A logo for a university

Description automatically generated

**Collaborative Application Development (CSC-40038)**

**Group Coursework**

**Submitted by:**

|  |  |
| --- | --- |
| **Group 10** | |
| Muhammad Shaheer Rehan | Student ID: 23029762 |
| Chakaya Nyamvula | Student ID: 23009693 |
| Jamie Wilson | Student ID: 20015339 |
| Jyothesh karnam | Student ID: 23032448 |
| Thomas Goold | Student ID: 20021999 |
| Mohanned Younis | Student ID: 23039903 |
| Reynold Kwok | Student ID: 23003788 |

**Date of Submission:** February 29, 2024.

**School of Computing and Mathematics,**

**Keele University, UK**

**2024**

TABLE OF CONTENTS

[1. ABSTRACT 4](#_Toc160103470)

[2. DATA ANALYSIS 5](#_Toc160103471)

[1.1 Introduction 5](#_Toc160103472)

[1.2 Assumptions 5](#_Toc160103473)

[1.3 Important Statistical Metrics 6](#_Toc160103474)

[1.3.1 Arithmetic Mean 7](#_Toc160103475)

[1.3.2 Standard Deviation 7](#_Toc160103476)

[1.3.3 Z-Score Calculation 8](#_Toc160103477)

[1.4 Approach for Statistical Analysis 8](#_Toc160103478)

[1.5 Threshold Calculation 8](#_Toc160103479)

[1.6 Significance 9](#_Toc160103480)

[1.7 Data Analysis for Each Conference 10](#_Toc160103481)

[1.7.1 D19 10](#_Toc160103482)

[1.7.2 D21 11](#_Toc160103483)

[1.7.3 GP21 12](#_Toc160103484)

[1.7.4 MSE21 13](#_Toc160103485)

[1.7.5 NP21 14](#_Toc160103486)

[1.7.6 SRM22 15](#_Toc160103487)

[1.7.7 SRM23 16](#_Toc160103488)

[1.8 General Observations 17](#_Toc160103489)

[3. COMPARATIVE STUDY OF PREDICTION MODELS 18](#_Toc160103490)

[1.9 Introduction 18](#_Toc160103491)

[1.10 Literature Review 18](#_Toc160103492)

[1.10.1 Regression Models 18](#_Toc160103493)

[1.10.2 Time-Series Models 20](#_Toc160103494)

[1.10.3 Classification Models 23](#_Toc160103495)

[1.10.4 Logistic Regression 25](#_Toc160103496)

[1.11 Assumptions 25](#_Toc160103497)

[1.12 Implementation 26](#_Toc160103498)

[1.12.1 Regression Models 26](#_Toc160103499)

[1.12.2 Time Series Models 30](#_Toc160103500)

[1.12.3 Classification Models 32](#_Toc160103501)

[1.13 Results 32](#_Toc160103502)

[1.13.1 Regression Models 32](#_Toc160103503)

[1.13.2 Time Series Models 43](#_Toc160103504)

[1.13.3 Classification Models 50](#_Toc160103505)

[4. IMPUTATION OF MISSING DATA 52](#_Toc160103506)

[1.14 Introduction 52](#_Toc160103507)

[1.15 Literature Review 52](#_Toc160103508)

[1.15.1 Missing Data Mechanism 52](#_Toc160103509)

[1.15.2 Methods for Handling Missing Data 53](#_Toc160103510)

[1.16 Assumptions 55](#_Toc160103511)

[1.17 Implementation of Missing Data Handling Techniques 56](#_Toc160103512)

[1.18 Limitations of Different Missing Data Handling Techniques 58](#_Toc160103513)

[1.18.1 Listwise deletion 58](#_Toc160103514)

[1.18.2 Linear Regression 58](#_Toc160103515)

[1.18.3 Association Rules 58](#_Toc160103516)

[1.18.4 Mode Imputation 58](#_Toc160103517)

[1.18.5 Multiple Imputation 59](#_Toc160103518)

[1.19 Results 59](#_Toc160103519)

[5. DEVELOPED PREDICTION APPLICATION 61](#_Toc160103520)

[1.20 Objective 61](#_Toc160103521)

[1.21 Assumptions 61](#_Toc160103522)

[1.22 Booking Count Prediction 61](#_Toc160103523)

[1.23 Advertisement Effect Prediction 62](#_Toc160103524)

[1.24 Prediction of Actual Attendees 62](#_Toc160103525)

[1.25 Graphical User Interface 63](#_Toc160103526)

[6. FUTURE WORK 64](#_Toc160103527)

[1.26 Data Analysis 64](#_Toc160103528)

[1.27 Prediction Models 64](#_Toc160103529)

[1.28 Imputation 64](#_Toc160103530)

[1.29 Classification Models 65](#_Toc160103531)

[1.30 User Interface and Final Predictions 65](#_Toc160103532)

[7. CONCLUSION 66](#_Toc160103533)

[8. REFERENCES 67](#_Toc160103534)

# ABSTRACT

This report encompasses the development of a prediction model for conference bookings. Data from seven previous conferences is analysed and certain conclusions are drawn from it. Several machine learning prediction models are considered for the task and literature review is conducted to compare the models before implementation of each model to compare quantitative results. Based on the accuracy of the models, the best prediction model is chosen for the application. Imputation is conducted to fill out missing values in the conference files. This imputed data is used to train a classification model to make predictions of actual attendance based on the booking count. The selected prediction model is finally trained on the entirety of available data for each target audience to make predictions on new data input by the user for future conferences. A user interface is developed for receiving input from the user. Finally, a number of recommendations are provided for further development of the application.

# DATA ANALYSIS

## Introduction

Data from seven previous conferences was provided to train a prediction model and make it capable of predicting booking counts for future conferences. Before proceeding with building a model, analysing the provided data is essential to understand the scope of the application and make necessary assumptions and conclusions from the available data.

## Assumptions

For an application where raw data is provided and communication with organizers who compiled the data is not possible, certain assumptions must be made to make effective data analysis possible.

* The data is assumed to completely capture all booking instances for the conferences.
* The booking process is assumed to be online only, and the form to make a booking is assumed to be launched just before the first available booking date for each conference.
* An advertisement campaign is assumed to have been launched right before the start of the booking process. The effects of this campaign cannot be quantified and are assumed constant across all conferences.
* For dates containing no booking records in the provided data, it is assumed that no bookings were made.
* For weeks with an unusually high number of bookings, if the booking count exceeds a certain threshold, it is assumed that an advertisement campaign was launched on the first day (Monday) of that week. Any variations in booking count below the threshold are assumed to be due to natural randomness.
* The type of the advertisement campaign is beyond the scope of this project. It is assumed that all advertisement campaigns for a conference carry the same impact regardless of closeness to the event date.
* The effect of an advertisement campaign is assumed to wear off by the end of the week in which it was launched. If subsequent weeks have booking counts higher than the threshold value, it is understood that new advertisement campaigns were launched in each of those weeks.
* The bookings made on or after the event date are assumed to be for participants who gained access to the conference materials in some recorded format.
* Effects of other external factors (such as the COVID-19 pandemic in 2021) are ignored.

## Important Statistical Metrics

For data analysis, certain statistical terminologies need to be explored as they provide useful insights about the data and allow for important conclusions to be drawn. These statistical metrics are discussed in the following sub-sections.

### Arithmetic Mean

Arithmetic mean is the average value of a dataset. This is simply calculated by a sum of all the values in a list divided by how many values there are. This sum will result in a value which is the mean.

*Arithmetic Mean = sum of values / number of values*

The mean is an integral metric for statistics, experiments, and data analysis. (Gajendrakar, 2022)

### Standard Deviation

Standard Deviation is just the measure of how spread-out numbers are. Variance is the average of the squared differences from the mean. The square root of the variance gives standard deviation.

Standard Deviation is used in many areas. One of these is in statistics. In statistics, it measures the spread of data points from the mean and helps understand the distribution of a dataset. (Agarwal, 2019)

The formula for standard deviation is given as:

Where:

σ = the standard deviation

μ = the mean of the population

xi = each value from the population

N = size of the population

It is important to note that the denominator in the formula is N if the dataset is the entire population and is N-1 if the dataset is a sample of the population. In the case of the conference data, the data analysis is performed for the entire population of data.

### Z-Score Calculation

The Z-Score is the position of a raw score in terms of its distance from the mean when measured using standard deviation units. The Z-Score is positive when the value is above the mean and negative when it’s below the mean. (Mcleod & Saul, 2023)

Z-Scores are important because they can help to calculate the probability of a score occurring within a standard normal distribution. It can also be useful to compare two scores from different samples.

The formula for this is:

*Z = x – μ / σ*

Where:

X = the number of registrations for that week

μ = the mean of all registrations for all weeks in the dataset

σ = the standard deviation calculation for the dataset

Using this calculation will result in the Z-Score for each raw score in the dataset.

## Approach for Statistical Analysis

The dates in the conference data files are used to determine week numbers from the start of the booking process. The bookings are then grouped by week numbers to get the booking count per week.

The mean and standard deviation of the bookings per week are then determined for each dataset which help to understand what the typical booking activity is like and how variable it is. The Z-Score is then calculated for each week in the dataset to identify how much a week’s booking amount deviates from the mean.

Then, the statistical analysis is visualised by plotting the weekly booking amounts. It marks weeks with booking amounts above the threshold to show what weeks have significantly higher booking activity which could be a sign of a possible advertising campaign.

## Threshold Calculation

The purpose of calculating the threshold is to identify what weeks have significantly higher booking activity compared to other weeks. It identifies the outliers in the weekly booking count data that have remarkably high bookings. These weeks are then assumed to be affected by advertisement campaigns.

The way the threshold is calculated is by taking the mean of weekly bookings and adding the twice of the standard deviation. This means that any data outside the first two standard deviations will be understood to be affected by advertisements. Such data would have a Z score of more than 2 as well.

The formula for this is:

*T = μ + 2σ*

Where:

μ = the mean of week-wise bookings for the entire dataset of a conference

σ = the standard deviation of week-wise bookings for the dataset

## Significance

The statistical analysis of the data presents the raw information in a more interpretable format in the form of weekly booking counts. This allows for an in-depth understanding of booking patterns by showing the stakeholders how booking amounts can change over time as well as the significance of any changes relative to the usual patterns.

Using standard deviation and Z-Scores allows for the identification of any significant changes in booking activity, making it easier to spot trends and any outliers (which are deemed as advertisement weeks).

This analysis is crucial for the development of a prediction model because it highlights periods of high or low demand outside of the usual demand, helping to infer where advertisements were launched in the booking phase of the conferences. With this information, a prediction model can make more accurate predictions.

## Data Analysis for Each Conference

### D19

A graph with numbers and lines

Description automatically generated

For the D19 dataset, the mean bookings per week were found to be 62. The threshold for advertisements was found to be 187 bookings per week. In this dataset, there is only one week that climbs above the threshold line and is therefore the only week understood to be affected by an advertisement campaign.

### D21

A graph of a graph

Description automatically generated with medium confidence **** ****

For the D21 dataset, it is interesting to note that there were barely any bookings for most of the weeks in the start of the booking process. This may indicate a poor launch advertisement campaign, however, quantifying the impact of a launch advertisement is beyond the scope of this project. Another interesting observation is that all of the advertisement weeks can be found right at the end of the booking process, which may indicate a desperate attempt by the organizers to raise the booking count right at the end.

### GP21

A graph with numbers and lines

Description automatically generated

For the GP21 dataset, a trend quite like that of D21 can be seen, where the launch advertisement seems to have failed. Therefore, several advertisement campaigns had to be launched later to increase the booking count. Again, the bulk of the advertisement weeks seem to appear at the end of the booking timeline, indicating possibly poor planning on part of the organizers. A predictive model would help organizers keep track of their expected booking counts and launch advertisements sooner rather than waiting for the last few weeks.

These are the chart and table for the GP21 data set. The first week as only a few bookings but then after that, there are no booking at all for many weeks. There is then a peak that exceeds the threshold value which could be because of an advertising campaign. Then a few weeks later there is another few peaks above the threshold value.

### MSE21

A graph of a graph

Description automatically generated with medium confidence 

For the MSE21 dataset, the bookings were spread out more evenly across the registration weeks, and there seem to be no weeks without any bookings. It is interesting to note that the weekly booking count did not cross the threshold even once for this dataset, indicating that the bookings took place more regularly and did not need to be influenced by an external factor like an advertisement campaign. So, there were no advertisement campaigns launched for this conference once the registrations were opened.

### NP21

A graph of a number

Description automatically generated with medium confidence 

For the NP21 dataset, it seems clear that the organizers were more proactive and immediately launched advertisement campaigns once they saw the number of bookings in the first week were low. This seems to have set them back on track to meet their target bookings because after these two advertisement campaigns, there were several weeks of inactivity and no further advertisement campaigns were launched, until the very end when yet another advertisement campaign was launched.

This is the graph and table for the NP21 dataset. The first couple of weeks have high bookings which are above the threshold value which shows that a marketing campaign has likely been running. It then drops off and there are very low bookings for most of the weeks after the initial peaks. Near the end of the registration period the bookings start to climb again and exceed the threshold value once again. This shows there was likely another advertising campaign near the end of the registration period.

### SRM22

A graph of a bar graph

Description automatically generated 

For the SRM22 dataset, the organizers again seemed to be proactive and launched an additional campaign as early as the second week of the registration timeline. Afterwards, no further advertising campaigns were launched, indicating that the target bookings were achieved well before the event date.

### SRM23

A graph of blue and white bars

Description automatically generated with medium confidence

In the SRM23 file, the pre-launch advertisement campaign seems to have failed, leading to almost no bookings in the first few weeks. The organizers therefore launched an advertisement campaign soon enough to increase the booking count, and a second one a few weeks later.

## General Observations

The analysis of the graphical representation of weekly booking count date for each file reveals that there are many weeks that are above the threshold value. These weeks are assumed to have had a marketing campaign active. Any peaks below the threshold value are random spikes and not attributable to an advertisement campaign.

Tabular data for Z scores for each week was also attached to give an indication of how much the booking count of each week deviates from the mean booking count. Z scores above the threshold are highlighted in red. The values for the threshold, mean and standard deviation are also included in the tabular data for each file.

# COMPARATIVE STUDY OF PREDICTION MODELS

## Introduction

To be able to make a well-informed decision about the choice of machine learning model to use for the application, different options need to be explored. An initial literature review was conducted to check which models may be viable for the specific tasks needed in this project. These models were then explored in detail and implemented to compare the results and select the best option for the application.

## Literature Review

Before implementing a model, it is important to review the literature to get a better understanding of the model and how it works. This will enable effective implementation of the model and an expectation of how it might perform.

### Regression Models

Regression models are an example of supervised learning algorithms. They are trained using labelled datasets to study and map relationships between independent and dependent variables.

Once the mapping is achieved using sufficient training data, these models can then predict the dependent variables based on new values of the independent variables. Some popular regression models are discussed below. (Gallo, 2015)

#### Linear Regression

Linear regression is a statistical technique used to estimate the relationship between a dependent and one or more independent variables. A straight line, represented by the equation y = mx + c + e , is used to model the relationship between the independent and dependent variables. In the equation, m represents the slope, c is the intercept, and e is the error term. (Panchotia & Rajat, 2020)

The objective of linear regression is to find the linear relationship between the variables that best fits the data. Linear regression is performed to generalize the relationship between the variables, allowing the model to predict values of the dependent variable based on new input values of the independent variables. (Ogunleye, 2022)

A popular optimization approach for linear regression models is the Ordinary Least Squares (OLS) regression. The goal of this approach is to find the values of the regression coefficients (m and c) such that the sum of squared differences between the predicted and observed values is minimized.

Linear regression is a popular prediction model because it is simple and easy to understand. It is computationally efficient and can therefore handle large datasets. If the relationship between variables is close to linear, linear regression can effectively make predictions. OLS provides the best linear unbiased estimates of the regression coefficients. (Satyavishnumolakala, 2020)

This model assumes a linear relationship between the variables, so, in cases where this is not true, this model cannot capture the relationship properly. It may be too simplistic for some real-world applications and is highly sensitive to outliers in the data. This model also assumes homoscedasticity (i.e. it assumes variance remains same across the data). OLS linear regression may also overfit the training data if the dataset is not sufficiently large.

#### Elastic Net Regression

Elastic Net regression is a linear regression model that incorporates both the L1 (Lasso) and L2 (Ridge) regularization penalties to improve the OLS linear regression model.

Elastic net strikes a balance between the variable selection aspect of Lasso and the regularization and stability features of Ridge regression. This way, it addresses some of the shortcomings of the individual Ridge and Lasso regression algorithms .

The cost function of elastic net regression is a combination of the OLS term and the L1 and L2 regularization terms. The model attempts to determine the values of coefficients (hyperparameters) which would minimize the cost function.

Elastic Net regression offers the benefits of variable selection, stability, and robustness, and it can handle multicollinearity between the predictors. On the other hand, the incorporation of additional hyperparameters increases complexity of the model and necessitates careful tuning. The addition of the regularization terms also adds to the computational cost of the model. (Dhumne & Shruti, 2023)

#### Decision Tree Regression

Decision Trees can also be used to perform regression. A decision tree regression model can estimate a target variable on the basis of one or more independent features. This model uses the values of the features to recursively divide the dataset into subsets.

The entire dataset is represented by the root node, decision nodes represent the splits based on feature values, and leaf nodes are the terminal nodes which represent predicted values. (*What is a Decision Tree?* )

Decision trees are usually constructed from the top down; at every node, the data is split based on the feature values. This way, child nodes are created, and a tree-like structure is created. This process continues until the stopping conditions are met. These conditions may be maximum samples per leaf node, maximum depth, etc . (The Click Reader, 2021)

Decision trees make no assumptions of linearity in the variable relationships and can capture the non-linearity in these relationships. They are also robust to outliers in the data and are easy to visualize and understand. (Penkel & Engin, 2019)

However, this model can be prone to overfitting if the tree is deep. It is also unstable and small changes in the dataset can lead to different tree structures. Another factor to consider is that decision trees make locally optimal decisions, but these may not result in a globally optimal model.

### Time-Series Models

A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Time series analysis has two main goals.

First, it's about modelling the random processes that are behind the patterns we see in data collected over time. Second, it's used to make predictions about future data points based on the past data and possibly other related information. In simpler terms, it's like trying to understand the story behind a sequence of events and using that understanding to guess what comes next. (D.Cryer & Jonathan, 2008)

Examples of cases where a time series model may be used include annual rainfall in an area, monthly average temperatures in an area, monthly sales of a product, GDP etc.

Time series offers the advantages of prediction/forecasting data with a time element, helps in monitoring features such as variability, seasonality and trends in data and helps to draw conclusions on data. (Bhattacharyya, 2023)

#### ARIMA

The Autoregressive Integrated Moving Average Model, or ARIMA, is a more advanced version of the ARMA model. It combines two approaches: Autoregressive (AR), which predicts future values based on past values, and Moving Average (MA), which uses past forecast errors. Together, these make a model that can analyse and predict trends over time. As acronym indicates, ARIMA (p, d, q) captures the key elements of the model: (S. Siami-Namini et al., 2018)

Autoregression (AR) is a type of model that predicts future data points by looking at past ones. It examines how current values are related to previous values (called lagged observations) up to a certain number (p).

Integrated (I) refers to making a time series stable (stationary) over time by calculating the differences between data points at various times (d). This helps in preparing the series for accurate forecasting. (Etienne, 2019)

Moving Average (MA) is a method that focuses on the relationship between current observations and the errors from previous forecasts. It uses this information to smooth out the data by averaging over several past observations (q).

Some advantages of ARIMA are that it relies only on past data from the time series to make future predictions and performs well on short term forecasts. It can handle time series data that change over time (non-stationary data). (Bora & Neha, 2021)

Disadvantages of the model are that it's hard to accurately forecast when trends will change direction. Choosing the (p,d,q) values for the model involves a lot of guesswork and subjectivity. It is computationally expensive and is not suitable for time series with seasonal patterns.

#### SARIMAX

SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors) is an extension of the ARIMA model that takes into account seasonality and exogenous variables. SARIMAX models exhibit exceptional forecasting performance, making them one of the most used statistical models for forecasting. (Forecasting SARIMAX and ARIMA models.)

SARIMAX (p,d,q) (P,D,Q)s are the key elements of the model. The components of the SARIMAX model are similar to that of an ARIMA model with an additional part for the seasonality denoted by (P, D, Q) s where s stands for the seasonal period. (Ampountolas & Apostolos, 2021)

SARIMAX as model has some advantages which include how easy it is to understand SARIMAX models are built on clear math rules, making them simpler to grasp. The model's parts show how past data connects to future guesses. (S & Rahul, 2023)

It deals with seasons, SARIMAX models can deal with data that follows a yearly or seasonal trend. They can predict patterns that repeat over time, making them good for forecasting such data. The model also does not need a huge amount of data, SARIMAX can work okay with not too much data. It mostly looks at past info and does not need too many examples to make good guesses.

The limitations of the model are parameter selection where figuring out SARIMAX parameters can be hard and can take a lot of time. It needs domain knowledge and many iterations to find the best model parameters. Moreover, SARIMAX is not designed to handle intricate feature engineering tasks. Its primary focus is on past values of the main data and exogenous factors only.

#### Exponential Smoothing

Exponential smoothing methods are appropriate for non-stationary data (i.e., data with a trend and seasonal data). Smoothing methods use weighted averages to make forecasts, where past observations influence the prediction. The weights can be the same for all (like in a moving average) or decrease exponentially, meaning more recent data has more influence than older data. (Etienne, 2019)

Types of exponential Smoothing include Simple Exponential Smoothing (SES). It is used when there are few data points, irregular data, or no seasonality or trend. SES is computationally efficient and easy to implement, making it suitable for real-time forecasting or situations where data availability is limited.

Holt’s Linear Smoothing is another type that is used when there are trends in data, but no seasonality. Double Exponential Smoothing extends simple exponential smoothing by incorporating a trend component and the level component. It is suitable for time series data that exhibits a trend. (Malkari & Nikhil, 2023)

Holt's Linear trend method assumes a trend that keeps increasing or decreasing at the same rate into the future, which isn't realistic for long-term forecasts since trends don't usually continue indefinitely. Holt’s Damped Trend method addresses this by introducing a damping parameter, making the trend eventually level off to a constant value, effectively smoothing out the future trend to make it more realistic.

Exponential smoothing may struggle with time series data that have complicated patterns or unexpected changes, like sudden shifts in trend or level, outliers, or sharp changes in seasonal patterns. The accuracy of forecasts can greatly depend on the selection of smoothing parameters (α for the level, β for the trend, and γ for the seasonality). Determining the best values for these parameters often involves trial and error.

Exponential smoothing is computationally efficient, adaptive and works well with data that has trend or seasonality.

### Classification Models

Classification models are a type of machine learning model used to categorise data into different classes or groups based on input features. These models are widely used in various fields, including image recognition, spam detection, medical diagnosis, and financial forecasting.

#### Use Of the Classification Models

For the task at hand (prediction of booking count), we believe that none of the classification models would be applicable. The main reason for this would be that the prediction we are trying to make, the number of bookings, is a continuous variable, not a categorical variable. However, this model has the potential to be used for another part of the application.

The primary goal of a classification model in our case will be to use the previously imputed dataset to train the new model to predict the chance, between 0 and 1, of participants attending the event. The classification models will use past observations from the dataset, specifically the created date column, which states the date on which the customer has booked their ticket(s).

This information could then be used to predict the percentage of those who have booked tickets for an event on a specific date and then actually turn up. We will be looking at multiple different models, to figure out what is our best approach.

#### Decision Trees

Decision Trees split the dataset continuously on feature values (e.g. if the guest turned up or not), to create a tree-like structure. At each branch or node of the tree, a decision is made based on a feature’s value, leading to a binary split.

This method is quite intuitive, meaning it could possibly handle the categorical features that the dataset portrays. A decision tree is easier to explain to an audience, as its layout is easy to visualise and follow. Non-linear features can be captured, and outliers don’t affect the model unlike the previous method. (Magee, 1964)

The main problems tend to be that the model overfits the data, which when you have lots of cells without data, the model will become skewed. Small variations in the data can lead to completely different tree structures, making them unstable and resulting in elevated levels of variance. An example would be predicting whether a customer will buy a product based on age, income, and purchase history.

#### Random Forest

Random Forest is similar to Decision Trees, the difference being that the random forest model constructs multiple trees during training. Each tree in the forest operates independently and outputs a class prediction. The mode is then taking from all of the outputs.

This model is significantly more accurate than its counterpart, by reducing overfitting due to taking the mode of all individual trees. This also makes the model significantly more robust to any noise or outliers, even missing values. However, due to the forest comprising of multiple decision trees, it’s harder to draw conclusions and interpret results behind individual predictions. (Breiman & L, 2001)

This model consumes larger quantities of memory due to there being multiple trees, so can slow down results, making the project less time efficient. The major negative of the model would be that it can be biased towards the majority class, which in our case would be ‘yes’.

The way to counteract this would be by implementing class weights or using sampling techniques. However, given the time constraints of the project, this may not be possible.

### Logistic Regression

For binary data such as the target variable (attendance status), there are only two options, such as spam/not spam, true/false, or in our case, yes/no. Given that this is non-negotiable, the choice of potential algorithms shrinks.

Logistic Regression is a linear model used for binary classification. It estimates the probability that a given input belongs to the positive outcome, yes, using the logistic function.

The function itself a mathematical S-shaped curve used to model growth processes; in this case machine learning for classification tasks. It is defined by the formula *f*(*x*)=, where *e* is the base of the natural logarithm, x0, is the value of the sigmoid's midpoint, and *k* is the steepness of the curve. This function outputs values between 0 and 1, making it particularly useful for representing probabilities in logistic regression. (Logistic Regression.)

An example application of this model is teachers predicting whether a student will pass or fail, based on previous results, attendance, and study hours. The major positive to this method would be the efficiency of the model, as training and prediction is fast.

Our dataset has a lot of missing data, so for the time being this model wouldn’t be the best choice. The missing data would be even more disastrous to the model if you consider them outliers, as this algorithm is disproportionately influenced by outliers, therefore affecting predictions.

## Assumptions

To implement prediction models on the available conference data, certain assumptions must be made. These assumptions are listed below:

* The booking count is only relevant before the event date. Any bookings made on or after the event date are assumed to be for online attendance or online viewing of recorded material. Therefore, these bookings are immaterial and irrelevant for the event organizers in terms of operational considerations.
* The effect of an advertisement campaign is assumed to be the same for all days of the week in which the campaign was launched.
* The effect of advertisements is assumed to be the same on all days and does not depend on other factors. For example, the effect of advertisements is assumed not to be different on weekdays vs weekends, and it is assumed not to be impacted by closeness to the event date itself.
* People from a particular target audience are assumed to show similar booking patterns across different conferences, so conference data can be grouped by target audience.

## Implementation

The primary objective of the project was to develop a predictive framework capable of forecasting event registrations for conferences tailored to distinct audience segments, including IT managers, education managers, property managers, and education property managers.

This forecasting aimed to analyze the patterns of registrations leading up to the start date of each conference, with the ultimate goal of predicting the total number of final registrations.

By understanding registration patterns and predicting total attendance, the application serves a strategic business function, allowing decision-makers to determine if and when further advertising is needed.

Accurate forecasts are essential for meeting attendance goals, optimizing marketing budgets, and improving the planning and execution of these events.

### Regression Models

The three regression models introduced in the literature review were implemented to evaluate their predictive performance for the project. To implement these models, it was important to first clean the data before each model could process it to make predictions. Once the data was prepared in the appropriate format, each model was implemented using the scikit-learn library in python and the results were compared.

For implementation of prediction models, the data was split into testing and training data randomly, such that 80% of the data was for training while the remaining 20% was for testing. This split was loosely driven by the Pareto principle.

#### Data Processing

Seven csv files were provided, each containing booking and attendance data for a unique conference. The conferences could be grouped based on the target audience, of which there were four in total: IT Managers, Property Managers, Education Property Managers, and Education Managers.

A fifth target audience was added, titled “Others”, which had the data of all of the files combined. This would allow for predictions of conferences for target audiences other than the original four listed above.

For the regression models, the booking date and booking reference for each row across all seven files was determined to be the relevant data. These columns were therefore imported into a data frame using the pandas library of python.

It was assumed that the booking process was active from the first date in each csv file up until the event date for each file. The dates missing from the csv file in this date range were therefore assumed to have had no bookings. Records for these missing dates were added to the data frames to ensure a consistent time range.

The entries (hereon to be referred to as bookings) on or after the event date for each conference were removed. These bookings are assumed to be of no practical significance to an event organizer.

Columns for weeks and days corresponding to the dates were then added. It was ensured that the weeks always start on a Monday regardless of the first date in a file. If the first booking was made on a day other than Monday, the program would select the Monday before the first booking day as the starting day for the first week. This week would be labelled as Week 1.

The bookings were grouped by week numbers to produce a count of bookings per week. The Z score for bookings per week was then determined to identify outliers. These were the weeks with Z scores greater than 2 and were deemed to have been affected by advertisements.

The machine learning models being used require a substantial volume of data to work effectively. Grouping the bookings by week reduced the size of data per target audience greatly. Therefore, day numbers were also determined, and the bookings were then grouped by the day numbers in a separate data frame.

Another column with binary values was added for Advertisement Status. The value would be 1 for the days that fell in the weeks with Z scores greater than 2. The data frames based on day numbers were then merged by target audience.

To implement the models, the days data frame was split into two data frames, one containing the day numbers and the advertisement status to form the feature matrix, and the other containing the booking counts.

#### OLS Linear Regression

The LinearRegression function from the scikit-learn library of python was used to perform OLS linear regression. 10-fold cross validation was carried out to estimate the accuracy and fit metrics of the model by taking the mean of the metrics found from all the folds.

The cross-validation code performs the testing and training data splits 10 times, taking the data from different sections of the dataset for each split. To plot the developed regression model against the actual data, a testing and training data split was manually performed at the end to prepare the data for the plot.

For this final split, the data was stratified on the Advertisement Status column to ensure both the training and testing datasets had days where the advertisement status was 1 (in case the main data frame had any advertisements).

The intercept term was set as 0 for the OLS linear regression model on the logical grounds that Day 0 i.e., the day before the booking process was started, should have 0 bookings.

After training the model on the training dataset, it was tested against the entire testing data, testing data with Advertisement Status as 1, and testing data with Advertisement Status as 0. The latter two tests were performed to determine the effectiveness of advertisements for each target audience.

#### Elastic Net Regression

For elastic net regression, the ElasticNetCV function from scikit-learn was used. This function has a built-in cross validation feature which was used to perform 10-fold cross validation. The goal for cross validation for this model was to determine the best value of alpha and the L1 ratio (hyperparameters) for each target audience data frame.

Alpha represents the amount of penalization that will be applied to the model. The L1 ratio determines the relative effect of the lasso and ridge penalties. If L1 ratio is 1, the penalty will be a lasso penalty, while for an L1 ratio of 0, the penalty will be a ridge penalty. For any other value, it will be a combination of the ridge and lasso penalties.

While adding another layer of cross validation to get more accurate metrics for the model is possible, it would add unnecessary complexity to the model. Tuning the hyperparameters with cross validation already maximizes the performance of the model in terms of accuracy (adding small bias to reduce variance through the best values of the regularization terms).

The second layer of CV would make the metrics for the model more robust, but it would increase computational costs. Therefore, the metrics obtained from just one prediction can suffice for this case as well.

The intercept term for the model is set to 0, as for the OLS regression model. During the splitting of data, the data was stratified on the Advertisement Status column just like in the OLS model. As with the OLS model, the model was tested against all testing data, and testing data with Advertisement Status as 1 and then 0, to establish the effectiveness of advertisements.

#### Decision Tree Regression

Decision trees are notorious for overfitting, so, to prevent this, the GridSearchCV function from scikit-learn was used to perform 10-fold cross validation to select the best combination of parameters for the decision tree. The parameters in question are the maximum depth of the tree, minimum number of samples for a split on an internal node, and minimum number of samples required on a leaf node.

Grid search cross validation ensures that the selected parameter values perform well on unseen data. This helps to strike a balance between overfitting and underfitting for decision trees.

As a proof of concept, cross validation was performed again on the testing and training split to find the mean values of the model accuracy metrics. The argument about unnecessary complexity made against this approach in the previous section still stands and this step was taken in this model purely for testing.

The data was stratified during splitting as for the previous models. However, the concept of intercept does not generally apply to regression trees because they try to approximate the data locally. As with the other two models, the effect of advertisements was also quantified.

### Time Series Models

The three time-series models introduced in the literature review were implemented to evaluate their predictive performance for the application.

#### Data Preprocessing

Data preprocessing played a pivotal role in setting the foundation for accurate and meaningful analysis. The initial step involved the combination of datasets corresponding to the target audience: `D19` and `D21` datasets for IT managers, `NP21` and `GP21` for property managers, and `SRM22` and `SRM23` for education managers.

This aggregation was essential to ensure that the analysis comprehensively represented the registration trends across different audience segments. Following the combination of these datasets, a crucial preprocessing step was the calculation of daily registrations. This involved transforming the raw registration data into a daily and weekly format, allowing for a more granular and insightful analysis of registration patterns over time.

An innovative aspect of the data preprocessing was the calculation of the Z-score for week-wise registrations. This statistical measure was used to identify significant deviations from the average registration rate, which were indicative of external influences on registration behavior, such as advertising campaigns.

If the Z-score exceeded a threshold of 2, it was inferred that an advertising campaign had likely taken place, influencing registration rates. To integrate this insight into the predictive modeling, a binary indicator was created, where `0` represented the absence of an advertising campaign, and `1` indicated its presence.

This binary indicator served as an exogenous factor in the forecasting model, acknowledging the impact of advertising on registration trends.

The final data set not only included the calculated daily registrations and the binary indicator for advertising campaigns but also ensured that the full date range of the data was represented.

#### Modelling approach

The modeling approach within the project was both comprehensive and methodical, employing three distinct forecasting models to identify the most effective methodology for predicting event registrations: ARIMA, SARIMAX, and Exponential Smoothing.

Each model was chosen for its unique capabilities in handling time series data, with a particular focus on capturing underlying trends, seasonal patterns, and the impact of external factors.

The ARIMA and SARIMAX models were both configured using an automated process to identify optimal parameters. This process, known as auto\_arima, systematically tested a range of parameter combinations to determine the best settings for each model based on the historical registration data. This automated parameter selection was crucial for enhancing model accuracy without manual trial and error.

A distinguishing feature of the SARIMAX model, in contrast to ARIMA, was its incorporation of advertisement campaigns as an exogenous factor. This model was specifically designed to assess how external advertising efforts influenced registration numbers, providing valuable insights into the effectiveness of marketing strategies on attendee numbers.

For model training and evaluation, the dataset was split into an 80/20 partition, with 80% of the most historical data allocated for training and the remaining 20% of the most recent data reserved for testing. This split ensured that the models were trained on a substantial historical dataset, allowing them to learn the underlying patterns effectively, while the recent data served as a robust benchmark for evaluating the models' forecasting performance.

### Classification Models

The three classification models introduced in the literature review were implemented to evaluate their predictive performance for predicting the actual attendees for a conference.

#### Data Processing

Implementing the classification models on the conference dataset provided a predicted number of participants who attended after making a booking. For this, the data needed to be split into training data and a testing data set.

We believe that the best split is a 20% testing and 80% training split, loosely driven by the Pareto principle, but that is just a thumb rule used by practitioners. This split is however backed up by multiple studies. (V. Roshan & Joseph, 2022)

The main hurdle in implementation was the fact that many booking records had missing data for the Attendance column. This data was essential to make the actual attendance predictions and was therefore imputed. The imputation process will be explained in the next chapter.

Once the data was imputed and then split, the model was then trained to understand the relationship between the variables, which was done using Python’s built in functions. Each model's accuracy was then evaluated and compared.

## Results

The results for all the aforementioned prediction models were compared to determine the best-performing model from each category. The best models from the regression and time-series models would then be compared with each other to make the final choice of the booking prediction model for the application.

For the regression and time-series models, the accuracy is expressed in terms of the relative mean absolute error to enable a comparison for all the models.

The best-performing model from the classification models would be selected for use in prediction of attendee counts.

### Regression Models

The performance of the regression models was tested for each target audience. The important metrics were fit quality (given by the R-square score), accuracy (given by the % relative mean absolute error), and the advertisement effect (expressed as a percentage).

The fit quality describes how well a particular model fits to the testing dataset. A score of 1 indicates a perfect fit, if there can be such a thing, and a score of 0 (or sometimes even a negative score) indicate a poor fit.

The size of available data plays a crucial role in how well a model fits the data, and so, this metric should be taken with a grain of salt.

The accuracy score is given by the mean absolute error (MAE) normalized by the range of values of the target variable in the testing dataset. The reason for choosing the MAE instead of the mean squared error (MSE) as the accuracy metric is that MSE places a significantly larger penalty on large errors.

These errors are usually caused by outliers in the data. As the datasets are inherently laced with outliers because of the advertisements (as dictated by the assumptions), and these outliers cannot be removed from the data as they provide a useful insight, penalizing large errors more severely than small errors feels counterintuitive.

In contrast, the MAE penalizes all errors equally, and therefore is deemed as the more appropriate metric for the accuracy of the models.

The advertisement effect % is a useful metric in terms of making predictions for the end user but serves no particular purpose for comparison of the models, other than the fact that a ludicrously high effect % can indicate that a model is overpredicting the impact of advertisements. This would indicate overfitting for the model.

A summary of these metrics for each target audience is presented in the following table for each regression model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Prediction Model** | **Target Audience** | **Fit Quality** | **Accuracy (%)** | **Advertisement Effect (%)** |
| **OLS Linear Regression** | IT Managers | 0.12 | 90.74 | 607.15 |
| Property  Managers | 0.13 | 93.31 | 1136.06 |
| Education Property Managers | -0.38 | 70.7 | 0 |
| Education Managers | -0.65 | 80.63 | 365.02 |
| Others | 0.12 | 94.04 | 378.3 |
| **Elastic Net Regression** | IT Managers | 0.09 | 90.99 | 78.55 |
| Property  Managers | 0.16 | 92.18 | 116.5 |
| Education Property Managers | -0.08 | 79.76 | 0 |
| Education Managers | 0.08 | 94.28 | 140.97 |
| Others | 0.07 | 95.32 | 22.18 |
| **Decision Tree Regression** | IT Managers | 0.05 | 92.36 | 844.65 |
| Property  Managers | 0.02 | 94.59 | 1572.12 |
| Education Property Managers | -0.37 | 70.71 | 0 |
| Education Managers | -1.14 | 79 | 376.01 |
| Others | 0.16 | 94.96 | 432.48 |

The low fit quality values may seem alarming at first, however, the reality of the matter is that the data available was simply not sufficient to fit a regression model with a good fit quality. Another reason for this was the sparsity of the data i.e. the number of entries (days) with no bookings. The fit quality is expected to significantly improve with an increase in available data.

Despite the poor fit quality across the board, the models do an acceptable job at accurately predicting the number of bookings for new data. A comparison of accuracy scores shows that Elastic Net Regression is the best-performing model.

An analysis of the Advertisement Effect scores also backs up the choice of Elastic Net Regression because the other models are predicting unrealistically high scores for the Advertisement Effect. The reason for these unrealistic predictions is likely to be overfitting in the models.

Elastic Net Regression, due to the inclusion of regularization terms, can better account for the booking count spikes due to advertisements in the training data and therefore present better predictions and more realistic scores of Advertisement Effect.

Graphical representation of each model for each target audience is presented in the following sub-sections.

#### OLS Linear Regression

A graph of a graph showing the results of a performance

Description automatically generated with medium confidence

A graph of a graph

Description automatically generatedA graph with black dots and a blue line

Description automatically generatedA graph with black and red dots

Description automatically generatedA graph with red and blue dots

Description automatically generated

The red points represent the predicted data points with Advertisement Status as 1. These points correspond to spikes in the regression line. This makes sense because when an advertisement is deemed to be active, the booking count should be higher on that day. Higher the Advertisement Effect score determined by the model for a target audience, larger the spikes on the graph.

#### Elastic Net Regression

A graph of a number of people

Description automatically generated with medium confidenceA graph of a number of people

Description automatically generated with medium confidenceA graph of a graph with a line

Description automatically generated with medium confidenceA graph of a graph showing the results of a graph

Description automatically generated with medium confidenceA graph of a graph with black dots

Description automatically generated with medium confidence

The red dots are the predicted data points where Advertisement Status was active. It is interesting to note that for some target audiences, there were no spikes despite the presence of advertisements in the data points. This is because in these models, cross validation determined the value of alpha to be significantly higher than for the other target audiences. Alpha dictates the size of the regularization penalty, and a higher alpha means that the coefficients become much smaller.

This helps prevent overfitting and make more realistic predictions. It is expected that if more data were available, the spikes in the other graphs would also eventually be straightened by the elastic net model.

#### Decision Tree Regression

A graph with a line and a black dot

Description automatically generated with medium confidenceA graph of a tree

Description automatically generatedA graph with black dots and a blue line

Description automatically generatedA graph of a graph

Description automatically generated with medium confidenceA graph of a tree

Description automatically generated with medium confidence

The regression trees are prone to overfitting. Despite the use of ten-fold cross validation to minimize overfitting, it is apparent from the results of the regression tree model for each target audience that this model is indeed overfitting to the data.

The rises and falls in the regression line can be explained by the fact that decision trees treat the data points locally. So, the regression line may experience a “spike” or a rise even when the advertisement status is 0.

#### Conclusion

Based on the results produced by the regression models, Elastic Net Regression stands out as the best-performing model. Therefore, this model will be selected from the regression models and will be compared with the best choice from the Time Series models to determine the best prediction model for this project.

### Time Series Models

For the evaluation of the forecasting models within the project, the Mean Absolute Error (MAE) was the important parameter. It is pivotal in quantifying the accuracy of the predictive models, providing a clear measure of their predictive performance.

Mean Absolute Error (MAE) measures the average magnitude of errors in the predictions i.e. how far off your predictions are on average, in the units of what you are predicting. A lower MAE is naturally better. The MAE was normalized using the range of test data and then was expressed as a percentage for ease of interpretation.

The percentage accuracy is expressed in terms of the MAE. The results by target audience are shown in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Target Audience** | **MAE%** | **% Accuracy** |
| **ARIMA** | IT Managers | 15.74% | 84.26% |
| Property Managers | 13.68% | 86.32% |
| Education Managers | 17.20% | 82.80% |
| Education Property Managers | 77.48% | 22.52% |
| General | 14.75% | 85.25% |
| **SARIMAX** | IT Managers | 16.21% | 83.79% |
| Property Managers | 12.95% | 87.05% |
| Education Managers | 16.74% | 83.26% |
| Education Property Managers | 77.48% | 22.52% |
| General | 13.46% | 86.54% |
| **Exponential Smoothing** | IT Managers | 644.54% | - |
| Property Managers | 238.48% | - |
| Education Managers | 876.86% | - |
| Education Property Managers | 2280.93% | - |
| General | 422.99% | - |

According to the table above across all the target audiences, the SARIMAX model outshines both ARIMA and Exponential Smoothing in forecasting accuracy for event registrations, as evidenced by its markedly lower RMSE and MAE values.

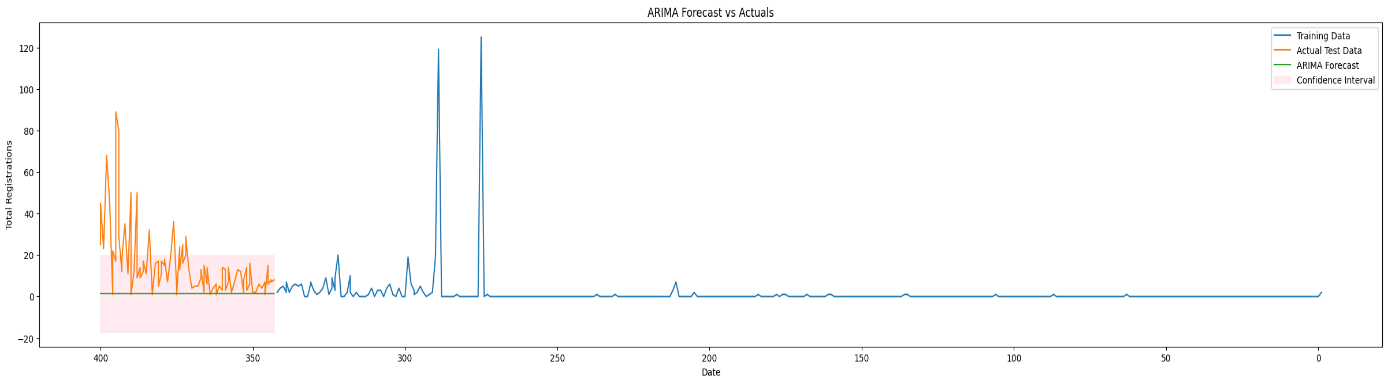
This indicates not only its ability to closely predict actual registration numbers but also the significant value added by considering external factors like advertising campaigns in the model.

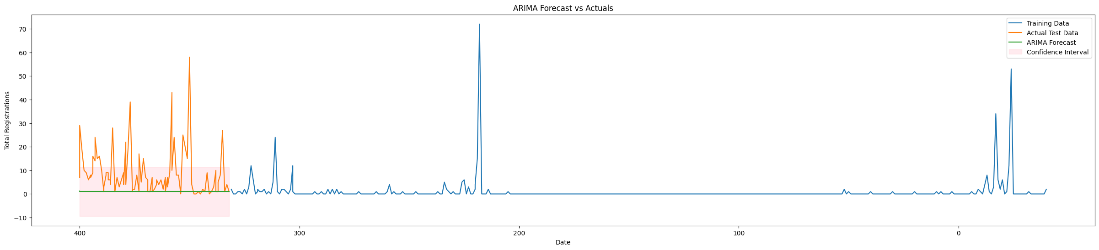
The ARIMA model, while more accurate than Exponential Smoothing, does not perform as well as SARIMAX, underscoring the importance of external factors in enhancing forecasting accuracy.

The Exponential Smoothing model's poor performance, particularly its high MAE percentage, suggests it may not be suitable for this specific forecasting task, likely due to its inability to incorporate external factors effectively. The exponential smoothing model’s poor performance made it difficult to compute the accuracy score due to significantly high MAE scores.

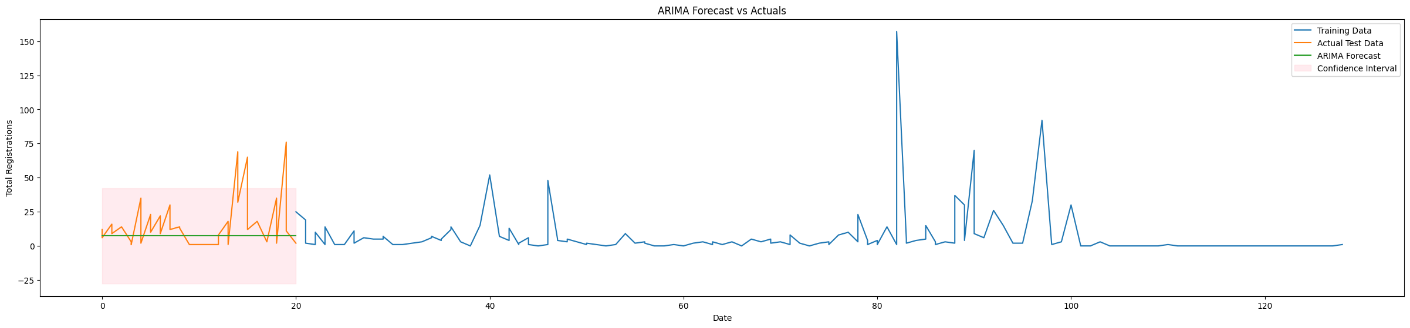
The graphs below depict the results of the forecasts:

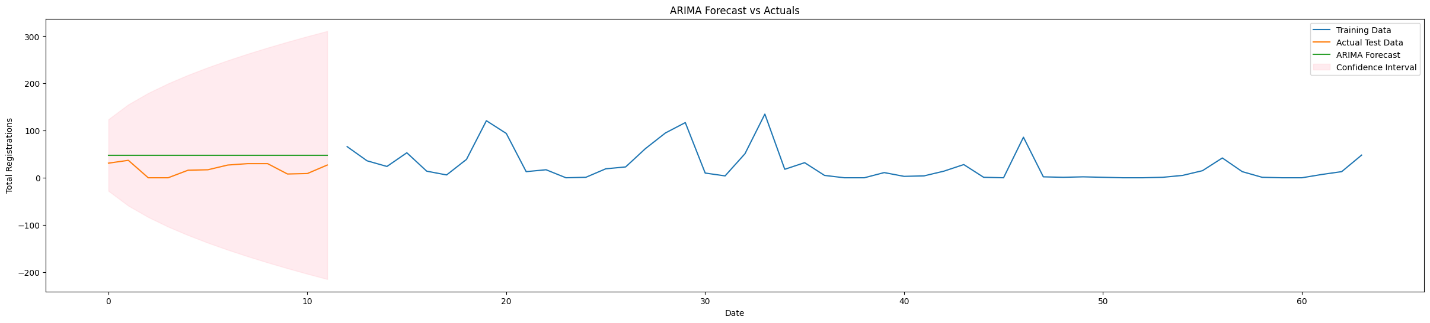
#### ARIMA

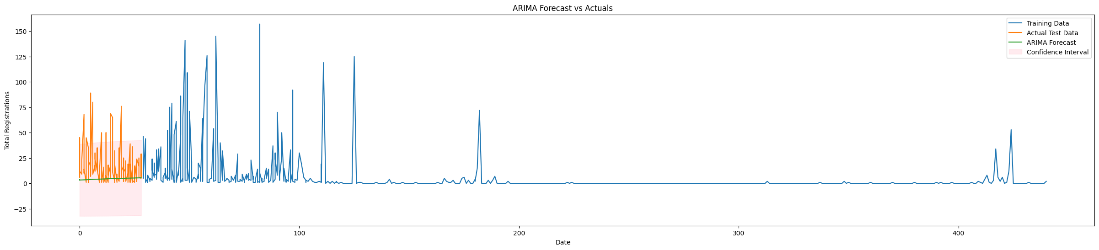
IT Managers



Property Managers

Education Managers

Education Property Managers

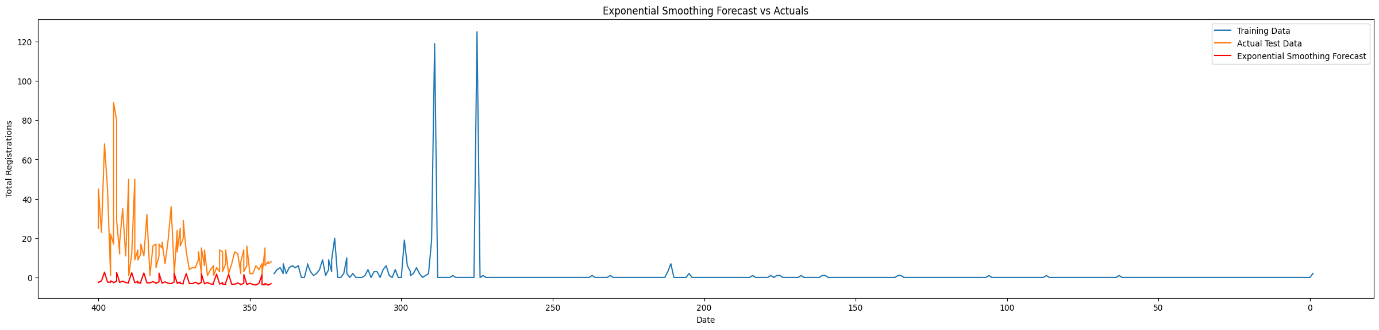


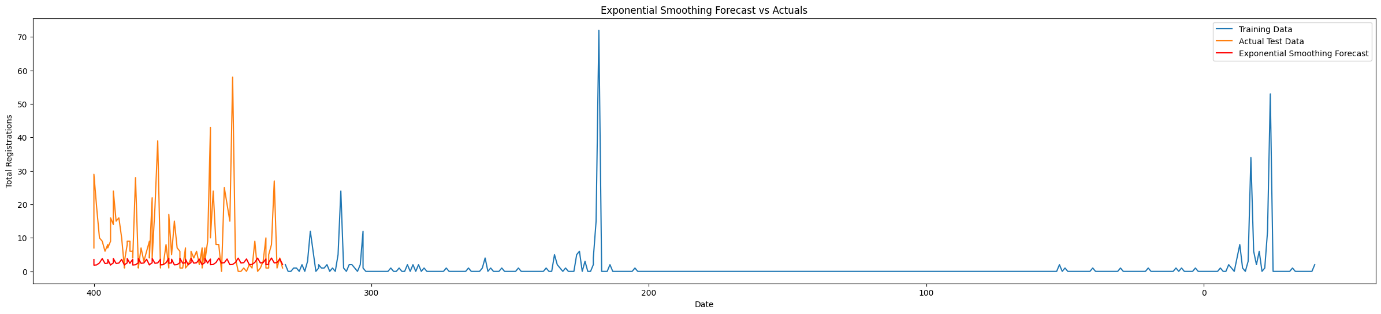
General

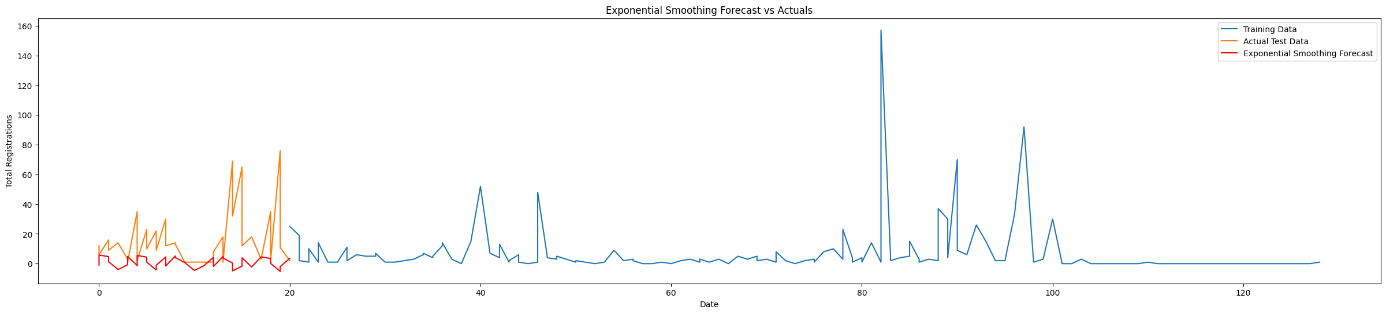
In conclusion, the ARIMA model demonstrates a limited capacity to accurately project future values as evidenced by the graphs. Although the model achieves, relatively commendable MAE and RMSE values, indicating a degree of accuracy in predicting data points close to the actual values, it falls short in capturing the underlying trends within the data.

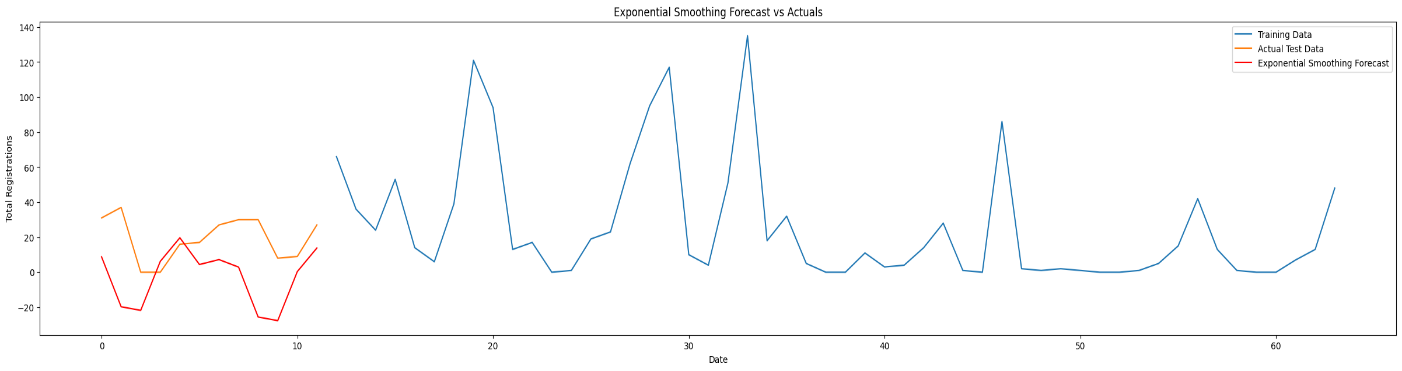
This can be caused by the limitation of the model to incorporate external factors such as marketing campaigns. Therefore, while ARIMA model can be useful in time series analysis of the data, its limitations in trend prediction highlight the need for model adjustments or consideration of alternative modelling approaches.

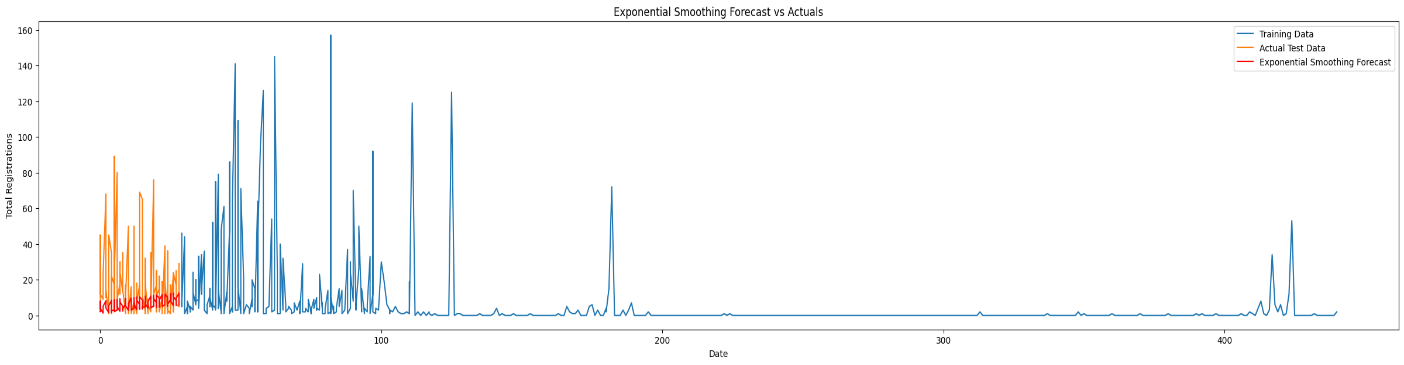
#### Exponential Smoothing

IT Managers

Property Managers

Education Managers

Education Property Managers

General

As seen from the graphs and RMSE and MAE values, the exponential smoothing model falls short in effectively forecasting the data. This inadequacy stems primarily from its inability to accurately capture the underlying trends within the data, as well as its failure to incorporate external factors such as advertisements.

The method’s simplistic approach, while beneficial for certain types of data, evidently does not suffice for the complexity and dynamics of our data. Given these shortcomings, more sophisticated modelling techniques would better suit the objective.

#### SARIMAX

1. IT managers

A graph with blue lines

Description automatically generated

1. Property managers.

A graph with blue lines

Description automatically generated

1. Education managers

A graph showing a graph

Description automatically generated with medium confidence

1. Education Property managers

A graph with blue lines

Description automatically generated

1. General

A blue line graph on a white background

Description automatically generated

In conclusion, the SARIMAX model, used for forecasting event registrations, demonstrates good predictive performance in terms of error metrics, with low mean absolute error (MAE) values indicating accurate predictions on the test data.

However, the graphical representation reveals that while the model is generally reliable, it struggles to capture the true variability and spikes in registration data. The actual registrations exhibit significant fluctuations that the model does not fully replicate, particularly for outlying (advertisement) peaks.

While the model provides a solid base for understanding general trends and preparing for average registration volumes, it may not be sufficient for precise resource allocation during peak periods, additional model refinement may be required to improve the accuracy of predictions.

#### Conclusion

Of the Time Series models, SARIMAX was found to be the best choice, however, it struggled to capture some of the trends of the data.

### Classification Models

To test the accuracy of the fully imputed datasets, we use accuracy metrics such as the F1 score and Accuracy score.

However, the results for both Logistic Regression and Decision Trees were remarkably similar, so to gauge which model shows higher accuracy, the Decision Matrix is used to show the breakdown of which imputations were correct, in terms of yes and no.

This left Logistic Regression as the better model, as it obtained a higher amount of correctly picked no values. It's worth noting that the accuracy metrics for Decision Trees and Random Forest were identical to two decimal places, but the matrix result was worse for the Random Forest model.

A screenshot of a white sheet with black text

Description automatically generated

**FINAL IMPLEMENTATION**

Given the time constraints, the model cannot be used to its full ability, so for now, the percentage of attendees that have been imputed from a given dataset (all yes values) is being calculated as a percentage of total bookings.

This percentage is now being applied to a new dataset as a blanket value, so whatever the model predicts as total bookings, it then takes a percentage of them bookings and that is the predicted actual attendance.

The actual method that was to be used will be in future work. The accuracy of each model is calculated by multiplying the accuracy value of both imputations.

A white table with black text and yellow text

Description automatically generated

#### Conclusion

In conclusion, the task requires the careful consideration of various classification models. While logistic regression, decision trees, and random forest each offer distinct advantages and drawbacks, Logistic Regression emerges as the most appropriate choice given the constraints of the project and quantitative results given.

Logistic regression, a classic choice for prediction tasks, presents an efficient and straightforward approach. Performance will not be compromised by the dataset's missing data, as methods to predict or ‘fill in’ the data will be used.

On the other hand, random forest shows high accuracy and robustness to noise, thanks to its ensemble nature and ability to handle missing values. Nevertheless, its computational demands and potential bias towards the majority class pose challenges, particularly within the project's time constraints.

Decision trees also offer a balanced solution that aligns with the project's goals and limitations. However, their tendency to overfit and high variance cause Decision Trees to be less applicable.

# IMPUTATION OF MISSING DATA

## Introduction

When conducting data analysis, it is common to encounter missing data in the dataset. Presence of missing data can pose challenges and potentially impact the validity and accuracy of the prediction. Several approaches were invented to handle the missing data to convert the incomplete dataset to a complete case for analysis. In this section, the missing data mechanism and methods for handling missing data will be discussed. Next, the selection of imputation methods in this project and the result will be explained. Limitation of some methods and assumptions were addressed as well.

## Literature Review

### Missing Data Mechanism

Understanding the missing data mechanism can facilitate the decision of the appropriate techniques for handling missing data. Otherwise, the imputation result could be highly biased if the method does not suit the missing data mechanism.

#### Missing completely at random (MCAR)

Missing data is classified as MCAR when it is unrelated to observed responses or specific values. (Rouzinov & Serguei, 2022)

An example of this concept, describing a scenario in which a researcher, while reviewing raw data forms, randomly removed some part of the data. In this case, the removed data forms were randomly removed and were not related to specific values or observed responses. Additionally, if data is absent due to design or samples being lost in transit, it is also classified as MCAR. (Donges, ) (Bennett & Derrick, 2009) (Kang & Hyun, 2013)

#### Missing at random (MAR)

Similar to the MCAR data, MAR data is not related to the variable itself, however, it is related to the observed response. Ahypothetical example outlined that participants of various age groups are assigned to two groups and asked to provide details, for example their household income. When examining the rate of missing data, that is related to household income, differences may emerge among subpopulations. For example, if there is a suspicion that the different is because men tend to be less inclined to share their income compared to women, then this means that the missing data is related to the observed response, which mean the household income variable is considered to be Missing at Random (MAR). (Bar, 2017)

#### Missing not at random (MNAR)

The data is considered not missing at random (MNAR) when it is directly related to the collected data. If the absence of missing data is not random and is linked or connected to the collected data, for example, some individuals refusing to fill parts of the survey, then the missing data can be classified as MNAR. (Bar, 2017)

### Methods for Handling Missing Data

Missing data can lead to problems that can affect the operation of constructing effective prediction systems.

There are two commonly used approaches in practice to deal with missing values: marginalization, which is a simple method of deleting the missing values. This approach will reduce the dataset size and decrease the quality of the research. Imputation, which involves substituting the missing values with estimated values. (Cheng & Ching-Hsue, 2019)

The choice between imputation techniques is influenced by the data type and the data mechanism. We spoke earlier about the data mechanism. Data types can be numerical, categorical, or a mixture of both. Categorical data consists of nominal and ordinal data, and numerical data comprises real, continuous, and discrete data. (T. Aljuaid & S. Sasi, 2016)

Handling missing values can happen during the data mining process or before it begins. Methods that handle missing values during the process are called Embedded methods. Methods that handle missing values before the data mining process are called Pre-replacing methods. (T. Aljuaid & S. Sasi, 2016)

#### Regression imputation

As we mentioned before, in imputation, we use existing variables to estimate missing data. There are several types of regression imputation methods, for example, Linear Regression imputation and Logistic Regression imputation. Depending on the mechanism of the dataset and the datatype of the missing values, each technique has its advantages and scenarios where it can be used, but in general, this method produces unbiased results if the data are MCAR or MAR. Formatting...

#### Linear Imputation

This method assumes that there is a linear relationship between the dependent variable and the independent variable or variables. However, in some scenarios, the relationship is not linear, and replacing missing values using this method will bias the model.

#### Logistic Imputation

This method is generally a great choice for imputing categorical data, which in our case is what we need. It does this by leveraging logistic regression to predict the missing values based on the observed data. As previously mentioned, the missing data has been classified as MCAR, which is needed for logistic regression so that a bias is not introduced.

#### Mode Imputation

Mode imputation is the oldest and most basic method for categorical data imputation. This technique imputes the missing value of the associated attribute using the mode of the attribute.

The Mode method is considered to be a pre-replacing method, and it can be used with different data types such as numerical or categorical, and with only the MCAR mechanism. It is also considered to be fast and simple to use in contrast to other methods.

#### K-Nearest Neighbour (KNN)

The K Nearest Neighbour (KNN) algorithm was proposed by Cover and Hart and is regarded as one of the most widely used techniques in data mining. The KNN algorithm is used for data imputation by estimating missing values based on the averages of neighbouring data points that are selected using a distance metric, such as Euclidean distance.

KNN provides fast and accurate ways of estimating missing values. In comparison to traditional statistical methods, machine learning methods like KNN have been found to perform better.

In KNN imputation, missing values are replaced with values that correspond to the nearest neighbours, based on Euclidean distance. These missing values are imputed by considering a specified number of instances that are most similar to the instance of interest. The KNN model is regarded as a pre-replacing method, and it can be used with different data types such as numerical, categorical or mixed data, as well as different mechanisms like MCAR, MAR, and MNAR.

#### Multiple Imputation

This method is similar to Regression imputation; however, instead of having one value as an estimation of the missing data, we have a set of values. Then, these values are analysed using standard statistical methods. In addition to that, multiple imputation restores the natural diversity of the missing values.

#### Listwise or case deletion

One of the most used approaches when it comes to dealing with missing data. This approach performs best in large datasets with relatively few numbers of missing values. This method fits well with MCAR data and will result in unbiased estimation.

#### Missing Values Imputation Using Association Rules

This method is simple probabilistic statement. For example, “IF A=1 AND B=1 THEN C=1” explains the co-occurrence of certain events. The conditional probability is considered as the “confidence” of the rule. The best approach is to generate the association rules with complete data set. An alternative yet practical approach is to treat the missing values as special values to obtain a complete dataset. Since association rules cannot be easily generated from numeric attributes, this method can be more suitable for symbolic attributes.

## Assumptions

* The presence of missing data in the dataset was explained to be occurred due to human error. As a result, the missing data in the dataset is considered as not related to the variable itself or any observed responses. The missing data is characterised as MCAR for further handling. This mechanism can be applied to datasets with partially recorded attendance.
* Another reason for the existence of missing values is the willingness of revealing the actual attendance. Some organisers might not want to reveal the attendance to third parties due to commercial reasons. In such cases, the missingness could be related to the missing variable. Hence, this type of missing data is considered as MNAR. This mechanism can be applied to the datasets with no attendance recorded.
* In some datasets, there were some booking activities registered after the event date. As the project target is to predict the number of bookings until the event date, assuming only those bookings registered on or before the event date were relevant. Any bookings made after the event date are assumed to be for online attendance or online viewing of recorded material. Therefore, these bookings were considered as immaterial and irrelevant in terms of operational considerations.
* An assumption is made that the behaviour of the same target audience would be similar. Datasets are grouped and combined based on the same target audience for the imputation purpose.

## Implementation of Missing Data Handling Techniques

After research of the imputation methods was completed, feasible methods were chosen to be implemented in this project. Listwise deletion, Linear regression, Logistic regression, Association Rules, Mode imputation, and the KNN method were discussed.

#### Listwise deletion

In the received dataset, there are 128 out of 6407 (1.998%) registrations have a creation date later than the event date. As some of the events allowed online access to the event material after the event date, this is the reason why there are registration after the event date.

However, our goal is to predict the registration number till the event date, the missing attendance of those registered after the event is not relevant to main objective. Therefore, these registration data can be omitted from the dataset.

#### Linear Regression

This model is used to predict the percentage of attendees who attend the event. It is trained and tested only on the completed rows of the CSV files. These completed rows are split into 80% for training the model and 20% for testing.

The number of days until the event date is used as a feature, and the "Percentage Attended" is the target to train the model. Three different metrics are used to evaluate the performance of the model: R-squared to measure fit quality, Mean Absolute Error (MAE), and percentage of the relative mean absolute error to measure the accuracy of the model.

#### Logistic Regression

The model would take some feature attribute as the predictors, and calculate the probability of the outcome variable. In the datasets, the model could take the “Date Before Event” and the “Attendee Status” as the predictors to impute the data in the “Atttended” column.

It is worth noting that the results produced were of a higher accuracy when combining both “Days Before Event” and “Attendee Status”. However, to unify the approach of all imputation methods, and the approach in future predictions, only the “Days Before Event” was accounted for training the models.

#### Missing Values Imputation Using Association Rules

In the received dataset, a booking reference can register on behalf of more than one person. Assuming the attendees tend to attend or not attend the event together with the group, the missing attendance of some of these attendees can be imputed by this rule. By finding the most common attendance value of these group attendees, the missing data can be imputed by the majority value within the group.

#### Mode Imputation

There are two approaches with the mode imputation. The first method is to impute the missing data with the mode of the whole data set. In all data files, the mode of the target attribute “Attended” is “Yes”. The other method is to separate the dataset by classes, “Attendee Status” in this project. The mode of each class was founded and imputed to the missing values according to the “Attendee Status”.

#### KNN

The “Created Date” of the booking activity in each file was compared with the event date and the “DateBeforeEvent” was calculated. The result was passed to the KNN model as a predictor variable for the target variable “Attended”. Therefore, when a new data was passed to the model, it would look to the similar “DateBeforeEvent” and find the majority outcome to impute the missing value.

## Limitations of Different Missing Data Handling Techniques

### Listwise deletion

In the received dataset, 128 out of 6407 (1.998%) registrations have a creation date later than the event date. Some of the events allow online access to the event material after the event date. However, our goal is to predict the registration number until the event date; the missing attendance of those registered after the event is not relevant to the main objective. Therefore, these registrations data can be omitted from the dataset.

### Linear Regression

When it comes to evaluating the effectiveness of the linear regression model, R-squared is used. The R-squared value ranges from zero to one. A value close to or equivalent to one indicates that our model fits well. The R-squared values we obtained for our target audience were all below 0.3, indicating the poor performance of the model. Which indicate that the linear regression is not a suitable method to impute missing data for the 'Attended' column. The results for the linear regression R-squared, Mean Absolute Error and Relative Mean Absolute Error percentage can be found in the appendix.

### Association Rules

In the database, there were 283 of booking references which consisted of 753 attendee references. In these attendee references only 209 missing data could be applied with the rules, which covered 5.33% of the missing data. With the remaining missing data being unhandled, this imputation method is not a feasible method for this project.

### Mode Imputation

This method was founded not suitable during the implementation, as some datasets were assumed to be MNAR. Bias would be resulted in the imputation as the method did not account for the uncertainty of the missing variable. Nevertheless, there is no sufficient amount of “No” recorded in the datasets, leading to a same mode “Yes” resulted in both the “Attending” and “Cancelled” of the “Attendee Status”.

### Multiple Imputation

Although the chosen methods are suitable for handling the missing data in the project, multiple imputation was recommended when dealing with MNAR data. Due to the time restraint, the implementation of multiple imputation could not be accomplished.

## Results

Before starting the imputation process, data processing and cleaning were performed to ensure that the dataset could be used for model training and imputation. Each dataset was read into the pandas data frame for processing in the Python program.

A new column “Date Before Event” representing the number of days from the event day was added, so that we could correlate the date from two events. Since some of the datasets contained plenty of missing data or no data was recorded at all, the dataset was grouped by the type of target audience into three groups: D19 & D21, GP21 & NP21, and SRM22 & SRM23.

Then, all datasets were combined together into a single group to impute the missing data in MSE21. Additionally, registrations with a creation date after the event date were removed from the dataset.

The k value of the KNN model needed to be determined before training the imputation models. A parameter test was conducted to determine which value provides the best accuracy score.

After finalising the parameters for the imputation models, the data in each dataset was split into 80% training data and 20% testing data. The 80% training data was fitted into the model to train up for the imputation and the performance of the models was tested with the 20% testing data. F1 Score and accuracy were used to evaluate the performance of the missing data imputation.

There was an imbalance of “Yes”/ “No” values in the datasets, where the numbers of “Yes” were more than “No”. Therefore, the SMOTE method was implemented to oversample the datasets and balance the amount of “Yes” and “No”.

The method generated dummy data of “No”, so that a balanced training dataset could be fitted to the models. Then, the performance of the models was tested again.

The following table shows the performance of the logistic regression and KNN method after implementing SMOTE.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression** | | |  | **KNN Method** | | |
| **Target Audience** | **Accuracy** | **F1** | **Confusion Matrix** | **Accuracy** | **F1** | **Confusion Matrix** |
| All | 57.32% | 70.22% | [33,56] | 67.09% | 77.95% | [42,35] |
| [145,237] | [120,274] |
| IT Managers | 76.57% | 85.01% | [29,20] |  | 68.88% | 78.13% | [38,3] |
| [47,190] | [86,159] |
| Property Managers | 45.90% | 61.18% | [2,2] | 39.34% | 54.32% | [2,2] |
| [31,26] | [35,22] |
| Education Property Managers | N/A | N/A | N/A | N/A | N/A | N/A |
| Education Managers | 60.80% | 69.57% | [20,6] | 56.80% | 69.66% | [9,17] |
| [43,56] | [37,62] |

Although the KNN method got a better accuracy and F1 score when all the datasets were combined into one dataset, the Logistic Regression had a better accuracy and F1 score overall when the target audience could be specified.

Therefore, the Logistic Regression imputation was chosen to impute the missing data in all the datasets. Each dataset became completed and was passed to train up the model for the attendance prediction of the upcoming events.

# DEVELOPED PREDICTION APPLICATION

## Objective

The required objective of the project was to see if a forecast of registrations for a conference could be made based on the way registrations have been made so far.

This application achieves this objective and goes a couple of steps further, as explained in the following sections.

## Assumptions

Certain assumptions must be made for the prediction model to work properly.

* The model will assume that the bookings made up to the current date followed a consistent trend. The distribution of bookings over individual days for the new data cannot be filled out by the model at this stage.
* There were no advertisement campaigns launched up to the “current date” that will be provided to the model to make predictions for any new conferences.
* The type of advertisement is not relevant to the application. It assumes that the type of advertisement has no effect.

## Booking Count Prediction

To make predictions for the booking count, the machine learning model was chosen from the best options discovered during the comparative study. The standout options were Elastic Net Regression from the regression models and SARIMAX from the time series models.

Their accuracy scores are compared in the following table for each target audience to decide the model for the application.

|  |  |  |
| --- | --- | --- |
|  | **Elastic Net Regression** | **SARIMAX** |
| **Target Audience** | **Accuracy (%)** | **Accuracy (%)** |
| IT Managers | 90.99 | 83.79% |
| Property Managers | 92.18 | 87.05% |
| Education Property Managers | 79.76 | 22.52% |
| Education Managers | 94.28 | 83.26% |
| Others | 95.32 | 86.54% |

The Elastic Net Regression model outperformed the SARIMAX model across all of the target audiences. Moreover, it offers the added benefit of being able to predict % effect of advertisements on the booking count (as explained in the Comparative Study chapter). Therefore, this was the model selected for the application.

For predicting booking counts for new conferences, the elastic net model for each target audience was now trained on all the data available instead of splitting the data into training and testing data.

The model takes the type of target audience, registration start date, event date, current date, and the bookings made up to the current date. It then predicts the total number of bookings that are expected to occur by the event date.

While the accuracy of the model is good based on the testing done, it is expected to improve and make better predictions if more data is made available for it to be trained on.

## Advertisement Effect Prediction

The elastic net model also uses the data from the previous conferences to estimate the effectiveness of an advertisement campaign for each target audience.

The application displays the predicted no. of bookings with and without advertisements to allow the user to make better-informed decisions before dedicating resources to an advertisement campaign.

## Prediction of Actual Attendees

While having a predicted booking count is useful, the real number that an event organizer would be concerned with is how many people turn up to a conference. If the number of people who show up is significantly lower than the number of bookings, it would result in wastage of resources.

Therefore, this application also gives a prediction of how many people will attend the event after making the booking. We expect this prediction to be particularly useful in a real-word scenario.

This prediction is made using the logistic regression model from the classification models. It takes the predicted booking count as the input, and based on the already imputed training data for the relevant target audience, makes a prediction of how many people will attend the conference.

## Graphical User Interface

A graphical user interface (GUI) was also developed so that the end users could easily interact with the application. The GUI was developed using tkinter in python.

# FUTURE WORK

There is scope for further development in this application, and some of these future developments are listed below.

## Data Analysis

A more detailed data analysis could help understand the effect of different factors. One such factor could be the age of participant. The more detailed the data that is provided for development of this application, the more accurate and realistic the predictions will be.

## Prediction Models

After performing the comparative study of prediction models, it was apparent that some initial ideas of which models would be better were not necessarily correct. Therefore, a comparative study of even more machine learning could benefit the application in terms of selecting an even more accurate model.

Zero-inflated models such as the zero inflated negative binomial or the zero inflated poisson model might help make more accurate predictions due to the large number of days with 0 bookings in the data.

Another approach could be to study models that can handle sparsity in the data better, such as a sparse linear model, or by including a sparsity related penalty to the prediction model.

## Imputation

An improvement of the imputation could be implementing the multiple imputation. As some of the missing data were considered as MNAR, multiple imputation could be implemented to lower the bias created by the imputed data.

The splitting and resampling of data could be randomised multiple times before fitting into the models, so multiple imputed datasets could be generated to compare the variance among the datasets. By doing so, the uncertainty of the missing data could be accounted for, and the performance of the imputation could be improved.

Another improvement could be to include z scores as an independent variable as well for imputation of missing values. Z scores could be converted into binary values where 1 would be above our threshold and 0 below it. It could also impact the accuracy and fit of our imputation models.

## Classification Models

Now that the model was decided, it could be used to test new data and impute missing values. The model would be trained for each individual event type, then used on data passed on from the regression ran before this stage, to impute the chance of a customer attending the event.

This value is multiplied by the amount of booking on the specific day and summed to give the predicted total attendees. This can then be turned into a percentage value of the total number of tickets booked.

## User Interface and Final Predictions

We believe that if the end user provides a csv file for the conference up to the current date, the model would be able to make a more accurate prediction and offer more detailed insights to the user. The functionality to handle such an input from the user therefore can added to the application in the future.

# CONCLUSION

A prediction model was developed for conference bookings. Data from seven previous conferences was analysed and used to train and test different machine learning models. Elastic net regression was found to be the most accurate and was selected for making predictions of booking count for future conferences. Imputation was conducted to fill out missing values in the conference files and the imputed data was fed to a classification model to predict how many people who register for a future conference will actually attend the event. A user interface was developed for receiving input from the user. Finally, a number of recommendations were provided for further development of the application.

# REFERENCES

'Forecasting SARIMAX and ARIMA models', *Skforecast Docs,*Available at: <https://joaquinamatrodrigo.github.io/skforecast/0.7.0/user_guides/forecasting-sarimax-arima.html.>

'Logistic Regression', Available at: <https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf.>

*What is a Decision Tree?*Available at: <https://www.ibm.com/topics/decision-trees> (Accessed: .

Agarwal, T. (2019) *Standard Deviation.*Available at: <https://www.wallstreetmojo.com/standard-deviation-examples/> (Accessed: .

Ampountolas and Apostolos (2021) 'Modeling and Forecasting Daily Hotel Demand: A Comparison Based on SARIMAX, Neural Networks, and GARCH Models', *MDPI,*Available at: <https://www.mdpi.com/2571-9394/3/3/37.>

Bar, H. (2017) 'Missing data — mechanisms and possible solutions / Datos ausentes: mecanismos y posibles soluciones', *Culture and Education,*Available at: <https://doi.org/10.1080/11356405.2017.1365426.>

Bennett and Derrick, A. (2009) 'How can I deal with missing data in my study?', Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.2001.tb00294.x.>

Bhattacharyya, R. (2023) *Time Series Analysis.*Available at: <https://www.wallstreetmojo.com/time-series-analysis/> (Accessed: .

Bora and Neha (2021) *Understanding ARIMA Models for Machine Learning.*Available at: <https://www.capitalone.com/tech/machine-learning/understanding-arima-models/> (Accessed: .

Breiman and L (2001) 'Random Forests', *Machine Learning,*Available at: <https://doi.org/10.1023/A:1010933404324.>

Cheng and Ching-Hsue (2019) 'A Distance-Threshold kNN Method for Imputing Medical Data Missing Values', *Journal of Advances in Computer Networks,*Available at: <https://api.semanticscholar.org/CorpusID:257396142.>

D.Cryer and Jonathan (2008) *Time Series Analysis With Applications in R*. Second edn. Springer Texts in Statistics.

Dhumne and Shruti (2023) 'Elastic Net Regression detailed guide', *Medium,*Available at: <https://medium.com/@shruti.dhumne/elastic-net-regression-detailed-guide-99dce30b8e6e.>

Etienne, B. (2019) *Time Series in Python — Exponential Smoothing and ARIMA processes.*Available at: <https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788> (Accessed: .

Gajendrakar, P. (2022) *Mean in Statistics.*Available at: <https://www.wallstreetmojo.com/mean/> (Accessed: .

Gallo, A. (2015) 'A Refresher on Regression Analysis', .

Kaiser and Jiří (2014) 'Dealing with Missing Values in Data', *Journal Of Systems Integration,*Available at: <https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Dealing+with+Missing+Values+in+Data&btnG=.>

Kang and Hyun (2013) 'The prevention and handling of the missing data', *Korean Journal of Anesthesiology,*Available at: <https://api.semanticscholar.org/CorpusID:18033167.>

Magee, J.F. (1964) 'Decision Trees for Decision-Making', .

Malkari and Nikhil (2023) 'Exponential Smoothing: A Method for Time Series Forecasting', *Medium,*Available at: <https://medium.com/@nikhilmalkari18/exponential-smoothing-a-method-for-time-series-forecasting-7ea35ca2c781.>

Mcleod and Saul. (2023) 'Z-Score: Definition, Formula, Calculation & Interpretation', *Simply Psychology,*-10-06. Available at: <https://www.simplypsychology.org/z-score.html> (Accessed: .

Min and S. Kim (2023) 'Comparative R Squared', *SSRN,*Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3790200.>

noe (2023) *How to calculate accuracy of a logistic regression? -.*Available at: <https://datascience.stackexchange.com/questions/118212/how-to-calculate-accuracy-of-a-logistic-regression> (Accessed: .

Ogunleye, J.O. (2022) *Predictive Data Analysis Using Linear Regression and Random Forest* IntechOpen.

Panchotia and Rajat (2020) 'Predictive Modelling Using Linear Regression', Available at: <https://medium.com/swlh/predictive-modelling-using-linear-regression-e0e399dc4745.>

Penkel and Engin (2019) 'Estimation of soil moisture using decision tree regression', *Springer Link,*Available at: <https://link.springer.com/article/10.1007/s00704-019-03048-8.>

Rouzinov and Serguei (2022) 'Regression-Based Approach to Test Missing Data Mechanisms', *MDPI,*Available at: <https://www.mdpi.com/2306-5729/7/2/16.>

S. Siami-Namini, N. Tavakoli and A. Siami Namin (2018) *A Comparison of ARIMA and LSTM in Forecasting Time Series.*pp. 1394.

S and Rahul (2023) 'Time Series Forecasting: A Comparative Analysis of SARIMAX and XGBoost Algorithms', *Medium,*Available at: <https://ogre51.medium.com/time-series-forecasting-a-comparative-analysis-of-sarimax-and-xgboost-algorithms-7cac99685564.>

Satyavishnumolakala (2020) 'Linear Regression -Pros & Cons', Available at: <https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0.>

Sharma and Natasha (2023) *Understanding and Applying F1 Score: A Deep Dive with Hands-On Coding.*Available at: <https://arize.com/blog-course/f1-score/> (Accessed: .

T. Aljuaid and S. Sasi (2016) *Proper imputation techniques for missing values in data sets.*pp. 1.

The Click Reader (2021) 'Decision Tree Regression Explained with Implementation in Python', *Medium,*Available at: <https://medium.com/@theclickreader/decision-tree-regression-explained-with-implementation-in-python-1e6e48aa7a47.>

V. Roshan and Joseph (2022) 'Optimal ratio for data splitting', Available at: <https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11583.>

Zou and Hui (2005) 'Regularization and Variable Selection Via the Elastic Net', Available at: <https://academic.oup.com/jrsssb/article/67/2/301/7109482#400951354.>

# APPENDIX

The code