geoconvention
Calgary • Canada • May 15-17 2023

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

Marcelo Guarido, David J. Emery, and Kristopher A. Innanen
CREWES – University of Calgary

Summary

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 wireline 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 0.82 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.

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 work, 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: PDF to Table

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.

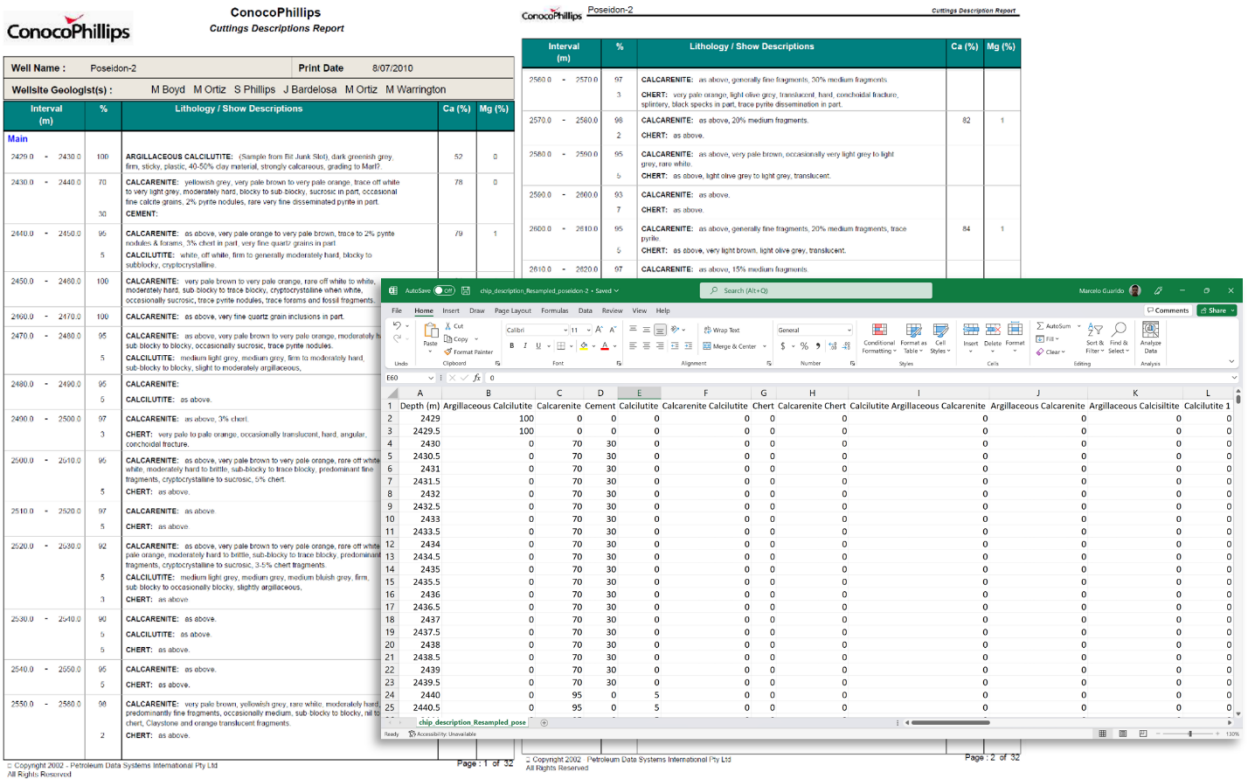


Figure 1: Cutting description from the Poseidon-2 well in Australia converted to a table.

Figure 1 is the result of our PDF-to-table routine on the Poseidon-2 mudlogging in Australia. Cutting descriptions are detected, extracted, and reorganized using NLP, and the generated table was resampled to the same depth sample rate of the wireline logs of the same well.

Predicting DTC using Wireline and Cutting Descriptions 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.

A baseline model using linear regression was created by randomly splitting the data (with 5684 observations) 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 model was trained using only GR, RDEP, and Depth to predict DTC. Figure 2 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, we have a robust DTC prediction.

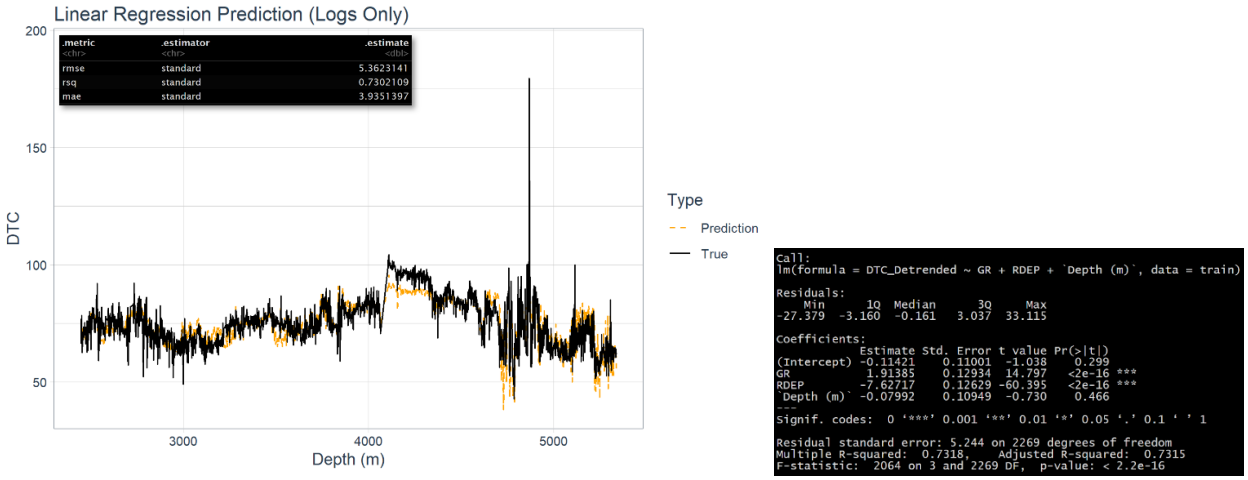


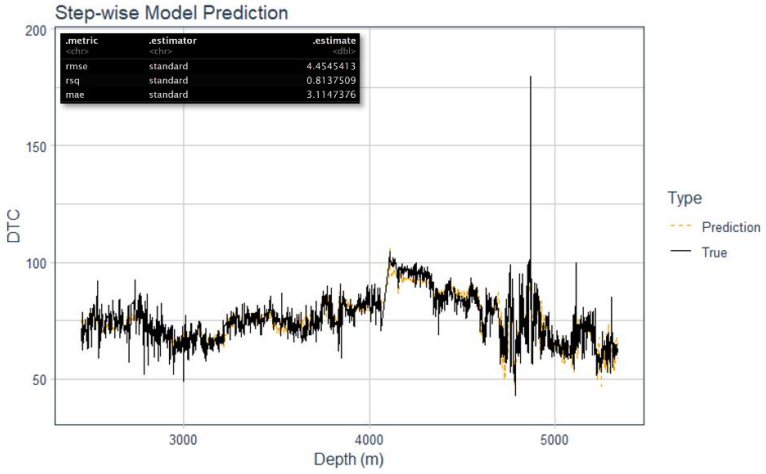
Figure 2: Baseline linear regression model using only RDEP, GR, and Depth to predict DTC.

The next step was to train a linear model using all the newly extracted cutting descriptions as features. However, at first, most of the logs were statistically insignificant in predicting DTC. We used a *stepwise regression* (Johnsson, 1992) to select only important features that 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). Figure 3 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

Figure 3: Feature selection using stepwise regression.

Applying the new model to the validation set generated the predictions in Figure 4. 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.




```
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`      0.26086    0.08708   3.472 0.000653 ***
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    0.11238   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
```

Figure 4: Stepwise regression prediction of DTC.

At last, we used a non-linear model (XGBoost regressor) with the same features as the stepwise one, as we expected that the relationship between wireline logs and cutting description to DTC is non-linear. Figure 5 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.

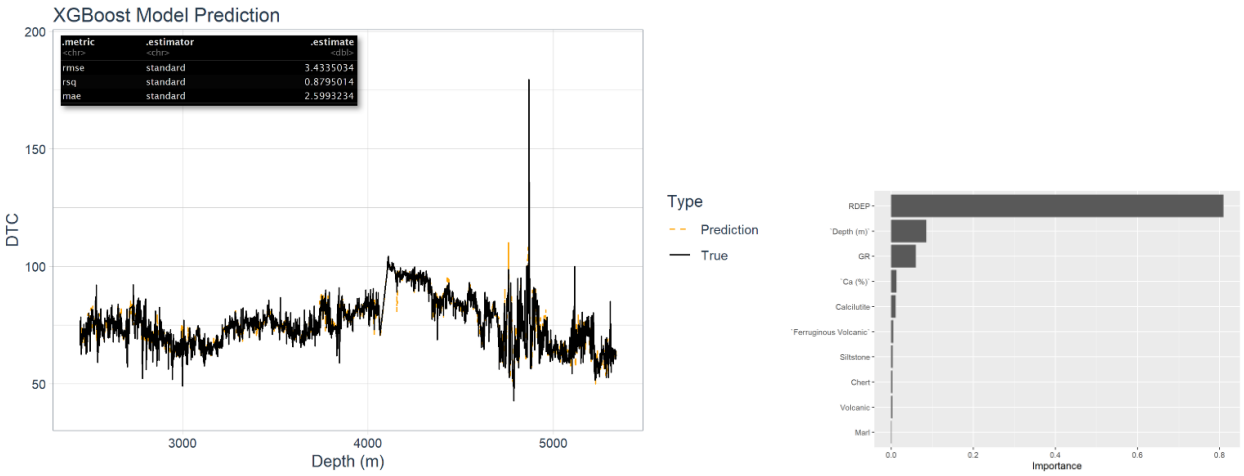


Figure 5: XGBoost predictions and most important features.

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 Figure 4, 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). Figure 5 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.

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 wireline 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 0.82 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.

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*.

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.