Predictive Analytics

To create prediction models, basically it is reasonable to use time series data from more than just one year. As the activities and tasks completed priorly are based on only the data of the year 2017, this is continued in this part and it is forgone to use data of multiple years. For further insights and more precise predictions, the code presented in the respective files can be applied to a dataset containing the data of multiple years. In such a case, of course, the code to clean the data should be applied to the dataset first.

As a start to the prediction analysis it was necessary to create a useful data basis. Naturally, this includes the data for all the rides in the year of 2017. From this a new data frame by adding a new column containing the hour of each start\_time, reducing the data frame content to the columns start\_time and hour and then grouping by hour. The result is a table showing the number of rides per hour. As it comprises 8,497 rows, there is not data contained for every hour in 2017 (would be 24 hours x 365 days = 8,760 rows of data).

Furthermore, it seemed reasonable to include weather data as an obvious choice for a potential predictor to the usage of bikes. Hourly weather data were then sorted by its datetime and checked for missing values. The result was that instead of 8,760 rows of data (24 hours x 365 days) there was only 8,690 rows of data.

To treat the problem of missing data points and uncomplete data frames (less than 8,760 data rows), a data frame is created with timestamps of every hour in 2017 as an index. As the first timestamp in the weather data is midnight of the 1st January 2017, this can be used as a starting point for the “date-time list”. Then the existing weather data are assigned and it is checked for null values. Such null values were filled with data from the respective previous row, because a significant change of the temperature within one hour is assumed to be unlikely. A problem with this procedure is missing or null values in a few consecutive rows (a block of more than one row has null values). Afterwards, the rides’ data is added to the data frame as well. Missing ride values are filled with a zero as it seems that there were no rides for this time / during that hour.