

**A logo on a grid

AI-generated content may be incorrect.**

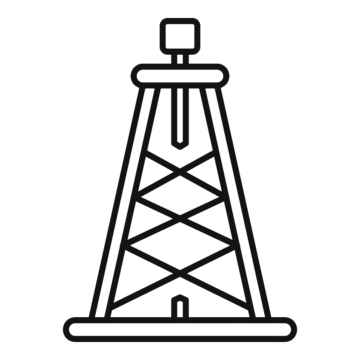
Katherine Grey

Bellevue University

DSC 680 Project 2

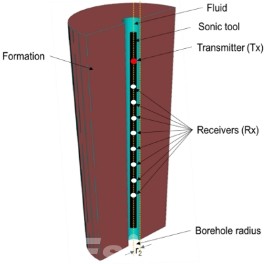
**Sonic Log Predictions:**

**Using XGBoost and Wavelet Transformation**



Date of Publication

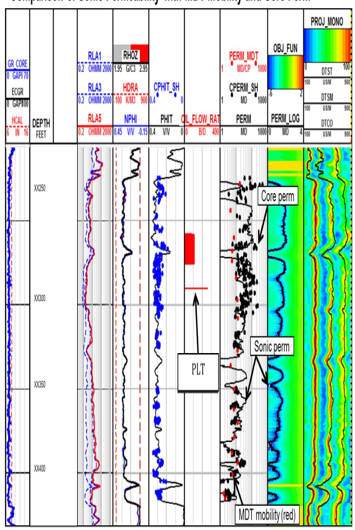
**May 4, 2025**



Business Problem

The importance of accurate sonic log data in the drilling and extraction of oil fields are vitally important. This information directs companies on the most timely, cost-effective way to maximize oil production. The problem is these logs are costly to produce, often in the tens of thousands of dollars for each bore, and given the complexity some key values may be missing. Given the impact inaccurate or missing information can be a new method utilizing the abilities of machine learning to interpret and find relationships between non-linear data is vital. This project aims to find a machine learning model combined with hypertooned feature engineering to bridge this gap.

Back Ground/ History

Sonic logs are vital for determining important details such porosity, permeability, lilthology, elastic properties and other features of reservoirs. This information can reveal key details that aid in the design and optimization of drill into and extraction of oil reservoirs. While this can be used in other boring operations this project specifically is looking at data from oil and gas well bores. Generating sonic logs can be extremely costly reaching into the tens of thousands of dollars for one log. It is because of these steep cost that some or all of this information is not included in a well suite, despite the risk of not having it oil and gas companies might be willing to accept drilling without this vital information. (allistus Nero, 2023) Often when this information is missing some of the data ( there are several things measured) the information is obtained by transforming common acquired logs like density, gamma levels, etc. The desired information is then extrapolated based on experimental formula built between the acquired data and the desired data.

The vital parts of the sonic logs is the time of flight information of two types of high frequency sound waves (Ultrasound), the compressional wave or transverse (DTC) and the shearwave (DTS). When these waves are produced and transmitted into the surrounding ground the will strike barriers and reflect back and be recorded by the probe that is sent down the bore. The time of flight or Delta T of these waves is what reveals vital information about things like porosity in the rock. While relationships have been found through experiments and this is then applied to new log features this older method is often inaccurate, leading to the potential of millions of dollars in lost revenue. There have been several studies on the adaption of ML in predicting both DTC and DTS showing great success in accuracy using XGBoost and with and without wavelet transformation.

The Data (EDA)

The data for this project was obtained from a synthetic sonic lag generation contest hosted on Kaggle. There are three data sets in this project, train, test and real data. They are comprised of 8 features and have 10,000+ rows. The features are:

* CAL - Caliper, unit in Inch,
* CNC - Neutron, unit in dec
* GR - Gamma Ray, unit in API
* HRD - Deep Resistivity, unit in Ohm per meter,
* HRM - Medium Resistivity, unit in Ohm per meter,
* PE - Photo-electric Factor, unit in Barn,
* ZDEN - Density, unit in Gram per cubit meter,
* DTC - Compressional Travel-time, unit in nanosecond per foot,
* DTS - Shear Travel-time, unit in nanosecond per foot,

A graph of different types of data

AI-generated content may be incorrect.The data contained several instance of values of -999, which took the place of missing data, these were iterate thru and replaced with NaN instead as deleting them or changing to 0 could have impact on the models. A histogram of the train set (below) showed skewed distributions and outliers. This was also shown in box plots before and after applying log function to the variables, as these are logarithmic relationships. Having prepped the information pandas *pd.corr()* function was used to look for correlations. In addition three different correlation matrixes were created. The original intent was to use *ppscore* but there was version issues (fixed later). A Pearson and Kendall relationship method were used the former being best for linear relationships and the later at slight more complex non-linear. The library ppscore although now deprecated for python versions after 3.11.0 is designed to look and find complex non-linear relationships and find correlations among the features.

|  |  |  |
| --- | --- | --- |
|  |  |  |

A chart with numbers and symbols

AI-generated content may be incorrect.

This matrixes so how effective using the right tool can be while it is typical to use pandas function to create correlations and it offers various methods to use the two used both showed correlations where ppscore showed there where none.

Methods

There are two approaches in this project on how the data is manipulated before being fed into the ML model. There was only one ensemble method used in this project. The reason for only using the ensemble method XGBoost is based after several papers and one in particular that looked at a comparison of non-ensemble like SVM and ensemble like RF and XGBoost. (allistus Nero, 2023), which concluded that ensemble methods work the best. The other approach was to process the data as is ( with minor adjustments for outliers, and missing information), and to use wavelet transformation.

As the DTS and DTC are sonic wave forms by performing wavelet transforms. The pwy.wavelet tool was used to create a function to take the whatever feature ( in this case the DTC, DTS and CAL) and return Daubechie coefficient, as they have proven to be effective with vibroacoustic and facilitate finely tuned examination of temporal and special characteristics of dynamic waves within elastic materials.

A screenshot of a graph

AI-generated content may be incorrect.

It is the hope that by doing this greater relationships can be found than on the raw data alone. The model was then trained using the train and test data set and the “real data”, both before wavelet transformation.

A group of blue and orange graphs

AI-generated content may be incorrect.

After wavelet transformation

A group of graphs showing different types of data

AI-generated content may be incorrect.

Analysis

The analysis and grading of the models performance will be based on Root Mean Square (RMSE). This information will then be compared to two other studies that did similar methodologies which are being used as a benchmark (Placeholder2) (allistus Nero, 2023).

The model before using wavelet transformation:

A comparison of a graph

AI-generated content may be incorrect.

The RSME for the DTC pre wavelet is 4.32, for the DTS is 23.11 and the overall is 16.62.

After wavelet transformation:

A comparison of graphs with blue and red dots

AI-generated content may be incorrect.

THE RSME for the DTC is 5.38 and DTS is 10.35 and overall RSME 8.72. These results were in line with other studies of ML for sonic log predictions.

Conclusions

Regardless of using wavelet transformation or not the models perform accurately at predicting both the DTC and DTS. However using wavelet transformation showed that it performed significantly better, for DTS and overall performance. This can be do to the fact that by doing a transformation can reveal discrete details that can show and find stronger correlations between other features in the dataset. The importance of accuracy given cost is import and it would be recommend that the extra step of wavelet transformation be undertaken to return the most accurate predictions.

Limitations of this project are as always the quality of the data. Many logs do not collect the amount of data that was in the ones used for this model. It is possible to use logs from other projects in similar areas or conditions to help fill gaps this could lead to predictions that are less than ideal. However some information is better than no information.

The challenges faced in the implementation was that many of the features that aided in the coding of this project are not supported in Python 3.13 and no replacements were readily available. This led to having to create virtual environments to be able to use older versions of Python.

The use of this model and sonic logs in general are not just for oil and gas operations. There is need in several different industries other than drilling where sonic logs are used, pile driving for transmission towers or bridge supports come to mind.

There is no current implementation plan for this project. As more work needs to be done on the coding to create a pipeline feature that can also deal with sonic log data that is missing features and not just data.

Ethics

The ethics of sonic log predictions in the context of well logging and subsurface data analysis involve considering accuracy, and the potential for unintended consequences when using machine learning to predict missing or incomplete sonic log data. Key ethical considerations include ensuring model accuracy, transparency in model development, and responsible use of predicted data to avoid misinterpretations or detrimental decisions.

# Works Cited

The data, code and markdown file can be found [HERE](https://github.com/LadyKate7390/Kate_G_DS_Portfolio/tree/main/Sonic%20Wave%20Velocity)

allistus Nero, A. A. (2023). *Prediction of compressional sonic log in the western (Tano) sedimentary basin of Ghana, West Africa using supervised machine learning algorithms, Heliyon,.*

MohammadRasool Dehghani, S. J. (2024). *Comparing the performance of machine learning methods in estimating the shear wave transit time in one of the reservoirs in southwest of Iran.* Nature.