SAPIENZA
UNIVERSITÀ DI ROMA

Short-term water demand forecasting using traditional and deep learning models

Name: Poompavai Chandrprakash(1853976)

Course: Masters in Data Science

Advisor: Prof. Simone Scardapane

Co-Advisors: Dr. Pietro Dicosta

Outline

- Introduction
- Objectives
- Implementation
- Project Modules
- Results
- Conclusions

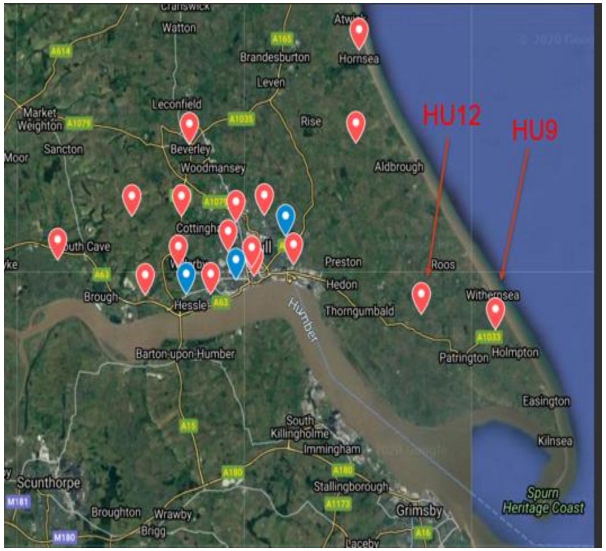
Introduction

- In today’s world, we could see enormous amounts of big data in every domain. The volume of data that one must deal with has exploded to unimaginable levels in the past decade.
- Private companies and research institutions capture terabytes of data about their users’ interactions, business, social media, and sensors from cell phones and automobiles. This leads to the evolution of big data analytics.
- Big Data Analytics largely involves collecting data from different sources and making it an available way for the consumer.
- Time series forecasting is one such domain, which uses enormous amounts of sensor data with the application of artificial intelligence, and data science. Analysing time series data with effective visualization can help to produce various insightful inferences.

Objectives

- The main objective of the thesis is to perform a comparative study on time series forecast of water demand using traditional and deep learning models, for planning the business in terms of seasonality, annual patterns production capacity, and expansion over a longer period, which in turn drives a long-term business strategy (e.g., plans to launch a facility or store internationally and expand into new markets).
- Two modules were performed as univariate and multivariate time series analysis.

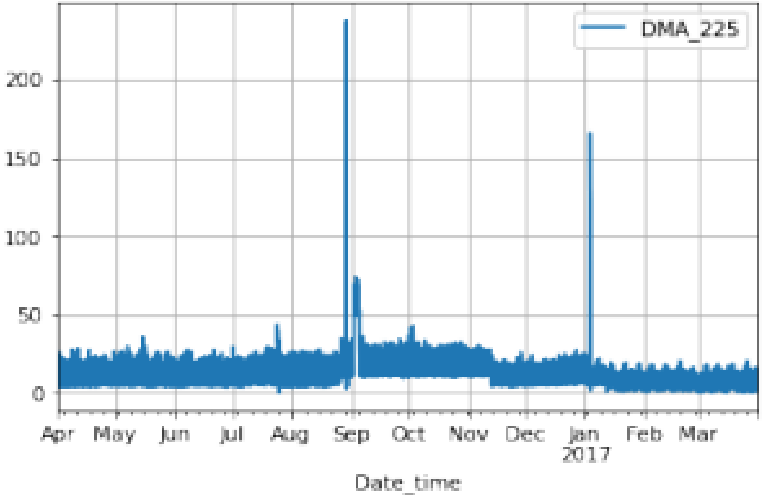
Dataset

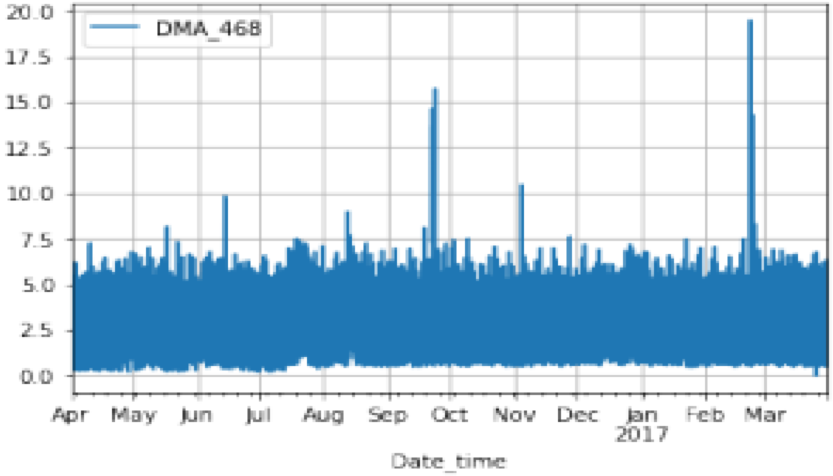


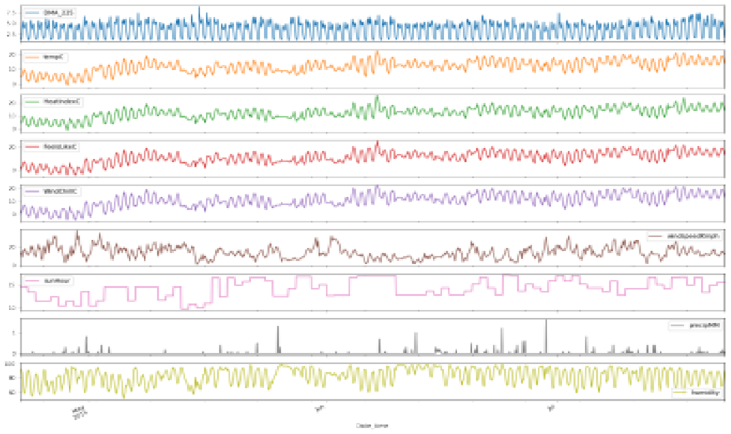
- Yorkshire dataset - Records flow of water every 15 minutes to each of the Distribution Management Areas (small clusters) across the region (DMA's). We considered two DMA's namely DMA 225 and DMA-468, in the HU1 location of the country. HU1 is a postcode area, located in the Hull postcode-town region, within the county of Yorkshire.
- Weather dataset - External weather factors are considered. WorldWeartherOnline is a company providing accurate and relevant weather data. The historical weather API provides hourly past weather for worldwide locations since July 2008.
- Univariate time series: One year of the Yorkshire dataset was considered (April 2016-March 2017 time intervals of 15 minutes).
- Multivariate time series: So online weather data and Yorkshire data were combined as the dataset, three months of data were considered. (April mid to July mid, time intervals of 15 minutes).
- Weather dataset link: https://www.worldweatheronline.com/developer/api/local-city-town-weather-api.aspx
- Water dataset link: https://datamillnorth.org/dataset/yorkshire-water-leakage-dma-15-minute-data

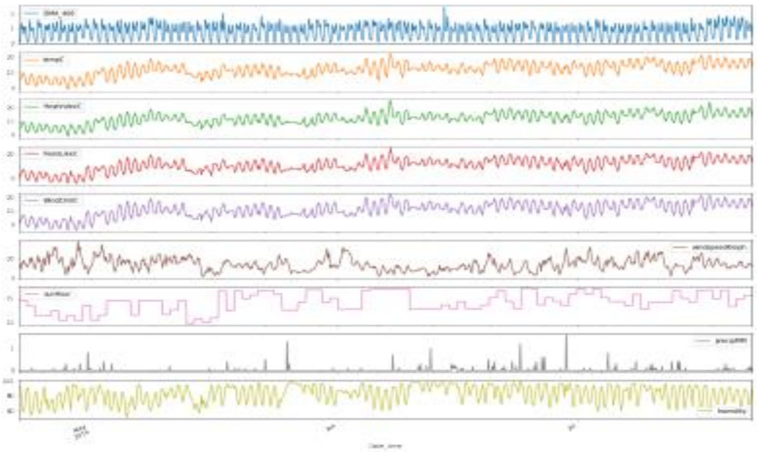
Data Plots

Below are the data plots of Univariate – DMA 225, DMA 468 ; Multivariate - DMA 225, DMA 468 respectively.









Models

Baseline persistence

It is a point of reference for all other modelling techniques on the problem.

Sarimax

Extension of ARIMA , explicitly supports univariate ts data with seasonality.

SARIMA(p, d, q)(P, D, Q, m)

Facebook prophet

Handles multiple seasonality, outliers. It is sum of three functions of time plus an error term: growth/trend g(t), seasonality s(t), holidays h(t), error term e(t) as below :

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

LSTM

Kind of RNN, which can learn long-term dependencies. Work on vanishing gradient problems and they have three gates namely input, output and forget gates. The cell plays an important role in saving and releasing information

Comparison

Traditional models

- Missing values can really affect the performance of the models.
- They cannot recognize complex patterns in the data.
- They usually work well only in few-step forecasts, not in long term forecasts.
- Example: AR, MA, ARIMA, VAR etc.

Deep learning models

- They are not affected by missing values. (e.g RNN).
- They can find complex patterns in the input time series.
- They can perform on long term forecasts.
- RNNs can model a sequence of data so that each sample can be assumed to be dependent on previous ones.
- Example: RNN, LSTM, GRU etc.

Project Modules

Univariate time series analysis

- Prediction: Hourly, Monthly, Weekly.
- DMA: 225,468.
- Models: LSTM, Fb prophet, Sarimax, baseline persistence model.

Multivariate time series analysis

- Prediction: Hourly.
- DMA: 225,468.
- Models: LSTM, Fb prophet, baseline persistence model.

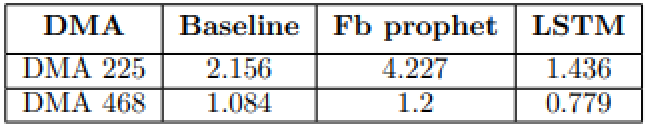
Accuracy metric

- Root Mean Square Error (RMSE/rmse): It is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. The smaller the value, the better the model's performance.

Results

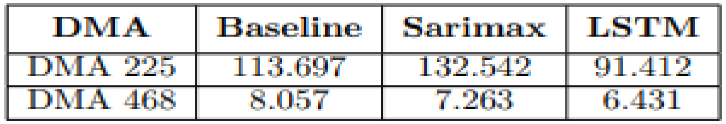
The below tabular column lists the rmse values of the models for the respective predictions.

Univariate hourly prediction



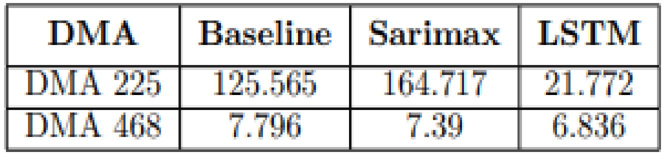
DMA	Baseline	Fb prophet	LSTM
DMA 225	2.156	4.227	1.436
DMA 468	1.084	1.2	0.779

Univariate weekly prediction



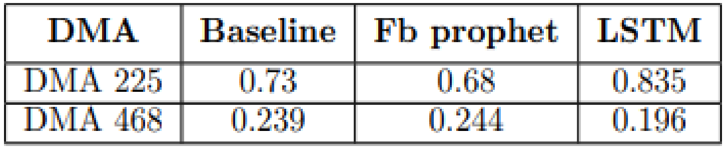
DMA	Baseline	Sarimax	LSTM
DMA 225	113.697	132.542	91.412
DMA 468	8.057	7.263	6.431

Univariate monthly prediction



DMA	Baseline	Sarimax	LSTM
DMA 225	125.565	164.717	21.772
DMA 468	7.796	7.39	6.836

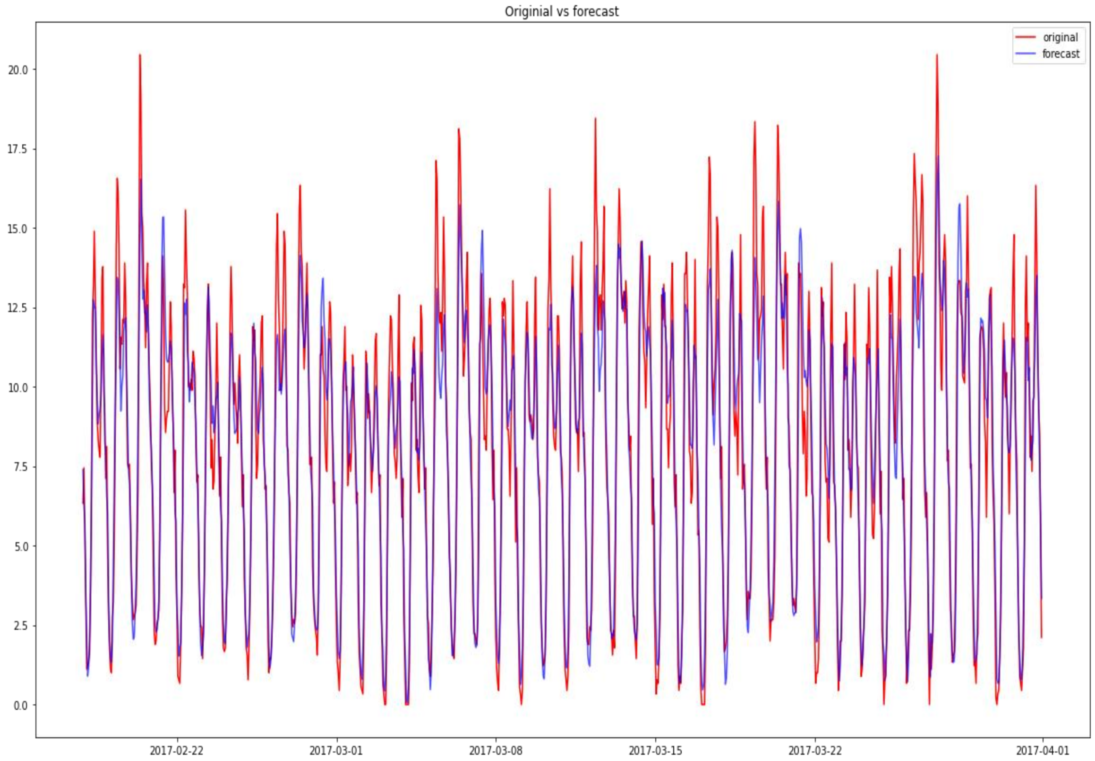
Multivariate hourly prediction



DMA	Baseline	Fb prophet	LSTM
DMA 225	0.73	0.68	0.835
DMA 468	0.239	0.244	0.196

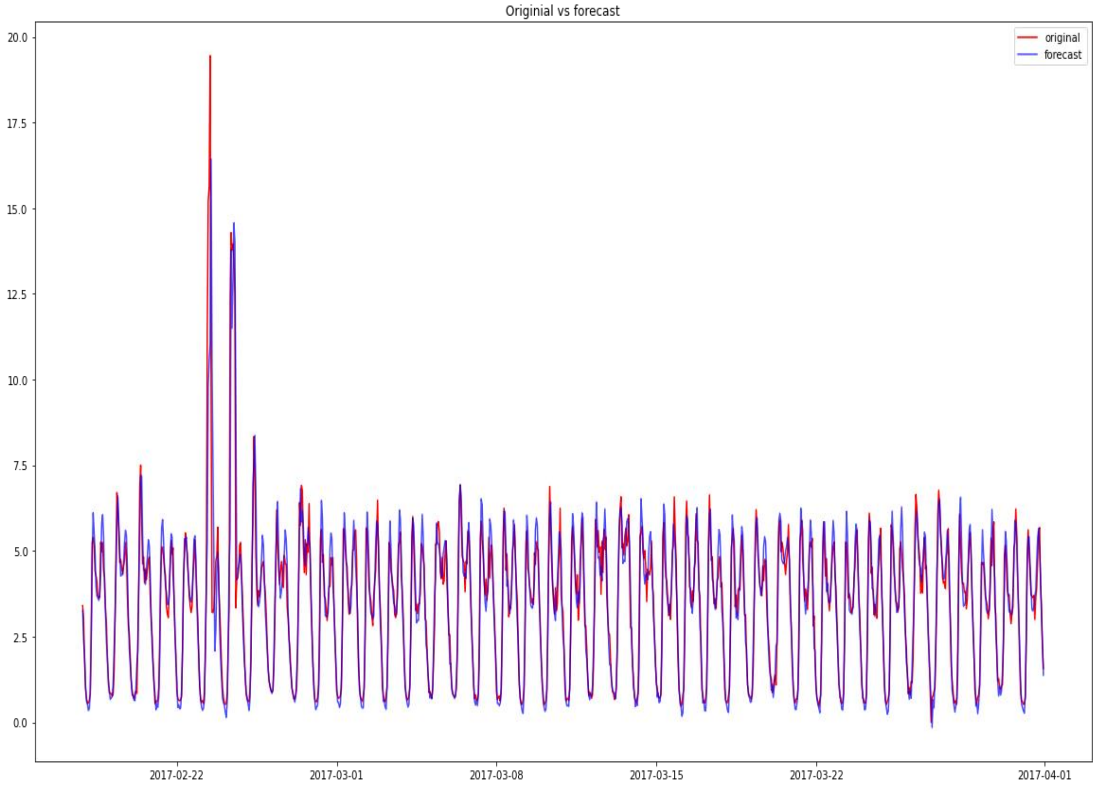
Univariate hourly DMA 225

- To perform univariate hourly prediction on DMA 225, we considered the last 1059 points as the test set and the remaining portion of data to train set.
- In the LSTM model, the look-back period was considered as 24 hours and output was predicted for 24 hours.
- LSTM performed better by achieving a lesser rmse score as 1.436.



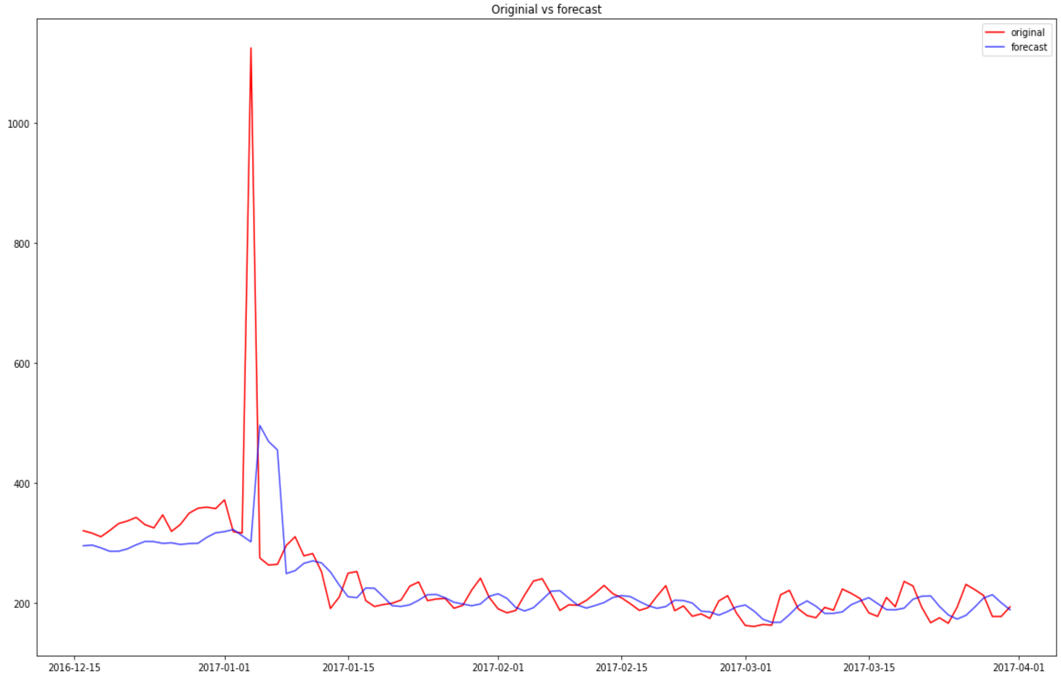
Univariate hourly DMA 468

- To perform univariate hourly prediction on DMA 468, we considered the last 1059 points as test set and the remaining portion of data to train set.
- In the LSTM model, the look-back period was considered as 24 hours and output was predicted for 24 hours.
- LSTM performed better by achieving a lesser rmse score as 0.7799.



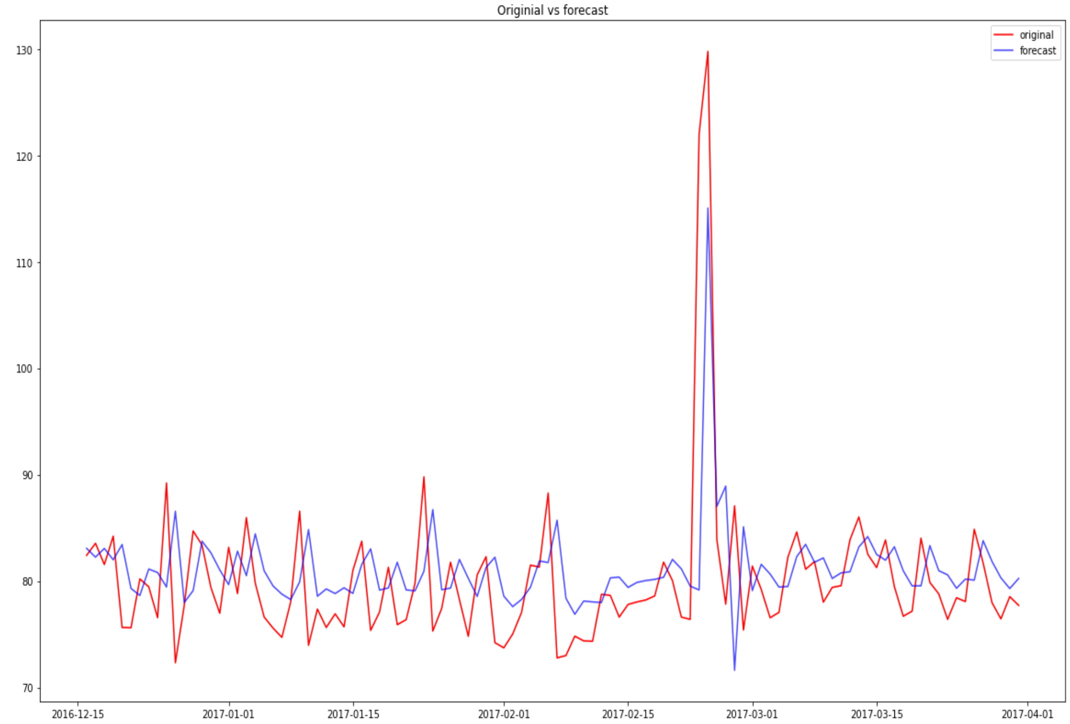
Univariate weekly DMA 225

- To perform univariate weekly analysis on DMA 225 data were re-sampled on a daily basis and the length of the dataset was 365. 70% of the dataset was considered as a train set and 30% was considered as a test set.
- In LSTM, the lookback period of 3 days was considered to perform the weekly prediction.
- LSTM performed better by achieving a root mean square value of 91.41 as the result



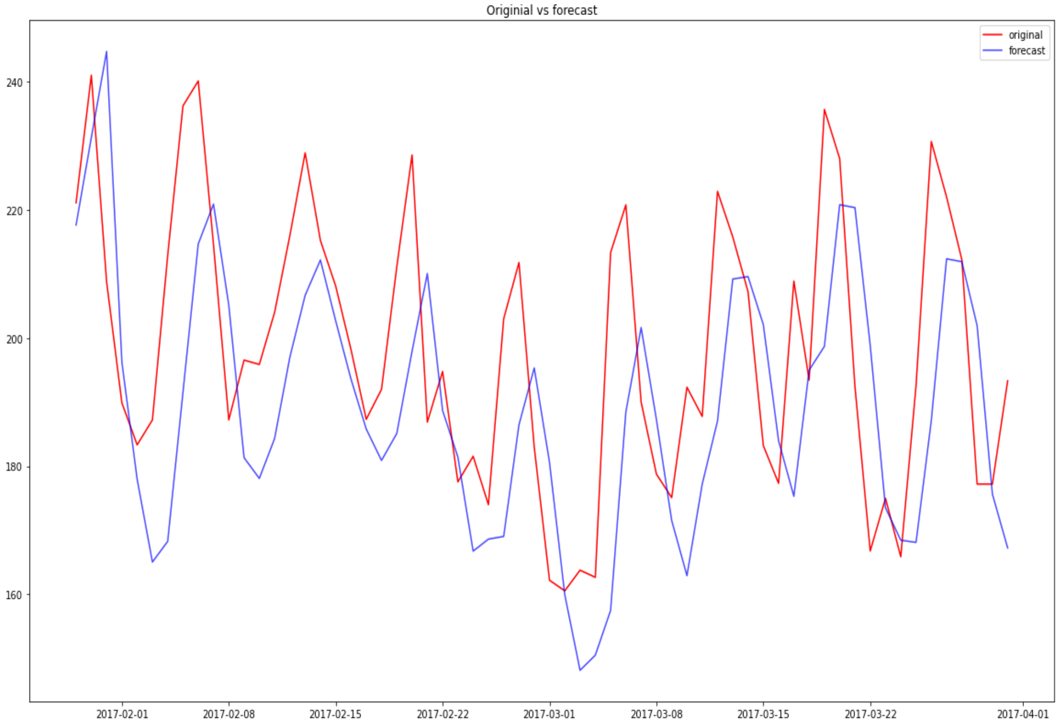
Univariate weekly DMA 468

- To perform univariate weekly analysis on DMA 468 data were re-sampled on daily basis and the length of the dataset was 365. 70% of the dataset was considered as a train set and 30% was considered as a test set.
- In LSTM, the lookback period of 3 days was considered to perform the weekly prediction.
- LSTM performed better by achieving a root mean square value of 6.431 as the result.



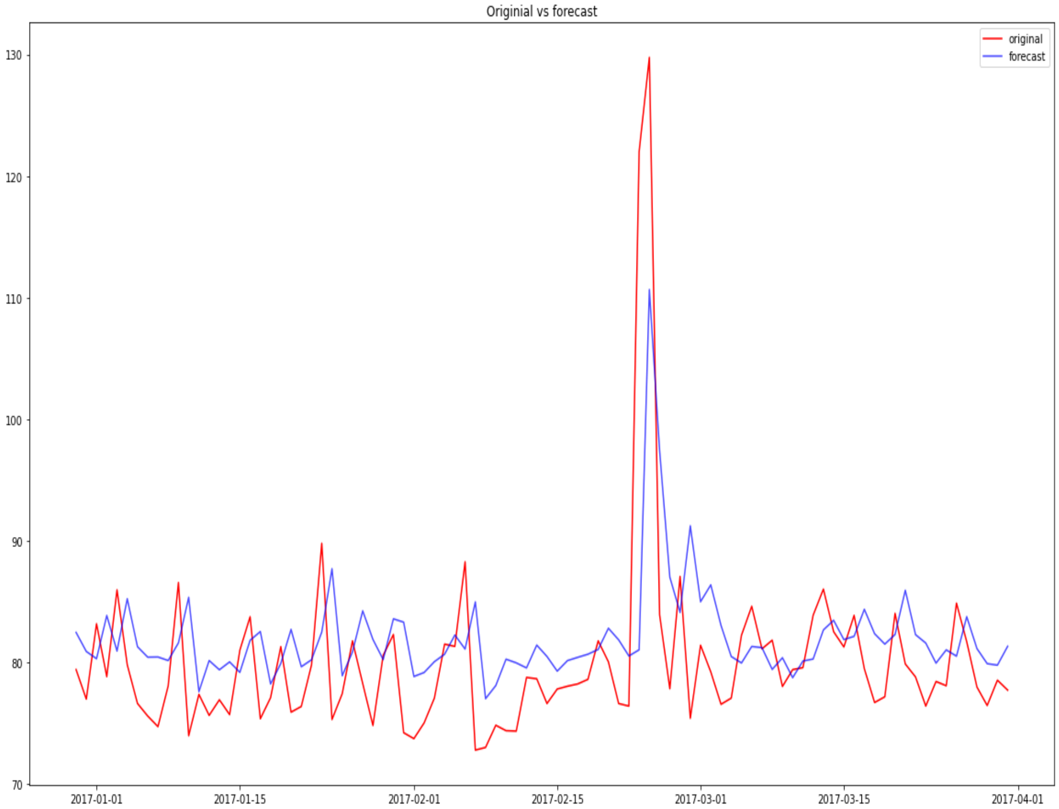
Univariate monthly DMA 225

- To perform univariate monthly analysis on DMA 225 was re-sampled on a daily basis and the length of the dataset is 365. The last 3 months of the dataset was considered as a test set and the remaining was considered as a train set.
- In LSTM, the look-back period of 27 days was considered to perform the monthly prediction. Hence,
- LSTM performed better by achieving a root mean square value of 21.772 as the result.



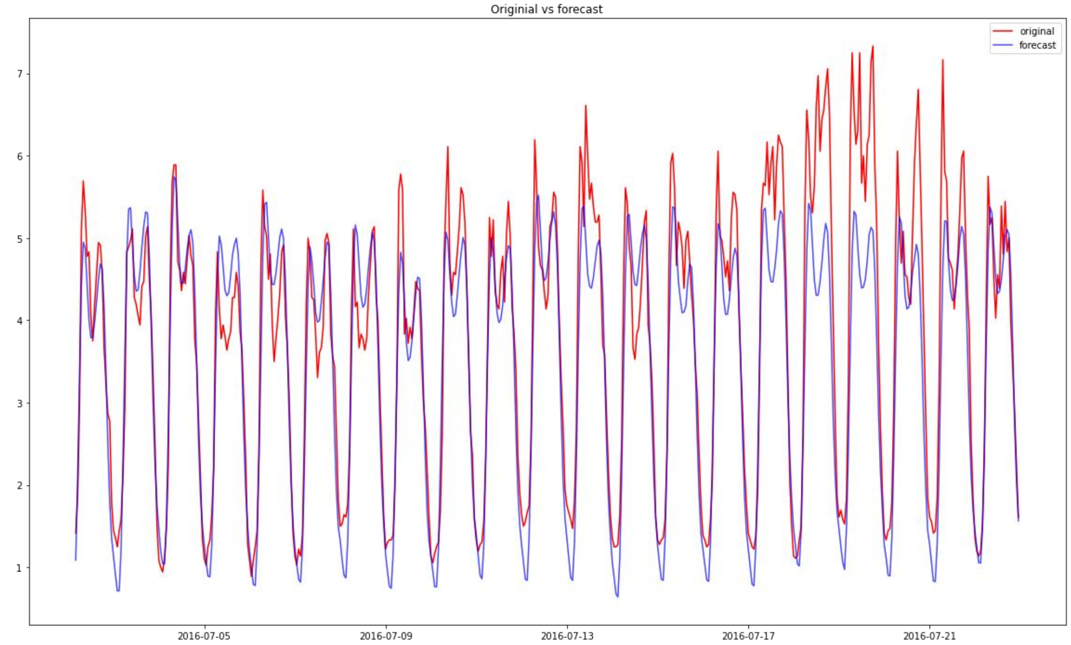
Univariate monthly DMA 468

- To perform univariate monthly analysis on DMA 468 was re-sampled on a daily basis and the length of the dataset was 365. The last 4 months of the dataset was considered as a test set and the remaining amount of data is considered as a train set.
- In LSTM, the look-back period of 27 days was considered to perform the monthly prediction.
- LSTM performed better by achieving a root mean square value of 6.836 as the result.



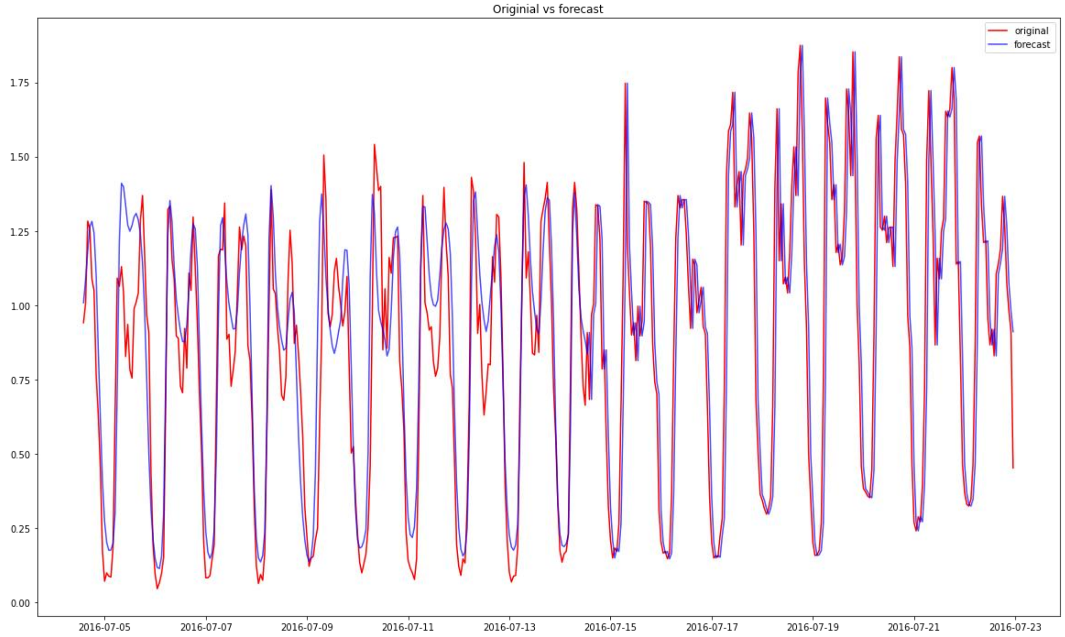
Multivariate hourly DMA 225

- To perform multivariate hourly analysis on DMA 225, we considered the last 500 points as test set and the remaining portion of data to train set.
- Facebook prophet performed better by achieving a rmse score of 0.68 as the result.



Multivariate hourly DMA 468

- To perform multivariate hourly analysis on DMA 468, the last 442 data points were considered as test sets and the remaining were considered as a train set.
- In the LSTM model, the look-back period was considered as 24 hours and output was predicted for 24 hours
- LSTM performed better by achieving a rmse score of 0.196 as the result.



Conclusion

- This research project objective was to perform a comparative study between models to develop short term water demand forecasts. Models such as LSTM, Fb prophet, Sarimax, Baseline were carried out.
- LSTM performed better with lesser rmse values eventually in many use cases. The reason for lower rmse values is because of its capability to use previous sequential data, by this way the model can learn from long term observations of sequence data. This will pave the way for time series forecasting.
- In a nutshell, to get an efficient prediction, LSTM can be implemented with deep learning, because of its properties to face the time series challenges and traditional models such as Sarimax are more efficient to producing good results for a smaller quantity of the dataset.
- In future research, we hope to implement and consider different models, to perform short term / long term water demand forecasting with some other external factors with respect to the water demand in each DMA. This will lead the client/organisation to plan business strategy in an effective way and achieve their final goals.