2018

From Business Understanding to Deployment: An application of Machine Learning Algorithms to Forecast Customer Visits per Hour to a Fast-Casual Restaurant in Dublin

Odunayo David Adedeji

*Dublin Institute of Technology, Ireland*

Follow this and additional works at: [https://arrow.dit.ie/scschcomdis](https://arrow.dit.ie/scschcomdis)

Part of the Computer Engineering Commons

**Recommended Citation**

From Business Understanding to Deployment: An application of Machine Learning algorithms to Forecast Customer Visits per hour to a Fast-Casual Restaurant in Dublin

Odunayo David Adedeji

D17124385

A dissertation submitted in partial fulfilment of the requirements of Dublin Institute of Technology for the degree of M.Sc. in Computing (Data Analytics)

2018
I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute’s guidelines for ethics in research.

Signed: Odunayo David Adedeji

Date: 12 September 2018
ABSTRACT

This research project identifies the significant factors that affect the number of customer visits to a fast-casual restaurant every hour and proceeds to develop several machine learning models to forecast customer visits. The core value proposition of fast-casual restaurants is quality food delivered at speed which means they have to prepare meals in advance of customers’ visit but the problem with this approach is in forecasting future demand, under estimating demand could lead to inadequate meal preparation which would leave customers unsatisfied while over estimation of demand could lead to wastage especially with restaurants having to comply with food safety regulations whereby heated food not consumed within 90 minutes has to be discarded. Hourly forecasting of demand as opposed to monthly or even daily forecasting is important to help the manager of the fast-casual restaurant optimize resources and reduce wastage.

Approaches to forecasting demand can be broadly categorized into qualitative and quantitative methods. Quantitative methods can be further divided into time series and regression-based methods. The regression-based approach which is used for this study enabled the researcher to gather data on several factors hypothesized to have an impact on the number of customer visits to the fast-casual restaurant every hour, carry out an experiment to test for the significance of these factors and to develop several predictive machine learning models capable of predicting the number of customer visits every hour.

The results of the experiments carried out shows that hour, day, public holidays, temperature, humidity, rain and windspeed are significant factors in predicting the number of hourly customer visits. Multiple linear regression, regression tree, random forest and gradient boosting machine learning algorithms were also trained to predict the number of customer visits with the Gradient boosting algorithm achieving the lowest Mean Absolute Percentage Error (MAPE) of 18.82%.

Keywords: Forecasting, Machine learning, Restaurant, Predictive Modelling, Regression
ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my family for their financial and emotional support throughout my education.

I would also like to thank my supervisor Brian Leahy for his gentle but precise guidance throughout this project.
TABLE OF CONTENTS

ABSTRACT ................................................................. ii
ACKNOWLEDGEMENTS ......................................................... iii
TABLE OF CONTENTS ......................................................... iv
TABLE OF FIGURES ......................................................... viii
TABLE OF TABLES ........................................................... x
ABBREVIATIONS ............................................................... xi

1 INTRODUCTION ................................................................... 1
  1.1 Introduction ........................................................................ 1
  1.2 Background ......................................................................... 1
  1.3 The Restaurant .................................................................... 3
  1.4 Research Problem ............................................................... 4
  1.5 Forecasting vs Prediction ..................................................... 5
  1.6 Research Question and Hypothesis ....................................... 5
  1.7 Research Aims and Objectives ............................................. 6
  1.8 Research Methodologies .................................................... 7
  1.9 Scope and Limitations ......................................................... 9
  1.10 Outline of the Thesis ......................................................... 9

2 LITERATURE REVIEW ....................................................... 11
  2.1 Introduction ......................................................................... 11
  2.2 Forecasting ........................................................................ 11
  2.3 Time Series Data ............................................................... 12
  2.4 Approaches to Forecasting .................................................. 14
    2.4.1 Qualitative Forecasting ............................................... 16
    2.4.2 Quantitative Forecasting .............................................. 18
      2.4.2.1 Regression methods ............................................. 18
      2.4.2.2 Time series methods .......................................... 19
        2.4.2.2.1 Naïve Forecasting ......................................... 19
        2.4.2.2.2 Simple Average Forecasting ............................ 20
        2.4.2.2.3 Simple Moving Average ................................. 20
        2.4.2.2.4 Weighed Moving Average ......................... 21
# Table of Contents

1. **Introduction** ........................................................................................................... 1
2. **Literature Review** .................................................................................................... 10
3. **Methodology** ........................................................................................................... 22
   - Data Collection
   - Data Preprocessing
   - Feature Engineering
   - Model Selection
4. **Results and Analysis** ............................................................................................... 35
5. **Conclusion** .............................................................................................................. 51
6. **Appendix** ............................................................................................................... 53

**References** .................................................................................................................. 57

**Appendix** ..................................................................................................................... 60
TABLE OF FIGURES

Figure 1.1: Valuation of Ireland’s Foodservice Market .......................................................... 1
Figure 1.2: CRISP DM Lifecycle ................................................................................................ 8
Figure 2.1: Time Series Plot of USA stock market on March 16 1999 ....................................... 13
Figure 2.2: Different time series plots showing various or no combinations of seasonality,
trend and cyclic behaviour ........................................................................................................ 14
Figure 2.3: Process of Subjective Adjustment ........................................................................... 17
Figure 2.4: KDD Process Overview ......................................................................................... 24
Figure 2.5: SEMMA Process .................................................................................................... 25
Figure 2.6: Overview of CRISP-DM activities ......................................................................... 26
Figure 2.7: Process of building a model and making predictions ................................................ 28
Figure 2.8: Simple linear regression example .......................................................................... 30
Figure 2.9: Various regression models ...................................................................................... 31
Figure 2.10: Decision tree with labelled nodes ......................................................................... 32
Figure 2.11: Differences between bagging and boosting ............................................................. 36
Figure 2.12: Process of creating a random forest model ............................................................. 38
Figure 2.13: Classification of factors affecting demand .............................................................. 40
Figure 3.1: Research Onion ....................................................................................................... 46
Figure 3.2: Research direction of induction and deduction methods ......................................... 49
Figure 3.3: Experimental Setup of this Research ..................................................................... 57
Figure 3.4: Process of creation of final dataset ......................................................................... 59
Figure 3.5: Histogram of Customer Visits ................................................................................... 63
Figure 3.6: Histogram of Rain .................................................................................................... 64
Figure 3.7: Histogram of Rain ..................................................................................................... 64
Figure 3.8: Histogram of Sunshine ............................................................................................ 65
Figure 3.9: Histogram of Windspeed ......................................................................................... 66
Figure 3.10: Histogram of Humidity ......................................................................................... 66
Figure 3.11: Scatter plot matrix of continuous variables .............................................................. 67
Figure 3.12: Correlation Coefficient of continuous variables ..................................................... 68
Figure 3.13: Boxplot of hourly customer visits .......................................................................... 69
Figure 3.14: Daily effect of on customer visits .......................................................................... 70
Figure 3.15: Holiday effect on customer visit ............................................................................ 70
Figure 4.1: Regression Tree ........................................................................................................... 77
Figure 4.2: Validation Error of regression tree .............................................................................. 77
Figure 4.3: Error Plot of Random Forest with 500 trees ................................................................. 79
Figure 4.4: Residual Plot .................................................................................................................. 81
Figure 4.5: Screenshot of Web Portal to generate predictions ......................................................... 81
Figure 5.1: MAPE of all models developed ..................................................................................... 86
TABLE OF TABLES

Table 1.1: Difference between Fast-casual and Full-Service Restaurants .......................... 3
Table 2.1: Impact of Forecasting on various areas of an organization .................................. 11
Table 3.1: Summary of research choices for this study .......................................................... 52
Table 3.2: Description of Variables ...................................................................................... 60
Table 3.3: Data Quality of Continuous variables................................................................. 61
Table 3.4: Data Quality of Categorical Variables ................................................................. 62
Table 3.5: Descriptive Statistics for continuous variables................................................... 62
Table 4.1: Regression Equation Details ................................................................................ 72
Table 4.2: Parameters linear regression .............................................................................. 76
Table 4.3: Regression Tree Parameters ................................................................................. 76
Table 4.4: Random Forest Parameters ............................................................................... 78
Table 4.5: Incnodepurity of Random Forest Predictors ......................................................... 80
Table 4.6: Parameters Gradient Boosting ............................................................................. 80
Table 5.1: RMSLE and MAPE values of all predictive models developed ......................... 84
Table 5.2: MAPE values with interpretation ........................................................................ 85
ABBREVIATIONS

ABT  Analytics Base Table
ARIMA  Auto Regressive Integrated Moving Average
CRISP-DM  Cross Industry Standard Process for Data Mining
HSE  Health Service Executive
MAE  Mean Absolute Error
MAPE  Mean Absolute Percentage Error
MSE  Mean Squared Error
RMSE  Root Mean Squared Error
RMSLE  Root Mean Squared Logarithmic Error
SAS  Statistical Analysis System
SEMMA  Sample Explore Modify Model Assess
SPSS  Statistical Package for Social Sciences
1 INTRODUCTION

1.1 Introduction

This chapter introduces forecasting in the Restaurant Industry, the problem this study looks to address and the rationale for undergoing this research. To guide the study hypotheses, aims and objectives are also discussed at a high level and then the chapter concludes with a description on the methodology and limitations of the study.

1.2 Background

A 2017 report by the Irish Food Board estimates Ireland’s Foodservice market to be worth 7.8 billion euros with limited service restaurants which includes Quick Service, Fast Casual, and “Food on the go” establishments accounting for 35% of the overall value. Fig 1.1 shows the market share percentage of the various segments of Ireland’s food service market.

![Figure 1.1: Valuation of Ireland's Foodservice Market](image-url)

Source: (Irish Food board, 2017)
Three major segments of the restaurant industry are described below:

- **Fast Food**: These restaurants are also referred to as quick service restaurants and they are establishments with a specific focus on speed of service. Jefferey et al. (2006) described fast food restaurants as outlets whose main characteristics are fast service, meals being paid for before eating, restricted menu and meal preparation options, no waiter and option to have the meal within the restaurant or take it out.

- **Fast Casual**: These restaurants are very similar to the fast food restaurants as the emphasis is on the speed of service and a limited menu, but they often offer a menu containing food made with better quality ingredients hence are a bit more expensive than fast food restaurants.

- **Full Service**: These restaurants often have menus with wide ranging options which includes appetizers, main courses, desserts and drinks. They also offer full table services with a number of wait staffs attending to customers.

For the purpose of this study the, Table 1.1 below highlights the key differences between fast-casual restaurants and full-service restaurants.
Fast Casual Restaurant | Full Service Restaurant
---|---
Advanced Booking | Some Fast Food Restaurants offer a delivery service where meals are ordered online or via telephone with expectation meal is delivered immediately | Tables can be booked well in advance
Meal Preparation | Bulk of meal preparation is done in advance with already processed meals that must be served to the customer or wasted. | Food is usually freshly prepared
Table Service | Customers approach the counter to order | Customers get seated and are approached by a waiter
Payment | Payment is made at the counter while ordering and before eating | Payment is collected by the waiter after the meal.
Menu | Usually Limited Menu | Menu is more robust

Table 1.1: Difference between Fast-casual and Full-Service Restaurants

1.3 The Restaurant

Boojum limited is a fast-growing fast-casual restaurant chain that specializes in high quality and affordable Mexican burritos. The first Boojum restaurant was opened in Belfast in 2007 and since then, the restaurant chain has grown to 15 restaurant outlets spread out across Ireland and Northern Ireland with outlets in Dublin, Limerick, Galway and Cork. All Boojum restaurants are open every day between 11:30am and 10pm though there are a couple of exceptions with one restaurant being closed on weekends and a couple of other restaurants closing later than 10pm. The restaurant
outlet used for this study follows the regular operation hours as it is open every day between 11:30am and 10pm.

Boojum Limited is very passionate about food and their business with their burritos gaining more popularity over the years. A review done by a leading Irish blog\(^1\) which compared burritos made from the top burrito restaurants in Ireland put Boojum burritos as the best burritos in Ireland in terms of ingredients combination, wrapping expertise and speed of service. Boojum Limited was also recognized as one of the best places to work in Ireland for the year 2018. In addition to that the management strives to differentiate from their competitors by offering very high-quality food at a reasonable price, they also offer their customers the option to customize their burritos as well as a click and collect service.

1.4 Research Problem

As the restaurant industry continues to grow, restaurant operators are under pressure to sustain their growth and increase revenue. A problem that most fast-casual restaurants face is in forecasting demand. To guarantee speed of service, fast-casual restaurants have to prepare meals in advance and meals not consumed within 90 minutes has to be discarded in order to adhere to food safety regulations by the HSE. Being able to accurately forecast number of customer visits per hour would help in reducing wastage and would have a direct positive impact on the revenue of the restaurant. Unlike other full-service restaurants, fast-casual restaurants operate under a higher level of uncertainty as most of the customers they serve daily come into the restaurant without any form of booking in advance.

According to Liu et al. (2001), the profitability of a restaurant depends on the ability of the manager to predict labour requirements, inventory levels and food preparation scheduling in an accurate and timely manner, any type of forecasting to be done by a fast-casual restaurant at the store level directly depends on the number of customers the restaurant is going to service during a particular time frame. Due to the perishable nature and short shelf lives of products sold by fast-casual restaurants, accurate demand forecasting becomes even more important.

\(^1\) http://www.dailyedge.ie/best-burrito-in-dublin-2480636-Dec2015/
Most researches into forecasting build models to predict monthly, quarterly or yearly demand; while this is important for long term planning, it is of no real use to the manager of a fast-casual restaurant for the immediate optimization of resources to achieve maximum profitability in the day to day operations of the fast-casual restaurant. Also forecasting demand in the form of expected revenue puts an extra layer of complexity in determining what the actual demand is going to be within a particular time frame, forecasting customer visits provides more actionable information for reducing wastage.

1.5 Forecasting vs Prediction

Prediction and forecasting are two words used frequently in the related research reviewed for this study to indicate the process of assigning a value to future demand. The difference and similarity between prediction and forecasting has been heavily debated on a popular analytical platform\(^2\) with most arguments reiterating the two words are more similar than different, A response from a user Ralph summarized both the similarity and differences of both words in one statement.

“Forecasting would be a subset of prediction. Any time you predict into the future it is a forecast. All forecasts are predictions, but not all predictions are forecasts, as when you would use regression to explain the relationship between two variables.”

Prediction in machine learning terms is simply the assignment of a value to a target variable (Kelleher et al.,2015) while forecasting deals specifically with future values and therefore labelling the research in this study as prediction or forecasting makes no major difference.

1.6 Research Question and Hypothesis

The aim of this research study is to identify the significant factors that affect the number of visits a fast-casual restaurant would get every hour and develop machine learning models capable of automatically predicting the number of customers that would be served each hour for which the restaurant is opened with a minimum amount of error. Mean Absolute Percentage Error (MAPE) is a measure which indicates the accuracy

\(^2\) https://www.analyticbridge.datasciencecentral.com/forum/topics/difference-between-prediction
of a forecasting model, Lewis (1982) suggests that a forecast with MAPE value less than or equal to 50% is a reasonable model. Therefore, to guide this study, the research problem can be formalized as a research question which is presented below:

“What are the significant factors that affect the number of hourly customer visits to a fast-casual restaurant and can these factors be used in building machine learning models to predict customer visits with a mean absolute percentage error of less than 50%?”

The following hypotheses can be deduced from the research question.

- Null Hypothesis 1 \((H_01)\): Day of the week, hour of the day, rain, temperature, humidity, windspeed and public holidays are not significant in predicting number of customer visits per hour to a fast-casual restaurant.

- Alternate Hypothesis 1 \((H_{a1})\): Day of the week, hour of the day, rain, temperature, humidity, windspeed and public holidays are significant in predicting number of customer visits per hour to a fast-casual restaurant.

- Null Hypothesis 2 \((H_02)\): Multiple linear regression, decision tree, random forest, and gradient boosting models would achieve MAPE of greater than 50% when applied to predict the number of fast-casual restaurant visits per hour in a dataset of historical visits.

- Alternate Hypothesis 2\((H_{a2})\): Multiple linear regression, decision tree, random forest, and gradient boosting models would achieve MAPE of less than 50% when applied to predict the number of fast-casual restaurant visits per hour in a dataset of historical visits.

1.7 Research Aims and Objectives

The main aim of this research study is to develop multiple linear regression, decision tree, random forest and gradient boosting models with input features of day of the week, hour of the day, rain, temperature, sunshine, windspeed, humidity and public holidays that can predict the number of customer visits per hour with MAPE of less than 50%.

The Objectives of this research study are:
• To critically review relevant literature and other independent sources for forecasting, predictive modelling and applicable supervised machine learning models.

• To obtain a dataset of historical hourly customer visits from a fast-casual restaurant in Dublin and merge the dates and time to historical weather data from Met Éireann.

• To identify and resolve any data quality issues found that would affect the outcome of the experiment.

• Carry out experiments to identify significant factors in predicting number of customer visits to the fast-casual restaurant per hour hereby testing hypothesis Ho1.

• Develop several predictive models using multiple linear regression, decision tree, random forest and gradient boosting algorithms.

• Evaluate the results by computing the mean absolute percentage for each predictive model and identifying the best model hereby testing hypothesis Ho2.

• Design and build a web app where the best predictive model identified would be deployed to display the predicted number of customer visits to a fast-casual restaurant for a particular hour.

• Identify the limitations of this research study and suggest areas of future research to build on this study.

1.8 Research Methodologies

The research methodology used in this is study is quantitative research which would involve conducting experiments on historical data obtained from a fast-casual restaurant to identify significant factors that affect the number of visits the fast-casual restaurant gets per hour and then building machine learning models capable of predicting the number of customer visits. Mean Absolute Percentage Error(MAPE) is
the measure that would be used to evaluate the machine learning models used and to identify which model performs best.

Microsoft excel which would be used for minor data wrangling and initial data exploration. SPSS would be used to explore the continuous variables individually and in combination with one another while R would be used for more advanced data wrangling and for building the machine learning models. Python would be used for deploying the best trained model.

This research study would follow the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology for data mining projects. CRISP-DM provides a very organized way for planning and achieving the objectives of a data mining project. CRISP-DM includes six phases which are business understanding, data understanding, data preparation, modelling, evaluation and deployment. Fig 1.2 shows the lifecycle of CRISP-DM.

![Figure 1.2: CRISP DM Lifecycle](Source: Wirth &Hipp,2000)
1.9 Scope and Limitations

The scope of this study is to build a machine learning model to predict the number of customer visits to a fast-casual restaurant in Dublin. Though sales amount is an important indicator of the demand of the restaurant, this study would not be forecasting sales amount but number of customer visits. The features that would be used in building the model are day of the week, hour of the day, rain, temperature, windspeed, humidity and public holidays. There are surely other factors like traffic, footfall, economic indicators, promotions and events that can also affect the number of customer visits to the fast-casual restaurant, but these other factors have not been included in this study due to lack of available data or complexities in obtaining the data.

The historical customer visits data used for this study is obtained from a single fast-casual restaurant in Dublin and spans over two years, this data was obtained directly from transactions on the till which represents customers that not only walked into the restaurant but actually made a purchase. Therefore, the scope of this research does not involve predicting the footfall to the restaurant but actual customer transactions. Historical weather data is obtained from Met Eireann and the weather station used is phoenix park which is about five kilometres away from the restaurant used for this study and as such might not reflect the actual weather conditions at the fast-casual restaurant, but it provides a reasonable estimate considering customers that patronize the restaurant come from different areas within Dublin.

1.10 Outline of the Thesis

Chapter two presents a comprehensive review of literature on approaches to forecasting in the restaurant industry, time series data, model evaluation and predictive modelling in general with a focus on multiple linear regression, regression tree, gradient boosting and random forest models.

Chapter 3 presents the design and methodology of the experiments that would be used in this study. The process and steps involved in creating the final dataset for this project including It details on the various features used for this experiment and how they are obtained is also discussed in this chapter. This chapter then explores the data to find patterns and potential data quality problems.
Chapter 4 discusses the implementation of the different experiments identified along with the results of each experiment. Multiple linear regression experiment is conducted to test for significant factors followed by a robust experiment developing several other machine learning algorithms to predict the number of hourly customer visits.

Chapter 5 discusses and analyses the results of the experiments carried out in chapter 4 by relating this to previous research in forecasting. The significant factors identified to have an impact on number of customer visits are discussed while the machine learning models built are analysed by comparing the Mean Average Percentage Error (MAPE) and the Root Mean Square Logarithmic Error (RMSLE) of each model.

Chapter 6 provides key summaries of the research, contributions to the area of forecasting for restaurants and limitations of the study. This chapter then concludes by providing areas for future research to be carried out.
2 LITERATURE REVIEW

2.1 Introduction

This chapter critically discusses research areas that are critical to this study. The chapter begins with a review of forecasting in the food service industry, the methods of forecasting and time series methods of forecasting. The second part of this chapter focuses on machine learning by beginning with a very broad introduction to data mining before a comprehensive review of the specific machine learning models that would be used in this study. The chapter concludes with a discussion on evaluation of forecasting models.

2.2 Forecasting

Forecasting is a critical tool for planning, strategy and every other form of decision making on future activities (Makridakis, 1988). Accurate forecasts provide valuable inputs into the major areas of a business such as marketing, sales, inventory and finance hereby helping the business to reduce expenses and maximize profitability. Rhamlow and Klimberg (2002) in their research estimated the impact forecasting has on key areas within an organization, this is shown in the table below:

<table>
<thead>
<tr>
<th>Area</th>
<th>Impact of forecasting (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Budgeting</td>
<td>85</td>
</tr>
<tr>
<td>Operation Decisions</td>
<td>67</td>
</tr>
<tr>
<td>Financial Decisions</td>
<td>65</td>
</tr>
<tr>
<td>Staffing</td>
<td>50</td>
</tr>
<tr>
<td>Contingency Planning</td>
<td>47</td>
</tr>
<tr>
<td>Investment Decisions</td>
<td>41</td>
</tr>
<tr>
<td>New Product</td>
<td>39</td>
</tr>
<tr>
<td>Ordering</td>
<td>31</td>
</tr>
</tbody>
</table>

Table 2.1: Impact of Forecasting on various areas of an organization

(Source: Rhamlow & Klimberg, 2002)
In the food service industry, forecasting demand is a very important first step in tackling problems like staffing and stock management which have a direct impact on the revenue of the restaurant within any period. Arunraj and Ahrens (2015) concluded in their research that food wastage and inadequate stocking is as a result of inaccurate forecasting. Inaccurate forecasts can lead to different problems, for example a restaurant manager that overestimates demand would schedule too much labour and order surplus food which would lead to excess costs. On the other hand, a restaurant manager that underestimates demand would schedule not enough labour and order inadequate food in anticipation of the demand, this would lead to slower service and potential loss of revenue. It is therefore important for restaurant managers that they are able to forecast demand as accurately as possible for reduce the cost of wastages and to maximize revenue.

Due to the perishable nature and short shelf lives of products sold by fast-casual restaurants like the restaurant considered in this study, accurate hourly demand forecasting rather than daily forecasting or weekly forecasting is even more important as the restaurant adheres to a food safety policy by which heated food not stored above 63°C must be consumed within 90 minutes or discarded (HSE, 2012). Over estimation of demand could lead to high amounts of wastage every hour while under estimation of demand could lead to reduced customer satisfaction.

### 2.3 Time Series Data

Time series is a collection of observations made sequentially through time (Chris, 2000). There are different examples of time series data like monthly sales of a product; the temperature at noon every day in Dublin and as in the dataset considered for this study, the number of customer visits to a fast-casual restaurant every hour. Fig 2.1 shows a plot of the Standard & Poor (S & P) 500 index for the U.S. stock market for 90 trading days starting on March 16 1999.
Figure 2.1: Time Series Plot of USA stock market on March 16 1999.

Source: (Chris, 2000)

Time series data often have several inherent patterns which are useful for analysing past behaviour and for generating forecasts, these patterns are described below:

- **Trend**: The time series data considered contains a trend if there is continuous increase or decrease in the demand over a time period. Trend can either be linear or non-linear.
- **Seasonality**: A time series data is said to exhibit seasonal traits if there are patterns that repeat at regular constant intervals like yearly, monthly or weekly.
- **Cyclic Behaviour**: This occurs when the data exhibits signs of increases and decreases at irregular intervals, cyclic behaviour is often discovered in time series data with minimum period of 2 years.

Fig 2.2 below from Hyndman and Athanasopoulos (2018) shows four different time series plots with various or no combinations of seasonality, trend and cyclic behaviour. The plot on the top left corner shows very strong annual seasonality together with some cyclic signs over 6-10 years while the plot on the top right corner exhibits no sign of seasonality but a very clear downward trend. The plot on the bottom left corner exhibits
strong seasonality as well as an upward trend while the plot on the bottom right corner exhibits no obvious seasonality, trend or cyclic behaviour.

Figure 2.2: Different time series plots showing various or no combinations of seasonality, trend and cyclic behaviour.

(Source: Hyndman & Athanasopoulos, 2018)

As would be seen later in this chapter, time series data forms the basis of time series forecasting models which require only historical data in a defined order to generate forecasts through extrapolation for the next forecasting period required.

2.4 Approaches to Forecasting

Forecasting can be broadly categorized into qualitative and quantitative forecasting. Quantitative forecasting uses historical data and normally involves developing a mathematical model to generate forecasts while Qualitative forecasting relies on expert judgement to gain insights into future demand (Meneghini et al., 2018). Some
forecasting approaches makes use of a combination of both qualitative and quantitative forecasting methods.

According to Reynolds et al. (2013), qualitative methods of forecasting have historically been the preferred option for the restaurant industry largely because restaurant managers lack the required skills and resources to carry out quantitative forecasting. Some studies (Johnson & Schmitt, 1974; Critchfield et al., 1978) found qualitative methods of forecasting to be better than quantitative methods. However there have been other studies (Armstrong, 1983; Lorek et al., 1976; Sanders & Manrodt, 2004) which presented results that showed quantitative methods of forecasting outperform qualitative methods.

Quantitative methods of forecasting can be further divided into time series and econometric modelling. Econometric modelling is a term used extensively in literature on forecasting (Cranage & Andrew, 1992; Kim & Gu, 2003; Reynolds et al., 2013) to describe a quantitative method of forecasting where regression equations are used to represent the relationship between several independent variables usually economic variables and a dependent variable to generate forecasts. In the research done by Kim and Gu (2003) and Reynolds et al. (2013), the independent variables considered were from an economic perspective e.g. unemployment rate, consumer price index etc., and was therefore appropriate to class the forecasting model used in the research as an econometric model. However, the research done by Kokkinou (2013) used independent variables such as temperature and day of the week but also classed the forecasting model as an econometric model. For clarity in this review, any quantitative method which involves expressing the relationships between a set of independent variables and the outcome to be predicted are referred to as Regression methods. There have been a few studies which looked to compare time series and regression methods of forecasting. Cranage and Andrew (1992) found time series box-Jenkins model to be better than regression models in forecasting the monthly sales for a restaurant though they acknowledged regression models respond more quickly to changing conditions. In contrast, the research done by Forst (1992) found a multiple linear regression model provided better results than a time series Box-Jenkins model in forecasting of a restaurant’s weekly sales.
Song and Li (2008) performed a comprehensive review of forecasting tourism demand. Out of the 121 empirical studies they reviewed, they found all the studies except 2 studies used the quantitative forecasting method. They also discovered that 72 of the studies used time series forecasting methods while 71 of the studies used different variations of regression forecasting methods. Another review paper by Li et al. (2005) provided a survey of 84 papers published between 1990 and 2004 and which used regression methods for predicting tourism demand, 22 papers they reviewed specifically compared the performance of several quantitative methods, they concluded that no single forecasting method consistently outperforms other methods.

There have been very limited research specifically into hourly prediction of demand for a fast-casual restaurant, several researches into forecasting suggest that the forecasts generated from their models can be easily adjusted to suit any level of aggregation required(Cranage & Andrew, 1992; Kokkinou, 2013). While this might be true for approaches where a forecast for a more granular duration (e.g. daily forecasts) is combined to form the forecast for a greater duration (e.g. weekly), it would be much more difficult and error prone to generate daily forecasts for example from a weekly forecast. Therefore, this study aims to address this gap in the literature by building a quantitative model to forecast hourly demand directly.

2.4.1 Qualitative Forecasting

Qualitative forecasting methods are largely based on subjective inclinations by the individual carrying out the forecasting. Qualitative techniques are generally used by restaurant managers because it’s quite easy and requires no significant tool, resource or time to use unlike quantitative forecasting. In some scenarios where there is no historical data, qualitative approaches to forecasting may become ideal.

The two common types of qualitative forecasting are executive opinion and the Delphi method with the latter being more popular. In the executive opinion method, the subjective views of managers or experts on forecasts are collected and averaged while the Delphi technique uses an approach where a panel of experts are interviewed independently, and forecasts are collated by an external body. Rowe and Wright (1999) performed a study where they applied Delphi methods to forecasting, their findings suggest that Delphi groups outperforms statistical methods.
The major limitations of qualitative forecasting are bias which is introduced because forecasts are based on subjective opinions which might not be accurate. Eroglu and Croxton (2010) affirm that the motivation and the personality of the individual performing the qualitative forecast are major sources of bias. Furthermore, Sanders and Manrodt (2003) acknowledge that individuals lack the capacity to comprehend and process significant amounts of information. Qualitative methods are also difficult to replicate, for example if an employee in charge of forecasting was not available then a different person would have no defined methodology to replicate the process used for generating forecasts.

A common use of qualitative methods of forecasting in the literature is for the judgemental adjustment of quantitative methods to create a hybrid model for the generation of forecasts. Webby and O’Connor (1996) suggest in their research that subjective adjustment involves using a statistical forecasting method to generate a forecast first before then adjusting it based on contextual factors, they propose the methods of integrating forecasts as: (i) model construction, (ii) combination of forecasts, (iii) subjective decomposition and (iv) subjective adjustment. Fig 2.3 below illustrates the process of subjective adjustment:

![Figure 2.3: Process of Subjective Adjustment](Source: Webby & O’Connor, 1996)

According to the research done by Davydenko and Fildes (2013), subjective adjustments should only be done to a quantitative forecast when it is necessary to consider contextual factors that could not be included in the quantitative generation of the forecasts. Lawrence et al. (2006) acknowledged the importance of human judgement in generating forecasts but also warned of the risk of the introduction of
bias. Armstrong (2006) reiterates the risk of bias introduced by subjective methods in scenarios where managers overestimate demand to motivate employees and sales people underestimating so they can easily achieve their targets.

2.4.2 Quantitative Forecasting
Quantitative forecasting methods involves using well-structured processes for analysing historical data to generate forecasts (Armstrong, 1983). Quantitative methods can easily be replicated with each replication generating similar or near similar forecasts. The more historical data available, the more accurate quantitative forecasts become. Quantitative forecasting methods can be further divided into regression and time series methods.

2.4.2.1 Regression methods
Regression methods involve generating forecasts based on the effect of several observable factors on the demand. Regression forecasting models use a set of regression equations to represent a regression relationship between the dependent variable and a set of independent variables (Reynolds et al. 2013). This makes it ideal for the application of supervised machine learning algorithms. The machine learning algorithms to be used for this study are reviewed later in this chapter.

Fildes (1985) states that regression methods are superior to other forecasting methods because they are typically aggregate linear or almost linear models with well-defined stochastic structures. In the simplest form, a regression equation links a dependent variable Yi linearly to a number of independent variables which are specified in the equation 2.1 below as:

\[ Y_i = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + e_i \]  

(2.1)

Where:

\( Y_i \): Dependent Variable

\( X_i \): Independent Variable

\( \beta_0 \) and \( \beta_i \): Coefficients of the equation

\( e_i \): Error Term
The coefficients of the equation 2.1 can be estimated by using the least squares approach to minimize the error. Fildes (1985) proposes the following order when developing a regression model: (i) define the system; (ii) adapt from theory-based model to data model; (iii) collect and refine data; (iv) specify the functional form and stochastic structure of the error term; (v) conduct misspecification tests; (vi) conduct specification tests; (vii) evaluate the effects of uncertain exogenous variables; (viii) compare ex post and ex ante forecasts with the base-line (naive) forecasting model; and (ix) use the model.

There have been several studies which used linear regression equations to generate forecasts. Ingenito and Trehan (1996) developed a model to forecast the GDP, they started with 34 independent variables but eventually settled on using just monthly employment and consumption data for their final model as addition of too much variables led to overfitting. Hua and Templeton (2010) also used a regression model to investigate key growth factors for restaurants in the USA using data from publicly traded restaurant organizations, they found that leverage, market share and annual changes in size were significant factors in forecasting the one-year head annual growth of the restaurant.

2.4.2.2 Time series methods

Time series method of forecasting involves generating forecasts solely based on the use of historical data available with the assumption that the trend present in the historical data would continue. Some of the commonly used time series forecasting methods used in the literature are naïve forecasting, simple average, moving average, simple exponential smoothing, holt’s linear trend method, holt winters and Autoregressive Integrated Moving Average (ARIMA)

2.4.2.2.1 Naïve Forecasting

This is a simple method in which the most recent observation is taken as the forecast of the period being considered. Naïve forecasting model is usually used as a baseline model which indicates the least performance a forecasting model should overcome. Naive forecasting assumes that the forecast of the next time is equal to the last observed period and this can be represented in the equation below:
\[ \hat{y}_{t+1} = y_t \]  \hspace{1cm} (2.2)

Where:

\( \hat{y}_{t+1} \): Forecast of next period

\( y_t \): Actual value of previous period

### 2.4.2.2 Simple Average Forecasting

In this method, the average of all observed periods is taken as the forecast for the next time period. In other words, the average of all points previously observed is taken as the forecast of the next time period and this can be represented in the equation below:

\[ \hat{y}_{t+1} = \frac{1}{t} \sum_{i=1}^{t} y_i \]  \hspace{1cm} (2.3)

Where:

\( \hat{y}_{t+1} \): Forecast of next period

\( t \): Total number of time periods

\( y_i \): Value of period \( i \)

### 2.4.2.3 Simple Moving Average

For large time series dataset with wide ranging time periods, taking the average of all time periods as done in the simple average forecasting method would not be ideal. The moving average method addresses this problem by taking the average of a window of previous time periods as the forecast of the next time period with the rationale that the average of recent time periods are better indicators for the next time period. Simple moving average method is widely used as it is easy to implement and provides more reasonable accuracy compare to the naïve and simple average forecasting methods. Miller et al. (1991) concluded in their research that the lesser time periods considered for the moving average, the more responsive the model becomes, they created and test a simple moving average model with \( n=3,7 \) and 14, the results showed the model with \( n=3 \) to be the best performing model. The equation for a simple moving average model
is presented in equation 2.4 below where the forecast for the next value in a time series is based on the average of a number ‘p’ of previous values.

\[
\hat{y}_l = \frac{1}{p} (y_{i-1} + y_{i-2} + y_{i-3} \ldots + y_{i-p}) \tag{2.4}
\]

Where:
\[
\hat{y}_l : \text{Forecast of next period}
\]
\[
P: \text{Number of previous periods to be considered}
\]
\[
y_{i}: \text{Value of period } i
\]

### 2.4.2.2.4 Weighed Moving Average

Another variation to the moving average time series forecasting model is the weighed moving average method. In contrast to the simple average method where each previous ‘p’ observations have equal weights. Weighed moving average assigns different weights to each observation based on its perceived importance and is represented in equation 2.5 below:

\[
\hat{y}_l = \frac{1}{p} (w_1 \times y_{i-1} + w_2 \times y_{i-2} + w_3 \times y_{i-3} \ldots + w_p \times y_{i-p}) \tag{2.5}
\]

Where:
\[
\hat{y}_l : \text{Forecast of next period}
\]
\[
P: \text{Number of previous periods to be considered}
\]
\[
y_{i}: \text{Value of period } i
\]
\[
w_p : \text{Weight assigned to period } p
\]

### 2.4.2.2.5 Simple Exponential Smoothening

This method combines the strengths of the simple moving average and weighed moving average methods by making use of all the data but applying different weights to each data point. Forecasts for the simple exponential smoothening method are calculated using weighted averages of the entire datasets where the weights assigned to each data point increases exponentially for data points more recent, the oldest observations in the
dataset have the smallest weights assigned to them to limit their impact on the generated forecast. Simple exponential smoothening for the next time period can be represented in equation 2.6 below:

\[ \hat{y}_{t+1} = \alpha Y_t + \alpha (1-\alpha)Y_{t-1} + \alpha (1-\alpha)^2Y_{t-2} + \ldots \]  

(2.6)

Where:

\( \hat{y}_{t+1} \): Forecast of next period

\( \alpha \): Smoothening Parameter

\( Y_t \): Value of Period t

The smoothening parameter \( \alpha \) has values between 0 and 1, when \( \alpha \) is equal to 1 then the simple exponential smoothening model becomes equivalent to the naïve mode. When \( \alpha \) is greater than 0.5, then higher weight is assigned to more recent observations and therefore the forecasting model is more responsive. However when \( \alpha \) is less than 0.5 then the model become less responsive as higher weight is assigned to farther observations which then reduces the effect of outliers on the generated forecast.

2.4.2.2.6 Double Exponential Smoothening

Most time series data have trend associated with it and so moving average and simple exponential smoothening methods does not work well with time series data that contains high variations. Trend is a pattern that is observed in the time series data over a period of time, the trend could either be increasing or decreasing. Double exponential smoothening method is a variation of simple exponential smoothening method that includes a smoothed trend factor (Miller et al., 1990), this is well suited for datasets with visible trend where demand is either decreasing or increasing and is given by equation 2.7 below:

\[ \hat{y}_{t+1} = l_{t-1} + T_{t-1} \]  

(2.7)

Where:

\( \hat{y}_{t+1} \): Forecast of next period

\( l_t \): Level at time t
$T_t$: Trend at time $t$

As can be seen from equation 2.7, the forecast generated from double exponential smoothening is a combination of the level and the trend at time $t$.

### 2.4.2.2.7 ARIMA Model

Autoregressive Integrated Moving Average (ARIMA) also known as the Box-Jenkins model makes use of auto regressive relationships (Andrew et al., 1990) and aims to describe correlations available in the data with each other. ARIMA models consider seasonality and stationarity available in the data and can be created using the following approaches proposed by Kokkinou (2013): (i) select a model based on the presence of stationarity and seasonality in the data; (ii) parameter estimation and (iii) model evaluation. Andrew et al. (1990) concluded in their research that the Box-Jenkins model is more effective for detecting long term trends while the exponential smoothening models are more effective for detecting short term trends. Shukla and Jharkharia (2013) also developed an ARIMA model to forecast the daily demand of vegetables in an Indian market for which they achieved a MAPE of 30%.

### 2.5 Knowledge Discovery

The ultimate aim of this research study is to extract valuable insights from data combined from various sources, this process can be referred to as knowledge discovery. Frawley et al. (1992) defined knowledge discovery as the ‘the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data’. Fayyad et al. (1996) proposed a framework called Knowledge Discovery in Databases (KDD) which revolves around the end to end process of extracting useful patterns from large databases, the KDD process includes five major steps which are: data selection, pre-processing, transformation, data mining and evaluation. De Martino et al. (2000) describes knowledge Discovery in Database (KDD) as a combination of different technologies such as data warehousing, database management, machine learning, visualization and decision support. An Overview of the KDD process is showed in Fig 2.4 below:
The objective of the KDD process is to get knowledge from data which usually happens in the data mining and evaluation phases, but these phases are only a small fraction of knowledge discovery project, other steps such as data processing and transformation are critical.

2.6 Data Mining

The Knowledge Discovery in Database (KDD) process presents data mining as a single phase in the knowledge discovery process though several researchers have used the terms data mining and KDD process synonymously (Cabena et al., 1997; Chapman et al., 2000; Kurgan & Musilek, 2006). However, data mining process must be managed as a combination of different steps to achieve best results. There are several methodologies for managing the end to end process of data mining which are explained in the sections below:

2.6.1 SEMMA

The SEMMA methodology is made up of five steps: sample, explore, modify, model and assess. It is a methodology developed by the Statistical Analysis System(SAS) institute and integrated into SAS tools like enterprise miner to provide sequential steps for data mining tasks. Steps of SEMMA are outline in Fig 2.5 below:
The major limitation of SEMMA is its attachment to a particular set of data mining tools by SAS and the absence of a step to use the knowledge discovered (Mariscal et al., 2010).

### 2.6.2 CRISP-DM

Cross Industry Standard Process for Data mining (CRISP-DM) defines a process model that provides a framework for data mining projects that is independent of both the industry and the technology used (Wirth & Hipp, 2000). A poll carried out by Kdnuggets\(^3\) shows CRISP-DM as the most popular methodology for datamining, analytics or data science projects. CRISP-DM breaks down the lifecycle of a data mining project into six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. From the CRISP-DM lifecycle diagram shown in Fig 1.2, the arrows between various phases show that the sequence of the phases is not fixed meaning there could be iterations between phases during a data mining project. The outer circle in the diagram with arrows also indicate that the entire data mining process is continuous as the deployment of a data mining project could unravel new business opportunities or problems and therefore trigger a new data mining project.

CRISP-DM describes detailed activities that have to be carried out for a typical data mining project. Fig 2.6 below shows an overview of CRISP-DM activities:

---

Wirth and Hipp (2000) gave an outline of each phase of the CRISP-DM lifecycle summarized below:

- **Business Understanding:** This is the first phase of CRISP-DM and it focuses on gaining a business perspective of the project objectives and converting this into a data mining problem definition.

- **Data Understanding:** This involves initial collection and exploration of the data. Data quality issues and some first insights into the data can also be identified here. The data understanding and business understanding phases are closely linked as defining data mining objectives requires some knowledge of the data that is available.

- **Data Preparation:** This phase covers all the tasks involved in creating the final Analytics Base Table (ABT) that would be used for creating models. Machine learning algorithms require data to be in different formats, all data pre-
processing steps carried out come under this phase. Other tasks include data cleaning, attribute selection and construction of new attributes.

- **Modelling**: In this phase several models are created and applied to the dataset, the parameters of each model are also tuned optimally to achieve the best results. The modelling phase is closely related to data preparation as several ideas or opportunities discovered during modelling would require going back to the data preparation stage to amend the analytics base table.

- **Evaluation**: This phase involves comparison of several models created in the modelling phase by not only considering a quantitative metric but also confirming the models achieve the business objectives identified in the business understanding phase.

- **Deployment**: In this phase, the knowledge derived from modelling and evaluation phases is presented in a way usable by the business. It might involve the simple task of generating a report or the more complex task of implementing a repeatable data mining process.

### 2.7 Predictive Analytics

Predictive analytics is the process of creating and using models that make predictions based on the patterns extracted from historical data (Kelleher et al., 2015). Predictive analytics is used widely across different industries where it helps to reduce churn, derive insights about customer behaviour, increase sales conversion and understand purchasing patterns (Linoff & Berry, 2011). Predictive analytics usually involves training a model on historical data in order to be able to make predictions, machine learning is used to train the models.

#### 2.7.1 Supervised Machine Learning

Machine learning is the automated procedure of extracting patterns from data (Kelleher et al., 2015). Machine learning can broadly be categorized into two: supervised machine learning and unsupervised machine learning. Supervised machine learning techniques involves training a model on a historical data set where the target variable to be predicted is known, this type of dataset is frequently known as labelled data. In contrast
unsupervised learning techniques have no labelled data. This research study aims to predict the number of customer visits to a fast-casual restaurant every hour based on a set of features and would therefore use supervised machine learning techniques. The steps involved in creating a model and making predictions is illustrated in Fig 2.7 below:

![Figure 2.7: Process of building a model and making predictions](Source: Kelleher et al.,2015)

### 2.7.2 Machine learning Predictions for forecasting

Several studies have applied regression methods to generate forecasts (Kokkinou, 2013; Reynolds et al.,2013) but very few studies have explored the potential of other machine learning algorithms different from simple linear regression for forecasting
with even fewer studies performing experiments to compare the forecasting performance of several machine learning algorithms. Yu et al. (2015) developed a support vector regression model to forecast the sales of newspapers and magazines, they concluded in their research that the support vector model outperformed the linear regression model. Kong and Martin (1995) also deployed a neural network model to forecast sales of a food product using data from a Victorian food wholesaler and found the neural network generated more accurate forecasts than the simple linear regression method. Therefore, it is worth exploring the potential of other machine learning algorithms in forecasting demand.

2.7.3 Machine Learning Algorithms for Regression
For machine learning problems where the target variable to be predicted is a continuous value as is the case in this study, the problem is termed a regression problem and machine learning techniques suited to predicting continuous values are used. For other problems where the target variable is a binary or categorical variable, the problem is termed a classification problem. The various regression models that would be used in this study are reviewed later in this chapter.

2.7.3.1 Linear Regression
The linear regression model assumes that the relationship between a target variable ‘y’ and a set of independent variables ‘x1, x2, x3……xk’ are linear. The linear regression model aims to discover the line of best fit for the distribution of the target variable with the other independent variables, this line determines the intercept and the coefficients of the independent variables which provides the equation to predict unknown instances of the target variable (Han et al.,2011). A simple linear regression model with only one independent feature can be represented as an equation shown in equation 2.8 below:

\[ Y = w_0 + w_1 x_1 + e \]  (2.8)

Where:

Y: Target, response or dependent variable

w0: Intercept
$w_1$: Slope  \\
$x_1$: Predictor or independent variable  \\
e: Error term

A graph showing an example of a simple linear regression is shown in Fig 2.8 below:

![Graph showing simple linear regression example](image)

*Figure 2.8: Simple linear regression example*  
(Source: Wang, 2016)

In most scenarios, there is usually more than one independent or predictor variables and therefore equation 2.8 can be modified to cater for more than one independent variable which becomes equation 2.9 shown below:

$$Y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n + e$$  \hspace{1cm} (2.9)

Where:

$Y$: Target, response or dependent variable  \\
$w_0$: Intercept  \\
$w_n$: Slope  \\
x_n: Predictor or independent variable  \\
e: Error term

The value of $w_n$ indicates the strength of the relationship between a particular independent variable $x_n$ and the target variable $Y$. A value close to 0 signifies a weak
relationship and a value much greater than 0 indicates a strong relationship to the target variable which could be positive or negative.

It is near impossible to create a regression model that would be a perfect fit. There are usually several lines that would fit the distribution of the target variable and the independent variables, it is therefore important to find the line with the best fit. A suitable approach for finding the line of best fit is to explore the residuals.

The plot to the left of Fig 2.9 shows different possible regression models capturing the relationship between rental price and size while the plot on the right shows the best regression model with the dotted lines indicating the residual errors.

![Figure 2.9: Various regression models](image)

(Source: Kelleher et al., 2015)

### 2.7.3.2 Decision Trees

Decision tree is a supervised machine learning technique which makes use of a branching method to split data and cater for every possible outcome, when this is visualized the split looks like a tree, hence the name decision trees. The operation of a decision tree is based on extracting rules from the data, it is these rules that gives the decision tree algorithm a ‘tree like’ shape when visualized. Decision trees mimic the human way of thinking which is visible in not only the results but also in all the decisions made in arriving at the result, this makes the decision tree algorithm easy to explain with simple words that all professionals can understand.
A decision tree begins with a root node, a couple of interior nodes depending on the number of input features and then terminates with leaf nodes. The connections between two nodes are called branches. The root node and each interior node indicates conditions that must be tested for each descriptive feature and the number of levels in the descriptive feature determines the number of branches from the root or interior node while the leaf node at the end gives the final prediction. Using a decision tree for making predictions involves testing the value of the independent variable at the starting node of the tree, the outcome of the test would then determine which of the interior nodes the process would go to, these tests would be repeated until the process reaches a leaf node at which point a prediction is made. (Kelleher et al., 2015). Fig 2.11 below from McTiernan (2016) shows a sample decision tree with labels for the different types of node.

![Decision tree with labelled nodes](Image)

*Figure 2.10: Decision tree with labelled nodes*

(Source: McTiernan, 2016)

The decision tree algorithm works by splitting the dataset based on select descriptive features, it does this by making two important considerations: The first consideration is to decide which of the descriptive features is the most informative for the root node and subsequent parent nodes while the second consideration is to decide how the dataset would be split into different branches. A popular approach for determining
which feature within a set of descriptive features is the most informative is by selecting 
the feature with the highest information gain. According to Kelleher et al. (2015), 
Information Gain is a measure of the decrease in the entire entropy of a set of instances 
which is achieved by testing on an input feature and can be computed using the 
following processes: (i) Calculate the entropy of the entire dataset with respect to the 
target variable. (ii) Take each descriptive feature, split the instances in the dataset using 
their values and the add the entropy of each subset (iii) Subtract the remaining entropy 
calculated in step (ii) from the entire entropy of the dataset calculated in step (iii) which 
would give the information gain illustrated in equation 2.10 below.

Entropy is the measure of the impurity of the elements contained in a set and the entropy 
of an entire dataset can be calculated using equation 2.10 below:

\[ H(t, D) = - \sum_{l \in \text{levels}(t)} (P(t = l) \times \log_2(P(t = l))) \]  

(2.10)

Where:

\( H(t, D) \): The entropy of the entire dataset \( D \) based on the target feature \( t \)

\( \text{levels}(t) \): The set of levels in the domain of the target feature.

\( P(t = l) \): The probability of a randomly selected instance having the target level \( l \).

Information gain of a descriptive feature, \( d \) can be obtained by subtracting the 
remaining entropy of that feature from the entropy of the entire dataset and which is 
represented in equation 2.11 below:

\[ IG(d, D) = H(t, D) - \text{rem}(d, D) \]  

(2.11)

Where:

\( IG(d, D) \): The information gain of a descriptive feature \( d \)

\( H(t, D) \): The entropy of the entire dataset based on the target variable \( t \)

\( \text{rem}(d, D) \): The remaining entropy of a descriptive feature \( d \)

Using information gain as a feature selection measures comes with some limitations, 
the major limitation being the preference of information gain to select a feature with 
many levels since that feature will split the data into a lot of sub sets which are likely 
to be pure sets without consideration for the relationship between the independent 
variable and the target variable (Kelleher et al., 2015). An approach to reduce this bias
is to divide the information gain of a particular descriptive feature by the information used to calculate the value of the feature, this measure is called information gain ratio and is given by equation 2.12 below:

\[ G(d, D) = \frac{IG(d,D)}{\sum_{l \in \text{levels}(d)}(P(d=l) \times \log_2(P(d=l)))} \] (2.12)

Where:

GR (d, D): The information gain ratio of a descriptive feature d
IG (d, D): The information gain of a descriptive feature d
P (d = l): The probability of a randomly selected instance having the level l of the descriptive feature

Note that the divisor of equation 2.12 is simply the entropy of the dataset with respect to the descriptive feature.

Yet another measure that can be used as an alternative to information gain is the Gini Index which computes how frequently the levels of instances of the target feature in a dataset would be misclassified if only the distribution of the target levels were used in making the predictions. Gini index is presented in the equation 2.13 below:

\[ \text{Gini}(t, D) = 1 - \sum_{l \in \text{levels}(t)} P(t = l)^2 \] (2.13)

Where:

Gini (t, D): The Gini index of a dataset with respect to the target feature t
P (t = l): The probability of an instance in a dataset having a target level l

Most of the discussions thus far has centred around using decision trees with a dataset that contains categorical descriptive and target features but in many real-world scenarios, the dataset may include continuous descriptive features or even a target continuous feature as is the case in this study. To include a continuous feature in a decision tree, a threshold within the range of the continuous feature has to be defined and this threshold would then be used to split the instances in the dataset depending on whether the value of the instance is greater or lesser than the specified threshold. A continuous feature would normally have a large range and it would be complicated to
test on each possible threshold. An approach to overcome this challenge is to sort the instances in the dataset in ascending order according to the value of the descriptive feature, the points at which the level of the target feature changes are considered as potential threshold points and the information gain or information gain ratio is calculated for all possible threshold points identified to determine the most optimal threshold point for splitting the dataset. Unlike categorical input features, continuous features can be included at different nodes down the path of the decision tree though they would use different threshold values at each point. (Kelleher et al., 2015).

From Kelleher et al. (2015), if the target variable to be predicted is a continuous value, then the decision tree is referred to as a regression tree. In making a prediction for a continuous target variable, the decision tree displays the average of the target variable for the instances that are in the leaf node after splitting using all the descriptive features in the dataset. The error of a regression tree when applied to make a prediction can be gotten by subtracting the mean of the instances at the leaf node from the actual value of that training instance, it is therefore more suitable to construct regression trees in a way that reduces the error between the predicted and actual value of the target feature. Instead of using entropy as a measure for selecting the best descriptive feature to split the dataset on, the decision tree algorithm when predicting a continuous feature uses a measure of variance to select the best descriptive feature which is presented in mathematically in equation 2.14:

\[
\text{var}(t, D) = \frac{\sum_{i=1}^{n}(t_i - \bar{t})^2}{n-1}
\]

Where:
(t, D): Variance of a dataset D at a node based on continuous target values t
n: The number of instances in D
\(\bar{t}\): Mean of the target variable in dataset D
t_i: Actual target value for each instance in D
2.7.3.3 Ensemble Models

While the focus of most machine learning algorithms is to develop and select the best model based on accuracy or other bias for a problem, aggregating the outputs of a set of models could provide more accurate results with the motivation being that combining a set of models could exploit the strengths and reduce the errors of the final prediction especially when each model generates predictions independently or sequentially as is the case with ensemble models. A machine learning model that consists of other models is referred to as a model ensemble.

There are two major approaches to developing ensemble models which are boosting and bagging. The boosting technique iteratively develops models which are then added to the ensemble with the newer models focusing more on instances that other models have either misclassified in the case of a categorical target feature or instances with large errors in the case of a continuous target feature. The iterative addition of models to the ensemble stops when a predefined number of models have been created.

Bagging also known as bootstrap aggregating uses a process where each model in the ensemble is trained by using a random sample where each random sample has the same...
size and where replacement is used when sampling which means one instance can be part of one or more samples. Each sample is referred to as a bootstrap sample and each bootstrap sample has at a model trained on it. The bagging technique is ideal for decision trees as decision trees are extremely sensitive to changes, a change in the dataset for example could mean different features are deemed more important with each different sample and therefore split on at the root node of the tree which would result in a different tree structure every time. Also, when bagging is used with decision trees, not only are the instances(rows) sampled, the sampling is extended to the features(columns) as well so that each bootstrap sample makes use of a random subset of the descriptive variables in the dataset (Kelleher et al., 2015). Fig 2.11 shows the major differences between bagging and boosting.

2.7.3.3.1 Random Forest
Random forest model is essentially a model ensemble as it consists of different decision trees. The sampling of the descriptive features in a dataset is referred to as subspace sampling and this further strengthens the independence of decision tree in the ensemble model. The random forest model is a combination of bagging, subspace sampling and decision trees and this model makes predictions by returning the class with the most votes as in the case of a categorical target feature or returning the median as in the case of a continuous feature. Fig 2.12 below shows the process of creating an ensemble model using bagging and subspace sampling.
2.7.3.3.2 Gradient Boosting

Unlike in the random forest model where predictions of each model are made independently, each model in the gradient boosting algorithm makes predictions in a sequential manner such that each subsequent model focuses on instances with large errors. The Gradient boosting algorithm is an ensemble model which generates predictions based on a group of weak prediction models which could be decision trees\(^4\). The gradient boosting algorithm uses an intuition whereby patterns found in residuals are leveraged repetitively to make a model with weak predictions much better. At the

\(^4\) https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d
stage where there is no more recognizable pattern of residuals, modelling should be discontinued in order to prevent overfitting.

2.8 Factors that affect Demand

In order to apply machine learning algorithms to predict the number of hourly customer visits to a fast-casual restaurant as is the case in this study, a set of descriptive features hypothesized to affect hourly demand must be gathered. Arunraj and Ahrens (2015) discussed several studies and the factors they considered to build a forecasting model to forecast daily demand of a product. Aburto and Weber (2007) considered climate, school vacation, festivals holidays, payments and price features in their Artificial neural network model to predict demand for a Chilean supermarket. Ramanthan and Muyldermans(2010) performed a detailed study on the factors that affect the sales of a top soft drink company in the United Kingdom, they considered factors such as promotion, promotion duration, promotion size, temperature, discount, festivals, ranks of products and holidays in understanding the underlying demand structures of different product families. In the same vein, Ali et al. (2009) examined the impact promotion features, percentage of discount and price in forecasting using a regression tree. Hasin et al. (2011) identified promotions, holidays, festivals, climate and brand loyalty as factors that affect demand, they used these factors as input into an artificial neural network which they used for forecasting. Žliobaitė et al. (2012) considered school vacation, festivals, holidays, pressure, temperature, rain and promotions as predictors for their forecasting model while Sharma and Sharma (2012) used temperature, holidays, sale of alternate products and day of the week as input features for the artificial neural network used in their study.

Emphasis of the importance of weather related features in forecasting is prevalent in the literature. In addition to some of the studies already discussed where weather related features were used, Mirasgedis et al. (2014) created a forecasting model which included meteorological features such as sunshine duration, relative humidity, precipitation and temperature, other features used where tourist activity, month of the year, season of the year and festivals. Agnew and Thornes(1995) also investigated the impact of weather on the food and drinks industry in the United Kingdom, they divided the factors that affected demand into internal and external factors, the internal factors considered were advertisement campaigns, promotions, changes in retail outlets and management
structures, merchandizing and changes in price while external factors considered were changes in weather conditions, economic variables, political condition, social trends and holidays.

Fig 2.13 below is from Arunraj and Ahrens (2015) and it shows the classification of demand influencing factors. Weather and seasonality factors form the basis of the factors that would be considered in building models for this current study.

![Classification of factors affecting demand](image_url)

(Source: Arunraj & Ahrens, 2015)

### 2.9 Evaluating Forecasting models

After building a model to forecast future demand, it is important to assess the accuracy of the forecasting model to decide if the model is good enough for the business objective identified. Evaluation of a model is usually done by computing a metric which assesses how accurately predictions made by the model matches up to the actual value of the target variable being predicted. Evaluating a model usually involves splitting out the data available into training and test set which is also known as an out of sample set, the aim of this split is to prevent selecting a model which has been overfitted to the data and also to pick the model with the best chance of generalizing well to unseen data when the model is deployed. There are several metrics used for evaluating the accuracy of a forecasting model and they can be broadly categorized into scale dependent and scale independent metrics. Scale dependent metrics include Root Mean Square Error, Mean Squared Error and Mean Absolute Error and are in the same
unit as the target variable to be measured while the scale independent metrics include R squared, Root Mean Squared Logarithmic Error and Mean Absolute Percentage Error.

2.9.1 Mean Squared Error (MSE)

The Mean Squared Error (MSE) is a metric which is calculated by subtracting the predicted values from the actual values, squaring the values, adding this up for all instances and then dividing it by the number of instances in the dataset. The Mean Square Error measures the average difference between the actual and predicted values and it is presented mathematically in equation 2.15. The Mean Squared Error is also scale dependent and therefore its value is in the same unit as the target variable, but it is sometimes still difficult to interpret the values gotten from MSE in relation to the target variable due to the squared difference used in calculating it. MSE takes on values between 0 and infinity with smaller values indication better overall model performance.

\[
\sum_{i=1}^{n}(t_i - M(d_i))^2 \over n \quad (2.15)
\]

Where:

\( t_i \) : Actual value of an instance

\( M(d_i) \) : Predicted value made by the model

\( n \) : The number of instances in the dataset being evaluated

2.9.2 Root Mean Absolute Error (RMSE)

Root Mean Square Error (RMSE) values are also in the same unit as the target variable just like the MSE but aims to offer a metric more closely interpretable to the unit of the target variable by introducing a square root to counter the squared error of the MSE. However, by introducing the squared root, instances with large errors are penalised hereby causing an impact on the final RMSE value obtained for the model. Equation 2.16 shows the formula for RMSE.

\[
\sqrt{\sum_{i=1}^{n}(t_i - M(d_i))^2 \over n} \quad (2.16)
\]

Where:

\( t_i \) : Actual value of an instance
M(dᵢ) : Predicted value made by the model

n: The number of instances in the dataset being evaluated

### 2.9.3 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) aims to counter the short comings of both the MSE and RMSE by taking out the squared term and squared root. MAE uses an absolute value of the difference between the actual and predicted values instead which disregards the negative signs of this difference. The mathematical representation of the MAE is shown in equation 2.17.

\[
\sum_{i=1}^{n} \frac{\text{abs}(tᵢ-M(dᵢ))}{n}
\]  

(2.17)

Where:

\( tᵢ \): Actual value of an instance

M(dᵢ) : Predicted value made by the model

n: The number of instances in the dataset being evaluated

abs: Absolute Value

### 2.9.4 R² Coefficient

Sometimes it is better to have a measure that is independent of the scale of the target variable in cases where the individual developing or evaluating the models has no domain knowledge to interpret results in terms of the target variable and also to have a universal metric with a standard range to compare machine learning models. The R² Coefficient is a scale independent measure which compares how well a model performs on a test set over a baseline model which returns the average prediction of the test set as prediction for each instance. The formula for calculating the R² coefficient is given in equation (2.18)

\[
1 - \frac{\frac{1}{2} \sum_{i=1}^{n} (tᵢ-M(dᵢ))^2}{\frac{1}{2} \sum_{i=1}^{n} (tᵢ-tm)^2}
\]  

(2.18)

Where:

\( tᵢ \): Actual value of an instance

M(dᵢ) : Predicted value made by the model
n: The number of instances in the dataset being evaluated

t_m: The mean value of the dataset

2.9.5 Mean Absolute Percentage Error (MAPE)
The Mean Absolute Percentage Error (MAPE) is also independent of the scale of target variable and it expresses the difference between the actual and predicted values as a percentage with lower percentages indicating a good model. The MAPE is a very popular metric used for evaluation of forecasting models specifically as it provides an even easier way to express the performance of a forecasting model. The MAPE is calculated by calculating the absolute value of the percentage difference between the actual and predicted values and then dividing it by the total number of instances. The MAPE can be expressed in Equation 2.19 given below:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{abs(t_i-M(d_i))}{abs(t_i)} \times 100
\]

(2.19)

Where:

t_i: Actual value of an instance
M(d_i): Predicted value made by the model
n: The number of instances in the dataset being evaluated
abs: Absolute Value

2.9.6 Root Mean Square Logarithmic Error (RMSLE)
The Root Mean Square Logarithmic Error (RMSLE) is computed by taking the log of the difference between the actual values and the predictions made by the model. Unlike the RMSE, RMSLE does not penalize huge difference between the actual and predicted values and it is also independent of the scale of the target variable. RMSLE can be calculated using Equation 2.20 below:

\[
\sqrt{\frac{\sum_{i=1}^{n} (\log(t_i + 1) - \log(M(d_i) + 1))^2}{n}}
\]

(2.20)

Where:

t_i: Actual value of an instance
M(d_i): Predicted value made by the model
n: The number of instances in the dataset being evaluated

log: Logarithmic Value
3 DESIGN/METHODOLOGY

3.1 Introduction

As seen from the literature review, quantitative methods of forecasting have been the dominant method over the past decade and therefore would form the basis of the design of this research. The main aim of this study is to identify the significant factors that affects the number of customer visits to a fast-casual restaurant and to use these features to build a machine learning model capable of predicting the number of customer visits with mean absolute error less than 50%, this chapter presents the design of the experiment and the methodology used to achieve this objective. The research options based on the research onion identified for this study and rationale for selecting each option are discussed. The data mining elements of this research makes it an ideal candidate for the use of the CRISP-DM methodology which offers a robust and structured methodology for data mining projects, the design of this research according to each phase is also discussed. This chapter concludes with a detailed description and exploration of the data used for this study.

3.2 Research Onion

The research onion was proposed by Saunders et al. (2007) and it shows the different stages that a standard research would go through as shown in Fig 3.1. Each layer of the research onion provides description on a more comprehensive stage of the research process (Saunders et al.,2007). On the outside of the research onion are the broad research philosophies which are epistemology, ontology and axiology which then dissolves to more specific philosophical stances in the first layer.
The second layer contains the research approach to be used while in the third layer, a research strategy is selected and specific research choices suitable for the research would be identified. Time frame considered for the research study is identified in the fifth layer while the sixth and final layer presents the data collection methodology.

3.2.1 Philosophies (Outer Layer)

Epistemology, ontology and axiology guide all the layers of the research onion from outside the layer itself.

Epistemology focuses on facts regarded as acceptable knowledge within a field after exhaustive tests have been carried out. It is most suited to scientific research since facts can only be accepted when proved beyond reasonable doubt.

Ontology focuses on the nature of the reality and researchers using ontology reflect their opinion on the way the world works, how the society is developed and how this affects all other things. Ontology examines the differences between reality, what we think is reality and how this affects the behaviour of individuals.
Axiology Philosophy enables the researcher to gauge and recognize the personal bias that might be introduced in the review and analysis of the research instead of trying to ignore the bias or making efforts to counter balance the effect of the bias. How the bias shaped the research study and its impact on the results needs to be documented.

3.2.1.1 Selected Philosophy - Epistemology
As this research study falls under the scientific field, epistemology is the most suitable choice. There are already known facts about the approaches to quantitative forecasting and this study aims to apply some of the quantitative techniques to a specific problem in order to uncover new insights and information.

3.2.2 Philosophical Stance (Layer 1)
The philosophical stance provides justification for how the research will be undertaken and also contains principles about the type of problem being investigated. The different research philosophical stances under epistemology are positivism, realism and interpretivism.

Positivism is a philosophical stance which is derived from epistemology. In the positivism research philosophy, research questions and hypotheses that can be measured and tested against established theories are derived, there is an emphasis on replication of the research by other scientists and it is well suited to researches that involve statistical techniques.

Realism is a philosophical stance also derived from epistemology and closely related to positivism, but realism acknowledges the imperfections of scientific methods and encourages continuous research with the believe that some theories can be revised. Realism usually involves the use of several research methods in search of an outcome that is more reliable.

Interpretivism is concerned with techniques which focus on individual’s involvement in social and cultural life and therefore under this philosophical stance, researchers have to interpret certain aspects of the study leading to human interest. Data collection in this philosophical stance is mostly through interviews and observations and there is therefore a preference for qualitative methods over quantitative methods.
3.2.2.1 Selected Philosophical Stance – Positivism

This study would make you use of the positivism philosophy as the hypotheses that some factors are significant factors in predicting number of customer visits to a fast-casual restaurant and that machine learning algorithms can predict the number of customer visits with MAPE of less than 50% can be tested using statistical methods.

3.2.3 Research Approach (Layer 2)

In the deductive approach, the researcher begins with a hypothesis and sets out to carry out experiments to gather evidence to either accept or reject the hypothesis. The hypothesis can either be gotten from theory or some logical assumption in the particular field of study. To summarize, for the deductive approach, the researcher moves from theory to hypothesis then to observation and finally confirmation.

In contrast to the deductive approach, the researcher moves in the other direction usually to create new theory. So, the researcher begins with a research question then moves on to observation and then to description and analysis before a new theory is eventually created. The inductive approach is therefore more suitable approaches for fields with very sparse researches and theories. Fig 3.2 shows the evolvement of the deductive and inductive approach.
3.2.3.1 Selected Research Approach - Deductive

For this study, the deductive research approach is well suited as the hypotheses are derived based on already existing theories and then a research approach to test the hypotheses is generated (Silverman, 2013). The deductive approach particularly fits well with the positivism research philosophy which specifies the definition of a research question or hypothesis at the beginning of the study and the use of statistical techniques to test the hypotheses (Sneider & Larner, 2009).

3.2.4 Research Strategies (Layer 3)

The research strategy details how the researcher intends to carry out the tasks of the research (Saunders et al., 2007). The different types of research strategies are experiment, survey, case study, action research grounded theory, ethnography and archival research.

According to Saunders et al. (2007), experimental research involves developing a research procedure which aims to examine results of an experiment against expected results and usually considers a fixed set of factors. Experiments test the causal effects of independent variables on a dependent variable and have a well-defined structure which enables replication.
Survey makes it possible for a researcher to collect a large volume of data to analyse several research questions and it usually involves taking a representative sample of the population (Bryman, 2012). Surveys can either produce qualitative data or quantitative data which can be statistically analysed for causative variables.

The case study research strategy involves deep analysis of one or more cases so as to identify important features and extrapolate generalizations. Case study also identifies and acknowledges the impact culture and context has on the difference between two or more cases (Silverman, 2013).

The action research strategy takes a practical approach to providing a solution to a specific problem within a particular field. There is a high degree of collaboration between the organization and the researcher in searching for the most appropriate solution to the problem. Action research begins with defining a clear objective then diagnosing the problem before identifying several actions to solve the problem and finally confirming this corresponds to the best approach.

The grounded theory strategy makes use of inductive methods to derive patterns from data as a pre-requisite for the study. Data collection is usually by interviews with several qualitative techniques like transcribing, coding and grouping applied to the data. Grounded theory is popular in the social sciences.

The roots of ethnography can be traced to anthropology which involves studying other people from a detached view although in ethnography the researcher is required to take an immersive approach into the community or case being researched. The main aim of ethnography is to gain an understanding on the differences meaning and importance from the perspective of the community.

Archival research strategy focuses on collecting data from already existing datasets or other archived documents (Flick, 2011) which enables the researcher to deeply explore and analyse changes in a particular field over a period of time in order to identify the current state of the art in solving a particular problem.

3.2.4.1 Selected Research Strategy - Experimental

The experimental research strategy involves analysing a set of causal variables and setting up a rigid experiment that can be replicated and is thus ideal for this study as this study involves testing the significance of several factors such as day of the week,
hour of the day, rain, temperature, sunshine, windspeed and humidity and public holidays in predicting the number of customer visits to a fast-casual restaurant and building a predictive machine learning model with the features.

3.2.5 Research Choices (Layer 4)
Research choices specifies how qualitative or quantitative methods are to be used on the research. Qualitative and Quantitative methods are not explicitly indicated on the research onion but it a fundamental consideration in every layer of the research onion. The research choices include the mono method, mixed method and the multi method. As hinted by the name, the mono method involves using either qualitative or quantitative data and not a combination of both while the multi method research choice makes use of multiple quantitative or qualitative data but only one choice. Lastly, in the mixed method, the researcher uses quantitative and qualitative methods in both data collection and for carrying out analysis in the research.

3.2.5.1 Selected Research Choice- Mono Quantitative
As the current research study would involve only collecting quantitative data on customer visits, weather and holiday before making use of quantitative methods to statistically analyse the data, the mono quantitative method is the most suitable research choice.

3.2.6 Time Horizon (Layer 5)
The time horizon refers to a time frame within which data is to be collected for the project (Saunders et al.,2017). There are two types of time horizon mentioned in the research onion which are cross sectional and longitudinal. Cross sectional design involves gathering data at a specific point in time, this is suitable for scenarios where studying a particular phenomenon at that point in time forms the basis of the study while longitudinal time horizon focuses on collection of data on specific variables repeatedly over an extended period of time with a particular interest in how the variable changes over time.

3.2.6.1 Selected Time Horizon - Longitudinal
The data used for this study is time series data which includes the number of customer visits to the fast casual every hour from the beginning of 2016 to the end of 2018. Therefore, the time horizon considered for data collection is longitudinal.
3.2.7 Data Collection and Analysis (Layer 6)

Data collection is an important part of any research initiative and the approach of data collection is largely dependent on the choices made by the researcher in previous layers of the research onion.

Data collected for research can be divided into two broad categories: primary and secondary data. Primary data is data which is obtained directly from the source and on which analysis is going to be carried on directly while secondary data constitutes data that has been gotten from other researchers or data that has already been statistically analysed.

3.2.7.1 Data Type – Primary and Secondary

The data on number of customer visits per hour obtained from the restaurant can be considered as primary data as it was obtained directly from the source without any form of prior analysis, however the weather data obtained from Met Eireann can be categorized as secondary data.

3.2.8 Summary Research Onion

Table 3.1 shows summary of the selected research options for this study according to the research onion.

<table>
<thead>
<tr>
<th>Research Onion layers</th>
<th>Selected Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philosophy</td>
<td>Epistemology</td>
</tr>
<tr>
<td>Philosophical Stance</td>
<td>Positivism</td>
</tr>
<tr>
<td>Approach</td>
<td>Deductive</td>
</tr>
<tr>
<td>Strategy</td>
<td>Experimental</td>
</tr>
<tr>
<td>Choices</td>
<td>Mono Quantitative</td>
</tr>
<tr>
<td>Time Horizon</td>
<td>Longitudinal</td>
</tr>
<tr>
<td>Data Type</td>
<td>Primary and Secondary</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of research choices for this study
3.3 Research Design by CRISP-DM

Crisp-DM is a robust methodology for the end to end management of data mining projects which makes it ideal for this research. It is important to note that though the phases of CRISP-DM are sequential, they are not fixed and it's usual to move between phases, for example while modelling it might become necessary to convert the dataset which would be a task under the data preparation phase.

The activities to be carried out in each CRISP-DM phase for this research study is briefly explained below:

- **Business Understanding:** This is derived from a comprehensive review of forecasting and its benefits as well as interaction with management of the fast-casual restaurant on the major problem of inadequate forecasting which leads to high wastages per hour as they have to

- **Data Understanding:** Dataset of hourly customer visits for two years obtained from a branch of Boojum Ltd, a fast growing fast-casual restaurant in Dublin. An initial exploration of the dataset would be done to identify potential data quality problems and gain an understanding of the breadth of analysis that can be done.

- **Data Preparation:** In this phase, the dataset would be further processed. Continuous variables would be normalized, and Categorical variables would be one hot encoded. Feature engineering would be done to create new attributes that would enhance the performance of the model. The processed dataset would then be split into train and test sets.

- **Modelling:** Multiple Linear Regression, Regression tree, Random Forest and Gradient Boosting models would be trained with the train dataset. Different parameters would be used for each model and wherever possible cross validation techniques would be used.

- **Evaluation:** All the models built would be evaluated based on the Mean Average Percentage Error(MAPE) and Root Mean Square Error(RMSLE) of
each model when applied to predict number of customer visits in test set.

- Deployment: After the models have been evaluated, the model with the best performance would be deployed to a web app to demonstrate how forecasts can be generated.

### 3.4 Experimental Set up

To address the aim of this study which is to identify the significant factors that affect the number of visits a fast casual restaurant gets and to build several predictive machine learning models to predict the number of visits, two separate experiments would be carried out in order to gather enough evidence to either accept or reject the hypotheses identified at the beginning of this experiment. Before carrying out both experiments, an exploratory analysis of the entire dataset would be done to uncover and fix potential data quality issues as well as gaining initial insights into the data set using data visualizations.

#### 3.4.1 Experiment 1 to identify significant factors.

This experiment is to test for alternate Hypothesis 1 which states that:

"*Day of the week, hour of the day, rain, temperature, windspeed, humidity and public holidays are significant in predicting number of customer visits per hour to a fast-casual restaurant.*"

The experiment that would be carried out to test for hypothesis is by applying a multiple linear regression algorithm to the entire dataset.

#### 3.4.2 Experiment 2 to develop predictive models

This experiment is to test for alternate Hypothesis 2 which states that:

"*Multiple linear regression, decision tree, random forest and gradient boosting models would achieve mean absolute percentage error of less than 50% when applied to predict the number of fast-casual restaurant visits per hour in a dataset of historical visits.*"
For this experiment, the dataset would first be split into train and test sets in the ratio 70% to 30%. The train sets would be used in developing and training a multiple linear regression, decision tree, random forest and gradient boosting model. The parameters of each model would be tuned, and the best version of each model would be selected.

### 3.4.3 Deployment

In order to demonstrate predictive model identified to have the best RMSLE and MAPE values would then be deployed to a web app which would generate predictions based on values assigned to the predictors used for building the predictive model.

### 3.4.4 Evaluation

The results of each experiment would be analysed to decide on if there is enough evidence to accept or reject the null Hypothesis.

#### 3.4.4.1 Evaluation Experiment 1

The probability value(p-value) of significance for each dependent variable obtained from the results of the multiple linear regression algorithm applied to the entire dataset would be used for evaluating the significance of each dependent variable in relation to the target variable; number of customer visits. A probability value less than 0.05 indicates a significant relationship between that factor and the target variable and hence alternate hypothesis 1 would be accepted if all the factors have a probability value of less than 0.05.

#### 3.4.4.2 Evaluation Experiment 2

After all the predictive models have been developed and trained using 70% of the data, each model would then be applied to a test set to generate predictions for all the instances in the test sets. The predictions generated by the model would then be compared to the actual number of customer visits using RMSLE and MAPE metrics. As discussed in Chapter 2, the MAPE is a scale independent evaluation metric which
is widely used for evaluating forecasting models because of its easy interpretation and is selected for this study in order to compare with previous research in the field of forecasting. RMSLE is a popular metric used for most regression-based machine learning on Kaggle\(^5\) which has a large collection of machine learning projects set up as competitions. RMSLE would also be computed for each model in order to compare this project with other similar projects on Kaggle.

\(^5\) [https://www.kaggle.com/](https://www.kaggle.com/)
3.6 Tools

Several tools and software would be used in this research and this can be broken down into Data Wrangling, Data Exploration, Modelling and Deployment.

3.6.1 Data Wrangling

Data Wrangling is the process of cleaning, transforming and aggregating raw data into a desired format which is suitable for analysis. As the final dataset used for this study contains data from different sources and in different formats, data wrangling is an important part of this research. Microsoft Excel provides a quick interface to quickly inspect the available data and fix minor data issues while R\(^6\) which is an open source

---

\(6\) https://www.r-project.org/
programming language very popular in the statistics community would be used for the more sophisticated data wrangling tasks.

### 3.6.2 Data Exploration

Data Exploration is usually the first step in building machine learning models as it enables the researcher to gain insights about the data and potential data quality problems usually using data visualizations. SPSS\(^7\) is a statistical software package that provides an interactive interface for carrying out analysis on data without the need for coding. SPSS would be used to analyse the continuous variables in the dataset both independently and in relation with each other. R through the ggplot2\(^8\) package would be used for a more customized analysis using boxplots to visualize the effect of categorical variables like day and hour on the number of categorical variables.

### 3.6.3 Data Modelling

This is where the actual development and training of the machine learning models is done. Most of the model building would be done in R. Multiple linear regression algorithm is available by default in R but the other machine learning models would be done using some packages such as rpart\(^9\) for decision trees and randomForest\(^10\) for the random forest model. As model deployment would be through python, the GradientBoostingRegressor\(^11\) from scikit learn for python would be used for developing the gradient boosting algorithm. Python\(^12\) is a general-purpose programming language with a growing popularity in the analytics community largely because of its versatility and ease of use.

### 3.6.4 Deployment

Deployment would involve demonstrating how a selected machine learning model can be used to generate predictions. Flask\(^13\) is a web development framework for python which makes it possible to deploy a machine learning model onto a web app.

---

7 https://www.ibm.com/analytics/spss-statistics-software
8 https://cran.r-project.org/web/packages/ggplot2/index.html
9 https://cran.r-project.org/web/packages/rpart/index.html
10 https://cran.r-project.org/web/packages/randomForest/index.html
12 https://www.python.org/
13 http://flask.pocoo.org/
3.7 Dataset

The dataset used from this study involves combining various data from different sources. First hourly customer visits data for 2 years (2016 and 2017) was obtained directly from a fast-casual restaurant in Dublin then historical weather data is to match the time and date from the hourly visits data was obtained from Met Eireann\textsuperscript{14} and finally data on public holidays was obtained online from a public data source\textsuperscript{15}. Fig 3.3 below shows the process of creating the final dataset used in this study.

![Diagram](image)

*Figure 3.4: Process of creation of final dataset*

The final dataset for this study contains 7942 rows and Table 3.2 gives the description of each variable.

\textsuperscript{14} https://www.met.ie/
\textsuperscript{15} https://www.officeholidays.com/countries/ireland/2016.php
<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer_Visits</td>
<td>This is the target variable and is a continuous value. It refers to the number of customers that visited the fast-casual restaurant and completed a purchase.</td>
</tr>
<tr>
<td>Hour</td>
<td>This is a categorical variable with 11 levels ranging from 11am to 9pm. Each hour represents sales made from the beginning of that hour to within a minute of the next hour so for example the 11am level represents sales made between 11:00 and 11:59.</td>
</tr>
<tr>
<td>Day</td>
<td>This is a categorical variable with 7 levels representing every day of the week: Sunday, Monday, Tuesday, Wednesday, Thursday, Friday and Saturday.</td>
</tr>
<tr>
<td>Holidays</td>
<td>This indicates that a specific day and by extension, a specific hour is a holiday. Important to note that some days like Father’s Day and mother’s days have also been included even though these days are not public holidays. Holidays has just two levels: Yes and No</td>
</tr>
<tr>
<td>Precipitation</td>
<td>This is a continuous value and is measured in millimetres and it indicates the amount of precipitation for a specific hour.</td>
</tr>
<tr>
<td>Temperature</td>
<td>This is a continuous variable measured in degree Celsius and it represents the air temperature for a specific hour.</td>
</tr>
<tr>
<td>Sunshine</td>
<td>This is a continuous variable measured in hours and it represents the number of hours of continuous sunshine for a specific hour.</td>
</tr>
<tr>
<td>Windspeed</td>
<td>This is a continuous variable measured in knot and it represents the mean wind speed for a particular hour.</td>
</tr>
<tr>
<td>Humidity</td>
<td>This is a continuous variable measured in percentage and it represents the relative humidity of a specific hour.</td>
</tr>
</tbody>
</table>
3.8 Data Quality

The data was inspected for potential data quality issues and to give initial insights into the dataset. A data quality report was generated using the data quality package in R. Table 3.3 shows details on the continuous variables while Table 3.4 shows details of the categorical variables. It can be seen from both Table 3.3 and Table 3.4 that there are no missing values. From Table 3.3, the minimum value for rain, temperature, sunshine and customer visits is 0, this could be considered normal for all these variable except for the target variable customer visits as it means there were no customers within that time frame. Consultation with the manager of the restaurant where the data was obtained confirmed this was a very rare occurrence and usually happened when the store had to close due to technical faults, therefore instances where the value for number of customer visits is equal to 0 were considered outliers and were removed from subsequent analysis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>non-missing</th>
<th>missing</th>
<th>missing percent</th>
<th>unique</th>
<th>mean</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>0.08</td>
<td>0</td>
<td>13.2</td>
</tr>
<tr>
<td>Temperature</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>256</td>
<td>11.36</td>
<td>0</td>
<td>26.1</td>
</tr>
<tr>
<td>Sunshine</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0.23</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Windspeed</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>11.51</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>Humidity</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>65</td>
<td>76.21</td>
<td>35</td>
<td>100</td>
</tr>
<tr>
<td>Customer_Visits</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>129</td>
<td>51.43</td>
<td>0</td>
<td>169</td>
</tr>
</tbody>
</table>

Table 3.3: Data Quality of Continuous variables

From Table 3.4, all the categorical variables appear to have no problems, the variable DateTime acts as a unique identifier and is equal to the total number of instances in the dataset as expected while the variables Hour and Day rightly correspond to 11 and 7 which is the number of hours the restaurant is open for everyday and the number of days in a week respectively. Holiday correctly has two unique values indicating if the particular hour is a holiday.
### Table 3.4: Data Quality of Categorical Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>n.non.miss</th>
<th>n.miss</th>
<th>n.miss.percent</th>
<th>n.unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>DateTime</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>7942</td>
</tr>
<tr>
<td>Hour</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Day</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Holiday</td>
<td>7942</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

3.9 Data Exploration

The final dataset contains 9 independent variables which are rain, temperature, sunshine, windspeed, humidity, hour, day and holiday with one dependent variable customer visits. Each variable was deeply explored individually and in relation to the target variable beginning with the continuous variables.

3.9.1 Continuous Variables

The dataset contains six continuous variables which included the target variable customer visits and five continuous independent variables: Rain, Temperature, Sunshine, Windspeed and Humidity.

### Table 3.5: Descriptive Statistics for continuous variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>3rd Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Visits</td>
<td>1</td>
<td>37</td>
<td>55</td>
<td>52.35</td>
<td>22.54</td>
<td>67</td>
<td>169</td>
</tr>
<tr>
<td>Rain</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.076</td>
<td>0.38</td>
<td>0</td>
<td>13.2</td>
</tr>
<tr>
<td>Temperature</td>
<td>-2.4</td>
<td>7.5</td>
<td>11.5</td>
<td>11.39</td>
<td>5.02</td>
<td>15.2</td>
<td>26.1</td>
</tr>
<tr>
<td>Sunshine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.24</td>
<td>0.35</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>Windspeed</td>
<td>1</td>
<td>8</td>
<td>11</td>
<td>11.47</td>
<td>5.07</td>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>Humidity</td>
<td>35</td>
<td>67</td>
<td>77</td>
<td>76.1</td>
<td>12.02</td>
<td>86</td>
<td>100</td>
</tr>
</tbody>
</table>
3.9.1.1 Customer Visits

Fig 3.5 is a histogram which shows the distribution of the target variable: customer visits. The distribution is fairly normal with a slight skew to the right. Table 3.5 gives descriptive statistics of all the continuous variables in the dataset including the target variable customer visits. The number of customer visits ranges from 1 to 169 and the average number of customer visits for the period considered is 52 and the standard deviation from the mean is 23.

![Figure 3.5: Histogram of Customer Visits.](image)

3.9.1.2 Rain

As seen in Fig 3.6, the distribution of the rain variable is heavily skewed to the right, this is because most values of this variable are equal to 0 as the minimum, 1st Quartile, median and 3rd quartile are all equal to zero which means that there was no rain recorded for that specific hour. Rain ranges from 0 to 13.2 and has a mean value of 0.076 with a standard deviation of 0.38 from the mean.
3.9.1.3 Temperature

The distribution of the temperature variable across the dataset as seen in Fig 3.7 is fairly normal with the lowest temperature and highest temperature recorded during this period being -2.4 and 26.1 respectively. The mean temperature is 11.39 Degrees Celsius with a standard deviation of 5.02 degree Celsius from the mean.
3.9.1.4 Sunshine

The variable sunshine ranges from 0 to 1 hour and has a very skewed distribution as seen in Fig 3.8, it can also be observed from the diagram that most instances in the dataset have a dataset of 0. The mean value for this variable is 0.23 hours with a standard deviation of 0.35 hours from the mean.

![Histogram of Sunshine](image)

**Figure 3.8: Histogram of Sunshine**

3.9.1.5 Windspeed

From Fig 3.9, the windspeed variable has a fairly normal distribution with the minimum and maximum windspeed for the period considered being 1 knot and 35 knots respectively. The mean windspeed is 11.47 knots with a standard deviation of 5.07 knots from the mean.
3.9.1.6 Humidity
Humidity indicates the relative humidity at a particular hour and is measured in percentage. The humidity in the dataset ranges from 35 to 100 percent and the mean humidity is 76.1 percent with a standard deviation of 12.02 percent from the mean. From Fig 1, it can be observed that the distribution of humidity in the dataset is skewed right.
3.9.2 Relationship between Continuous Variables

After exploring each continuous variable independently, the next step was to investigate the relationship between continuous variables. It is important to see if any of the independent variables has a strong relationship with the dependent variable customer visits but is also important to see observe the relationship between the several independent variables as a strong relationship between two independent variables can lead to a condition called multicollinearity which is unfavourable. Fig 3.11 shows a scatterplot matrix of all the continuous variables in the dataset. Due to the high number of instances and variance in the dataset, it is difficult to identify the strengths of the relationship between the continuous variables. Therefore, to confirm the strength of relationships between the continuous variables, the correlation coefficient between each pair of continuous variables was computed. The correlation coefficient is a numeric measure that indicates the strength of the linear relationship between two continuous variables and it ranges between -1 and 1 with values approaching 1 and -1 indicating a very strong positive or negative relationship while values farther indicates a weak relationship.

![Scatter plot matrix of continuous variables](image)

**Figure 3.11: Scatter plot matrix of continuous variables**

Fig 3.2 shows that all the continuous variables in the dataset have a very weak relationship with the target variable customer visits with humidity having the strongest
relationship with customer visits resulting in a correlation coefficient of just -0.2. The strongest relationship among continuous variables in the dataset is between humidity and sunshine with a correlation coefficient of 0.5.

3.9.3 Categorical Variables
There are three categorical variables in the dataset which are hour, day and holiday. Unlike the continuous variables, exploring the categorical variables individually gives no insights, so the relationship of each categorical variable with the target variable: customer visits is explored directly instead.

3.9.3.1 Hour Effect on Customer Visits
Fig 3.13 shows the box and whisker plot for hourly customer visits. It can be clearly observed that there is an effect on the number of customer visits based on the hour. There is a visible peak at 12pm and 1pm which is most likely due to lunch, a steady decline can be observed after that up on till 6pm where it rises again before declining all the way till the last hour at 9pm. In summary median customer visits are higher at 1pm which is the classic lunch time, this is followed by 12pm and 2pm which could be
as a result of people turning up early or late for lunch. This trend repeats itself with dinner time though with a lesser overall median, peak dinner time is at 6pm and this represents the fourth highest median after 1pm, 12pm and 2pm; the next highest median customer visits are at 5pm and 7pm which as was with lunch time could be because of people coming in for an early or late dinner. It is also worth noting that the lowest median customer visits are at the opening and closing hours of the restaurant. The box and whisker plot is also useful for detecting outliers which are rare occurrences that can affect the outcome of a mathematical forecasting model.

![Boxplot of hourly customer visits](image)

3.9.3.2 Daily Effect of Customer Visits

Fig 3.14 shows the daily effect of customer visits. It can be observed that median customer visits are lowest on Sunday with Monday and Saturday also experiencing relatively lower median customer visits. There is a visible trend with median customer visits gradually increasing as the week progresses from Sunday with the peak median customer visits happening on Friday before declining on Saturday.
3.9.3.3 Holiday Effects on Customer Visit

From Fig 3.16, the median customer visits are higher on days that are not holidays and lower on days classified as holidays. It might be interesting to split the holiday variable into more categories corresponding to what type of holiday it is, the extreme level of aggregation of this study means only very few instances would be classed as specific holidays which would not be an informative feature for building a model.
4 IMPLEMENTATION AND RESULTS

4.1 Introduction

This chapter discusses details about the experiments carried out to answer the research questions and test the hypothesis identified for this study. The results of each experiment would be presented and discussed briefly as well.

4.2 Significant Factors- Experiment 1

In sections 3.9, the various factors identified that could have an effect on the number of customer visits every hour to the fast food restaurant were analysed individually and in relation to the target variable. Though some initial insights into the importance of these factors was investigated using scatterplots, correlation and through box plots; carrying out a statistical test gives more conclusive evidence as to how important each of the factors are.

A multiple linear regression model was created to investigate the significance of the relationship between number of customer visits and factors such as rain, temperature, humidity, sunshine, windspeed, hour, day and holidays. A significant regression equation was found \((F = 888.2, p < 0.001)\) with an adjusted \(R^2\) of 0.7144 indicating the factors explain 71.44\% of the variability in the dependent variable customer visits.

Table 4.1 shows details of the regression equation which includes the coefficient estimates and P-value for each independent variable. The P-value indicates the likelihood that the relationship between an independent variable and the dependent variable is not due to chance, a p-value of less than 0.005 is generally regarded a strong threshold at which point the null hypothesis which states the relationship is due to chance can be rejected.
<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>T value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>34.07288</td>
<td>1.5055</td>
<td>22.707</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Rain</td>
<td>-1.94682</td>
<td>0.37345</td>
<td>-5.201</td>
<td>2.04e-07</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.26103</td>
<td>0.03010</td>
<td>-8.672</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Sunshine</td>
<td>-0.43094</td>
<td>0.47447</td>
<td>-0.908</td>
<td>0.36377</td>
</tr>
<tr>
<td>Windspeed</td>
<td>-0.15127</td>
<td>0.02765</td>
<td>-5.471</td>
<td>4.63e-08</td>
</tr>
<tr>
<td>Humidity</td>
<td>-0.07143</td>
<td>0.01474</td>
<td>-4.846</td>
<td>1.28e-06</td>
</tr>
<tr>
<td>Hour12pm</td>
<td>50.92708</td>
<td>0.64267</td>
<td>79.243</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour13pm</td>
<td>64.65888</td>
<td>0.64387</td>
<td>100.423</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour 14pm</td>
<td>49.99929</td>
<td>0.64432</td>
<td>77.591</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour15pm</td>
<td>42.95709</td>
<td>0.64448</td>
<td>66.654</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour16pm</td>
<td>38.76847</td>
<td>0.64432</td>
<td>60.170</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour17pm</td>
<td>40.69304</td>
<td>0.64623</td>
<td>62.970</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour18pm</td>
<td>44.19495</td>
<td>0.64926</td>
<td>68.069</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour19pm</td>
<td>35.88537</td>
<td>0.65262</td>
<td>54.987</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour20pm</td>
<td>21.42969</td>
<td>0.66014</td>
<td>32.462</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>Hour21pm</td>
<td>7.57271</td>
<td>0.66585</td>
<td>11.373</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>DaySun</td>
<td>-17.03486</td>
<td>0.50797</td>
<td>-33.535</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>DaySun</td>
<td>-17.03486</td>
<td>0.50797</td>
<td>-33.535</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>DayMon</td>
<td>-10.35051</td>
<td>0.51716</td>
<td>-20.014</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>DayTue</td>
<td>-8.77607</td>
<td>0.51001</td>
<td>-17.208</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>DayWed</td>
<td>-5.31922</td>
<td>0.50867</td>
<td>-10.457</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>DayThu</td>
<td>-1.56650</td>
<td>0.50595</td>
<td>-3.096</td>
<td>0.00197</td>
</tr>
<tr>
<td>DaySat</td>
<td>-8.85724</td>
<td>0.51025</td>
<td>-17.359</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>HolidayYes</td>
<td>-5.77373</td>
<td>0.86642</td>
<td>-6.664</td>
<td>2.85e-11</td>
</tr>
</tbody>
</table>

Table 4.1: Regression Equation Details

Table 4.1 shows that the impact of rain, temperature, windspeed and humidity and number of customer visits is statistically significant with the p-value for each approaching zero. Despite the fact that all the continuous variables are all only weakly correlated to customer visits as seen in Fig 3.12. All continuous variables except sunshine have a statistically significant impact on the number of customer visits most
likely due to the high number of instances in the dataset (7803). Among the continuous variables, humidity has the largest impact with a coefficient estimate of -0.07143. The p-value of the relationship between sunshine and number of customer visits is 0.36377 which is above the threshold p-value of 0.005 and which implies the relationship is not statistically significant as there is not enough evidence to reject the null hypothesis.

To include a categorical variable as a predictor for a multiple linear regression model, a dummy variable has to be created for all but one level in the categorical variable. The dummy variable usually takes a binary form of 0 and 1 with 0 indicating the absence of that level on a particular instance and 1 indicating its presence. As can be observed from Table 4.1, the categorical variables; hour, day and holidays were converted to dummy variables. It is important to note that one level is omitted in each categorical variable which serves as a reference;“11am”, “Fri “and “No” are the reference levels for the categorical variables Hour, Day and Holiday respectively. It can be seen from Table 4.1 that all the dummy variables derived from the categorical variables have a statistically significant relationship with the number of customer visits with a p-value of less than 0.005. As expected and observed from the box plot diagrams; Hour13pm, DayFri and HolidayNo are the levels within the various categorical variables with the highest coefficient estimate and therefore the highest impact on number of customer visits. It is important to note that DayFri and HolidayNo are not explicitly included in Table 4.1 because they are reference levels but their identification as the most important level can still be ascertained as the coefficient estimates of all other levels within their respective categorical variables has a negative sign which means that a prediction for a day that is a Friday and a day which is not a holiday would be higher than all other days as there would be no subtraction on that basis.

4.3 Data Preparation for modelling

To prepare the dataset for modelling, a couple of tasks were carried out to improve the quality of the data and to split the dataset into train as well as test sets

4.3.1 Outliers Removal

Outliers are specific instances with values which deviate from the distribution of other values in the dataset. Machine learning algorithms when used for predictive modelling can be sensitive to outliers which could lead to inaccurate models. It was therefore
important to handle outliers in the dataset before developing the machine learning models for this study. An outlier can be mathematically expressed as a value that is greater than 1.5 times the inter quartile range below the first quartile or above the second quartile. An instance is an outlier if greater than value obtained in equation 4.1 or less than the value obtained in equation 4.2.

\[
Q3 + 1.5(IQR) \quad (4.1)
\]

\[
Q1 - 1.5(IQR) \quad (4.2)
\]

Where:

Q3: Third Quartile

Q1: First Quartile

IQR: Interquartile range given by: \(IQR = Q1 - Q3\) \(4.3\)

Number of customer visits is the target variable to be predicted in this study and therefore removing outliers based on the values of customer visits is most appropriate. Using the equations 4.1, 4.2 and 4.3 whilst applying the quartile values already obtained from Table 3.5

\[
IQR = Q1 - Q3 = 67 - 37 = 30
\]

Therefore, an instance is an outlier if number of customer visits is greater than:

\[
Q3 + 1.5(IQR) = 67 + 1.5(30) = 112
\]

Or less than:

\[
Q3 - 1.5(IQR) = 37 - 1.5(30) = -8
\]

The lower boundary for outlier identification is a negative value which does not conform to the expectations of the data as there can be no negative number of customers. Equation 4.2 was therefore modified to categorize an instance as an outlier if it is less than equation 4.4.

\[
1.5(IQR) - Q1 \quad (4.4)
\]

Therefore, the lower boundary to cut off an outlier can be give as:

\[
1.5(IQR) - Q1 = 1.5(30) - 37 = 8
\]
Therefore, an instance is considered an outlier if the value for customer visits is less than 8 or greater than 112. Just 126 instances conformed to the rules to identify outliers which represents less than 2% of the data. After removal of outliers, there were 7,677 instances left in the dataset.

4.3.1.2 Training and Test Split

The data was randomly split into training and test sets corresponding to 70% and 30% of the entire data respectively. Each machine learning model was trained on the training data and the test data was used as an out of sample dataset to avoid overfitting and forms the basis of evaluating the accuracy of the machine learning model on data not seen or used for training it. After splitting the dataset randomly, the training and test sets had 5374 and 2303 instances respectively.

4.4 Modelling - Experiment 2

As discussed in chapter 3, four different machine learning models would be developed to predict the number of customer visits. The following sections gives details on each machine learning model developed.

4.4.1 Multiple Linear Regression

A multiple linear regression model is usually the base line for machine learning problems where the target variable is a continuous value. This model was used in section 4.2 to determine the significant factors in the entire dataset. In this experiment, the multiple linear regression model was created using only the training set and then applied to the test set to generate predictions which would form the basis of evaluating the performance of the model by computing RMSLE and MAPE for the model. The baseline lm() function in R was used and all the independent variables in the dataset were selected. Table 4.2 shows details on the setting of the parameters.
### Table 4.2: Parameters linear regression

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Selected Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>formula</td>
<td>Customer_Visits ~ Rain + Temperature + Sunshine + Windspeed + Humidity + Hour + Day + Holiday</td>
</tr>
<tr>
<td>data</td>
<td>Train_Set</td>
</tr>
<tr>
<td>subset(Optional)</td>
<td>Not used</td>
</tr>
<tr>
<td>weights (Optional)</td>
<td>Not used</td>
</tr>
</tbody>
</table>

### 4.4.2 Regression Tree

A regression tree model was created using the rpart package for R. Fig 4.1 shows the visualization of the regression tree created, it can be observed that nodes are split on just two predictive features; hour and day.

### Table 4.3: Regression Tree Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Selected Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>formula</td>
<td>Customer_Visits ~ Rain + Temperature + Sunshine + Windspeed + Humidity + Hour + Day + Holiday</td>
</tr>
<tr>
<td>data</td>
<td>Train_Set</td>
</tr>
<tr>
<td>weights</td>
<td>Not used</td>
</tr>
<tr>
<td>subset</td>
<td>Not used</td>
</tr>
<tr>
<td>na.action</td>
<td>Default</td>
</tr>
<tr>
<td>method</td>
<td>Anova</td>
</tr>
<tr>
<td>model</td>
<td>Default</td>
</tr>
<tr>
<td>x</td>
<td>Default</td>
</tr>
<tr>
<td>y</td>
<td>Default</td>
</tr>
<tr>
<td>parms</td>
<td>Not used</td>
</tr>
</tbody>
</table>
Table 4.3 shows the parameters set to develop the regression tree model using the rpart package.

To optimize the performance of the regression tree, the validation error was plotted and examined to see how it reduced over the training time. From Fig 4.2 it can be observed there is a continuous decline in the error over time with no observable period where the error stabilizes or increases. Therefore, pruning the tree did not make the model performance better on the test set.
4.4.3 Random Forest

As discussed in chapter 2, the random forest model is a model ensemble which is a combination of different regression trees trained on a subset of instances and predictive features in the dataset. A random forest model was developed using the training data and default parameters for the randomForest model R package. As can be seen from Fig 4.3 which is the plot of how the error declines along with the number of trees, the error decline becomes stable at about the 300th tree. The default number of trees for the random forest algorithm package used for this analysis is 500 trees, this parameter was then set to 300 trees to optimize the algorithm. Table 4.4 shows some of the parameter settings for the final random forest model used.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Train_Set</td>
</tr>
<tr>
<td>Subset</td>
<td>Not used</td>
</tr>
<tr>
<td>X.formular</td>
<td>Customer_Visits ~ Rain + Temperature + Sunshine + Windspeed + Humidity + Hour + Day + Holiday</td>
</tr>
<tr>
<td>Importance</td>
<td>Default</td>
</tr>
<tr>
<td>ntree</td>
<td>300</td>
</tr>
<tr>
<td>replace</td>
<td>Default</td>
</tr>
</tbody>
</table>

Table 4.4: Random Forest Parameters
The random forest model is also a useful model in identifying important predictors. The importance function returns the incnodepurity which is a measure of the mean squared error for choosing the best predictors to split the data and with higher values of incnodepurity indicating a more important predictor. Table 4.5 shows the incnodepurity of all the predictor variables in the dataset in order of importance and this corroborates strongly to the statistical test done in section 4.2 to determine important variables with hour and day being the most important predictors to determine number of customer visits per hour.
### 4.4.4 Gradient Boosting

The gradient boosting regressor from scikit learn for the python programming language was used to build the gradient boosting algorithm for predicting the number of customer visits. The gradient boosting algorithm is prone to overfitting due to it repetitive leveraging of patterns in residuals so applying the developed gradient boosting algorithm to a test set not used for training gives a better indication of the true accuracy of the model and how well it would generalize to unseen data. Table 4.6 shows the parameters of gradient boosting algorithm while Fig 4.4 shows the residual plot.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Selected Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>loss</td>
<td>ls</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.1</td>
</tr>
<tr>
<td>n_estimators</td>
<td>100</td>
</tr>
<tr>
<td>Max_depth</td>
<td>3</td>
</tr>
<tr>
<td>criterion</td>
<td>friedman_mse</td>
</tr>
<tr>
<td>Min_samples_split</td>
<td>2</td>
</tr>
<tr>
<td>Min_samples_leaf</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table 4.6: Parameters Gradient Boosting*
4.5 Model Deployment

To demonstrate how the best performing machine learning could be used for generating forecasts. A web app was developed using flask and Heroku which would display the predicted number of customer visits based on the values assigned to the predictor variables used in building the model. Fig 4.6 below shows a screenshot of the live web app.

![Residual vs. Predicted](image)

*Figure 4.4: Residual Plot*

![Screenshot of Web Portal to generate predictions](image)

*Figure 4.5: Screenshot of Web Portal to generate predictions*
5 ANALYSIS, EVALUATION AND DISCUSSION

5.1 Introduction

In this chapter the results obtained from the various experiments are discussed in the context of the research question set at the beginning of this study and in relation to several similar studies. The machine learning models developed are evaluated using RMSLE and MAPE values to determine which model performs best in predicting the number of customer visits for each hour. The limitations of the studies are also discussed.

5.2 Significant Factors

Variables related to weather which include rain, temperature, sunshine, humidity and windspeed were all found be weakly correlated to number of customer visits with humidity having the highest correlation coefficient of just 0.2 to number of customer visits but the effect of all these weather variables except sunshine was found to be significant nonetheless as seen in Table 4.1. It can also be seen from Table 4.1 that the dummy variables derived from hour and day have a significant effect on the number of customer visits, this is corroborated by analysing Table 4.5 which ranks the importance of all the variables used to create the random forest model, hour of the day and day of the week have the highest impact on predicting hourly number of visits.

The most significant variable is hour which indicates the current hour while the second most significant variable is Day which indicates the day of the week. Both hour and Day are easily observable factors which does not require any secondary data. Also, if a time series model like ARIMA was applied to predict the number of fast casual visits, it can be argued that hour and day would also be critical as time series model relies entirely on historical time series data and like the ARIMA model also on detecting trend and seasonality in the data. However, unlike time series models, machine learning models used for this study does not require the time series order of the data to be maintained, instead data is collected on each factor that has a relationship with the target variable and each data point is treated as being independent to the other data.
points. This makes it easy to statistically report on just how significant each factor has on a target variable.

Based on the results of the multiple linear regression experiment and the results presented in Table 1, there is enough evidence to reject the null hypothesis 1 in favour of alternate hypothesis 1 which states that:

“Day of the week, hour of the day, rain, temperature, humidity, windspeed and public holidays are significant in predicting number of customer visits per hour to a fast-casual restaurant.”

Kokkinou (2013) used two weather related variables; temperature and precipitation along with dummy variables derived from day of the week to develop a forecasting model. The results of their experiment showed only temperature and dummy variables for Saturday and Sunday had a statistically significant effect on predicting daily visits, this contrasts to the current study where not only temperature but also precipitation(rain) along with all days of the week have a significant effect on number of customer visits. It is important to note the difference in level of aggregation which might have had an impact on the results of both studies as this study predicts hourly demand while the research done by Kokkinou predicted daily demand.

Arunraj and Ahrens (2015) developed a hybrid model for generating daily forecasts where a Seasonal Auto Regressive Moving Average Model(SARIMA) was first developed with the predictions generated included as a predictor along with other variables such as month, holidays, promotion and weather-related variables for the multiple linear regression model. They excluded temperature and humidity to avoid multicollinearity and also excluded days of the week as they deemed it already accounted for in the SARIMA model and hence their study did not report the significance of temperature, humidity and day of the week. However, in contrast to this study, they found sunshine duration has a significant impact on daily forecasts at the 5% level.

5.3 Predictive Machine Learning Models

Several machine learning models were created to predict the number of customer visits to the fast-casual restaurant every hour. Table 5.1 shows the RMSLE and MAPE value of each model when applied to the test set. It can be seen that the gradient boosting
model has the lowest RMSLE and MAPE values and is therefore the best performing model in predicting the number of customer visits per hour to the restaurant.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSLE</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear Regression</td>
<td>0.2522</td>
<td>20.76</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.2612</td>
<td>22.12</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.2404</td>
<td>20.51</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.2318</td>
<td>18.78</td>
</tr>
</tbody>
</table>

Table 5.1: RMSLE and MAPE values of all predictive models developed

Using regression methods of forecasting instead of time series methods has various advantages such as ability to identify potentially significant factors affecting the dependent variable. Another advantage of regression and machine learning approach is the ability to disregard the order of the data and treat each data point independently of the other data points, this is unlike most time series methods which relies on the order of the data to generate a forecast for the next period, this helps to randomize the data as required which helps to develop models that generalize better to unseen data.

Several studies (Kokinnou, 2013; Reynolds et al.,2013) that have applied regression methods in their research studies make use of only the multiple linear regression algorithm. While the multiple linear regression algorithm is the baseline machine learning model for most machine learning projects with a continuous target variable, there are other machine learning algorithms that can be used for prediction and as seen in this study other machine learning algorithms use more sophisticated techniques with potential to generate results which outperform the multiple linear regression model.

There is no universal range for evaluating forecasting models in terms of RMSLE or MAPE values, each case would have different acceptable ranges based on the business problem being assessed and the level of accuracy required to tackle the specific business problem. However, Lewis (1982) proposed the interpretation of forecasting models based on their MAPE presented in Table 5.2.
The gradient boosting model achieved the lowest MAPE when applied to the test set of 18.78% which falls between the 10-20 range, the model can therefore be interpreted as a good forecast by comparing to the MAPE values from Table 5.2.

From the MAPE and RMSLE results of all the machine learning models created when applied to the test set and presented in Table 5.1, it can be seen that all models developed achieve MAPE values of less than 50%, this is visualized in Fig 5.1. Therefore, there is enough evidence to reject null hypothesis 2 in favour of alternate hypothesis 2 which states that:

“Multiple linear regression, decision tree, random forest and gradient boosting models would achieve mean absolute percentage error of less than 50% when applied to predict the number of fast-casual restaurant visits per hour in a dataset of historical visits”.

<table>
<thead>
<tr>
<th>MAPE</th>
<th>INTERPRETATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;10</td>
<td>Highly accurate forecasting</td>
</tr>
<tr>
<td>10-20</td>
<td>Good forecasting</td>
</tr>
<tr>
<td>20-50</td>
<td>Reasonable forecasting</td>
</tr>
<tr>
<td>&gt;50</td>
<td>Inaccurate forecasting</td>
</tr>
</tbody>
</table>

(Source: Lewis, 1982)
5.4 Model Deployment.

Several studies have noted that ease of use is a very important factor considered when individuals choose to adapt a methodology for forecasting, they go on to note that judgemental forecasting methods and time series methods like simple moving average remain popular because they are easy to use. It was important to demonstrate how forecasts could be easily generated using machine learning techniques and displayed on a web app. Deployment of the model using flask was part of the objectives of this research study achieved.

5.5 Limitations

Machine learning models requires collection of historical data with which different algorithms would be trained and evaluated but a lesser recognized fact is that the same data used for developing the machine learning models is required for the model to generate predictions when deployed, therefore there has be a mechanism for the immediate capture of the data to generate future predictions. This is a problem especially in cases where part of the data required is from a secondary source as is the case in this study where weather data was obtained from Met Eireann., there was no
API by which present and future weather data could be queried and therefore the model deployed required a rather inconvenient workaround where users have to enter values for weather related variables like temperature, humidity and sunshine.

There are several other factors that could affect the number of customer visits not considered for this study like events, promotions, economic related variables, impact of competitors and many others. Most times it is impossible to consider these factors for machine learning problems largely because there is no way to measure or capture the data but information of some of the factors could be available to the restaurant managers which is why Davydenko and Fildes(2013) suggests judgemental adjustments to quantitative models when specific contextual factors could not be included in the quantitative model.

Finally, this research only considered a single branch of a fast-casual restaurant chain and only two years’ worth of data was used for the experiments. This makes it difficult to make generalizations about the results obtained from this study.
6 CONCLUSION

6.1 Introduction

This chapter provides a summary of this research study by reiterating the major results of the experiments carried out. The research questions, hypotheses and objectives identified at the beginning of the study would be discussed as well as how the results contribute to the body of the knowledge. Limitations identified while carrying out the study would be discussed along with suggestion on areas of future work and research to expand on the results of the study.

6.2 Research Problem Overview

Fast-casual restaurants unlike their full-service counterparts operate under a high level of uncertainty with very little to no booking in advance and the bulk of their customers walking in randomly. This makes forecasting challenging especially with food safety legislation whereby heated food not consumed after 90 minutes has to be discarded. The major objective of this study which has been achieved was to identify the significant factors that affect the number of customer visits to a fast-casual restaurant every hour and to use these features in building machine learning models capable of predicting the number of hourly customer visits. The objectives achieved as part of this research study are summarized below:

- Extensively reviewed literature on the various methods for forecasting in general and specifically for the restaurant industry as well as a review of machine learning methods suitable for regression.

- Obtained a dataset which contained number of customer visits for each hour from a fast-casual restaurant located in Dublin and combined this with weather data from Met Eiran as well as data on holidays.
• Cleaned the data obtained and ensured final data used for experiments was of a high quality to prevent bias in the results obtained.

• Implemented a multiple linear regression experiment to determine the significant factors affecting the number of customer, hereby gathering enough statistical evidence to reject null hypothesis 1 in favour of alternate hypothesis 1 which states that Day of the week, hour of the day, rain, temperature, windspeed, humidity and public holidays are significant in predicting number of customer visits per hour to a fast-casual restaurant.

• Developed multiple linear regression, decision tree, random forest and gradient boosting algorithms and tuned parameters for each algorithm to obtain best results.

• Evaluated all the developed models by calculating MAPE and RMSLE values for each model when applied to a test set and gathered enough evidence to reject null hypothesis 2 in favour of alternate hypothesis 2 which states that multiple linear regression, decision tree, random forest and gradient boosting models would achieve mean absolute percentage error of less than 50% when applied to predict the number of fast-casual restaurant visits per hour in a dataset of historical visits.

• Designed a web app where the best predictive model identified was deployed to display the predicted number of customer visits to a fast-casual restaurant for a particular hour.

• Identified potential limitations of the results obtained and made suggestions on future research to expand on the finding of this study.

6.3 Experimentation and Evaluation

The aim of this research study has been to identify the significant factors that impact hourly customer visits to a fast-casual restaurant and to use the features to build predictive machine learning models capable of predicting the number of hourly
customer visits to the fast-casual restaurant. The research followed the CRISP-DM methodology which provides a robust end to end process for implementing data mining projects. The two major experiments conducted for this study were the multiple linear regression model developed to statistically test for significant factors and the development and comparison of the multiple linear regression model to the decision tree, random forest and gradient boosting algorithms by computing the RMSLE and MAPE values for each model when applied to the test set.

After an in dept review of the literature on forecasting and predictive modelling, a dataset containing the number of hourly customer visits for two years was obtained from a single branch of Boojum limited, a fast growing fast-casual restaurant chain that sells burritos. This dataset was then combined with weather data from Met Eireann as well as data for all public holidays within the two-year period considered.

Data understanding and preparation stages involved improving the overall quality of the data by removing instances with null values and outliers. Each variable in the dataset was then explored individually and in relation to the target variable, customer visits using tables and data visualizations to gain initial insights into the dataset. To prepare for modelling and evaluation, the cleaned dataset was split into training and tests set with the training and test sets containing 70% and 30% of the entire data respectively.

A multiple linear regression model was created using the entire dataset to obtain the probability value of the significance of the relationship between each independent variable and the number of customer visits per hour to confirm the relationship was not just by chance. Modelling involved developing multiple linear regression, decision tree, random forest and gradient boosting models using only the training data, several parameters were adjusted for each model to obtain the best version for each algorithm. Evaluation of the models was done based on the RMSLE and MAPE values of each algorithms when applied to a separate test set not used for training.

Finally, the best performing model which was the random forest model was deployed to a web app which required the user to enter values for each variable used in developing the model in order to generate and display a prediction for the number of customers expected to visit the restaurant for the selected hour.
6.4 Contribution to the body of Knowledge

Forecasting in the restaurant industry has been a widely researched area though there have been very few recent researches and fewer researches specifically focusing on fast-casual restaurants. This study has made some worthwhile contributions to the body of knowledge summarized below:

- This study not only focused on the development of forecasting models but also on other areas designed according to the CRISP-DM methodology which includes business understanding, data understanding, data preparation, modelling, evaluation and deployment. This study therefore applied an expansive methodology that can be applied to other scenarios.

- As part of this study, significant factors that affect the number of visits a fast-casual restaurant gets very hour were identified with the most significant of the factors identified being easily observable factors like hour of the day and day of the week.

- One important advantage of regression-based forecasting methods over their time series counterparts discovered in this study is the ability to define every data point to be predicted with one or more factors which makes it possible to disregard the order or time at which the data was collected. The ability to randomize data is important in prediction problems in order to prevent overfitting and to evaluate a forecasting model on a random test set which gives an indication of how the model will perform when deployed.

- This study applied other various regression-based machine learning algorithms different to the tradition multiple linear regression algorithm and discovered that the gradient boosting as well as the random forest model outperformed the traditional multiple linear regression model by achieving lower MAPE and RMSLE values when applied to the test set.

- Most studies reviewed in the area of forecasting finished the research at the evaluation stage, this study is different as it demonstrates how the best model
selected after evaluation can be deployed making it easy to access future predictions.

**6.5 Future Work and Research**

The current study has uncovered a business case for hourly forecasts for fast-casual restaurant useful in the daily operations of the restaurant and for reducing wastage. Due to data availability only, number of customer visits has been forecasted. A more detailed study could be conducted to include forecasting the amount of specific menu items that would be used by the restaurant every hour, this would provide even more actionable information to the restaurant manager for ordering and stocking. Exploration of the dataset could also uncover some insights about the effect of weather, time or season on the demand of certain menu items.

Similar experiments could also be conducted to include a larger number of fast-casual restaurants and a longer time period to determine if the factors considered in this study remain significant and if a machine learning model can be developed to predict the outcomes whilst achieving reasonable MAPE and RMSLE values.

Time series methods like exponential smoothening and the ARIMA model are very popular quantitative forecasting techniques not utilized in this study. An experiment could be carried out to compare the performance of the time series methods and regression-based methods. The mode of evaluation should be done such that the test sets should be the same for both methods.
7 BIBLIOGRAPHY


McTiernan, K. (2016) Investigation into the Predictive Capability of Macro-Economic Features in Modelling Credit Risk for Small Medium Enterprises, 163.


8 APPENDIX

Appendix A – Holidays 2016

<table>
<thead>
<tr>
<th>Day</th>
<th>Date</th>
<th>Holiday</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday</td>
<td>January 01</td>
<td>New Years Day</td>
<td></td>
</tr>
<tr>
<td>Sunday</td>
<td>March 06</td>
<td>Mother’s Day</td>
<td>Not a National Holiday</td>
</tr>
<tr>
<td>Thursday</td>
<td>March 17</td>
<td>St Patrick’s Day</td>
<td></td>
</tr>
<tr>
<td>Friday</td>
<td>March 25</td>
<td>Good Friday</td>
<td>Not a public holiday. Banks are closed. Some schools and businesses may close</td>
</tr>
<tr>
<td>Monday</td>
<td>March 28</td>
<td>Easter Monday</td>
<td>Monday after Easter Sunday</td>
</tr>
<tr>
<td>Monday</td>
<td>May 02</td>
<td>May Bank Holiday</td>
<td>first Monday in May</td>
</tr>
<tr>
<td>Monday</td>
<td>June 05</td>
<td>June Bank Holiday</td>
<td>first Monday in June</td>
</tr>
<tr>
<td>Sunday</td>
<td>June 19</td>
<td>Father’s Day</td>
<td>third Sunday in June. Not a public holiday</td>
</tr>
<tr>
<td>Monday</td>
<td>August 01</td>
<td>August Bank Holiday</td>
<td>first Monday in August</td>
</tr>
<tr>
<td>Monday</td>
<td>October 31</td>
<td>October Holiday</td>
<td>last Monday in October</td>
</tr>
<tr>
<td>Sunday</td>
<td>December 25</td>
<td>Christmas Day</td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>December 26</td>
<td>St Stephen’s Day</td>
<td></td>
</tr>
<tr>
<td>Tuesday</td>
<td>December 27</td>
<td>Christmas Day (in lieu)</td>
<td>Next working day after December 25</td>
</tr>
<tr>
<td>Wednesday</td>
<td>December 28</td>
<td>Bank Holiday</td>
<td>Not a public holiday. Banks closed</td>
</tr>
</tbody>
</table>

Appendix B – Holidays 2017

<table>
<thead>
<tr>
<th>Day</th>
<th>Date</th>
<th>Holiday</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>January 01</td>
<td>New Years Day</td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>January 02</td>
<td>New Years Holiday</td>
<td>Banks are closed. Not an official public holiday but will be widely observed</td>
</tr>
<tr>
<td>Friday</td>
<td>March 17</td>
<td>St Patrick’s Day</td>
<td></td>
</tr>
<tr>
<td>Sunday</td>
<td>March 26</td>
<td>Mother’s Day</td>
<td>Not a National Holiday</td>
</tr>
<tr>
<td>Friday</td>
<td>April 14</td>
<td>Good Friday</td>
<td>Not a public holiday. Banks are closed. Some schools and businesses may close</td>
</tr>
<tr>
<td>Monday</td>
<td>April 17</td>
<td>Easter Monday</td>
<td>Monday after Easter Sunday</td>
</tr>
<tr>
<td>Monday</td>
<td>May 01</td>
<td>May Bank Holiday</td>
<td>first Monday in May</td>
</tr>
<tr>
<td>Monday</td>
<td>June 05</td>
<td>June Bank Holiday</td>
<td>first Monday in June</td>
</tr>
<tr>
<td>Sunday</td>
<td>June 18</td>
<td>Father’s Day</td>
<td>third Sunday in June. Not a public holiday</td>
</tr>
<tr>
<td>Monday</td>
<td>August 07</td>
<td>August Bank Holiday</td>
<td>first Monday in August</td>
</tr>
<tr>
<td>Monday</td>
<td>October 30</td>
<td>October Holiday</td>
<td>last Monday in October</td>
</tr>
<tr>
<td>Monday</td>
<td>December 25</td>
<td>Christmas Day</td>
<td></td>
</tr>
<tr>
<td>Tuesday</td>
<td>December 26</td>
<td>St Stephen’s Day</td>
<td></td>
</tr>
<tr>
<td>Wednesday</td>
<td>December 27</td>
<td>Bank Holiday</td>
<td>Not a public holiday. Banks closed</td>
</tr>
</tbody>
</table>
Appendix D

Normal Q-Q Plot of Temperature

Expected Normal vs. Observed Value
Appendix E
Appendix F

Normal Q-Q Plot of Windspeed

Expected Normal vs. Observed Value
Appendix G

Normal Q-Q Plot of Humidity