Peak Oil Production

John Tian, Arnan Bawa, Andrew Chu, Ashley Song

Goal:

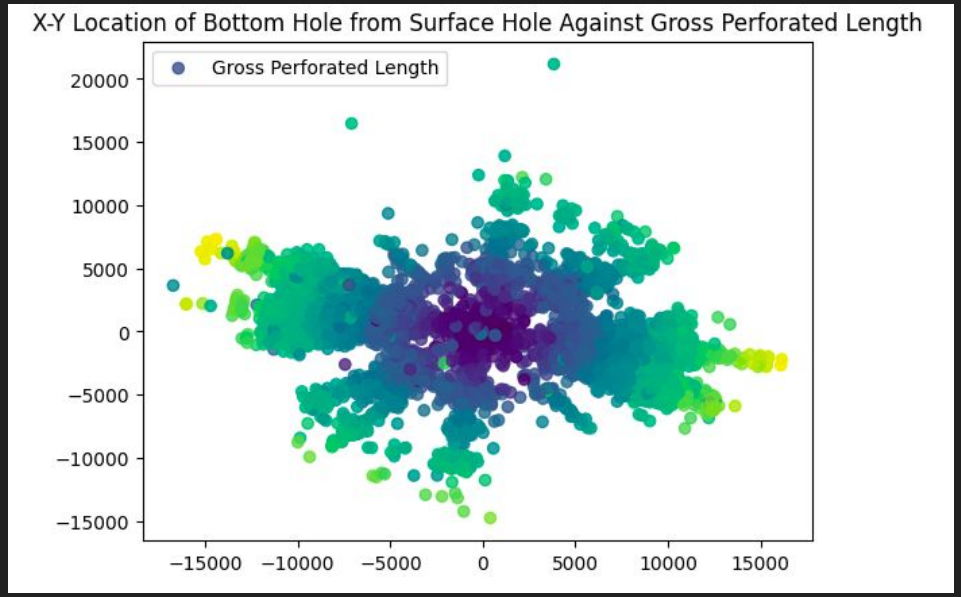
To create a model to predict the peak oil production rate for oil wells given various features

Structure of the Data

- Huge amounts of missing values, including a considerable amount for the target feature
- Range of values in each column vary wildly -> standardization crucial

Visualization

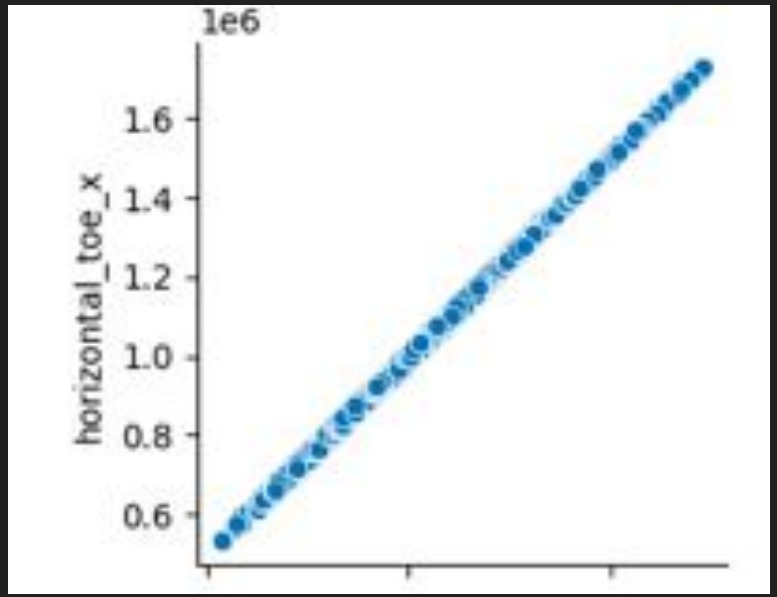
- Consider physical structure of wells



More Visualization

- High amount of collinearity -> consider nonlinear models later

Horizontal Toe X vs Surface X



Surface X

Cleaning the Data

- KNN Imputation on every numeric feature apart from the target feature
- MICE Imputation on the target feature using KNN Imputed values
- MICE Imputation on label-encoded categorical features
- One-hot encoding to properly handle categorical features

Modelling

- Feature engineering
  - distance from surface to bottom
  - length of heel to toe of well
  - fluid intensity
  - etc…

- AutoML Methods (AutoGluon) to efficiently compare optimized models
  - WeightedEnsemble L2 ranked most effective

Analysis

- Ran model on ~400 features, many of which deemed trivial by AutoGluon's feature_importance method
  - PCA or t-SNE for dimensionality reduction in the future?
- Ensembling/boosting methods outperformed deep learning models
  - Deep learning overkill for tabular data?
- Feature engineering substantially improved performance (30% reduction in RMSE)
  - Importance of "augmenting" existing data