University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

Time Series Analysis for Hospital Admission Forecasting: A Comparative Study of Predictive Models

Neha Ingavale
2211488

Supervisor: **Maldonado Felipe**

November 24, 2023
Colchester

Abstract

This dissertation uses a synthetic dataset that replicates real world admission patterns to study the accuracy of different time series models for forecasting hospital admissions. A range of models is covered in the study including basic time series method like moving average and exponential smoothing advanced neural network models like LSTM and RNN typical time series techniques like ARIMA and SARIMA and modern machine learning techniques like Random Forest and Gradient Boosting.

Also Facebook Prophet advanced forecasting model are included in the study by the research. Using two datasets one with outliers and the other without is a key component of our research. This makes it possible to compare the effectiveness of each model and their sensitivity to unusual data points.The need for efficient hospital admission management which has become more challenging as a result of ageing populations chronic illnesses and unexpected public health emergencies like the COVID-19 pandemic is the driving force behind the project.

Following the implementation and assessment of various modelsÂ the analysis includes initial preprocessing such as the identification of outliers and the assessment of data skewness. The findings show the advantages and disadvantages of each model and demonstrate how well they adapt to shifting hospital admission trend. By comparing predictive modelling in hospital admissions and concentrating on the models response to epidemics this research aims to advance healthcare management and analytics. Policymakers and hospital managers can use this information to allocate and manage resources more effectively.

Keywords: Time Series Analysis, Hospital Admission Forecasting, ARIMA, SARIMA, LSTM, RNN, Facebook Prophet

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

The growing demand for hospital services is creating more challenges for the healthcare industry worldwide. Factors like the ageing population increase in chronic diseases and unexpected public health crises like the COVID-19 pandemic are all increasing this rise. Because of these issues managing hospital admissions effectively has become important. Worldwide overcrowding is a common issue that hospitals must deal with. An increasing number of difficulties result from this situation such as longer wait times a decrease in the standard of care and increased strain on medical personnel and health-care resources. Beyond the immediate treatment of patients the effect of overcrowding also affect more general aspects of healthcare management and operational effectiveness at these facilities.

Managing hospital admission efficiently is very important because it is closely linked to the quality of patient care and how well healthcare systems work. Being able to predict how many people will be admitted to the hospital and doing it accurately has become a big challenge these days. If hospitals can forecast admissions correctly, they can plan better for the number of patients coming in. This means they can use their beds more effectively have the right amount of staff and maintain high standards of patient care. All this leads to better results for patients and makes the hospital run more smoothly.

Predicting hospital admissions accurately is a really complex task. The trickiness comes from how the things that affect admission rates can change and be hard to predict. Usual ways of guessing these numbers have big problems when thinking of factors like seasonal sickness or unexpected health crise suddenly changing the pattern of who is coming into the hospital. Plus even though there is way better at time series analysis and using things like machine learning using them right in hospitals is still something experts are figuring out. These modern models have to get the balance right between being correct and being able to shift when things change. And then there is the big question of how to handle private patient info in these models, which just makes it all even more complicated to use them the right way.

There are many of research out there on predicting hospital admissions. A lot of studies have come up with all sorts of method. They have used some advanced models like LSTM, CNN, and various hybrid model but also some advanced time series models like ARIMA/SARIMA. Some of these models are really good at factoring in extra things and detailed patient info which makes them more accurate in certain cases. But they often need a lot of data and sometimes getting that much data is not possible or it might not be right to use it due to privacy issues.

For real time prediction models like those using XGBoost algorithms, these models perform very well. Such kind of models can be very useful but these models are not much used in different healthcare environments yet. A big issue with these models is whether they can work well for all kinds of patients and in different hospitals. Also, because these models depend on data that is constantly being updated this can be an issue in keeping them accurate all the time.

1.2 Objective

The main goal of this dissertation is to make a valuable contribution to healthcare management field. This was done by using a statistical technique called time series analysis to make predictions about the number of patients who will be admitted to a hospital. To conduct this research a fabricated dataset is used. This dataset is designed to replica the original data that was used in a previous study called **"Machine Learning to Effectively Aid Bed Management in Kettering General Hospital."** [6] Using a synthetic

dataset helps address ethical concerns related to using real patient information. It also allows for a detailed examination of different method for making predictions. **The research has two main objectives**. First it aims to assess how well various time series models work in predicting hospital admissions under normal circumstance. Second it aims to test the performance of these models during times of public health crises which are represented by unusual data points in the dataset. The focus is to understand model performance during typical admission patterns and unexpected surge such as those experienced during a pandemic.

1.3 Methodological Approach

In this research there is a mix of methods being used. Starting with some basic time series algorithms like Moving Average and Exponential Smoothing to get a basic idea of the trend. Then moving towards some more complicated time series algo like ARIMA/SARIMA to understand detailed analysis. Also have implemented some machine learning model like RNN, LSTM, Random Forest and Gradient Boosting good for analyzing more complex changing pattern in data like the number of hospital visit. First the basic data processing step are done to handle the outliers as data also consider the sudden increase in patient admissions due to any reasons. Then the data is changed to weekly frequency to analyze it better. This help the analysis and gives a deeper understanding of how admissions happen.
In the end models is compared to see how these models work seeing which ones do a good job in a controlled setup. This way one can see which models are really up to the task and which ones might need a bit more work. The research aim to improve knowledge about predicting hospital admission using time series analysis. This is does by comparing different prediction method on synthetic dataset. This help to understand the advantages and disadvantages of these method on different scenarios of dataset which can benefit healthcare management.

CHAPTER 2

Literature Review

It is a fascinating field of research to see how hospital admissions for non-COVID conditions are affected by the initial COVID-19 lockout. This study which came out in 2022 provides insight into the impact that respiratory disorder, cardiovascular illnesses and cancer had in England, Scotland, and Wales. The utilisation of large datasets from OpenSAFELY for England, EAVEII for Scotland and the SAIL Databank for Wales makes this study unique. [1] It provide a thorough examination of weekly hospital admission rates following them from the time of the epidemic until October 25, 2020. Through the utilisation of a controlled interrupted time series analysis and statistical models such as Generalised Least Squares and Ordinary Least Squares Estimation the researchers give an in depth understanding of the patterns and percentage variation in admission rate post lockdown.

First off there was a big drop in the number of people getting admitted to hospitals for the conditions they were looking at. What is more this drop was even bigger when it came to planned admissions compared to the emergency ones. And this was not just a temporary thing even after the lockdown the admission rates stayed lower than before the pandemic. But the study does point out some issue like it is hard to tell if this was because people needed less healthcare or if it was more about disruptions in the services. For this dissertation this research is very important. This is a great example of how time series analysis may be used to understand hospital admission trends better. Plus it shows why it is important to think about outside factors and be careful about how you research studies like this.[1]

A significant study that stands out in the field of hospital admission prediction utilising advanced machine learning algorithm was carried out in the Madrid region of Spain. Aiming to predict daily hospital admission for respiratory and circulatory problems among individuals over 65 a group that according to the World Health Organisation accounts for a sizable share of death cases for these ailments the study was published in 2020. The study used novel approach combining Long Short Term Memory Networks (LSTM) and Convolutional Neural Networks (CNNs) and it ran from January 1, 2001 to December 31, 2013. To adequately capture the intricate spatiotemporal interactions between environmental factors such as hospital admission rates and air pollution pollen concentration this methodology was selected.

The average Root Mean Square Error (RMSE) for respiratory patients was 15.06 while for circulatory cases it is 13.98. The corresponding coefficients of determination (R2) was 0.88 and 0.84. These results are noteworthy. So what this study shows is how good CNN and LSTM algorithms are at predicting hospital admissions. But they did mention some limits like how much and how good the data is. Also they could notlook at some types of pollen and pollution because they just did not have that data. This research is super useful for researching ahead the time series analysis of hospital admission. It points out the advantages and limitations of using forecast model and it is like a starting point for this research especially since it uses neural network for figuring out hospital admissions over time. The way thing were done and the findings from this give a peek into how machine learning can be used in healthcare especially for predicting the number of new patients might visit the hospital based on different environmental factors.[2]

A major study carried out at the pediatric emergency department (PED) of the Lille Regional Hospital Centre in France is highly relevant to time series of patient admissions. The main goal of this study was to create forecasting models for the daily attendance at the emergency room. This study was conducted in the year 2012. This study is especially important because it predicted patient arrivals using the well-known Autoregressive Integrated Moving Average (ARIMA) method of time series analysis. The dataset gives their investigation a strong basis because it included a complete year's worth of daily patient attendance. As the ARIMA model is used in this work it shows how time series analysis can be used to accurately forecast hospital admissions which is

an important aspect of healthcare planning and management. The study demonstrated how to use ARIMA models effectively but it also raised questions about some of its inherent shortcomings including the way it uses historical data and its potential difficulty in adjusting to sudden changes in admission patterns. These findings are especially helpful for research ahead since they provide a methodological baseline and point out areas that need more investigation in the predictive modelling of hospital admissions.[3]

The fourth paper which contributes significantly to the field of hospital admission forecasting is a study that used predictive modelling to address hospital crowding at a large scale tertiary hospital in Chongqing, China. The important research was of the creation and use of a hybrid model that combines the Nonlinear Autoregressive Neural Network (NARNN) and Autoregressive Integrated Moving Average (ARIMA) model. A very Large dataset was used by the researcher that included daily and monthly patient admission data from January 2010 to October 2016. The methodology mentioned in the paper involved comparing the predictive power of the NARNN, SARIMA and hybrid SARIMA NARNN algorithm for new patient admission.

The study found that although the combined model did better in general especially in reducing error measured by things like RMSE, MAE, and MAPE, it was not always performed as well as than the SARIMA or NARNN model in all respect. Plus the study showed seasonal pattern in admission with difference seen in various month underscoring the complexity of hospital admission prediction. These finding is especially important for predicting new admissions because they put light on the difficulties that come after combining linear and nonlinear methods to forecast hospital admissions. This knowledge will be useful in creating more complex and potent predictive models for use in healthcare setting.[4]

Research from the UK teaching hospital provides an interesting viewpoint on the challenges of predicting hospital admissions. The goal of this study was to create a real-time prediction pipeline that would predict emergency admissions by using real time electronic health records from the hospital emergency department. The study showed the effectiveness of machine learning in this setting by achieving Area Under the Receiver Operating Characteristic (AUROC) scores ranging from 0.82 to 0.90 using

a collection of XGBoost classifiers on a dataset of 109,465 emergency department visit. The model achieved a mean absolute error far lower than the conventional benchmark essentially aggregating patient level probabilities to calculate the number of admission. The study which was carried out in Covid-19 pandemic period showed how models could adjust to changing operating conditions while maintaining their functionality. These observations are critical for predicting the number of hospital admissions because they highlight the potential of machine learning especially XGBoost classifiers in accurate and on time hospital admission forecasting. The study shows a good way to handle both health records right as they come in and also how it works well in a tough and dynamic situation which is a useful example for prediction of hospital admission this can help in understanding the advantages and disadvantages of using advanced analytic to predict hospital admission.[5]

A few important point for improving time series analysis were studied from the literature on hospital admission forecasting. The UK study shows the necessity for models that can adjust to unanticipated event and emphasises role of external factors on admission rates during COVID 19 lockdown. Also the paper by the Madrid study which show a best way is to combine conventional time series method with modern algorithms for machine learning such as CNN and LSTM to increase accuracy in challenging situation. The ARIMA technique research from the Lille Regional Hospital Centre highlights the need for more flexible approaches by showing the shortcoming of standard models in terms of how well they adjust to sudden change.
Comparatively the study from Chongqing, China show the advantage of mixing linear and nonlinear models to handle complex admission patterns with hybrid model which are NARNN and ARIMA. The last paper uses the model XGBoost classifiers for real time predictions which is the study of the UK teaching hospital that highlights the benefit of machine learning models in dynamic environments. To precisely forecast hospital admission these studies combine and support a multidimensional approach to time series analysis that integrate multiple method and ensure adaptability.

CHAPTER 3

Dataset Description

The dataset used in this dissertation is synthetic dataset which is used to generate fake real world hospital admission patterns to avoid the ethical and privacy concerns associated with using actual patient data. The dataset was taken from the research of the NHSx skunkworks project on bed allocation which can be found at: https://github.com/nhsx/skunkworks-bed-allocation where the Python code was used to create this dataset. The objective of this project was to help the NHS to manage its hospital and allocate beds more effectively. As the dataset is synthetic it is maintaining patient privacy while accurately reflecting admission trends has become possible.

The NHSx skunkworks project came up with a technique that was used for generating this fake data. This replica of data they generated include data that are common in hospital environments such as trends, seasonality, and random fluctuation. This dissertation synthetic dataset was carefully generated with a Python script that is meant to replicate real hospital admission situation. The idea is to make sure this data is good for testing and checking out different time series and machine learning models so it is set up to really duplicate what happen in hospital. The important thing in the script is to provide a detailed but still changeable dataset for studying how beds allocation can be done in the NHS. The way it creates different fields in the data shows the variety one would see in real hospital info. Basic things like naming folders or deciding how many entries to have, which is great for all kinds of research needs can also be modified. The dataset end up with two CSV files (patient_df.csv and historic_admissions.csv) and

two JSON files (hourly_elective_prob.json and specialty_info.json) this makes up the final dataset.[28] All of these files build a good foundation for doing in depth analyses and forecasts of patient admissions. Out of which historic admission data was used for analysis. **The way the synthetic data is generated is completely independent of the time series analysis done for this dissertation.**

The dataset got records of hospital admissions for a certain period. Every entry in this set has a timestamp showing when the admission happened this column is named 'ADMIT_DTTM'. To make it easier for time series analysis this timestamp column was split into three separate columns for the year, month, and day of each admission. This breakdown of the column 'ADMIT_DTTM' helps to understand the admission patterns over different times like daily, monthly, or yearly trend. There is one more column named 'Total' which gives info of a total number of admissions at that timestamp. Even though these variable are synthetic they are set up to look a lot like real hospital data.

The Structure of Dataset:

- Unnamed: 0: This column is like an index column, providing a unique identifier for each row in the dataset. Which dropped later as it is of no use. The data type of the column is int64 (integer type)

- ADMIT_DTTM: This column contain timestamps of admissions formatted as date and time. The timestamps in the dataset start from January 1, 1855 at 00:00 hours suggesting that the data covers a historical period. As it is synthetic data. The datatype of the column is: datetime64[ns] (datetime type)

- Total: This column represents the total number of admissions at each given timestamp. The datatype of the column is Total: int64 (integer type)

Before starting with the analysis some data preprocessing steps were done to make sure it was good quality and fit the research goal. These steps included cleaning up the data to get rid of any outliers converting data into weekly frequency and feature engineering.

This synthetic dataset is an important resource for improving the understanding of hospital admission trend. As it also contains outliers which shows this data also have sudden increases in admission. Its use in this dissertation provides insights that could

guide future research and real world application in hospital services and healthcare analytics as well as showing the potential of handling the dynamic trends with potential models.[6]

CHAPTER 4

Methodology

This section will outline the detailed step required for conducting the dissertation analysis and constructing the model. It will provide a thorough explanation of each method process and model used along with the reason behind their selection.

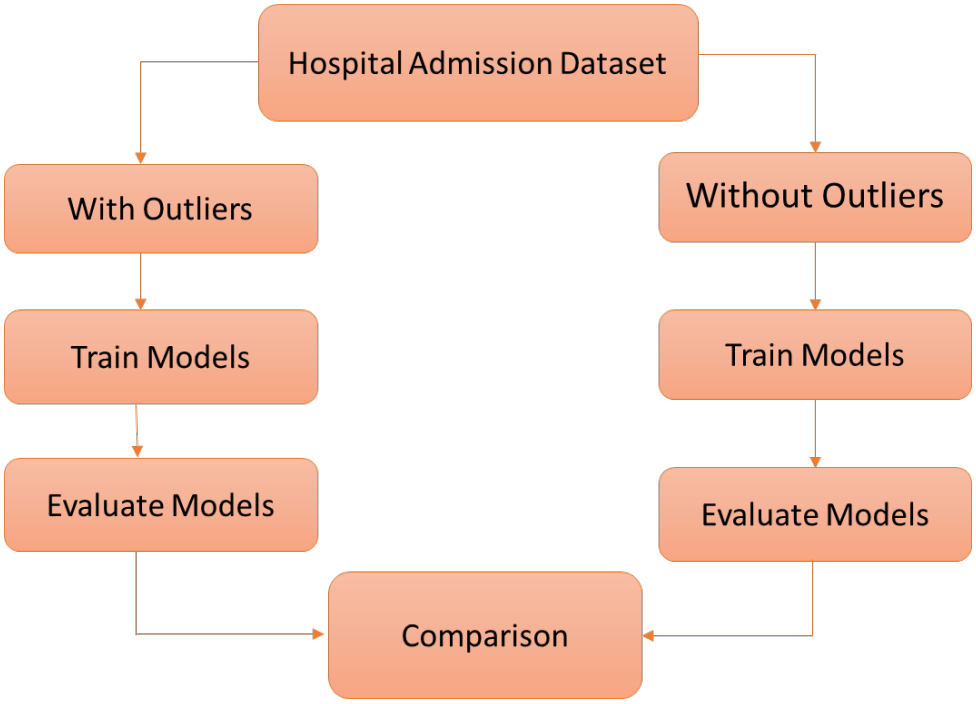

Figure 4.1: Methodology Flowchart

4.1 Data Preprocessing

4.1.1 Feature Engineering

After importing the data first step is to check the shape of the dataset which indicate that it has 87,673 rows and 2 columns. The next step is to start with data preprocessing which started with converting the ADMIT_DTTM column into datetime format. This conversion was important because it made it possible to perform increasingly complex time based manipulations and analysis. The date and time of each hospital admission were represented by timestamps formatted as strings in the original ADMIT_DTTM column of the dataset. The timestamps in the synthetic dataset indicated a historical period coverage, starting at 00:00 hours on January 1, 1855.

- Splitting timestamp:

The Year, Month, and Day columns were created separately from the ADMIT_DTTM column for an additional time analysis. This dissection made it possible to look into admission pattern on a yearly, monthly, and daily basis. Understanding the perception of hospital admissions such as observing any seasonal pattern or certain day with unusually high admission rates and identifying this kind of analysis.

- Dropping Redundant Column

The 'Unnamed: 0' column in the original dataset used as an index column, providing distinct identifier for every row. But this column was removed from the dataset because it was unnecessary for analysis. The dataset was simplified and analytical efforts focused on important variables by keeping only significant data.

4.1.2 Setting Index and Preliminary Checks

The dataset index was then set to ADMIT_DTTM. This rearrangement was planned to take advantage of timebased indexing an effective pandas library feature for time series analysis in Python. Later in the research resampling operation which were essential for transforming the data into a weekly frequency became simple when the datetime was given as the index.

The preliminary analysis of the dataset was done to verify the dataset accuracy and suitability for additional research. In this analysis null values were checked and the statistical characteristics of the dataset were summarised. An expected feature of a well constructed synthetic dataset is the absence of null values which the check verified. The dataset had 87,673 rows with a mean of 22.81 standard deviation of 4.80 and range of 6 to 46 admissions in the Total admissions column this is according to the statistical summary obtained from some preliminary insights into the data. The duplicate value check was done next on the dataset which due to the nature of the dataset representing time series data could indicate duplicate entries. The first instance of the duplicate was supposed to store and remove any further ones by setting keep = first in the code. But there is no duplicate entries pointing to a simple and well maintained data set generation.

4.2 Basic Trend and Pattern Analysis

4.2.1 Seasonal Decomposition Analysis

A seasonal decomposition of the hospital admissions data was carried out to find the basic trends. The trend, seasonal, and residual components of the dataset are divided into three discrete parts using this technique.

Long term progression in the dataset can be seen by the trend component which displays patterns that develop over time. Seasonality shows regular systematic change that occur at regular intervals such days of the week, months, or quarters. Mostly called as noise in the data the residual component is unpredictability not explained by the trend or seasonal component. Understanding the underlying qualities of the admissions data through this decomposition is important for targeting more in depth research and helping to choose the best forecasting model.[27]

4.3 Transformation to Weekly Frequency and Wavelet Transform

4.3.1 Converting to Weekly Data

The data was changed from a daily to weekly frequency to capture the more macroscopic trend in hospital admission. This conversion was done by summing up all of the admission for every week. The study could reduce the day to day fluctuation and allow the identification of more significant trend and pattern that appear over longer time periods by adding up the daily admissions into weekly total. For hospital operations the weekly aggregate is particularly important since it fit in better with staffing and resource planning which frequently use a weekly scheduling system.

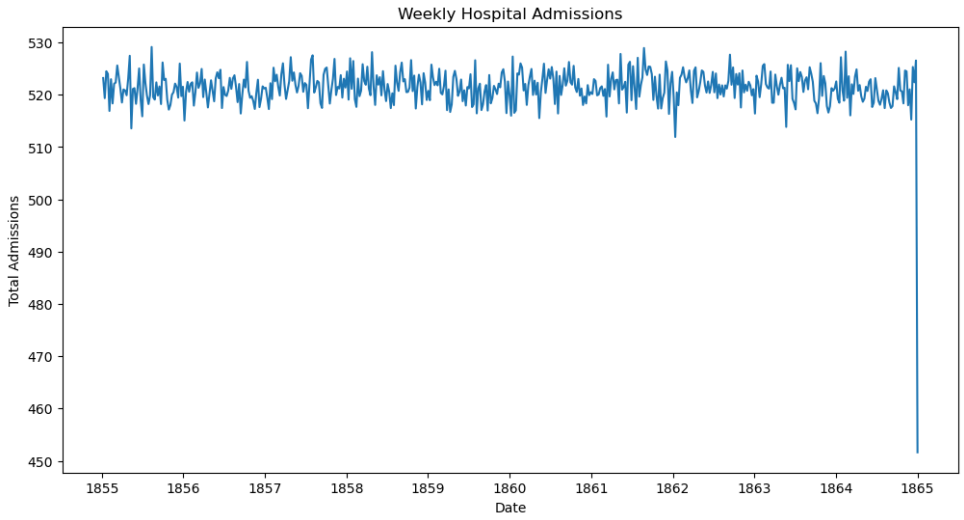

Figure 4.2: Weekly Hospital Admission

There were **two reasons** for moving to a weekly frequency: First reason is converting into weekly frequency made the presence of deeper trends more visible by lowering the noise in the data that could be related to daily changes.

Second it reduced the possibility of errors that may occur from particular occurrences or anomalies on particular days of the week. The conversion to weekly frequency also made it easier to compare admissions data with other sources that often provide information every week like public health reports or policy changes. This allowed for more integrated study of admissions data related to these factors.

4.3.2 Wavelet Transform Analysis

The next operation done on the dataset is that the weekly time series data was subjected to the Wavelet Transform. Wavelet Transform is advanced mathematical method in which data is break down into different frequency components to analyse time series on different scales.[18]

Because it can capture both the time related and frequency related information the Wavelet Transform is very useful in time series analysis. This matters a lot when you are dealing with data that changes in different ways over time like how often people are admitted to a hospital. The Wavelet Transform makes it possible to identify both localized and global trends and patterns in the data when it comes to hospital admission time series. Because it is really good at picking up on things like disease outbreaks, new health policies, and public health actions this method can spot both quick and slow changes that usually happen in healthcare. It can also find complicated patterns in the data like repeating cycles of hospital admission that other methods might miss but Wavelet Transform can identify it. This allows for an indepth understanding of the changes that take place in hospital admissions. This enhanced understanding plays a crucial role in supporting the predicting accuracy of later models.

4.4 Identifying and Handling Outliers

Outliers can have a significant impact on statistical analysis and predictive modelling in time series data. In this study the Interquartile Range (IQR) method was utilized to identify outliers in hospital admissions data. This method calculates the IQR as the difference between the data 75th percentile (Q3) and 25th percentile (Q1). Points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR were classified as outliers. [29] This approach provided a statistically robust method for outlier detection.[19]

4.4.1 Data Skewness Analysis

Following outlier detection skewness in the data was assessed. Initially a logarithmic transformation was applied to address the skewness as this technique is effective in normalizing right skewed data. However upon further analysis it was found that the log transformation did not significantly improve the skewness or the model's performance. The observation that the log transformation had limited impact led to the decision to forgo this step. Instead two distinct datasets were prepared: **one with outliers and the other without outliers.** This approach allowed for a direct comparison of model performances under different data conditions providing insights into each model's sensitivity to extreme values.

4.5 Correlation and Autocorrelation Analysis

4.5.1 ACF and PACF plots

Autocorrelation and partial autocorrelation functions are key tools in time series analysis. They are really useful for figuring out the best kind of model that captures the pattern and trends in time series data. The selection and modelling of predictive model in the context of hospital admission forecasting can be affected by understanding the correlation between observations at various time lag.

An Autocorrelation Function (ACF) plot of the weekly admissions data was made to study the relationship between the data over various time period. This ACF plots shows the dataset correlation with its own historical values across a range of time lags. [20] In simpler words the plot shows the relationship between the total hospital admissions for today as well as those for yesterday the day before and so on. The ACF plot is very important as it allows to observe both the overall trend of how the data is related across time and any potential seasonal trends.

At the same time, in the code, a Partial Autocorrelation Function (PACF) plot was also created. Now, this is different from the ACF because you see the ACF looks at everything in between a time lag and present, but the PACF focuses on the connection at a specific time lag but it takes into account and adjusts for the previous time lags.

Understanding what these plots mean is really important for building the model. For

example in PACF plot if there is sudden drop off it suggests that an Autoregressive (AR) model would be good. But when if the ACF plot is more like slowly going down that could be a sign that maybe a Moving Average (MA) component is what is needed. For hospital admission data which can be affected by many things like disease outbreaks or changes in policies, these plots are useful. They show how past admissions can affect future ones. This makes them key tools for choosing the right model and adjusting it to fit the data correctly. [17]

By carefully looking at the ACF and PACF plots, one can determine the most suitable time series model for predicting hospital admissions. This involves determining if there is a need to adjust the data to make it stay consistent over time this is common practice in forecasting. Also it is about seeing how past events in the data are connected to what might happen in the future which is important for making good predictions. One can understand that these autocorrelation analyses are used to make sure that the forecasting models picked are solid and accurately reflect the time related patterns in the hospital admissions data.[17]

4.6 Implementation of Time Series Algorithms

4.6.1 Model Training and Testing

To check how well the time series models work, the dataset was split into two parts: 80% which is commonly done before training and testing the model, and the remaining 20% for testing how accurately the model predicts. As there are two datasets one with outliers and one without both datasets were separately split.

4.6.2 Moving Average and Exponential Smoothing Implementation

Moving average and exponential smoothing models are the basic methods in time series forecasting and they are very suitable for hospital admission data which frequently has seasonal and trend components. The aim of creating good predictive models for hospital admission was the main reason for using these algorithm. This helps ensure resources are used well and patient care is managed better.

- **Moving Average Model**

The Moving Average method is a fundamental time series forecasting technique. It calculates the mean of the observations inside the given window and the forecast for the following period is based on this mean. The technique focus attention on longer term trends or cycles while reducing short term fluctuation.[7]
The training data was used in a Moving Average model which figures out the average number of admissions over a certain period. Since the data in this study is cyclical a window size of four weeks was used to catch the monthly trends in admissions. Then prediction for the test period is made using the latest value of the moving average from the training set. The model's assumption that the latest trend will keep going is seen in the forecast's repetition across the test set. A standard for the model's forecasting ability was established by comparing how the moving average forecast did against the actual admissions in the test data.
Formula

$$\overline{Y}_{t+1} = \frac{1}{n} \sum_{i=t-n+1}^{t} Y_i$$

Where:

- $\overline{Y}_{t+1}$ is the forecast for the next period
- n is the window size
- Y_i are the observations.

[7]
The window size was set to 4, and the prediction for the next period (like the next week) is just the average of how many people were admitted in the past four weeks. This way of forecasting works well for hospital admissions because it can pick up on the regular ups and downs of how many patient come in especially if there is a pattern that repeats every week.

- **Exponential Smoothing Model**

Another model used was exponential smoothing which is different from the moving average because it gives past observations weights that decrease exponentially. This

means that the prediction is more impacted by recent admissions than by earlier data. Exponential Smoothing is a method where recent data is given more weight meaning it is seen as more important while older data is considered less important. This method works well for predicting things like hospital admissions because what is happening recently can often tell you more about what will happen in the future.

Formula

$$\overline{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\overline{Y}_t$$

Where:

- $\overline{Y}_{t+1}$ is the forecast for the next period
- Y_t is the current observation
- $\overline{Y}_t$ is the current period's forecast
- α is the smoothing level.

[8]

After trying out different parameters the smoothing level parameter was set at 0.2 to balance the weight given to recent versus older admissions data. Forecasts for the test set time were created using the fitted model. Similarly to the moving average model the exponential smoothing forecast performance was evaluated to provide information on how accurate and dependable it is in predicting hospital admissions.[8]

- **Forecast Evaluation**

The forecasts generated by each model were meticulously evaluated on both datasets one with outliers and the other without using performance metrics adjusted to assess forecast accuracy. This comprehensive analysis not only understand the advantages and limitations of each model but also provided crucial insights into their applicability to the specific dynamics of hospital admission data.

An important part of the study was putting these time series algorithms into practice because they gave the more complicated model that came after a baseline comparison. To plan and manage healthcare services well it is really important to understand the

patterns over time of hospital admissions. This can be done by using methods like exponential smoothing and moving average forecasts as basic approaches.

4.6.3 Machine Learning Models Implementation

- **Random Forest**

The Random Forest algorithm is a kind of advanced machine learning method that belongs to a group called ensemble learning. What it does is create a bunch of decision trees when it is learning and then it uses the average of what all these trees say to make its predictions. This method is really good for things like guessing how many people will be admitted to a hospital.[9] Random Forest is great at understanding complicated and non straight forward connections in data. This is really useful for things like hospital admission patterns where lots of different factors come together in ways that are not easy to predict. Random Forest reduces the possibility of overfitting, which is a common problem in time series analysis by averaging the predictions from several trees.[21] This model is good at dealing with the different kinds of data you often find in healthcare records like numbers and categories and can process them well. Planning and understanding are improved by Random Forest findings about the variables (such as day of the week and time of year) that have significant impact on hospital admission.

Formula

$$\text{Random Forest Prediction} = \frac{1}{T}\sum_{t=1}^{T}\text{Decision Tree}_t(x)$$

Where:

- T is the total number of decision trees,
- $\text{Decision Tree}_t(x)$ is the prediction of the t-th decision tree for input x.

[9]

When looking at past admission data for predicting hospital admissions Random Forest is used to consider things like date, time, and maybe outside factors like local events or weather. This method works really well for this because it can handle complicated links between these variable. Due to its way of grouping it is able to make predictions that are more right and can deal with the noise and ups and downs that are in time series data.

The model was set up and learned from the dataset paying special attention to features like 'Year', 'Month', and 'Day'. These are really important for understanding the time related patterns in hospital admission. After training the model's predictions were evaluated against actual admission data to see how good it is at forecasting. In summary Random Forest's ability to adapt its strength and how easy it is to understand makes it a really strong tool for time series analysis in healthcare especially for predicting patient admissions where being accurate and knowing which variables are important is key.

- **Gradient Boosting**

Gradient Boosting is a quite advanced model of machine learning and it is part of a group known as ensemble methods. What it does is build a model step by step using lots of decision trees. Each new tree works on fixing the errors the previous ones made. This method is really good for both predicting numbers (regression) and sorting things into categories (classification) including forecasting over time.[21]
Formula

$$\text{Gradient Boosting Prediction} = \sum_{t=1}^{T} \alpha_t \cdot \text{Decision Tree}_t(x)$$

Where:

- α_t is the weight assigned to the t-th tree,
- Decision Tree$_t(x)$ is the prediction of the t-th tree.

[10]
This method is super flexible and can adapt to complicated and changing patterns in data which is perfect for things like predicting how many people will go to the hospital. It is especially good at understanding complex relationships in data like in healthcare where lots of different things can affect how many patients there are. By fixing errors one by one it gets more accurate and makes fewer mistakes overall. Just like Random Forest Gradient Boosting can also show which parts of the data matter most for its predictions which helps to understand what really affects hospital admissions.[23]
For forecasting hospital admissions, Gradient Boosting looks at past data and learns from the patterns and seasonal changes. Its ability to keep improving with each step

makes it a great fit for the ever changing world of healthcare. In this study, it was used to understand the connection between time related features (like 'Year', 'Month', 'Day') and the number of people admitted to the hospital. The model learned from some of the data and then made predictions on a test set. These predictions were then checked to see how well the model can forecast future admissions. Hence Gradient Boosting's precision its skill in handling tough patterns and its clear results make it a really useful tool for predicting things over time in healthcare. It offers a sophisticated way to guess patient admissions helping hospitals plan better and improve their services.

4.6.4 ARIMA and SARIMA

ARIMA (Autoregressive Integrated Moving Average): ARIMA is a popular statistical way of looking at and predicting data over time. It is popular for handling different kinds of time series patterns.

SARIMA (Seasonal ARIMA): SARIMA adds a seasonal bit to the ARIMA model, making it good for time series data that has seasonal changes.

Components of ARIMA: Autoregressive (AR): This part of the model looks at the connection between an observation and a certain number of past observations (p). It is based on the idea that current values are connected with past values in time series.

Integrated (I): This part involves differencing the time series data one or more time (d) to make it stationary meaning the data keep a constant mean and variance over time. Differencing helps in removing trends or seasonal structure.

Moving Average (MA): This component models the connection between an observation and a residual error from a moving average model applied to lagged observation (q).[24]

Components of SARIMA: SARIMA includes the non-seasonal (p, d, q) and seasonal (P, D, Q, s) part. The seasonal elements are similar to the non-seasonal ones but are related to the seasonal period of the data. Seasonal Autoregressive (SAR): Capture the connection between an observation and a certain number of lagged observations in the seasonal period (P). Seasonal Differencing (SD): Involves differencing the series at seasonal intervals (D). Seasonal Moving Average (SMA): Models the connection between an observation and a residual error from a moving average model at seasonal lags (Q). Seasonal Period (s): The length of the seasonal cycle in the data (like 12 for monthly data with yearly seasonality).

ARIMA Model Formula

ARIMA(p,d,q)
Where:

- p is the number of lag observations included (AR part),
- d is the number of times the data have been differenced (I part),
- q is the size of the moving average window (MA part).

[11]

SARIMA Model Formula

The SARIMA model is denoted as SARIMA(p,d,q)(P,D,Q)s, where:

p : Number of autoregressive terms (AR part).

d : Degree of differencing (I part).

q : Number of moving average terms (MA part).

P : Number of seasonal autoregressive terms.

D : Degree of seasonal differencing.

Q : Number of seasonal moving average terms.

s : Length of the seasonal cycle.

[11]

ACF and PACF Analysis:

AR (p): Number of autoregressive terms. Indicated by significant spikes in the PACF plot. MA (q): Number of moving average terms. Suggested by significant spikes in the ACF plot. These two terms after looking into ACF and PACF plots were used for ARIMA and SARIMA.

In hospital admissions forecast, these models are very useful for a couple of reasons: Handling Seasonality with SARIMA: Hospital admissions often have seasonal patterns (like flu season). SARIMA can effectively model and predict these seasonal patterns. Capturing Trends with ARIMA: If hospital admissions data show trends but not seasonality ARIMA is more fitting. It can identify underlying trends in admission rates over time.

ARIMA Model Fitting:

An ARIMA model, with the order, decided from the ACF and PACF plots, was then fitted. For example, an ARIMA(0, 0, 1) model means no autoregressive terms no differencing and one moving average term. Both the dataset with and without outliers was fitted using the same parameters.

Evaluation of ARIMA:

The model's forecasts were evaluated against a test dataset of admission data to check its accuracy and how reliable it is in predicting future admissions. A comparison of model performance in the presence of outliers and without outliers was done. For which different metrics were used MSE and MASE.

Grid Search for Parameter Optimization:

A grid search method was used to find the best mix of parameters for the SARIMA model. This method goes through lots of different parameter combinations systematically finding the best one based on the Akaike Information Criterion (AIC) Schwarz Bayesian Criterion (SBC), and the LjungBox Qtest.[25]

SARIMA Model Fitting and Evaluation:

The parameters found from the grid search were chosen and the SARIMA model was then fitted to the hospital admission data one with outliers and one without. These were the parameters a SARIMA(1, 1, 0)(1, 1, 1, 52) model means first-order autoregression and differences in both the non-seasonal and seasonal parts, with a seasonal period of 52 weeks. The model's accuracy in forecasting was checked using the same performance measures as the ARIMA model. Both models provide a strong way to understand and predict hospital admissions, helping healthcare places get ready for future patient numbers. Their skill in using both past patterns and seasonal changes makes them very valuable in healthcare for planning resources and running things efficiently.

4.6.5 Neural Network Models: LSTM and RNN

LSTM and RNN are types of neural networks particularly adept at handling sequential data making them suitable for time series analysis like hospital admission forecasts.

RNN: Recurrent Neural Networks are designed to recognize patterns in data sequences such as time series by maintaining internal memory of previous inputs. RNNs are known for their ability to connect previous information to the present task. The basic

RNN has a simpler structure where the hidden state at time t is calculated as:

RNN Formula

The recurrent neural network (RNN) cell formula is given by:

$$h_t = \tanh(W_h \cdot h_{t-1} + W_x \cdot x_t + b)$$

Where:

h_t : the current hidden state,

h_{t-1} : the previous hidden state,

x_t : the current input,

W_h, W_x : weight matrices,

b : the bias term.

[12]

LSTM: Long Short-Term Memory Networks are a special kind of RNN, capable of learning long-term dependencies. LSTMs are particularly effective because they overcome the vanishing gradient problem common in traditional RNNs. LSTM units include several gates the forget gate, input gate, and output gate each with its function and corresponding mathematical formulas.[11]

Implementation of LSTM and RNN

Data Normalization

The data was normalized using MinMaxScaler to scale it within a range of 0 to 1. LSTM models in particular, are sensitive to the scale of the input data and normalization helps in speeding up the learning process and improving model performance.

Preparing Data for LSTM and RNN

The scaled data was separated into features and targets. For LSTM and RNN, the input needs to be reshaped into a 3D format, as LSTM expects data in the form of [samples, time steps, and features]. X_train, X_test, y_train, and y_test were prepared by splitting the dataset, maintaining a typical split of 80% for training and 20% for testing. X_train_w, X_test_w, y_train_w, and y_test_w were prepared by splitting the

dataset without outliers, maintaining a typical split of 80% for training and 20% for testing.

LSTM Model Structure

The LSTM model was constructed with two LSTM layers each with 50 units indicating the model's complexity and capacity to learn from the data. LSTM's ability to retain information over long sequences makes it ideal for forecasting based on past admission trends.

A Dense layer with one unit was added for the output, predicting the continuous value of hospital admissions.

RNN Model Structure

The RNN model used SimpleRNN layers with a similar structure to the LSTM model but with SimpleRNN units which are less complex than LSTMs.

Training and Prediction:

The models were compiled with a mean squared error loss function and optimized using the Adam optimizer. Both models were trained over 100 epochs with a batch size of 32 indicating how many samples are processed before the model is updated. Predictions were made on the test data.

Evaluation

The LSTM and RNN models predictions were evaluated using metrics like MSE, RMSE and others to assess their accuracy in forecasting hospital admissions.

Application in Hospital Admission Forecasting: Due to its complex structure with gates LSTM is particularly suited for hospital admission data with long-term dependencies. It can remember patterns like seasonal trends effectively.

As RNN is simpler than LSTM. RNNs are effective for shorter-term dependencies which can be useful in predicting admissions based on recent trends. In the realm of hospital admissions, LSTM and RNN models are used to predict future admissions based on historical data. Their ability to capture and learn from the temporal sequence of admissions data makes them highly effective for this application.

The LSTM model, with its capacity to remember information for long periods, is particularly suitable for hospital admission data that might have long-term dependencies (e.g., seasonal patterns, and long term trends).

RNNs, while simpler can still effectively model the patterns in admission data especially

when the focus is on short term dependencies.
Both LSTM and RNN models are powerful tools in the predictive analytics of healthcare data. Their implementation in forecasting hospital admissions can significantly aid in proactive resource management, staff allocation, and policy planning in healthcare facilities.

4.6.6 Facebook Prophet

Facebook Prophet is a modern, open-source forecasting tool designed for handling time series data that shows strong seasonal patterns. It is particularly user-friendly and robust against missing data and outliers, making it a popular choice for a variety of applications including forecasting hospital admissions.
Components of Prophet Model Prophet is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data.
Mathematical Foundation of Prophet The underlying model of Prophet is a decomposable time series model with three main components:
Trend (g(t)): Captures non-periodic changes. Seasonality (s(t)): Represents periodic changes (e.g., weekly, monthly, yearly). Holidays (h(t)): Accounts for holidays or events.
The mathematical formula for the Prophet model is:

Time Series Decomposition Formula

The time series decomposition formula is given by:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Where:

$y(t)$: the predicted value,

$g(t)$: the trend function,

$s(t)$: represents seasonal changes,

$h(t)$: accounts for effects of holidays or events,

ϵ_t : the error term.

[13]

Implementation and Evaluation with Raw Data

Raw Data Approach: Initially, the Prophet model was applied directly to the raw hospital admission data. This approach tests Prophet's ability to handle the inherent outliers and fluctuations in the raw data without any preprocessing.

Model Fitting and Forecasting: The model was trained on the unaltered dataset, where the 'ADMIT_DTTM' was set as the 'ds' (datestamp) and the total admissions as the 'y' variable. After fitting the model predictions were made for the test period, and the models performance was evaluated.

Performance Assessment:

The forecasted values were compared against the actual admissions in the test data to assess the accuracy of the model. Metrics like Mean Squared Error (MSE) were used to quantify the models performance.

Implementation and Evaluation without Outliers Data

Outlier-free Data Approach: To compare results, the same Prophet model was then applied to the data without outliers.

Model Fitting and Forecasting without Outliers Data: The training process was repeated without outliers data, maintaining the same structure for the Prophet model. Predictions were generated for the corresponding test period using the outliers free data.

This comparative approach between outliers data and without outliers data highlights the effectiveness of the Prophet model in different data conditions providing valuable insights for healthcare analysts and decision makers. It underscores the model's adaptability and accuracy in forecasting hospital admissions, crucial for strategic planning in healthcare facilities.

4.6.7 Model Comparison and Evaluation

Comparative Analysis

In the methodology a comprehensive comparative analysis was conducted to evaluate and contrast the performance of various time series forecasting models. The models implemented include Moving Average, Exponential Smoothing, Random Forest, Gradient Boosting, ARIMA, SARIMA, LSTM, RNN, and Facebook Prophet. Each model has its unique approach to handling time series data making this comparison critical in understanding their relative effectiveness in the context of hospital admissions forecasting.

Performance Metrics

To evaluate and compare the models the following performance metrics were used, each offering a different perspective on the models forecasting accuracy:

Mean Squared Error (MSE):

Mean Squared Error (MSE) Formula

The Mean Squared Error (MSE) is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where:

Y_i : Actual value at the i-th point in the dataset,

$\hat{Y}_i$: Predicted value at the i-th point,

n : Total number of observations in the dataset.

[14]

Measures the average squared difference between actual and forecasted values. Lower MSE values indicate better model performance.

Mean Absolute Error (MAE) Formula

The Mean Absolute Error (MAE) is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

Where:

$\left|Y_i - \hat{Y}_i\right|$: Absolute difference between the actual and predicted values at the i-th point,

n : Total number of observations.

[14]
Represents the average absolute difference, providing a straightforward interpretation of average error magnitude.

Root Mean Squared Error (RMSE) Formula

The Root Mean Squared Error (RMSE) is given by:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

Where:

MSE : Mean Squared Error as calculated above,

RMSE : brings the error metric back to the same scale as the data.

[14]
The square root of MSE offers error in the same units as the data facilitating comparison across different models.

Coefficient of Determination (R^2) Formula

The coefficient of determination is given by:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Where:

Y_i : Actual value,

$\hat{Y}_i$: Predicted value,

$\bar{Y}$: Mean of actual values.

[14]

Indicates the proportion of variance in the observed data that is predictable from the model.

1. Moving Average and Exponential Smoothing Forecasts:

MAE and RMSE: Provide clear insights into the average errors and the scale of errors, respectively.

MASE: Useful for comparing these simpler models to a naïve benchmark.

2. Random Forest and Gradient Boosting Models: MSE and RMSE: Important for understanding the magnitude of prediction errors. R-squared (R^2): Helps to evaluate how well these complex models capture the variance in the data.

3. ARIMA and SARIMA: MAE and RMSE: Provide insights into the average and squared errors, critical for time series data. MASE: Particularly relevant for these linear models as it compares them against a simple lagged model.

4. LSTM and RNN: MSE and RMSE: Crucial for evaluating the performance of these neural network models, especially given their sensitivity to the scale of input data. Accuracy: Can be a useful metric although it should be interpreted cautiously due to the models non-linear the nature.

5. Facebook Prophet: MAE and RMSE: Important to understand the average and scaled errors especially given Prophet's ability to handle outliers. MAPE: Useful for providing an intuitive percentage-based error measure, helping in communicating the model's performance. Prophet's robustness to outliers makes MSE and RMSE reliable metrics for evaluating its performance.

For each model the selection of metrics depends on the specific characteristics of the model and the nature of the time series data. While MSE and RMSE are universally applicable and provide a consistent basis for comparison MAE, MASE, and MAPE offer different perspectives that are useful in certain contexts. For instance, MASE is more informative for linear models like ARIMA and SARIMA, while RMSE and MSE are

critical for evaluating the prediction accuracy of machine learning models like Random Forest and Gradient Boosting. In the case of LSTM and RNN where the models are more complex and non-linear traditional metrics like MSE and RMSE are essential but their interpretation might require additional context. Lastly for models like Prophet which are designed to handle outliers and missing data effectively metrics like MAE and MAPE can provide a more intuitive understanding of model performance.

CHAPTER 5

Results and Discussion

5.1 Statistical Summary and Preliminary Observations

Null values: Being a synthetic dataset, it had no missing values, ensuring data completeness. A standard procedure for forward filling was prepared to handle any potential missing data, thus maintaining the continuity and integrity of the analysis. **Statistical Summary:** The dataset, comprising 87,673 entries reflected a moderate level of hospital traffic, with an average of approximately 22.81 admissions per day. The standard deviation for total admissions was 4.80, indicating some daily variability. Admission data spanned from 1855 to 1865, offering a decade's worth of information for analysis. The minimum and maximum daily admissions were 6 and 46 respectively highlighting a wide range in patient influx. Quartile values for admissions (19, 23 and 26 for the 25th, 50th and 75th percentiles, respectively) provided further insights into the admission distribution.

	Total	Year	Month	Day
count	87673	87673	87673	87673
mean	22.81	1859.50	6.52	15.73
std	4.80	2.87	3.45	8.80
min	6	1855	1	1
25%	19	1857	4	8
50%	23	1860	7	16
75%	26	1862	10	23
max	46	1865	12	31

Table 5.1: Descriptive Statistics

5.2 Wavelet transform

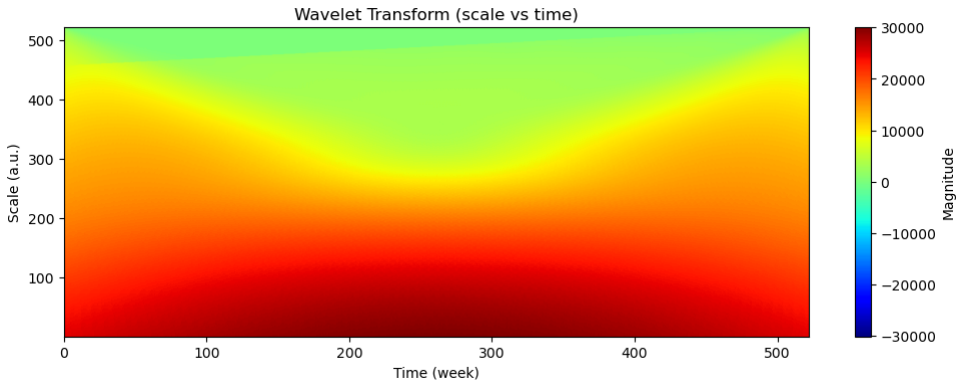

Figure 5.1: Wavelet Transform

The plot suggests that the variation in hospital admissions is not uniform over time. This could reflect external factors affecting admissions such as seasonal illnesses policy changes or other temporal influences. The continuous wavelet transform (CWT) reveals that most of the energy is concentrated at lower frequencies (higher scales) which indicates the presence of long-term trends in hospital admissions. These trends could correspond to demographic changes seasonal effects or healthcare policy impacts over a longer horizon. The intensity of colours (magnitude of wavelet coefficients) appears to diminish at higher scales. This suggests that while there are long-term trends their impact on the variability of admissions is not as strong as the short-term fluctuations

which are more pronounced at lower scales (higher frequencies). The higher magnitudes at lower scales indicate significant short-term fluctuations in weekly hospital admissions. This could be due to weekly patterns such as more admissions on specific days of the week or due to irregular events such as epidemics or natural disasters. The lack of distinct bright spots or strong colour contrasts at any scale across the time axis suggests the absence of sudden sharp changes in admission patterns that persist over a significant period. This could imply relative stability in the admission process without abrupt shifts that would require immediate attention in the forecasting model. Understanding the dominant scales (frequencies) can be critical for developing accurate forecasting models. Models that can account for both long-term trends and short-term fluctuations are likely to yield better predictions for future admissions. To explore the long-term trends and short-term trends implementation of simple methods like Moving average and Holt-winter some advanced models like ARIMA/SARIMA, LSTM, RNN, ML models like Random Forest, Gradient boost, and Facebook Prophet were implemented.[18]

5.3 Seasonal Component analysis

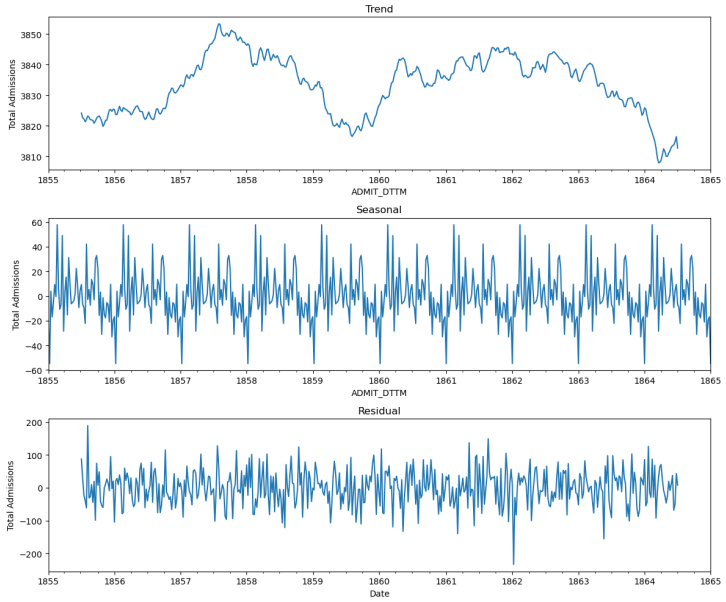

Figure 5.2: Seasonal Decomposition

The seasonal plot indeed shows a repeating pattern, indicating the presence of a seasonal cycle in hospital admissions. The regularity of peaks and troughs at consistent intervals

confirms a strong seasonal influence. This suggests that hospital admissions are affected by factors that vary cyclically possibly on a weekly or monthly basis. The amplitude (height) of these peaks and troughs can give insights into the strength of this seasonal effect. Higher peaks and deeper troughs indicate more pronounced seasonal variations. Identifying the exact nature of these seasonal factors (e.g., specific health issues prevalent during certain times of the year or behavioural patterns like increased accidents during weekends) would require additional context or domain specific knowledge.

Trend Component Analysis:

The trend component helps identify any long-term changes in the level of hospital admissions over the observed period. Depending on the plot, if there is a clear upward or downward trend it would indicate a long term increase or decrease in admissions. A relatively flat trend would suggest stability in admission numbers over time.

Residual Component Analysis:

The residuals, which represent the portion of the data not explained by the trend and seasonal components, ideally should show no pattern and resemble random noise. The presence of significant outliers in the residuals could indicate extraordinary events or anomalies that the seasonal and trend components do not account for. Considering the fact from the plot outliers analysis is done.[27]

5.4 Analysis of Outliers

Outlier Detection Methodology:

To identify outliers in hospital admissions data, the Interquartile Range (IQR) method was employed a robust statistical technique. This approach involved calculating the IQR for the 'Total' admissions within the dataset. Specifically, the first quartile (Q1) determined and third quartile (Q3) values, and defined outliers as data points that fell below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. This method is particularly effective for identifying extreme values that differ significantly from the central tendency of the data.

Identified Outliers:

The analysis revealed several weeks with unusually high or low admission numbers. The outliers were characterized by a notable deviation from the typical range of weekly admissions, indicating periods of extraordinary hospital activity. These outliers included

specific weeks such as:

ADMIT_DTTM	Total
1855-05-13	3658
1855-08-12	4010
1861-08-25	4004
1862-01-12	3613
1865-01-01	3337

Table 5.2: Outliers

Impact of Log Transformation:

Initially, to address the skewness in the data and potentially mitigate the impact of outliers log transformation was applied to the 'Total' admissions data. However upon reevaluating the data post transformation, the observation was that the outliers remained present. This indicated that the log transformation, while useful for normalizing the data distribution was not sufficient to diminish the impact of these extreme values. Also, after log transformation data was left skewed.

Creation of Separate Datasets:

Given the persistence of outliers post-transformation, two distinct datasets were created for further analysis: one including the outliers and one excluding them. This approach allows to assess the influence of outliers on the modeling efforts. By comparing the performance of various statistical and machine learning models on these two datasets aim was to understand the robustness of different models in the presence and absence of outliers.

The presence of outliers in hospital admissions data poses significant challenges and opportunities for data analysis. Findings for analysis highlight the importance of careful outlier detection and consideration in predictive modelling. The comparative analysis of model performance on datasets with and without outliers offers valuable insights into the data's characteristics and the suitability of various modelling approaches for healthcare data analytics.

Skewness of data

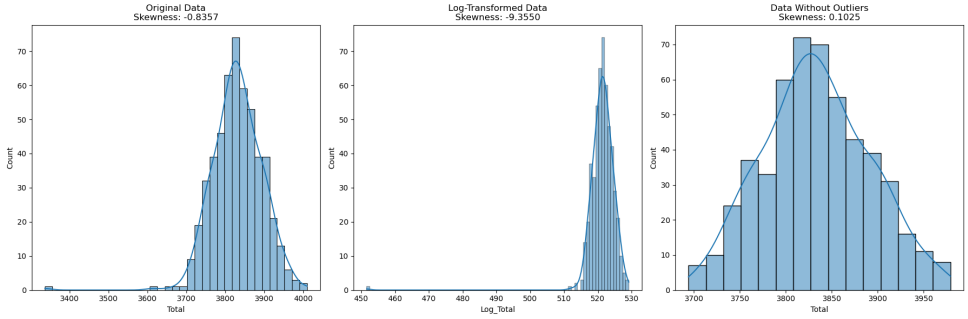

Figure 5.3: Skewness

From the analysis of the plot, an initial observation revealed a left-skewed distribution with a skewness value of -0.8357. To address this skewness log transformation was applied. However, this transformation unexpectedly intensified the skewness to -9.3550 suggesting an increased asymmetry in the data, likely due to the compression of the lower end data points. This led to a further investigation into the role of outliers in skewing the distribution. Upon removing the identified outlier entries the skewness was significantly reduced to 0.1025 indicating a more symmetric and normal like distribution. This progression of findings from initial skewness through the impact of the log transformation to the reduction of skewness after outlier removal highlights the intricate interplay between data transformations outlier presence and distribution shape in the admissions dataset. It underscores the importance of nuanced data preprocessing in healthcare analytics where understanding and adjusting for skewness and outliers is crucial for accurate analysis and modelling.

5.5 Stationarity and Autocorrelation Analysis

Stationarity Test Results:

The Augmented Dickey-Fuller (ADF) test was applied to the hospital admissions data after outlier removal to assess stationarity. The test yielded an ADF statistic of -22.8749, decisively below the critical value threshold of -3.4431 at a 1% significance level. With a p-value of 0.0, the null hypothesis of a unit root presence, and hence non-stationarity, was firmly rejected. This implies that the adjusted time series data is stationary showing constant statistical properties over time and no long term trends or seasonal effects.

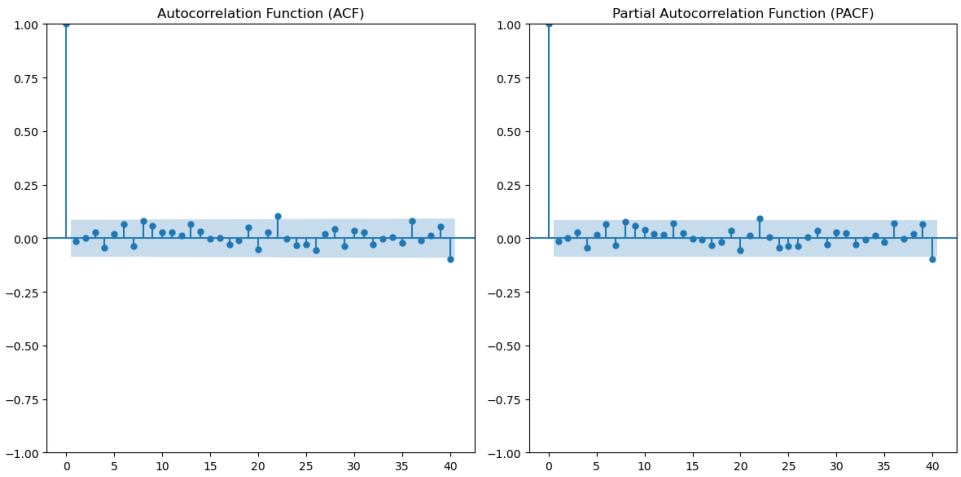

Figure 5.4: ACF and PACF

Also stationary test was done on a dataset with outliers and data seemed to be stationary hence there is no need to do differencing.[26]

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) Analysis:

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots further support the findings of stationarity. The ACF plot illustrates that autocorrelations quickly drop off after the initial lag remaining within the confidence interval for subsequent lags. This indicates no significant long term autocorrelation patterns. Similarly, the PACF plot shows that partial autocorrelations are also within the confidence bounds, suggesting that any correlation with past values does not persist beyond the immediate lags. These plots provide visual confirmation that the data, devoid of outliers, does not exhibit prolonged memory aligning with the principles of a stationary time series.

5.6 Interpreting Model Outcomes

5.6.1 Moving Average and Exponential Smoothing Forecast

Moving Average Forecasting:

The Moving Average model, which forecasts future values as the average of a fixed number of prior observations was applied to both the original dataset and the dataset adjusted for outliers. Using a window size of 4 the model exhibited slightly better performance metrics when trained on the dataset without outliers. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were notably lower at 3400.48 and 58.31 respectively compared to 5825.79 and 76.33 when trained on the original dataset. This suggests that the presence of outliers in the data can inflate the model's forecast error highlighting the importance of outlier treatment for improving forecast accuracy.

Exponential Smoothing Forecasting:

Similarly, the Simple Exponential Smoothing model, which applies exponentially decreasing weights to past observations was evaluated. When trained on the outlier adjusted dataset the model achieved a Mean Squared Error (MSE) of 3429.92 and a Root Mean Squared Error (RMSE) of 58.57. Conversely the same model on the original dataset resulted in MSE and RMSE values of 5827.87 and 76.34, respectively. The more favourable performance metrics on the adjusted dataset further reinforce the impact of outliers on forecasting accuracy.

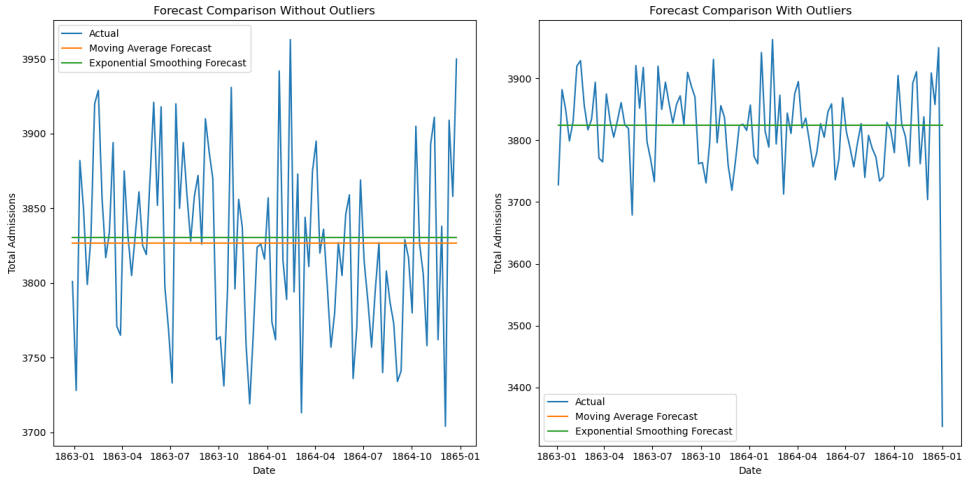

Figure 5.5: Moving average and Exponential Smoothing analysis

Comparative Analysis and Implications:

The comparative analysis of the Moving Average and Exponential Smoothing models on the two datasets underscores the influence of outliers on the predictive performance. The presence of outliers appears to degrade the model's ability to accurately forecast future admissions as evidenced by higher error metrics. This analysis affirms the necessity of careful data preprocessing including outlier removal to enhance the reliability of time series forecasting models in healthcare settings.

5.6.2 Machine Learning Models Forecast

Random Forest Forecasting:

The Random Forest algorithm, a robust ensemble machine learning method was employed to forecast hospital admissions. When trained on the original dataset the model yielded a Mean Squared Error (MSE) of 6460.31 and a Root Mean Squared Error (RMSE) of 80.38, with an accuracy of 98.45%. However, the performance improved on the dataset without outliers resulting in a lower MSE of 4132.15 and RMSE of 64.28, while the accuracy marginally increased to 98.63%. This improvement upon outlier removal underscores the influence of outliers on the model's prediction accuracy, suggesting that the Random Forest model is sensitive to extreme values in the data.

Gradient Boosting Forecasting:

Similarly, the Gradient Boosting Regressor, another powerful ensemble technique was applied to the datasets. On the original data, the model's MSE and RMSE were 6829.51 and 82.64 respectively with an accuracy of 98.43%. Training on the outlier-free data yielded better performance, with an MSE of 4357.67 and RMSE of 66.01, and a slight increase in accuracy to 98.60%. Like the Random Forest the Gradient Boosting model showed enhanced performance metrics when outliers were excluded from the dataset.

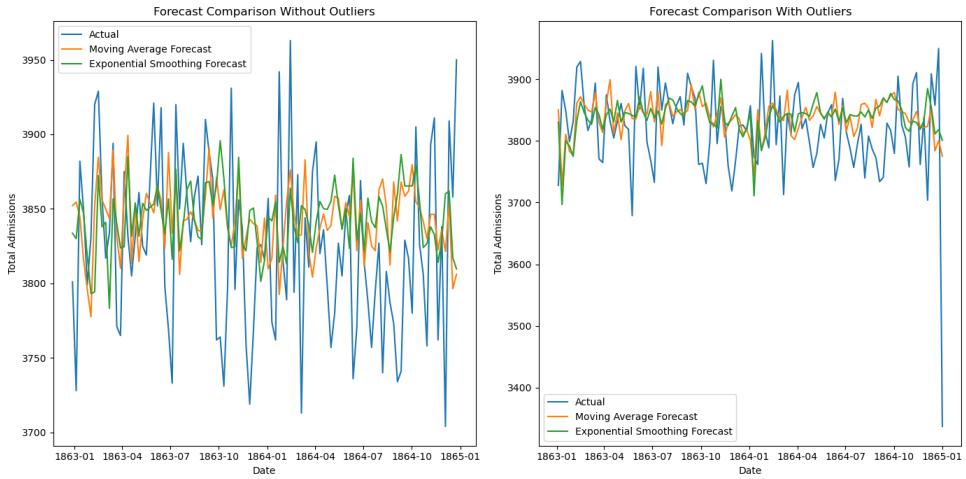

Figure 5.6: Random Forest and Gradient Boosting analysis

5.6.3 ARIMA and SARIMA Forecast

ARIMA Model Analysis:

The ARIMA (AutoRegressive Integrated Moving Average) model was employed to forecast hospital admissions data. For the dataset without outliers the ARIMA model (1, 0, 1) indicated an excellent fit with an R-squared value of 1.0, which typically signifies a perfect fit but may also imply overfitting. The model yielded a Mean Squared Error (MSE) of 3494.05 and a Root Mean Squared Error (RMSE) of 59.11. However, the model's performance on the original dataset, using the order (0, 0, 1), showed higher error values with an MSE of 6035.65 and an RMSE of 77.69, alongside a negative R-squared value suggesting a less satisfactory model fit. This contrast in performance metrics highlights the impact of outliers on the ARIMA model's forecasting ability.

SARIMA Model Analysis:

Given the identified seasonal component from the decomposition plot, the Seasonal ARIMA (SARIMA) model was also considered appropriate. The SARIMA model, incorporating a seasonal order of (1, 1, 1, 52) to account for yearly fluctuations, was applied to the datasets. For the dataset without outliers, the SARIMA model showed improved performance with an MSE of 4032.67 and an RMSE of 63.50. The model's MSE and RMSE were 5908.57 and 76.87, respectively, with a negative R-squared value, which further substantiates the previous findings regarding the influence of outliers.

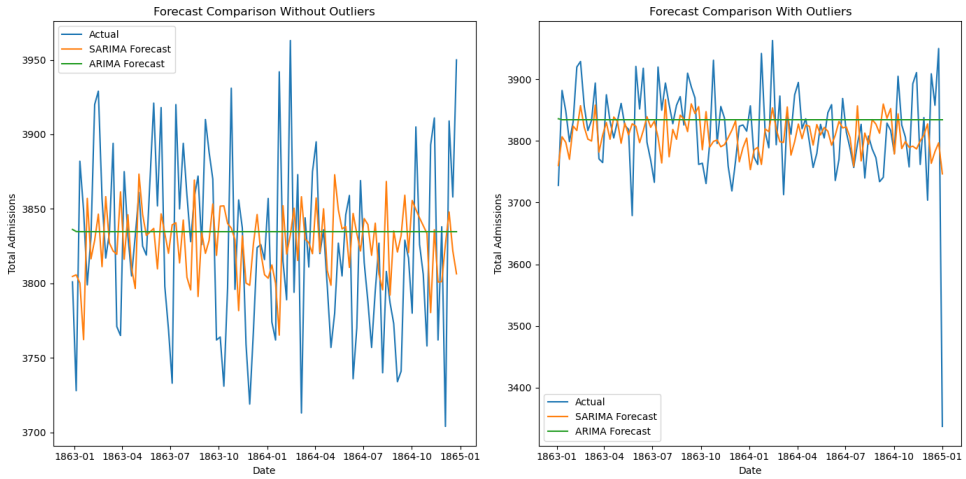

Figure 5.7: ARIMA and SARIMA Analysis

Comparative Insights and Model Suitability:

The comparative performance of the ARIMA and SARIMA models underscores the importance of accounting for seasonal patterns in the data and the detrimental effects of outliers on model accuracy. While both models performed better on the dataset without outliers, the SARIMA model's incorporation of seasonality appears to provide a more nuanced fit, reflecting the underlying seasonal behaviour observed in the admissions data.

5.6.4 LSTM and RNN Forecast

LSTM Model Performance Analysis:

Long Short-Term Memory (LSTM) networks, renowned for capturing long-term dependencies in time series data, were employed in this study. The LSTM models were trained on datasets both with and without outliers, revealing distinct performance characteristics. For the dataset without outliers, the LSTM model yielded a Mean Squared Error (MSE) of 0.0467 and a Root Mean Squared Error (RMSE) of 0.2162. However, the model's performance on the dataset with outliers displayed an MSE of 0.0097 and an RMSE of 0.0987. Intriguingly, the model's R-squared values were highly negative in both cases, suggesting potential overfitting or inadequacies in capturing the data's underlying patterns.

RNN Model Performance Analysis:

Recurrent Neural Networks (RNN), another class of neural networks suitable for se-

quential data, were also evaluated. The RNN model trained on the dataset without outliers exhibited an MSE of 0.0489 and an RMSE of 0.2211, while the model trained with outliers showed an MSE of 0.0093 and an RMSE of 0.0962. Like the LSTM models, the RNN models also demonstrated highly negative R-squared values, highlighting potential challenges in the models' predictive abilities for this specific dataset.

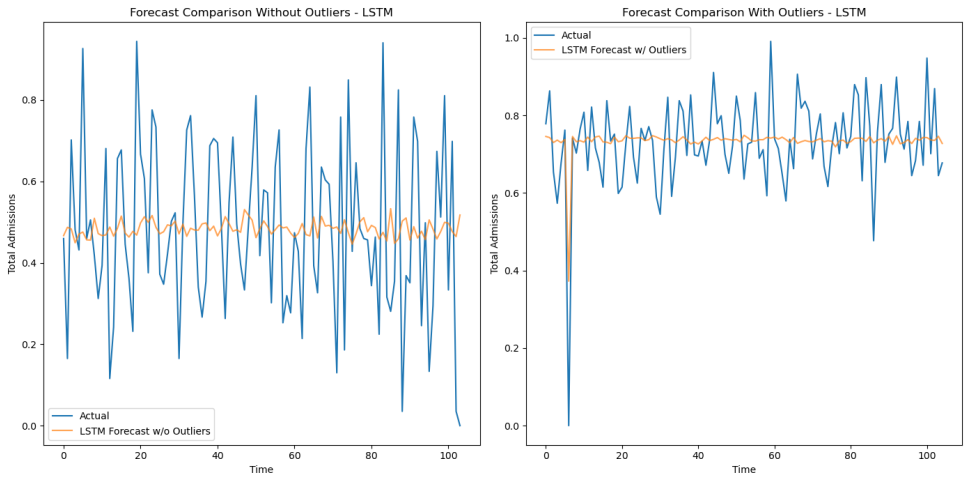

Figure 5.8: LSTM Analysis

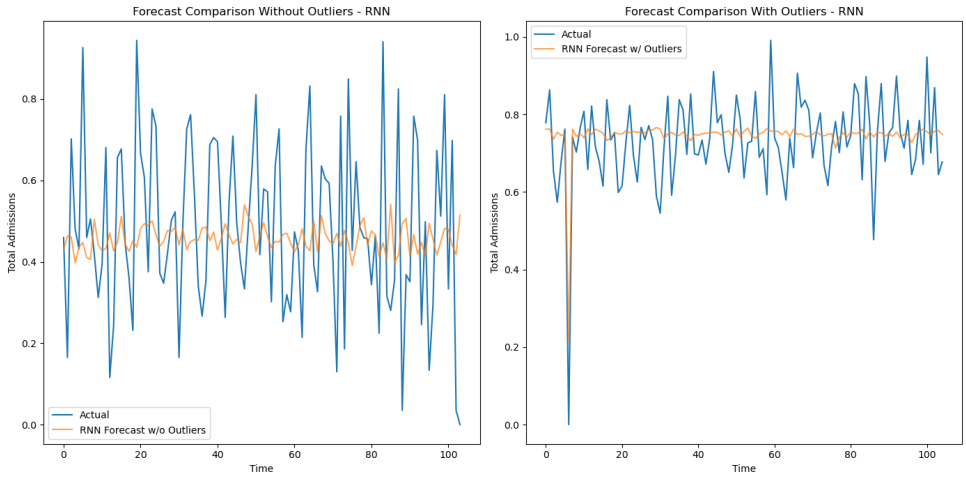

Figure 5.9: RNN Analysis

Comparative Insights and Model Suitability:

The LSTM and RNN models demonstrated better performance in terms of lower MSE and RMSE on the dataset with outliers. This counterintuitive result might be attributed to the complexity of these models and their sensitivity to the scale and nature of the input data. The negative R-squared values across all models signal a critical need for

further investigation into model configuration, training, and perhaps data preprocessing steps.

5.6.5 Facebook Prophet Forecast

Prophet Model Implementation and Evaluation:

Facebook's Prophet, a robust forecasting tool designed for time series data with strong seasonal patterns, was applied to both the original and outlier-adjusted hospital admissions datasets. The data was first aggregated into weekly frequency and then split into training and test sets. The training data was fitted to the Prophet model, and predictions were made for the corresponding test data periods.

Forecasting on Original Dataset:

For the original dataset, the Prophet model yielded a Mean Squared Error (MSE) of 6962.36, a Mean Absolute Error (MAE) of 60.48, and a Root Mean Squared Error (RMSE) of 83.44. Despite an R-squared value of 1.0, which typically indicates a perfect fit, the model's high MSE and RMSE suggest that it may not have captured the underlying admissions patterns effectively.

Forecasting on Dataset Without Outliers:

When applied to the dataset without outliers, the Prophet model demonstrated a slightly improved performance with an MSE of 6755.03, an MAE of 59.10, and an RMSE of 82.19. The similar trend in performance metrics indicates that the model's ability to forecast hospital admissions was moderately affected by the presence of outliers.

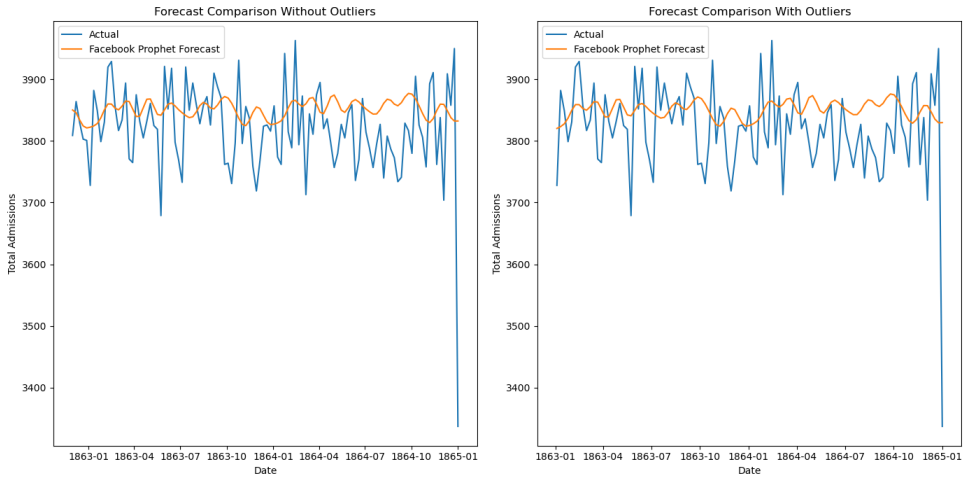

Figure 5.10: Facebook Prophet Analysis

5.6.6 Comparative Insights and Model Suitability

The results reveal that while Prophet is well-suited for capturing seasonal trends its performance in forecasting hospital admissions data was modest.

Comparative analysis of all methods:

Model	MAE	RMSE
Moving Average without outliers	47.048	58.314
Moving Average with outliers	52.386	76.327
Exponential Smoothing without outliers	47.506	58.566
Exponential Smoothing with outliers	52.384	76.341

Table 5.3: Moving Average and Exponential Smoothing Model Performance Metrics

Model	MSE	MAE	RMSE	R_squared
Random Forest with outliers	6460.31	80.38	58.50	-0.114
Random Forest without outliers	4132.15	64.28	52.55	-0.216
Gradient Boosting with outliers	6829.51	82.64	53.50	-0.178
Gradient Boosting without outliers	4357.67	66.01	59.08	-0.282

Table 5.4: Random Forest and Gradient Boosting Model Performance Metrics

Model	MAE	RMSE
ARIMA without outliers	48.262	59.110
ARIMA with outliers	53.658	77.689
SARIMA with outliers	56.074	76.867
SARIMA without outliers	52.681	63.503

Table 5.5: ARIMA and SARIMA Model Performance Metrics

Model	MSE	RMSE	R-squared (R^2)
LSTM without outliers	0.04673	0.21618	-103.936
LSTM with outliers	0.00973	0.09866	-114.113
RNN without outliers	0.04887	0.22105	-106.695
RNN with outliers	0.00925	0.09620	-127.505

Table 5.6: LSTM and RNN Performance Metrics

Model	MAE	RMSE
Prophet without outliers	59.104	82.189
Prophet with outliers	60.480	83.441

Table 5.7: Prophet Models Performance Metrics

The Moving Average and Exponential Smoothing models show decent accuracy without outliers but are highly sensitive to outliers as indicated by the significant increase in their error metrics. Random Forest on the other hand demonstrates a balance of accuracy and robustness slightly outperforming Gradient Boosting in both scenarios it maintains lower error rates even in the presence of outliers highlighting its resilience. ARIMA slightly edges out SARIMA, particularly in terms of better handling outliers.
LSTM and RNN, despite low Mean Square Error (MSE) values exhibit highly negative R-squared values in all scenarios raising concerns about their fit to the dataset. The Prophet model exhibits steady performance, demonstrating some robustness against outliers. But Random Forest does the best in these kinds of predictions, especially because it handles outliers really well. Looking at all this it is super important to think about how accurate models are and how they deal with outliers when you are doing time series forecasting like for hospital admissions.[16]

CHAPTER 6

Conclusion and Future Scope

Various forecasting models were explored and applied to hospital admissions data, both with and without the presence of outliers. The models encompassed traditional statistical approaches like Moving Average and Exponential Smoothing, advanced machine learning techniques including Random Forest and Gradient Boosting, time-series specific models like ARIMA and SARIMA, neural network-based models such as LSTM and RNN, and the robust Facebook Prophet model.

Model Performance Insights:

- With Outliers: The Facebook Prophet, LSTM and Random Forest model exhibited superior performance in terms of lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) on the original dataset. This suggests LSTM's capability to handle complex patterns even in the presence of noise and outliers.

- Without Outliers: The Simple moving average, ARIMA, and Random Forest model showed improved accuracy, indicating its effectiveness in a more controlled data environment. The reduction in error metrics for the dataset without outliers emphasizes the impact of data quality on model performance.

6.1 Evaluation of Forecasting Models Against Research papers

In the model, the Moving Average and Exponential Smoothing performed well without outliers, aligning with the challenges of handling outliers as discussed in the UK study during the COVID-19 lockdowns. The Random Forest and Gradient Boosting models showed better performance without outliers and also somewhat with outliers, similar to the findings in the Madrid study, which highlighted the importance of **combining traditional time series methods with advanced machine learning techniques like CNNs and LSTMs.**[1]

Model ARIMA and SARIMA models also performed better without outliers, echoing the observations from the Lille Regional Hospital Centre study about the ARIMA model's limitations in adjusting to sudden changes. The LSTM and RNN models in the study, with their low RMSE, align with the Madrid study's success in using LSTM for admission predictions, showcasing the potential of neural networks in this domain.

The Facebook Prophet model's robustness to outliers in the study mirrors its design philosophy for adaptability, similar to the Chongqing study.

Overall, models showcase the effectiveness of various forecasting methods in different scenarios, emphasizing the importance of data preprocessing and the potential of combining multiple approaches for more accurate predictions, as supported by the multidimensional approach in the literature.

6.2 Future Scope

The methodologies and insights from this study can be extended to other areas of healthcare analytics, such as patient flow management, resource allocation, and epidemic outbreak prediction. To enhance the performance of predictive models, it is crucial to consider external factors, such as **climatic changes and epidemic patterns**, that significantly impact hospital admissions. Integrating these variables could lead to more nuanced predictions and, subsequently, more effective healthcare strategies.

From the results it is clear that **Random Forest and Prophet model might be an effective combination for future studies with hospital admission prediction.** As the Mean Absolute Error (MAE) of 58.531 with outliers and 52.545 without Random Forest reveals how it may stay accurate even if it has outliers. Given hospital admissions becoming erratic this is a crucial factor to keep into consideration. However, when outliers are considered, the Prophet model's mild rise in MAE from 59.104 to 60.480 shows considerable uniformity in handling data anomalies. The combined strategy encounters a chance to offer an enhanced and subtle prediction system through integrating Prophets inline effectiveness in collecting trends and seasonal variations in Random Forest's reliability that is illustrated through its somewhat lower MAE scores and the capacity to deal with complex data.

Also, the **Simple Moving Average model gives almost same results as Random Forest in terms of MAE.** But Random Forest is particularly adept at capturing complex non-linear relationships within the data which is a common characteristic in hospital admission data due to various influencing factors like seasonal trends healthcare policies and patient demographics. This capability is generally beyond the scope of a Simple Moving Average model which assumes a linear relationship and is primarily effective for data with consistent unvarying trends.

Hence, using these prediction tools in hospital systems can help make better decisions based on data use resources wisely and improve patient care. Future studies could look into **combining different models** to get better and more trustworthy results. Also, can implement new methods in machine learning and deep learning to find even better ways to forecast patient inflow in hospitals.

Bibliography

[1] Syed Ahmar Shah,Sinead Brophy,John Kennedy,Louis Fisher,Alex Walker, Brian Mackenna, Helen Curtis, Peter Inglesby,
Impact of first UK COVID-19 lockdown on hospital admissions: Interrupted time series study of 32 million people eClinical Medicine, vol. 49, 2022. https://doi.org/10.1016/j.eclinm.2022.101462

[2] Navares, R., Aznarte, J.L, *Deep learning architecture to predict daily hospital admissions* SpringerLink 32,2020. https://doi.org/10.1007/s00521-020-04840-8

[3] Kadri, F., Harrou, F., Chaabane, S. et al.
Time Series Modelling and Forecasting of Emergency Department Overcrowding. SpringerLink 38, 107,2014
https://doi.org/10.1007/s10916-014-0107-0

[4] Lingling Zhou, Ping Zhao, Dongdong Wu, Cheng Cheng and Hao Huang,
Time series model for forecasting the number of new admission inpatients BMC Medical Informatics and Decision Making, vol. 21,2018.
https://doi.org/10.1186/s12911-018-0616-8

[5] Zella King1, Joseph Farrington, Martin Utley, Enoch Kung, Samer Elkhodair, *Machine learning for real-time aggregated prediction of hospital admission for emergency patients* Nature Journal, 2022.
https://doi.org/10.1038/s41746-022-00649-y

[6] *Machine learning to effectively aid bed management in Kettering General Hospital* NHS AI Lab Skunkworks, 2021.
https://nhsx.github.io/skunkworks/bed-allocation

[7] Ivan Svetunkova, Fotios Petropoulosb, *Old dog, new tricks: a modelling view of simple moving averages* ResearchGate, 2017.
https://www.researchgate.net/publication/320130719_Old_dog_new_tricks_a_modelling_view_of_simple_moving_averages

[8] Chris Chatfield, Anne B. Koehler, J. Keith Ord and Ralph D. Snyder, *A New Look at Models for Exponential Smoothing* ResearchGate,Vol 50, 2001.
https://www.jstor.org/stable/2681090

[9] Breiman, L., *Random Forests. Machine Learning* SpringerLink,Vol 45, 2001.
https://doi.org/10.1023/A:1010933404324

[10] Jerome H. Friedman, *Greedy function approximation: A gradient boosting machine* ResearchGate, Vol 29, Oct 2001.
https://www.jstor.org/stable/2681090

[11] Harvey, A.C., *ARIMA Models*, SpringerLink, 1990.
https://doi.org/10.1007/978-1-349-20865-4_2

[12] Rumelhart D. E., Williams, R. J., *Learning representations by back-propagating errors. Nature*, Nature,323, 1986.
https://doi.org/10.1038/323533a0

[13] Taylor, S. J., Letham, B., *Forecasting at Scale*, The American Statistician,72, 2018.
https://doi.org/10.1080/00031305.2017.1380080

[14] James, Gareth, et al, *An introduction to statistical learning*, Springer,Vol 112, 2013.
https://ds.amu.edu.et/xmlui/bitstream/handle/123456789/14323/An%20Introduction%20to%20Statistical%20Learning%20-%20441%20pages.pdf?sequence=1&isAllowed=y

[15] Aurea Grane, Helena Veiga, *Wavelet-based detection of outliers in financial time series*, ScienceDirect,Vol 54, 2010.
https://doi.org/10.1016/j.csda.2009.12.010

[16] Nowrouz Kohzadi, Milton S. Boyd, Bahman Kermanshahi, Iebeling Kaastra, *A comparison of artificial neural network and time series models for forecasting commodity prices*, ScienceDirect,Vol 10, 1996.
https://doi.org/10.1016/0925-2312(95)00020-8

[17] RJ Hyndman, G Athanasopoulos, *Forecasting: principles and practice*, academia,Vol 10, 2018.
https://www.academia.edu/download/64659947/Athanasopoulos,%20George_%20Hyndman,%20Rob%20J.%20-%20Forecasting_%20Principles%20and%20Practice%20(2018).pdf

[18] Donald B. Percival, Andrew T. Walden, *Wavelet Methods for Time Series Analysis*, Technology and Medicine, London,2006.
http://staff.washington.edu/dbp/PDFFILES/5-Lund.pdf

[19] Menelaos Pavlou, Gareth Ambler, Rumana Z. Omar, Andrew T. Goodwin, Uday Trivedi, Peter Ludman, Mark de Belder , *Outlier identification and monitoring of institutional or clinician performance: an overview of statistical methods and application to national audit data*, BMC Health Services Research,2023.
http://staff.washington.edu/dbp/PDFFILES/5-Lund.pdf

[20] D Taylor, *Time-series analysis. Use of autocorrelation as an analytic strategy for describing patterns and change.*, SageJournals,1990.
10.1177/019394599001200210

[21] Colin, *Random Forests in Time Series Analysis*, GitHub,25 October 2023,
https://colinchpy.github.io/2023-10-25/09-08-05-789999-random-forests-in-time-series-analysis/

[22] Zhiyuan He, Danchen Lin, Thomas Lau, Mike Wu, *Gradient Boosting Machine: A Survey*, PointZeroOneTechnology,25 October 2023.
https://ar5iv.org/abs/1908.06951

[23] Davide Boldini, Francesca Grisoni, Daniel Kuhn, Lukas Friedrich, Stephan A Sieber, *Practical guidelines for the use of gradient boosting for molecular property prediction*, Journal of Cheminformatics, 2023.
https://pubmed.ncbi.nlm.nih.gov/37641120/

[24] Box, G.E.P. and G.M. Jenkins, *Time Series Analysis, Forecasting and Control (Holden-Day)*, GoogleScholar, 1970.
https://books.google.com/books?hl=en&lr=&id=rNt5CgAAQBAJ&oi=fnd&pg=PR7&ots=DL1-zUoWTF&sig=H4P_h0quX6whueRuvzwy7O6pQAs

[25] *How to Grid Search SARIMA Hyperparameters for Time Series Forecasting.*, Machine Learning Mastery,
https://machinelearningmastery.com/how-to-grid-search-sarima-model-hyperparameters-for-time-series-forecasting-in-python/

[26] Rizwan Mushtaq, *Augmented Dickey Fuller Test*, SSRN,
https://www.zbw.eu/econis-archiv/bitstream/11159/126879/1/EBP084690321_0.pdf

[27] Kasun Bandara, Rob J Hyndman, and Christoph Bergmeir, *MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns*,2021 Cornell University,
https://doi.org/10.48550/arXiv.2107.13462

[28] *Skunkworks bed allocation Fake Data Generation*, NHSx,
https://github.com/nhsx/skunkworks-bed-allocation/blob/main/fake_data_generation/README.md

[29] Moore, D. S., *Introduction to the Practice of Statistics*,2009 WH Freeman and company,
https://ds.amu.edu.et/xmlui/bitstream/handle/123456789/14700/Introduction%20to%20the%20Practice%20of%20Statistics%20%20-%201010%20pages.pdf?sequence=1&isAllowed=y