Final Year Project

# PREDICTING FIRE BRIGADE CALL OUTS FOR DUBLIN CITY

Matt Grogan

Student ID: 16312653

A thesis submitted in part fulfilment of the degree of

**BSc. (Hons.) in Computer Science with Data Science**

Supervisor: Dr. Colm Ryan

GitHub Link: https://github.com/mattGrogan1/Final_Year_Project


UCD School of Computer Science

University College Dublin

May 11, 2020.

# Contents

# List of Graphs



# List of Tables



***Important***

The Literature Review stage sections (1-7) have been edited from how they appeared in December 2019. These sections have been rewritten in the past tense as all work was completed as of May 2020. Sections 3 and 4 have been supplemented with additional graphs. Section 5 has greater written detail in relation to the Fire Brigade, Ambulance, Weather and Event files, as well as, the inclusion of additional multiple tables. Section 6 provides an updated regression model implementation. Here, the linear regression model was discarded in favour of the Poisson regression model. Section 7 provides an updated Gantt Chart which takes into account the revised project deadlines due to the Covid-19 global pandemic.

# Section 1

# Abstract

It is vital that emergency Fire Brigade and Ambulance services in major cities are able to manage the number of call outs they are receiving throughout a year. It is essential that they are able to predict the number of call outs they will receive so that they can plan for the future and have sufficient resources in their busiest stations, so to keep up with the demand of emergency situations. It is critical that the men and women who provide this service work at a high rate as they are responsible for saving the lives of the city's occupants. The demand for Fire Brigades and Ambulances varies throughout the year, depending on the month , week, day, and even, time of the day. Although there have been many research papers on determining Ambulance service call outs, there are little to no analysis on predicting Fire Brigade emergency call out incidents. In 2016, Dublin City Council released to the public, the emergency Fire Brigade and Ambulance service call out details from the year 2011 to the year 2015, a timespan of five years. For this project, features will be identified that affect the number of call outs of the Dublin City Fire Brigade and Ambulance service receive, i.e. is the location in the city a major factor or does the volume of call outs differ, as a result of the weather conditions. A regression model will then be created that can be used to predict the number of Fire Brigade and Ambulance call outs for a 999 call on a certain date.

# Section 2

# Project Specification

The overall goal of this project is to build a regression model that can be used to predict the number of Fire Brigade call outs in Dublin on a certain date within a given timeframe.

### 2.1 Core Goals

The core goals require that I carry out the following tasks:

- Analyse and clean the raw data made available by Dublin City Council, paying close attention to any changes in underlying data that occur over the time period.
- Perform statistical analyses to identify features that influence fire brigade call outs, e.g. temporal features, such as the time of the day, the day of the week, the month of the year, and location, Rathfarnham vs. North Strand
- Develop a reasonable baseline method for predicting fire service call outs.
- Using simple temporal features, develop improved prediction methods.
- Investigate whether incorporating additional features, such as, aspects of weather and events taking place in Dublin, can improve the prediction of fire brigade call outs.

### 2.2 Advanced Goals

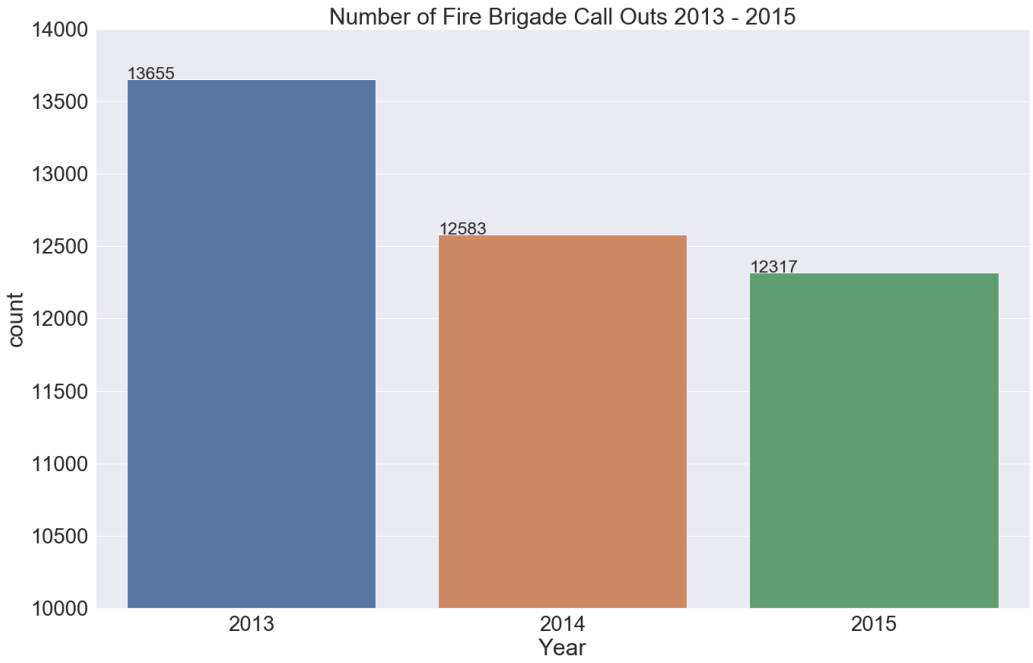
In addition to the core goals, there are advanced goals I must also familiarise myself with. The advanced goals require that I undertake the following tasks:

- Apply the same pipeline to predicting specific types of fire brigade call outs, e.g. domestic fires.
- Apply the same pipeline to predicting fire brigade call outs at a more local level, such as, fire brigade call outs to Tallaght.
- Apply the same pipeline to predicting ambulance call outs and determine if the same features are predictive of both fire brigade and ambulance requirements.

# Section 3

# Introduction

### 3.1 Background

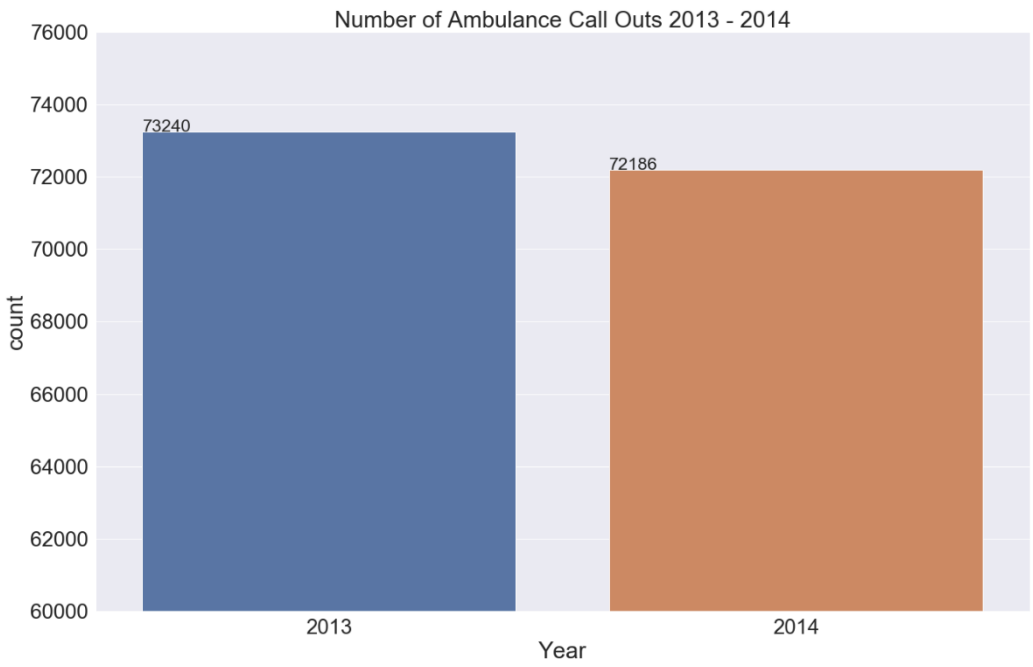



*Graph 1: Number of Fire Brigade Call Outs per Year 2013 - 2015.*

Dublin City's Fire Brigade and Ambulance services respond to emergency calls involving fires, as well as crises such as, flooding and road accidents. The service covers Dublin City Centre, North County Dublin, South County Dublin, and Dun Laoghaire/Rathdown (Dublincity.ie, 2019). The Dublin Fire Brigade and Ambulance service have a responsibility to the 1.3 million inhabitants of Dublin to provide assistance that is reliable and fast. The number of call outs a Fire Brigade or

Ambulance receives is especially important. If a Fire Brigade or Ambulance crew is occupied by many emergencies at once, this is urgent and could factor into whether an individual survives or loses their life. Predicting the number of call outs would aid the Dublin Fire Brigades and Ambulances to manage their resources more efficiently. They would be able to plan ahead for the future, determine which stations are busiest and assign their resources accordingly to the busiest stations. Therefore, it is essential that the Dublin Fire Brigade and Ambulance services are able to predict the number of call outs they are going to receive.

Between 2013 and 2015, there were 38,555 emergency 999 calls requesting the Dublin Fire Brigade and between 2013 and 2014, there were 145426 calls to the Dublin Ambulance service. Of the total Fire Brigade calls, 13655 were made in 2013, 12,583 were made in 2014, with the remaining 12,317 made in 2015. This represents a 7.85% decrease between 2013 and 2014 and a 2.11% decrease between 2014 and 2015. Overall, emergency calls requiring the assistance of Dublin Fire Brigade have declined roughly 9.8% over the three-year period. Similarly, there was a decrease from 73240 Ambulance call outs in 2013, to 72186 call outs in 2014. This is a slow and steady decrease. This decline could be attributed to many factors. Nowadays, the public are much more aware of preventing fires and know how to react in the event of an emergency. Also, items are created nowadays with the intention of being fireproof, as opposed to many years ago when safety was not a feature that was important. This would result in a decline in emergency calls requiring Fire Brigade and Ambulance assistance and could explain the year on year decrease present across all years.




*Graph 2: Number of Ambulance Call Outs per Year, 2013 - 2014.*

Since the turn of the twenty-first century, climate change has had a major effect on weather systems. In Ireland, this has resulted in abnormal weather patterns. Within the last number of years, the country has experienced inclement weather conditions, particularly in the Winter months. In particular, the winter of 2013 had storms every

one to three days (National Directorate, 2014). During the Summer, the country has entertained warm, Mediterranean clement weather. These severe events can cause huge impacts on a society. These extreme weather changes can affect built up areas, such as Dublin City, and can cause great challenge to the men and women of the Dublin Fire Brigade and Ambulance services. Weather may affect the estimation of call outs and it is essential that the Dublin Fire Brigade and Ambulance are aware of this. Different weather patterns will have a different effect on the number of call outs the services will receive.

Dublin is the capital city of Ireland and as a result, major events and festivals are going to occur. There are certain occasions that happen in Dublin annually that can have an impact on the Fire Brigade service. Every year, Dublin city allows many Soccer, GAA and Rugby sporting events to occur in Croke Park and The Aviva Stadium. These events close many streets within the city centre throughout the day and night. This combined with the increased presence of people in the capital attending the sporting events, can impact the service provided by Dublin Fire Brigade. Another major event that is held annually in the city and surrounding areas, is the KBC Dublin Marathon. This major athletic event is held on the last Sunday of October. It attracts thousands of runners to the capital. In the years 2013, 2014 and 2015, 12,317, 12,087, 12,920 runners participated in the 42-kilometre run (Marathonguide.com, 2013; 2014; 2015). When this racing event is occurring, the roads around the city centre and suburbs are closed. As a result, Fire Brigades and Ambulances may struggle to respond to call outs due to the increase in traffic along the open roads.

### 3.2 Outline of Report

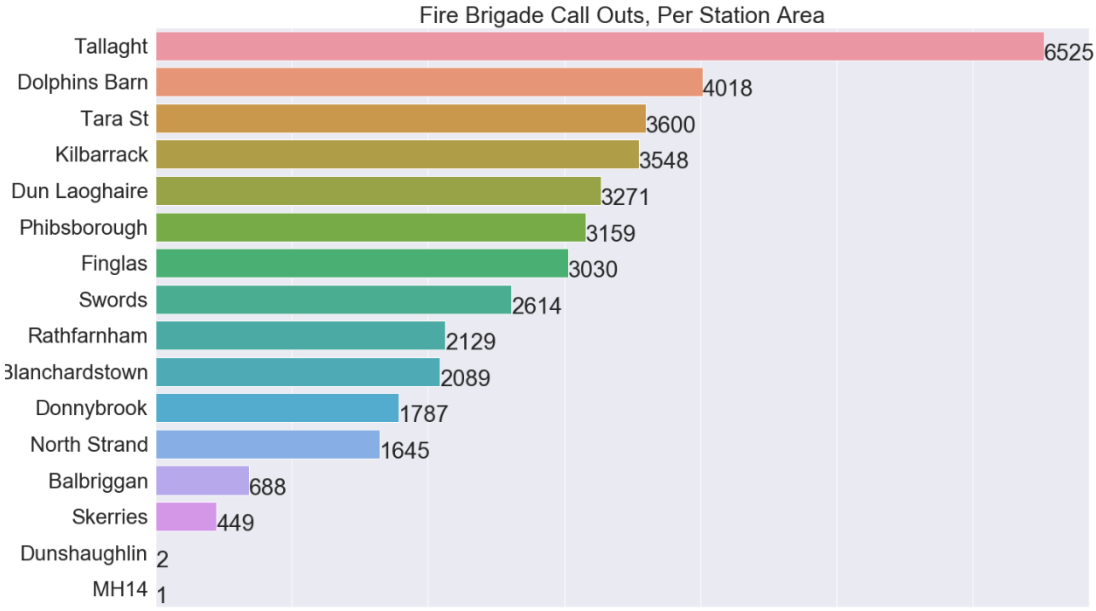
This report is organised as follows. Section 4 discusses related ideas and works that researchers have written on to predict Fire Brigade and Ambulance call outs and whether temporal features and aspects of weather affect the service. Section 5 gives an overview of the data to be used in this study and any considerations that were taken when preparing this report are discussed here. Section 6 provides an overview of the approach that will be used to analyse the dataset and discusses the regression models that will be implemented to predict Fire Brigade and Ambulance call outs. Section 7 outlines the future work that will carried out as part of this project. A Gantt chart will be provided here for clarification. Section 8 provides an explanation of the implementation of the project. Section 9 evaluates the regression models produced using a variation of evaluation techniques. Concluding remarks will be outlined in section 10, while section 11 will provide the bibliography.

# Section 4

# Related Works and Ideas

### 4.1 Previous Studies

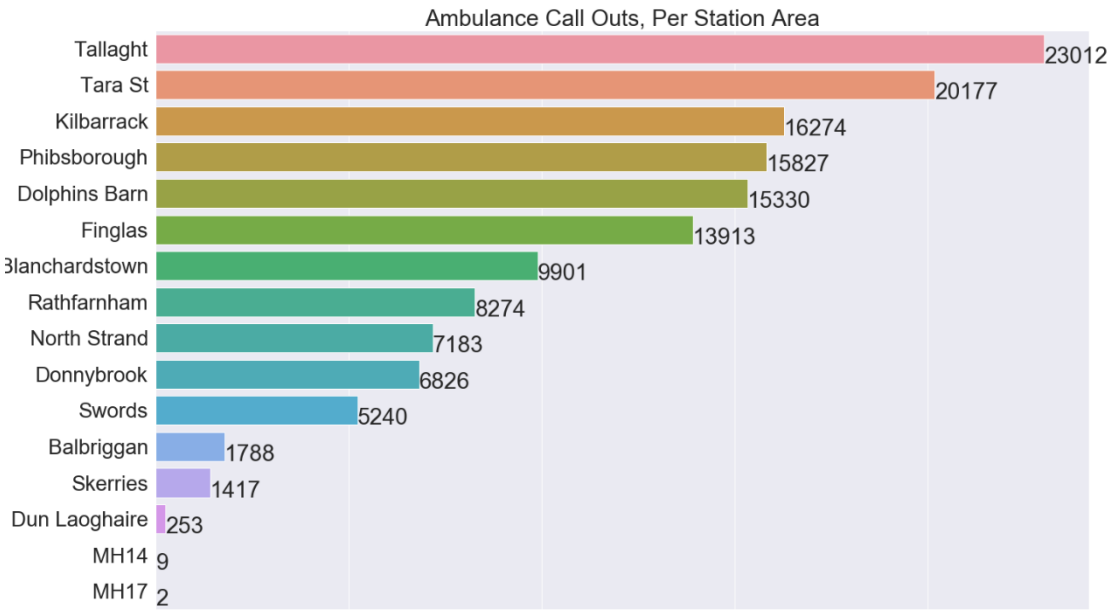
Predicting fire brigade call outs is a relatively unresearched area in that there are little to no previous studies. For reasons unknown, researchers have chosen to focus more so on a city's ambulance service, as opposed to their fire brigade service. Of those that do focus on Fire Brigade call outs, the main area of research is dedicated to analysing the fires themselves and the reasons as to why fires occur. Investigations of this type have been the focus of researchers for quite some time. Duncanson, Woodward and Reid (2000) studied fires in New Zealand and discovered that fires were more prevalent in socio-economically deprived areas. If this study was to be related to Dublin, it could be noted that fires, and as a result, fire brigade call outs, may be higher in areas of Dublin that qualify as disadvantaged. This would mean that call outs in stations in North County Dublin and West County Dublin would have more call outs than the South County Dublin and Dun Laoghaire/Rathdown as these are areas of high affluence and wealth, according to the Dublin City Socio-Economic Profile (2011). This idea by Duncanson, Woodward and Reid is proven to be true (Figure 2). According to the graph below, four out of the five areas with the highest number of call outs in Dublin are located in North County Dublin, West Country Dublin, and the City Centre. These areas are Tallaght, Dolphin's Barn, Tara Street and Kilbarrack, respectively.




*Graph 3: Fire Brigade Call Outs, Per Station Area.*

While there is little research into estimating predictions in relation to a fire brigade, studies involving the forecasting of call centre calls is a much more researched area of focus that has been given a considerable amount of time. Call centres are very similar to fire brigades with respect to the fact that, both are active 24 hours a day, and both involve calls that relate to the general public. Therefore, their target market is essentially the same. Bianchi, Jarrett and Choudary Hanumara (1998) employed Box-Jenkins ARIMA modelling to estimate and analyse incoming calls to call centres in 117 area codes in the United States at a given time. Whereas, Tych et al. (2002) carried out analysis on hourly telephone calls to call centres in Manchester and the Northeast of England using a Dynamic Harmonic Regression model. Furthermore, this model was tested against an ARIMA model and provided much better results. ARIMA, which is an acronym for Auto Regressive Integrated Moving Average, is a group of models that is able to explain a time series according to past

values. Similar methodology, that also provides a good estimate of time series, is that of harmonic regression. Harmonic Regression is a form of regression model where the predictor variable are modelled as trigonometric functions. Pedregal and Trapero (2006) employed the use of a dynamic harmonic regression model to predict the hourly price time series of electricity markets.




*Graph 4: Ambulance Call Outs, Per Station Area.*

Predicting demand on a variety of subjects is a topic that contains a rich collection of research papers. As early as the 1980's, regression models were created to estimate demand in many areas. An early study carried out by Kamenetzky, Shuman and Wolfe (1982) tested demand models by suggesting that variables represented by population and employment were of most importance. Bowdish et al. (1992) used regression analysis to predict the expected numbers on Race Day at the Indianapolis Motor Speedway competition. However, the researchers found the model used was too basic to predict the demand. Catalina, Virgone and Blanco (2008) tested many regression models and discovered that polynomial regression was an efficient forecasting tool for their problem. Analysis of previous papers prove that it is sufficient to suggest that a regression model once implemented, will successfully estimate the number of fire brigade call outs in Dublin on a given data within a given time frame.

It is widely assumed by researchers that the volume of an area being studied and focusing on a pre-determined time period follows a Poisson distribution or process. This claim is backed by Henderson (2005) where he states that the Palm-Khintchine theorem highlights that in a scenario, such as a call centre, where there are a high number of possible callers, when combined these prospective callers are grouped together, with each having a very small probability of calling at a particular moment in time. Likewise, Zhu and McKnew (1992) proposed that the variation of demand during six four hour durations for the Emergency Ambulance Services Centre in Shanghai can be completely shown by means of a Poisson Distribution. In contrast to these studies, Brown et al. (2005) suggests that instead of following a constant for the Poisson Distribution, an inhomogeneous Poisson Process should be used in terms of modelling. This notion of a varying Poisson Distribution is further clarified in a

further study by Brown, Weinberg and Stroud (2006). The researchers use a Bayesian Markov Chain Monte Carlo technique to fit this idea of an inhomogeneous Poisson Process to forecast the future arrivals at a North American private bank. They are able to point predict arrivals at the call centre, as the approach followed ensures counts are easily accessibly obtained. This approximation is suitable for high rates of demand.

Many researchers have carried out statistical analysis in order to derive important information from datasets. As there are little studies into predicting fire brigade call outs, most of the data used as part of this analysis relates to Emergency Medical Services. Budge et al (2010) forecast EMS travel time distributions for Calgary, according to the dependence on distance using a nonparametric regression model while Matteson et al. (2010) predict EMS arrival times per hour in Toronto using time series models containing a dynamic latent factor structure. This structure was a very important aspect, as the volume of calls expected per hour is certain to be low. This feature contrasts the previously mentioned study found above, by Brown, Weinberg and Stroud (2006). Earlier research into modelling EMS calls in Calgary, Alberta, Canada, carried out by Channouf et al. (2006) developed time series models to predict daily and hourly call demand. Here, the authors created autoregressive and ARIMA models to compare demand accuracy. Aladdini (2010) estimated EMS response times using linear regression models using the standard deviation as the function of the average response time.

Rare research relating to Fire Brigades and the implementation of prediction was carried out in early studies by Kolesar in which he developed models involving data collected from the New York City Fire Department, which is the largest fire department in the world. Kolesar (1975) proposed a model using parameters, to estimate the expected fire brigade travel time, according to the area of the fire. This model provided good estimates for regions of New York City, such as, the Bronx. Further studies by Kolesar, Walker and Hausner (1975) presented a model that derived that there was no significant effect on the demand for the New York City Fire Brigade depending on the time of the day.

### 4.2 Effect of Weather Systems

In recent years, climate change has altered weather systems around the world, affecting areas from Australia, to Canada, and the United Kingdom. There is no research on how weather effects the estimation of the number of fire brigade call outs. Instead, studies that have investigated weather effects focus on Emergency Medical Services, such as ambulances. Change of weather has proven to have a drastic effect on the demand for Emergency Medical Services, and in turn, an effect on the demand for assistance required from a fire brigade. Cold weather, for example, extreme flooding, snow and ice, such as that experienced in Ireland during the Winter of 2010, can result in demand for EMS increasing. Studies carried out in London by Thornes (2013) found that when air temperature dropped by 2 degrees Celsius the demand for EMS increased. While no study exists relating to the Fire Brigade service, it can be assumed that demand would also increase. While there are few studies focusing on the impact of cold weather, there is many that investigate the effects of warm weather. Research on the effect of extreme heat on the ambulance service in Toronto by Dolney and Sheridan (2006) highlighted that between 1999 and 2002, the demand for ambulance calls increased by 10% on days that were considered oppressively hot

while studies by Nitschke et al. (2011) in Adelaide, Australia discovered that the number of call outs increased by 10% during the heatwave of 2008, and by 16% during the heatwave of 2009. Comparing these studies in London, Toronto and Adelaide to the city of Dublin, the demand for EMS is certain to increase during extreme warm and cold weather systems.

# Section 5

# Data Considerations

The data for this paper was obtained from Data.gov.ie, © 2015 licensed under CC BY 4.0. It was located by searching Fire Brigade and Ambulance Call Outs in the search bar and navigating to the relevant page.. This website was created to make data held by public bodies in Ireland, freely available and easily accessible online. For purposes of conducting research into predicting the number of Fire Brigade and Ambulance call outs on a given date at a certain time, it is essential that the dataset used is static, that is, that it is no longer being updated. In fact, this dataset is static as it is composed of historical data. Dublin City Council released data ranging from the years 2011 to 2015, inclusive. It contained information relating to call outs responded to by the Ambulance and Fire Brigade services active in Dublin County. The data can be downloaded onto a local computer and stored in the desired location. The data contained within the Fire Brigade and Ambulance Call Outs section contains five data resources:

1. A csv file relating to Fire Brigade and Ambulance Call Outs in 2011.
2. A csv file relating to Fire Brigade and Ambulance Call Outs in 2012.
3. A csv file relating to Fire Brigade Call Outs between 2013 and 2015.
4. A csv file relating to Ambulance Call Outs between 2013 and 2015.
5. A xlsx file containing descriptions of the headings within each csv file.

### 5.1 Dataset Discrepancies

Initial research aimed to include all datasets containing information to do with the Dublin Fire Brigade and Ambulance call outs between the years 2011 and 2015. However, upon analysis of all the datasets, this proved to be difficult as there are some noticeable differences. It was therefore determined that the older data from 2011 and 2012 would not enhance the regression models to be implemented.

Firstly, while there is a Description column in the 2013 – 2015 data, as well as in the 2012 data, there is none in the 2011 dataset. This would make analysis difficult as this column describes the type of call outs each Fire Brigade responded to. Another noticeable difference between the 2011 dataset and the others is that, for the Station Area column, only the Dublin area code was listed, whereas for the 2013 – 2015 and 2012 data, the region was listed. Therefore, the 2011 data was excluded to ensure maximum accuracy. In addition, the Description column was different between the remaining two datasets. In the 2013 – 2015 data, Description described the type of

accident that the service were called to in plain English, whereas, the 2012 data used the NATO phonetic alphabet to distinguish between accidents. To ensure that maximum accuracy is maintained, the 2012 csv file was excluded. Therefore, for purposes of this research, the dataset containing information spanning the years 2013 to 2015 was used.

### 5.2 Dataset Description

The datasets relating to the number of Fire Brigade and Ambulance call outs between 2013 and 2015 contained 38,555 rows and 220828 rows, respectively. There were seven fields within the Fire Brigade dataset and eight fields within the Ambulance dataset. Table 1 represents the meaning behind each of these fields.

*Table 1: Description of Columns within Fire Brigade and Ambulance Datasets.*

| COLUMN | MEANING |
|---|---|
| DATE | The date that the fire occurred. |
| STATION_AREA | The location of the fire brigade that responded to the fire. |
| DESCRIPTION | The description of the type of fire. |
| MONTH | The month the fire occurred. |
| DAY_OF_WEEK | The day of the week the fire occurred. |
| DAY | The day of the month the fire occurred. |
| HOUR | The hour of the day the fire occurred. |
| CLINICAL_STATUS | The type of Ambulance call out (Unique to Ambulance Dataset). |

This data was organised as such, so that it is recognised at structured data. However, there was an aspect of data cleaning that was completed. As many columns were decided to not be useful in predicting Fire Brigade and Ambulance service call outs, they were removed from the dataset. Table 2 describes the columns that were removed from the datasets.

*Table 2: Description of Columns Removed from Fire Brigade and Ambulance Datasets.*

| COLUMN | MEANING |
|---|---|
| ORD | The time the first service arrived. |
| MOB | The time the first service was on route to the scene. |
| IA | The time the first service arrived at the scene. |
| LS | The time the first service left the scene. |
| AH | The time the first service arrives at the hospital. |
| MAV | The time the last service is mobile and available. |
| CD | The time the incident was closed. |

Common practice to manage missing values is to either remove them, use the values either above or below the missing value to assign to the blank space, or fill them in with the average of the relevant column. For purposes of this study, the missing values were removed from the dataset to ensure maximum accuracy when

predicting the number of call outs as the dataset only would only contain rows that contain useful, reliable information.

### 5.3 Additional Feature Dataset Descriptions

The weather dataset was used to incorporate weather features into the regression models was also structured. The weather files were accessed from VisualCrossing.com and located by entering the required location as Dublin. CSV files were downloaded containing nineteen columns. Table 3 shows the five columns that were chosen to be used as part of this research.

*Table 3: Description of Included Weather Feature Columns.*

| COLUMN | MEANING |
|---|---|
| DATE | The date that the weather feature occurred. |
| TEMPERATURE | The mean temperature on a date. |
| WIND_SPEED | The speed of the wind experienced on a date. |
| PRECIPITATION | The level of rainfall that occurred on a date. |
| CLOUD_COVER | The level of visibility experienced on a date. |

In addition to these five columns, there were many other columns that I chose not to include. The decision to not include many of these columns was due to my belief that they would not affect the volume of Fire Brigade and Ambulance call outs in an as strong manner as the above five features would, for example, I strongly believed the precipitation on each date throughout the three years would have a much greater impact on call outs, as opposed to the chill of the wind on a given date. Table 4 shows the columns that were discarded as part of this research.

*Table 4: Description of Discarded Weather Feature Columns.*

| COLUMN | MEANING |
|---|---|
| MAXIMUM_TEMPERATURE | The maximum temperature on a date. |
| MINIMUM_TEMPERATURE | The minimum temperature on a date. |
| WIND_CHILL | The temperature of the wind felt on a date. |
| HEAT_INDEX | The warmth when humidity was considered. |
| SNOW_DEPTH | The level of snow that fell on a date. |
| WIND_GUST | The strength of the wind on a date. |
| RELATIVE_HUMIDITY | The measure of humidity present on a date. |

In regards to events occurring in Dublin and determining whether they will impact the estimate number of call outs on particular date, four major types of recurring events were chosen that occurred each year. GAA matches, Rugby matches, Soccer matches, and two of the yearly Dublin Marathons were selected, as these events could potentially have an effect on the number of call outs on the day's in which they occurred. To gather the dates that each of the sporting events occurred on, I located the relevant Wikipedia pages relating to the matches and input the dates into a csv file, along with the type of event that occurred on that date. For the dates of the two marathon events, a Google Search of the marathon along with the year was conducted. The dates that were shown were input into the same csv file that contained the sporting events. Table 4 shows the dates for each of the events that were chosen

to be used as part of this research. The dates that each of the events relating to GAA, Rugby, Soccer matches, as well as, the Marathons, are shown in Tables 5, 6, 7 and 8, respectively.

*Table 5: GAA Events Occuring in Dublin 2013 – 2015.*

| DATE | | |
|---|---|---|
| 02/02/2013 | 01/02/2014 | 07/02/2015 |
| 02/03/2013 | 01/03/2014 | 07/03/2015 |
| 10/03/2013 | 29/03/2014 | 28/03/2015 |
| 16/03/2013 | 13/04/2014 | 12/04/2015 |
| 23/03/2013 | 26/04/2014 | 25/04/2015 |
| 14/04/2013 | 27/04/2014 | 26/04/2015 |
| 28/04/2013 | 20/07/2014 | 12/07/2015 |
| 14/07/2013 | 02/08/2014 | 01/08/2015 |
| 27/07/2013 | 03/08/2014 | 02/08/2015 |
| 03/08/2013 | 24/08/2014 | 08/08/2015 |
| 04/08/2013 | 31/08/2014 | 23/08/2015 |
| 25/08/2013 | 21/09/2014 | 30/08/2015 |
| 01/09/2013 | | 05/09/2015 |
| 22/09/2013 | | 20/09/2015 |

*Table 6: Rugby Events Occuring in Dublin 2013 – 2015.*

| DATE | | |
|---|---|---|
| 10/02/2013 | 02/02/2014 | 14/02/2015 |
| 09/03/2013 | 08/02/2014 | 01/03/2015 |
| 09/11/2013 | 08/03/2014 | |
| 16/11/2013 | 08/11/2014 | |
| 24/11/2013 | 16/11/2014 | |

*Table 7: Soccer Events Occuring in Dublin 2013 – 2015.*

| DATE | | |
|---|---|---|
| 26/03/2013 | 05/03/2014 | 29/03/2015 |
| 02/06/2013 | 25/05/2014 | 07/06/2015 |
| 07/06/2013 | 03/09/2014 | 13/06/2015 |
| 06/09/2013 | 11/10/2014 | 07/09/2015 |
| 15/10/2013 | 18/11/2014 | 08/10/2015 |
| 15/11/2013 | | 16/11/2015 |

*Table 8 Marathon Events Occuring in Dublin 2013 – 2015.*

| DATE | | |
|---|---|---|
| 03/06/2015 | 02/06/2014 | 01/06/2015 |
| 28/10/2013 | 27/10/2014 | 26/10/2015 |

# Section 6

# Outline of Approach

### 6.1 Regression Model Implementation

Before any regression models were implemented, statistical analysis was required so to determine which features were best suited for the particular regression models. In order to estimate the number of Fire Brigade and Ambulance call outs, the use of a Poisson regression model and a negative binomial regression model was employed. The Poisson model was used as the baseline against which the predictions from the Negative Binomial model were compared. A Negative Binomial regression model was used because we were attempting to predict a count on the number of call outs experienced by Dublin Fire Brigade and Ambulance. (Famoye (2005) states that both Poisson and Negative Binomial models are suitable when attempting to predict count data. Negative Binomial regression is a generalisation of the Poisson regression model and differs in that, it does not follow the assumption that variance = mean. This is because the variance of the Negative Binomial regression model is a quadratic function of the mean (Ver Hoef and Boven, 2007). A major advantage of using a Negative Binomial model over a Poisson regression model is that the former is able to handle data that is over dispersed. This occurs when the variance is likely to be much greater than the mean (Chang, 2005). Overdispersion can be modelled using Negative Binomial or by Zero-Inflated Negative Binomial regression. However, for purposes of this study, as there were no excess zeros within the dataset, Zero-Inflated Negative Binomial regression was not suitable.

The Negative Binomial regression model consists of variables. y is a vector which contains a count of call outs seen on days 1 through n. Therefore, y = [y1,y2,y3…yn} where yn is the number of call outs that occurred on day n. The regression/explanatory variables are explained by a matrix X, of size m x n, where m = the rows in the dataset that contain n regression variables. $\lambda$ is a vector of size m x 1, which represents the rate of events. The vector contains n rates [$\lambda$0, $\lambda$1, $\lambda$2,… $\lambda$n], where n = the counts in y. This vector is calculated by the Negative Binomial model during the training phase. A new parameter alpha is required by the negative binomial regression model because it does not follow the assumption that the variance = mean. For purposes of this research, the Negative Binomial model used was the NB2 model. While there is a similarly named NB1 model, the difference between the NB1 model and the NB2 model is that, although both are derived from the Poisson-Gamma distribution, the NB1 model is less flexible than the NB2 model, i.e. the NB1 model does not fit training data as well as the NB2 model can. It is for this reason that the NB2 model is the more commonly used Negative Binomial model.

Implementing the NB2 regression model required four steps. Firstly, a Poisson regression model was fit on the datasets representing the number of call outs by Dublin Fire Brigade and Dublin Ambulance. A Poisson model was used so to provide the rates of events vector $\lambda$. Secondly, an auxiliary Ordinary Least Squares (OLS) regression model was fitted on the dataset. This model provided a value for alpha (represented by the letter *'a')*. Next, the alpha value was used to fit the NB2 regression model to the data. The dataset was split into training and testing. I decided to use 70% of the data for training and the remaining 30% for testing as Moreira, Carvalho and Horvath (2018) state that the average amount of data used for training is usually 63.2% with the remaining 36.8% set aside for testing. By dividing the data in a 70:30 ratio, this ensured a more balanced division when implemented through the dataset. To divide the data into training and testing, it was decided to randomly assign 70% of the data to the train set as this ensured a more random sampled approach. This resulted

in the remaining 30% of the data that was not assigned to training, being considered as the test set, which would be used to predict the number of Fire Brigade call outs.

Lastly, the NB2 regression model fitted with the alpha value made predictions about the volume of expected call outs for given days. The predictions generated at this stage, were compared to the predictions that were generated by the Poisson regression model.

### 6.2 Inclusion of Additional Features

In order to investigate whether including additional features, such as, weather and events could improve prediction methods, csv files relating to the weather for each day of the years 2013 to 2015 were merged together to provide an overall weather file, while, a csv file was created containing dates representing each of the four types of events previously explained. These files were incorporated into the Fire Brigade and Ambulance datasets by joining on similar dates. After joining of all the files, the Negative Binomial models were fitted on the datasets and predictions were generated that were then compared against the initial Negative Binomial model to determine whether including additional features had an effect on overall prediction accuracy on both, the Fire Brigade regression models and the Ambulance regression models.

# Section 7

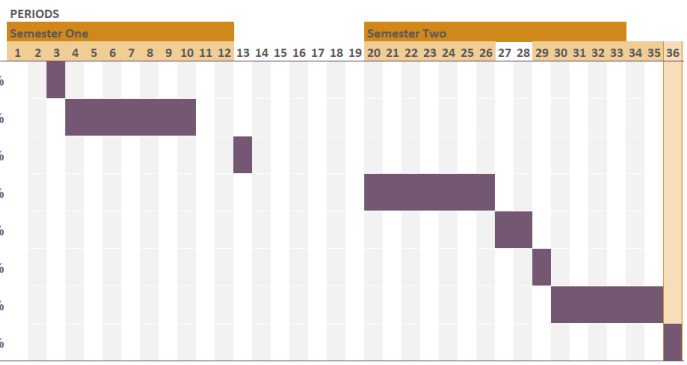
# Project Work Plan

The goals of this project required that the data released by Dublin City Council be cleaned and analysed. This task was completed as part of the Literature Review stage, as shown on the accompanying Gantt chart. The datasets have had their missing values handled by removing all rows with missing values and were merged into one dataset containing all relevant values. Statistical analysis was performed on the variables to isolate the features that will best help predict the Fire Brigade and Ambulance call outs. Upon completion of this research, a baseline Poisson regression model and a Negative Binomial regression model were created to predict Fire Brigade and Ambulance call outs. Simple temporal features, such as, the month of the year, date in the month, and day of the week were used to develop improved prediction methods. An investigation was carried out to determine if incorporating features, such as weather, or events occurring in Dublin, would help improve the prediction of the regression models. This pipeline was further used to predict specific types of fire brigade call outs, while also being used to predict Fire Brigade call outs at a local level.

### 7.1 Gantt Chart

To complete the goals of this project, a Gantt Chart was created to show the work plan of the project. Graph 5 shows the Gantt Chart. The chart highlights the work that has already been completed, shown through the deep purple colour. On returning to University College Dublin after Christmas, it was the aim that the first seven weeks would be dedicated solely to performing the statistical analysis, creating the linear regression model and the negative binomial model and applying these models to predict the number of call outs. One week was assigned to drawing conclusions from the models that were fitted on the data. To provide a considerable amount of time to ensure the write up of the results and conclusions found are of a sufficient standard, three weeks were for preparing the final report, with the hope of finishing this before the submission date during the final week of University. Due to the Covid-19 worldwide pandemic, the submission date was pushed back to the equivalent term week 36. This change is reflected in the Gantt Chart.

| ACTIVITY | PLAN START | PLAN DURATION | ACTUAL START | ACTUAL DURATION | PERCENT COMPLETE |
|---|---|---|---|---|---|
| Receive Project | 3 | 1 | 3 | 1 | 100% |
| Literature Review | 4 | 7 | 4 | 7 | 100% |
| Presentation | 13 | 1 | 13 | 1 | 100% |
| Model Implementation | 20 | 7 | 20 | 7 | 100% |
| Testing/Re-iterating | 27 | 2 | 27 | 2 | 100% |
| Drawing Conclusions | 29 | 1 | 29 | 1 | 100% |
| Preparing Final Report | 30 | 6 | 30 | 6 | 100% |
| Final Submission | 36 | 1 | 36 | 1 | 100% |




*Graph 5: Gantt Chart.*

### 7.2 Evaluation

The evaluation stage of this paper was crucial in determining if the Negative Binomial model made more accurate predictions than that of the Poisson regression model. To find out whether this was the case or not, a measure of the goodness of fit of the Negative Binomial model was carried out. This would show how well Poisson regression model and the negative binomial model described and fit the data. The statistics that were focused on were the Log-Likelihood, the Deviance and Pearson $Chi^2$, the $R^2$ Score and the Mean Absolute Percentage Error. In addition, a Pearson $Chi^2$ Test was performed to calculate the critical values of all regression models.

# Section 8

# Core Contribution

### 8.1 Forward Stepwise Regression Implementation

Before implementation of the Poisson and Negative Binomial models could begin, statistical analysis was carried out, in order to ensure that the models were fit with the best predictor variables available from the datasets. As there were many possible variable sets composed within the data, it was not feasible to consider every combination when constructing the models. In fact, when given a set of $p - 1$ candidate variables, where p is the number of candidate variables, there were $2^{p-1}$ alternative models that could be constructed. Therefore, it was decided to implement stepwise regression to build the models. Two forms of stepwise regression were considered: Backward Elimination and Forward Stepwise Regression. Backward Elimination is implemented by starting with a model consisting of all available predictor variables. Variables are removed one by one depending on the Akaike Information Criterion (AIC) and the model that minimises the AIC is considered the most appropriate. Akaike's Information Criteria is a measure of how well the potential model fits the data less a penalty term for how complex the model is. Forward Stepwise Regression is implemented by beginning with a model consisting of no predictor variables. Variables are added to the model one by one depending on the Akaike Information Criterion. The model that results in a reduced AIC is chosen as the most suitable. Backward Elimination would ensure that no more than $p + 1$ models would have to be tested, whereas, Forward Stepwise Regression could result in more possible scenarios being tested. However, the latter does not have a predictor variable correlation problem where important independent variables are removed during the early stages of model construction (Lewis, 2014). Therefore, it was decided to implement a Forward Stepwise Regression approach, as it would result in a more significant model than that, that would be created with Backward Elimination.

To apply Forward Stepwise Regression to the dataset, RStudio was chosen as a suitable application as it can run the programming language R, which is particularly useful when carrying out initial data analysis. To begin, the dataset was loaded into the application. Pre – processing was applied to the data, changing the features MONTH, DAY_OF_WEEK, DAY and HOUR into Factors, consisting of 12 levels, 7 levels, 31 levels and 24 levels, respectively. This ensured that each level present in the feature was considered independently, when either a Poisson model, or Negative Binomial model was fit on the data.

### 8.1.1 Creation of Poisson Baseline Model

The Poisson model consisted of elements of time as well as the STATION_AREA and DESCRIPTION features. These were fitted prior to the Forward Stepwise Regression, which would determine whether or not, these features were significant to the final Poisson model. To successfully run the Forward Stepwise Regression, an intercept only model was set up. This model contained an intercept but no regression coefficients. A Poisson model consisting of dependant variable COUNT and independent variables STATION_AREA + DESCRIPTION + MONTH + DAY_OF_WEEK + DAY + HOUR was also created. Forward Stepwise Regression was applied and the recommended final model with the lowest Akaike Information Criterion was generated. This result, along with the reduced AIC is shown below.

| **POISSON MODEL** | **AIC** |
|---|---|

| | |
|---|---|
| COUNT ~ MONTH + DAY + DAY_OF_WEEK + DESCRIPTION + HOUR + STATION_AREA | 337700 |

Forward Stepwise Regression included all features in the final Poisson model. At no point during the Forward Stepwise Regression was it decided that there would be a decrease in Akaike Information Criterion by adding no more explanatory variables.

### 8.1.2 Formation of Initial Negative Binomial Model

A Negative Binomial model was required to compare the baseline Poisson model against. This would determine whether or not a Poisson model, or a Negative Binomial model was more suitable for this dataset. To begin, a Negative Binomial intercept only model was initiated. This model included an intercept and had no explanatory variables present. A Negative Binomial model was created with the features STATION_AREA + DESCRIPTION + MONTH + DAY_OF_WEEK + DAY + HOUR as independent variables. Forward Stepwise Regression was applied and the final recommended model with the lowest Akaike Information Criterion was generated. This result is shown below.

| **NEGATIVE BINOMIAL MODEL** | **AIC** |
|---|---|
| COUNT ~ MONTH + DAY + DAY_OF_WEEK + DESCRIPTION + STATION_AREA + HOUR | 291400 |

Forward Stepwise Regression recommended a Negative Binomial model that included all features. At no point during the Forward Stepwise Regression was it decided that there would be a decrease in Akaike Information Criterion by adding no more explanatory variables. Evidently, there is a decrease in the Akaike Information Criterion between the Negative Binomial model and the Poisson model, suggesting that it is perhaps, already clear that the Negative Binomial model is much more suitable than the Poisson model.

### 8.1.3 Inclusion of Additional Features in Negative Binomial Model

A Negative Binomial model including some weather features and a Negative Binomial model including event features were used to compare against the original Negative Binomial model in order to determine whether including either of these additional features improved the model or results in a worse model. A Negative Binomial model with STATION_AREA + DESCRIPTION + MONTH + DAY_OF_WEEK + DAY + HOUR was created, along with the following weather features, TEMPERATURE + WIND_SPEED + PRECIPITATION + CLOUD_COVER. In a similar manner, a Negative Binomial model consisting of the features STATION_AREA + DESCRIPTION + MONTH + DAY_OF_WEEK + DAY + HOUR + EVENT was created. Forward Stepwise Regression was applied to these Negative Binomial models and the recommended models with the lowest Akaike Information Criterion were generated. The results are shown below.

| NEGATIVE BINOMIAL MODEL – WEATHER FEATURES | AIC |
|---|---|
| COUNT ~ MONTH + DAY + DAY_OF_WEEK + CLOUD_COVER + DESCRIPTION + WIND_SPEED + TEMPERATURE + PRECIPITATION + STATION_AREA + HOUR | 289400 |

| NEGATIVE BINOMIAL MODEL – EVENT FEATURE | AIC |
|---|---|
| COUNT ~ MONTH + DAY + DAY_OF_WEEK + DESCRIPTION + EVENT + STATION_AREA + HOUR | 291400 |

Forward Stepwise Regression suggested two Negative Binomial models which included all the initial features in each respective model. The Negative Binomial model with weather features obtained a lower AIC than the model without weather features, while the model with event features maintained the same AIC as the initial Negative Binomial model. At no point during the Forward Stepwise Regression was it decided that there would be a decrease in Akaike Information Criterion by adding no more explanatory variables.

### 8.2 Train – Test Split

To implement the Poisson and Negative Binomial models, open source application Jupyter Notebook was chosen. While the models could have been executed on R, I chose to switch to Jupyter Notebook as I was more confident in my programming skills with Python than I was with R. The Jupyter Notebook system is capable of running Python code and can be used for machine learning and visualisation techniques. To begin, the dataset was loaded representing the Fire Brigade call outs for Dublin into a DataFrame. As the requirement was to predict the number of call outs on a date throughout the year, the index for the DataFrame was set to DATE. When the dataset was loaded, the default type for the columns representing features MONTH, DAY_OF_WEEK, HOUR and DAY was integer, so these features were transformed into object types.

Before I can fit any model on the data, the dataset was split into 70% training and 30% testing. Early on when implementing this approach, it was discovered that errors would occur when attempting to train the Negative Binomial models if X_train and X_test sizes were not equal, i.e. sometimes shapes of (27052, 141) for the X_train and (11503, 138) for the X_test would occur. Because the 141 and 138 parts of the X_train and X_test were not equal, this would result in errors. To counteract this, a while loop was incorporated which would compare the shapes of the X_train and X_test. If the shapes were not equal, the random division of the dataset into 70% training and 30% testing would be run again until the shapes of the X_train and X_test were equal, i.e. until X_train (27052, 139) and (11503, 139) aligned.

### 8.3 Implementation of the Poisson Baseline Model

The regression expression of the Poisson baseline model was set up, i.e. COUNT ~ MONTH + DAY + DAY_OF_WEEK + DESCRIPTION + HOUR + STATION_AREA, in patsy notation. Patsy is a package for python that is useful in building regression matrices with a set of variables. It helps describe the features that

are to be contained within the required model, which, in this case, is the Poisson model. Essentially, patsy was being made aware of the fact that the dependant variable was COUNT, and that the independent variables were MONTH, DAY, DAY_OF_WEEK, DESCRIPTION, HOUR and STATION_AREA. Two matrices, X and Y, were created which would represent the X_train, X_test, y_train and y_test splits. The patsy notation aided in creating these matrices. Then, the GLM class from statsmodels was used to train the Poisson baseline model on the training set. Predictions corresponding to the test set were generated. The results were be compared against the Negative Binomial model results during the evaluation stage.

### 8.4 Execution of Negative Binomial Model

Patsy Notion was used to the set up the regression expression corresponding to the model generated in RStudio for the Negative Binomial model which incorporated time features, i.e. COUNT ~ MONTH + DAY + DAY_OF_WEEK + DESCRIPTION + STATION_AREA + HOUR. Patsy notation was used to aid the creation of the X and X matrices that correspond to the X_train, X_test, y_train and y_test that were used in the Negative Binomial model. The statsmodels GLM class was used to train a Poisson model. This was only required in order to generate predictions from the Negative Binomial model.

The values of the vector of fitted rates $\lambda$ were added into the newly created column in the training set, as were the values of the dependent variable of the Ordinary Least Squares regression. Patsy was used here to create the model specification for the Ordinary Least Squares regression. Patsy was told that AUX_OLS_DEP is the dependent variable and is explained by the values of the vector of fitted rates. This corresponded to the regression expression AUX_OLD_DEP ~ BB_LAMBDA -1. The -1 used in this expression was to ensure that patsy does not use an intercept of the regression. The Ordinary Least Squares regression model was then fit, and the value of alpha was printed. The value of alpha was 0.069149

It needed to be determined whether this value of alpha was significant or not. If alpha was not significant, and therefore, set as 0, then the Negative Binomial model would be defunct. This is due to the Negative Binomial model's variance function which states that the variance = mean = alpha$(\text{mean})^2$. If alpha were to be set as 0, the Negative Binomial model's variance function would reduce itself into the Poisson model's variance function, which is variance = mean. This would result in the Negative Binomial model failing in its attempt to do a better job of fitting the training data set than the Poisson model. The t-score of alpha was 89.432759. The critical t-value was calculated at a 99% significance level with degrees of freedom = 26821 and a value of 1.64491 was obtained. This was less than the t-statistic of alpha. Therefore, it was concluded that the value of alpha was statistically significant, and the Negative Binomial model was trained on the train set.

### 8.5 Implementation of the Negative Binomial Models with Additional Features

Patsy notation was used to set up the regression expressions which corresponded to the models generated in RStudio which incorporated weather features and event features. These model expressions were COUNT ~ MONTH + DAY +

DAY_OF_WEEK + CLOUD_COVER + DESCRIPTION + WIND_SPEED + TEMPERATURE + PRECIPITATION + STATION_AREA + HOUR, and COUNT ~ MONTH + DAY + DAY_OF_WEEK + DESCRIPTION + EVENT + STATION_AREA + HOUR. . Patsy notion aided in creating the X and Y matrices for each of these models so that predictions could be generated from the Negative Binomial models. The GLM class from statsmodels was implemented on each of these expressions in order to train two Poisson models that would aid in generating predictions. The results generated from each of these Negative Binomial models were compared against the results generated from the initial Negative Binomial model.

### 8.6 Advanced Goals Application

#### 8.6.1 Implementation of Pipeline to Specific Types of Call Outs

In order to predict the number of call outs for specific types of fires, Forward Stepwise Regression was applied to determine a baseline Poisson model, as well as three Negative Binomial models; one with only time element features, the second incorporating some weather features, and the third incorporating events. The COUNT feature in the Fire Brigade dataset was overwritten, to represent the number of times a specific type of fire occurred. From analysing the different types of fires within the dataset, it was decided to investigate how well the Poisson and Negative Binomial models would predict types relating to Car and Alarm, with counts of 3188 and 7064, respectively. The dataset was filtered to only include these two types of fires. This was done so to apply the pipeline previously described onto a subset representing specific types of fires. Before applying Forward Stepwise Regression on any models, the Poisson model and the Negative Binomial model were set up with COUNT as the dependent variable and STATION_AREA, MONTH, DAY_OF_WEEK, DAY and HOUR as the independent variables. For the Negative Binomial model which would incorporate weather features, TEMPERATURE, PRECIPITATION, VISIBILITY and WIND_CLOUD_COVER were included as independent variables, in addition to those previously mentioned. For the Negative Binomial model which would incorporate event features, EVENT was included as an independent variable, along with those in the original Negative Binomial model. After completion of Forward Stepwise Regression on all the required models, the following recommended models were generated as they each acquired the lowest Akaike Information Criterion.

| **POISSON MODEL** | **AIC** |
|---|---|
| COUNT ~ MONTH + HOUR + STATION_AREA + DAY + DAY_OF_WEEK | 156592215 |

| **NEGATIVE BINOMIAL MODEL** | **AIC** |
|---|---|
| COUNT ~ MONTH + HOUR + STATION_AREA + DAY_OF_WEEK | 743714 |

In relation to the Negative Binomial model with time features, the value of alpha was 0.073816 and the t-score of alpha was 72.724923. The t-value statistic at a 99% significance level with 7152 degrees of freedom was calculated to be 1.645067. This

concluded that the value of alpha was quite significant as it was considerably less than the t-score of alpha.

| NEGATIVE BINOMIAL MODEL – WEATHER FEATURES | AIC |
|---|---|
| COUNT ~ MONTH + HOUR + PRECIPITATION + STATION_AREA + WIND_SPEED + TEMPERATURE + DAY_OF_WEEK | 743481 |

| NEGATIVE BINOMIAL MODEL – EVENT FEATURE | AIC |
|---|---|
| COUNT ~ MONTH + DAY + DAY_OF_WEEK + EVENT + STATION_AREA + HOUR | 743709 |

After applying Forward Stepwise Regression, it became evidently clear that some features were not present in the recommended, final models. This is because including certain features would have resulted in an increase in the Akaike Information Criterion for that respective model. When the pipeline was applied to predicting specific types of call outs, the initial negative binomial model and the Negative Binomial model incorporating event features excluded DAY while the Negative Binomial model incorporating weather features excluded DAY and CLOUD_COVER. This suggested that these features were not significant to determining specific types of fires.

### 8.6.2 Implementation of Pipeline to Call Outs at a Local Level

In order to predict the number of call outs for specific types of fires, Forward Stepwise Regression was applied to determine a baseline Poisson model, as well as three Negative Binomial models; one with only time element features, the second incorporating some weather features, and the third incorporating events. The COUNT feature in the Fire Brigade dataset was overwritten, to represent the number of times a call out occurred at a Fire Brigade station in Dublin. From analysis, it was discovered that the two most busy stations in Dublin within the dataset were Tallaght and Dolphin's Barn, with counts of 4427 and 2766, respectively. The dataset was filtered to only include these two station areas. This was done so to apply the pipeline previously described onto a subset of the data representing call outs at a local level. Before applying Forward Stepwise Regression on any models, the Poisson model and the Negative Binomial model were set up with COUNT as the dependent variable and DESCRIPTION, MONTH, DAY_OF_WEEK, DAY and HOUR as the independent variables. For the Negative Binomial model which would incorporate weather features, TEMPERATURE, PRECIPITATION, CLOUD_COVER and WIND_SPEED was included as independent variables, as well as those previously mentioned. For the Negative Binomial model which would incorporate event features, EVENT was included as an independent variable, along with those in the original Negative Binomial model. After completion of Forward Stepwise Regression on all the required models, the following recommended models were generated as they each acquired the lowest Akaike Information Criterion.

| POISSON MODEL | AIC |
|---|---|

| | |
|---|---|
| COUNT ~ DESCRIPTION + HOUR + DAY + MONTH + DAY_OF_WEEK | 25470000 |

| **NEGATIVE BINOMIAL MODEL** | **AIC** |
|---|---|
| COUNT ~ DESCRIPTION + HOUR | 671900 |

In relation to the Negative Binomial model with time features, the value of alpha was 0.044834 and the t-score of alpha was 135.806798. The t-value statistic at a 99% significance level with 7411 degrees of freedom was calculated to be 1.645059. This concluded that the value of alpha was quite significant as it was considerably less than the t-score of alpha.

| **NEGATIVE BINOMIAL MODEL – WEATHER FEATURES** | **AIC** |
|---|---|
| COUNT ~ DESCRIPTION + HOUR + WIND_SPEED | 671883 |

| **NEGATIVE BINOMIAL MODEL – EVENT FEATURE** | **AIC** |
|---|---|
| COUNT ~ DESCRIPTION + HOUR | 671885 |

Once again, it was evidently clear that some features were not present in the recommended, final models. When the pipeline was applied to predicting call outs at a local level, the initial negative binomial model excluded MONTH, DAY, and DAY_OF_WEEK as possible features. The Negative Binomial model incorporating weather features excluded these features in addition to TEMPERATURE, PRECIPITATION and CLOUD_COVER. The Negative Binomial model incorporating event features excluded MONTH, DAY, and DAY_OF_WEEK, while also surprisingly excluding EVENT. This highlighted that EVENT was not considered a key feature when the pipeline was applied to call outs at a local level. Essentially, this rendered this model into the initial Negative Binomial model.

### 8.6.3 Implementation of Pipeline to Ambulance Call Out Dataset

The pipeline was applied to the Ambulance service call out dataset in order to determine whether the same features are predictive of both the Fire Brigade and Ambulance data. Before applying Forward Stepwise Regression to the Ambulance dataset, some pre-processing was carried out. Firstly, it was discovered that the Ambulance Dataset was too large to run on the laptop being used. There was not enough space in the RAM to run the dataset in its entirety. A work around was applied which involved having to change the paging size of the laptop and then restarting the laptop. However, this approach would only correctly work when the Ambulance dataset was reduced to include only dates between 1st January 2013 and 31st December 2014, inclusive. So, the year 2015 was removed in order to ensure that the Poisson and Negative Binomial models were able to be applied on this dataset. This work around approach is included within the README file if the error appears on any laptop investigating the research within this project.

In order to predict the number of call outs for Ambulance call outs, Forward Stepwise Regression was applied to determine a baseline Poisson model, as well as three Negative Binomial models; one with only time element features, the second incorporating some weather features, and the third incorporating events. A COUNT feature was created in the Ambulance dataset, to represent the number of times a call out for an Ambulance occurred on each date of the year. Before applying Forward Stepwise Regression on any models, the Poisson model and the Negative Binomial model were set up with COUNT as the dependent variable and STATION_AREA, CLINICAL_STATUS, MONTH, DAY_OF_WEEK, DAY and HOUR as the independent variables. For the Negative Binomial model which would incorporate weather features, TEMPERATURE, PRECIPITATION, CLOUD_COVER and WIND_SPEED were included as independent variables, in addition to those previously mentioned. For the Negative Binomial model which would incorporate event features, EVENT was included as an independent variable, along with the features in the original Negative Binomial model. After completion of Forward Stepwise Regression on all the required models, the following recommended models were generated as they each acquired the lowest Akaike Information Criterion.

| **POISSON MODEL** | **AIC** |
|---|---|
| COUNT ~ DAY_OF_WEEK + MONTH + DAY + HOUR + CLINICAL_STATUS + STATION_AREA | 1258966 |

| **NEGATIVE BINOMIAL MODEL** | **AIC** |
|---|---|
| COUNT ~ DAY_OF_WEEK + MONTH + DAY + HOUR + CLINICAL_STATUS | 1243272 |

In relation to the Negative Binomial model with time features, the value of alpha was 0.002717 and the t-score of alpha was 50.533032. The t-value statistic at a 99% significance level with 101634 degrees of freedom was calculated to be 1.644869. This concluded that the value of alpha was quite significant as it was considerably less than the t-score of alpha.

| **NEGATIVE BINOMIAL MODEL – WEATHER FEATURES** | **AIC** |
|---|---|
| COUNT ~ DAY_OF_WEEK + MONTH + DAY + TEMPERATURE + CLOUD_COVER + PRECIPITATION + HOUR + CLINICAL_STATUS | 1240251 |

| **NEGATIVE BINOMIAL MODEL – EVENT FEATURE** | **AIC** |
|---|---|
| COUNT ~ DAY_OF_WEEK + MONTH + DAY + EVENT + HOUR + CLINICAL_STATUS | 1242519 |

Again, it became evidently clear that some features were not present in the recommended, final models. When the pipeline was applied to predicting Ambulance call outs, all three Negative Binomial models exclude STATION_AREA, with the Negative Binomial model incorporating weather features excluding WIND_SPEED

as well. This suggested that STATION_AREA and WIND_SPEED were not significant features when attempting to predict the number of Ambulance call outs.

# Section 9

# Evaluation

Evaluating the Poisson and Negative Binomial models was crucial in determining which model was most successful in predicting call outs. To ensure accurate evaluation of all models, a subset of regression evaluation techniques were selected. The regression models relating to the Fire Brigade call out data were evaluated first, followed by the regression models associated with the Ambulance call out data. This allowed for cross examination to determine any similarities that were present in both datasets. For clarity, graphs were provided, along with an explanation relating to each of the following methods:

1. Log-Likelihood
2. Deviance and Pearson $Chi^2$ with a Pearson $Chi^2$ Test
3. $R^2$ Score
4. Mean Absolute Percentage Error

These methods were chosen so that an accurate goodness-of-fit was determined for each of the models. All models were critiqued according to these techniques, one method at a time.

### 9.1 Log-Likelihood

*Table 9: Log-Likelihood Values for Fire Brigade and Ambulance Models.*

| MODEL | VALUE |
|---|---|
| **FIRE BRIGADE** | |
| Poisson | -117520 |
| Negative Binomial | -101510 |
| Negative Binomial - Weather | -100830 |
| Negative Binomial - Events | -101050 |
| **AMBULANCE** | |
| Poisson | -439640 |
| Negative Binomial | -434260 |
| Negative Binomial - Weather | -433170 |
| Negative Binomial - Events | -433990 |

The first technique used for evaluation was the Log-Likelihood. The Maximum Log-Likelihood was generated by the Maximum Likelihood Estimation. This technique fixes the values of the model coefficients to optimal values and was executed by statsmodels when we trained the Poisson and Negative Binomial models. Generally, the Log-Likelihood values are negative, which was the case with our models. Log-Likelihood values are to be compared against each other, as a single

Log-Likelihood value cannot be interpreted by itself. Graph 6 shows the Log-Likelihood values of the Fire Brigade regression models and Graph 7 shows the values of the Ambulance regression models. Comparing the Fire Brigade models against the Ambulance models allows us to deduce some similarities. Both Poisson models obtained the worst Log-Likelihood values when fitted on the two datasets. The Log-Likelihoods were improved when the initial Negative Binomial models were fit on the data. In terms of the Log-Likelihood, incorporating weather and event features into the Negative Binomial model proved successful across both datasets, with both models obtaining better values than the respective initial Negative Binomial model. In fact, the Negative Binomial model incorporating weather features achieved the best Log-Likelihood.

### 9.2 Deviance and Pearson Chi$^2$

Deviance is a measure of the linear model with regard to a perfect model. A value reported for Deviance is always larger than or equal to zero. A value of 0 indicates a perfect model has been achieved. Therefore, low values are more desirable than high values.

*Table 10: Deviance and Pearson Chi$^2$ values for Fire Brigade and Ambulance Models.*

| MODEL | DEVIANCE | PEARSON CHI$^2$ |
|---|---|---|
| **FIRE BRIGADE** | | |
| Poisson | 88461 | 89100 |
| Negative Binomial | 21615 | 21700 |
| Negative Binomial - Weather | 21470 | 21600 |
| Negative Binomial - Events | 21616 | 21700 |
| **AMBULANCE** | | |
| Poisson | 153780 | 156000 |
| Negative Binomial | 98703 | 101000 |
| Negative Binomial - Weather | 98855 | 101000 |
| Negative Binomial - Events | 98684 | 100000 |

The Pearson's Chi$^2$ values are used within the Pearson Chi$^2$ test. This test indicates how well the data distribution fits with the distribution that would be expected if the variables were independent. High Pearson Chi$^2$ values indicate a poor fit on the data, whereas low Pearson Chi$^2$ values indicate a good fit. Graph 8 shows the Fire Brigade regression model Deviance and Pearson Chi$^2$ values and Graph 9 shows the Ambulance regression model values. The Poisson model performed the worst across both datasets, obtaining the poorest Deviance and Pearson Chi$^2$ values. A drop in Deviance was evident in both graphs, which indicated that the application of the Negative Binomial model was much more successful than the Poisson model. The inclusion of additional features improved both, the Deviance and Pearson Chi$^2$ values across both datasets. Weather features had the greatest impact upon the Fire Brigade data, whereas, event features improved the values of the Ambulance dataset. In fact, weather features resulted in the worst values for Deviance of all the Negative Binomial models, which highlighted that different features had an impact on the different datasets.

### 9.2.1 Pearson Chi$^2$ Test

*Table 11: Pearson Chi$^2$ Test Results for Fire Brigade and Ambulance Models.*

| MODEL | DF | CRITICAL VALUE |
|---|---|---|
| **FIRE BRIGADE** | | |
| Poisson | 26683 | 27064.113 |
| Negative Binomial | 26683 | 27064.113 |
| Negative Binomial - Weather | 26679 | 27060.084 |
| Negative Binomial - Events | 26679 | 27060.084 |
| **AMBULANCE** | | |
| Poisson | 101543 | 99425.436 |
| Negative Binomial | 101558 | 99439.974 |
| Negative Binomial - Weather | 101555 | 99437.066 |
| Negative Binomial - Events | 101554 | 99436.097 |

As the reported values of Pearson Chi$^2$ and Deviance were considerably high for all models across both datasets, a quantitative determination of the goodness of fit at a 95% confidence level was made. The results of the Fire Brigade regression models are shown in Table 8 and the results of the Ambulance regression models are shown in Table 9. Some interesting facts became known from examination of these tables. Across both datasets, the Poisson model's critical value was considerably smaller than the reported respective, Deviance Pearson Chi$^2$ values. Therefore, these models, while showing moderate fits for the test sets, fit very poorly on the training set. The Negative Binomial models for the Fire Brigade data had much larger critical values than their respective Deviance and Pearson Chi$^2$ values. As a result, this indicated a better fits on the data. However, the Negative Binomial models for the Ambulance dataset obtained critical values higher than the reported Deviance, but lower than the Pearson Chi$^2$, which suggested that these models only provided a moderate fit on the data. Nevertheless, The Pearson Chi$^2$ Test highlighted that the inclusion of additional weather and event features across both datasets, improved the fit of the Negative Binomial models.

## 9.3 $R^2$ Score

*Table 12: $R^2$ Scores for Fire Brigade and Ambulance Models.*

| MODEL | VALUE |
|---|---|
| **FIRE BRIGADE** | |
| Poisson | 0.4575 |
| Negative Binomial | 0.4635 |
| Negative Binomial - Weather | 0.4812 |
| Negative Binomial - Events | 0.4629 |
| **AMBULANCE** | |
| Poisson | 0.3443 |
| Negative Binomial | 0.3444 |
| Negative Binomial - Weather | 0.3581 |
| Negative Binomial - Events | 0.3478 |

The $R^2$ Score measures just how well the regression model fits the dataset, i.e. how close the predicted values compare to the actual values. Values are always between 0 and 1, with 0 suggesting that the regression model explains none of the

variation present and 1 suggesting that the regression model explains all of the variation present. Generally, higher values are recommended. Graph 9 shows the $R^2$ Scores of the Fire Brigade regression models and Graph 10 shows the R2 Scores of the Ambulance regression models. The $R^2$ Scores of the Fire Brigade models were slightly higher than the equivalent models relating to the Ambulance dataset. Across both datasets, the Poisson model obtained the lowest $R^2$ Scores. While implementing the Negative Binomial models slightly increased the respective $R^2$ Scores, it was the addition of weather features that obtained the highest scores throughout both datasets. This suggested that the weather features appeared to have a positive effect on the amount of variation that could be explained by the model's independent variables.

It could be argued that the $R^2$ Scores of all the regression models above are considerably low as they all attained scores of less than 0.5. Research conducted by Moksony (1999) suggests that low $R^2$ Scores are not as detrimental to the goodness of fit as first presumed. Due to the $R^2$ equation being affected by many differing factors, Moksony recommended that it does not capture the strength of the impact of the independent variables or indicate how well the model fits the data. As a result of these findings, it was decided that the $R^2$ Score was not sufficient in determining how well the models fit the Fire Brigade data and so, another goodness of fit measure was required.

### 9.4 Mean Absolute Percentage Error

*Table 13: Mean Absolute Percentage Error Values for Fire Brigade and Ambulance Models.*

| MODEL | VALUE |
|---|---|
| **FIRE BRIGADE** | |
| Poisson | 26.8636 |
| Negative Binomial | 26.1646 |
| Negative Binomial - Weather | 25.4559 |
| Negative Binomial - Events | 26.1831 |
| **AMBULANCE** | |
| Poisson | 6.7253 |
| Negative Binomial | 6.7237 |
| Negative Binomial - Weather | 6.5998 |
| Negative Binomial - Events | 6.7033 |

The Mean Absolute Percentage Error (MAPE) is a measure of how accurate a forecasting system is. This is the most common measure used to estimate error in models and is particularly useful when the size of a variable is significant in evaluating the accuracy of predictions (Khair, Fahmi, Hakim and Rahim, 2017). Lower values are more desirable with this statistical measure as they represent a low amount of error between the actual and predicted values in a regression model. Graph 11 shows the MAPE values obtained by the Fire Brigade regression models and Graph 12 shows the Ambulance regression model values. The Poisson models obtained the worst values across both datasets. The addition of Negative Binomial models certainly improved the MAPE values obtained, with weather features achieving the lowest MAPE throughout the two datasets. As previously mentioned, event features appeared to have more of an impact on the Ambulance call outs. This is evident here as the MAPE obtained was reduced compared to the initial Negative Binomial model.

Whereas, event features worsened the MAPE in the Fire Brigade regression model. Analysing both sets of regression models, the Fire Brigade regression models achieved reasonable forecasting (>20 & <30) whereas, the Ambulance models obtained excellent estimating (<10). It is evidently clear that the regression models performed much better on the Ambulance data, than the Fire Brigade data.

### 9.5 Use of Pipeline to Predict Specific Types of Call Outs & Call Outs at Local Level

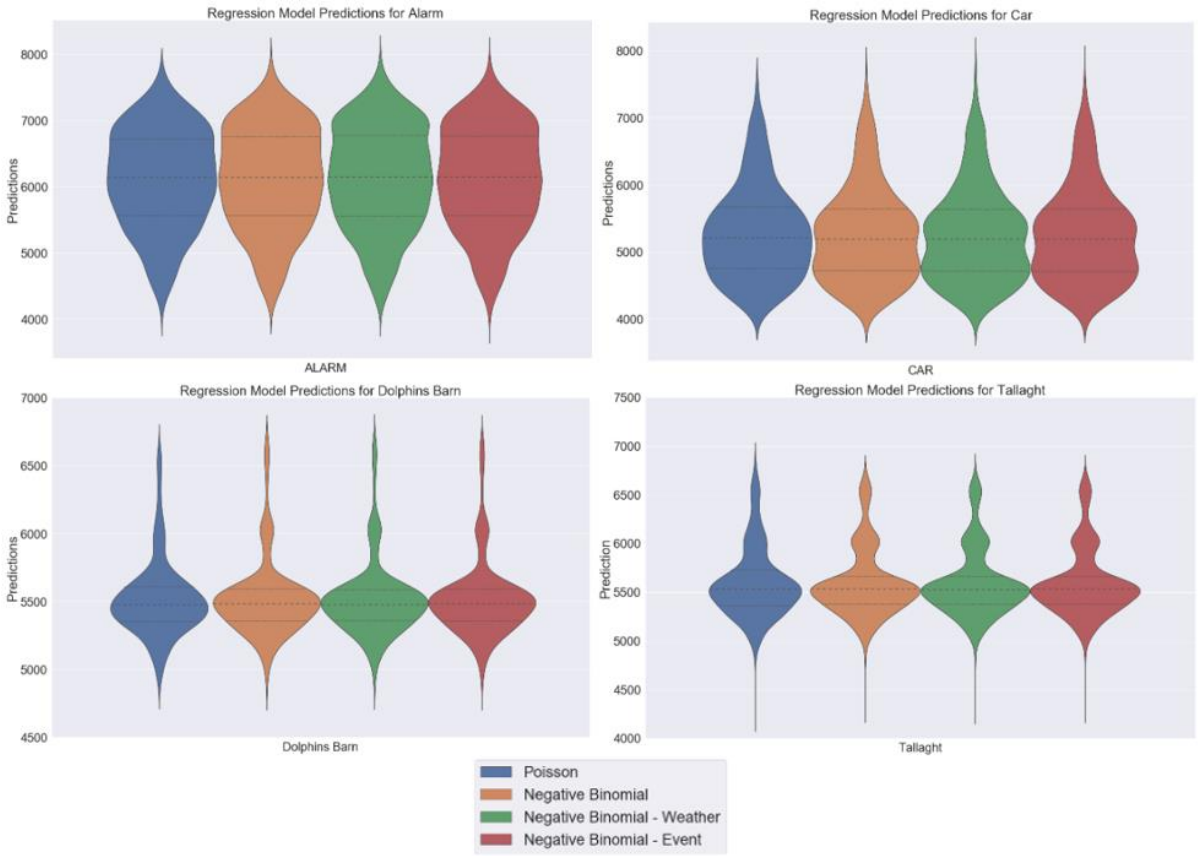
To evaluate how well the regression models predicted call outs relating to specific types of fires, as well as, call outs at a local level, each of the model's Mean Absolute Percentage Error was examined, along with the predictions each model made. This highlighted how accurate the regression models were at forecasting Fire Brigade call outs relating to Cars and Alarms, as well as, Fire Brigade call outs to Tallaght and Dolphin's Barn. Table 10 shows the MAPE values obtained by the models when predicting call outs relating to types of fires, and Table 11 shows the MAPE values for the models when predicting call outs a local level.

*Table 14: Mean Absolute Percentage Error Values for Specific Types of Fires and Call Outs at a Local Level.*

| MODEL | VALUE |
|---|---|
| **ALARM** | |
| Poisson | 14.1032 |
| Negative Binomial | 14.0395 |
| Negative Binomial - Weather | 14.0205 |
| Negative Binomial - Events | 14.0607 |
| **CAR** | |
| Poisson | 65.6680 |
| Negative Binomial | 65.0893 |
| Negative Binomial - Weather | 64.9705 |
| Negative Binomial - Events | 65.0056 |
| **DOLPHINS BARN** | |
| Poisson | 37.2728 |
| Negative Binomial | 37.2705 |
| Negative Binomial - Weather | 37.2461 |
| Negative Binomial - Events | 37.2705 |
| **TALLAGHT** | |
| Poisson | 14.3203 |
| Negative Binomial | 14.3146 |
| Negative Binomial - Weather | 14.3092 |
| Negative Binomial - Events | 14.3146 |

It is quite clear that the regression model performed quite poorly when attempting to predict types of fires relating to Cars and call outs to Dolphins Barn. Yet, at the same time, were able to achieve considerably good forecasting when estimating call outs for Alarms and call outs to Tallaght. To further analyse the regression models, violin plots were employed to take a closer look at the predictions made. Graph 13 shows four violin plots, the first row shows the predictions for Alarm and Car, and the second row shows the predictions for Dolphins Barn and Tallaght. Taking into account call outs relating to Alarm and Car had actual counts of 3188 and

7063, and call outs to Tallaght and Dolphins Barn had counts of 6525 and 4018, there were visible issues from the violin plots. In all instances, the quartile ranges showed that the majority of predictions being made by all regression models were nowhere near the actual counts. Therefore, all regression models were overpredicting their estimations and this favoured the type of fire and station area with the higher count. This overestimation did not favour the lower counts for call outs relating to Alarms and Dolphins Barn, resulting in the inaccurate MAPE witnessed.




*Graph 13: Dispersion of Predictions for Specific Types of Fires and Call Outs at a Local Level*

# Section 10

# Conclusion

### 10.1 Concluding Remarks

Some noticeable patterns were observed throughout this research. In terms of a baseline model, the Poisson model followed the pattern that was to be expected, that is, that it would attain the worst performance across all evaluation criteria. The results obtained by the Poisson model highlighted that the variance does not equal the mean for both, the Fire Brigade and Ambulance datasets. This is what was intended to occur as the purpose of the Negative Binomial model was to show that this model's variance

equation function was better suited to this dataset, rather than the Poisson model's variance function.

From analysis of the three Negative Binomial models, it is evidently clear that the addition of features such as, weather and events, improved the predictions made by the Negative Binomial model. However, overall, there was extraordinarily little improvement. Saying that, the Negative Binomial model which incorporated weather features regularly outperformed the other regression models, highlighting that weather features are should be considered when estimating Fire Brigade and Ambulance call outs. Similarly, while the event features did not perform as well as weather features, they did generally, outperform the Negative Binomial with no additional features. This indicates that event features should also be considered when attempting to predict call outs relating to the Ambulance and Fire Brigade services.

In terms of Fire Brigade call outs at a local level and specific types of fires, the results were far from ideal. The overestimation present across all four models skewed the results. For call outs relating to specific types of fires, Alarm call outs were estimated more correctly, at the expense of Car call outs. In the same manner, when predicting call outs at a local level, call outs to Tallaght were estimated more correctly in comparison to call outs to Dolphin's Barn.

Focussing on the evaluation criteria employed, the Log-Likelihood, Deviance and Pearson $Chi^2$ values obtained were all considerably high. In an ideal scenario, these figures would have been lower. Similarly, when configuring the Mean Absolute Percentage Error for Fire Brigade regression models, the errors would have been lower, in a perfect scenario. Ideally, the Fire Brigade MAPE would have followed the reporting of the Ambulance MAPE errors, i.e. less than 10, therefore, excellent forecasting.

Looking back at this research project, focusing in particular of my mindset at the Literature Review stage, I would have imagined that I was expecting to receive more desirable outcomes across all aspects of the regression implementation. From reading past papers on predicting count data, the results obtained by those researchers was far more accurate and allowed for more definite final remarks than those that I achieved. I was surprised as to how generally poor the results were from all my Evaluation criteria. I was certainly expecting much higher $R^2$ Scores and much more desirable Mean Absolute Percentage Error values, especially after applying the Forward Stepwise Regression to obtain the best model in each scenario. These unsatisfactory results would not have been considered when research into this project first began in late 2019.

### 10.2 Recommendations for Future Research

While the Negative Binomial regression models implemented throughout this project generated results that were able to be used explained, and most importantly, compared against the Poisson model, and to each other, they did not perform as well had expected. Even though they consistently outperformed the Poisson model in terms of fitting the Fire Brigade and Ambulance datasets, it can be argued that the Negative Binomial models were sub-optimal. For future research into predicting Fire Brigade and Ambulance call outs in Dublin city, the employment of more advanced regression techniques are recommended . A Random Forest Regression model could be applied

to the datasets, as this method uses a measure of predictor variable performance and a measure of the internal structure of the dataset to estimate quite accurate predictions (Liaw and Wiener, 2002). Similarly, a Long-Short Term Memory Neural Network could be implemented. This form of Neural Network makes effective use of model parameters by addressing their computational efficiency (Sak, Senior and Beaufays, 2014). Both the Random Forest and Long-Short Term Memory Neural Network would be suitable methods for future research into predicting Fire Brigade and Ambulance call outs in Dublin city.

# Section 11

# Bibliography


1. Aladdini, K. (2010). *EMS Response Time Models: A Case Study and Analysis for the Region of Waterloo*.
2. Bianchi, L., Jarrett, J. and Choudary Hanumara, R. (1998). *Improving forecasting for telemarketing centers by ARIMA modelling with intervention*.
3. Bowdish, G., H. Cordell, W., Bock, H. and F. Vukov, L. (1992). *Using regression analysis to predict emergency patient volume at the Indianapolis 500 mile race*.
4. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). *Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective*.
5. Brown, L., Weinberg, J. and Stroud, J. (2006). *Bayesian Forecasting of an Inhomogeneous Poisson Process with Applications to Call Center Data*.
6. Budge, S., Ingolffson, A. and Zerom, D. (2010). *Empirical Analysis of Ambulance Travel Times: The Case of Calgary Emergency Medical Services*.
7. Catalina, T., Virgone, J. and Blanco, E. (2008). *Development and validation of regression models to predict monthly heating demand for residential buildings*.
8. Chang, L. (2005). *Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network*.
9. Chanouf, N., L'Ecuyer, P., Ingolfsson, A. and Avramidis, A. (2007). *The application of forecasting techniques to modelling emergency medical system calls in Calgary, Alberta*.
10. D. Kamenetzky, R., J. Shuman, L. and Wolfe, H. (1982). *Estimating Need and Demand for Prehospital Care*.
11. Dolney, T. and Sheridan, S. (2006). *The relationship between extreme heat and ambulance response calls for the city of Toronto, Ontario, Canada*.
12. Dublin City Socio-Economic Profile. (2011). *Dublin City Socio-Economic Profile Indicator Catalogue*.
13. Dublincity.ie. (2019). *Dublin Fire Brigade | Dublin City Council*. [online] Available at: http://www.dublincity.ie/main-menu-your-council-about-dublin-city-council-council-departments/dublin-fire-brigade
14. Duncanson, M., Woodward, A. and Reid, P. (2000). *Socioeconomic deprivation and fatal unintentional domestic fire incidents in New Zealand 1993–1998*.
15. Famoye, F., 2005. *Count Data Modeling: Choice Between Generalized Poisson Model And Negative Binomial Model*.

16. G. Henderson, S. (2005). *Should we model dependence and nonstationarity, and if so how?*
17. Khair, U., Fahmi, H., Hakim, S. and Rahim, R., 2017. *Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error*.
18. Kolesar, P. (1975). *A Model for Predicting Average Fire Engine Travel Times*.
19. Kolesar, P., Walker, W. and Hausner, J. (1975). *Determining the Relation between Fire Engine Travel Times and Travel Distances in New York City*.
20. Lewis, M., 2014. *Stepwise Verses Hierarchical Regression: Pros And Cons*.
21. Liaw, A. and Wiener, M., 2002. *Classification and Regression By Randomforest*.
22. Matteson, D., McLean, M., Woodward, D. and Henderson, S. (2011). *Forecasting Emergency Medical Service Call Arrival Rates*.
23. Moksony, F., 1999. *Small Is Beautiful: The Use and Interpretation Of R2 In Social Research*.
24. Moreira, J., Carvalho, A. and Horvath, T., 2018. *A General Introduction To Data Analytics*.
25. Nitschke, M., Tucker, G., Hansen, A., Williams, S., Zhang, Y. and Bi, P. (2011). *Impact of two recent extreme heat episodes on morbidity and mortality in Adelaide, South Australia: a case-series analysis*.
26. Pedregal, D. and Trapero, J. (2006). *Electricity prices forecasting by automatic dynamic harmonic regression models*.
27. Sak, H., Senior, A. and Beaufays, F., 2014. *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*.
28. Thornes, J. (2013). *The Impact of Extreme Weather and Climate Change on Ambulance Incidents and Response Times in London*.
29. Tych, W., J. Pedregal, D., C. Young, P. and Davies, J. (2002). *An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system*.
30. Ver Hoef, J. and Boveng, P. (2007). *Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Over Dispersed Count Data?*.
31. Zhu, Z., McKnew, M. and Lee, J. (1992). *Effects of Time Varied Arrival Rates: An Investigation in Emergency Ambulance Service Systems*.