Machine Learning Approach To Log-based Lithology Interpretation

Promise Ekeh

Introduction

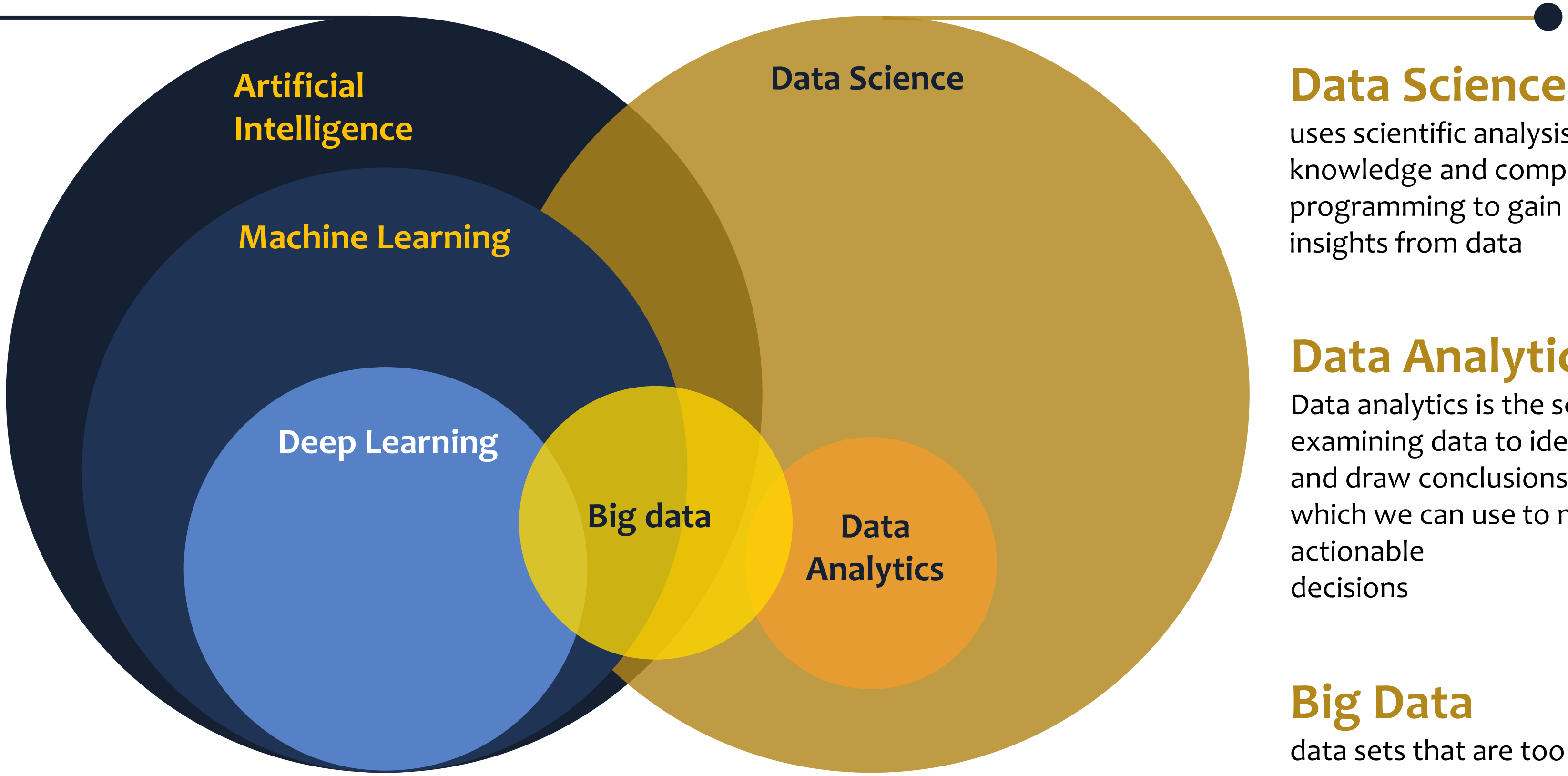
- **Lithology** is the description of general physical characteristics, composition or type of rock
- It gives information about the rock's depositional and diagenetic history
- Lithology interpretation is critical across all phases of exploration through to development and production. Its importance includes:
    - Reservoir characterization for flow unit identification
    - Petrophysical Evaluation - porosity, water saturation (Sw), and permeability
    - Identifying fluid type from Neutron density tools

Introduction – AI vs DS

Artificial Intelligence
Technology for machines to interpret, understand, learn and make intelligent decisions.
Machines with human intelligence

Machine Learning
Use algorithms to make computer learn from data without being programmed to do so
- Supervised
- Unsupervised
- Reinforcement

Deep Learning
A category of machine learning that emphasizes training the computer about the basic instincts of human beings.



Data Science
uses scientific analysis, statistical knowledge and computer programming to gain meaningful insights from data

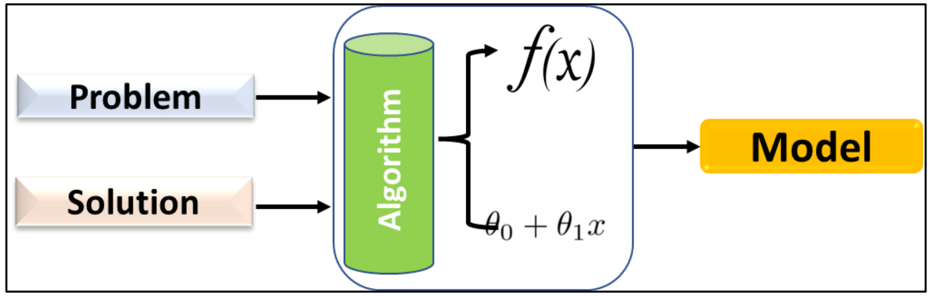
Data Analytics
Data analytics is the science of examining data to identify trends and draw conclusions from them, which we can use to make actionable decisions

Big Data
data sets that are too large or complex to be dealt with by traditional data-processing application software

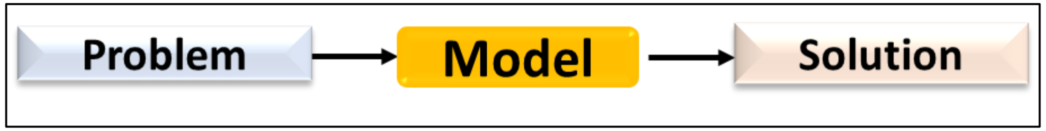
Introduction – ML

Machine Learning (ML) is a branch of Artificial Intelligence (AI) in which computers learn patterns/trends from data and make predictions without necessarily being programmed to do so

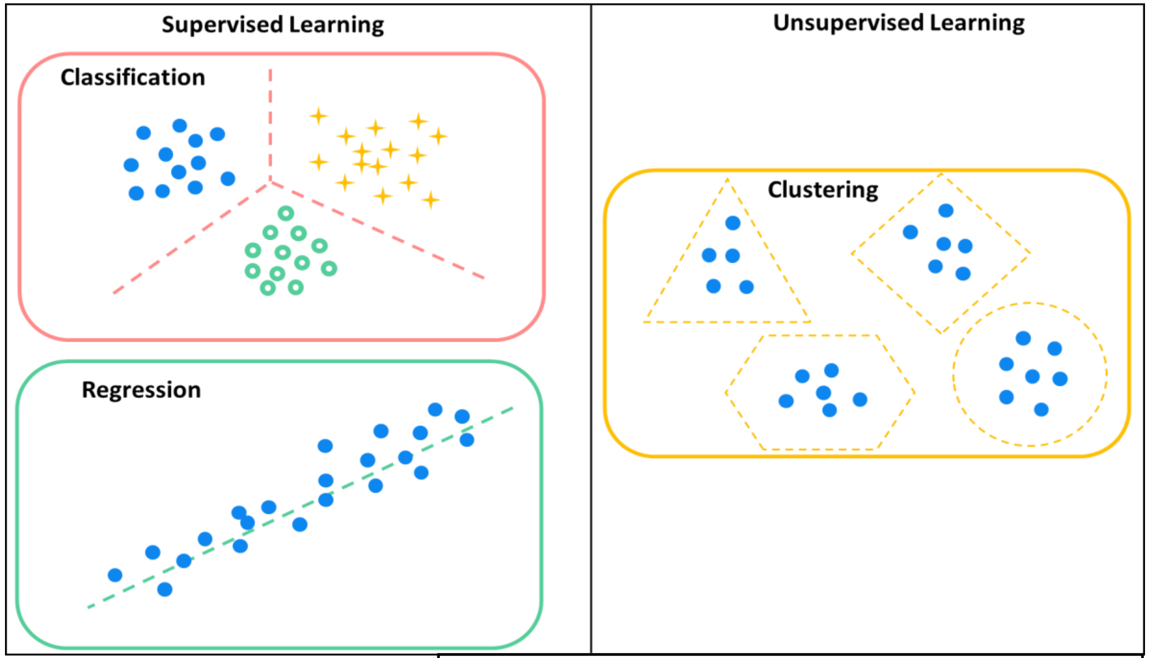
- **Training model**



- **Making predictions**



- **Types of ML**



Reinforcement
Algorithm provided with a set of actions, parameters, and results. Through trial and error the algorithm will learn to obtain optimal results by reinforcing its own behaviour.

Although enterprise AI applications can simplify the complexity, subject matter expertise plays an important role in successful deployments

Hybrid Approach = **Knowledge Driven** + **Data Driven**

Knowledge Driven: Based on a users pre-existing understanding of the data, context, discipline, etc. Algorithm aims to learn from the data.

Data Driven: Based on the input of no prior knowledge of the data. The algorithm strives to understand context and learn.

Introduction – Aim & Objectives

Aim

to produce model to automate lithology interpretation from well logs, which will serve as input into predictions of facies from multi-attribute ML analysis of seismic cubes within the north sea

Objectives

- Comprehensive data pre-processing
- Apply different ML classifiers to obtain the best predictive model
- Train model with basic logs suites that will be widely adopted.

DATA

- 2020 Force Database
  - Well logs with labelled lithologies
  - Penalty matrix

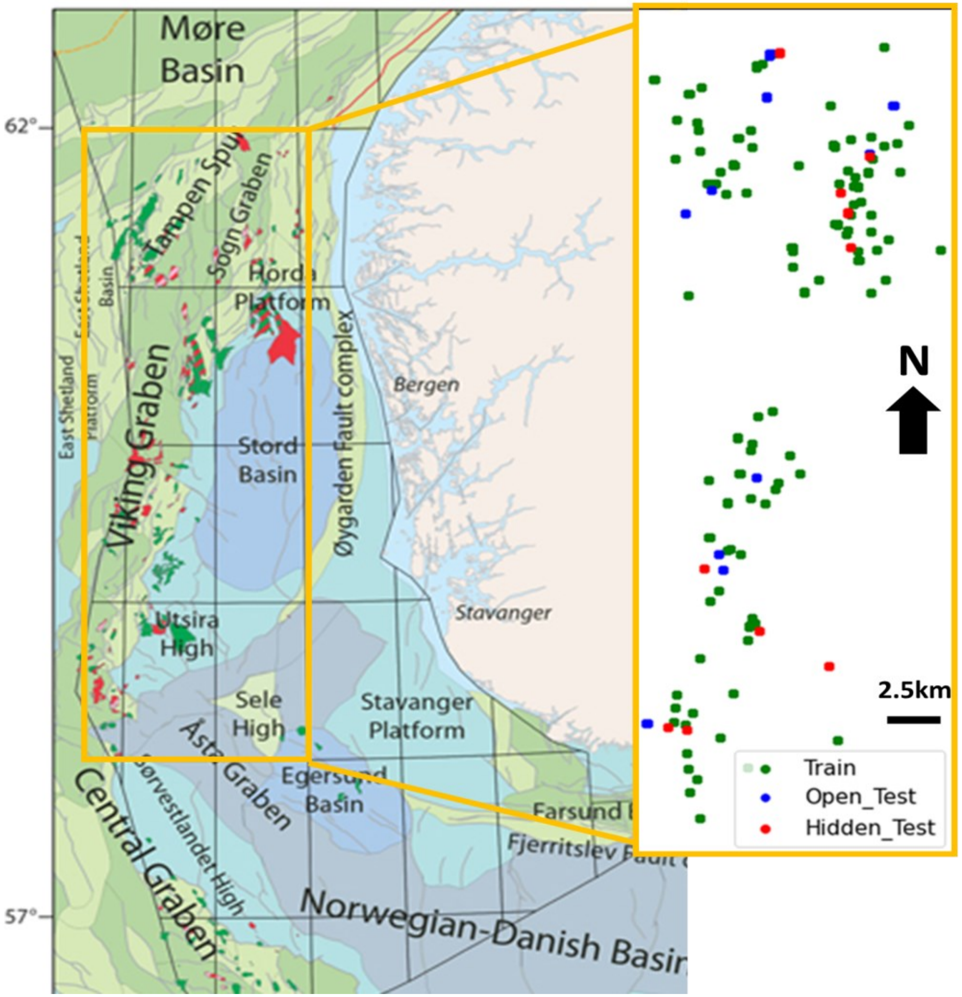
118 Wells

- **20 Well logs** →
- **69 Formations**
  14 Groups
- **12 Labelled Lithologies** →

1. Gamma ray	GR
2. Deep resistivity	RDEP
3. Med resistivity	RMED
4. Shallow resistivity	RSHA
5. Flushed Zone Resistivity	RXO
6. Comp. waves sonic	DTC
7. Shear wave sonic log	DTS
8. Caliper	CALI
9. Density	RHOB
10. Density correction	DRHO
11. Spontaneous Potential	SP
12. Neutron Porosity	NPHI
13. Bit Size	BS
14. Photoelectric absorption	PEF
15. Rate of penetration	ROP
16. Mud Weight	
17. Differential Caliper	DCAL
18. Avg. Rate of penetration	ROPA
19. Micro Resisitivity	RMIC
20. Spectral Gamma Ray	SGR

1. Shale
2. Sandstone
3. Shale/sandstone
4. Limestone
5. Marl
6. Tuff
7. Chalk
8. Halite
9. Coal
10. Dolomite
11. Anhydrite
12. Basement

- Diskos (NPD)- composite and mudlog report



Workflow

Exploratory Data Analysis

- Correlation Metric
- Missing values
- Log Distribution(Pairplots & boxplots)
- Lithology Distribution

Pre-processing

- Handling Outlier
- Removing badholes
- Feature Selection
- Feature/Label encoding
- Data Splitting

Feature Engineering

- Vshale
- Resampling

Model Training

- Random Forest
- XGBOOST
- CATBOOST

Model Optimization

- Hyperparameter Tunning
- Stratified Cross validation

Evaluation

- Accuracy Score
- Confusion Matrix
- Visual Comparison

Tool

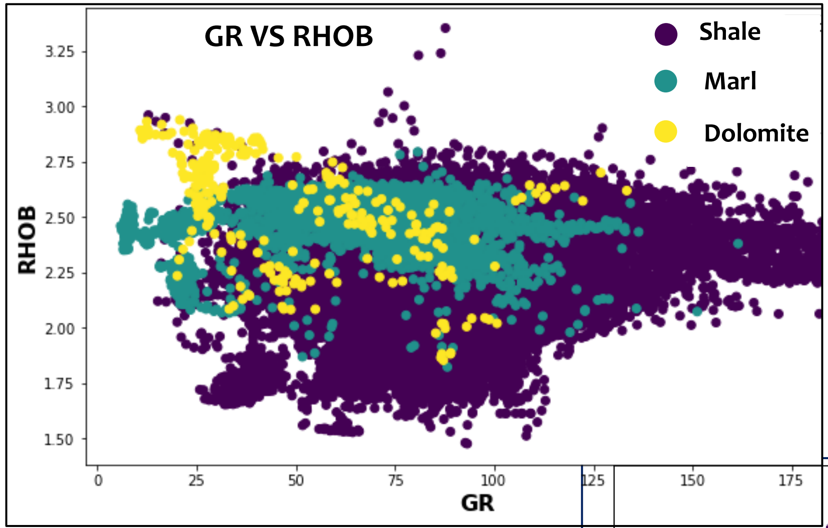
Anaconda/ Jupyter Notebook

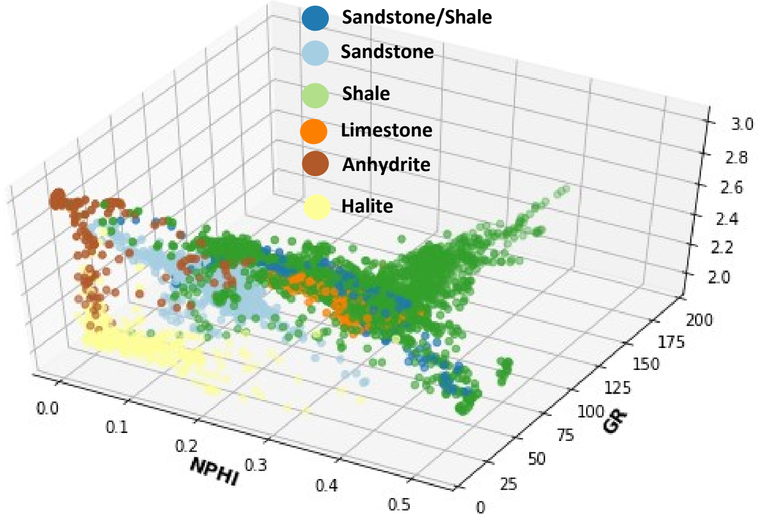
Python version 3.8.3

Workflow – EDA

Exploratory Data Analysis

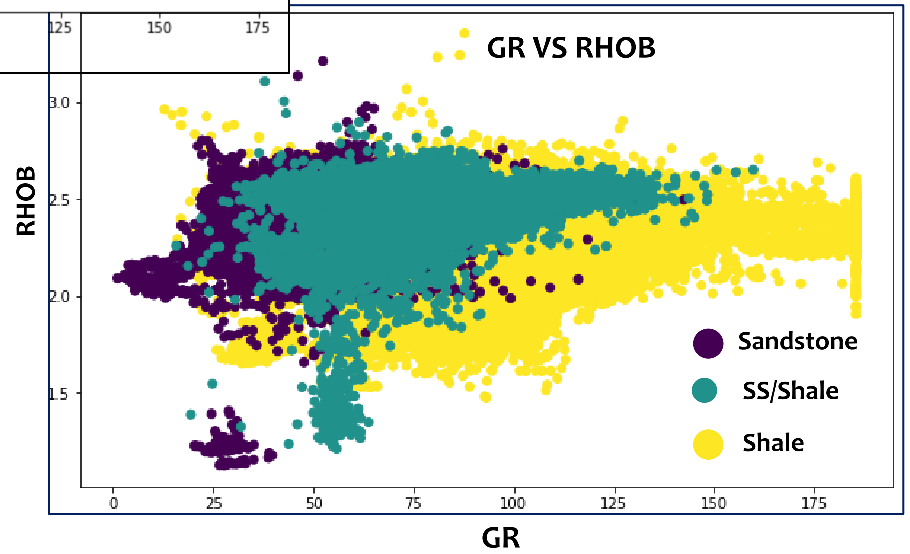
- Correlation Metric
- Missing values
- Log Distribution(Pairplots & box plots)
- Lithology Distribution



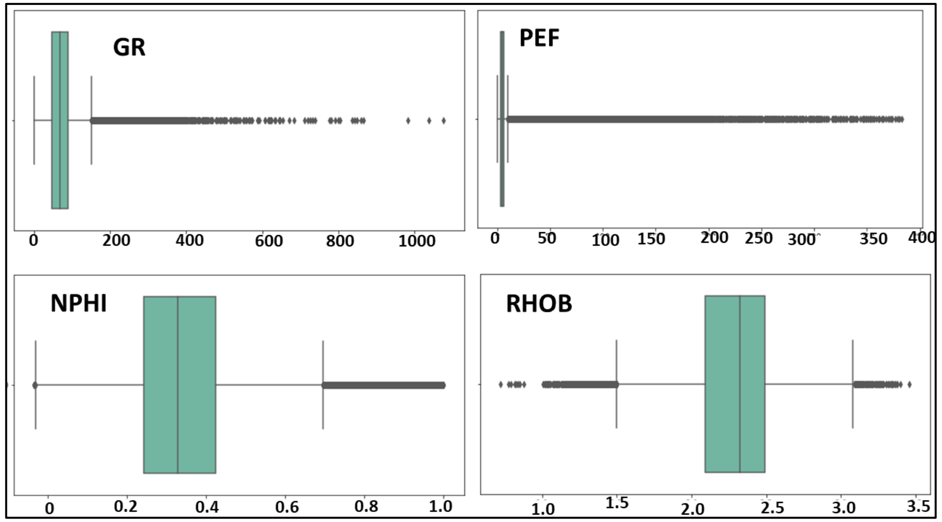


Correlation Matrix

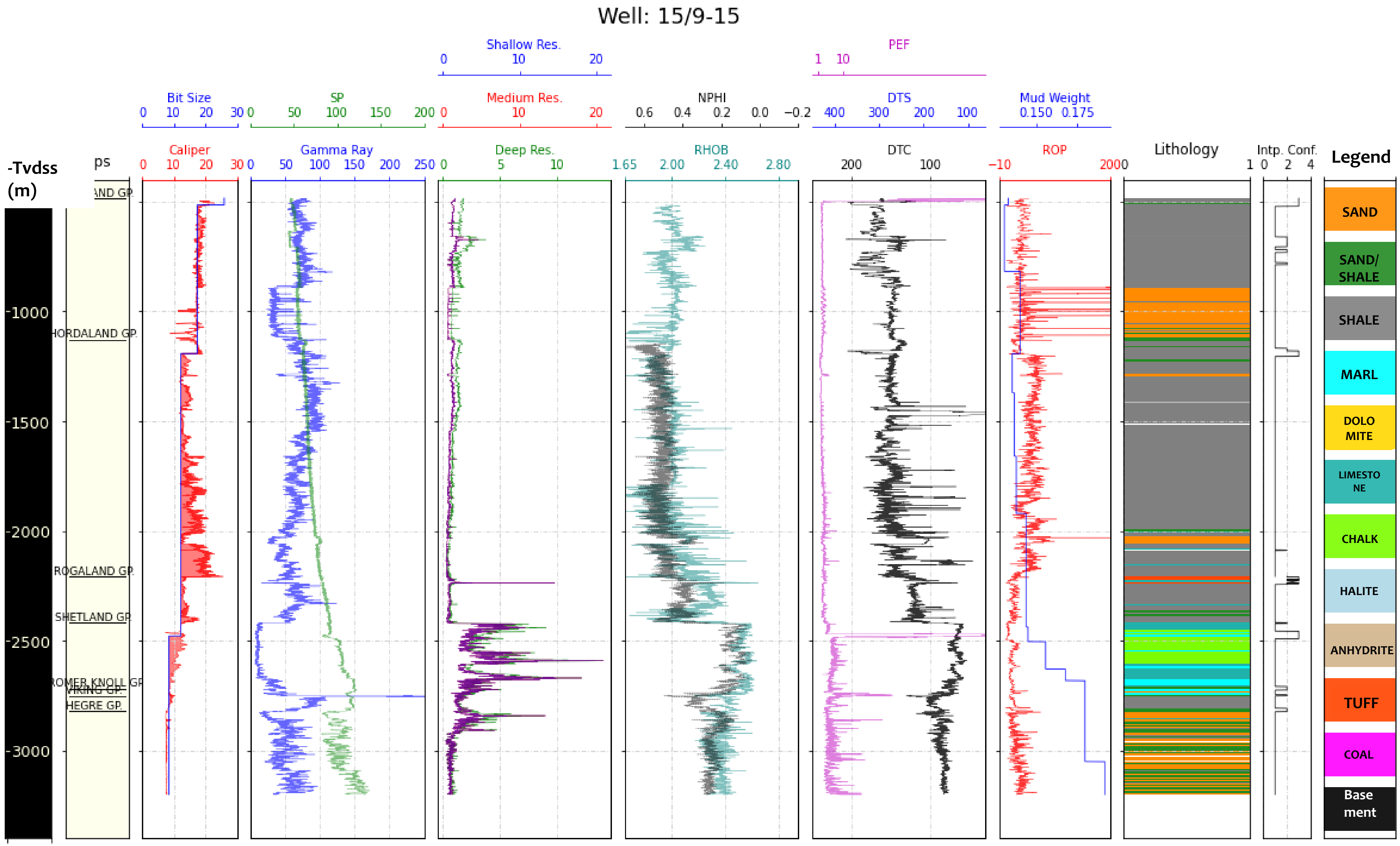
	Z_LOC	DEPTH_MD	CALI	RHOB	GR	NPHI	BS	PEF	DTC	SP	DRHO	RMED
Z_LOC	1.00											
DEPTH_MD	-0.99	1.00										
CALI	0.58	-0.58	1.00									
RHOB	-0.74	0.74	-0.57	1.00								
GR	-0.09	0.10	-0.24	0.17	1.00							
NPHI	0.55	-0.54	0.45	-0.62	0.22	1.00						
BS	0.59	-0.59	0.90	-0.48	-0.14	0.35	1.00					
PEF	-0.03	0.03	0.05	-0.09	-0.08	0.02	0.08	1.00				
DTC	0.78	-0.78	0.59	-0.83	0.02	0.79	0.54	0.03	1.00			
SP	-0.03	0.03	-0.03	0.06	0.10	-0.11	0.09	-0.01	-0.02	1.00		
DRHO	-0.00	0.00	0.00	0.00	0.00	-0.03	0.00	-0.00	-0.05	-0.05	1.00	
RMED	-0.04	0.06	-0.04	-0.01	-0.02	-0.05	-0.09	0.06	-0.10	-0.02	-0.00	1.00



Box plot- Data Distribution



Workflow – EDA



PS: Image from python

Workflow – Pre-processing

Pre-processing

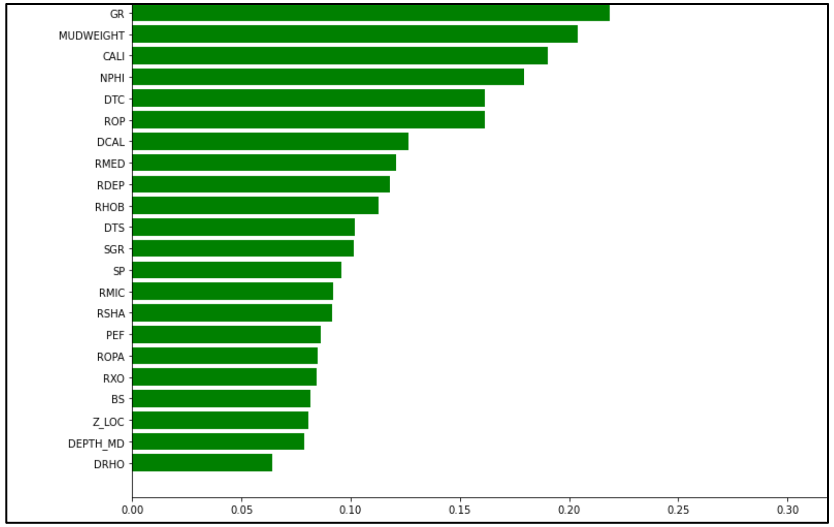
- Handling Outlier
- Removing badholes
- Feature Selection
- Feature/Label encoding
- Data Splitting

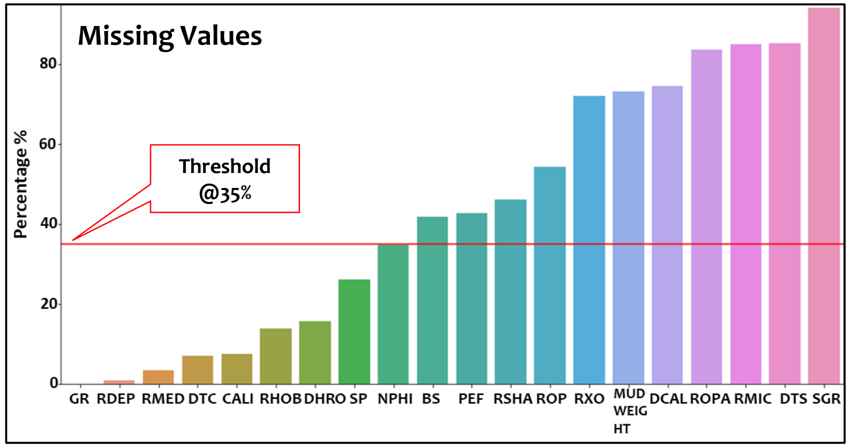
Dataset

Training — 108 Wells

Test — 10 Wells

Log Importance to Lithology distribution





Label encoding

Shale	0
Sandstone	1
Shale/sandstone	2
Limestone	3
Marl	4
Tuff	5
Chalk	6
Halite	7
Coal	8
Dolomite	9
Anhydrite	10
Basement	11

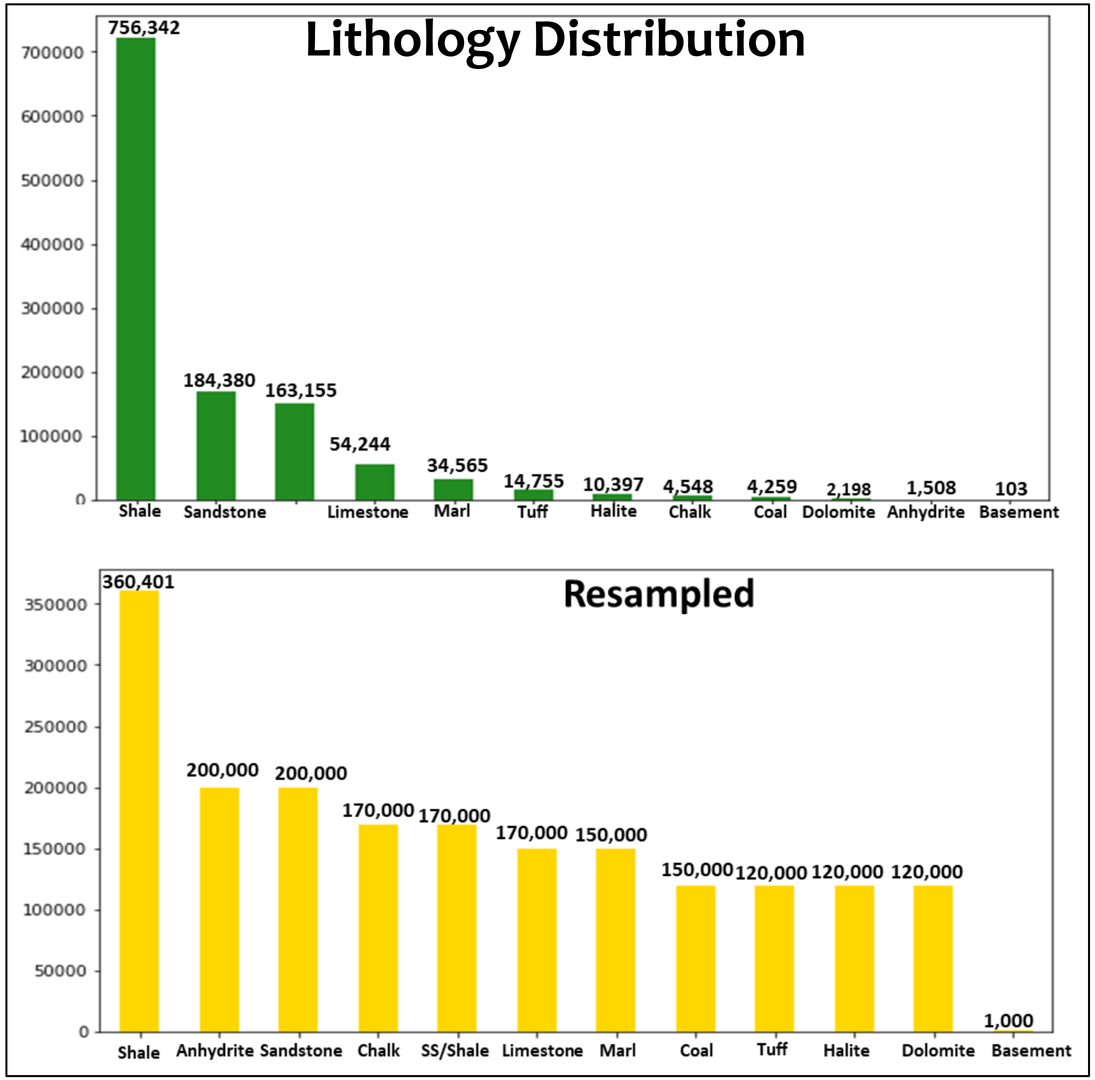
Workflow – Feature Engineering

Feature Engineering

- Vshale
- Resampling

$$V_{sh} = \frac{GR_{sh} - GR}{GR_{sh} - GR_{cl}}$$

- **Resampling**

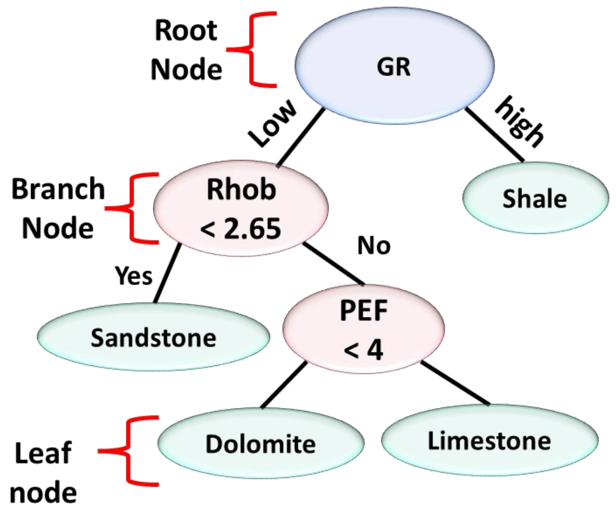


Workflow – ML Algorithms

Model Training

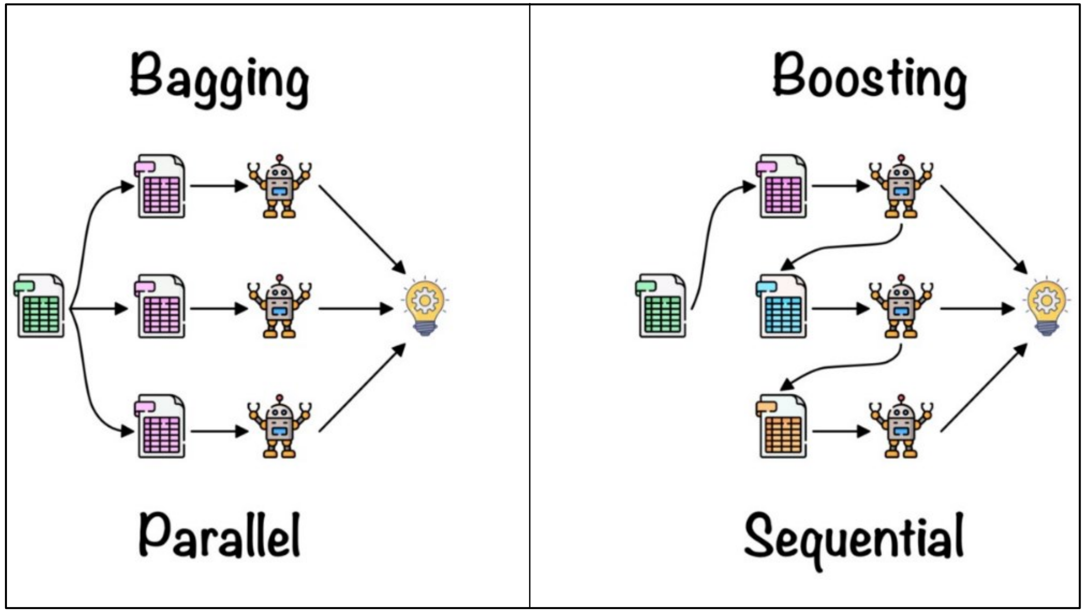
- Random Forest
- XGBOOST
- CATBOOST

Decision trees (DT)



Ensemble of DT: combining multiple tree. Making predictions from a collection of trees rather than an individual tree.

Random Forest **CATBOOST & XGBOOST**



Random Forest

- It is flexible to use
- Randomly selects features

CATBOOST

- Does not require much parameter tuning
- Can handle categorical values

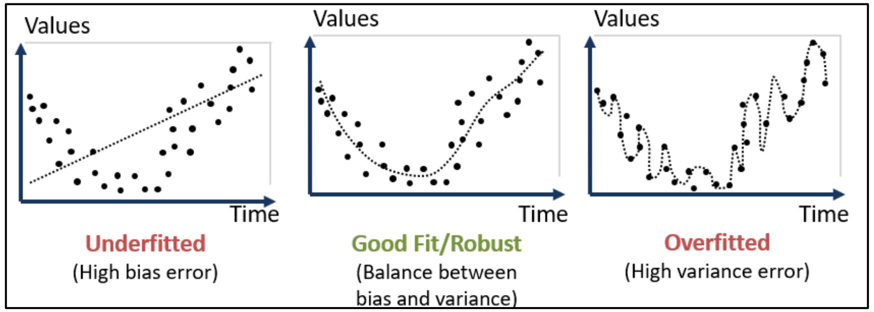
XGBOOST

- Fast
- Can handle missing values
- regularization techniques that reduce overfitting
- Tree Pruning within algorithm

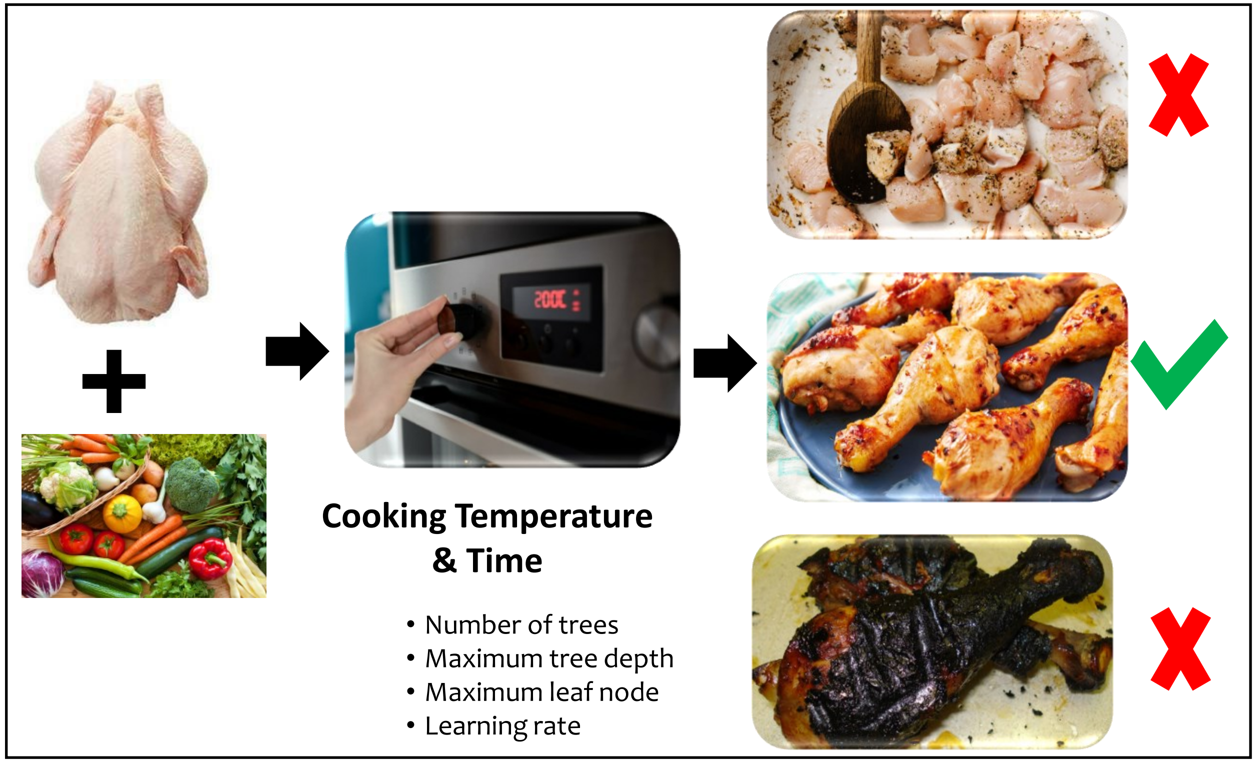
Workflow – Model Optimisation

Model Optimization

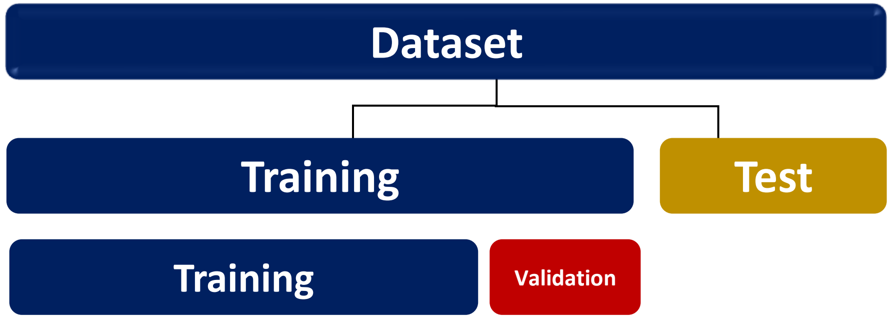
- Hyperparameter Tunning
- Stratified Cross validation



Hyper parameter tuning



Cross fold validation



	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Workflow – Evaluation

Evaluation

- Accuracy Score
- Confusion Matrix
- Visual Comparison

Accuracy Score is the proportion of the total number of correct predictions, i.e. ratio of correct predictions to the number of total predictions.

Penalty Matrix

Label \ Prediction	Sandstone	SS/shale	Shale	Marl	Dolomite	Limestone	Chalk	Halite	Anhydrite	Tuff	Coal	Basement
Sandstone	0	2	3.5	3	3.75	3.5	3.5	4	4	2.5	3.88	3.25
SS/shale	2	0	2.38	2.75	4	3.75	3.75	3.88	4	3	3.75	3
Shale	3.5	2.38	0	2	3.5	3.5	3.75	4	4	2.75	3.25	3
Marl	3	2.75	2	0	2.5	2	2.25	4	4	3.38	3.75	3.25
Dolomite	3.75	4	3.5	2.5	0	2.63	2.88	3.75	3.25	3	4	3.63
Limestone	3.5	3.75	3.5	2	2.63	0	1.38	4	3.75	3.5	4	3.63
Chalk	3.5	3.75	3.75	2.25	2.88	1.38	0	4	3.75	3.13	4	3.75
Halite	4	3.88	4	4	3.75	4	4	0	2.75	3.75	3.75	4
Anhydrite	4	4	4	4	3.25	3.75	3.75	2.75	0	4	4	3.88
Tuff	2.5	3	2.75	3.38	3	3.5	3.13	3.75	4	0	2.5	3.25
Coal	3.88	3.75	3.25	3.75	4	4	4	3.75	4	2.5	0	4
Basement	3.25	3	3	3.25	3.63	3.63	3.75	4	3.88	3.25	4	0

$$S = -\frac{1}{N}\sum_{i=0}^{N} A_{\hat{y}_i y_i}$$

N =number of samples
yi =prediction for sample i,
$\hat{y}i$ =actual target for sample i, and A = penalty matrix.

Confusion Matrix

Predicted

Label	Sandstone	SS/shale	Shale	Marl	Dolomite	Limestone	Chalk	Halite	Anhydrite	Tuff	Coal
Sandstone	66%	13%	5%	9%	0%	5%	2%	0%	0%	2%	0%
SS/shale	24%	24%	44%	4%	0%	0%	0%	0%	0%	2%	1%
Shale	1%	2%	89%	6%	0%	1%	0%	0%	0%	2%	0%
Marl	1%	11%	22%	44%	1%	17%	2%	0%	0%	1%	0%
Dolomite	9%	14%	29%	13%	24%	1%	0%	0%	9%	0%	0%
Limestone	10%	1%	7%	14%	2%	59%	5%	0%	1%	1%	0%
Chalk	20%	0%	0%	0%	0%	22%	53%	0%	0%	5%	0%
Halite	1%	0%	0%	0%	0%	7%	0%	92%	0%	0%	0%
Anhydrite	0%	1%	0%	3%	4%	15%	0%	8%	69%	0%	0%
Tuff	20%	5%	36%	3%	0%	3%	0%	0%	0%	33%	0%
Coal	2%	4%	2%	0%	0%	0%	0%	0%	0%	0%	91

100 80 60 40 20 0

Results – Confusion matrix

Results for all 10 Test wells

XGB Model

Accuracy score= 75%
Penalty score= - 0.64

Predicted (columns) / Label (rows)

	Sandstone	SS/shale	Shale	Marl	Dolomite	Limestone	Chalk	Halite	Anhydrite	Tuff	Coal
Sandstone	66%	13%	5%	9%	0%	5%	2%	0%	0%	2%	0%
SS/shale	24%	24%	44%	4%	0%	0%	0%	0%	0%	2%	1%
Shale	1%	2%	89%	6%	0%	1%	0%	0%	0%	2%	0%
Marl	1%	11%	22%	44%	1%	17%	2%	0%	0%	1%	0%
Dolomite	9%	14%	29%	13%	24%	1%	0%	0%	9%	0%	0%
Limestone	10%	1%	7%	14%	2%	59%	5%	0%	1%	1%	0%
Chalk	20%	0%	0%	0%	0%	22%	53%	0%	0%	5%	0%
Halite	1%	0%	0%	0%	0%	7%	0%	92%	0%	0%	0%
Anhydrite	0%	1%	0%	3%	4%	15%	0%	8%	69%	0%	0%
Tuff	20%	5%	36%	3%	0%	3%	0%	0%	0%	33%	0%
Coal	2%	4%	2%	0%	0%	0%	0%	0%	0%	0%	91%

Random Forest Model

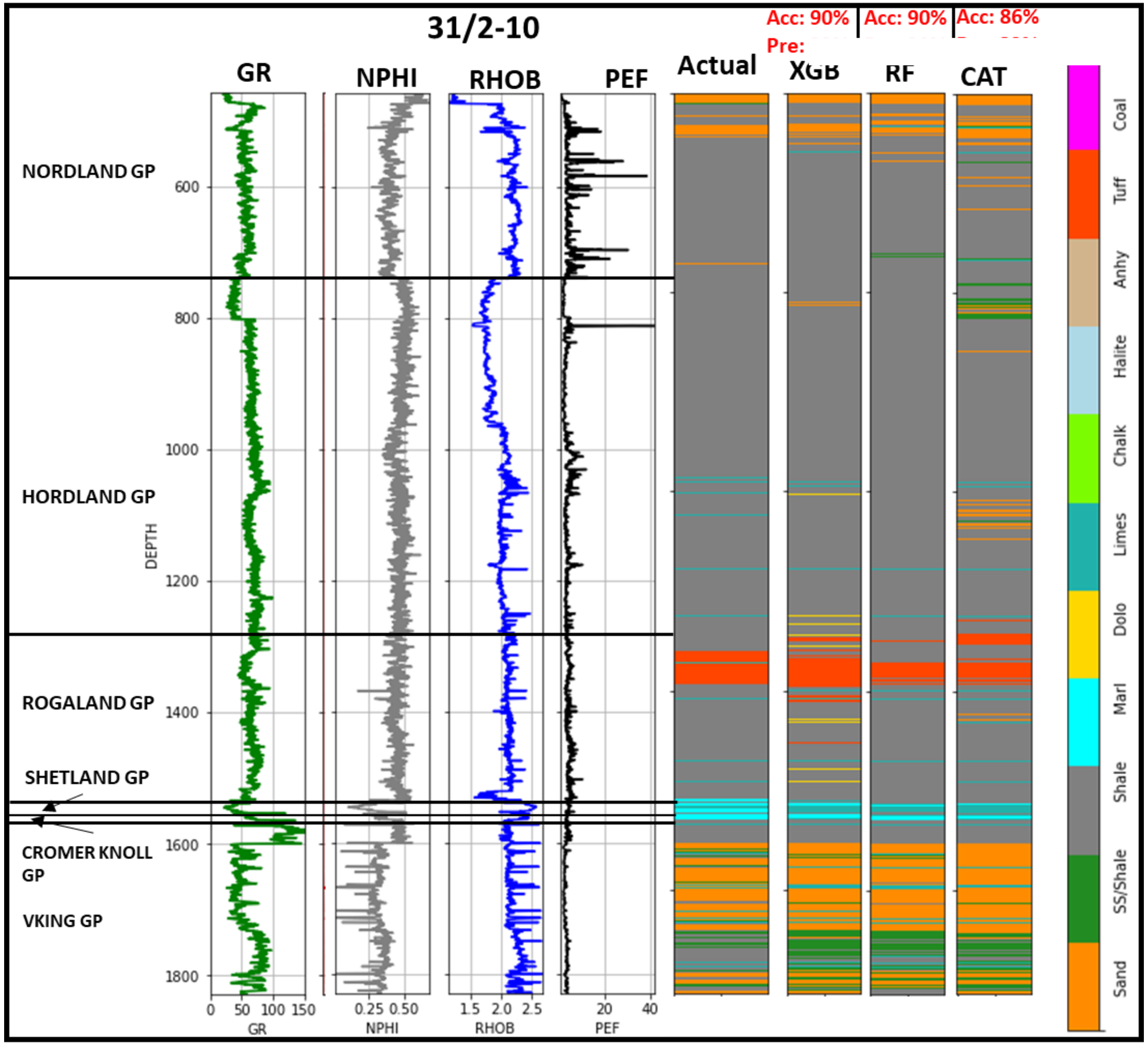
Accuracy score= 76%
Penalty score= - 0.61

Predicted

	Sandstone	SS/shale	Shale	Marl	Dolomite	Limestone	Chalk	Halite	Anhydrite	Tuff	Coal
Sandstone	63.55%	10.35%	9.79%	6.75%	1.80%	6.81%	0.77%	0.00%	0.16%	0.01%	0.01%
SS/shale	13.26%	27.88%	53.84%	3.77%	0.07%	0.64%	0.00%	0.00%	0.02%	0.37%	0.14%
Shale	0.79%	1.81%	92.79%	3.38%	0.12%	0.97%	0.02%	0.02%	0.02%	0.01%	0.06%
Marl	3.71%	18.49%	26.73%	23.27%	0.43%	24.77%	2.59%	0.00%	0.00%	0.00%	0.00%
Dolomite	1.39%	0.00%	43.55%	0.00%	0.35%	5.57%	0.00%	6.97%	42.16%	0.00%	0.00%
Limestone	9.73%	1.44%	9.57%	7.34%	1.50%	66.60%	3.58%	0.00%	0.06%	0.17%	0.00%
Chalk	11.46%	0.00%	0.00%	0.07%	0.00%	55.90%	32.56%	0.00%	0.00%	0.00%	0.00%
Halite	0.38%	0.00%	0.03%	0.15%	0.00%	10.60%	0.00%	88.80%	0.03%	0.00%	0.00%
Anhydrite	0.34%	0.00%	2.01%	1.84%	0.67%	10.05%	0.00%	9.88%	75.21%	0.00%	0.00%
Tuff	1.70%	0.00%	67.80%	0.11%	0.00%	9.35%	0.00%	0.00%	0.00%	21.04%	0.00%
Coal	15.98%	3.28%	12.70%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	68.03%

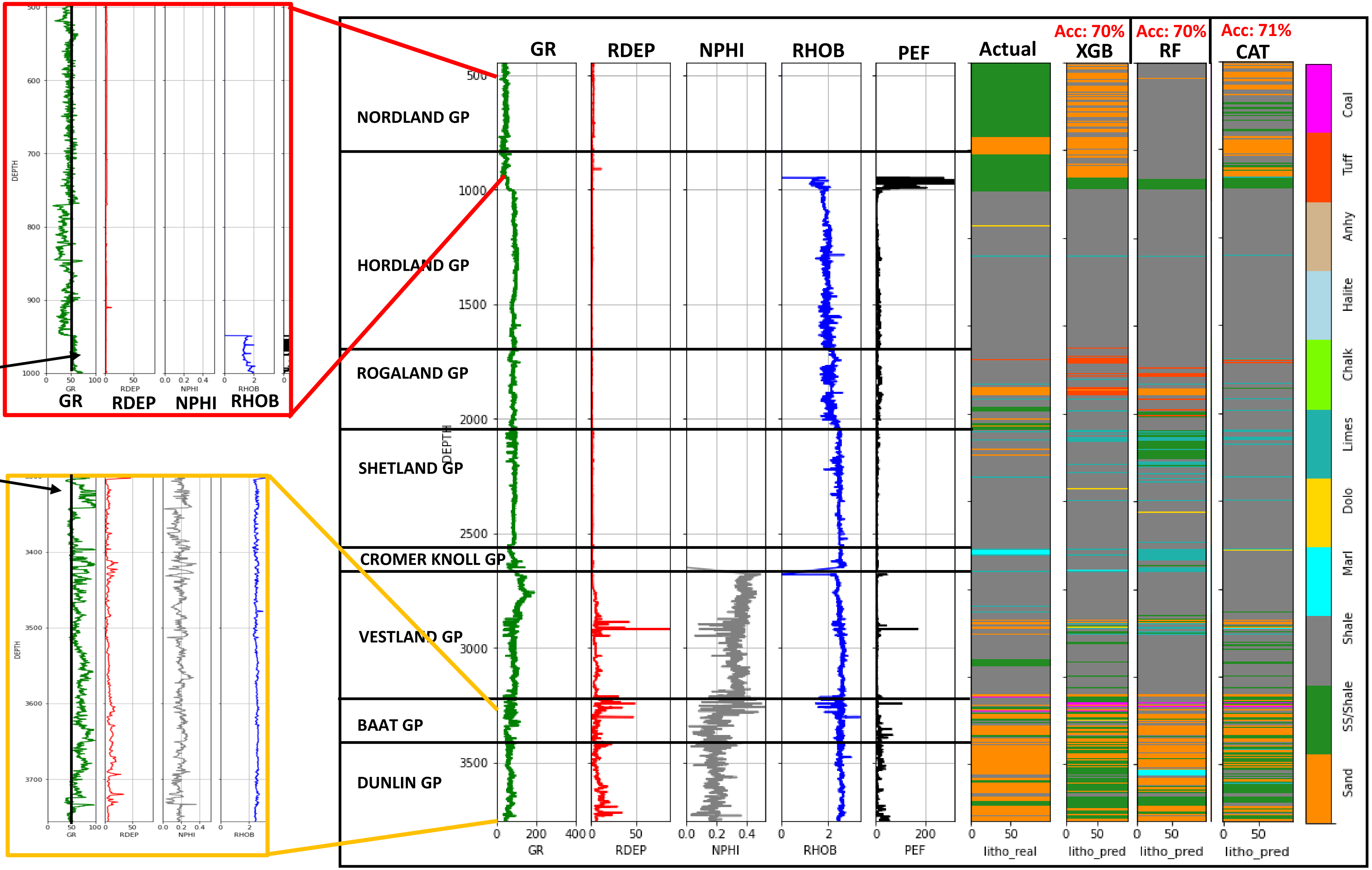
Results – Well 31/2-10

- 90% average performance
- Best performing well
- Misprediction of some thin packages
- Complete log run



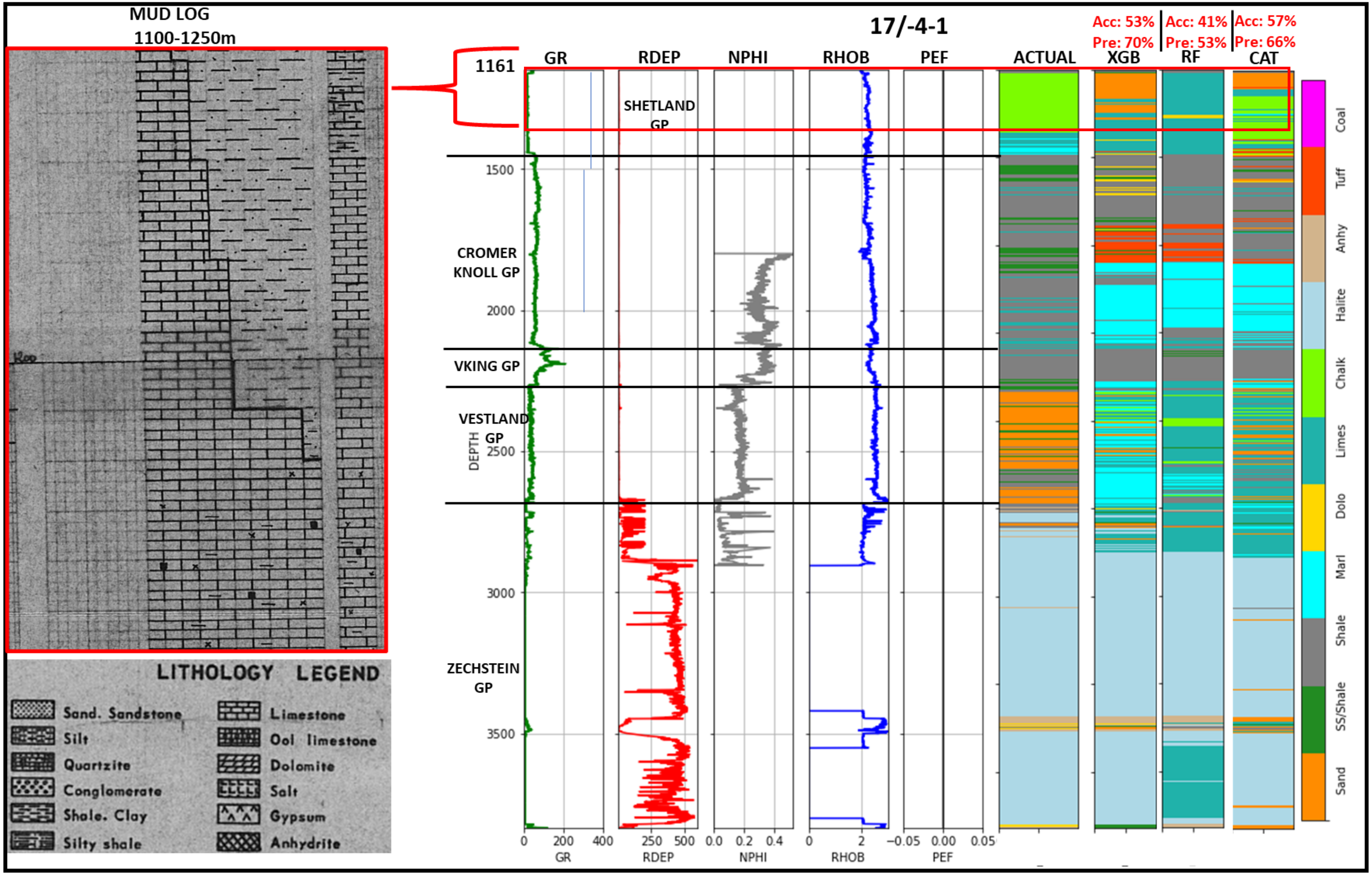
Results – Well 35/11-5

- Average performance (70%)
- Inconsistencies in labelled lithology
- Incomplete log run



Results – Well 17/4-1

- Least performing well
- Accuracy in lithology labelling?



Conclusion

- Improvement of FORCE models achieving higher prediction of lithologies
- Wrongly Labelled lithologies resulted in reduced model performance
- ML can effectively improve geological workflow by significantly
  - saving time & cost to carryout routine task
  - Eliminating human errors and sentiments
- Quality of data feed used in training model affects the output of the model
  - Quality over quantity
  - Complete vs incomplete log run
- Domain knowledge (Subject Matter expertise) is important in Artificial Intelligence

Human interpretation

2 hours per well

20 hours 10 wells → Half working week given 40 hours per week

↓

1200 minutes

Machine learning

8 minutes → 75% accuracy

Saves 900 mins(15hrs ~ 2 working days)

- In matured basins like the North Sea there are over 1000 wells, manually interpreting these wells could take a life time, but with ML this can be achieved in a very short time