Predicting Distribution of Dublin Bikes

Con O'Leary; olaoghac@tcd.ie

August 11, 2021

1 Data procurement

There are considerable time gaps in *Dublinbikes 2020 Q1 usage data*. I recognised this accidentally, when making a provisional graph that showed the percent bike occupancy of each station throughout the full duration of the data. The sections of the x-axis that are atopped by red show the final station to be plotted (plotted in red) fluctuating in percent occupancy; they show where there is data. Where there is a mesh of colours, is where the graphing library is adjoining lines across a data-less span. **The end of the second data-less span is the 27th of January** 2020, and so I processed the information from then onwards. The reason why I was adamant to exclude any dates before the 27th that did have full data is so that I would not have more data for certain days of the week than I have for others: I was worried this might lead to the models orientating themselves more around certain days.

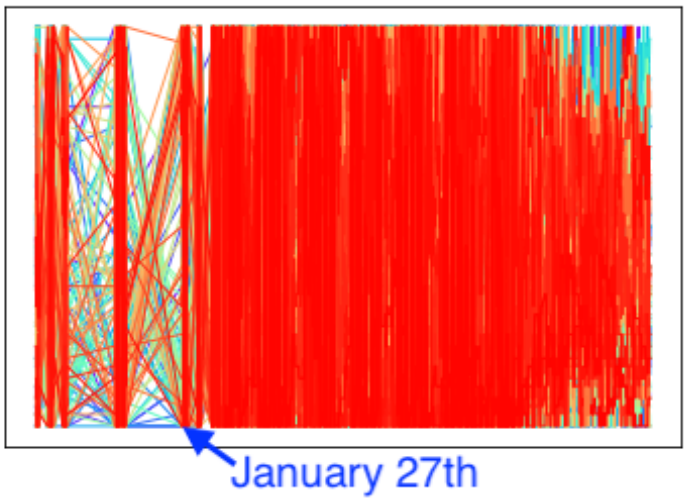

Figure 1: % bike occupancy for every station over the total duration

2 Baseline Approach

I first went about a linear regression. Noting the cyclical nature of the amount of bikes in a station, I derived polynomial features from the amount of bikes in respective stations over the total time. I disabled the shuffle parameter in my train-test split, as it seemed testing would be improper if, for example, we were looking for a prediction on the amount of bikes in station x in 30 minutes if the training included the amount of bikes in station x 25 and 35 minutes ago. Understandably, the x-axis of the training data spanning months and the x-axis of the testing data spanning weeks was not conducive at all, with any degree of polynomial features.

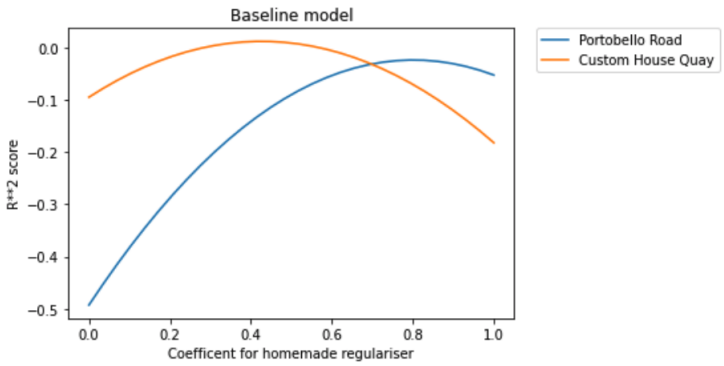

Figure 2: % bike occupancy for every station over the total duration

3 Approach 1

4 Approach 2

5 Evaluation

I believe **R**2** evaluation is appropriate for measuring success in this task. A prediction that is off by one bike is probably only going to be a fifth as much of an issue (to a hypothetical user availing of the model) as a prediction that is off by five bikes. As such, that R**2 evaluation measures how correct the variance of predictions are—and that variance accounts equally for the frequency of errors and the magnitude of errors—means that error in the model is represented proportionally to how much of an issue is presented to the objective of the task.

(i) -

(ii) -