A Predictive Model for Oil Peak Rate

Rice Datathon24 Chevron Track

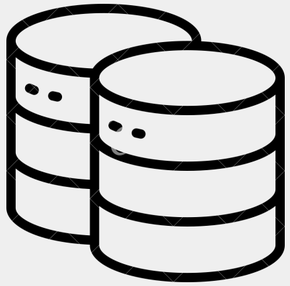
Oscar Wu , Chih-Jung Fang,
Naman Gupta , Rung-De Chu

PROJECT OBJECTIVE

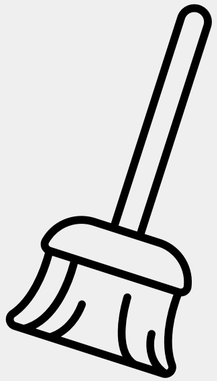
By exploring the oil peak rate, we aim to develop a model to optimize Chevron's resource management and strategy for asset development.

How can we predict the Oil Peak Rate ?

PROJECT WORKFLOW

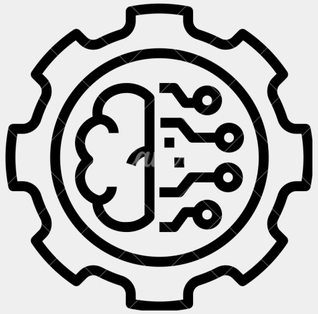


Data



Data Wragling

Data Exploration



Modeling

Objectives

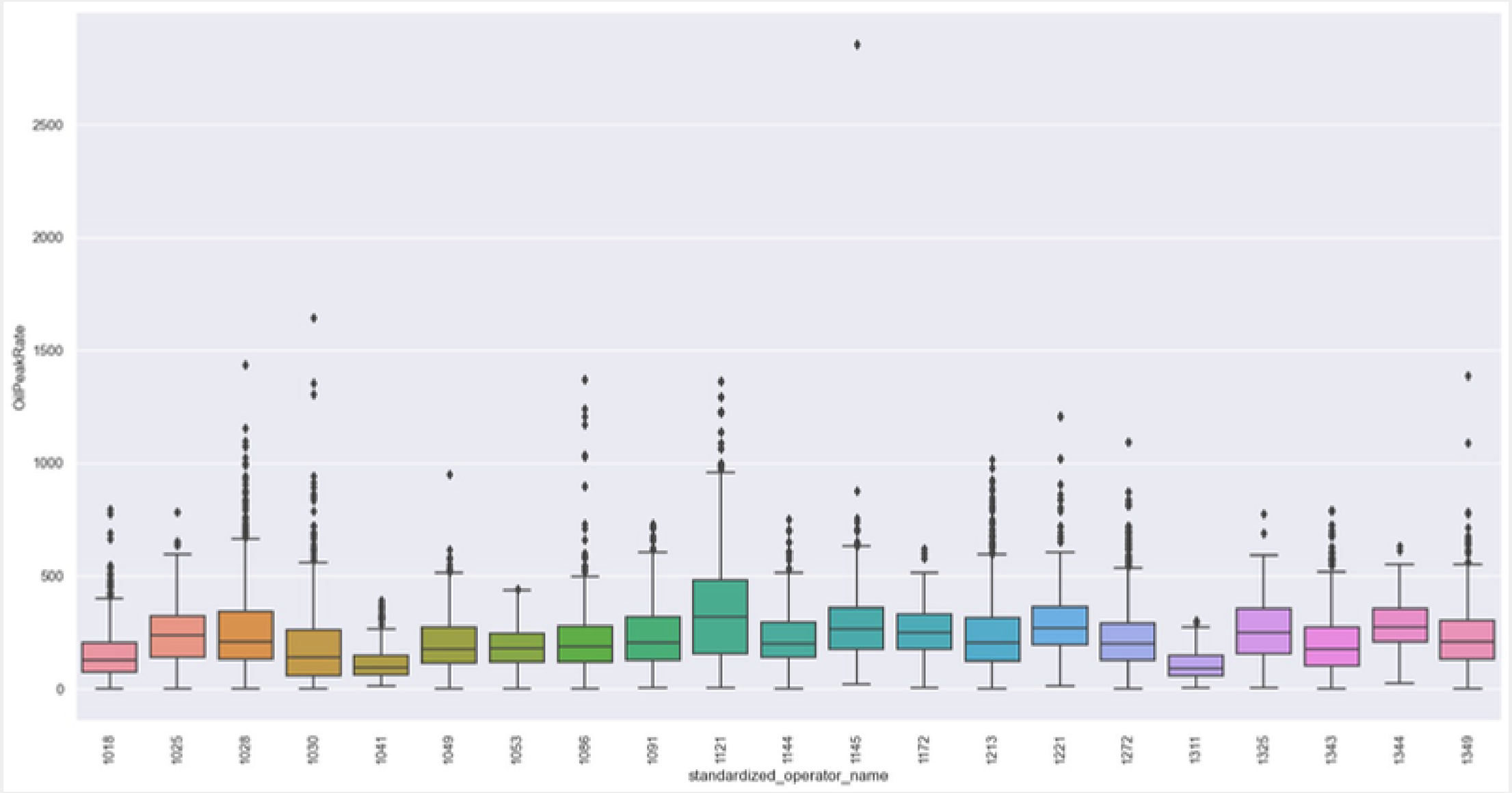
DATA WRAGLING

Remove irrelevant or unnecessary data

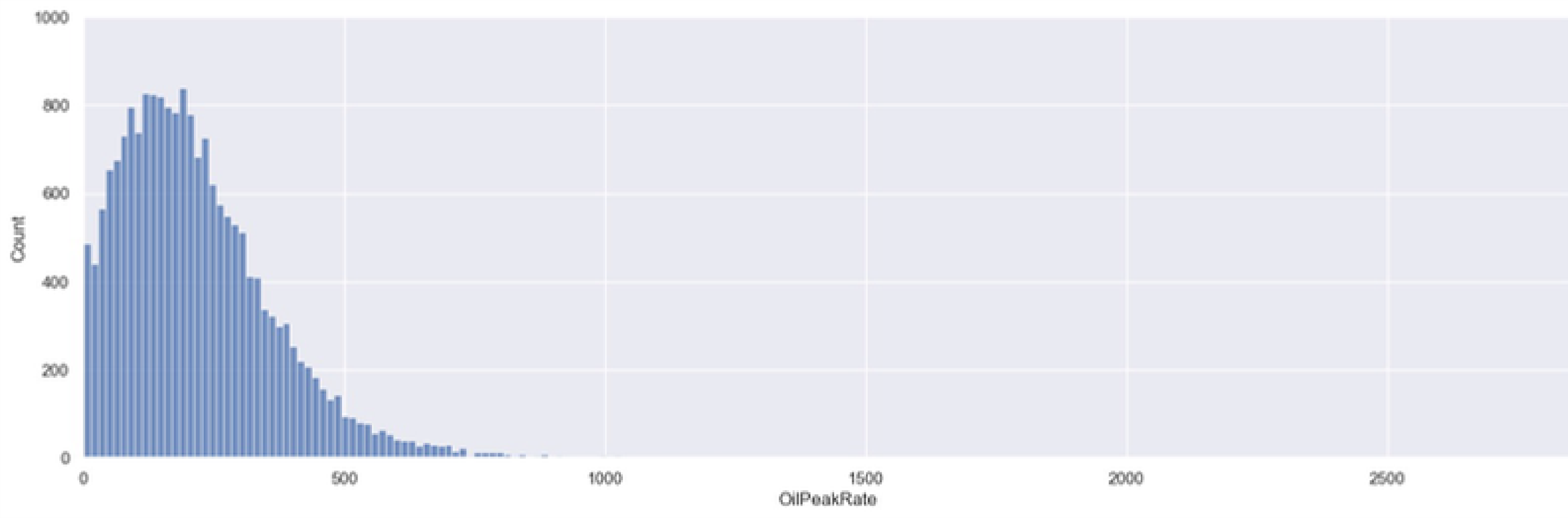
Use KNN imputer to process missing values

1.

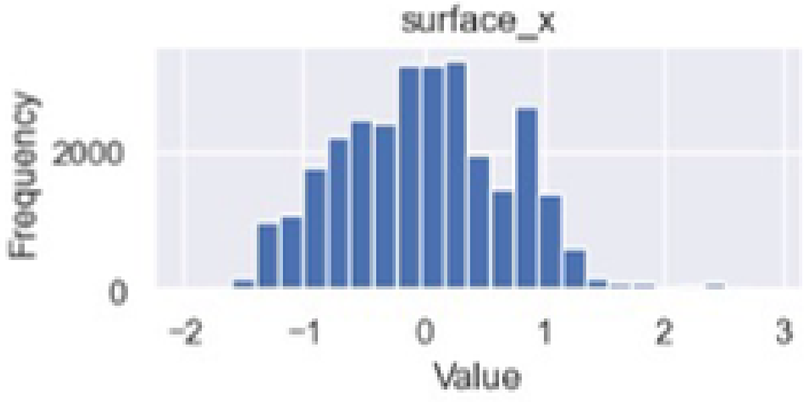
COMPARATIVE ANALYSIS OF OIL PEAK RATE BY OPERATOR

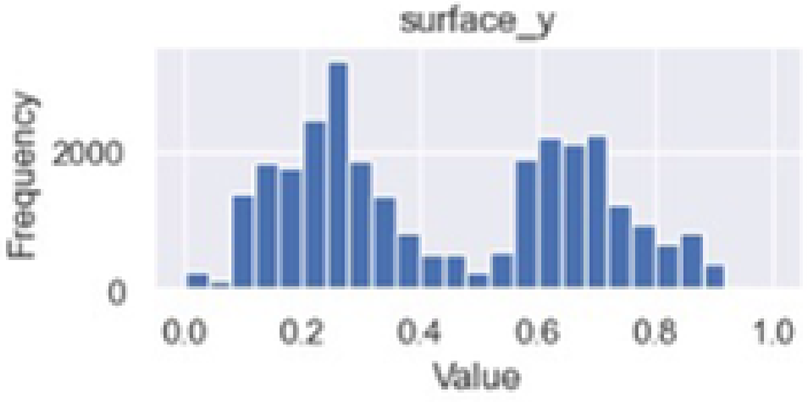


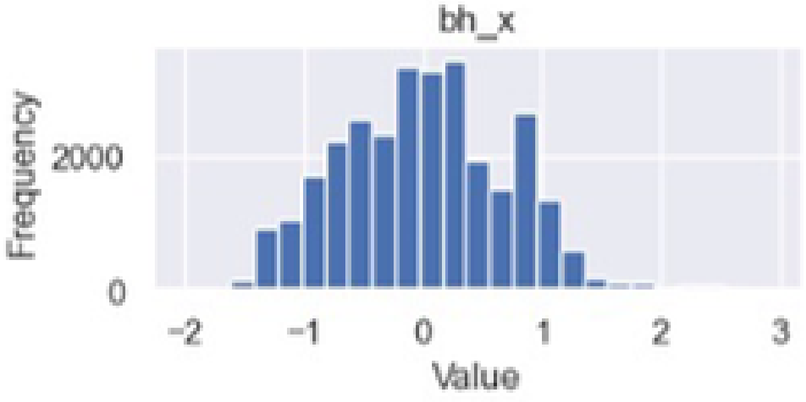
DISTRIBUTION OF OIL PRODUCTION PEAK RATES

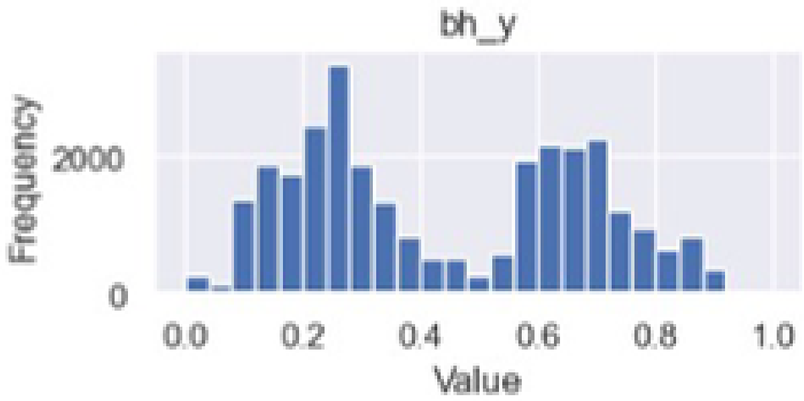


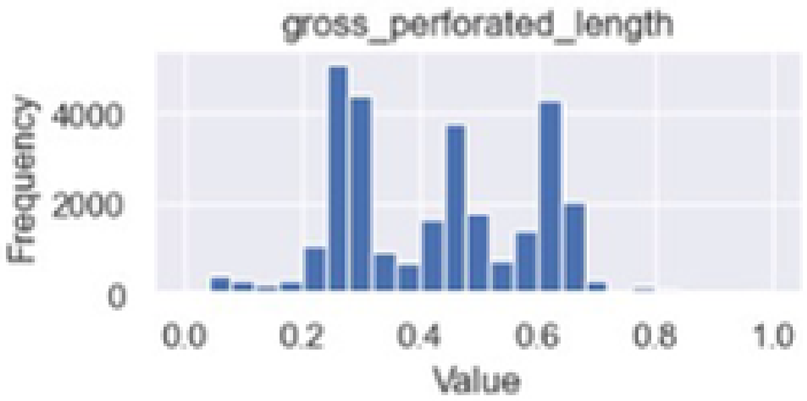
EXPLORATION OF NUMERICAL DATA -1

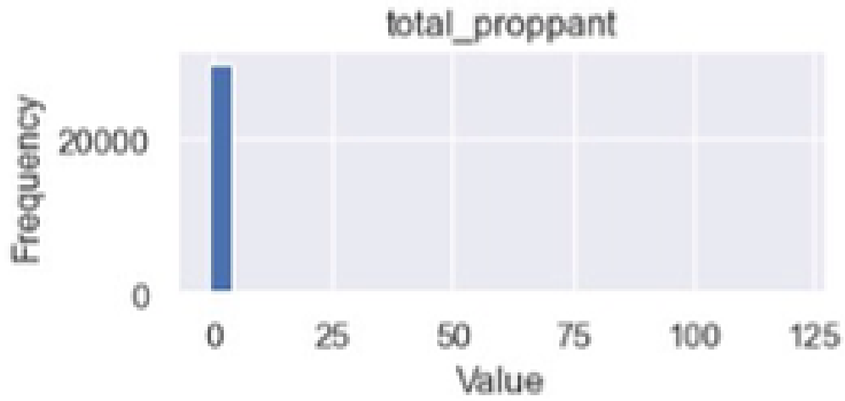




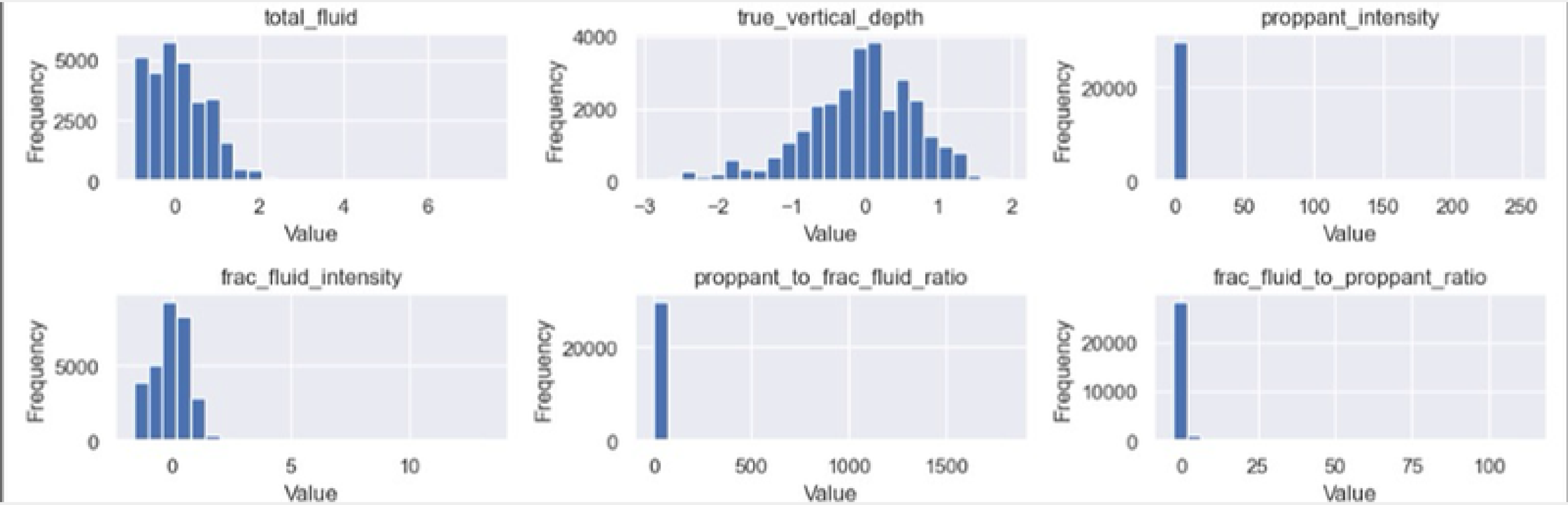




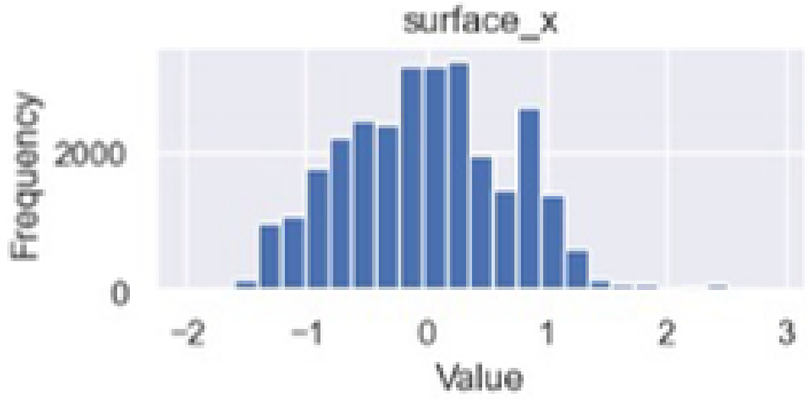


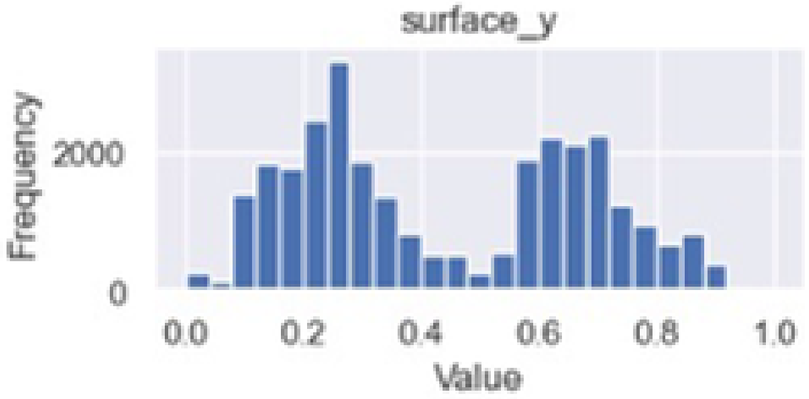


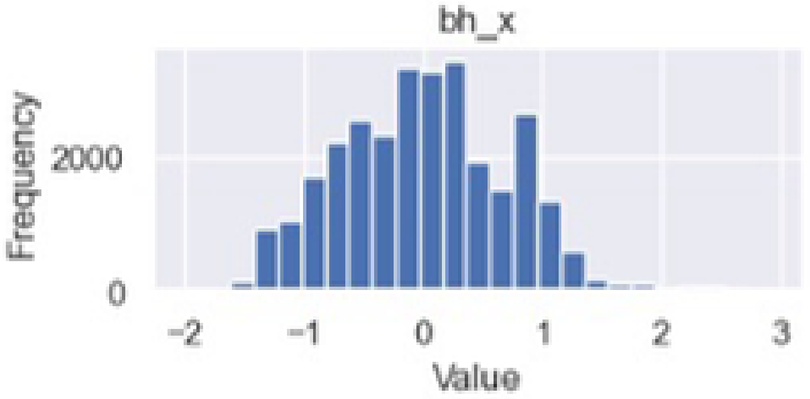
EXPLORATION OF NUMERICAL DATA -2

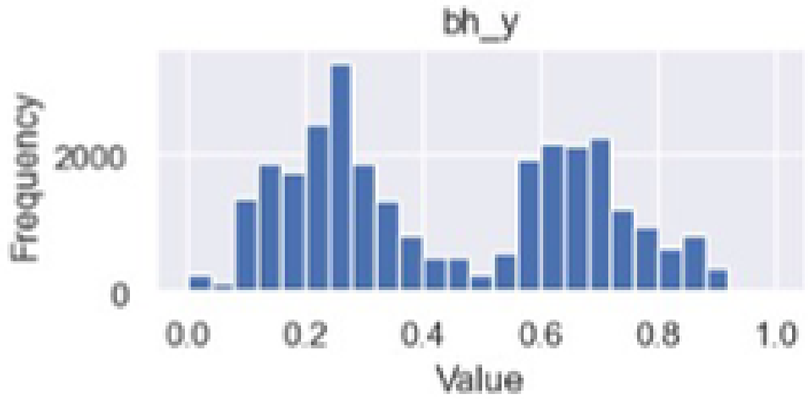


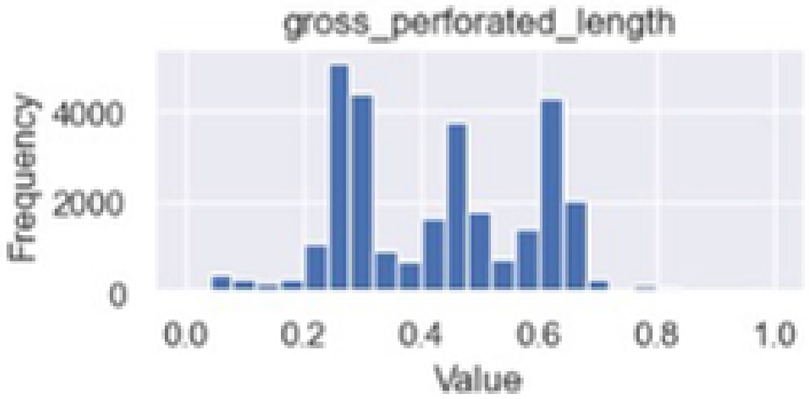
EXPLORATION OF NUMERICAL DATA -3

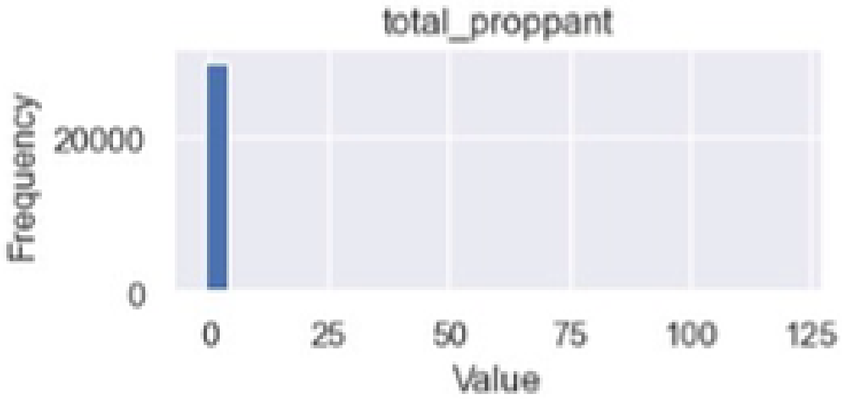












ADDING 4 NEW FEATURES

PROPPANT_EFFICIENCY

$$\text{proppant_efficiency_f} = \frac{\text{total_proppant}}{\text{true_vertical_depth}}$$

WELL_TRAJECTORY_LENGTH

$$\text{well_trajectory_length_f} = \sqrt{(\text{surface_x} - \text{bh_x})^2 + (\text{surface_y} - \text{bh_y})^2 + (\text{true_vertical_depth})^2}$$

FLUID_EFFICIENCY

$$\text{fluid_efficiency_f} = \frac{\text{total_fluid}}{\text{true_vertical_depth}}$$

WELL_INCLINATION

$$\text{well_inclination_f} = \arccos\left(\frac{\text{true_vertical_depth}}{\text{well_trajectory_length_f}}\right)$$

DATA SCALING

MinMax Scaler

preferred when the features contains less outliers.

Robust Scaler

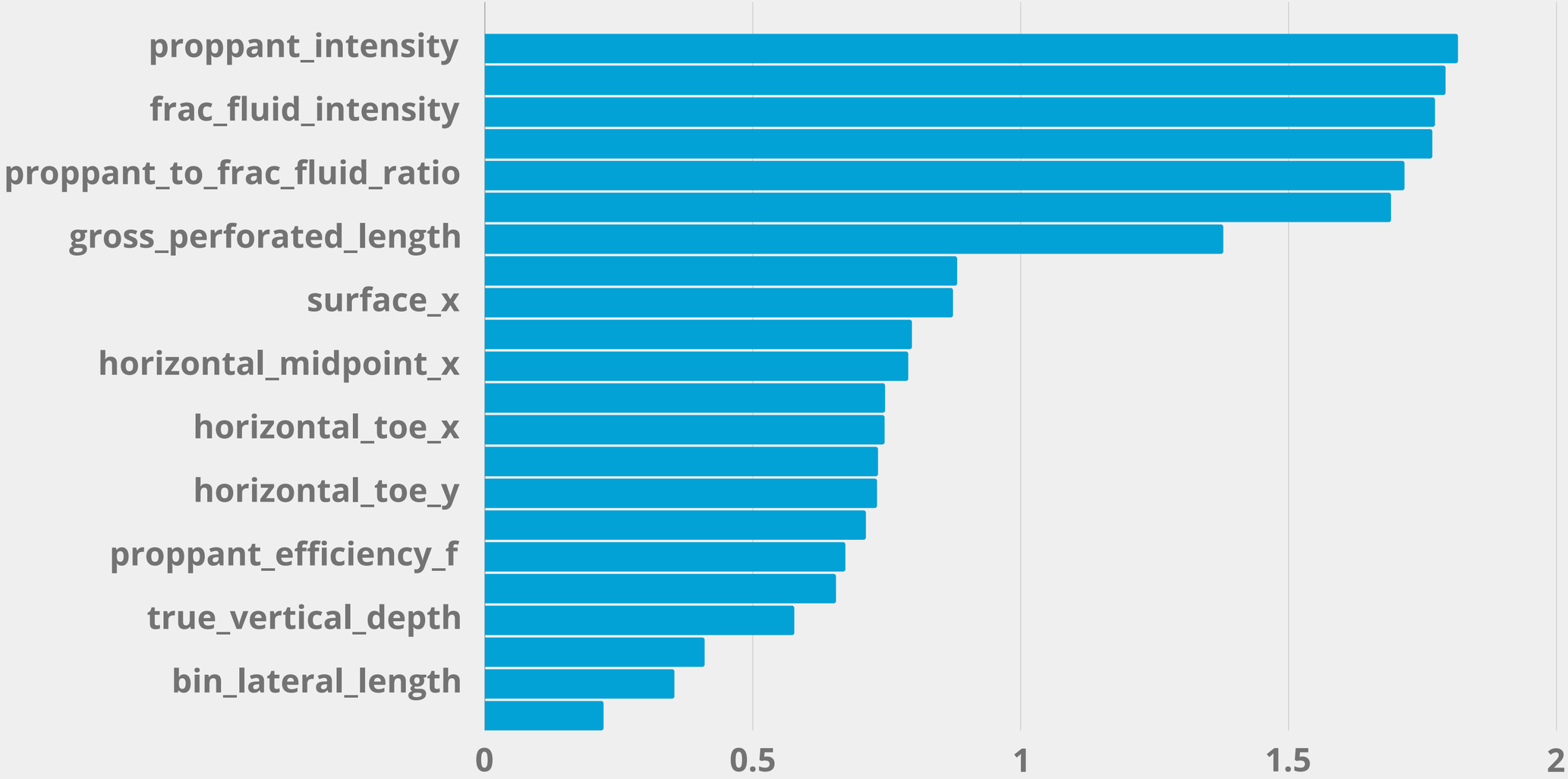
preferred when the features contains many outliers.

MODEL COMPARISON

MODEL	RMSE
LinearRegression	0.6935
CatBoostRegressor	0.4712
GradientBoostingRegressor	0.5898
LGBMRegressor	0.4953
RandomForestRegressor	0.4311
DecisionTreeRegressor	0.6257
ANN	0.5078

FEATURE IMPORTANCE

We utilized **mutual information scores** with s**ignificant correlation matrix values** to identify important features, which are then employed to tune the model.



MODEL TUNING

Model	RMSE Before Tuning	RMSE After Tuning	RMSE Difference
CatBoostRegressor	0.4712	0.4494	0.0218
LGBMRegressor	0.4953	0.4692	0.0262
RandomForestRegressor	0.4311	0.4284	0.0027

RandomizedSearchCV efficiently tunes machine learning models by randomly sampling parameter combinations, saving time and computational resources while effectively optimizing model performance.

THANK YOU