# **ABSTRACT**

This report includes a Business Problem that has been solved using Adaptive Business Intelligence techniques. In brief, the business problem stated in this report revolves around the restaurant sector where the objective is to predict the footfall in the restaurants in upcoming months based on data collected across various fields in the past months. Initially, exploratory data analysis was carried out to gain insights into the database and once a thorough understanding was established as to how the restaurant footfalls were affected based on the existing training data set, prediction was done on the test data set.

# **INTRODUCTION**

The introduction has been divided into the following subsections. Everything from the business problem description to the method and data set used has been categorically placed in the underlying sections to provide a better interpretation to this report. Also, the contribution made by each member has also been provided.

## **PROBLEM DESCRIPTION, IMPORTANCE AND IMPACT**

To begin with, restaurant owners often face the problem of not knowing how a restaurant will perform when it comes to attracting customers. Even when every other aspect that is responsible for a successful restaurant operation is met, there is always the question of footfalls that a restaurant manager or owner simply cannot guess. An owner may invest a lot of money in a restaurant but it may turn out that the restaurant is not performing enough to make a profit. It is also possible that a sudden peak in footfall was unforeseen which will also lead to loss simply because the restaurant was not capable enough to meet the demands. Either way, the owner has to bear a loss.

With that said, if there was a feasible way to predict a realistic and accurate footfall that a restaurant should expect, appropriate actions could be taken before hand in order to meet the increased or decreased demands. For instance, in the case of increase footfall, extra employees could be hired, the food production could be increased without fear of waste, etc.

If this system is deployed worldwide, it would not only serve the restaurants, but in the long run, huge amount of food materials that was otherwise discarded by restaurants could be saved.

## **METHOD AND DATASET USED**

Since we are trying to predict the footfalls in the restaurants, Time Series Forecasting was the most appropriate method as it is used for predictive modelling. ARIMA and Holt Winters was used for the prediction. Detailed theory about these methods will be discussed in the Technology Review Section.

The dataset used was taken from Kaggle which includes 8 distinct datasets. The records in these datasets were collected from 2 separate sources. The data for training and testing are separate and spans from 2016 to mid-2017. In brief, the 8 datasets include:

* Past footfall data (Primary training data).
* Bookings made online through source 1.
* Bookings made online through source 2.
* Location and genre of the restaurants from source 1.
* Location and genre of the restaurants from source 2.
* Store ID’s
* Special dates like holidays.
* Test data.

## **CONTRIBUTION**

Starting from the theory to design and development including the presentation and report has been a shared effort. Both of us have equally contributed to each section of this assignment. Given the size of the project, it was fairly easy for both of us to divide the responsibilities. For instance, there were numerous datasets for exploratory data analysis which was divided equally. Then each of us implemented one prediction algorithm. When it came to the report, both of us were present in a room and starting from introduction to conclusion, input was made by both. All in all, we have both contributed equally in this assignment.

# **THCHNOLOGY/TECHNIQUE REVIEW**

The technology adopted for this project is Time Series Forecasting. Given the fact that the final goal is to predict the footfall in a restaurant, which is solely a time-based scenario, this makes this problem a Time Series Forecasting Problem. Now we must understand the fact that Time Series forecasting in itself are a set of steps that can be implemented using different methods. Now, there are numerous methods that can be used for time series forecasting, but the methods that we have decided to implement for this Time Series Forecasting Business problem are ARIMA and Holt Winters. In the subsequent subsections, we will discuss the reason for this selection along with these technologies in detail and by doing so, a basic understanding can be achieved before we move on to the technical implementation.

## **TIME SERIES**

In a nutshell, when a dataset is providing valuable insights based on the time factor in the data, it would be wise to use Time Series Forecasting. To put that in a different perspective, when we have predictions that depend on time, i.e., absence of time factor will render the prediction redundant, Time Series Forecasting can be implemented. But, in the case of a prediction that has nothing to do with time, we can not use Time Series Forecasting.

The way Time Series forecasting works is that it evaluates and analyses the past data in order to predict the likelihood of an event in a given time in future. For instance, in the diagram below, we can see that the past data regarding furniture sales is used to predict the future furniture sales in the future. This is a simple case of time series forecasting.

Chart

Description automatically generated

**Figure 1: Example of Time Series Forecasting**

### **Components of Time Series**

Before moving on to the methods available for Time Series Forecasting, lets look at the components in order to better understand Time Series Forecasting in general. Initially, a model is developed for prediction. It does so by analyzing and identifying hidden patterns and relations in a dataset. Now, there are four components into which these patterns are classified.

1. Trend – Gradual changes in series over a period of time.

A picture containing text

Description automatically generated

**Figure 2: Trend – Non-Linear**

A picture containing diagram

Description automatically generated

**Figure 3:Trend - Linear**

1. Cycle – The variations in the data like curves over a long time.

A picture containing text

Description automatically generated

**Figure 4: Cyclical**

1. Seasonal – Patterns that occur on specific times.

Chart, line chart

Description automatically generated

**Figure 5: Seasonal**

1. Noise/Error – No specific conclusions can be observed.

Sometimes, we may also observe that these components are combined. The following set of figures will demonstrate such cases.

Diagram

Description automatically generated

**Figure 6: Combination of the Trend Cycle and Noise**

### **Steps in Time Series Forecasting**

* The Definition of the problem
* Data and Information Gathering
* EDA and Evaluation of the data
* Model Creation
* Prediction sand forecasting is done using the model
* The accuracy is evaluated

## **MODELS IN TIME SERIES**

There are numerous models that can be used to implement a time series problem. We have discussed some methods and models in this report and based on the data we have gathered, we have chosen the models ARIMA and Holt Winters.

A sequence of values over a series of time stamped values. Time series can take many shapes and forms, there can be as many time series sequences as it can be of any real numbers.

Time series is important area of statistics because it finds many applications is real world, for e.g. the rainfall over a period of time, weekly international flights and passengers.

Broadly, there are two types of time series.

1. **Univariate -** It refers to a time series that contains a single observational value recorded through a time, for example employment rate, carbon di oxide emission etc.

In other words, the response parameter is influenced by only one factor.

1. **Multivariate -** It contains a multiple observational value recorded through a time, but it relates to a single response variable, but the contributing factors are multiple. For example, temperature of a geographical area is based on multiple factors, Interest Rates, Inflation, and Consumption etc.

It is important to note that multivariate time series patterns are more common in real world, since in real world the business cases are often complex in nature, and a variety of factors are often responsible for a value of interest.

Even in case if restaurant forecasting, a multitude of factors as noted earlier are responsible. For example, location, time and holiday season are all contributing factors in the number of visits.

Models

The whole idea of time series forecasting is based on the assumption, that a value of interest in future has some relation with its values in the past.

Time series models are simple a mathematical equations and processes that can predict a value of interest based on its contributing factors in the past, as recorded in time sequenced manner.

Here are few modelling techniques that we considered for restaurant visit forecasting

### **Single Moving Average**

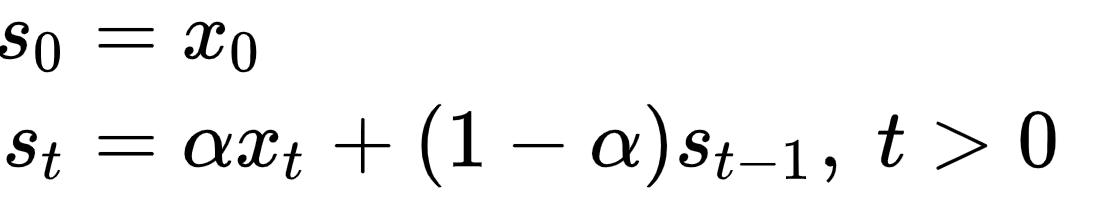
This is simplest form of time series forecasting, it says that a value in the future is simple and average of the past value, in a single period of time. This observation is further divided into a number of subdivisions, based on time. Observation from a single division is then averaged and is considered as a forecast of the future. As we move in future a number of more data is added and thus becomes another division, or period, which is then averaged, and therefore creating a moving average. This method is most suitable when a variable to be forecasted.



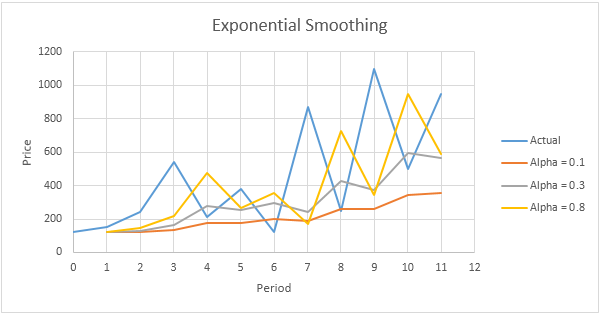
**Figure 7: Example of Simple Moving Average**

### **Exponential Smoothing**

It is an advanced version of Moving average, In SMA all the observations of the pasts were weighed equally. The underlying principle of exponential smoothing is to apply weights to the forecasted values such that the recent values are given more weights, and as we go further in the past, the weights are decreased exponentially. Recent past is more important than further past. The exponential functions whose exponent factor is called Alpha α, is used to assign weights to the values of the past and then the average is taken. This smoothens the time series and hence the name exponential smoothing.



**Figure 8: Formula of Exponential Smoothing**



**Figure 9: Example of Exponential Smoothing**

For restaurant visit forecasting this is a simplistic method and also we need to note that both Exponential smoothing and moving average are univariate time series, but in case of a real world business problem a number of factors would be ideally responsible and hence these are not a good choice for our problem. Also, it cannot be used when there is trend and seasonality in the data.

### **ARIMA**

It stands for Auto Regressive Integrated Moving Average. In reality this to a number of models or class of models. It is based on the autocorrelation in the data. But first we need to understand stationary and differencing time series.

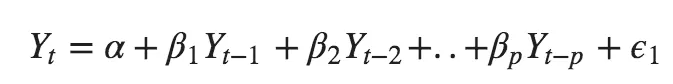
A stationary pattern is one whose parameters are not dependent over the time at which the data is being observed. And therefore, seasonal data, i.e. a fluctuating data at a fixed time is not a stationary time series. Contrast can be drawn with cyclical data. A good case can be crime prediction.

Differencing the data means that one set of observation can be deducted from another set of observation within the same series and the difference is more or less same.

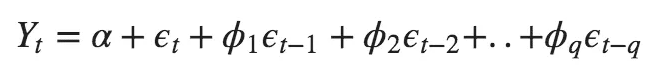
ARIMA is mostly used when there is non-stationary data. Additionally, when there is moving average and also a trend.

So first we difference the time series in order to make the series stationery, and then combine the Auto Regressive as well as Moving Average techniques. Additionally, we do a step called Integration

Auto Regressive or AR only model can be represented by the equation,



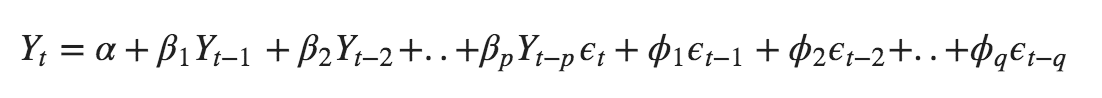
Pure Moving Average or MA only model can be represented by the equation,

****

This model can be summarized as

**Predicted Value = Constant + Linear combination Lags of Y (maximum p lags) + Linear Combination of Lagged forecasts error (maximum q lags)**

The final equation of ARIMA is,



For programming we need to determine the following terms

**p** Order of the AR term

**q** Order of the MA term

**d** No. of differencing needed for stationary series

In R ‘forecast’ and ‘tseries’ package is required for using ARIMA Modelling

### **Holt Winters**

For restaurant visit forecasting this is one the better methods to forecast the required parameters. Holt-Winters is a method of modelling 3 factors of the time-series:

1. Typical value (basically average)
2. Trend over time
3. Cyclical pattern (seasonality)

It employs exponential smoothing method to encode values of the past and uses it to forecast “typical” values.

Since it uses these three aspects, hence it is called triple smoothing method as well.

The model asks for many parameters for each smoothing (ɑ, β, γ), the size of a typical season, and the number of periods within each season.

Seasonality simply means a fixed length of time where a pattern repeats fully. So, if a pattern repeats itself in 6 periods of time, the season is 6 periods. The most important part in Holt-Winters is to use selecting these three parameters (trend, values and seasonality) for accurate forecasting.

### **Suitability of Holt Winters in our case**

* The data as explored in EDA stage shows that there is a seasonality, although we do not know the season period.
* Also, it is seen that there is a trend in our data, because for visitors each genre of food the mean values range between 10 and 100 visitors. In each of the category, long-term trend looks stable. There is an upward trend in "Okonomiyaki" and "Creative Cuisine, while the popularity of "Asian" genre is seen to be declining from 2016.

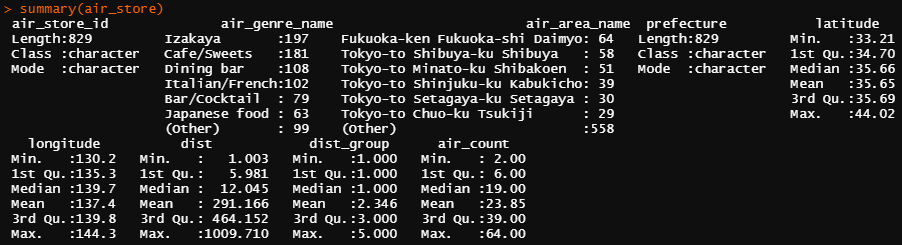
# **SOLUTION DESIGN AND DEVELOPMENT**

The primary goal of this project is prediction, but for that to take place, we must have a firm understanding of the variables that are contributing to that prediction. That includes a thorough exploratory data analysis through which we can quantify the extent of influence that each parameter has over the data and the relation between the features.

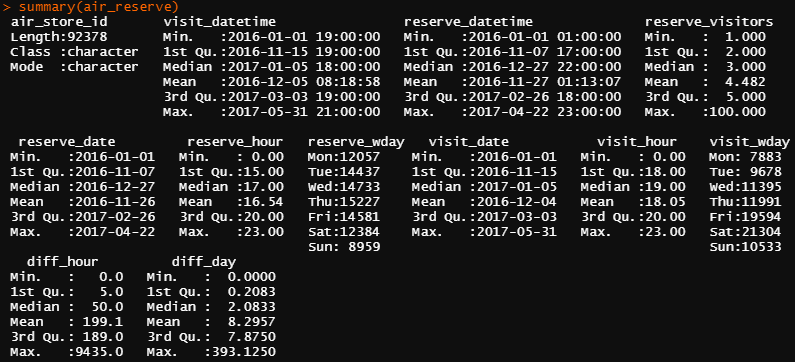
The following sections will include the entire project starting from the data sets to how the prediction was made.

## **Overview of the Dataset and its structure**

The following screenshots will provide the summary of each datasets used in this project.



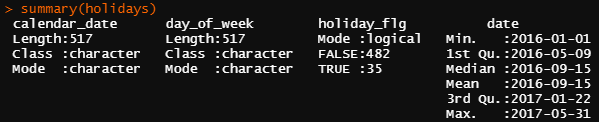
**Figure 10: Summary of air\_store**



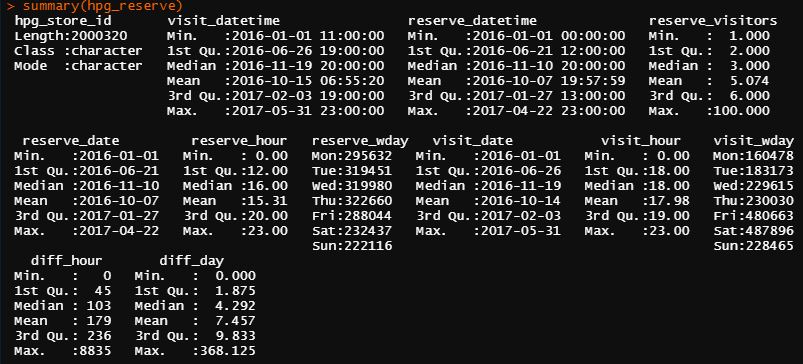
**Figure 11: Summary of air\_reserve**



**Figure 12: Summary of air\_visits**



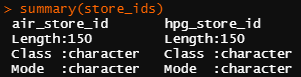
**Figure 13: Summary of holidays**



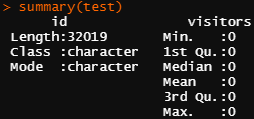
**Figure 14: Summary of hpg\_reserve**



**Figure 15: Summary of hpg\_store**



**Figure 16: Summary of store\_ids**



**Figure 17: Summary of the test data set**

Before moving on to the next steps, lets look at the architecture of the entire project. The figure below will demonstrate the overall structure and process of the prediction and the figure that follows will provide with a much more detailed view of the project.

Diagram, text

Description automatically generated

**Figure 18: Steps taken for the prediction**

Now, that we have the basic steps, the following figure will magnify each step in order to paint a much detailed picture of the entire process.

Graphical user interface

Description automatically generated

**Figure 19: Steps for prediction in detail**

## **Programming Language, Platform and Libraries used**

The project was made using the programming language R and RStudio was used as a platform to code the project.

The following table shows the libraries used and its purpose.

|  |  |  |
| --- | --- | --- |
| Function | Libraries | Purpose in Project |
| Visualization | ggplot2 | Basic library for creating plots. |
| scales | For automatically making breaks and labels for legends and axes. |
| grid | Add Details using ridges in feature relations. |
| gridextra | For arranging numerous grids in a single page. |
| rcolorbrewer | Extend the color palette for visualization. |
| corrplot | Universal use for finding correlation between features. Has been used multiple times during EDA in this project. |
| Data Manipulation | dplyr | General dataset manipulation library. |
| data.table | For faster aggregation of data. |
| Input/Output | readr | In this case, to read csv files. |
| Data Wrangling | tidyr | Make the data tidy |
| String Manipulation | stringr | Feasibility in using string functions |
| Factor Manipulation | forcats | Handling and modifying variables. Used when playing with EDA and changing each parameter to look at results. |
| Specific Visualization | ggrepel | Geoms for ggplot2. |
| ggextra | Used ggMarginal function to plot histogram in reservation vs visits. |
| ggforce | Facet zoom view used for relation between mean visitors and time slope. |
| viridis | Color package. |

## **EDA on each feature**

Now that we have the layout of each steps, we will further dive in to each one to understand the purpose and execution. Please note that during the analysis, we will only use the training data and once the prediction is over, we will use the test data to confirm the accuracy.

1. The frequency of visits to each air restaurants can be obtained by taking the total number of visitors each day and median of the visitors each day.

Chart, bar chart

Description automatically generated

**Figure 20: Data related to visits on the air store**

* We can see that the frequency of visitors per day and per visit is maximum 20 but we also observe that it goes up to 100 at certain times.
* Friday, Saturday and Sunday have the most frequency whereas Monday and Tuesday have least.
* In the long term chart, we see a periodic pattern.
* In terms of months, we see maximum visitors in December.

1. To get the relation between the reservation and final visit, we check the frequency of reservations made each day and then find out the timing of these visits with respect to the reservation.

Chart, histogram

Description automatically generated

**Figure 21: Reservations vs Visits**

* We can see that in 2016, not many reservations were made using air. However, we see a spike in reservations by the end of 2016 and the trend continued throughout 2017.
* Most reservations were made for the time of Dinner which is in the evening.
* The most distinct pattern can be seen in the time difference between reservation and visit which was 24 hours. This tells us that instead of booking a seat a few hours before, the customers book it a day ahead in advance.

1. Just like air reservations, we look at hpg reservations.

Chart

Description automatically generated

**Figure 22: hpg reservation data**

* The pattern here is orderly with the exception of December 2016 where the frequency of reservations peaked.
* Just like the air reservations, here the dinner reservations are more than the other mealtimes and also there is a 24 hour difference between reservation and visit.

1. Next, we can classify the restaurants into respective genres of cuisines with respect to the air restaurants.

Chart, bar chart

Description automatically generated

**Figure 23: Classification with respect to genre**

* Tokyo is the densest area after Fukukoka.
* In terms of genre, Izakaya has the highest number of restaurants followed by sweets and cafes with Asian restaurants occurring the least.

1. We do the same for hpg restaurants.

Chart, bar chart

Description automatically generated

**Figure 24: hpg restaurant genres and their frequency**

* Unlike the air restaurants, we see that international restaurants are the second highest just after Japanese restaurants.

1. For holidays, we find the total number and its distribution in the range of our prediction.

Chart, bar chart

Description automatically generated

**Figure 25: Holiday Data**

* The holidays amount to 7 percent of the data.

## **Relation between Features**

At this point, we have the knowledge of each dataset and what they provide. But for predicting the footfall, we must find out the relation between each feature and its effect on the footfall.

1. Frequency of visitors with respect to cuisine.

A picture containing calendar

Description automatically generated

**Figure 26: Visitors with respect to cuisine**

* The mean lies between 10 to 100 customers per cuisine each day. We can also observe that the Okonomiyaki and Creative cuisine shows growth and Asian shows decline at the later part of 2016.
* Although the number of Asian restaurants is less, their demand is more.

1. Now we find out if the holidays have any impact on the footfall. We do so by analyzing the data for days that are holiday vs the days that are not.

Chart, box and whisker chart

Description automatically generated

**Figure 27: Effect of Holidays**

* We do not see any effect of holidays in the footfall.
* We also observe that the holidays during the weekend has minimum impact on the footfall. However, the impact is much more on Mondays and Tuesdays.

1. Now, we find out the frequency of visitors with respect to restaurant in a particular area. To do this, we refer to the results found in our initial exploratory data analysis where pubs were the most popular type of restaurants. Now, we can assume that the visitors will be more for pubs and by doing so, we can also assume that the customers that visit these pubs will also visit other nearby restaurants.  
   For the feature visualization, we plot a graph with specific cuisines for each area for both air and hpg datasets.

Chart, scatter chart

Description automatically generated

**Figure 28:Visitors with respect to area in air**

* The above graph shows us that in certain places, there are different kinds of restaurants whereas in some, there is only one type.
* We also see that restaurant like Izakaya and Café are much more occurring than others.

1. The same can be mapped for hpg restaurants.

Chart, scatter chart

Description automatically generated

**Figure 29: Visitors with respect to area in hpg**

* Here, the occurrence is must more prominent as the number of cuisines are more.
* We also see that the places with fewer restaurants have more traffic.
* Tokyo appears to have a variety of genres.
* Japanese and International genres are present all over the map.

1. Now that we have the frequency of restaurants specific to certain genre in an area, a much more informative graph can be obtained using boxplot.

Chart, scatter chart

Description automatically generated

**Figure 30: Cuisine and its count in each area**

* Jitter plot is used to make the dots separate and not overlay on each other.
* We can see that only some cuisines have medians of more than 2.
* Cafes and sweets have 26 counts in a specific area.
* Here, the minimum occurrence of a particular cuisine is 2.

1. Now, in the case of hpg restaurants,

Chart

Description automatically generated

**Figure 31: Occurrence per hpg area**

* The minimum occurrence of a particular type of restaurant is 1.
* The median for Japanese is more than 10 in an area.

1. Earlier, we have seen the relation between the visits and reservation, now we will find out how many reservations actually turned in to a visit. For this, we take the reservation data and match it with visitors count. In other words, we calculate the total number of reservations made in each restaurant and match it with the visit data set.

Chart, scatter chart

Description automatically generated

**Figure 32: Reservation vs Visits**

* The frequency of reservations and visits spike around 20.
* We see that the walk-in visitors were also abundant.
* We also see that many who made reservations did not visit.

## **Feature Extraction and Engineering**

We must now extract new parameters from the existing parameters in order to predict the footfall in the restaurant which is our final goal.

1. Firstly, we observe the week and month parameter from the date of visits.

---------------------------------------------------------------------

# **EVALUATION**

## **Using ARIMA**

The forecast will be for the same time period that we had first began with, i.e., 23rd April – 31st May. Let us call it prediction length. The forecast will be made using auto.arima tool. The ts tool will be used to convert the data into objects and tsclean will be used to clean the data and remove any outliers.

Initially, we will try to predict the final 39 days of the training dataset, then we will get the time series for a particular air\_store\_id. The count of visitors will be transformed using log1p and the all visit\_dates data will be joined.

Now, we will put the data into training and validation groups. The ts function will be used to clean the data and remove any outliers and then the week frequency will be added. Since we have a small dataset, computing time will not be an issue. Then the prediction will begin and confidence intervals will be set.

A picture containing graphical user interface

Description automatically generated

**Figure 33: ARIMA Forecast**

* The prediction is shown in dark blue color and the confidence interval in light blue.

## **Using Holt\_Winters**