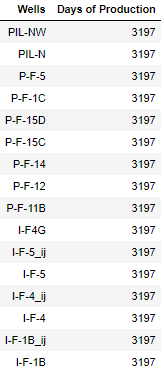
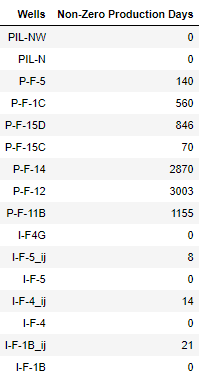
Keenan Flynn

4161 Project 2

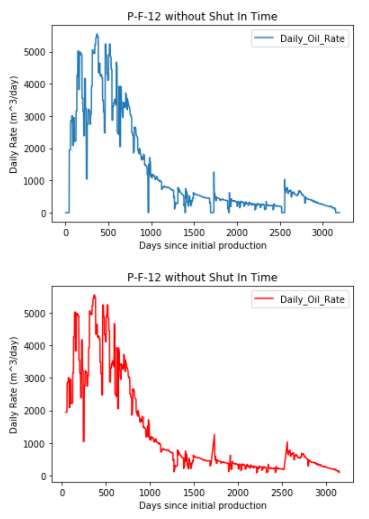
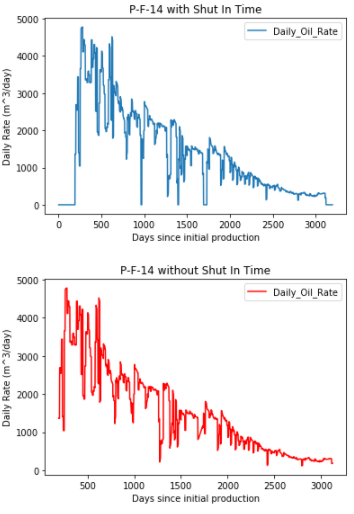
Volve Field Time Series Exploration, Analysis, Machine Learning (LSTM)

The Volve Field lies off the coast of Norway in the North Sea. This region is known for offshore oil production platforms which will be the focus of this project. The Volve Field data set is a time series excel workbook with 16 wells corresponding to individual sheets, each with more than 50 attributes. Time was uniform across each well, with a starting time on January 1, 2008 and a final time of October 1, 2016. Data was given for each day, equivalent to 3197 data points for each well.

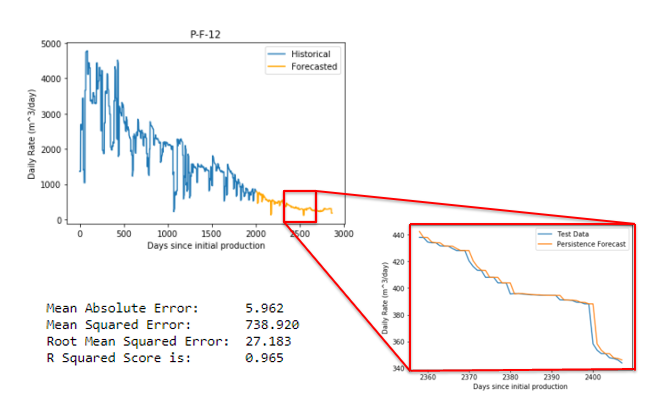
The focus of the study was narrowed down to a single attribute, Daily Oil Production Rate (m^3/day), because this has the largest impact on the economics for oil wells. For most wells, Daily Oil Rate was frequently zero, meaning that the first step in the Data Exploration process was to find which wells were viable to be examined. 2 wells were chosen to be studied, the P-F-12 and P-F-14 because they had the highest amount of Non-zero Production Days.

The P-F-12 and P-F-14 were then visualized through time series plots. Each well can be seen below in 2 separate graphs. The first graph includes ‘shut in’ time or zero time and the second is the production without shut in time. The data set for each well will be analyzed without shut in time. This allows for a smoother curve

After Data Exploration, analysis was run on Well P-F-12. P-F-14 will be saved for LSTM validation. P-F-12 was then modeled through a Persistence Forecast. The Persistence Rorecast is where the observation from the prior time step is used to predict the observation at the current time step.

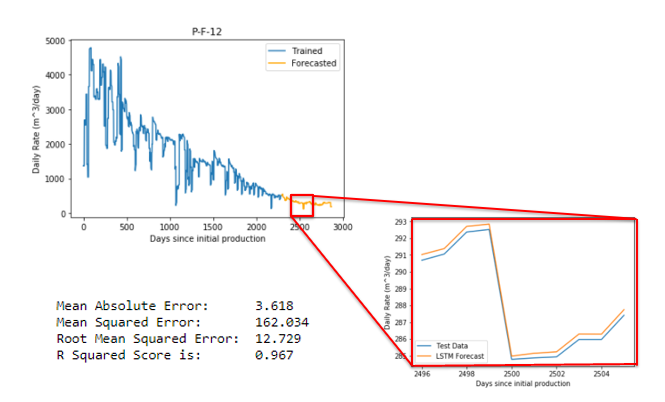


The Persistence Forecast has several limitations. The biggest drawback is that it has a horizon of only 1 time step. Ideally, we would like a model that can predict further into the future. The forecast is very good in sections that have little change, but not so great in sections that have large vertical changes. This can be seen above at around 2400 days. In this case the forecast has a very good R squared score and a RMSE of 27. In the next model we will try to lower the RMSE.

LSTMs or Long Short-Term Memory networks are good at forecasting time series data. This is due to their nature as Recurrent Neural Networks (RNNs). These types of neural networks have the ability to learn and remember over long sequences. This will be a good model to forecast our oil production data.

The first step in LSTM is to turn our time series data into a supervised learning problem. The LSTM need an input value and an output value for every time step. We can achieve this by setting the input as the data from the previous time step and the output as the current time step. After this the data needs to become stationary, meaning that we need to remove the decreasing trend of the data. This will give a more accurate model. This can be achieved by differencing the data, or subtracting the previous time step from the current time step. Finally, the data must be scaled from -1 to 1 for the activation functions of the LSTM to work. This was done through SciKit Learn’s MinMaxScaler function. It is important to note that each of these transformations also has a inverse function so that we can get meaningful results from the model.

After data transformation, the LSTM model was fit. The data was split into training and testing data with a batch size of 1. For the purposes of time only 300 epochs and 4 neurons were used. The model outputs were inversely transformed and then plotted as seen below.

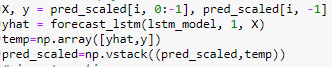


The LSTM forecast shows a very good R squared score and the RMSE has been lowered to 12.7 compared to the Persistence forecast. The mean squared error has also been significantly decreased. One worry that I have with the model is that the R squared is very high. I was also unable to explain how the LSTM model would have a broader horizon than Persistence. Let’s take a closer look at the code given from the shampoo model.



This code is from the for loop that is forecasting the test data (orange line). ‘yhat’ is created from the forecast\_lstm line, which in turn uses data from X. X is the output value from the last time step in the testing data, meaning that every LSTM forecast only has a horizon of 1 because it sees the old test data. I believe that this is why the model looks so similar to the test data. It would be useful to see how the horizon could be increased, however it could not be done in allotted time.

After fitting the LSTM model on P-F-12, I tried to validate my results by forecasting well P-F-14. The same methods were used with some modification. The scaling function had to be changed so that it returned an entire dataset rather than a train/test split. A new array was created called pred\_scaled with an initial value being the same as the very last scaled value from the historical data. I then started forecasting pred\_scaled in the loop and appending the LSTM predicted value back onto pred\_scaled. In this way I could try to increase my horizon and forecast the wells production based on no prior knowledge.



To summarize my results the model predicts a sharp sudden increase of production, which is not realistic or probable. There is still some work to do but I am happy with where the progress that I made, as the code went through many trials and errors. I was able to see the usefulness of the algorithm and hope to keep learning more about it as I continue in my studies.

