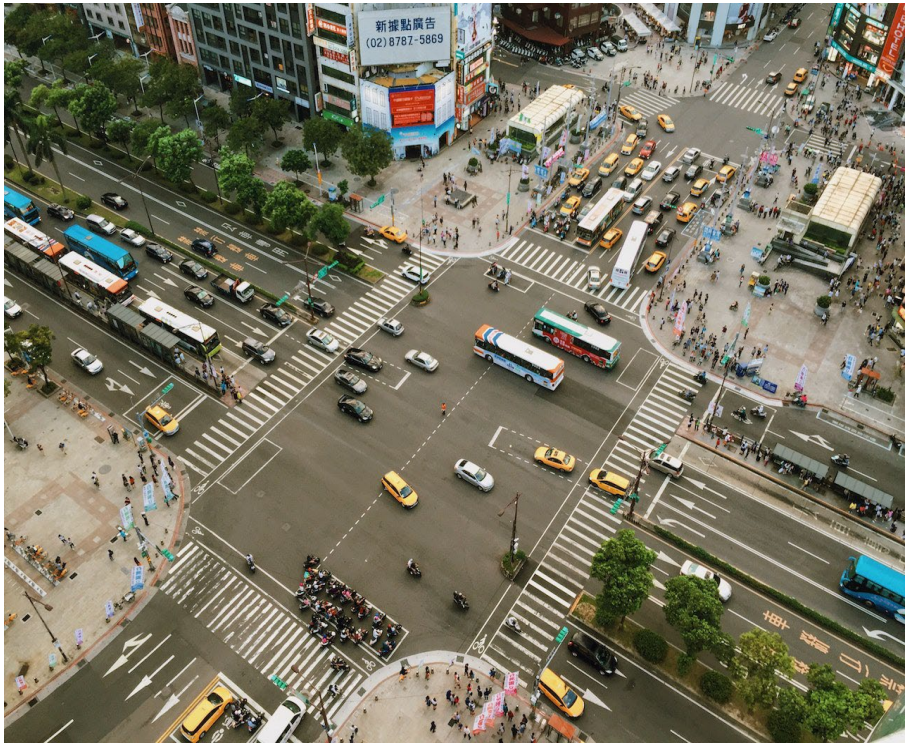
Traffic: From Chaos to Order



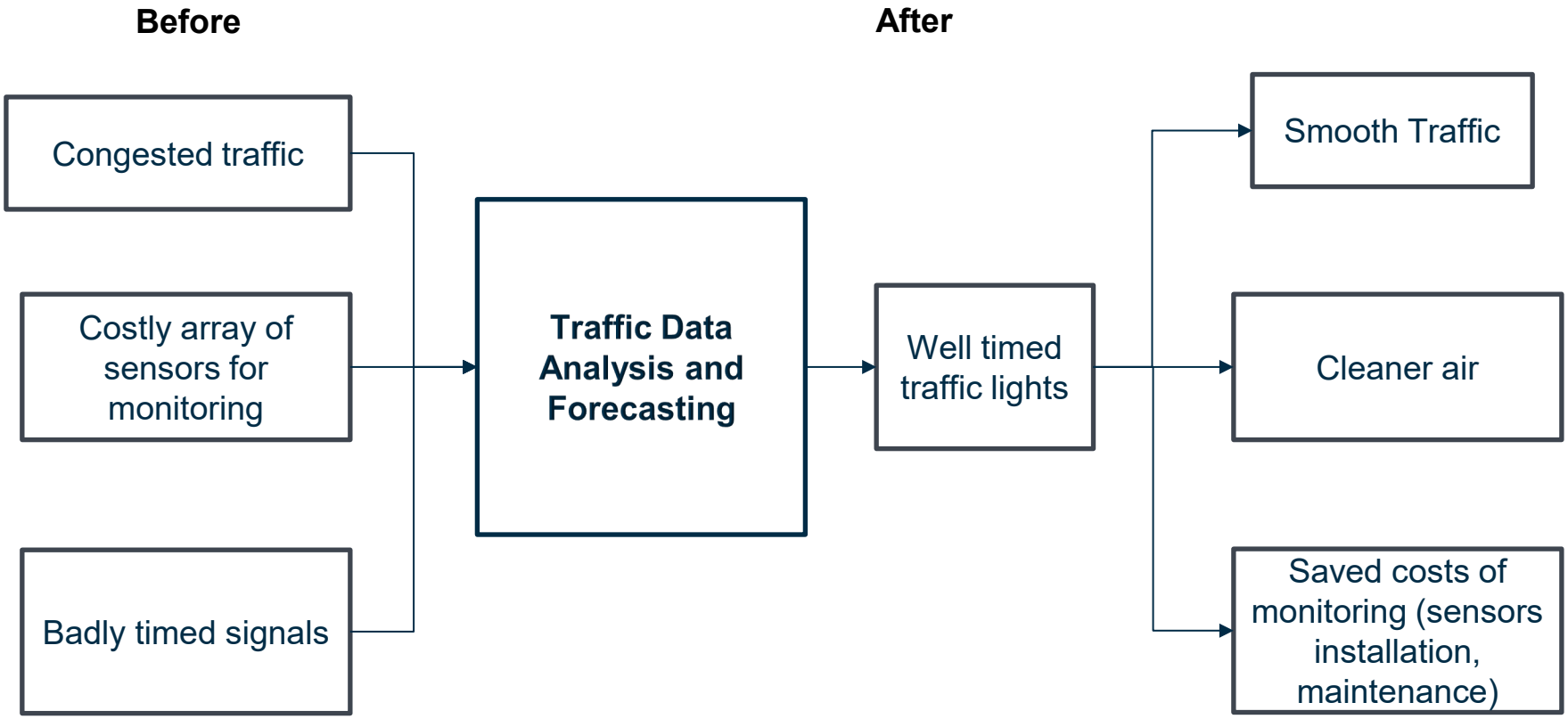
Tired of..

- Being stuck in your car all day?
- Losing your hearing to automotive horns blaring?
- Your health and mood being affected by vehicle fumes caused by traffic jams?
- The sight of a clogged intersection?

Then read on..

Solution?

- Well timed traffic signals! – A large impact on traffic flow



GOAL OF THIS STUDY

- Understand the flow of traffic through Data studies
- Forecast traffic flow (with the ultimate goal of timing traffic signals) with 80% or more accuracy

What we did

- Explored the trends in the data to understand it better
- Prepared and modeled the data for time series forecasting
- Compared the performance of three kinds of models: ARIMA, XGBoost and a Linear Regressor

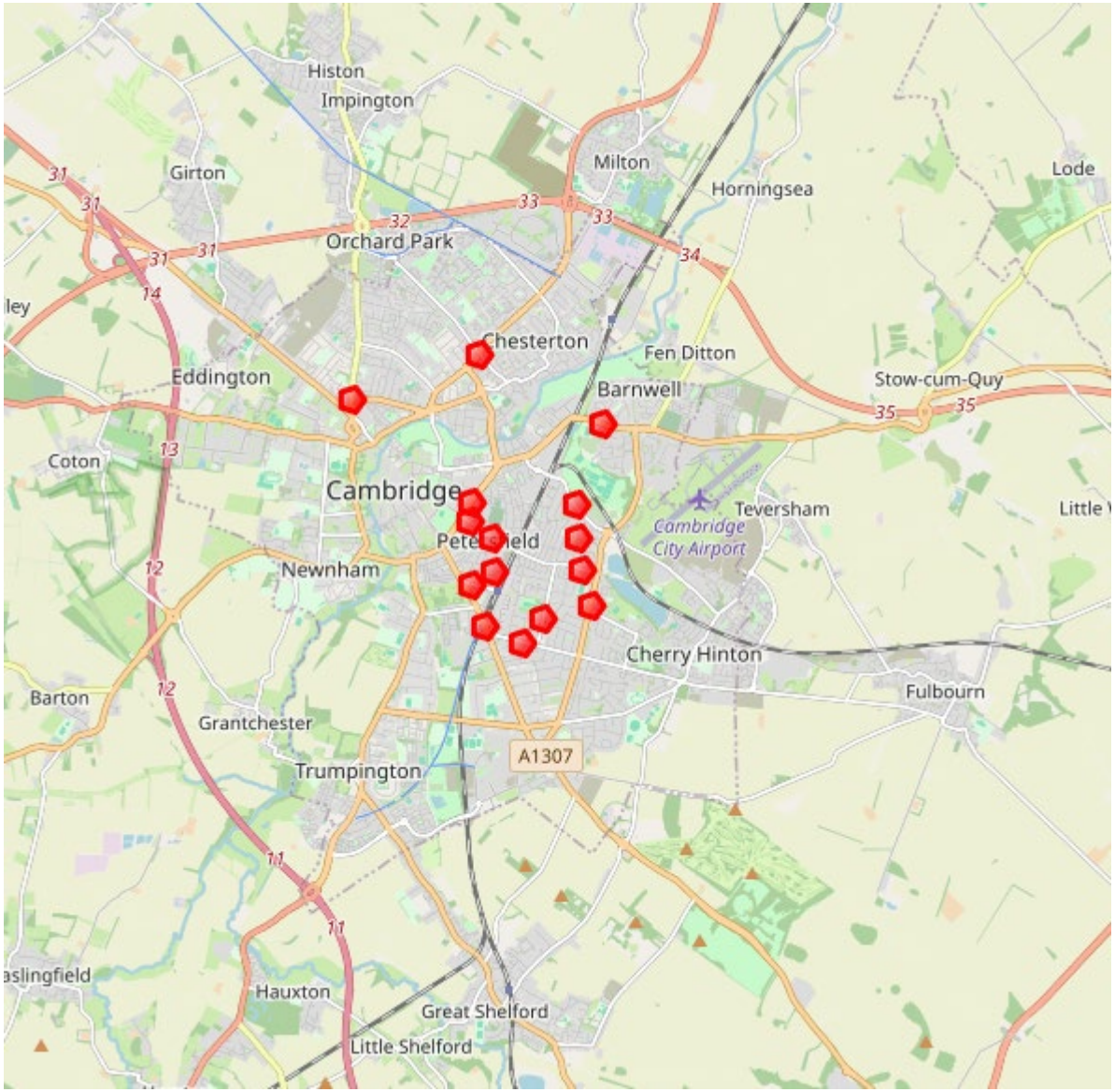
What we found

- The Linear Regressor model displayed better forecasting capability than the ARIMA model and at least as good as the XGB Model for the dataset
- The Linear Regressor is also incredibly faster to train than the others and also meets the accuracy criteria, making it a good choice for forecasting

What was explored?

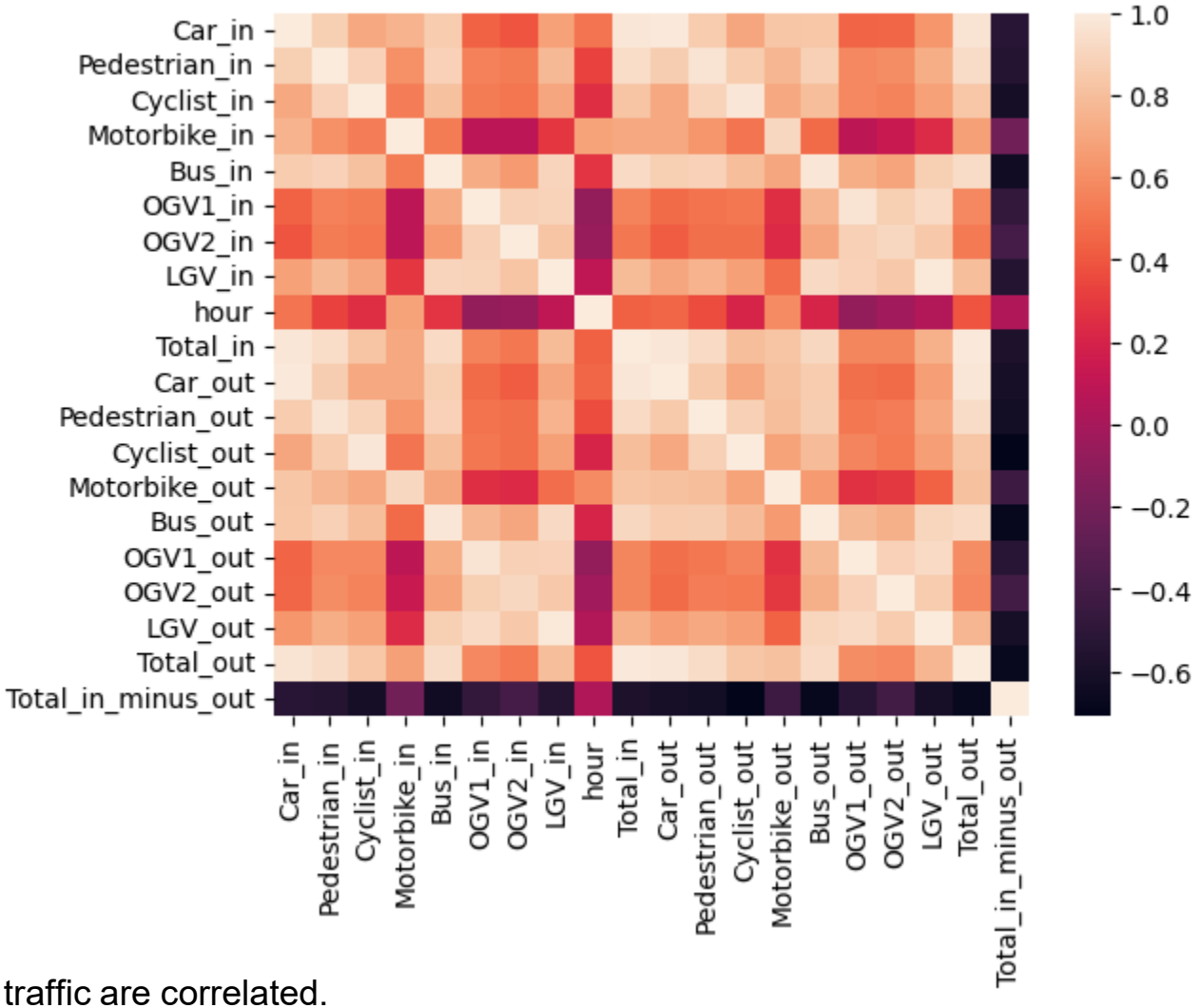
- Data from the Mill Road Project was used for this study
- This is hourly data for the flow of vehicles and pedestrians at the Mill Road area near Cambridge University, UK
- This data will be analyzed to develop a model that will predict the traffic for the next hour based on a combination of expected traffic from several days of data at the time, and recent traffic hourly data.
- This output can then be used to time the traffic signal appropriately.

Map of sensors



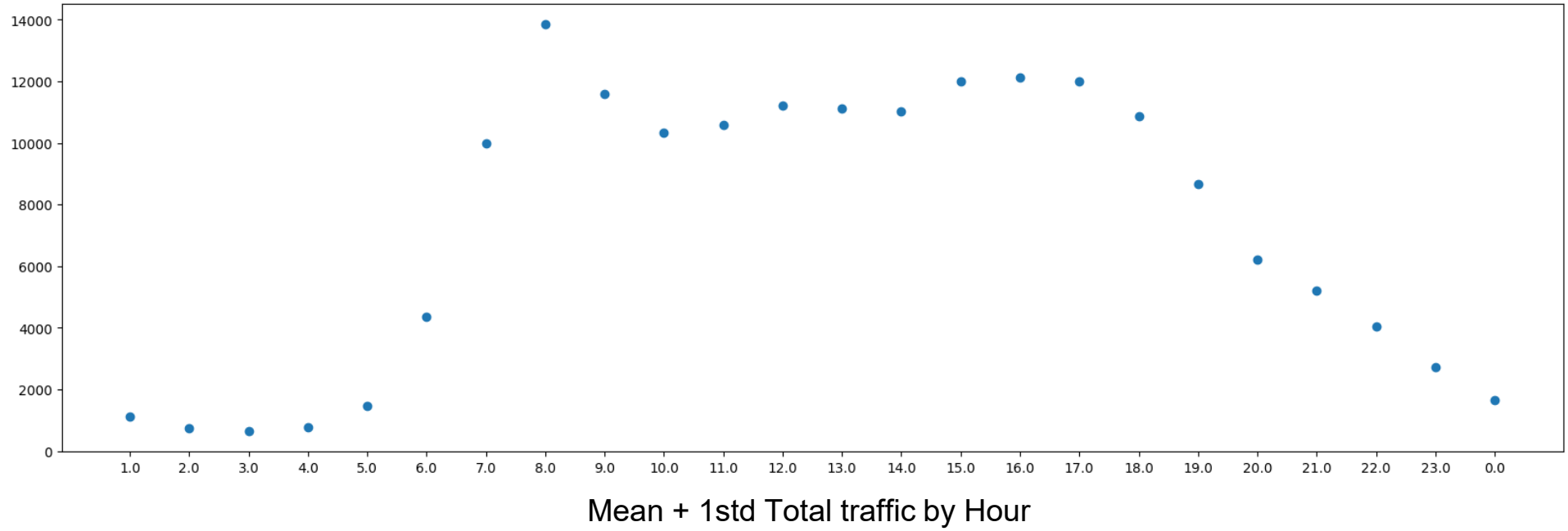
- The sensors are distributed in the area around Mill Road
- The sensors together monitor most roads leading into or out of the area

Correlations of different kinds of traffic?



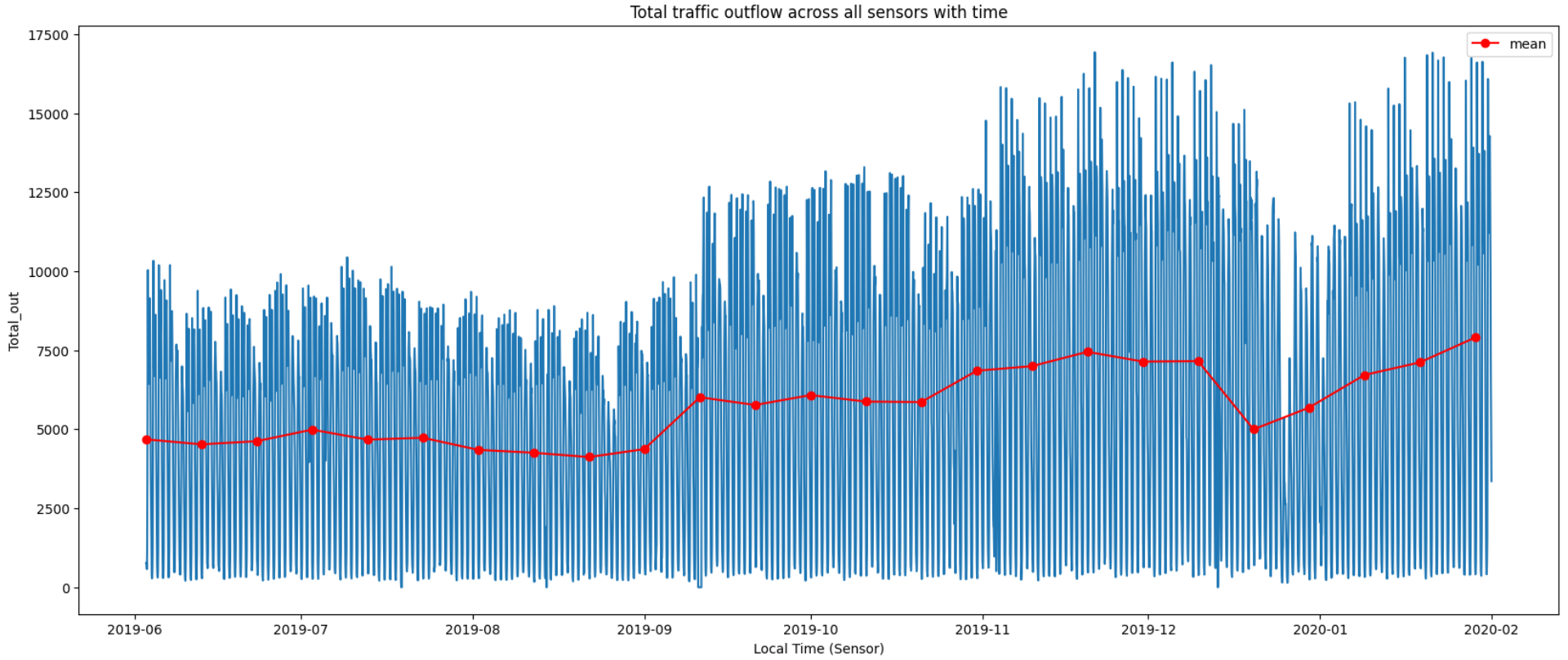
- A lot of the traffic are correlated.
- The traffic congestion can be measured by '*Total_out*' which is the sum of all different kinds of traffic out from the area
- The strong correlation of many other variables to Total_out also bolsters our choice

Traffic by Hour



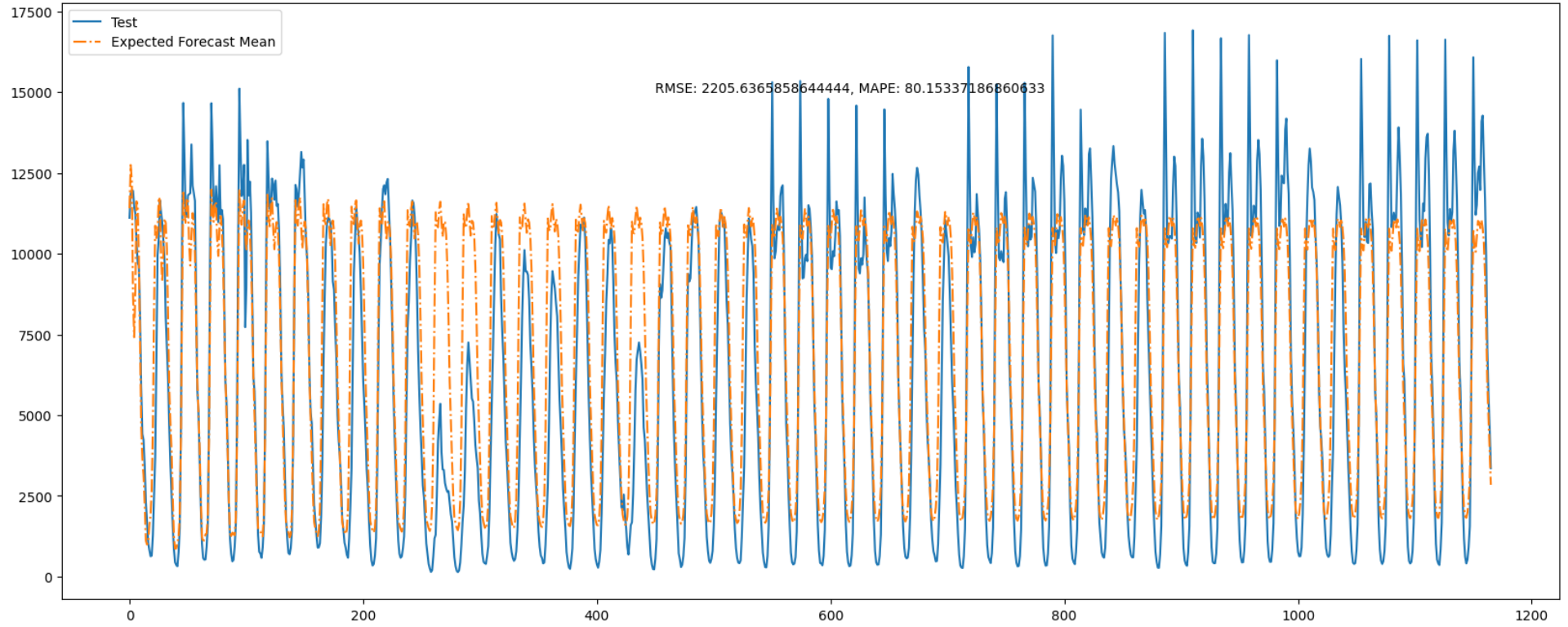
- Traffic is low at early mornings and late evenings
- It reaches a peak around 8 am and remains steady from 10 to about 5 pm, after which the traffic steadily declines
- This is expected traffic pattern considering the times of day people are most busy with work, school, travel etc..
- These values are later used to adjust the ARIMA model to try get better accuracy

Traffic by Hour and Day across all sensors



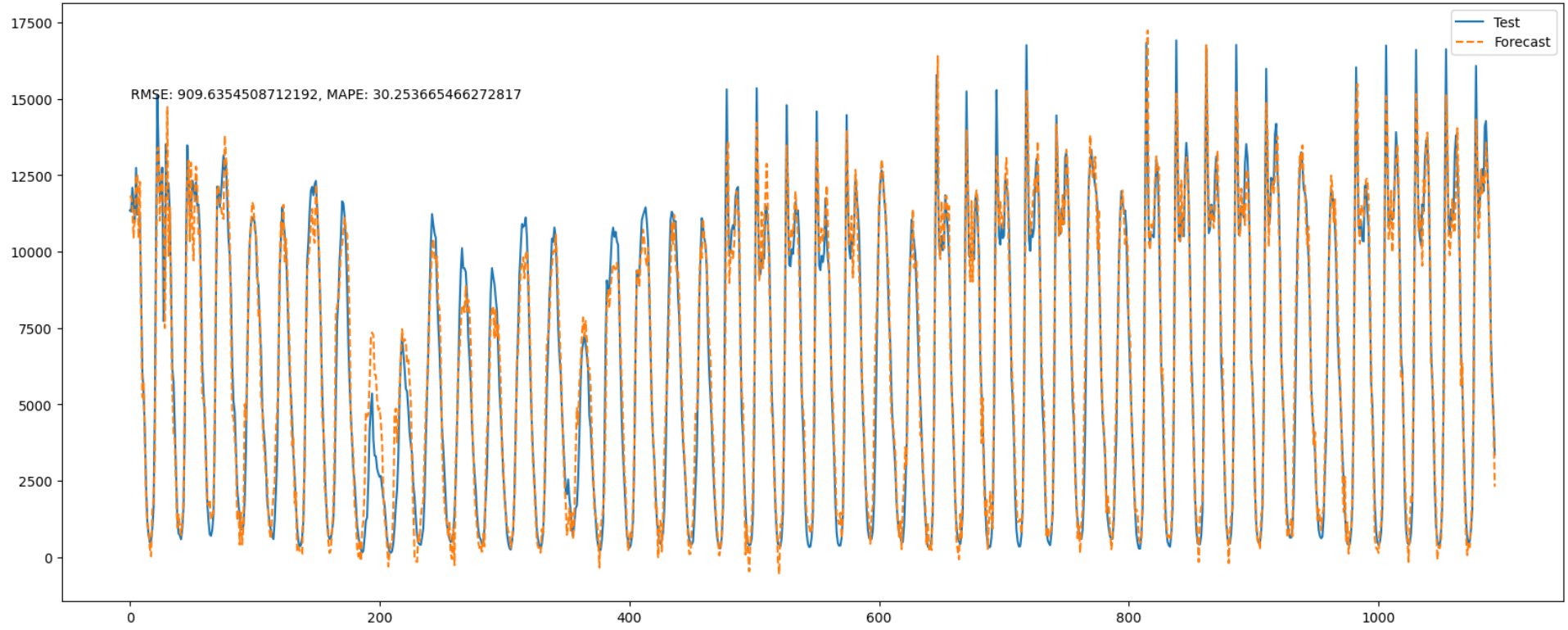
- There is a trend in observed traffic flow.
- During the Christmas holiday season in December, the traffic drops, then picks up again into the following year
- This data is non-stationary and would need to be accounted for in an ARIMA model

Model Performance: mean of ARIMA and Hour aggregate forecast



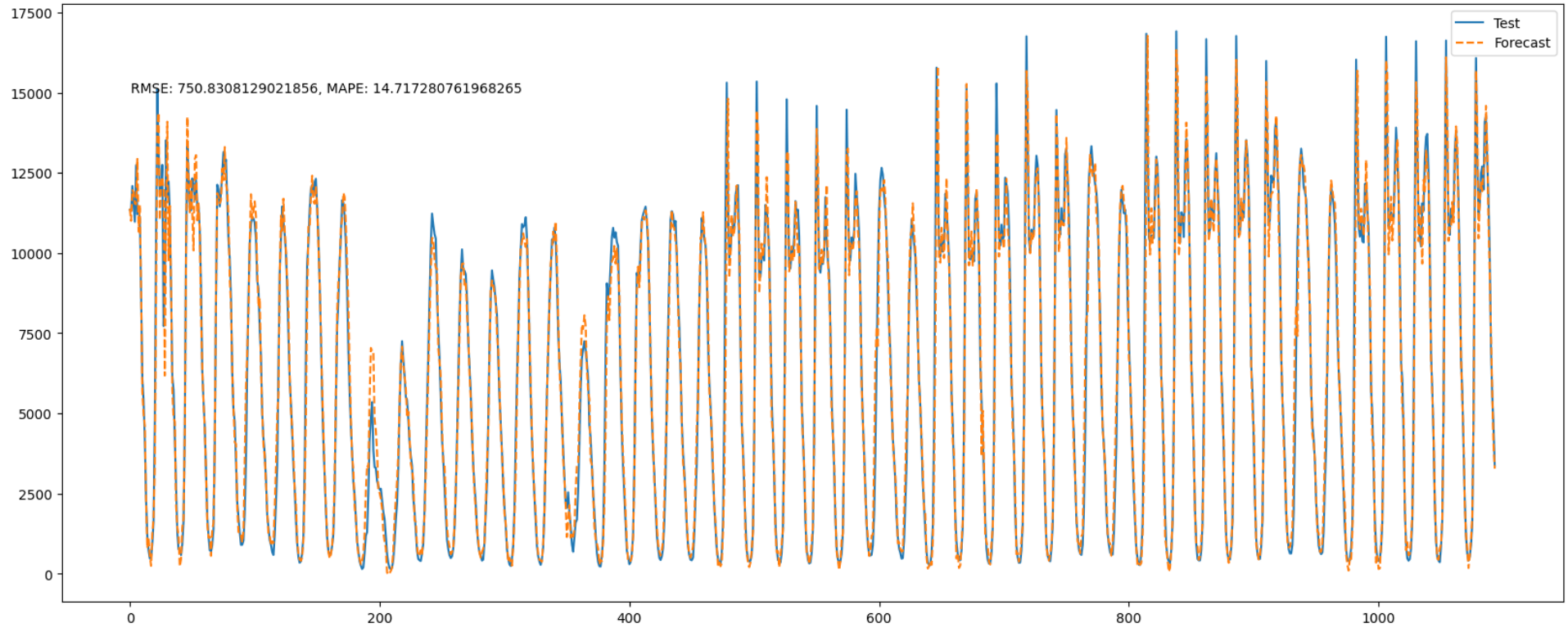
- The RMSE is about 2206 and MAPE is 80.
- Training time around 10-11 mins
- The model is better at predicting general peaks and valleys but struggles at following seasonal trends in the data, such as the holidays at Christmas

Model Performance: XGBoost Regressor



- The RMSE is about 910 and MAPE around 30.
- Training time around 4-7 mins
- The model is able to follow most trends in the data, including seasonal trends
- It is faster to train and more accurate than the ARIMA+hour based adjustment

Model Performance: Linear Regressor



- The RMSE is about 751 and MAPE around 15.
- Training time around .003 mins
- The model is able to follow most trends in the data, including seasonal trends
- It is faster to train and more accurate than both the ARIMA and XGB Regression

Model Performance Matrix

Model	Training Time [minutes]	MAPE	RMSE
ARIMA (27,1,5)	0.5	189	4080
ARIMA (27,1,50)	11	115	2517
ARIMA (27,1,50) with expected sensor value	11	80	2205
XGBoost Regressor	6.5	30	910
Linear Regressor	0.00333	15	751

- Since Linear Regressor is faster to train as well as more accurate than XGB and ARIMA, in addition to meeting the 80% accuracy target, it seems to be the best model to forecast the hourly traffic data

Recommendations

- Use traffic flow simulation software timed as per the model forecast, in conjunction with historical timing, to match the traffic flow on test data. Several flow softwares such as SUMO, VISSIM and AIMSUN are available for this purpose
- Implement this new timing system in real time once testing produces good results
- Continue to monitor the traffic and check the performance of the model for future data beyond the test data
- Retune the model if necessary

Challenges

- Determining the parameter to model traffic with was the first challenge and involved some research.
- The time complexity of the ARIMA model was a significant hindrance to effective ARIMA modeling; achieving models with higher 'q' values took unreasonable amounts of time.
- Finding the optimum set of parameters for the XGB pipeline took a lot of trial and error.