Summary of Methodologies Used & Results

Introduction

The extraction of oil, gas, and/or water, whether from a single well or an entire field, fluctuates over time. It is crucial to possess a reliable qualitative and quantitative approach to comprehend and forecast production for the industry. With the data collected in the past years , creating a robust regression model to do the prediction is possible. Our approach involved systematic data cleaning alongside the exploration of diverse models, including linear regression, decision trees, random forests, voting regression, and neural networks. We removed unnecessary features, handled missing data using both removal and algorithmic filling methods, removed outliers, and selected significant features using algorithms. This iterative and comprehensive approach allowed us to refine and optimize our predictive models continually.

After rigorous evaluation, the Random Forest model emerged as our preferred choice. Its ability to capture intricate patterns, handle non-linearity, and offer robust performance with large datasets aligned seamlessly with the project goals. The RMSE of 83.26/74.15/91.53/79.21 reflects the possibility in predicting oil peak rates, marking a significant milestone in our endeavor.

Method

4 datasets: Delete all row with the empty value;
Fill in all the blanks using KNN algorithm;
Select some features with the (z-score>2);
Drop the columns having large amounts of missing values, and fill the others with the KNN algorithm.

5 models: Linear Regression, Decision Tree, Random Forest Regression, KNN Regression, Voting Regression, Neural Network models.

Fine-Tuning those models to achieve better performance.

Result

Models	RMSE(testset)			
	EmptyValFILLED	DropedRowsWithEmptyVal	Selected Features	DropedColsWithEmptyVal
LinearRegression	97.85	93.56	102.76	175.93
DecisionTree	98.77	90.65	101.04	95.16

RandomForest	83.26	74.15	91.53	79.21
KNN	95.58	85.44	111.25	92.99
Voting Regression	88.6	83.66	98.137	87.93
Neural Network	95.64	91.37	102.2	95.14

Conclusion

In this study, we introduced various models for predicting oilPeakRate, each applied to distinct datasets. RandomForest outperformed other models in both preprocessing and feature extraction. The dataset contains numerous missing values, potentially impacting model performance. Future data collection efforts could enhance our model. Notably, features like true_vertical_depth exhibit correlations with our target value, offering valuable insights for future exploration.