MSc in Data Analytics

Author: D. Smith

e-mail: [REDACTED]

Student ID: [REDACTED]

Table of Contents

Abstract

A use of Statistical Analysis and Machine Learning techniques to identify the impact the day of the week has on the habits of pedestrians; Localised to Capel Street, Dublin for the year 2015.

Dublin City Council has been measuring the footfall around its city's streets for years. Capel Street receives thousands of pedestrians and cyclists daily. I use data available to the public to analyse the trends of people visiting Capel Street throughout the entire year of 2015 using Statistical Analysis and Machine Learning algorithms.

This study concludes that Machine Learning algorithms can correctly identify Saturdays based purely on daily Footfall counts alone. Further research is needed to see if the other days of the week can have as much success.

Introduction

I was tasked with choosing a dataset that has been collated by the Dublin City Council in the domain of infrastructure and transport. I am then to analyse and apply machine learning techniques to this dataset.

The dataset I have chosen to work is the pedestrian counter for Capel Street found within "Pedestrian Footfall DDC". This dataset is described to have the results of pedestrian counters for Capel Street for the year 2015. This caught my eye as Capel Street has been in the news lately for its recent pedestrianisation (Kelly, 2022) and it has also been named as one of the coolest streets in the world (Burns, 2022). That is why I chose that one in particular.

Footfall is a measurement of how many people enter or leave a designated area within a particular timeframe. Hour, day, week, etc. There are a multitude of systems that can measure footfall but the Dublin City Council uses a PYRO-Box people counter. These devices count people that come within its range by detecting their body temperature. It can detect both pedestrians and cyclists (marketplace.intelligentcitieschallenge.eu, n.d.) It is a useful metric for county and city councils to obtain as the data collected can be used to help manage the flow of pedestrians and vehicles in the area.

My goal is to see if we can analyse trends from this data and to see what the Dublin City Council can learn from this. I want to know the impact the day of the week has on pedestrian habits and then to apply machine learning algorithms to see if accurate predictions can be made. If the Dublin City Council can know what days are the busiest, then any infrastructural repairs or maintenances can be planned and prepared for accurately.

Programming

All my work will be explored programmatically using the programming language Python in a Jupyter Notebook file. The notebook will be included with this report. The code will be commented and variable names will be named in a camel case format. Considerations will be taken to ensure that the variable names make sense to increase its readability.

Exploratory Data Analysis

Data Acquisition

The data was acquired from data.gov.ie. This website is a portal in which various Irish public sectors publish open and transparent data for everyone to use. The datasets provided here are free to use and redistribute. This makes it ideal for this report as the datasets are required to be included with the report as per instructions (Data.gov.ie, 2018).

Data Cleaning

The data was first checked for null values; There were none. The data contained 365 rows, one for each day of the year, as expected. There were 4 rows containing the date, and various results from the pedestrian counters. Results including how many enter and leave the area, and the total amount. The data was then was checked to see what types of data I was dealing with. Refer to Table 1 for a data dictionary of the data.

Column	Data Type	Description	Example
Date	Object	The date in a YYYY-MM-DD format	2015-08-16
Capel_Street	Integer	The combined total of the two columns "IN" and "OUT".	5757
IN	Integer	A count of the amount of footfall entering the area.	3162
OUT	Integer	A count of the amount of footfall leaving the area.	2595

Table 1 - Data Dictionary

I wanted to ensure that the footfall total was accurate for every entry in the dataset. By only checking the head or tail of it means that I could be dealing with inaccurate data. I wrote a short script in Python to check that the total values were the sums of the "IN" and "OUT" columns. The values were indeed accurate.

I felt that "Capel_Street" was not a good column title. It is not obvious what this value is supposed to be so I renamed it to "Footfall". Then I changed the IN and OUT columns to lowercase for consistency purposes.

The Date column being an object type also needed to be adjusted. I converted it to a datetime object. After this I added two additional columns to display the day of the week and the month. This task was made easier due to the previous step of converting the date column to a datetime object as I was able to use the functions that that class has access to (doc.python, 2022).

Data Visualisation

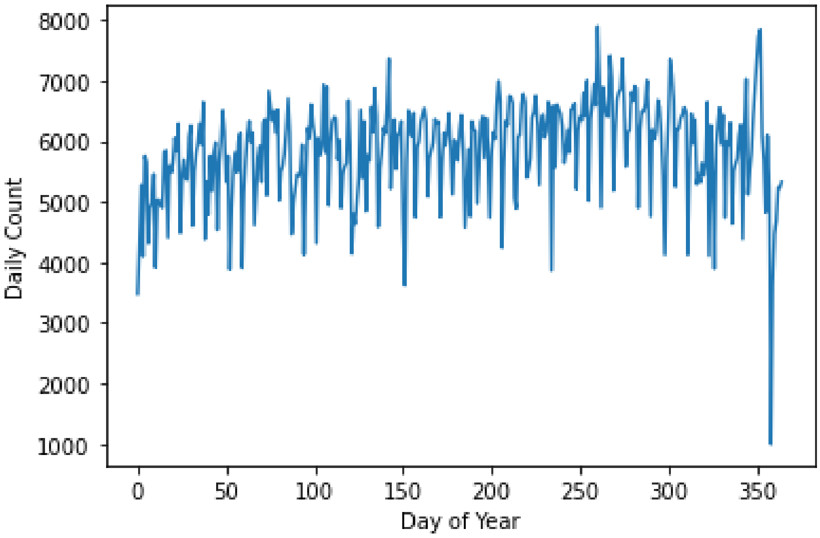


Figure 1 - Line Graph of Daily Counts throughout the year

Figure 1 shows a line graph of the footfall throughout the entire year. It looks erratic with loads of obvious peaks and dips. This implies to me that the day of the week has an impact on the footfall in the area. This needed to be explored further and I chose to do this with a scatterplot.

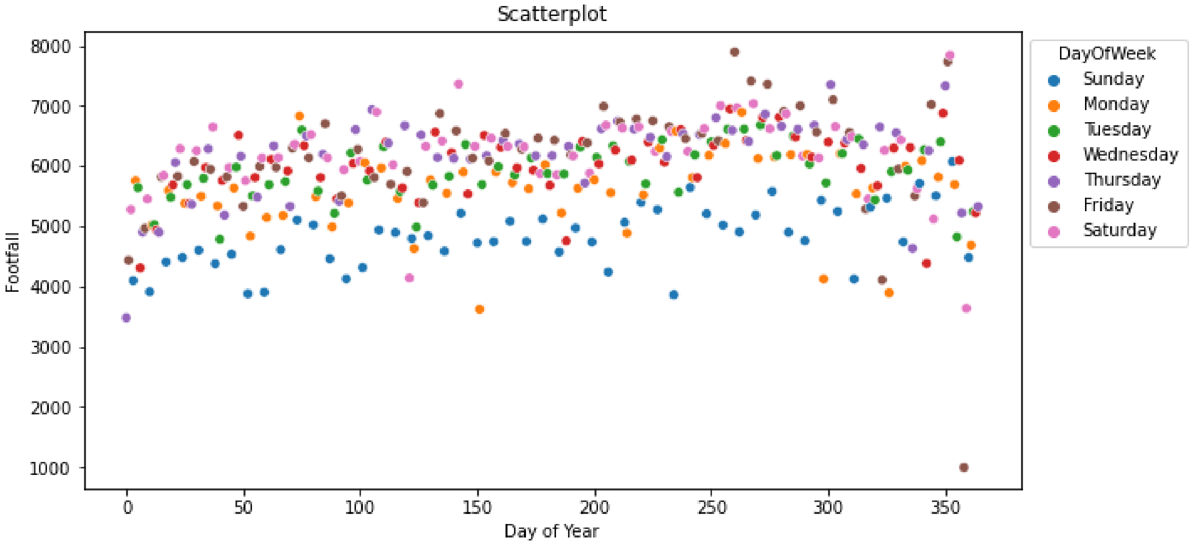


Figure 2 - Scatterplot of daily Footfall over the year

In Figure 2 we can see the impact the day has more clearly. The Sunday values represented by the blue dots clearly dominate the bottom of the graph. Sunday is therefore the least popular day to be travelling through Capel Street. Mondays are more in the centre but there are some that are a lot lower than expected. I expect that these are bank holidays. Saturday and Friday then appear to be the most popular day. One last thing to note from this is the obvious outlier towards the end of the year. Which I suspect is the 25th of December due to it being Christmas and most people spend time with their families that day. The vast majority of shops are closed that day so people would not have much reason to travel to Capel Street.

Below are a few bar charts to show the mean values of daily visitors to Capel Street throughout some selected months. Refer to the Jupter Notebook file attached for more.

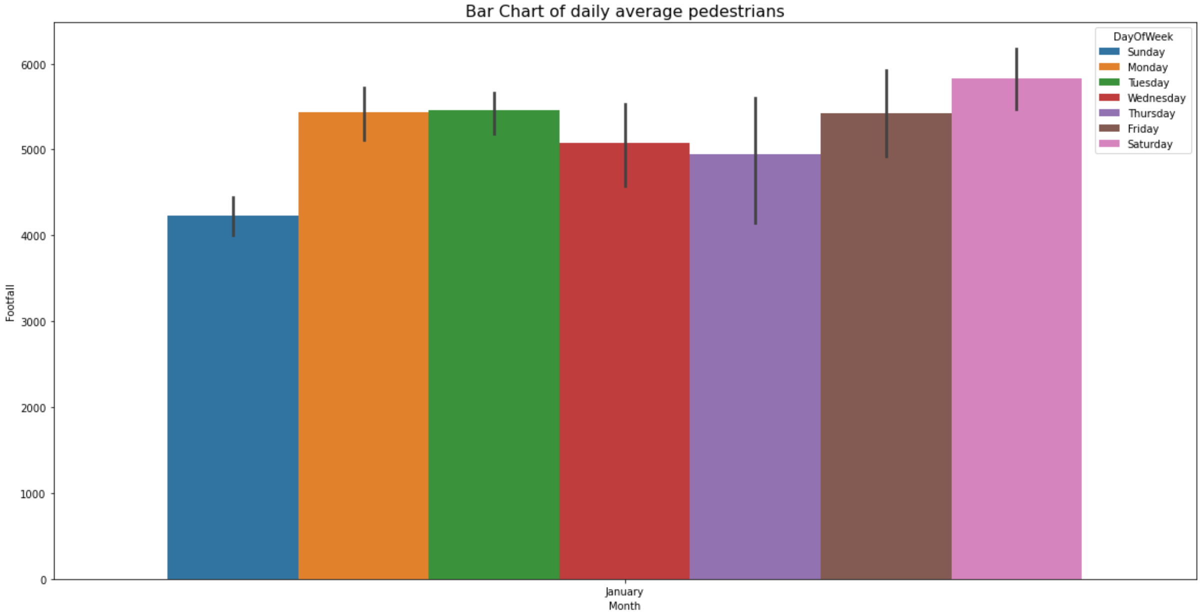


Figure 3 - Bar Chart of mean number of visitors for month January

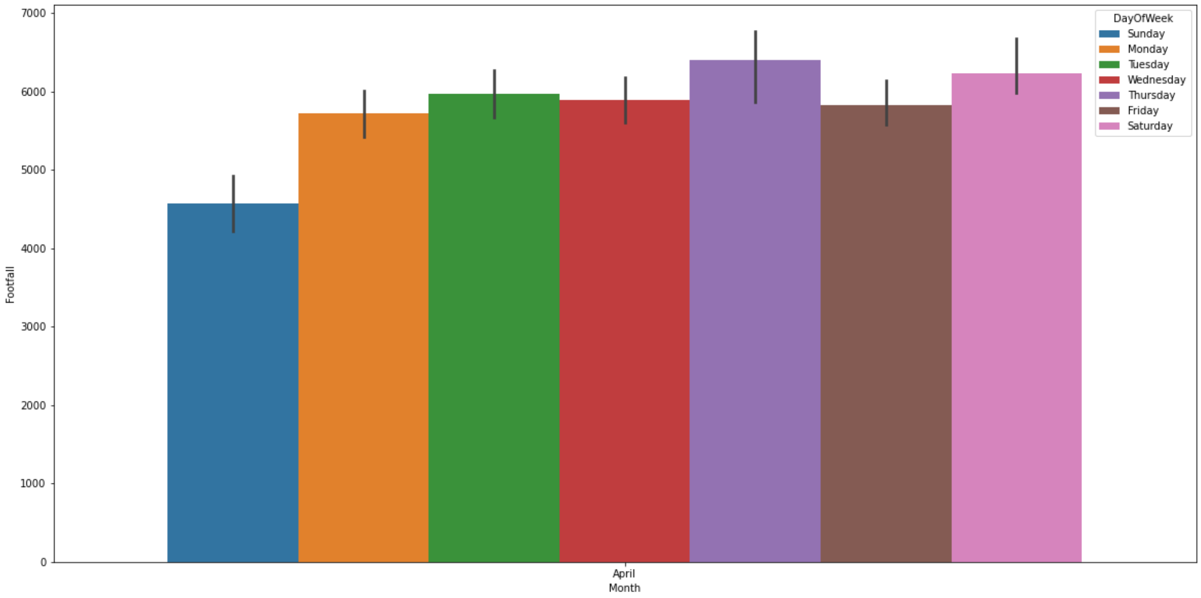


Figure 4 - Bar Chart of mean number of visitors for month April

If I was able to access to footfall counters for other years, I would able to do something with this in the Machine Learning Section. Unfortunately, those records are not yet available to the public.

Preparation for Machine Learning

Machine Learning is known to have issues when dealing with categorical data. Before sending my data through it, I opted to convert the Month and Day columns to integers. A method known as Integer Encoding.

Before	After
Monday	0
Tuesday	1
Wednesday	2
Thursday	3
Friday	4
Saturday	5
Sunday	6

Table 2 - Integer Encoding results for Day

Before	After
January	1
February	2
March	3
April	4
May	5
June	6
July	7
August	8
September	9
October	10
November	11
December	12

Table 3 - Integer Encoding results for Months

I also used a scaler to scale the data in the datasets to try get the best results possible. Scaling the data also helps to reduce the computational effort required to crunch the numbers. Which leads to less waiting to see the results.

Statistical Analysis

Data Spread and Central Tendency

Understanding the spread of data is important to a statistical analysis. Without this understanding, we cannot judge how accurately some calculations, such as the mean or median, represent the data. Extreme outliers have the potential to give a misleading understanding to these calculations.

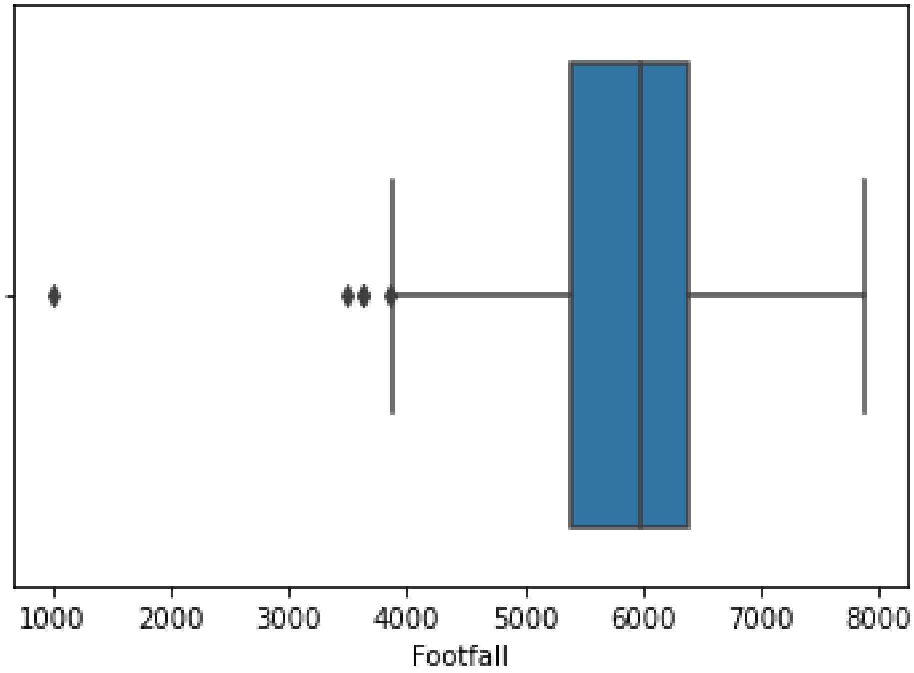
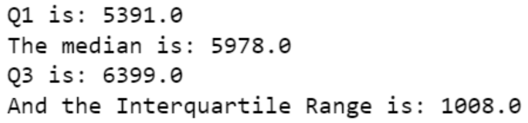


Figure 5 - Box and Whisker Plot



```
Q1 is: 5391.0
The median is: 5978.0
Q3 is: 6399.0
And the Interquartile Range is: 1008.0
```

Figure 6 - Code Output of Quartiles and Median

Figure 5 represents a box and whisker plot. Box and whisker plots are a visual representation of the spread and centres of the data. The data is split into quartiles and this data ranges from the smallest value to the largest value. They are useful visual representation of data because with just a glance you can quickly gleam from it the min value, first quartile, median, third quartile and max value. Outliers can also be seen (Meyers et al., 2009).

Fifty percent of our data is between the values of 5,391 and 6,399. Box and Whisker plots are also useful as we can make an educated guess as to how the histogram will look. I should expect to see that it is skewed to the left as the median is closer to the third quartile value rather than the first quartile. There are also the outliers that will give a left-skewed histogram that characteristic tail at the beginning of the graph.

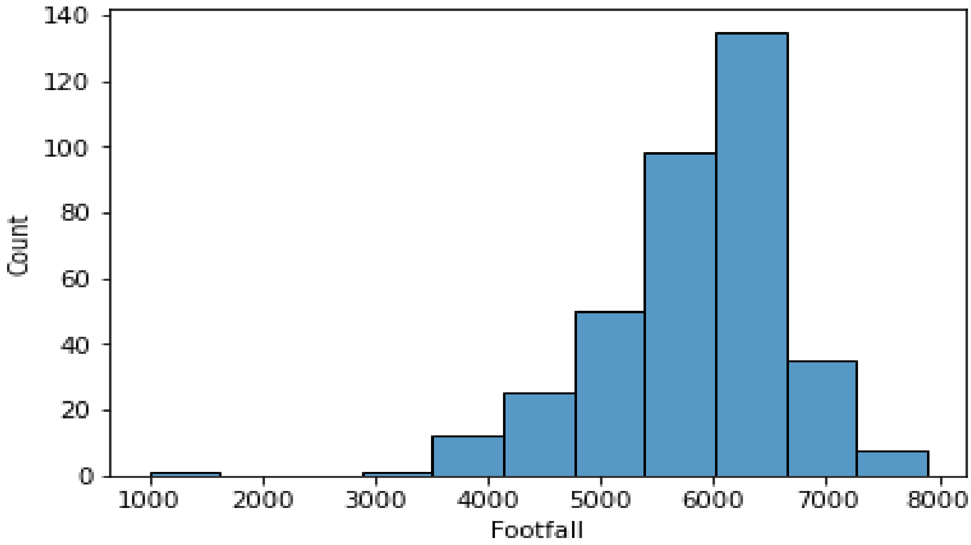
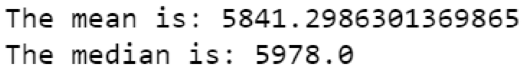


Figure 7 - Histogram of daily footfall throughout the year

Referring to Figure 7, the histogram was indeed skewed to the left. It is a unimodal histogram due to it only having one peak. Histograms are a good graphical representation of the centre of the data. From this I can tell that the data centres roughly between 6,000 to 6,600. Left-skewness, more often than not, indicates that the median is greater than the mean.



```
The mean is: 5841.2986301369865
The median is: 5978.0
```

Figure 8 - Code Output of Mean and Median

Which it is.

Poisson Distribution

A Poisson Distribution is a discrete probability distribution that can be used to predict the likelihood, k, of the number of events, n, occurring within a given interval of time. (Groenevelder, 2020) For mapping this out on a graph, I will take the mean value from earlier and use it as my mu value.

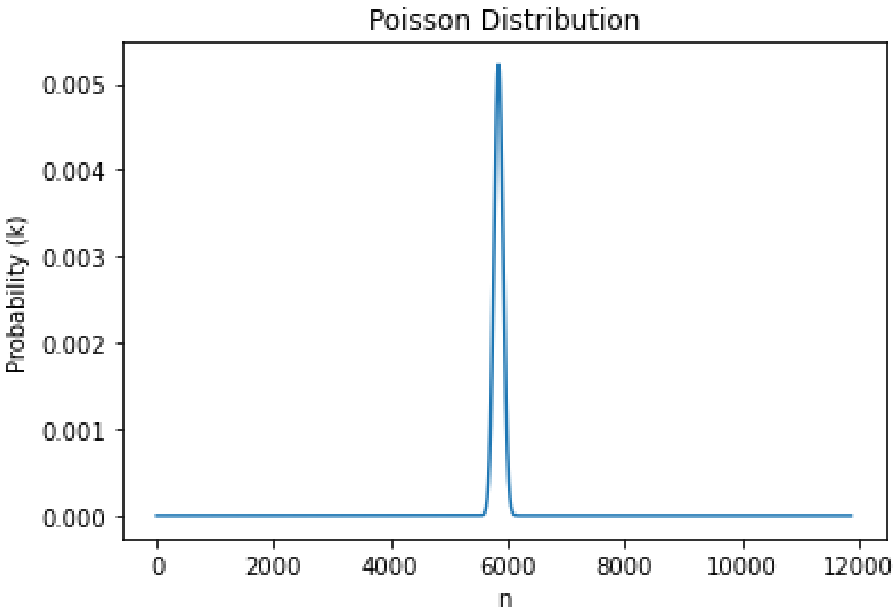


Figure 9 - Poisson Distribution (μ = ~5841.3)

Initially, I was not expecting this result. I thought the curve would be more like a bell-shape and the probabilities near the mean would higher than half a percent. Thinking about it, however, it makes sense. All the probabilities under the curve, when added together, must be equal to one. When the mu value and the number of events is so high, the probability of the event happening gets lower. Although the values towards the ends of the curve are practically zero, they are not zero. Just an incredibly small number. Its lack of a bell-shape can be explained by the fact that a Poisson Distribution uses the mu value as its variance. Unlike a Normal Distribution in which the mean and the variance are two different parameters.

Refer to the attached Jupyter Notebook file for some calculations I did using this graph to estimate probabilities based on input values for n.

Normal Distribution

A Normal Distribution is a continuous probability distribution. It follows a bell-shape and it is symmetric. It is centred around the mean and its width, or variance, is determined by the standard deviation (Weiss, 2015). The mean, median, and mode are all the same in a Normal Distribution. I took the mean and standard deviation from my dataset to plot out a normal distribution curve for these values. There are calculations for some probabilities in the attached Jupyter Notebook file.

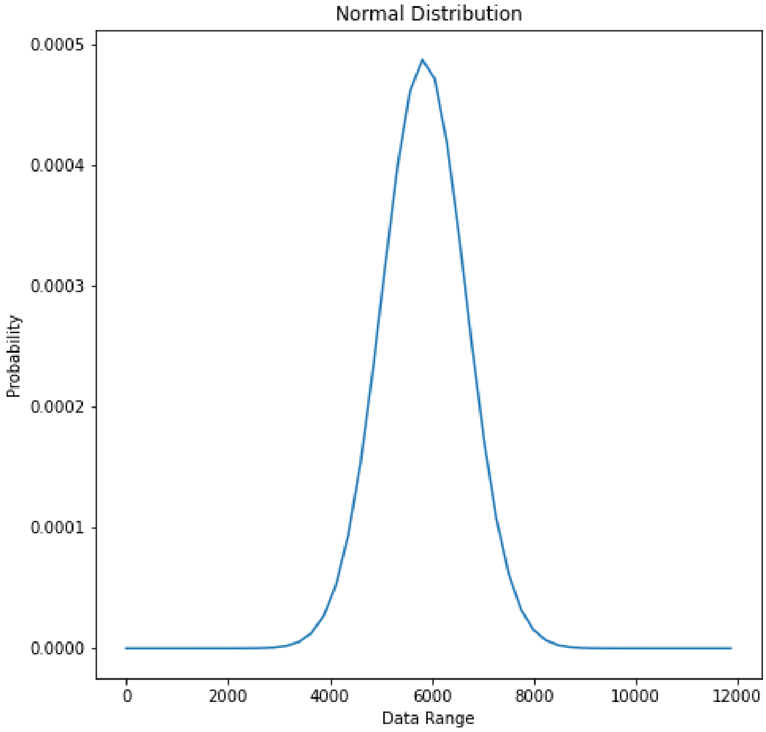


Figure 10 - Normal Distribution (μ = ~5,841.3; σ = ~817.6)

I noticed that the Normal and Poisson distributions shared similarities so I plotted them together to compare the results. As in, they were both looked symmetrical around the mean of the dataset. Poisson distributions are known to become more symmetrical, like Normal Distributions, as the mean gets higher. In some cases where the Poisson's mean is high, it approximates a Normal Distribution (Grace-Martin, 2018). Certainly, it cannot be used as an approximation in this case.

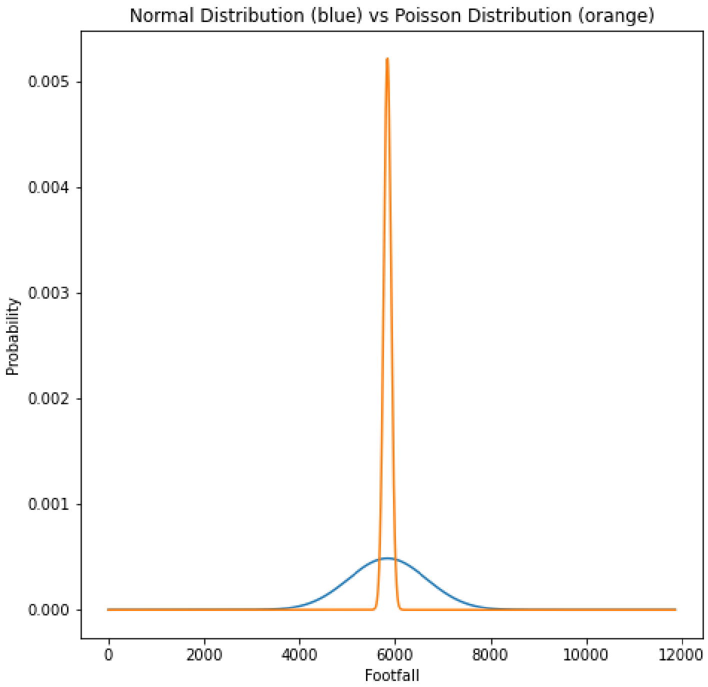


Figure 11 - Poisson against Normal

Machine Learning

Supervised or Unsupervised or Semi-supervised Learning?

I have chosen to undertake a supervised machine learning approach to this project as it makes the most sense in my opinion for what I am trying to accomplish with my project. I want to try take the daily footfall counts and see I can classify the day of the week with it. Supervised machine learning can be defined as the use of labelled datasets to train a machine learning algorithm to correctly classify data or accurately predict results (IBM, 2020). This definition fits exactly what I am trying to do.

Unsupervised Learning uses the algorithms to identify patterns in data that is not labelled or classified (Pratt, 2020). All my data was labelled so I did not take this path.

Semi-supervised Learning can be described as a learning problem to which a small number of data is labelled and the rest of the data is unlabelled (Brownlee, 2020). Then the model attempts to solve this problem by applying algorithms to identify the unlabelled data. All my data is labelled so I did not take this approach as it did not seem logical.

Methodology

For my project management framework, I had decided to follow the CRISP-DM (Cross Industry Standard Process for Data Mining) model. The other main models, KDD and SEMMA, would both be good for this project too. CRISP-DM is known to be applicable across multiple industries. I also chose this methodology as it has similarities to the software development methodology Agile Scrum. This is something that I have had experience with in the past so I felt it would be a perfect fit for me to adopt this approach.

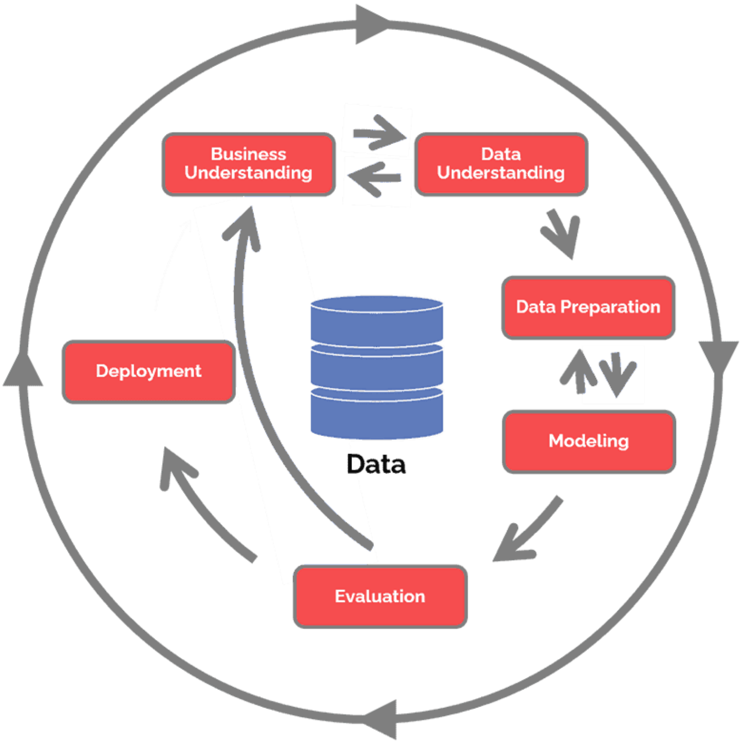


Figure 12 - CRISP-DM (Hotz, 2022)

I had also noticed that CRISP-DM has the option of going back-and-forth between the Data Preparation and Modelling stages. The other project management frameworks lacked this choice. I liked this as it allowed me to adjust the data with minimal effort and it did not force me into completing a full cycle each time I wanted to do so.

This is how I adopted this methodology to the project.

Business Understanding: This phase started with the formulation of my question that I wanted to ask. Which was if it was possible to use machine learning techniques to study the habits of pedestrians on a day-to-day basis.

Data Understanding: This is where I checked to see if the datasets that I needed to accomplish my goals were available for me to use. Due to being constrained to data that had been collected by the Dublin City Council, and in the domain of Transport and Infrastructure, it was very possible that my question was not answerable with what was there. As you can see, there is a link that goes to-and-fro this phase the previous one. If the datasets are not available, or they are too dirty, it can shape the question you want to answer. It is a cyclic process that I had to go through until I was satisfied.

This phase also included the processes of cleaning the data and gaining insight to what it contained. This is all covered in the earlier sections of this report. Namely the Exploratory Data Analysis and Statistical Analysis sections.

Data Preparation: Here is where the data is prepared and shaped to be suitable for the model. The data I used featured categorical data which is not easy for some machine learning algorithms to work with. I transformed all the categorical data by using the process of Integer Encoding. The data also contained columns that I did not need and these were removed and/or not included. The data I was left with was then scaled with a scaler to help improve the results of the model as much as possible.

Modelling: At this step the most suitable model is picked that will best help to try answer the question. Once the model is chosen, the data then needs to be split for testing, training, or validation purposes. The model is then built. It is evaluated for effectiveness which can mean comparing to other models to see what is best. The model I picked was SVM to help predict classification problems.

Evaluation: The results of my model had to be evaluated. I had to see if it could do what I wanted it to do or if there were any additional steps that I could take in order to improve the model. This is phase where you must decide if you must iterate the whole process again or to continue to deployment. This is where I adjusted the hyperparameters of the chosen model to help improve its accuracy scores.

Deployment: I am not deploying this model for real world use as it is an academic exercise. However, this is where I must write up this report and make a record of all my work. The final results of the work need to be written down. This phase of the process is also where a review and reflection take place. It is where I will ask myself what could I have done differently to improve the end results if things do not pan out well. Any lessons that I learned during the entire project, that could affect how I will tackle future projects, will be published here.

SVM

Support Vector Machines (SVM) are a supervised learning model that are commonly used in solving regression and classification problems. I chose SVM as the main model for my classification problem as SVMs are well-known to perform well with small datasets (Pasupa & Sunhem, 2016). I only have 365 sets of data; One for each day of the year 2015.

I split the data for testing and training purposes. I used a thirty percent split. The model was built and then used on my testing set. The results were not good. I was getting accuracy scores somewhere around eighteen to twenty-five percent. The next step I took was to try improve these scores.

GridSearchCV is a cross-validation technique that performs hyperparameter tuning in order to find the optimal parameters to use for a given model. I ran this technique and discovered what the best parameters to use were. With this done, I was now getting accuracy scores in the thirties. While this improved the accuracy score of my model, it is still not a great result. It is as good as I can get it though.

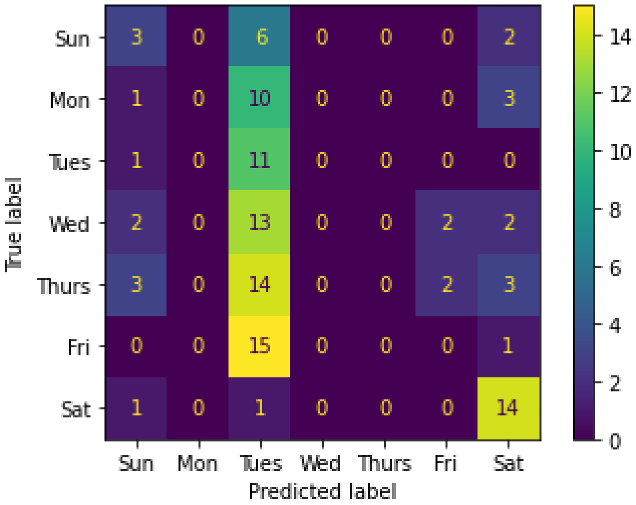


Figure 13 - Confusion Matrix of the model before Hyperparameter Tuning

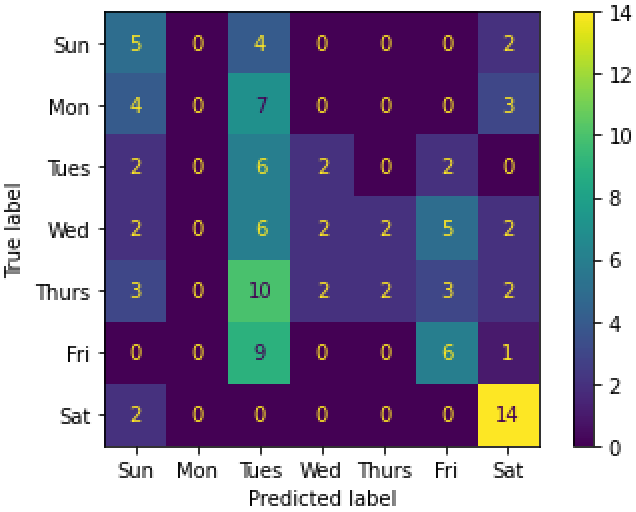


Figure 14 - Confusion Matrix of the model after Hyperparameter Tuning

The figures above are Confusion Matrices of the model before and after Hyperparameter Tuning. Confusion Matrices are used for conveying the True Positives, False Positives, True Negatives, and False Negatives that the model made. The diagonal squares from the top-left corner to the bottom-right corner represent the true positive values for those days. Meaning, the days that were correctly identified by the model. Although fewer Tuesdays were correctly identified, gains where possible were made on every other day; Monday excluded. It is clear to see that Saturday is an easy day for the model to get accurately classify and that Tuesdays and Sundays do okay. The model seems to dislike Mondays as much as people do.

Precision and Recall scores are found in the attached Jupyter Notebook file.

Decision Tree Classifier

Although I had read that SVMs are known to perform well with smaller datasets, I thought it was poor practice to just assume that it would out-perform other classification models without testing it first-hand. I then moved on to building a Decision Tree Classifier model so I could compare the results.

Referring to the Jupter Notebook, the results seemed poorer. Accuracy scores for the preliminary model was giving me results from the high-teens and sometimes crossing the twenty thresholds. I then started optimising the decision tree model by testing different parameters. The best results I could get was by setting the default criterion from "gini" to "entropy" and specifying a max depth of either 3 or 4. The preliminary model was certainly overfitting the model by not having a cap on depth. After all that, I was getting scores in the mid-twenties to low thirties.

While the SVM model did indeed perform better overall, the difference between them was not huge. Accuracy score was only about 4 percent in the difference.

Diagrams for the decision trees are found in the Jupyter Notebook. They were excluded from this report as I was not able to format it in a way that the reader could read it easily.

Also, to note: As the score is not great, which samples the model uses for testing/training will have an impact on the accuracy scores. Therefore, I reported rough ranges for the accuracy scores as if the reader runs my Jupyter notebook file themselves, they may see slightly different results.

I felt I had done all I could to improve the accuracy score. Hyperparameters were tuned; More data could not be added as it not was available; Two classification models were used; Data was scaled.

Presentation of Findings to Dublin City Council

Although the models failed to get excellent results for correctly identifying every day of the week, it was able to accurately classify Saturdays most of the time. What I can take from this is that extra consideration needs to be taken of this day to properly to manage the flow of pedestrians and cyclists through Capel Street. If the pedestrian counters for subsequent years were to be made available to the public, the machine learning models could be improved upon further. However, I acknowledge that there is a limit to this. Due to the outbreak of COVID-19, and the restrictions of public movement that the government imposed on the populace of the country, only the years of 2016 to 2019 would be useful. The data from 2020 to 2021 would be tainted by these restrictions. That is only speaking for this model though. The results from those years would still be interesting to see and could be used to derive other models. For example, the impact COVID-19 has had on footfall to the area. Have the footfall numbers returned yet from before pandemic level would be another basis for a study.

Reflection

I am not surprised by the results that the machine model did not perform greatly. I was working with a such a small dataset and of course that the day of the week is not the only factor. I have shown that it is *a* factor but not *the* factor. I would guess that the weather has a sizeable impact. Who wants to go out on foot when it is raining? What I can say is the Dublin City Council needs to ensure that extra consideration is taken on Saturdays for managing the flow of pedestrians and cyclists. That is the day when the numbers are highest and the day that the machine learning models were able to predict the most accurately.

If I was to do this project from the start, instead of trying to identify every day of the week I would split it into weekdays and weekends. Those results should be way more accurate. However, as part of my classes I am used to dealing with machine learning outcomes where the accuracy score was high. Gaining experience in a situation where things are not working out is a good thing.

Overall, it was an interesting academic exercise. I have never had something like this to do in my undergrad; An assignment where all the modules were combined into one project. It was neat to see during the work how the modules flowed one after another.

References

Brownlee, J. (2020) *What is semi-supervised learning*, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/what-is-semi-supervised-learning/ (Accessed: November 12, 2024).

Burns, S. (2022) *Dublin's Capel Street named among coolest in the world*, *The Irish Times*. The Irish Times. Available at: https://www.irishtimes.com/ireland/dublin/2022/08/29/dublins-capel-street-named-among-coolest-in-the-world/ (Accessed: November 11, 2024).

Data.gov.ie (no date) *Data.Gov.IE*. Available at: https://data.gov.ie/ (Accessed: November 3, 2024).

Datetime - basic date and time types¶ (no date) *datetime - Basic date and time types - Python 3.11.0 documentation*. Available at: https://docs.python.org/3/library/datetime.html (Accessed: November 9, 2024).

Grace-Martin, K. (2018) *Differences between the normal and Poisson distributions*, *The Analysis Factor*. Available at: https://www.theanalysisfactor.com/differences-between-normal-and-poisson-distributions/#:~:text=A%20Poisson%20distribution%20with%20a%20high%20enough%20mean%20approximates%20a,you%20nothing%20about%20the%20other. (Accessed: November 14, 2024).

Groenevelder, R. (2020) "Poisson Distribution" in *An Introduction to Probability and Statistics Using Basic.* CRC Press, pp. 90-92.

Hotz, N. (2022) *What is CRISP DM?*, *Data Science Process Alliance*. Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed: November 9, 2024).

IBM Cloud Education (2020) *What is supervised learning?*, *IBM*. Available at: https://www.ibm.com/cloud/learn/supervised-learning#:~:text=Supervised%20learning%2C%20also%20known%20as,data%20or%20predict%20outcomes%20accurately. (Accessed: November 10, 2024).

Kelly, O. (2022) *Capel Street becomes Dublin's longest traffic-free street*, *The Irish Times*. The Irish Times. Available at: https://www.irishtimes.com/ireland/dublin/2022/05/20/capel-street-becomes-dublins-longest-traffic-free-street-1.4884011/ (Accessed: November 11, 2024).

marketplace.intelligentcitieschallenge.eu. (n.d.). *Smart Mobile Pedestrian Counters for Events*. [online] Available at: https://marketplace.intelligentcitieschallenge.eu/en/solutions/smart-mobile-pedestrian-counters-for-events (Accessed 22 Oct. 2024).

Meyers, L.S., Gamst, G. and Guarino, A.J. (2009) "3A Data Screening," in *Applied Multivariate Research: Design and interpretation*. Thousand Oaks: Sage, pp. 53–54.

Pasupa, K. and Sunhem, W. (2016) “A comparison between shallow and deep architecture classifiers on small dataset,” *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)* [Preprint]. Available at: https://doi.org/10.1109/iciteed.2016.7863293.

Pratt, M.K. (2020) *What is unsupervised learning?*, *SearchEnterpriseAI*. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/unsupervised-learning#:~:text=Unsupervised%20learning%20refers%20to%20the,are%20neither%20classified%20nor%20labeled. (Accessed: November 10, 2024).

Weiss, N.A. (2015) “6. The Normal Distribution,” in *Introductory statistics*. Harlow: Pearson, p. 284.