# Group ID - MSc in Data Analytics

Author: M. Brabazon

e-mail: [sba22216@student.cct.ie](mailto:sba22216@student.cct.ie)

Student ID: sba22216

**Abstract**

*The goal is to predict cycle volume as a percentage of overall (cycle and traffic) volume for a known day of the week, and with given weather conditions.*

**Introduction**

This project follows the CRISP-DM project management framework to gather and analyse data relating to Dublin City Council traffic and active travel. Its objective is to present insights to Dublin City Council that may assist them in the area of traffic management by way of moving towards active travel.

**Project Management Framework**

Consideration was given to three different project management frameworks for use in this project: KDD, SEMMA and CRISP-DM. Of the three frameworks, KDD is the most rigid in that it requires that every stage in the process be completed before iterating through the entire process again. Given that both the business area and the data was unfamiliar to the author, the analysis would involve much exploration, discovery and reassessing at most stages in the project. For these reasons, KDD was deemed unsuitable for this project.

It has been argued that SEMMA and CRISP-DM are very similar (Azevedo, A. and Santos, M.F. 2008), and perhaps it is merely a matter of preference which is selected for any particular project. However, CRISP-DM was selected because it provides clearer guidelines in its phases, and it allows the project to commence with no knowledge of the topic or data. Understanding is built by iterating through the process steps as necessary.

In accordance with the CRISP-DM framework, the data analysis was carried out in the following phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

To illustrate the process flow throughout the data analysis, this paper mirrors the phases of the data analysis.

Although the phases are clearly defined above, overlapping did occur throughout the project. Particularly, between the data understanding phase and data preparation phase. This was largely due to the fact that it was necessary to reduce the size of the datasets early in the project due to limited hardware resources. Initial data preparation steps that were required for the reduction of data were covered in the data understanding phase to maintain the process flow. Once the data was combined into one manageable dataset, the phases were more closely followed, with iterations a necessary.

**1. Business Understanding**

What insights or recommendations may be beneficial to Dublin City Council in the area of transport? What resources are available to the researcher and how may the data be acquired?

Ireland’s 2021 Climate Act commits the country to reducing its emissions by over 50% by 2030, and to net-zero emissions by 2050.

Dublin City Council investment in active transport infrastructure.

Weather effects on commuter behaviour.

With a broad understanding that Dublin City Council is committed to active transport, emissions reduction and the development of greener infrastructure, the next thing to understand was what resources does DCC have for collecting data that may provide the basis for useful research. An initial search for ‘transport and infrastructure’ on: data.gov.ie, under Dublin City Council, returned a variety of data, including volume data for traffic and drawings for strategic cycle network. Further searches on data.gov.ie and data.smartdublin.ie found complete sets of data for traffic volume and cycle counts, along with location data for cycle counters. A full list of resources can be found in the appendix.

The objective was to predict the proportion of cycle volume to overall volume on a given day of the week, and known weather conditions

**2. Data Understanding**

With a list of data resources identified, the next phase was to gain an understanding of the data and how it could be used to achieve the objectives set out in the previous phase. To start, the locations data were downloaded so that the data collection sites could be visualised to enable more specific selection of the required data.

|  |  |  |
| --- | --- | --- |
|  | **Source** | **Local file** |
| **SCATS GPS coordinates** | data.smartdublin.ie | scats\_site\_locations.csv |
| **Cycle Counter Locations** | data.gov.ie | cycle-counts-2020.csv |

Table Data downloads

*2.1 Selection of count data sites*

An initial look at the Cycle Counter and SCATs locations revealed that there is a very limited number of cycle counters to choose from. These counters were plotted ono a map, along with the SCATs locations. (see Figure 1)

A picture containing map of all traffic and cycle counters

Description automatically generated

Figure All Cycle Counters and SCATs sites in Dublin City Centre

From this figure, a corresponding SCATs traffic counter was chosen for each Cycle Counter on the map (see Figure 2)

A picture containing map of all cycle counters and the corresponding traffic volume counters

Description automatically generated

Figure All Cycle Counters with paired SCATs traffic counters

*2.2 Download of source data*

Firstly, three sets of data for cycle counts were downloaded: 2020, 2021 and 2022. Unwanted columns and rows were removed. The 'Date & Time' values checked for consistency and converted to datatime datatypes, as these values would be the basis of merging the data with the other sets.

At this point, the count data for each of the Cycle Counters was checked to ensure that data was present, with enough observations for the analysis. It was discovered that the cycle counter on Guild Street had not been in operation after 2020. For this reason, Guild Street and its paired traffic volume counter were eliminated from the data.

The table below shows the remaining Cycle Counters and the SCATs sites to which they were paired. Note: the Guild Street cycle counter has been omitted due to lack of data available for this counter

| **Cycle Counter** | **Import as** | **SCATs Counter** | **Import as** |
| --- | --- | --- | --- |
| Charleville Mall | CycVolCharleville | 282 | TraVolCharleville |
| Grove Road Totem | CycVolGroveRd | 157 | TraVolGroveRd |
| North Strand Rd N/B | CycVolNorthStrand\_nb | 194 | TraVolNorthStrand\_nb |
| North Strand Rd S/B | CycVolNorthStrand\_sb | 194 | TraVolNorthStrand\_sb |

Table Traffic and cycle counters

Secondly, the weather data was downloaded from Met Éireann (met.ie/climate/available-data/historical-data). Dublin Airport was selected as the site for weather downloads as it had more features available than other nearby weather stations.

Lastly, when downloading the SCATs data, it was discovered that each of the 29 monthly data files were approximately 500MB in size. For this reason, these files were not stored locally. Instead, a data file containing the download links to all of the SCATs data files was compiled and saved locally. This list was then accessed programmatically and each data file was read in, adding only the data for the selected SCATs sites to a Pandas DataFrame.

*2.3 Combination of data*

Before combing the data into a single dataset, the SCATs data needed to be reduced further so that it would contain only one observation per hourly observation in the other datasets. The SCATs data contained multiple observations for each hour. These observations are for each detector associated with each SCATs site. It was decided to combine these detector counts by grouping the observations by hour and site and calculating the average counter reading. However, it was noticed that many of the detectors had very few or no observations. Factoring these into the calculation would reduce the average count for each site observation significantly. To avoid this count reduction, detectors with less than 100 counts per month were removed from the data. The remaining data was then grouped as outlined above.

The observation time (hourly) was used as the index to merge the three sets of data. As such, the datetime format wad homogenised across all three datasets. Finally, a check for duplicates was performed on this column and duplicates were removed.

The three datasets were merged and the resulting dataset was saved to one manageable data file: ‘data/cycle\_traffic\_weather.csv ‘. This dataset was then be used for the rest of the Data Analysis.

*2.4 Data Exploration*

The first step to understanding the features within the dataset was to visualisation all of the features at once. A simple histogram of all features was chosen to start (see figure 3).

Diagram, engineering drawing

Description automatically generated

Figure Initial visualisation of features

The histograms showed some characteristics of the features and raised some questions regarding validity and distributions. For example, with CycVolGroveRd, why is are there counts recorded of over 1000 when almost all of the counts appear to be less than 500? Could this indicate the presence of outliers? To investigate this further, a box plot was created for that feature (see figure 4).

Chart, box and whisker chart

Description automatically generated

Figure Box Plot for CycVolGroveRd

The figure shows that there are potentially many outliers. However, when plotted in the same way, the other cycle counts show similar results. Firstly, it was noted that the low mean was heavily influenced by the fact that there are a lot of hourly counts of zero or very small numbers, probably recorded at night. The fact that the results are consistent across all cycle counters suggested that the values are valid and may not need to be dealt with.

*2.5 Statistical Description of the Data*

Before carrying out a more detailed description of the data, it was first necessary to deal with the missing data. A quick search for nulls revealed that there was a number of missing values for the volume counters (*See Table 3 below*). To address these, the steps in phase 3.1 below were carried out next.

|  |  |
| --- | --- |
| obs\_time | 0 |
| CycVolCharleville | 30 |
| CycVolGroveRd | 3 |
| CycVolNorthStrand\_nb | 35 |
| CycVolNorthStrand\_sb | 659 |
| TraVolGroveRd | 703 |
| TraVolNorthStrand | 726 |
| TraVolCharleville | 675 |

Table Missing Values

With the null values filled, the data features were then examined further to learn more about their characteristics.

**Poisson Distributions**

At first glance, the distributions for the north strand traffic counter looks as though it could possibly fit a Poisson distribution, but with the exception of the high zero count. This was plotted along with the wind speed (which looks more like the Poisson distribution but is not count data). The figures below show both plots. As expected, the windspeed data fits the qq-plot for Poisson distribution quite well. The traffic counter data does not fit quit so well but this was also expected given the high number of zero counts.

Chart, line chart

Description automatically generated

**Normal Distributions**

The figure below, on the left, shows the windspeed distribution which appears to be lognormal. The figure on the right shows the distribution after the log function has been applied to the windspeed values. The figure on the left is closer to the normal distribution. This type of method of transformation can be applied to feature that need to be normalised for algorithms to work. Such transformations were not necessary in this project.

**Chart, histogram

Description automatically generated**

*2.6 Check data validity*

Further checks were then carried out for other data that may contain errors or may not make sense. For example, negative values for any of the count features or weather features such as rain or clamt (cloud amount).

**3. Data Preparation**

*3.1 Missing data*

It was discovered, in section 2.5 above, that some features contained missing values. The number of missing values for the volume counters was noted to be between 3 and 4%. While this is not an extremely large percentage of the observations, and an argument could be made for the removal of these observations, it was decided that an effort would be made to impute these missing values.

Firstly, the sequence of the missing values was checked to see if they were randomly dispersed throughout the data or if randomness could not be assumed. Whether or not the data is missing at random will impact the effectiveness of the chosen imputation method (Xiao-Bai Li, 2009). The check revealed that the majority of missing values were in sequence. For example, a cycle counter or traffic volume counter will have no recorded observations for a number of consecutive days. The help understand the nature of the missing values, one counter was selected (CycVolNorthStrand\_sb). The volume counts for the other cycle counters was plotted for one of the time periods when the CycVolNorthStrand\_sb observations were missing (see Figure 3).

Chart, histogram

Description automatically generated

Figure Missing Values CycVolNorthStrand\_sb Missing, week 19, 2022

For comparison, the same features were plotted for the same week in the previous year, 2021, to observe the trends in the cycle count features (see Figure 4). Traffic volume count plots revealed similar trends.

Chart, histogram

Description automatically generated

Figure Cycle Volume counts, cycle volume counts, week 19, 2021

Due to the obvious trends in this feature over time and in relation to the other features, the univariate approaches to imputation, including SimpleImputer, back-fill, forward-fill and univariate interpolation were quickly ruled out. From the multivariate methods available, the K-Nearest Neighbour method was selected because it requires no training and it is easy to implement. There is also a small number of features, so this will not present a problem to the KNN approach.

The KNN imputation method was performed with various values for the n\_neighbors hyperparameter. With a value as low a 1, it was observed that the method performed well as the pattern began to resemble that observed in Figure 4. Values of 2 and 3 represented the 2021 trend mor closely and the value of three was eventually selected. Figure 5 below shows the same period as Figure 3 above, but with the imputed values added.

Chart, histogram

Description automatically generated

Figure Imputed Values, cycle volume counts, week 19, 2022

Similar results were observed for the traffic volume counts after imputation. Given that the amount missing data for features was low at 3 to 4%, the visual inspection from the plot in Figure 5 above was deemed sufficient evidence a successful imputation of missing values.

*3.2 Feature Amendments*

#### **Get sum of North Strand Rd Cycle counters (Northbound and Southbound)**

The cycle counter data contained in both CycVolGraveRd and CycVolCharleville contain data for both northbound and southbound bicycles. However, CycVolNorthStrand\_nb and CycVolNorthStrand\_sb features contain this data separately. These two features we simply summed and recorded in the new feature: CycVolNorthStrand. The north and southbound features were then dropped.

#### **Drop potentially problematic features**

A closer look at the ‘sun’ feature revealed that this value which is described by Met Éireann as “Sunshine duration (hours)”, is calculated on an hourly basis. For an hour of constant sunshine, the value would be 1.0 whereas an hour of dark cloud would have a value of 0.0. However, every hour of darkness has a value of 0.0 so this could lead to unexpected results. Also, taking the average of this value, when grouping the observations by day, would be less representative of the weather conditions at different times of the year. Therefore, this feature was dropped from the data.

Similarly, the ‘wddir’ “Predominant Wind Direction (degree)” was dropped. This was measured in degree and could easily become meaningless when averaged for daily grouping.

**Create combined columns for cycle volume and traffic volume using mean()**

Next, all three cycle counter features were merged together by creating a new feature: CycleVolume which was assigned the mean of the three cycle counter values for each observation. The individual counter features were then dropped. The same process was performed on the three traffic counters, creating the new feature: TrafficVolume.

*3.3 Group data to daily*

With the problematic features removed, the data was then grouped by day. The mean of all input and output features was calculated for the grouping. This reduced the number of observations for machine learning to 882. This was considered sufficient.

*3.4 Additional Features*

**Create the new CycleProportion feature**

The CycleProportion feature was created for each row as follows:

*CycleVolume / (CycleVolume + TrafficVolume)*

This new feature provided the target values for the machine learning. It was the change in the proportion of cycle volume to overall volume that the machine learning models will attempt to predict. Both the CycleVolume and TrafficVolume features were then dropped.

**Add new categorical feature for day of week**

It was observed earlier that both the cycle and traffic counts vary consistently depending on the time of day. The day of the week also appears to have an influence in the volumes, albeit to a lesser extent. For this reason, a day\_of\_week categorical feature was added to the data: a numeric value between 0 and 6, stored as datatype string to allow for encoding in preparation for machine learning.

**4. Modelling**

Supervised learning was selected as the machine learning approach for this project. This selection was made because the data contains target / output values from which the machine learning models can learn by example. With the supervised machine learning approach, the data is split into training and testing subsets. When the learning is complete, an attempt is made to predict the target / output values for the test subset. The results are then compared to the actual observed outputs in the test subset. This comparison is the basis for measuring the success of the model.

*4.1 Model Selection*

The objective was to predict the proportion of cycle volume to overall volume on a given day of the week, and known weather conditions. This is a regression problem with a continuous numeric target. For supervised learning with numeric prediction, the options include: Decision Trees, Linear Regression, Random Forest, Neural Network and Gradient Boosting Tree.

*4.2 Benchmark Model – Decision Tree*

As a benchmark, the Decision Tree model was selected. This would run quickly and provide some insights quickly. Namely, are predictions about the output able to be made from the inputs. And which input features influence the prediction most. The benchmark model was used to provide an initial estimate of how well the inputs can be used to predict the outputs. It is not intended to perform as well as the main learning algorithms. It allowed issues to be corrected easily without investing too much time running the more processor-intensive algorithms. Issues could be detected when, for example, receiving a negative R2 value, meaning that the prediction is less accurate than the mean. This would not necessarily mean that there is an error in the data preparation, but it would, in most cases, signal the need for further investigation.

The Decision Tree model has high variance, which was apparent when re-running the model, even without changing the data or hyperparameters. The values returned such as mean squared error and R2 would vary considerably between runs. Decision Tree models are also susceptible to overfitting. The models selected for the training both reduce this variance.

*4.4 Model 1 – Random Forest*

The Random Forest reduces the variance, as it uses Ensemble Learning and is based on the Bagging algorithm. Essentially, it creates many different trees based on randomly selected subsets of the training data. It builds and combines many small trees. This approach helps to prevents overfitting. Random Forest models take a lot longer to run but they avoid many of the downfalls associated with Decision Tree models. If there was a large number of features in the data, these may have needed to be reduced in some way or the Random Forest model may have been deemed unsuitable to this project. This is because the time taken for this model to execute is greatly affected by larger numbers of features. As the number of features in the data is small, this was chosen as the first model.

*4.5 Model 2 – Gradient Boosting*

For comparison, a second model was selected: Gradient boosting. Similar to Random Forest, Gradient Boosting uses many trees. However, how the trees are built is different in each model. Gradient Boosting builds trees based on the previous existing trees. It selects week learners and designs new trees to improve its learning. Random Forest however, builds trees independently of each other. Another difference between the two models is that Gradient Boosting calculates the results as the learning is progressing, whereas Random Forest combines them at the end, by average in the case of regression. It is generally understood that Gradient Boosting can perform badly when there is a lot of statistical noise present in the data. However, the data used in the machine learning phase of this project was calculated using averages and so this drawback of Gradient Boosting was not considered to be concerning.

*4.6 Comparisons*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAE | MSE | Root MSE | R2 |
| Benchmark | 0.0427 | 0.0033 | 0.0578 | 0.2731 |
| Random Forest | 0.0377 | 0.0026 | 0.0513 | 0.4289 |
| Gradient Boosting | 0.0376 | 0.0026 | 0.0516 | 0.4218 |

*4.7 Hyperparameter Tuning*

The model chosen for further learning was Gradient Boosting. It could also have been Random Forest as both models performed similarly.

After running GridSearchCV with a variety of hyperparameter values. The following values were returned for the best application of the Gradient Boosting algorithm:

Best parameters are: {'criterion': 'friedman\_mse', 'learning\_rate': 0.05, 'max\_depth': 2, 'max\_features': 'sqrt', 'n\_estimators': 250}

**5. Evaluation**

*5.1 Evaluation Results*

The parameter values returned by the cross validation method were used to run the Gradient Boosting model again. The results were as follows

mean\_absolute\_error 0.03716312704926352

mean\_squared\_error 0.002642587251261331

root\_mean\_squared\_error 0.05140610130384652

r2 0.4257378571760516

This produced only a minor improvement over the initial run of the model. This warrants additional attempts to tune the hyperparameters (time permitting).

*5.2 Determine Next Steps*

Chart, bar chart

Description automatically generated

**6. Deployment**

*6.1 Report*

Weather conditions are an important factor in predicting commuter behaviour in Dublin city and should be taken into account when designing smart traffic management systems.

## **Acknowledgements**

The idea and parts of the formatting for this template came from the Association for Learning Technology (UK) ALT-C 2004 Research Paper Format Template.

## **References**

Azevedo, A. and Santos, M.F. (2008) ‘KDD, SEMMA and CRISP-DM: a parallel overview’. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=edsrca&AN=rcaap.com.recipp.recipp.ipp.pt.10400.22.136&site=eds-live

Xiao-Bai Li University of Massachusetts Lowell *et al.* (2009) *A bayesian approach for estimating and replacing missing categorical data*, *Journal of Data and Information Quality*. Available at: https://dl.acm.org/doi/10.1145/1515693.1515695 (Accessed: November 16, 2022).

Statistical Analysis

Overview

Chart, line chart

Description automatically generated

The scatter plots do not show much correlations between the variables, with the exception of some apparent correlation between ‘Quantity (tonnes)’ and Value (k Euro)’.

Both Quantity and Value variables appear distributions appear to be bimodal, with the major modes tending towards the higher ends of the respective scales. Both of the growth variables appear to have distributions close to normal. The possibility of normal distribution in the Growth variables is explored in detail below.

Distributions

Chart, histogram

Description automatically generated

Quantity Growth appears closer to normal, but the Value Growth distribution also displays a bell-shape.

Chart, scatter chart

Description automatically generated

Probability plots show that the points fall close to the line, indicating that both distributions are likely to be normal.

Finally, a Shapiro -Wilk test results are as follows:

Quantity Growth: ShapiroResult(statistic=0.9766903519630432, pvalue=0.682663083076477)

Value Growth: ShapiroResult(statistic=0.964688241481781, pvalue=0.34873512387275696)

With both p-values above 0.05, the null hypothesis that these data come from normal distributions cannot be rejected. The conclusion is that both distributions can be considered close to normal.

OUTCOME

Enough similarities between quantity and Value to choose either

Need more features to predict for each country