
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Department of Political Science


# Introduction to Machine Learning HT: Homework 2 - Group Project


Group 6:
Shekhar Kedia, 23351315
Stefan Keel, 23366536
Yana Konshyna, 23359606


Lecturer: Dr. Giovanni Di Liberto

April 19, 2024

[0]Replication files are publicly accessible at github.com/keeleek42/MLproject.

# School of Social Sciences and Philosophy: Assignment Submission Form

| | |
|---|---|
| **Student Names:** | Shekhar Kedia, Stefan Keel |
| **Student ID Numbers:** | 23351315, 23366536 |
| **Programme Title:** | MSc. in Applied Social Data Science |
| **Student Name:** | Yana Konshyna |
| **Student ID Number:** | 23359606 |
| **Programme Title:** | Postgraduate Diploma in Applied Social Data Science |
| **Module Title:** | Introduction to Machine Learning HT |
| **Assessment Title:** | Homework 2 - Group Project |
| **Lecturer (s):** | Dr. Giovanni Di Liberto |
| **Date Submitted** | April 19, 2024 |

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar

I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write

Signed: Shekhar Kedia, Stefan Keel, Yana Konshyna Date: April 19, 2024

# Contents

# List of Figures

# List of Tables

# 1 Introduction and Data

The task was to provide a basis for decision-making to improve the Dublin Bikes station set-up, model an improved set-up, and display and explain the outcomes.

Observations for 2023 were used to account for probable seasonal, holidays, events, and weather impacts. The analysis of hourly average usage over the full year, the peak hours as well as the differences between weekends and weekdays resulted in a distinguishable pattern.

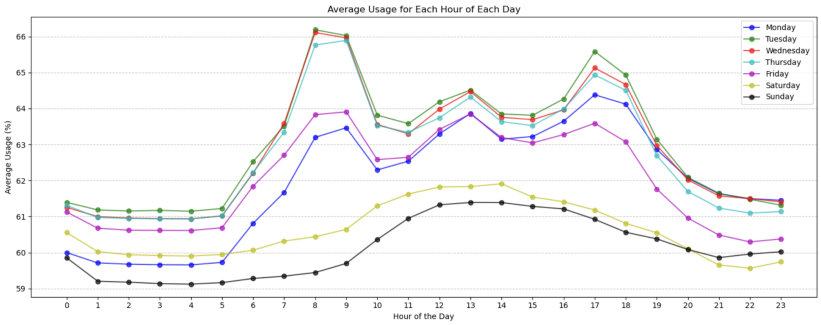



Figure 1.1: Full Hourly Average Usage for each Weekday

The creation of a subset focused on main traffic hours. The combination of traffic data from TomTom (Dub) and expertise from Yana Konshyna, who has extensive experience in radio advertising and is well-acquainted with peak traffic hours, provided the decision basis. The analysis of daily usage per hour was used as a test to confirm the chosen peak hours.

The peak hours are:

- Monday - Friday: 7-10 am / 12-1 pm / 5-8 pm
- Saturday - Sunday: 10-5 pm

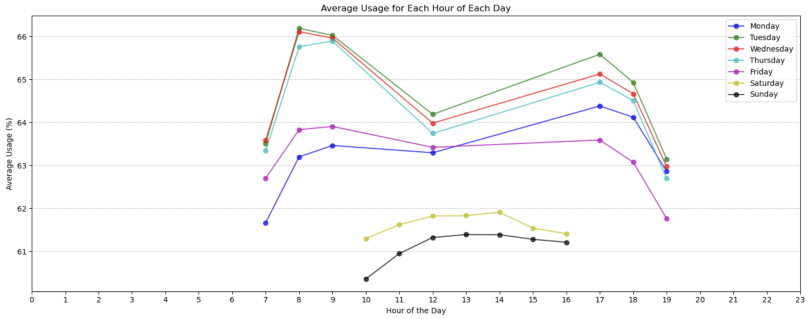



Figure 1.2: Peak Hourly Average Usage for each Weekday

It was found that all days have a similar pattern and that the split Monday - Friday and Saturday - Sunday was necessary. Further analysis of the dataset resulted in the exclusion of test station (507) and closed stations to create a valid dataset.

Quarterly data has been analysed as well but no significant differences were found.

Summarising, the initial dataset for the year 2023 had 1,994,400 rows. Focusing on the peak hours for all valid stations, the data was concentrated to a total of 581,066 rows.

# 2 Task 1: Remove and Expand

We ran various machine learning models to understand which model makes better predictions of usage level given the set of features. The **daily average usage level** calculated for each station was used as outcome variable (Y) and **Station ID, Available Bikes** and **Day of Week** were used as feature set (X). Various models have been tested to determine the best one according to the MSE.

Table 2.1: Model Performance Comparison

| Model | Train MSE | Test MSE |
|---|---|---|
| Ridge Base | 285.866 | 296.716 |
| Lasso Base | 286.127 | 297.123 |
| ElasticNet Base | 285.986 | 296.967 |
| Linear Regression | 285.866 | 296.716 |
| Random Forest Regressor | 77.431 | 174.639 |

**Random Forest Regressor** provided the best MSE. Cross-validation was carried out to check for robustness.

The model was used and **five** stations were identified that showed very low usage. As a second validation, a manual inspection of the stations in the dataset confirmed the model findings. The following stations were identified for exclusion:

Table 2.2: Stations to remove/reduce

| Station Name | ID |
|---|---|
| BROADSTONE | 116 |
| MOUNTJOY SQUARE EAST | 111 |
| FITZWILLIAM SQUARE EAST | 89 |
| HARDWICKE PLACE | 61 |
| GRANGEGORMAN LOWER (CENTRAL) | 104 |

A heatmap with all stations and their usage was created and three clusters, where either stations' capacity should be enhanced or new ones should be built, are circled in black.

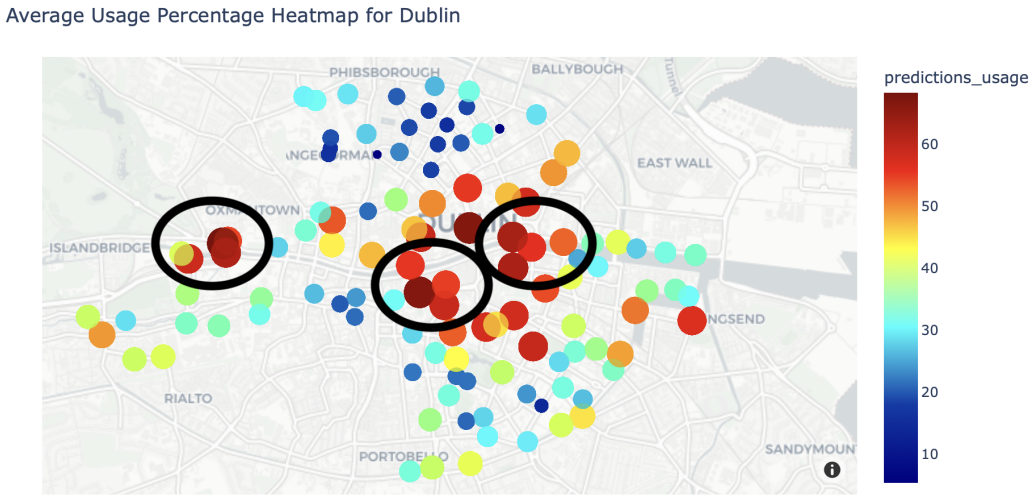



Figure 2.1: Heatmap with Stations and Areas to improve

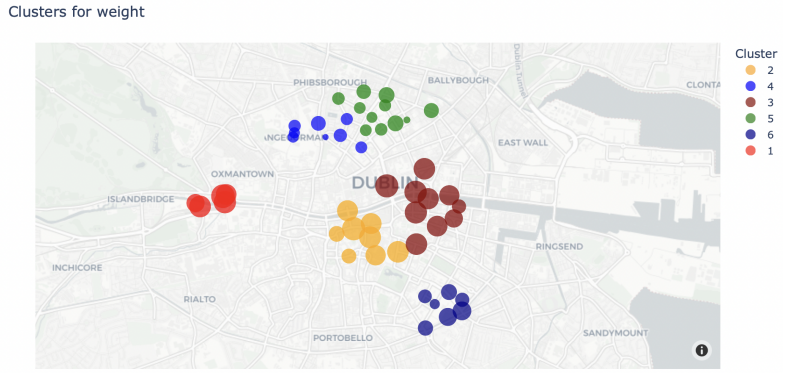
# 3 Task 2: Prediction

From the previous task, five stations were identified for exclusion and with the heatmap three clusters, where new stations were identified to be beneficial for the user experience, were circled. A high usage between 50 to 60 % was chosen to be optimal as there are still bikes available and it would not be a "dead" station.

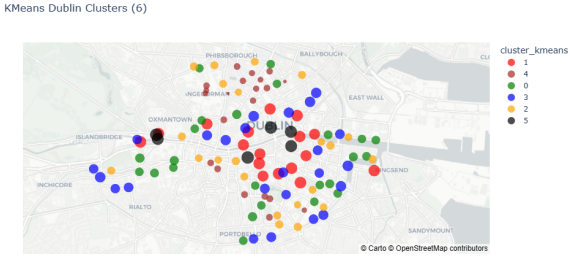
A **cluster-weighted model** was created to understand the load of each station and to accordingly make predictions for future usage. The understanding is, that the biggest effect of adding or removing a station is on the nearby stations. To achieve this, subsequent steps were undertaken:

- Calculate the total demand for bikes in the cluster (average demand in a day) and then calculate weights for each station in the cluster. The weights will help to understand how much of the total demand of the cluster is met by a particular station.
- Add a new station using Google Streetview in the high-usage cluster having a total capacity of bikes = average capacity of all stations in the cluster. A reduction in initial weights for all stations is expected due to the addition of the new station.
- The first step is applied to the low-usage clusters where the five stations from task one will be excluded. The initial weights for stations is expected to increase due to deletion of a station.
- Lastly, the revised or New weights, Cluster ID and Available Bikes are used as input features to predict the new daily usage percentages for all the stations.

KMeans clustering with Usage level, Longitude and Latitude (3.1b) was also done. Instead, the manual clustering technique, where stations adjacent to high and low usage stations, that need to be expanded, were added, was identified to be more feasible (3.1a) to predict the effect of changes in stations.




(a) Manual Clusters




(b) KMeans Clustering

Figure 3.1: Clusters with KMeans

After the exclusion of the five stations, the addition of three new stations, which location was done by assessing possible installation site using Google Streetview, and the calculation of the initial and new weight of the stations within the clusters (a metric), a prediction of the usage was made. Different models were tested which resulted in choosing Random Forest based on the best MSE.

Table 3.1: Model Performance Comparison

| Model | Train MSE | Test MSE |
|---|---|---|
| Ridge Base | 140.673 | 132.315 |
| Lasso Base | 141.298 | 131.738 |
| ElasticNet Base | 143.273 | 134.165 |
| Linear Regression | 140.658 | 132.294 |
| Random Forest Regressor | 10.302 | 71.914 |

Predicting the time course for the anticipated usage changes resulting from the removal of five stations and the installation of three new ones, a notable shift in usage patterns emerges. Stations categorized as 'low-usage' witness a surge in activity, likely attributed to redistributed demand. Conversely, stations previously categorised as 'high-usage' experience an improved distribution of activity, due to the increased capacity available in the cluster. This optimal solution enhances the general user experience and also reduces maintenance costs by dismantling unused stations.

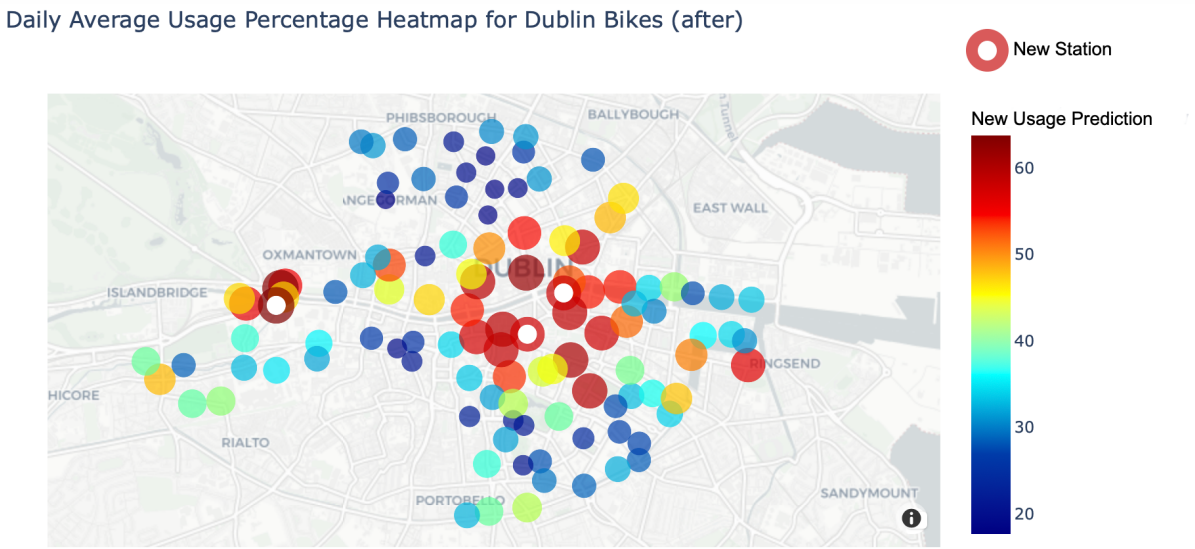



Figure 3.2: New Daily Average Usage Percentage Heatmap for Dublin Bikes.

## 3.1 Discussion

The goal is to maximise the usage of the stations whilst providing a constant amount of bikes to enhance user experience. It is of utmost importance to find the equilibrium between user experience and cost-effectiveness for Dublin Bike Company. For the future, it would be advisable to analyse the movements of the bikes to see how to further optimise the infrastructure.

# Bibliography

Dublin traffic report | TomTom Traffic Index. URL https://www.tomtom.com/traffic-index/dublin-traffic/.