Ideation and modelling approach on the test project shared for SOTA OG

1. Theory / Fundamental Concepts –

a. Business Objective:

The gas well has sensors that feedback data at every step of the pipeline. The usual pipeline steps are reservoir *pressure, tubing pressure, casing pressure, choke point, position, line pressure, and combine point*. The data captured at each step of the pipeline is utilised for a real-time dashboard and for making future predictions based on historical data.

2. Understanding Dataset –

a. Data Understanding:

As stated above, there are variables that are related to the state of gas flow at that timestamp, like *gas metre temperature, gas metre static pressure, and gas metre differential pressure,* and there are other variables associated with data points at the pipeline level.

b. Descriptive Understanding and EDA

One month of data with variables (bottom hole pressure, tubing pressure, casing pressure, line pressure, gas meter pressure, temperature, etc.) is being shared, where each data point is spaced over a 10-minute interval. We will filter the data when *line pressures* of both gas wells are within the range of 530–630 psi.

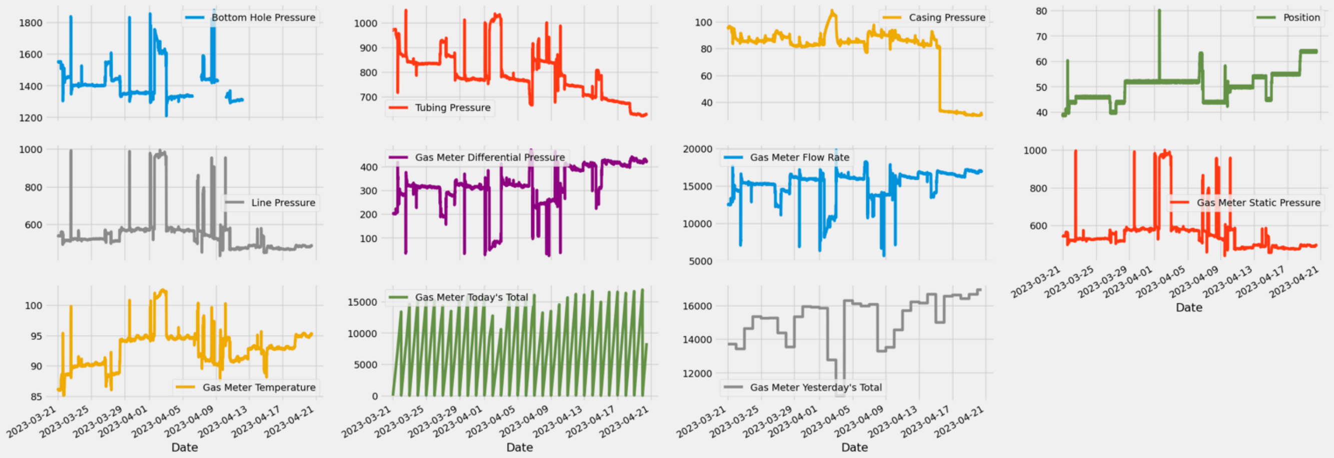
Some pointers from the EDA of the dataset -

- *Bottom Hole Pressure* has some null values; we may impute them or aggregate them based on our approach to data sampling (downsampling to hours or days).

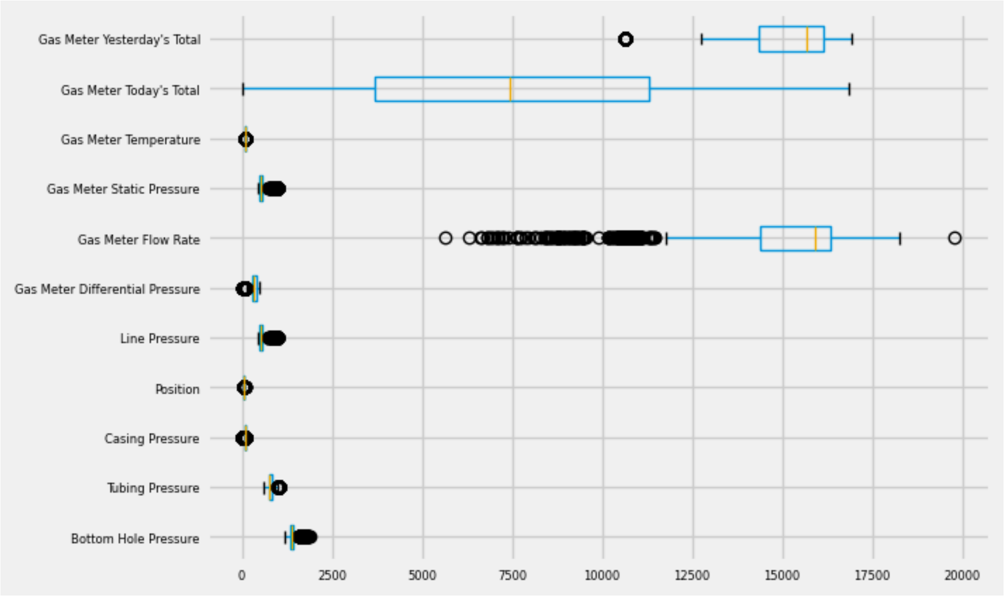
- The rest of the columns floats (numeric); however, the temperature, cumulative total, and total columns have different ranges of data and units. Thereby, it is necessary to deal with different unit data by either scaling using the MinMax Scaler or

- The below plots are based on raw dataset with date as index, i.e., the dataset has not been resampled for days or hours and no aggregation have been performed for exploratory data analysis.

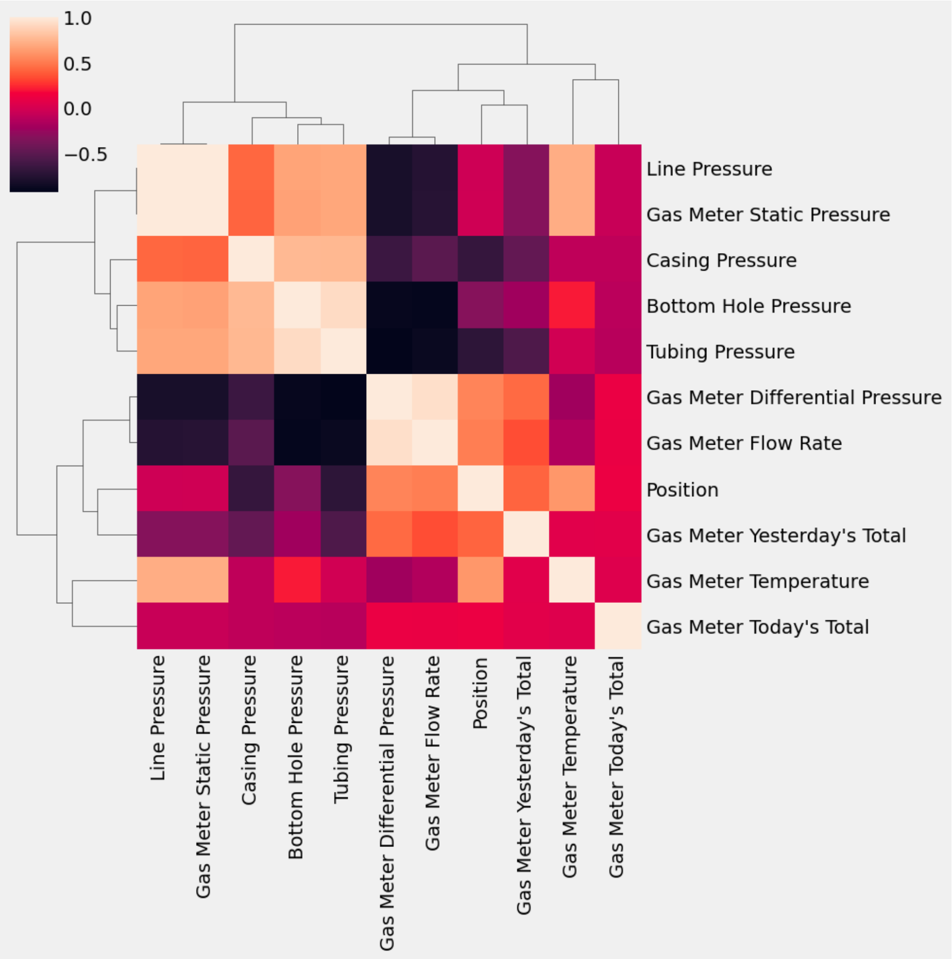
```
# Facet plots of the one of the gas well dataset
```



```
# Interesting box plot for the Gas Meter’s Today’s Total and Yesterday’s Total
```



```
# Cluster Map on correlation matrix
```



c. Questions

i. Is *bottom hole pressure and reservoir pressure* the same thing? As in the dataset, we have a field for bottom hole pressure. Can I assume it to be the same?

ii. The dataset has column '*Position'*," and I am assuming it is the same as choke position. Looking at the preliminary data, this is the choke position, but the variable name is just 'Position'.

iii. Choking points influence the values of *reservoir pressure, tubing pressure, casing pressure, and line pressure.* So, is it safe to assume that choking position is the most important feature in the dataset that influences each of the other independent variables mentioned and the dependent variable (gas flow rate), which we want to predict? However, we can further confirm from the correlation matrix.

iv. The combine point has '*junction pressure*', which by way of engineering principles influences the gas flow rate in both gas wells, as there is no variable for such in dataset, can we calculate it by formula and include it in modelling?

v. Can we define gas flow rate as "the amount of gas being passed over the 10-minute interval from a pipeline? Hence, if we aggregate hourly, then for 1 hour of gas flow rate = (10 minutes * 6)60 minutes, and the gas flow rate will be the sum of all the gas flowed through that hour, i.e., summation of 6 datapoints over that hour? What is the unit of gas flow rate in the dataset?

3. Modelling Approach 1 –

a. Problem Framing:

We can use choke position, and other variables associated with gas meter readings and prepare a multivariate time series forecasting (Encoder – Decoder LSTM Forecast Model] model case where 't-1' timestamp values for these variables helps identify the output variable at 't' timestamp.

Here the output variable we are trying to predict is the combine gas flow rate (sum of gas flow rate of well 'x' and well 'y') based on different choke position of well 'x' and 'y' as well other features such gas meter temperature, gas meter static pressure, gas meter differential pressure.

So, choke position of well 'x', gas flow rate of well 'x', choke position of well 'y' and gas flow rate of well 'y' at timestamp of lag (t-1) can be used to predict the combine gas flow rate at time stamp 't'.

We can then pull all the unique values of choke position between 0 to 100 for both well 'x' and well 'y', along with paired gas flow rate for each well and store in the processed output file with gas flow rate of well 'x' and well 'y'.

b. Data Preparation:

We can downsample the data from 10-minute interval to days and aggregate the gas on daily total, which is available to us in dataset as a variable for both 't-1' and 't' time sample. *'Gas Meter Today's Total' and 'Gas Meter Yesterday's Total'.*

The input sequence thus obtained can be used by LSTM to learn and extract. The size of input or lag could be 2 Days to 7 Days or any prior number of days. The choice of this defines how the input data should be prepared for training, and test data evaluation.

However, we can also ideate and try the downsample data to hourly as well and aggregate gas flow rate over that hour and follow the same approach as above on deciding the number of lags for input we can to consider for our model to learn and extract relationship.

These choices define a few things:

- How the training data must be prepared in order to fit the model.
- How the test data must be prepared in order to evaluate the model.
- How to use the model to make predictions with a final model in the future.

We will also scale the data based on MinMax Scaler as some of the units in dataset are not same

c. Evaluation Metric:

As every data point is numeric (float) with unit of gas flow rate (cubic meter/second), it would be ideal to have evaluation metric of MAE or *RMSE*. We will use RMSE and calculate the error in prediction to actual.

We will use the efficient *Adam* implementation of stochastic gradient descent and fit the model for certain number epochs and a specified batch size [These two choices also depend on the data preparation, if we go with Daily aggregate or hourly aggregate].

d. Train and Test Split:

If we prepare the data based on daily gas flow rate for each of features/variables of gas pipeline, we can fit the model on 3 weeks of data and evaluate on remaining 1 weeks of data. We can alter this approach based on the downsample aggregation we choose in the data preparation stage.

e. Algorithm: *Vanilla LSTM*

Long Short-Term Memory network, or LSTM for short is a type of RNN and is quite successful in time series forecasting because it overcomes the challenges involved in training a recurrent neural network, resulting in stable models. In addition to harnessing the recurrent connection of the outputs from the prior time step.

LSTM recurrent neural networks are capable of automatically learning features from sequence data, support multiple-variate data, and can output a variable length sequence that can be used for multi-step forecasting.

We can develop a model with a single hidden LSTM layer with 200 units (let’s try this and test accuracy). The LSTM layer is followed by a fully connected layer with 100 nodes (let’s try this and test accuracy) that will interpret the features learned by the LSTM layer.

Finally, an output layer will directly predict the output gas flow rate combined.

We can try with different architecture, or different framework such prophet or CNN-LSTM, ConvLSTM

4. Modelling Approach 2 –

The choice of modelling here differs by the fact that we will model the gas flow rate individually for each of gas well and finally aggregate and extract the choking position.

However, there could be some underlying relationship which will not be captured in this approach, as independent models will produce gas flow rate based on choking position of one gas well, while we need to model it for both the gas well together as they could both influence each other.

We can try this approach with same data preparation, evaluation metric, train – test split and algorithm to check results. Here, just clarifying the problem framing clearly, rest of the steps in modelling remains same for this approach.

a. Problem Framing:

We can use the choke position, and other variables associated with gas meter readings and prepare a multivariate time series forecasting (Encoder – Decoder LSTM Forecast Model] use case where ‘t-1’ timestamp values for these variables helps identify the output variable at ‘t’ timestamp.

Here the output variable we are trying to predict is the individual gas flow rate based on different choke position of well ‘x’.

So, choke position of well ‘x’, gas flow rate of well ‘x’, at timestamp of lag (t-1) can be used to predict the gas flow rate of time stamp ‘t’ for well ‘x’.

The task is repeated for well ‘y’ as well.

This will tell us for different choke position of well ‘x’ and well ‘y’ what are the gas flow rate of well ‘x’ and gas flow rate of ‘y’. We can then add the gas flow rate of well ‘x’ and gas flow rate of well ‘y’.

Finally, we can then pull all the unique values of choke position between 0 to 100 for both choke ‘x’ and choke ‘y’ with associated gas flow rate of well ‘x’ and well ‘y’ and store in the processed output file with gas flow rate of well ‘x’ and well ‘y’.

Please provide your valuable inputs and comments with different color text