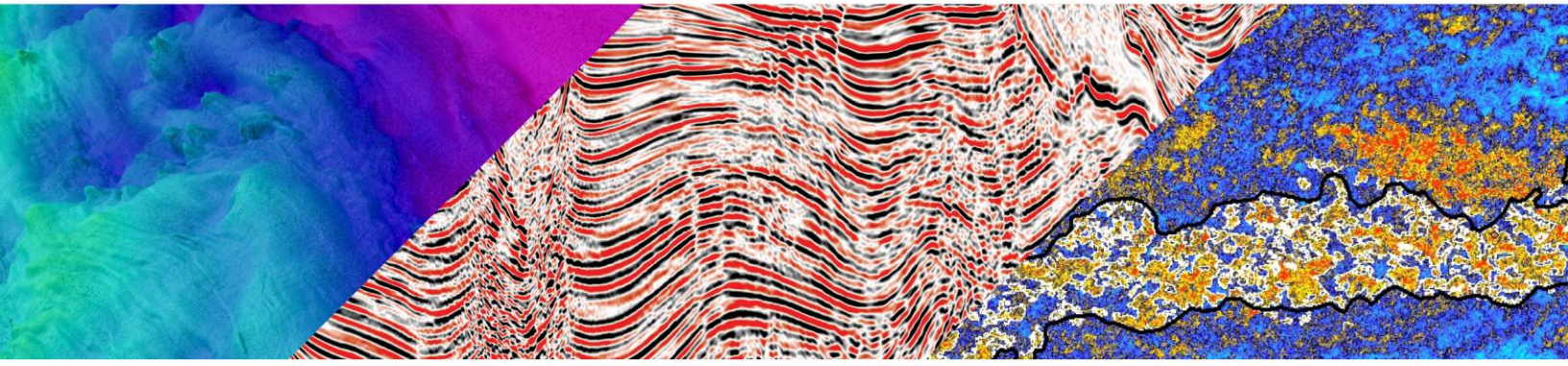
ARLAS White Paper

10451 Clay Road
Houston, TX 77041 USA
Tel: +1 713 860 2100
Fax: +1 713 334 3308

www.tgs.com



Contents

Analytics Ready LAS (ARLAS)

The primary objective of the Analytics Ready LAS (ARLAS) project is to produce complete well logs that will provide insight for different types of subsurface intelligence problems such as lithofacies picking, velocity model building and production prediction.

A basin-scale machine learning pipeline was developed to predict density, gamma ray, neutron porosity, deep resistivity and compressional sonic. 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 [1-3] with the capacity for modeling data driven piece-wise target-feature interactions.

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. Assuming we have a tree model with K trees, $x_i$represents the input feature vector, $f_k$represents each tree function and $\hat{y}_i$ is the predicted target curve. The subscript i enumerates each training sample in the training dataset.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

The inputs for ARLAS model training are different combinations of feature and target curves. For example, when predicting sonic from gamma and resistivity, the target sample will be recorded sonic and the feature vector will consist of gamma, resistivity, spatial locations and other relevant categorical features.

The cost function that we minimize is the mean squared error between predicted and target curves. During each round of training, a new tree function is built in a manner that minimizes the mean squared error cost function summed over all samples in the training dataset. At every iteration, the resulting model is used to predict on a validation set, and the model which exhibits lowest validation set error is deployed.

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

Training data selection

	Features				
Well Count	GR	Den	Res	Neut	Son
17		x			
18,453	x				
152	x	x			
109					x
7		x			x
125	x				x
6	x	x			x
41				x	
8		x		x	
640	x			x	
219	x	x		x	
1				x	x
27	x			x	x
6	x	x		x	x
60,555			x		
95		x	x		
25,130	x		x		
1,073	x	x	x		
3,540			x		x
35		x	x		x
3,696	x		x		x
31	x	x	x		x
127			x	x	
37		x	x	x	
20,369	x		x	x	
2,541	x	x	x	x	
59			x	x	x
8		x	x	x	x
3,798	x		x	x	
2,009	x	x	x	x	x

Objective:

Select a set of uniformly distributed well logs with good depth coverage of all five curve classes.

Method:

The current ARLAS pipeline performs basin-scale model training and log predictions. All well logs located within the boundaries of a basin trend are selected and downloaded.

For all LAS files, we use some SQL selection rules to determine the mnemonic that is used to represent density, gamma, neutron porosity, deep resistivity and compressional sonic curve classes.

Given a set of LAS files and available mnemonics, we divide the logs into sets containing 1, 2, 3, 4 and 5 curves. If there are enough well logs containing all five curve classes, this set of LAS files will be used in our training and validation set. Otherwise, this set of LAS files will be used for training models which predict a target from 4 curves and a selection of different sets of LAS containing 4, 3 and 2 curves will be augmented to this set to train the other models.

Fig. 1 shows the number of LAS files containing 1, 2, 3, 4 and 5 curve classes for EagleFord and the highlighted boxes are the files selected for training and validation. Spatial distribution of the well logs is also accounted for in the selection of LAS files for model training.

FIGURE 1 NUMBER OF LAS FILES WITH 1, 2, 3, 4 AND 5 CURVE CLASSES IN EAGLEFORD BASIN

Data Cleanup

Objective:

Obtain a clean training set with each curve class represented in the same units and amplitude ranges within reasonable limits for the geology of the basin

Method:

We perform different types of preconditioning to the training dataset, including:

- Classifying wellbore cave-in
- Standardizing curve units
- Gamma and Neutron porosity normalization
- Separation of recordings from disparate tools
- Curve amplitude capping to remove anomalous data

Fig. 2 illustrates amplitude distributions for the five curves classes and yellow rows indicate capping ranges for the EagleFord basin.

	DEN	GR	NEUT	RES	SON
0.00%	-32768.00	-150.47	-0.75	-0.24	-2618.00
0.50%	1.76	6.68	0.00	0.25	0.33
1.00%	1.87	12.98	0.01	0.29	35.69
5.00%	2.06	30.03	0.07	0.47	58.67
10.00%	2.13	39.84	0.15	0.61	70.63
20.00%	2.21	51.38	0.22	0.82	80.65
30.00%	2.28	59.48	0.25	1.04	86.70
40.00%	2.34	66.02	0.28	1.32	92.33
50.00%	2.40	71.90	0.31	1.66	97.76
60.00%	2.45	77.61	0.34	2.08	103.31
70.00%	2.50	83.72	0.37	2.73	110.13
80.00%	2.54	90.80	0.40	4.06	119.46
90.00%	2.59	101.18	0.45	8.18	133.74
97.00%	2.68	117.59	0.52	35.19	154.18
98.00%	2.70	123.19	0.54	60.11	160.13
99.00%	2.74	134.12	0.58	149.90	171.93
99.10%	2.74	135.88	0.59	174.45	174.25
99.20%	2.75	137.94	0.59	202.74	176.97
99.30%	2.75	140.22	0.60	246.72	180.38
99.40%	2.76	142.87	0.60	313.94	184.81
99.50%	2.77	146.17	0.63	415.66	190.51
99.60%	2.79	149.98	0.69	602.25	198.25
99.70%	2.82	155.55	0.93	997.48	208.83
99.80%	2.86	166.42	27.19	1560.26	229.34
99.90%	2.93	206.58	36.03	1973.71	250.83
100.00%	994.44	1674.01	63.46	19348522.00	3624.00

FIGURE 2 AMPLITUDE DISTRIBUTION FOR EACH CURVE CLASS IN THE EAGLEFORD BASIN

Model Training

Objective:

Obtain regression models to predict target curves based on different combinations of 4, 3, 2 and 1 available curves at a given spatial location.

Method:

For each target curve, we train 15 models to predict the target when different combinations of 4, 3, 2 and 1 curves are present. We tested different types of feature engineering on the curves including smoothing, derivative, logarithm, compound terms, etc.

We trained 75 curve prediction models, 75 5th percentile regression models and 75 95th percentile regression models for a total of 225 models per basin. We train and perform hyperparameter tuning of the models on GPU's.

Model Validation

Objective:

Obtain expected errors when using the trained models for curve predictions within a given basin.

Method:

We perform model validation on a set of testing wells that were not used in training. The error rates computed from these wells are reported as our expected errors for a given model for each basin. Fig. 3 shows an example of spatial distribution of mean prediction errors for density computed on test wells in the EagleFord basin. The table in Fig. 4 provides the mean percentage errors for all curve prediction models in the EagleFord basin.

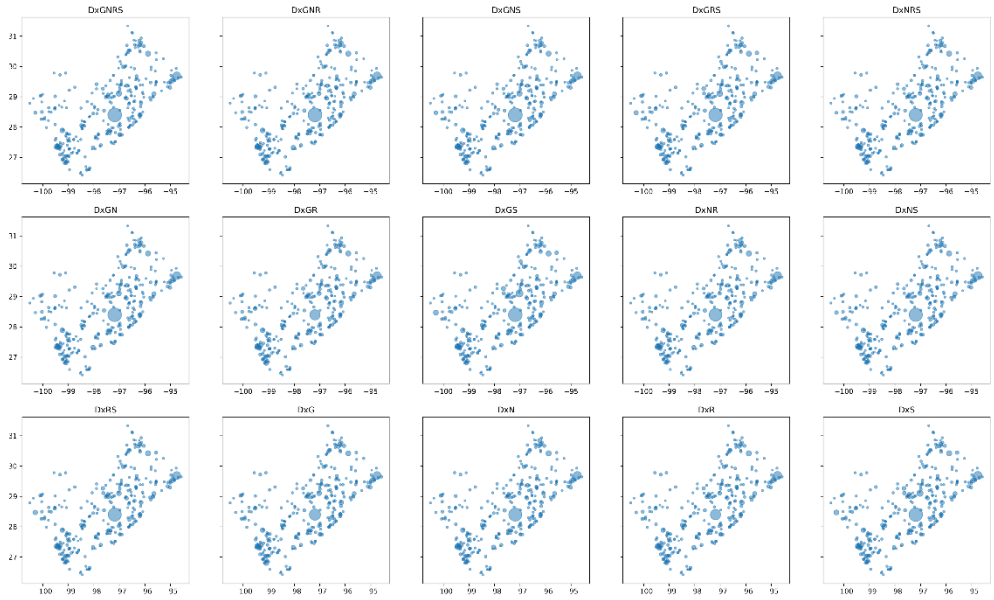


FIGURE 3 SPATIALLY DISTRIBUTED VALIDATION ERRORS FOR DENSITY PREDICTION MODELS IN EAGLEFORD BASIN

Features	D	G	N	R	S	DG	DN	DR	DS	GN	GR	GS	NR	NS	RS	DGN	DGR	DGS	DNR	DNS	DRS	GNR	GNS	GRS	NRS	DGNR	DGNS	DGRS	DNRS	GNRS
Target																														
D		14	13.3	13.7	13.8					12.6	13.2	13.3	12.8	13	13.3							12.3	12.5	13	12.6					12.3
G	19.9		18.4	18.5	18.6		17.8	18.6	18.5				17.6	17.9	17.8				17.5	18	18				17.2				17.6	
N	17.4	16.5		15.3	14.2	15.8		15.1	14		14.5	13.7			13.8		13.9	13.4			13.7			13.4				13.3		
R	11.1	10.3	9.6		9.1	11	10.7		9	9.4		8.9		8.9		9.6		8.8		8.6			8.6				8.6			
S	14.5	14.5	13.2	12.8		13.4	12.3	12		12.9	12.7		11.9			12	11.7		11.2			11.9				11				

FIGURE 4 MEAN PERCENTAGE ERRORS OF PREDICTION FOR ALL EAGLEFORD BASIN MODELS

ARLAS Generation

Objective:

Obtain Analytics Ready LAS files containing completed well logs with standardized units and mnemonics for density, gamma ray, neutron porosity, deep resistivity and compressional sonic.

Method:

All well logs from a given basin are preprocessed with data cleanup steps. Curve gaps and missing curves are predicted using trained models.

References

[1] Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

[2] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[3] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).