Analysis of Time Series & Logistic Regression

Temitope Oladimeji – x23187204

National College of Ireland
MSc in Data Analytics
Statistics for Data Analytics
CA – Semester 2 2023/2024

***Abstract* —** This research project comprises two main components: time series analysis and logistic regression. In the first part, the temporal patterns and suitability of a selected variable was explored using time series techniques, investigating the presence of trends, seasonality, and potential cycles within the data. Implementing statistical tests and visualisations, the stationarity of the time series, and its suitability, were also assessed to guide further modelling for forecasting purposes.

In the second part, logistic regression was utilized to investigate the influence of various predictor variables on a binary outcome. The model was developed using logistic regression estimates to facilitate the understanding of the relationships between the factors and the target variable, and its predictive performance was evaluated on a subset of the data. As determined by statistical tests, key predictors were identified based on their significance ensuring adherence to logistic regression assumptions.

The primary aim of this study is to provide insights and a comprehensive understanding of both of these models.

Keywords — ARIMA/SARIMA, exponential smoothing, simple time series, forecast, binary, logistic regression, KDD, machine learning

1. INTRODUCTION

1.1. Time Series

Time series can be defined as an ordered sequence of quantitative observations on a variable [1]. Widely applied in areas such as stock market analysis, sales forecasting, and econometrics, the analysis of time series deals with detecting observable regularities in a variable [2], to improve predictive performances.

Analysing time series data involves plotting the data over time to identify patterns such as trends and seasonality, and making predictions about future values. This pattern of data reveals how the time series has behaved in the past and serves as the basis for forecasting. Providing there is enough past data, as stated in [3], forecasts can capture "genuine patterns and relationships". This is especially useful in fields such as the financial market to forecast future values of an exchange rate [1]. Other forecasting applications, from [3], include: strategic planning, determining future resource requirements, and scheduling processes.

Components of a time-series data include:

a) Trend – long term general direction
b) Cycles – patterns of highs and lows; usually of more than a year
c) Seasonal effects – shorter cycles, typically of less than a year
d) Irregular fluctuations – rapid changes in data, occurs in even shorter time frames than season effects

When it comes to complex forecasting models such as exponential smoothing and ARIMA, it is essential for the time series to be stationary. This makes it easier to model as it contains no trend, cyclical, or seasonal effects. Another benefit is to obtain meaningful statistics about the data [4], providing useful descriptions of future behaviour.

Non-stationary time series can become stationary via the differencing approach. This eliminates/reduces trend and seasonality by removing changes in the level of a time series. Differencing also calculates differences between consecutive observations which stabilizes both the mean and variance of the time series. Essentially, if the time-series plot shows roughly the same mean and variance through time without any significant seasonality, it is assumed to be covariance stationary.

A common goal of analysing data is its decomposition into separate components to generate insights [5]. Time series data with seasonal aspect can be decomposed into a trend, seasonal, and an irregular component [6]:

i. Trend component captures changes in level over time
ii. Seasonal component captures effects due to time of year
iii. Irregular component captures influences that are not described by time and seasonal effects

Seasonal decomposition can be described in two ways: additive model and multiplicative model. The additive model is most appropriate where seasonal fluctuations do not depend upon the level of time series:

$$Y_t = trend_t + seasonal_t + irregular_t$$

(Eq 1: Additive model)

$$Y_t = trend_t \; x \; seasonal_t \; x \; irregular_t$$

(Eq 2: Multiplicative model)

A better method than the ratio-to-moving average method is using Loess, which handles non-linear relationships. However, this method does not produce a formula that can be extrapolated for forecast.

1.1.1. Simple Time-Series Model

Simple time series models do not involve complex patterns, and as such are considered as baseline models. Estimating a trend in a time-series and using that trend to predict future values is the simples method of forecasting.

i. Average method – forecast of all future values are equal to the mean of historical data
ii. Naïve method – all forecast are set to the value of last observation; used for non-stationary data
iii. Seasonal naïve method – all forecast are set to be equal to the last observed value from the same season of the year
iv. Drift method – amount of change over time is set to the average change seen in historical data

1.1.2. Exponential Smoothing Model

This fits a time series that has a constant level and an irregular component at time i.e. no clear trend or seasonal pattern. An advantage of this model is that it yields good short term predictions.

i. Double exponential model (holt exponential smoothing) fits the time series with both a level and a trend.

ii. Triple exponential model (holt-winter's exponential smoothing) fits a time series with level, trend, and seasonal components.

1.1.3. ARIMA Models

ARIMA models are designed to fit stationary time series. Its predicted values are a linear function of recent actual values and recent errors of prediction. In order to achieve constant variance, it is recommended to log transform the values.

As mentioned in section 1.1, achieving stationarity is important for an ARIMA model. Augmented Dickey-Fuller (ADF) test evaluates this assumption, where a significant result suggests stationarity. The ACF plot can be used to test for autocorrelation whereby an appropriate model would have normally distributed residuals, and the autocorrelation of residuals should be zero for every lag.

1.2. Logistic Regression

Logistic regression can be used to rank the relative importance of predictor variables in explaining the target variable. The dependent variable can be dichotomous [7], in which case it would be labelled as binary, or it could contain multiple levels i.e. multinomial logistic regression.

Widely used to predict binary outcomes, such as whether a customer will default on a loan, or the prediction of lightning strikes [7], some of the advantages of logistic regression model include: its ability to handle both quantitative and qualitative predictor variables, and its applicability to binomial or multinomial regression.

The fitting of a logistic regression model involves finding the best linear combination of independent variables to maximize the likelihood of obtaining the observed outcome. This is known as the maximum likelihood estimation.

Assumptions of a logistic regression model:

i. Dependent variable should be mutually exclusive
ii. There should be absence of multicollinearity
iii. Predictors do not have to be normally distributed, linearly related to the target variable, or of equal variance within each group
iv. There should be absence of outliers
v. Independence of errors

The assessment involves various diagnostic tests such as:

i. Wald test - which checks the significance of each predictor
ii. Cox and Snell R-squared - which shows the proportion of variation in the dependent variable predicted by the predictor variables
iii. Model Deviance – where a lower value indicates a better-fitted model
iv. Hosmer and Lemshow test – which assesses the model fit, with a p-value < 0.05 indicating poor fit.

2. RELATED WORK

2.1. Time Series Analysis

In the work by Krispin [8], which focuses on time series analysis applied to natural gas consumption in the US, the dataset spans the period from 2000 to 2018 and is organized on a quarterly basis.

One notable aspect addressed in [8] is the reformatting of time and date variables prior to any implementation. The book emphasizes the importance of this step, detailing the potential implications if they are not appropriately handled. This aligns with the approach taken in our analysis, where we also recognized the significance of reformatting time variables for accurate analysis and forecasting.

Furthermore, the book provides guidance on creating time series objects, highlighting the significance of each parameter, including the start and end points, and the frequency, the latter reflecting the quarterly nature of the data. The implementation of our own data draws inspiration from these insights. While we build upon the steps outlined in [8], our work extends beyond the scope covered in the book. Specifically, we build and evaluate complex models such as Exponential Smoothing and ARIMA, evaluating the optimum model for each.

Nonetheless, the principles established in the book have proved beneficial in shaping our approach to time series analysis.

2.2. Logistic Regression

Our understanding of logistic regression is complemented by the insightful work in the "Practical Guide to Logistic Regression" [9]. This book explores the fundamental logic of the model as well as its appropriate applications. Serving as a guide to our analysis, the book also focuses on various aspects of the logistic model, including the steps in building a model and the interpretation of its coefficients.

The book further details the steps involved in predicting probabilities and offers guidance on the fitting and evaluation of a logistic regression model. Our own investigation into, and analysis of, logistic regression makes use of the foundational principles laid out in the book. The use of various plots and the confusion matrix table gives us additional insights into the functionality of the model.

3. METHODOLOGY

3.1. Time Series Methodology – Investigation of suitable models

3.1.1. Data Understanding and Data Pre-processing

The file used for the time series analysis was in an excel format, and was a daily time series of historical weather data reported by Met Eireann from one of the weather stations in Ireland.

The dataset had 29,889 rows, commencing from 1st January 1942, to 31st October 2023, and included variables of daily measurements such as Air Temperature, Precipitation, Wind Speed, Evaporation etc.

When read from a csv file, the date column was interpreted as a factor with levels representing the unique dates. Preprocessing of the dataset included the conversion of the date column into actual date objects. Additional cleaning also involved the extraction of both the date and assigned variables, and the parameter definition for the time series.

```
'data.frame':   29889 obs. of  9 variables:
 $ date                                  : Date, format: "1942-01-01" "1942-01-02" "1942-01-03" ...
 $ maxtp.Maximum.Air.Temperature...degrees.C. : num  9.7 9.9 11.2 9.2 3.5 5.1 7.1 7.1 4.5 5.3 ...
 $ mintp.Minimum.Air.Temperature...degrees.C. : num  6.8 7.9 8.9 2.7 -0.8 0.7 0.5 1.4 0.7 -2.8 ...
 $ gmin.Grass.Minimum.Temperature...degrees.C.: num  4.7 6.7 7.2 3.4 0 -3.7 -1 0.2 0.9 -4.1 ...
 $ rain.Precipitation.Amount...mm.       : num  0 0.1 1.5 3.5 0.6 0 0 0 0.2 0 ...
 $ cbl..Mean.CBL.Pressure.hpa.           : num  1020 1016 1007 1002 1013 ...
 $ wdsp.Mean.Wind.Speed...knot.          : num  17.2 15.2 14 17 13 9.7 10.3 9.3 11.8 4 ...
 $ pe.Potential.Evapotranspiration...mm. : num  1.1 0.7 0.5 0.6 0.6 0.4 0.2 0.2 0.5 0 ...
 $ evap.Evaporation..mm.                 : num  1.4 0.9 0.6 0.7 0.7 0.5 0.2 0.2 0.7 0.1 ...
```

Figure 1: Structure of the dataset

As there were only 7 missing values out of the 29889 observations, deleting the rows containing missing values seemed a reasonable option. Given the proportion of missing data relative to the size of the dataset, common practice is to delete missing values. This also avoids introducing bias associated with imputation.

3.1.2. Exploratory Data Analysis (EDA)

Noise removal involves handling missing values that are likely to be noise. This is an important step to avoid misinterpretation of the data.

Visualizing outliers was done via box plots, and the IQR statistical test identified and removed outliers via the upper and lower bound thresholds.

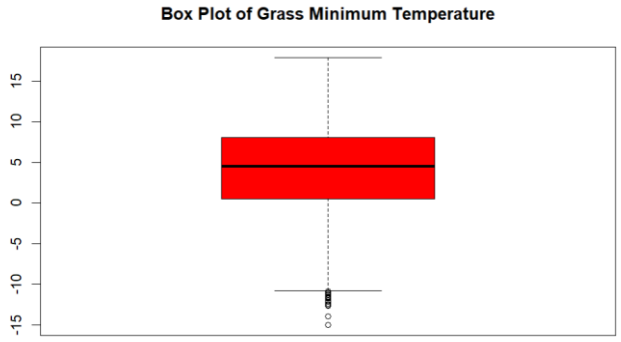

Figure 2: Box Plot of Variable before IQR Test

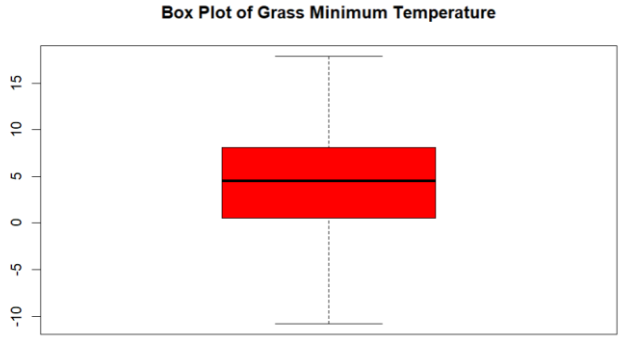

Figure 3: Box Plot of Variable after IQR Test

3.1.3. Assessment of the raw time series

The time series was defined with a frequency of 12 where each observation corresponded to a month. The series was across a span of the 82 years, as shown in the figure below:

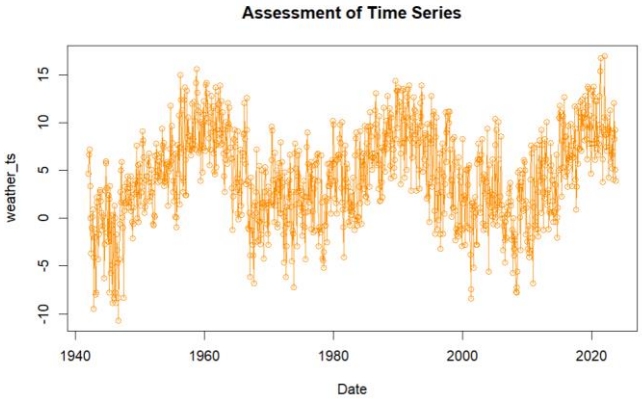

Figure 4: Raw Time Series

The plot shows a clear upward trend over the past 80 years suggesting that the measured variable is increasing over time. A notable aspect of the plot is its cyclical nature, with periods of faster and slower growth, displaying an overall positive trend.

There is also some seasonal variation, with lower values in the winter and higher values in the summer. This is due to the fact that the measured variable (minimum grass temperature) is affected by temperature.

The model was split where data from 2019 to 2022 was used to fit the model, and data from 2023 was used to evaluate the performance of the fitted models.

Prior to building any model, a test for stationarity is advisable. This was done using the Augmented Dickey-Fuller Test where a significant p-value indicates non-stationarity. Achieving stationarity is especially useful for complex models such as the ARIMA, which is designed to fit stationary time series.

```
        Augmented Dickey-Fuller Test

data:  train
Dickey-Fuller = -5.2605, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

Figure 5: Result of the ADF test for stationarity

3.1.4. Model Selection and Fitting

3.1.4.1. Simple Time Series

The first model built was the simple time series. It was built using four different methods: average method, naïve method (random walk), seasonal naïve method, and drift method.

The forecast set for these methods were 7 months in 2023, with a comparative analysis on which model performed better using quantitative metrics such as the MAE. Section 4 details the results.

3.1.4.2. Exponential Smoothing

The fitting of the exponential smoothing focused on yielding good short term predictions.

- The double exponential smoothing (holt-exponential smoothing) fits a time series with both a level and a trend.
- The triple exponential smoothing (holt-winter's exponential smoothing) fits a time series with level, trend, and a seasonal component.

As the time series itself contained seasonal aspects, the triple exponential smoothing method was expected to yield the best result, shown in section 4.

Using the 'ets' function to investigate the best model, the accuracy of the intermediate models were investigated to check the accuracy and evaluate its performance.

3.1.4.3. ARIMA

As no differencing was required, given the stationarity of the time series as proven by the result of the ADF test, the ACF and PACF plots were visualized to investigate the appropriate parameters to build the models with.

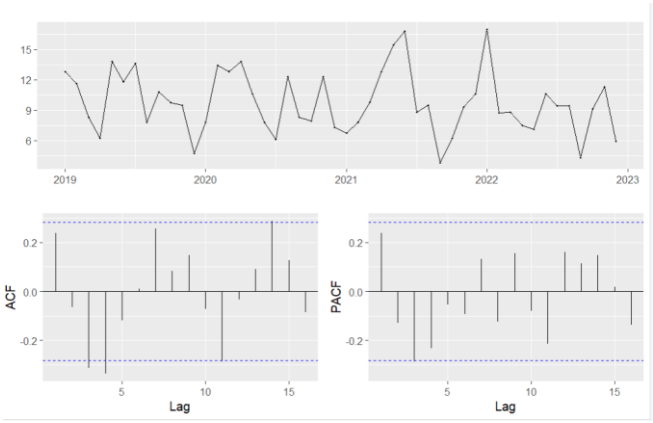

Figure 6: ARIMA model

The ACF and PACF plots show that the first lag of the CF is significant, while the first lag of the PACF is not. This is indicative of a first-order autoregressive component (AR) in the data.

The autocorrelations decrease fairly quickly after the first lag, suggesting the data may not be integrated. This means an I value of 0. The PACF shows a significant peak at the first lag, and other lags, suggesting a moving average component (MA) in the data.

3.2. Logistic Regression – Discussion of Modelling Process

3.2.1. Data Understanding and Data Pre-processing

The file used for the logistic regression analysis contained attributes of 100 participants, with the variables:

- Age
- Weight
- Gender
- ftiness_score
- cardiac_condition

caseno	age	weight	gender	fitness_score	cardiac_condition
Min. : 1.00	Min. :30.00	Min. : 50.00	Female:37	Min. :27.35	Absent :65
1st Qu.: 25.75	1st Qu.:34.00	1st Qu.: 69.73	Male :63	1st Qu.:36.59	Present:35
Median : 50.50	Median :39.00	Median : 79.24		Median :42.73	
Mean : 50.50	Mean :41.10	Mean : 79.66		Mean :43.63	
3rd Qu.: 75.25	3rd Qu.:45.25	3rd Qu.: 89.91		3rd Qu.:49.27	
Max. :100.00	Max. :74.00	Max. :115.42		Max. :62.50	

Figure 7: Descriptive Statistics for the dataset

The data had no missing values, but parsimony was achieved by removing the 'caseno' variable as the variable simply tells how many rows the dataset has. The 'gender' variable was one-hot encoded, subsequently creating two new variables.

3.2.2. Exploratory Data Analysis (EDA)

The next step was to visualize the distribution of the target variable and detect outliers (see figure 8).

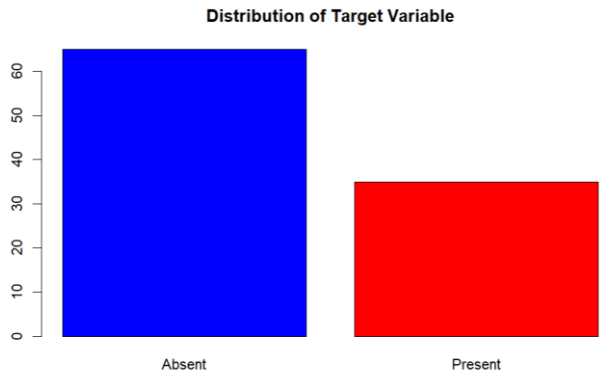

Figure 8: Target Variable Distribution

To solve this imbalance, an oversampling technique was used, made possible by the SMOTE package in R. Oversampling created synthetic samples for the 'present' class which was the minority class. An advantage to this is the preservation of data compared to under-sampling; conversely, synthetic samples could introduce noise in the data because those introduced samples might not accurately represent the true distribution.

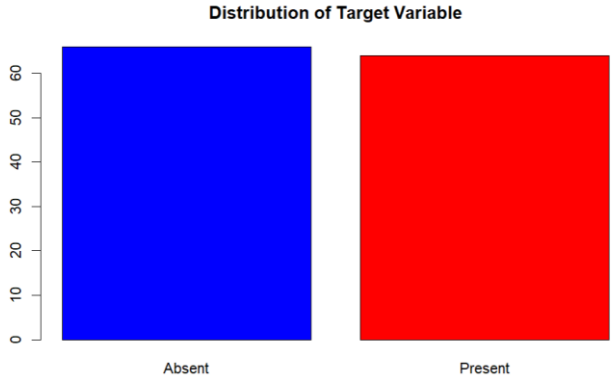

Figure 9: Oversampling method used to balance the distribution

Pre-processing the data also involved the handling of outliers.

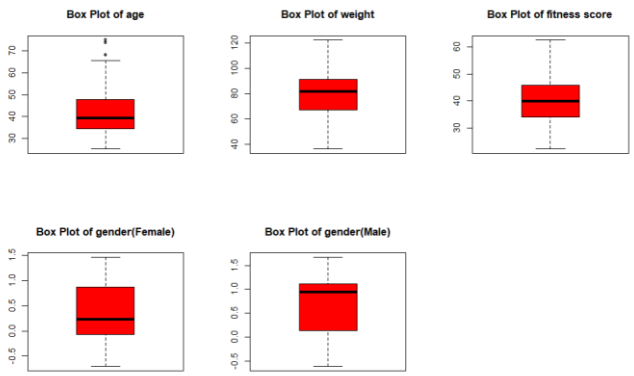

Figure 10: Box Plot before IQR test

The IQR statistical test was performed to identify and remove the outliers. The dataset had very little outliers as shown in figure 10, but to ensure an adequate model, cleaning of the data is still a necessity, and as such, removing outliers remains an important step.

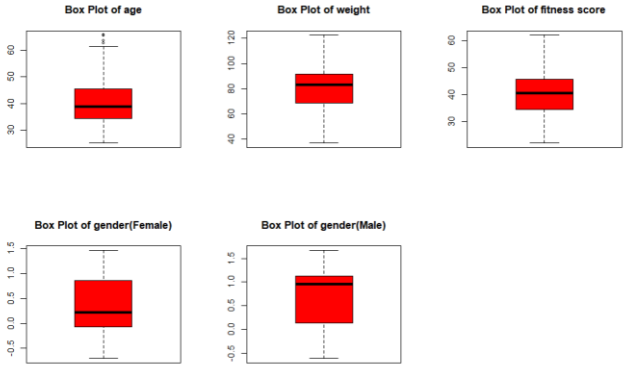

Figure 11: Box Plot after IQR test

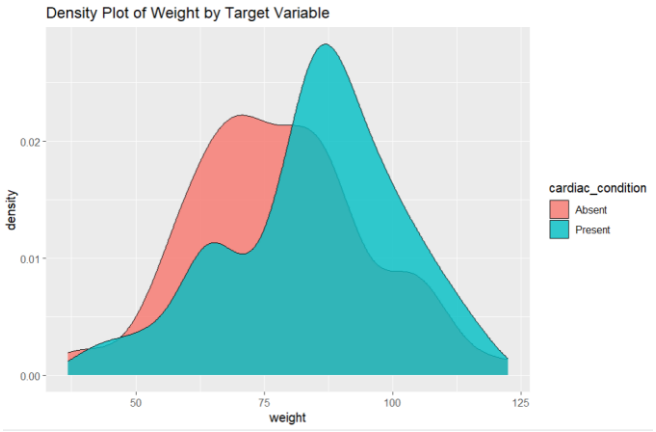

Figure 12: Density Plot of weight by target variable

The density plot shows the distribution of weight by target variable, which in this case is the presence or absence of a cardiac condition. The darker colour at the weight value indicates a higher point density at that value.

The plot shows that the distribution of weight is different for people with and without cardiac conditions. People with cardiac conditions tend to be heavier than people without cardiac conditions. This is evident from the fact that the density plot for people with cardiac conditions is shifted to the right compared to the density plot for people without cardiac conditions.

An overlap is also present in the distribution, meaning there are some people with cardiac conditions who are not overweight or obese, and there are some people without cardiac conditions who are overweight or obese.

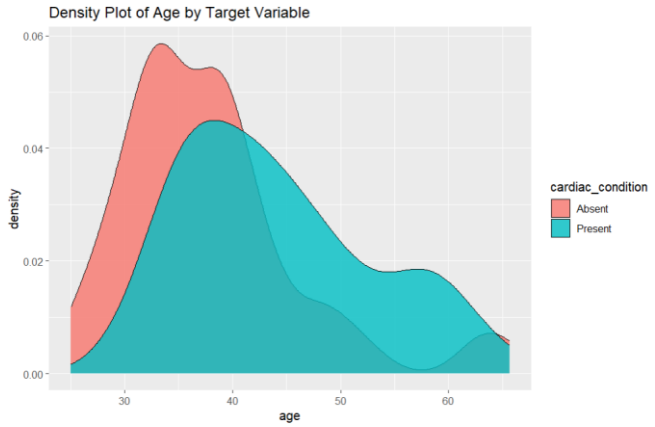

Figure 13: Density Plot of age by target variable

The plot shows that people with cardiac conditions tend to be older than people without cardiac conditions. Essentially, it suggests that age is associated with the presence of a cardiac condition.

3.2.3. Model Building & Evaluation

Prior to fitting the model, the predictor variables were normalized to scale the data to a specific range. This ensured the absence of extreme values or outliers represented as noise while evaluating the model.

For the model building, the seed was set for reproducibility. The model was then split into training and testing sets at a 70:30 ratio in which the former was used to fit the model, while the latter was used to evaluate the performance of the model.

Various tests were performed such as the collinearity, deviance, homser & lemshow, and cox-snell tests. A confusion matrix table was generated for further analysis which assessed the accuracy, precision, recall, and F1 score metrics.

4. RESULTS

4.1. Time Series – Evaluation of the final model

4.1.1. Simple Time Series Model Results

Forecasting Method	RMSE	MAE	MAPE	MASE	ACF1
Mean Average	3.05	2.46	29.62	0.67	0.24
Naïve Method	3.74	3.01	36.01	0.81	-0.27
Seasonal Naïve	4.46	3.70	41.04	1.00	0.34
Drift Method	3.73	3.01	35.65	0.81	-0.27

Table 1: Accuracy comparison for the simple time series model

The mean average method has the lowest RMSE, MAE, MAPE, MASE, and ACF1 (auto-correlation of forecast errors), which means that it is the most accurate forecasting method. The ACF1 for all of the methods is close to 0, suggesting no strong correlation exists between the forecast errors and the actual values.
As the mean proved to be the best method, it was assessed using the testing data.

Forecasting Method	RMSE	MAE	MAPE	ACF1
Testing set on the mean	2.48	2.08	33.70	-0.16

Table 2: Evaluation of the average method model

4.1.2. Exponential Smoothing Model Results

Type of Model	RMSE	MAE	MAPE	MASE	ACF1
Simple Exponential Model	3.05	2.46	29.63	0.67	0.24
Double Exponential Model	3	2.37	28.73	0.64	0.23
Triple Exponential Model	2.68	2.2	25.4	0.59	0.24

Table 3: Comparison of the three exponential smoothing models

As stated in section 3.1.4.2., the double exponential smoothing fits a time series with a level and a trend, while the triple exponential smoothing fits a time series that contains seasonality.

As the raw time series (figure 4) showed evidence of seasonality, the triple exponential smoothing was able to capture this seasonal pattern, compared to the double exponential smoothing model.

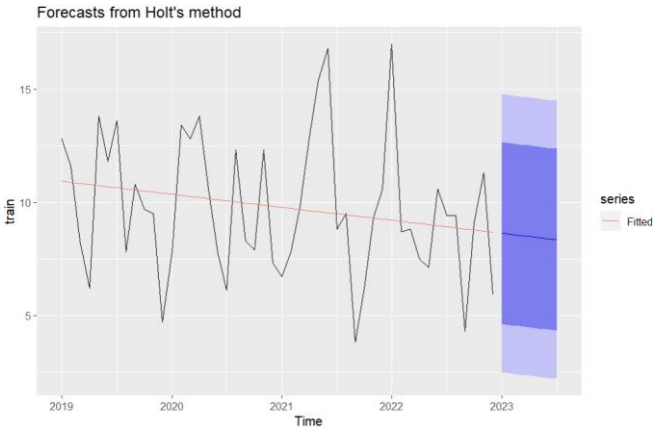

Figure 14: Holt's method

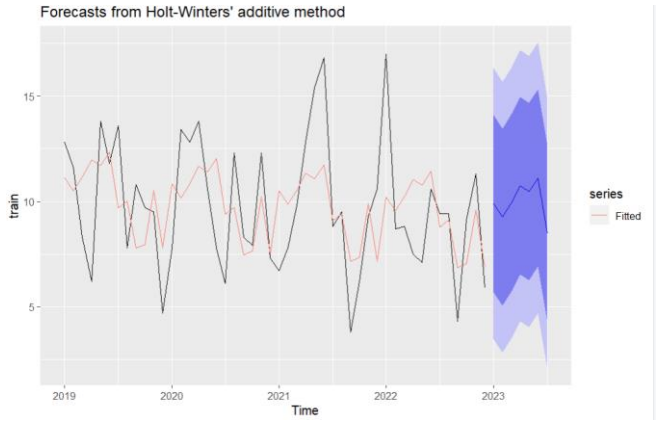

Figure 15: Holt-Winter's method

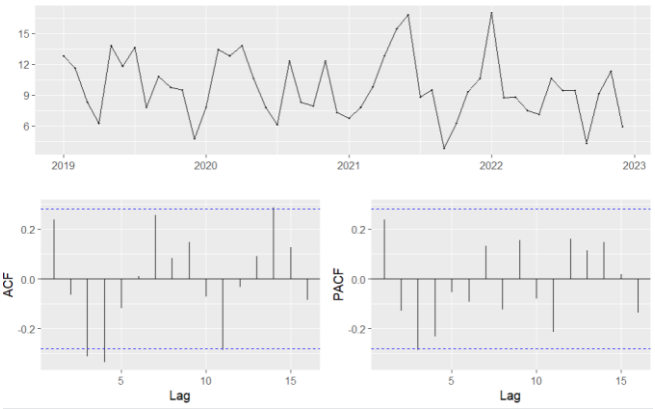

Figure 16: ARIMA model – ACF and PACF plots

The model was also fitted using the 'ets' function with the parameters: ANN, AAN, and AAA.

Models	AIC	RMSE	MAE	MAPE	MASE	ACF1
ANN	298.9745	3.05	2.46	29.63	0.67	0.24
AAN	301.3425	3	2.37	28.73	0.64	0.23
AAA	314.3311	2.68	2.2	25.4	0.59	0.24

Table 4: Fitting the models using the 'ets' function

The model with the lowest AIC value (lower the AIC, the better the model) had the AAA parameters, and it was evaluated using the testing set.

Model	AIC	RMSE	MAE	MAPE	ACF1
ANN	47.1972	2.48	2.08	33.71	-0.16

Table 5: Evaluation of the ANN model

4.1.3. ARIMA Model Results

The time series required no differencing, as required for the ARIMA fitting. The ACF and PACF plots were analysed to obtain the "p,d,q parameters" needed to build the model.

The auto-regressive component (p) is chosen based on the lag where the PACF plot cuts off. As there was no differencing, the (d) component remains 0. The MA order (q) is chosen based on the lag where the ACF plot cuts off.

As the first lag of the PACF is not significant, it meant the AR component was first-order i.e. less than 1. Zero was the value taken, and 1 was selected for the MA component. As the differencing was 0, the first model built was a (0,0,1) model.

The (0,0,1) model had an AIC of 246.19. Closely related models were also built to investigate the most suitable:

i. (1,0,0) – AIC of 246.44
ii. (0,0,0) – AIC of 247.37
iii. (2,0,0) – AIC of 247.55
iv. (1,0,1) – AIC of 248.15

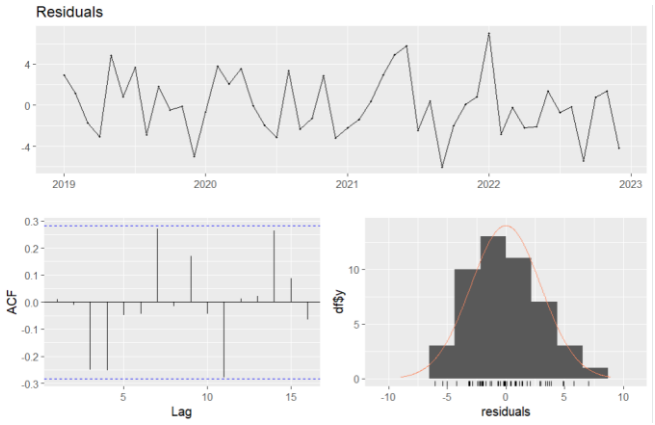

Figure 17: Residuals of the time series model

Residuals are the difference between the actual values and the predicted values. The graph shows that the residuals are generally distributed around zero, with a few outliers. This suggests that the time series model is a good fit for the data, and it captures the underlying structure of the data.

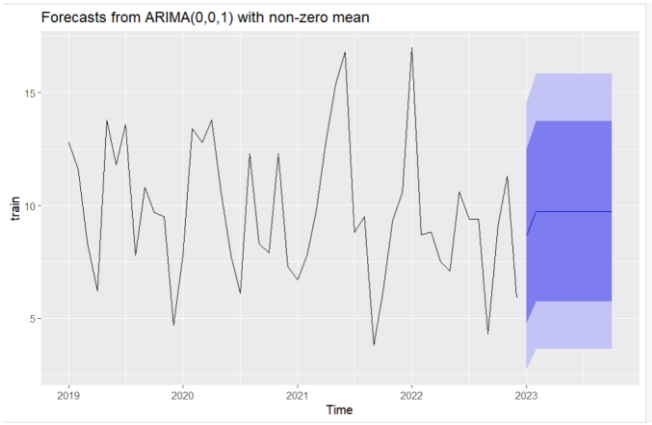

Figure 18: Forecast for the ARIMA model

The result of this graph is the prediction of an increase in forecast over time at a constant rate.

The fitted model was also evaluated with a resulting accuracy of the following metrics:

Model	RMSE	MAE	MASE	ACF1
(0,0,1)	2.952	2.384	0.645	0.011

Table 6: Evaluation of the ARIMA (0,0,1) model

4.2. Logistic Regression Results

The summary of the logistic model revealed an AIC of 112.72, which represents the measure of the quality of the model. Generally, the lower the AIC, the better the model.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.0699     2.0481  -1.987   0.0469 *
age            3.7304     1.2186   3.061   0.0022 **
weight         1.6007     1.4362   1.115   0.2651
fitness_score  1.2151     1.3316   0.912   0.3615
genderFemale   0.1334     1.5805   0.084   0.9327
genderMale     1.8435     1.5511   1.189   0.2346
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.60  on 86  degrees of freedom
Residual deviance: 100.72  on 81  degrees of freedom
AIC: 112.72

Number of Fisher Scoring iterations: 4
```

Figure 19: Logistic model summary

The intercept (-4.0699) represents the log odds of the outcome event occurring when the predictor variables are equal to zero. For example, the coefficient for the age variable indicates that the log odds of the event occurring increases by 3.7304 for a one-unit increase in age when all variables are constant. A 100.72 residual deviance shows how well it measures the goodness of fit of the model.

To improve the model, backward step-wise regression was applied. This iterative process aims to reduce the AIC of the model by removing non-significant variables.

```
Step:  AIC=108.19
cardiac_condition ~ age + genderMale

             Df Deviance    AIC
<none>            102.19 108.19
- genderMale  1   107.80 111.80
- age         1   114.06 118.06
```

Figure 20: Backward step-wise regression – final model

```
  Chi-squared: 7.269
           df: 8
      p-value: 0.508

Summary: model seems to fit well.
```

Figure 20: Hosmer & Lemshow Test

A confusion matric table was generated which contained essential metrics such as accuracy, precision, recall, and the F1-score.

Confusion Matrix	
Accuracy	**0.78**
Sensitivity	**0.81**
Specificity	**0.75**
PPV	**0.72**
NPV	**0.83**

Table 7: Confusion Matrix table

5. CONCLUSION

Key findings and insights were observed, highlighting the adequacy and performance of the final models for both time series and logistic regression.

Key findings:

(a) Time Series Analysis:

- The mean average method proved to be the most accurate forecasting method for the simple time series model, with low RMSE, MAE, MAPE, MASE, and ACF1.
- Exponential smoothing models, particularly the triple exponential model, successfully captured the seasonality in the time series data, outperforming other models.
- The ARIMA (0,0,1) model demonstrated good fit and accuracy in predicting the increasing trend in the time series data.

(b) Logistic Regression:

- The logistic regression model predicted the binary outcome effectively, with an accuracy of 78%.
- The ‘Age’ variable was shown to be a significant predictor, indicating a high association the presence of a cardiac condition.
- The final model, obtained through backward step-wise regression, demonstrated improved AIC and model fit.

Limitations & Future Work:

(a) Time Series Analysis:

- Exploring advanced variants of the ARIMA and exponential smoothing models could capture complex seasonal patterns more accurately.
- Future work could implement time series cross-validation to obtain more realistic estimates over time.
- Complex models assume stationarity. It is important that the data exhibits non-stationary pattern before fitting these models.
- The choice of frequency might have an impact on detecting and modelling seasonality. Some models are better suited for certain frequencies to capture either short-term or long-term patterns. Future work could involve building models at different frequencies to find the optimal one.

(b) Logistic Regression:

- Stepwise regression can sometimes lead to overfitting, and the final model can be highly dependent on the order in which the variables are removed. Regularization methods to counter this, such as Lasso, might be considered for future analysis.
- To provide a more accurate estimate of the model's generalization performance, implementing k-fold cross-validation might be a useful addition to the modelling process.
- Logistic regression performs best with predictors that have strong correlation with the target variable. Future work could address the multicollinearity issue with step-wise regression, resulting in accurate coefficient estimates and identifying important predictors to be used for analysis.

7. REFERENCES

[1] H. Madsen, *Time Series Analysis* (Chapman & Hall/CRC Texts in Statistical Science). CRC Press, 2007.

[2] G. Kirchgässner and J. Wolters, *Introduction to Modern Time Series Analysis*. Springer, 2008.

[3] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[4] K. Neusser, *Time Series Econometrics* (Springer Texts in Business and Economics). Springer International Publishing, 2016.

[5] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining* (Intelligent Systems Reference Library). Springer International Publishing, 2014.

[6] T. J. Hastie, *Generalized Additive Models*. CRC Press, 2017.

[7] S. Menard, *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications, 2010.

[8] R. Krispin, *Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R*. Packt Publishing, 2019.

[9] J. M. Hilbe, *Practical Guide to Logistic Regression*. CRC Press, 2016.