Development of a machine learning model for the extrapolation of short-term bicycle counts with the inclusion of meteorological data

Master‘s Thesis of Danil Belikhov

Mentoring:
Dr.-Ing. Simone Weikl
Mario Ilic, M.Sc.
Georgios Grigoropoulos, M.Sc.

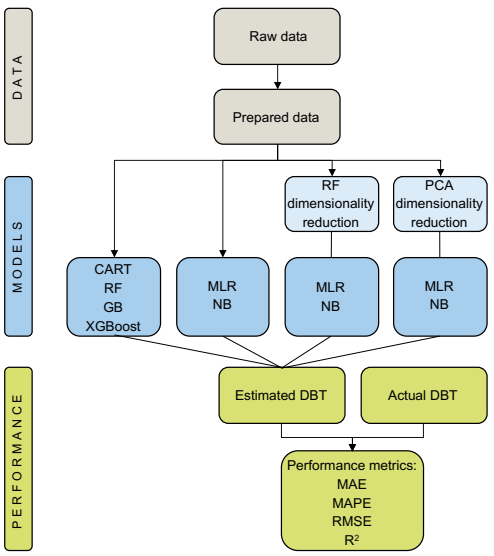

Figure 1. Research Architecture

Theoretical Background

Planning and building new cycling infrastructure requires the presence of cycling volume data and uptrend in it. To have a reliable database with all-day cycling traffic counts or an average daily volume of cyclists for the desired area it is necessary to set up automatic counting facilities or do counting manually. To reduce the manual workload and costs of traffic counters' installation, a method of extrapolation of short-term bicycle counts based on machine learning models can be implemented.

The city of Munich has a database with cycling traffic counts for recent years measured by six permanent bicycle traffic measurement stations and additional cycling counts measured manually. The city of Munich developed an extrapolation method that uses manual counting data and provides results that could be improved using other machine learning methods and additional meteorological data shown in this work.

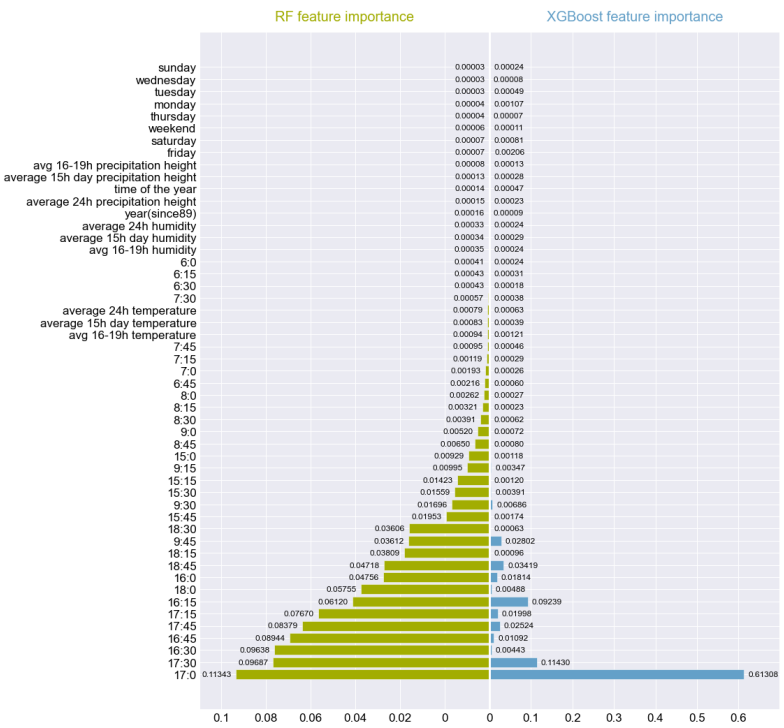

Figure 2. RF and XGBoost feature importance comparison

Methodology

This study explored different machine learning models and their combinations as shown in **Figure 1**, as well as different input data structures to extrapolate short-term bicycle traffic counts taking weather conditions into account and getting the most accurate output. A comprehensive literature review was done that showed the current activities in this field. Three types of input data were prepared for feeding the models. The features included morning peak hours counts, evening peak hours counts, meteorological data, year, and day of the week data while the target data had an average daily volume of cyclists. Two feature reduction techniques were applied and tested: Random Forest and Principal Component Analysis (PCA). Besides the use of additional weather data, a single-location case study was conducted where the best-performed models were used to predict bicycle traffic on a dataset containing only one location.

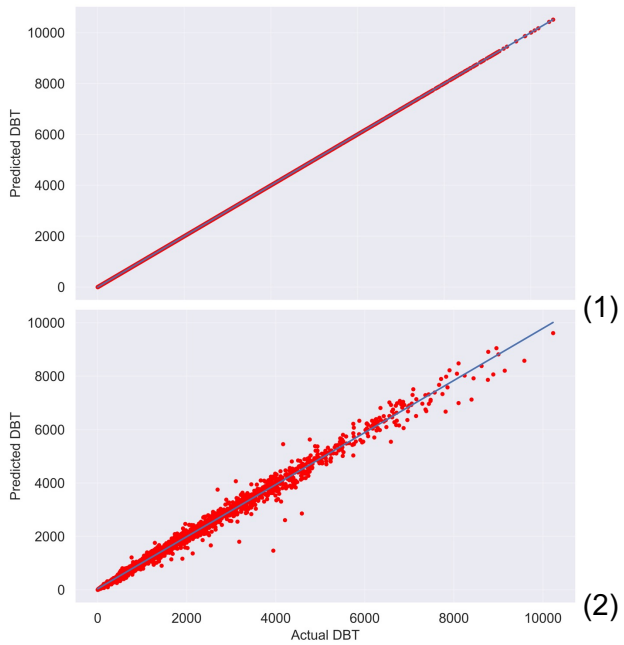

Figure 3. Scatter plots of the actual and predicted DBT for training (1) and testing (2) datasets of the Random Forest (RF) model

Key Findings

Some models had similar accuracy but some of them perform better. It was important what kind of input data to use for model training to have the most precise result and not overfit the model. Based on available data the most important variables for all models are the evening peak hours counts from 16:00 to 19:00. The morning peak hours counts and the meteorological data were less important for a good prediction result (**Figure 2**). The best-performed models were Random Forest with an accuracy of 91% and Extreme Gradient Boosting with 86% accuracy which differs a lot from the results of the initial method used by the city of Munich (**Figure 3**). Models performed better when training on a single-location dataset and more precise weather data but that leads to overfitting. There is a high potential for exploring data from other cities since these models performed well on the current data. Also, more advanced deep learning models can be tried out, as well as the determination of the smallest amount of the input data that will be needed to produce sufficient prediction results.