Angewandte KI in der Medizin

Project Nr. 4

Highway Traffic Prediction

WS 22/23

Due at 31.01.2023

1. Project Data

The dataset contains about 48 thousand entries. The data describes the number of vehicles that passed one of four different junctions for each hour. So every entry in the dataset consists of four features, which are the following:

1. DateTime
2. Junction
3. Vehicles
4. ID

DateTime shows the hour and date in which the vehicles were count. Junction contains a value from range one to four, that points out, for which of the four junctions the vehicles got count. Vehicles is just the amount of vehicles that passed a junction in an hour. ID identifies the row uniquely.

DateTime	Junction	Vehicles	ID
2015-11-01 02:00:00	1	10	20151101021

So, the above entry says that on first november of 2015 on 2AM on the first junction, 10 vehicles got count in this hour. Before training both regression models a time series plot for every junction is made to show the trend of how many cars hourly passed the junction in the time the data was measured.

The data source is: https://www.kaggle.com/code/jaymineshkumarpatel/traffic-prediction/data

2. Project Model

To predict the hourly amount of vehicles for each junction for the coming year, two kinds of regression models are getting used. One is a Random Forest Regressor (RFR) and the other one is a Support Vector Machine Regressor (SVR).

To train both machine learning regression models, these features are used: Year, Month, Day and Hour. The amount of cars is the labeled value. To get the features, the timestamp from the original dataframe gets splitted into them, while every feature gets its own column in a new data frame from which the model will learn from. For predicting the following year a similar data frame is used with just a different time range to predict (2017-07-01 to 2018-06-30). The hourly prediction of the highway traffic is then getting added as another column (Predicted_Vehicles) to the prediction data frame.

Also the data is standardized (z-score standardization) with an StandardScaler.

The models themselves are trained with a 70/30 test-train ratio splitted data. To see how good the predicted data are, the following metrics are used:

- MAE
- MSE
- RMSE
- Huber Loss
- MAPE

The models are getting trained for each junction first, then the data gets predicted on the corresponding regression model for the coming year.

3. Project Results

The data gets predicted for each junction. So there are four data frames containing the predicted hourly count of vehicles each assigned to one of the four junctions. The predicted cars passing the junctions for the following year have been written to a separate .csv file for each junction. The name convention for these .csv files goes after the below:

➔ next_year_prediction_junction_{junction_number}.csv

The project was done with scikit-learn.