‘More Bikes’: Machine-learning experiments to predict rental-bike availability

Tim Lawson

November 27, 2023

Abstract

TODO

1 Introduction

The task is to predict the number of available bikes at each of 75 rental stations over a period of three months. It is divided into four sub-tasks:

1. Predict with a separate model for each rental station.
2. Predict with a single model for all rental stations.
3. Predict with a set of linear models of the number of available bikes at each of 200 other rental stations.
4. Predict with any or all of the above.

The predictions are evaluated by the mean absolute error (MAE) between the predicted and true values of the target variable (the number of available bikes).

2 Data analysis

For sub-tasks 1 and 2, the data comprises a set of 75 time series of the number of available bikes at each rental station, with a fixed hourly interval.

A table of the feature categories and summary statistics.

Exploratory data analysis yielded the following observations:

- The precipitation feature is zero for all data points, i.e., it is missing.

For sub-task 1, the features that are properties of the rental station are constant and thus not useful for prediction.

The pairwise correlations between the features were computed over all 75 time series, with the following observations:

- The ‘short’ and ‘full’ average numbers of bikes and differences in the average numbers of bikes are fully correlated, i.e., either the ‘short’ or ‘full’ versions of these features are redundant.
- The average and maximum wind speed are strongly positively correlated, i.e., one of these two features is redundant.

A heatmap of the pairwise correlations between the features.

3 Subtasks

3.1 Separate models for each station

The first subtask is to predict the number of available bikes with a separate model for each station. An intuitive baseline is to predict a constant value, e.g., the average number of available bikes.