ML group project BDP

Brendan Walsh, Darragh Kane O' Toole, Philip Kruger

March 2023

The github for this project can be found here.

Part I
Data Processing

We decided to use the data from all stations in Dublin. We used the following features:

- Bikes in use at this time last week
- Bikes in use at this time last month
- Bikes in use this day last year
- Temperature this hour
- Rainfall this hour
- Wind this hour
- Binary Pandemic (Yes/No)
- Binary Workday (Yes/No)

Some of these factors are more relevant than others, however this weighting will be sorted out by the ML models. We gathered the weather information from Dublin Airport; it's available on the same website as the bicycle data. We calculated the number of bikes in use by subtracting the number of bikes available at that time from the number of bikes available at the time that day when the most bikes were available (some time at night). This is necessary since bikes are lost, stolen, or more are added by Dublin county council. As such, it's not perfect, however, it's the best we have from the available data. The binary covid is 1(Yes) from the first measures on March 27th 2020 until when all restrictions where removed on 6th March 2022. Workday is according to weekdays and official bank holidays.
Some graphs which show the pattern of this data including the weekly and daily cycles can be found here.

Part II
Answers

1 Question 1

1.1 Assessing the impact of the pandemic on the city-bike usage;

Overview

- The pandemic massively reduced the use of bikes compared to the previous year
- This is shown by the graph, particularly useful is the pink line which shows the previous years usage over the pandemic usage.
- Although using the exact previous year is a very simple method, it indicates the vast drop off in bicycle use during the pandemic

3 Linear regressions where also performed to get a coefficient for our covid dummy variable. The full equations can be found below in the appendix. The first one simply takes all our variables with the exception of past bike usage. The second is identical however it is performed on a subset of the data which excludes the night[1]. The third regression includes the past bike usage which removes the relevance of the covid dummy variable however it has the highest r squared of the three regressions. The R squared for our regressions are very bad they are 0.18 for the first, 0.21 for the second and 0.44 for the third. This is probably due to not separating the data into seasonal (daily, weekly, etc) components. However, the coefficient for covid (which does not depend on any cyclical processes appear to be rather good[2]. The coefficient for covid for regression 1 is -26.7 which is very close to the mean difference of bike usage during pandemic and non pandemic times which is -27.33. Furthermore, for regression 2 the coefficient is -43.8 and the difference in the mean of pandemic and non pandemic times during the day is -42.9. The three regressions mapped on the original data can be seen below. It is worth noting that while the regressions look VERY off, it is less so than it first appears as due to the scale of the x axis the peaks are far more apparent than the rest, so on average the actual values are closer to the regression estimate than the graphs appear to portray.

2 Question 2

We used seven different models - Linear Regression, Ordinary Least Squares, Lasso Regression, Ridge Regression, Random Forests, Support Vector Regression, and K Nearest Neighbours (KNN) Regression to determine how our covid variable and others affected bicycle usage. To account for non-linear relationships between our weather variables (rain and temp), we squared, cubed, and exponential those two variables. Ultimately, we ran 28 models. The weakest of our 28 models was K Nearest Neighbours with an average Mean Squared Error of 4663.849 and an average R-Squared of 0.160794. The strongest of our 28 models was Random Forests with an average Mean Squared Error of 4118.252 and an average R-Squared of 0.2589679. The best fitting of the random forests models, surprisingly, was the multiple regression, not the polynomial regression. However, these differed after three decimal places. This model provides feature importance figures of 0.43197317, 0.05032412, 0.24372927, and 0.27397343 demonstrating temperature followed by covid as the most consequential in determining bicycle use. Overall, from all our models, we see the covid dummy variable and rain to negatively impact bicycle usage, and increasing temperature and workday to dramatically increase bicycle usage.

The Dataset was plotted again with a simple additive term from the beginning of the pandemic period. The first graph uses the coefficient from the first regression and the second graph uses the coefficient from the second. These projection do not appear to bring the bike usage during pandemic time up to the pre covid time period. This is since these graphs show the density of the max usage more than average usage. However, these projection do bring the average bike use up to pre pandemic levels. A good illustration of this can be found in the graph below we can see that orange projection is slightly above the blue original except during the high usage time of rush hour and lunch. This is since we are using an additive term from a regression which is applying the average effect and can thus not reach the peaks.

3 appendix

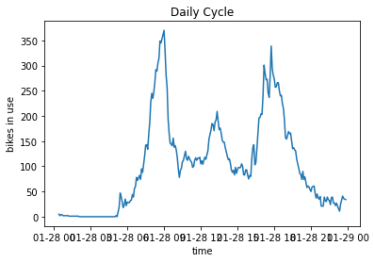
1. bikes = 12.409827829205625 + 4.009318521489726 (temp) + -10.804690902778974 (rain) + 28.14334492667443 (workday) + -26.784976345583253 (covid) + 1.8738942778853254 (wind)

[1] This is because both during covid and non covid times the ridership during the night is low to non existent

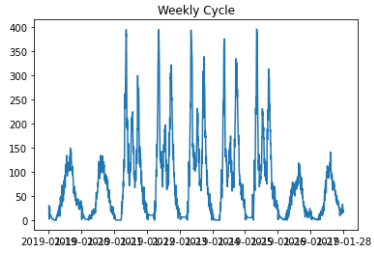
[2] This is not true for model 3 where the covid explanitory power of the covid dummy variable is probably subsumed into bike usage 1 week and month ago

2. bikes = 53.62117296751988 + 3.5851503391942616 (temp) + -12.600695012797239 (rain) + 37.058611808909255 (workday) + -43.84066475817235 (covid) + 1.6391592130255948 (wind)
3. bikes = -16.236957954611455 + 2.128572641174578 (temp) + -9.501432754760447 (rain) + 16.299968239470523 (workday) + -6.217248937900877e-14 (covid) + 0.9627233018504205 (wind) + 0.26642792530659776 (week) + 0.13847001714669194 (year) + 0.10788809117422904 (month)

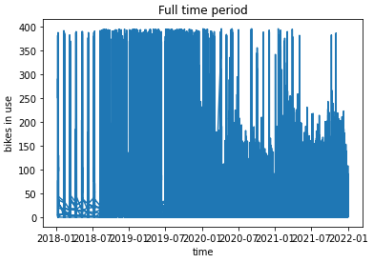
4 Graphs



(a) This shows the bike usage with peaks for rush hours and lunch

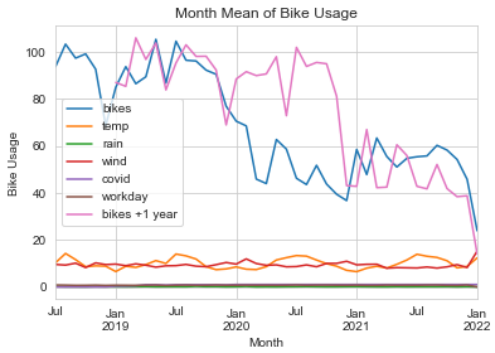


(b) The weekly cycle with 2 weekends and the workweek in between

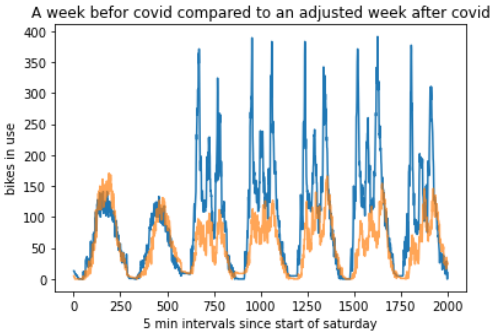


(c) The full timeperiod. The drop for covid can clearly be seen

Figure 1: The data[3]



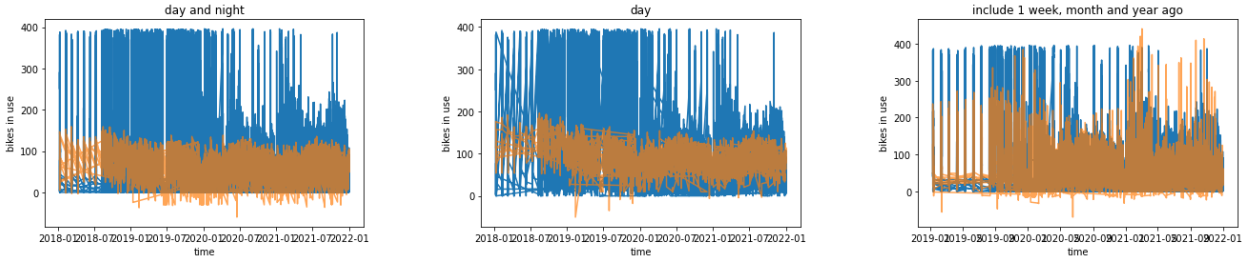
(a) features plotted



(b) comparing a non covid week in 2019(blue) with our projection of a similar week (orange) in 2020 without the effect of the pandemic

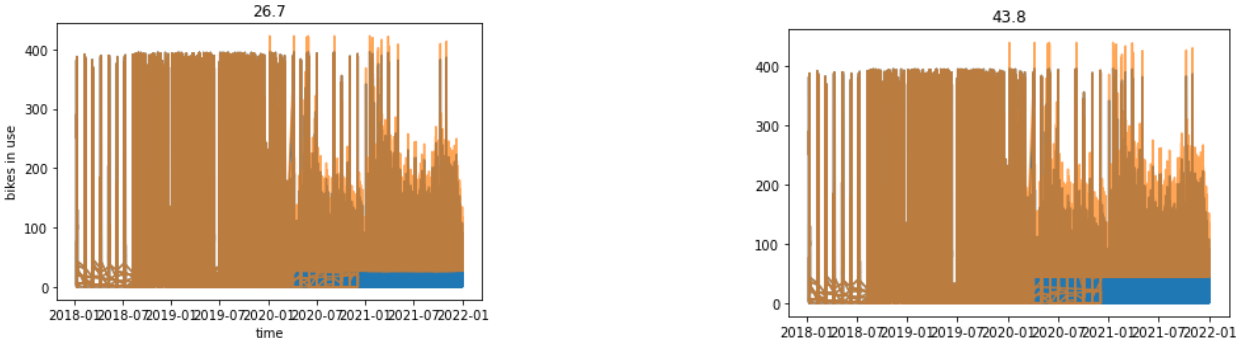
Figure 2: Regressions

Note: these graphs are not officially part of the report but i've included them since they're interesting in showing difficult it is to represent the entire time period in 1 graph



(a) Regression 1 (b) Regression 2 (c) Regression 3

Figure 3: Regressions



(a) Projection 1 (b) Projection 2

Figure 4: Regressions