# A Synopsis

# for

# B.Tech in Computer Science and Engineering

# 

**FACULTY OF TECHNOLOGY**

**KALINGA UNIVERSITY, NAYA RAIPUR, CHHATTISGARH**

# 

# A Synopsis

# On

# Feature Engineering using Signal Processing to predict DTC and DTS values

# Kalinga University

# Of

# Bachelor of Technology

# In

# Computer Science

# By

# Bigit Krishna Goswami

# Harsha Prasad Sah

# Satadal Deka

# Under the Guidance

# Mr. Pawan Sir

# Assistant Professor

# DEPARTMENT OF COMPUTER SCIENCE

**KALINGA UNIVERSITY, NAYA RAIPUR, CHHATTISGARH**

## Session: 2022-2023

## List of Contents:

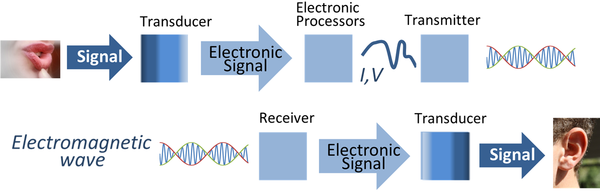
|  |  |  |
| --- | --- | --- |
| Chapter | Title | Page No. |
| 1 | Introduction |  |
| 2 | Literature review |  |
| 3 | Problem Identification |  |
| 4 | Methodology |  |
| 5 | Technologies Used |  |
| 6 | Result & Screenshots |  |
| 7 | References |  |

Introduction

In this project we aim to find out the effect of Wavelet Transformation using Signal Processing techniques on DTC and DTS velocities predictions. DTC and DTS, these are the sonic log outputs, that the compressional and shear wave velocities.

Signal Processing:

Signal processing is an electrical engineering subfield that focuses on analyzing modifying and synthesizing signals, such as sound, images, and scientific measurements. Signal processing techniques are used to optimize transmissions, digital storage efficiency correcting distorted signals, subjective video quality and to also detect on pinpoint components of interest in a measured signal.



The above image is an overview of Signal transmission using electronic signal processing. Transducers convert signals from other physical waveforms o electric current or voltage waveforms, which then are processed, transmitted as electromagnetic waves, received and converted by another transducer to final form.

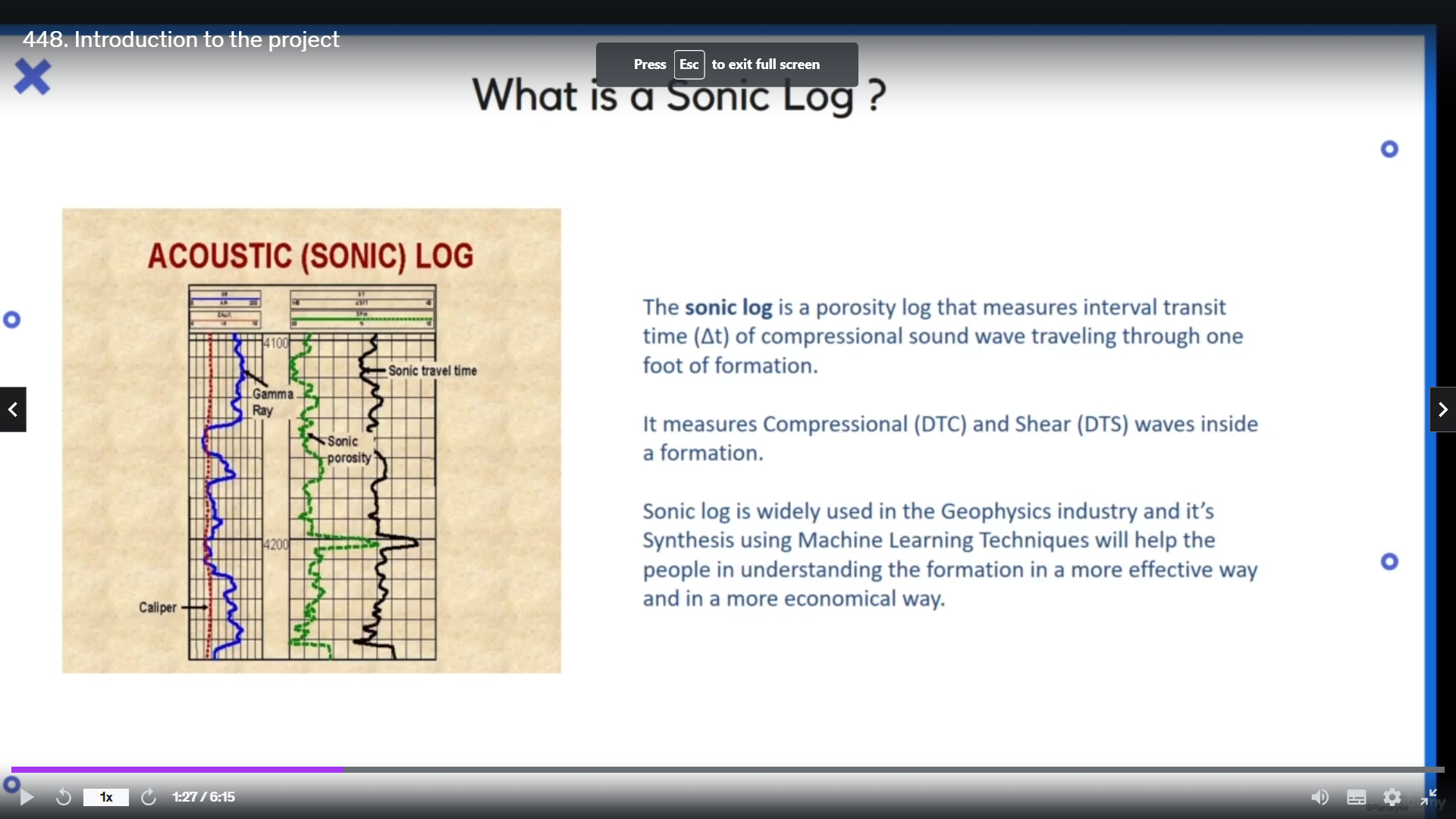
Literature Review

Sonic Log:

The sonic log is a porosity log that measures interval transit time ([Δt) of compressional sound wave travelling through one foot of formation.](https://www.bing.com/ck/a?!&&p=466877e0096c970dJmltdHM9MTY3MTA2MjQwMCZpZ3VpZD0zYTljYzgzOC01YWVlLTZiMjctMTkxMy1kOTQxNWI3NTZhYWUmaW5zaWQ9NTUwMg&ptn=3&hsh=3&fclid=3a9cc838-5aee-6b27-1913-d9415b756aae&psq=delta+symbol&u=a1aHR0cHM6Ly93d3cuYXZhbnRpeGxlYXJuaW5nLmNhL21pY3Jvc29mdC13b3JkL2hvdy10by1pbnNlcnQtb3ItdHlwZS10aGUtZGVsdGEtc3ltYm9sLWluLXdvcmQtJWNlJWI0LW9yLSVjZSViNC8&ntb=1" \t "_blank)

It measures Compressional (DTC) and Shear (DTS) waves inside a formation.

Sonic log is widely used in the Geophysics industry and it’s Synthesis using Machine Learning Techniques will help the people in understanding the formation in a more effective way and in a more economical way.



Wavelet Transformation:

Wavelets: A little Wave

Wavelet series: In mathematics, a wavelet series is a representation of a square-integrable function by a certain orthonormal series generated by a wavelet.

Wavelet Transformation can be viewed as the projection of a signal into a set of basic functions named wavelets. It is an application of Short Time Fourier Transform. There are two types of Wavelet Transform:

* Continuous Wavelet Transform
* Discrete Wavelet Transform

This is used in many domains involving signals i.e. waves which can be either sound waves, electrical waves, Sonic Log waves, etc.

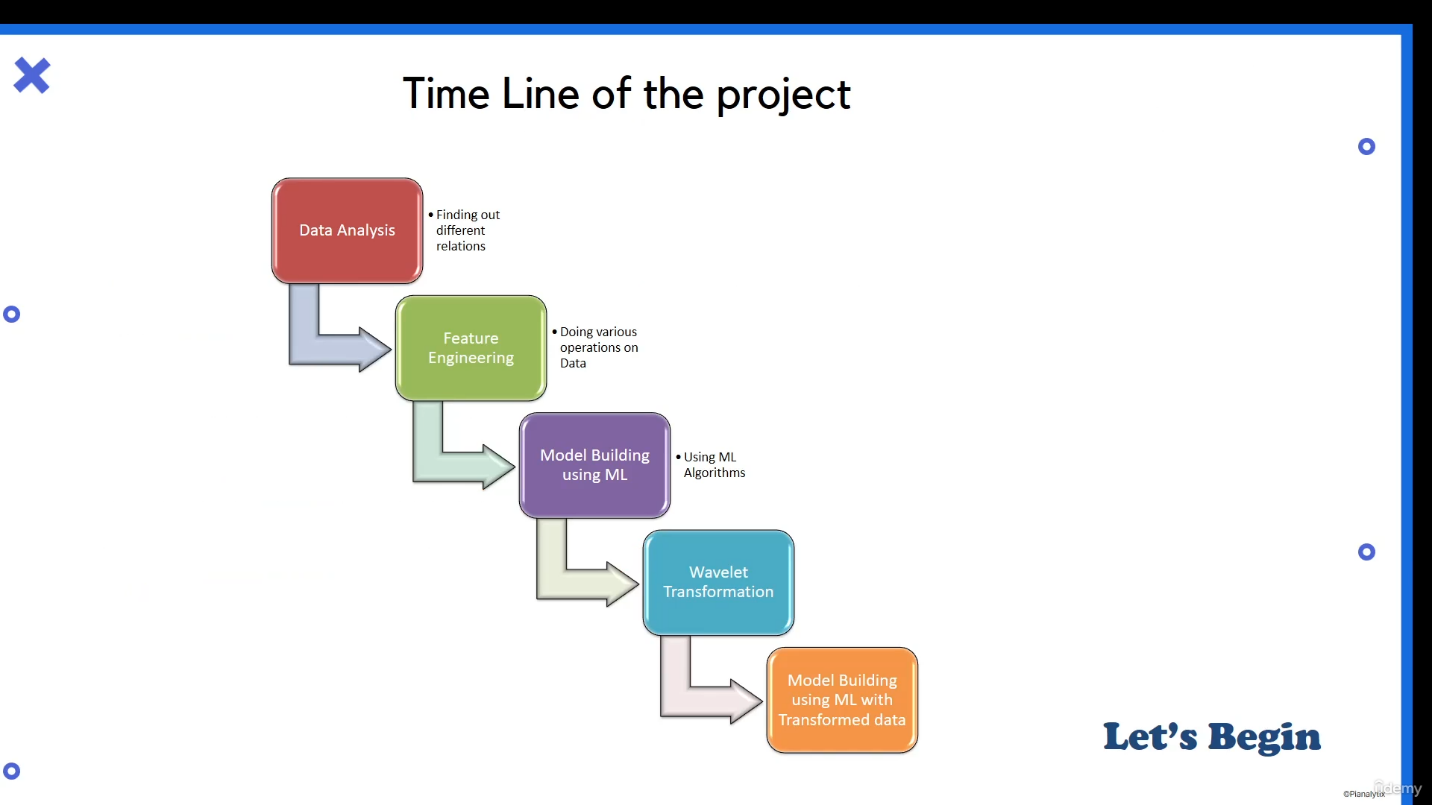
Problem Identification

In this project we will observe the effects of Wavelet Transformation on Log data for Sonic Log prediction.

Methodology

The project will go through the following steps:

1. Sonic log prediction using normal log data
2. Performing Wavelet Transformation on log data
3. Sonic log prediction using transformed data and observing the results.



DTC and DTS prediction using XG Boost Regression model:

XGBoost is an open-sourse software library that implements optimized distributed gradient boosting machine learning algorithms under the Gradient Boosting framework.

XG Boost, which stands for Extreme Gradient Boosting, is a scalble, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

This model will work as mentioned in the following steps:

* Imports
* Data exploration and Preparation
* Model Development
* Model Prediction and Evaluation

Data is taken from Synthetic Sonic Log Curves Generation Contest. I was a machine learning contest, held in the field of petroleum engineering.

Description of the data:

* 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,

Data Analysis and Preparation:

In this project, firstly a histogram is plotting based on the data using python commands in jupyter notebook. Data can be visualized in many forms such as histogram plot, line-plot, etc. The spikes present in the graphs represent the outliers, as sometimes, the machine can have a wrong reading, so these are termed as outliers. Also, since our data can’t have negative value as the graph is represented in the positive gradient of graph representation. In other words, we made the outliers as null values.

The data is then plotted as a box plot as by doing this it plots all our data in a statistical manner. It will give us the 75 percentile of the data, mean, median and the 25 percentile. The dots represent the outliers. By this, we can come to know about the outliers from our data. This is done using a simple command, i.e. sns.boxplot, which is present in the seaborn library.

The values are further converted into nan values and logarithmic values, which intern makes our data visualization way easier. Some of the outliers are left untouched which are nearby the highest percentile values. Next comes the step of finding correlations, which helps us find correlation between the values, for which we used a score called as PPS.

PPS or Predictive Power Score is an asymmetric, data-type-agnostic score that can detect linear or non-linear relationships between two columns. The score ranges from 0 (no predictive power) to 1 (perfect predictive power). It can be used as an alternative to the correlation (matrix).

Building Model:

In our project we have to predict two parameters DTC and DTS. For the purpose of making this process easier, we will make separate datasets for them. Then we make our independent and dependent variables. We divide our data into train data and test data. We also used shuffle, to randomly take values, and we created train data and test data separately for DTC and DTS. In our project, we will use XG Boost model with tuned Hyperparameters for prediction.

XG Boost is an ensemble technique which comes under the boosting technique. The basic principal of a boosting ensemble technique is we take one base model, we learn something from it and then, we do the changes in the same model and we proceed further till our changes are completely done. XG Boost stand for Extreme Gradient Boosting model.

Since, in this project our main aim is using a feature engineering technique, so, we utilized the basic hyperparameters which are generally used for similar models. We make two separate model for DTC and DTS separately. Here, we fit our model with train data and predicting it, which helps us to come to know about the accuracy. The better the score achieved by the model, the more good that model is, performance wise. If there is no over-fitting or if there is no underfitting, then we’ll see what is the future importance. This will play a vital role in the upcoming section in our project. Based on this, we’ll be doing the feature transformation or wavelet transformation of our data. CNC log is the most important feature in a prediction level. The is a feature present inside the ‘xgb’ library. We’ll plot this graph for both the models separately.

All the processes followed up till now using the training data, will be performed on the test dataset as well.

Evaluation of ML Model:

Previously, in the test data we had only 7 columns, since in our test data, there are no DTC and DTS data present. Now, we import the real test data, under file name “real\_test\_result.csv”. Earlier, the DTC and DTS values obtained were our predicted values, that was predicted from our model. Now, we’ll check our real test values. We can say this as our validation data.

The important thing to note here is about our function “result\_plot()”. The purpose of this function is to display the results in the form of r2\_square, and a scatter plot and a line plot. We pass our data in the form of lists and using the function we’ll be plotting our results, where 2 plots are from line plot and the remaining 2 are scatter plot. Here, the important note is, we used r square and rmac for plotting our graphs.

Alternative:

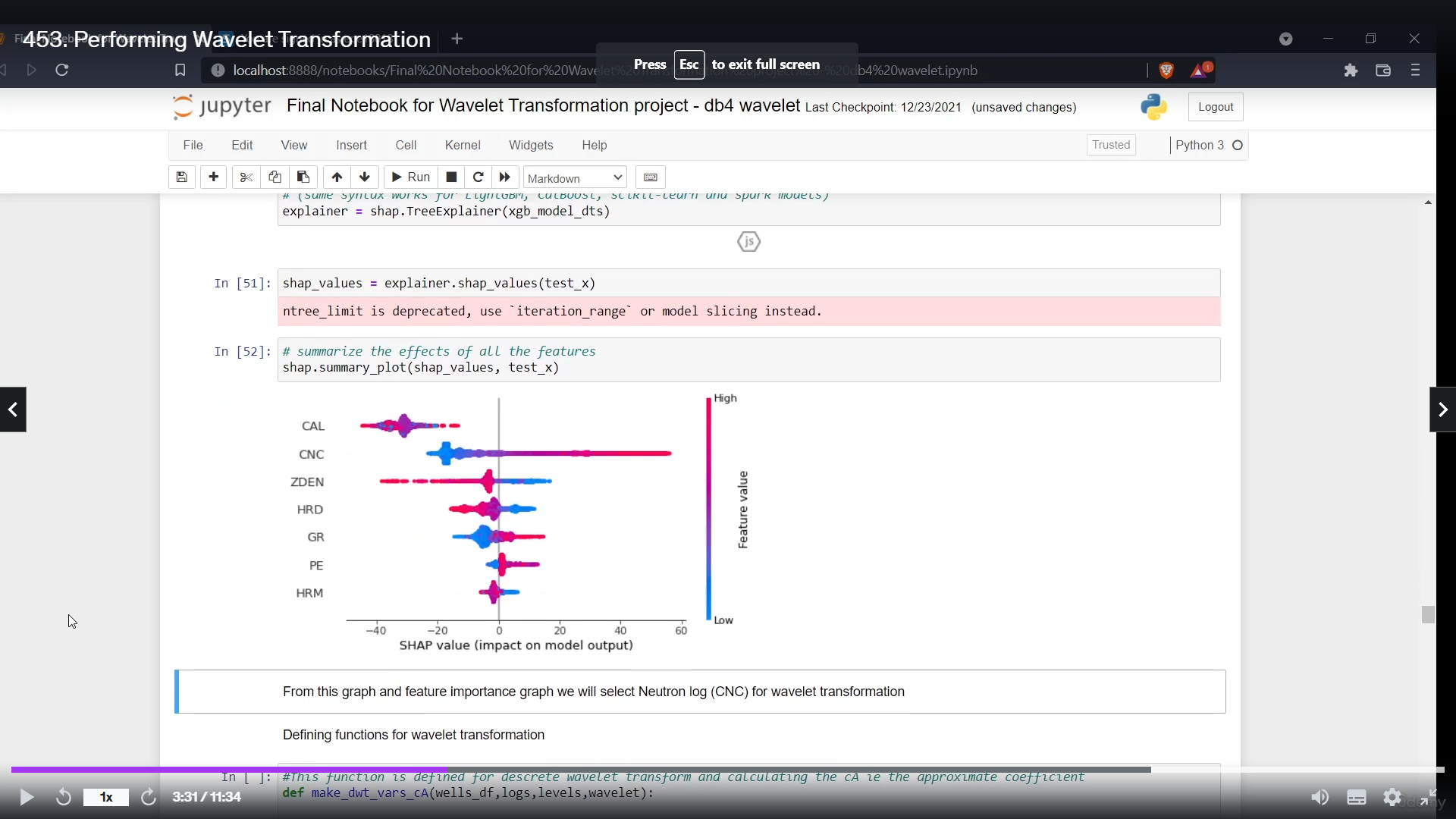
We could also use PyCaret library, which is a kind of automated machine learning library. We didn’t use it here, as being in a disadvantage of not being that effective as the model used by us. Till here, we had come to a point where we can say, that our model performs better predictions for DTC as compared to DTS.

Performing Wavelet Transformation:

In this part, we will be performing transformation in only one of the feature. The reason behind that is, when we perform the wavelet transformation, it creates a huge number of features, for example, for one log, it will decompose that into various levels and it will create two different coefficients, one is the approx. coefficient and second one is the detailed coefficient. These are the values of the broken down signals, which our transformation signal has created and there are different levels.

Since, it creates multiple layers of a single feature, so we focus on only one feature while using this, the feature which has the highest importance. Performing this on all the features will create a lot of features and leading to problems of dimensionality or over-fitting of a model, generalization of a model, etc. and it will become harder to find and select required features. So, here we are performing this in only one feature from both the categories, having the highest value as per our plots.

For this, we will be using the SHAP value graph. It also gives us the important value. As we can see for DTC value CNC is having the higher value and earlier also we say the CNC is the highest important feature for DTC prediction.



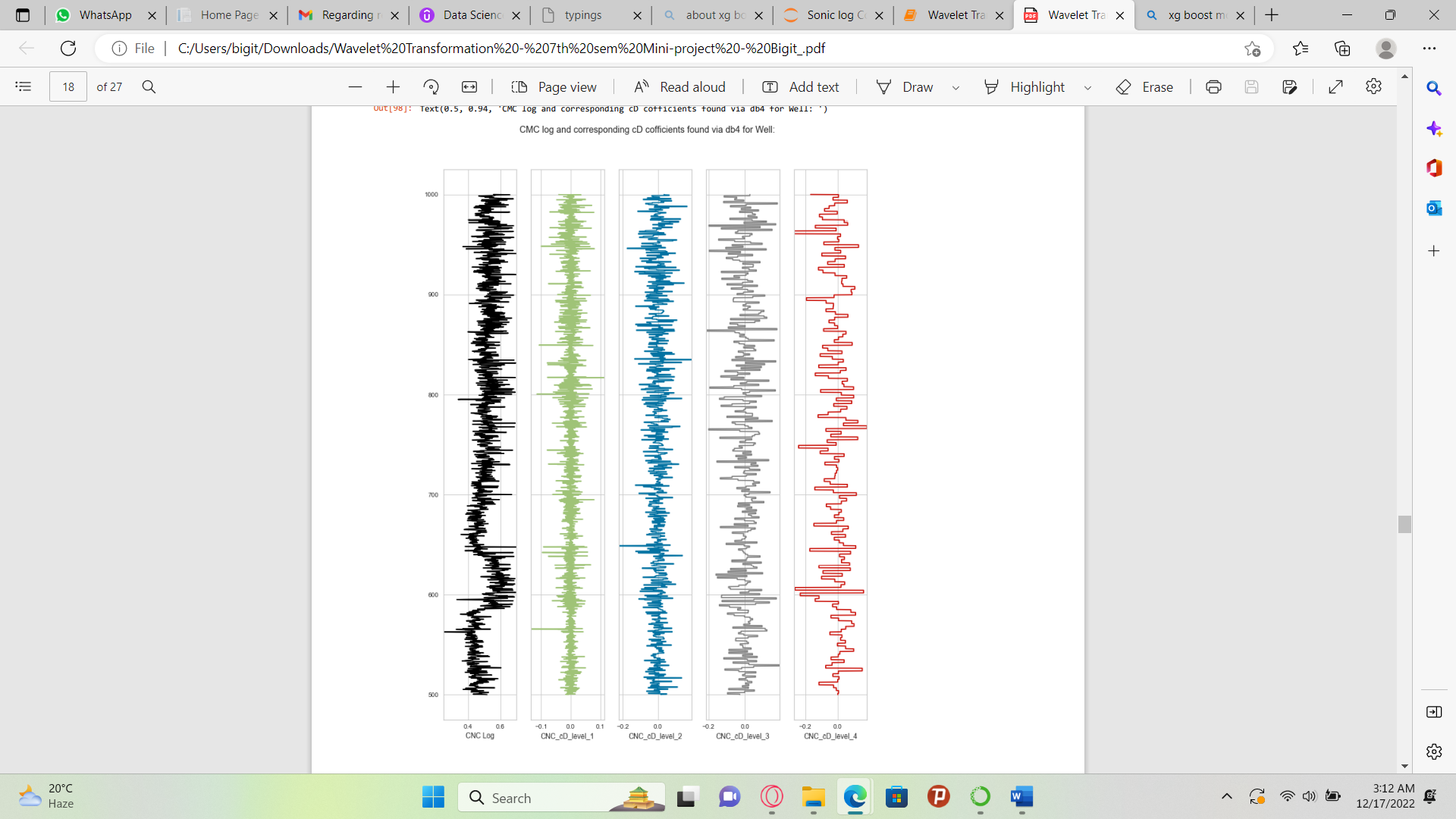
After that comes the wavelet transformation function. This function will do the wavelet transformation. We have made two functions mainly differentiating by use of approximate coefficient in the first one and the detailed coefficient on the other one.

Our function will group our data on the basis of depth. It will create a new data frame ‘temp\_df’ in which the depth will be present. Now, it will read through the logs, level wise. Level means, that many times it will decompose our signal. We can classify our wavelet in many forms. In our project, we’ll be using db4 wavelet. We take the values produced by it, concatenate it with our current data frame. Here, we are interpolating our wavelet values on the basis of our depth. We create a new feature of all the transformations that we have created on the basis of ‘level’.

Now we’ll create a new data frame. It is always good to create a new data frame, rather than performing the functions on the same dataset. Then, dropping all the null values.

At last, we will see through the db4 wavelet transformation. Here, we’ll be call the user defined functions. We are the values obtained by functions values into a list, concatenating them back into our data frame.

We can also view our output transformation created till now. Here is a glimpse:



Here we see, in our CNC log, at the beginning it was kind of mixed. In level 1, it became kind of labelled. In level 2, 3, 4 it became much more precise level wise. The data became more discrete, it is more well washed as compared to the first CNC log.

Evaluation using Transformed Data:

All the steps from making our independent variables and dependent variables, till doing our train test mostly remains same, through-out our train test and evaluation. The only difference being the transform data. Now, here we’ll be predicting using our test data. Here we’ll also be trying wavelet on Caliper log for DTS prediction as well as DTC prediction. The result show that a slight increase in the DTC value for prediction and a more drastic change in the prediction value for the DTS.

Technologies Used

In this project, we made use of the Jupyter Notebook. Jupyter Notebook is an open-course web-based notebook environment that supports three main computational engines or jupyter languages, namely Julia, Python, and R, making the word “Jupyter”.

From basic computing principles, a notebook interface is a coding environment that allows for the development of code, as well as documentation of the logic behind the code. Simply, it’s an environment that facilitates literate programming. At launch in 1987, the notebook interface was a system of documenting mathematical metaphors to support mathematical applications that undertook the computation.

Import Libraries and Dataset

The basic or ground level libraries used for importing, manipulating, visualizing data and processing techniques are pandas, numpy, seaborn and matplotlib. Also imported train\_test\_split to split the data along side that, for accuracy matrix we imported mean\_squared\_error and r2\_score. The main library used, which is xgboost. It is an implacable version of our XG Boost Regression model, that is being used in this project.

References