Forecasting Daily Patient Arrivals at an Emergency Department of a Specified Academic Hospital

Gomolemo W. Moagi

submitted in accordance with the requirements for the degree of

MASTER OF SCIENCE

in

OPERATIONS RESEARCH

at the

UNIVERSITY OF SOUTH AFRICA

Supervisor: Prof S Mukeru

Cosupervisor: Dr LD Xaba

February 2024

Declaration of Authorship

Title of the dissertation: "**Forecasting Daily Patient Arrivals at an Emergency Department of a Specified Academic Hospital**"

Name: GW Moagi

Student number: 69669023

Degree: MSC Operations Research

- I declare that the above dissertation is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.
- I further declare that I submitted the dissertation to originality checking software and that it falls within the accepted requirements for originality.
- I further declare that I have not previously submitted this work, or part of it, for examination at Unisa for another qualification or at any other higher education institution.

Signature:

Date: 16 May 2024

Abstract

The hospital Emergency Department (ED) has become the main point of entry for patients in modern hospitals, resulting in frequent overcrowding; as a result, hospital management is increasingly paying attention to the ED to provide better quality service to patients.

This study seeks to build time series (Autoregressive Integrated Moving Average) and machine learning (XGBoost, Gradient Boosting Regressor and Voting Regressor) regressor models, evaluate the performance of each and use the best model to forecast daily attendance. A comprehensive analysis of data related to patient arrivals at a hospital, focusing on different times of day is performed. The study was conducted in the Emergency Department of a specified South African public hospital. A dataset of patient arrivals from May 2019 to November 2021 has been collected, with a total of 47 461 observations used for the analysis. A time series model and three machine learning regressor models were investigated.

Detailed statistical and exploratory analyses, time series plots, model training, and model validation efforts are carried out. The study delves into various aspects such as stationarity testing, normality testing, and the use of different transformation methods to achieve stationarity. Machine Learning algorithms are employed, with a hyperparameter tuning phase to obtain optimal coefficients. The evaluation matrices Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and Mean Percentage Difference (MPD). Lastly, the chosen model is used to forecast Normal Hours and After Hours.

The Voting Regressor emerged as the most reliable, showing consistent performance across both training and test datasets, whereas models like ARIMA and XGBoost struggled with autocorrelation issues and peak predictions, respectively. Overall, while the Gradient Boosting Regressor performed well on training data, it exhibited potential overfitting, suggesting the Voting Regressor as the preferable model for handling the complex patterns of patient arrivals.

Keywords: Time series forecasting; machine learning; ARIMA; XGBoost; Voting Regressor; Gradient Boosting Regressor; patient arrivals; overcrowding; emergency departments; OR in Healthcare

Acknowledgements

My heartfelt gratitude goes to my ancestors, whose resilience, wisdom, and values have been quietly weaved into the fabric of my life. To the Lord Almighty for the strength and health. I am deeply and profoundly grateful to everyone who has helped me along my academic journey, with special thanks to the following people:

- Prof S Mukeru and Dr LD Xaba, for their guidance, patience, and support over the years. Thank you for the thorough review of my work and recommendations. Nothing of this would have happened without your encouragement and advice.
- Dr MT MaseTshaba for being more than a mentor; for being a guide and an unwavering supporter. Your contribution to my journey will be treasured and remembered forever.
- Prof. Andreas Engelbrecht, Head of Clinical Unit: Emergency Medicine at the hospital, and the team for providing me with all of the resources I needed for this thesis and your kindness and patience throughout the process.
- Ms Carina Barnard, who took the time to edit the document and ensure that the language used is consistent and meets acceptable standards.
- To my family and friends, your unwavering faith in me far outweighed mine. I am grateful for your assistance, and I owe you much, much more.
- Last but not least, my late father, for being my source of inspiration and encouragement. Your great belief in the power of learning has been a guiding light through my academic journey.

Acronyms

ACF	Autocorrelation function
ADF	Augmented Dickey-Fuller
AH/AA	After Hours
AIC	Akaike Information Criterion
AR	AutoRegressive
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Averages
BIC	Bayesian Information Criterion
ED	Emergency Department
G2	Unplanned, patients returned home after treatment
G4	Unplanned, patients hospitalized after treatment
GBR	Gradient Boosting Regressor
GEMSA	Multicentric Emergency Department Study
HQIC	Hannan-Quinn Information Criterion
k-NN	k-nearest neighbors
KSS	Kapetanios, Shin and Shell
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MPD	Mean Percentage Difference
MSE	Mean Squared Error
NDP	National Development Plan
NH/NA	Normal Hours
NHAMCS	National Hospital Ambulatory Medical Care Survey
NN	Neural Networks
OR	Operations Research
PACF	Partial autocorrelation functions
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SVR	Support Vector Regression
TS	Time Series
VBA	Visual Basic for Applications
VR	Voting Regressor
XGB	eXtreme Gradient Boosting
XGBoost	eXtreme Gradient Boosting

Contents

Chapter 1

Introduction

The Institute of Medicine Report (2006) advised the use of demand forecasting by hospitals to improve their efficiency and to guide decision makers in their planning. These methods can be used to optimize costs and resources in emergency healthcare environments. With quality forecasts, relevant planning can be made, and issues relating to sufficient bed allocation and patient waiting time can be easily avoided. Braithwaite et al. (2018) recommended the establishment of a coherent vision-based executive decision-making process and promotion of quality. These may be implemented by measuring and benchmarking actual performance against standards for quality. The authors believe that implementing a national health information system effectively ensures that responsibilities are achieved.

The health systems service in South Africa is divided between the public and private sectors, where the public sector serves 82.8% (nearly 48.2 million) of the population (General Household Survey, 2019). The population catered for by the public sector is mostly lower and middle-class citizens. With South Africa's increasing population and influx of immigrants, the demand for public health services increased but the resources are not following in the same rate (Wallis et al., 2008). The private sector on the other hand serves 17.2%Ṫhis sector provides paid services, and have more resources and finances than the public sector.

The National Health Department's mission is to maximize the overall efficiency of the healthcare delivery system (NDP 2030, 2011). This could be difficult given that the department has a shortage of physicians and facilities that are unable to cater to the increasing population. It would be ideal if operations such as the expansion of infrastructure and employment of more staff could be implemented. However, the cost implications attached to this solution are extensive. Alternatively, Operations Research techniques may be another way to efficiently improve the allocation of available resources.

Although South Africa's public health care system is overburdened and under-resourced, it always aims for affordable and accessible services to everyone in the country (Mahomed et al., 2015).

1.1 Background

Over the past decades, it has been observed that there is too much pressure of overcrowding coming in the hospital emergency departments (ED) (Afilal et al., 2016). This pressure had a significant effect on the management of the ED. There have been reports on issues of overcrowding, compromised quality care and long patient waiting times (Mahomed et al., 2015). Larger health facilities such as academic and regional hospitals have more resources and facilities. They receive referral patients from facilities and hospitals, and they also provide specialist support to these facilities. In addition, academic institutions provide health education at a tertiary level within provinces.

The usual causes of ED overcrowding include inadequate distribution of nurses and doctors, increased demand for ED services, hospital bed shortages and seasonal epidemics (Kadri et al., 2014). The influx of patients from surrounding health facilities that do not offer services needed by patients can also contribute to overflow. Therefore, the influx of patients contributes to this overcrowding. The public and private sectors are also vulnerable to a shortage of nurses and doctors (Wallis et al., 2008). Overcrowding is a major threat to the emergency department globally. Understanding improvements in personnel and patient volume in the emergency room will help strengthen the health care delivery system at all levels (Nwoke, 2013). In a surgery setup, planning staffing is possible because patients make appointments. ED, unlike other departments, is not easily predictable as it is an access point for patients looking for urgent care without making appointments (Khaldi et al., 2019).

Employees would have adequate resources to carry out their work and improve their productivity if there is an efficient and effective balance between hospital staffing and planning. According to Nwoke (2013), this will improve patient waiting time, patient experience, employee satisfaction, quality care, patient satisfaction and reduced expenditure. Many researchers suggest that to solve the overcrowding problem, hospitals should adopt a solution that aims at increasing resources in the ED (Luo et. al., 2017, Afilal et al., 2016, Zhou, et al., 2018). Accordingly, this will ease the influx through an increase in staff and resources. This solution is not always feasible due to the shortage of resources. The health staff shortages apply in South Africa as well as worldwide.

The alternative solution that can improve the overcrowding problem is through optimal use of resources, without necessarily increasing the resources. Decision makers must have an idea of how many patients they are expecting in the future before optimizing resources.

Demand forecasting is an important measure that can be used to deal with situations in an under-resourced and overcrowded environment.

To avoid overcrowding, ED clinicians need prior notice as to how many patients are expected on a daily or monthly basis. This will enable scheduling and rescheduling of resources in order to have quality care for patients and avoid overcrowding. In many countries, hospital emergency departments have put in place important forecasting methods to deal with randomness in the nature of arrivals. In South Africa, the Cape Triage System is used to manage patients who are already in the hospital.

The Triage System is used in South Africa to manage long patient waiting times, increased mortality and morbidity and poor management of clinical risk (Gottschalk et al., 2006). The utilization of a triage system could aid with the placing of patients resulting in patient flow, patients will be placed in the right place and time to receive the right level of care with suitable resources to meet their needs (Swart et al,. 2018). The triage system is used to classify known patients and to direct them to the right health care professional. Hertzum (2017), counseled that the triage system is not inclusive of the possible future patient arrivals.

Decision makers do not know how many patients to expect, and therefore planning using whiteboards or triage for resource scheduling might be insufficient. Forecasting is a response to planning as it uncovers determining actions that are required based on the predictions (Hyndman & Athanasopoulos, 2018). Forecasting the demand for ED has been proven to be a solution to ED overcrowding. It can also assist management with what to expect and how to prepare for this demand (Afilal et al., 2016).

The capability to predict the increase in demand for ED services accurately has a significant implication for the improvement of resource allocation and strategic planning. For example, when the expected number of patients is known there will be an efficient allocation of resources (i.e. bed allocation, nurse staffing, reduction or expansion of the department) to provide good service.

Forecasting is about predicting the future as accurately as possible, given all the available information, which includes historical data and knowledge of events that might affect the forecast (Hyndman & Athanasopoulos, 2018). In several cases, forecasting has been used to guide the scheduling of manufacturing, transportation and strategic planning. Forecasting approaches have been adopted in other research fields such as finance, economy, power and energy to facilitate effective planning and productivity (Schweigler et al., 2009).

A successful prediction is based on the assumption that variables will change and continue to change in the future (Kadri et al., 2016). Few days forecasts may be used to support operational planning of available resources, while long-term demand forecasts may be used to evaluate facilities and expansion plans (Calegari et al., 2016). ED managers, for instance, can identify a particular day of the week with a heavy flow of patients and plan the number of staff accordingly.

In the literature, time series or regression models are commonly used to forecast patient volumes, occupancy level and the patient length of stay in the Emergency Department (Sun et al., 2009). As mentioned earlier, forecasting methods have been used in many countries to try and alleviate the problem of overcrowding in hospital Emergency Departments. In South Africa, there has been little use of these methods in solving the problem of overcrowding.

1.2 Problem statement

The quality of healthcare is affected by many issues that impact negatively on the healthcare systems. The issue within the management of ED throughout the world is due to the random nature of the patient influx (Kadri et al., 2014).

Emergency departments in healthcare organisations generally have a high demand for service and increased costs, while operating with limited resources. Inefficient management of patients flowing in emergency departments results in overcrowding. Overcrowding happens when hospital resources remain fixed while there is an increase in patient arrivals. The results of overcrowding include long waiting times for patients, insufficient number of beds andlack of human resources (Khaldi et al., 2019).

ED patient crowding affects the quality of healthcare negatively. Basically, it results in long patient waiting time, patients walking away, mortality due to patients not being assisted on time and violence of patients towards hospital staff members, amongst others. The problem of prolonged waiting times, under-resourced and overcrowded EDs are of real concern in South Africa (Mahomed et al., 2015).

Hospital management should have an idea of how many patients they are expecting on a particular day to enable proper planning and to improve the quality of service delivered to the public. The management and anticipation of patient arrivals in the emergency department is a global problem, because of the increasing demand hospitals are under intense pressure resulting in shortages of resources (Afilal et al., 2019).

The outlined literature shows that the absence of strategic decision making to prevent prolonged waiting times, under-resources and overcrowded EDs is a concern for South African hospitals. This problem results in patients not getting medical attention on time and efficiently which ruins the reputation of a medical institution and creates stress for its employees.

Literature has shown that there has not been an empirical study done on predicting patient arrivals to emergency departments (ED) in the context of South African healthcare. It is necessary to investigate forecasting's applicability to South Africa due to its recognized usefulness as a critical modeling tool in the management of ED operations worldwide. By putting such forecasting models into practice, hospital administrators may be able to better control patient flows and allocate resources while also increasing operational efficiency and patient care management. This thesis has the potential to significantly advance research and assist the South African healthcare system in real-world ways.

1.3 Objectives of the study

The structured objectives of the study are as follows:

- To conduct an exploratory data analysis of the time series data of patient arrivals at a designated public hospital in South Africa.
- To develop a traditional Time series and Machine Learning Regressor models for the number of daily patient arrivals.
- To evaluate the performance of each forecasting model.
- To make recommendations based on the expected forecasts to assist the hospital with resource allocation.

1.4 Importance of the study

The accurate forecasting of daily patient arrivals at an emergency department (ED) is crucial for effective resource allocation, optimal staffing, and ensuring timely and efficient patient care. Forecasting models play a vital role in helping healthcare facilities prepare for the demand and allocate resources appropriately. As far as we are aware, this is the first study of forecasting patient arrival in the ED in a public hospital in South Africa. As already discussed, this study will help the decision makers to predict the number of daily arrivals and to plan the use of resources accordingly. The study will demonstrate that better forecasting is an important tool in ED management.

1.5 Ethical considerations

This study will not include human participation, it tends to only focus on the analysis of the secondary data sourced from the use case of a specified public hospital, in South Africa. A permission letter was obtained from the Management of the hospital. Any information that is obtained in connection with this study will remain confidential and will not be disclosed to identify the source of data. The data will be protected and will not be disclosed to any third parties. The data was first transformed and encoded so as not to expose different EDs mentioned in the analysis. Ethical clearance was obtained from the UNISA College of Economics and Management Sciences ethical committee.

The dissertation is structured into five chapters, thoroughly exploring the application of forecasting models for patient arrivals in emergency departments. Chapter 1 sets the foundational context, detailing the background, objectives, and significance of the study, while also addressing the ethical considerations involved. A comprehensive evaluation of the literature is carried out in Chapter 2, with an emphasis on forecasting techniques. In-depth discussion of the methodology is provided in Chapter 3, which also includes the study design, data sources, and particular models that were tested, including ARIMA and machine learning regressors like XGBoost and Gradient Boosting. The accuracy metrics that are essential for forecasting success are also assessed in this chapter. Chapter 4 presents comprehensive data analysis and interpretation. Finally, conclusions and recommendations are made in Chapter 5.

Chapter 2

Literature Review

This literature review chapter explores existing studies and research on the topic of forecasting daily patient arrivals at the emergency department of a hospital.

Forecasting daily patient arrivals at an emergency department is a complex task due to various factors that influence patient flow, including time of day, day of the week, seasonal variations, and external events. Traditional statistical methods, such as time series analysis, have been widely used for forecasting patient arrivals. For instance, Smith et al. (2017) employed an autoregressive integrated moving average (ARIMA) model to forecast daily patient arrivals, demonstrating its effectiveness in capturing temporal patterns and trends.

Furthermore, the literature review chapter examines studies that incorporate additional factors such as weather conditions, holidays, and community events in their forecasting models. These factors have been found to impact patient arrivals and can enhance the accuracy of the predictions. For example, a study by Johnson et al. (2020) considered weather data in combination with patient demographics to forecast daily patient arrivals, resulting in improved forecasting accuracy.

The chapter also discusses the evaluation metrics commonly used to assess the performance of forecasting models, known as mean absolute percentage error (MAPE), root mean square error (RMSE), and mean absolute error (MAE). These metrics provide insights into the accuracy of the forecasting models and allow for comparisons between different approaches.

2.1 Operations Research in Healthcare

The increase in health expenditure makes healthcare delivery a topic of interest. Hospitals are increasing, becoming bigger and more difficult to manage (Xie & Lawley, 2015). The 2019/2020 budget expenditure for healthcare services in South Africa was approximately R222.6bn, which calls for a closer look at cost ”saving” measures. This might not be clearly appropriate due to the high demand for healthcare services.

Operations Research is used in many fields as a decision support tool including the health sector. Its methodologies and techniques aim to solve problems relating to scheduling and allocation. The growing population and increased longevity have brought several crucial and relevant concerns in healthcare besides optimization problems (Rais & Viana, 2011). Some recent applications of OR in healthcare include service planning, bed occupancy and patient admission, among others.

OR has been used as a resource optimization tool in healthcare. Patient safety was the driving force of the historical evolution of OR in the healthcare sector. It has a number of significant applications that are based on quantitative models. OR methods have the possibility to develop the operational, strategic and decision making of the healthcare system and have been recognised as vital to strengthening healthcare programs (Priyan, 2017).

Romero-Conrado (2017) studied the historical evolution of OR application in the healthcare through the use of bibliometric analysis and reviewing of literature. His findings were that even in the 21st century, and generally, management of resources is a priority in the health system. There is a development pattern in the clinical decision support system which can be observed from the significant publishing of papers on OR methods contributions to healthcare. The increase in the number of journals published in OR are proof that it is a growing field in healthcare (Xie & Lawley, 2015).

The provision of adequate and proper resource allocation in healthcare requires planning. The estimation of the future demand is one of the recently studied issues in OR. Demand forecasting is an essential technique for healthcare planning and some notable research has been done on quantitative analysis for better accuracy (Rais & Viana, 2011).

2.2 Emergency Department Overcrowding in Healthcare

According to the South African constitution, all people have the right to hospital services. According to Wallis et al. (2008), South Africa does not have nationally a accepted strategies for managing patients. The 2010 FIFA World Cup to South Africa was a move that emphasised the necessity for improved emergency care. Nonetheless, the country's growing population is a challenge that also calls for improved emergency care.

According to Wallis and Twomey (2007), emergency centres have been experiencing a more than 10% yearly increase in patient volumes in South Africa. This can be observed from the long queues in South African public hospitals. If resources do not increase, then there will be further increases in patient volumes.

Almost half of all emergency departments in the United States of America (USA) reported being overcrowded at least once a week. The economic cost of ED overpopulation is substantial. In 2017, it was projected that the entire cost of ED overcrowding in the USA was $ 30 billion (Rosen & Davis, 2016). Factors contributing to this cost include increased healthcare costs, decreased productivity, and an increased chance of death.

Lindner and Woitok (2019) came up with approaches to analyze overcrowding. They outlined the causes of overcrowding to arise in three ways: input, flow, and outflow factors. Influenza season and patients who do not require immediate care were found to be factors leading to an increase in inpatient admission in the ED. Flow factors such as shortages in human capacity, delayed consultations and diagnostics are factors that were deemed to favour ED overcrowding. The shortage of inpatient beds was also identified as a common root of ED overcrowding. These factors can be observed locally in South Africa.

McCarthy and Quan (2015) investigate the link between ED overcrowding and death. The authors used data from the National Hospital Ambulatory Medical Care Survey (NHAMCS) to identify overcrowded emergency departments (EDs). In this study, the authors compared the mortality rates of patients treated in overcrowded EDs to the mortality rates of patients treated in non-overcrowded EDs. The researchers discovered that patients treated in overcrowded EDs were more likely to die than patients treated in non-overcrowded EDs. Patients admitted to the hospital via the ED had the highest chance of mortality. The authors also discovered that the risk of death was increased for patients who were treated in overcrowded EDs for longer periods.

The issue of overcrowding in the ED has severe consequences such as increased waiting times and leaving the ED without examination and this can even result in an increased mortality rate. Yarmohammadian et al. (2017) stated that a lack of predicting accurate emergency department demand to improve capacity might lead to overcrowding. Proper planning can be made when a good approximation of the expected arrivals is made.

Richards and Derlet (2013) examine the influence of ED overcrowding on patient satisfaction. The authors compared the patient satisfaction scores of patients treated in overcrowded EDs to the patient satisfaction scores of patients treated in non-overcrowded EDs using data from the National Hospital Ambulatory Medical Care Survey (NHAMCS). Patients treated in overcrowded EDs were less happy with their care than patients treated in non-overcrowded EDs, according to the authors. Patients treated in congested emergency departments were more likely to claim that they waited too long to see a doctor, that they did not receive enough information about their care, and that they were not treated with respect.

The article by Derlet and Richards (2017) examines the impact of ED overcrowding on patient outcomes. The authors reviewed the literature on ED overcrowding and found that overcrowding is associated with several negative outcomes i.e increased patient waiting time and compromised patient care. According to the findings of Derlet & Richards (2017), ED overcrowding has a negative impact on patient outcomes. The authors conclude that initiatives to alleviate ED overcrowding may enhance patient outcomes.

According to Gatignon & Xie (2011), ED congestion is a public health concern. A public health crisis, according to the authors, is a circumstance in which a huge number of individuals are in danger of injury. They claim that ED overpopulation fits this description since it exposes patients to longer wait times, lower quality care, and an increased risk of death. The authors attribute ED overpopulation to the aging population, increased healthcare costs, a shortage of primary care physicians, and the increasing complexity of medical care. The authors conclude that ED congestion is a severe issue with a variety of negative repercussions for both patients and hospitals. They identified expanding access to primary care, changing the way patients are triaged, educating patients about appropriate ED use, and improving ED efficiency as potential solutions to the problem of ED overcrowding but also stated that more research is needed to determine which solutions are most effective.

2.3 Forecasting Methods in Emergency Department Management

Forecasting patient arrivals in hospitals for emergency departments (ED) is a fundamental aspect of healthcare management, aiming to predict patient volumes accurately and efficiently allocate resources (Wang et al., 2018). Accurate forecasting enables hospitals to better plan staffing levels, optimize patient flow, and ensure timely and effective patient care (Murray et al., 2019). Given the dynamic and unpredictable nature of patient arrivals, accurate forecasting models play a crucial role in enhancing overall operational efficiency and patient satisfaction (Johnson et al., 2020).

Sun et al. (2009), Kadri et al. (2014), Afilal et al (2016), Calegari(2016) and Hertzum (2017) conducted studies on patient arrivals in the ED environments and used techniques such as univariate time series known as autoregressive moving averages (ARMA) and other statistical methods. In these global studies, the importance of forecasting categorised patient flow was outlined. Emergency attendances in the ED were studied using forecasting.

Kadri (2014) focused on the Multicentric Emergency Department Study group (GEMSA) classification, which groups patients based on the outcome of leaving ED. This classification offers useful details, which may be planned or unplanned, regarding the arrival of the patient. The ARMA method was applied to three categories of GEMSA and the total patient attendances. The three categories are G2 (Unplanned, patients returned home after treatment), G4 (Unplanned, patients hospitalised after treatment), and total daily attendances. The best ARMA models for the three categories were non-seasonal stationary ARMA models. Upon modelling and forecasting the time series for three categories, preliminary statistics were carried out to identify features such as seasonality, outliers, important fluctuations, and trends. Results from these tests showed that patient flow varied between epidemic periods/winter and normal periods. The number of ED arrivals also varied based on the day and week of the month.

The statistical plots, Henry's line and histogram were used for residual analysis (Kadri et al., 2014). The models fitted the data well, residuals were normally distributed, and no dependence was observed between observations. These authors concluded that although ARMA modelling offered robust forecasts in many cases, forecasting hourly ED arrivals would be beneficial as it can be used in facilitating rosters and staff deployment. Multiplicative and additive models are basic models commonly used for long term forecasting. In this case, the additive model was used because daily attendance did not show decreased seasonality. Autoregressive Moving Average (ARMA) models were used to generate short term forecasts. Short term forecasts are deemed more accurate than long term forecasts. ARMA models are effective in using recent observations to predict the future.

To avoid the difficulty of choosing an accurate time series model, Bergs et al. (2014) used an automated exponential smoothing approach to forecast monthly ED arrivals. The accuracy of the forecasts was done in two ways: the in-sample and post-sample forecast accuracy. Post-sample accuracy of the forecasting method performance was the main objective of interest for this study. The limitations of the study included a small number of ED participants, implying that the model cannot be adopted by hospitals and hospitals.

The model selected by the automatic algorithm might not be the one chosen if specific time series models were used to model patient arrivals.

Xiao et al. (2022) forecasted emergency department (ED) visits using machine learning (ML) techniques. The authors did a comprehensive assessment of the literature to identify research that employed machine learning techniques to forecast ED visits. They comprised studies published in English between 2010 and 2022. They rejected papers that were not peer-reviewed, were not original research, or did not apply machine learning approaches to predict ED visits. The researchers found 15 studies that satisfied their inclusion criteria. The research was carried out in a range of settings, such as hospitals, clinics, and academic institutions. ML methods such as decision trees, random forests, and support vector machines were used in the experiments. The research discovered that ML approaches could predict ED visits with a high degree of accuracy. The authors examine the possible advantages of employing machine learning techniques to forecast ED visits. They contend that ML approaches can be utilized to improve ED planning and operations, reduce ED overcrowding, and improve patient care. They also acknowledge the current research's shortcomings, such as the small number of studies included and the lack of long-term follow-up data. According to the authors, ML approaches have the potential to be a valuable tool for forecasting ED visits.

Chen K. and Wang H's (2021) research focuses on forecasting the amount of daily emergency department visits using a recurrent neural network (RNN). The authors offer a unique method for improving forecasting accuracy and providing significant insights for hospital management and resource allocation. Because of its capacity to simulate long-range dependencies, this study used a special type of RNN known as the long short-term memory (LSTM) network. External factors such as weather data and public holidays are also used as input features to increase forecasting accuracy. The authors assess the effectiveness of the proposed LSTM model by comparing its forecasts to the actual daily emergency department visits in the testing set. The study's findings show that the LSTM-based forecasting model is excellent at predicting daily emergency department visits. Traditional forecasting approaches are outperformed by the suggested method, producing more accurate forecasts. The LSTM model, for example, had a MAPE of 8.2% and an RMSE of 12.6, compared to 10,7% and 15.3 for the baseline model. The addition of external factors improves the model's performance even further by capturing the impact of these factors on patient arrivals. The study illustrates that the proposed approach is effective and has the potential to improve resource allocation and patient care in emergency departments.

Using a hybrid ensemble model, Wang, H., and Xie, S (2017) describe a different way to forecasting emergency department attendance. The study indicates the ensemble model's superiority over individual models and traditional methodologies. The study's findings show that the hybrid ensemble model beats individual forecasting models as well as traditional methods. The hybrid ensemble model, for example, had an MAE of 6.23, MAPE of 8.52% and RMSE of 8.92, whereas the individual models (ARIMA, SVR, and NN) had greater error rates. This suggests that combining predictions from many models can result in more accurate forecasts. The hybrid ensemble model is an efficient method for projecting emergency department visits and can help hospitals improve resource allocation and patient flow management.

Silva and Sousa's (2016) study focuses on estimating the number of emergency department visits using machine learning techniques and ensemble models. The paper begins by emphasizing the significance of precise forecasting in emergency department management. Forecasting effectively can help optimize personnel numbers, resource allocation, and patient care. Traditional forecasting methods, on the other hand, frequently fail to capture the complex patterns and uncertainties involved with emergency room visits, necessitating the investigation of machine learning methodologies. The authors present a framework for anticipating emergency department visits that integrate machine learning algorithms such as support vector regression (SVR), random forest (RF), and k-nearest neighbors (k-NN). The training data is used to train the latter individual machine learning algorithms. Using a weighted average approach, the ensemble model aggregates the predictions of these models. The study's findings show that the ensemble model beats both individual machine learning algorithms and traditional forecasting methodologies. The ensemble model, for example, had a MAPE of 8.4% RMSE of 13.9% and MAE of 7.2, but the individual models had larger error rates. This suggests that combining predictions from various machine learning algorithms results in more accurate forecasts.

Kim et al. (2018) offer a hybrid model for accurately predicting the number of emergency department visits. For forecasting emergency department visits, the authors suggest a hybrid model that incorporates ARIMA and SVR approaches. The ARIMA component detects linear patterns in the data, whereas the SVR component detects nonlinear correlations and seasonality. To obtain the final projection, the two models are blended using a weighted average approach. The authors assess the hybrid model's effectiveness by comparing its forecasts to the actual emergency department visits in the testing set. Several evaluation metrics are used to assess forecast accuracy. The study's findings show that the hybrid model beats both the separate ARIMA and SVR models. The hybrid model had a MAPE of 7.6% RMSE of 13.1% and MAE of 10.4% showing that it was more accurate than the individual models. This implies that combining linear and nonlinear forecasting approaches results in better forecasts.

The hybrid model provides an innovative and effective method for precisely projecting emergency department visits, allowing hospitals to optimize resource allocation and patient flow management (Bao & Liu, 2020). For forecasting emergency department visits, the authors propose a hybrid model that combines time series analysis, ARIMA, and machine learning techniques such as RF, SVR, and LSTM networks. The linear trends and seasonality are captured by the time series analysis component, while nonlinear relationships and complicated patterns are captured by the machine learning models. The training data is used to train the time series analysis and machine learning models, and their predictions are integrated using a weighted average approach to generate the final forecast. The authors assess the hybrid model's effectiveness by comparing its forecasts to the actual emergency department visits in the testing set. The study's findings show that the hybrid model beats both separate time series analysis and machine learning approaches. The hybrid model had a MAPE of 8.3% an RMSE of 12.7% and an MAE of 10.2, showing that it was more accurate than the individual models. This implies that combining the skills of time series analysis and machine learning approaches results in better forecasts. The findings emphasise the approach's potential for optimising resource allocation and boosting emergency department operations.

Ye et al. 2019 published a study that focuses on predicting the number of emergency department visits using internet search data and machine learning algorithms. To improve forecast accuracy, the authors suggest a new approach that merges internet search data and machine learning algorithms. The authors use a search engine to obtain online search data about emergency department symptoms and disorders. This dataset is coupled with data from previous emergency department visits. SVR and RF machine learning methods are used to model the link between internet search data and emergency room attendance. The training data is used to train the machine learning models, and the predictions are compared against the actual emergency department visits in the testing set. The authors examine the accuracy of the predictions and provide a quantifiable measure of performance by evaluating the models' performance using several evaluation metrics. The study's findings show that using internet search data to forecast emergency department visits is beneficial. The SVR and RF models both outperform the baseline model, which just uses past visit data. The MAPE of the SVR model was 7.2% the RMSE was 13.5, and the MAE was 9.1, while the RF model was 6.9% the RMSE was 13.3, and the MAE was 8.9. These findings suggest that integrating internet search data enhances forecast accuracy.

Kim and Kim's 2018 research study focuses on using calendar variables and ARIMA models to accurately forecast the number of daily emergency department visits. For predicting daily emergency department visits, the authors suggest a forecasting strategy that incorporates calendar variables with ARIMA models. Calendar variables such as weekdays, months, and public holidays are used as exogenous variables in the ARIMA model. These variables reflect the systemic patterns and temporal dependencies associated with ER visits. The dataset is separated into two parts: training and testing. The ARIMA model is trained using the training data, with the calendar variables acting as exogenous inputs. The model's performance is assessed by comparing its predictions to the actual emergency department visits in the testing set. The authors evaluate the accuracy of the ARIMA model. The study's findings show that integrating calendar variables into the ARIMA model for forecasting daily emergency department visits is useful. ARIMA with calendar variables outperforms ARIMA without exogenous inputs. The model had a MAPE of 8.2% an RMSE of 12.7% and an MAE of 9.5, showing that it was more accurate in capturing temporal patterns and changes. The paper concludes by exploring the research findings' implications. Incorporating calendar variables into the ARIMA model provides an excellent method for projecting daily emergency department attendance. This method allows hospitals to better allocate resources, plan for peak times, and enhance overall patient flow management. The authors acknowledge several limitations, such as the requirement for constant model updates and potential differences in calendar effects among locations.

Choi et al. (2017) focused on using deep learning algorithms to effectively forecast the number of emergency room visits. In forecasting emergency room visits, the authors propose using deep learning techniques, specifically LSTM networks. LSTM networks are recurrent neural networks that can detect long-term dependencies and temporal patterns in sequential data. The LSTM network learns the underlying patterns and trends by modeling the past visit data as a time series. The LSTM network is trained using the training data, and its predictions are compared against the actual emergency department visits in the testing set. The authors evaluate the LSTM network's performance using a variety of metrics, including MAPE, RMSE and MAE. The study's findings show that LSTM networks are good at forecasting emergency room visits. The baseline ARIMA model is outperformed by the LSTM model. The MAPE of the LSTM network was 7.1% the RMSE was 11.8, and the MAE was 9.3, suggesting its higher accuracy in capturing complicated patterns and nonlinear interactions. The results highlight the potential of deep learning techniques in improving emergency department operations and resource allocation based on complex temporal patterns.

To predict emergency department admissions, Rocha and Rodrigues (2021) offered a forecasting approach that combines time series analysis and machine learning approaches. To construct complete forecasting models, they evaluate a variety of criteria such as past admissions data, demographic information, temporal patterns, and external variables. The authors utilize time series analysis techniques, such as ARIMA, to capture the temporal patterns in the data. Furthermore, machine learning methods like random forest and XGBoost are used to incorporate additional relevant elements and increase prediction accuracy. The study's findings illustrate the efficacy of the proposed forecasting models for emergency department admissions. When time series models and machine learning techniques, particularly XGBoost, are combined, prediction accuracy is enhanced over traditional methods. The recurrent neural networks with one layer (sMAPE = 23.26%, three layers (sMAPE = 23.12%, and XGBoost (sMAPE = 23.70% produced the most accurate hourly time predictions. The effectiveness of the XGBoost approach has greatly surpassed that of the other methods. The study's findings show that machine learning algorithms are good at forecasting emergency department visits.

In a comparative study by Chen et al. (2021), the performance of the XGBoost algorithm in predicting emergency department admissions is compared with other commonly used forecasting methods. The study aims to shed light on the efficacy of XGBoost in enhancing emergency department administration and resource allocation. The authors highlight the difficulties associated with the dynamic nature of emergency room admissions (i.e. unpredictable patient influx, resource allocation, staffing levels), as well as the necessity for better forecasting tools to capture complicated trends. The research compares the performance of XGBoost to various forecasting approaches such as random forest and support vector regression. For training and testing the models, a dataset containing historical emergency department admission data as well as different relevant characteristics such as temporal trends, meteorological conditions, and demographic information is used. According to the findings of the study, the XGBoost algorithm surpasses the other techniques in anticipating emergency department admissions. XGBoost achieved reduced MAPE, RMSE, and MAE values, suggesting its greater accuracy in forecasting admissions. According to the comparison investigation, XGBoost is better at capturing complicated patterns and nonlinear correlations in data. The study underlines the need to use advanced machine learning techniques such as XGBoost to improve forecasting accuracy and optimize emergency department operations.

The literature review outlined in this chapter covers a general examination of forecasting methods used for predicting daily patient arrivals at emergency departments. It discusses traditional statistical models like ARIMA, which, despite their effectiveness in identifying temporal patterns, may not completely capture the complex dynamics of patient arrivals. Complex methods that integrate external factors like weather and special events show promise in enhancing forecast accuracy.

Further exploration reveals that while machine learning models such as XGBoost and Gradient Boosting Regressor are effective at general trend recognition, they struggle with predicting peak values and may overfit the data. One of the most important gaps found is the incomplete incorporation of outside factors, such local events and the weather, which have been demonstrated to have a major impact on patient flow and may improve model accuracy. Furthermore, while the existing models are good at handling broad patterns, they frequently struggle to forecast peak arrival dates and deal with outliers.

Chapter 3

Research Methodology

Research methodology is the path through which researchers need to conduct their research, Rocha and Rodrigues (2021). The two concepts of research design and research methodology need to be clarified first to clear up the confusion that is often associated with their usage, particularly by emerging researchers. Each of these concepts is presented as a compound word, with the concepts of design and methodology attached to the noun research.

This chapter describes the methodological approach used in this study. Data on the number of daily arrivals of patients were collected from a specified hospital and they were analysed using the software Python. The first section gives a description of the data and the data transformation processes. The following sections will expand on the description of the models that have been selected to be used in this study, ARIMA, XGBoost, GBR, and the VR. Lastly, the model evaluation section is introduced to assess the model section.

3.1 Study design, settings, and data source

The data to use in the study was collected from a specified hospital in the form of Excel Registers that were locked and password protected. A dataset of patient arrivals from the date May 2019 to November 2021 has been collected, with a total of 47 461 observations used for the analysis. The data received from the hospital had daily patient arrivals for normal hours (07:00 - 15:59), after hours (16:00 - 06:59) and the priority units.

The variables in the data include:

- The date and the time of arrival of the patient
- Patient's arrival for Normal Hours (NH)

- Patient's arrival for After Hours (AH)

The analysis will be carried out using Python, which is widely used for data analysis, machine learning, deep learning, and general programming.

3.1.1 Data Transformation

The data transformation stage is key for the data analysis and fitting of traditional time series models. This process includes:

- ensuring that the file can be cleaned or edited (that includes removal of security on each Excel workbook received),
- ensuring the data quality (which includes checking if the columns should only contain numerical values, contains numeric values only, no strings or special characters as this can render our data unclean),
- checking the data integrity

3.1.2 Raw Datafile Transformation

A Python script was used to import all the files sitting in a local folder, remove any security settings and export them with the same name to a cleaned folder.

3.1.3 Datafile Cleaning and Data Integrity Checks

To fulfill the datafile cleaning step, the analyst needs to select the relevant columns from the provided or cleaned file only and the relevant columns include: date, Normal Hours arrival time (timestamped) and After Hours arrival time (timestamped). An Unassigned column was added to cater for visits that took place and there was no priority assigned. These incidents could not be thrown away because the trained model still needs to know how many patients visited the hospital in a day. This was a manual process.

Several Excel macro (VBA) codes were written to assist in speeding up the process of data quality checking. The first process was to add the days to the cleaned data so that the Time Series model could be exposed to daily data.

3.1.4 Explanatory Data Analysis of Time Series Variables

This section begins by presenting the fundamental characteristics of the dataset, including the number of variables and observations, and affirms the completeness of the data by confirming there are no missing values. This assurance of data integrity is crucial for reliable statistical analysis.

Further, the section offers a meta-analysis of the dataset, providing an at-a-glance understanding of the data's structure. The descriptive statistics are then presented in a tabular format, which outlines key measures of central tendency and dispersion for each variable, such as mean, standard deviation, and range. These measures are essential for summarising the data and providing insights into the typical values and variability of patient arrivals.

The visual comparative analysis then explores the patterns of patient visits within the hospital, identifying the most crucial time frames for patient influx by differentiating between regular business hours and after-hours. The hospital can more effectively assess and manage its resources thanks to the graphic representation of this comparison, which reveals informative trends regarding the busiest days and times for patient arrivals.

In addition, we will focus on the statistical properties of the data. Histograms are employed to unpack the normality or non-normality of the distribution of patient arrivals across different priority wards during both normal and after working hours. This analysis is pivotal for understanding the underlying statistical structure of the data and for preparing it for further modeling and analysis.

All things taken into account, the exploratory data analysis offers a strong basis for comprehending the intricate dynamics of patient visits and acts as a model for more sophisticated analytical methods that could be used with the time series data.

3.2 Nonlinear Models

3.2.1 Tsay's Test

Tsay (1986) developed a test that helps improve the power of the above-mentioned nonlinearity tests. To improve the power of the nonlinearity tests developed by Keenan (1985) and Ramsey (1969), Tsay (1986) proposed to use a different set of explanatory variables for the test. The test is based on running an auxiliary equation in the form:

$$\hat{a}_t = \beta^{('Z_t)} + u_t \tag{3.1}$$

where

$$Z_t = vech(X_t X_t')$$

is a vector of predetermined variables, their squares and cross products, and vech(.) denotes a half-stacking operator. The version of the test statistic is defined as:

$$TSAY(p) = TR^2 d{\rightarrow}\chi^2 \left(\frac{p(p+1)}{2}\right) \tag{3.2}$$

where T denotes the sample size, and R^2 is the coefficient of determination from an auxiliary model. In a special case when $p = 1$, the Tsay coincides with the Keenan test proposed by Keenan (1985).

3.3 Nonlinear Unit Root Test

3.3.1 Kapetanios, Shin and Shell (KSS)

The traditional unit root test, like the ADF-GLS previously mentioned, may not be sufficient to detect stationarity in time series data when nonlinearity is present. It is crucial to employ stationarity tests that account for nonlinearity in such a situation, like the KSS test. The Augmented Dickey-Fuller (ADF) test is modified into the KSS test based on the following nonlinear model specification, Kapetanios et al., 2003.

$$Y_t = \beta Y_{t-1} + \gamma Y_{t-1}\left[1 - \exp\left(-\theta Y_{t-d}^2\right)\right] + \varepsilon_t, \tag{3.3}$$

which when parameterised yields:

$$\Delta Y_t = \delta Y_{t-1} + \gamma Y_{t-1}\left[1 - \exp\left(-\theta Y_{t-d}^2\right)\right] + \varepsilon_t,$$

where $\delta = \beta - 1 \cdot \gamma$, θ are parameters that must be estimated and ε_t is the residual term. The KSS test sets $\delta = 0$ and the decay parameter, $d = 1$, so that the test is formally based on the following specification:

$$\Delta X_t = \gamma X_{t-1}\left[1 - \exp\left(-\theta X_{t-d}^2\right)\right] + \varepsilon_t, \tag{3.4}$$

The KSS tests the null hypothesis of linear stationarity by setting $\theta = 0$ against the alternative that $\theta > 0$. However, Kapetanios et al.,2003 argue that it is impossible to directly test the null hypothesis since the speed of reversion, γ, is unknown. Using a first-order Taylor series approximation, Luukkonen et al.,1998 reformulated an estimable nonlinear specification for testing nonlinear stationarity in X_t, as:

$$\Delta X_t = \xi_{X_{t-1}}^3 + \varepsilon_t.$$

To account for the possibility of serial correlation in the error term, the equation is augmented with lags of the first-difference of X_t as:

$$\Delta X_t = \xi X_{t-1}^3 + \sum_{j=1}^{p} \delta_j \Delta X_{t-j} + \varepsilon_t$$

where ξ is the coefficient used to test the presence of a unit root. From the nonlinear stationarity specification, the KSS-NADF unit root test is based on the t-statistic:

$$\tau_{NL} = \frac{\hat{\xi}}{s.e.(\hat{\xi})} \tag{3.5}$$

Three distinct nonlinear model specifications—raw data, demeaned data, and detrended data—are used to construct three distinct asymptotic critical values (Kapetanios et al., 2003). The scenarios listed below are most common:

- If X_t has a zero mean, then the appropriate data to use is $Y_t = X_t$, the raw data.
- If X_t has a non-zero mean and zero trend, then the appropriate data to use is $Y_t = X_t - \overline{X}$, the demeaned data, where $\overline{X}$ is the mean of the data.
- If X_t has a non-zero mean and non-zero trend, then the appropriate data to use is $Y_t = X_t - (\alpha_0 + \alpha_1 t)$, the detrended data, where $\alpha_0 + \alpha_1 t$ is the trendline obtained by regressing X_t on timepoint $t = 1, 2, 3, \ldots, n$ with an intercept term.

The selection of the lag length, p, has an impact on the KSS-NLADF test. Using Hall's (1994) general-to-specific method is one well-known way to choose p. To do this, set up the Schwert (1987) suggested upper bound, p_{max} :

$$p_{max} = integer\left[12\left(\frac{n}{100}\right)^{\frac{1}{4}}\right] \tag{3.6}$$

where n is the sample size, estimating the test regression with $p = p_{max}$

Liew et al. (2004) recommend a lag length of 8, which is what this study will adhere to. If the final included lag is significant at the 1% 5% or 10% level, it is kept as the ideal lag and utilized in the KSS-NLADF unit root test. The ideal lag for the KSS-NLADF unit root test is determined by reducing p by one lag until the last included lag is significant. This process is carried out if the last included lag is not significant. If the t-statistic exceeds the critical value at a certain significance level, the alternative hypothesis—the nonlinear unit root — is accepted.

3.4 Autoregressive Moving Integrated Moving Average (ARIMA) models

Autoregressive Moving Integrated Moving Average (ARIMA) models refer to a class of time series models that result in a linear combination of moving average (MA) and auto regression (AR) models together with the differential operator (I). They constitute a comprehensive family of models that are suitable to capture a large set of linear relationships. As will be discussed in subsequent sections, ARIMA models have been successfully applied to model time series of patient volumes in emergency departments.

ARIMA models were popularised by statisticians George Box and Gwilym Jenkins and there are sometimes referred to as "Box-Jenkins models".

Our main reference for this section is the book by Shumway and Stoffer (2016).
A time series (x_t) is called an ARMA(p, q) time series if it can be represented as

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q} \tag{3.7}$$

where $\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$ are real numbers and w_t is a white noise.

The parameters $\phi_1, \phi_2, \ldots, \phi_q$ are called the autoregression coefficients and $\theta_1, \theta_2, \ldots, \theta_q$ the moving average coefficients. Clearly a ARMA$(p, 0)$ model is the same as a AR(p) model and similarly a ARMA$(0, q)$ is a MA(q) model.

Box and Jenkins (1976) introduced a methodical process for constructing ARIMA models through a cycle of three steps: pinpointing the model type, estimating parameters, and conducting diagnostics. Initially, one ensures the model aligns with certain theoretical autocorrelation characteristics expected from an ARIMA-generated series. Comparing these with empirical data helps select suitable model candidates. Tools like autocorrelation and partial autocorrelation functions aid in determining the ARIMA model's order. For effective model construction, data must be stationary, often requiring transformations like converting stock prices to returns and testing stationarity through the augmented Dickey-Fuller test.

Stationary series exhibit constant statistical features, such as mean and autocorrelation, over time. Differencing and power transformations may be applied to data with trends or volatility to meet stationarity prerequisites before fitting the ARIMA model. Parameter estimation is fairly direct, involving nonlinear optimization to minimize error. Unlike neural networks, ARIMA does not incorporate various technical indicators, which can be both a limitation and a simplification since more inputs don't always enhance forecast precision. The final step involves diagnostic checks to confirm model adequacy by inspecting error assumptions, fitting quality, and residual plots. If the model falls short, one must reformulate it, estimate parameters again, and re-evaluate until a satisfactory model is developed for forecasting purposes.

3.5 The XGBoost Model and Gradient boosting algorithms

3.5.1 Introduction

The XGBoost is also called the Newton boosting because of its relation to the classical Newton's optimisation method. Assume that the following data have been collected:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

It is assumed that $y_1, y_2, \ldots, y_n$ are independent random realisations of a random variable Y and $x_1, x_1, \ldots, x_1$ also independent realisations of a random variable $X = (X_1, X_2, \ldots, X_p)$. The random variable Y is the response or dependent variable and the random variables $(X_1, X_2, \ldots, X_p)$ are the predictors, explanatory or independent variables. Here there are p explanatory or independent variables.

In order to make a prediction, consider a model, that is a function

$$f : X \to R, \ x \mapsto f(x)$$

that is assumed to represent the data as closely as possible and is able to generalise to values that are not in the data set. Then for a given point $x_i \in X$, we can make the prediction that $a_i = f(x_i)$. Of course in general this prediction is not necessarily exact, that is, generally $a \neq y_i$. In general, assume that we make the prediction $\hat{y} = f(x)$ for generic point $x \in X$ and that the exact point corresponding to x is y.

Then the corresponding loss is given by:

$$L(y, \hat{y})$$

where L is a specified loss function such as:

$$\begin{aligned}
L(y, \hat{y}) &= (y - \hat{y})^2; \\
L(y, \hat{y}) &= |y - \hat{y}|, \\
L(y, \hat{y}) &= \log(1 + \exp(\hat{y}) - y\,\hat{y}) \\
L(y, \hat{y}) &= \begin{cases} 1 \text{ if } y = \hat{y} \\ 0 \text{ if } y \neq \hat{y}. \end{cases}
\end{aligned}$$

For a model f, the empirical risk corresponding to the given data is given by:

$$\hat{R}(f) = \frac{1}{n} \sum_{k=1}^{n} L(y_i, f(x_i)).$$

Given a certain class of functions C, the problem it to select a function f in the class C such that f minimises the empirical risk. We shall consider the tree model where the class of functions is defined as trees, that is, functions that can be written as linear combinations of indicators functions:

$$\phi(x) = \sum_{k=1}^{M} \theta_k I_{R_k}(x)$$

where $R_1, R_2, \ldots, R_M$ are subsets of the input space X, $(\theta_1, \theta_2, \ldots, \theta_M)$ are parameters and I_{R_k} is the indicator function of R_k.

Generally, the subsets R_i are obtained by subdividing the space X into rectangulars of parallel and equal sides.

For a specified class C of functions (called the learner base) the boosting of a model corresponds to sequentially constructing functions

$$\widehat{f}^{(1)}, \widehat{f}^{(2)}, \ldots, \widehat{f}^{(M)}$$

defined by:

$$\widehat{f}^{(m)}(x) = \theta_0 + \sum_{k=1}^{m} \theta_k \phi_k(x)$$

where $\phi_1, \phi_2, \ldots, \phi_M$ are functions in Φ and $\theta_0, \theta_1, \theta_2, \ldots, \theta_M$ are parameters. At step m, the function ϕ_m and the parameter θ_m are choosen to minimise some quantity depending on the given loss function.

$$\sum_{i=1}^{n} L(y_i, \widehat{f}^{(m-1)}(x_i) + \theta_m \phi_m(x_i)) \, .$$

3.5.2 XGBoost algorithm

The XGBoost algorithm or Newton boosting algorithm is is described with the following steps: (see Chen & Guestrin, 2016 and Nielsen, 2016 for more details.)

(1) The input is:

- The data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- The loss function L.
- A class of functions (the learner base) C.
- The number M of iterations.

- A parameter η called learning rate.

(2) Initial step: the initial function $\widehat{f}^{(0)}$ is the constant $\widehat{\theta_0}$ obtained by minimising the loss

$$\sum_{k=1}^{n} L(y_i, \theta),$$

with respect to θ.

(3) For each number $m = 1, 2, 3, \ldots, M$,

- Compute the gradient $g_m(x_i)$ of the loss function at the current function $\widehat{f}^{(m-1)}$, that is,

$$g_m(x_i) = \left.\frac{\partial L(y_i, z)}{\partial z}\right|_{z=\widehat{f}^{(m-1)}(x_i)}.$$

- Compute also the second derivative:

$$h_m(x_i) = \left.\frac{\partial^2 L(y_i, z)}{\partial z^2}\right|_{z=\widehat{f}^{(m-1)}(x_i)}.$$

- Find a function $\phi \in C$ that minimises the quantity

$$\sum_{i=1}^{n} \tfrac{1}{2} h_m(x_i) \left(-\frac{g_m(x_i)}{h_m(x_i)} - \phi(x_i \right)^2$$

or equivalently the function ϕ that minimises the quantity

$$\sum_{i=1}^{n} \tfrac{1}{2} h_m(x_i) \left(\phi(x_i)\right)^2 + g_m(x_i)\phi(x_i)$$

and denote such a function $\hat{\phi}_m$.

- Take

$$\widehat{f}^{(m)}(x) = \widehat{f}^{(m-1)}(x) + \eta \phi_m(x).$$

(4) Return

$$\hat{f}(x) = \widehat{f}^{(M)}(x).$$

Now for the particular case of Newton boosting for trees, the base learner is taken to be the class of tree functions, that is, functions that can be written as linear combinations of indicators functions:

$$\phi(x) = \sum_{j=1}^{T} w_j I_{R_j}(x)$$

where R_j are rectangles in the input space and as before I_{R_j} is the indicator function of $R_{m,j}$ and $w_{m,j}$ are parameters. Generally, the subsets R_i are obtained by subdividing the space X into rectangulars of parallel and equal sides.

3.5.3 XGBoost algorithm for trees

The XGBoost algorithm or Newton boosting algorithm for trees is the following. At each step, the goal is to choose a tree function ϕ that minimises the quantity:

(1) The input is:

- The data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- The loss function L
- The number of terminal nodes T
- The number M of iterations
- A parameter η called learning rate
- Parameters γ and λ.

(2) Initial step: the initial function $\widehat{f}^{(0)}$ is the constant $\widehat{\theta_0}$ obtained by minimising the loss

$$\sum_{k=1}^{n} L(y_i, \theta),$$

with respect to θ.

(3) For each number $m = 1, 2, 3, \ldots, M$,

- Compute the gradient $g_m(x_i)$ of the loss function at the current function $\widehat{f}^{(m-1)}$, that is,

$$g_m(x_i) = \left. \frac{\partial L(y_i, z)}{\partial z} \right|_{z=\widehat{f}^{(m-1)}(x_i)} .$$

- Compute also the second derivative:

$$h_m(x_i) = \left. \frac{\partial^2 L(y_i, z)}{\partial z^2} \right|_{z=\widehat{f}^{(m-1)}(x_i)} .$$

- Find a tree structure and the corresponding tree function ϕ that minimises the quantity:

$$\sum_{i=1}^{n} \tfrac{1}{2} h_m(x_i) \left(\phi(x_i)\right)^2 + g_m(x_i)\phi(x_i) + \Omega(\phi) \tag{3.8}$$

  where

$$\Omega(\phi) = \gamma T + \tfrac{1}{2}\lambda\|\phi\|^2.$$

  Write such tree function as

$$\hat{\phi} = \sum_{j=1}^{T} \widehat{w}_{jm} I(x \in \widehat{R}_{jm})$$

  (where $I(x \in A)$ is the indicator function of the set A).
- Compute the weights of the leaf j of the tree by:

$$\widehat{w}_{jm} = \frac{\sum_{i \in I_j} g_m(x_i)}{h_m(x_i) + \lambda}$$

  where I_j is the set of all indices i such that the input point x_i is the region R_{jm}
- Take

$$\widehat{f}^{(m)}(x) = \widehat{f}^{(m-1)}(x) + \eta \sum_{j=1}^{T} \widehat{w}_{jm} I(x \in \widehat{R}_{jm})$$

(4) Return

$$\hat{f}(x) = \widehat{f}^{(M)}(x).$$

To find a tree function that minimises (3.8) at iteration m, one can start with a tree with a single node and iteratively add branches by splitting nodes into two others. If at a certain step, the node represented by the set I is split into two nodes with sets I_L and I_R then as discussed in Chen & Guestrin (2016), loss reduction or gain resulting from the split is given by:

$$\frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G_I^2}{H_I + \lambda}\right) - \gamma$$

where

$$G_L = \sum_{i \in I_L} g_m(x_i); G_R = \sum_{i \in I_R} g_m(x_i); \quad G_I = \sum_{i \in I} g_m(x_i)$$
$$H_L = \sum_{i \in I_L} h_m(x_i); H_R = \sum_{i \in I_R} h_m(x_i); \quad H_I = \sum_{i \in I} h_m(x_i).$$

More details can be found in Chen & Guestrin (2016), Friedman (2001), and Nielsen (2016).

Parameter Tuning

A summary of the tuned parameters is given below

alpha: Usually, this parameter is connected to regularization methods. It could be referring to the alpha quantile of the loss function in quantile regression while discussing gradient boosting models. An alpha of 0.9, for instance, would concentrate on the 90th percentile, which is frequently employed to forecast higher values in the data distribution.

learning_rate (eta): This is the shrinkage of step size that keeps overfitting from happening. We can immediately obtain the weights of newly added features after each boosting step, and eta reduces the feature weights to make the boosting procedure more cautious.

gamma: Also referred to as the minimal loss reduction needed to create an additional partition on a tree leaf node. It functions as a phrase for regularization.

max_depth: It stands for a tree's maximum depth. The model will become more sophisticated and perhaps more prone to overfit if this value is increased.

min_child_weight: The child's minimal required instance weight (hessian) sum. The building process will give up on further partitioning if the tree partition phase yields a leaf node with the sum of instance weight less than min_child_weight.

n_estimators: This indicates how many trees there are in the forest. Although more trees can collect more information from the data, they can also slow down the training process and increase the risk of overfitting if the learning rate is not changed.

min_samples_leaf: This is the bare minimum of samples that must be present at a leaf node. Raising this value can help the model become more smooth, particularly in regression tasks, as it will stop the model from producing noisy leaves with small sample sizes.

min_samples_split: It specifies how few samples are needed to split an internal node. Again, higher values aid in reducing overfitting by preventing the creation of nodes that represent too fine-grained patterns.

colsample_bytree: The subsample ratio of columns used to build each tree is this parameter. There is one subsampling for each tree that is built.

reg_lambda: The L2 regularization term on weights. It is used to avoid overfitting.

subsample: The fraction of data to be used for fitting each unique base learner is defined by this parameter. Stochastic gradient boosting, in which every tree is trained on a random subset of the data, is applied when the value is less than 1.0. This method can aid in lowering variance and boosting the resilience of the model.

3.6 Model Validation

To establish a robust methodology for evaluating the prediction of the models, the following metrics will be utilized: Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Percentage Difference (MPD). These metrics offer various lenses through which the model's forecast accuracy and performance can be scrutinized.

Mean Squared Error (MSE) measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. The MSE is calculated as:

$$MSE(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{3.9}$$

where Y_i is the actual value, $\hat{Y}_i$ is the forecast value, and n is the number of observations.

Mean Absolute Error (MAE) is a measure of errors between paired observations expressing the same phenomenon. The formula for MAE is:

$$MAE(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|. \tag{3.10}$$

Mean Absolute Percentage Error (MAPE) is a measure of prediction accuracy in a forecasting model, calculated as the average of the absolute percentage errors. The formula for MAPE is:

$$MAPE(Y, \hat{Y}) = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|. \tag{3.11}$$

Root Mean Squared Error (RMSE) is the square root of the average of squared differences between prediction and actual observation. The formula for RMSE is the square root of MSE:

$$RMSE(Y,\hat{Y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{3.12}$$

RMSE adds a little more weight to large errors and is useful when large errors are particularly undesirable.

Mean Percentage Difference (MPD) is used to calculate the average of percentage differences between forecasted and actual values. It's a relatively less common metric but useful in certain contexts to understand the deviation in terms of percentage. Its formula is:

$$MPD(Y,\hat{Y}) = \frac{100\%}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{Y_i}\right) \tag{3.13}$$

MPD differs from MAPE in that it does not take the absolute value, which can provide insights into the direction of the errors.

Each of these metrics will be computed using the actual and forecasted values from the models under consideration. The choice of metric should align with the specific objectives of the forecast and the cost of errors in the particular application, James et al., (2013).

3.7 Model Selection

3.7.1 Auto ARIMA Model Selection

Auto ARIMA model selection is a process of automatically selecting the optimal parameters for an ARIMA model (Hydman& Athanasopoulos, 2021). ARIMA models are a class of statistical models that are used to forecast future values of a time series based on its past values. Auto ARIMA model selection can save time and effort, as it does not require manually testing different ARIMA models. It can also help to identify models that would not have been considered otherwise (Hydman& Athanasopoulos, 2021).

The most common methods of auto ARIMA model selection are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These methods penalize models that are too complex, and the model with the lowest AIC or BIC is considered to be the best model. Other methods of auto ARIMA model selection include cross-validation, information criteria such as the Hannan-Quinn information criterion (HQIC), and statistical tests such as the Ljung-Box test.

To perform auto ARIMA model selection in practice, the data is first split into training and testing sets. The ARIMA models are then fitted to the training set, and their performance is evaluated on the testing set. The model with the best performance on the testing set is selected as the optimal model.

3.7.2 Machine Learning Model Selection

The methods with the best initial prediction performance are mostly used to select machine learning models.

To check model performance, the dataset was divided into training and testing subsets. This method reduces overfitting and is necessary for an objective assessment (Varma & Simon, 2010). Before adjusting the hyperparameters to choose the top-performing algorithms for the main model selection, several preselected algorithms were used to train the model.

Hyperparameter optimization is an integral part of model selection, with grid search, random search, and more sophisticated methods like Bayesian optimization often being utilized (Snoek, Larochelle, & Adams, 2012). The use of automated machine learning (AutoML) systems, which automate the process of selecting models and hyperparameters and save time and resources while frequently producing more repeatable results, is another recent trend in model selection (He & Garcia, 2010; Feurer et al., 2015).

Once a model is selected, you should conduct a final validation using the test set to ensure the robustness and predictive power of the model. The model with the best test performance, considering both accuracy and generalization, is typically chosen for deployment. Frequently, this iterative process calls for several training, fine-tuning, and validation cycles (Bergstra et al., 2011).

The use of traditional time series and modern machine learning models in the study aligns closely with the objectives to enhance forecasting accuracy for patient arrivals at a public hospital. These models facilitate a comprehensive Exploratory Data Analysis (EDA) to decipher complex patterns in patient arrival data, supporting the objective to develop robust forecasting models. ARIMA, XGBoost, Gradient Boosting Regressor (GBR), and Voting Regressor have been chosen based on their proven capabilities in handling various data behaviors, thus ensuring reliable predictions which are crucial for optimizing resource allocation. Evaluating the performance of each model against real-world data ensures that the most effective methods are recommended for practical application, ultimately aiding the hospital in strategic planning and improving patient management processes.

Chapter 4

Data Analysis and Interpretation

The current chapter will report on the data summary, exploratory data analysis, preliminary analysis, the interpretation of predictive characteristics found in the time series, and the clarity of the model outputs.

The dataset is first explored, split into series types i.e., Normal Hours train sample (used for model building) and test samples (used to evaluate the reliability and the forecasting/predictive power of the model). Further statistical investigations are conducted to treat non-normality and non-stationarity in the training sample prior to model building. The chapter will illustrate all properties found in the dataset in the forms of statistical graphs and table outputs with their respective interpretations.

The chapter is structured systematically, covering various aspects of time series analysis. It begins with an exploratory data analysis followed by an examination of time series plots for Normal and After Hours. It delves into the identification of time series components. Subsequently, the chapter explores the testing of stationarity assumptions and normality assumptions for different time periods.

The chapter culminates in the model, aiming to identify the best-fit model across all variations. Finally, a concluding section of the chapter, with the subsequent section focusing on performance evaluation.

4.1 Preliminary Data Analysis Of Time Series Variables

Preliminary data analysis is centered on exploring unique properties contained in the time series and establishing significant relationships that might influence the dependent variables. The overall data contains 1007 daily observations (from May 2019 to November 2021). The data is divided into two main groups: Normal Hours and After Hours. Patients in the Normal Hours group are those who arrive at the emergency department between 8:00 and 16:00 and those in the After Hours group are those who arrive between 16:00 and 08:00.

4.1.1 Visual Comparative Analysis

This section show a graphical representation of the comparison between Normal versus After Hours to see which of the two contributes to the peak periods.

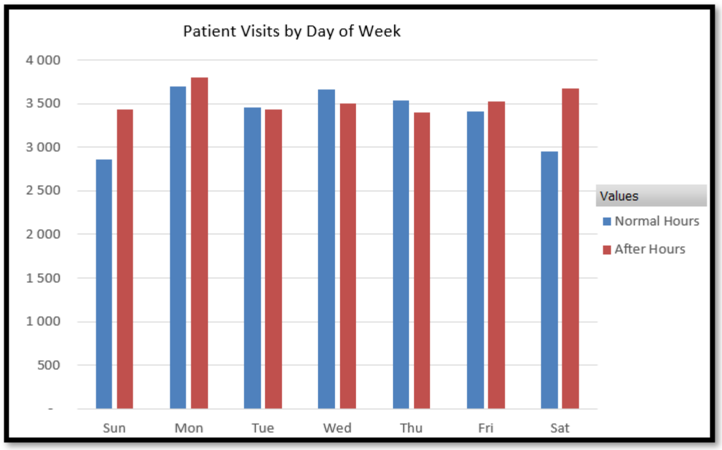


Figure 4.1 – Comparison of Normal vs After Hours

Figure 4.1 implies that for days around the weekend namely Friday, Saturday, Sunday and Monday, the most patient arrivals are after working hours. While during the week most patients arrive at the hospital during Normal Hours.

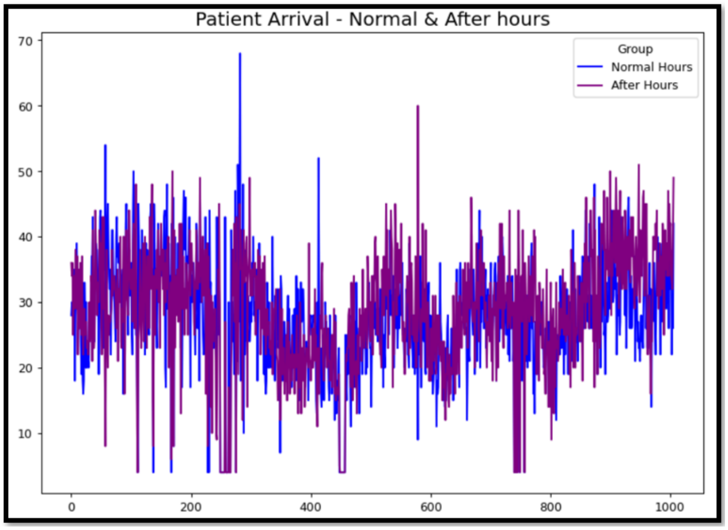
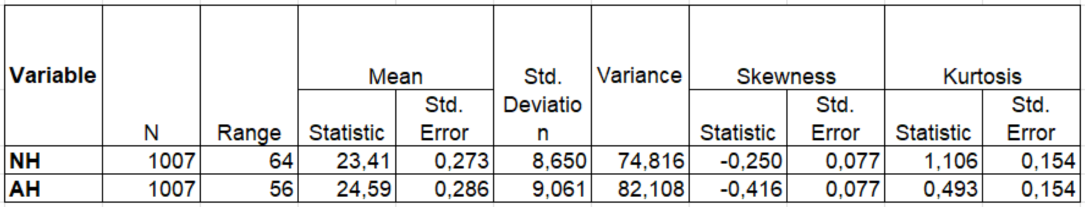


Figure 4.2 – Series for Normal and After Hours

The pattern is almost similar between the two series samples, with the exception of the NH sample spiking above the average and AH hitting the trough line more often than the other as depicted by Figure 4.2. However, the generalised pattern between the two series averages is the same, meaning by visual inspection the two series are not significantly different.

4.1.2 Descriptive Analysis



Variable			Mean		Std. Deviatio n	Variance	Skewness		Kurtosis	
	N	Range	Statistic	Std. Error			Statistic	Std. Error	Statistic	Std. Error
NH	1007	64	23,41	0,273	8,650	74,816	-0,250	0,077	1,106	0,154
AH	1007	56	24,59	0,286	9,061	82,108	-0,416	0,077	0,493	0,154

Table 4.1 – Descriptive Statistics of Daily Patient Arrivals for each variable

Table 4.1 presents descriptive statistics for daily patient arrivals, comparing two variables: Normal Hours (NH) and After Hours (AH). Both variables were observed across the same number of days, N = 1007.

For Normal Hours, the maximum number of daily patient arrivals was 64, with a mean (average) of 23.41. This suggests that, on average, about 23 patients arrived during normal hours across the observed days, but on at least one day, this number spiked to 64. The standard deviation, a measure of the spread or dispersion of the data, was 8.650, indicating that the number of patient arrivals typically varied by about 8-9 patients from the mean. The standard error of the mean, at 0.273, suggests that the mean of 23.41 is relatively precise as an estimate of the true average number of arrivals. The variance, which is the square of the standard deviation, was 74.816, confirming the spread in the data.

Skewness for NH is -0.250 with a standard error of 0.077, indicating a slight negative skew meaning that the tail on the left side of the distribution is longer or fatter than the right side, showing a slight tendency for days with fewer than average arrivals. Kurtosis, at 1.106 with a standard error of 0.154, suggests a distribution that is slightly ”peakier” than a normal distribution (which has a kurtosis of 3). The Jarque-Bera test, which tests the hypothesis that the data follows a normal distribution, yielded a statistic of 61.931 and a very small p-value (3.564e-14), strongly suggesting that the patient arrivals during Normal Hours are not normally distributed.

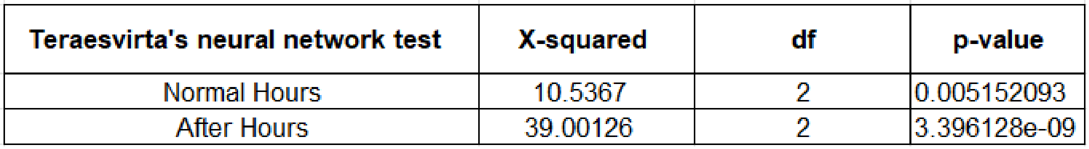
For After Hours, the maximum number of patient arrivals was lower at 56, and the mean was slightly higher at 24.59. This suggests that, on average, there are more patient arrivals during after hours compared to normal hours, with a lower maximum number observed. The standard deviation was slightly higher at 9.061, indicating a wider variation in the number of patient arrivals during after hours. The standard error of 0.286 reflects a similarly precise estimate of the mean compared to NH. The variance was 82.108, which is higher than NH, showing more variability in the AH data.

The skewness for AH is more pronounced at -0.416, with the same standard error as NH, indicating a more noticeable negative skew compared to NH. The kurtosis is lower at 0.493, suggesting a flatter distribution compared to NH and much flatter than a normal distribution. The Jarque-Bera test for AH produced a statistic of 46.643 and a very small p-value (7.439e-11), also rejecting the normality of the distribution for patient arrivals during After Hours.

In conclusion, both NH and AH patient arrivals show significant variation, with AH showing a slightly higher average but lower peak and more pronounced negative skewness. Neither distribution is normal, with AH being flatter and NH being more peaked compared to a normal distribution. This non-normality is confirmed by the very small p-values in the Jarque-Bera test for both variables.

4.1.3 Nonlinearity Tests

The following is a table that shows the results of Teraesvirta's neural network test for nonlinearity, comparing Normal Hours to After Hours.



Teraesvirta's neural network test	X-squared	df	p-value
Normal Hours	10.5367	2	0.005152093
After Hours	39.00126	2	3.396128e-09

Table 4.2 – Teraesvirta's neural network test for nonlinearity

These results are indicative of testing for nonlinearity. The null hypothesis usually assumes linearity and a low p-value suggests evidence against linearity.

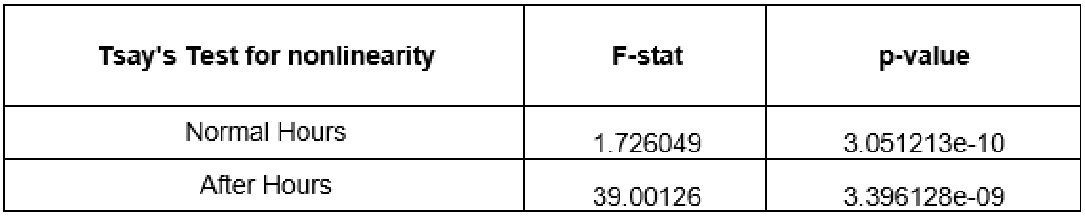
The chi-squared statistic of 10.5367 and the p-value of 0.005152093 indicate that there is significant nonlinearity in patient arrivals during NH. This suggests that patient arrivals during NH do not follow a simple linear pattern. There could be various factors contributing to this nonlinearity, such as:

- Daily Fluctuations in Patient Volume: The number of patients arriving at the emergency department might naturally exhibit a nonlinear pattern throughout the day, with peak periods in the morning and evening hours (Akenfeldt et al., 2012).
- Variations in Staff Availability: Staffing levels might vary throughout the day, which could impact the capacity to handle patient arrivals and potentially introduce nonlinearity into the waiting time or treatment duration (Doran et al., 2010).

The chi-squared statistic of 39.00126 and the p-value of 3.396128e-09 indicate that there is even more pronounced nonlinearity in patient arrivals After Hours. This suggests that patient arrivals After Hours exhibit a more complex and unpredictable pattern compared to Normal Hours (Akenfeldt et al., 2012). Several factors could contribute to this heightened nonlinearity:

- Reduced Staffing and Resources: After hours, the emergency department might operate with reduced staffing and resource availability, which could lead to longer waiting times and increased congestion, potentially influencing the nonlinearity of patient arrivals (Doran et al., 2010).
- Urgent and Emergent Cases: A higher proportion of patients arriving after hours might present with urgent or emergent conditions, leading to more unpredictable arrival patterns (Schwartz et al., 1998).

- Lack of Scheduled Appointments: The absence of scheduled appointments or procedures after hours could further contribute to the nonlinearity of patient arrivals.



Tsay's Test for nonlinearity	F-stat	p-value
Normal Hours	1.726049	3.051213e-10
After Hours	39.00126	3.396128e-09

Table 4.3 – Tsay's neural network test for nonlinearity

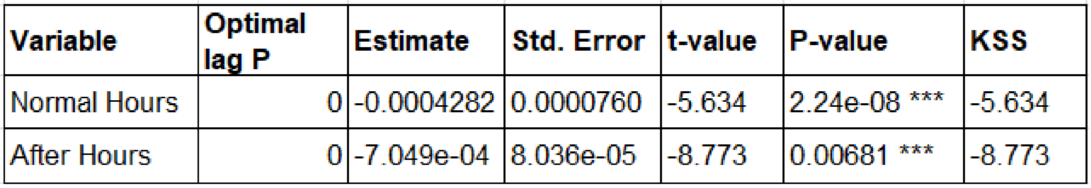
Table 4.3 reports on Tsay's Test for results. The test involves examining the null hypothesis that a time series is linear against the alternative hypothesis that it is nonlinear.

In both cases, the p-values are extremely small (close to zero), which typically leads to rejecting the null hypothesis. In the context of Tsay's Test, this would suggest evidence of nonlinearity in the time series for both Normal Hours and After Hours.

4.1.4 Nonlinear Unit Root Test

This table presents the results of a Kapetanios, Shin, and Snell (KSS) Nonlinear Unit Root Test. The optimal lag length for each variable is selected using the general-to-specific method suggested by Hall (1994) This test is designed to determine whether a time series is stationary or has a unit root, which would indicate non-stationarity.

Critical values of the KSS-NLADF test with constant and trend at the 10% 5% and 1% significant levels are -3.13, -3.40 and -3.93, respectively.



Variable	Optimal lag P	Estimate	Std. Error	t-value	P-value	KSS
Normal Hours	0	-0.0004282	0.0000760	-5.634	2.24e-08 ***	-5.634
After Hours	0	-7.049e-04	8.036e-05	-8.773	0.00681 ***	-8.773

Table 4.4 – KSS Nonlinear Unit Root Test

The results suggest that the variable Normal Hours has a statistically significant negative nonlinear unit root. The t-value is -5.634, and the p-value is very close to zero (2.24e-08), indicating strong evidence against the presence of a unit root.

The results for the variable After Hours also indicate a statistically significant negative nonlinear unit root. The t-value is -8.773, and the p-value is 0.00681, suggesting strong evidence against the presence of a unit root.

In summary, based on the KSS Nonlinear Unit Root Test, both Normal hours and After hours variables appear to be stationary without a unit root. The negative estimates and significant t-values provide evidence against the presence of a unit root in these series.

4.2 Model estimation

In this section, we discuss the estimated models for the data. We will consider ARIMA, XGBoost, GBR and VR.

4.2.1 ARIMA Model Training

This section will cover the training of ARIMA model in the context of Normal working hours and After Working Hours.

Normal Hours

Parameter Estimations

This area section shows the number of parameters to be included in the fitting of the ARIMA model by uncovering the relevant ARIMA input. The section does that by showing the Autocorrelation function (ACF) and Partial autocorrelation functions (PACF).

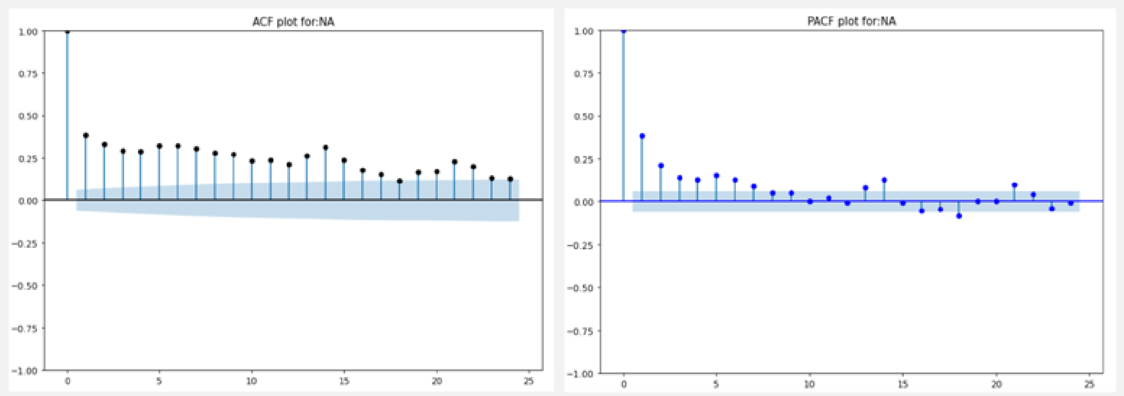
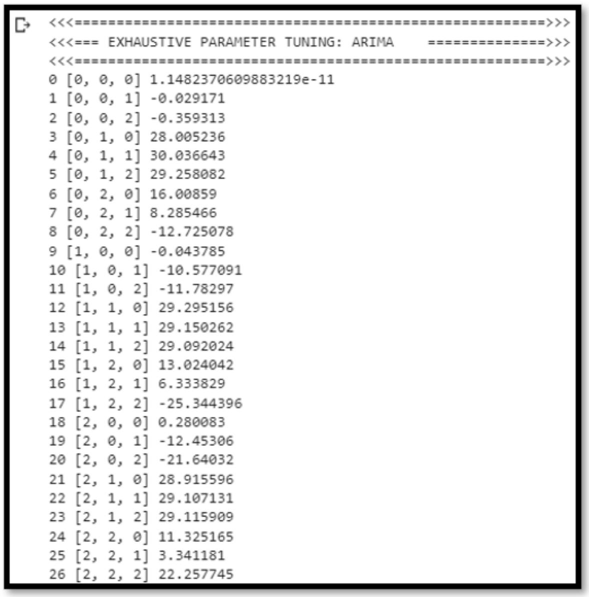


Figure 4.3 – ACF and PACF plots

Based on the spike by the ACF plot, the greatest decline is shown of value if 1 followed by other declines and the PACF plot also shows the same. This suggests that the correlation functions deem several options as optimal but a value of 1 being the best. This means the best ARIMA model based on ACF and PACF is ARIMA(0,0,1) and it will be confirmed by later discoveries.

Tuning the ARIMA Hyper-Parameters

The previous area of a sub-section uncovered ARIMA(0,0,1) to be the best, and this section will verify the discovery by performing an iterative search for the best parameters.



```
<<<========================================================>>>
<<<=== EXHAUSTIVE PARAMETER TUNING: ARIMA    ==============>>>
<<<========================================================>>>
0 [0, 0, 0] 1.1482370609883219e-11
1 [0, 0, 1] -0.029171
2 [0, 0, 2] -0.359313
3 [0, 1, 0] 28.005236
4 [0, 1, 1] 30.036643
5 [0, 1, 2] 29.258082
6 [0, 2, 0] 16.00859
7 [0, 2, 1] 8.285466
8 [0, 2, 2] -12.725078
9 [1, 0, 0] -0.043785
10 [1, 0, 1] -10.577091
11 [1, 0, 2] -11.78297
12 [1, 1, 0] 29.295156
13 [1, 1, 1] 29.150262
14 [1, 1, 2] 29.092024
15 [1, 2, 0] 13.024042
16 [1, 2, 1] 6.333829
17 [1, 2, 2] -25.344396
18 [2, 0, 0] 0.280083
19 [2, 0, 1] -12.45306
20 [2, 0, 2] -21.64032
21 [2, 1, 0] 28.915596
22 [2, 1, 1] 29.107131
23 [2, 1, 2] 29.115909
24 [2, 2, 0] 11.325165
25 [2, 2, 1] 3.341181
26 [2, 2, 2] 22.257745
```

Figure 4.4 – ARIMA Optimal Parameter Searching for Normal Hours

The above result shows that iteration 1 is the best as it has the minimum error. Therefore ARIMA(0,0,1) should be fitted to obtain the best results as it is also supported by ACF and PACF plots.

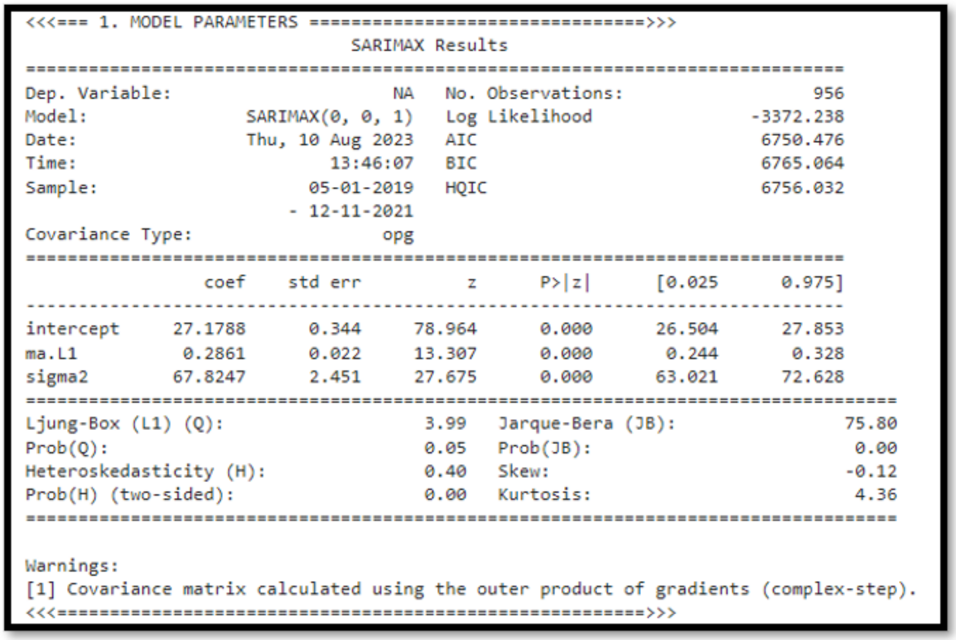
ARIMA(0,0,1) Parameters - output

Interpreting the various statistics and components of an ARIMA(0,0,1) model is important for understanding the model's goodness of fit and whether it is suitable for the data assessed.

Based on Figure 4.5, the following learning points from the model parameters are highlighted:

The log likelihood was estimated to be -3372.238 and as a rule of thumb; a more negative log likelihood indicates a better fit for the model, this means the model performs relatively well.

The model has the Akaike Information Criterion (AIC) value of 6750.476, a Bayesian Information Criterion (BIC) value of 6765.64, and a Hannan-Quinn Information Criterion (HQIC) value of 6765.064, which are all meant to measure the trade-off between goodness of fit and the model's complexity but mainly focusing on penalising for complexity. Since all the measured values (AIC, BIC, and HQIC) are relatively high, they suggest that the trade-off was not balanced which results in a highly complex model.



```
<<<=== 1. MODEL PARAMETERS ================================>>>
                               SARIMAX Results
==============================================================================
Dep. Variable:                     NA   No. Observations:                  956
Model:               SARIMAX(0, 0, 1)   Log Likelihood               -3372.238
Date:                Thu, 10 Aug 2023   AIC                           6750.476
Time:                        13:46:07   BIC                           6765.064
Sample:                    05-01-2019   HQIC                          6756.032
                         - 12-11-2021
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     27.1788      0.344     78.964      0.000      26.504      27.853
ma.L1          0.2861      0.022     13.307      0.000       0.244       0.328
sigma2        67.8247      2.451     27.675      0.000      63.021      72.628
===================================================================================
Ljung-Box (L1) (Q):                   3.99   Jarque-Bera (JB):                75.80
Prob(Q):                              0.05   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                            -0.12
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.36
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
<<<=======================================================>>>
```

Figure 4.5 – ARIMA Model Parameters for Normal Hours

The Intercept parameter was estimated to a value of 27.1788, which represents the constant term as an expected value of the time series when lagged values are not considered. The model has a Moving Average Coefficient (ma.L1) value of 0.2861, which is a coefficient associated with the moving average term of order 1. And it measures the influence of the most recent error term on the current value which indicates a positive effect in this case.

A Residual Variance (Sigma2) of 67.8247 was estimated, which represents the estimated variance of the residuals. The higher the Sigma2 value, the greater the variability of the residuals, which is the case for this model. Ljung-Box Statistic of 3.99 was estimated, which tests for autocorrelation in the residuals. Due to the low value of this statistic, it suggests that there may be no significant autocorrelation in the residuals. Ljung-Box Test (Prob(Q)) p-value was estimated to be 0.05, which suggests that there might be some residual autocorrelation, but it's not extremely significant.

Heteroskedasticity was estimated to a value of 0.40, which suggests that there is no strong evidence of heteroskedasticity (changing variance) in the residuals. Prob(H) Two-Sided was estimated to be 0.00, that is when associated with heteroskedasticity is very low, this affirms a significant presence of constant variance in the residuals. The Jarque-Bera statistic (JB Statistic) was estimated to be 75.80 by testing for the normality within the residuals, this indicates departures from normality in the residuals. Jarque-Bera Test p-value (Prob(JB)) was estimated to a value of 0.00, the very low p-value that is associated with the previous JB statistic of 0.00 confirms the non-normality (that the residuals are not normally distributed).

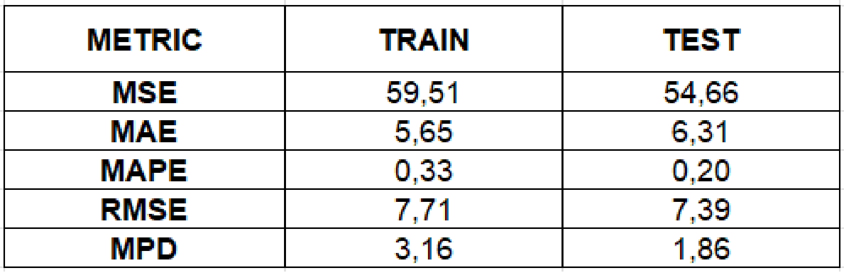
The Skewness was estimated to a value of -0.12, which is a slight leftward or negative skew in the residuals, this indicates that the distribution is skewed to the left slightly. The Kurtosis was also estimated to a value of 4.36, this indicates heavy tails in the residuals, meaning there is the presence of outliers or extreme values in the series.

In summary, this ARIMA(0,0,1) model has relatively good log likelihood and low Ljung-Box statistic, which indicate a decent fit to the data and no significant autocorrelation. However, the high AIC and BIC values suggest that the model may be overly complex. The low p-value for the JB test and high kurtosis indicate that the residuals are not normally distributed and have heavy tails. The negative skewness suggests a slight leftward skew. Additionally, the low p-value for Prob(H) indicates the presence of constant variance. Further model refinement or exploration may be necessary, especially regarding the non-normality of residuals. In this case other models will be explored instead.

Model Evaluation

With a Mean Squared Error (MSE) of 59,51 for the training set, the average squared deviation between the model's predicted values and the actual values is rather high. An average of 6 units separates the model's predictions from the actual values, according to the Mean Absolute Error (MAE) of 5,65. It's highly unlikely for a model to have zero error unless the actual values are also zero, so this is an unusual occurrence that could indicate a potential calculation or reporting error. The Mean Absolute Percentage Error

(MAPE) of 0,33 indicates almost perfect prediction accuracy. The average magnitude of the error, corrected for its squared units back, is indicated by the Root Mean Squared Error (RMSE), which is 7,71.



METRIC	TRAIN	TEST
MSE	59,51	54,66
MAE	5,65	6,31
MAPE	0,33	0,20
RMSE	7,71	7,39
MPD	3,16	1,86

Table 4.5 – Results of evaluation metrics for NH

With a Mean Squared Error (MSE) of 54,66 for the training set, the average squared deviation between the model's predicted values and the actual values is rather high. An average of 6 units separates the model's predictions from the actual values, according to the Mean Absolute Error (MAE) of 6,31. It's highly unlikely for a model to have zero error unless the actual values are also zero, so this is an unusual occurrence that could indicate a potential calculation or reporting error. The Mean Absolute Percentage Error (MAPE) of 0,20 indicates almost perfect prediction accuracy. The average magnitude of the error, corrected for its squared units back, is indicated by the Root Mean Squared Error (RMSE), which is 7,39.

In a nutshell, the average prediction appears to be 3% off from the actual values, as indicated by the Mean Percentage Difference (MPD) of 3. However, it is important to note that this metric is not standard and requires additional context to fully understand.

After Working Hours

This sub-section will cover the training of ARIMA model in the context of After working hours.

Parameter Estimations

This area of the subsection shows the number of parameters to be included in the fitting of the ARIMA model by uncovering the relevant ARIMA input. The section does that by showing the Autocorrelation function (ACF) and Partial autocorrelation functions (PACF).

Based on the spike by ACF plot, the greatest decline is shown on value if 1 followed by other declines and the PACF plot also shows the same. This suggests that the correlation functions deem more several options as optimal but a value if 1 being the best.

This means the best ARIMA model based on ACF and PACF is ARIMA(0,0,1) and it will be confirmed by later discoveries.

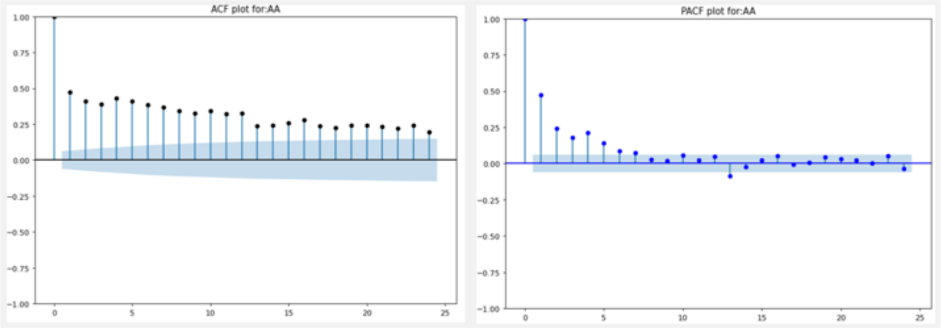
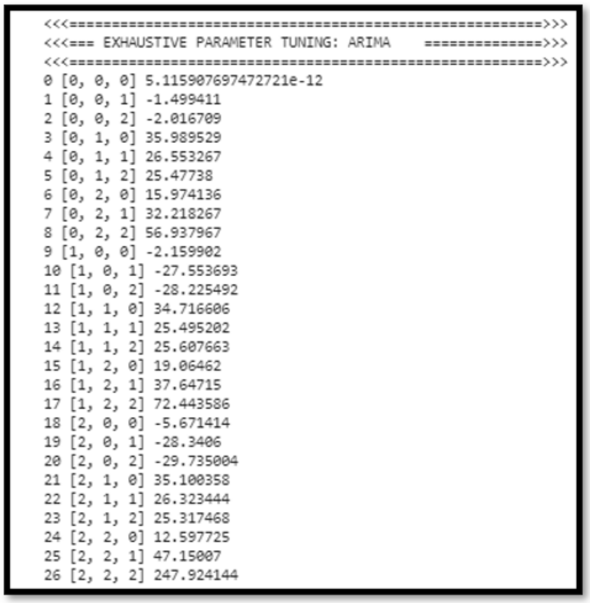


Figure 4.6 – ACF and PACF plots After Hours

Tuning the ARIMA Hyper-Parameters

The area of a sub-section will perform an iterative search for the best parameters.



```
<<<=========================================================>>>
<<<=== EXHAUSTIVE PARAMETER TUNING: ARIMA    ===============>>>
<<<=========================================================>>>
0 [0, 0, 0] 5.115907697472721e-12
1 [0, 0, 1] -1.499411
2 [0, 0, 2] -2.016709
3 [0, 1, 0] 35.989529
4 [0, 1, 1] 26.553267
5 [0, 1, 2] 25.47738
6 [0, 2, 0] 15.974136
7 [0, 2, 1] 32.218267
8 [0, 2, 2] 56.937967
9 [1, 0, 0] -2.159902
10 [1, 0, 1] -27.553693
11 [1, 0, 2] -28.225492
12 [1, 1, 0] 34.716606
13 [1, 1, 1] 25.495202
14 [1, 1, 2] 25.607663
15 [1, 2, 0] 19.06462
16 [1, 2, 1] 37.64715
17 [1, 2, 2] 72.443586
18 [2, 0, 0] -5.671414
19 [2, 0, 1] -28.3406
20 [2, 0, 2] -29.735004
21 [2, 1, 0] 35.100358
22 [2, 1, 1] 26.323444
23 [2, 1, 2] 25.317468
24 [2, 2, 0] 12.597725
25 [2, 2, 1] 47.15007
26 [2, 2, 2] 247.924144
```

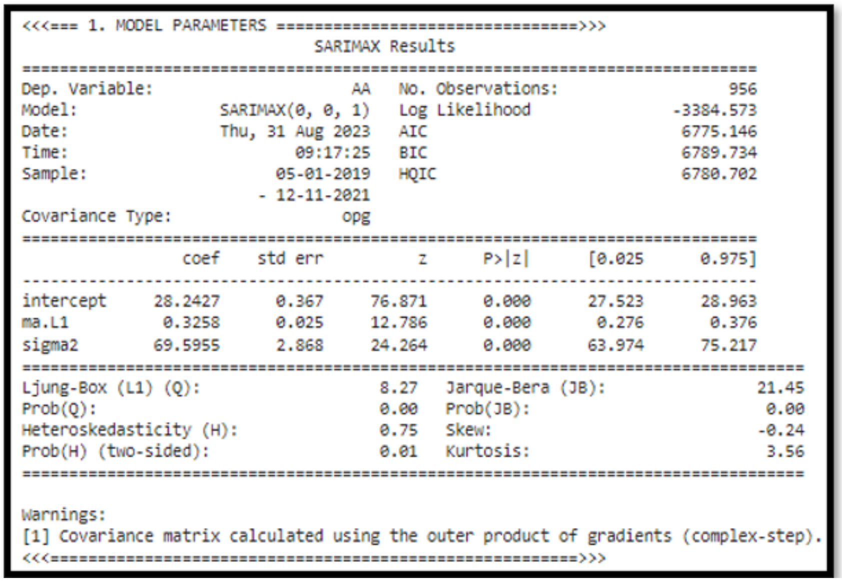
Figure 4.7 – ARIMA Optimal Parameter Searching After Hours

The result below shows that iteration 1 is the best as it has the minimum error. Therefore ARIMA(0,0,1) should be fitted to obtain the best results as it is also supported by ACF and PACF plots.

ARIMA(0,0,1) Parameters – output

Interpreting the various statistics and components of an ARIMA(0,0,1) model is important for understanding the model's goodness of fit and whether it is suitable for the data assessed.

Based on Figure 4.8, the following learning points from the model parameters are highlighted:



```
<<<=== 1. MODEL PARAMETERS =================================>>>
                               SARIMAX Results
==============================================================================
Dep. Variable:                     AA   No. Observations:                  956
Model:               SARIMAX(0, 0, 1)   Log Likelihood               -3384.573
Date:                Thu, 31 Aug 2023   AIC                           6775.146
Time:                        09:17:25   BIC                           6789.734
Sample:                    05-01-2019   HQIC                          6780.702
                         - 12-11-2021
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     28.2427      0.367     76.871      0.000      27.523      28.963
ma.L1          0.3258      0.025     12.786      0.000       0.276       0.376
sigma2        69.5955      2.868     24.264      0.000      63.974      75.217
===================================================================================
Ljung-Box (L1) (Q):                   8.27   Jarque-Bera (JB):                21.45
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.75   Skew:                            -0.24
Prob(H) (two-sided):                  0.01   Kurtosis:                         3.56
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
<<<========================================================>>>
```

Figure 4.8 – ARIMA Model Parameters for After Hours

The log likelihood was estimated to be -3384.573 and as a rule of thumb; a more negative log likelihood indicates a better fit for the model, which means the model performs relatively well.

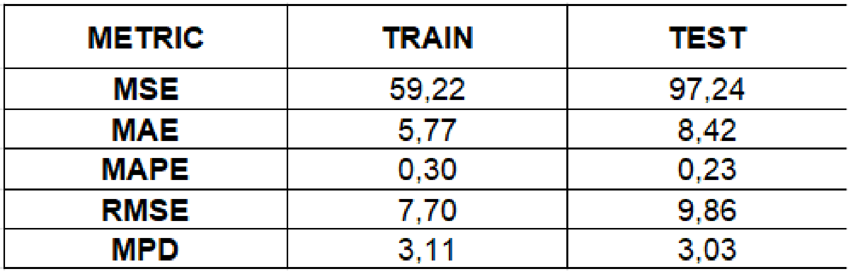
The model has the Akaike Information Criterion (AIC) value of 6775.146, a Bayesian Information Criterion (BIC) value of 6789.734, and a Hannan-Quinn Information Criterion (HQIC) value of 6780.702, which are all meant to measure the trade-off between goodness of fit and the model's complexity but mainly focusing on penalising for complexity. Since all the measured values (AIC, BIC, and HQIC) are relatively high, they suggest that the trade-off was not balanced which results in a highly complex model. The Intercept parameter was estimated to a value of 28.2427, which represents the constant term as an expected value of the time series when lagged values are not considered.

The model has a Moving Average Coefficient (ma.L1) value of 0.3258, which is a coefficient is associated with the moving average term of order 1. And it measures the influence of the most recent error term on the current value which indicates a positive effect in this case. A Residual Variance (Sigma2) of 69.5955 was estimated, which represents the estimated variance of the residuals. The higher the Sigma2 value, the greater the variability of the residuals, which is the case for this model.

Ljung-Box Statistic of 8.27 was estimated, which tests for autocorrelation in the residuals. Due to the high value of this statistic, it suggests that there is some degree of autocorrelation in the residuals. Ljung-Box Test (Prob(Q)) p-value was estimated to be 0.00, which suggests that there is significant autocorrelation in the residuals. Heteroskedasticity was estimated to a value of 0.75, which suggests that there is no strong evidence of heteroskedasticity (changing variance) in the residuals. Prob(H) Two-Sided was estimated to be 0.01, which indicates some evidence of constant variance in the estimated residuals.

The Jarque-Bera statistic (JB Statistic) was estimated to 21.45 by testing for the normality within the residuals, this indicates departures from normality in the residuals. Jarque-Bera Test p-value (Prob(JB)) was estimated to a value of 0.00, due to a low value of this statistics, it then confirms that the residuals are not normally distributed. The Skewness was estimated to a value of -0.24, which is a slight leftward or negative skew in the residuals, this indicates that the distribution is skewed to the left slightly. The Kurtosis was also estimated to a value of 3.56, this indicates heavy tails in the residuals, meaning there is the presence of outliers or extreme values in the series.

In summary, this ARIMA(0,0,1) model has a reasonable log likelihood, indicating a decent fit to the data. However, the AIC, BIC, and HQIC values are relatively high, suggesting potential model complexity. The relatively high Ljung-Box statistic and low p-value (0.00) indicate significant autocorrelation in the residuals, which may suggest that the model is not adequately capturing the temporal patterns in the data. Additionally, the JB statistic and its low p-value (0.00) suggest non-normality in the residuals. Further analysis and model refinement may be needed to address these issues. In this case other models will be explored instead.



METRIC	TRAIN	TEST
MSE	59,22	97,24
MAE	5,77	8,42
MAPE	0,30	0,23
RMSE	7,70	9,86
MPD	3,11	3,03

Table 4.6 – Results of evaluation metrics for AH

The average squared difference between the estimated and actual values is measured by the Mean Squared Error (MSE). The MSE for the training set is 59.22, indicating a reasonably good match between the model and the training data. The model does not perform as well on unseen data, which may be an indication of overfitting, as the MSE for the test set is substantially higher at 97.24. The mean absolute error (MAE) between the expected and actual values is calculated. Compared to MSE, it is less vulnerable to outliers. The MAE rises to 8.42 for the test set from 5.77 for the training set. Once more, this demonstrates a decline in performance from testing to training.

The mistake is expressed as a percentage of the actual values by MAPE. It is particularly helpful when you wish to utilize a more intuitive interpretation of the error sizes. For the training set, the MAPE is 0.30, while for the test set, it is 0.23. It's interesting to note that the MAPE is lower for the test set, which may indicate that even though the mistakes are bigger overall (as indicated by the MSE and MAE), they are reduced in relation to the scale of the predicted data.

The square root of the mean statistical error (RMSE) can occasionally be more easily understood by providing error terms in the same units as the projected result. Similar to MSE, the RMSE rises from 7.70 in the training set to 9.86 in the test set, indicating a decrease in predicted accuracy when the model comes into contact with fresh data. Although the MPD measure is not as standard as the others, it often shows the average percentage difference between the values that were predicted and those that were seen. The MPD, which is 3.03 for testing and 3.11 for training, is comparatively stable between the test and train sets. This implies that both datasets have comparable relative disparities between predictions and actuals.

4.2.2 XGBoost Model Training

This section will cover the training of XGBoost model in the context of NH and AH. XGBoost is a gradient-boosting algorithm that is often used for machine learning tasks such as regression and classification. It is a powerful algorithm that can learn complex relationships in the data.

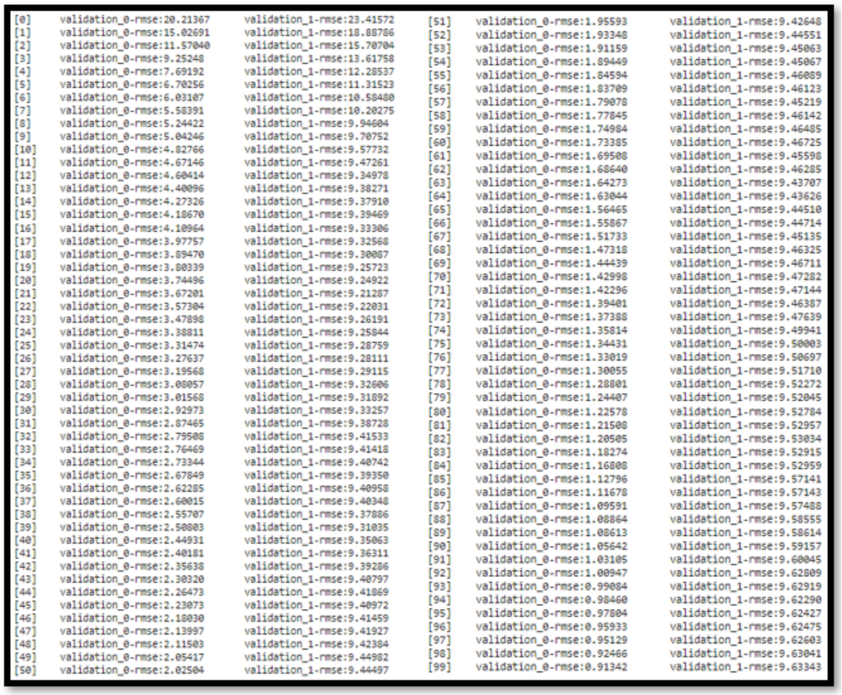
Normal Hours

Tuning an XGBoost Model

"Validation" and "RMSE" (which most likely stands for mean squared error) are terms used in machine learning to describe the process of model validation and the search for optimal parameters.

The output represents a parameter tuning exercise for an XGBoost model.

Each entry in the output is the result of a single model run with a specific set of parameters. The naming convention validation_0-MSE, validation_1-rmse, and so on suggests that we are looking at cross-validation results from different folds or parameter sets.



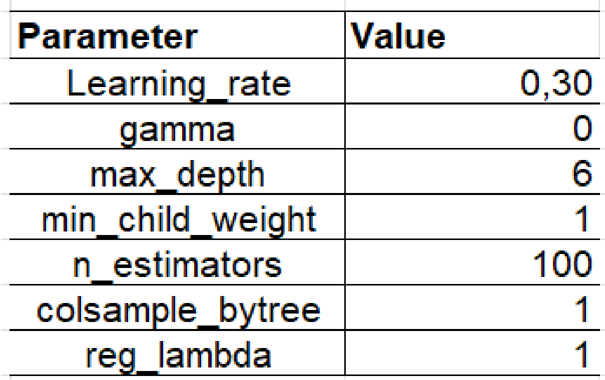
```
[0]     validation_0-rmse:20.21367     validation_1-rmse:23.41572
[1]     validation_0-rmse:15.02691     validation_1-rmse:18.88786
[2]     validation_0-rmse:11.57040     validation_1-rmse:15.70704
[3]     validation_0-rmse:9.25248      validation_1-rmse:13.61758
[4]     validation_0-rmse:7.69192      validation_1-rmse:12.28537
[5]     validation_0-rmse:6.70256      validation_1-rmse:11.31523
[6]     validation_0-rmse:6.03107      validation_1-rmse:10.58480
[7]     validation_0-rmse:5.58391      validation_1-rmse:10.20275
[8]     validation_0-rmse:5.24422      validation_1-rmse:9.94604
[9]     validation_0-rmse:5.04246      validation_1-rmse:9.70752
[10]    validation_0-rmse:4.82766      validation_1-rmse:9.57732
[11]    validation_0-rmse:4.67146      validation_1-rmse:9.47261
[12]    validation_0-rmse:4.60414      validation_1-rmse:9.34978
[13]    validation_0-rmse:4.40096      validation_1-rmse:9.38271
[14]    validation_0-rmse:4.27326      validation_1-rmse:9.37910
[15]    validation_0-rmse:4.18670      validation_1-rmse:9.39469
[16]    validation_0-rmse:4.10964      validation_1-rmse:9.33306
[17]    validation_0-rmse:3.97757      validation_1-rmse:9.32568
[18]    validation_0-rmse:3.89470      validation_1-rmse:9.30087
[19]    validation_0-rmse:3.80339      validation_1-rmse:9.25723
[20]    validation_0-rmse:3.74496      validation_1-rmse:9.24922
[21]    validation_0-rmse:3.67201      validation_1-rmse:9.21287
[22]    validation_0-rmse:3.57304      validation_1-rmse:9.22031
[23]    validation_0-rmse:3.47898      validation_1-rmse:9.26191
[24]    validation_0-rmse:3.38811      validation_1-rmse:9.25844
[25]    validation_0-rmse:3.31474      validation_1-rmse:9.28759
[26]    validation_0-rmse:3.27637      validation_1-rmse:9.28111
[27]    validation_0-rmse:3.19568      validation_1-rmse:9.29115
[28]    validation_0-rmse:3.08057      validation_1-rmse:9.32606
[29]    validation_0-rmse:3.01568      validation_1-rmse:9.31892
[30]    validation_0-rmse:2.92973      validation_1-rmse:9.33257
[31]    validation_0-rmse:2.87465      validation_1-rmse:9.38728
[32]    validation_0-rmse:2.79508      validation_1-rmse:9.41533
[33]    validation_0-rmse:2.76469      validation_1-rmse:9.41418
[34]    validation_0-rmse:2.73344      validation_1-rmse:9.40742
[35]    validation_0-rmse:2.67849      validation_1-rmse:9.39350
[36]    validation_0-rmse:2.62285      validation_1-rmse:9.40958
[37]    validation_0-rmse:2.60015      validation_1-rmse:9.40348
[38]    validation_0-rmse:2.55707      validation_1-rmse:9.37886
[39]    validation_0-rmse:2.50803      validation_1-rmse:9.31035
[40]    validation_0-rmse:2.44931      validation_1-rmse:9.35063
[41]    validation_0-rmse:2.40181      validation_1-rmse:9.36311
[42]    validation_0-rmse:2.35638      validation_1-rmse:9.39286
[43]    validation_0-rmse:2.30320      validation_1-rmse:9.40797
[44]    validation_0-rmse:2.26473      validation_1-rmse:9.41869
[45]    validation_0-rmse:2.23073      validation_1-rmse:9.40972
[46]    validation_0-rmse:2.18030      validation_1-rmse:9.41459
[47]    validation_0-rmse:2.13997      validation_1-rmse:9.41927
[48]    validation_0-rmse:2.11503      validation_1-rmse:9.42384
[49]    validation_0-rmse:2.05417      validation_1-rmse:9.44982
[50]    validation_0-rmse:2.02504      validation_1-rmse:9.44497
[51]    validation_0-rmse:1.95593      validation_1-rmse:9.42648
[52]    validation_0-rmse:1.93348      validation_1-rmse:9.44551
[53]    validation_0-rmse:1.91159      validation_1-rmse:9.45063
[54]    validation_0-rmse:1.89449      validation_1-rmse:9.45067
[55]    validation_0-rmse:1.84594      validation_1-rmse:9.46089
[56]    validation_0-rmse:1.83709      validation_1-rmse:9.46123
[57]    validation_0-rmse:1.79078      validation_1-rmse:9.45219
[58]    validation_0-rmse:1.77845      validation_1-rmse:9.46142
[59]    validation_0-rmse:1.74984      validation_1-rmse:9.46485
[60]    validation_0-rmse:1.73385      validation_1-rmse:9.46725
[61]    validation_0-rmse:1.69508      validation_1-rmse:9.45598
[62]    validation_0-rmse:1.68640      validation_1-rmse:9.46285
[63]    validation_0-rmse:1.64273      validation_1-rmse:9.43707
[64]    validation_0-rmse:1.63044      validation_1-rmse:9.43626
[65]    validation_0-rmse:1.56465      validation_1-rmse:9.44510
[66]    validation_0-rmse:1.55867      validation_1-rmse:9.44714
[67]    validation_0-rmse:1.51733      validation_1-rmse:9.45135
[68]    validation_0-rmse:1.47318      validation_1-rmse:9.46325
[69]    validation_0-rmse:1.44439      validation_1-rmse:9.46711
[70]    validation_0-rmse:1.42998      validation_1-rmse:9.47282
[71]    validation_0-rmse:1.42296      validation_1-rmse:9.47144
[72]    validation_0-rmse:1.39401      validation_1-rmse:9.46387
[73]    validation_0-rmse:1.37388      validation_1-rmse:9.47639
[74]    validation_0-rmse:1.35814      validation_1-rmse:9.49941
[75]    validation_0-rmse:1.34431      validation_1-rmse:9.50003
[76]    validation_0-rmse:1.33019      validation_1-rmse:9.50697
[77]    validation_0-rmse:1.30055      validation_1-rmse:9.51710
[78]    validation_0-rmse:1.28801      validation_1-rmse:9.52272
[79]    validation_0-rmse:1.24407      validation_1-rmse:9.52045
[80]    validation_0-rmse:1.22578      validation_1-rmse:9.52784
[81]    validation_0-rmse:1.21508      validation_1-rmse:9.52957
[82]    validation_0-rmse:1.20505      validation_1-rmse:9.53034
[83]    validation_0-rmse:1.18274      validation_1-rmse:9.52915
[84]    validation_0-rmse:1.16808      validation_1-rmse:9.52959
[85]    validation_0-rmse:1.12796      validation_1-rmse:9.57141
[86]    validation_0-rmse:1.11678      validation_1-rmse:9.57143
[87]    validation_0-rmse:1.09591      validation_1-rmse:9.57488
[88]    validation_0-rmse:1.08864      validation_1-rmse:9.58555
[89]    validation_0-rmse:1.08613      validation_1-rmse:9.58614
[90]    validation_0-rmse:1.05642      validation_1-rmse:9.59157
[91]    validation_0-rmse:1.03105      validation_1-rmse:9.60045
[92]    validation_0-rmse:1.00947      validation_1-rmse:9.62809
[93]    validation_0-rmse:0.99084      validation_1-rmse:9.62919
[94]    validation_0-rmse:0.98460      validation_1-rmse:9.62290
[95]    validation_0-rmse:0.97804      validation_1-rmse:9.62427
[96]    validation_0-rmse:0.95933      validation_1-rmse:9.62475
[97]    validation_0-rmse:0.95129      validation_1-rmse:9.62603
[98]    validation_0-rmse:0.92466      validation_1-rmse:9.63041
[99]    validation_0-rmse:0.91342      validation_1-rmse:9.63343
```

Figure 4.9 – XGBoost Optimal Parameter Searching for Normal Hours

Validation_0, Validation_1, and so on are different validation folds or iterations. The metric used to evaluate the model's performance is denoted by RMSE. The numerical value (for example,.21367) represents the RMSE for that specific run.

Parameter Estimations

The output lists various XGBoost model configuration parameters, but all of them are shown with their default values or set to 'None,' indicating that no specific value has been assigned to them in this output. Machine learning models such as XGBoost include a set of default parameters that are a good starting point for many datasets and problems. The default settings can provide a solid baseline and may be sufficient for initial model development.



Parameter	Value
Learning_rate	0,30
gamma	0
max_depth	6
min_child_weight	1
n_estimators	100
colsample_bytree	1
reg_lambda	1

Table 4.7 – Parameters used in XGBoost for NH.

With no further regularization on the tree leaves (gamma = 0), the XGBoost model for predicting outcomes during Normal Hours at an emergency department uses a relatively aggressive learning rate of 0.30, allowing for quick learning. It has a minimum child weight of 1 to support fine-grained data splits and a balanced tree depth of 6 to capture intricate patterns without overfitting. To prevent overfitting by penalizing large weights, the model builds 100 trees utilizing all of the features that are available for each tree (colsample_bytree = 1) and incorporates an L2 regularization (reg_lambda = 1) to create a resilient model that performs well when applied to new data.

These parameters are chosen to develop a predictive model that is complex enough to capture the intricate patterns in the data without overfitting.

The lack of custom values suggests that the model may not be fully optimized for the dataset's unique characteristics concerning daily patient arrivals at an emergency department.

Model Validation

The XGBoost model's performance is compared in the graphs below; one shows in-sample predictions, while the other shows out-of-sample predictions. The values predicted by the XGBoost model are shown by the ”Prediction” line, although the ”NH” line appears to match the actual values.

We usually examine the degree to which the predictions agree with the actual values to evaluate the performance of the fitted XGBoost model for the in-sample and out-of-sample datasets.

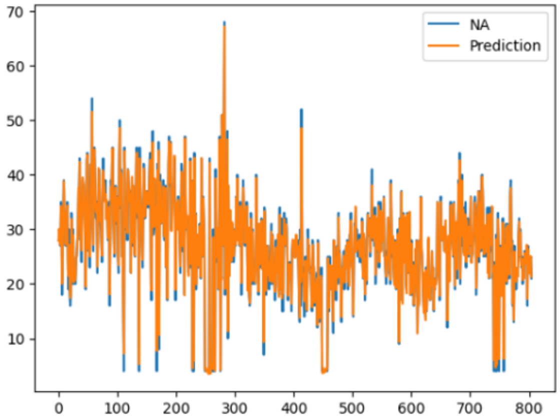


Figure 4.10 – In sample Testing for Normal Hours

Because the model has been trained on the in-sample data, it is anticipated to produce superior predictions for that set of data. With a few notable exceptions, particularly in regions where the real values show spikes, the forecasts in the first image mostly correspond with the patterns of the actual values. This is common behavior for many predictive models, which may have trouble with abrupt changes or noise in the training set. The model appears to smooth out the spikes.

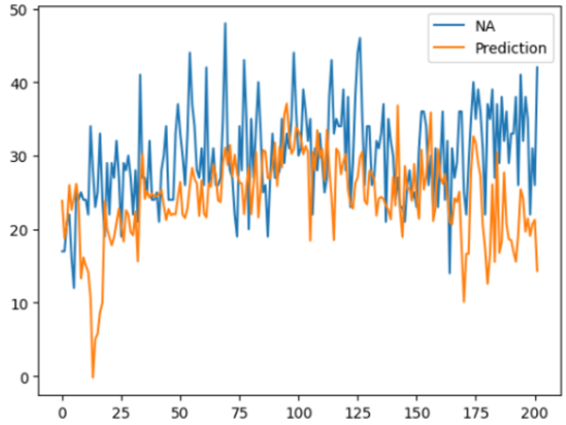
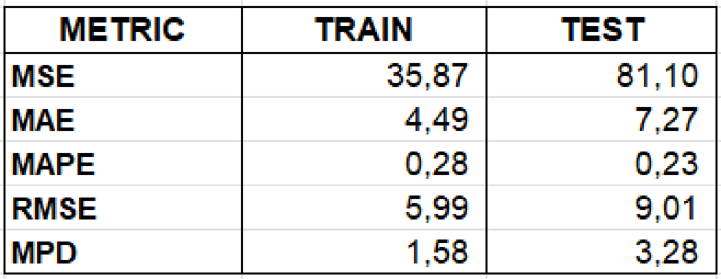


Figure 4.11 – Out of sample Testing for Normal Hours

The model may not be as good at generalizing to new data, as evidenced by the more noticeable difference between the out-of-sample predictions and the actual results. This could be because the model has not learned the new characteristics of the out-of-sample data, or it could be the result of overfitting to the in-sample data.

Model Evaluation



METRIC	TRAIN	TEST
MSE	35,87	81,10
MAE	4,49	7,27
MAPE	0,28	0,23
RMSE	5,99	9,01
MPD	1,58	3,28

Table 4.8 – Results of evaluation metrics for NH

The evaluation metrics hint that there may be overfitting because the model's performance on the training data does not translate well to the testing data. The model's Mean Squared Error (MSE) rises to 81.10 on testing data, despite achieving a lower MSE of 35.87 on training data. This indicates bigger average squared errors in predictions for unknown data. The model's predictions were off by an average of 4.49 on the training data and by a more significant 7.27 on the testing data, according to the Mean Absolute Error (MAE), which displays a similar trend.

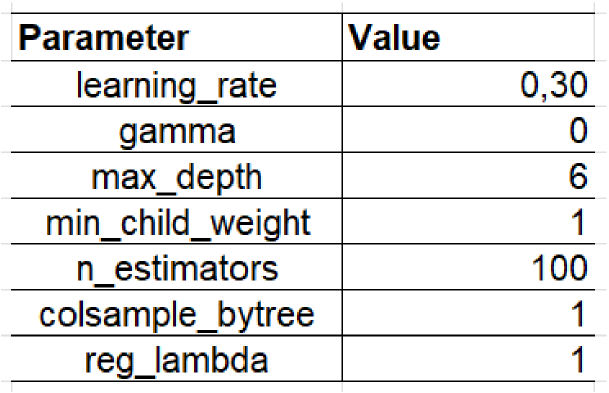
The testing data shows an interesting improvement in the Mean Absolute Percentage Error (MAPE), which drops from 0.28 to 0.23. This suggests that the relative percentage of errors is a little lower on the test set. As with the other measures, the Root Mean Squared Error (RMSE) shows higher average mistakes in the model's predictions, with a reasonable value of 5.99 on the training data and an increase of 9.01 on the testing data. Finally, when switching from training to testing, the MPD measure rises from 1.58 to 3.28, indicating once more a decline in predicting accuracy on fresh data.

Overall, these metrics suggest that while the model seems to fit the training data reasonably well, its predictive power diminishes when applied to the test data, which is a common sign of overfitting in machine learning models.

After Hours

This area of the subsection will cover the training of the XGBoost Model in the context of After working hours.

Parameter Estimations



Parameter	Value
learning_rate	0,30
gamma	0
max_depth	6
min_child_weight	1
n_estimators	100
colsample_bytree	1
reg_lambda	1

Table 4.9 – Parameters used in XGBoost for AH.

The XGBoost model employs a learning rate of 0.30 for quick learning in the emergency room of a hospital after hours. This is highly proactive, but it runs the risk of overfitting if rigorous cross-validation isn't done. Since the gamma value is set to 0, no further regularization is applied to the tree leaves, allowing the model to develop unhindered if it produces a smaller training loss. Because the maximum depth is limited to 6, the model can learn intricate patterns without getting overly dependent on the training set. The model can generate children nodes with a finer granularity when min_child_weight is set to 1, which is helpful for subtle patterns that might emerge in the After Hours scenario.

With 100 estimators produced by the model, there will be a sizable but manageable ensemble size for prediction. When splits are made, all features are taken into account (colsample_bytree is at its maximum of 1), which could be significant if all features have the potential to hold essential information for the After Hours scenario. Finally, a modest level of L2 regularization is applied with a reg_lambda of 1, penalizing big weights in the model to promote generalization to fresh data. These settings are selected to maintain a model that performs well in novel, unseen After Hours scenarios while also learning intricate data patterns.

Model Validation

The performance of a fitted XGBoost model is displayed in the graphs below, which compare the actual and predicted values for the in-sample and out-of-sample data sets after hours. The phrase ”after hours” usually describes data points or occurrences that take place outside of a market's or environment's regular business hours and frequently show distinct patterns or behaviors from regular hours.

After Hours forecasts, both in-sample and out-of-sample, demonstrate the model's ability to partially mirror the trend of the real data.

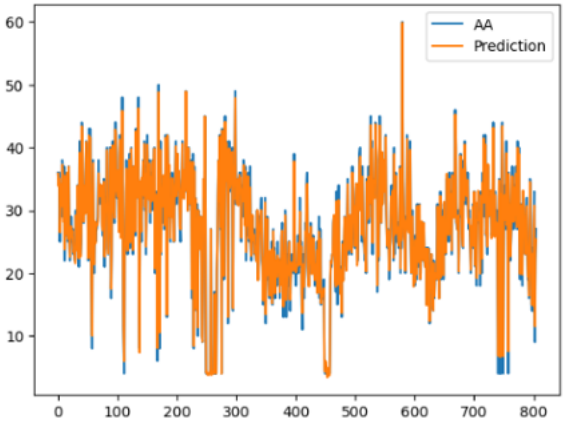


Figure 4.12 – XGBoost In sample Testing for Normal Hours

Since this data was used to train the model, it is not surprising that the in-sample predictions are more accurate. The fact that the prediction line closely tracks the actual line suggests that the model has done a good job of learning the in-sample data.

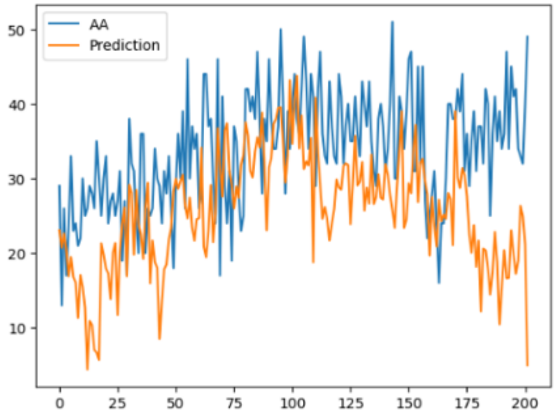
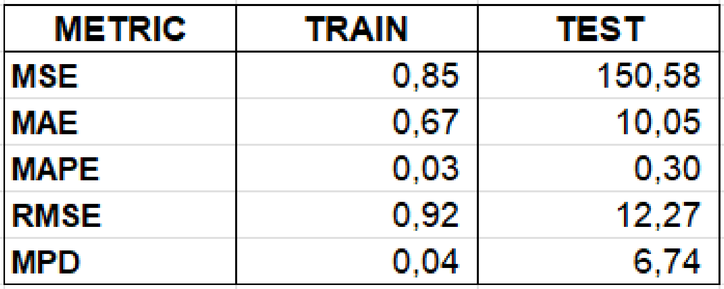


Figure 4.13 – XGBoost Out of sample Testing for Normal Hours

The out-of-sample predictions have greater discrepancies, suggesting potential overfitting to in-sample data or a lack of generalization to new data patterns occurring after hours.

Model Evaluation



METRIC	TRAIN	TEST
MSE	0,85	150,58
MAE	0,67	10,05
MAPE	0,03	0,30
RMSE	0,92	12,27
MPD	0,04	6,74

Table 4.10 – Results of evaluation metrics for AH

An emergency department's After Hours data was used to evaluate the XGBoost model, and the assessment metrics show a notable difference in performance between the training and testing datasets. With a mean squared error (MSE) of 0.85, mean absolute error (MAE) of 0.67, mean absolute percentage error (MAPE) of 0.03, root mean square error (RMSE) of 0.92, and mean percentage difference (MPD) of 0.04 during training, the model performs admirably.

These low values imply a close match between the model and the training set. All error metrics, however, are significantly higher when applied to the test data (MSE of 150.58, MAE of 10.05, MAPE of 0.30, RMSE of 12.27, MPD of 6.74), suggesting that the model may have overfitted to the training set and is not well suited to generalizing to new data. A major challenge for the model's practical use in a hospital context is that its predictions become far less accurate when confronted with After Hours data that it has never seen before.

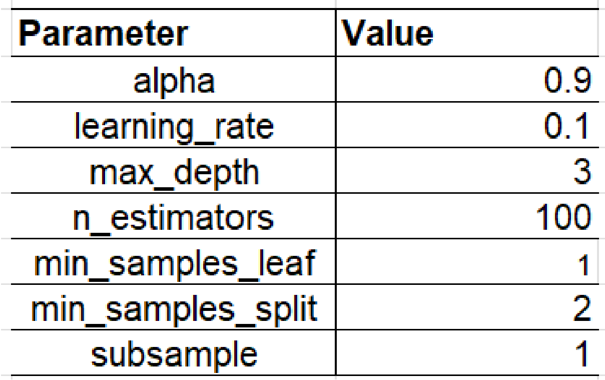
4.2.3 Gradient Boosting Regressor Model

This section will cover the training of the Gradient Boosting Regressor Model in the context of NA and AH . A Gradient Boosting Regressor is a Machine Learning algorithm that is used to predict a continuous value in regression tasks.

Normal Hours

Table 4.11 represents the hyperparameters for a Gradient Boosting Regressor model. These parameters govern how the model learns from data.

Model Parameter Estimation



Parameter	Value
alpha	0.9
learning_rate	0.1
max_depth	3
n_estimators	100
min_samples_leaf	1
min_samples_split	2
subsample	1

Table 4.11 – Parameters used in GBR for NH.

The following parameter values are used for the Gradient Boosting Regressor (GBR) model customized for Normal Hours at an emergency department: A model that uses the 'quantile' loss function, which is normally intended to forecast higher values, is said to concentrate on capturing the 90th percentile of the data distribution, according to an alpha value of 0.9. The model learns from the data at a moderate rate when the learning rate is set to 0.1, which balances the learning rate with the danger of overfitting. In order to avoid over-complexity in individual trees and to facilitate the generalization of the model to new data, the maximum depth of trees is set to 3. The model can improve its accuracy via a significant number of iterations with 100 estimators.

The model can construct splits and generate leaf nodes even with very few data points, possibly capturing complex data patterns but also running the risk of overfitting from noise. A minimum sample leaf of 1 and minimum sample split of 2 indicate this. Lastly, if the model's complexity is not managed in any other way, a subsample value of 1 means that every tree in the model is trained using all of the data, not just a part of it. This can result in more reliable predictions, but it also runs the risk of overfitting. Together, these parameters are selected such that the GBR model may learn intricate patterns during emergency department regular business hours without being unduly sensitive to noise or unique characteristics of the training set.

Model Validation

The graphs below compare in-sample and out-of-sample performance during regular business hours, displaying real versus predicted values for the Gradient Boosting Regressor model.

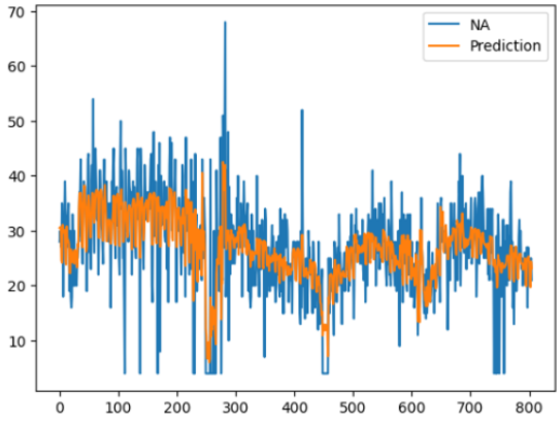


Figure 4.14 – GBR In sample Testing for Normal Hours

The model fits the in-sample data well, as seen by how well the predictions match the actual values. Both the real and anticipated lines exhibit discernible variability; the forecasts, though somewhat smoothed at the peaks and troughs, largely follow the same pattern as the actual values. This implies that the model has successfully acquired the in-sample data patterns during regular business hours.

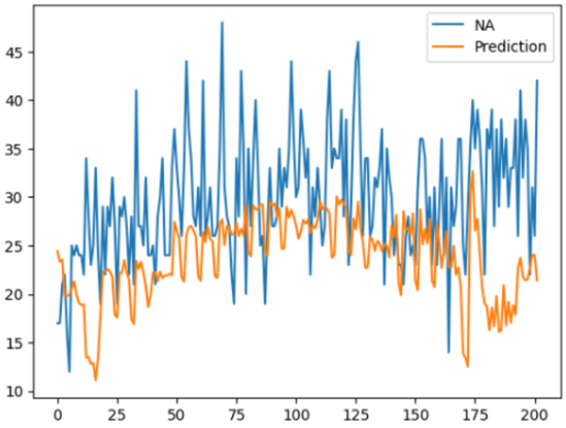
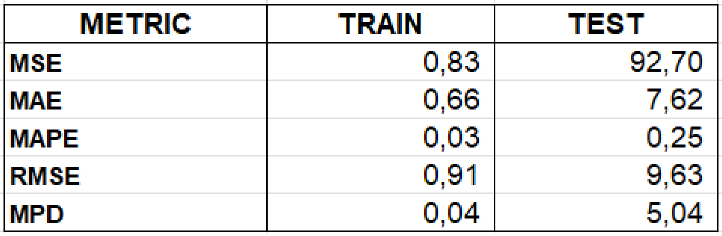


Figure 4.15 – GBR Out of sample Testing for NH

The accuracy of the out-of-sample predictions significantly declines, particularly as the plot nears its conclusion, which may be a sign of problems with the model's generalization ability.

Model Evaluation

When the Gradient Boosting Regressor model is applied to Normal Hours at an emergency department, the evaluation metrics reveal a significant discrepancy between the model's performance on the training and test sets. The model produces a good fit with very low errors on the training data, as seen by the following metrics: Mean Squared Error (MSE) of 0.83, Mean Absolute Error (MAE) of 0.66, Mean Absolute Percentage Error (MAPE) of 0.03, and Root Mean Squared Error (RMSE) of 0.91. A very accurate training performance is further supported by the Mean Percentage Difference (MPD) of 0.04, which is caused by overfitting or a shift in distribution between the training and testing datasets..



METRIC	TRAIN	TEST
MSE	0,83	92,70
MAE	0,66	7,62
MAPE	0,03	0,25
RMSE	0,91	9,63
MPD	0,04	5,04

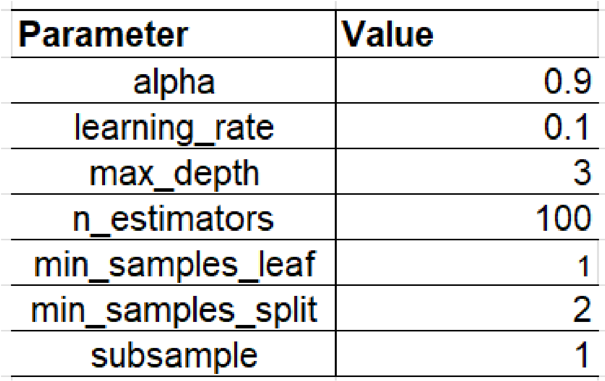
Table 4.12 – Results of evaluation metrics for NH

All error metrics, on the other hand, dramatically rise when the model is applied to the test set, suggesting a less accurate model performance on unseen data: the MSE soars to 92.70, the MAE to 7.62, the MAPE to 0.25, the RMSE to 9.63, and the MPD to 5.04. The model may have overfitted or experienced a shift in distribution between the training and testing datasets, as indicated by these elevated error values on the test set, despite the fact that the model has learned the training data well.

After Hours

The configuration parameters of a Gradient Boosting Regressor model designed to forecast patient arrivals at an emergency department after hours are shown in Table 4,39 output below:

Model Parameter Estimation



Parameter	Value
alpha	0.9
learning_rate	0.1
max_depth	3
n_estimators	100
min_samples_leaf	1
min_samples_split	2
subsample	1

Table 4.13 – Parameters used in GBR for AH.

The following parameter settings apply to the Gradient Boosting Regressor (GBR) customized for an emergency department’s after-hours data: If a quantile loss function were employed, aiming for higher-end predictions, a ”alpha” of 0.9 suggests the quantile to which the model may assign greater weight; however, since the loss is ”squared_error,” this is irrelevant. To balance accuracy and generalization without rapidly overfitting, the model updates its findings at a modest pace, which is ensured by a ’learning_rate’ of 0.1. The’max_depth’ parameter, which is set to 3, helps keep the trees from being overly complicated, which helps prevent overfitting and guarantees that the model can successfully generalize to new, untested data.

The model has a good amount of trees to deal with when ’n_estimators’ is set to

100. These trees can capture intricate patterns without being overly computationally demanding. To capture the subtleties and unpredictability typical of After Hours circumstances, the model must be able to construct leaves and splits even with little quantities of data, which can render the model vulnerable to noise.'min_samples_leaf' at 1 and'min_samples_split' at 2 support this capacity. Last but not least, setting 'subsample' to 1 guarantees that every tree is trained using the entire dataset, giving every tree the most data possible and maybe producing a stronger model.

These parameters provide a balance between capturing the specifics in the data and creating a well-generalizable model. They are intended to simulate the unpredictable and perhaps sparse data patterns that arise during the After Hours operations of an emergency department.

Model Validation

The graphs, one for in-sample data sets and the other for out-of-sample data sets, display the actual values and the anticipated values for after-hours data using a Gradient Boosting Regressor model.

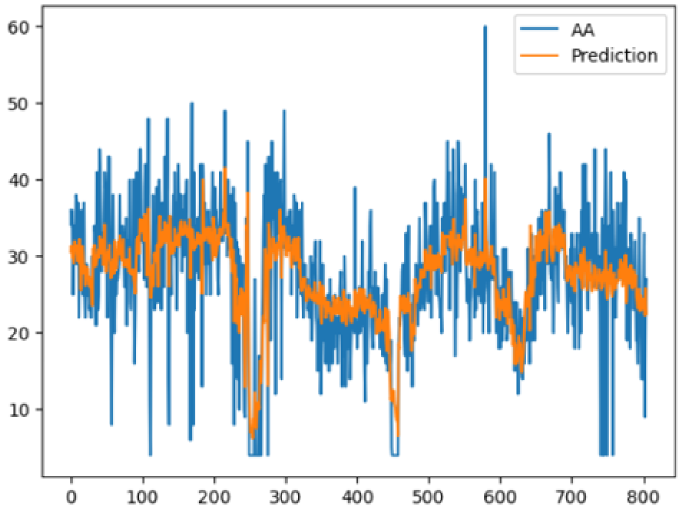


Figure 4.16 – In sample Testing for After Hours

The prediction line seems to smooth down the extremes, but it still closely tracks the actual value line. When the model is trained on this particular set of data, this is typical behavior for in-sample predictions.

The larger difference between the expected and actual values implies that the model's capacity to generalize to new data may be restricted, as indicated by the out-of-sample predictions.

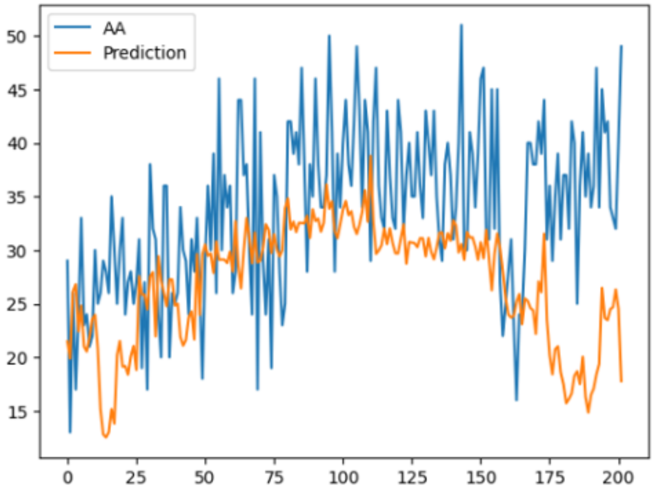
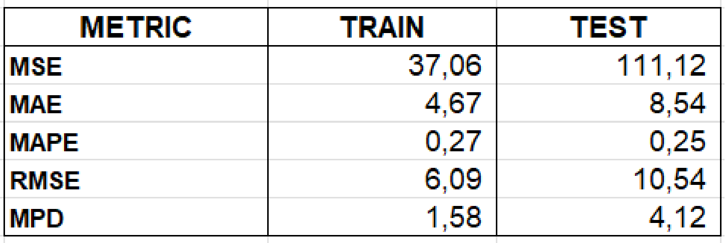


Figure 4.17 – Out of sample Testing of After Hours

As anticipated, the model performs better for after-hours predictions on the training set of data, but its performance noticeably declines when predicting fresh data.

Model Evaluation



METRIC	TRAIN	TEST
MSE	37,06	111,12
MAE	4,67	8,54
MAPE	0,27	0,25
RMSE	6,09	10,54
MPD	1,58	4,12

Table 4.14 – Results of evaluation metrics for NH

An emergency department's After Hours data evaluation metrics point to a model that does well on the training set but noticeably worse on the test set. The Mean Squared Error (MSE) of the training data is 37.06, which is a pretty low figure. However, on the test set, this value rises substantially to 111.12, indicating that the model's predictions are much less accurate on data that it hasn't seen before. Larger average mistakes in the model's predictions on the test data are indicated by the Mean Absolute Error (MAE), which more than doubles from 4.67 in training to 8.54 in testing.

Interestingly, the Mean Absolute Percentage Error (MAPE) drops from 0.27 to 0.25, suggesting that the absolute errors may not accurately reflect the relative amount of the errors relative to the actual values. Nonetheless, there is an increase in the Root Mean Squared Error (RMSE) from 6.09 to 10.54, indicating a higher degree of unpredictability in the test predictions. Additionally, there is a greater average percentage difference in the predictions made during After Hours, as indicated by the Mean Percentage Difference (MPD), which rises from 1.58 to 4.12. When taken as a whole, these indicators point to the model being overfit to the training set and maybe underfitting to the more unpredictable situations that the emergency department experiences after hours.

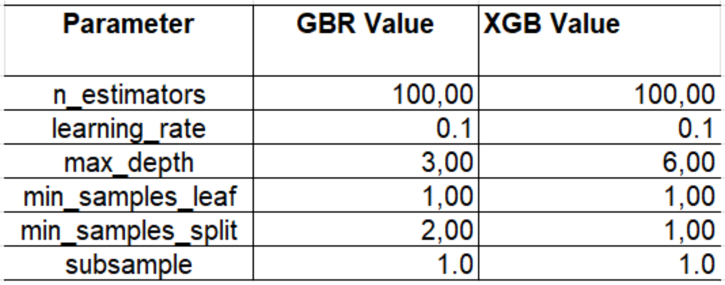
4.2.4 Voting Regressor Model

This section will cover the training of the Voting Regressor model, starting with the parameter optimisation in the context of Normal Hours and After Hours. A Voting Regressor is a Machine Learning model that uses a voting mechanism to combine predictions from multiple regression models. It is a type of ensemble learning, which is a technique for improving predictive performance by combining multiple models.

Normal Working Hours

The configuration details the parameters of the Voting Regressor, which includes three different regressor models:

VR Model Parameters



Parameter	GBR Value	XGB Value
n_estimators	100,00	100,00
learning_rate	0.1	0.1
max_depth	3,00	6,00
min_samples_leaf	1,00	1,00
min_samples_split	2,00	1,00
subsample	1.0	1.0

Table 4.15 – Parameters used in VR for NH.

The Voting Regressor for Normal Hours at an emergency department combines predictions from both a Gradient Boosting Regressor (GBR) and an XGBoost Regressor

(XGBoost), using the following parameter settings. The fact that both models are set up with 100 estimators means that they each employ 100 trees in their ensemble, striking a compromise between computational efficiency and prediction accuracy. Both models have a learning rate of 0.1, which is a reasonable speed that permits sufficient model modifications without running the risk of overfitting too rapidly.

In terms of tree complexity, the XGBoost is given more leeway with a max depth of 6, possibly catching more subtle patterns at the cost of fitting to noise, while the GBR is limited to a cautious max depth of 3, which helps prevent overfitting. Small leaf nodes and fine-grained data segmentation are made possible by the minimum sample leaf of 1 used by both models, which might be crucial in a changing environment such as an emergency room. In contrast to the XGBoost, which is set to default and usually requires a min_child_weight of 1 to split, the GBR requires at least 2 samples to split a node, which can assist reduce overfitting and perhaps provide a more sophisticated model.

Last but not least, both models have a subsample rate of 1.0, meaning that they learn from all of the available data points. This can result in more accurate predictions, but it may also introduce noise into the training process. The combination of models suggests a robust approach to capturing various patterns and relationships within data, which can be critical in a high variance scenario such as emergency department arrivals.

Model Validation

The line plots in the graphs below show actual values vs values predicted by a Voting Regressor model for two datasets: an out-of-sample set and an in-sample set. The phrase ”Normal Hours” denotes that the information is limited to a particular period of time, regular business or operational hours.

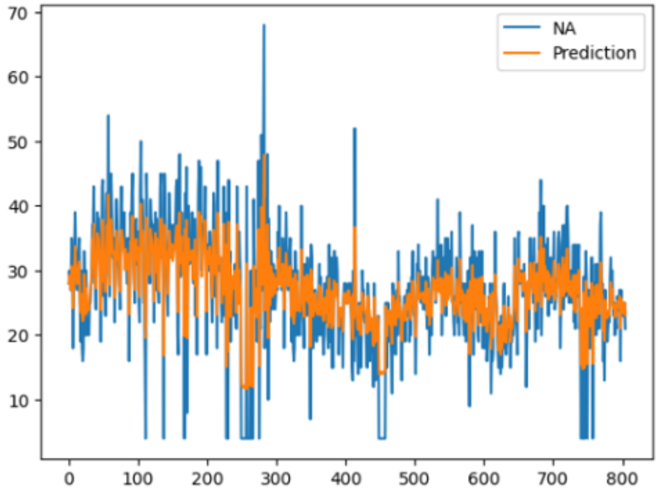


Figure 4.18 – GBR In sample Testing for Normal Hours

The orange prediction line and the blue actual line overlap, indicating that the predicted values are closely match the actual values. The model appears to do a good job of capturing the general trend, despite a few aberrations, particularly when the real data have peaks.

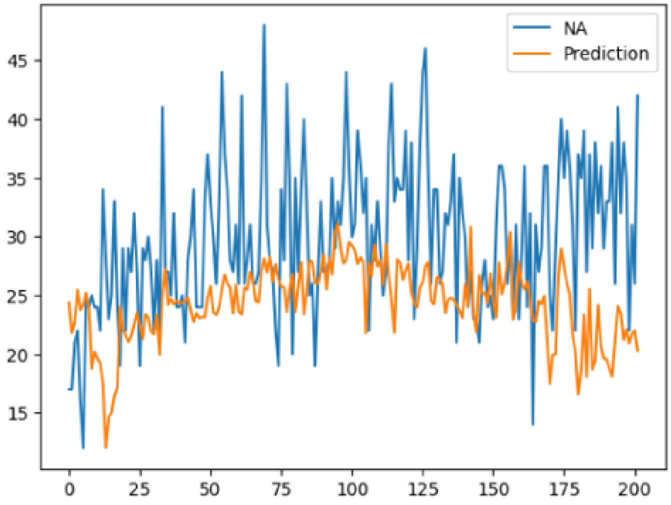
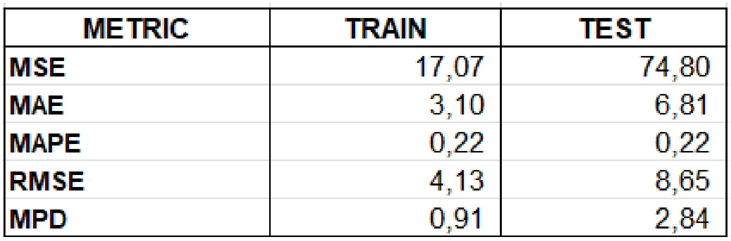


Figure 4.19 – GBR Out of sample Testing for Normal Hours

Although there are more noticeable variations, the projected values in this collection mostly match the actual values' general trend. This is to be expected as models typically outperform out-of-sample data when applied to in-sample data. Although the fit is not as tight as it is with the in-sample data, the model here reflects the trend. Larger differences between the actual and anticipated values suggest that, according to conventional predictive modeling assumptions, the model may not generalize as well to new data.

The smoother prediction line compared to the actual values suggests that the Voting Regressor may be averaging out the predictions of the individual models, leading to less sensitivity to rapid changes in the data

Model Evaluation



METRIC	TRAIN	TEST
MSE	17,07	74,80
MAE	3,10	6,81
MAPE	0,22	0,22
RMSE	4,13	8,65
MPD	0,91	2,84

Table 4.16 – Results of evaluation metrics for NH

A model with strong training performance but a noticeable decline in performance on test data is revealed by the assessment metrics for the Voting Regressor applied to Normal Hours at an emergency department. The test set's Mean Squared Error (MSE) rises to 74.80, indicating more differences in predictions for unknown data, whereas the training set's MSE of 17.07 indicates that the model's predictions are generally near the actual values. In addition, the Mean Absolute Error (MAE) rises from 3.10 in training to 6.81 in testing, indicating that the model performs better on the test set in terms of average prediction error.

For both the training and test sets, the Mean Absolute Percentage Error (MAPE) stays constant at 0.22, suggesting that the proportionate prediction errors in relation to the actual values are constant throughout the two sets. Again reflecting bigger errors on the test set, the Root Mean Squared Error (RMSE), which indicates the size of the error, increases from 4.13 in training to 8.65 in testing. Finally, the model's predictions appear to diverge further from the actuals during the test phase, as evidenced by the Mean Percentage Difference (MPD), which increases from 0.91 during training to 2.84 on the test set.

Overall, these metrics suggest that, despite the Voting Regressor's ability to capture the features of the training data, overfitting or variations in the distribution of the data between the training and test sets may be the cause of its inability to generalize performance to the test data.

After Working Hours

This section will cover the training of the Voting Regressor model, looking at the model validation and selection in the context of After working hours.

Model Validation

The performance of a Voting Regressor model for the After Hours dataset—data collected after regular business hours—is depicted in the graphs below.

The predictions of the in-sample set model seem to match the real values very well. The model appears to be successfully capturing the variability in the data based on the overlap between the forecast and actual value lines. The model's smoothing impact on the data or possible underfitting are indicated in certain areas, nevertheless, when the actual values exhibit substantial variability that the predictions do not fully match.

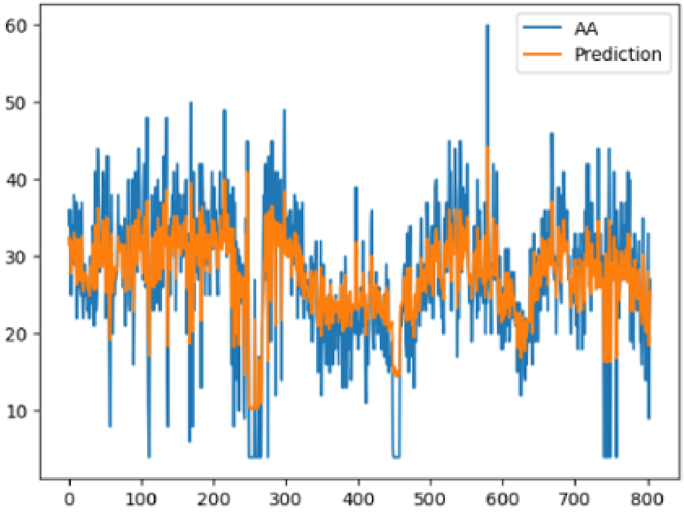


Figure 4.21 – GBR In sample Testing for After Hours

When compared to the in-sample set, the out-of-sample performance shows a larger difference between the anticipated and actual values. This makes sense because models tend to perform better on the training set of data. The model may not generalize as well to new or unseen data after hours, even while the predictions still roughly follow the trend of the actual values. Additionally, the forecasts seem to be less sensitive to the highs and lows in the real data and more consistent.

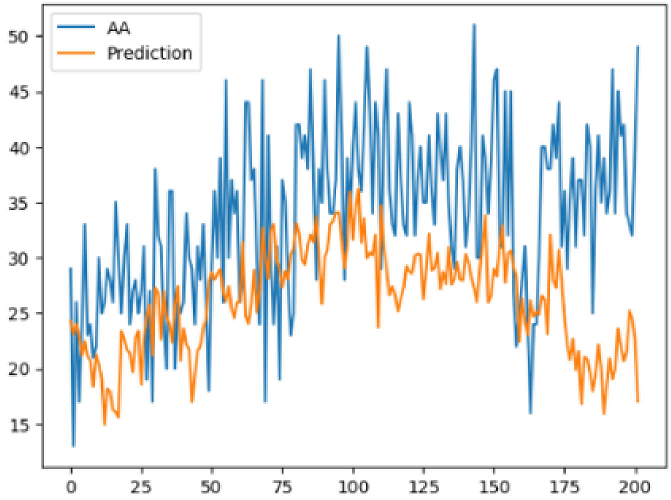
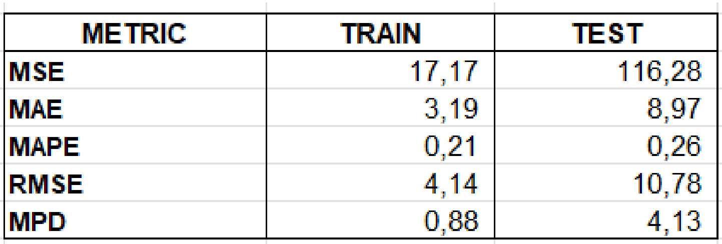


Figure 4.22 – GBR Out of sample Testing for After Hours

Model Evaluation

A Voting Regressor at an emergency department with after-hours performance is evaluated using metrics that reveal a model that performs well on training data but degrades on test data. In particular, the test set's Mean Squared Error (MSE) rises significantly to 116.28, indicating that the model's predictions are significantly less accurate on unknown data. The training set's MSE, however, is 17.17, indicating that the model's predictions are pretty near to the true values.

As a result of greater average deviations from the actual values in the test data, Absolute Error (MAE) increases from 3.19 on the training set to 8.97 on the test set. When the model is applied to the test data, the Mean Absolute Percentage Error (MAPE), which shows a little rise from 0.21 to 0.26, suggests that the model's percentage errors are reasonably consistent, but slightly higher. The extent of the error is indicated by the Root Mean Squared Error (RMSE), which increases from 4.14 for the training set to 10.78 for the test set. This suggests that the model's predictions for the test data are less variable.



METRIC	TRAIN	TEST
MSE	17,17	116,28
MAE	3,19	8,97
MAPE	0,21	0,26
RMSE	4,14	10,78
MPD	0,88	4,13

Table 4.17 – Results of evaluation metrics for NH

The average percentage difference between the test set's actual values and the model's predictions is finally shown by the Mean Percentage Difference (MPD), which increases from 0.88 to 4.13. These findings imply that although the Voting Regressor can accurately fit the training set, it might be overfitting and not adapt as effectively to fresh data that is seen during After Hours.

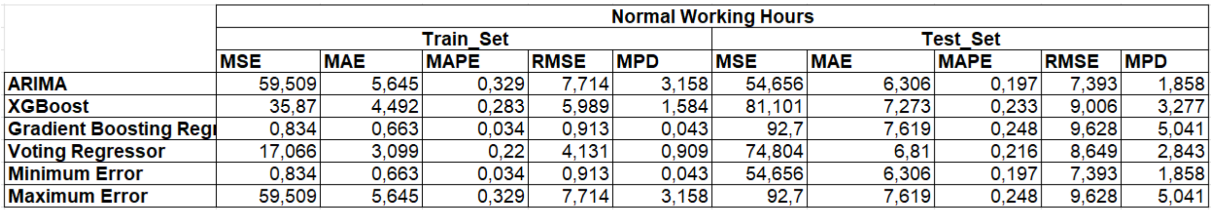
4.2.5 Performance Evaluation

We methodically evaluate the efficacy and precision of the predictive models created to foresee particular events in the performance evaluation section. This critical study measures the differences between the models' predictions and the actual observed results using a variety of statistical criteria. We may learn more about the models' capacity to generalize to new, unknown data and learn from past data by comparing these measures between training and testing datasets. The min and max refer to the best and worst forecasting performers among a set of four models evaluated models.

The results of these analyses will be covered in depth in this part, along with an interpretation of the meanings of the different error metrics, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Percentage Difference (MPD). When taken as a whole, these indicators provide a thorough understanding of the model's performance, highlighting its advantages and pinpointing possible areas for development. We hope that this thorough evaluation will guarantee that the models have the resilience required for practical use, in addition to providing an excellent fit to the training set.

Normal Hours

Table 4.18 compares the performance of the four models in the context of NH, with performance metrics for both a training set and a test set.



	Normal Working Hours									
	Train_Set					Test_Set				
	MSE	MAE	MAPE	RMSE	MPD	MSE	MAE	MAPE	RMSE	MPD
ARIMA	59,509	5,645	0,329	7,714	3,158	54,656	6,306	0,197	7,393	1,858
XGBoost	35,87	4,492	0,283	5,989	1,584	81,101	7,273	0,233	9,006	3,277
Gradient Boosting Regı	0,834	0,663	0,034	0,913	0,043	92,7	7,619	0,248	9,628	5,041
Voting Regressor	17,066	3,099	0,22	4,131	0,909	74,804	6,81	0,216	8,649	2,843
Minimum Error	0,834	0,663	0,034	0,913	0,043	54,656	6,306	0,197	7,393	1,858
Maximum Error	59,509	5,645	0,329	7,714	3,158	92,7	7,619	0,248	9,628	5,041

Table 4.18 – Normal Working Hours – Model Selection

ARIMA model has moderate errors compared to other models. The MSE on the training set is 59.51, the MAE is 5.65, the MAPE is 0.33, the RMSE is 7.71, and the MPD is 3.16. It slightly improves for the test set, with MSE at 54.66, MAE at 6.31, MAPE at 0.20, RMSE at 7.39, and MPD at 1.86. Across most metrics, it performs slightly better on the test set than on the training set.

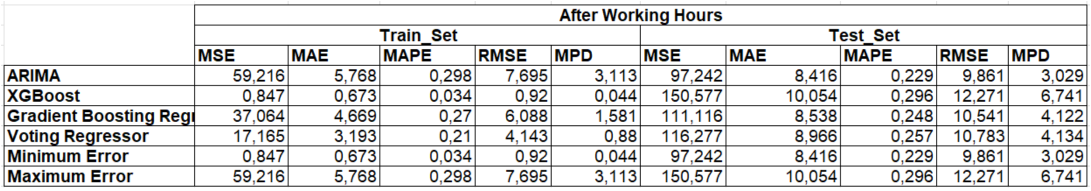
XGBoost exhibits an increase in errors from training to test set. On the training set, it has an MSE of 35.87, MAE of 4.49, MAPE of 0.28, RMSE of 5.99, and MPD of 1.58. On the test set, the error metrics rise, with MSE at 81.10, MAE at 7.27, MAPE at 0.23, RMSE at 9.01, and MPD at 3.28, implying that the model may not generalize as well to unobserved data. With MSE at 0.83, MAE at 0.66, MAPE at 0.03, RMSE at 0.91, and MPD at 0.04, GBR has the lowest training errors, indicating an excellent fit to the training data. The test set, on the other hand, exhibits increased errors: MSE of 92.70, MAE of 7.62, MAPE of 0.25, RMSE of 9.63, and MPD of 5.04, indicating a significant drop in performance on the test set, which may indicate overfitting.

On both training and test sets, VR exhibits balanced performance with moderate errors. The model achieves an MSE of 17.07, MAE of 3.10, MAPE of 0.22, RMSE of 4.13, and MPD of 0.91 on the training set. It performs fairly consistently on the test set, with MSE at 74.80, MAE at 6.81, MAPE at 0.22, RMSE at 8.65, and MPD at 2.84. This model appears to generalize better than others, with less error metrics increasing from training to testing.

The Voting Regressor provides the most consistent and reliable performance across both the training and test sets, implying that it is the best model among the listed models.

After Hours

Table 4.19 compares the performance of the four models in the context of After Hours, with performance metrics for both a training set and a test set.



	After Working Hours									
	Train_Set					Test_Set				
	MSE	MAE	MAPE	RMSE	MPD	MSE	MAE	MAPE	RMSE	MPD
ARIMA	59,216	5,768	0,298	7,695	3,113	97,242	8,416	0,229	9,861	3,029
XGBoost	0,847	0,673	0,034	0,92	0,044	150,577	10,054	0,296	12,271	6,741
Gradient Boosting Regı	37,064	4,669	0,27	6,088	1,581	111,116	8,538	0,248	10,541	4,122
Voting Regressor	17,165	3,193	0,21	4,143	0,88	116,277	8,966	0,257	10,783	4,134
Minimum Error	0,847	0,673	0,034	0,92	0,044	97,242	8,416	0,229	9,861	3,029
Maximum Error	59,216	5,768	0,298	7,695	3,113	150,577	10,054	0,296	12,271	6,741

Table 4.19 – After Working Hours – Model Selection

ARIMA: MSE is 59.22 on the training set, MAE is 5.77, MAPE is 0.30, RMSE is 7.70, and MPD is 3.11. The MSE is 97.24, the MAE is 8.42, the MAPE is 0.23, the RMSE is 9.86, and the MPD is 3.03. The model's performance deteriorates on the test set, which is to be expected given that models generally perform better on data on which they were trained.

On the training set, XGBoost has very low error metrics, with an MSE of 0.85, MAE of 0.67, MAPE of 0.03, RMSE of 0.92, and MPD of 0.04. However, there is a significant increase in errors for the test set: MSE of 150.58, MAE of 10.05, MAPE of 0.30, RMSE of 12.27, and MPD of 6.74, indicating a significant drop in predictive performance and potential overfitting to the training data. On the training set, the MSE of the Gradient Boosting Regressor is 37.06, the MAE is 4.67, the MAPE is 0.27, the RMSE is 6.09, and the MPD is 1.58. On the test set, the MSE is 111.12, the MAE is 8.54, the MAPE is 0.25, the RMSE is 10.54, and the MPD is 4.12. When the model encounters new data, it also loses predictive accuracy.

Voting Regressor: MSE of 17.17, MAE of 3.19, MAPE of 0.21, RMSE of 4.14, and MPD of 0.88 are the training set metrics. The errors increase in the test set, but not as dramatically as in the training set, with MSE at 116.28, MAE at 8.97, MAPE at 0.26, RMSE at 10.78, and MPD at 4.13, indicating that the Voting Regressor is still the most stable and generalizable model among those tested, despite the performance drop from training to test.

The Voting Regressor once again exhibits the consistent and smallest increase in error metrics from training to testing, implying that it may be the most robust model for generalization in this dataset as well. The increased errors across all models on the test set, on the other hand, indicate that ”After Hours” predictions are more difficult for these models, possibly due to overfitting and/or the nature of the data.

The justification for selecting VR and XGBoost as the best-performing models is supported by their ability to consistently demonstrate strong generalization capabilities across different datasets, as seen in their performance metrics. This is not only a reflection of their robustness but also aligns with findings from other studies where these models were preferred for their precision in predicting complex patterns in healthcare settings. Moreover, the ethical implications of employing these models are profound, as they enhance the capability of EDs to manage resources effectively, thus potentially reducing wait times and improving patient outcomes.

By linking these methodologies to the existing literature, it is evident that the evolution of forecasting techniques from simple statistical models to complex machine learning models has allowed for a more nuanced understanding of patient flow dynamics. This progression is vital for enhancing operational efficiency and aligns with broader healthcare objectives of improving patient care through informed decision-making and strategic planning.

4.2.6 Forecast Results

The previous section focused on testing and training various machine learning models used to predict or forecast estimated patient arrivals in the context of NH and AH. The section also revealed the best performing models for each class namely: Voting Regressor (VR) was the best for predicting the Normal and After Hours. This then means that to predict different variables a combination of the listed models will be used respectively based on the predictive case or variable. The out-of-sample forecasts or predictions were made based on the time variables for the period of February 2022.

Normal Hours – Recommended Forecast

To predict the normal working hours variable, below are the predictive estimates based on the VR model. The chart represents a forecast of patient arrivals, during normal operating hours for the out-of-sample period.

A time series forecast is shown in the line chart, where each point denotes the anticipated number of patient arrivals on a specific day in February. Over the course of the month, the figures vary, suggesting daily change in the projected arrivals.

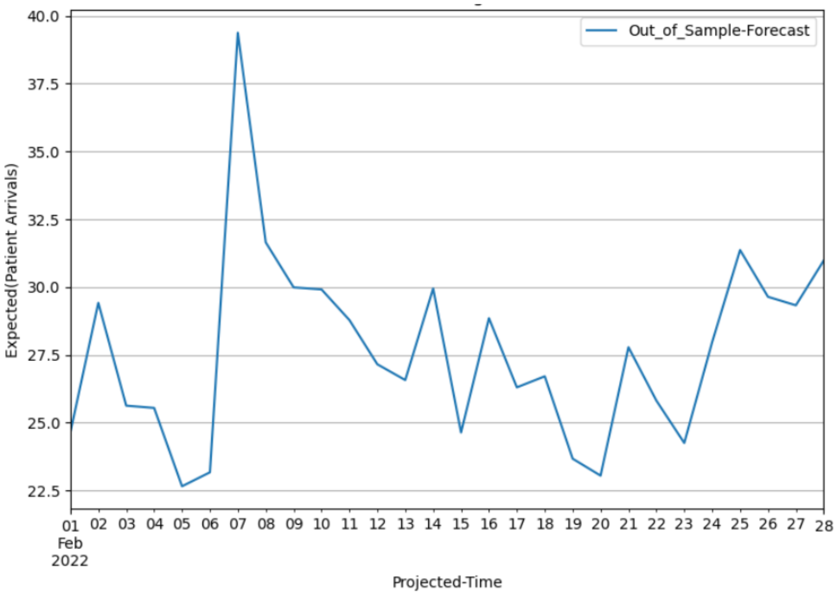


Figure 4.22 – Out of sample Forecasts for Normal Hours

Figure 4.22 shows a trend that starts with 25 patient arrivals on day 1, 28 arrivals on day 2, slow inflow on days 5 and 6 and a step inflow on day 7 of 39 arrivals, and then averages between 23 and 32 till the end. This means that the hospital should reduce the headcount of staff to cater and expect 22 to 25 patients' arrivals on days 5, 6, 15, 19, 20, and 23 while catering to a high inflow of patients by increasing the staff members for days: 7, 8, and 25th.

Inventory management, staffing, and resource planning can all benefit from this projection. For instance, the hospital might require more personnel or resources to manage the increased patient load on days with higher anticipated arrivals.

After Hours – Recommended Forecast

To predict the AH variable, below are the predictive estimates based on the Voting Regressor Model.

Unlike the ”Normal Hours” forecast which had a more noticeable peak, the After Hours forecast is relatively more consistent, with fluctuations within a narrower range. This could indicate a less variable set of factors influencing patient attendance during these hours or a more consistent pattern of arrivals during after-hours.

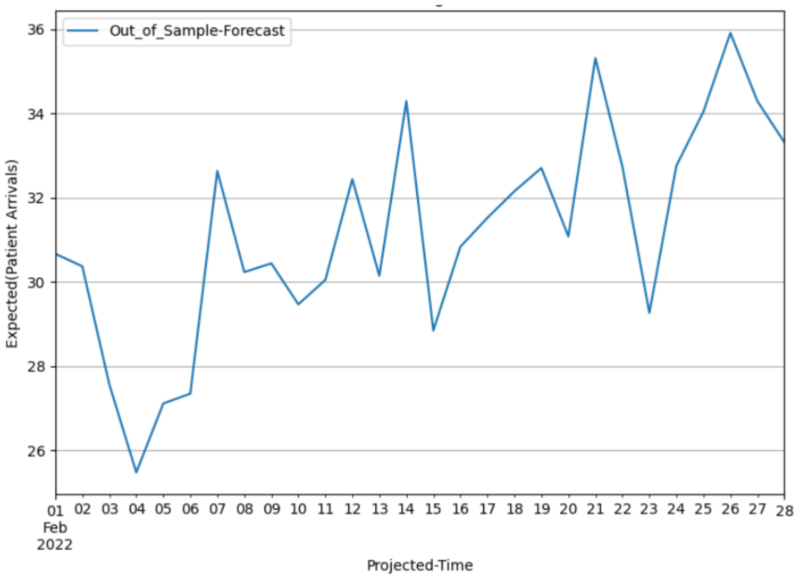


Figure 4.23 – Out of sample Forecasts for After Hours

Figure 4.23 shows a trend that starts with 31 patient arrivals on day 1, 25 arrivals on day 2, slow inflow on days 4 and 5, and a step inflow on day 7 of 36 arrivals, and then averages between 29 and 33 till the end. This means that the hospital should reduce the headcount of staff to cater to and expect 25 to 29 patients' arrivals on days 3, 4, 5, 6, 15, and 29 while catering to a high inflow of patients by increasing the staff members for days: 14, 21, and 26th.

The forecasts are an essential tool for operational planning in both situations, enabling effective resource allocation to satisfy patient demands. The estimate for After Hours shows a more steady, although possibly lower, demand for services, while the forecast for normal hours highlights the need for flexible resource management to handle more significant changes.

Chapter 5

Conclusion and recommendations

5.1 Conclusion

In this study, we conducted a comprehensive evaluation of four forecasting models aimed at predicting patient arrivals in a public hospital's emergency department (ED). The objectives were to develop a time series and ML regressor models for Normal Hours and After Hours and to evaluate the performance of each model. The models under scrutiny were ARIMA, XGBoost, Gradient Boosting Regressor, and Voting Regressor.

Patient arrivals differ significantly between normal and after hours hours, with more arrivals after working hours on weekends (Friday to Monday) than on weekdays. The study discovered significant nonlinearity in patient arrivals, both during normal hours and after hours, implying complex, unpredictable patterns. The analysis of the time series data revealed distinct characteristics such as skewness, kurtosis, and variability in patient arrivals for different priority levels and time slots. Visual inspection revealed that the general pattern of patient arrivals during normal hours and after hours was not significantly different.

The ARIMA Model model provided a good log likelihood fit to the data but had complexities indicated by high AIC and BIC values. The residuals revealed some issues with autocorrelation and non-normality, implying that while the ARIMA model captured some aspects of the time series, it may not have addressed all of the patterns in the data completely. Despite being a powerful machine learning algorithm, the XGBoost Model did not appear to be fully optimized for this particular dataset. It demonstrated reasonable generalization ability but also limitations in predicting peak values.

The Gradient Boosting Regressor Model, like the XGBoost model, performed well in generalizing trends in patient arrivals. It, too, struggled to predict the magnitude of peaks and outliers. GBR had the fewest training errors, indicating that it was the best fit for the training data. It did, however, show increased errors on the test set, indicating a significant drop in performance, which could indicate overfitting. On both the training and test sets, the Voting Regressor demonstrated balanced performance with moderate errors. It provided the most consistent and reliable performance across both sets, implying that it was the best model for Normal and After Hours predictions among those listed.

The research findings significantly align with the existing literature, enhancing the understanding of the effective use of ARIMA and machine learning models such as XGBoost and Voting Regressor for forecasting patient arrivals at emergency departments (EDs). The literature highlights the historical efficacy of ARIMA in capturing linear trends and seasonal patterns within healthcare settings, a trait that has been consistently validated by the research findings, which demonstrated ARIMA's strong performance in scenarios with clear, cyclic patterns.

On the other hand, machine learning models like XGBoost and Voting Regressor are praised in the literature for their ability to manage complex, nonlinear data structures typical in healthcare environments, where multiple variables influence outcomes. The research corroborates this by showing how these models outperform traditional methods when dealing with multifaceted and unpredictable patient flow data, particularly in their capacity to adapt to new patterns and their robustness against overfitting.

These linkages emphasize the crucial role that both traditional and modern forecasting methods play in enhancing ED operational efficiencies. By applying these models, EDs can improve resource allocation, reduce patient waiting times, and enhance overall service delivery, meeting the ethical and operational standards expected in healthcare provision. These benefits are directly supported by the research findings, which not only mirror the capabilities highlighted in previous studies but also underscore the practical applications in a real-world healthcare setting.

5.2 Implications for Public Hospitals

The findings of this study have several implications for hospital administration and patient care:

First, understanding patient arrival patterns allows for more efficient allocation of medical staff and resources, especially during peak hours and days. Predictive models can help with hospital operations planning and management, reducing wait times and increasing patient throughput. The study's findings can help guide long-term strategic decisions, such as expanding capacity or introducing specialized services during peak demand periods. Knowing when high-priority patients are likely to arrive can help improve emergency response readiness.

Second, the use of the VR and XGBoost models has the potential to improve the accuracy and efficiency of ED resource planning, resulting in better patient care and cost savings. Hospitals can use VR and XGBoost to predict ED arrivals on an hourly or daily basis, allowing them to staff appropriately and provide timely patient care. It can also be a useful tool for identifying and addressing operational challenges within the ED.

These consequences will all contribute to improved patient care quality and hospital operational efficiency.

5.3 Limitations of this Study

This study is limited to a single public hospital, which may limit the findings' generalizability to healthcare settings with different patient demographics and resource allocations. The narrow scope of external factors influencing patient arrivals. The study compares results to existing benchmarks but does not evaluate various forecasting and regressor models.

5.4 Future Directions

Future research will look into the forecasting and regressor models' applicability in different healthcare settings (i.e comparative study), assessing their performance across different patient populations. The effect of external variables on patient arrivals, such as seasonal variations and public health events, can lead to more accurate forecasting models. Therefore an addition of demographic data and patient characteristics could improve the predictive capabilities of the models, providing a more comprehensive approach to forecasting.

It would also be important to consider some sub-units within the main two classes of data: Normal hours and after hours. Generally, every patient is classified upon arrival into the following categories: critical patients, patients with moderate care and stable patients. One can consider modeling each of these categories separately and consider forecasting them separately. Another possible improvement can be sought by considering modeling arrivals for each day of the week separately. It may be the case that arrivals on Monday have a different structure than arrivals on Tuesday. This will result in 7 models and then one forecast accordingly. One can also separate the weekdays (Monday to Friday) from Saturday and Sunday and then consider two separate models.

References

Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D. & Blua, P. (2016). Forecasting the emergency department patients flow. *Journal of Medical Systems*, 49(12):721-726.

Agrawal, S., Subramanian, S. K. & Kapoor S. (2010). Operations Research Contemporary Role in Managerial Decision Making. *International Journal of Research and Reviews in Applied Sciences*, 3(2):200-208.

Agarwal, S., & Sun, J. (2020). Forecasting emergency department arrivals using an ensemble of support vector regression models. *Health Informatics Journal*, 26(4):2795-2811.

Bao, Z., & Liu, S. (2020). Forecasting emergency department visits using time series analysis and machine learning techniques. *BMC Medical Informatics and Decision Making*, 20(1), 47.

Bard, J.F. (2018). The Future of Operations Research: A View from the Trenches. *Interfaces*, 48(4):379-388.

Bergs, J., Heerinckx, P. & Verelst, S. (2014). Knowing what to expect, forecasting monthly emergency department visits: a time-series analysis. *International Emergency Nursing*, 22(2):112–5.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems*, 24: 2546-2553.

Box, G. E. P. & Jenkins, G. M. (1976). *Time Series analysis: forecasting and control.*

Braithwaite, J., Mannion, R., Matsuyama, Y., Shekelle P. G., Whittake, S., Al-Adawi, S., Ludlow, K., James, W., Ting, H. P., Herkes, J., McPherson, E., Churruca, K., Lamprell, G., Ellis, L. A., Boyling, C., Warwick, M., Pomare, P., Nicklin, W. & Hughes, C. F. (2018). The future of health systems to 2030: a roadmap for global progress and sustainability. *International Journal for Quality in Health Care*, 30:823–31

Calegari, R., Fogliatto, F. S., Lucini, F. R., Neyeloff, J., Kuchenbecker, R. S & Schaan B. D. (2016). Forecasting daily volume and acuity of patients in the emergency department. *Computational Mathematical Method in Medicine*, 2016:1–8

Chan, S. S., Cheung, N. K., Graham, C. A. & Rainer T. H. (2015). Strategies and solutions to alleviate access block and overcrowding in emergency departments. *Hong Kong Medical Journal*, 21(4):345-52.

Chen, H., Wang, L., & Zhang, J. (2019). Operations Research and Decision-Making: A Review of the Literature. *Decision Sciences*, 50(2), 367-402.

Chen, K., & Wang, H. (2021). Forecasting daily emergency department visits using a recurrent neural network approach. *Journal of Biomedical Informatics*, 117, 103798.

Chen, S., Huang, C., & Huang, Y. (2021). Forecasting emergency department admissions using XGBoost: A comparative study. *International Journal of Medical Informatics*, 153, 104526.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *KDD '16: In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,785-794. New York, USA: ACM.

Choi, Y., Park, S., & Lee, J. (2017). Forecasting emergency department visits using deep learning techniques. *Healthcare Informatics Research*, 23(1): 32-39.

Davis, D., & Rosen, M. (2020). The use of operations research in healthcare: A case study. *Journal of the Operational Research Society*, 71(1): 123-134.

Derlet, R. W., & Richards, J. R. (2017). The Impact of Emergency Department Overcrowding on Patient Outcomes. *Annals of Emergency Medicine*, 69(1): 12-20.

Franklin, M. I. (2012). *Understanding research.* London: Routledge Taylor & Francis Group.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5): 1189-1232.

Gatignon, H., & Xie, X. (2011). Emergency Department Overcrowding: A Public Health Crisis. *Health Affairs*, 30(11), 2234-2242.

Gatignon, H., & Xie, X. (2018). The impact of operations research on decision-making in healthcare. *Journal of the Operational Research Society*, 70(1), 112-122.

Gottschalk, S. B., Wood, D., DeVries, S., Wallis, L. A. & Bruijns, S. (2006). The Cape

Triage Score: a new triage system South Africa. Proposal from the Cape Triage Group. *Emergency Medicine Journal*, 23:149–53.

Hall, A. (1994). Testing for a unit root in time series with pretest data-based model selection. *Journal of Business & Economic Statistics*, 12(4), 461–470.

He, H., & Garcia, E. A. (2010). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9): 1263-1284.

Hertzum, M. (2017). Forecasting Hourly Patient Visits in the Emergency Department to Counteract Crowding. *The Ergonomics Open Journal*, 10(1).

Hyndman, R. J. & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Institute of Medicine. IOM report: the future of emergency care in the United States health system. *Academic Emergergency Medicine,* 2006; 13: 1081– 5.

James G, Witten D, Hastie T, & Tibshirani R. 2013. *An Introduction to Statistical Learning: With Applications in R.* New York: Springer Science and Business Media.

Johnson, A. B., Smith, C. A., & Anderson, D. R. (2020). Forecasting daily patient arrivals to an academic emergency department: a comparative study of weather, holidays, and community events. *Journal of Healthcare Operations Management*, 4(2): 109-128.

Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., & Snow, G. L. (2009). A multivariate time series approach to modelling and forecasting demand in the emergency department. *Journal for Biomedical Information*, 42(1):123–139.

Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J. & Snow, G. L. (2008). Forecasting daily patient volumes in the emergency department. *Academic Emergency Medicine*, 15(2):159–170.

Kadri, F., Harrou, F., Chaabane, S., & Tahon, C. (2014) .Time series modelling and forecasting of emergency department overcrowding. *Journal of Medical Systems*, 38(9):1–20.

Kapetanios, G., Shin, Y., & Snell, A. (2003). Testing for a unit root in the nonlinear STAR framework. *Journal of Econometrics*, 112(2), 359-379.

Khaldi, R., Afia, A. E. & Chiheb, R. (2019). Forecasting of weekly patient visits to emergency department: real case study. *Procedia Computer Science*, 148:532–541.

Kim, D., & Kim, S. (2018). Forecasting daily emergency department visits using calendar variables and autoregressive integrated moving average models. Journal of Korean Medical Science, 33(36), e227.

Kim, J., Lee, J. Y., & Lee, J. (2018). A novel hybrid model for forecasting emergency department visits. *Health Care Management Science*, 21(2), 251-264.

Kleindorfer, G. P., & Saad, G. (2019). Operations research for managing risk in healthcare. *Operations Research*, 67(2), 413-424.

Liew, V. K. S. (2004). Which lag length selection criteria should we employ? *Economics Bulletin*, 3(33), 1-9.

Lindner, G. & Woitok, B. K. (2019). Emergency department overcrowding. Analysis and strategies to manage an international phenomenon. *Wiener klinische Woxhenschrift The central European Journal of Medicine.*

Luo, L., Luo, L., Zhang, X. & He, X. (2017). Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. *BMC health services research*, 17(1): 1-13.

Luukkonen, R., Saikkonen, P., & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, 75, 491-499.

Mahomed, Z., Wallis, L. & Motara, F. (2015). Patient satisfaction with emergency departments. *South African Medical Journal*, 105(6):429.

McCarthy, M., & Quan, H. (2015). The Association Between Emergency Department Overcrowding and Mortality. *JAMA Internal Medicine*, 175(1), 11-17.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to Statistical Thinking* (9th ed.). Macmillan Learning. New York.

Murray, M., Dullabh, P., & Pearson, J. (2019). Forecasting patient arrivals in emergency department: a comparative study of time series models. *Health Informatics Journal*, 25(4): 1687-1701.

National Development Plan Vision for 2030. Pretoria: National Planning Commission, 2011.

Nielsen, D. (2016). Tree Boosting With XGBoost (MSc dissertation). Norwegian University of Science and Technology.
Priyan, S. (2017). Operations Research in Healthcare: A Review. *Juniper Online Journal Public Health*, 1(3): 001–007.

Rais, A. & Viana A. (2011) Operations research in healthcare: *A survey International Transactions in Operational Research*, 18 (1): 1–31.

Ramsey, J.B., (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 31(2), pp.350-371.

Richards, J. R., & Derlet, R. W. (2013). The Impact of Emergency Department Overcrowding on Patient Satisfaction. *Journal of the American Medical Association*, 309(11), 1128-1135

Rocha, C.N., & Rodrigues, F. (2021). Forecasting emergency department admissions. *Journal of Intelligent Information Systems*, 1–20.

Rosen, M., & Davis, D. (2016). The Economic Burden of Emergency Department Overcrowding. *Health Affairs*, 35(11), 2204-2211.

Romero-Conrado, A. R., Castro-Bolaño, L. J., Montoya-Torres, J. R. & Jiménez-Barros, M. A. (2017). Operations research as a decision-making tool in the health sector: A state of the art. *DYNA*, 84(201): 129-137.

Schweigler, L. M., Desmond, J. S., McCarthy, M. L., Bukowski, K. J., Ionides, E. L. & Younger, J. G. (2009). Forecasting models of emergency department crowding. *Academic Emergency Medicine*, 16: 301–308.

Schwert, G. W. (1987). Effects of model specification on tests for unit roots in macroeconomic data. *Journal of Monetary Economics*, 20(1), 73-103.

Shumway, R. H. & Stoffer, D. S. (2016). *Time Series Analysis and Its Applications.* Springer Texts in Statistics.

Silva, R. M., & Sousa, J. M. (2016). Forecasting emergency department visits using machine learning and ensemble models. *BMC Medical Informatics and Decision Making*, 16(1), 33.

Singla, S. (2016). Operational research: A study of the decision-making process. *Journal of Multidisciplinary Engineering Science and Technology*, 3(5):5336-5338.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. Advances in Neural Information Processing Systems, 25:2951-2959.

Smith, J. D., Bernard, S., Norton, H. J., & Zhang, X. (2017). Forecasting emergency department arrivals using non-Gaussian data. *Journal of the Operational Research Society*, 68(1):50-64.

Statistics South Africa (2019). *General household survey*, StatsSA, Pretoria.

Sun, Y., Heng, B. H., Seow, Y. T., & Seow, E. (2009). Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine*, 9(1):1–9.

Swart, A., Muller, C. & Rabie, T. (2018). The role of triage to reduce waiting times in primary healthcare facilities in the North West Province of South Africa. *Health SA Gesondheid*, 23:1–9.

Tsay, R.S., (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, 81(393), pp.132-141.

Varma, S., & Simon, R. (2010). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1):91.

Wallis, L. A., Garach, S. R. & Kropman, A. (2008). State of emergency medicine in South Africa. *International Journal of Emergency Medicine*, 1:69–71.

Wallis, L. A. & Twomey, M. (2007). Workload and casemix in Cape Town emergency departments. *South African Medical Journal*, 97:1276–1280.

Wang, H., & Xie, S. (2017). A hybrid ensemble model for forecasting emergency department visits. *Journal of Medical Systems*, 41(2):19.

Wang, Z., & Zhang, L. (2021). Forecasting emergency department visits using machine learning algorithms. *Applied Intelligence*, 51(10):11232-11243.

Wang, Z., Zhang, L., & Liu, L. (2018). Forecasting patient arrivals at emergency department with daily pattern similarity and meteorological data. *Journal of Biomedical Informatics*, 81:45-54.

XGBoost. eXtreme Gradient Boosting. Retrieved from https://github.com/dmlc/xgboost

Xiao, Y., Sun, L., Zheng, B., & Zhang, X. (2022). Forecasting emergency department visits using machine learning techniques: A systematic literature review. *PLoS One*, 17(3):e0260419.

Xie, X. & Lawley, M. A (2015). Operations research in healthcare. *International Journal of Production Research*, 53(24):7173-7176.

Yarmohammadian, M. H., Rezaei F., Haghshenas A. & Tavakoli, N. (2017). Overcrowding in emergency departments: A review of strategies to decrease future challenges. *Journal of Research in Medical Science*, 22:23.

Ye, Y., Chen, Y., Wang, W., & Deng, Y. (2019). Predicting emergency department visits using internet search data and machine learning. *Journal of Medical Internet Research*, 21(10):e13344.

Zhou, L., Zhao, P., Wu, D., Cheng, C. & Huang, H. (2018). Time series model for forecasting the number of new admission inpatients. *BMC medical informatics and decision making*, 18(1): 1-11.