

CSU44061 Project Report: Weather and Cycling in Dublin

Conall Tuohy 18320949, Tanmay Kaushik 18308341, Liam Bartsch 18330219

December 3, 2021

Contents

1	Introduction LB, TK	2
2	Dataset and Features LB	2
3	Machine Learning Models TK, CT	4
4	Experiments/Results/Discussion CT, TK, LB	5
5	Summary CT, TK, LB	8
6	Contributions CT, TK, LB	8
7	GitHub Link	8

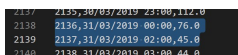
1 Introduction LB, TK

Over the last few years there has been a huge increase in bicycle rentals especially in densely populated areas. There are many companies which have set up cycle stations all over the city from which a person can rent a bike for a short or long duration. Looking at the huge popularity of this model, our team was really interested in seeing the effect of change in weather on this rental pattern. We know that weather plays an integral part in our everyday lives and understanding it's effect on our actions such as cycling can make the difference in our resiliency and situation preparedness. The problem of weather and it's relationship on our cycling habits is interesting because it is often overlooked and better understanding leads to improvement and better lives. This is the reason we used a combination of weather attributes in our dataset with different machine learning models which will be explained in detail below to see the effect of weather on bike rentals.

The input to our program are weather data such as precipitation amount, air temperature, mean wind speed, visibility, cycle counter (i.e. the number of bikes that have been counted by cycle counter at different locations) as well as the count of cyclist in and out of stations at different locations. We used a variety of techniques such as Linear Regression, Lasso regression, Ridge regression and *knn* classifier to output predictions in the count of cyclists depending on the weather conditions.

2 Dataset and Features LB

Our dataset consists of weather data and cycle data. We gathered the data from data.europa.eu and data.smartdublin.ie websites. We manually downloaded the relevant CSV files that are available at those URL's and renamed the files to something more readable and then read them in our python script. Originally there was more data than we needed in the csv's so our preprocessing consisted of collecting the relevant data points by eliminating the weather data before the 1st of January 2019 and after the 1st of October 2021 since we did not have cycle count data for outside that range. This ended up being 10 columns with 24 095 rows, a total of $10 \times 24\,095 = 240\,950$ data points. We had to remove one entry on the 31st of March at 01:00 for the weather manually since that was a daylight saving time that the cycle count did not have as shown in the screenshot below.



2137	2135,30/03/2019	23:00,112.0
2138	2136,31/03/2019	00:00,76.0
2139	2137,31/03/2019	02:00,45.0
2140	2138,31/03/2019	03:00,44.0

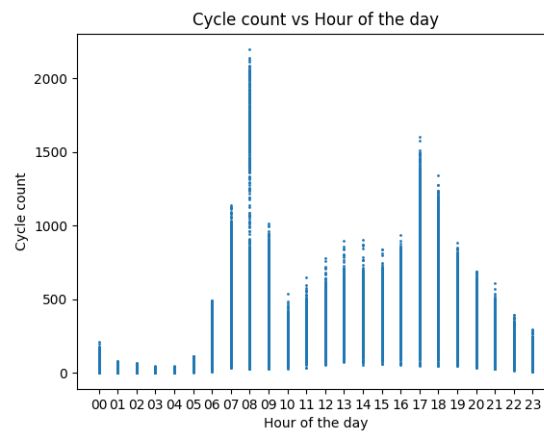
Figure 1: Screenshot of daylight saving time gap for cycling count

We also had to count the totals and discard the irrelevant columns in the cycle and weather data. We did not find it necessary to normalize or augment

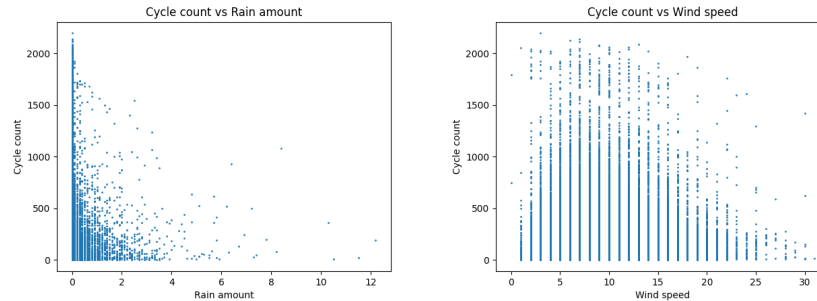
the data.

We decided to have our input features be the different weather metrics as well as the hour of the day to get our output feature which is the number of cyclists during that hour. Originally we thought to map it over seasons and days during the year, but since we only had data for 3 years and since the different seasons variables were accounted for in the weather data we ended up using the hours of the day as another input which allowed us to combine the different days, months and years together. Our rationale seemed logical since what we are interested in determining what impacts cyclist the most which would more likely be the hour of the day over than the month or the year.

We plotted the data in terms of scatter plots to see if there were any noticeably patterns that emerged visually at the start.



In the above plot, we can see that most cyclist cycle in between 08:00 and 09:00 in the morning which makes sense since that's generally when the day is considered to start and people cycle in to work. There is also a peak at between 17:00 and 18:00 as that when the day ends and people cycle home.



As we can see in the above two plots, there seems to be less cycling going on when there is more rain and faster wind speeds, indicating that cyclist tend to cycle less in heavy rainfall and heavy wind. There are other plots with the different input features but to keep this report concise we decided not to include them here. They can be seen in our GitHub repository under [Images](#) and are explained in more detail in the readme [Readme](#).

3 Machine Learning Models TK, CT

For all the models we decided to split the data into 80% training and 20% test data as this would give the model enough data to train.

1. Linear Regression

Linear regression will try to predict the number of cyclists (output feature) from our weather and time data (input features). It will try to find a linear relationship between the input features and output feature as a line that best fits the data. We then fit the model on the training data and asked the model to make a prediction using the test data. The biggest problem is if the data is not linear which won't work for this model.

2. Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression uses shrinkage to encourage models with fewer features. Lasso regression uses L_1 regularization to eliminate (shrink down to 0) some features and make the model simpler. This helps with over fitting as it shrinks down the features that contribute the least.

3. Ridge Regression

Ridge regression uses the L_2 regularization which does not eliminate features like Lasso, which subsequently makes it harder to interpret than Lasso regression. This is best used when we have a lot of features as is the case with our data.

4. knn Regression

knn regression can be very effective in regression problems. The way this

method works is by using the similarity of features to make it's prediction and predicting the k nearest numbers output.

5. Dummy Regressors

We used two dummy regressors to compare our models against, one was using the mean where it always predicted the mean of the data and the other was where it predicted the median of the data. The use of the mean was useful because it represents the central spread of all the data including extremes values such as very low values and very high values compared to the rest of the data. The median dummy regressor is not badly effected by extreme values which is why we chose it as our second baseline model.

4 Experiments/Results/Discussion CT, TK, LB

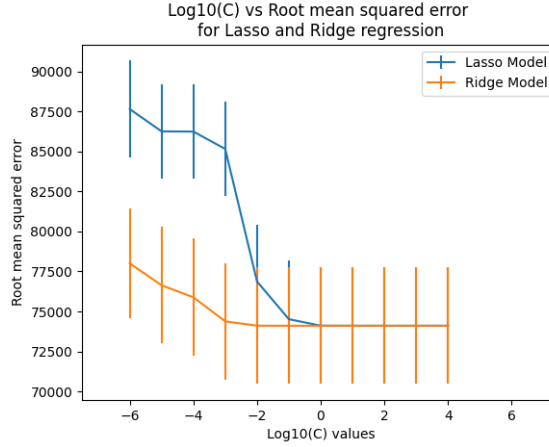
For each model except the dummy models we used 5-fold cross validation to split the training data into 5 folds, that were then used as training and testing data. For Lasso, Ridge and k nn regression, becuse we had to choose different hyperparameters we picked the best performing models (the ones with the smallest root squared mean error) to use as comparison, in our final analysis and when contrasting the models to each other. For each model, as well as the dummy models, we calculate the root squared mean error to be able to compare their performance between each other to accurately show which had the least room for error. The root squared mean error is our primary metric.

For linear regression there were no hyperparameters so after the cross validation we fit the model on the entire training data and then make our prediction using the updated model. We wanted to see the coefficients for our linear regression to see which features were most significant according to our model. The results are summarized in the table below but the most significant features seemed to be Date & Time, Rain, Temperature, Sun Duration and Cloud amount, all of which seem consistent with what we would guess might influence a cyclist's decision to cycle.

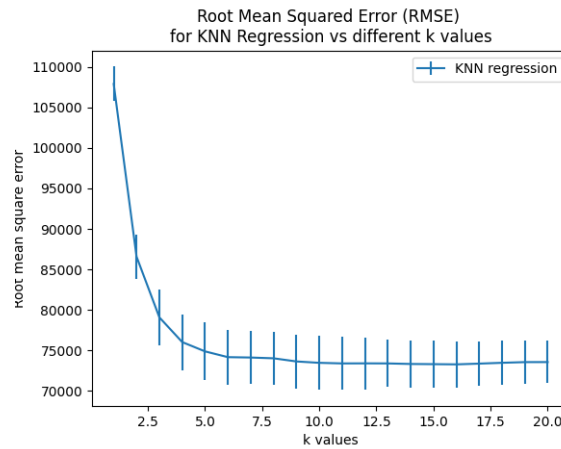
Input Features	Coefficient
Date & Time	9.482939
Rain	-6.936782
Temperature	6.638328
Humidity	-1.472087
Wind Speed	-1.938624
Wind Direction	-0.007346
Sun Duration	193.541913
Visibility	0.000310
Cloud Amount	9.045273

Figure 2: Table with the coefficient of each feature

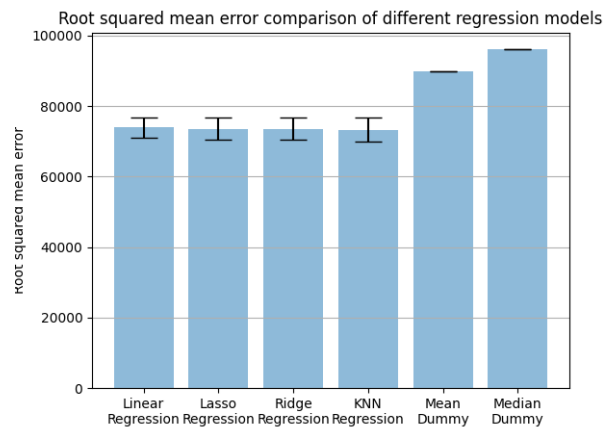
For Lasso regression we used cross validation to select the hyper parameter c by multiplying the value of c by 10 each time. For each different c value we then use 5 fold cross validation on the training data and kept track of the mean and standard deviation of the root squared mean error. We noticed that the model that performed the best was when $c = 10$. For Ridge regression a similar approach was used where we selected the c parameter by using cross validation and then on each model using 5 fold cross validation to then evaluate the model and the best performing model was when $c = 0.1$. We then plotted the different c values for Lasso and Ridge to see the root squared mean error evolution between the two models. We had to use $\log_{10}(C)$ to be able to view the significant changes in the models. As you can see in the graph below both models seem to minimize their errors around $\log_{10}(C)$ values of around -1 to 1 which is consistent with our best performing c values selection for each model ($c = 10$ for lasso and $c = 0.1$ for ridge).



For k nn regression we also used the same approach but instead of selecting the hyper parameter c we used cross validation to select the parameter k which determines how many nearest neighbors should be taken into account. We also used 5 fold cross validation to improve this model. As you can see on the graph below the error seems to be minimized at around $k = 10$ which is why we selected this as our best k nn model for the final comparison.



We then our two baseline dummy models for comparison (to make sure our models are actually predicting something sensible), one for the mean and the other for the median. The comparison of all models is shown in the plot below along with the error bars.



As we can see all of our 4 models (linear, lasso, ridge and knn) perform significantly better than our dummy models which means that our models are actually doing a better job than just predicting the mean or median results. However our 4 models seem to perform around the same which could mean that they aren't doing a great job in general and are over fitted.

5 Summary CT, TK, LB

Our results showed that our best performing model was the k nn regression model and this makes some sense because cyclist are humans after all which means that they tend to follow patterns and think alike leading to a lot of similar data points which helped our k nn model which relies on following the pattern of it's neighbors.

We have also seen that rain decreases the amount of cyclist present on the road, which is what we expected to see as it's more uncomfortable and difficult to cycle in the rain. The most frequent times to cycle were between 08:00 and 09:00 in the morning and between 17:00 and 18:00 in the evening which aligns with the standard work hours of the day. What we learned is that cyclists seem to include weather patterns such as rain in their decision making process of whether or not to cycle as can be noticed by the fewer amount of cyclists on the road with heavier rainfall. Our k nn model works best out of all of our tested models because it does the best job of thinking like it's neighbours which is not too different to how humans act and behave as if most cyclists decide not to cycle in the rain then a reasonable cyclist might also decide not to cycle.

We learned that cyclists do use different weather features to gauge whether or not to cycle but it is still unclear how accurate this is and how well does it scale in different locations or areas, further research would be required.

6 Contributions CT, TK, LB

Conall Tuohy = CT, Tanmay Kaushik = TK, Liam Bartsch = LB

All three of us worked equally on the project proposal. LB worked mostly on the processing of data and visualisations. CT and TK worked mostly on the Linear, Lasso, Ridge and K nn regression models. LB also worked on the Mean and Median dummy regressors. All three of us worked closely with each other to double check each other's work. Overall we all contributed equally to the project.

7 GitHub Link

[Weather and Cycling in Dublin Project](#)