**INTRODUCTION**

Production from unconventional petroleum reservoirs includes petroleum from shale, coal, tight-sand and oil-sand. These reservoirs contain enormous quantities of oil and natural gas but pose a technology challenge to both geoscientists and engineers to produce economically on a commercial scale. These reservoirs store large volumes and are widely distributed at different stratigraphic levels and basin types, offering long-term potential for energy supply. Most of these reservoirs are low permeability and porosity that need enhancement with hydraulic fracture stimulation to maximize fluid drainage. Production from these reservoirs is increasing with continued advancement in geological characterization techniques and technology for well drilling, logging, and completion with drainage enhancement. Currently, Australia, Argentina, Canada, Egypt, USA, and Venezuela are producing natural gas from low permeability reservoirs: tight-sand, shale, and coal (CBM). Canada, Russia, USA, and Venezuela are producing heavy oil from oil sands. USA is leading the development of techniques for exploring, and technology for exploiting unconventional gas resources, which can help to develop potential gas-bearing shales of Thailand.

In this project, I will investigate a dataset containing examples of the geological features of unconventional reservoirs such as 'Porosity (%)', 'Acoustic impedance (kg/m2s\*10^6)', 'Brittleness Ratio', 'Vitrinite Reflectance (%)' and understand their relationship among each other, their significance and their relationship with the production of the reservoir which is simply the cross-sectional area multiplied by the square root of permeability.

**DESCRIPTION OF THE DATASET**

The Dataset has the following features 'Porosity (%)', 'Acoustic impedance (kg/m2s\*10^6)', 'Brittleness Ratio', 'Vitrinite Reflectance (%)' and the target feature is the *Aroot(K)* as shown in Figure 1 Graphical user interface

Description automatically generated

Figure 1 : Sample of the geological dataset

**Porosity** is a measure of the void spaces in a material, and is a fraction of the volume of voids over the total volume, between 0 and 1, or as a percentage between 0% and 100%

**Matrix Perm** “Matrix Permeability” is one of the most important parameters for characterizing a source rock reservoir and for predicting hydrocarbon production. The low permeability value and the presence of induced fractures during core retrieval and transportation make the accurate measurement of the true permeability values for source rocks a significant challenge for the industry

**Acoustic impedance** are measures of the opposition that a system presents to the acoustic flow resulting from an acoustic pressure applied to the system.

**Brittleness Ratio** is the ratio of uniaxial compressive strength to tensile strength.

**TOC** “Total Organic Carbon” is the amount of carbon found in an organic compound and is often used as a non-specific indicator of water quality or cleanliness of pharmaceutical manufacturing equipment. TOC may also refer to the amount of organic carbon in soil, or in a geological formation, particularly the source rock for a petroleum play;

**Vitrinite Reflectance** is the proportion of incident light reflected from a polished vitrinite surface.

is a productivity metric obtained from rate transient analysis (RTA) in unconventional reservoirs and it is equivalent to kh in conventional reservoirs. AOK is simply the cross-sectional area multiplied by the square root of permeability

From the dataset we can observe that all features including the target / output feature is numeric and that there are no categorical features.

**OBJECTIVE**

* Understand the dataset and the features
* Feature engineer the features if required
* Predict the production of unconventional reservoirs using different regression models
* Evaluate the performance of regression models using 3 cross-validation splits
* Interpret feature importance
* Identify flaws in the model and formulate a plan.

**PLAN**

Python, seaborn, pandas and numpy will be the tools used to conduct the Exploratory Data Analysis, First we will use the .describe() function to understand the dataset more, know the mean, standard deviation, minimum values, maximum values, medians for each features.

Techniques to apply

* Histogram and density plots to check if there are skewed data
* Boxplots to check features that might have outliers
* Scatter plots will be used to understand the relationship between a feature and another
* Heatmap plot to check if there are any features that are linearly correlated

**INSIGHTS**

From the Histogram and Density plots shown in Figure 2, we can see that all features including the target feature are normally distributed, however Matrix Perm (nd) has a positive skew, Brittleness ratio and TOC has a negative skew

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Figure 2 : Histogram and Density plots of all features

From the Boxplot as shown below in Figure 3, Outliers exist in the Matrix Permeability feature as well as the Vitrinite Reflectance Feature, this can be treated by transforming the features.

A screenshot of a computer

Description automatically generated with low confidence

Figure 3 : Box plots of all features

From the pair plot shown in Figure 4, we can observe that Porosity, Matrix Perm are strongly correlated with the target feature Aroot(K), we can also observe that Porosity and Matrix Perm are linearly correlated, TOC and Porosity are also linearly correlated, this means that we can drop the Matrix Perm and TOC columns as they will not provide significant value

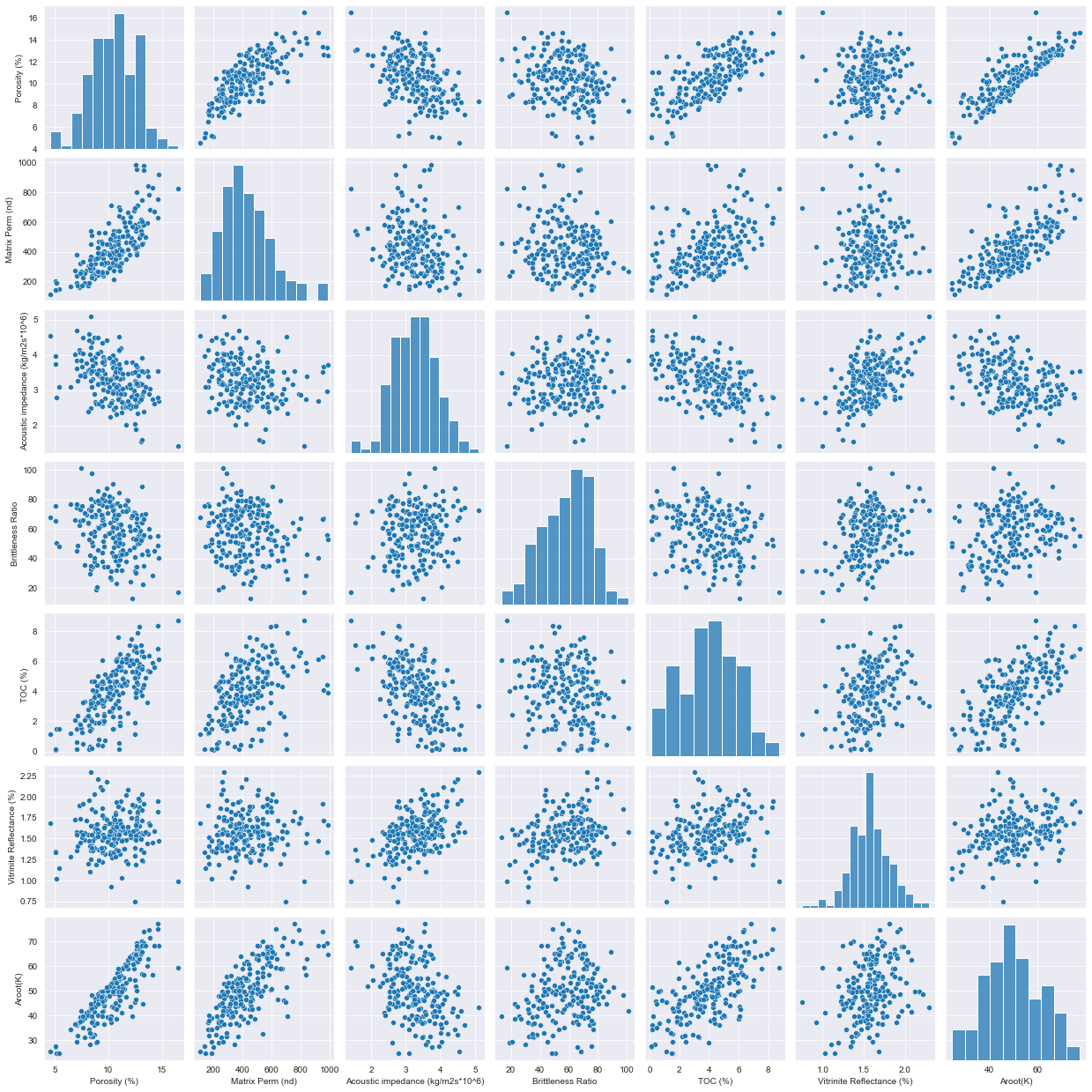
****

Figure 4: Pair plot of all features

**DATA CLEANING & FEATURE ENGINEERING**

As observed in the pair plot Porosity and Matrix Perm are linearly correlated, TOC and Porosity are also linearly correlated, However we want to quantify the Correlation thus we will use the pearson correlation, a threshold we will set is that features that have a person number above 0.7 will be considered linearly correlated and can be dropped from the dataframe.

Using the Heatmap function from the seaborn library as shown in Figure 5, we can observe that TOC and Porosity have a pearson correlation number of 0.71, and Matrix Perm and Porosity have a pearson correlation number of 0.76, this means these features are linearly correlated and we can drop any of these features.

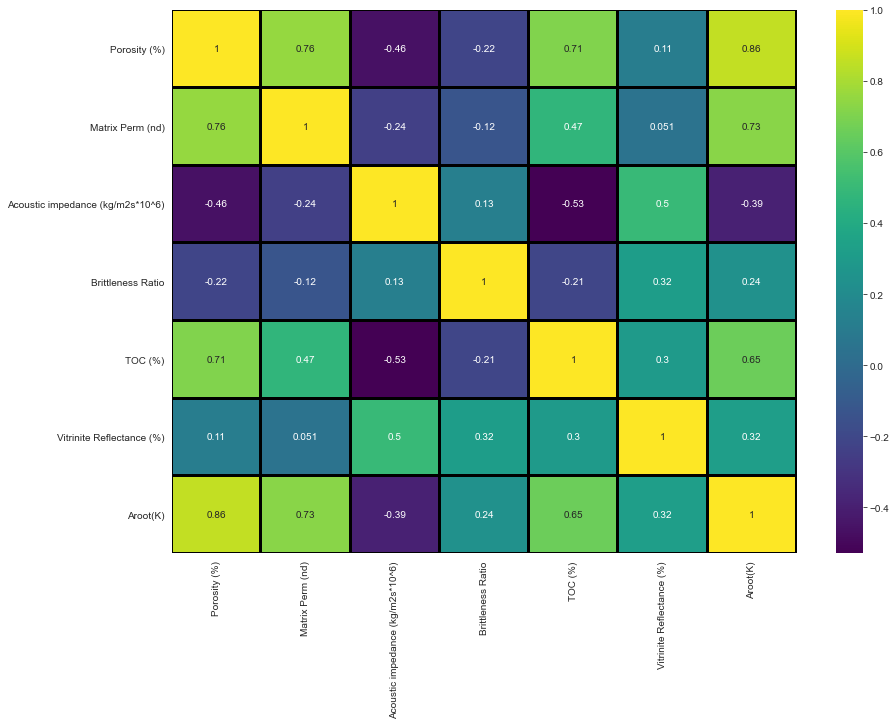


Figure 5: Heat map of all features

*Dropping Features*

We will drop the Matrix Permeability and the TOC feature as they are already linearly correlated with the Porosity feature

Table

Description automatically generated with medium confidence

We can observe that the new set of features are:

1. Porosity
2. Acoustic Impedance
3. Brittleness Ratio
4. Vitrinite Reflectance

*Scaling*

We can also observe that we need to scale the features as there are big differences between the values of Brittleness Ratio column and Vitrinite Reflectance column.

We will use the MinMax scaler from the sklearn library to scale the features so that they are between the values (0 and 1)

Graphical user interface

Description automatically generated

**REGRESSION MODELS**

***First, I have created the X and y dataframes that will be used to train different linear regression models***



***Following that I created sklearn pipelines Linear Regression, Lasso Regression and Ridge Regression which would scale the training set first.***

Text

Description automatically generated

***Following that I have used KFlod cross validation (3 splits) to make sure I have robust predictions and used cross\_val\_predict to predict the target feature using Linear Regression, Lasso Regression and Ridge Regression***



A picture containing text, indoor, orange, screenshot

Description automatically generated

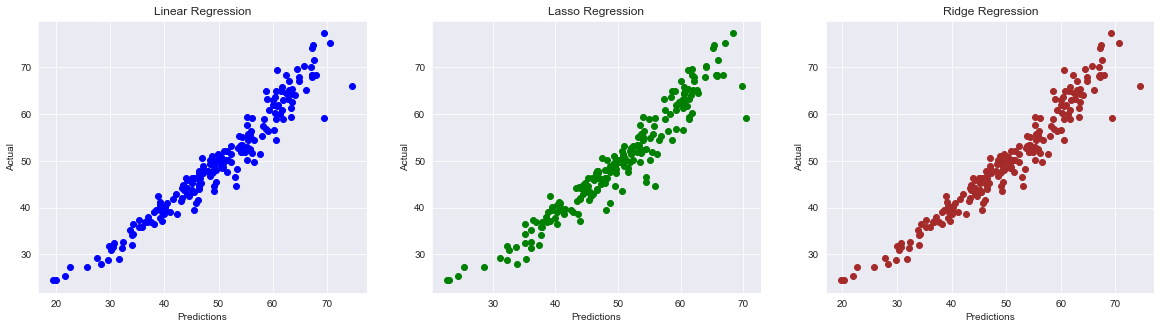
***The R2 Score of each regression model is as following***

Text

Description automatically generated

Where Linear Regression and Ridge regression obtained the best r2 score at 0.9337 and 0.9328 respectively.

***The Scatter plots is as following***



***To achieve better accuracy, I have implemented polynomial transformation of features in the pipeline.***

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

***The R2 Scores of the Regression models after using Polynomial features is as following***

Text

Description automatically generated with medium confidence

We can observe that r2 score increase significantly after implementing polynomial transformation.

***The scatter plots of the Regression models after using Polynomial features is as following***

