Final Report

Conor Dalton

December 2021

1 Predicting Bike Station Occupancy

1.1 Feature Engineering

1.1.1 Selection of Station

In choosing the stations I was going to use I decided to display them on a map. This would make it easier to choose stations which would have different behaviour. I decided to choose station 97 (Kilmainham Gaol) as it is the furthest from the city center. I also chose station 109 (Buckingham Street) as it is beside Connolly and I assumed it would have drastically different behaviour than my other station.

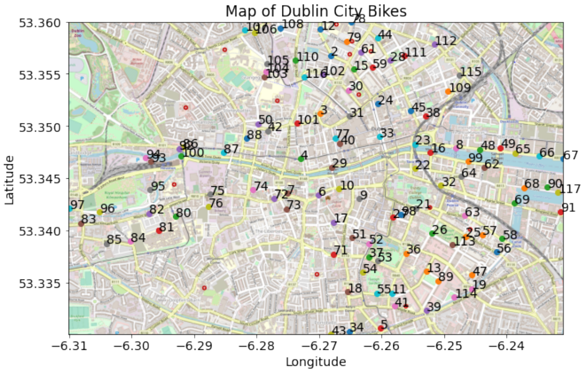

Figure 1: Map of Dublin Bike stations.

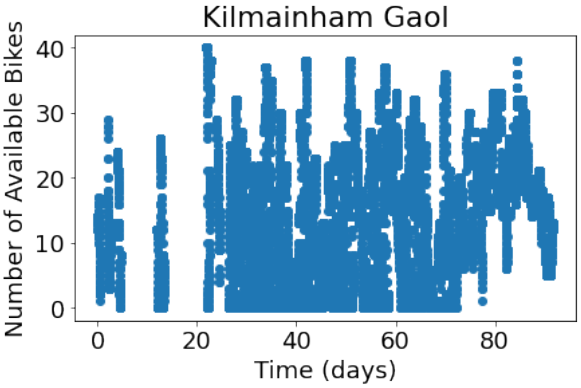

Figure 2: Map of Dublin Bike stations.

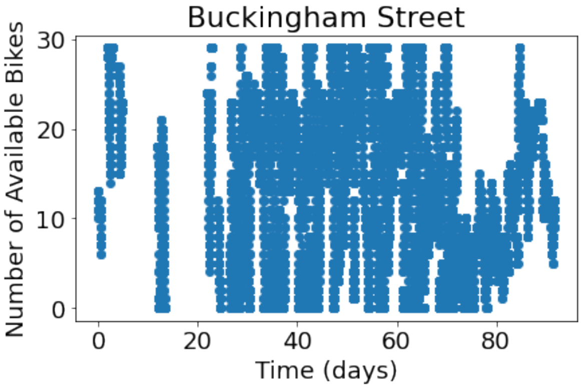

Figure 3: Map of Dublin Bike stations.

1.1.2 Data Cleaning

As can be seen in figures 2 and 3, much of the data in the first few weeks is missing. I decided to simply remove the first 27 days in both datasets as there was so much missing that it would likely make it more difficult to work with the data in a way that didn't negatively affect the performance of the models. It is likely that there were other missing data points, this is evident in that not every day has the same number of entries (between 287 and 289). This is unlikey to have any major impact on my predictions as it is very close to 288 (the number of 5 minute time blocks in a day).

1.1.3 Feature Selection

The features I took from the datasets are

1.2 Ridge Regression

I initially decided to use ridge regression to predict future values. I wanted to use some form of linear regression because I used them for a similar problem in my group project. Lasso regression is more likely than Ridge to reduce coefficients to zero but I was curious to see how even the lightly weighted parameters were affected so I decided to use ridge regression instead.

I used 5 fold cross validation to find the ideal value for alpha to use and I found that varying alpha had very little effect until alpha became quite large as can be seen in figures.... I decided to use a value of 1 for every model as this value gave a near minimum mean-squared error for every prediction model.

The metric I used to evaluate performance is mean squared error. I feel that this is a suitable metric as it tells us how far off predictions from a model are.

1.3 K Neighbours Regression

1.4 Evaluation

1.4.1 Method 1

1.4.2 Method 2

2 Shorter Questions

2.1 ROC Curve

An ROC curve is...

2.2 Situations Where Linear Regression will be Inaccurate

Linear regression is ...

2.3 SVM Classifier vs. Neural Net Classifier

An SVM Classifier... an Neural Net Classifier

2.4 Operation of a Convolutional Layer in a ConvNet

ConvNets are... Convolutional layers are. . .

2.5 K-fold Cross-Validation

Resampling is done to change amount of data.

A Appendix