Oil & Gas Case Study Project
Group 90

I. Description of the data

In this oil & gas case study, we are predicting “CumOil12Month” (dependent/response variable) or in another word: the amount of oil produced per well within the first 12 months of a wells productive life.

Data

In the dataset, we have 6098 observations with 23 variables, including the response variable. Independent variables include operator/company names (categorical variable: 65 unique values total), completion dates, reservoirs (categorical variable: 9 unique values total) and other quantitative variables. All independent variables could be categorized by three, whether it gives us information about 1) completion, 2) geology, and 3) well spacing.

1) In the completion group, the variables include operator, completion date, amount of proppant and fluid, and percent of components of the reservoir, as well as the total cost of the horizontal well (unit in millions of dollars).
2) In the geology group, the variables include reservoir, depth, porosity, and pressure of the reservoir.
3) In the well spacing group, the variables include horizontal and vertical distances to the nearest offset well (unit in Feet).

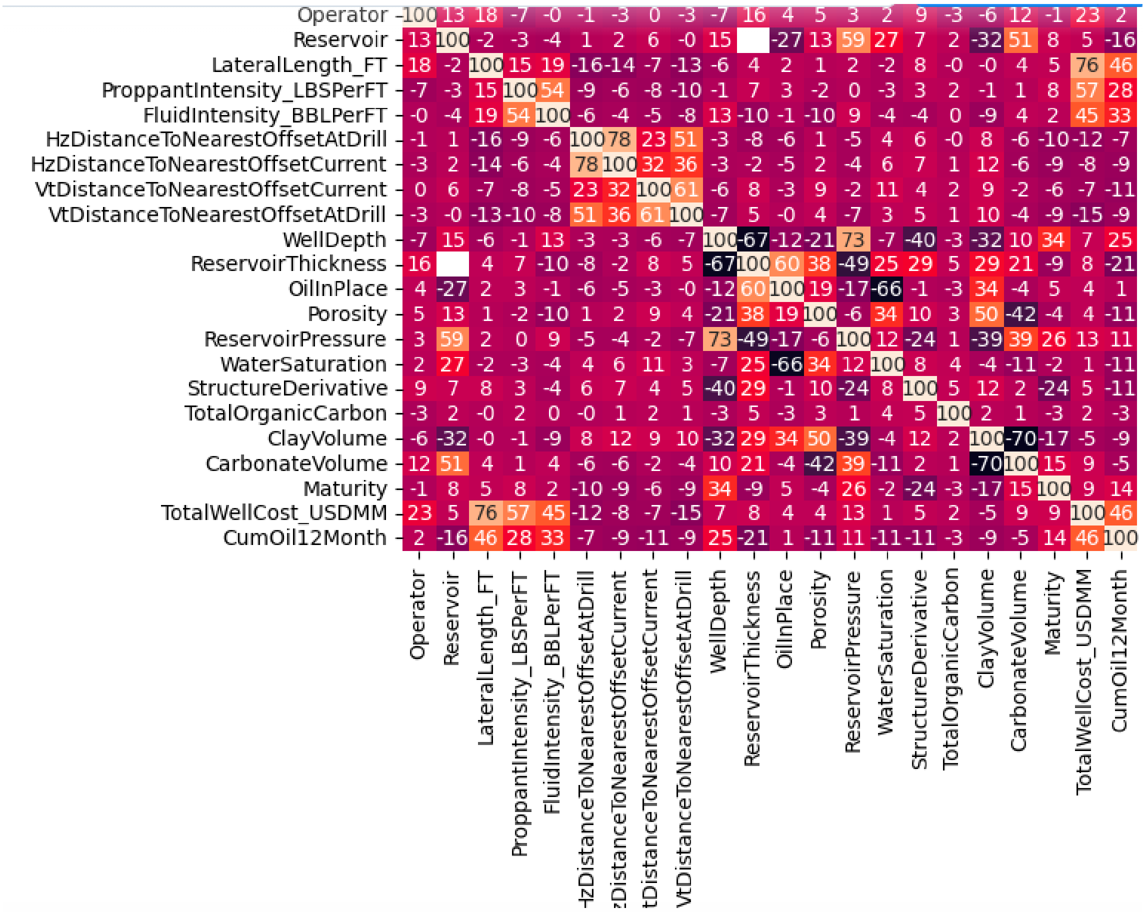
Initial explorations

As an exploratory analysis, we’ve plotted scatter plots and histograms to investigate distributions of each variable, created a correlogram to identify any correlation across variables, and time series plot for investigating trend (based on completion dates).

We are only showing the most interesting visualizations here as we generate more plots and prints to understand the dataset.

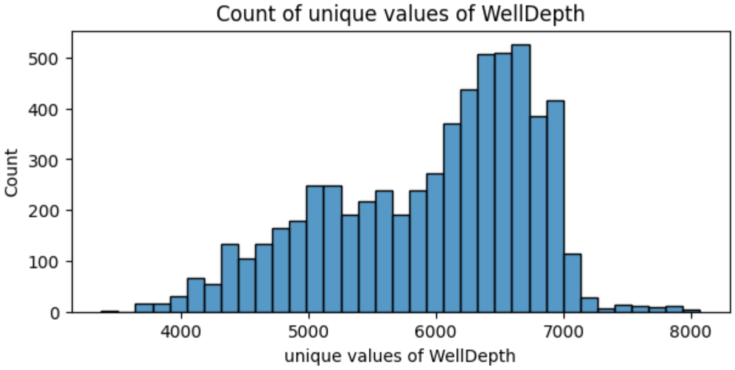
II. Visualizations and captions

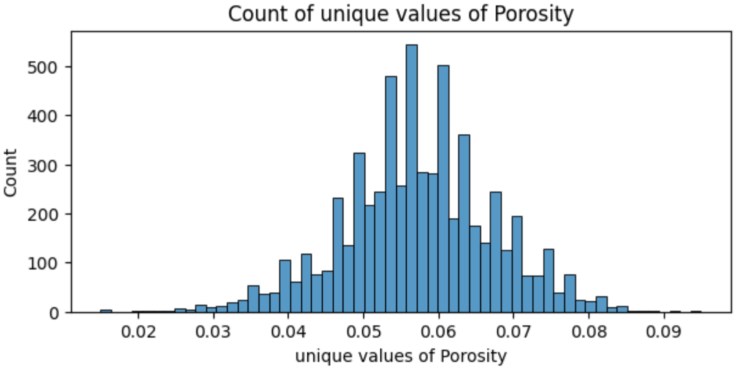
a) Correlogram (%)

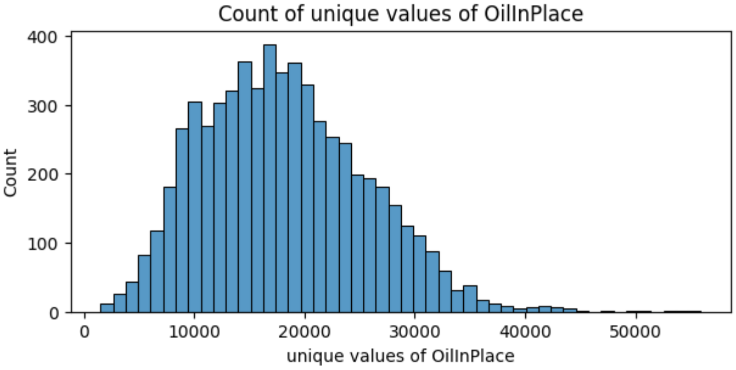


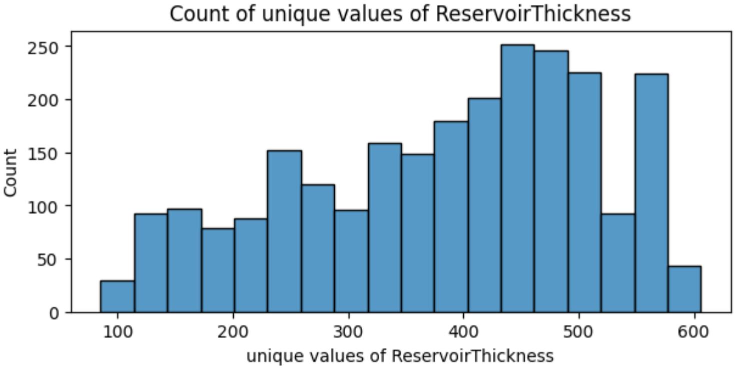
There are some variables showing high correlation (e.g., reservoir thickness & well depth, clayvolume & carbonate volume), so we would very much need models with the capacity to deal with overfitting.

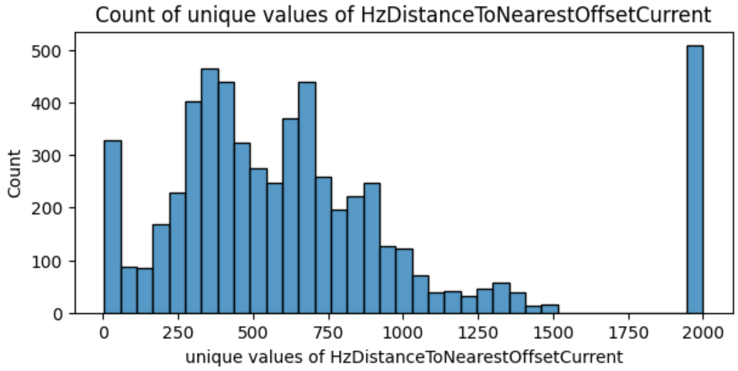
b) Histograms of some of the key variables





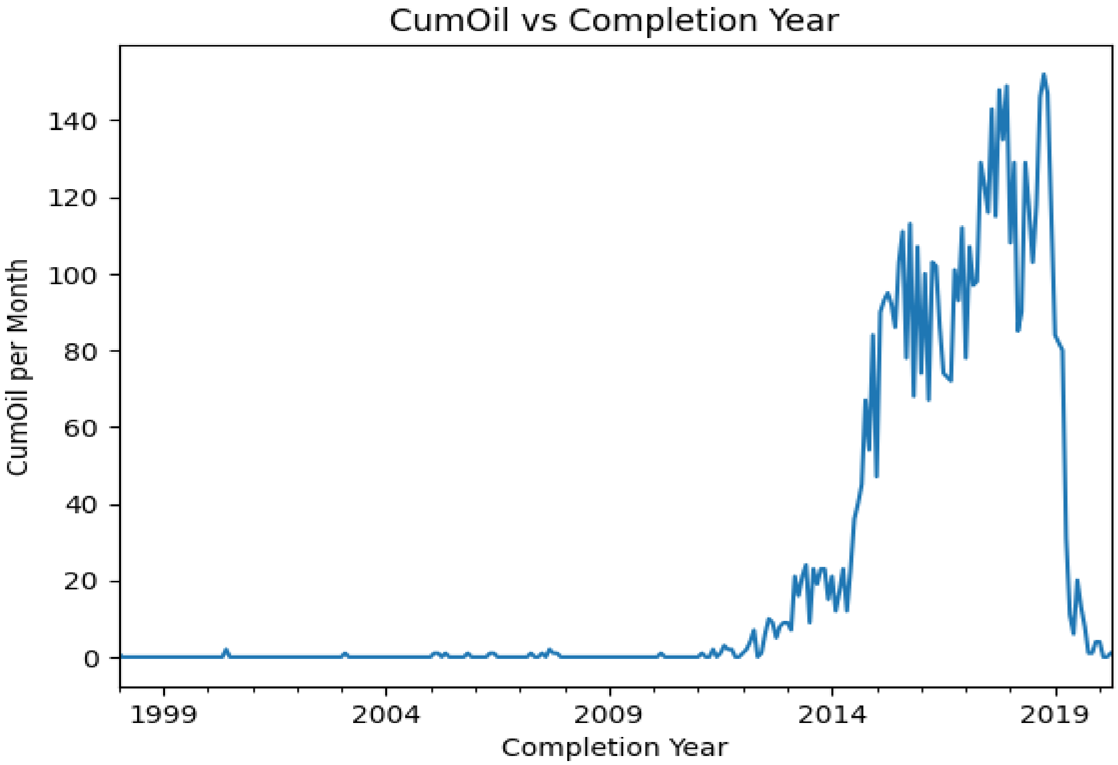






We observed a wide range of distributions across variables. Some look normally distributed, and some are not. We may need to apply transformation for some variables. We even have variables with concentrated density in the minimum and the maximum of the value range.

c) Completion date trend (# of obs per month)



Most of the observations are concentrated in the period of 2014-2019. We observe very sporadic occurrences before 2011. We might have to truncate data, or be strategic when choosing a train/test set.

III. Revised project question

We would like to develop a model to best predict the amount of oil produced per well within the first 12 months of a wells productive life (“CumOil12Month”) from the given predictors using the given data. In this case the project question stays the same after the exploration.

IIII. Baseline model

We train a multilinear regression model as a baseline model. In the first place, missing values in the dataset was an issue. We had to do a mode (which is not good for qualitative variables and we’ll fix that later) imputation on both the train and the test dataset in order to perform the regression. For each column we impute missing values with the mode of observations.

Below is the output of our baseline model:

Dep. Variable:	CumOil12Month	**R-squared (uncentered):**	0.853
Model:	OLS	**Adj. R-squared (uncentered):**	0.852
Method:	Least Squares	**F-statistic:**	1408.
Date:	Sat, 03 Dec 2022	**Prob (F-statistic):**	0.00
Time:	03:38:35	**Log-Likelihood:**	-60074.
No. Observations:	4878	**AIC:**	1.202e+05
Df Residuals:	4858	**BIC:**	1.203e+05
Df Model:	20		
Covariance Type:	nonrobust		

We can see that the adjusted R-squared of the model is 85.2%. The adjusted R squared is a measure of goodness of fit of the model, in-sample.

We also try to look at the correlation between the explanatory variables. The correlogram shows the presence of explanatory variables with a high level of correlation which means we should expect some overfitting. Finally the baseline model’s R-squared on the test set is around 38.20%, as opposed to 85.3% on the train set.

IV. References

1. Alkhammash EH. An Optimized Gradient Boosting Model by Genetic Algorithm for Forecasting Crude Oil Production. *Energies*. 2022; 15(17):6416. https://doi.org/10.3390/en15176416

2. Ibrahim NM, Alharbi AA, Alzahrani TA, et al. Well Performance Classification and Prediction: Deep Learning and Machine Learning Long Term Regression Experiments on Oil, Gas, and Water Production. *Sensors (Basel)*. 2022;22(14):5326. Published 2022 Jul 16. https://doi.org/10.3390/s22145326