Prediction of the cumulative well production for the first year using Machine Learning

Olga Brusova

Problem statement

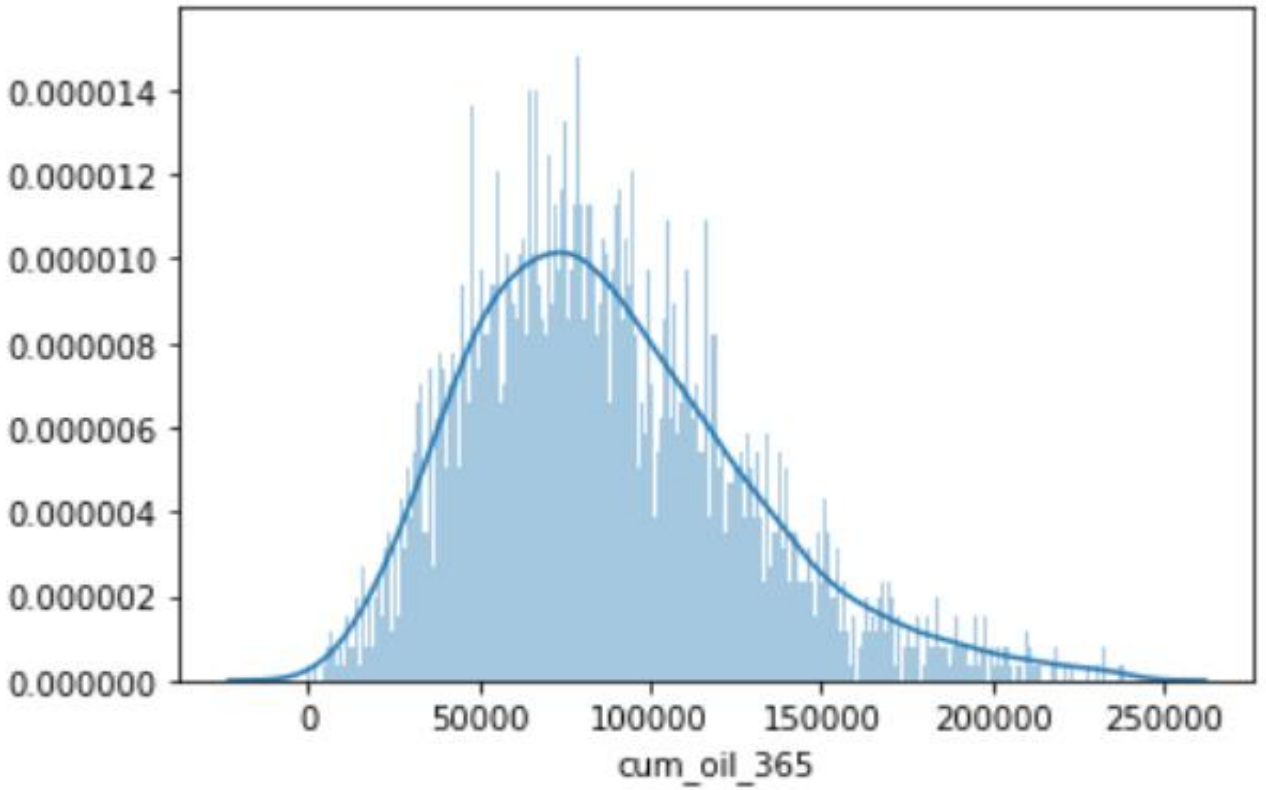
- The ultimate goal of any oil/gas company is to produce hydrocarbons as effectively as possible
- Oil production optimization is a very complex problem since there are many parameters that effect the overall production rates
- Finding the optimal combination for production controls is not straightforward and usually requires a lot of experience in the field
- ML can help to find a function that will predict oil production from controls

Target Variable

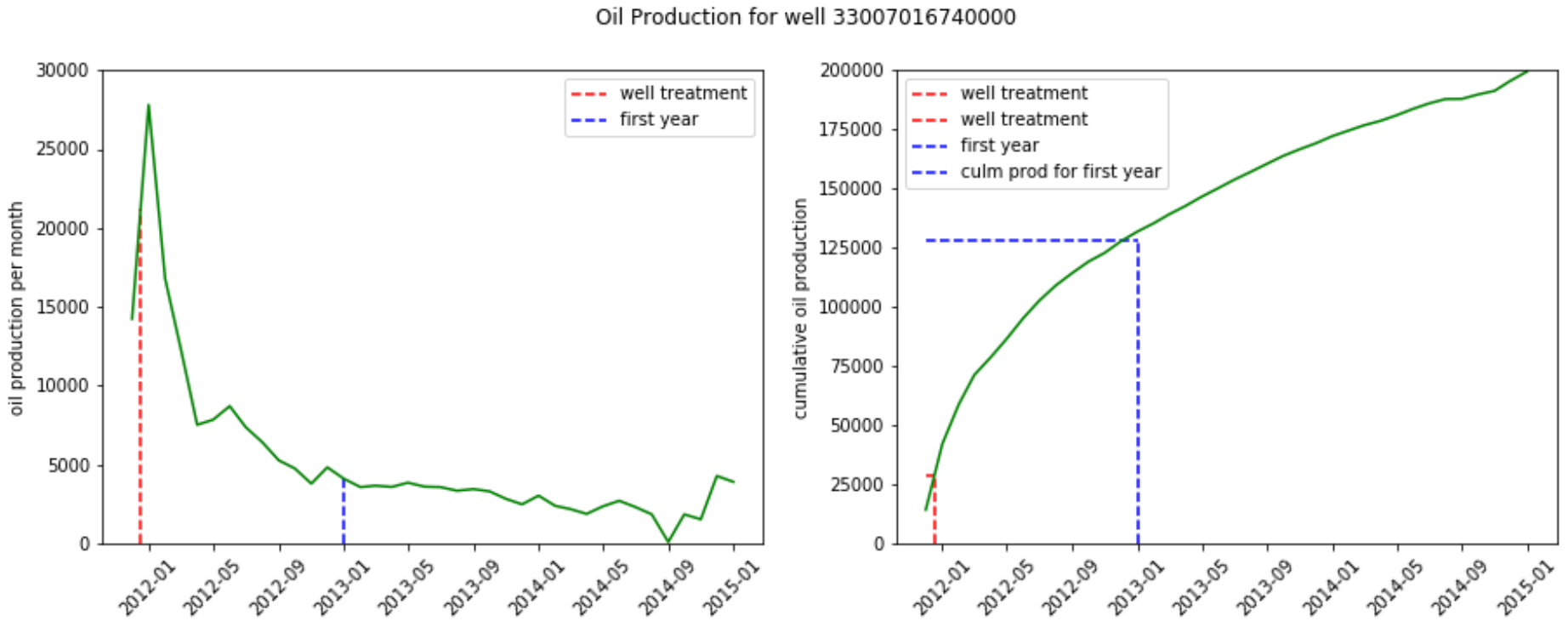
- We are seeking to predict cumulative oil production for a well for the first year of production (cum_oil_365)

- The histogram shows the distribution of cum_oil_365 for all the wells in our dataset

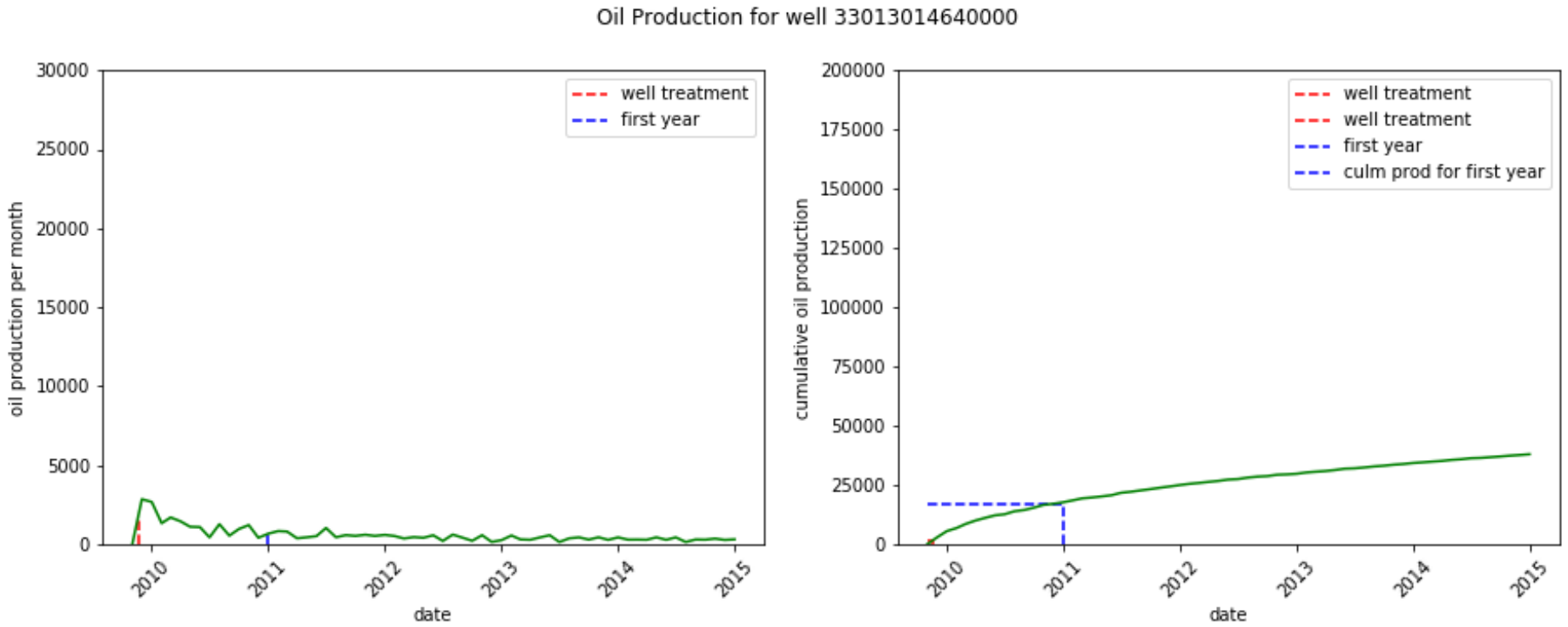
- The distribution is skewed to the right with a long tail of good producers



Example of Oil Production for a “good” producer



Example of Oil Production for a “bad” producer



Features

Oil production from a well depends on many aspects that include but not limited to the following:

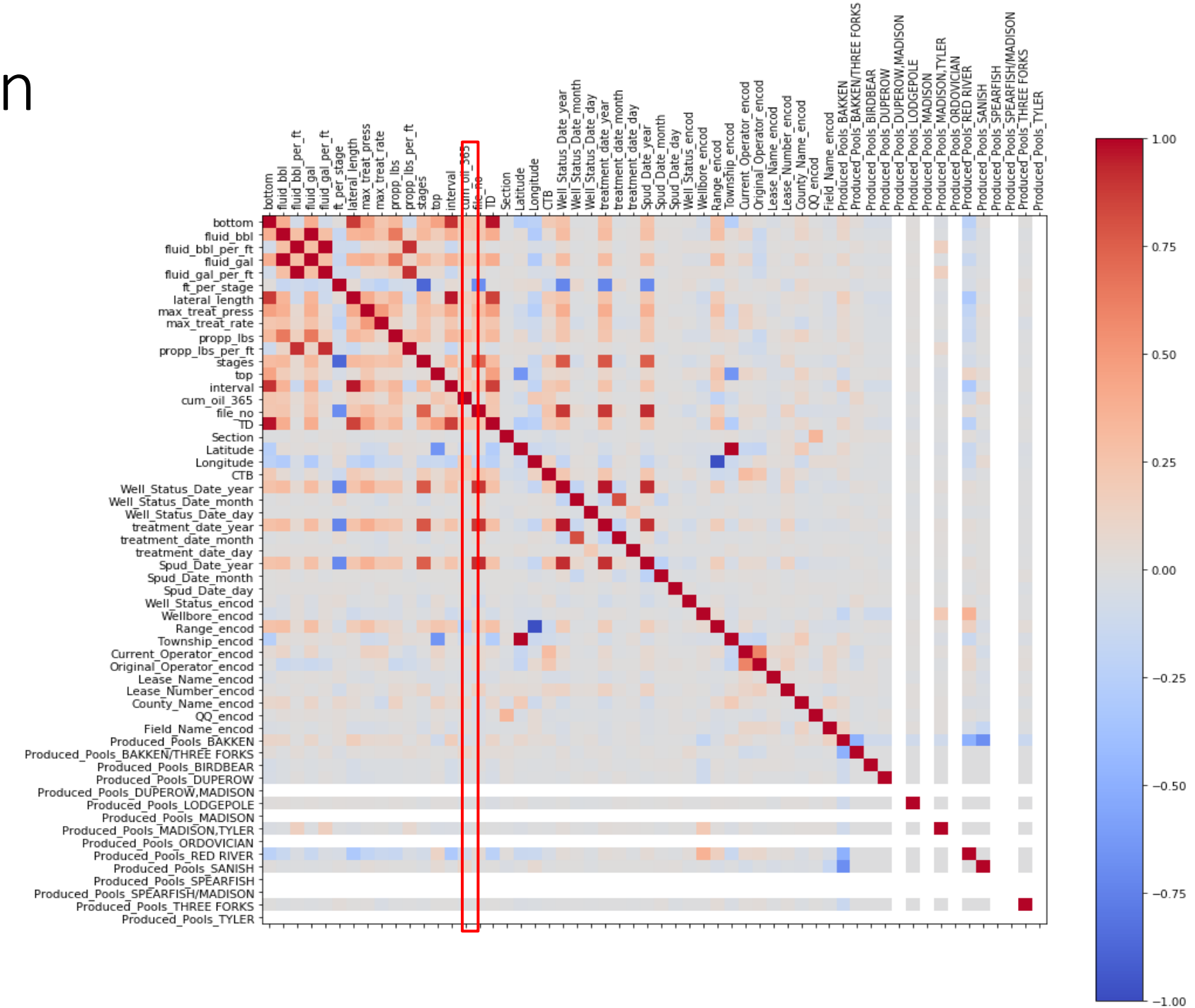
- Geology of the subsurface (producing interval, interval thickness, rock properties or producing and surrounding rocks, etc.),
- Well bore specifics (location, direction/deviation)
- Drilling parameters
- Completion parameters
- Perforation parameters, etc.

Features correlation with Production

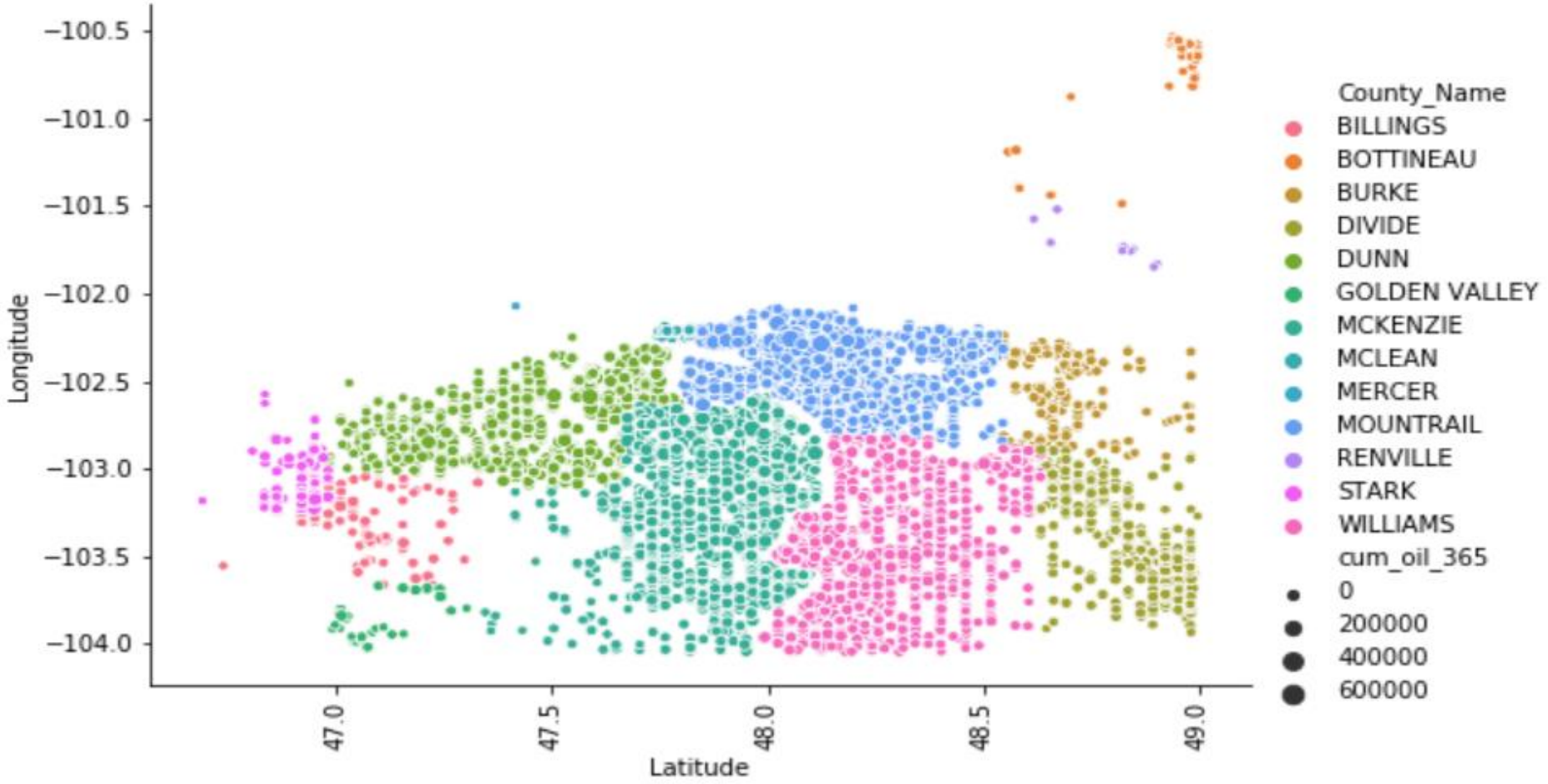
Many features show very week correlation with oil production and can be ignored:

- Dates: stud date, status, treatment
- Course location (section, township, etc)

Some features are empty and can be removed

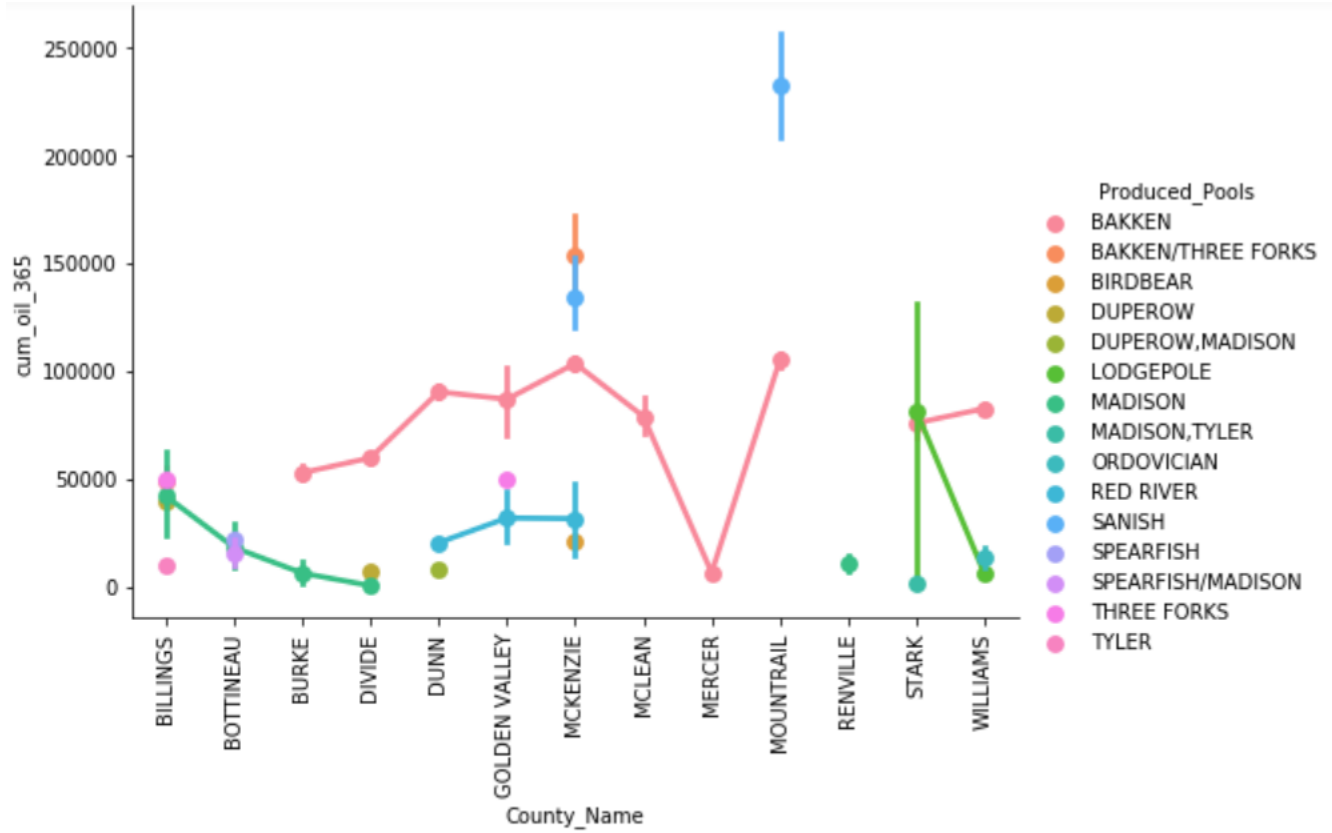


Well location correlation with Production



Production is highest in the northern and centrals parts of the field
Longitude and Latitude are important features

Well location correlation with Produced Interval

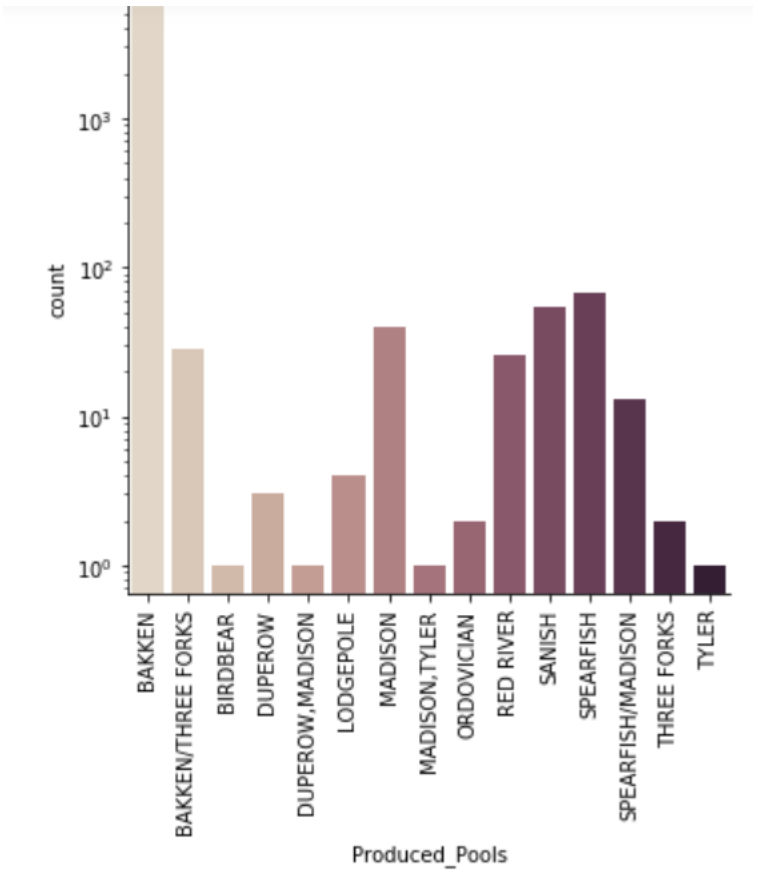


Producing formations (pool) has a great effect on overall well performance.

Some formations are correlated with excellent well performance, others are not producing well

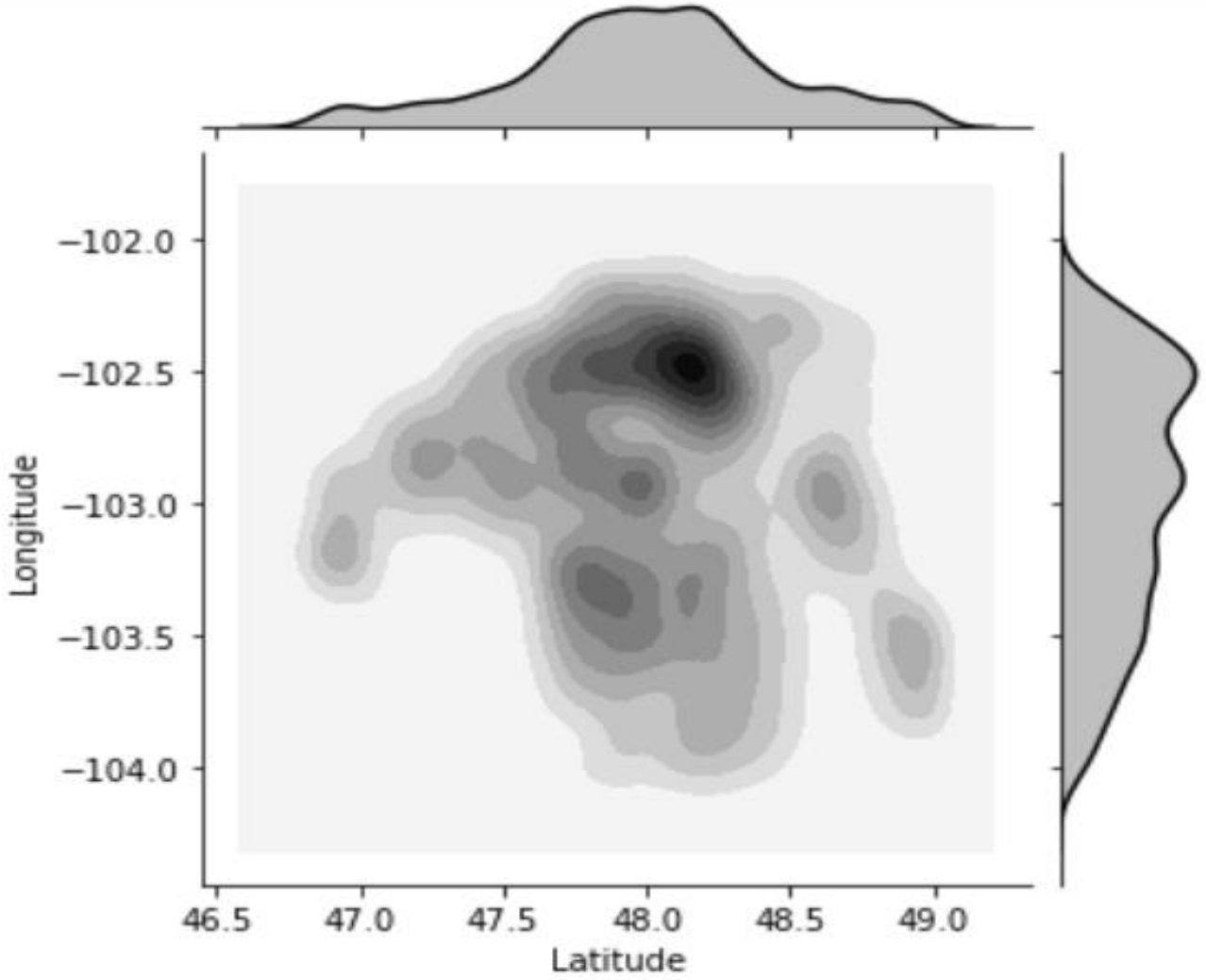
There also can be variation in well performance from the same producing interval (e.g., BAKKEN formation)

Relative Production from Different Intervals



BAKKEN formation is the dominant producing interval in the area.

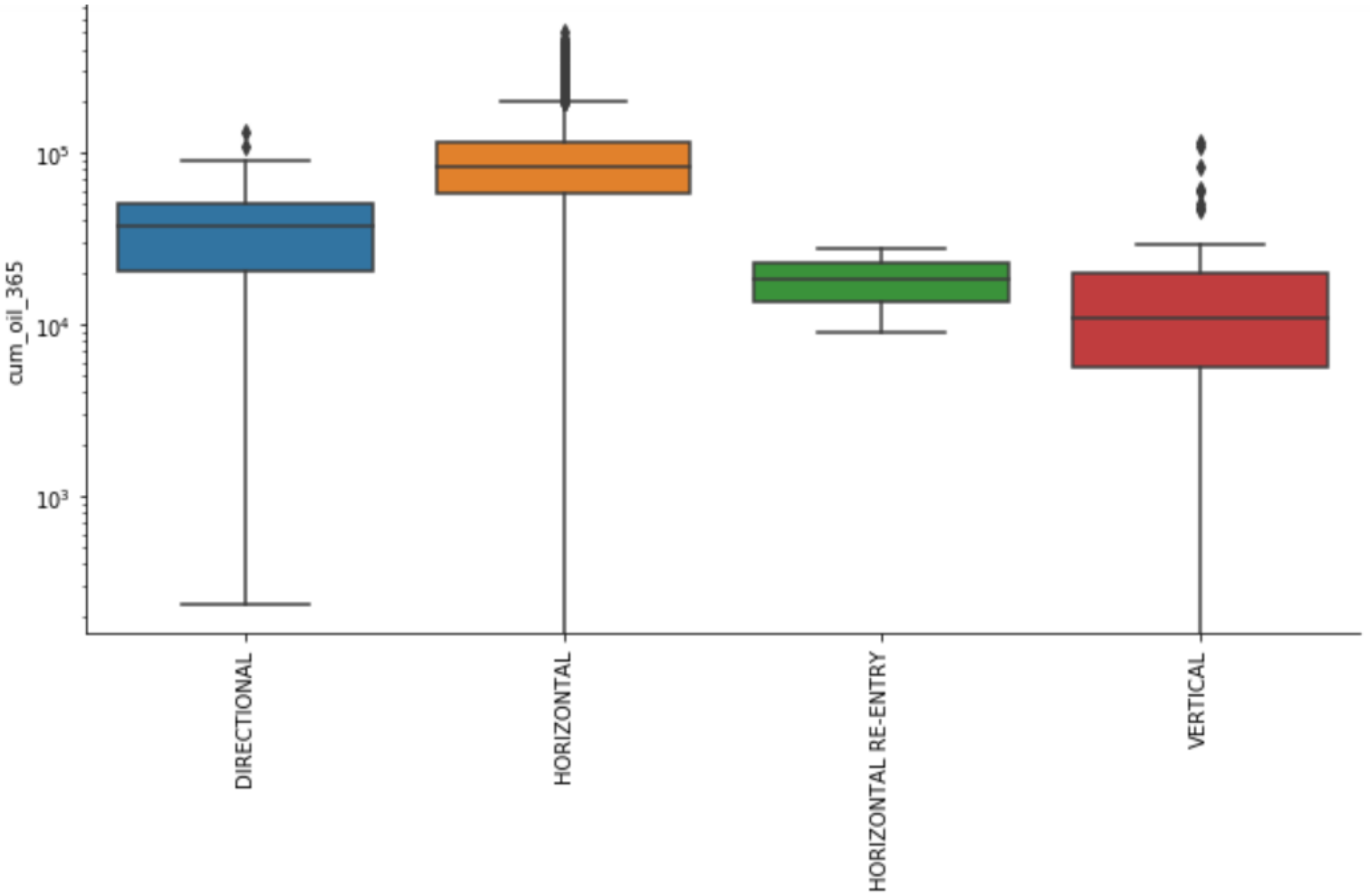
Best Production from Bakken Interval



BAKKEN formation is the dominant producing interval in the area.

Production is highest in the northern and centrals parts of the field
Longitude and Latitude are still important features

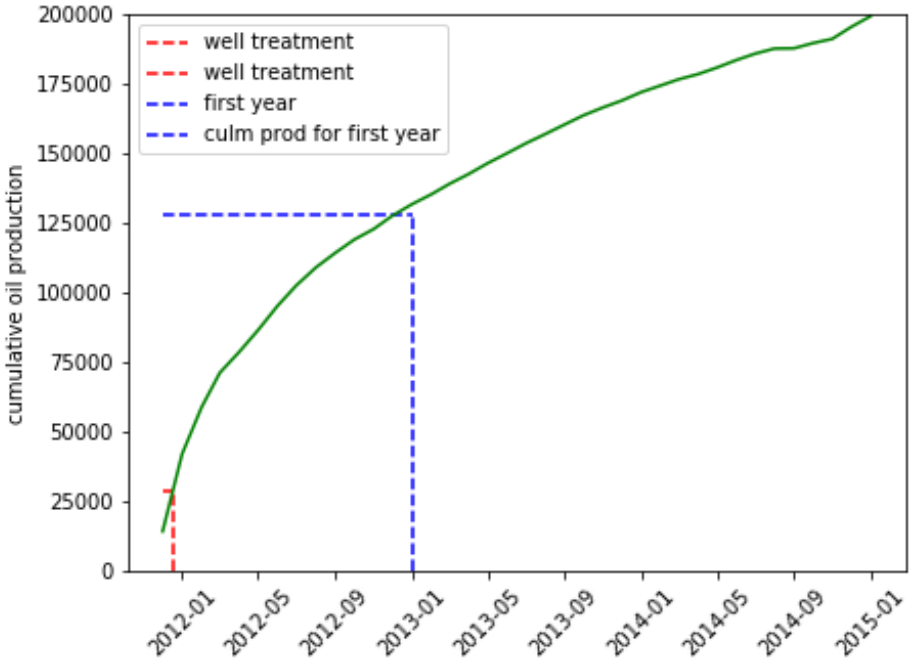
Well drilling direction vs. Production



Horizontal wellbores result in better production

Vertical wells are the worst producers on average

Example of Feature Engineering

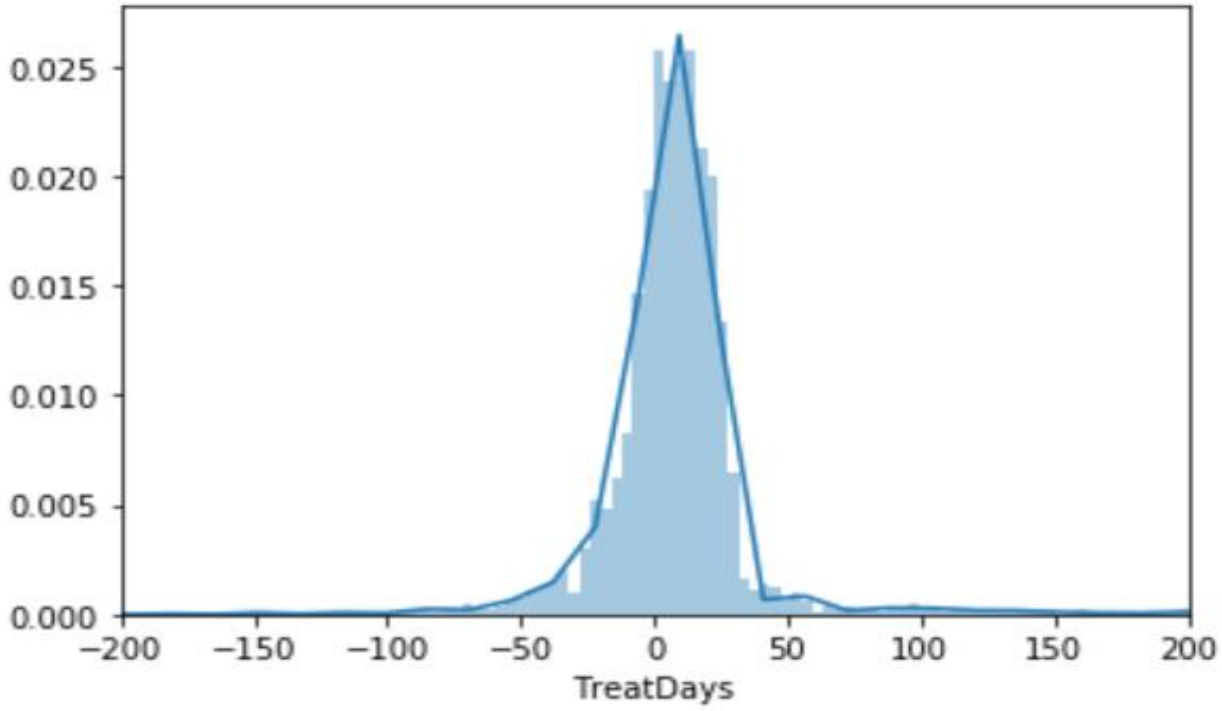


All the wells in the dataset has well treatment jobs

The date of the treatment may be an important variable that effect the overall cumulative well production for the first year

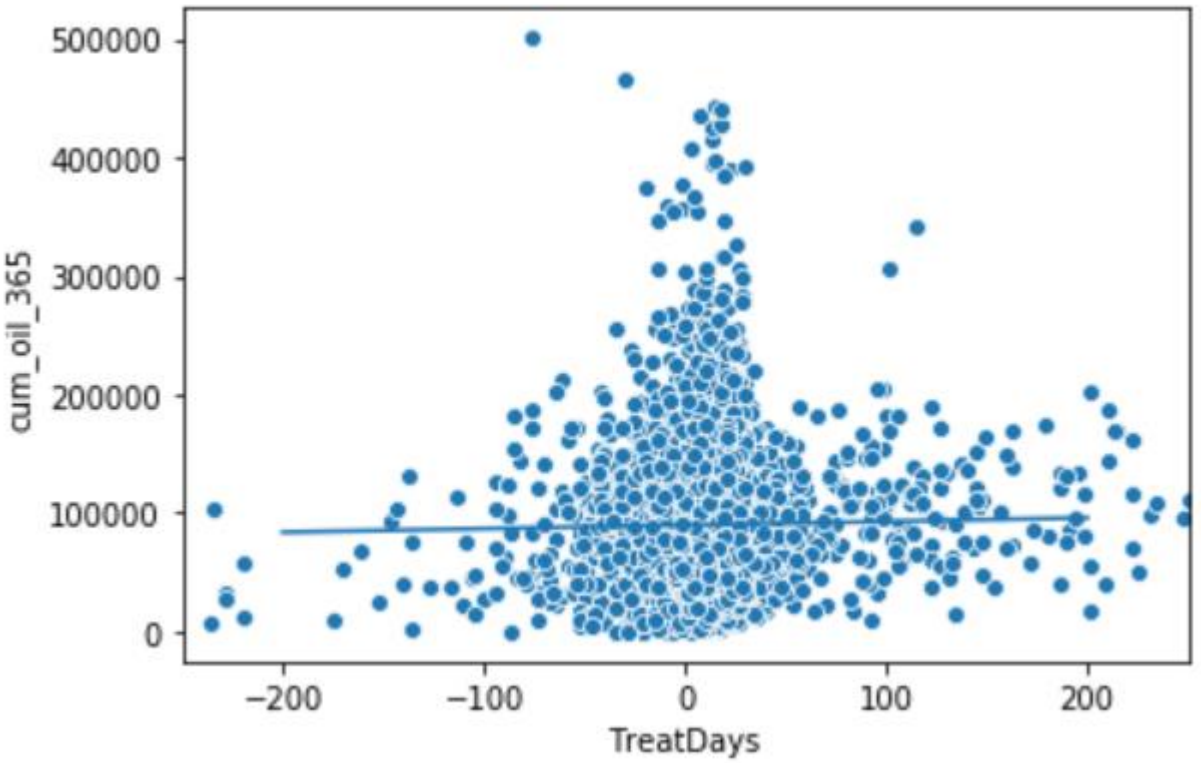
Hypotheses:

Timing of well treatment



Most wells have treatment job withing the first two months of production onset.
Almost half of the wells had treatment jobs preceding the production (negative values on the histogram).

Oil Production and the timing of well treatment



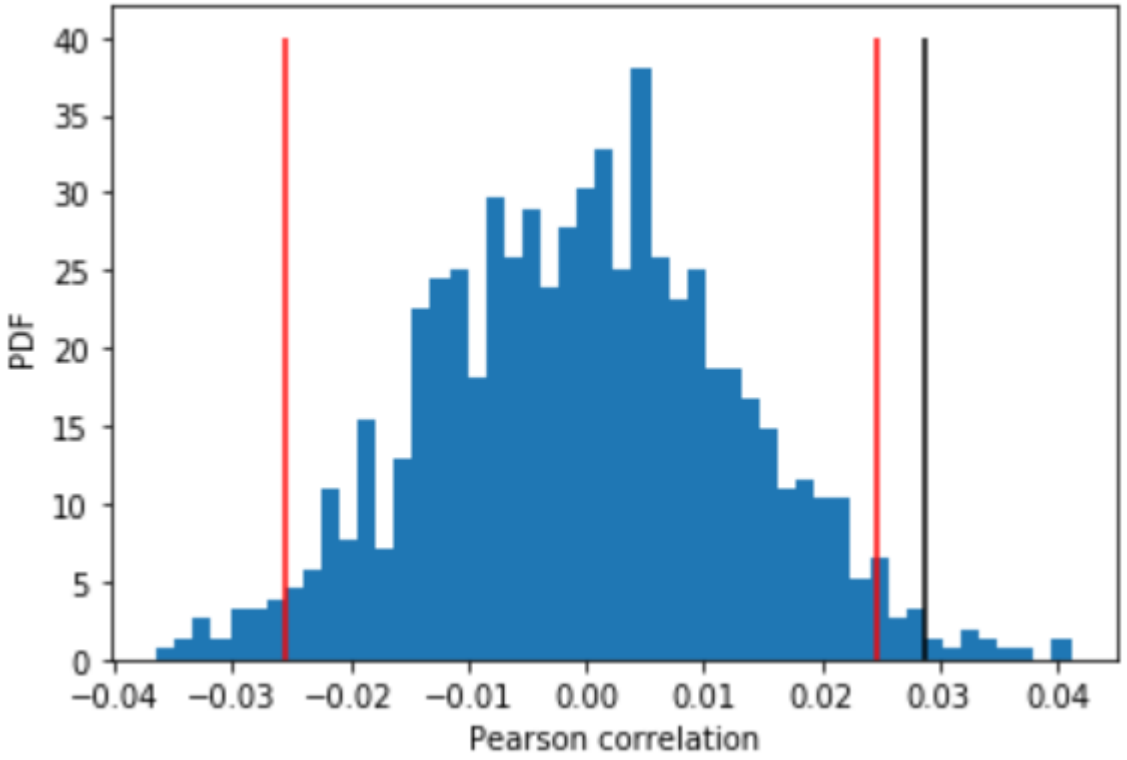
The two variables shows that they are not strongly correlated.

Blue line show the linear regression. The pearson coefficient is 0.028 which indicates no correlation.

Is this a statistically significant conclusion?

Our null hypothesis is that these two variable are not correlated. The observed correlation between TreatDays and cum_oil_365 may just be by chance.

Significance test using bootstrapping



1. Simulate the data assuming the null hypothesis is true. (permute the cum_oil_365 but leave the TreatDays values fixed). This simulates the hypothesis that they are totally independent of each other.
2. For each permutation, compute the Pearson correlation coefficient and assess how many of your permutation replicates have a Pearson correlation coefficient greater than the observed one.
3. The plot shows the distribution of pearson coefficients for permutation replicas and an observed one from the data as a back vertical line. 2.5% and 97.5% confidence intervals are shown in red vertical lines.
4. The p-value is 0.012. It is quite small and suggests that our hypothesis is likely to be true. Cumulative oil production for the first year does not depend on when the treatment of the well happened.

Regression Models with Hist Gradient Boost