Veikko Kovanen
veikko.kovanen@gmail.com
21.6.2018

BCG Gamma Data Science Competition 2018

This report describes briefly my solution to the BCG Data Science Competition where the challenge was to forecast hourly traffic volumes. My model is fully based on the historical TMS/LAM data provided by Liikennevirasto, although additional data sources was encouraged to utilize as well. I used all available data of the tree specified locations (Askisto, Mäntsälä and Kemijärvi), which was available from 2010 to 2018, and trained a Random Forest regression model. Using this model, I generated a forecast series for the specified target dates 22.6. – 26.6.2018, including hourly volumes for cars & vans, trucks and buses to both directions in the mentioned three locations.

Before committing on Random Forest (RF), I considered more traditional time series methods, such as ARIMA, as well as more sophisticated deep learning models like RNN. The former would be good due to its simplicity, whereas RNN would catch complicated long-term relationships between the inputs. One concern I had was that the prediction period includes the Midsummer's holidays that are potentially challenging for simple time series models. On the other hand, RNN would be laborious to implement as it involves various hyper parameters that potentially need tuning as well as long computation time. Thus, I decided to give a try for RF first. It turned out to perform very well so I did not bother to try other alternatives after that. I did however try to incorporate historical weather data, including daily rain and temperatures near the LAM stations, but this did not improve the performance, so I excluded it from the final solution.

The model was tested on test data which represented 30 % of the labeled data. This set was selected randomly, and it was not used for training the model. The model scored 0.97/1.00 using the coefficient of R^2 = (1 - u/v), where u is the regression sum of squares sum((y_true - y_pred)^2) and v is the residual sum of squares sum((y_true - mean(y_true))^2). As the model was not tuned and tested again on the same test data, there was no need for further validation with a separate validation data.

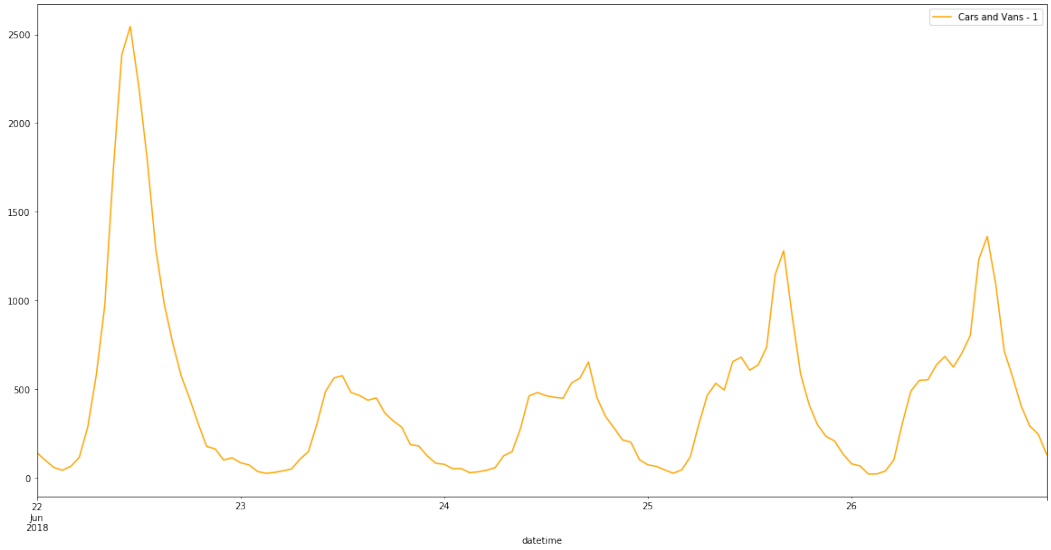

Figure 1. Forecast of cars and vans to direction 1 in Mäntsälä over the target period 22.6. - 26.6.

As the model has been trained with 10 years of hourly data from three LAM stations, the model generalizes well to these locations (Askisto, Mäntsälä, Kemijärvi). That is, the model can be used for predicting hourly volumes of any vehicle type over any future period as long as the location is one of the three. To extend the model to produce meaningful predictions to other locations, one has to retrain the model so that TMS data from the desired locations is included. Figure 1 presents a subset of the prediction as a graph.

Full forecast for the target period is in **prediction_22062018_26062019.csv** and the model code is in **model.py**. All source files and more detailed description of the work is at https://github.com/KovaVeikko/bcg-gamma (will be public after the deadline, on 22.6.)