Using Natural Language Processing to Convert Mud-Log Chip Descriptions to Useful Data Tables

Marcelo Guarido, David J. Emery, and Kristopher A. Innanen

ABSTRACT

We successfully created a natural language processing pipeline to extract mud-logging cutting descriptions from PDF files. We converted them to usable structured numerical tables that can be used to match with wireline logs or seismic sessions. The nature of the original tables required extensive preprocessing of the extracted object, including data manipulation, pattern recognition, missing values treatment, and resample. The extract and processed table were merged with well logs and used to predict DTC and provided important improvement of the predictions compared to the baseline model using wireline logs only, where the R2 improved from 0.73 to o.82 using a linear regression model. Feature selection with the stepwise regression generated an optimized model that kept the quality of the predictions and used logs and cutting descriptions with equal importance. Lately, an XGBoost regressor created a non-linear model to improve the predictions with an R2 of 0.88, relying more on the wireline logs. New tests were done on a train-validation split of 5% and 95% to avoid biased predictions. Both the stepwise and XGBoost regression predictions were less precise but still close to the actual values, showing the robustness of the methodology.

INTRODUCTION

Cutting descriptions are part of the mud-logging analysis (Whittaker, 1990), where samples from the subsurface are examined, and chip descriptions are added to a logging report. These analyses are widely used in the industry by crossing the description with other logging measurements. Mohamed A. El-Dakak et al. (2021) used cutting descriptions to match sand reservoir bodies with the wireline logs. A similar application was made by Sakurai et al. (2002), where the cutting descriptions were used to calibrate lithology models from the wireline logs. Vo Thanh & Lee (2022) used cutting descriptions as one of the tools for the facies and depositional analysis. Although important, the cutting descriptions are inside the mud-logging reports, in PDF format, and are not directly used as data tables but as a reference for the petrophysicists.

Tables can be extracted from PDF files and exported to data tables using *Natural Language Processing* (NLP), a field to identify words from texts and audio. Some applications include the use of NLP to identify severe injuries from HSE reports (Guarido & Trad, 2019), sentiment analysis to improve human-robot interactions (Atzeni & Reforgiato Recupero, 2020), text classification from extensive engineering reports in PDF for design and development (Abdoun & Chami, 2022), and converting tables from PDF to HTML files by using image classification techniques (Zhong, ShafieiBavani, & Jimeno Yepes, 2020).

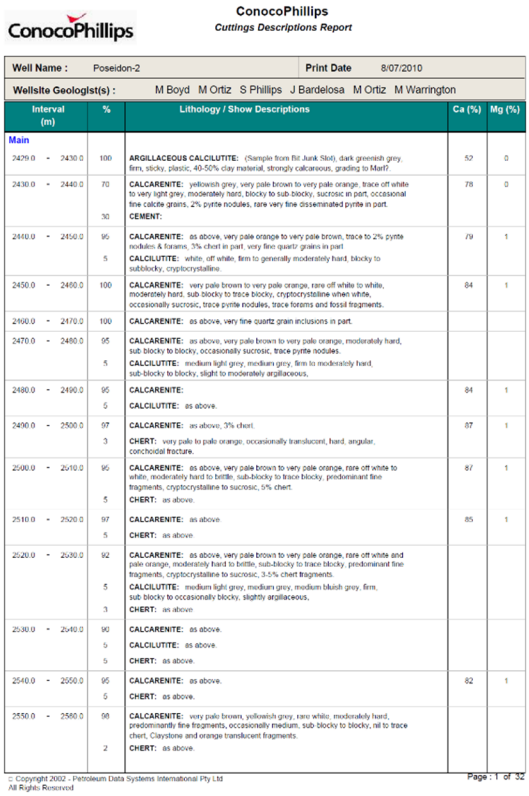
There are different tools to extract tables from PDF files, such as the *Tabulizer* library in R (Leeper, 2018). They work well on organized tables but start to fail on more

complex and unstructured ones, and preprocessing the data is required to convert more complex tables to usable ones.

In this report, we will automatically extract cutting descriptions from mud-logging reports from the ConocoPhillips Poseidon survey in Australia using, as a starting point, the Tabulizer package and NLP techniques to localize the names of the minerals and their quantity in the specific cut as well as oversampling the data to match the depth sample rating of wireline logs. The final goal is to create an application where users can automatically upload their mud-logging reports and export the cutting descriptions tables. Also, as a proof of concept (PoC), we will use the extracted cutting descriptions and logs that are commonly used on log-while-drilling (LWD), such as gamma-ray and resistivity, to generate synthetic sonic logs, using linear regression for a deeper analysis, and the XGBoost (Chen & Guestrin, 2016) for more precise predictions.

CUTTING DESCRIPTIONS REPORTS

Cutting descriptions are part of the mud-logging report of drilled wells worldwide. These reports are standard and widely used to match the interpretation of reservoir characterization by petrophysicists. However, such data is usually in PDF files, and they are used as supporting information but not matched directly with wireline logs and seismic sessions. We propose to convert such tables from PDF to CSV files automatically. Extracting tables from PDF files can be arduous, as some can be highly unstructured, and exhaustive preprocessing is required.



ConocoPhillips

Cuttings Descriptions Report

Well Name :	Poseidon-2	**Print Date**	8/07/2010
Wellsite Geologist(s) :	M Boyd M Ortiz S Phillips J Bardelosa M Ortiz M Warrington		

Interval (m)	%	Lithology / Show Descriptions	Ca (%)	Mg (%)
Main				
2429.0 - 2430.0	100	**ARGILLACEOUS CALCILUTITE:** (Sample from Rit Junk Slot), dark greenish grey, firm, sticky, plastic, 40-50% clay material, strongly calcareous, grading to Marl?	52	0
2430.0 - 2440.0	70	**CALCARENITE:** yellowish grey, very pale brown to very pale orange, trace off white to very light grey, moderately hard, blocky to sub-blocky, sucrosic in part, occasional fine calcite grains, 2% pyrite nodules, rare very fine disseminated pyrite in part.	78	0
	30	**CEMENT:**		
2440.0 - 2450.0	95	**CALCARENITE:** as above, very pale orange to very pale brown, trace to 2% pyrite nodules & forams, 3% chert in part, very fine quartz grains in part.	79	1
	5	**CALCILUTITE:** white, off white, firm to generally moderately hard, blocky to subblocky, cryptocrystalline.		
2450.0 - 2460.0	100	**CALCARENITE:** very pale brown to very pale orange, rare off white to white, moderately hard, sub blocky to trace blocky, cryptocrystalline when white, occasionally sucrosic, trace pyrite nodules, trace forams and fossil fragments.	84	1
2460.0 - 2470.0	100	**CALCARENITE:** as above, very fine quartz grain inclusions in part.		
2470.0 - 2480.0	95	**CALCARENITE:** as above, very pale brown to very pale orange, moderately hard, sub blocky to blocky, occasionally sucrosic, trace pyrite nodules.		
	5	**CALCILUTITE:** medium light grey, medium grey, firm to moderately hard, sub-blocky to blocky, slight to moderately argillaceous.		
2480.0 - 2490.0	95	**CALCARENITE:**	84	1
	5	**CALCILUTITE:** as above.		
2490.0 - 2500.0	97	**CALCARENITE:** as above, 3% chert.	87	1
	3	**CHERT:** very pale to pale orange, occasionally translucent, hard, angular, conchoidal fracture.		
2500.0 - 2510.0	95	**CALCARENITE:** as above, very pale brown to very pale orange, rare off white to white, moderately hard to brittle, sub-blocky to trace blocky, predominant fine fragments, cryptocrystalline to sucrosic, 5% chert.	87	1
	5	**CHERT:** as above.		
2510.0 - 2520.0	97	**CALCARENITE:** as above.	85	1
	3	**CHERT:** as above.		
2520.0 - 2530.0	92	**CALCARENITE:** as above, very pale brown to very pale orange, rare off white and pale orange, moderately hard to brittle, sub-blocky to trace blocky, predominant fine fragments, cryptocrystalline to sucrosic, 3-5% chert fragments.		
	5	**CALCILUTITE:** medium light grey, medium grey, medium bluish grey, firm, sub blocky to occasionally blocky, slightly argillaceous.		
	3	**CHERT:** as above.		
2530.0 - 2540.0	90	**CALCARENITE:** as above.		
	5	**CALCILUTITE:** as above.		
	5	**CHERT:** as above.		
2540.0 - 2550.0	95	**CALCARENITE:** as above.	82	1
	5	**CHERT:** as above.		
2550.0 - 2560.0	98	**CALCARENITE:** very pale brown, yellowish grey, rare white, moderately hard, predominantly fine fragments, occasionally medium, sub-blocky to blocky, nil to trace chert, Claystone and orange translucent fragments.		
	2	**CHERT:** as above.		

FIG. 1. Cutting description from the Poseidon-2 well in Australia.

Error! Reference source not found. is the first page of the cutting descriptions of the mud-logging from Poseidon-2 by ConocoPhillips. It is a perfect example of an unstructured table, and it is used in this report. It contains two rows of headers on the first page only, 1 row with the names of the columns, and five rows of data: the depth interval, the percentages of the listed minerals in the lithology descriptions column, the rate of Ca, and the ratio of Mg.

We use the *Tabulizer* package in R to convert the PDF table to a data frame. Due to the unstructured origin of the table, the extracted table is far from usable. A whole process of data regularization, with the use of *Natural Language Processing* (NLP), is required to create a table that can be matched with wireline logs.

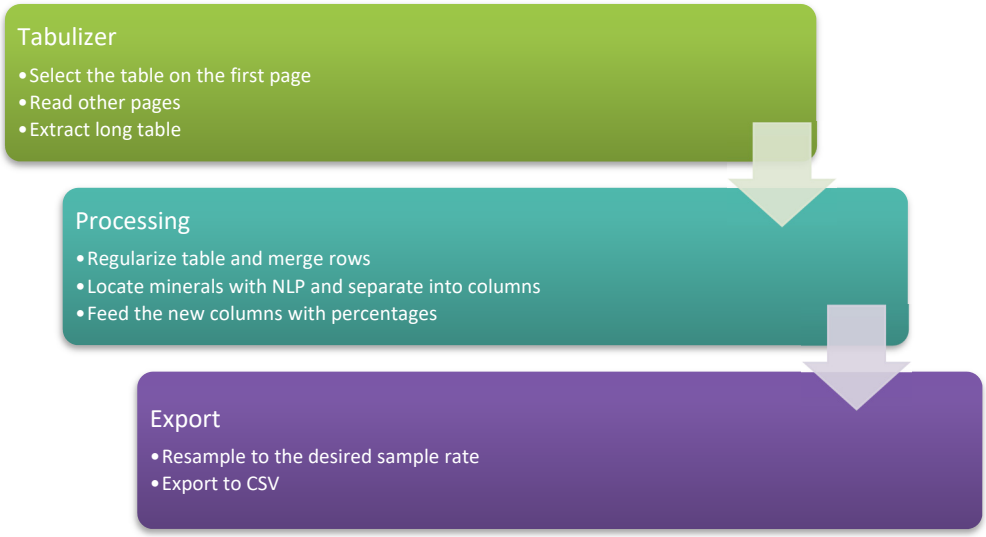
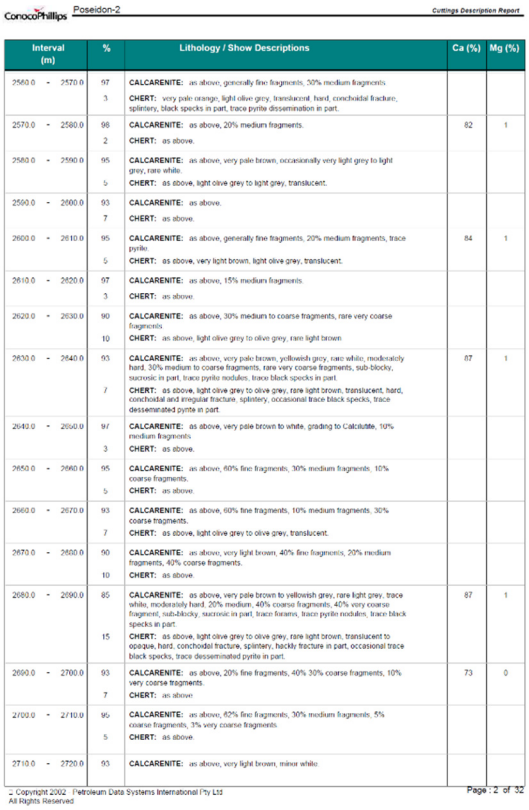


FIG. 2: PDF to CSV process.

FIG. 2. presents the process of reading a table in a PDF file and convert to a CSV one. Each step is detailed below:

- The first step is to manually select the table area of the first page (in a pop-up window), as it contains a table header.
- All other pages are read automatically, as they do not contain the table header, as the second page on FIG. 3.
- Page header, footnote, and page number are automatically removed.
- Each page is automatically preprocessed, separated, and concatenated in the final steps.
- As text descriptions are long, the extracted table divides descriptions into different rows so they are merged.

- Use NLP to recognize the name of minerals in the descriptions column, as the words are capitalized and followed by a colon and create a column for each character.
- Populate the new columns with the information from the "%" column.
- Resample the table to the desired sample rate.
- Export to CSV.



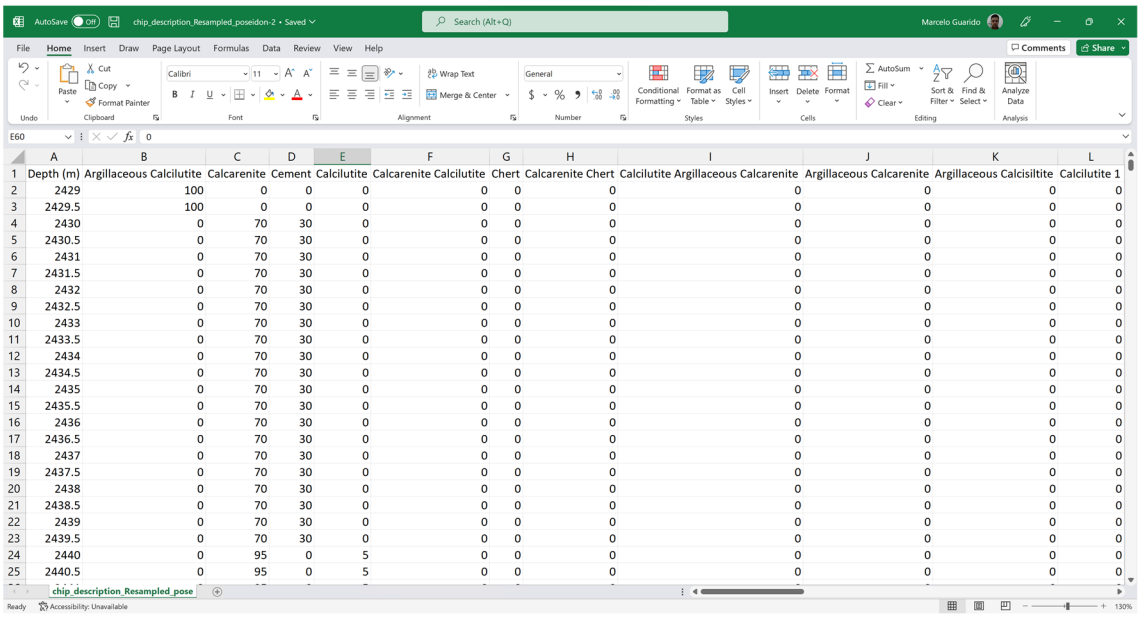
ConocoPhillips Poseidon-2 Cuttings Description Report

Interval (m)	%	Lithology / Show Descriptions	Ca (%)	Mg (%)
2560.0 - 2570.0	97	**CALCARENITE:** as above, generally fine fragments, 30% medium fragments.		
	3	**CHERT:** very pale orange, light olive grey, translucent, hard, conchoidal fracture, splintery, black specks in part, trace pyrite dissemination in part.		
2570.0 - 2580.0	98	**CALCARENITE:** as above, 20% medium fragments.	82	1
	2	**CHERT:** as above.		
2580.0 - 2590.0	95	**CALCARENITE:** as above, very pale brown, occasionally very light grey to light grey, rare white.		
	5	**CHERT:** as above, light olive grey to light grey, translucent.		
2590.0 - 2600.0	93	**CALCARENITE:** as above.		
	7	**CHERT:** as above.		
2600.0 - 2610.0	95	**CALCARENITE:** as above, generally fine fragments, 20% medium fragments, trace pyrite.	84	1
	5	**CHERT:** as above, very light brown, light olive grey, translucent.		
2610.0 - 2620.0	97	**CALCARENITE:** as above, 15% medium fragments.		
	3	**CHERT:** as above.		
2620.0 - 2630.0	90	**CALCARENITE:** as above, 30% medium to coarse fragments, rare very coarse fragments.		
	10	**CHERT:** as above, light olive grey to olive grey, rare light brown.		
2630.0 - 2640.0	93	**CALCARENITE:** as above, very pale brown, yellowish grey, rare white, moderately hard, 30% medium to coarse fragments, rare very coarse fragments, sub-blocky, sucrosic in part, trace pyrite nodules, trace black specks in part.	87	1
	7	**CHERT:** as above, light olive grey to olive grey, rare light brown, translucent, hard, conchoidal and irregular fracture, splintery, occasional trace black specks, trace disseminated pyrite in part.		
2640.0 - 2650.0	97	**CALCARENITE:** as above, very pale brown to white, grading to Calcilutite, 10% medium fragments.		
	3	**CHERT:** as above.		
2650.0 - 2660.0	95	**CALCARENITE:** as above, 60% fine fragments, 30% medium fragments, 10% coarse fragments.		
	5	**CHERT:** as above.		
2660.0 - 2670.0	93	**CALCARENITE:** as above, 60% fine fragments, 10% medium fragments, 30% coarse fragments.		
	7	**CHERT:** as above, light olive grey to olive grey, translucent.		
2670.0 - 2680.0	90	**CALCARENITE:** as above, very light brown, 40% fine fragments, 20% medium fragments, 40% coarse fragments.		
	10	**CHERT:** as above.		
2680.0 - 2690.0	85	**CALCARENITE:** as above, very pale brown to yellowish grey, rare light grey, trace white, moderately hard, 20% medium, 40% coarse fragments, 40% very coarse fragment, sub-blocky, sucrosic in part, trace foram, trace pyrite nodules, trace black specks in part.	87	1
	15	**CHERT:** as above, light olive grey to olive grey, rare light brown, translucent to opaque, hard, conchoidal fracture, splintery, hackly fracture in part, occasional trace black specks, trace disseminated pyrite in part.		
2690.0 - 2700.0	93	**CALCARENITE:** as above, 20% fine fragments, 40% 30% coarse fragments, 10% very coarse fragments.	73	0
	7	**CHERT:** as above.		
2700.0 - 2710.0	95	**CALCARENITE:** as above, 60% fine fragments, 30% medium fragments, 5% coarse fragments, 3% very coarse fragments.		
	5	**CHERT:** as above.		
2710.0 - 2720.0	93	**CALCARENITE:** as above, very light brown, minor white.		

FIG. 3. Second page of the cutting descriptions table.

All these steps work because the patterns are the same for all the pages, and an example of part of the final table can be seen in FIG. 4. We can observe some of the new columns created with the percentage values included. The number of new columns equals the number of different minerals listed in the descriptions. However, this method would work only for the ConocoPhillips table template. To design the web application, we need to include other templates; the user can select one.

With the table extracted and exported to a resampled CSV format, we can merge it to wireline logs from LAS files and use the percentages of the mineral as logs for different analyzes, such as facies classification, or create synthetic traces, for example.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Depth (m)	Argillaceous Calcilutite	Calcarenite	Cement	Calcilutite	Calcarenite Calcilutite	Chert	Calcarenite Chert	Calcilutite Argillaceous Calcarenite	Argillaceous Calcarenite	Argillaceous Calcisiltite	Calcilutite 1
2	2429	100	0	0	0	0	0	0	0	0	0	0
3	2429.5	100	0	0	0	0	0	0	0	0	0	0
4	2430	0	70	30	0	0	0	0	0	0	0	0
5	2430.5	0	70	30	0	0	0	0	0	0	0	0
6	2431	0	70	30	0	0	0	0	0	0	0	0
7	2431.5	0	70	30	0	0	0	0	0	0	0	0
8	2432	0	70	30	0	0	0	0	0	0	0	0
9	2432.5	0	70	30	0	0	0	0	0	0	0	0
10	2433	0	70	30	0	0	0	0	0	0	0	0
11	2433.5	0	70	30	0	0	0	0	0	0	0	0
12	2434	0	70	30	0	0	0	0	0	0	0	0
13	2434.5	0	70	30	0	0	0	0	0	0	0	0
14	2435	0	70	30	0	0	0	0	0	0	0	0
15	2435.5	0	70	30	0	0	0	0	0	0	0	0
16	2436	0	70	30	0	0	0	0	0	0	0	0
17	2436.5	0	70	30	0	0	0	0	0	0	0	0
18	2437	0	70	30	0	0	0	0	0	0	0	0
19	2437.5	0	70	30	0	0	0	0	0	0	0	0
20	2438	0	70	30	0	0	0	0	0	0	0	0
21	2438.5	0	70	30	0	0	0	0	0	0	0	0
22	2439	0	70	30	0	0	0	0	0	0	0	0
23	2439.5	0	70	30	0	0	0	0	0	0	0	0
24	2440	0	95	0	5	0	0	0	0	0	0	0
25	2440.5	0	95	0	5	0	0	0	0	0	0	0

FIG. 4. Extracted table from Poseidon-2 cutting descriptions.

USING CUTTING DESCRIPTIONS WITH WIRELINE LOGS

Extracting the cutting descriptions from a PDF file and converting them to a table is only part of the job: we need to test if the new logs can improve petrophysical analysis. As *proof of concept* (PoC), a sonic log (DTC) will be predicted with and without the cutting descriptions for the Poseidon-2 borehole.

The previous section extracted cutting descriptions from mud-logging PDF files and converted them to a numerical table for the Poseidon-2 well. Now, this table can be merged with LWD and wireline LAS files to include more information in the analysis, and FIG. 5 shows the merging outcome. The cutting descriptions table was resampled to the same sample rate as the wireline logs of 50cm. There are, after merging, 8 well logs, 26 mineral rates, and 2 percentage variables (Ca and Mg) as data. From the well logs, only 3 logs are vastly available for the depth interval: GR, RDEP, and DTC. Gamma-ray and resistivity are logs that are also available in several LWD surveys. So, the proposal is to predict synthetic DTC using GR and RDEP only, include the cutting descriptions, and evaluate the effectiveness of the cutting descriptions as log data.

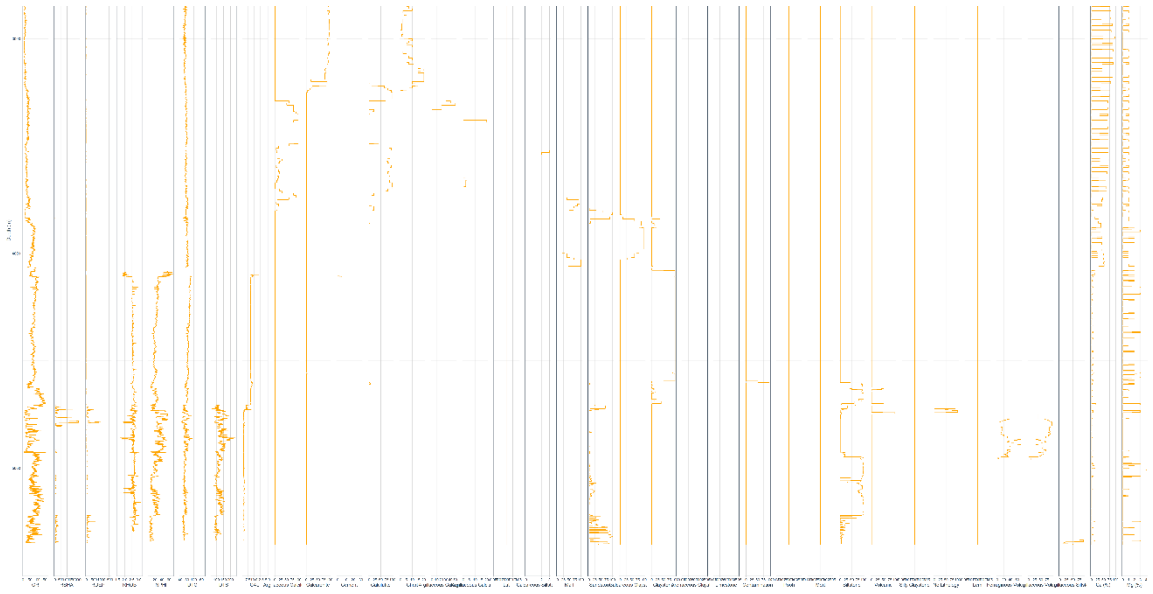


FIG. 5. Wireline logs and the extracted chip descriptions merged.

Baseline Model

In the first test, the data (with 5684 observations) were randomly split into 40% for training (2273 observations) and 60% for validation (3411 observations), and the data were standardized to have all variables at the same scale. This is a splitting sin, as the rows are correlated to the rows above and below, but the idea is to show how the cutting descriptions can improve the predictions. As we are predicting DTC, this is a regression problem.

The baseline model is a linear regression trained using only GR, RDEP, and Depth to predict DTC. FIG. 6 shows the training results. Assuming a significance level of 5% ($P < 0.05$), only GR and RDEP were statistically significant in predicting the target (p-value smaller than 0.05). The adjusted R^2 is 0.73, a reliable metric number, showing that using only these two logs.

```
Call:
lm(formula = DTC_Detrended ~ GR + RDEP + `Depth (m)`, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-27.379   -3.160   -0.161    3.037   33.115

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.11421    0.11001  -1.038    0.299
GR            1.91385    0.12934  14.797   <2e-16 ***
RDEP         -7.62717    0.12629 -60.395   <2e-16 ***
`Depth (m)`  -0.07992    0.10949  -0.730    0.466
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.244 on 2269 degrees of freedom
Multiple R-squared:  0.7318,    Adjusted R-squared:  0.7315
F-statistic:  2064 on 3 and 2269 DF,  p-value: < 2.2e-16
```

FIG. 6. Baseline linear regression model using only RDEP, GR, and Depth to predict DTC.

Predictions on the validation data are in FIG. 7. The R^2 is 0.73, similar to the training value. Visually the predictions (in orange) closely follow the actual values (in black), with some difference at depths 3500m, 4100m, and 4750m, showing that we can get great DTC values only from a handful of logs.

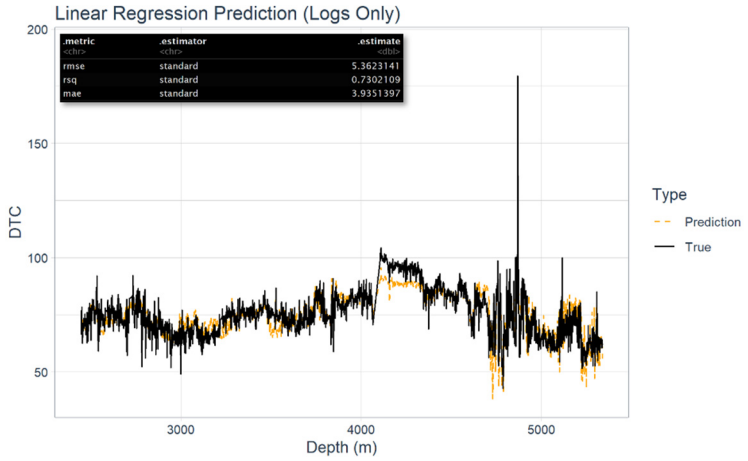


FIG. 7. Predictions using only logs with a 40/60 split.

Linear Regression

The next step is to include the cutting descriptions as new features for the linear regression, and FIG. 8 shows its summary. Not all cutting descriptions are statistically significant to predict the target, as their p-value are larger than 0.05, but several are. With the new R^2 of 0.82, we can already see that including the cutting descriptions is helping the linear model to reach more accurate predictions.

```
Call:
lm(formula = DTC_Detrended ~ . - Depth - Trend_DTC - DTC, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-24.2906  -2.3174  -0.0601   2.4379  22.1395

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        -0.07621    0.09041  -0.843 0.399360
`Depth (m)`                        -5.10753    0.43234 -11.814  < 2e-16 ***
`Argillaceous Calcilutite`         -0.28764    0.42570  -0.676 0.499306
Calcarenite                        -2.53034    0.75655  -3.345 0.000838 ***
Cement                              0.34663    0.08092   4.284 1.92e-05 ***
Calcilutite                         0.06983    0.44667   0.156 0.875787
`Calcarenite Calcilutite`          -0.47073    0.14862  -3.167 0.001560 **
Chert                              -0.44993    0.16017  -2.809 0.005010 **
`Calcarenite Chert`                -0.40398    0.15794  -2.558 0.010602 *
`Calcilutite Argillaceous Calcarenite` -0.05612  0.14630  -0.384 0.701306
`Argillaceous Calcarenite`         -0.27558    0.12137  -2.271 0.023270 *
`Argillaceous Calcisiltite`        -0.12594    0.08487  -1.484 0.137975
`Calcilutite 1`                     0.19880    0.11180   1.778 0.075495 .
`Calcareous Siltstone`             -0.13957    0.09888  -1.412 0.158227
`Argillaceous Calcilutite 1`        0.34530    0.09703   3.559 0.000381 ***
Marl                                0.68583    0.26453   2.593 0.009587 **
Sandstone                           0.24984    0.27965   0.893 0.371752
`Marl Calcilutite`                 -0.12855    0.13569  -0.947 0.343524
`Calcareous Claystone`             -0.08277    0.35293  -0.235 0.814598
Claystone                           2.97673    0.60937   4.885 1.11e-06 ***
`Claystone 1`                      -0.20035    0.17072  -1.174 0.240709
Contamination                       0.29400    0.11964   2.457 0.014075 *
Siltstone                          -0.24599    0.51090  -0.481 0.630218
Volcanic                            0.38974    0.12842   3.035 0.002434 **
`No Lithology`                      0.30034    0.13603   2.208 0.027349 *
`Ferruginous Volcanic`              1.10406    0.20897   5.283 1.39e-07 ***
`Argillaceous Volcanic`             0.30196    0.30407   0.993 0.320785
`Siltstone 1`                       0.08152    0.11029   0.739 0.459887
`Siltstone  Siltstone 1`            0.14090    0.06590   2.138 0.032627 *
`Argillaceous Siltstone`            0.31283    0.13287   2.354 0.018638 *
`Ca (%)`                           -1.68043    0.67561  -2.487 0.012945 *
GR                                  2.27713    0.13907  16.374  < 2e-16 ***
RDEP                               -4.72598    0.15345 -30.799  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.281 on 2240 degrees of freedom
Multiple R-squared:  0.8235,    Adjusted R-squared:  0.821
F-statistic: 326.7 on 32 and 2240 DF,  p-value: < 2.2e-16
```

FIG. 8. Linear regression model summary using logs and cutting descriptions.

Applying the new model to the validation set generated the predictions in FIG. 9. The new R^2 is 0.81 for the validation set, close to the training one, suggesting there is no overfitting, and the predictions (in orange) are closer to the actual values (in black). This

shows that using the extracted cutting descriptions from the PDF file of mud-logging help improve the sonic log (DTC) estimation.

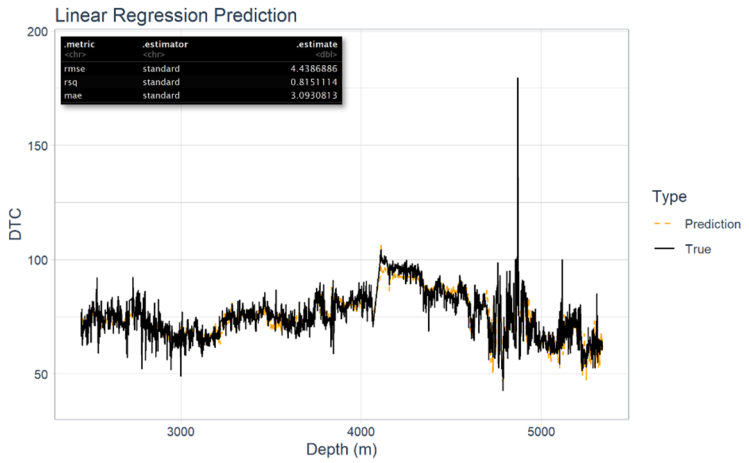


FIG. 9. Predictions with the new model using all the cutting descriptions variables.

Stepwise Regression

From FIG. 8, many input features are not statistically significant to predict DTC. They can be automatically removed using *stepwise regression* (Johnsson, 1992), which adds or subtracts a variable at each step to improve the *Akaike Information Criterion* (AIC). The AIC is a metric used to determine the best model of a set of models created in the same set of observations and is suited for model selection (Bozdogan, 1987). FIG. 10 shows the stepwise regression and how AIC decreases with adding a feature to the model, indicating an improved model. The summary only shows the results for significant features for a linear regression model.

Step <S3: AsIs>	Df <dbl>	Deviance <dbl>	Resid. Df <dbl>	Resid. Dev <dbl>	AIC <dbl>
	NA	NA	2272	232626.37	10522.201
+ RDEP	-1	164201.15914	2271	68425.21	7742.748
+ Claystone	-1	11905.59089	2270	56519.62	7310.256
+ GR	-1	3638.88705	2269	52880.73	7160.991
+ Siltstone	-1	4409.50446	2268	48471.22	6965.084
+ Chert	-1	1948.92161	2267	46522.30	6873.804
+ Calcilutite	-1	433.89126	2266	46088.41	6854.505
+ Marl	-1	411.73326	2265	45676.68	6836.108
+ Cement	-1	404.94381	2264	45271.73	6817.867
+ `Ferruginous Volcanic`	-1	407.18323	2263	44864.55	6799.330
+ `Depth (m)`	-1	370.41554	2262	44494.14	6782.486
+ `Ca (%)`	-1	1190.78662	2261	43303.35	6722.825
+ Calcarenite	-1	814.37080	2260	42488.98	6681.672
+ `Argillaceous Calcilutite 1`	-1	222.65322	2259	42266.33	6671.729
+ Volcanic	-1	179.73710	2258	42086.59	6664.043
+ `Calcarenite Calcilutite`	-1	173.82514	2257	41912.76	6656.635
+ Contamination	-1	107.77090	2256	41804.99	6652.783
+ `Calcarenite Chert`	-1	110.49629	2255	41694.50	6648.767
+ `No Lithology`	-1	99.43980	2254	41595.06	6645.340
+ `Argillaceous Calcarenite`	-1	96.34674	2253	41498.71	6642.069
+ `Argillaceous Siltstone`	-1	82.81256	2252	41415.90	6639.529
+ `Siltstone Siltstone 1`	-1	85.12202	2251	41330.77	6636.852
+ `Calcilutite 1`	-1	82.97880	2250	41247.80	6634.284
+ `Argillaceous Calcisiltite`	-1	39.81617	2249	41207.98	6634.089

FIG. 10. Feature selection using stepwise regression.

The stepwise regression results are in FIG. 11. The model has a similar R^2 as using all the features, but some are not statistically significant. The stepwise regression uses the

AIC metric to determine the best model, while the linear regression with the selected features uses another statistical method to define variable significance.

```
Call:
lm(formula = DTC_Detrended ~ RDEP + Claystone + GR + Siltstone +
    Chert + Calcilutite + Marl + Cement + `Ferruginous Volcanic` +
    `Depth (m)` + `Ca (%)` + Calcarenite + `Argillaceous Calcilutite 1` +
    Volcanic + `Calcarenite Calcilutite` + Contamination + `Calcarenite Chert` +
    `No Lithology` + `Argillaceous Calcarenite` + `Argillaceous Siltstone` +
    `Siltstone  Siltstone 1` + `Calcilutite 1` + `Argillaceous Calcisiltite`,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-23.7128  -2.3492  -0.0624   2.4954  22.0296

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.06767    0.09007  -0.751 0.452514
RDEP                           -4.63892    0.14619 -31.733  < 2e-16 ***
Claystone                       2.64004    0.18929  13.947  < 2e-16 ***
GR                              2.26257    0.13164  17.188  < 2e-16 ***
Siltstone                      -0.61817    0.17092  -3.617 0.000305 ***
Chert                          -0.35745    0.12527  -2.854 0.004363 **
Calcilutite                     0.29630    0.12718   2.330 0.019913 *
Marl                            0.73528    0.10253   7.171 1.00e-12 ***
Cement                          0.35096    0.08081   4.343 1.47e-05 ***
`Ferruginous Volcanic`          1.05934    0.13837   7.656 2.83e-14 ***
`Depth (m)`                    -4.89432    0.40638 -12.044  < 2e-16 ***
`Ca (%)`                       -2.69692    0.41170  -6.551 7.07e-11 ***
Calcarenite                    -1.87118    0.22906  -8.169 5.13e-16 ***
`Argillaceous Calcilutite 1`    0.28086    0.08208   3.422 0.000633 ***
Volcanic                        0.31937    0.08867   3.602 0.000323 ***
`Calcarenite Calcilutite`      -0.36713    0.10138  -3.621 0.000300 ***
Contamination                   0.23988    0.08804   2.725 0.006489 **
`Calcarenite Chert`            -0.30700    0.11659  -2.633 0.008517 **
`No Lithology`                  0.23056    0.09067   2.543 0.011064 *
`Argillaceous Calcarenite`     -0.23653    0.10175  -2.325 0.020184 *
`Argillaceous Siltstone`        0.25202    0.11258   2.238 0.025288 *
`Siltstone  Siltstone 1`        0.13833    0.06389   2.165 0.030479 *
`Calcilutite 1`                 0.21857    0.10546   2.073 0.038331 *
`Argillaceous Calcisiltite`    -0.12187    0.08268  -1.474 0.140588
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.281 on 2249 degrees of freedom
Multiple R-squared:  0.8229,    Adjusted R-squared:  0.821
F-statistic: 454.2 on 23 and 2249 DF,  p-value: < 2.2e-16
```

FIG. 11. Summary of the stepwise regression.

Moreover, the predictions on the validation set in FIG. 12 show that with fewer features, we can still have the same level of accuracy in predicting DTC.

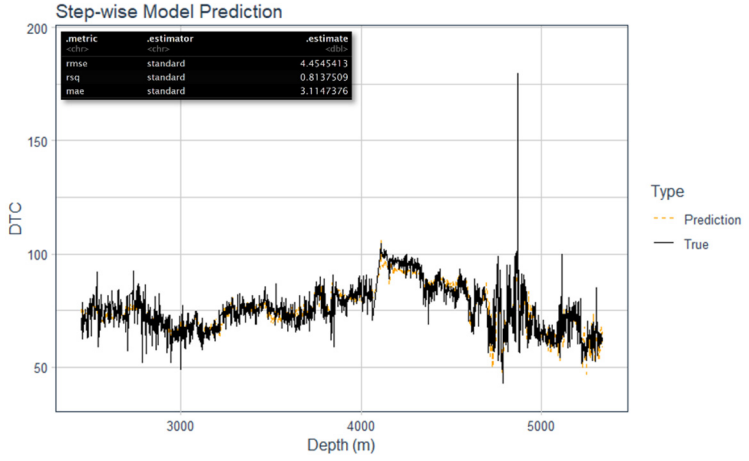


FIG. 12. Stepwise predictions on the validation set with a 40/60 split.

These results indicate that including cutting descriptions in petrophysical analysis can improve their outcome. We have considered a linear correlation between the features and the target. A non-linear regressor model could improve the situation even further in the predictions’ quality.

XGBoost Regressor

A non-linear regressor model may be able to improve the predictions even further. Using the same features selected by the stepwise regression, an XGBoost regressor was trained, and the predictions were estimated on the validation set. FIG. 13 shows predictions (in orange) closely matched with the actual values (in black), with R^2 of 0.88, a good improvement compared to a linear model.

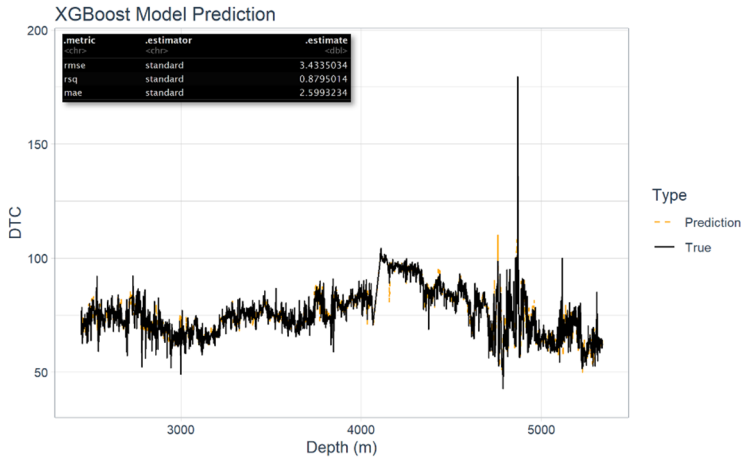


FIG. 13. XGBoost regression on the validation set with a 40/60 split.

Feature importance can be extracted from several statistical and machine learning models. In linear regression, using standardized features, the magnitude of the weights (or parameters) indicates the effect power of the feature. Larger the absolute values of the weights, the more significant the importance of the features of those weights. From FIG. 11, the stepwise regression modelling suggests that the most important features are RDEP, Depth, Ca, and Claystone. *Shapley Values* (Hart, 1989), used in game theory, can be used to understand which feature had the highest contribution for each observation to reach a predicted value. The XGBoost library provides the most important features by measuring the *gain* (their contribution to improving the accuracy in each branch of each tree). FIG. 14 shows that, for the XGBoost regressor, the wireline logs are the most important features, highly dominated by RDEP (also the most important feature in the stepwise regression). The cutting descriptions have more discrete importance, different than the stepwise regression. As a recommendation for the future, use *Shapley Values* in both models for a direct comparison.

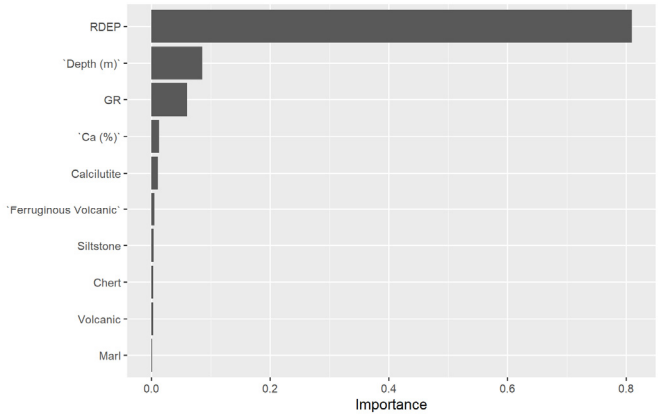


FIG. 14. Top 10 most important features used by the XGBoost regressor in 40/60 split data.

Our results indicate the significance of including cutting descriptions to estimate a synthetic DTC, but as pointed out previously, we are sinning with the random 40/60 train-validation split. The next step is testing a different train-validation ratio.

MODELLING WITH A 5/95 SPLIT

Randomly splitting log data into train-validation is not optimal as an observation is correlated to neighbouring observations. To minimize a biased prediction, we tried our methodology on a new ratio of train-validation split: 5% (284 observations) and 95% (5400 observations), respectively.

Stepwise Regression 2

FIG. 15 shows the summary of the stepwise regression model trained on 5% of the data. The number of statistically significant features has decreased, as most cutting descriptions are absent from the training data (zero for those observations). The weights suggest that the most important features are Depth, Ca, and RDEP.

```
Call:
lm(formula = DTC_Detrended ~ RDEP + Claystone + GR + Volcanic +
    Siltstone + Calcarenite + `Depth (m)` + `Ca (%)` + Marl +
    `Calcareous Claystone` + `No Lithology` + `Argillaceous Calcilutite 1`,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9835  -2.3522  -0.2917   2.4866  22.6639

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.2462     0.2477  -0.994  0.32109
RDEP                        -4.4025     0.4032 -10.919  < 2e-16 ***
Claystone                    1.1320     0.4448   2.545  0.01149 *
GR                           2.1850     0.3610   6.052 4.74e-09 ***
Volcanic                     0.6611     0.1510   4.377 1.72e-05 ***
Siltstone                   -1.3066     0.4190  -3.118  0.00202 **
Calcarenite                 -2.5688     0.4543  -5.654 3.97e-08 ***
`Depth (m)`                 -6.9181     1.1325  -6.109 3.48e-09 ***
`Ca (%)`                    -5.6660     1.0575  -5.358 1.80e-07 ***
Marl                         0.7436     0.2531   2.938  0.00359 **
`Calcareous Claystone`      -0.6084     0.3056  -1.991  0.04747 *
`No Lithology`              -0.2677     0.1534  -1.745  0.08210 .
`Argillaceous Calcilutite 1`  0.2833     0.1748   1.621  0.10618
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.08 on 271 degrees of freedom
Multiple R-squared:  0.8404,    Adjusted R-squared:  0.8334
F-statistic: 118.9 on 12 and 271 DF,  p-value: < 2.2e-16
```

FIG. 15. The new stepwise model with a 5/95 split.

Predictions in FIG. 16 have lost accuracy, with R^2 of 0.79 on the validation set. As the R^2 on the training set is 0.84, minor overfitting is suggested. Visually the predictions (orange) are still closely matched to the actual values (black), indicating the robustness of the model.

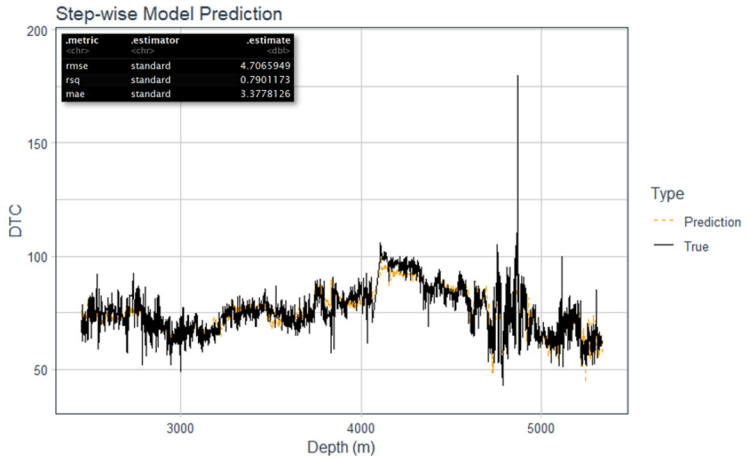


FIG. 16. Stepwise regression predictions on the validation set with a 5/95 split.

XGBoost Regression 2

Training an XGBoost regressor model on 5% of the data using the same variables selected by the stepwise regression model produced the predictions on the validation set in FIG. 17. The R^2 reduced to 0.86 but is still high. The predictions (orange) are closely matched to the actual values (black), only with some spikes showing up around 4000m.

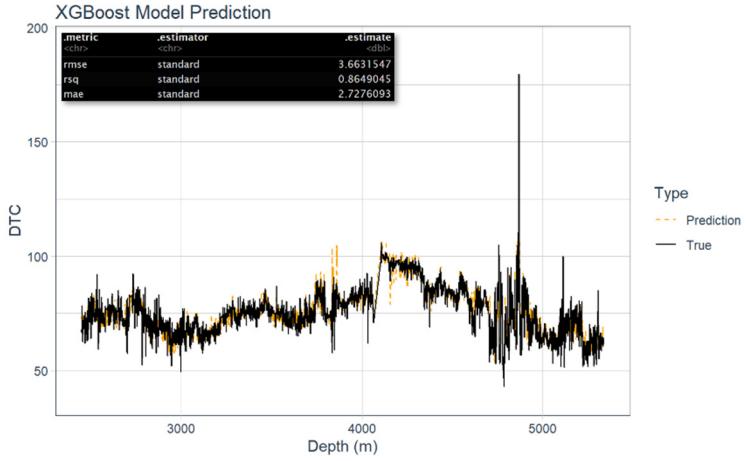


FIG. 17. XGBoost regression predictions on the validation set with a 5/95 split.

Features importance in FIG. 18 are like the ones in FIG. 14, where RDEP, Depth, and GR are the most important features. The wireline logs gain more importance by training the linear and XGBoost models on fewer data.

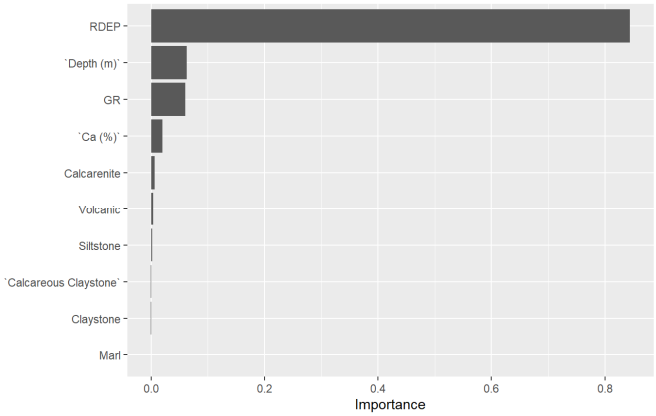


FIG. 18. XGBoost feature importance on the 5/95 split.

CONCLUSIONS

We presented a natural language processing pipeline to successfully extract mud-logging cutting descriptions from PDF files and converted them to usable structured numerical tables that can be used to match with wireline logs or seismic sessions. The nature of the original tables required extensive preprocessing of the extracted object, including data manipulation, pattern recognition, missing values treatment, and resample.

The extract and processed table were merged with well logs and used to predict DTC and provided important improvement of the predictions compared to the baseline model using wireline logs only, where the R^2 improved from 0.73 to o.82 using a linear regression model. Feature selection with the stepwise regression generated an optimized model that kept the quality of the predictions and used logs and cutting descriptions with equal importance. Lately, an XGBoost regressor created a non-linear model to improve the predictions with an R^2 of 0.88, relying more on the wireline logs.

New tests were done on a train-validation split of 5% and 95% to avoid biased predictions. Both the stepwise and XGBoost regression predictions were less precise but still close to the actual values, showing the robustness of the methodology.

ACKNOWLEDGEMENTS

We thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 543578-19. The 1st author was supported by the Canada First Research Excellence Fund, through the Global Research Initiative at the University of Calgary. We thank Soane Mota dos Santos for the suggestions, tips and productive discussions.

REFERENCES

Abdoun, N., & Chami, M. (2022). Automatic Text Classification of PDF Documents using NLP Techniques. *INCOSE International Symposium, 32*(1), 1320-1331.

Atzeni, M., & Reforgiato Recupero, D. (2020). Multi-domain sentiment analysis with mimicked and polarized word embeddings for human-robot interaction. *Future Generation Computer Systems, 110*, 984-999.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

El-Dakak, M. A., Abdelfattah, T. A., Diab, A. I., Kassem, M. A., & Knapp, C. C. (2021). Integration of borehole depth imaging and seismic reflection results in reservoir delineation: An example from The Alam El Bueib 3C field, Northern Western Desert, Egypt. *Journal of African Earth Sciences, 184*, 1464-343X.

Guarido, M., & Trad, D. O. (2019). Using natural language processing and machine learning to predict severe injuries classification in the oil and gas industry. *CREWES Research Report, 31*.

Hart, S. (1989). Shapley Value. In J. Eatwell, M. Milgate, & P. Newman, *Game Theory* (pp. 210–216). London: Palgrave Macmillan UK.

Johnsson, T. (1992). A procedure for stepwise regression analysis. *Statistical Papers, 33*, 21–29.

Leeper, T. J. (2018). Tabulizer: Bindings for Tabula PDF Table Extractor Library. *R package version 0.2.2*.

Rodriguez, L., Chiapello, E., Lambert, L., Leduc, J. P., M. L., & Sanchez, M. J. (2015). Quantitative and Comparative Evaluation of Mineralogy and TOC Analysis from Cores, Cuttings and Logs in Vaca Muerta Unconventional Shale Play. *SPE/AAPG/SEG Unconventional Resources Technology Conference*.

Sakurai, S., Grimaldo-Suarez, F. M., Aguilera-Gomez, L. E., & Rodriguez-Larios, J. A. (2002). Estimate of Lithology and Net Gas Sand from Wireline Logs: Veracruz and Macuspana Basins, Mexico. *Gulf Coast Association of Geological Societies Transactions, 52*, 871-881.

Vo Thanh, H., & Lee, K. (2022). 3D geo-cellular modeling for Oligocene reservoirs: a marginal field in offshore Vietnam. *Journal of Petroleum Exploration and Production Technology, 12*, 1–19.

Whittaker, A. (1990). *Mud logging handbook.* United States.

Zhong, X., ShafieiBavani, E., & Jimeno Yepes, A. (2020). Image-Based Table Recognition: Data, Model, and Evaluation. In A. Vedaldi, H. Bischof, T. Brox, & J. Frahm, *Computer Vision -- ECCV 2020* (pp. 564-580). Springer International Publishing.