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

Problem statement:

Many industries are faced with optimization problems where processes, methodologies, and actions are sought to be as effective as possible with balance between minimal cost and resources required, and best quality and performance. Oil and Gas industry is no exception. The ultimate goal of any oil/gas company is to produce hydrocarbons as effectively as possible. That usually requires production control of different reservoir fluid types (oil, gas, and water). While maximizing oil production, one might want to minimize income of reservoir brines.

Oil production optimization is a very complex problem, since there are many parameters that effect the overall production rates. Oil production from a well depends on many aspects that include but not limited to the following: geology of the subsurface (interval thickness, rock properties or producing and surrounding rocks, etc.), well bore specifics, completion and perforation parameters, etc. For a human it is a problem with too many degrees of freedom (tens and even hundreds of parameters). Finding the optimal combination for all these controls is not straightforward and usually requires a lot of experience in the field. Moreover, production controls may be quite different between different h/c fields. The same parameter that is associated with production increase in one field may hinder it in another.

This problem is ideal for Machine Learning. When trained on historical production data it is possible for an algorithm to find a function that will predict oil production from controls. This can be an extremely useful tool for new well planning as well as quality control of existing wells.

Description of the dataset

The dataset for this project is publicly available and consists of three .csv files containing production data and some of the control variable.

'Well-index.csv' file lists all the wells in the dataset and some of the meta data associated with each well such as location, operation company, well type, producing interval, etc. Each well in the dataset has a unique identifier (API).

In order to produce from a tight formation, wells are fractured hydraulically by injecting pressurized fluids and other additives. Information about well treatment procedures is given in ‘completion.csv’ file. It describes the amount and type of fluid injected, number of injection zones (stages), depth interval, etc. Each well in the database has API for reference.

‘Monthly-production.csv’ contains well production data for different fluids over some period. The rows in the dataset are associated with a reported month of production. Number of days that a well was producing in that month given in a separate column along with the volumes of fluids produced. I will be predicting a well production for the first year. This number can be derived from ‘monthly-production.csv’ for each well.

Although, the available dataset contains quite a few features that we can use to predict cumulative well production for the first year, these features are either quite general (e.g., metadata) or specific to engineering (e.g., treatment design). We are lacking any features related to the properties of the subsurface that our wells penetrate. These are important features since they describe the capacity and other properties of the reservoir that the wells are producing from. Lack of geologic features may not allow us to predict oil production accurately.

For this project I had the input data in three different .csv files. They had to be analyzed and combined into a since file/dataframe with features and lables.

First, I created the oil production metric that I’ll building a model for. I chose to predict well cumulative oil production for the first year (I called it cum_oil_365). I had to engineer this parameter from the data provided in ‘monthly-production.csv’ file. The file contains multiple records (rows) for each well. Each row reports production volumes for different fluids for each month as well as number of days the well was producing during this month. The task was to find how much oil each well produces for the first 365 days.

There were 8939 unique wells in the database originally. After dropping wells that do not have oil production reported I’m left with 6307 wells.

The next step was to prepare feature data for the wells that have production data. Well parameters and measurement are contained in two different datasets: ‘completion.csv’ and ‘well-index.csv’.

First, I loaded ‘completion.csv’ file into a pandas data frame. I dropped columns that I believe not important (e.g. file number) and assigned correct types to dates and categorical data.

Data from ‘well-index.csv’ file was also loaded into a data frame. Some categorical data (e.g. Well Status data) had “Confidential” values that I treaded as missing data and replaced with empty string for further processing.

After merging all three datasets on the API (unique well id) key I have 46 feature variables and 6440 records of which 5987 are unique wells. This implies that some wells have multiple records.

Removing duplicate rows resulted in 6433 records.

Next, I split the dataset into two. The first one has only wells with single record per well. There are 5727 of such wells. The rest of the data must be worked on more carefully. I cannot apply the same aggregation methods to all the features in the dataset with multiple rows per well. Some columns it makes sense to sum, some to average, others to make extreme values (min or max). The aggregation method also depends on the particulars of a well treatment (e.g., the treatment repeated for the same interval in the same zones/stages by pumping more fluids, more stages were created in the same interval, the new interval treated and opened for production that does not overlap with an old one, the new interval treated and opened for production that overlaps with an old one.). After carefully analyzing each scenario and applying appropriate aggregation methods I reduced 706 record to 260 and merged them back to the single record data frame.

There are three columns with dates in the dataset: Well_Status_Date, treatment_date, Spud_Date. Each column was split into 3 different columns of year, month, and day.

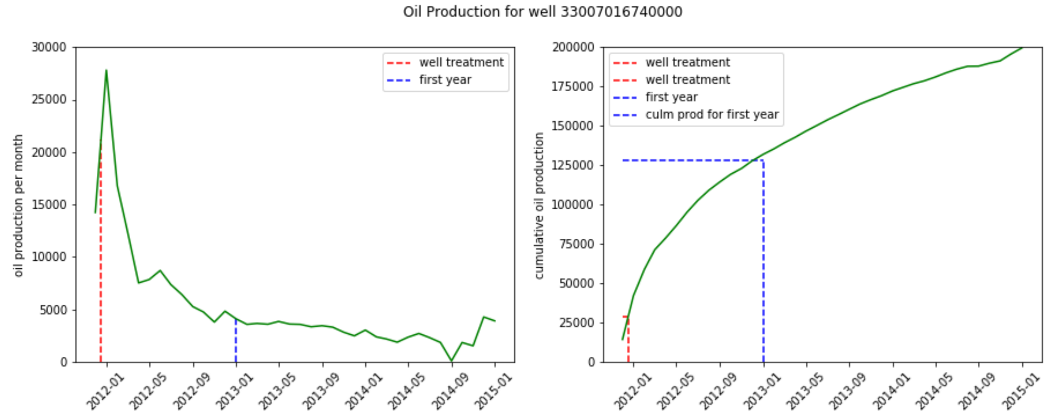
Next, I removed outliers for some of the variables if the value was different from the mean by three standard deviations. Data points with infinite or none values were replaced by the averages.

The next step in data prep is to encode categorical data. For majority of variables I chose simple encoding scheme by replacing a string in a category with a number. This is the case for any variable that has too many categories or if values in the variable are dominated by one category. The only categorical variable I chose to do one hot encoding is Produced_Pools (geological zone from which a well is producing).

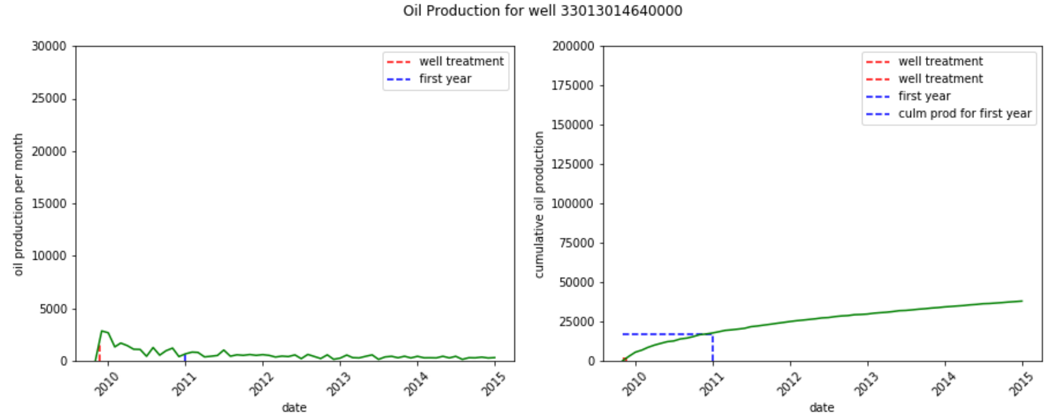
Initial findings from exploratory analysis

First, I need to visualize the oil production for different wells to get more familiar with the quantity I am trying to predict. The raw data is given as oil production for each month. I am simplifying the problem a little by predicting cumulative production for the first year. It is a good metric for overall well performance and can be used to separate good wells from bad ones.

The following plots shows the production profile for a good performing well. The left plot shows monthly oil production and the right plot shows the cumulative production over the same period. The blue line indicates the time stamp for the first year of production. The red line is the month of well treatment.

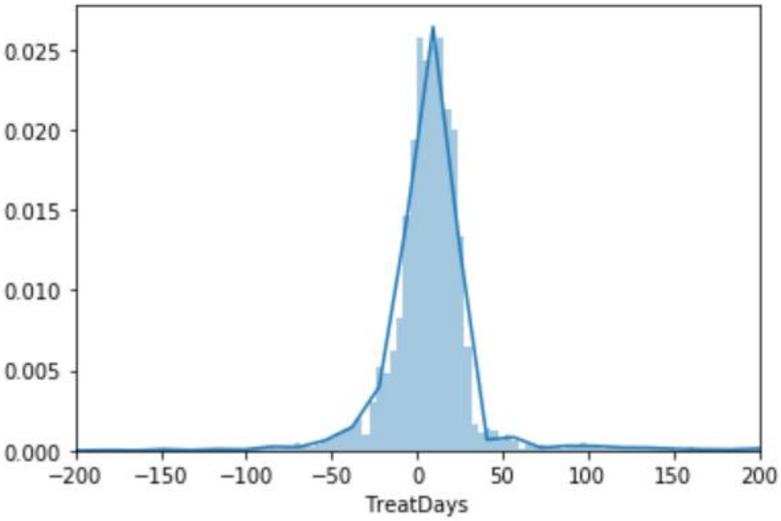


Similar plot is created for a low performing well.



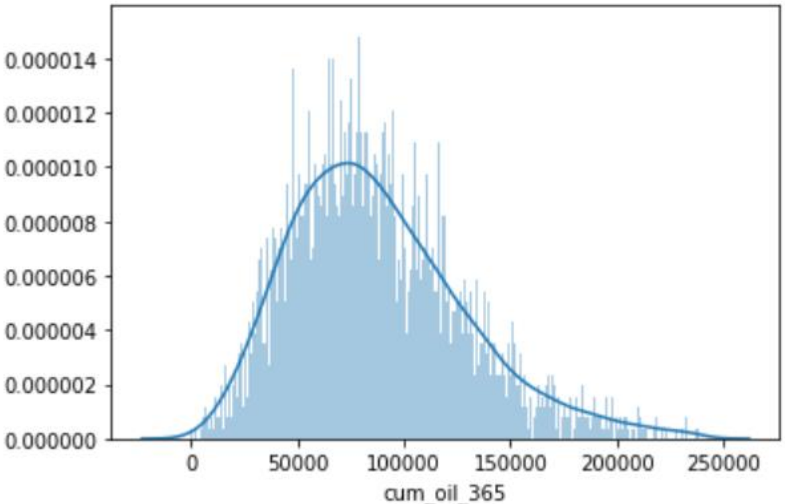
One interesting question to investigate based on these plots is theater or not well performance depends on how fast the well treatment was performed after the well was brough on production. The intuition tells us that the sooner well treatment is performed the better well production should be. I expect a positive correlation between number of days between first well production and well treatment and cumulation production for the first year.

First let’s plot a histogram of number of days between first well production and well treatment.



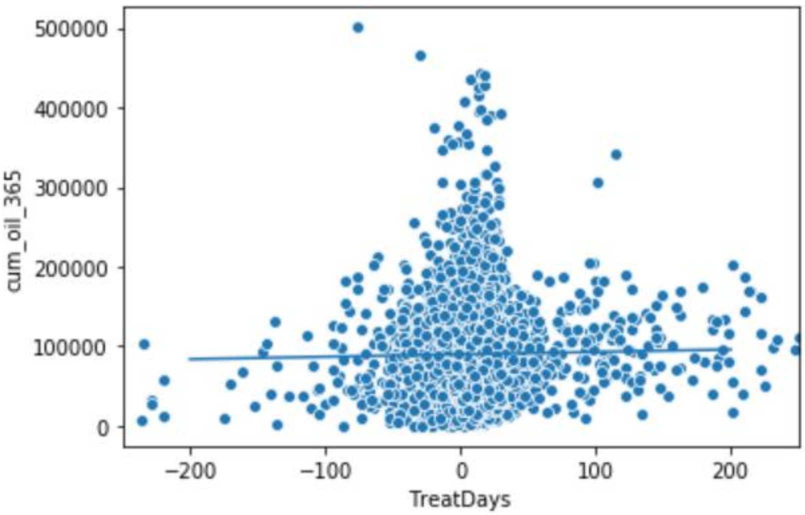
We can see that most wells have treatment job withing the first two months of production onset. It is interesting that almost half of the wells had treatment jobs preceding the production (negative values on the histogram).

The histogram of cumulative oil for the first year of production is well behaved distribution skewed towards the right-hand side.



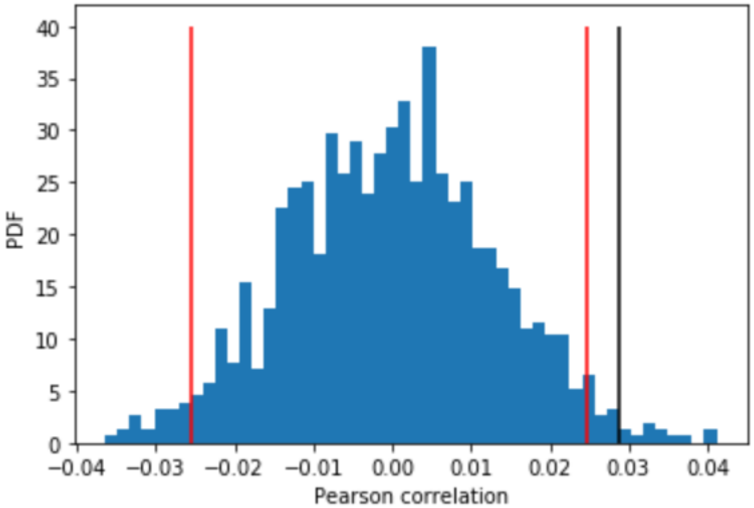
We can investigate if there is really a significant dependency between the cumulative production of the well during the first year and how fast the well treatment was performed on the well.

The cross plot of the two variables shows that they are not strongly correlated. Blue line show t he linear regression. The pearson coefficient is 0.028 which indicates no correlation.



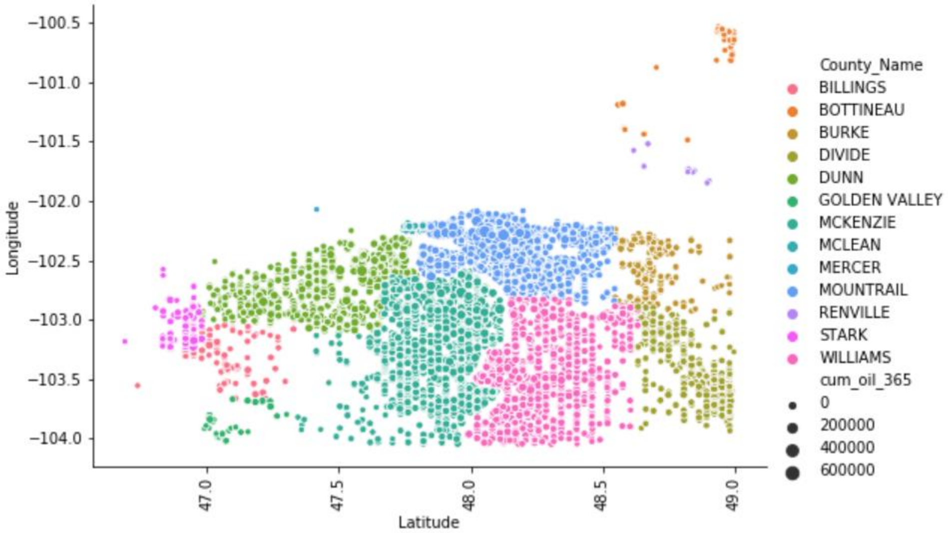
Our null hypothesis is that these two variable are not correlated. The observed correlation between TreatDays and cum_oil_365 may just be by chance.

To do the test, you need to simulate the data assuming the null hypothesis is true. To do so, we need to permute the cum_oil_365 but leave the TreatDays values fixed. This simulates the hypothesis that they are totally independent of each other. 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.
The plot below 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.

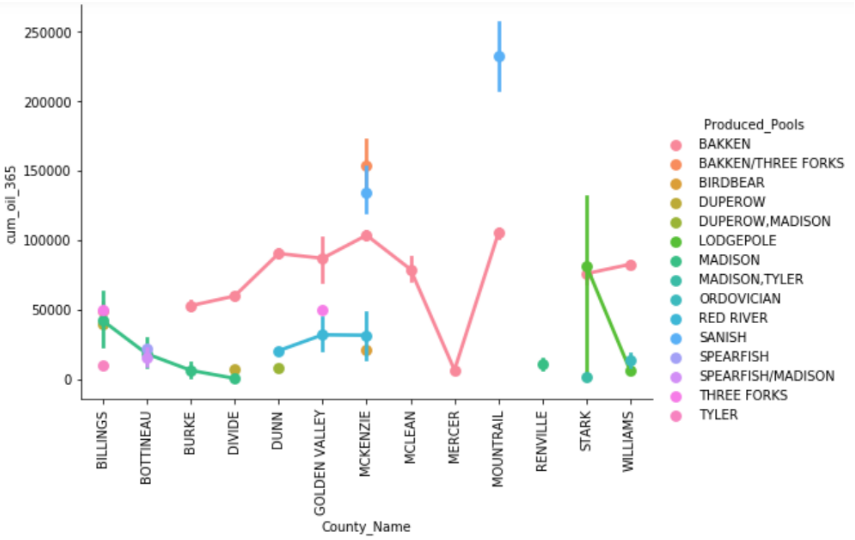


The p-value is 0.012. It is quite small and suggests that our hypothesis is likely to be true. Cumulative oil production for the first year does not depend on when the treatment of the well happened.

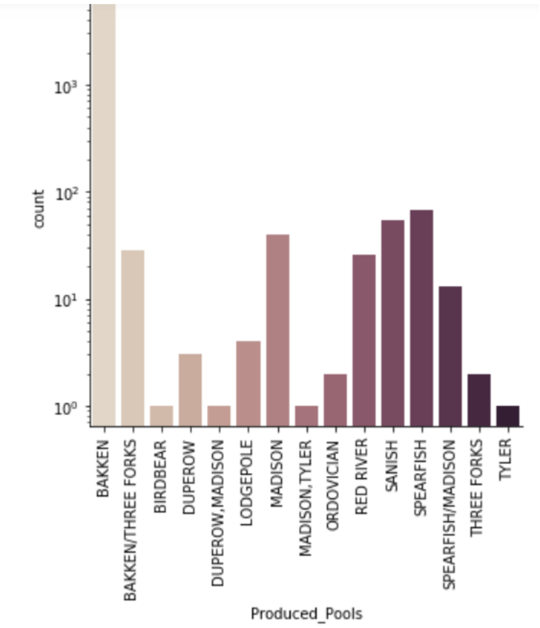
Other interesting observation can be made from exploratory analyses. The plot below shows the geographical position of all the wells, colored by the county name. Bigger circles indicate bigger cumulative production for the first year. This picture does not discriminate production interval.



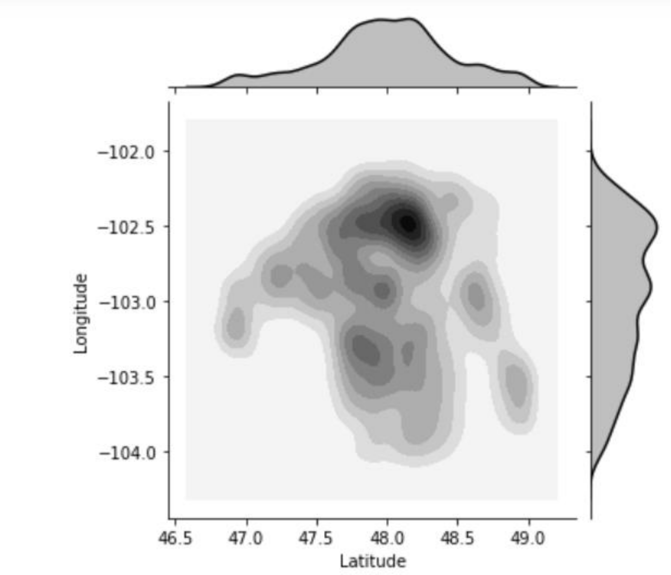
From the figure below we can see that producing intervals may play important role for production prediction. Some zones have shown to have better production than others.



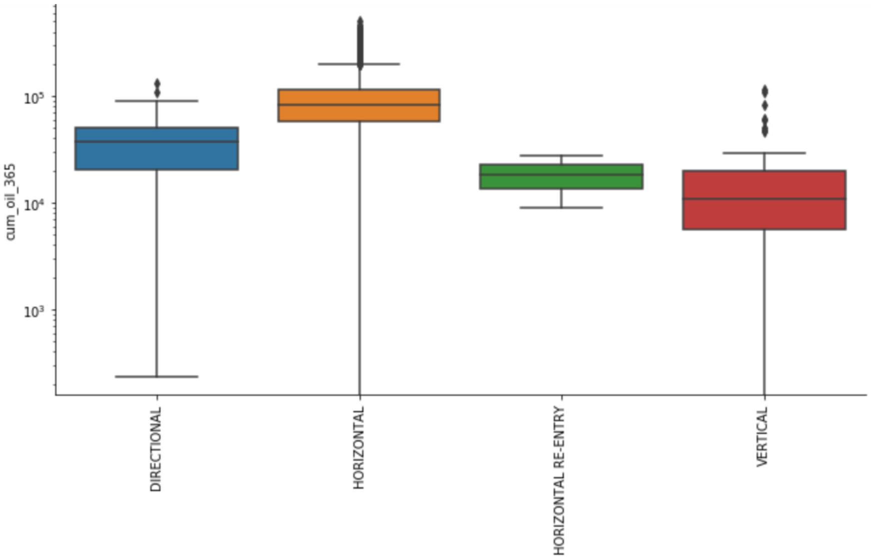
Next plot shows that most of the wells are producing from the BAKKEN interval.



BAKKEN production is highest in the north.



Wellbore direction is also important for increasing production. Vertical wells traverses the producing interval perpendicular and are the least productive. Horizontal wells are drilled parallel to the strata boundaries and are most productive as they intersect more of the productive interval.



From the above analyses we can see that lateral position and depth/interval of a producing well may play an important role in production profile.

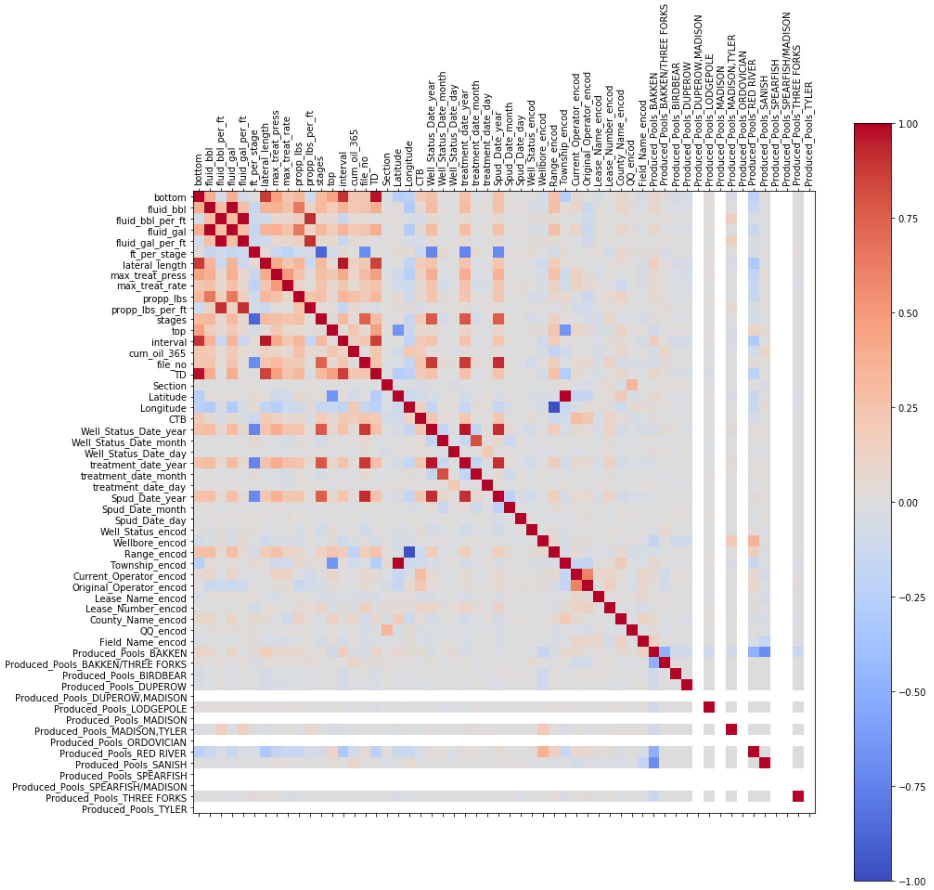
Results and In-depth analysis using machine learning

In the section I will explain the process of ML model building and optimization.

I start with a closer look at the features that I generated at the previous step. As seen from the correlation matrix below. Not all the features in our dataset will be useful for our model. The selection of the features for the model comes from the domain knowledge and experiments with the model prediction as we hold away one feature at a time.

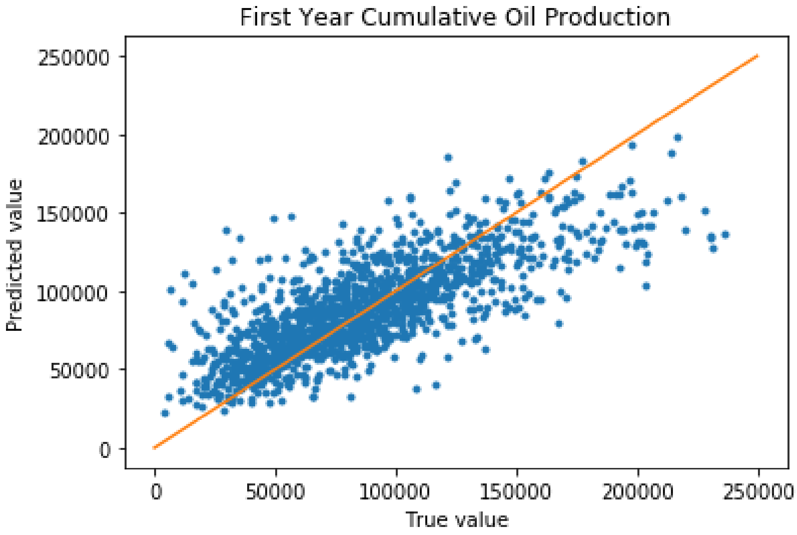
If I ran a random forest classifier with all mostly default set of parameters and all the features, I get RMSE of 28523. The best prediction for this dataset is around RMSE of 27000 RMSE. But experimenting with input features I was able to reduce RMSE to 27203 on validation set.

Almost all the features that I removed from the input should not be correlated with oil production. These include dates (completion, treatment, status check). Some variables are repeated (e.g., expressed in different measurement units). Location variables such as township and section names are redundant as we have better variables to describe position of each well in space (latitude and longitude). Lease names/numbers also should not have any relationship to production as well as total depth of a well. Empty variables were also dropped.



Once the feature space was defined, I experimented with a few regressors: random forest, XGBoost, Gradient Boost, and Hist Gradient Boost.

The last one gave the best model performance. I have achieved RSME of 27034 with vanilla Hist Gradient Boost regressor. The figure below shows the graph of true well performance vs. predic ted well performance. The solid line is 1:1 correlation (perfect model).



Closing Remarks

I was able to build a sensible model that predict cumulative production of a well for the first year of production based on some generic well parameters as well as some information on well location, well design, and well treatment. The models can be used qualitatively to identify good producers in the area and to avoid bad ones. In order to achieve quantitative accuracy, I will need additional information about the details of subsurface around the well's producing zone.