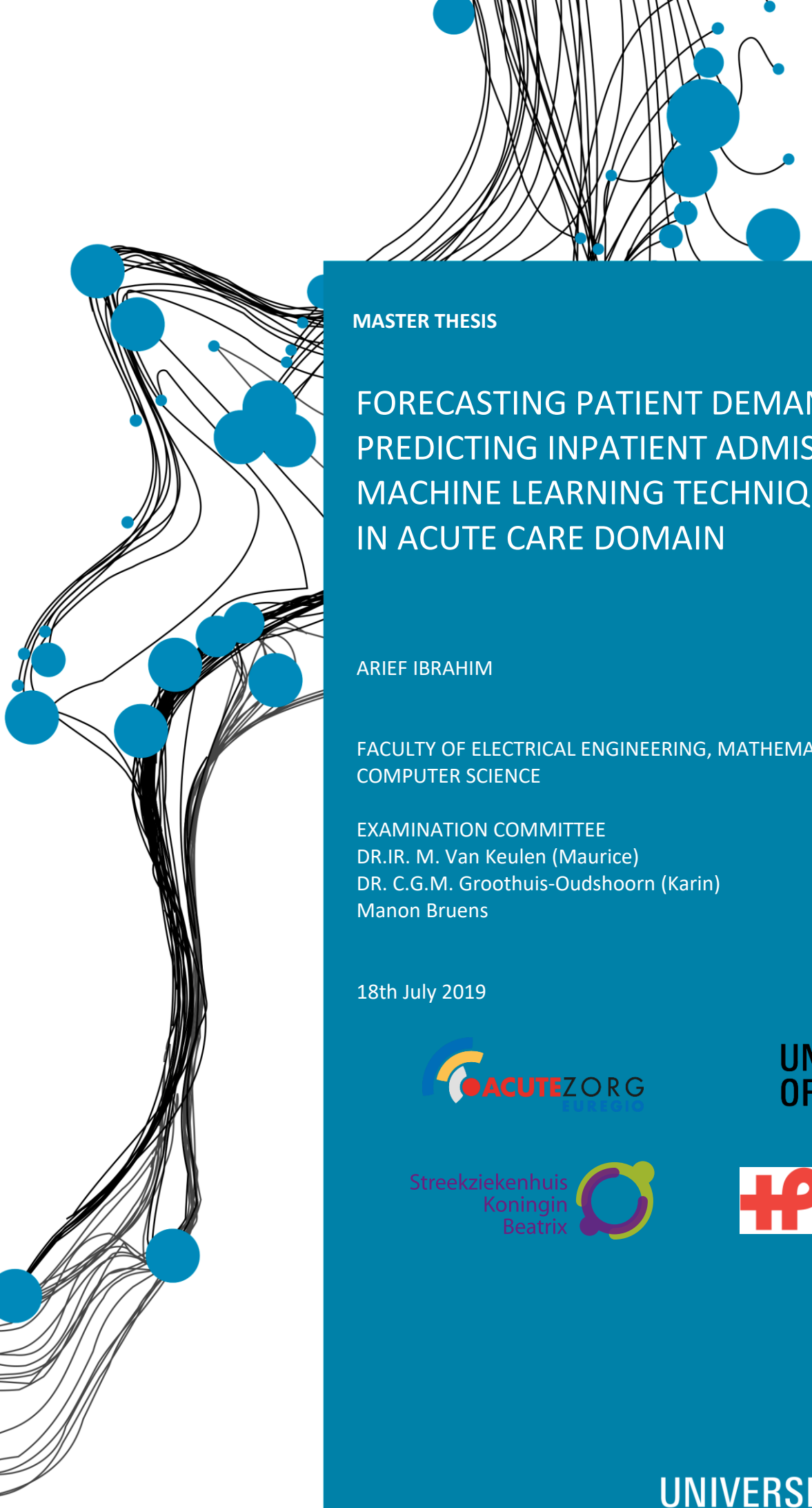


MASTER THESIS

FORECASTING PATIENT DEMAND AND PREDICTING INPATIENT ADMISSION VIA MACHINE LEARNING TECHNIQUES IN ACUTE CARE DOMAIN

ARIEF IBRAHIM

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS, AND COMPUTER SCIENCE

EXAMINATION COMMITTEE
DR.IR. M. Van Keulen (Maurice)
DR. C.G.M. Groothuis-Oudshoorn (Karin)
Manon Bruens

18th July 2019

ACUTEZORG EUREGIO

UNIVERSITY OF TWENTE.

Streekziekenhuis Koningin Beatrix

Huisartsenpost Oost-Achterhoek Voor Spoed

UNIVERSITY OF TWENTE.

Acknowledgment

This thesis marks the end of my journey in the master program of Business Information Technology (BIT) at the University of Twente. I have gained not only new knowledge but also different yet insightful perspectives and the most importance, the experience, during my master study. All these learning events, I believe, will enhance my career afterward. I deeply realize that I cannot achieve this significant milestone in my life without the support of the people around me.

First and foremost, I would like to express my gratitude to the Almighty, the One and Only. Believing in His blessing, mercy, love, and guidance, I have become more optimistic in every difficulty. I would also like to say thank you to my country, especially to the Ministry of Communication and Information of the Republic of Indonesia (KOMINFO), that have covered my living cost during my master study. Besides, I would like to say thank you to the Netherlands government through its StuNed program of Nuffic Nesso Indonesia that have funded my master study. It is such an honor for me to be one of the awardees of KOMINFO-StuNed scholarship. Moreover, I would like to thank my UT supervisors, Maurice, and Karin, and my company's supervisor, Manon (including Nancy), for all your valuable inputs, comments, and feedback to improve my thesis and support me in completing it on schedule.

At this moment, I would like to sincerely say thank you to my wife, Ratna Kusumastuti, who has been very supportive, patient, and full of love in accompanying her husband journey. Also, this thankfulness goes to my family: Mama, Bapak, Ka Rahma, Ka Lala, Ka Amey, Ka Dila, including my wife's family: Mama Dewi, Om Helmi, Mba Icha and Ka Cempaka for all your support and praying for my wife and me as well. I would like to express my gratitude for the support of my Indonesian friends in Enschede: Yaumi, Yani, Yasir, Dzul, Nden, and many others as well as my UT friends: Nivedita, Amit, Liu, Thomas, Dirk, Olivia, and Erick. With their support, their help, collaboration, and friendship, my thesis journey has become colorful. I will not forget to say thank you for all Indonesian bike community in Enschede who have shared memories in our touring in several cities in the Netherlands and also Germany.

And to other people that I cannot mention one by one, thank you for being a part of my journey during my study. I wish you all the best, and I hope we will meet again in the future.

Enschede, 12 July 2019

Arief Ibrahim

ABSTRACT

One of the challenges in managing acute care services, such as ED and GP-Post, is the increasing trend of patient demand. The inability to provide sufficient care services during such high demand period can lead to the overcrowding event. As a result of that, the patients get more suffered, the waiting time becomes longer, and the acute care management can also get financial loss. In the Netherlands, a study shows that almost 70% of Dutch ED managers consider their ED operation is at or even above the capacity several times a week. Similarly, a report also emphasizes the importance of effective solution in addressing the increasing demand for acute care, in particular, ED and GP-Post.

However, measuring the moment when the high patient demand turns into overcrowding event is a challenging problem. The main cause is no universal agreement on how to define and measure the overcrowding itself. In other words, different acute care locations most likely have a specific characteristic of overcrowding. Therefore, in this thesis, counting the daily number of patient demand or patient arrival is used as an indirect indicator of the overcrowding. The ability to forecast the number of incoming patients accurately on the next day can provide valuable information for acute care management in anticipating the overcrowding event. Among various methods, forecasting through machine learning (ML) method was used in this thesis for three reasons: (1) the effectiveness of ML methods which can be considered as a black box, (2) the ability of ML in providing the correlation and the importance level of external factors (e.g temperature and humidity), and (3) the ability of ML in predicting the future with a certain level of accuracy, error ranges, and confidence interval.

The Emergency Department (ED) and the general Practitioner Post (GP-Post) at Winterswijk in the Netherlands are selected as a case study to research and develop two forecasting tools of ED and GP-Post patient demand based on the internal historical data and also external data such as weather and pollen. Moreover, the stakeholders are also interested in predicting the probability of inpatient admission to the hospital through ML techniques. Apart from these two, analyzing the linear correlation between external factors with some particular patient groups (e.g., Age and Treatment group) also become interesting insight for the stakeholders. Based on these objectives, the main research question of this thesis is formulated as:

“To what extent can one utilize machine learning techniques in the acute care domain such as ED and GP-Post?”

The methodology used in this thesis is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM), in particular, the first five phases namely Business Understanding, Data Understanding, Data Preparation, Modelling, and Evaluation. These five phases can be broken down into more detail activities. In Business Understanding phase, three main activities were performed, namely (a) problem identification, (b) motivation, objective and scope, and (c) domain analysis. In Data Understanding phase, also three main activities were performed, namely (a) statistical summary & visualization, (b) time series analysis, and (c) features analysis. In the Data Preparation phase, four main activities were performed, namely (a) Aggregation, (b) Integration, (c) Feature Engineering, (d) Segregation. In the Modelling phase, four main activities were performed, namely (a) model building, (b) bias vs. variance analysis, (c) feature selection, (d) model optimization. In Evaluation, two main activities were performed, namely (a) model comparison, (b) result in analysis. Apart from the methodology, the literature gap was performed to identify a hybrid ML model using SARIMAX and Gradient Tree Boosting.

There were three primary results from this thesis. First, forecasting ED patient demand with a hybrid model, SARIMAX(0,0,0)x(1,0,1,7) and Gradient Tree Boosting, came up as the best model by MAPE 16.50%, RMSE 6.56, and MAE 5.25. To achieve this performance, only six features, namely Is_Weekday, GP_Post_WH_Opening, W_TX-1, ICPCcode_L-1, Is_Weekday-2, Is_Weekday-3, were required out of the initial 1132 features. Second, forecasting GP-Post patient demand with a hybrid model, SARIMAX (1, 0, 1) x (1, 0, 1, 7) and Gradient Boosting, also came up as the best model by MAPE 13.70%, RMSE 13.94, and MAE 9.43. To achieve this performance, only seven features, namely Is_Weekday, GP_Post_WH_Opening, U5-7, Is_Holiday-1, Is_Weekday-1, Is_Weekday-3, Is_Weekday-5, were required out of the initial 1132 features. Third, predicting patient admission with GradientBoostingClassifier yielded the best performance by accuracy 78%, precision 73%, recall 73%, F1 score 73%, ROC-AUC 77%. To achieve this performance, only 21 features were required out of the initial 38 features. The top three of 21 features are Treatment_CHI, Age, and Urgency_Green. The summary of these results is presented in Table 1.

Table 1: Summary result

Objectives	Best ML Model	Metrics Evaluation on Testing Dataset	Feature Importance
Forecasting ED patient demand	Hybrid model (SARIMAX and Gradient Tree Boosting)	MAPE: 16.50% RMSE: 6.56 MAE: 5.25	Calendar: Is_Weekday, Is_Weekday-2, Is_Weekday-3, GP_Post_WH_Opening Internal: ICPCcode_L-1 External: W_TX-1 (Max temperature of yesterday)
Forecasting GP-Post patient demand	Hybrid model (SARIMAX and Gradient Tree Boosting)	MAPE: 13.70% RMSE: 13.94 MAE: 9.43	Calendar: Is_Weekday, Is_Holiday-1, Is_Weekday-1, Is_Weekday-3, Is_Weekday-5, GP_Post_WH_Opening Internal: U5-7
Predicting Inpatient Admission	GradientBoostingClassifier	Accuracy: 78% Precision: 73% Recall: 73% F1 score: 73% ROC-AUC: 77%	The top three of 21 features are Treatment_CHI, Age, and Urgency_Green

To conclude and answer the main research question, several machine learning techniques have been applied to the two areas in acute care domain, namely (1) Input area, by forecasting the patient demand, and (2) Output area, by predicting the inpatient admission. In the Input, the result of this thesis showed that the overfitting problem at a single SARIMAX was resolved by applying feature selection technique with Lasso. However, the Hybrid model came up as the best ML forecasting model for ED and GP-Post. In the Output, the result of this thesis showed that GradientBoostingClassifier returned the best prediction, especially in recall score. The optimization through hyper-parameter techniques was able to improve the outcome prediction. Further improvement is even possible via ROC and Precision-Recall curve analysis.

Contents

List of Figures

List of Tables

List of Acronyms

AIC	Akaike information criterion
AR	Autoregressive
ARIMA	Autoregressive integrated moving average
AUC	Area under the curve
CRISP-DM	Cross-Industry Standard Process for Data Mining
ED	Emergency Department
FN	False Negative
FP	False Positive
ICMED	International Crowding Measure in Emergency Departments
GP	General Practitioner
LR	Linear Regression
MA	Moving Average
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MSE	Mean Square Error
NEDOCS	National Emergency Department Overcrowding Score
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SARIMAX	Seasonal ARIMA with exogenous variables
TN	True Negative
TP	True Positive

1. Introduction

This chapter discusses the motivation behind the research, the problem identification, the proposed method in addressing the problem, the research objectives, and the research questions.

1.1 Motivation

Hospitals around the world have been facing a similar major challenge in managing their acute care service such as the Emergency Department (ED), in particular dealing with overcrowding which occurs when the emergency services exceeding the available resources for patient care [1,2,3]. The lack of a proper ED treatment quality may severely affect the patient such as unnecessarily increased the length of stay (LOS), undesirably delayed for triage and treatment, prolonged transport and waiting, worsen the patient's condition, and even financial losses [4,5].

In the Netherlands, a study [6] in 2013 shows that patient's length of stay (LOS) in ED is counted in hours rather than in days as in many other countries, especially in the developing countries. The study [6] also points out that the overcrowding in the Dutch ED does not occur every day; it only occurs during a few hours per day. Even though it might look better than many other countries, the same study [6] indicates the opposite perception described by almost 70% of Dutch ED managers who perceive their ED operation is at or even above the capacity several times a week. Besides ED, Netherlands' healthcare systems have another similar entity called GP-Post. It is part of General Practitioner (GP), but it provides healthcare services for out of hours (5pm-8am). A recent report[1] in 2018 has an echo with the previously mentioned study that the increasing demand for acute care, which includes ED and GP-Post, requires increasingly effective solutions to be able to keep offering high quality and better care services with scarce human resources and equipment.

As the demand for acute care becomes more and more increasing, which potentially cause the overcrowding event, many studies have been conducted in addressing this problem. Various solutions have been offered to make the patient's flow better so that the hospital's management can effectively allocate their resources ahead of time. These solutions have different methodologies such as quality function deployment, failure-mode and effect analysis, simulation, queuing theory, and forecasting. Apart from various available methodologies, the challenge in addressing the overcrowding problem in healthcare services is that there is no universal agreement on how to define and measure the overcrowding. Even though several crowding measures have been developed [7-9] such as ICMED and NEDOCS, their usefulness is in question, as several studies reveal conflicting results [7, 10-13]. In addition, the transferability and scalability of these crowding measures need to be assessed further. This matter is raised because other researchers concern about the performance of the crowding measures, especially in small patient volume ED [8,14,15]. A more recent paper [16] even argues that predefined thresholds of crowding scales might not be optimally applied to all EDs. In other words, different ED in different locations most likely have a specific characteristic of overcrowding.

Instead of defining and measuring the overcrowding, counting the number of patient demand or patient arrival on a daily basis can be efficiently used as an indirect indicator of the overcrowding. The ability to forecast the number of incoming patients several days in the future can provide valuable information for acute care management in anticipating the overcrowding event. Therefore, in this thesis, forecasting with machine learning (ML) technique is selected as the main method because of three reasons. The first, machine

[1] (2019, January 10). Monitor acute zorg 2018: extra aandacht nodig voor vergrijzing .... Retrieved May 5, 2019, from https://www.nza.nl/actueel/nieuws/2019/01/10/monitor-acute-zorg-2018-extra-aandacht-nodig-voor-vergrijzing

learning has been popularly and effectively used to develop a forecasting or prediction model, especially for time series data. As opposed to the simulation method, as a comparison example, ML does not require detail information on each step in the process flow because ML can be considered as a black box which only requires input and output information to build the forecasting model. Secondly, machine learning can be utilized to analyze the correlation and the importance level of external factors, such as weather, to the patient demand against the prediction outcome. The third, machine learning ability in predicting the future with a certain level of accuracy, error ranges, and confidence interval can help for planning and allocating resources.

In this thesis, the Emergency Department (ED) and the general Practitioner Post (GP-Post) at Winterswijk in the Netherlands are selected as a case study to research and develop forecasting tools based on the internal historical data and also external data such as weather and pollen data. This forecasting tool is expected to function as an early warning alarm that can properly anticipate the overcrowding events at ED and GP-Post of Winterswijk. Besides, the stakeholders are interested in predicting the probability of inpatient admission to the hospital with classification methods in machine learning. Having an automate classifier tool can help the management to assess the quality of their service and operation and parallelly improve it as well. Apart from the two machine learning techniques, stakeholders also have high curiosity in analyzing a linear correlation between external factors (e.g., weather and pollen) with some particular patient groups (e.g., age groups, treatment groups) during a certain period such as weekdays-weekends or seasons. With these additional insights on hands, the management, practitioners, or even staffs at ED and GP-Post might have a better understanding of treating the patients in a specific situation (e.g., extreme heat temperature).

1.2 Research objective

As mentioned in the previous section, two entities, which are the ED and GP-Post at Winterswijk, are the subject of this thesis. Based on the above-given motivation, three research objectives can be summarized as follow:

Research objective-1: To forecast one day ahead of ED and GP-Post patient demand with machine learning techniques

GP-Post at Winterswijk is located at the same hospital building as the ED. However, ED and GP-Post are two separate entities, so their historical data are separated and unintegrated. Therefore, two different machine learning models will be built to forecast each of both.

Research objective-2: To predict ED's inpatient admission with a machine learning technique

Besides forecasting, ED's stakeholders are also interested in predicting the probability of inpatient admission to the hospital. Unlike the previous objective, predicting inpatient admission will not return a prediction number. Instead, it predicts by classifying two states, either admitted or not. Therefore, classification methods of machine learning will be used to build a classifier model.

Research objective-3: To analyze a linear correlation between external factors with some particular patient groups

Apart from the two machine learning techniques, stakeholders also have high curiosity in analyzing a linear correlation between external factors (e.g., weather and pollen) with some particular patient groups (e.g., age groups, treatment groups) during a certain period such as weekdays-weekends or seasons.

1.3 Research Question

Based on the background, motivation, and the objectives as discussed in section 1, and after studying the relevant literature in section 2, the main research question of this study can be formulated as follows:

"To what extent can one utilize machine learning techniques in the acute care domain such as ED and GP-Post?"

Furthermore, to completely answer the main research question, the following sub-questions are needed and stated:

Sub Question-1: Which ML methods can be applied to forecast patient demand at the ED?

Sub Question-2: Which ML methods can be applied to forecast patient demand at the GP-Post?

Sub Question-3: Which ML methods can be applied to predict the ED inpatient admission to the hospital?

Sub Question-4: What insights can be derived from exploratory data analysis (EDA) in relation to univariate time series analysis and correlation analysis?

Sub Question-5: Which ML model gives the best prediction result for subquestion 1, 2, and 3?

Sub Question-6: Which features can yield the optimal prediction for subquestion 1, 2, and 3?

1.4 Contribution

This section discusses the contribution of the current study, theoretically and practically, to the acute care domain in general and to particularly the ED and GP-Post.

1.4.1 Contribution to the theory

This research provides two primary contributions to the theory as follow.

- This thesis proposes and implements the hybrid ML model, which combines SARIMAX and Gradient Tree Boosting. To the best of my knowledge, this is the first hybrid model in building an ML forecasting model in the acute care domain. This approaches might be used, followed, or further extended for forecasting patient demand in the acute care domain in the future.
- This thesis also provides a simple yet applicable approach in reducing the number of features even further after performing feature selection with Lasso
- To the best of my knowledge, this is the first thesis that attempts to incorporate and analyze the correlation of pollen data from various plant species to the forecasting of ED and GP-Post patient demand. Therefore, the result of this study might be used as a comparison or benchmark for future similar studies.

1.4.2 Contribution to the practice

This research provides several contributions to the practice as follow.

- The forecasting ED and GP-Post patient demand tool can be implemented in the daily ED operation. The ED management can utilize the forecasting result for anticipating the possibility of overcrowding event on the next day. As a result of that, all the preparation required in term of facilities, human resources, and even budgeting can be planned and allocated accordingly.

- The inpatient admission prediction tool can be used as an early assessment tool for the admission status to the hospital. Besides, it can also be used as a benchmark in evaluating the quality of ED service and operation, which eventually, improve it as well.
- Apart from the two above, the research also provides the linear correlation result between the external factors and some particular patient groups. With these additional insights on hands, the management, practitioners, or even staffs at ED and GP-Post might have a better understanding of treating the patients in a specific situation.

2. Background and Related works

In this chapter, the background of acute care domain, including the description of ED and GP-Post, will be discussed. Moreover, the related works on forecasting in the emergency department and predicting inpatient admission to the hospital will be presented and summarized. Next, the literature gap is also performed to identify the unexplored machine learning techniques in the context of forecasting in acute care domain. Lastly, the basic machine learning concept will be discussed.

2.1 Acute care: ED and GP-Post

Acute care is the most time-sensitive service in the healthcare domain. According to WHO[2], it includes *"all promotive, preventive, curative, rehabilitative or palliative actions, whether oriented towards individuals or populations, whose primary purpose is to improve health and whose effectiveness largely depends on time-sensitive and, frequently, rapid intervention."* In term of its function, acute care is the central entity which consists of a range of clinical health-care functions, including emergency medicine, trauma care, pre-hospital emergency care, acute care surgery, critical care, urgent care, and short-term inpatient stabilization. The illustration of acute care with other clinical health-care functions can be found in Figure 1.

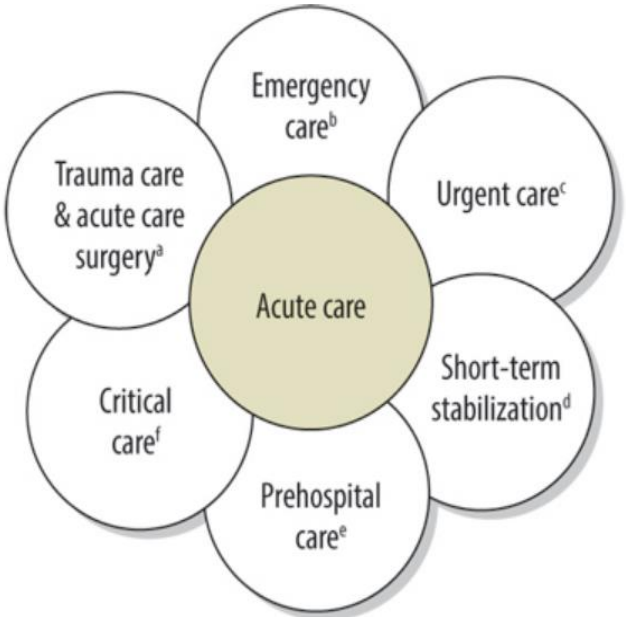


Figure 1: Acute care illustration as adopted from WHO

In the Netherlands, the ED and GP-Post are primarily intended to provide emergency care during out-of-hours[3]. Under the Dutch regulation, the patients who seek healthcare service are highly recommended to firstly consult with GP during working hours or with GP-Post during out-of-hours before they can go to the emergency department. After the assessment at GP or GP-Post, the patient can be referred to the ED for further treatment. The assessment at ED then decides whether the patients can go home or have to be

[2] (n.d.). WHO | Health systems and services: the role of acute care. Retrieved May 5, 2019, from https://www.who.int/bulletin/volumes/91/5/12-112664/en/

[3] (2017). Monitor Samenwerking spoedeisende hulp (seh) en huisartsenposten .... Retrieved May 5, 2019, from https://www.rijksoverheid.nl/documenten/rapporten/2017/09/01/monitor-samenwerking-spoedeisende-hulp-seh-en-huisartsenposten-hap

admitted as an inpatient to the hospital. An overview of simplified patient flow in the Netherland can be found in Figure 2.

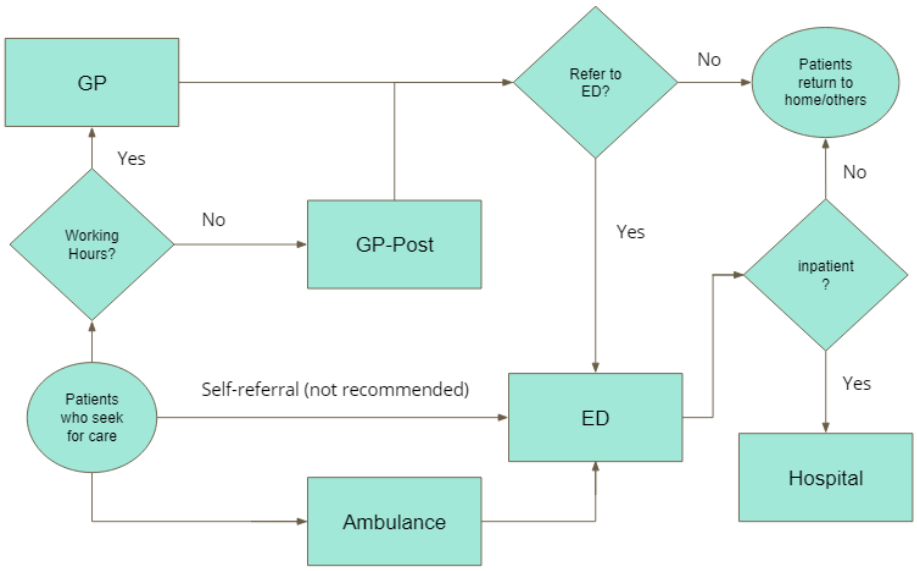


Figure 2: A simplified overview of patient flow in the Netherland

2.2 ED and GP-Post at Winterswijk

ED and GP-Post at Winterswijk in Oost-Achterhoek, which are selected as a study case in this thesis, was officially opened on 28 May 2015. Both are owned and managed by HuisartsenOrganisatie Oost-Gelderland (HOOG). The company has emerged from a merger between Archiatros (facility services from 2003), the Zorggroep van Apeldoorn (2008), Oost-Achterhoek (2007) and Zutphen (2007) and the GPs Apeldoorn, OostAchterhoek and Zutphen of the SDHS (2001). The structure of HOOG and its coverage[4] are described in Figure 3.

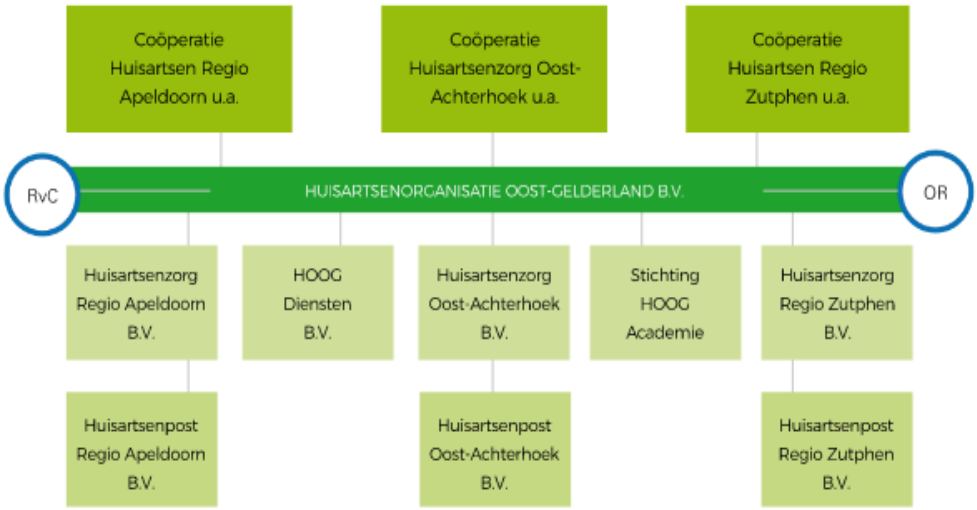
Although both are operated under the same company (HOOG), they manage their patient database separately and operate with a different working hour. ED operates every day 24-hours while GP-Post operates from 17.00 until 08.00 mornings on the next day. However, during the weekend, GP-Post operates 24-hours. ED and GP-Post are equipped with several facilities such as treatment rooms, children room, trauma room, and special room to treat patient a contagious disease.

2.3 Forecasting in the Emergency Department

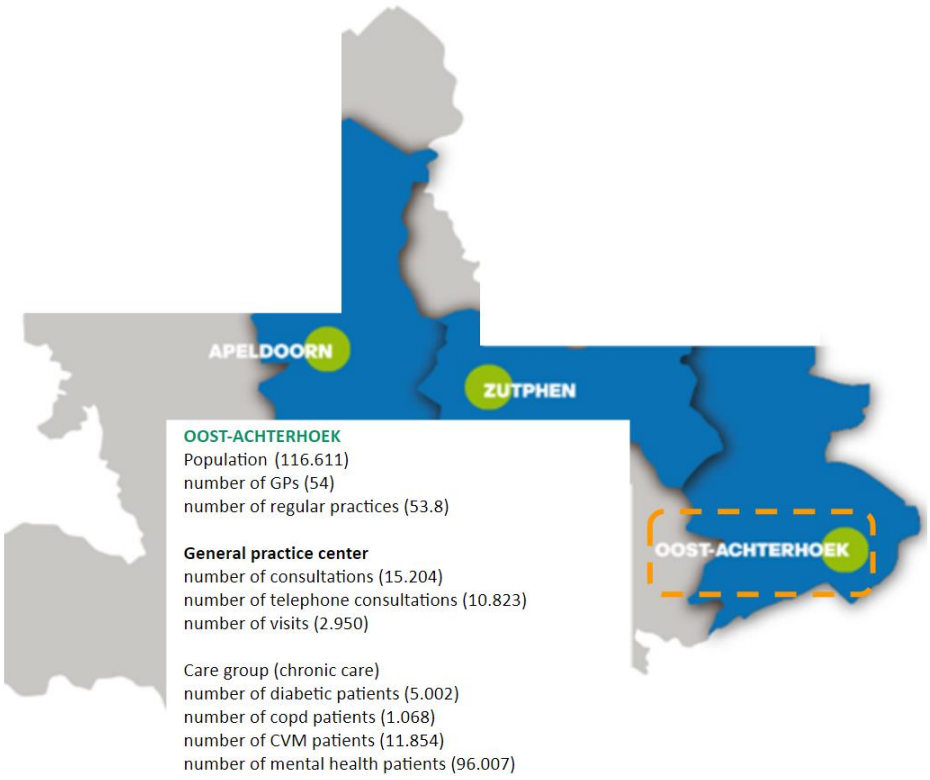
An exhaustive literature review on forecasting in the emergency department [21] categorizes the forecasting topics into three sections based on the patient's flow: Input, Throughput, and Output. The input is mainly dominated by the topics about forecasting patient demand or patient arrival, while the Throughput is mostly dealt with the topics about predicting patient's LOS, and the Output is typically covered by the topics about predicting the inpatient admission to the hospital. Although the mentioned research only discusses the emergency department, the three forecasting topics categorization can also be applied for GP-

[4] (n.d.). Jaarbericht – Hoogzorg. Retrieved May 12, 2019, from https://www.hoogzorg.nl/over-hoog/jaarverslag/

Post in the Netherlands. A simplified and contextualized illustration of the three forecasting topics can be found in Figure 4.



(a)



(b)

Figure 3: (a) HOOG structure (b) HOOG coverage

Based on the interview with stakeholders, the Input and the Output are selected as the main subject area in this thesis. The Input section is applicable for ED and GP-Post, so there will be two forecasting tools

for forecasting GP-Post's and ED's patient demand at a daily level. The Output section, which predicts the probability of inpatient admission to the hospital, can only be applied to ED.

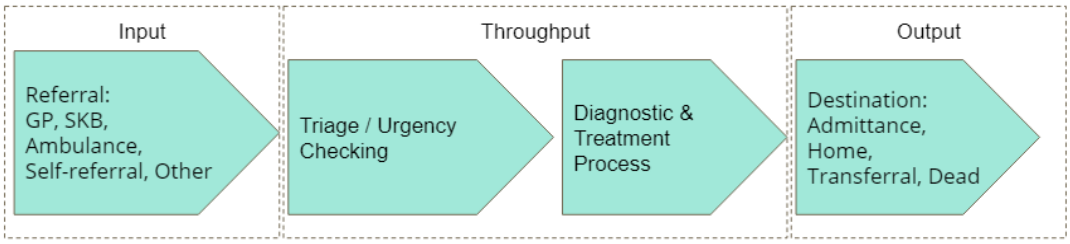


Figure 4: The three forecasting topics at ED and also applicable for GP-Post

2.3.1 Predicting inpatient admission to the hospital

Based on the literature study, the methods used by researchers for forecasting patient demand can be classified into two categories: the classical ML time series and the novel ML time series. The first category considers patient arrival as a sequence of time so that the previous values will function as the main baseline in predicting the future. The examples of classical time series models are ARIMA and its variations such as SARIMA which can also include exogenous variables. The second category includes the correlation between the patient arrival number and other independent variables such as festival or holiday events, weather, temperature, humidity, and so on. The example methods of machine learning are ANNs, SVM, decision trees, and Bayes networks. The list of relevant papers with the finding summary is presented in Table 2.

Table 2: The list of literature review of Forecasting Patient Demand

Authors (Year)	Objective	Methods	Features	Findings
Choudhury (2019) [22]	Hourly forecasting patients at ED	ARIMA, Holt-Winters, TBATS, Neural Network	Univariate Time Series	ARIMA (3, 0, 0)(2, 1, 0) was selected as the best fit model
Whit & Zhang (2019) [23]	Forecasting Arrivals and Occupancy Levels	SARIMAX	temperature and holiday effects	SARIMAX model yielded the best predicting power by exploiting both exogenous variables (temperature and holiday effects) and internal dependence. It suggests that some local related data might be useful for predicting the ED arrivals.

Ekström et al. (2018) [24]	Forecasting ED inflow	ARIMA, Gradient Boosting, Neural Network	day of week, day of month, year, month, and hour of the day	Gradient boosting as modelling method yielded the best results for forecasting the coming 72 hours of ED inflow
M Carvalho-Silva et al. (2018) [25]	Forecasting patient	ARIMA	the precipitation and the maximum temperature	The best model for the test period was the ARIMA (1,1,1)(1,0,1)
WC Juang et al. (2018) [26]	To construct an adequate model and to forecast monthly ED visits	ARIMA	Univariate Time Series	The ARIMA (0, 0, 1) model can be considered adequate for predicting future ED visits
Morten Hertzum (2017) [27]	Forecasting Hourly Patient Visits and ED Occupancy	ARIMA, linear regression, Naive models	Calendar Variables	Hourly patient arrivals can be forecasted with decent accuracy.
PatrickAboagye-Sarfo et al. (2015) [28]	To develop multivariate vector-ARMA (VARMA) forecast models for predicting emergency department (ED) demand and compare to the benchmark univariate autoregressive moving average (ARMA) and Winters' models.	VARMA, ARMA, Winters models	time (monthly)	The VARMA models provided a more precise and accurate forecast with smaller confidence intervals and better measures of accuracy in predicting ED demand in WA than the ARMA and Winters' method.
SS Jones et al. (2008) [29]	Forecasting daily patient volume	SARIMA, ES, Regression	Calendar variables	Regression-based models that incorporate calendar variables account for site-specific special-day effects and allow for residual autocorrelation provide a more consistently accurate approach to forecasting daily ED patient volumes.

Spencer S.Jones et al. (2008) [30]	A multivariate time series approach to modeling and forecasting demand in the emergency department	Multivariate VAR	ED arrivals, census, laboratory orders, radiography orders, CT orders, inpatient census, laboratory orders, radiography orders, CT orders	multivariate VAR models provided more accurate forecasts of ED census compared to standard univariate time series methods.

2.3.2 Predicting inpatient admission to the hospital

Another topic which also attracts researchers in ED domains is forecasting inpatient admission in the output process (Figure 4). Since the ED is ideally designed as the temporary place for an emergency patient, the quick yet accurate diagnose could improve the service quality of the hospital. Therefore, diverse popular forecasting methods can be applied to this case. The results do not only predict the probability of inpatient admission, but it can also be used for identifying the most relevant factors which affect the prediction. The list of collected research papers which focus on forecasting ED inpatient admission is summarized in Table 3.

Table 3: Literature Review of Predicting inpatient admission

Authors (Year)	Objective	Methods	Evaluation Matrics	Findings
Lucke et al., (2018) [31]	Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years	Multivariable logistic regression	ROC	The strongest independent predictors of hospital admission were age, sex, triage category, mode of arrival, the performance of blood tests, chief complaint, ED revisit, type of specialist, phlebotomized blood sample and all vital signs
Hong et al.,(2018) [32]	Predicting hospital admission at emergency department triage using machine learning	logistic regression (LR), gradient boosting (XGBoost), and deep neural networks (DNN)	AUC	Models trained on the full set of variables yielded an AUC of 0.91 for LR, 0.92 for XGBoost, and 0.92. An XGBoost model built on ESI level, outpatient medication counts, demographics, and hospital usage statistics yielded an AUC of 0.91 (95% CI 0.91–0.91).

O'donovan et al., (2017) [33]	Machine Learning Generated Risk Model to Predict Unplanned Hospital Admission in Heart Failure	Random Forests, Xgboost, and Treenet	risk score	the model correctly predicted outcomes for 12189 (84%) patients (c-statistic 0.77)
Leegon et al., (2006) [34]	Predicting Hospital Admission in a Pediatric Emergency Department using an Artificial Neural Network	ANN	ROC	The AUC for the training set was 0.909 slightly different with the test set 0.907 as well as the validation set 0.897.
Sun et al., (2011) [35]	Predicting Hospital Admissions at Emergency Department Triage Using Routine Administrative Data	Logistic Regression	ROC	Age, PAC status, and arrival mode were most predictive. (ROC) curve was 0.849 (95% confidence interval [CI] = 0.847 to 0.851)
Peck et al., (2012) [36]	Predicting Emergency Department Inpatient Admissions to Improve Same-day Patient Flow	expert opinion, naïve Bayes conditional probability, and a generalized linear regression model	ROC	Of the three methods considered, logit-linear regression performed the best, with a receiver operating characteristic (ROC) area under the curve (AUC) of 0.887, an R2 of 0.58
Golmohammadi (2016) [37]	Predicting hospital admissions to reduce emergency department boarding	Logistic regression, ANN	Confusion metrics	An admission prediction model based on demographic and clinical determinant factors can accurately estimate the likelihood of inpatient admission

2.4 Literature Gap

Even though many studies and scientific papers, as listed above, have already discussed forecasting patient demand with various ML methods, most of them developed several ML models separately and individually. The best model was selected after comparing the prediction result of each ML model. Apart from these typical approaches, a few recent papers [38-39] try to develop hybrid ML models, which are the combination of two or more ML models. The expectation in combining the ML models is to get more optimal prediction result rather than comparing and selecting the single best ML model.

In a study [38], a hybrid ML model is developed by combining ARIMA-LR. The rationale of the ARIMA-LR combination is the ability of ARIMA and also LR in capturing the seasonal trend and effect of predictors. The result of [38] shows that ARIMA-LR model outperformed several widely used models such as the generalized linear model (GLM), ARIMA, ARIMAX, and ARIMA–ANN. In another study [39], the experiment result shows a contradiction with the previous claim of ARIMA-LR superiority. Using ARIMA–ANN, the study [39] reveals that their model could outperform others which are LR, ARIMA, ANN, exponential smoothing,

and ARIMA-LR. Although there is the contradiction result between these two studies [38] [39], they both share the interesting commonality which is that hybrid ML model can make better forecasting patient demand in comparison to a single ML model. Hence, there are still plenty of other hybrid ML models that can be explored besides ARIMA-LR and ARIMA-ANN.

In this thesis, a hybrid model using SARIMAX and Gradient Tree Boosting will be developed to forecast patient demand. The selection of the two algorithms has several reasons and rationales. The first, SARIMAX, which is a variation of ARIMA (a classical model for time series forecasting), offers more parameters for capturing time series patterns such as trends and seasonality. Moreover, SARIMAX is able to accommodate external or exogenous data as its features prediction. In addition, a recent study [23] shows that SARIMAX has the best predicting power. The second, Gradient Tree Boosting or its variance XGBoost is a novel ML model, and it has widely become a top ML technique used among data scientist in industry[5]. Only one paper [24] attempted to build forecasting patient demand model using this kind of method. Interestingly, the same study [24] shows that the gradient boosting model yielded the best forecasting results. The third reason, the research in forecasting patient demand using hybrid SARIMAX-Gradient Tree Boosting has not yet found at the best of my knowledge. Finally, the idea of a hybrid model between a classical ML time series and a novel technique ML has become interesting topics because several researchers from the outside of healthcare domain have implemented similar approaches [40-43]. The more detail information on how to implement the Hybrid model with SARIMAX and Gradient Tree Boosting will be explained in Chapter 3.

In predicting inpatient admission, ML models based on ensemble learning such as AdaBoost, Gradient Boosting, Random Forest, and Extra Tree are selected for three main reasons. The first reason is similar to the previous paragraph in which ensemble algorithms such as Gradient Tree Boosting is a novel ML model, and it has widely become a top ML technique used among data scientist in industry. The second reason is the ability of ML models based on ensemble learning to handle complex non-linear patterns, which usually occur in a real dataset. The last reason is based on the previous literature study that shows the higher performance of this algorithm in comparison with others, in particular to linear regression models [32]. The more detail information on how to implement the ML models based on ensemble learning will be explained in Chapter 3.

2.5 Machine Learning

Machine learning is considered as a subset of artificial intelligence that uses many statistical techniques to give computer systems the ability to learn the pattern and rules from data without being explicitly programmed. In general, machine learning is divided into three main categories, as can found in Table 4 below. In this thesis, supervised learning is applied to forecasting patient demand and predicting inpatient admission. Forecasting patient demand can be classified as regression method while predicting inpatient admission is part of the classification. However, regression for patient demand has to be specially treated because unlike the normal or typical regression, the data is in time series format. The further description about regression with time series data and also classification will be explained in the chapter methodology.

[5] (2017)). XGBoost, a Top Machine Learning Method on Kaggle, Explained. Retrieved May 6, 2019, from https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html

Table 4: Machine Learning category, method, and algorithm

Category of Learning	Characteristics	Method	Example of Algorithms
Supervised Learning	- Labeled data - Direct feedback - Predict outcome/future	a. Regression b. Classification (Binary & Multi)	- Linear regression - Bayesian linear regression -Decision Tree - Decision forest - Support vector machine
Unsupervised Learning	- No labels - No feedback - Find hidden structure in data	a. Clustering	- K-means
Reinforcement Learning	- Decision process - Reward system - Learn a series of action	a. Markov decision process	- Monte Carlo - Q-Learning

3. Methodology

In this section, the methodology used in this thesis will be discussed. This methodology aims to address and answer the subquestions 1, 2, and 3 of the research questions.

3.1 CRISP-DM

To provide solid arguments for answering the research questions, having a solid, standard, and systematic approach and methodology will fundamentally help. Among the many available methodologies, CRISP-DM[6] is used as the main methodology and research framework throughout this thesis. The main reason is the flexibility of CRISP-DM methodology so it can be applied on cross-industry from various domains. It is also an open standard process model that describes common approaches so that it can accommodate some required adjustments.

CRISP-DM has six major phases, namely Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. As described in Figure 5: CRISP-DM Framework, these phases provide end-to-end steps as a guideline in conducting any typical data analytics or data mining project, including machine learning. Data preparation and Modelling are two phases which need several iterations until they reach the desired output model. In the evaluation phase, checking the machine learning result against a business requirement or business understanding is essential as part of validation.

Figure 5: CRISP-DM Framework

Besides the high-level framework illustrated above, the more detail, standardized, and actionable steps can be broken down from the CRISP-DM Framework. For example, in the Data Preparation phase, at

[6] (n.d.). crisp-dm - The Modeling Agency. Retrieved May 12, 2019, from https://www.the-modeling-agency.com/crisp-dm.pdf

least four steps need to be performed: collect initial data, describe data, explore data, and verify data quality. Using CRISP-DM as a baseline, the actionable steps for each phase is improved with several adjustments according to the context and necessity required in this thesis, as shown in Figure 6. Following the actionable steps will answer the subquestions 1, 2, and 3. The detail of each step will be explained in the following sections.

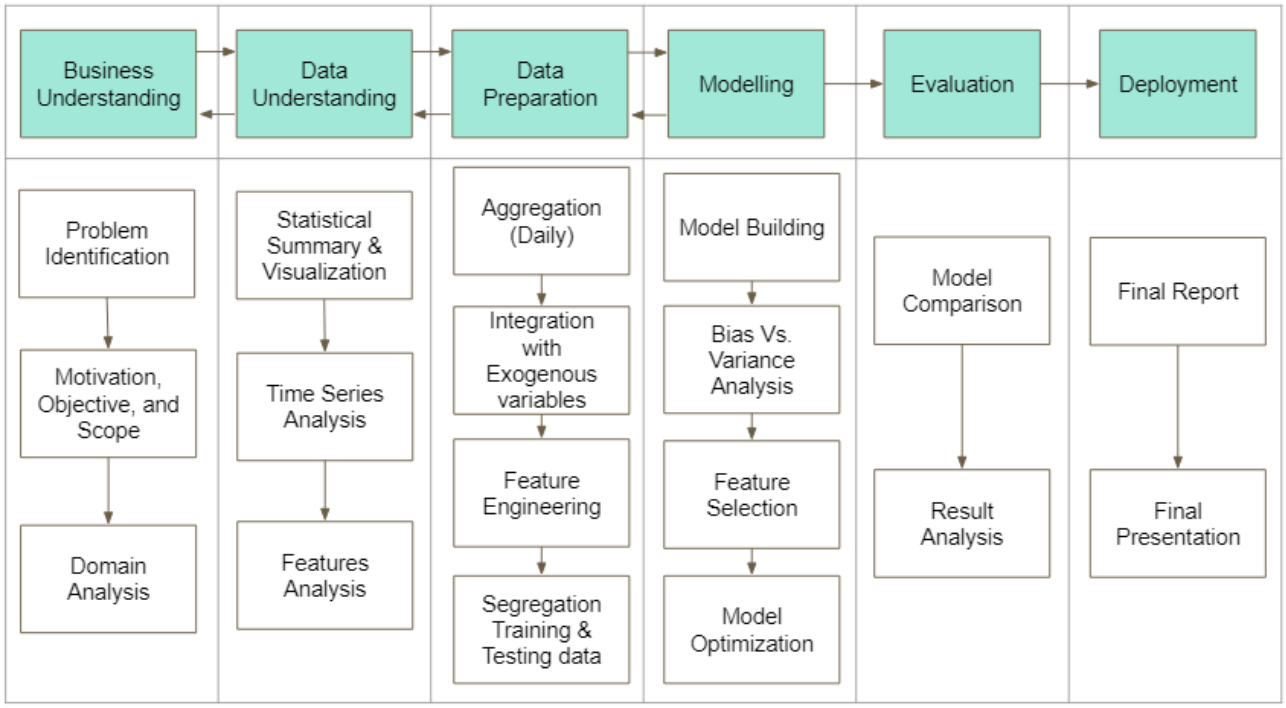


Figure 6: Improved and adjusted actionable steps based on CRISP-DM

3.2 Business Understanding

In Business Understanding phase, as described in Figure 6, there are three main steps need to be conducted. The first step is to identify the problem currently faced by the stakeholders. Once the problem has been identified, the motivation and objective of the research can be derived. Besides, the research scope also needs to be agreed and confirmed among all relevant parties.

One of the challenges in Business Understanding phase for ML project is how to identify and scrutinize the relevant information or features in the early stage. Putting all the information from the raw data might not be the best option. A study [44] proposes domain analysis for the development of prediction instruments. Through domain analysis, the same study argues that features with adequate predicting power can be identified.

In this thesis, domain analysis is performed in two methods: Literature review and Interview. Extensive literature review in chapter 2 was carried out by listing all relevant research papers in the acute care domain and categorizing them based on the research objective. In addition, the literature list table was summarized and presented by highlighting the features used on each paper and also the relevant findings. As a result of that, prospective features can be collected as the references for developing ML forecasting model. Moreover, literature review within the healthcare or acute care domain is beneficial for narrowing down the ML models among the plethora of ML algorithms. Besides the extensive literature review, an interview with domain experts is also conducted. The several aims of the interview are to identify the

potential features, to confirm the usefulness of exogenous features, to find out more empirical explanation about some doubtful data points, and also to understand the patient flow between ED and GP-Post.

3.3 Data Understanding

To properly understand the true characteristic of the data, several actions have to be performed through Exploratory Data Analysis (EDA). Looking at the statistical summary is the basic yet mandatory step in analyzing the data. Several important information can be revealed from the statistical summary of the data, such as the mean, median, and standard deviation. Statistic summary also includes the histogram to visualize the distribution of the data. Through visualization, any hidden pattern which might not be captured by numbers can be spotted on, such as the unusual data points or outliers. Besides, a specific analysis in relation to univariate time series will also be performed using the ADF test, and ACF and PACF plotting. The result of EDA will be discussed in detail in chapter 4.

Another important step in understanding the data is correlation analysis. In this study, Pearson-Correlation is used to analyze the correlation among variables or features. Pearson Correlation generates correlation coefficient between 1 and -1. The closer correlation coefficient to 1 indicates a strong and positive correlation between the two series variables. The closer correlation coefficient to -1 also implies a strong correlation between the two series variables, but it is in the negative or reverse way. The correlation coefficient between 1 and -1, especially the ones closer to 0, indicate a weak or even none correlation. The interpretation of the Pearson correlation coefficient is presented in Table 32 in the Appendix. The interpretation by Chan et al. [18] will be mainly used because it is more related to this research. The result of Pearson correlation will also be discussed in detail in Chapter 4.

Besides correlation, time series will be decomposed into several components for further analysis. Since the data format is in time series, analyzing its trend, cycle, season, or residual is also important steps. Understanding time series pattern in detail might significantly help in building a forecasting model with SARIMAX. The detail explanation of EDA is also discussed in Chapter 4.

3.4 Data Preparation

Before dataset can be thrown to ML algorithms for building a prediction model, several preparation steps need to be performed, depending on the format of the raw data. Since ED's and GP-Post's data were recorded based on the primary key, aggregating the primary key on a daily basis is required to calculate the total number of the patient. After that, the data has to be integrated with exogenous variables such as weather or pollen. The integration process is done by inserting more columns on the dataset. Besides weather, other relevant features are calendar variables such as weekend, weekday, holiday and so on.

Furthermore, as a part of the feature engineering process, lagged value-columns will be derived from each feature mentioned above by shifting rows seven times. This step is required to ensure that the respective feature's values are available (by using the past data) for forecasting patient demand on the next day or 1 day ahead. Also, the lagged value-columns are expected to provide more pattern and information about the data. As a result of the shifting action, there are rows in the dataset, which becomes blank or NULL. Removing these rows are necessary to avoid any error thrown by the ML model during the training process. Next step, some features, except Boolean and Dummy features, require normalization process through transformation by scaling each feature to a given range. Amongst many normalization techniques,

MinMaxScaler[7] from scikit-learn ML library is preferred because of its ability in handling the Non-Gaussian or Not normally distributed data. It essentially shrinks the value range such that eventually, the range is only between 0 and 1.

Finally, the dataset will be split into two part: training dataset and testing dataset. During the training process, the cross-validation technique will be used to avoid bias and overfitting [49]. Generally, cross-validation involves 5 steps: (1) shuffling the dataset randomly, (2) splitting the dataset into k folds, (3) split each fold as training and testing, (4) fit ML algorithm on each training fold set and evaluate it on testing fold set, (5) summarize the model performance by calculating the mean of model evaluation scores (e.g., mean square error or MSE). These five steps can be used in ML classification problem such as predicting inpatient admission. However, in classical time series modeling (e.g., using ARIMA and its variations), splitting between training and the testing dataset is not random, it is done in a sequential manner with the higher proportion on the training dataset and a smaller proportion on the testing dataset. Moreover, in analyzing the prediction result of classical time series modeling, the specific metric evaluation, namely AIC or Akaike's Information Criterion, is also used for two reasons. First, AIC is a default built-in metric evaluation in ML library used for forecasting time series which is relevant in forecasting ED and GP-Post patient demand. Secondly, this book [50] argues that AIC and other common evaluations (e.g., RMSE, MAE) will lead to the same model selection for the large time series, which is relevant to this study.

3.5 Modelling

Modeling is the crucial phase in the ML process. During model building, several ML algorithms are selected and configured in such a way to learn and generalize the pattern from the training dataset. As a result of the training, the ML model is created so it can be evaluated against the testing dataset. As already explained in Chapter 2, the hybrid model with SARIMAX and Gradient Tree Boosting was selected as the main algorithms for forecasting ED and GP-Post daily patient demand. Besides forecasting, ML classification with ensemble learning algorithms is also used to predict inpatient admission. The modeling approach on each algorithm will be further explained in the following section.

Another action step in the modeling phase is feature selection. It is the process to reduce the number of variables or features used in a model by selecting the most significant of them without sacrificing the model performance. According to this study [51], three main reasons to perform feature selection are the interpretability of a model, fasten model execution, and reduce overfitting. The feature selection approach on each ML algorithm will also be explained in the next section.

3.5.1 ML model for forecasting ED and GP-Post patient demand

To build an ML model for forecasting patient demand, the hybrid approach will be performed. It consists of several steps, as described in Figure 7. Firstly, time series dataset is trained with SARIMAX to get the forecasting model and residual values. The motivation in selecting SARIMAX as the main ML model is its ability in handling various types of time series, including time series with seasonal components. Unlike ARIMA, SARIMAX can be fed with external variables, so it makes SARIMAX more flexible and suitable for building a forecasting model with external variables. Secondly, Feature Selection with Lasso will be performed to reduce the insignificant features and improve SARIMAX performance. The next step, residual

[7](n.d.). sklearn.preprocessing.MinMaxScaler — scikit-learn 0.21.1 .... Retrieved May 23, 2019, from http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

values then will be forecasted by Gradient Tree Boosting. As a robust non-linear model, Gradient Tree Boosting is expected to capture the hidden and Non-linear correlation in residual which was unable to be predicted by SARIMAX. Finally, the forecasting model by SARIMAX is added with the residual forecasting by Gradient Tree Boosting. The theoretical background of each model is explained in the following section.

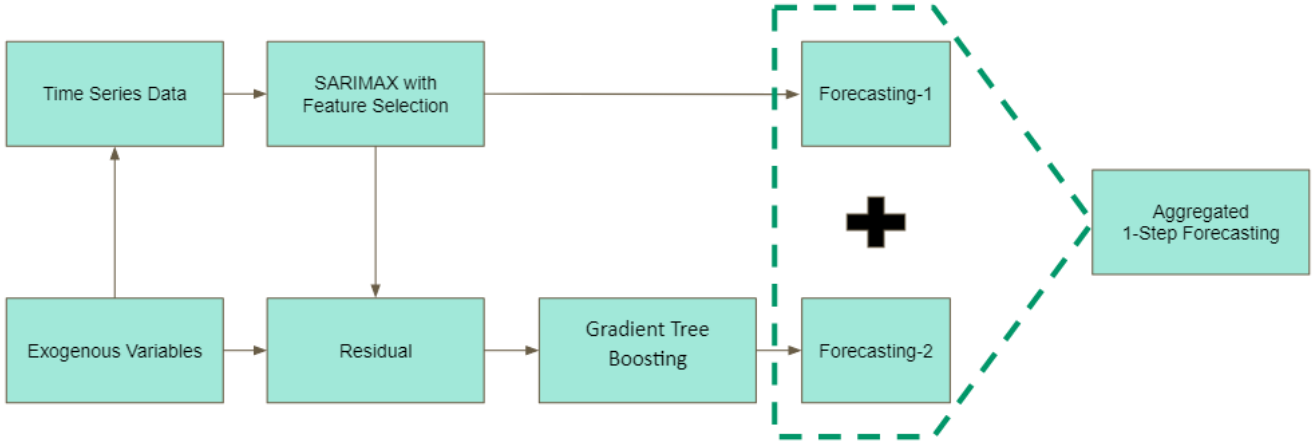


Figure 7: Hybrid SARIMAX-Gradient Tree Boosting

3.5.1.1 Forecast patient demand with SARIMAX

SARIMAX is a parametric time series in which it requires a priori knowledge and assumption about the data distribution of time series such as stationarity. Moreover, SARIMAX can be described as Seasonal(S) ARIMA with the addition of exogenous or external variable (X). Seasonal ARIMA, as the name suggested, is the extension of the ARIMA model with the seasonal component. Further, ARIMA or AutoRegressive Integrated Moving Average consists of three statistical components: (a) autoregression which is the linear function of lagged or passed value against itself, (b) integration which indicates the number of differences required to guarantee the stationarity, (c) moving average or lagged the forecast error.

Mathematically, part (a) can be formulated as below:

$$y_t = c + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \varepsilon_t$$

In a more general term, the above formula is referred to an AR(p) model or an autoregressive model of order p. The ε_t component is called white noise, error, or residual. Actually, this white noise can further be used to formulate part (c) as below:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

The second formula or part (c) is referred to an MA(q) model or a moving average model of order q. By "Integrating" or combine differencing between autoregression AR(p) and moving average MA(q), it forms a Non-Seasonal ARIMA which can be formulated as below:

$$y'_t = c + \theta_1 y'_{t-1} + \theta_2 y'_{t-2} + \ldots + \theta_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where y'_t is the differenced series, and it might have been differenced more than once [45].

The above equation is generally shortened and referred as ARIMA(p,d,q) model whereas p is the order of the autoregressive part, d is the degree of first differencing involved, and q is the order of the moving average part.

However, ARIMA(p,d,q) still does not include Seasonal and Exogenous, the two additional components which make it becomes SARIMAX. Generally, the notation for SARIMA or SARIMAX model can be written as ARIMA(p,d,q)(P, D, Q)m in which (p,d,q) is the Non-Seasonal part while (P, D, Q) is the Seasonal part with m as the frequency. It can also be written as SARIMAX(p,d,q)x(P,D,Q,m) which is used for this thesis.

According to the authors [46], SARIMA and SARIMAX models were formulated as two equations below:

$$\underbrace{(1-\Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p)}_{This\ part\ is\ AR(p)} \times \underbrace{(1-\beta_1 B^s - \beta_2 B^{2s} - \ldots - \beta_p B^{ps})}_{This\ part\ is\ ARs(P)} \times \underbrace{(1-B)^d}_{This\ part\ is\ I(d)} \times \underbrace{(1-B^s)^D}_{This\ part\ is\ Is(D)} y_t =$$

$$C + \underbrace{(1 - \Psi_1 B - \Psi_2 B^2 - \ldots - \Psi_q B^q)}_{This\ part\ is\ MA(q)} \times \underbrace{(1 - \theta_1 B^s - \theta_2 B^{2s} - \ldots - \theta_Q B^{Qs})}_{This\ part\ is\ MAs(Q)} \varepsilon_t$$

Where,

- B = backward shift operator (such that B • y(t) = y(t−1))
- AR(p) = autoregressive part of order p, ARs(P) = seasonal autoregressive part of order P
- MA(q) = moving average part of order q
- I(d) = differencing of order d
- Is(D) = seasonal differencing of order D
- MAs(Q) = seasonal moving average part of order Q
- C = intercept term
- t = time (e.g., the day for daily collision data)
- s = period of seasonal pattern appearing (e.g., 7 days for daily collision data)
- φ, β, ψ, and θ = model parameters to be estimated
- and ε_t = error term.

SARIMAX equation which includes γ as the parameter of xreg (external variables) to be estimated:

$$y'_{t,xreg} = y'_t + \gamma \bullet \varphi_p(B) \bullet \beta_p(B^s) \bullet (1 - B)^d \bullet (1 - B^s)^D x_t$$

where

$$\Phi_p(B) = (1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p)$$

$$\beta_p(B^s) = (1 - \beta_1 B^s - \beta_2 B^{2s} - \ldots - \beta_p B^{ps})$$

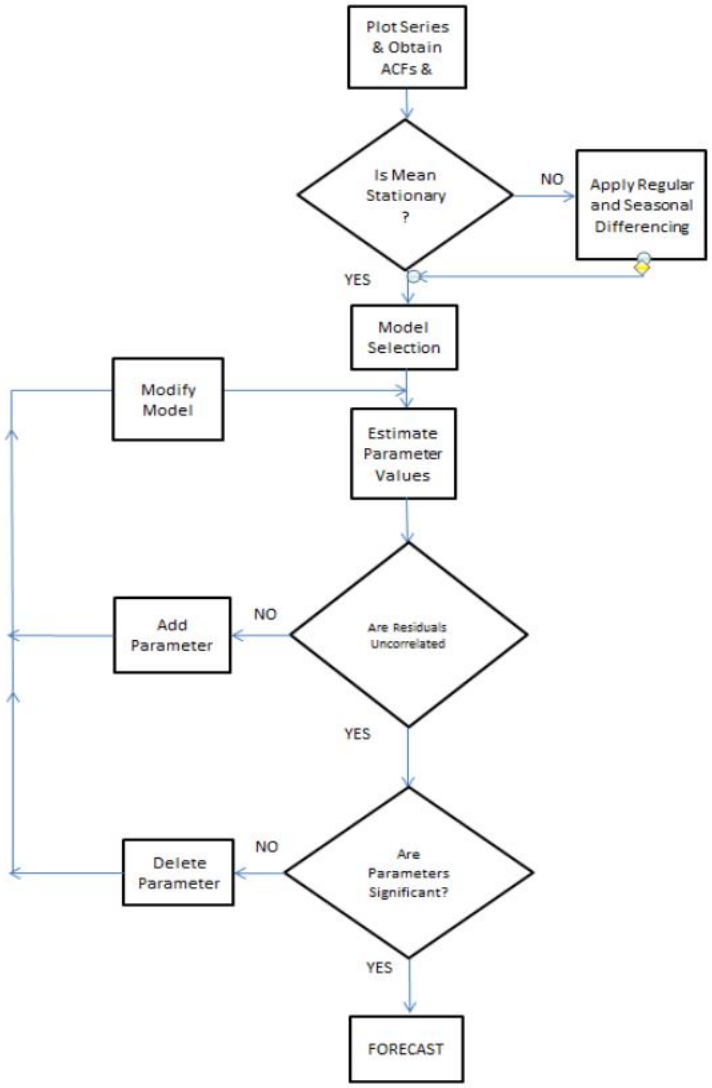


Figure 8: Box-Jenkins modeling flow[8]

In modeling with ARIMA, SARIMA, SARIMAX, and its other variations, several approaches have been developed by researchers. The most popular approach is Box-Jenkins modeling, as described in Figure 8. It is started by obtaining ACF of time series then it checks if the series data is stationary. Several techniques, such as regular and seasonal differencing, can be applied to non-stationary data. Initially, the estimation of some parameters (e.g., p, d, q) is required. Modifying parameters, either by adding or removing, need to be performed iteratively based on residual correlation analysis and statistical p-value. Once the model returns sufficient result, the final step is to forecast the future.

[8] "Build or Make your own ARIMA forecasting model? - Autobox Blog." 28 Jan. 2013, https://autobox.com/cms/index.php/blog/entry/build-or-make-your-own-arima-forecasting-model. Accessed 11 May. 2019.

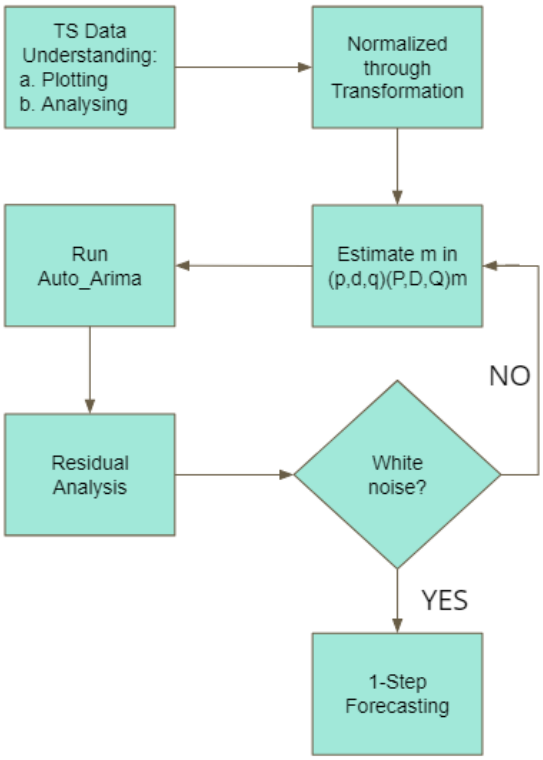


Figure 9: Single model SARIMAX approach

With the advancement of computing power and ML tools, it is possible to find the optimal parameter configuration of ARIMA, SARIMA, or SARIMAX automatically. Consequently, the approach to find the best configuration of SARIMAX is slightly different from the approach of Box-Jenkins modeling. As illustrated in Figure 9, the step of parameter estimation can be performed automatically. Moreover, in many cases, regular and seasonal differencing can also be automated. Although the Box-Jenkins approach still be relevant in this thesis to some extent, auto Arima function in a python library called pmdarima[9] is mainly used for heavy lifting task in searching for the most optimal configuration. As illustrated in Figure 9, the step of parameter estimation can be performed automatically. Moreover, in many cases, regular and seasonal differencing can also be automated. So, as adopted from this study [45], the automate searching for the most optimal configuration of SARIMAX is the preferred approach for this thesis. A more detail illustration on this approach will be explained in Chapter 5 Experiment Design and The Implementation.

3.5.1.2 Bias Vs. Variance Analysis

Bias can be simply defined as an error made by the ML algorithm in predicting the target output or the dependent variable. The high bias that occurs during the training implies that the ML algorithm is unable to identify the relevant relation between the features and the dependent variable on the training dataset. This phenomenon is called underfitting [66], which occurs when the ML model is unable to capture the underlying pattern of the data. On the other hands, the variance can be simply defined as an error made by the ML algorithm from the randomness in training. As a result of that, the ML model can extremely fit its prediction to the training dataset, but it poorly generalizes the unseen or the testing dataset. As opposed to

[9] "pmdarima: ARIMA estimators for Python — pmdarima 1.2 ... - alkaline-ml." https://www.alkaline-ml.com/pmdarima/. Accessed 11 May. 2019.

the previous underfitting term, this phenomenon is called overfitting. In other words, overfitting occurs when the model captures the noise and the outliers in the data along with the underlying pattern. Underfitting ML models usually have low variance and a high bias while the overfitting models usually have high variance and low bias as described in Figure 10.

Performing bias vs. variance analysis is important for at least two reasons. The first is to avoid the underfitting and overfitting phenomenon, which can reduce the validity of the ML model. The second is to find a trade-off point in selecting the ML model, as illustrated in Figure 10, where a simple model tends to have a high error, and complex model tends to have high variance.

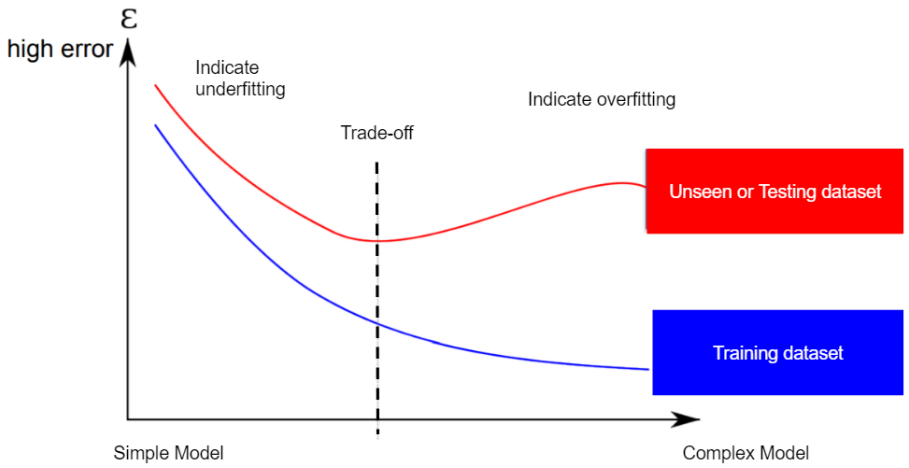


Figure 10: Bias Vs. Variance Trade-off

3.5.1.3 Feature Selection for SARIMAX with Lasso

Although ARIMA, SARIMA, SARIMAX and other of its variations provide coefficient values with statistically significant indicator, it is not considered as a proper approach for feature selection because statistically significant indicator mainly aims to test the hypothesis, not to select the features[10]. In other words, there is the case in which an insignificant coefficient can be useful for forecasting is possible to happen while the reverse case can occur as well. Therefore, in this approach, feature selection with Lasso is preferred for several reasons [51].

First, Lasso can be used to remove the insignificant coefficients. In doing so, it also reduces the variance without necessarily increasing the bias. Besides, eliminating the unimportant coefficients can automatically simplify the model dimension, in particular, the model with many features. As a result, how a complex model works can be humanly understood and interpreted. Furthermore, a simplified model with a few numbers of features can potentially avoid overfitting better. Lastly, in comparison with other feature selection method especially in regularization method such as Ridge, research [20] explains that Lasso is more preferable in reducing the number of features because it is mainly intended to shrink the coefficient of less important variables or features to zero. Moreover, in this thesis, the more accurate prediction of patient numbers is more emphasized over features explanation to avoid overcrowding event. Hence, Lasso is primarily used as a Feature Selection method, while another statistical method called the variance inflation factor (VIF) is also used to cross-check the presence of multicollinearity effect.

[10] (2011). Statistical tests for variable selection | Rob J Hyndman. Retrieved May 22, 2019, from https://robjhyndman.com/hyndsight/tests2/

The simplified steps in performing Feature Selection is described in Figure 11. Firstly, Lasso is defined as the feature selection model. After that, the fitting function is called to build a prediction model by feeding the total features and one target variable. Once the fitting is done, the coefficient values of each feature can be generated in which the values can vary from negative values, zero, or positive values. Zero coefficient values on certain features indicate that the respective features have no contribution to the prediction values based on Lasso. Therefore, taking out these zero coefficient features will generally not affect the prediction power of the ML model. However, the remain non-zero coefficient features can still make a complex ML model with many features without necessarily improving the ML model performance.

Reducing further the remain non-zero coefficient features is done by iteratively executing SARIMAX model from only one feature until covering all significant coefficient features (Non-Zero weight features). During iteration, RMSE is calculated for training and testing dataset. The list of RMSEs will be used to plot the trade-off between selecting a simple model with a few features or complex model with lots of features. Performing this step is important for at least three reason: (1) Select an optimal number of features from the initial Lasso-Feature Selection with the lowest RMSE, (2) Find the best SARIMAX model in relation to the optimal number of features, (3) Avoid multicollinearity among the selected features to make the ML model more robust. If the RMSE result does not significantly change by adding more features, then the fewer features are preferred for simplicity, interpretability, and robustness. A more detail illustration on this approach will be explained in Chapter 5 Experiment Design and The Implementation.

3.5.1.4 Forecast SARIMAX's Residue with Gradient Tree Boosting

Gradient Tree Boosting is a type of decision trees in a machine learning technique for regression and classification problems. Gradient Tree Boosting applies the ensemble learning technique which aims to create a single strong learner by combining various weak learners to optimize the prediction result. By applying such technique, GBRT has several advantages[11] such as its ability in handling heterogeneous features, its robustness to handle outliers through robust loss functions, and arguably versatile and flexible for solving various types of ML problems[12].

Forecasting time series with Gradient Tree Boosting requires data restructuring from time series data type into a supervised data type. It can be done by adding dummy variables or features to capture DateTime related domain such as lagged values. Once the data is already in a supervised data format, the general approach in building an ML regression model can be followed to execute any kinds of ML algorithms, including Gradient Tree Boosting.

---

[11] (n.d.). 1.11. Ensemble methods — scikit-learn 0.21.1 documentation. Retrieved May 22, 2019, from http://scikit-learn.org/stable/modules/ensemble.html

[12](n.d.). XGBoost, a Top Machine Learning Method on Kaggle, Explained. Retrieved May 22, 2019, from https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html

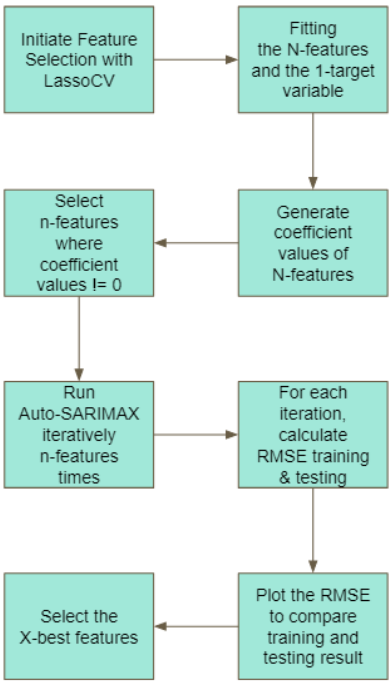


Figure 11: Steps of Feature Selection with Lasso

3.5.1.5 Feature Selection for Gradient Tree Boosting with Feature Importance

Besides using Lasso for feature selection as already explained above, various feature selection techniques are available. Instead of looking at coefficient significance, other ML models, especially in decision trees family, have an in-built attribute which provides feature importance score. By setting an arbitrary threshold or through the iteration process, several important or significant features can be selected while conversely, the insignificant features can be taken out from the model. Based on the literature study, several papers also use similar approach [52-54].

However, unlike Lasso, the result of feature selection with feature importance score of Gradient Tree Boosting will be used for solely improving the ML model performance. The feature selection result will not be used for interpreting the selected features against the prediction outcome for several reasons. Firstly, as illustrated in Figure 7, the Gradient Tree Boosting result has a high dependency on another process because it primarily aims to forecast the residue of SARIMAX. Secondly, the mentioned papers [52-54] have more emphasis on the improvement result made by ML model after feature selection rather than firmly concluding the effect of selected features against the prediction outcome. Another paper [67] tried comparing linear (e.g., Lasso), and Non-linear (e.g., Gradient Boosting) feature selection methods highlights that both eventually yield similar values after adding more features. Hence, making a conclusive statement that only selected features affecting the prediction outcome based on feature importance score of Gradient Tree Boosting is problematic with many uncertainties as already explained.

The steps in performing feature selection with Gradient Tree Boosting is quite similar to Lasso in Figure 11. However, Gradient Tree Boosting uses GridSearchCV in finding the optimal parameters. Besides, Gradient Tree Boosting does not require iteration to check trade-off between bias and variance since Gradient boosting is fairly robust to over-fitting. The feature selection steps with Gradient Tree Boosting is described in Figure 12.

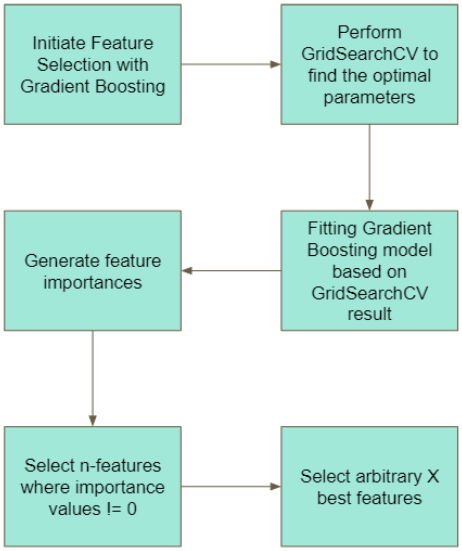


Figure 12: Steps of Feature Selection with Gradient Tree Boosting

3.5.2 ML models for predicting inpatient admission

The purpose of ensemble learning algorithms is to combine the predictions of several base estimators so that it improves the robustness over a single estimator[13]. Generally, ensemble learning can be classified into two main groups, namely bagging and boosting. There are two key principles in bagging or averaging method, namely, (1) to build several estimators independently and then (2) to average their predictions. By combining several estimators and averaging their predictions are usually better than any of the single base estimator because its variance is reduced. Conversely, in boosting methods, base estimators are built sequentially instead of independently as in bagging method. Consequently, boosting methods try to reduce the bias of the combined estimator. The motivation of boosting methods is to produce a powerful ensemble prediction by combining several weak models.

ML models based on ensemble learning, in particular, Gradient Boosted Regression Trees (GBRT), will be used for predicting or classifying two binary states of inpatient admission: admitted (1) or Not-admitted (0). Four different ensemble ML models consist of a Boosting method, and Bagging methods will be built and compared. AdaBoost and Gradient Boosting are part of Boosting methods while Random Forest and Extra Trees are part of Bagging methods. Based on several measurement metrics, such as accuracy and recall, the best ML model will be selected. Furthermore, the selected ML model is optimized to improve the prediction result through a grid search.

3.6 Evaluation

In the evaluation phase, the result of ML models will be analyzed and compared through various evaluation metrics. Forecasting patient demand and predicting inpatient admission use the different evaluation metrics because the former one is part of the regression method, while the latter is part of the classification method. The explanation of both is presented in the subsection below.

---

[13] (n.d.). 1.11. Ensemble methods — scikit-learn 0.21.1 documentation. Retrieved May 22, 2019, from http://scikit-learn.org/stable/modules/ensemble.html

3.6.1 Evaluation Metric for Forecasting Patient Demand

There are numerous evaluation metrics to measure the performance of their ML model. In a recent study [47], at least 17 evaluation metrics were used by ML researchers. Among these 17, MAPE, RSME, and MAE are ranked in the top 3 as the most popular evaluation metrics. Mean Absolute Percent Error or MAPE measures the magnitude of the error compared to the magnitude of the data in a percentage format. So, the lower percentage of MAPE indicates the better performance of the ML model. MSE works by squaring the difference between the measured and the prediction. It also disregards the difference between over-prediction and under-prediction. Similar to MAPE, the lower and closer MSE value to 0 implies the better accuracy of ML model prediction. Another version of MSE is RMSE or Root Mean Squared Error. It is essentially derived by simply squaring root MSE or Mean Square Error. Other metrics, Mean Absolute Error (MAE), measures in the absolute value of the difference between the measured and the prediction. The ML model can be considered as a good prediction model if MAE score is relatively small and close to 0. Besides the mentioned evaluation metrics, AIC (Akaike's Information Criterion) is also commonly used for measuring time series forecasting.

In this thesis, AIC, MAPE, RMSE, and MAE are used for the main evaluation metrics due to their simplicity, interpretability, and popularity. Table 5 below provides the mathematical formula of the four-evaluation metrics, including their pros and cons.

Table 5: Performance Evaluation metrics

Name	Formula	Pros(+)/Cons(-)
AIC (Akaike's Information Criterion)	$\mathrm{AIC} = -2\log\mathcal{L} + 2\mathrm{p}$ where $\mathcal{L}$ is the maximized likelihood using all available data for estimation and P is the number of free parameters in the model[14]	+ Commonly used in time series + in-built and default metric in forecasting library - Applicable for evaluating the training data
Root Mean Square Error (RMSE)	RMSE = $\sqrt{MSE}$=$\sqrt{mean\ (y_t - \hat{y}_t)^2}$	+ same scale as observations - unscaled - sensitive to outliers
Mean Absolute Error (MAE)	MAE = mean $(\lvert(y_t - \hat{y}_t\rvert)$	+ less sensitive to outliers + relatively simple - unscaled
Mean Absolute Percentage Error (MAPE)	MAPE = mean $(100 \times \frac{\lvert(y_t - \hat{y}_t\rvert}{y_t})$	+ scaled - need $y_t > 1$ - higher penalty on positive error - need a similar range for compared time series

[14] (2010, October 4). Cross-Validation - Rob J Hyndman. Retrieved May 22, 2019, from https://robjhyndman.com/hyndsight/crossvalidation/

3.6.2 Evaluation Metric for Classifying Inpatient Admission

Several metrics are available for measuring the performance of machine learning models on classification. One of them is confusion metric, which is the relationship table between machine learning prediction against the actual class or predefined label. The relationship consists of four parameters, namely: True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). Many measures or metrics can be derived based on these four parameters, but the four most popular ones are Accuracy, Precision, Recall, and F1-Score. The example of confusion metric used in this thesis is defined as described in Table 6, while the description and the formula of four parameters are provided in Table 7.

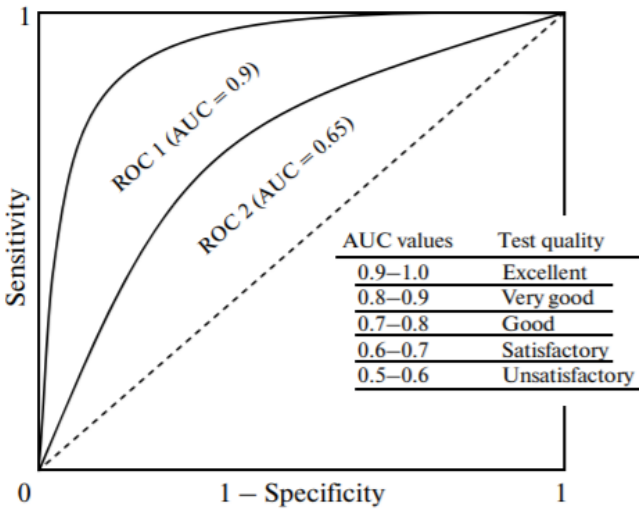
Table 6: The confusion metrics

The Confusion Metrics Composition		ML Output Prediction	
		Positive (1)	Negative (0)
The actual Label	True (1)	True Positive (TP)	False Negative (FN)
	False (0)	False Positive (FP)	True Negative (TN)

Table 7: Confusion Metrics types and formula

Metrics	Description	Formula
Recall	The proportion of all correct results returned by the model	TP/(TP+FN)
Precision	The proportion of true results overall positive results	TP/(TP+FP)
Accuracy	Measures a classification model as the proportion of true results to total cases	(TP+TN)/(TP+FN+FP+TN)
F1-Score	F-score is calculated as the weighted average of precision and recall between 0 and 1, where the ideal F-score value is 1	2*(Recall * Precision) / (Recall + Precision)

Another metric is AUC-ROC; AUC stands for Area Under Curve while ROC stands for Receiver Operating Characteristic. AUC-ROC measures the area under the curve plotted with true positives on the y-axis and false positives on the x-axis. AUC-ROC is useful because it provides a single number for comparing models of different types. The different values of AUC-ROC curves based on a study [48] are illustrated in Figure 13 below.



AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

Figure 13: ROC-AUC curves with different AUC values

3.7 Deployment

Reporting and Presentation are the two main activities in the deployment phase.

4. Exploratory Data Analysis (EDA)

In this section, Exploratory Data Analysis (EDA) will be performed through two types of analysis, which consist of univariate time series analysis and correlation analysis. The result of this section will answer the subquestions 4 in the research questions.

4.1 EDA for ED and GP-Post patient demand

The ED's initial raw data is the historical data of patient arrival from 2012 to 2018. It consists of 85,049 rows and 11 columns in which patient ID becomes the primary key. The list of ED columns can be found in Table 8.

Table 8: The ED raw data columns

Column Name	Format
ED Number	Numbers
Date of Birth	Date
Gender	Categorical: Man, Woman
Place of Residence	Categorical: Utrecht, Eindhoven, etc
Arrival Date Time	Date Time
Leaving Date Time	Date Time
Length of Stay	String
Urgency	Categorical: Green, Yellow, Blue, Red, Other
Treating Specialism	Categorical: CHI, INT, KIN, etc
Referral	Categorical: Ambulance, Hospital, etc
Destination	Categorical: Home, Admittance, Dead, Transferral

Unlike ED raw data, GP-Post raw data has a shorter period, only available from 2013 until 2017, which consists of 149,726 rows and 11 columns. Moreover, GP-Post raw data does not have leaving time, length of stay, Referral, and Destination. Two column names, Urgency, and ICPC code have a comparable function as in ED's Urgency and Treating Specialism, with different categorical values. The list of GP-Post columns can be found in Table 9.

Table 9: GP-Post raw data columns

Column Name	Format
Caller ID	Number
GP-Post number	Number
Year of Birth	Date

Gender	Categorical: Man, Woman
Consultation Type	Categorical: Consult, Phone, Visit, Other
Self-referral	Categorical: Yes, No
Arrival Date	Date
Arrival Time	Time
Holiday	Categorical: Yes, No
Urgency	Categorical: U0, U1, U2, U3, U4, U5
ICPC code	Categorical: Alphabetical code (D73, K01, R74, etc)

Weather data is one of the external data sources used for forecasting ED and GP-Post patient demand. Weather daily historical data were downloaded from knmi.nl[15]. Hupsel station (code: 283) is selected because it is the nearest station around the hospital and GP-Post location in Winterwijk area. Various weather information such as temperature, humidity, wind speed, and many others with a total of 39 variables, are provided by knmi.nl. However, only 29 variables will be used for building ML models due to the quality of their daily data completeness. Besides weather data, the daily spreading amount of pollen on the air is also included. The pollen data that consists of forty-four various plants in the Netherlands can be downloaded from elkerliek.nl[16]. Out of forty-four, only forty were used because of the quality of their daily data completeness. The full list of weather and pollen variables can be found in Appendix file.

After data pre-processing, the new dataset is formed with the dimension of 1454 rows and 1132 columns plus one target variable column. The ED and GP-Post features are the same. Therefore, the feature list of ED and GP-Post dataset can be found in Table 10.

Table 10: ED and GP-Post patient demand dataset

Column	Description
Patient arrival	The target variable of ED or GP-Post daily patient demand
Internal: ED and GP-Post related data	Features extraction of ED-related data: Treatment, Referral, Urgency, etc., including their lagged values. Also, The GP-Post related data: Urgency, ICPC code, etc., including their lagged values
Dummy: Calendar	The created variables based on the calendar such as Is_holiday, Is_weekday, sine and cosine of the cyclic period, etc., also including their lagged values

[15] (n.d.). KNMI - Daggegevens van het weer in Nederland - KNMI projects. Retrieved May 26, 2019, from http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi

[16] (n.d.). Pollentellingen - Elkerliek Ziekenhuis. Retrieved May 26, 2019, from https://www.elkerliek.nl/Elkerliek/Hooikoorts/Pollentellingen.html

External: Weather data	The daily weather-related data such as temperature, humidity, etc., also including their lagged values
External: Pollen daily data	The daily pollen data from various plants: Alnus, Betula, Corylus, etc., also including their lagged values

4.1.1 ED Univariate Time Series Analysis

ED patient arrival is the dependent or target variable, which is in daily time series format. Although the period of ED raw data is started from 2012 until 2018, only 2013 until 2017 is used as the dataset to fit the dimension of the external data period. The plot of daily ED patient demand can be found in Figure 14 below.

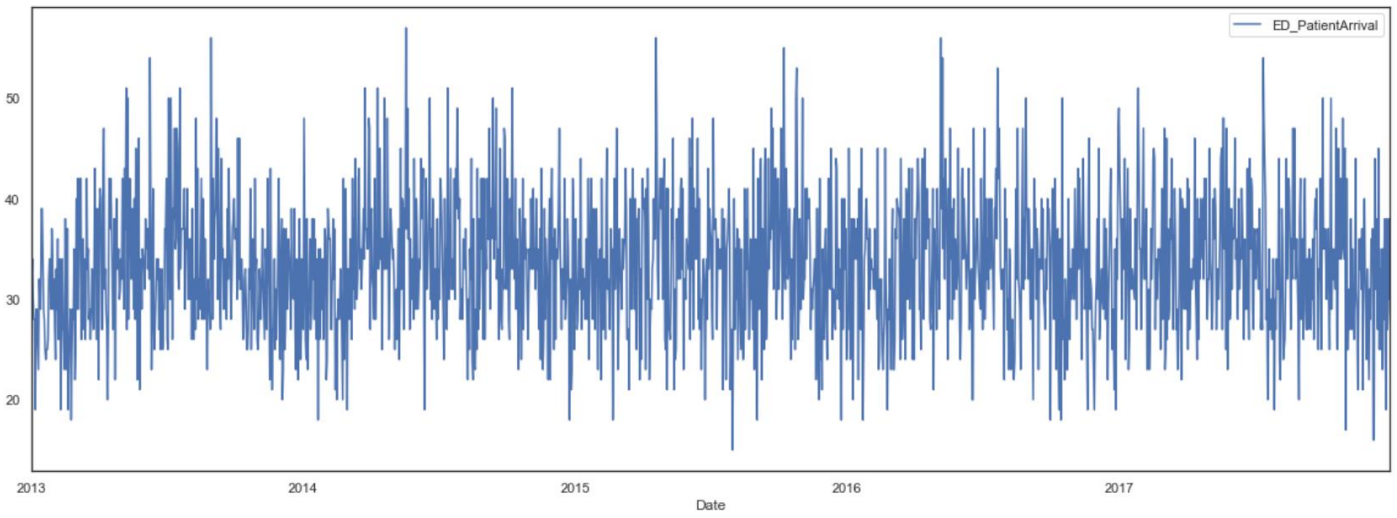


Figure 14: Daily ED patient demand

Although there is some fluctuation with up and down values along the 2013-2017 period, the range values are relatively stable from 20 to 50 patient number per day. Besides, it hardly spots any increasing or decreasing trends in the figure. However, slightly seasonal pattern from year to year can still be recognized from Figure 14. The more detail characteristic of ED patient demand can be found in Table 11 in which the summary statistic is provided not only for the whole years (2013-2017) but also for each year. The mean of each year in Table 11, which is stable around 33, is the reflection of flat-trend in Figure 14. Moreover, the standard deviations and the distribution of Min-to-Max values do not significantly change each year. These also suggest a similar pattern of ED patient demand from year to year.

Table 11: Statistical Summary of ED Patient Demand Dataset

Year	Count	Mean	Std	Min	25%	50%	75%	Max
All	1826	33.38	6.77	15	28	33	38	57
2013	365	32.63	6.56	18	28	32	37	56
2014	365	33.80	6.69	18	29	33	38	57

2015	365	33.62	6.88	15	29	33	39	56
2016	366	33.27	6.90	18	29	33	38	56
2017	365	33.58	6.78	16	29	33	38	54

However, looking at the histogram of ED patient demand in Figure 15 help in understanding the data characteristic better than simply conclude it from statistical summary in Table 11. The overall distribution data seems to follow the normal distribution (Figure 15). However, the frequency plot in each year as described in Figure 32 until Figure 36, shows quite different distribution pattern in each year. In other words, even though the daily average number of ED patient demand relatively flat around 33 from 2013 to 2017 as provided in Table 11, the actual daily demand distribution of each year is quite dynamic. The changes in ED patient demand might be influenced by various external factors such as weather or special events (e.g., holiday or festival day).

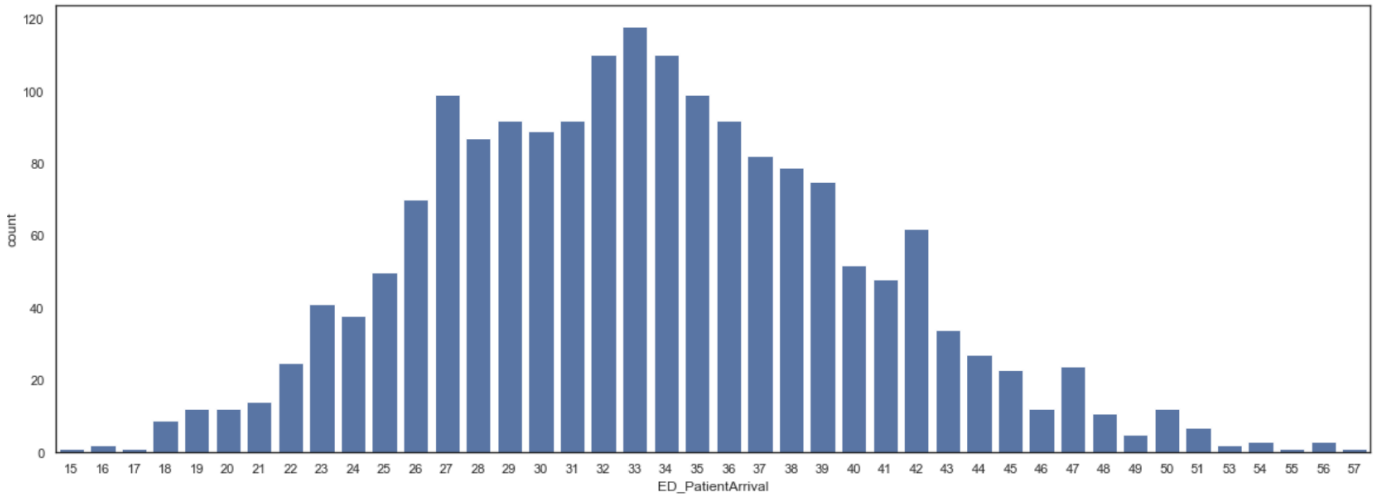


Figure 15: The frequency plot of ED patient demand in 2013-2017

Besides general statistical description, ED daily patient demand can also be analyzed through the decomposition of its time series components. Typically, the time series consists of four components: Observe, Trend, Seasonal, and Residual. These four-time series components can be classically modeled through the additive model, in which the four components are simply added, or multiplicative model, in which the four components are multiplied. In Figure 16, ED daily patient demand in 2017 was decomposed by the additive model. The Observe shows the actual time series value, which in this case, it seems to not suggest any specific patterns. Similarly, the Trend does not show long term positive or negative tendency; instead, it always fluctuates. Conversely, Seasonal indicates a strong pattern which might occur weekly. The Residual is the remainder or left-over values when the seasonal and the Trend have been subtracted from the data.

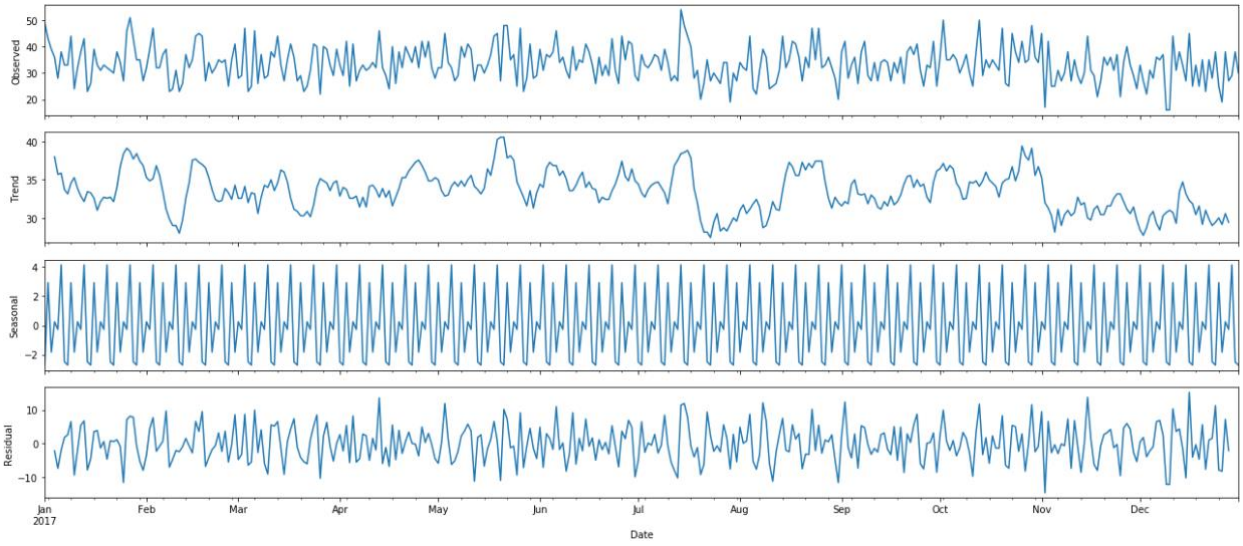


Figure 16: ED patient demand Time Series Decomposition with an additive model

Although classical time series decomposition above provided several insights, its simple method has limitations and problems [50]. Firstly, the classical time series decomposition merely assumes a fixed yearly pattern in the seasonal component, so it is unable to capture seasonal changes over time. Secondly, it is not robust to the unusual values caused by external factors such as the weather. Thirdly, the estimation of the trend tends to over-smooth the up and down events as can be found in the Figure above. Therefore, more advanced techniques are required than looking only at classical time series decomposition.

Another technique in time series analysis is to check the stationarity of the time series data. Stationarity indicates that the time series properties such as mean and variance do not depend on the time at which the series is observed. In other words, mean and variance are constant over time in stationary time series. White noise is an example of a stationary series because it looks similar at any point in time. As a result of that, a white noise series will have no predictable patterns in the long-term, and its plot looks relatively horizontal with constant variance. On the other hands, the existence of trend or seasonality will make time series Non-Stationary because they affect and change the values of time series at different times.

Besides looking at Figure 14 which seems to suggest the stationarity of ED patient demand series due to no visible trend, a more appropriate approach in checking stationarity is to perform the statistical test, called Augmented Dickey-Fuller test or ADF test. ADF test generates p-value that can indicate the stationarity if the p-value is less than the significant threshold level. Conversely, the p-value more than threshold value suggests that the series is Non-Stationary. The result in Figure 17 shows that p-value is zero or close to 0, which means it is less than 5% (0.05) significant threshold level. Moreover, taking a look at the test statistic and critical value yields the same conclusion. The statistic value ends up being -7.365, which is much smaller than the 5% critical value of -2.863. Therefore, the ADF test result provides statistical evidence that confirms the previous prediction based on the visualization analysis that ED patient demand series was indeed stationary.

```
TestStationaryAdfuller(Label_Y.ED_PatientArrival.values, 0.05)
```

```
Test Statistic  : -7.365
p-value  : 0.000
#Lags Used  : 20.000
Number of Observations Used  : 1805.000
Critical Value (1%)  : -3.434
Critical Value (5%)  : -2.863
Critical Value (10%)  : -2.568
A p-value below a threshold (0.05 or 5%) suggests we reject the null hypothesis so it is stationary
```

Figure 17: ADF test of ED patient demand

Even though the ADF test result showed a stationary series, it does not automatically imply that ED patient demand is white noise. Through the visualization analysis of ACF (Auto-Correlation Function) and PACF (Partial Auto-Correlation Function) plots in Figure 18, several other insights can be extracted. The ACF plot provides the information about the linear relationship between lagged values of itself while The PACF plot is a plot of the partial correlation coefficients between the series and lags of itself that is not captured by correlations at all lower-order-lags. ACF plot in Figure 18 clearly shows significant repetitive correlation on multiple lags, especially in every seven lags or perhaps indicating seven lag days. This can be an indication that the ED patient demand series has seasonal components. A similar case occurs in the PACF plot, which dominantly has a significant correlation in every seven lags. Besides, both ACF and PACF shows a drop-off below a significant level almost at the same point, after 70 or 80. This might also suggest the existence of the integration between Auto Regression model and Moving Average model or well-known as ARIMA model to predict the time series. The modeling for time series forecasting will be further detail discuss in chapter 5.

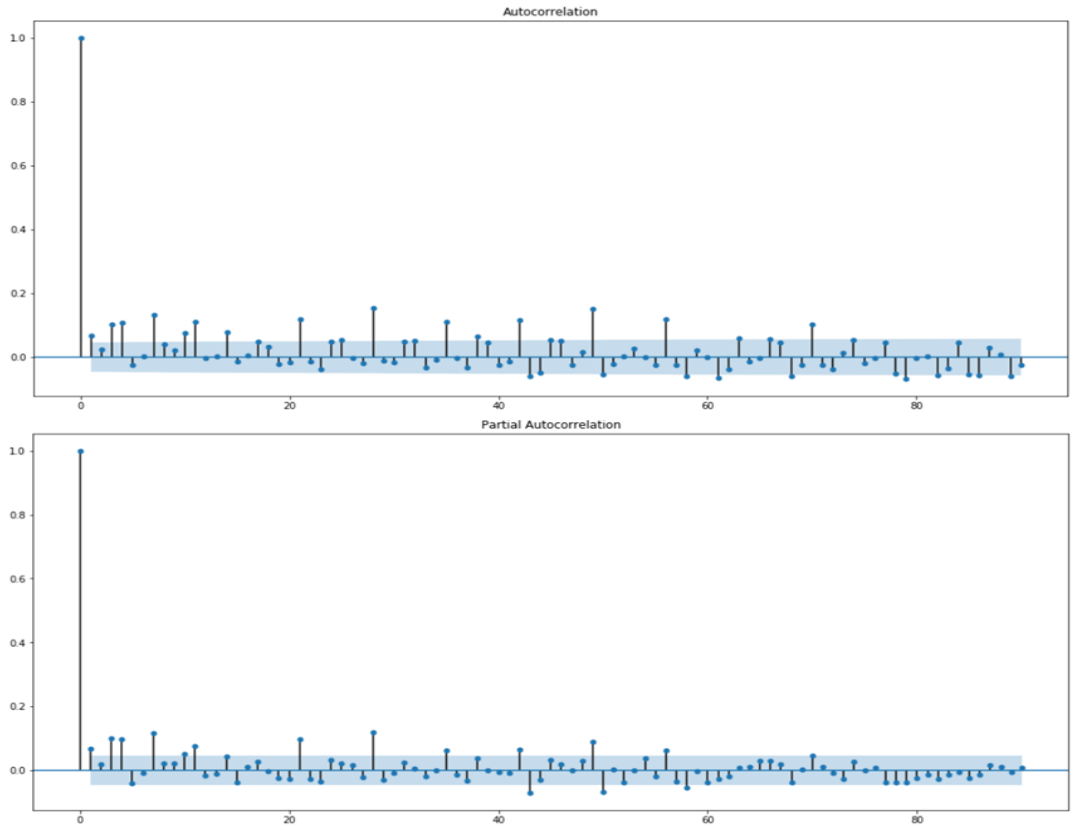


Figure 18: ACF (Autocorrelation) and PACF (Partial Autocorrelation) plots of ED patient demand

4.1.2 GP-Post Univariate Time Series Analysis

Unlike ED daily patient demand plot (in Figure 14), which does not show a specific pattern, GP-Post patient demand plot in Figure 19 easily indicates the visible separation among the value along the 2013-2017 period. This separation is caused by different working hour duration of GP-Post during weekday and weekend. In the weekday, GP-Post operates from 17.00 to 08.00 morning in the next day. In the weekend, Saturday and Sunday, GP-Post operates 24 hours. As a result of that, more patient demand naturally occurs during the weekend, which is depicted by the separation group on the top and the bottom in Figure 23. This separation is also reflected by the histogram in Figure 20.

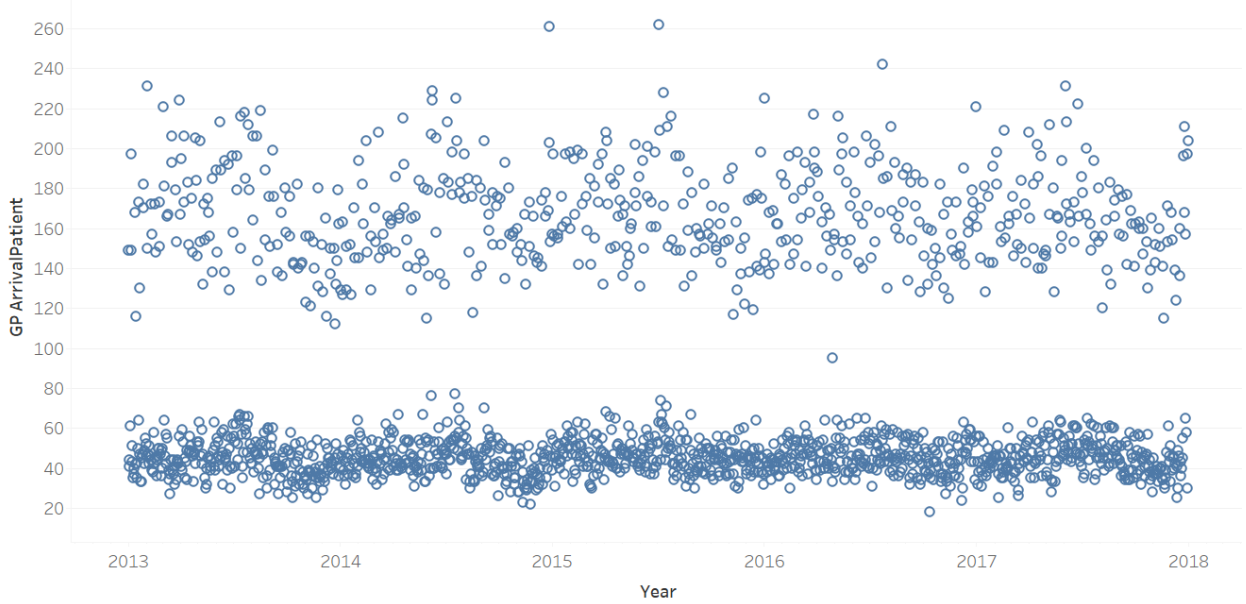


Figure 19: GP-Post daily patient demand

In total average, there is 82 daily patient demand for GP-Post with a standard deviation of 6.77. In weekday, the daily average is only around 48 patients while on the weekend, the daily average is up to 167 patients. The summary statistic of GP-Post patient demand is presented in Table 12. Besides, Table 12 also provides the summary statistic of GP-Post patient demand split by weekdays and weekend.

Table 12: GP-Post statistical summary

Statistic Summary	Count	Mean	Std	Min	25%	50%	75%	Max
GP Post Patient Demand Total	1826	82	6.77	18	42	50	143	262
GP Post Patient Demand Weekday	1304	47.91	19.81	18	40	45	51	224
GP Post Patient Demand Weekend	522	167.13	24.75	122	150	165	183	262

While the average number of GP-Post patients during the weekend is much higher, more than 20% higher than a weekday, however, the maximum number of patients in weekday and weekend is relatively close, 224 and 262. It shows the possibilities of having high demand patients are not only exclusive during the weekend, but also weekdays. However, analyzing through the weekdays, patient demand distribution of

25%, 50%, and 75% in Table 12 indicates that the case of high demand patient during weekdays might only happen occasionally.

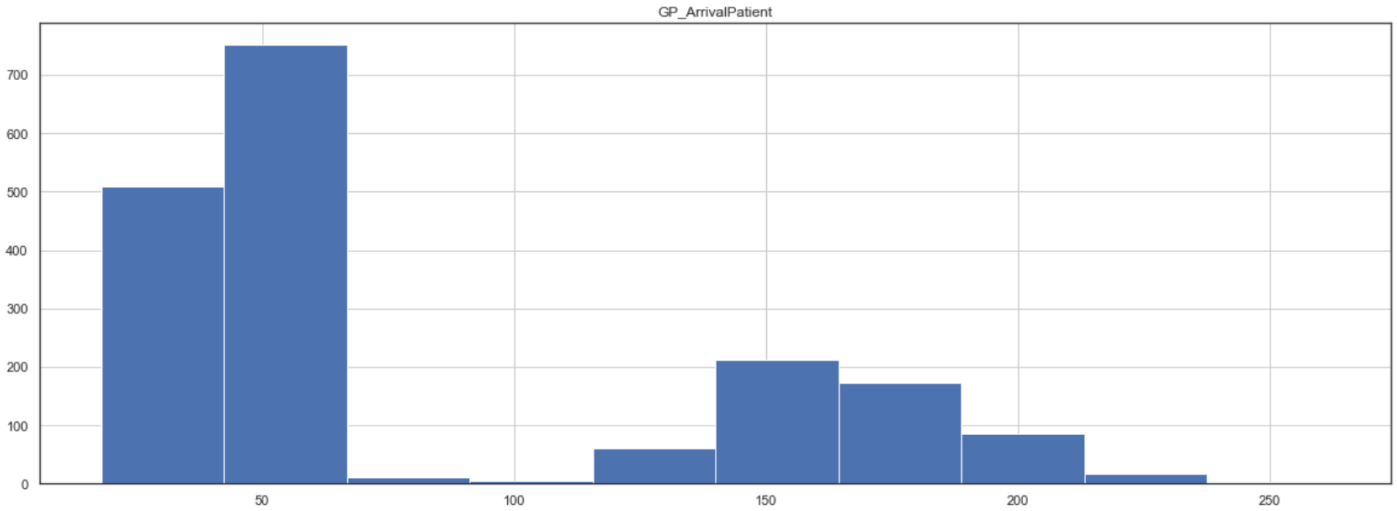


Figure 20: The histogram of GP-Post patient demand

Similar to ED patient demand, GP-Post patient demand of 2017 can be decomposed through classical additive time series model into four components: Observe, Trend, Seasonal, and Residual as described in Figure 21. Also, the GP-Post decomposition plot does not indicate having long term positive or negative trend, instead relatively flat with several exceptions. In term of Seasonal pattern, GP-Post plot has smoother seasonal pattern than ED. Residual component of GP-Post also shows less fluctuated in comparison with ED's Residual.

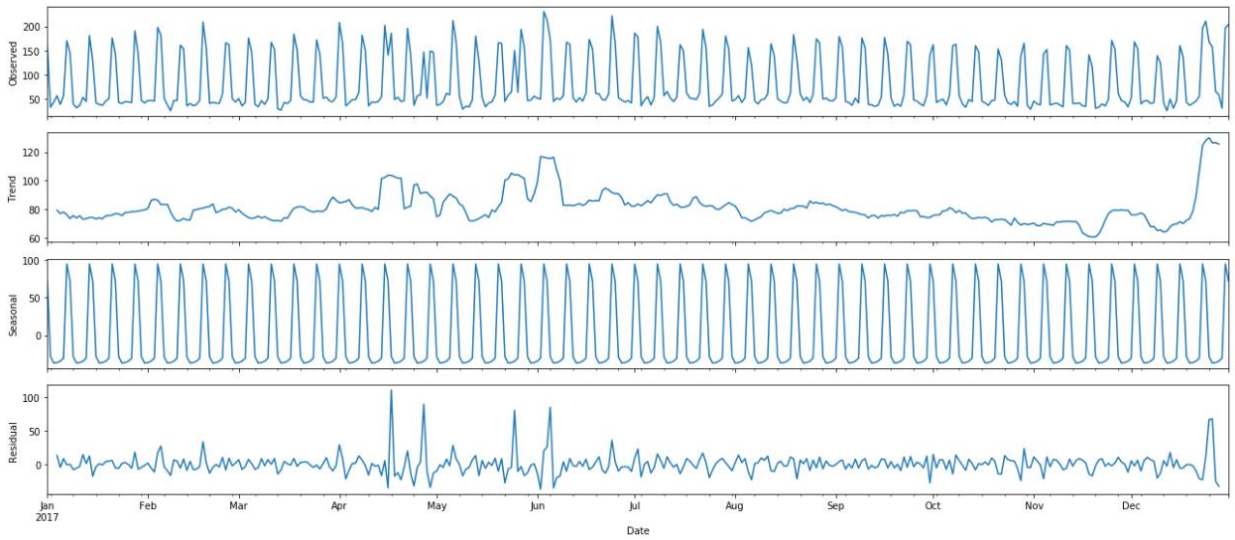


Figure 21: GP-Post Time Series Decomposition with an additive model

Plotting result in Figure 19 or Figure 21 shows very limited information to infer the stationarity of GP-Post patient demand. Therefore, the ADF test is required to check the stationarity. The result of the ADF test is presented in Figure 22. Since the p-value is less than 5% and Test Statistic -6.929 is much lower than the 5% of Critical Value of -2.863, it shows strong evidence that GP-Post time series is already stationary.

```
TestStationaryAdfuller(Label_Y.GP_ArrivalPatient.values, 0.05)
```

```
Test Statistic  : -6.929
p-value  : 0.000
#Lags Used  : 23.000
Number of Observations Used  : 1802.000
Critical Value (1%)  : -3.434
Critical Value (5%)  : -2.863
Critical Value (10%)  : -2.568
A p-value below a threshold (0.05 or 5%) suggests we reject the null hypothesis so it is stationary
```

Figure 22: ADF Test of GP-Post patient demand

Besides knowing the stationarity of the GP-Post patient demand, the next action is to analysis through ACF and PACF plots as provided in Figure 23. In ACF, a noticeable pattern is a large number of significant lags slowly degrade as the lag increases until eventually shrink under the significant level. In PACF, A few significant lag observations that abruptly drop as the lag increases. ACF and PACF plots indicate a strong autocorrelation component with multiple lags. Although it is possible to estimate a model from analyzing ACF and PACF alone, a more exhaustive approach with SARIMAX which includes external variables will be performed in Chapter 5.

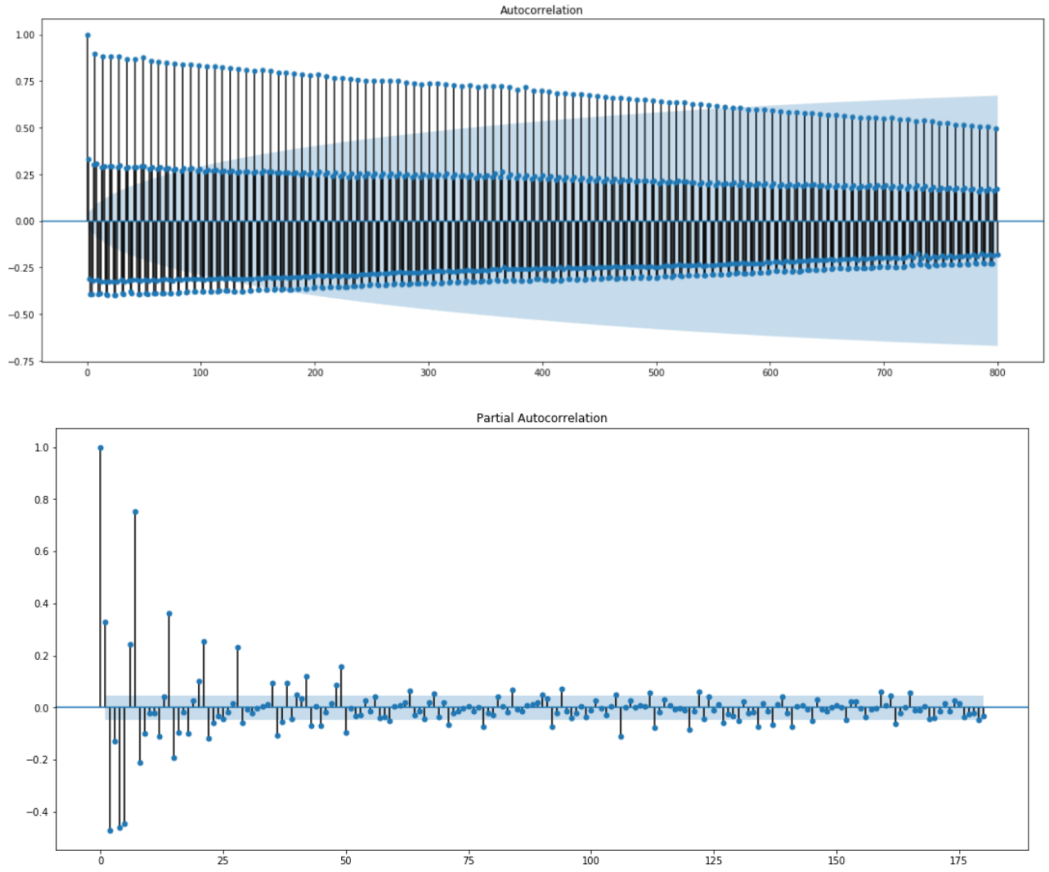


Figure 23: ACF (Autocorrelation) and PACF (Partial Autocorrelation) plots of GP-Post patient demand

4.1.3 Weather (Temperature and Humidity) correlation with ED and GP-Post patient demand

Naturally, the weather will affect the tendency of patient demand for health care. Heavy rain or snow on a particular day potentially decrease the number of patient demand. Besides rain and snow, a study [55]

analyzes the association between heat and ED patient demand in Australia, Botswana, Netherlands, Pakistan, and the USA. In the Netherlands, the study shows different result between regions and pointed out the association between hot days with a higher proportion of ED visits, especially in children categories. However, the study offers more than one possible reason to explain the mentioned association such as school holiday, dehydration and electrolyte imbalance in children, and the higher chance that exertional heat stroke primarily affects younger active populations. Another study [57] shows similar result by reviewing 72 relevant literature about a relationship between HFMD (Hand, foot, and mouth disease) and weather-related data such as temperature, relative humidity, wind speed, rainfall, atmospheric pressure, and sunshine. The rationale behind the research is the common tendency for HFMD patients to spike during summer months, so a relationship between HFMD and weather patterns might be valid. It finds that HFMD, which predominantly affected children under five years of age, has a statistically significant relationship with temperature and humidity. Conversely, no significant relationship is identified between HFMD and the remain weather factors: precipitation, wind speed, and sunshine. However, more than 80% of these findings are originated in China, and the remains were mostly in Pacific area countries such as Japan, Vietnam, South Korea, Singapore, and Thailand.

Based on above mention information, the relationship or linear correlation between patient from different age categories in ED and GP-Post and the weather-related data (Temperature and Humidity) is analyzed through Pearson Correlation method with several scenarios: splitting by weekday and weekend, categorizing based on season, and taking out the holiday dates. Selecting or filtering the dataset with these scenarios would significantly reduce the sample size. However, according to this paper [19], to detect at least correlation coefficient of 0.5 with a significant result ($p < 0.05$) with sufficient power (80%), the ***minimum*** required sample size for this specification is only 29. In other words, the sample size used in this research, even after filtering with the several mentioned scenarios, is still sufficient to perform Pearson correlation with the mentioned specification.

Looking at the overall total daily patient number of GP-Post (GP_ArrivalPatient) and ED (ED_PatientArrival) in Figure 37 in the Appendix, both indicated a poor or lack correlation with temperature and humidity as referring to Person Correlation interpretation in Table 32. Slightly fair correlation coefficients occurred during weekends (Figure 38). Interestingly, the better fair correlation with temperature in the weekends-summer (Figure 40) for GP_ArrivalPatient was not followed by ED_ PatientArrival. During weekends-autumn (Figure 42), ED_PatientArrival seemed to be more sensitive with temperature than GP_ArrivalPatient with a bit higher correlation coefficient. In correlation with humidity, breaking the data into weekday and weekend or splitting it by season did not indicate any significant correlation.

Another insight from Figure 38 was a positive and fair correlation between temperature and both age group 5-19 and 20-65 in GP-Post during weekends. This result became interesting with the fact that actually, temperature correlation with total GP-Post patient (GP_ArrivalPatient) was lower on weekends than on weekdays, but the temperature correlation to the GP-Post age groups (5-19 and 20-65) showed the opposite result. All these findings, especially on the weekends, are aligned with the previously mentioned study [55] that pointed out the association between hot days with patient demand. However, instead of affecting the children group (0-4), the findings in this thesis suggest more effect of temperature on other age groups: 5-19 and 20-65 in GP-Post.

Breaking the dataset into four seasons and also split each season between weekdays and weekends, as provided in Figure 39 until Figure 46 in the Appendix, offered several insightful information, in particular

during summer as suggested by the study mentioned above [57]. For example in Figure 40, there was a fair correlation between temperature parameters (W_TX-Max, W_TG-Average, W_TN-Min) and the patient number of some age groups: 0-4, 5-19 and 20-65 in GP-Post during summer. In autumn (Figure 42), more positive and fair correlation with temperature diversely occurred, not only with age groups in GP-Post (5-19 and 20-65) but also with age groups in ED (5-19 and 20-65). However, the positive and fair correlation in both Figure 40 and Figure 42 were specifically calculated for only the weekends of the respective season. So, the increased correlation coefficient possibly caused by (1) the overall growing trend in total patient demand during the summer or autumn or (2) the weekend factor that made higher tendency for patients to seek health care services. This explanation was supported by the fact that the same increasing pattern occurred in Figure 40 and Figure 42 between the temperature parameters with the total number of GP-Post patient (GP_ArrivalPatient) and ED patient (ED_PatientArrival). As also pointed out by the study mentioned above [55], the school holiday might be the possible causality factor besides the correlation with the temperature. As opposed to the result in this literature [57], the correlation with Humidity is not found as a significant factor to ED and GP-Post based on group-Age patients in almost conditions except during all weekends (Figure 38) and winter (Figure 44).

Besides age groups, the temperature and humidity correlation against various treatment groups in GP-Post (ICPC code A, B, D, etc.) and ED (CAR, CHI, LON, etc.) were also analyzed. Similarly, the Pearson correlation was used to calculate correlation coefficient with several scenarios: splitting by weekday and weekend, categorizing based on season, and taking out the holiday dates (Figure 47 until Figure 56 in the Appendix). Generally, both temperature and humidity had a poor correlation with the treatment groups in all scenarios with several exception such as ICPC code S, ICPC code R. The patient numbers of ICPC code S (patient with skin related diseases) had positive fair to moderate correlation with temperature by the range of 0.44 to 0.69, where the highest occurred in weekends period (Figure 48). Conversely, the ICPC code R (patient with respiratory-related diseases) had a fair negative correlation with the temperature by the range of -0.30 to -0.54 (Figure 47 and Figure 48). Among all the seasons, the weekends in spring (Figure 56) indicated quite significantly fair correlation for ICPC code S and ICPC code R with the temperature. In the correlation with humidity, mostly ICPC code S had a poor correlation except in Figure 48, where ICPC code S indicates a fair negative correlation with average humidity (W_TG). Similar to the mentioned findings, other studies have also analyzed the effect of temperature and humidity on patients with skin problem [58-60] and patients with the respiratory problem [61-62].

4.1.4 The spread of pollen correlation with ED and GP-Post patient demand

Besides weather-related data, other external data used for analyzing ED and GP-Post patient demand is pollen. A study [56] discusses the allergic reaction induced by pollen in Europe and highlights the recent finding on the respiratory allergic diseases in relation to the pollen. For example, Betula or Birch, which is one of the pollen species considered as the most allergenic tree pollen, is frequently associated with the nasal symptom. Another finding showed that sensitization to Parietaria Judaica pollen, which is the main allergenic genus of the Urticaceae (nettle) family, noticeably increased the risk of developing asthma. A more recent finding in this paper [63] highlights Poaceae pollen as the leading pollutant and the main cause of pollen allergy (with skin or respiratory) across the world. It also points out the possible effect of the climate change on plant phenology, especially Poaceae, that eventually has an implication on pollen concentration.

Following the same approach with the previous sub-sections, the Pearson correlation was used to calculate correlation coefficient with several scenarios: splitting by weekday and weekend, categorizing

based on season, and taking out the holiday dates (Figure 57 until Figure 66 in the Appendix). Among various types of pollen provided in Figure 57, Urtica, Betula, and Poaceae are selected as examples. Generally, Urtical had a fair and positive correlation with ICPC code S, either in weekdays or weekends (Figure 57). However, breaking down the correlation by season, the same fair correlation occurred only during Autumn. Although the fair level correlation was identified between Urtica and ICPC code S, this finding was not aligned with the mentioned study [56] where Urtica was more associated with Asthma than with patient with the skin-related problem (ICPCcode_S). In Betula, the fair and sufficient correlation with any types of treatment groups were hardly found in all periods. In Poaceae, the fair and positive correlation only found during all weekends and Autumn-Weekends with ICPCcode_S (Figure 58 and Figure 62). This finding was aligned with the previously mentioned paper [63], especially the correlation between Poaceae and patient with skin problem (ICPCcode_S).

4.2 EDA for the prediction of ED inpatient admission

The ED's raw data for predicting ED inpatient admission is the same raw data used for ED patient demand forecasting as listed in Table 9. However, daily aggregation is not required because the date time dimension will not be included in predicting the ED inpatient admission. In other words, classifying whether or not the patient will be admitted to the hospital must not depend on the date and time. Instead, using the characteristic of each patient, such as age or urgency level, will logically affect and make more relevant contribution to the output prediction. As a result of that, the categorical columns in the ED raw data such as gender, treatment, urgency level and so on, will be counted as a binary number (1 or 0) as features in describing the characteristic of each patient. The only exception is the age of each patient, which still in the integer number format.

The distribution of each feature columns in the dataset for predicting ED inpatient admission is presented in Figure 24. Several insights can be derived from Figure 24. Firstly, most of the binary distribution is highly imbalance except for Gender_man, Gender_vrouw, Referral_GP, Urgency_Geel, and Urgency_Groen. Secondly, knowing that the presence of the high imbalance distribution data can help in deciding the splitting methods used in building the ML model. For instance, stratified random splitting method between the training and the testing dataset might be a relevant method to ensure the balanced distribution between the training and the testing dataset. Thirdly, the target variable, which is Label_Admittance, also shows imbalance distribution where the admitted patients (1) are much less than the non-admitted patients (0). With this information, using recall as the main evaluation metrics might be more relevant than accuracy or precision. All these findings will be discussed further in chapter 5, 6, and 7.

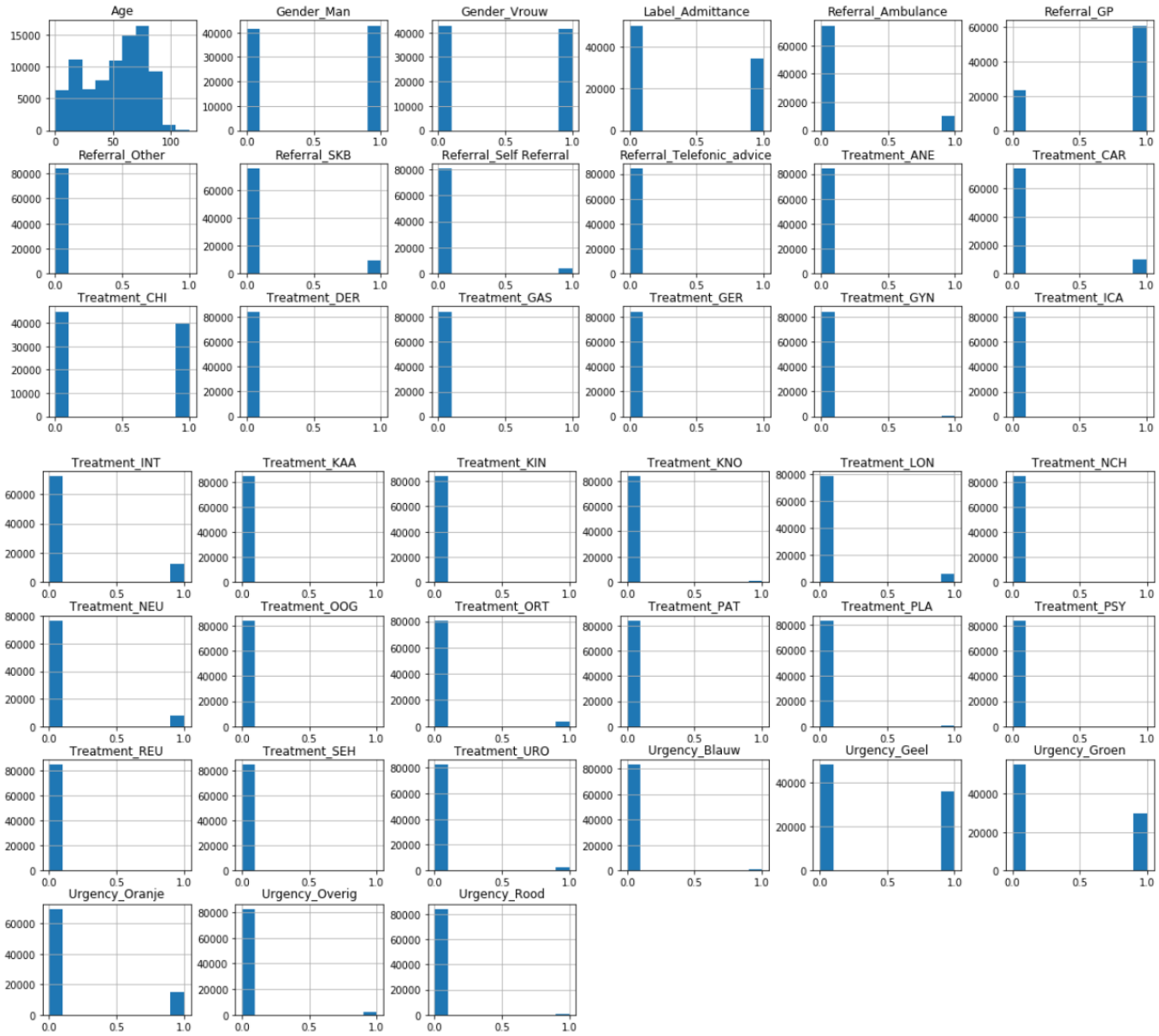


Figure 24: The histogram of inpatient admission dataset

5. Experiment Design and The Implementation

In this chapter, the experiment design and the implementation consist of two mains section, namely (1) Forecasting ED and GP-Post patient demand, (2) Predicting ED inpatient admission. In the first section (1), three scenarios of the experiment design and the implementation will be presented: (a) Single SARIMAX model, (b) Single SARIMAX model with Feature Selection, (c) A Hybrid model. In the second section (2), two scenarios of the experiment design and the implementation will be presented: (a) Predicting with four types of ensemble algorithm, (b) Tuning the selected model with Hyper parameter.

5.1 Forecasting ED and GP-Post patient demand

Splitting the dataset into training and testing is the required step before ML modeling process. The ED training dataset is started from index 08-01-2013 until 31-12-2016 while the remaining dataset, which is from index 01-01-2017 until 31-12-2017, is used for the testing dataset. The dimension of the ED training dataset for the dependent variable is 1454 rows and one target variable column while the feature or exogenous variable of training dataset consists of 1454 rows with 1132 features columns. The dimension of the ED testing dataset for the dependent variable is 365 rows and one target variable column while the external features or exogenous variables of the testing dataset consists of 365 rows with 1132 feature columns.

Similar to ED dataset, the GP-Post training dataset is also started from index 08-01-2013 until 31-12-2016 and the GP-Post testing data set is from index 01-01-2017 until 31-12-2017. The dimension of the GP-Post training dataset for the dependent variable is 1454 rows and one column while the feature or exogenous variable of training dataset consists of 1454 rows with 1132 columns. The dimension of the GP-Post testing dataset for the dependent variable is 365 rows and one column while the feature or exogenous variable of GP-Post testing dataset consists of 365 rows with 1132 columns.

5.1.1 Scenario-1: A single SARIMAX model

In scenario-1, forecasting patient demand with a single SARIMAX was implemented through pmdarima library[17] version 1.2.0 on python as a programming language. Pmdarima is essentially a Python & Cython wrapper of various statistical and machine learning libraries such as statsmodels and scikit-learn. However, unlike stasmodels library, pmdarima operates by generalizing all ARMA, ARIMA, SARIMAX models into a single class. It also provides an auto_arima() function which operates a bit like a grid search. The flowchart of SARIMAX model implementation with auto_arima() function is described in Figure 25.

In Figure 25, auto_arima function executed various sets of p and q (also P and Q for seasonal models) parameters on the training dataset, then select the best model in order to minimize the AIC. To select the differencing terms, auto_arima used a test of stationarity (such as an augmented Dickey-Fuller test) and seasonality (such as the Canova-Hansen test) for seasonal models. Next, the best SARIMAX model suggested by auto_arima() was analyzed to check the presence of white noise. The best SARIMAX model with white noise residue on the training dataset was evaluated by fitting it against the testing dataset. All the metric evaluation result, AIC, MAPE, RMSE, and MAE were calculated and generated, including the mean and standard deviation of residue.

[17] (n.d.). pmdarima: ARIMA estimators for Python — pmdarima 1.2 ... - alkaline-ml. Retrieved May 24, 2019, from https://www.alkaline-ml.com/pmdarima/

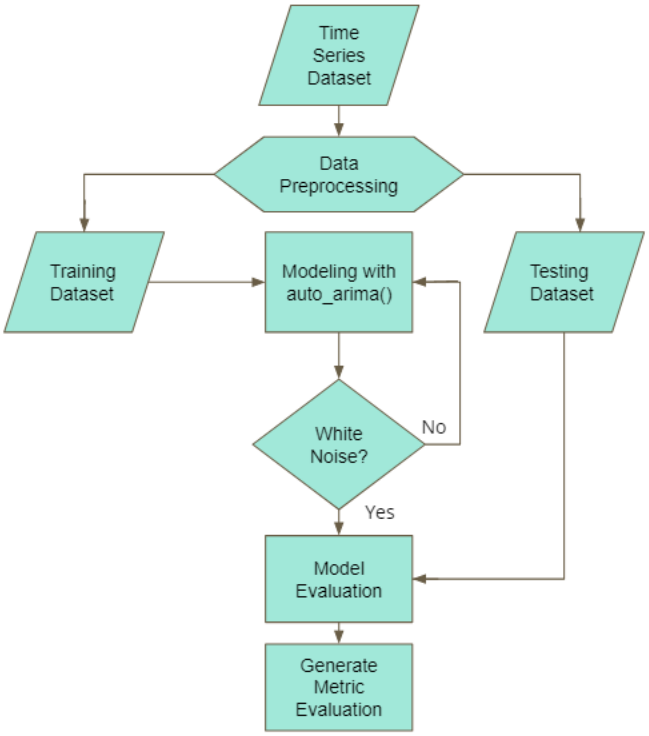


Figure 25: Flowchart of implementing scenario-1

Amongst many input parameters[18] provided by auto_arima function, the relevant parameters, as listed in Table 13 below, are selected and used for the training process.

Table 13: SARIMAX parameters

Parameters	Description	Input Value
y	The time-series to which to fit the ARIMA estimator	Dependent variable
exogenous	If provided, these variables are used as additional features in the regression operation	Features columns such as weather, pollen, dummy variable, etc
start_p	The starting value of p, the order (or number of time lags) of the auto-regressive ("AR") model	Default=2
d	The order of first-differencing. If None (by default), the value will automatically be selected based on the results of the test (i.e., either the Kwiatkowski–Phillips–Schmidt–Shin, Augmented Dickey-Fuller or the Phillips–Perron test will be conducted to find the most probable value).	None

[18] (n.d.). pmdarima.arima.auto_arima — pmdarima 1.0.0 ... - alkaline-ml. Retrieved May 24, 2019, from http://www.alkaline-ml.com/pmdarima/1.0.0/modules/generated/pmdarima.arima.auto_arima.html

start_q	The starting value of q, the order of the moving-average (“MA”) model	Default=2
max_p	The maximum value of p	Default=5
max_d	The maximum value of d, or the maximum number of non-seasonal differences	Default=2
max_q	The maximum value of q	Default=5
start_P	The starting value of P, the order of the auto-regressive portion of the seasonal model.	Default=1
D	The order of the seasonal differencing. If None, the value will automatically be selected based on the results of the seasonal_test parameter	None
start_Q	The starting value of Q, the order of the moving-average portion of the seasonal model.	Default=1
max_P	The maximum value of P	Default=2
max_D	The maximum value of D	Default=1
max_Q	The maximum value of Q	Default=2
max_order	If the sum of p and q is >= max_order, a model will not be fit with those parameters but will progress to the next combination. If max_order is None, it means there are no constraints on maximum order.	None
m	The period for seasonal differencing, m refers to the number of periods in each season.	m=7
seasonal	Boolean to indicate whether to use seasonal on ARIMA model	True
stationary	Boolean to indicate whether the time-series is stationary	True
information_criterion	The information criterion used to select the best ARIMA model	AIC
alpha	Level of the test for testing significance	Default=0.05
test	Type of unit root test to use in order to detect stationarity if stationary is False and d is None.	test=’adf’

stepwise	Whether to use the stepwise algorithm outlined in Hyndman and Khandakar (2008) to identify the optimal model parameters. The stepwise algorithm can be significantly faster than fitting all (or a random subset of) hyper-parameter combinations and is less likely to over-fit the model.	True
trend	The trend parameter	None
with_intercept	Whether to include an intercept term	None

5.1.2 Scenario-2: A single SARIMAX model with Feature Selection

In scenario-2, several additional processes were added to the prior scenario-1 flowchart, as described in Figure 26. Feature selection with Lasso was performed as part of data pre-processing. As a result of that, Lasso selected N-best features based on the coefficient weight of each feature. As explained in Chapter 3, the N-best features can be reduced further by running auto_arima iteratively and ascendingly based on each weight of N-best features. On each iteration, RMSE was calculated on training and testing dataset. Once all the features had been iterated, the best model with the lowest RMSE was selected.

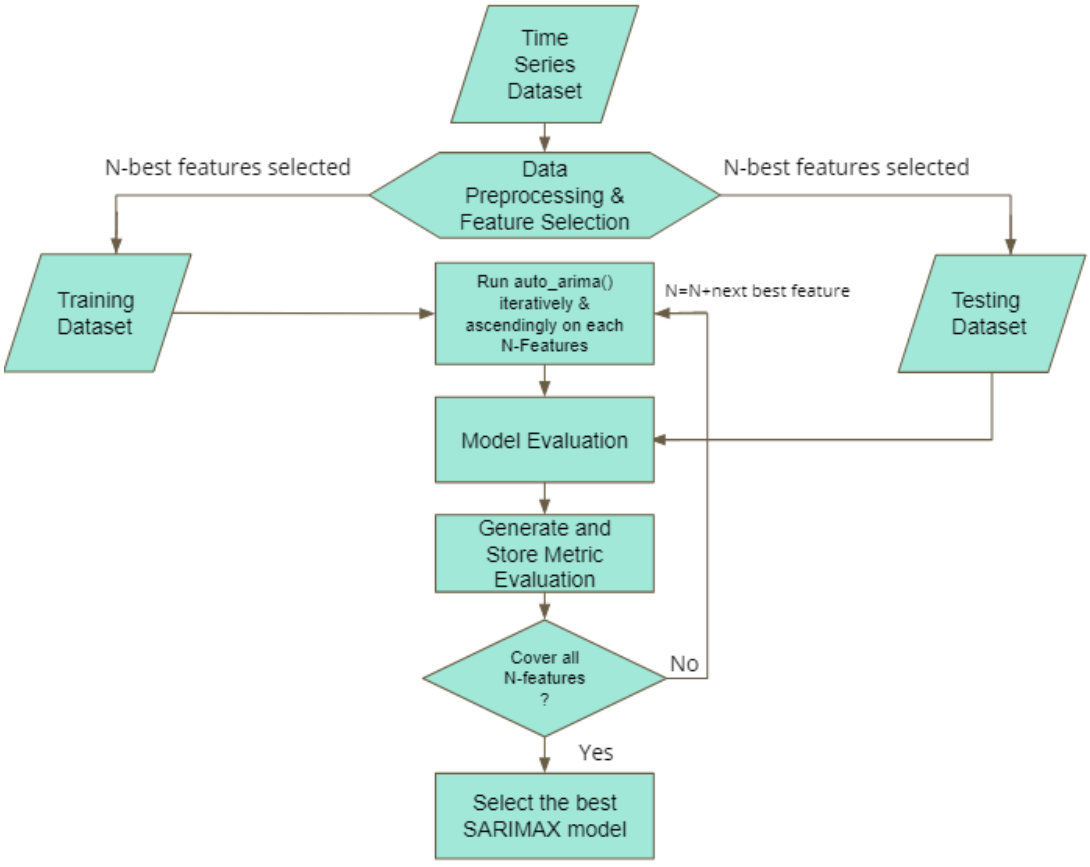


Figure 26: Flowchart of implementing scenario-2

In the Feature Selection process, LassoCV[19] from scikit-learn library is used. The Feature Selection with Lasso process will be evaluated by applying the new and smaller dimension dataset to the training and testing process with SARIMAX model. There are two ways of implementing Lasso. The first, manually build Lasso linear model with iterative fitting along a regularization path. Then, the coefficient values of all features can be ranked and sorted to select the significant features and to remove the insignificant features. The second, the feature selection process can be automated through Meta-transformer for selecting features based on importance weights. The implementation of this method is via SelectFromModel[20] function in the feature_selection class of scikit-learn library. The second way is more preferred for practicality. The relevant parameter list of SelectFromModel is presented in Table 14 below.

Table 14: SelectFromModel parameters

Parameters	Description	Input Value
estimator	The base estimator from which the transformer is built	Estimator = LassoCV
threshold	The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded.	threshold = -np.inf (since max_features is used, threshold is set as -np.inf)
max_features	The maximum number of features selected scoring above the threshold	max_features = A number with the lowest RMSE in Figure 26

5.1.3 Scenario-3: A Hybrid Model

In scenario-3, the best SARIMAX model from scenario-2 was still being used as the main model for forecasting the time series. After fitting the training dataset, the residue of the best SARIMAX model then was forecasted by Gradient Tree Boosting. The combination prediction of SARIMAX model and Gradient Tree Boosting, as illustrated in Figure 27, formed a Hybrid model. Eventually, the Hybrid model was evaluated against the testing dataset, and all relevant metric evaluation was calculated and generated.

The Gradient Tree Boosting is implemented through Gradient Boosting for regression or GradientBoostingRegressor[21] from scikit-learn library. On top of that, GridSearchCV[22] is used for the exhaustive search over specified GradientBoostingRegressor parameter values. Moreover, GridSearchCV

[19] (n.d.). 3.2.4.1.3. sklearn.linear_model. LassoCV — scikit-learn 0.21.1 .... Retrieved May 24, 2019, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html

[20] (n.d.). sklearn.feature_selection.SelectFromModel — scikit-learn 0.21.1 .... Retrieved May 24, 2019, from http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html

[21](n.d.). 3.2.4.3.6. sklearn.ensemble.GradientBoostingRegressor — scikit-learn .... Retrieved May 24, 2019, from http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[22](n.d.). sklearn.model_selection.GridSearchCV — scikit-learn 0.21.1 .... Retrieved May 24, 2019, from http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

also enables cross-validation to optimize the grid searching with param_grid parameter. The relevant GridSearchCV parameters are listed in Table 15 below.

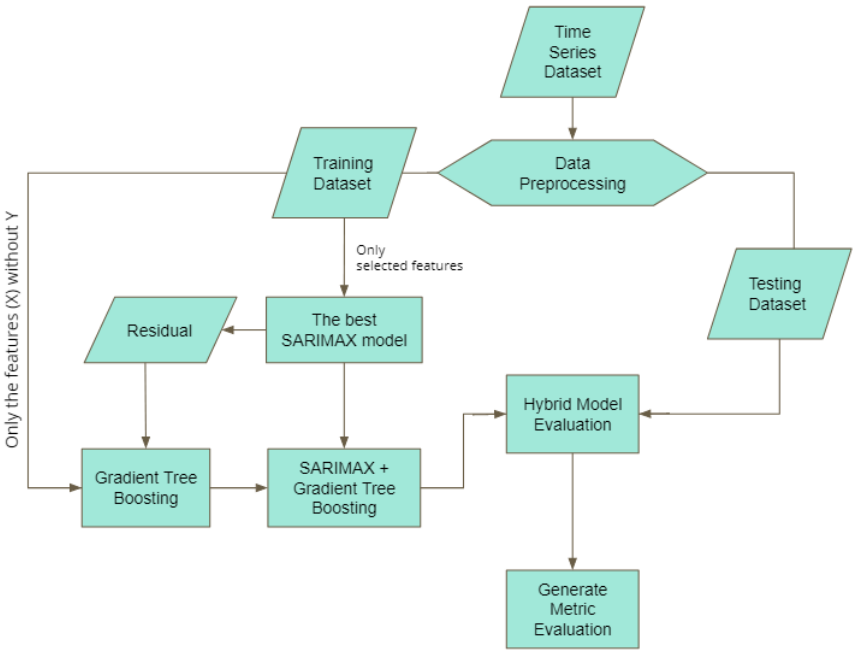


Figure 27: Flowchart of implementing scenario-3

Gradient Tree Boosting has various parameters that need to be configured to find the best ML model performance. Running all combination with brute force mode will take very long execution time. Instead, in this research, performing plot analysis on each parameter of gradient tree boosting was preferred. As a result of that, the range values of each parameter can be reduced and restricted. Using the smaller value ranges of each parameter, GridSearchCV then was used to search the find the optimal parameter configuration of Gradient Tree Boosting. The parameter configuration of GridSearchCV was presented in Table 15. After finding the best GradientBoostingRegressor model with GridSearchCV, Feature Selection was performed by evaluating the feature importance. Similar to Feature Selection in Lasso, the feature selection process of GradientBoostingRegressor was automated through SelectFromModel function in the feature_selection class of scikit-learn library.

Table 15: GridSearchCV parameters

Parameters	Description	Input Value
estimator	The base estimator from which the transformer is built	Estimator= GradientBoostingRegressor
scoring	A single string input to evaluate the predictions on the test set.	Scoring= 'neg_mean_squared_error'
cv	Determines the cross-validation splitting numbers.	cv = 5

param_grid	Dictionary with parameters names (string) as keys and lists of parameter settings to try as values, or a list of such dictionaries, in which case the grids spanned by each dictionary in the list are explored. This enables searching over any sequence of parameter settings.	{
		'learning_rate': np.linspace(0.01, 1, 10, endpoint=True),
		'n_estimators': [5, 25, 50, 100, 200], 'max_depth' : [3, 5, 10, 15], 'min_samples_split': np.linspace(0.01, 0.1, 3, endpoint=True), 'min_samples_leaf': np.linspace(0.2, 0.5, 3, endpoint=True)
		}

5.2 Predicting ED inpatient admission with Gradient Boosting Classification

The implementation of inpatient admission prediction consists of two main scenarios, namely (a) Predicting with four types of ensemble algorithm and (b) Tuning the selected model with Hyper parameter. In (a), four ensemble algorithms were set using their default parameter configuration then run them on the training dataset with cross-validation method. The implementation of cross-validation is through cross_val_score[23] function in scikit library. Besides, the pipeline function in scikit library is also used to assemble several steps that can be cross-validated all together while setting lots of parameters of four different ML ensemble algorithms. The relevant parameter list of cross_val_score is presented in Table 16 below.

After comparing their metric evaluation results, one best model was selected for further optimization process GridSearchCV function. Although GridSearchCV can help in finding the best configuration, running all different combination with brute force mode will take very long execution time. Instead, in this research, performing plot analysis on each parameter of gradient tree boosting, as described in Figure 91 until Figure 96, was preferred. As a result of that, the range values of each parameter can be reduced and restricted. Using the smaller value ranges of each parameter, GridSearchCV then was used to search the find the optimal parameter configuration of Gradient Tree Boosting.

[23] (n.d.). sklearn.model_selection.cross_val_score — scikit-learn 0.21.1 .... Retrieved May 25, 2019, from http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

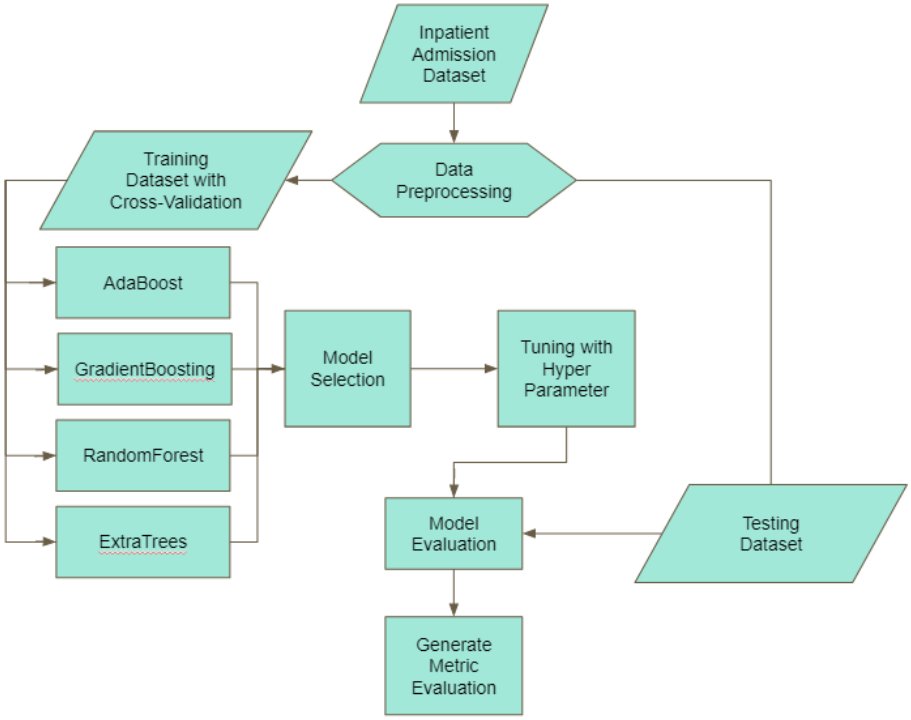


Figure 28: Flowchart of implementing inpatient admission prediction

Table 16: coss_val_score parameter

Parameters	Description	Input Value
estimator	The object to use to fit the data	Estimator = pipeline of AdaBoost, Gradient Boosting, Random Forest and Extra Trees
X	The features data to fit	X = 80% of feature dataset
Y	The target variable to try to predict	Y = 80% of label dataset
scoring	A score options	scoring=MSE
cv	Determines the cross-validation splitting numbers. For binary label, StratifiedKFold is used	cv=10

6. Result and Discussion

In this chapter, the result of the modeling phase will be presented into three main sections: (1) Result of forecasting ED patient demand, (2) Result of forecasting GP-Post patient demand, (3) Result of predicting inpatient admission. At the end of this chapter, there will be discussion section which summarizes the finding in (1), (2), and (3) to answer the subquestion 5 and 6 in the research question. Moreover, the comparison with other studies will also be analyzed in the discussion section.

6.1 Result of Forecasting ED patient demand

This section consists of three different results from three type of modeling, namely: a single SARIMAX model, a SARIMAX model with Feature Selection, and a Hybrid model. Each model generates two metrics evaluation results based on two different datasets: training and testing. Firstly, the metric evaluation result of the training and testing generated by each model will be presented. After that, the comparison between the training result and the testing result as well as the comparison among the result of three different models will be analyzed and discussed.

6.1.1 Forecasting ED patient demand: Result of a single SARIMAX model

After running various order combination with auto_arima function, SARIMAX(1,0,1)x(0,0,0,7) with AIC 9811.79 comes up as the best model. Besides AIC, other metrics evaluation, namely MAPE, RMSE, MAE, can be derived and manually calculated using SARIMAX(1,0,1)x(0,0,0,7). The complete metric evaluation results of SARIMAX(1,0,1)x(0,0,0,7) for training and testing dataset, including the residue information, are presented in Table 17.

In the training dataset, all the metrics evaluations indicate an excellent forecasting performance of SARIMAX(1,0,1)x(0,0,0,7) with their relative small error scores. Taking MAE by 2.48 as an example, it implies that SARIMAX(1,0,1)x(0,0,0,7) can predict the number of ED patient demand with the range of plus 2.48 or minus 2.48 to the actual value. Considering the standard deviation of ED patient demand is 6.7, the MAE range of ± 2.48, along with small error metrics values of MAPE and RMSE, strongly indicates the excellent accuracy of SARIMAX(1,0,1)x(0,0,0,7) prediction.

Besides relying on these error metrics numbers, plotting between the prediction values against the actual values also supports this argument. The close proximity between the prediction values against the actual values is described in Figure 68. Moreover, Figure 68 also indicates the ability of SARIMAX(1,0,1)x(0,0,0,7) in accurately predicting the actual value of ED patient demand across any value ranges either the ranges close to the minimum, the middle, or the maximum.

Similarly, the line graph in Figure 67 shows how close the prediction values against the actual values. The small range of residue value in Figure 67 also describes the close proximity of both. More comprehensive visualization of the residue is provided in Figure 69. On the top left, the residues fluctuate around a mean of zero and have a uniform variance. On the top right, the density plot shapes a normal distribution with mean zero. On the bottom left, most of the dots on the QQ plot fall perfectly in line with the red line. Lastly, on the bottom right, the correlogram or ACF plot shows the residual errors are not autocorrelated. All these indications suggest that residue are not significantly different from the characteristic of white noise. In other words, SARIMAX(1,0,1)x(0,0,0,7) has considerably captured all the relevant pattern in the data that were required for making the prediction, and it remained the uncorrelated and random residue or white noise.

Table 17: Result of a single SARIMAX(1, 0, 1)x(0, 0, 0, 7) model

SARIMAX(1, 0, 1)x(0, 0, 0, 7)	Training (Dataset 2013-2016)	Testing (Dataset 2017)
AIC	9811.79	-
MAPE	7.81%	38.99%
RMSE	3.24	16.77
MAE	2.48	12.81
Mean of Residue	0.00	0.15
Std of Residue	3.20	16.79

However, applying a trained model SARIMAX(1,0,1)x(0,0,0,7) on testing dataset yields unexpected results. The error rates jump up more than a hundred times from the training results, as found in Table 17. For example, MAPE in testing dataset increases more than 400% while RMSE and MAE also surge nearly 14 points and 10 points respectively. The mean of residue and its standard deviation between training and testing also suffers from the same problem. Scatter plot between the prediction against the actual in Figure 71 clearly shows the inability of SARIMAX(1,0,1)x(0,0,0,7) to predict the actual values in the training dataset accurately. Similarly, the line plot in Figure 70 strongly indicates the failure of SARIMAX(1,0,1)x(0,0,0,7) in predicting the testing dataset. Surprisingly, the model even fatally estimated the number of patient demand by returning negative values on several occasions, such as in late March 2017 and in the middle of April 2017. The analysis and explanation further on this problem will be discussed in the last section.

6.1.2 Forecasting ED patient demand: Result of SARIMAX model with Feature Selection

Applying Lasso on 1132 features significantly reduced a large number of features by returning only 52 significant features. These 52 features are significant because their coefficient values are not zero. In other words, Lasso only selects 52 features which have a contribution to the forecasting through their coefficient values. Even though these 52 selected features have a contribution to the model performance, actually the amount of their contribution varies with diverse weight (Table 33 in Appendix). So, their individual contribution to the ML model performance also varies depending on the relative weight values.

Consequently, reducing the number of features further is possible by following the steps, as stated in Chapter 3-Methodology, particularly in Feature Selection section. As visualized in Figure 29, the small number of features initially caused a high RMSE for training and testing. After that, RMSE immediately dropped to its lowest value until eventually recovered to relative stable RMSE value regardless of more features being added. So, Figure 29 suggests that adding more features do not necessarily improve the ML model performance.

Based on Figure 29, only six features are selected out of the initial 52 features because they yielded the best performance in the testing dataset. These six features are Is_Weekday, GP_Post_WH_Opening, W_TX-1, ICPCcode_L-1, Is_Weekday-2, Is_Weekday-3. The result of a VIF statistical test to check the presence of multicollinearity among these six features can be found in Table 34 in the Appendix. All the VIF

score are less than 10, which is still acceptable as the threshold of VIF score is 10 [64]. In other words, these six selected features do not indicate having significant multicollinearity.

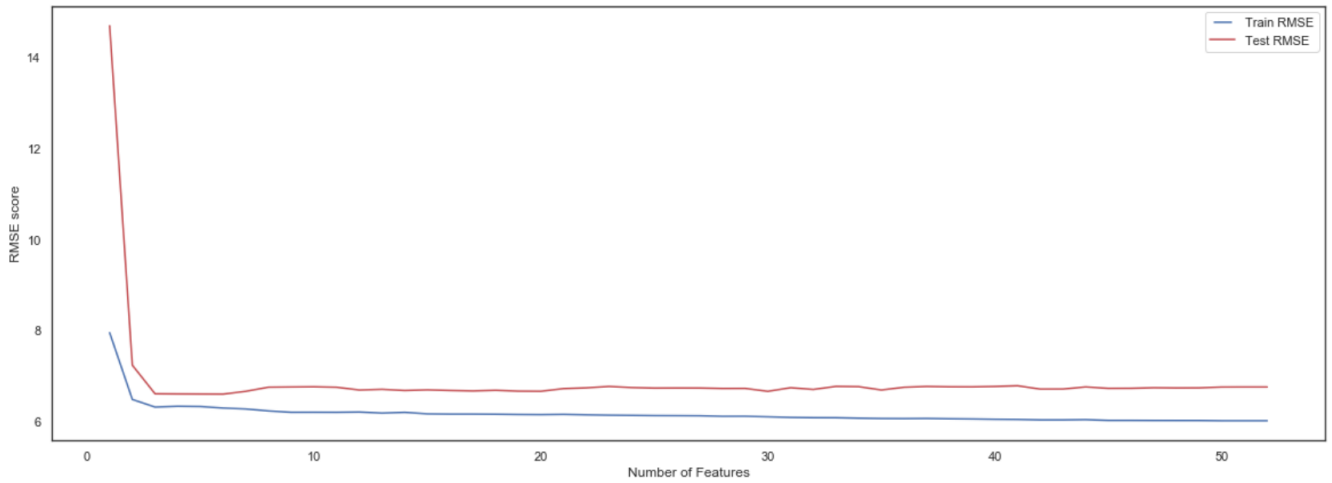


Figure 29: Trade-off plot in forecasting ED patient demand

The best SARIMAX model using only six selected six features is SARIMAX(0,0,0)x(1,0,1,7). The error metrics of SARIMAX(0,0,0)x(1,0,1,7) is presented in Table 18. Although the overall training error metrics in Table 18 are worse than training error metrics in Table 17, the performance of SARIMAX(0,0,0)x(1,0,1,7) on the testing dataset is significantly better, and the result gaps between the training and testing are much closer. Besides, the mean and standard deviation of residue on both training and testing in Table 18 are small. The relatively small difference between error metrics in Training and Testing indicates the ability of the ML model to learn from only six features in the training dataset and then generalize them to the testing data.

The coefficient and p-values of the six features are presented in Table 19. The p-values indicate that the first four features in Table 19 are statistically significant. Is_Weekday has the highest positive coefficient, so it has the most positive effect on the number of ED patient demand. Conversely, GP_Post_WH_Opening has a negative effect, as indicated by its negative coefficient. The further analysis of the selected features are provided in the next discussion section, and the complete list of the coefficient and p-values, including AR and MA components, can be found in Table 35 in Appendix.

Table 18: Result of SARIMAX(0, 0, 0)x(1, 0, 1, 7) model with Feature Selection

SARIMAX(0, 0, 0)x(1, 0, 1, 7)	Training (Dataset 2013-2016)	Testing (Dataset 2017)
AIC	9489.52	-
MAPE	15.84%	16.54%
RMSE	6.28	6.59
MAE	5.01	5.28
Mean of Residue	0.14	0.14
Std of Residue	6.29	6.60

The forecasting result of SARIMAX(0,0,0)x(1,0,1,7) is plotted in Figure 72. In Figure 72, the ML model can follow the daily fluctuation of ED patient demand with slightly up and down prediction values on many occasions. However, the ML model seems to inaccurately predict the extremely high spikes or the extremely down spikes. Figure 72 also reflects the worse performance of SARIMAX(0,0,0)x(1,0,1,7) on the training dataset than the previous SARIMAX order (Figure 67). Nevertheless, SARIMAX(0,0,0)x(1,0,1,7) is far more superior in predicting the unseen testing data as visualized Figure 75. In Figure 74, the histogram of the training residue follows the normal distribution, which indicates white noise. However, the correlogram suggests that the residue is not random as shown by several spikes whose values more than significant level. Hence, the correlogram in Figure 74 indicates the presence of residue that is not captured by SARIMAX(0,0,0)x(1,0,1,7).

Table 19: ED six selected features

Variable	Six selected features	Description	Coefficient	P>|z|
x1	Is_Weekday	Indicator variable, 1=weekday and 0=Not-weekday	17.73	0.00
x2	GP_Post_WH_Opening	Indicator variable, 1=GP-Post working hour (08.00-17.00) during holiday, 0=GP-Post working hour (17.00-08.00)	-4.02	0.00
x3	W_TX-1	Max Temperature of yesterday	6.54	0.00
x4	ICPCcode_L-1	Number of ICPC code L patient Yesterday	12.27	0.00
x5	Is_Weekday-2	Is_Weekday of two days ago	9.00	0.15
x6	Is_Weekday-3	Is_Weekday of three days ago	9.61	0.13

6.1.3 Forecasting ED patient demand: Result of a Hybrid model

Table 20 shows the metric evaluation result after performing the Hybrid model in the training dataset. In comparison with the result in Table 18, the Hybrid model slightly improves the performance in the training dataset by reducing only 0.01 point of RMSE and MAE. However, this improvement does not apply for MAPE, which increases by 0.03 point instead. In the testing dataset, the Hybrid model performance outperforms the SARIMAX(0, 0, 0)x(1, 0, 1, 7) in all metric evaluations. However, the improvement is relatively small, with only 0.03-0.04 points. Among 1132 features, only five are identified as significant after selected by feature importance method of GradientBoostingRegressor. These five selected features are Gender_Man-1, Age_66-74_year-7, W_DDVEC-6, W_Q-7, dayofyear_sin365-6. Visualizing the forecasting values of Hybrid model in Figure 77 and Figure 78 indicates the ability of the Hybrid model in following seasonal upward and downward pattern in the long term. However, similar to Figure 72 and Figure 75, the Hybrid model in Figure 77 and Figure 78 still struggles in predicting certain days with extremely high or deficient demand.

Table 20: Result of a Hybrid model

Hybrid Model	Training (Dataset 2013-2016)	Testing (Dataset 2017)
MAPE	15.87%	16.50%
RMSE	6.27	6.56
MAE	5.00	5.25

6.2 Result of Forecasting GP-Post patient demand

This section consists of three different results from three type of modeling, namely: a single SARIMAX model, a SARIMAX model with Feature Selection, and a Hybrid model. Each model generates two metrics evaluation results based on two different datasets: training and testing. Firstly, the metric evaluation result of the training and testing of each model will be presented. After that, the comparison between the training result and the testing result as well as the comparison among the result of three different models will be analyzed and discussed.

6.2.1 Forecasting GP-Post patient demand: Result of a single SARIMAX model

Auto_arima function returns SARIMAX (1, 0, 2)x(0, 0, 0, 7) as the best model predictor with AIC 11685.63 . Table 21 provides the metric evaluation result for both training and testing, including the residue information. Looking at the error metrics of the testing result, particularly RMSE and MAE, both values indicate that SARIMAX (1, 0, 2)x(0, 0, 0, 7) can predict better than the interval of standard deviation which is 6.7. MAPE value less than 10% is also a good indication of SARIMAX (1, 0, 2)x(0, 0, 0, 7) performance. Analyzing through visualization in Figure 79 and Figure 80 clearly shows the close proximity between the prediction values against the actual values. Also, the residue diagnosis in Figure 81 strongly indicates the white noise pattern with random residue values, normal distribution histogram and density, correlogram points which are always under the significant range, and QQ plot which shows the fitness of the forecasting values. With all these indications, SARIMAX (1, 0, 2)x(0, 0, 0, 7) performance on the testing dataset is sufficiently accurate.

Table 21: Result of a single SARIMAX (1, 0, 2)x(0, 0, 0, 7) model

SARIMAX(1, 0, 2)x(0, 0, 0, 7)	Training (Dataset 2013-2016)	Testing (Dataset 2017)
AIC	11685.63	-
MAPE	8.09%	48.11%
RMSE	6.16	40.59
MAE	4.71	28.66
Mean of Residue	-0.02	9.04
Std of Residue	6.16	39.63

However, a similar problem, as found in forecasting ED patient demand also occurs in forecasting GP-Post patient demand. There is a huge difference between error metrics in training and testing, including the difference of mean and standard deviation of residue as found in Table 21. MAPE, RMSE, and MAE in the testing result have increased about six times larger than in training result. The mean of residue has also jumped up 9 points while the standard deviation of residue has multiplied more than six times from the training result. Also, the visualization of the line plot in Figure 82 and the scatter plot in Figure 83 shows the fatal prediction error by returning the negative values. With this kind of errors, SARIMAX(1,0,2)x(0,0,0,7) is unable to learn or to generalize the pattern of the training data through its features. The further analysis and explanation of this problem will be discussed in the last section.

6.2.2 Forecasting GP-Post patient demand: Result of SARIMAX model with Feature Selection

Similar to ED feature selection process, GP-Post feature selection also used Lasso to reduce the number of features. Lasso can identify the 132 significant features out of total 1132 features, reduced almost 90% of the original numbers. The list of 132 features can be found in Table 36. Following the steps in the Feature Selection Section in Chapter 3-Methodology, these numbers can be reduced further by running auto_arima function iteratively and incrementally based on the weight of 132 features. As described in Figure 30, there is a trade-off point where using a fewer number of features will significantly increase the error (RMSE) while adding more features was no longer improve the performance. Combining with collinearity analysis and variance inflation factor (VIF) test to take out the multicollinearity variables, the seven features were selected with tolerable VIF test score. These seven selected features are Is_Weekday, GP_Post_WH_Opening, U5-7, Is_Holiday-1, Is_Weekday-1, Is_Weekday-3, Is_Weekday-5. The VIF test result of these features, as found in Table 37, shows their VIF factors are still under the threshold value (10). Feeding these feature to auto_arima function returned the best order SARIMAX(1,0,1)x(1,0,1,7) by AIC 11776.18.

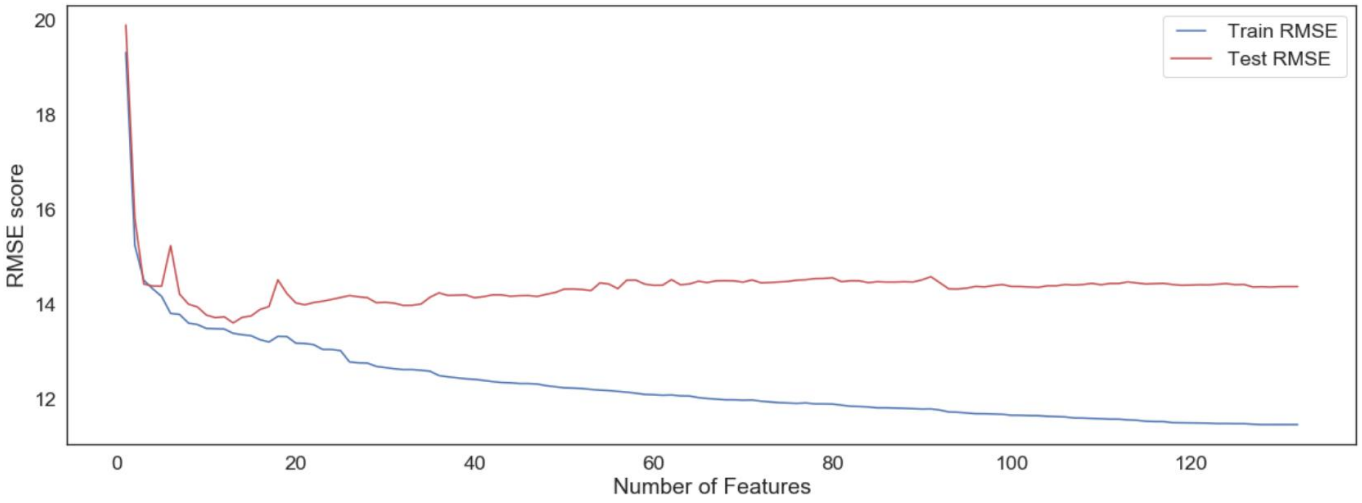


Figure 30: Trade-off plot in forecasting GP-Post patient demand

The complete result of applying SARIMAX(1,0,1)x(1,0,1,7) with only seven features can be found in Table 22. Similar to the ED forecasting, the training result of GP-Post forecasting are worse in all error metrics than before the performing feature selection process. However, SARIMAX(1,0,1)x(1,0,1,7) learns better in generalizing the pattern in the testing dataset with only seven features than SARIMAX(1,0,2)x(0,0,0,7) in the

prior section. This generalization is reflected by the small variance between training and testing in Table 22, while the opposite condition occurs in Table 21.

Table 22: Result of SARIMAX (1, 0, 1)x(1, 0, 1, 7) model with Feature Selection

SARIMAX(1, 0, 1)x(1, 0, 1, 7)	Training (Dataset 2013-2016)	Testing (Dataset 2017)
AIC	11776.18	-
MAPE	13.60%	13.75%
RMSE	13.77	14.20
MAE	9.45	9.52
Mean of Residue	0.23	-0.15
Std of Residue	13.78	14.22

The coefficient and p-values of the seven features are presented in Table 23. The p-values indicate that all the seven features are statistically significant with different prediction power as indicated by their coefficient values. As opposed to the ED result in Table 19, Is_Weekday has a negative coefficient value or has a negative effect on the number of GP-Post patient demand. This is reasonable since GP-Post operating hours are shorter in weekdays, so the number of patients becomes fewer in comparison to the weekend where GP-Post operating hours are longer; hence, the number of patients become more. Another opposing result is that GP_Post_WH_Opening has positively affected GP-Post patient demand while it has a negative effect on ED patient demand (Table 19). The further analysis of the selected features is provided in the next discussion section.

Table 23: GP-Post seven selected features

Variable	Seven Selected Features	Description	Coefficient	P>|z|
x1	Is_Weekday	Indicator variable, 1=weekday and 0=Not-weekday	-77.88	0.00
x2	GP_Post_WH_Opening	Indicator variable, 1=GP-Post working hour (08.00-17.00) during holiday, 0=GP-Post working hour (17.00-08.00)	74.79	0.00
x3	U5-7	Number of U5 patients seven days ago	12.51	0.00
x4	Is_Holiday-1	Indicator variable of yesterday, 1=holiday and 0=Not-holiday	18.25	0.00
x5	Is_Weekday-1	Is_Weekday of yesterday	48.97	0.00
x6	Is_Weekday-3	Is_Weekday of three days ago	65.18	0.00
x7	Is_Weekday-5	Is_Weekday of five days ago	61.25	0.00

The results of the visualization in Figure 84 until Figure 88 are consistent with the metric error evaluations as discussed before. In Figure 84, the training result of a SARIMAX SARIMAX (1, 0, 1)x(1, 0, 1, 7) model with feature selection was able to follow the up and down weekly pattern even though it is not as accurate as a single SARIMAX in Figure 79. Similarly, the training scatters plot in Figure 85 is not as close as

the plot in Figure 80. The result in Figure 84 and Figure 85 are reasonable since the error metrics on the testing dataset of the first mentioned are higher than the latter. Conversely, the testing visualization in Figure 87 shows the superiority and stability of a SARIMAX (1, 0, 1)x(1, 0, 1, 7) model with feature selection where it can tightly fit the actual pattern of GP-Post patient demand in 2017 as opposed to the testing plot of a single SARIMAX only as described in Figure 82. The testing scatters plot in Figure 88 aligned with Figure 87 in which it shows values pattern closer to the diagonal line in comparison to the testing scatter plot in Figure 83.

6.2.3 Forecasting GP-Post patient demand: Result of a Hybrid model

The metric evaluation result of a hybrid model is presented in Table 24. In comparison with the result in Table 22, the Hybrid model slightly improves the performance in the training dataset by reducing 0.34 of MAPE and only 0.01 point of RMSE and MAE. In the testing dataset, Hybrid model performance is better in all metric evaluations than the SARIMAX(1, 0, 1)x(1, 0, 1, 7) with seven selected features. However, the improvement is considerably small, with only 0.05-0.26 points.

Among 1132 features, 196 features were initially identified as significant by GradientBoostingRegressor with hyperparameter configuration. However, as presented in Figure 31, reducing the 196 features further is possible by following a similar approach as in the feature selection section in Chapter 4-Methodology. As a result of that, eight features are selected as the optimal number as adding more features was no longer improve the performance while reducing more made the performance worse. These eight selected features are GP_Post_Age_75_plus-7, ICPCcode_L-7, ICPCcode_S-2, W_TG-1, W_Q-3, W_DR-1, W_EV24-2, dayofyear_sin365-1.

Table 24: Metric Evaluation after Hyper-Parameter Tuning

GradientBoostingRegressor (with hyperparameter)	Training (Dataset 2013-2016)	Testing (Dataset 2017)
MAPE	13.26%	13.70%
RMSE	13.19	13.94
MAE	9.09	9.43

Figure 89 points out the ability of the Hybrid model following the seasonal upward and downward pattern in the long term. However, Figure 89 still struggled in predicting certain days with an extremely high and extremely low number of patient demand. For example, in the late of July 2016 (Figure 89), the hybrid model fails in caching the spike of patient demand. Another example in the same year about the second week of October shows that it could not predict the lowest demand point. Nevertheless, the forecasting plot of the hybrid model in the testing dataset as described in Figure 90, ensures that any prediction still falls within a significance level, with a 95% confidence interval.

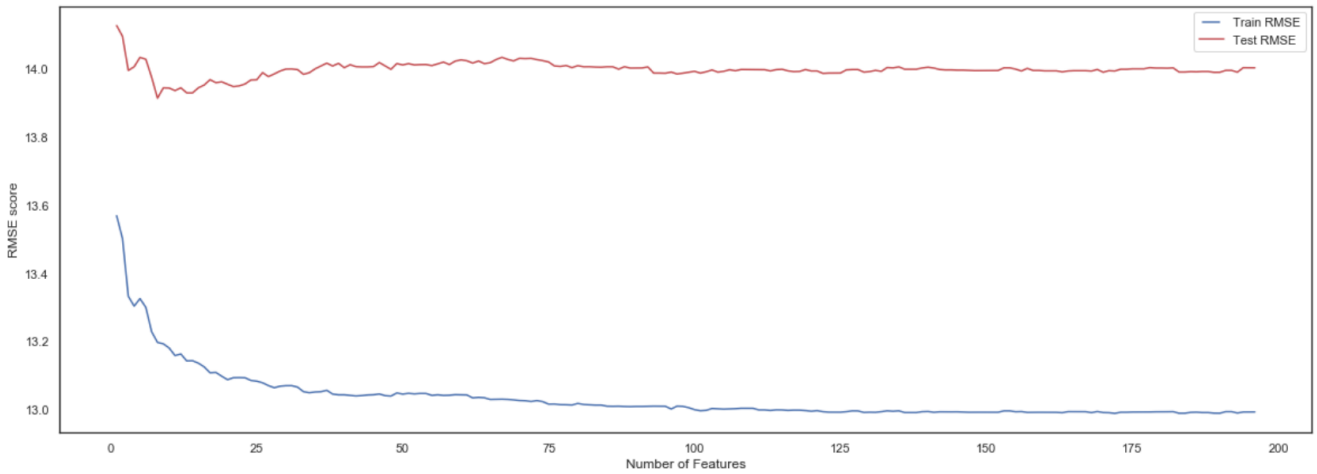


Figure 31: Trade-Off Plot in forecasting GP-Post patient demand with a Hybrid model

6.3 Model building for predicting ED inpatient admission to the hospital

As opposed to the two previous main sections (1) and (2), this section consists of four different results from four type of modeling: AdaBoostingClassifier, GradientBoostingClassifier, RandomForestClassifier, and ExtraTreesClassifier. The best model then is selected based on their metrics evaluation on the training dataset. Further, the optimization of the best model is performed, and its result on testing and testing are presented.

6.3.1 Result of inpatient admission

The performance result of four ensemble algorithms in predicting ED inpatient admission is presented in Table 25. The performances of four ensembles using the default configuration were tight with small variance in all five metrics evaluation. In general, the boosting classifier method, namely AdaBoostingClassifier, and GradientBoostingClassifier, outperformed the begging methods, namely RandomForestClassifier and ExtraTreesClassifier. AdaBoostingClassifier performed slightly better than GradientBoostingClassifier in the Precision score.

Table 25: Metric Evaluation of ED inpatient admission with the default configuration

ML Classifier Models	Accuracy	Precision	Recall	F1 Score	ROC-AUC
AdaBoostClassifier	78%	74%	71%	72%	85%
GradientBoostingClassifier	78%	73%	73%	73%	85%
RandomForestClassifier	76%	71%	70%	70%	81%
ExtraTreesClassifier	76%	71%	67%	69%	80%

Amongst all the classification algorithms, GradientBoostingClassifier returned the highest recall value, 0.73 or 73%. GradientBoostingClassifier also outperformed AdaBoostingClassifier in term of F1 score, which is the harmonic combination between Precision and Recall. Therefore, GradientBoostingClassifier was selected as the classifier model for further optimization. The result of optimization via hyper-parameter on the training dataset is provided in Table 26, which slightly improved ROC-AUC value by 0.01 point.

Table 26: Optimization with Hyper parameter

GradientBoostingClassifier (Training Dataset)	Accuracy	Precision	Recall	F1 Score	ROC-AUC
GradientBoostingClassifier	78%	73%	73%	73%	86%

The result of the testing dataset is provided in Table 27, where all the metrics evaluations returned the same result as the training, except for ROC-AUC, which dropped by almost 0.1 points. Plotting ROC curve and Precision-Recall curve are described in Figure 97 and Figure 98. Besides, only 21 out of 39 features have weight values, and they are selected as feature importance based on GradientBoostingClassifier as presented in Figure 99 and Table 39. Further analysis of these results will be discussed in the next section.

Table 27: Testing Result of predicting inpatient admission

GradientBoostingClassifier (Testing Dataset)	Accuracy	Precision	Recall	F1 Score	ROC-AUC
GradientBoostingClassifier	78%	73%	73%	73%	77%

6.4 Discussion

This section will discuss several valuable findings based on previous results. The finding in relation to ED and GP-Post will be discussed together in one section due to their similar and comparable results. The last section will discuss the finding in predicting inpatient admission

6.4.1 Discussion on the Result of Forecasting ED and GP-Post patient demand

The first finding to be discussed and analyzed is the similar phenomenon occurred in Forecasting ED and GP-Post patient demand with a single SARIMAX model, which was the huge different error metric result on the training and the testing dataset. The visualization analysis also indicated that SARIMAX(0, 0, 0)x(1, 0, 1, 7) of ED and SARIMAX(1, 0, 2)x(0, 0, 0, 7) of GP-Post extremely fit only on the training dataset while both SARIMAX performed poorly on the testing dataset. These results indicate the overfitting phenomenon, as explained in Chapter 3, section Bias Vs. Variance Analysis. A recent study [65] which used ARIMA (another version of SARIMAX) also suffered from the same overfitting problem while the two-hybrid model namely [38] and [39], did not report this problem.

Another explanation of the overfitting, as pointed out by research [20], it is the side effect of multicollinearity where there is a high correlation among features. As discussed in Chapter 4, the correlation among the features, such as Age group against temperatures, occurred in variation level. The more significant correlations logically occur within the same group feature, such as in weather features between Max Temperature against Average Temperature. One of the solutions proposed to overcome such a problem is to perform regularization methods such as Lasso [20]. After performing feature selection with Lasso, only six out of 1132 features are selected in section 6.1.2 and only seven out of 1132 features selected in section 6.2.2. Moreover, the result in section 6.1.2 and section 6.2.2 between training and testing indicated that the overfitting was no longer found. So, performing Lasso is not only for feature selection purpose, but it is also useful for removing the overfitting as a side effect of multicollinearity, as also stated in [20].

The next interesting finding from the previous sections, either in the case of ED or GP-Post, is that a hybrid model outperformed a single SARIMAX and SARIMAX with feature selection. The superiority of a hybrid model in this thesis aligns with the result of the similar researches using a hybrid model such as ARIMA-LR [38] and ARIMA–ANN [39], including other research from the outside of healthcare domain [40-43]. However, the improvement done by the hybrid model in this thesis is relatively small with less than 1%.

Another finding, especially in correlation with exogenous or external features, reveals that only W_TX-1 (Max Temperature of yesterday) came up as a statistically significant feature in forecasting ED patient demand (Table 19) while none of the exogenous features found as significant in forecasting GP-Post patient demand (Table 23). However, to make a conclusive statement that “only W_TX affecting ED patient demand while none of the external features affecting GP-Post patient demand” is problematic for several reasons. Firstly, the initial number of selected features with Lasso is 52 features of ED (Table 33) and 132 features of GP-Post (Table 36), which includes not only W_TX but also other external features of weather and pollens. The effect of weather-related data such as temperature, humidity, and precipitation in Table 33 and Table 36 to the patient demand is consistent with the result of several studies [23, 25]. However, in this thesis, further feature selection was performed. As a result of that, only a few remain features came up after further optimization process, as explained in section 6.1.2 and 6.2.2. Secondly, the feature selection method using Lasso itself has a limitation where it is primarily designed to reduce the number of features, so Lasso tends to pick only one feature among a group features with high correlation [20]. This also explains why some high correlated features with W_TX, as pointed out in Chapter 4, did not come up in the few remain

list. Lastly, the pairwise analysis using Pearson Correlation in Chapter 4 also highlighted that besides W_TX, there were fair correlations of some external features against ED (e.g., Figure 42, Figure 52) and GP-Post patient demand (e.g., Figure 40, Figure 52).

The last finding worth to discuss is the presence of GP_Post_WH_Opening as a significant feature for both ED and GP-Post. This feature was the last feature to be included only after performing the domain analysis, as suggested by this study [44]. Another paper [23] also suggested that some local related data or variables might be useful for predicting the ED arrivals. So, including GP_Post_WH_Opening improved the ML model performance. As GP_Post_WH_Opening feature provides the flag or status on which day GP-Post operates during working hours, it helps the ML model to predict the number of patients accordingly. Interestingly, the effects of GP_Post_WH_Opening against ED and GP-Post are not the same. It negatively affected ED while conversely it positively affected GP-Post. This might be interpreted that the number of ED patient demand tends to decrease while GP-Post increases instead during the holiday where GP-Post operates in working hours (GP_Post_WH_Opening=1).

6.4.2 Discussion on the Result of Predicting ED inpatient admission

The decision in selecting GradientBoostingClassifier as the best model in Table 25 was mainly based on Recall value. Unlike the Precision, Recall score indicated the proportion of true results overall the output predictions done by the ML model, using the formula TP/(TP+FN). Since the primary interest in this project is to know how well the ML model predicted the admitted patient (Label=1), then recall score was more representative than the precision score. The formula of Precision and Recall in relation to the actual condition and ML output prediction is presented in Table 28.

Table 28: Precision and Recall in this thesis

PRECISION = TP / (TP+FP) RECALL = TP / (TP+FN)			ML Output Prediction	
			Admitted	Not-Admitted
			Positive (1)	Negative (0)
Actual Condition (Based on Label_Admittance column)	Admitted	TRUE (1)	True Positive (TP)	False Negative (FN)
	Not-Admitted	FALSE (0)	False Positive (FP)	True Negative (TN)

Based on Table 28, having 100% Precision might imply that the value of False Positive (FP) was 0 because of TP/(TP+0) = 100%. However, 100% of precision score did not inform about how many True Positive (TP) were correctly predicted because as long as FP=0, any value of TP would produce 100% precision. Conversely, recall value included False Negative (FN) as the denominator in its formula, TP/(TP+FN), to indicate how sensitive ML model predicts the Admitted or TRUE (1) condition. So, having FN value closer to 0 (or high recall) was better than having FP close to 0 (or high precision) in the context of this research. In other words, the ML performance never mistakenly predicted the number of the actual inpatient admission, as more represented by Recall score, is more crucial than being able to predict 100% of the Non-admitted patient as more represented by Precision score.

The relation between precision and recall can be analyzed further by looking at the ROC curve (Figure 97) and Precision-Recall curve (Figure 98). ROC curve in Figure 97 shows the possibility to further improve the performance of GradientBoostingClassifier by adjusting the threshold so that the values of TP and FP will

also be adjusted accordingly. By default, the threshold value is 0.5, which means that all the scoring values above 0.5 are interpreted as "positive" (1) while everything below 0.5 is "negative" (0). GradientBoostingClassifier did not just predict "positive" (1) or "negative" (0) as a prediction outcome, but actually it generated a probability scores which are a real number between 0 and 1, and then used the threshold setting to decide if the prediction is a "positive" (1) or a "negative" (0). So, adjusting the threshold also implies the changes in the balance between precision and recall. In other words, there is a trade-off between Precision and Recall as further described in Figure 98. Improving the recall scores at the x-axis has a consequence of decreasing precision at the y-axis. Ideally, selecting the right threshold should analyze further the average error cost per prediction. However, since the cost of FP and FN has not yet identified in this research, the threshold value remains as default 0.5.

Another finding to discuss is about the feature importance in Figure 99 and Table 39 as generated by GradientBoostingClassifier. Based on the importance values, only 21 out of 38 features have feature importance values, while others are zero. The top three most important features based on the importance values in Figure 99 are Treatment_CHI (Patient with surgery treatment), followed by Age, and Urgency_Green as the third. Although Figure 99 ranks them based on their importance values, making a firm conclusion that, "it is the fixed ranking list and only these 21 features are the most significant features in predicting inpatient admission", has several problematic issues. First, the feature importance of GradientBoostingClassifier is sensitive to its parameters. In other words, different parameter configurations can generate different feature importance score and rank. Second, the relevant papers [52-54] did not firmly conclude the effect of selected features against the prediction outcome; rather, they more emphasized on the improvement result made by ML model after feature selection. Moreover, among the similar and relevant literature listed in Table 3, only the first one [31] directly associated its result with its selected features while the others only mentioned the presence of certain features. Therefore, a better conclusion for interpreting the feature important result in Figure 99 is to state that 21 out of 38 features yielded the optimal prediction of inpatient admission using GradientBoostingClassifier as indicated in Table 26 and the result of the testing dataset is provided in Table 27.

7. Conclusion, Limitations, and Recommendations

In this chapter, the conclusion is provided to answer the research questions based on the implementation result. Besides, this chapter highlights the relevant limitations of the study. Finally, the recommendations are also given to improve a similar project in the future.

7.1 Conclusion

This section concludes the report by offering answers to the sub-questions and the main project question as discussed below. Furthermore, to completely answer the main research question, the following sub-questions are needed and stated:

Sub Question-1: Which ML methods can be applied to forecast patient demand at the ED?

Among various available ML methods, Literature gap in chapter 2 explained the motivation of selecting a hybrid model which combined SARIMAX and Gradient Tree Boosting. As a benchmark, two scenarios were built namely a single SARIMAX model without feature selection and a single SARIMAX model with Feature Selection. Chapter 3 further described the methodology and framework in applying and implementing them on this thesis. In the hybrid model, SARIMAX functioned as the primary model in forecasting time series with external variables (weather and pollen) while Gradient Tree Boosting aimed to capture and predict the non-linear pattern and correlation in the SARIMAX's residue which was not optimally predicted by SARIMAX. Feature Selection method with Lasso was also used not only to simplify the ML model but also to optimize the result at the same time by selecting only the relevant features before feeding them to SARIMAX model.

Sub Question-2: Which ML methods can be applied to forecast patient demand at the GP-Post?

Similar to the answer on SubQuestion 1, forecast GP-Post patient demand also used a hybrid model where SARIMAX was the main model, and Gradient Tree Boosting was the predictor of SARIMAX's residue. SARIMAX and Feature Selection with Lasso were also used while a single SARIMAX model without Feature Selection was still presented in Chapter 6 as a benchmark.

Sub Question-3: Which ML methods can be applied to predict the ED inpatient admission to the hospital?

Four ensemble methods were used for predicting ED inpatient admission to the hospital. Among these four, GradientBoostingClassifier returned the best performance in all metric evaluation, especially the recall score. Applying hyper-parameter techniques via GridSearchCV on GradientBoostingClassifier improved the performance slightly, especially in AUC score.

Sub Question-4: What insights can be derived from Exploratory data analysis (EDA) in relation to univariate time series analysis and correlation analysis?

In chapter 4-Exploratory Data Analysis (EDA), two types of analysis were performed: univariate time series analysis and correlation analysis. In univariate time series analysis, two insights were derived. First, the ADF test showed that the ED and GP-Post daily patient demand was stationary. Second, ACF and PACF plotting indicated the correlation between the ED and GP-Post daily patient demand with their lagging values. In correlation analysis, the insights were distributed in three main interests: (1) the temperature and humidity correlation with Age Groups of ED and GP-Post, (2) the temperature and humidity correlation with Treatment Groups of ED and GP-Post, (3) the spread of pollen correlation with Treatment Groups of ED and GP-Post.

In (1) generally, the correlation on weekends tended to strengthen positively with temperature and negatively with humidity, of which the 5-19 and 20-65 were the most sensitive Age groups. Similarly, in (2), the correlation on weekends tended to stronger positively with temperature and negatively with humidity, of which ICPCcode_S and ICPCcode_R were the most sensitive treatment groups. In (3), selecting Urtica, Betula, Parietaria as the examples, showed that Urtica had a fair and positive correlation with ICPCcode_S during all-weekdays, all-weekends, autumn-weekdays, and autumn-weekends, Betula did not have an indication of having fair correlate with any of the treatment groups in all period, and Poaceae had a positive and fair correlation with ICPCcode_L (all-weekends, summer-weekends) and ICPCcode_S (autumn-weekends).

With all these insights, the research objective-3, which is to analyze a linear correlation between external factors with some particular patient groups, has been met and achieved.

Sub Question-5: Which ML model gives the best prediction result for subquestion 1, 2, and 3?

In the case of ED, the summary result of each model was presented in Table 29 below. A single SARIMAX model without Feature Selection suffered from the overfitting problem as indicated by the big difference between the metric error evaluation during training and the testing. Performing Lasso in the Feature Selection process improved the stability of SARIMAX model in generalizing the pattern during training and testing so that the variance error results on both were significantly reduced. Not only reducing the variance but implementing Lasso also improved the SARIMAX performance on the testing. However, in term of performance, the hybrid model outperformed SARIMAX with Lasso.

Table 29: The Summary result of ED patient demand forecasting

ED	SARIMAX (1,0,1)x(0,0,0,7)		SARIMAX(0,0,0)x(1,0,1,7) with Feature Selection		Hybrid Model of SARIMAX(0,0,0)x(1,0,1,7) and Gradient Tree Boosting	
Evaluation Metric	Training	Testing	Training	Testing	Training	Testing
AIC	9811.79	-	9489.52	-	-	-
MAPE	7.81%	38.99%	15.84%	16.54%	15.87%	16.50%
RMSE	3.24	16.77	6.28	6.59	6.27	6.56
MAE	2.48	12.81	5.01	5.28	5.00	5.25

In the case of GP-Post, the summary result of each model was presented in Table 30 below. Overfitting still found as the main issue with a single SARIMAX model in forecasting GP-Post patient demand. Lasso was indeed successful in avoiding the overfitting issue, reducing the number of features and eventually improving the SARIMAX performance. Further minor optimization was achieved by using a hybrid model. Hence, a hybrid model also came up as the best ML model for forecasting GP-Post patient demand.

Table 30: Summary result of GP-Post patient demand forecasting

GP-Post	SARIMAX (1, 0, 2) x (0, 0, 0, 7)		SARIMAX (1, 0, 1) x (1, 0, 1, 7) with Feature Selection		Hybrid model of SARIMAX (1, 0, 1) x (1, 0, 1, 7) and Gradient Tree Boosting	
Evaluation Metric	Training	Testing	Training	Testing	Training	Testing
AIC	11685.63	-	11776.18	-	-	-
MAPE	8.09%	48.11%	13.60%	13.75%	13.26%	13.70%
RMSE	6.16	40.59	13.77	14.20	13.19	13.94
MAE	4.71	28.66	9.45	9.52	9.09	9.43

In the case of predicting inpatient admission, GradientBoostingClassifier came up as the best classifier. The result of five different metric evaluation was presented in Table 31.

Table 31: Testing Result of predicting inpatient admission with GradientBoostingClassifier

GradientBoostingClassifier (Testing Dataset)	Accuracy	Precision	Recall	F1 Score	ROC-AUC
GradientBoostingClassifier	78%	73%	73%	73%	77%

With all these results, the research objective-1, which is to forecast one day ahead of ED and GP-Post patient demand with machine learning techniques, and objective-2, which is to predict ED's inpatient admission with a machine learning technique, have been met and achieved. In addition to objective-1, there are two options of ML models can be used for forecasting ED and GP-Post, SARIMAX with Feature Selection or the hybrid model because the performance of both is almost the same.

Sub Question-6: Which features can yield the optimal prediction for subquestion 1, 2, and 3?

The features for subquestion 1 and 2 were analyzed through their weight in Feature Selection process via Lasso and also through their statistical significance as generated by the SARIMAX model. In the case of subquestion 1, initially, 52 features were selected. In the case of subquestion 2, 132 features were the initial number of selected features. Further optimization revealed that only six features and seven features were optimally required and selected to forecast ED and GP-Post patient demand, respectively. The six selected features for ED (Table 19) are Is_Weekday, GP_Post_WH_Opening, W_TX-1, ICPCcode_L-1, Is_Weekday-2, Is_Weekday-3 while the seven selected features for GP-Post (Table 23) are Is_Weekday, GP_Post_WH_Opening, U5-7, Is_Holiday-1, Is_Weekday-1, Is_Weekday-3, Is_Weekday-5. The features for subquestion 3 were analyzed through feature importance of GradientBoostingClassifier. As a result, 21 out of 38 features were selected and presented in Figure 99 and Table 39.

Main Research Question: ***To what extent can one utilize machine learning techniques in the acute care domain such as ED and GP-Post?***

The application of machine learning in the acute care domain was mainly categorized into three areas: Input, Throughput, and Output. The input is mainly dominated by the topics about forecasting patient demand or patient arrival, while the Throughput is mostly dealt with the topics about predicting patient's LOS, and the Output is typically covered by the topics about predicting the inpatient admission to the hospital. In this thesis, the Input and the Output were selected as the primary research area.

In the Input, two forecasting daily patient demand tools for ED and GP-Post were built using ML regression techniques. Literature study and Literature Gap Analysis in Chapter 2 identified SARIMAX as the main method for building the time series forecasting model with external variables. The result showed that SARIMAX did not robust to the overfitting problem. Applying other ML technique, which was Feature Selection with Lasso, can help to not only overcome the overfitting problem but also reduce the complexity of the ML model by selecting the most relevant features. Other than that, Literature Gap Analysis also found out that the Hybrid model recently becoming one of the popular ML forecasting techniques but have not been extensively explored in the acute care domain. Gradient Tree Boosting as the novel ML technique in handling the non-linearity problem was selected to forecast the residue generated by SARIMAX model. The expectation of combining SARIMAX and Gradient Tree Boosting was to effectively forecast time series data with external variables and optimally forecast its residue so that the performance can be significantly improved. Three experiment scenarios (single SARIMAX, SARIMAX with feature selection, hybrid) were designed, implemented, and compared. The result of this thesis showed that the Hybrid model of SARIMAX and Gradient Tree Boosting was the best forecasting for forecasting ED and GP-Post patient demand.

In the Output, the ML classification techniques were used and developed to predict the inpatient admission to the hospital. Four ensemble techniques were used to learn 38 characteristics of patients and classify them into two binary states: 1 (admitted) or 0 (not admitted). The result showed that GradientBoostingClassifier returned the best prediction, especially in term of recall scores. The optimization through hyper-parameter techniques was able to improve the outcome. Further improvement is still possible with threshold adjustment via ROC and Precision-Recall curve.

7.2 Limitations

During the study, several limitations occurred. The list of limitations is provided below.

- Data limitation

The historical dataset used during the study is only from 2013 until 2017. It is then split into training (2013-2016) and testing (2017). So, the results of this study do not fully cover and reflect the more current dataset, which is 2018.

- Knowledge limitation

Due to limited time, the researcher mainly focuses on a hybrid model using SARIMAX and Gradient Tree Boosting. Although study literature and literature gap have been conducted before selecting the relevant ML model, a huge number of the ML models and their combinations make finding another much better ML models are widely open

- Interpretability of ML model limitation

One of the main objectives in this study is to forecast the patient demand so that the result can be used for anticipating the overcrowding event. So, the researcher mainly focuses on improving forecasting performance by reducing the error metrics. Lasso is purposely selected to not only reducing the number of features but also to improve the ML model performance by lowering the error rate. Consequently, the expected external features, such as pollen, can get eliminated during the improvement process.

In carrying out this research, the following validity threats were considered.

- Since the ML model was only tested on 2017 dataset, using it for forecasting the near future such as 2019 or 2020 might yield bigger bias or prediction error.
- The proposed ML hybrid models were tested on a local and personal computer. Implementing or deploying it for business operation might cause several problems such as long execution time
- The complete features consist of the ED's and GP-Post's features. However, in reality, ED and GP-Post have two separate databases, so the features of ED and GP-Post are not integrated

7.3. Recommendations

This section provides several recommendations for further implementation or a similar project in the future so that the result can be improved.

1. Including more recent data (e.g., 2018 or 2019) is recommended to improve the forecasting validity of the ML model.
2. If the main objective of the next research is getting better accuracy or lowering the bias, then another novel ML methods such as Deep Learning might improve the result. However, it has a drawback in which interpreting the ML model and result is quite difficult
3. If the main objective of the next research is assessing the effect of certain external features against the outcome prediction, then using different feature selection methods, other than Lasso, should be considered. Elastic Nets is more recommended than Ridge because Elastic Nets is designed to resolve the trade-off between L1 (Lasso) and L2(Ridge). So, Elastic Nets combines the strengths of both Ridge and LASSO while at the same time minimizing the negative impact of either of these procedures.
4. In real day to day business operation, the implementation of hybrid models has to be tested further. The logical consequence of adding more models is the ML model becomes more complicated either in the operational or during maintenance. Time execution on experimenting (training and testing), the implementation and the real business operation will naturally increase, so hybrid ML model tends to perform slower. With these cons, simpler and faster ML is more recommended for the real business operation. A single SARIMAX model with Feature Selection can be a good option.
5. Since the feature importance list in this thesis shows the dependency between ED-related features and GP-Post related features, it will make sense to integrate these two data in the real business operation. Incorporating ED and GP-Post data management does not only help in analyzing and modeling the ML model in the future, but it also makes the day to day operation more efficient.

References

[1] Skinner, J., Higbea, R., Buer, D., & Horvath, C. (2018). using predictive analytics to align ED staffing resources with patient demand: A hospital in Grand Rapids, Mich., used management theory and data analysis to design and implement a much more precise model for setting staffing levels in its emergency department. *Healthcare Financial Management*, *72*(2), 56-62.

[2] Lowthian, J. A., Curtis, A. J., Cameron, P. A., Stoelwinder, J. U., Cooke, M. W., & McNeil, J. J. (2011). Systematic review of trends in emergency department attendances: an Australian perspective. *Emergency Medicine Journal*, *28*(5), 373-377.

[3] Pines, J. M., Hilton, J. A., Weber, E. J., Alkemade, A. J., Al Shabanah, H., Anderson, P. D., ... & Kumar, V. A. (2011). International perspectives on emergency department crowding. *Academic Emergency Medicine*, *18*(12), 1358-1370.

[4] Hoot, N. R., LeBlanc, L. J., Jones, I., Levin, S. R., Zhou, C., Gadd, C. S., & Aronsky, D. (2008). Forecasting emergency department crowding: a discrete event simulation. *Annals of emergency medicine*, *52*(2), 116-125.

[5] Zhang, X. L., Zhu, T., Luo, L., He, C. Z., Cao, Y., & Shi, Y. K. (2013, July). Forecasting emergency department patient flow using Markov chain. In *2013 10th International Conference on Service Systems and Service Management* (pp. 278-282). IEEE.

[6] Van Der Linden, C., Reijnen, R., Derlet, R. W., Lindeboom, R., Van Der Linden, N., Lucas, C., & Richards, J. R. (2013). Emergency department crowding in The Netherlands: managers' experiences. *International journal of emergency medicine*, *6*(1), 41.

[7] Boyle, A., Coleman, J., Sultan, Y., Dhakshinamoorthy, V., O'keeffe, J., Raut, P., & Beniuk, K. (2015). Initial validation of the International Crowding Measure in Emergency Departments (ICMED) to measure emergency department crowding. *Emerg Med J*, *32*(2), 105-108.

[8] Jones, S. S., Allen, T. L., Flottemesch, T. J., & Welch, S. J. (2006). An independent evaluation of four quantitative emergency department crowding scales. *Academic Emergency Medicine*, *13*(11), 1204-1211.

[9] Weiss, S. J., Derlet, R., Arndahl, J., Ernst, A. A., Richards, J., Fernández-Frankelton, M., ... & Brautigan, M. (2004). Estimating the degree of emergency department overcrowding in academic medical centers: results of the National ED Overcrowding Study (NEDOCS). *Academic Emergency Medicine*, *11*(1), 38-50.

[10] Hoot, N. R., Zhou, C., Jones, I., & Aronsky, D. (2007). Measuring and forecasting emergency department crowding in real time. *Annals of emergency medicine*, *49*(6), 747-755.

[11] Raj, K., Baker, K., Brierley, S., & Murray, D. (2006). National Emergency Department Overcrowding Study tool is not useful in an Australian emergency department. *Emergency Medicine Australasia*, *18*(3), 282-288.

[12] Weiss, S. J., Ernst, A. A., & Nick, T. G. (2006). Comparison of the national emergency department overcrowding scale and the emergency department work index for quantifying emergency department crowding. *Academic Emergency Medicine*, *13*(5), 513-518.

[13] Garcia-Romero, M., Rita-Gáfaro, C. G., Quintero-Manzano, J., & Angarita, A. B. (2017). NEDOCS vs subjective evaluation,¿ Is the health personnel of the emergency department aware of its overcrowding?. Colombia Médica, 48(2), 53-57.

[14] Hwang, U., McCarthy, M. L., Aronsky, D., Asplin, B., Crane, P. W., Craven, C. K., ... & Rathlev, N. K. (2011). Measures of crowding in the emergency department: a systematic review. *Academic Emergency Medicine*, *18*(5), 527-538.

[15] Phillips, J. L., Jackson, B. E., Fagan, E. L., Arze, S. E., Major, B., Zenarosa, N. R., & Wang, H. (2017). Overcrowding and its association with patient outcomes in a median-low volume emergency department. *Journal of clinical medicine research*, *9*(11), 911.

[16] Van, M. D. L., Van, M. L., Gaakeer, M. I., Richards, J. R., Derlet, R. W., & Van, N. D. L. (2018). A different crowd, a different crowding level? The predefined thresholds of crowding scales may not be optimal for all emergency departments. *International emergency nursing*, *41*, 25-30.

[17] Dancey, C. P., & Reidy, J. (2007). Statistics without maths for psychology. Pearson Education.

[18] Chan, Y. H. (2003). Biostatistics 104: correlational analysis. Singapore Med J, 44(12), 614-9.

[19] Bujang, M. A., & Baharum, N. (2016). Sample size guideline for correlation analysis. World, 3(1).

[20] Schreiber-Gregory, D. N. (2018). Ridge Regression and multicollinearity: An in-depth review. Model Assisted Statistics and Applications, 13(4), 359-365.

[21] Gul, M., & Celik, E. (2018). An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, 1-22.

[22] Choudhury, A. (2019). Hourly Forecasting of Emergency Department Arrivals–Time Series Analysis. *Available at SSRN 3311030*.

[23] Whitt, W., & Zhang, X. (2019). Forecasting arrivals and occupancy levels in an emergency department. *Operations Research for Health Care*.

[24] Ekström, A., Nordberg, M., & Eriksson, O. (2018). Shorter waiting time, better emergency healthcare: Forecasting Stockholm's emergency department visits. *Model Assisted Statistics and Applications*, *13*(4), 377-385.

[25] Carvalho-Silva, M., Monteiro, M. T. T., de Sá-Soares, F., & Dória-Nóbrega, S. (2018). Assessment of forecasting models for patients arrival at emergency department. *Operations Research for Health Care*, *18*, 112-118.

[26] Juang, W. C., Huang, S. J., Huang, F. D., Cheng, P. W., & Wann, S. R. (2017). Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. *BMJ open*, *7*(11), e018628.

[27] Hertzum, M. (2017). Forecasting hourly patient visits in the emergency department to counteract crowding. *The Ergonomics Open Journal*, *10*(1).

[28] Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F. M., Preen, D. B., Stewart, L. M., & Fatovich, D. M. (2015). A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. *Journal of biomedical informatics*, *57*, 62-73.

[29] 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.

[30] 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 of biomedical informatics*, *42*(1), 123-139.

[31] Lucke, J. A., de Gelder, J., Clarijs, F., Heringhaus, C., de Craen, A. J., Fogteloo, A. J., ... & Mooijaart, S. P. (2018). Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. *Emerg Med J*, *35*(1), 18-27.

[32] Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *PloS one*, *13*(7), e0201016.

[33] O'donovan, F., Brecht, T., Kekeh, C., Su, Z., Boussios, C., Menon, V., ... & Fonarow, G. (2017). Machine Learning Generated Risk Model to Predict Unplanned Hospital Admission in Heart Failure. *Circulation*, *136*(suppl_1), A16855-A16855.

[34] Leegon, J., Jones, I., Lanaghan, K., & Aronsky, D. (2006). Predicting hospital admission in a pediatric Emergency Department using an Artificial Neural Network. In *AMIA... Annual Symposium proceedings. AMIA Symposium* (Vol. 2006, pp. 1004-1004). American Medical Informatics Association.

[35] Sun, Y., Heng, B. H., Tay, S. Y., & Seow, E. (2011). Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine*, *18*(8), 844-850.

[36] Peck, J. S., Benneyan, J. C., Nightingale, D. J., & Gaehde, S. A. (2012). Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine*, *19*(9), E1045-E1054.

[37] Golmohammadi, D. (2016). Predicting hospital admissions to reduce emergency department boarding. *International Journal of Production Economics*, *182*, 535-544.

[38] Xu, Q., Tsui, K. L., Jiang, W., & Guo, H. (2016). A hybrid approach for forecasting patient visits in emergency department. *Quality and Reliability Engineering International*, *32*(8), 2751-2759.

[39] Yucesan, M., Gul, M., & Celik, E. (2018). A multi-method patient arrival forecasting outline for hospital emergency departments. *International Journal of Healthcare Management*, 1-13.

[40] Li, P., & Zhang, J. S. (2018). A new hybrid method for China's energy supply security forecasting based on arima and xgboost. *Energies*, *11*(7), 1687.

[41] Moeeni, H., & Bonakdari, H. (2017). Forecasting monthly inflow with extreme seasonal variation using the hybrid SARIMA-ANN model. *Stochastic environmental research and risk assessment*, *31*(8), 1997-2010.

[42] Boualit, S. B., & Mellit, A. (2016, November). SARIMA-SVM hybrid model for the prediction of daily global solar radiation time series. In *2016 International Renewable and Sustainable Energy Conference (IRSEC)* (pp. 712-717). IEEE.

[43] Xu, S., Chan, H. K., & Zhang, T. (2019). Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. *Transportation Research Part E: Logistics and Transportation Review*, *122*, 169-180.

[44] van der Spoel, S., Amrit, C. A., & van Hillegersberg, J. (2016, March). The role of domain analysis in prediction instrument development. In *Big Data Interoperability for Enterprises (BDI4E) Workshop 2016*.

[45] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[46] Chen, Y., & Tjandra, S. (2014). Daily collision prediction with SARIMAX and generalized linear models on the basis of temporal and weather variables. *Transportation Research Record*, *2432*(1), 26-36.

[47] Parmezan, A. R. S., Souza, V. M., & Batista, G. E. (2019). Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*, *484*, 302-337.

[48] Metz, C. E. (1978, October). Basic principles of ROC analysis. In *Seminars in nuclear medicine* (Vol. 8, No. 4, pp. 283-298). WB Saunders.

[49] Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), 2079-2107.

[50] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[51] Fonti, V., & Belitser, E. (2017). Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics.

[52] Mohan, C., & Nagarajan, S. (2019). An improved tree model based on ensemble feature selection for classification. TURKISH JOURNAL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES, 27(2), 1290-1307.

[53] Ubing, A. A., Jasmi, S. K. B., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 10(1), 252-257.

[54] Yang, J., Yao, D., Zhan, X., & Zhan, X. (2014, June). Predicting disease risks using feature selection based on random forest and support vector machine. In International Symposium on Bioinformatics Research and Applications (pp. 1-11). Springer, Cham.

[55] van der Linden, N., Longden, T., Richards, J. R., Khursheed, M., Goddijn, W. M., van Veelen, M. J., ... & van der Linden, M. C. (2019). The use of an 'acclimatisation' heatwave measure to compare temperature-related demand for emergency services in Australia, Botswana, Netherlands, Pakistan, and USA. PloS one, 14(3), e0214242.

[56] D'Amato, G., Cecchi, L., Bonini, S., Nunes, C., Annesi-Maesano, I., Behrendt, H., ... & Van Cauwenberge, P. (2007). Allergenic pollen and pollen allergy in Europe. Allergy, 62(9), 976-990.

[57] Coates, S. J., Davis, M. D., & Andersen, L. K. (2019). Temperature and humidity affect the incidence of hand, foot, and mouth disease: a systematic review of the literature–a report from the International Society of Dermatology Climate Change Committee. International journal of dermatology, 58(4), 388-399.

[58] Engebretsen, K. A., Johansen, J. D., Kezic, S., Linneberg, A., & Thyssen, J. P. (2016). The effect of environmental humidity and temperature on skin barrier function and dermatitis. Journal of the European Academy of Dermatology and Venereology, 30(2), 223-249.

[59] Fleischer Jr, A. B. (2019). Atopic dermatitis: the relationship to temperature and seasonality in the United States. International journal of dermatology, 58(4), 465-471.

[60] Sargen, M. R., Hoffstad, O., & Margolis, D. J. (2014). Warm, humid, and high sun exposure climates are associated with poorly controlled eczema: PEER (Pediatric Eczema Elective Registry) cohort, 2004–2012. Journal of Investigative Dermatology, 134(1), 51-57.

[61] Watson, M., Gilmour, R., Menzies, R., Ferson, M., McIntyre, P., & New South Wales Pneumococcal Network. (2006). The association of respiratory viruses, temperature, and other climatic parameters with the incidence of invasive pneumococcal disease in Sydney, Australia. Clinical infectious diseases, 42(2), 211-215.

[62] Liu, Y., Guo, Y., Wang, C., Li, W., Lu, J., Shen, S., ... & Qiu, X. (2015). Association between temperature change and outpatient visits for respiratory tract infections among children in Guangzhou, China. International journal of environmental research and public health, 12(1), 439-454.

[63] García-Mozo, H. (2017). Poaceae pollen as the leading aeroallergen worldwide: a review. Allergy, 72(12), 1849-1858.

[64] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

[65] Janardhanan, D., & Barrett, E. (2017, December). CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models. In 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST) (pp. 55-60). IEEE.

[66] Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. Cognition, 118(1), 2-16.

[67] Xu, Z., Huang, G., Weinberger, K. Q., & Zheng, A. X. (2014, August). Gradient boosted feature selection. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 522-531). ACM.

Appendix A – Yearly Frequency Plot of ED patient demand

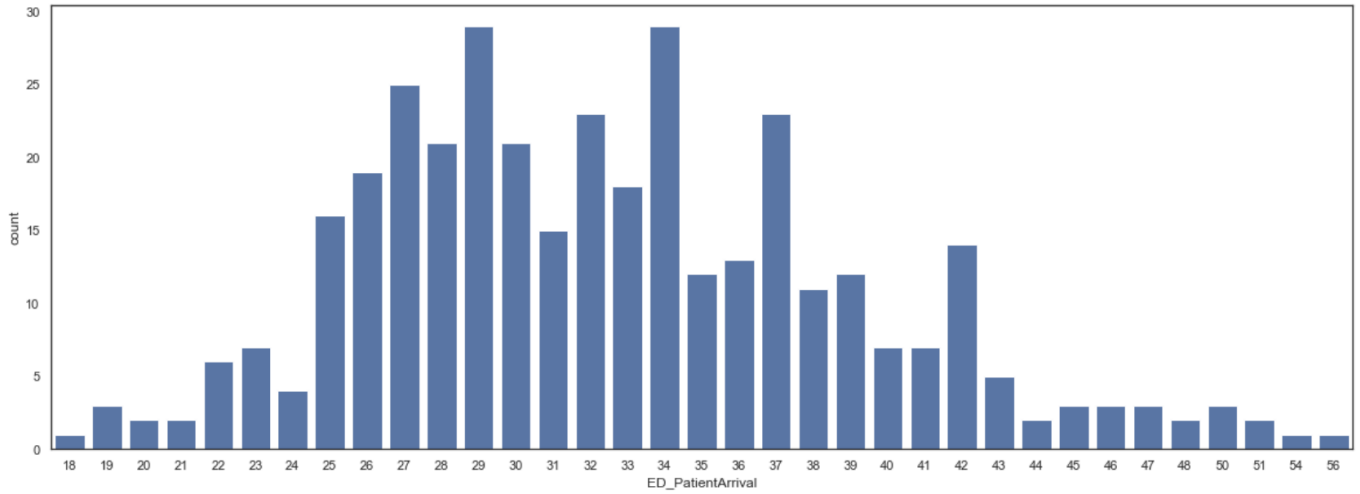


Figure 32: Histogram of ED patient demand in 2013

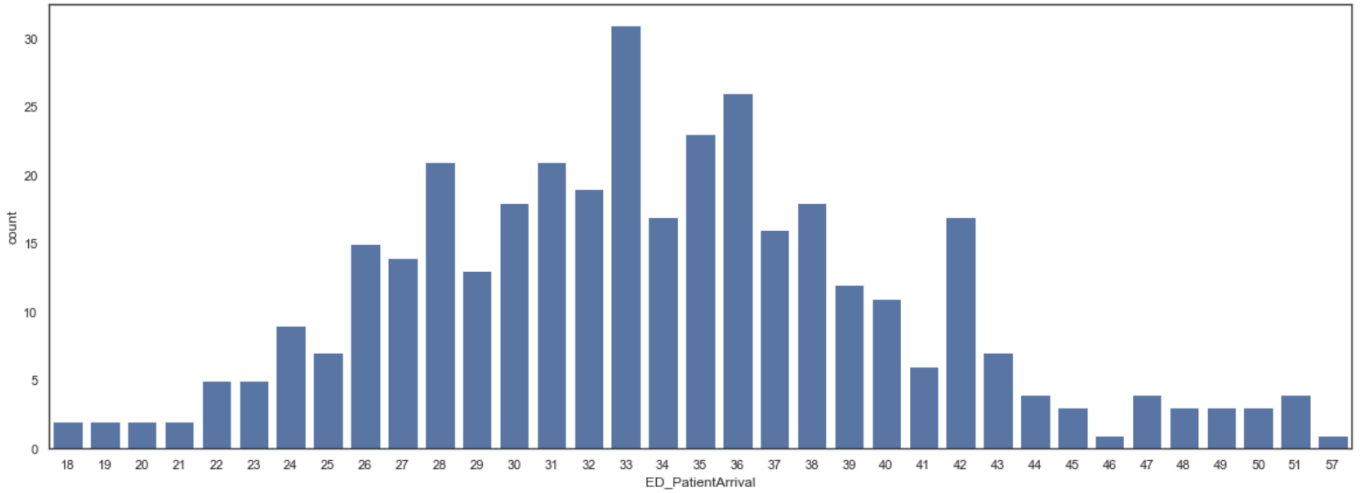


Figure 33: Histogram of ED patient demand in 2014

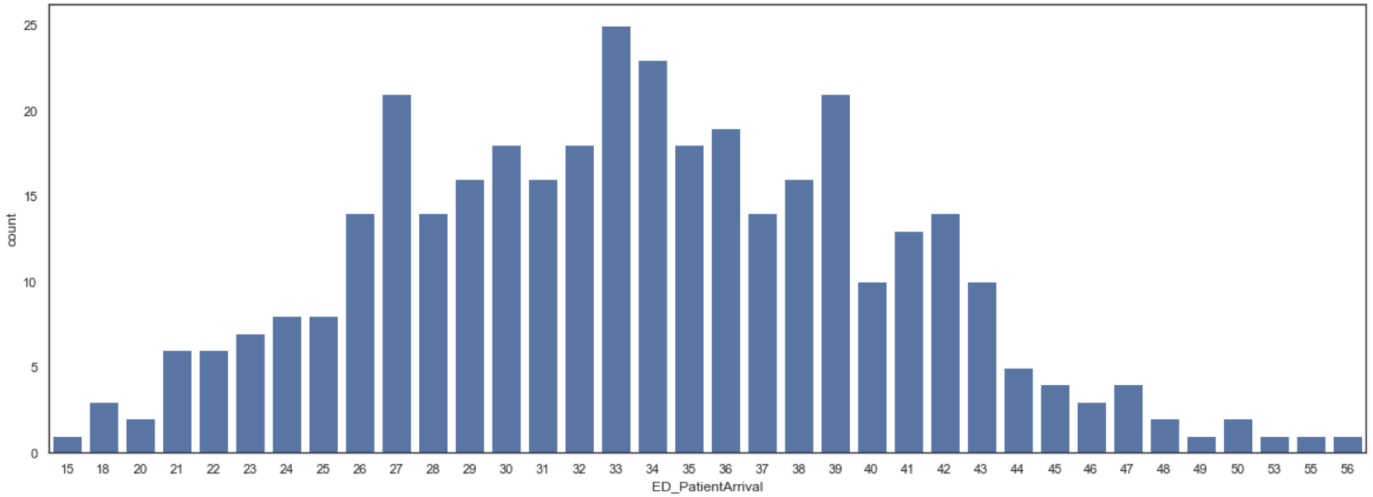


Figure 34: Histogram of ED patient demand in 2015

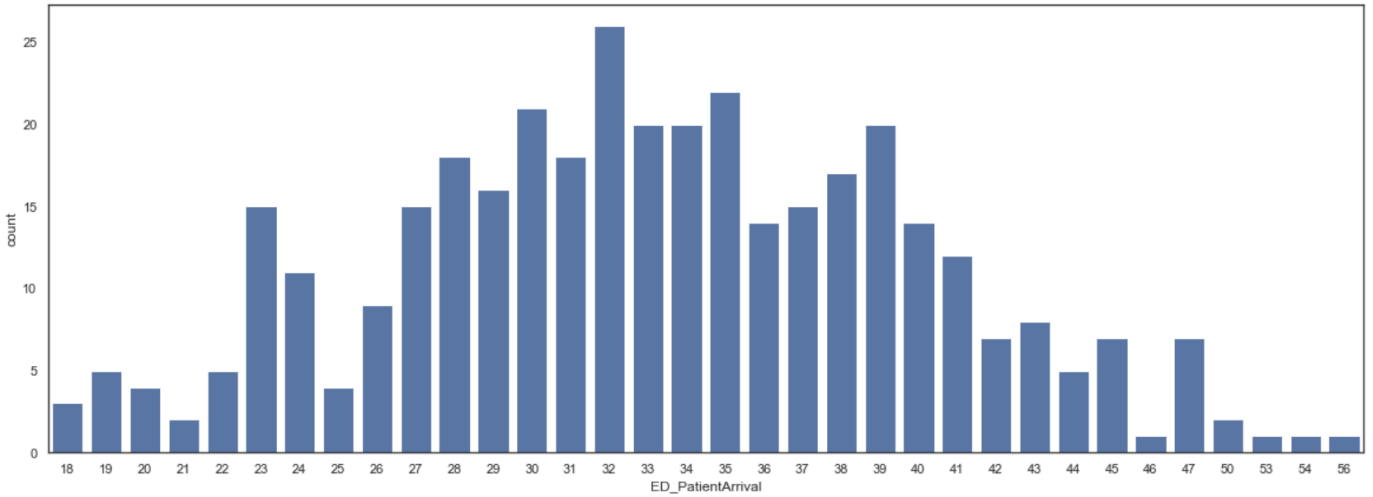


Figure 35: Histogram of ED patient demand in 2016

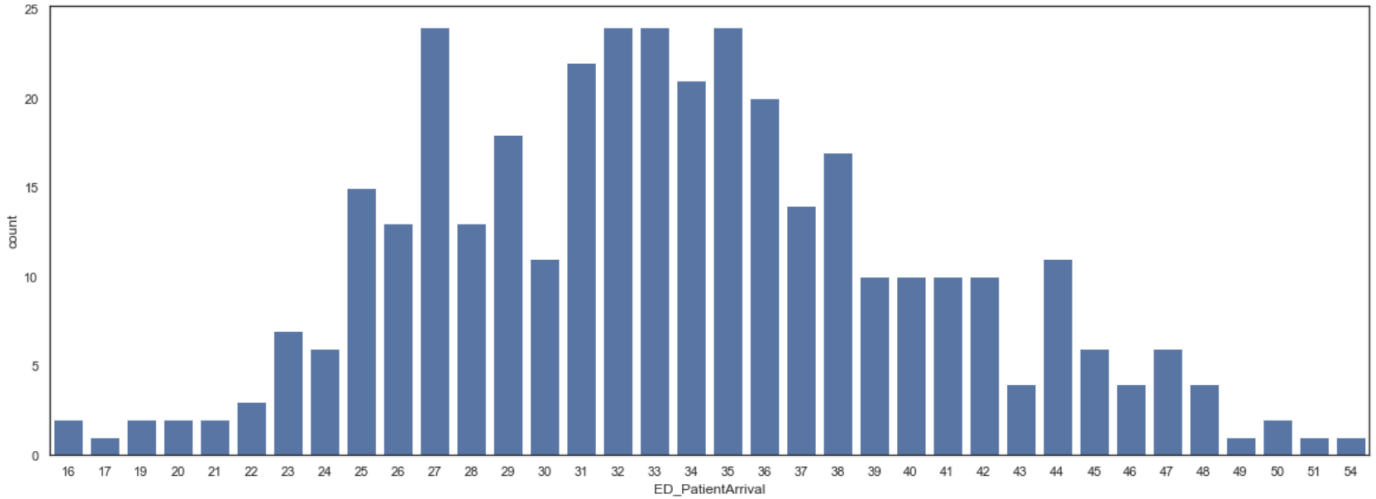


Figure 36: Histogram of ED patient demand in 2017

Appendix B – Linear Correlation Analysis

Table 32: Pearson Correlation Interpretation by Dancey & Reidy [17] and Chan et al. [18]

Correlation coefficient		Dancey & Reidy (Psychology)	Chan YH (Medicine) -This is used in this thesis-
+1	-1	Perfect	Perfect
+0.9	-0.9	Strong	Very Strong
+0.8	-0.8	Strong	Very Strong
+0.7	-0.7	Strong	Moderate
+0.6	-0.6	Moderate	Moderate
+0.5	-0.5	Moderate	Fair
+0.4	-0.4	Moderate	Fair
+0.3	-0.3	Weak	Fair
+0.2	-0.2	Weak	Poor
+0.1	-0.1	Weak	Poor
0	0	Zero	None

Features Descriptions

Features Name	Description
W_TX	Maximum Temperature
W_TG	Average Temperature
W_TN	Minimum Temperature
W_UX	Maximum Relative Humidity
W_UG	Average Relative Humidity
W_UN	Minimum Relative Humidity
ICPCcode_A	General and Unspecified
ICPCcode_B	Blood, Blood Forming, Organs and Immune Mechanism
ICPCcode_D	Digestive
ICPCcode_F	Eye
ICPCcode_H	Ear
ICPCcode_K	Cardiovascular
ICPCcode_L	Musculoskeletal
ICPCcode_N	Neurological
ICPCcode_P	Psychological
ICPCcode_R	Respiratory
ICPCcode_S	Skin
ICPCcode_T	Endocrine/Metabolic and Nutritional
ICPCcode_U	Urological
ICPCcode_W	Pregnancy, Childbearing, Family Planning
ICPCcode_X	Female Genital
ICPCcode_Y	Male Genital
ED_Treatment_CAR	Cardiology
ED_Treatment_CHI	Surgery
ED_Treatment_LON	Pulmonology
ED_Treatment_NEU	Neurology

Temperature & Humidity Correlation with Age groups in ED and GP-Post

All weekdays

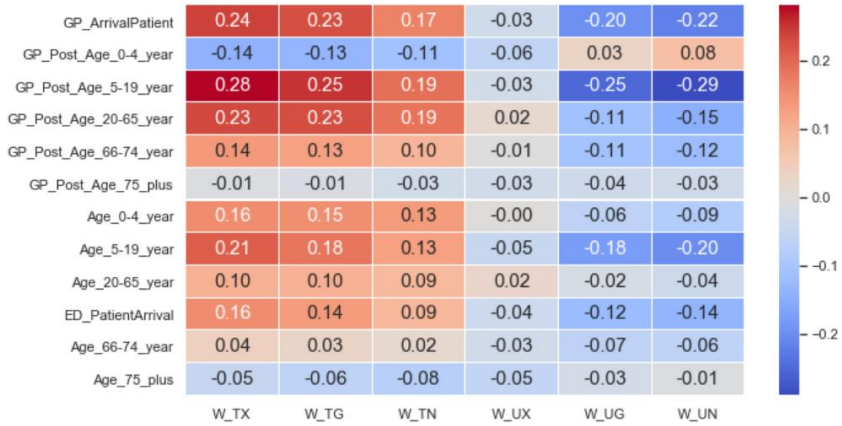


Figure 37: Temperature & Humidity Correlation with Age Groups for all weekdays

All weekends

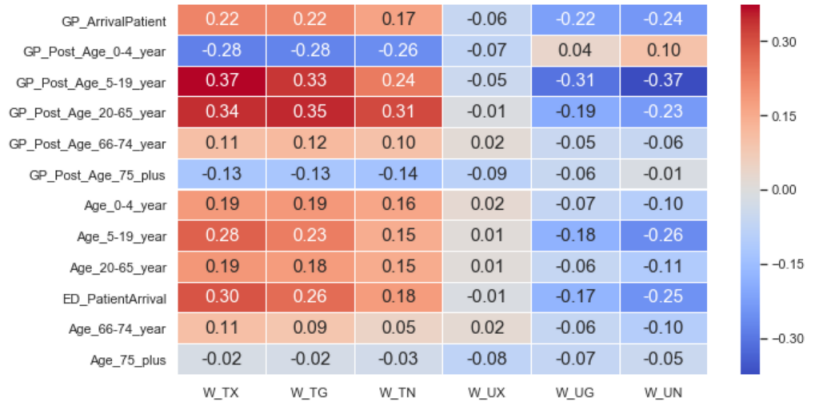


Figure 38: Temperature & Humidity Correlation with Age Groups for all weekends

Summer on Weekdays

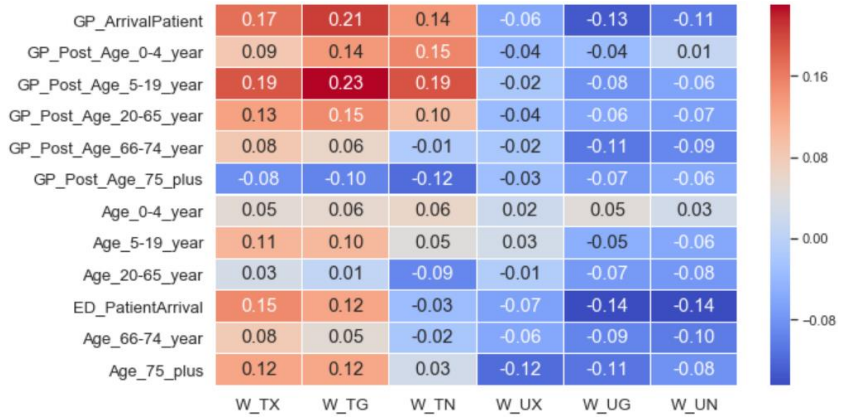


Figure 39: Temperature & Humidity Correlation with Age Groups for all weekdays in Summer

Summer on Weekends

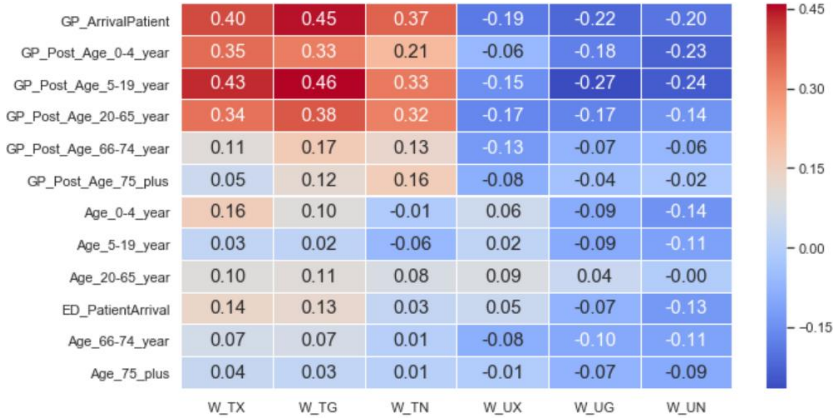


Figure 40: Temperature & Humidity Correlation with Age Groups for all weekends in Summer

Temperature & Humidity Correlation with Age groups in ED and GP-Post

Autumn on Weekdays

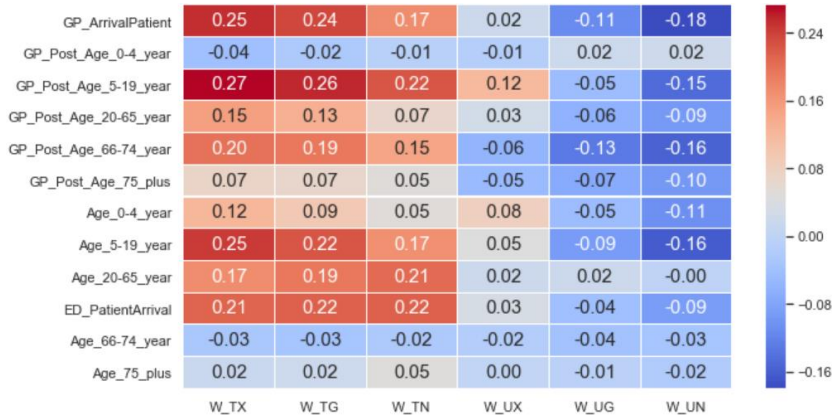


Figure 41: Temperature & Humidity Correlation with Age Groups for all weekdays in Autumn

Autumn on Weekends

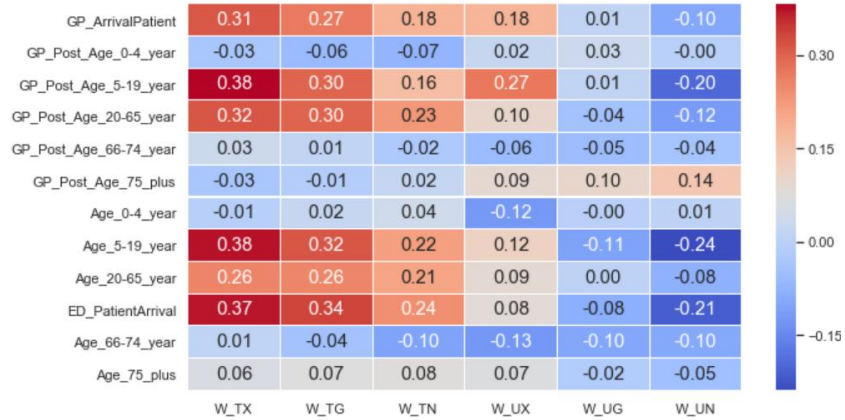


Figure 42: Temperature & Humidity Correlation with Age Groups for all weekends in Autumn

Winter on Weekdays

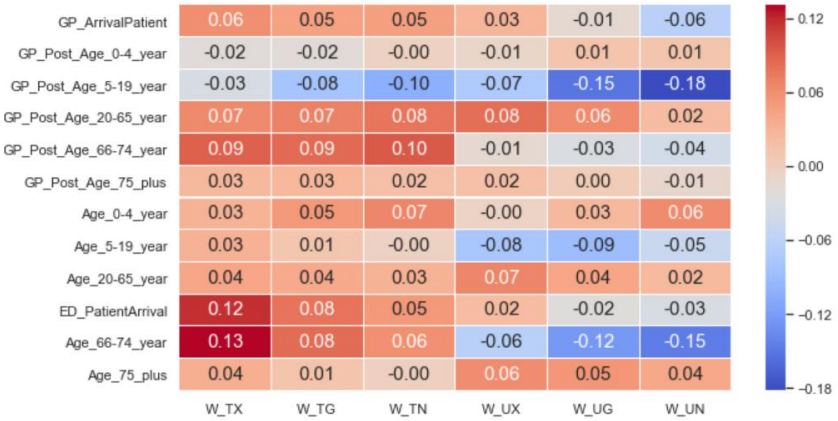


Figure 43: Temperature & Humidity Correlation with Age Groups for all weekdays in Winter

Winter on Weekends

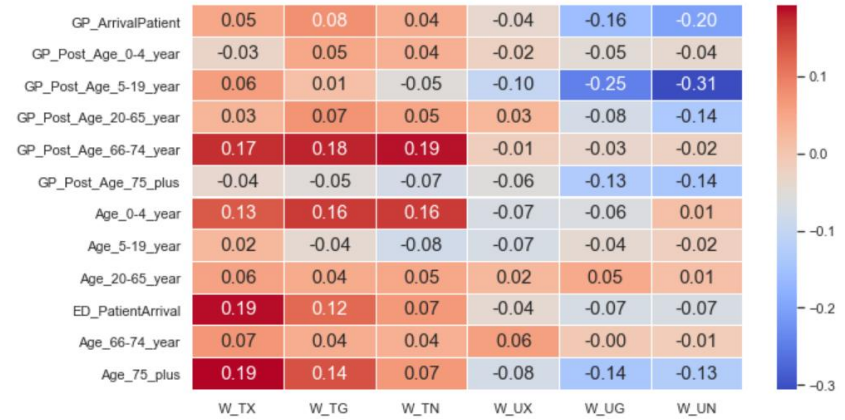


Figure 44: Temperature & Humidity Correlation with Age Groups for all weekends in Winter

Temperature & Humidity Correlation with Age groups in ED and GP-Post

Spring on Weekdays

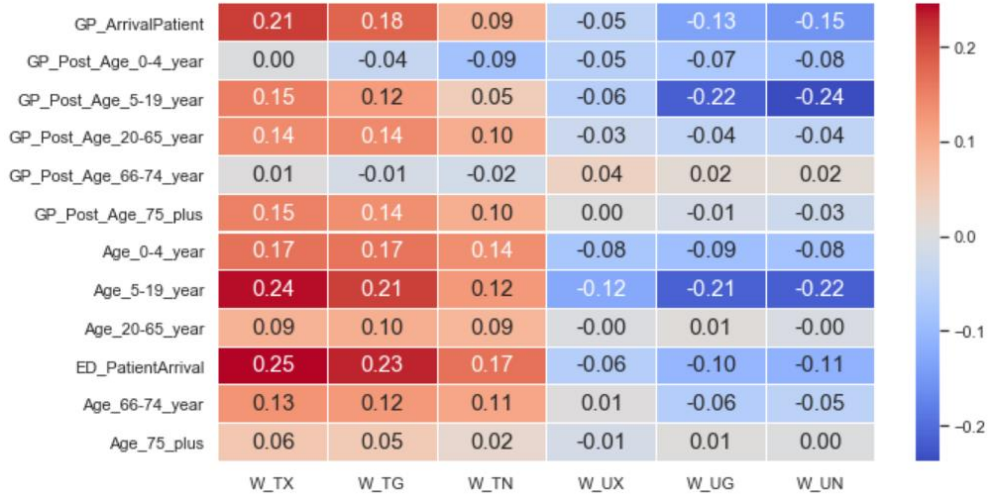


Figure 45: Temperature & Humidity Correlation with Age Groups for all weekdays in Spring

Spring on Weekends

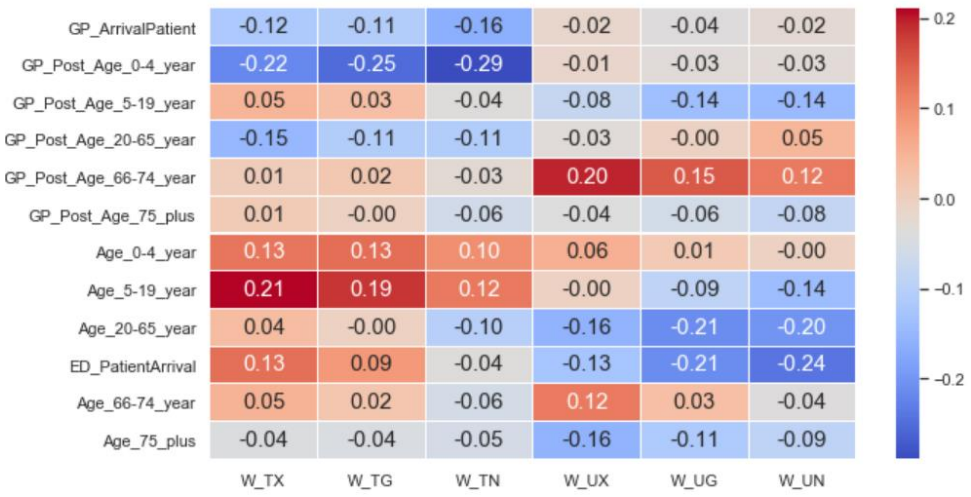


Figure 46: Temperature & Humidity Correlation with Age Groups for all weekends in Spring

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

All weekdays

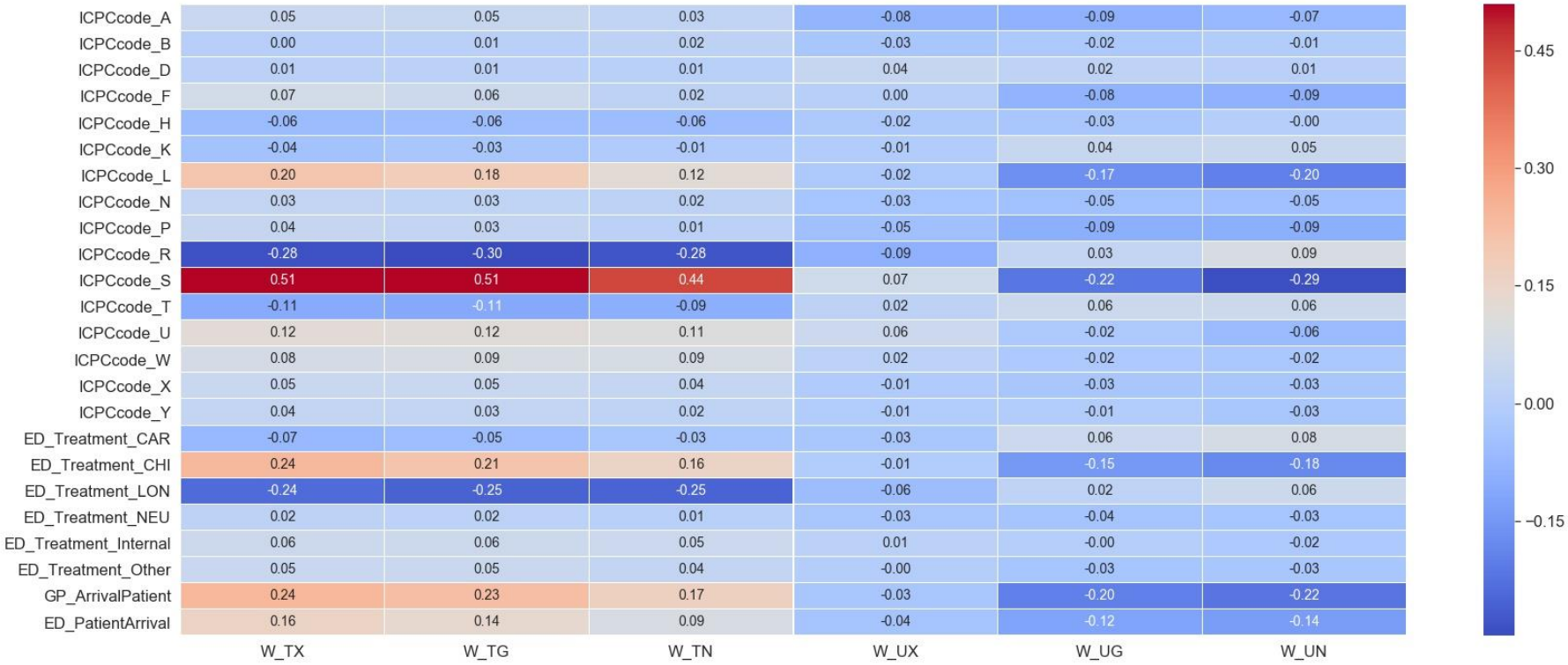


Figure 47: Temperature & Humidity Correlation with Treatment Groups for all weekdays

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

All weekends

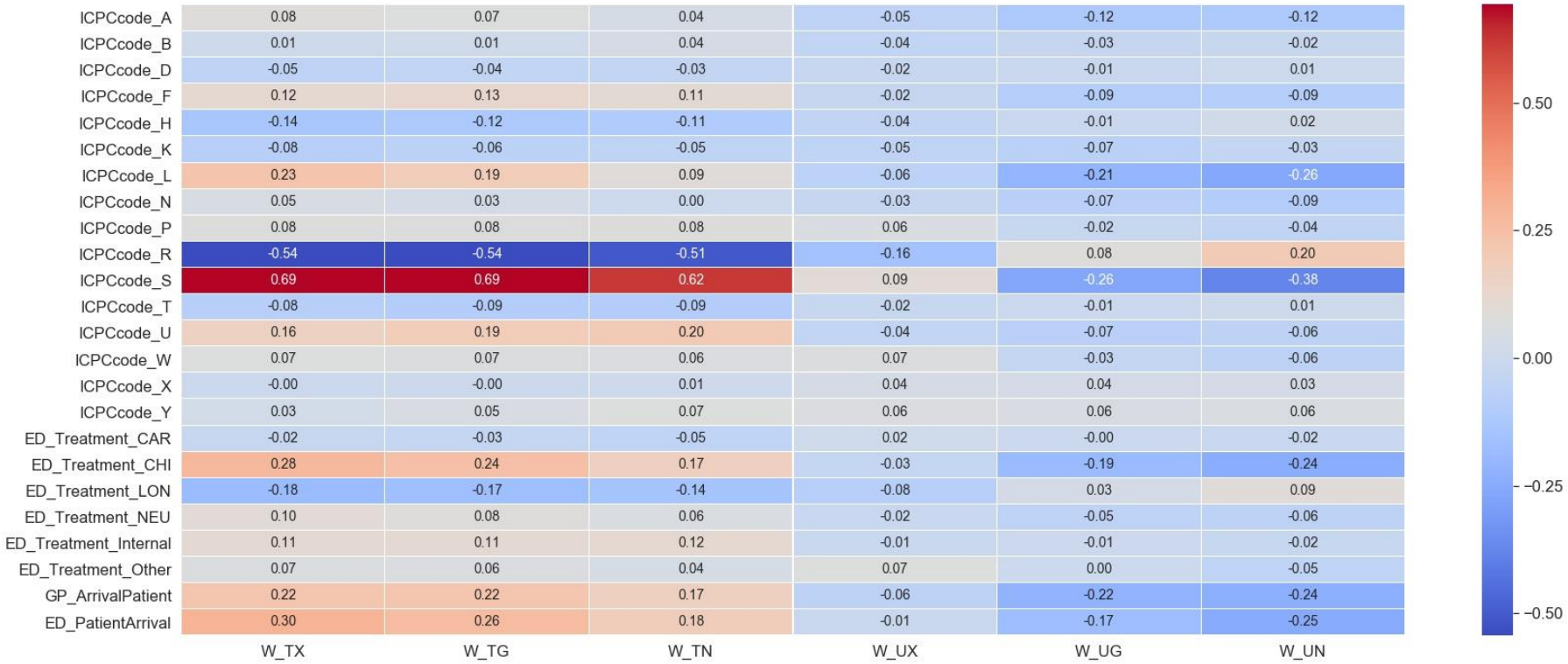


Figure 48: Temperature & Humidity Correlation with Treatment Groups for all weekends

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Summer-Weekdays

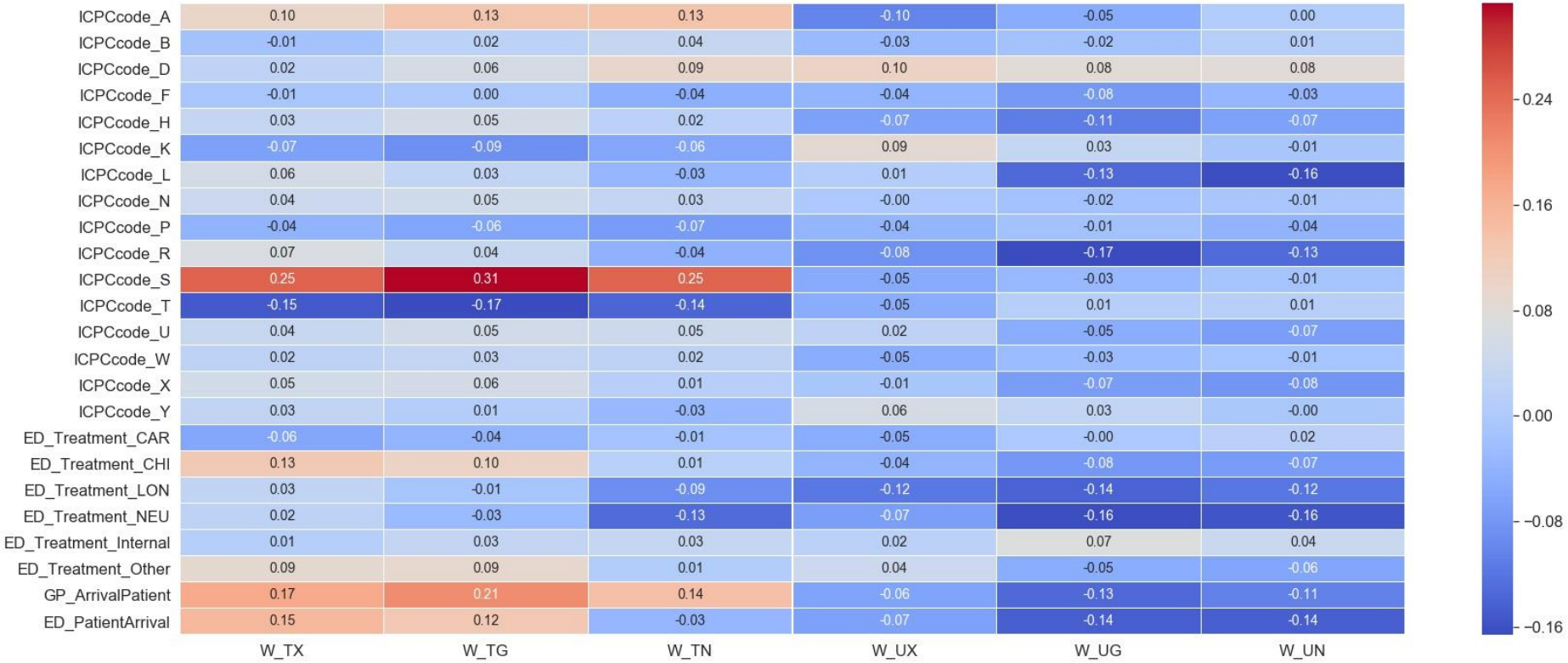


Figure 49: Temperature & Humidity Correlation with Treatment Groups for all weekdays in Summer

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Summer-Weekends

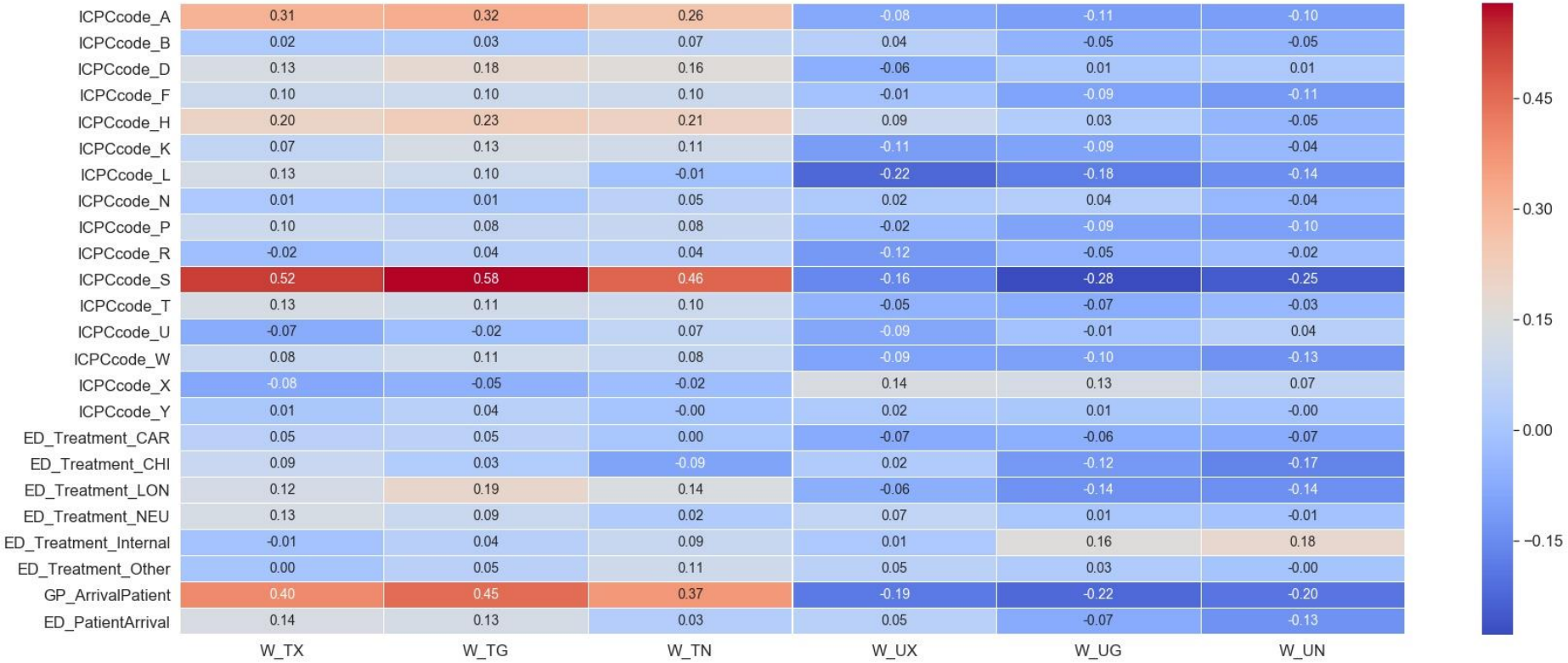


Figure 50 : Temperature & Humidity Correlation with Treatment Groups for all weekends in Summer

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Autumn-Weekdays

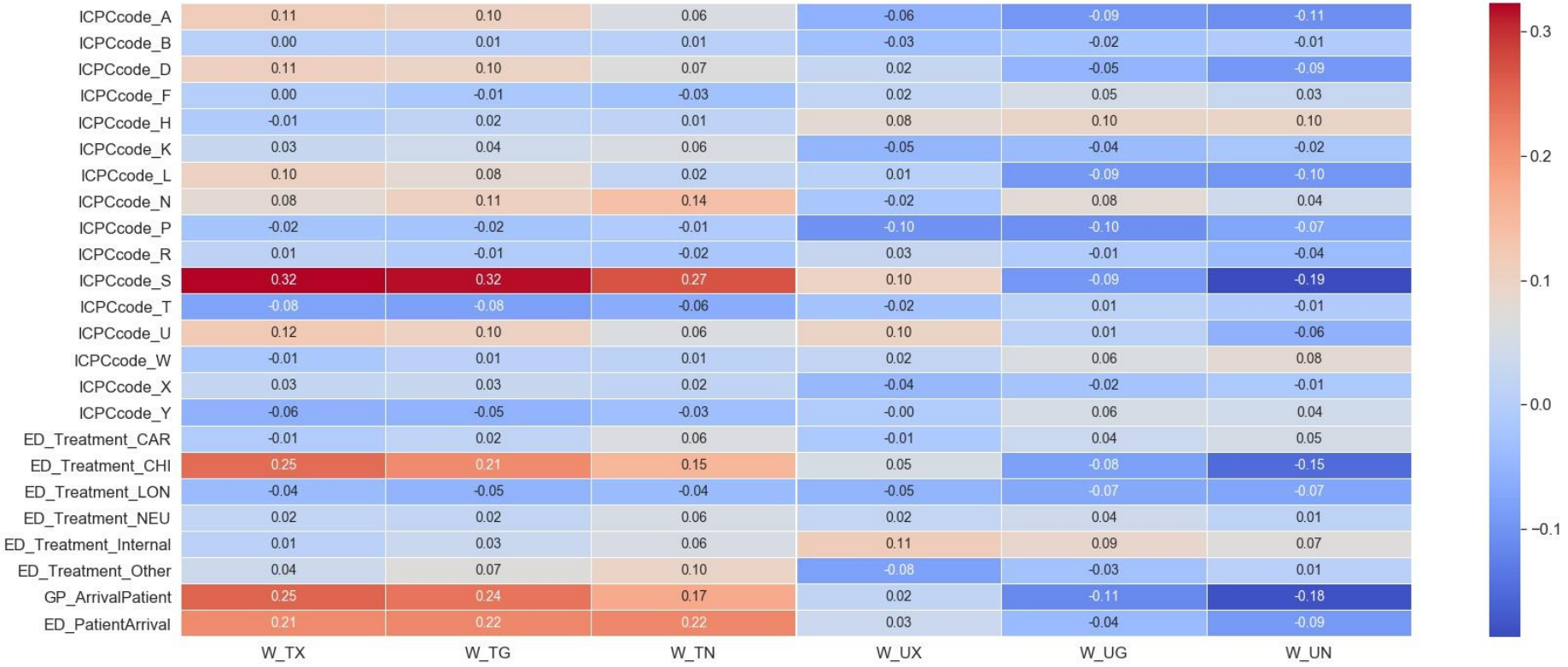


Figure 51 : Temperature & Humidity Correlation with Treatment Groups for all weekdays in Autumn

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Autumn-Weekends

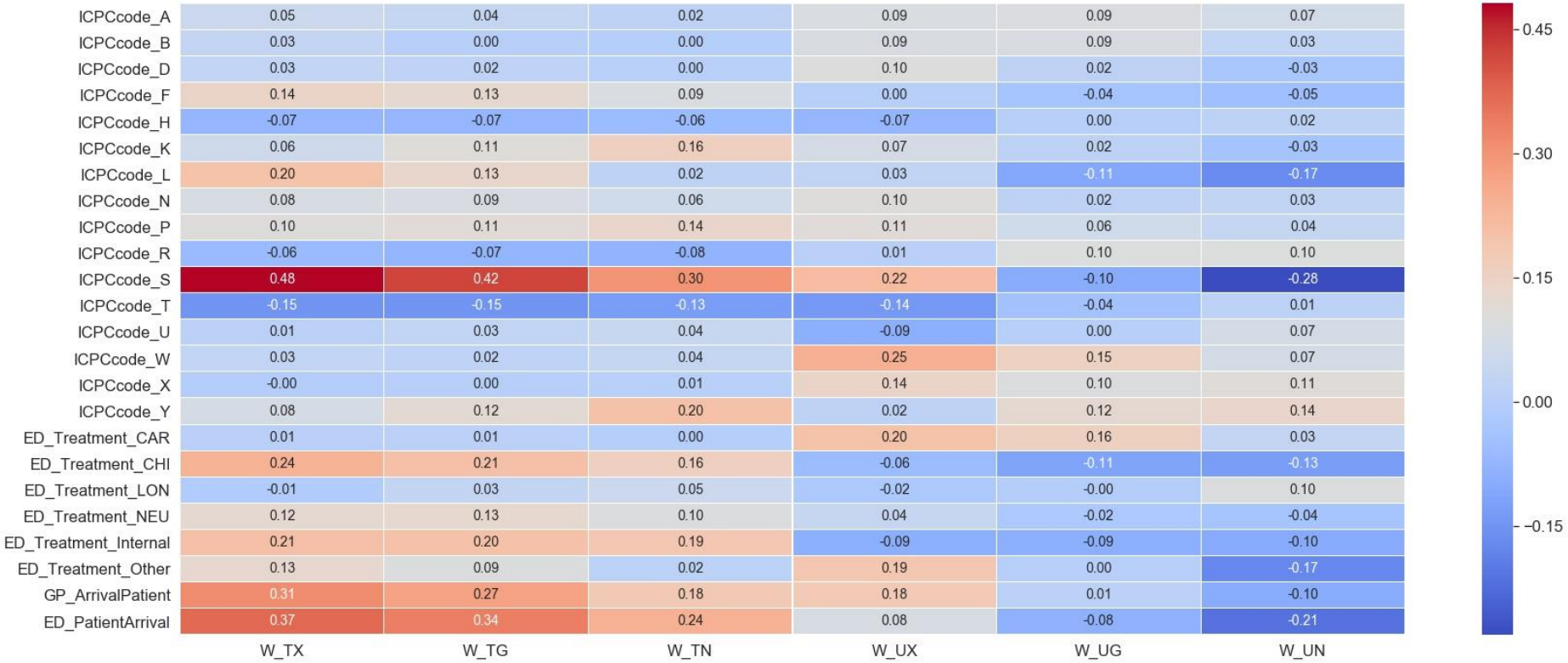


Figure 52: Temperature & Humidity Correlation with Treatment Groups for all weekends in Autumn

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Winter-Weekdays

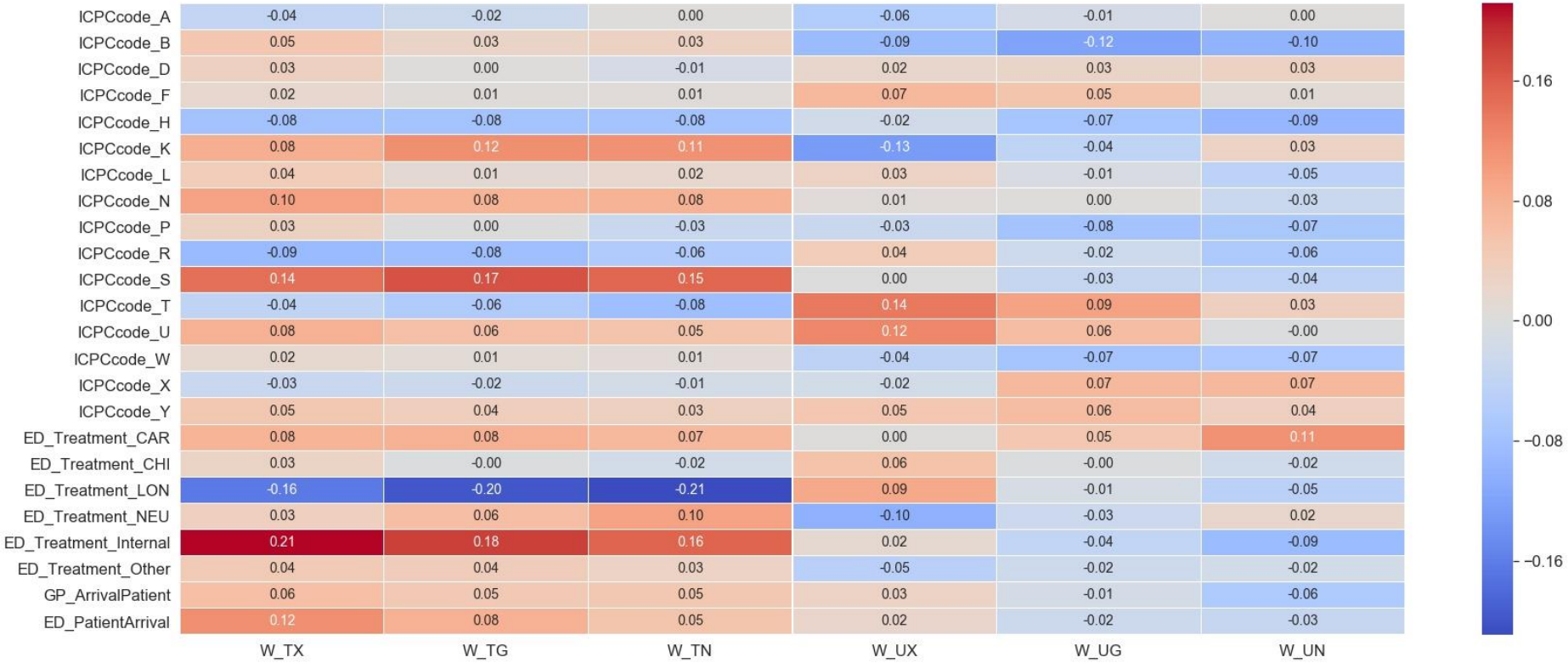


Figure 53: Temperature & Humidity Correlation with Treatment Groups for all weekdays in Winter

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Winter-Weekends

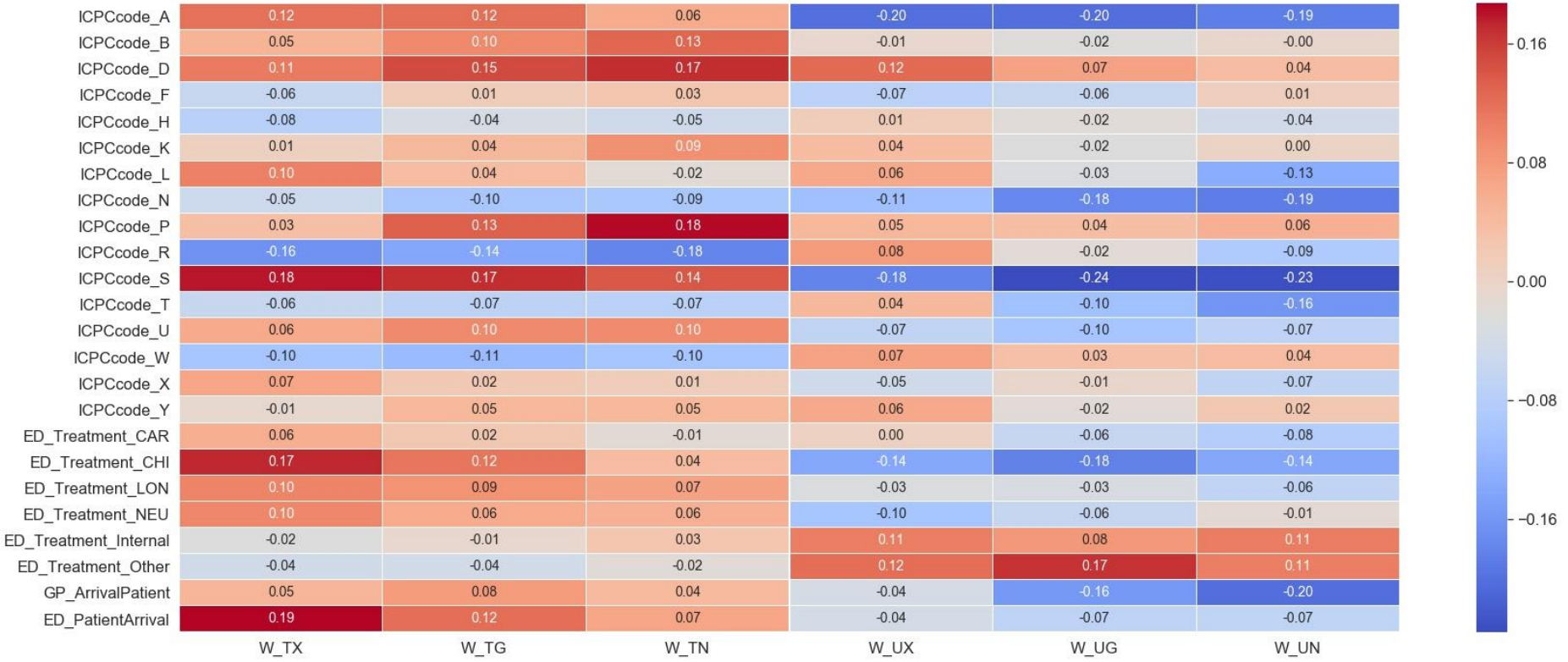


Figure 54: Temperature & Humidity Correlation with Treatment Groups for all weekends in Winter

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Spring-Weekdays

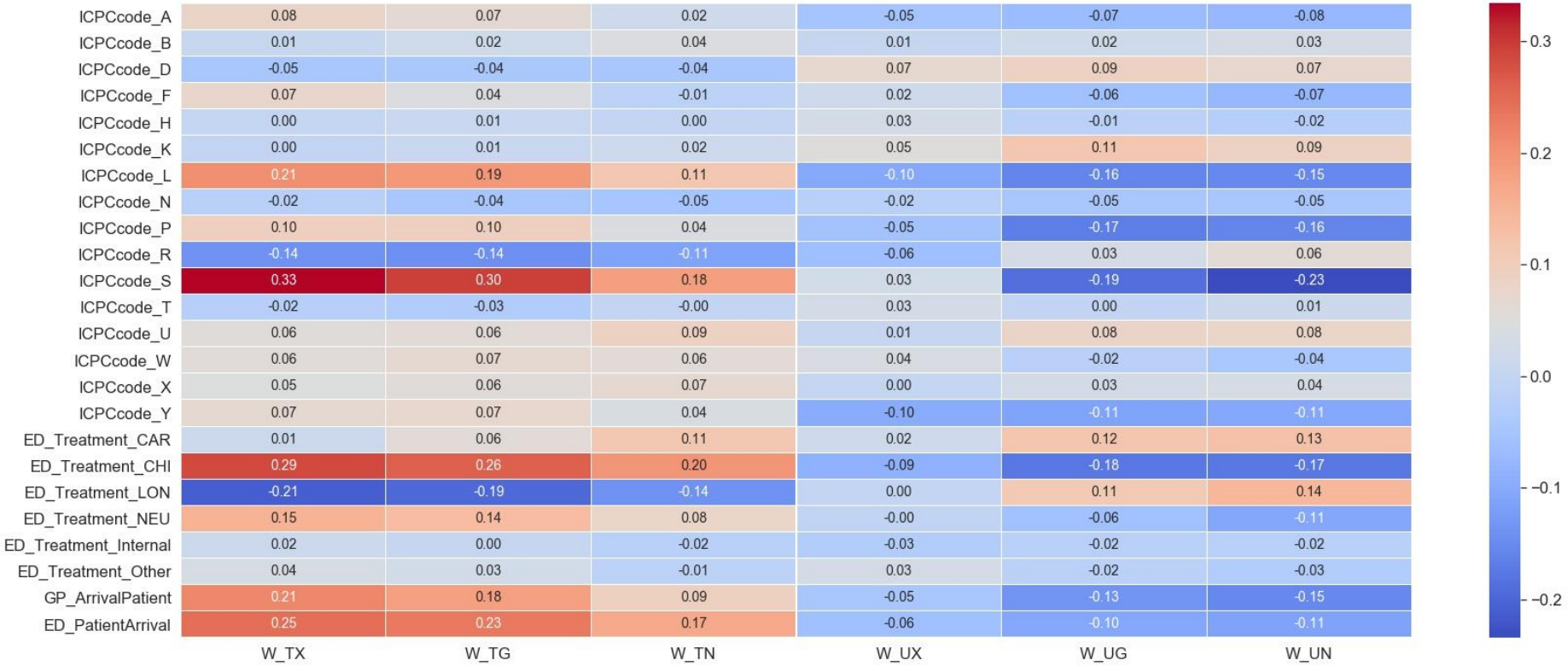


Figure 55: Temperature & Humidity Correlation with Treatment Groups for all weekdays in Spring

Temperature & Humidity Correlation with Treatment groups in ED and GP-Post

Spring-Weekends

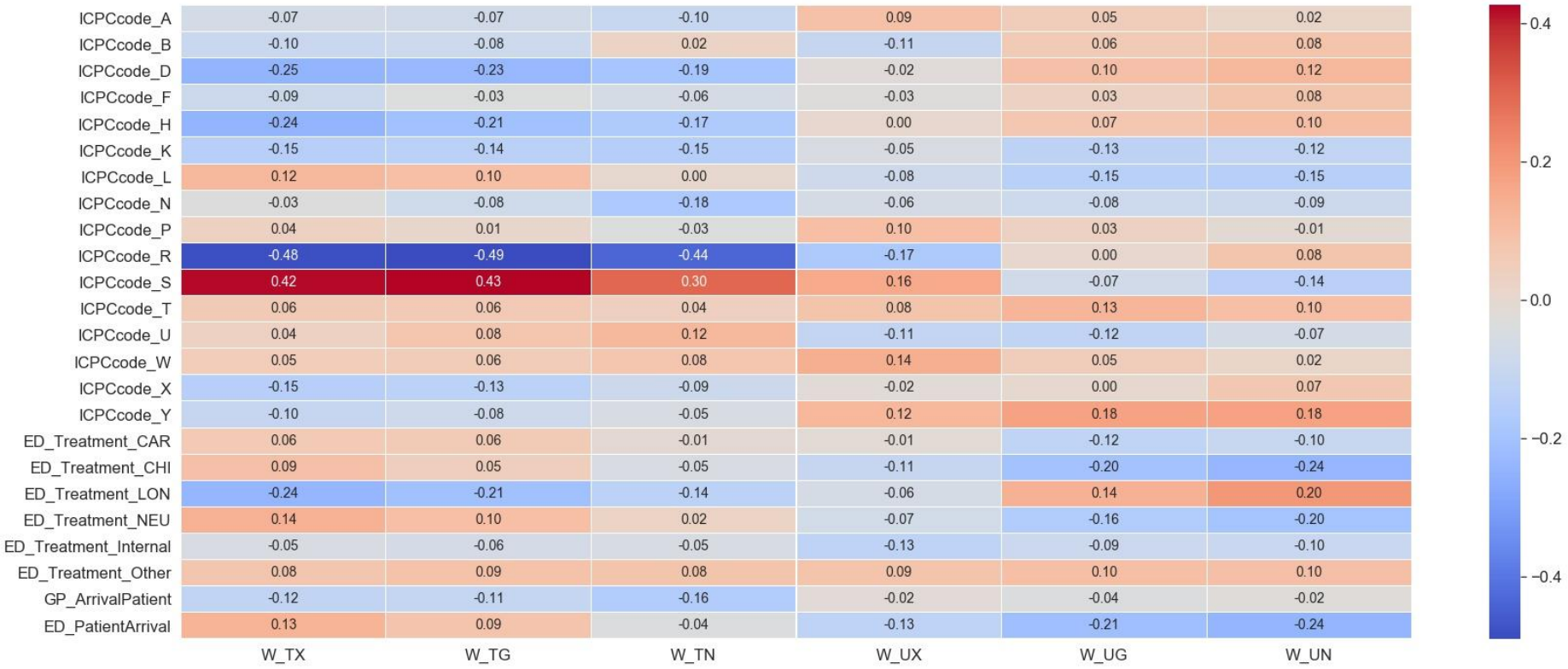


Figure 56: Temperature & Humidity Correlation with Treatment Groups for all weekends in Spring

Pollen Correlation with Treatment groups in ED and GP-Post

All weekdays

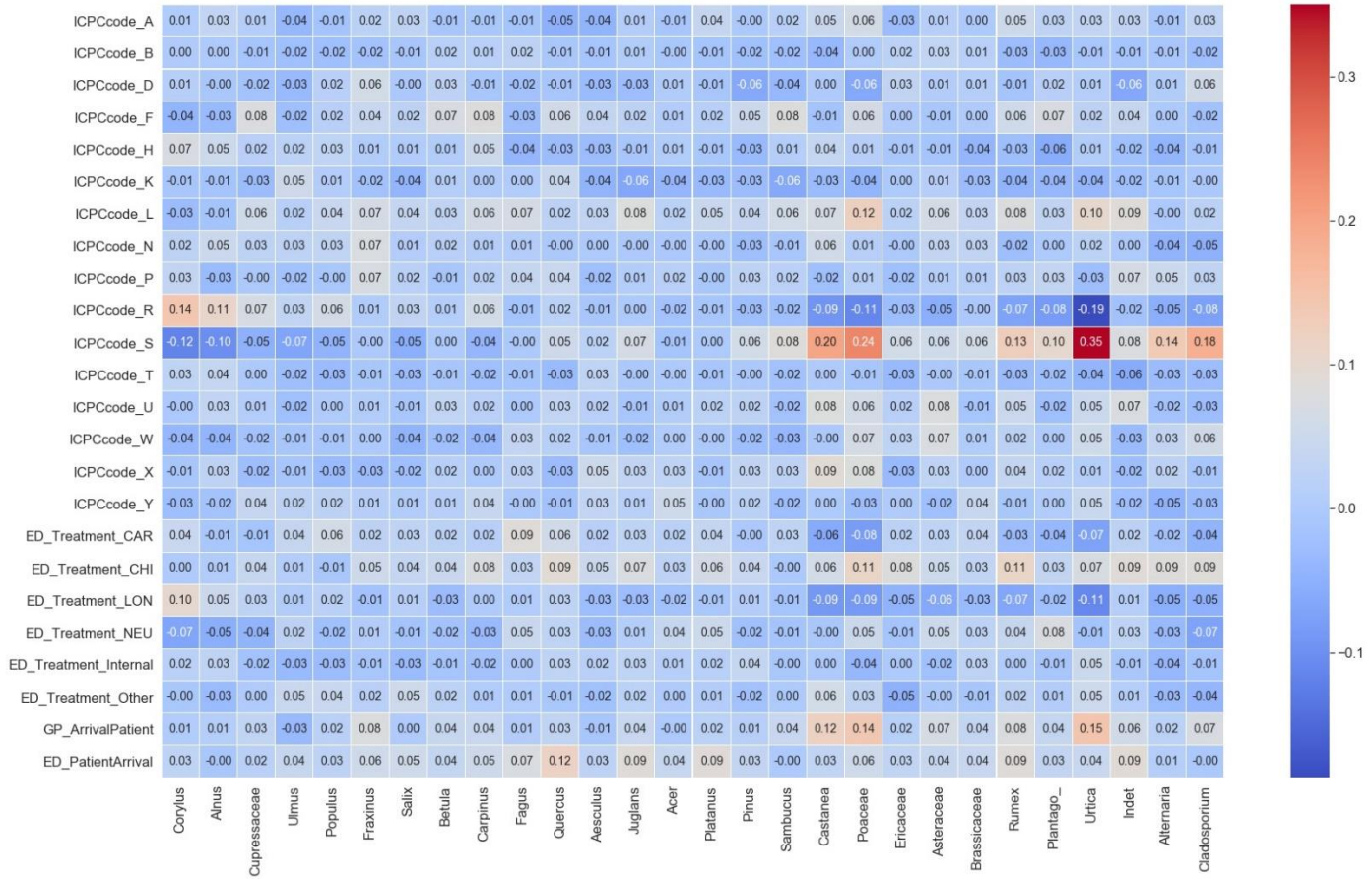


Figure 57: Pollen Correlation with Treatment Groups for all weekdays

Pollen Correlation with Treatment groups in ED and GP-Post

All weekends

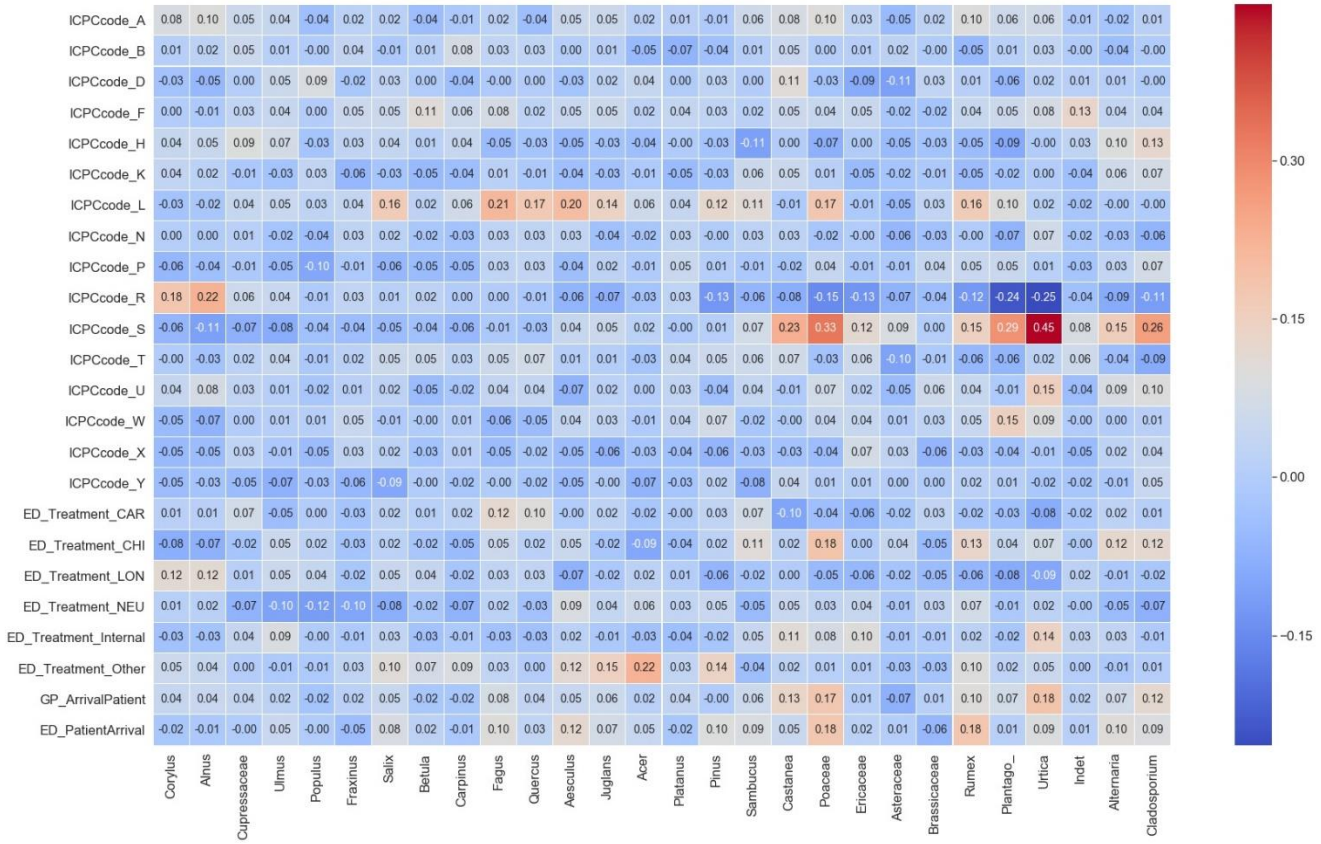


Figure 58: Pollen Correlation with Treatment Groups for all weekends

Pollen Correlation with Treatment groups in ED and GP-Post

Summer-Weekdays

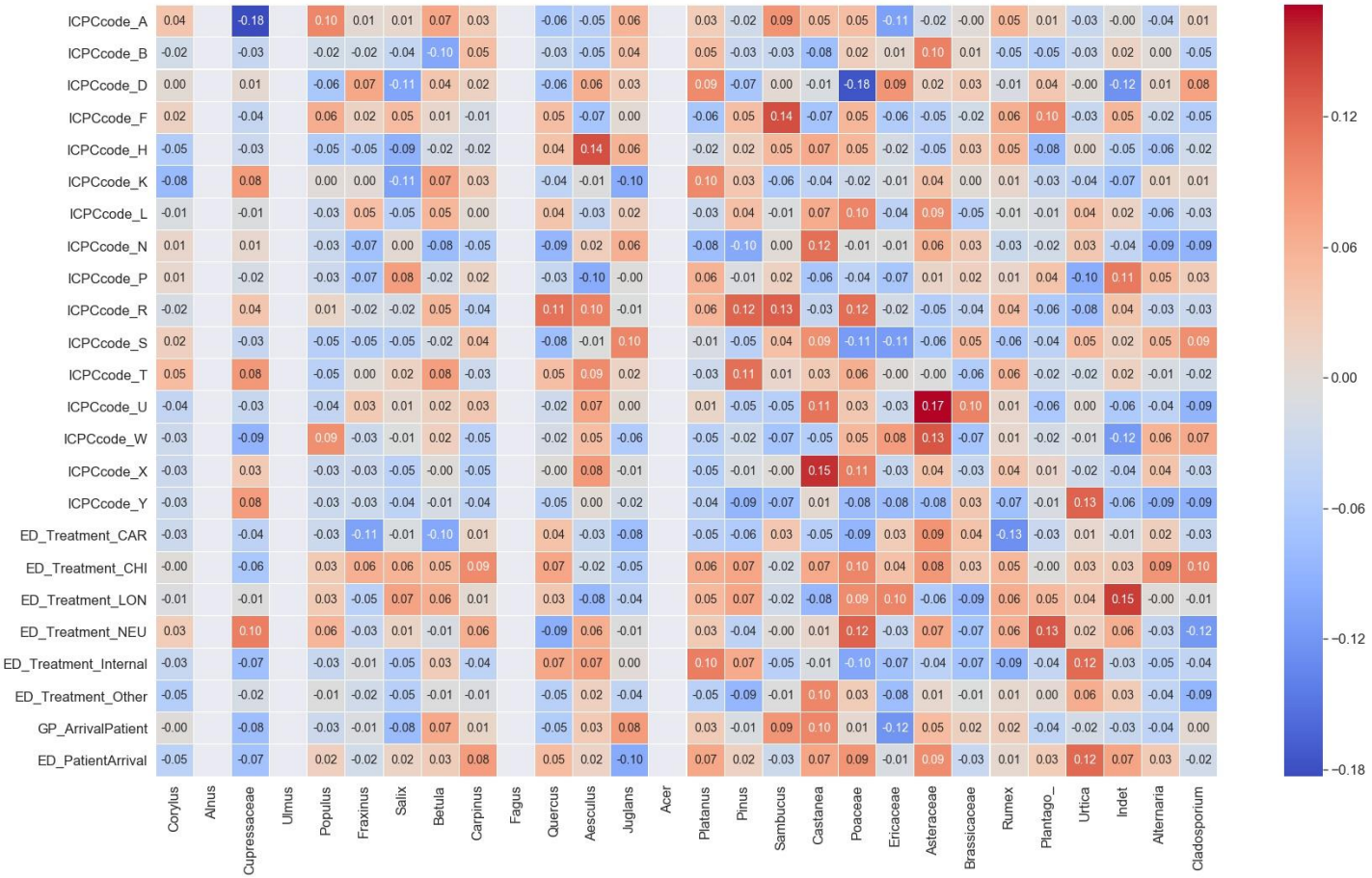


Figure 59: Pollen Correlation with Treatment Groups for all weekdays in Summer

Pollen Correlation with Treatment groups in ED and GP-Post

Summer-Weekends

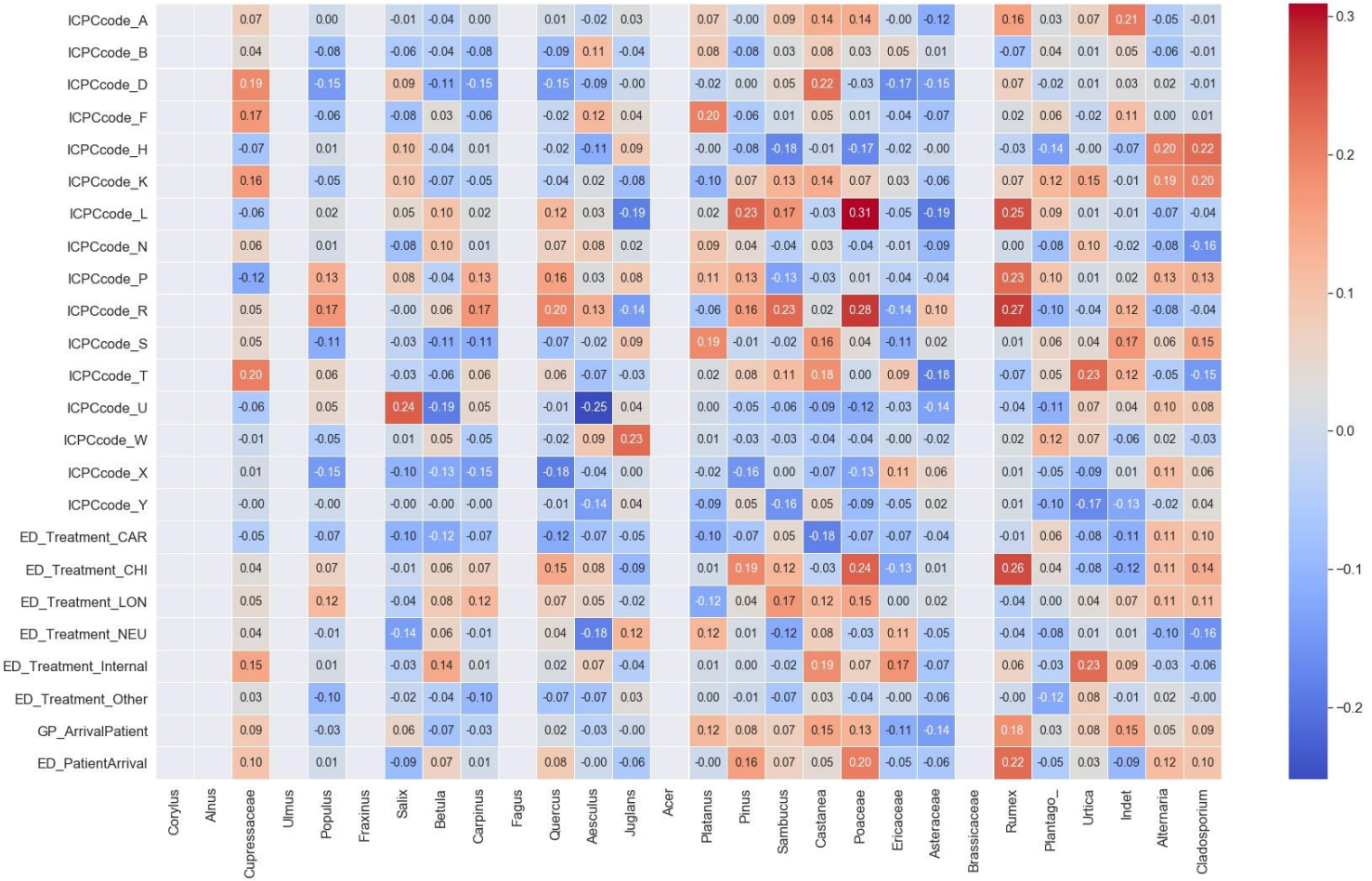


Figure 60 : Pollen Correlation with Treatment Groups for all weekends in Summer

Pollen Correlation with Treatment groups in ED and GP-Post

Autumn-Weekdays

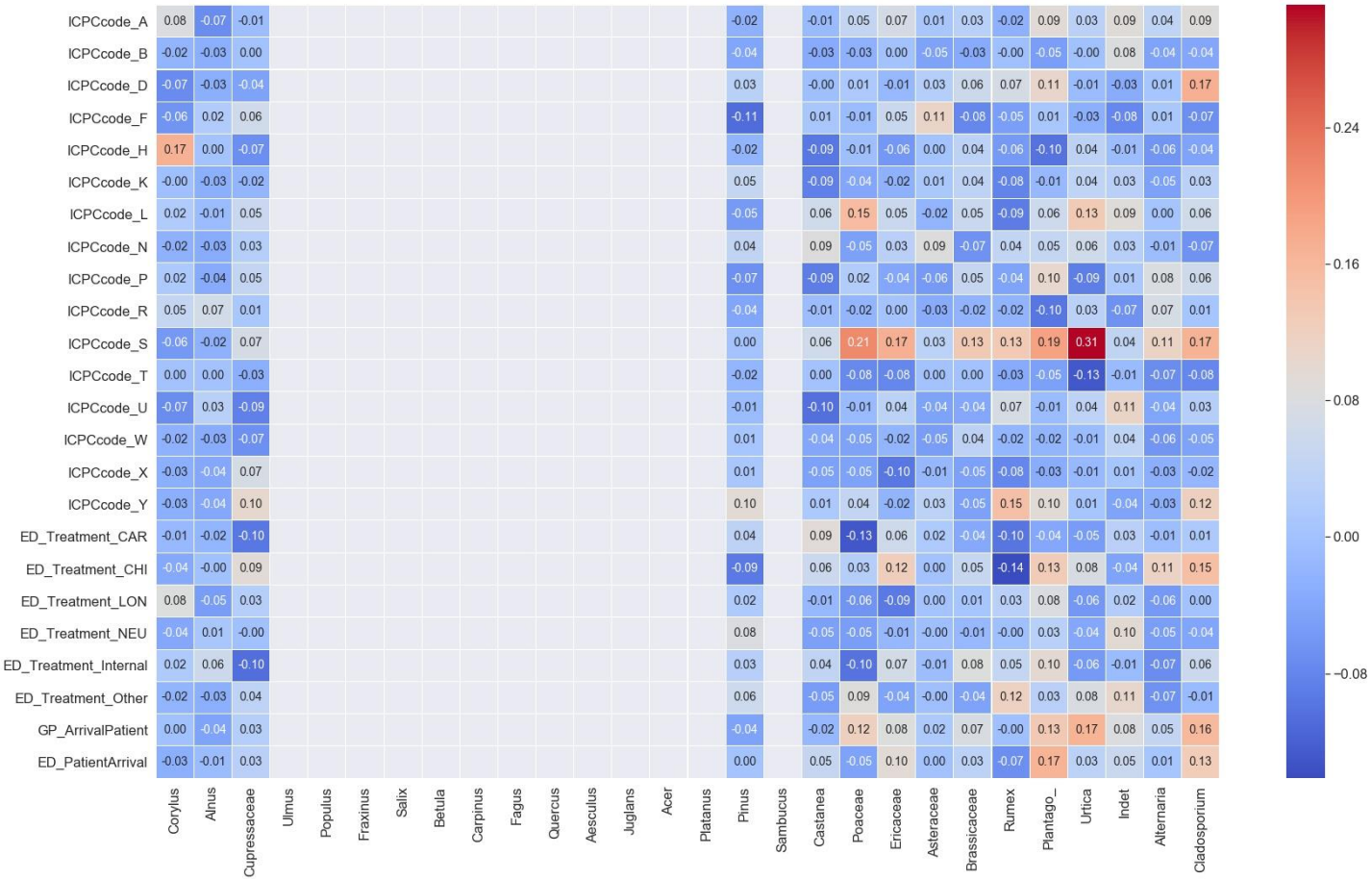


Figure 61 : Pollen Correlation with Treatment Groups for all weekdays in Autumn

Pollen Correlation with Treatment groups in ED and GP-Post

Autumn-Weekends

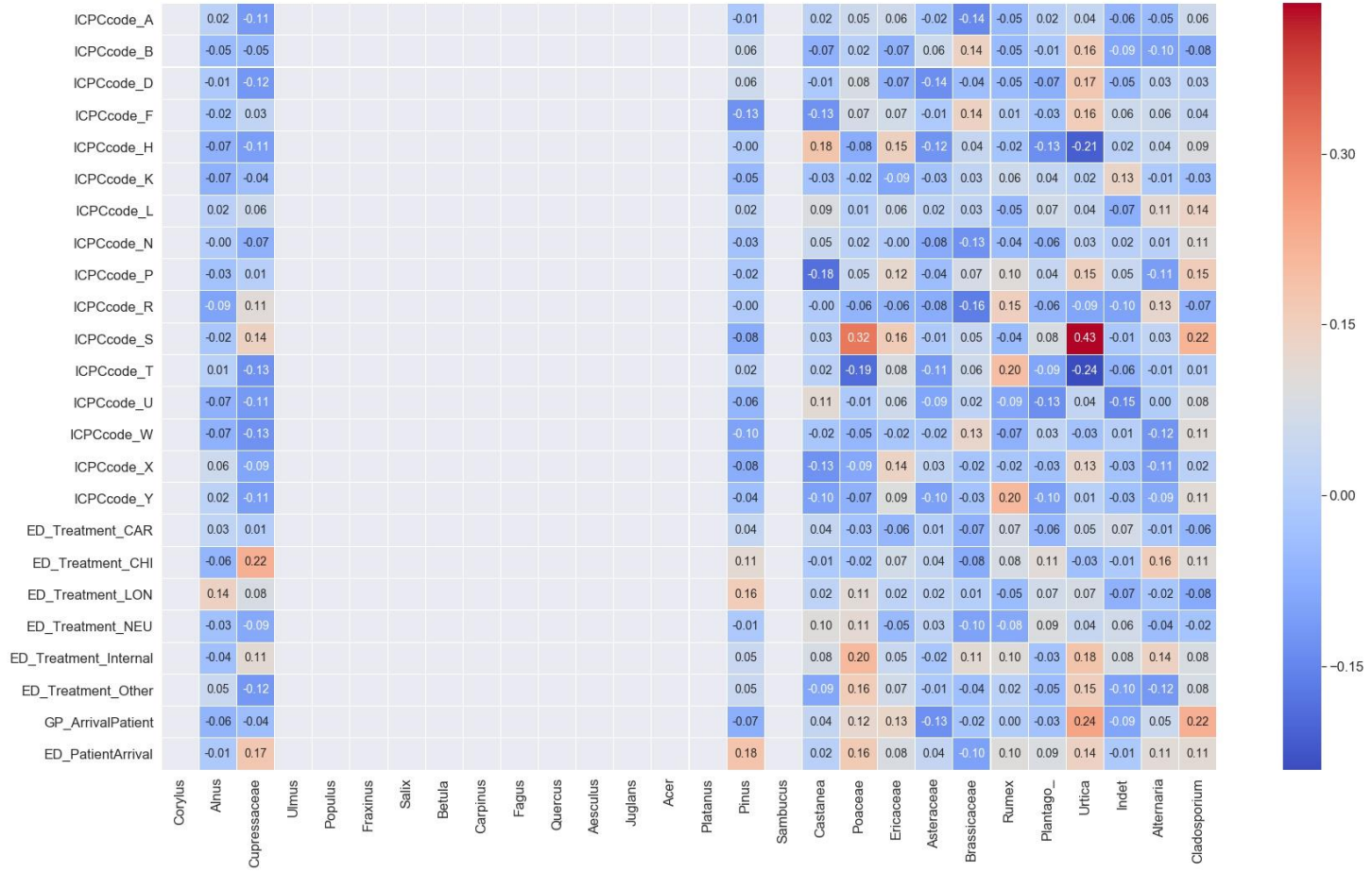


Figure 62: Pollen Correlation with Treatment Groups for all weekends in Autumn

Pollen Correlation with Treatment groups in ED and GP-Post

Winter-Weekdays

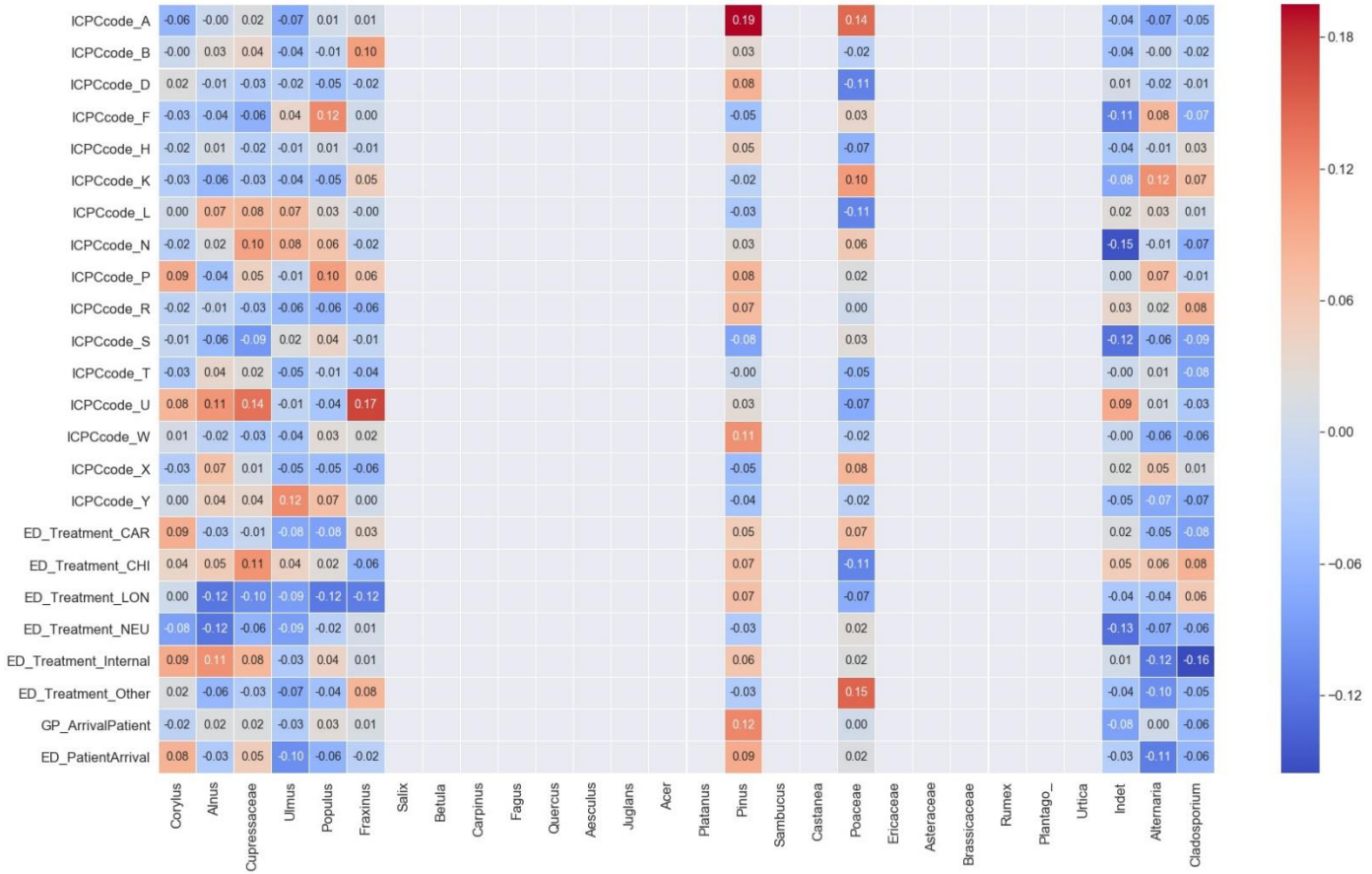


Figure 63: Pollen Correlation with Treatment Groups for all weekdays in Winter

Pollen Correlation with Treatment groups in ED and GP-Post

Winter-Weekends

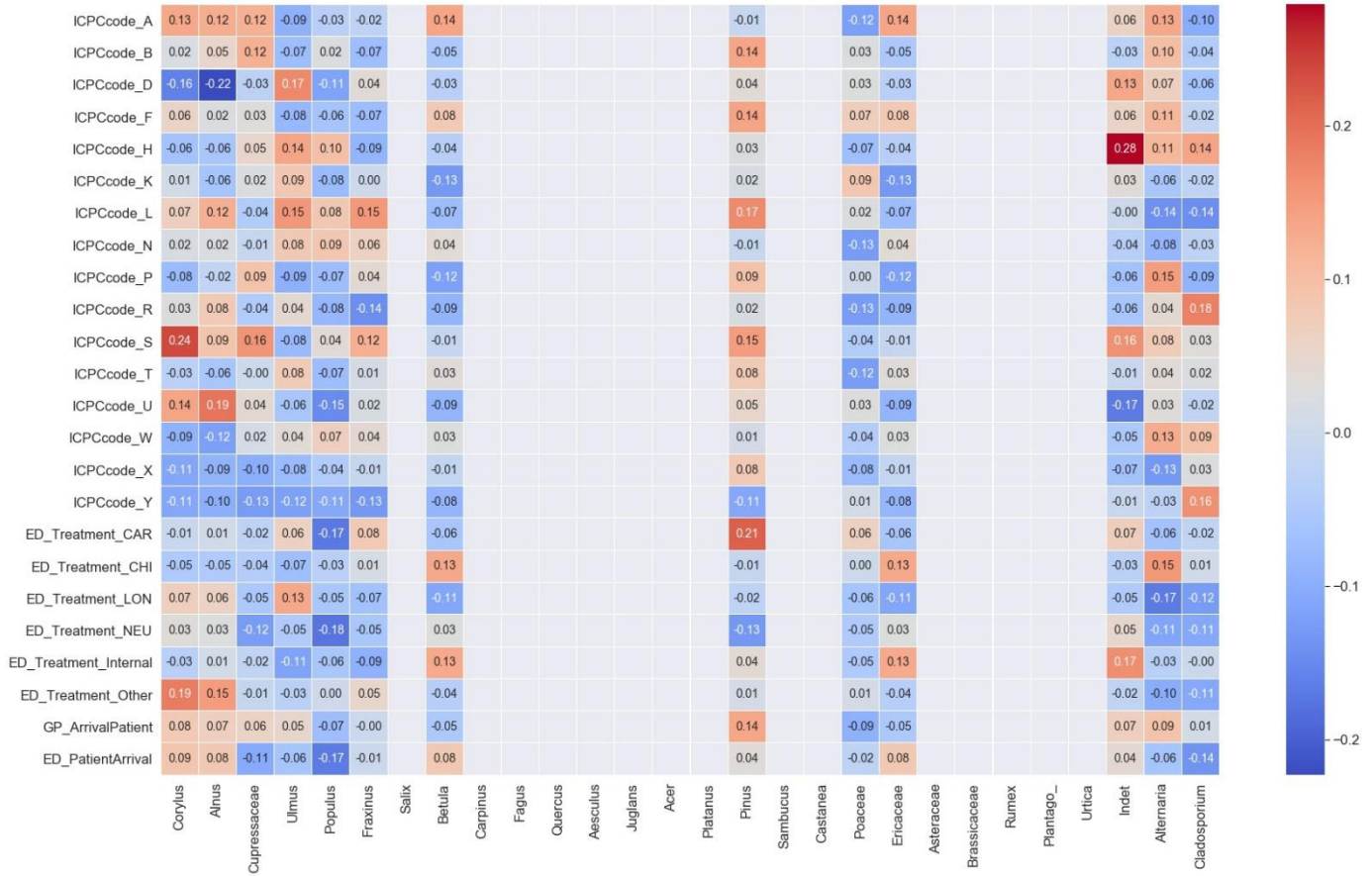


Figure 64: Pollen Correlation with Treatment Groups for all weekends in Winter

Pollen Correlation with Treatment groups in ED and GP-Post

Spring-Weekdays

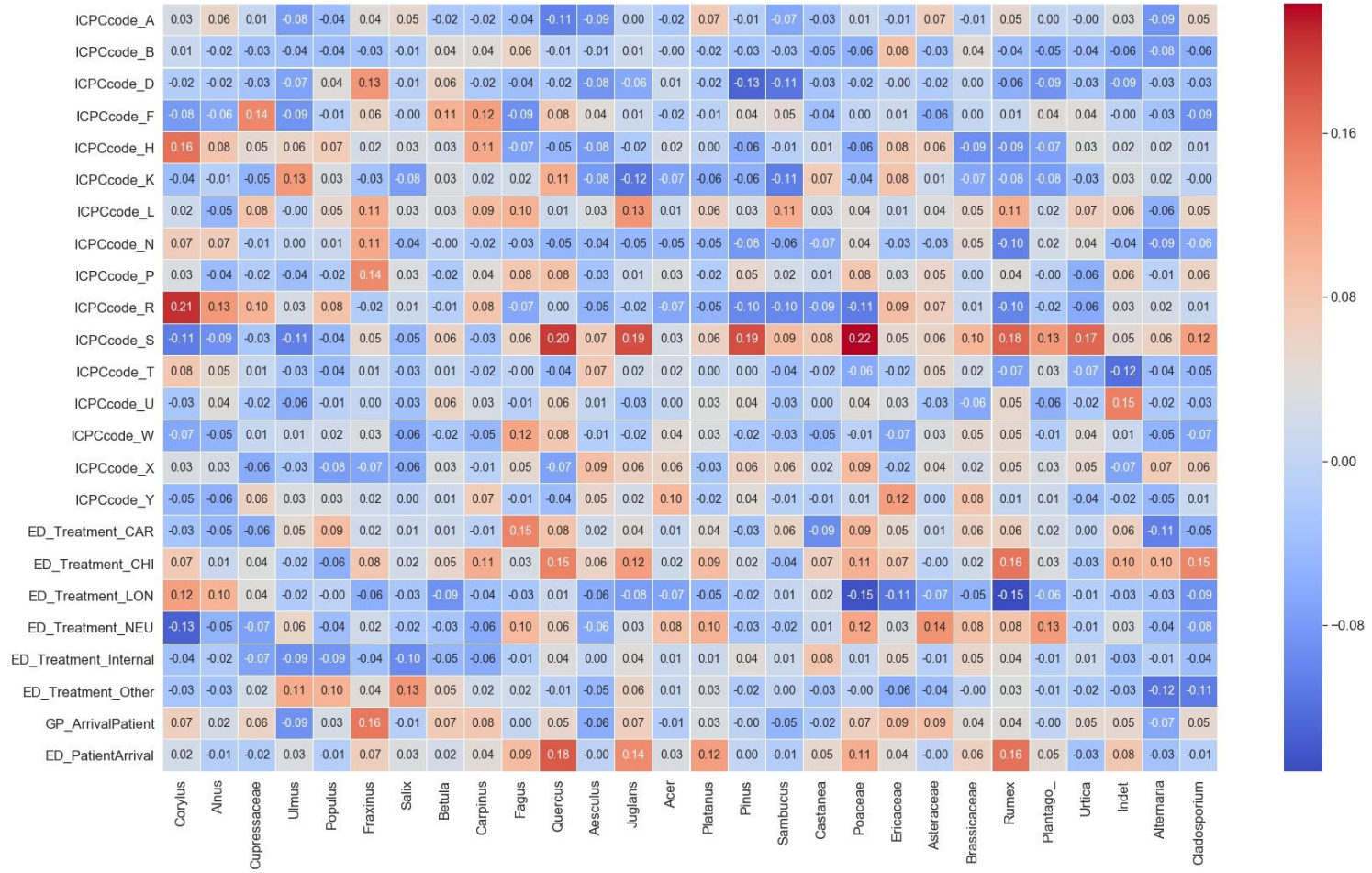


Figure 65: Pollen Correlation with Treatment Groups for all weekdays in Spring

Pollen Correlation with Treatment groups in ED and GP-Post

Spring-Weekends

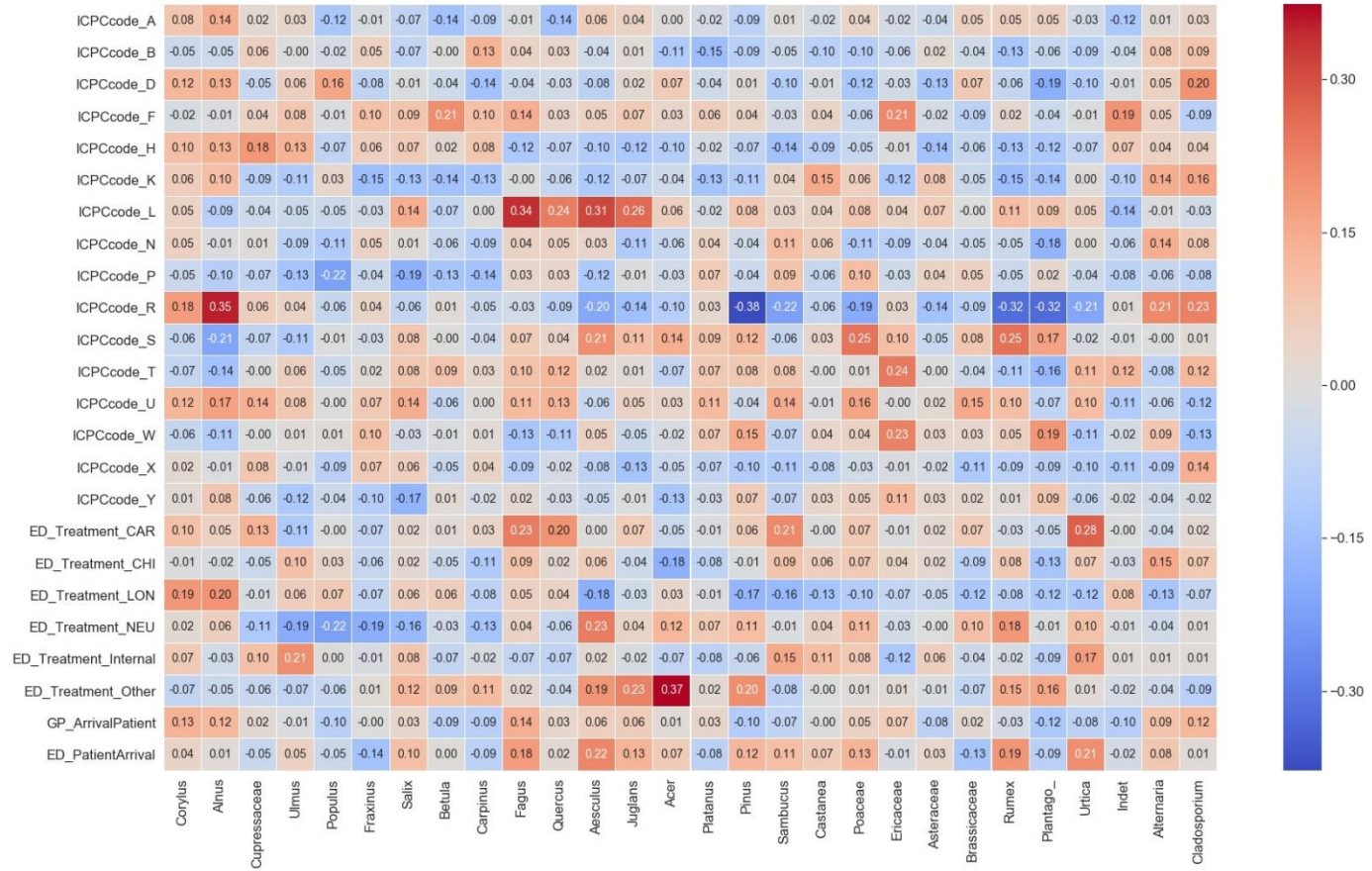


Figure 66: Pollen Correlation with Treatment Groups for all weekends in Spring

Appendix C – Forecasting ED patient demand

Forecasting ED patient demand: Training result of a single SARIMAX model

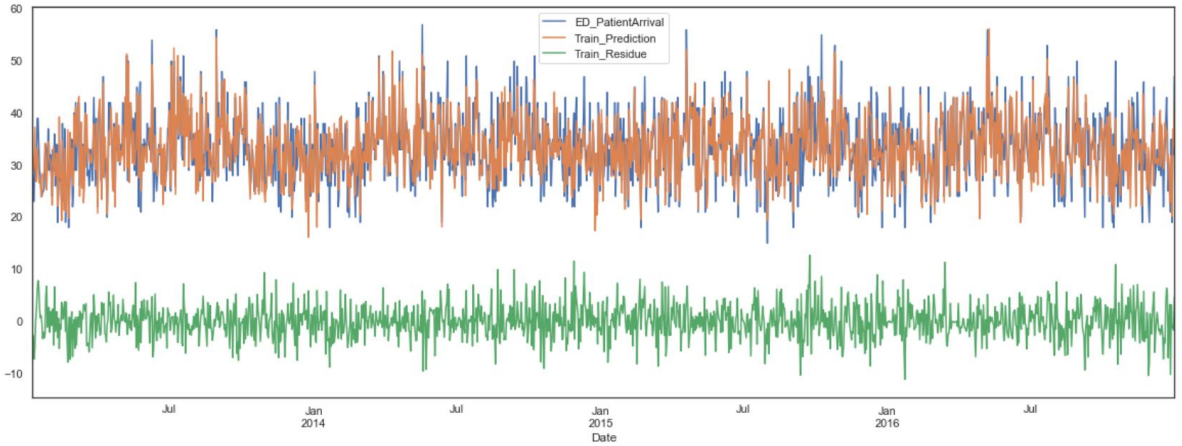


Figure 67: ED training line plot of a single SARIMAX(1, 0, 1)x(0, 0, 0, 7)

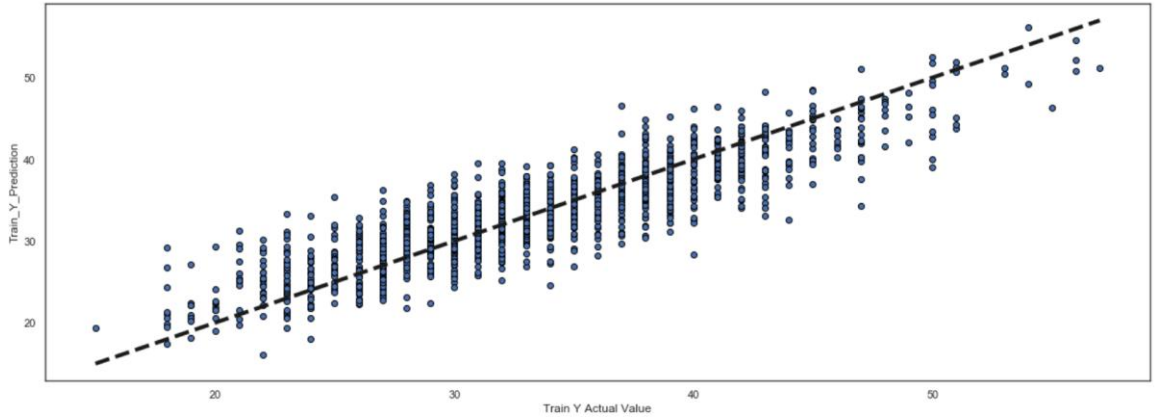


Figure 68: ED training Scatter plot of a single SARIMAX(1, 0, 1)x(0, 0, 0, 7)

Forecasting ED patient demand: Training result of a single SARIMAX model

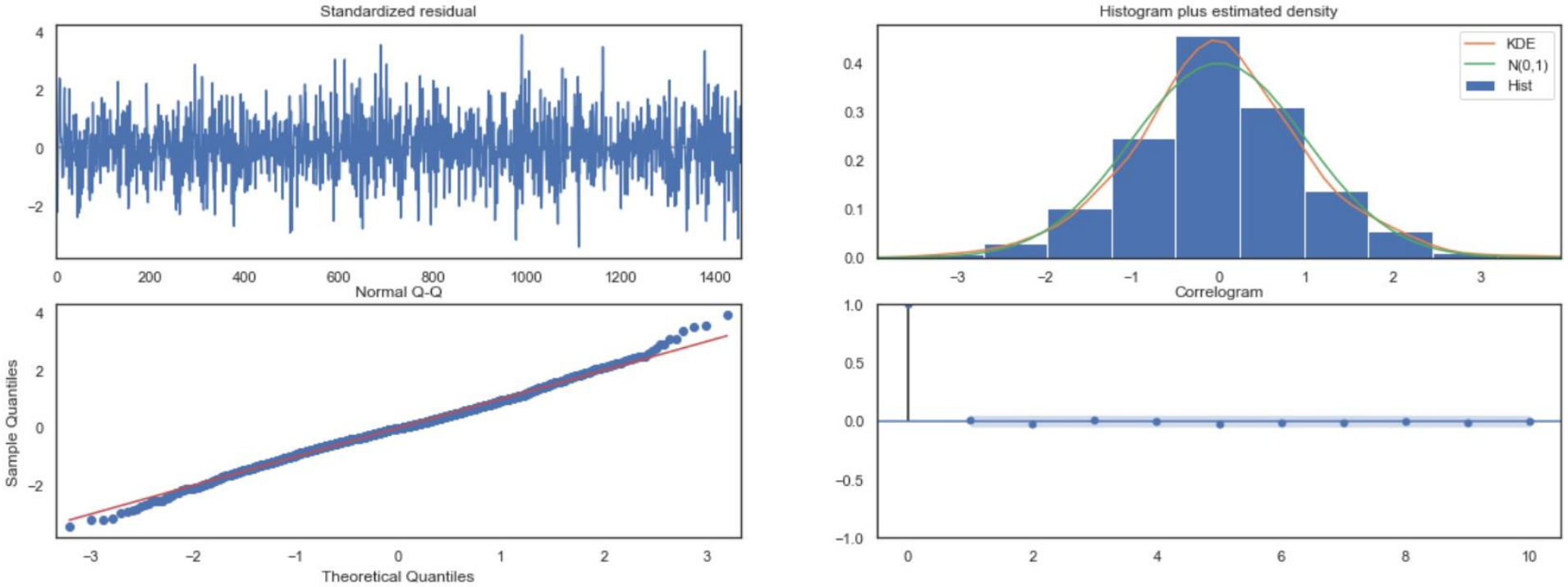


Figure 69: ED Plot Diagnosis of SARIMAX(1, 0, 1)x(0, 0, 0, 7)

Forecasting ED patient demand: Testing result of a single SARIMAX model

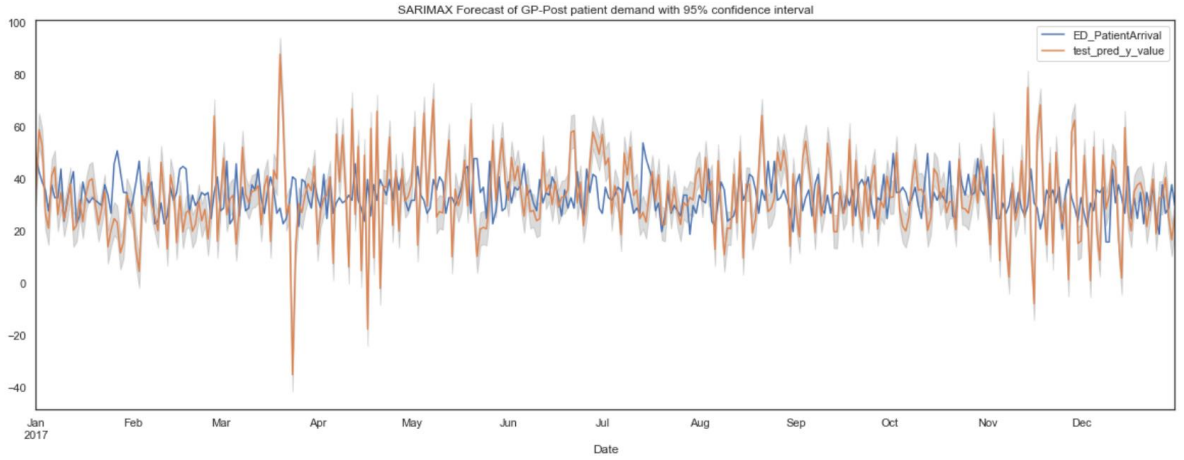


Figure 70: ED testing line plot of a single SARIMAX(1, 0, 1)x(0, 0, 0, 7)

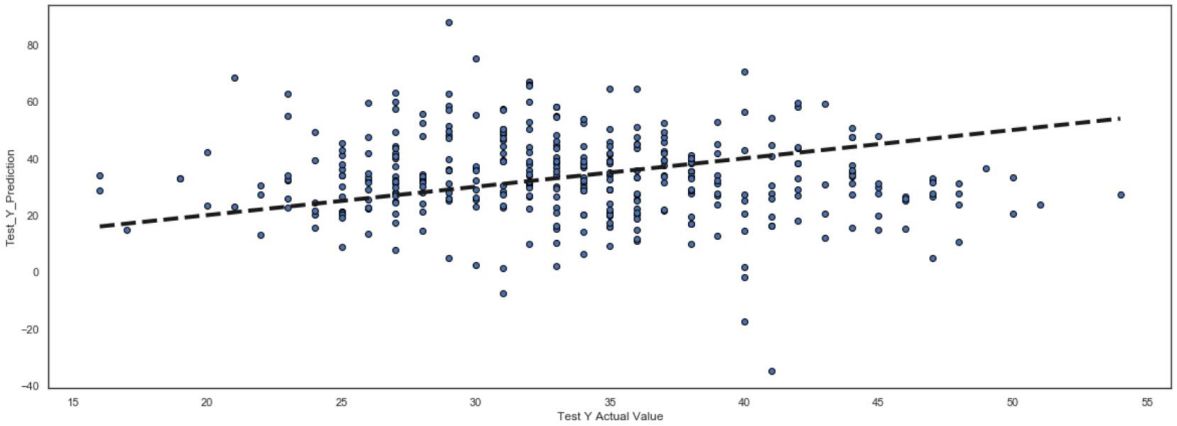


Figure 71: ED testing scatter plot of a single SARIMAX(1, 0, 1)x(0, 0, 0, 7)

Forecasting ED patient demand: Training result of a SARIMAX model with Feature Selection

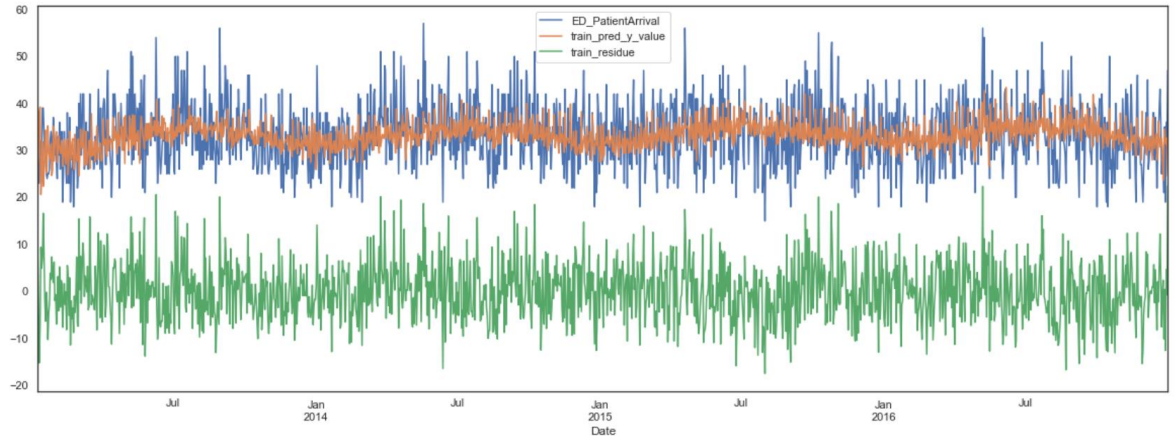


Figure 72: ED training line plot of a SARIMAX(0, 0, 0)x(1, 0, 1, 7) with feature selection

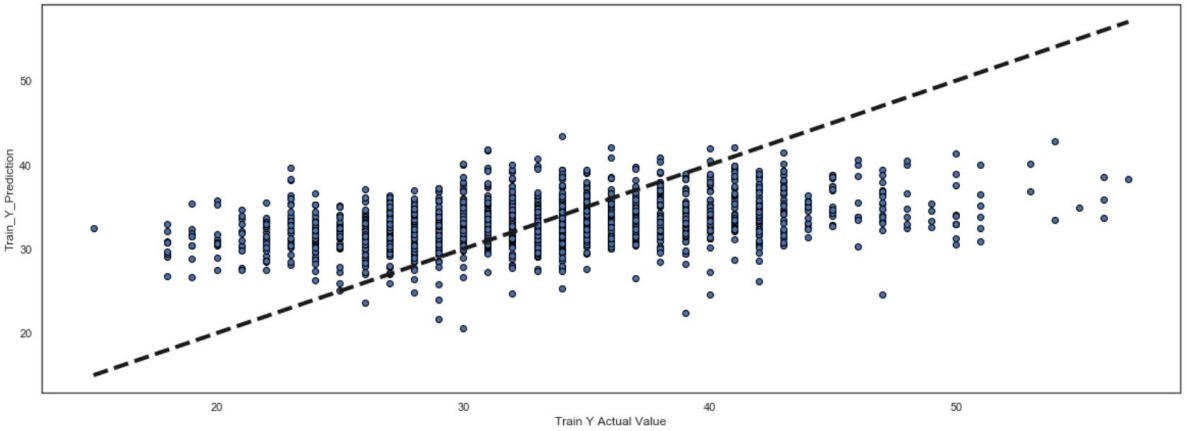


Figure 73: ED training Scatter plot of a SARIMAX(0, 0, 0)x(1, 0, 1, 7) with feature selection

Forecasting ED patient demand: Training result of a SARIMAX model with Feature Selection

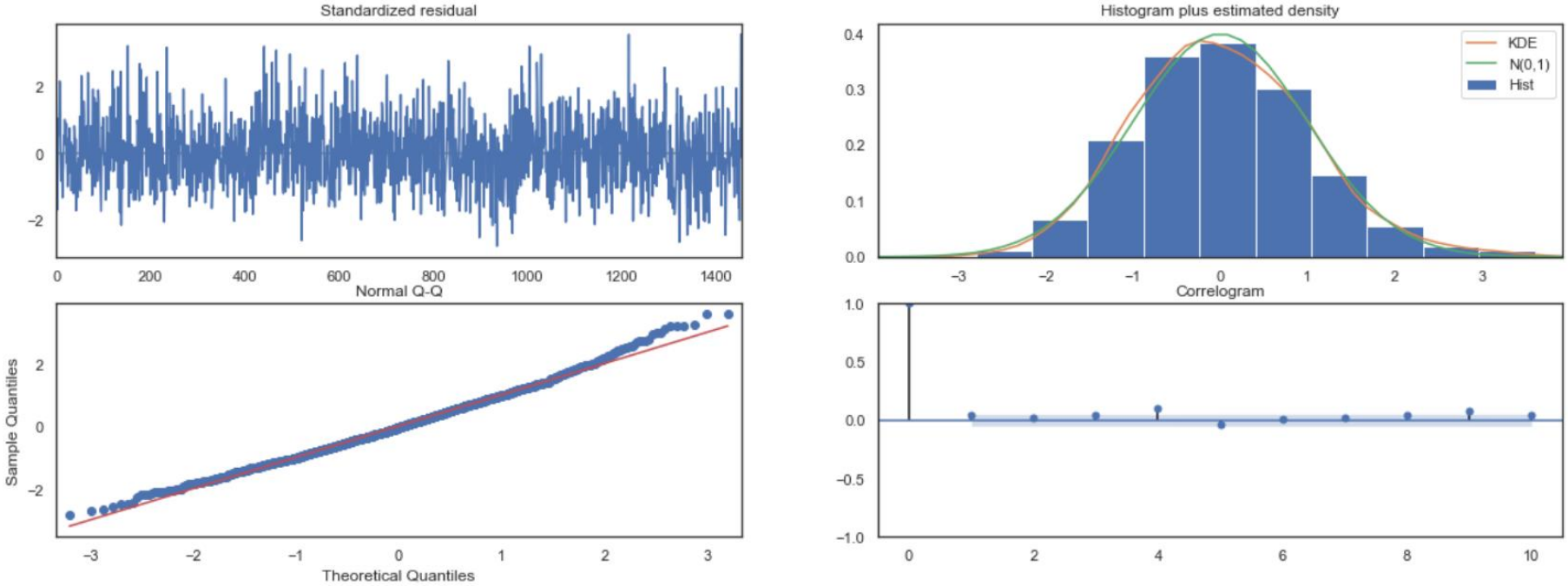


Figure 74: ED Diagnosis Plot of SARIMAX(0, 0, 0)x(1, 0, 1, 7)

Forecasting ED patient demand: Testing result of a SARIMAX model with Feature Selection

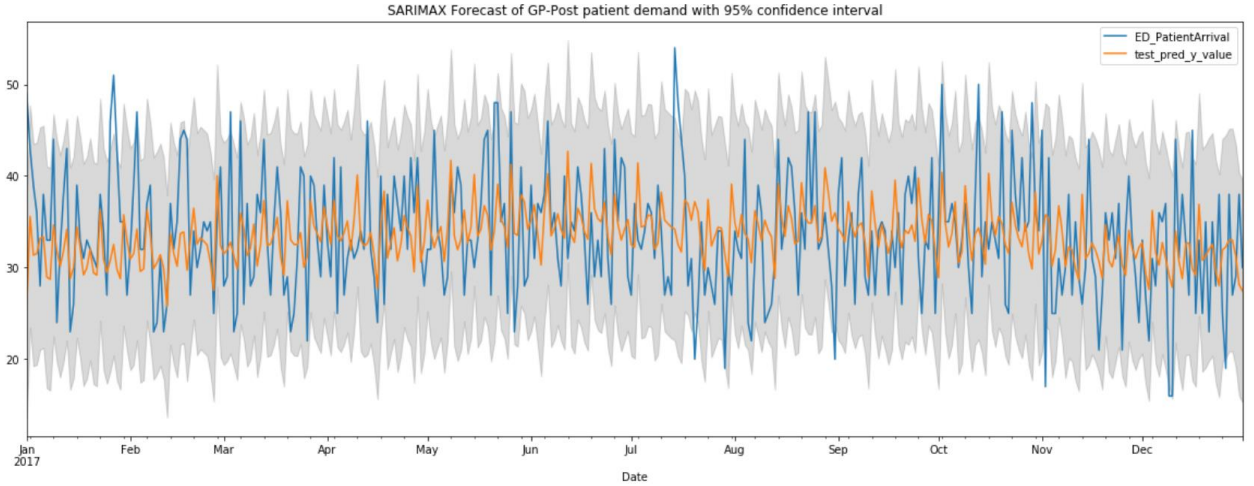


Figure 75: ED testing line Plot of a SARIMAX(0, 0, 0)x(1, 0, 1, 7) with feature selection

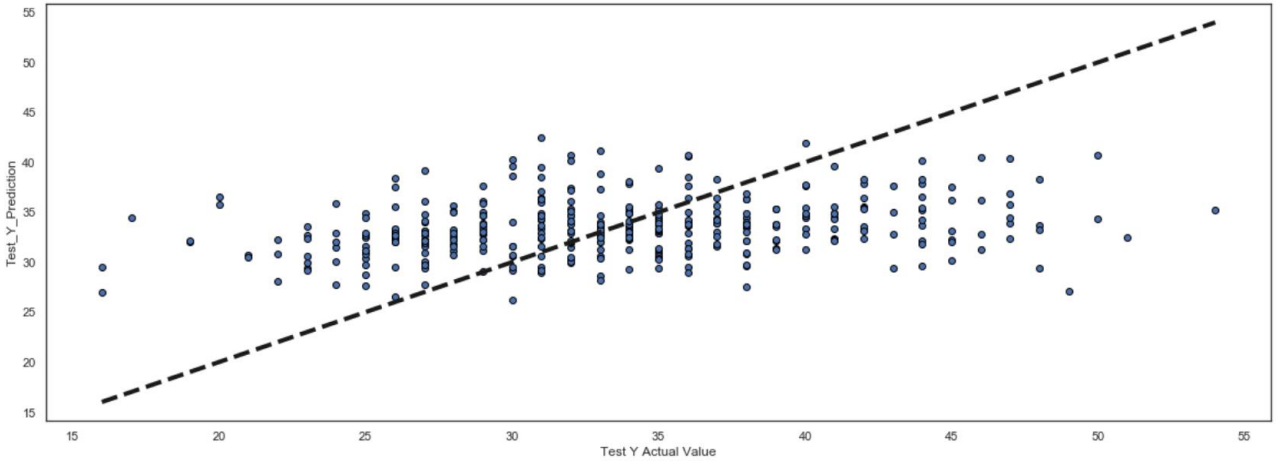


Figure 76: ED testing Scatter plot of a SARIMAX(0, 0, 0)x(1, 0, 1, 7) with feature selection

Forecasting ED patient demand: Result of a SARIMAX model with Feature Selection

Table 33: The initial 52 features of ED Feature Selection with Lasso

No	Features	Coef_Value	Absolute	Weight	No	Features	Coef_Value	Absolute	Weight
1	ICPCcode_L-1	3.44	3.44	6.42	27	W_RHXH-1	-0.33	0.33	0.62
2	Is_Weekday	2.71	2.71	5.06	28	W_RHXH-6	0.28	0.28	0.53
3	W_TX-1	1.97	1.97	3.69	29	quarter_cos4-7	-0.28	0.28	0.52
4	Is_Weekday-3	1.54	1.54	2.88	30	Artemisia-3	-0.25	0.25	0.48
5	Is_Weekday-2	-1.41	1.41	2.63	31	Treatment_NEU-6	-0.24	0.24	0.45
6	GP_Post_WH_Opening	-1.41	1.41	2.63	32	W_FHNH-3	0.21	0.21	0.40
7	is_month_start	-1.32	1.32	2.47	33	W_UXH-3	0.19	0.19	0.36
8	Referral_GP-4	1.19	1.19	2.22	34	W_FHXH-1	-0.16	0.16	0.30
9	Is_Spring-1	0.97	0.97	1.82	35	W_TNH-2	-0.16	0.16	0.29
10	Is_Weekday-7	0.73	0.73	1.36	36	U1-1	0.16	0.16	0.29
11	quarter_sin4-3	-0.72	0.72	1.34	37	W_RHXH-5	-0.14	0.14	0.26
12	Is_Weekday-6	-0.71	0.71	1.32	38	W_DDVEC-6	0.13	0.13	0.24
13	Is_Spring-7	0.65	0.65	1.21	39	W_RHXH-3	-0.13	0.13	0.24
14	W_T10NH-1	-0.55	0.55	1.03	40	W_FXXH-5	0.09	0.09	0.16
15	Is_Autumn-3	0.54	0.54	1.01	41	W_FHNH-1	-0.07	0.07	0.14
16	W_RHXH-2	-0.48	0.48	0.91	42	W_UXH-2	-0.06	0.06	0.11
17	Is_Holiday	-0.48	0.48	0.89	43	W_UXH-1	-0.06	0.06	0.11
18	Output_Home-4	0.47	0.47	0.88	44	W_TNH-7	-0.05	0.05	0.09
19	W_T10NH-2	-0.47	0.47	0.87	45	ICPCcode_Y-2	0.05	0.05	0.09
20	W_TNH-4	-0.46	0.46	0.86	46	day_cos_monthly-6	0.04	0.04	0.08
21	Age_20-65_year-1	0.43	0.43	0.81	47	W_UXH-7	-0.04	0.04	0.08
22	W_T10NH-3	-0.42	0.42	0.79	48	W_SP-4	0.04	0.04	0.07
23	day_cos_monthly-7	0.41	0.41	0.77	49	W_SQ-2	0.03	0.03	0.05
24	W_RHXH-7	-0.39	0.39	0.72	50	U1-6	-0.02	0.02	0.03
25	Age_5-19_year-2	0.37	0.37	0.69	51	quarter_sin4-2	-0.01	0.01	0.01
26	W_TNH-3	-0.37	0.37	0.69	52	is_month_end-1	0.00	0.00	0.00

Table 34: VIF factor of ED six selected features

Variable	Six selected features	VIF Factor
x1	Is_Weekday	2.60
x2	GP_Post_WH_Opening	1.04
x3	W_TX-1	5.64
x4	ICPCcode_L-1	2.95
x5	Is_Weekday-2	3.69
x6	Is_Weekday-3	4.66

Table 35: The ED complete list of coefficient and p-values of SARIMAX with Feature Selection

Variable(Features)	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
x1 (Is_Weekday)	17.7259	4.032	4.396	0	9.823	25.629
x2 (GP_Post_WH_Opening)	-4.0182	0.768	-5.234	0	-5.523	-2.513
x3 (W_TX-1)	6.5363	0.883	7.399	0	4.805	8.268
x4 (ICPCcode_L-1)	12.273	1.964	6.247	0	8.423	16.123
x5 (Is_Weekday-2)	9.0017	6.285	1.432	0.152	-3.318	21.321
x6 (Is_Weekday-3)	9.6127	6.328	1.519	0.129	-2.791	22.016
ar.S.L7	1	8.26E-05	1.21E+04	0	1	1
ma.S.L7	-0.9903	0.009	-115.115	0	-1.007	-0.973
sigma2	38.7482	1.43	27.102	0	35.946	41.55

Forecasting ED patient demand: Training and Testing result of a Hybrid model

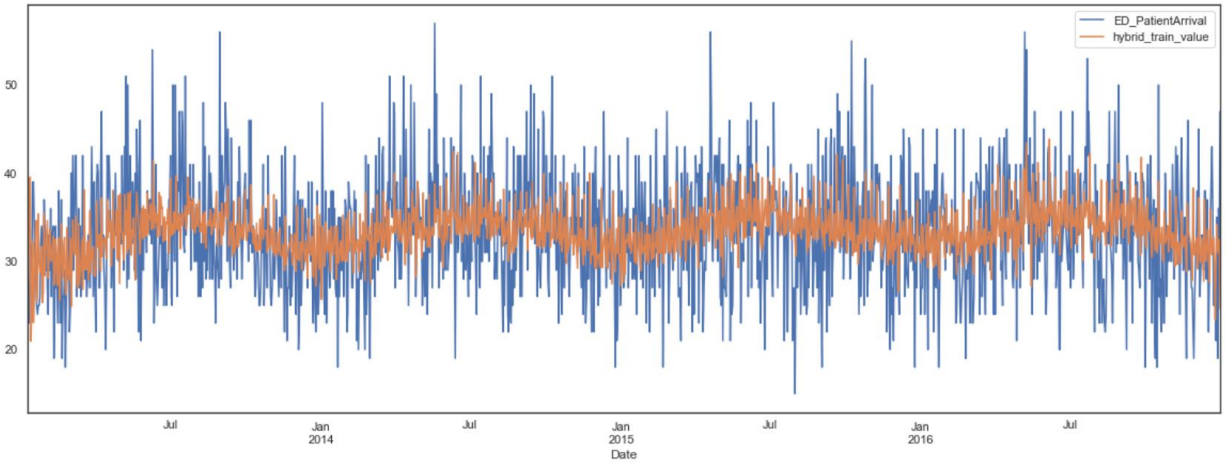


Figure 77: ED Hybrid training line Plot

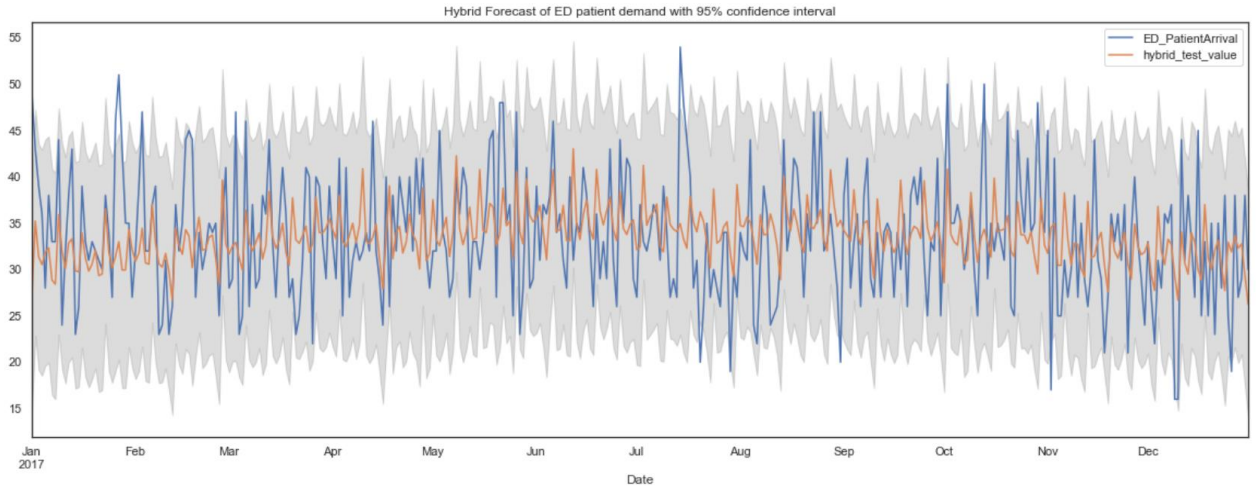


Figure 78: ED Hybrid Testing line Plot

Appendix D – Forecasting GP-Post patient demand

Forecasting GP-Post patient demand: Training result of a single SARIMAX model

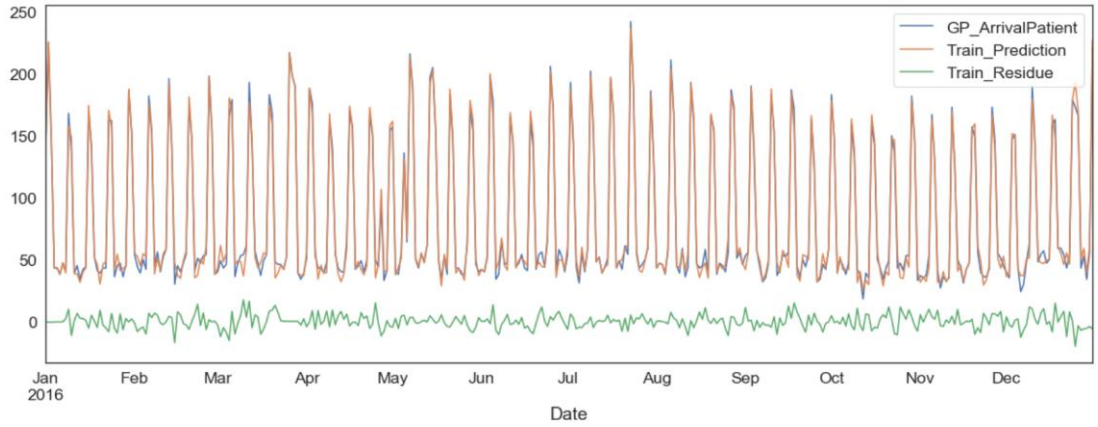


Figure 79: GP-Post training line plot of a single SARIMAX (1, 0, 2)x(0, 0, 0, 7) model

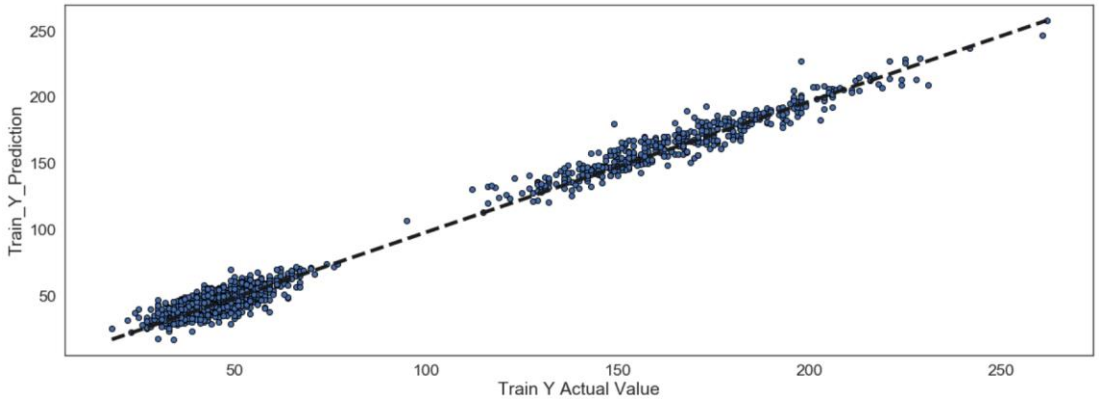


Figure 80: GP-Post training Scatter plot of a single SARIMAX (1, 0, 2)x(0, 0, 0, 7) model

Forecasting GP-Post patient demand: Training result of a single SARIMAX model

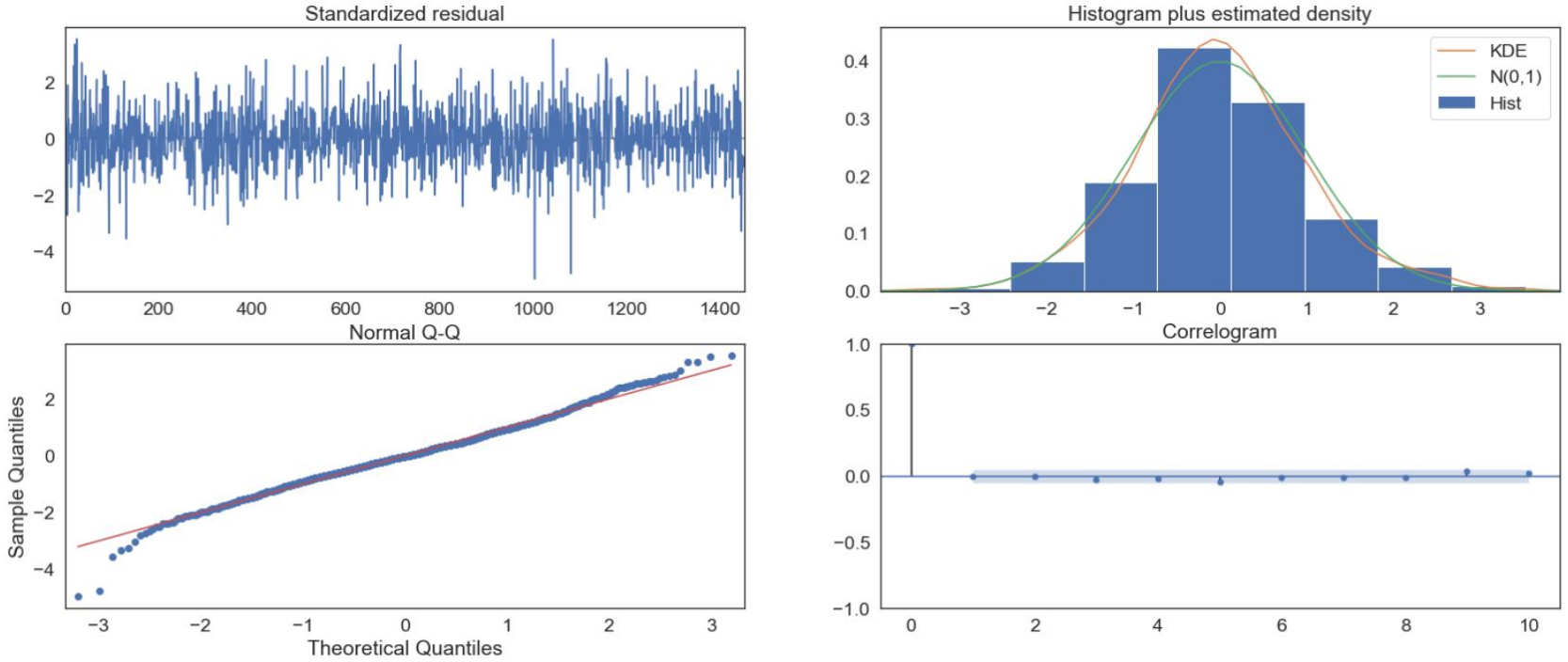


Figure 81: GP-Post Plot Diagnosis of a single SARIMAX (1, 0, 2)x(0, 0, 0, 7) model

Forecasting GP-Post patient demand: Testing Result of a single SARIMAX model

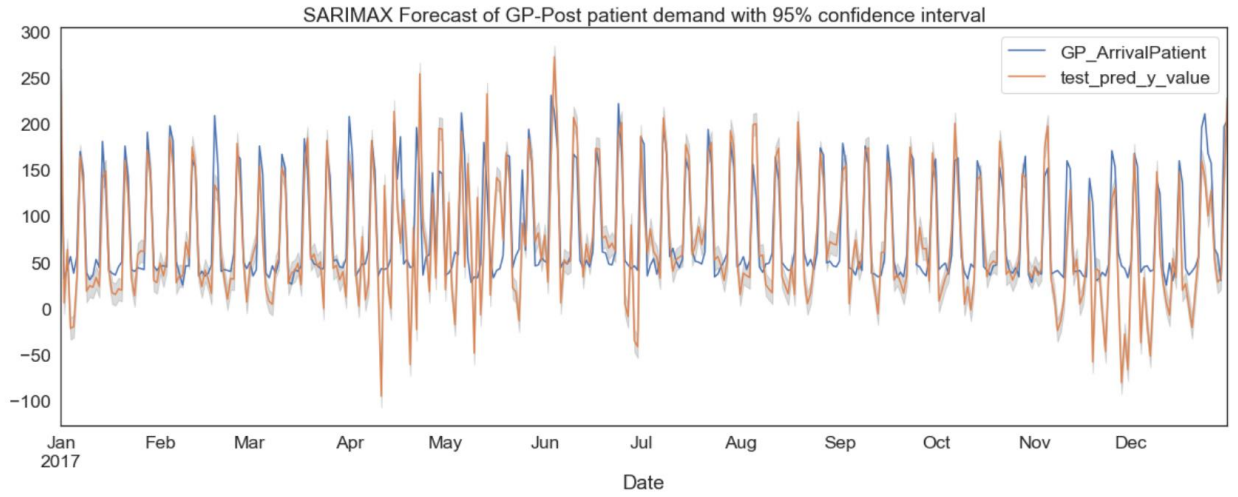


Figure 82: GP-Post testing line plot of a single SARIMAX (1, 0, 2)x(0, 0, 0, 7) model

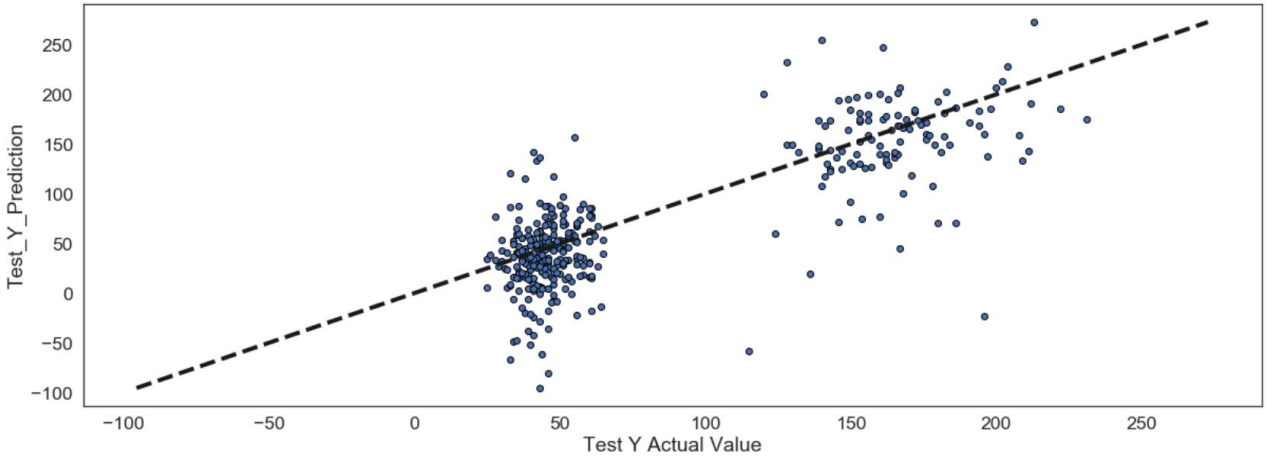


Figure 83: GP-Post testing Scatter plot of a single SARIMAX (1, 0, 2)x(0, 0, 0, 7) model

Forecasting GP-Post patient demand: Training result of a SARIMAX model with Feature Selection

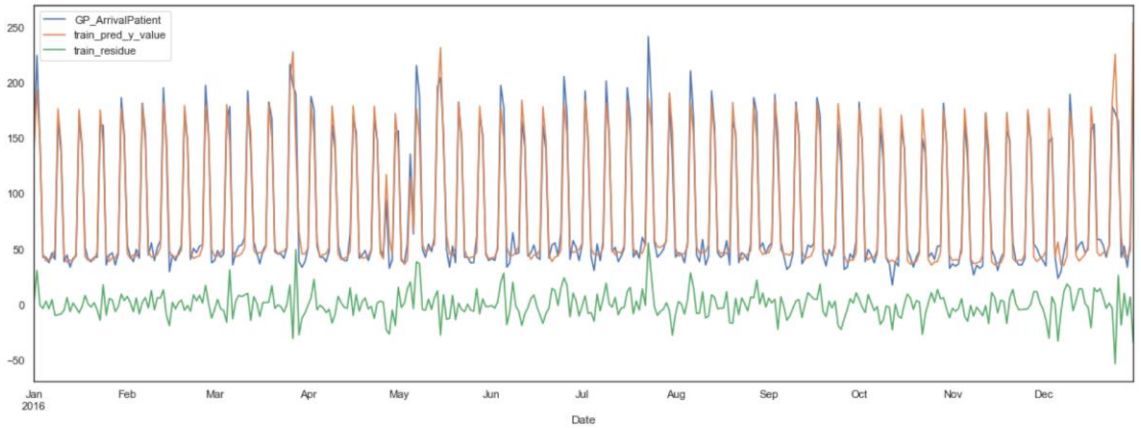


Figure 84: GP-Post training line Plot of a SARIMAX(1, 0, 1)x(1, 0, 1, 7) with feature selection

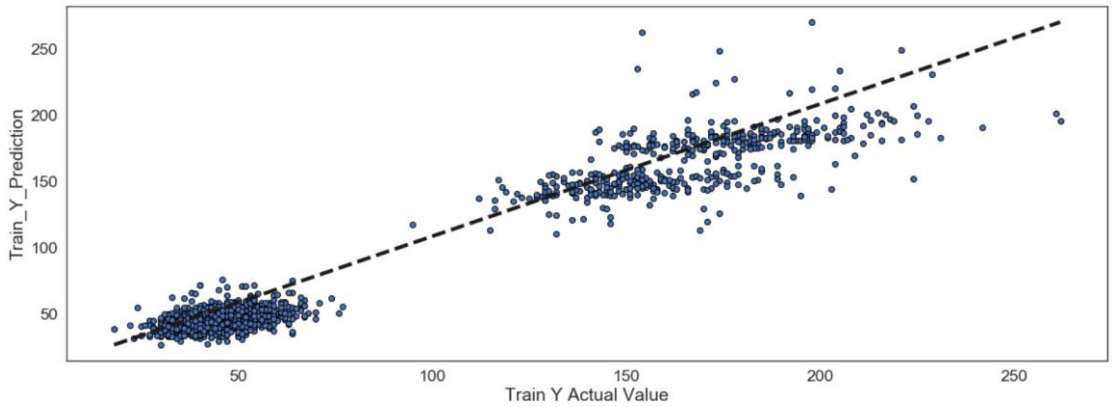


Figure 85: GP-Post training Scatter plot of a SARIMAX(1, 0, 1)x(1, 0, 1, 7) with feature selection

Forecasting GP-Post patient demand: Result of a SARIMAX model with Feature Selection

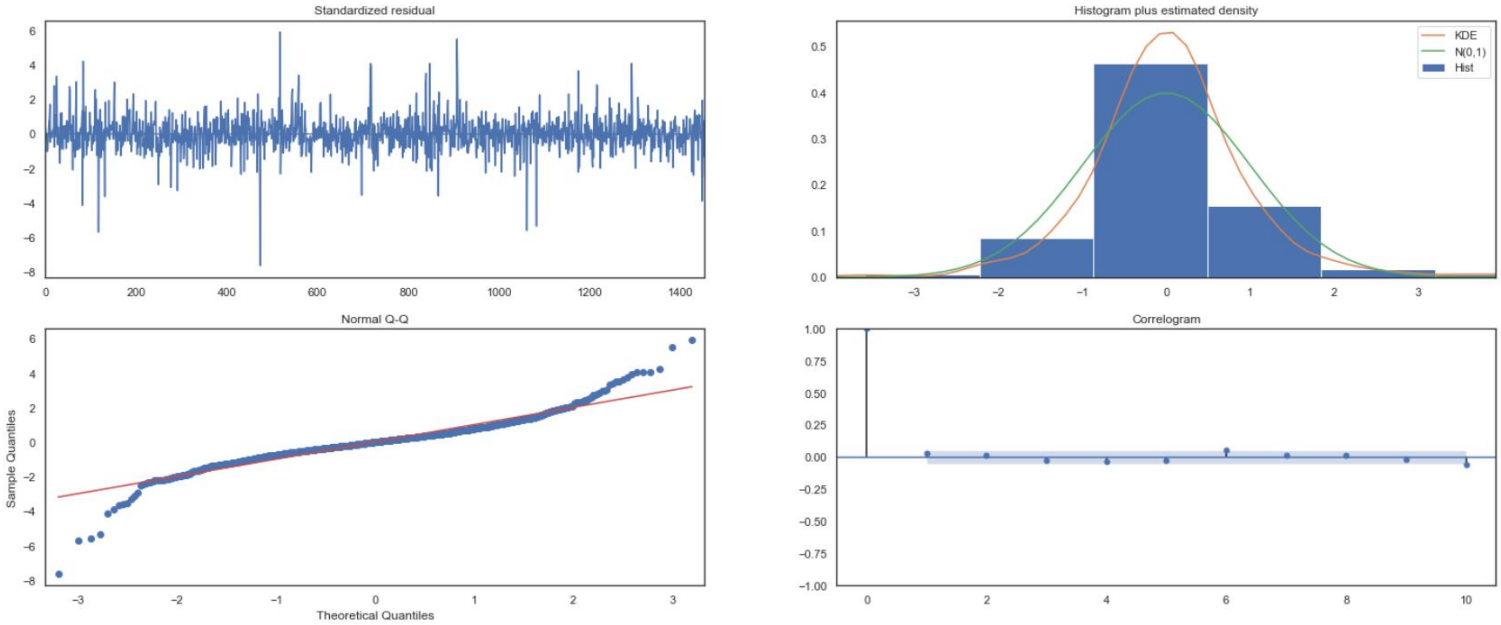


Figure 86: GP-Post Diagnosis Plot SARIMAX(1, 0, 1)x(1, 0, 1, 7) with feature selection

Forecasting GP-Post patient demand: Testing result of a SARIMAX model with Feature Selection

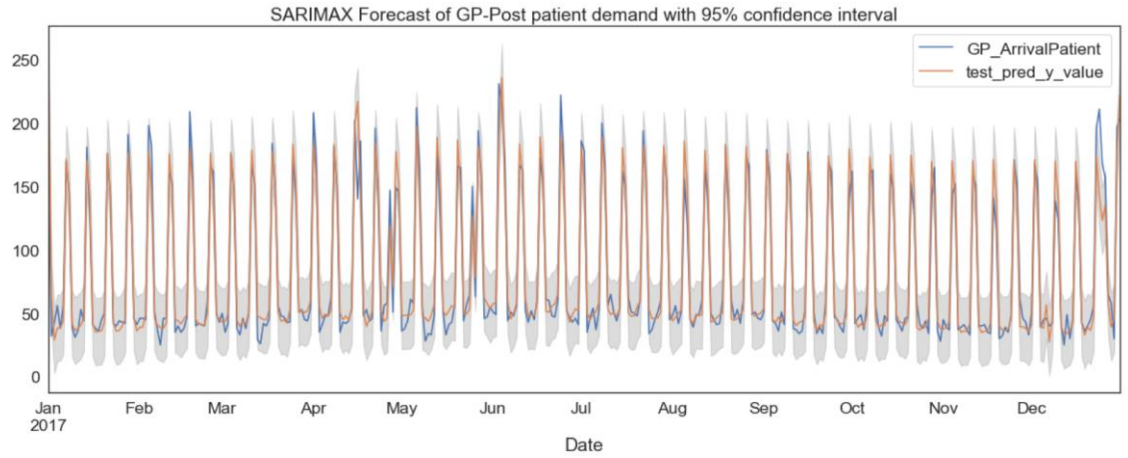


Figure 87: GP-Post testing Plot of SARIMAX(1, 0, 1)x(1, 0, 1, 7) with feature selection

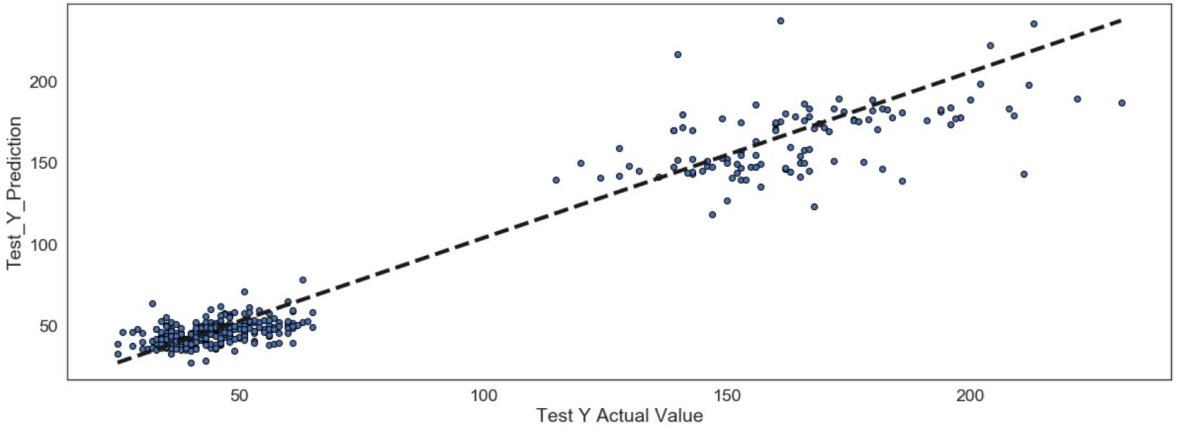


Figure 88: GP-Post testing Scatter plot of SARIMAX(1, 0, 1)x(1, 0, 1, 7) with feature selection

Forecasting GP-Post patient demand: Result of a SARIMAX model with Feature Selection

Table 36: The initial 132 features of GP feature selection with Lasso

No	Features	Coef	Absolute	Weight	No	Features	Coef	Absolute	Weight	No	Features	Coef	Absolute	Weight
1	Is_Weekday	-91.17	91.17	0.21	46	Is_Weekday-4	-1.63	1.63	0.00	91	W_SP-1	-0.59	0.59	0.00
2	GP_Post_WH_Opening	69.56	69.56	0.16	47	Is_Holiday	1.63	1.63	0.00	92	ED_Referral_Ambulance-1	-0.57	0.57	0.00
3	Is_Weekday-5	25.13	25.13	0.06	48	ICPCcode_N-6	-1.60	1.60	0.00	93	day_sin_monthly-5	-0.57	0.57	0.00
4	Is_Weekday-3	19.18	19.18	0.04	49	is_quarter_start-5	1.55	1.55	0.00	94	W_T10NH-6	-0.54	0.54	0.00
5	Is_Weekday-1	19.18	19.18	0.04	50	ED_Treatment_LON-3	1.51	1.51	0.00	95	Age_75_plus-2	-0.52	0.52	0.00
6	Is_Holiday-1	16.40	16.40	0.04	51	GP_Post_Age_20-65_year-7	1.49	1.49	0.00	96	W_UXH-6	0.48	0.48	0.00
7	U5-7	7.81	7.81	0.02	52	ED_Urgency_Geel-3	1.48	1.48	0.00	97	W_RHXH-3	-0.46	0.46	0.00
8	W_EV24-3	6.36	6.36	0.01	53	is_month_start-6	1.44	1.44	0.00	98	ED_Referral_SKB-5	0.44	0.44	0.00
9	Is_Holiday-2	-6.26	6.26	0.01	54	W_RHXH-7	1.42	1.42	0.00	99	Age_66-74_year-6	0.43	0.43	0.00
10	GP_Post_Age_0-4_year-2	6.23	6.23	0.01	55	ED_Output_Admittance-3	1.42	1.42	0.00	100	W_FXX-6	-0.43	0.43	0.00
11	weekofyear_sin52-7	6.22	6.22	0.01	56	Is_Spring-2	-1.31	1.31	0.00	101	day_sin_monthly-6	-0.40	0.40	0.00
12	ICPCcode_R-7	5.58	5.58	0.01	57	W_SP-7	-1.25	1.25	0.00	102	W_RHXH-4	-0.40	0.40	0.00
13	W_TX-1	5.41	5.41	0.01	58	W_TNH-7	1.24	1.24	0.00	103	Is_Summer-7	0.39	0.39	0.00
14	ICPCcode_U-6	5.22	5.22	0.01	59	ICPCcode_H-1	1.12	1.12	0.00	104	ED_Urgency_Blauw-1	0.37	0.37	0.00
15	ICPCcode_F-7	4.74	4.74	0.01	60	ED_Urgency_Blauw-3	1.11	1.11	0.00	105	W_UXH-1	0.34	0.34	0.00
16	quarter_cos4-3	4.32	4.32	0.01	61	W_FHXH-2	-1.09	1.09	0.00	106	W_UXH-7	-0.28	0.28	0.00
17	GP_Post_WH_Opening-5	-3.88	3.88	0.01	62	ED_Urgency_Blauw-5	1.08	1.08	0.00	107	U1-4	0.26	0.26	0.00
18	weekday_sin7-3	3.72	3.72	0.01	63	W_TNH-4	-1.08	1.08	0.00	108	ED_Urgency_Overig-4	-0.26	0.26	0.00
19	month_cos12	-3.71	3.71	0.01	64	ED_Treatment_LON-5	1.05	1.05	0.00	109	Age_20-65_year-3	0.24	0.24	0.00
20	ED_Treatment_LON-6	3.67	3.67	0.01	65	ED_Urgency_Geel-2	-1.02	1.02	0.00	110	ICPCcode_A-2	0.23	0.23	0.00
21	ICPCcode_S-7	3.58	3.58	0.01	66	ED_Treatment_NEU-6	1.01	1.01	0.00	111	ED_Urgency_Oranje-4	0.22	0.22	0.00
22	ED_Referral_GP-1	3.48	3.48	0.01	67	ED_Output_Transferral-7	0.98	0.98	0.00	112	W_T10NH-3	-0.22	0.22	0.00
23	Is_Winter-6	3.30	3.30	0.01	68	ED_Output_Transferral-4	-0.96	0.96	0.00	113	Age_0-4_year-4	0.20	0.20	0.00
24	Is_Autumn-7	-3.15	3.15	0.01	69	W_SQ-3	0.94	0.94	0.00	114	W_SQ-5	0.19	0.19	0.00
25	U1-7	2.93	2.93	0.01	70	ED_Treatment_Internal-5	0.90	0.90	0.00	115	ED_Urgency_Overig-1	0.19	0.19	0.00
26	Is_Weekday-6	-2.88	2.88	0.01	71	W_TN-1	0.88	0.88	0.00	116	W_DDVEC-1	-0.18	0.18	0.00

27	Is_Spring-3	-2.87	2.87	0.01	72	Is_Autumn	0.88	0.88	0.00	117	W_UXH-5	-0.16	0.16	0.00
28	weakday_cos7-1	-2.82	2.82	0.01	73	ED_Urgency_Groen-7	-0.88	0.88	0.00	118	Is_Spring-7	-0.15	0.15	0.00
29	GP_Post_WH_Opening-7	-2.75	2.75	0.01	74	W_FHNH-1	0.87	0.87	0.00	119	Is_Holiday-4	-0.14	0.14	0.00
30	W_RHXH-2	-2.57	2.57	0.01	75	W_FHNH-4	0.83	0.83	0.00	120	W_SQ-6	0.13	0.13	0.00
31	W_RHXH-1	-2.56	2.56	0.01	76	W_DDVEC-4	0.83	0.83	0.00	121	W_FHNH-7	-0.13	0.13	0.00
32	is_month_end	-2.54	2.54	0.01	77	ICPCcode_N-7	0.83	0.83	0.00	122	ED_Output_Transferral-5	0.11	0.11	0.00
33	Is_Weekday-7	-2.39	2.39	0.01	78	ED_Referral_Self_Referral-5	-0.82	0.82	0.00	123	W_DDVEC-6	0.11	0.11	0.00
34	W_TXH-3	2.24	2.24	0.01	79	Is_Holiday-5	-0.82	0.82	0.00	124	Age_75_plus-5	0.10	0.10	0.00
35	ICPCcode_D-1	2.22	2.22	0.01	80	ED_Treatment_CAR-7	0.81	0.81	0.00	125	ED_Gender_Man-1	0.09	0.09	0.00
36	Is_Autumn-4	2.00	2.00	0.00	81	W_UXH-4	-0.80	0.80	0.00	126	Is_Winter-7	0.09	0.09	0.00
37	quarter_sin4	-1.98	1.98	0.00	82	ICPCcode_T-6	-0.80	0.80	0.00	127	Age_0-4_year-3	-0.04	0.04	0.00
38	ED_Urgency_Overig-3	-1.97	1.97	0.00	83	ED_Referral_Other-7	-0.72	0.72	0.00	128	Artemisia-7	-0.03	0.03	0.00
39	W_UNH-5	1.95	1.95	0.00	84	Is_Holiday-6	-0.70	0.70	0.00	129	W_T10NH-1	0.03	0.03	0.00
40	Is_Holiday-3	1.89	1.89	0.00	85	ICPCcode_Y-4	-0.69	0.69	0.00	130	Is_Spring-4	0.00	0.00	0.00
41	W_FHVEC-1	-1.85	1.85	0.00	86	weakofyear_cos52-5	-0.68	0.68	0.00	131	is_month_end-7	0.00	0.00	0.00
42	ED_Treatment_LON-4	1.77	1.77	0.00	87	W_UXH-2	0.67	0.67	0.00	132	is_quarter_end-6	0.00	0.00	0.00
43	ED_Urgency_Rood-1	1.71	1.71	0.00	88	W_UNH-3	0.63	0.63	0.00					
44	W_RHXH-6	-1.71	1.71	0.00	89	ICPCcode_T-5	-0.63	0.63	0.00					
45	Age_0-4_year-1	1.70	1.70	0.00	90	ED_Treatment_LON-7	0.63	0.63	0.00					

Table 37: GP-Post VIF test result

Variable	Seven Selected Features	VIF_Factor
x1	Is_Weekday	7.576491
x2	GP_Post_WH_Opening	1.15806
x3	U5-7	9.735467
x4	Is_Holiday-1	1.159655
x5	Is_Weekday-1	5.29852
x6	Is_Weekday-3	5.082313
x7	Is_Weekday-5	4.723685

Table 38: The GP-Post complete list of coefficient and p-values of SARIMAX with Feature Selection

Variables (Features)	coef	std err	z	P>|z|	[0.025	0.975]
x1 (Is_Weekday)	-77.8818	8.223	-9.471	0	-93.999	-61.765
x2 (GP_Post_WH_Opening)	74.7898	0.889	84.101	0	73.047	76.533
x3 (U5-7)	12.5091	3.583	3.491	0	5.486	19.532
x4 (Is_Holiday-1)	18.2473	1.132	16.118	0	16.028	20.466
x5 (Is_Weekday-1)	48.9703	6.101	8.027	0	37.012	60.928
x6 (Is_Weekday-3)	65.18	7.692	8.474	0	50.104	80.256
x7 (Is_Weekday-5)	61.2473	7.775	7.877	0	46.008	76.486
ar.L1	0.9515	0.018	53.008	0	0.916	0.987
ma.L1	-0.8539	0.025	-33.877	0	-0.903	-0.805
ar.S.L7	0.9999	0	3291.706	0	0.999	1
ma.S.L7	-0.9913	0.009	-108.465	0	-1.009	-0.973
sigma2	186.8632	3.69	50.64	0	179.631	194.095

Forecasting GP-Post patient demand: Result of a Hybrid model

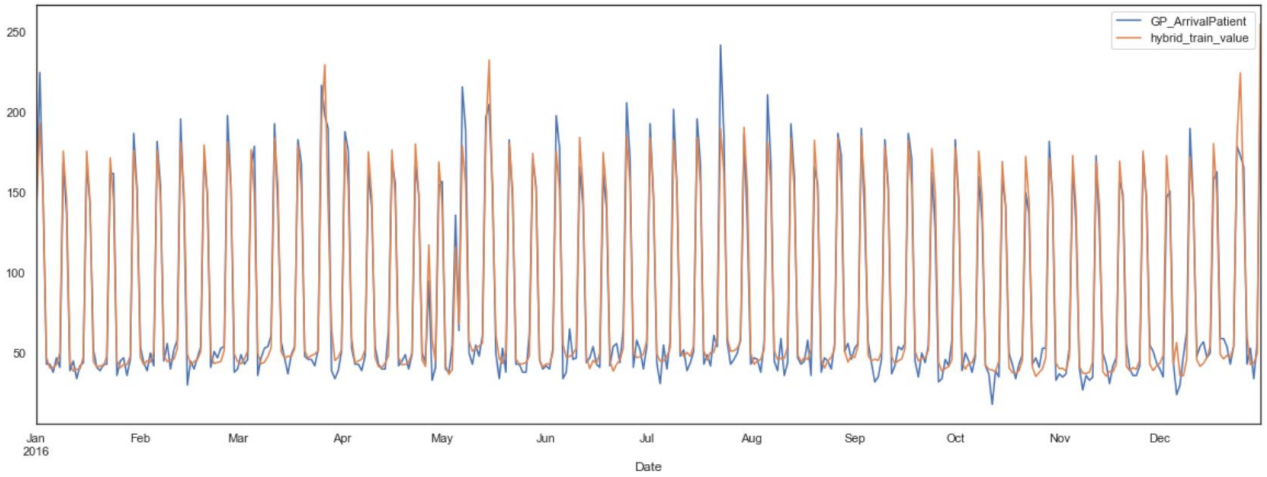


Figure 89: GP-Post Hybrid Training line plot

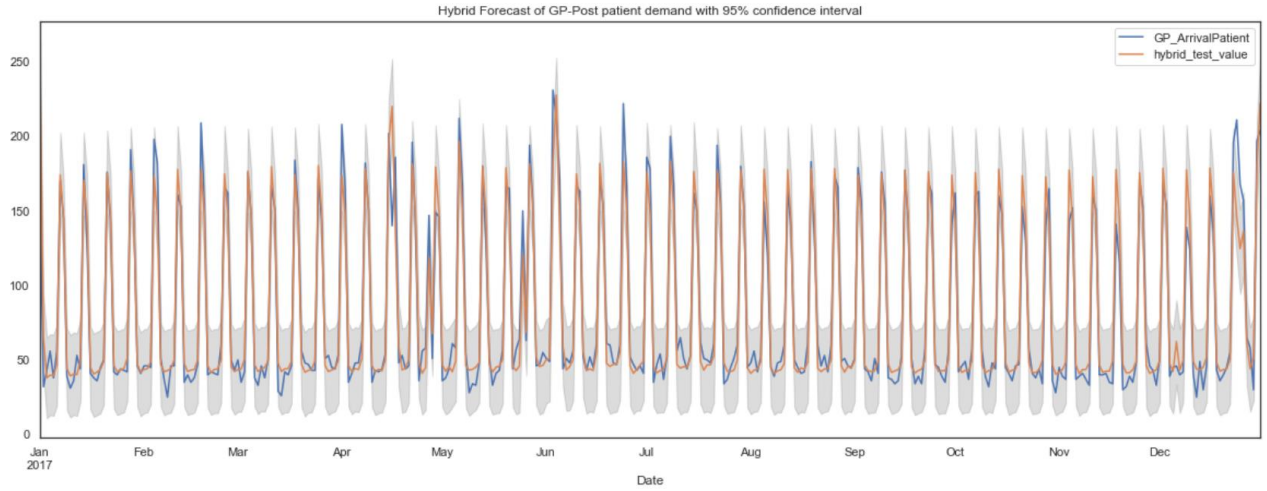


Figure 90: GP-Post Hybrid Testing line plot

Appendix E – Predicting ED Inpatient Admission

Inpatient Admission to the Hospital: Hyper Parameter Analysis

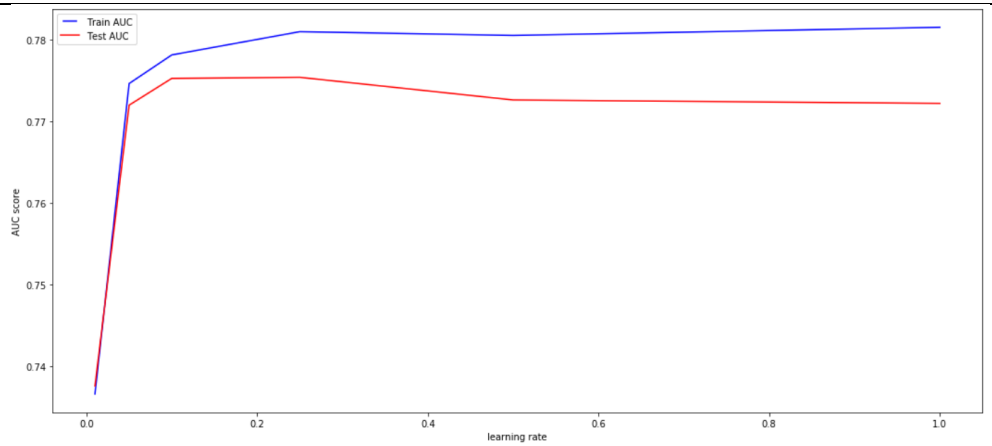


Figure 91: Learning rate

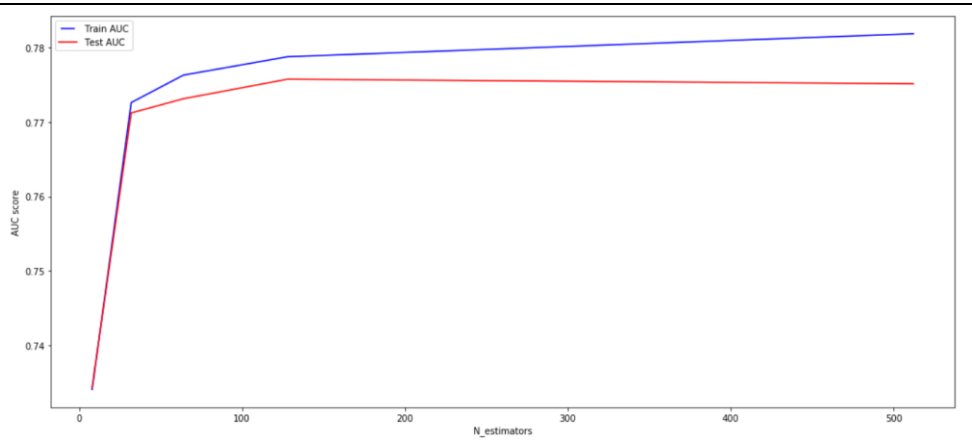


Figure 92: n_estimator

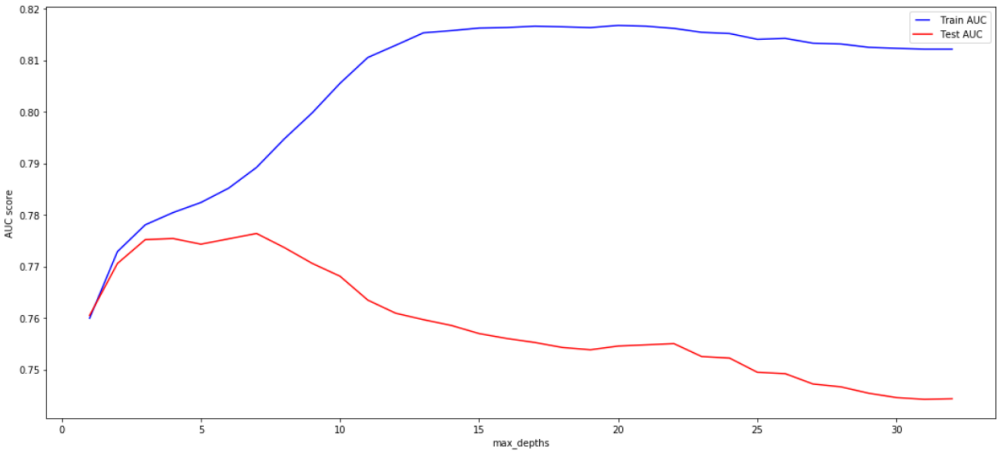


Figure 93: max_depths

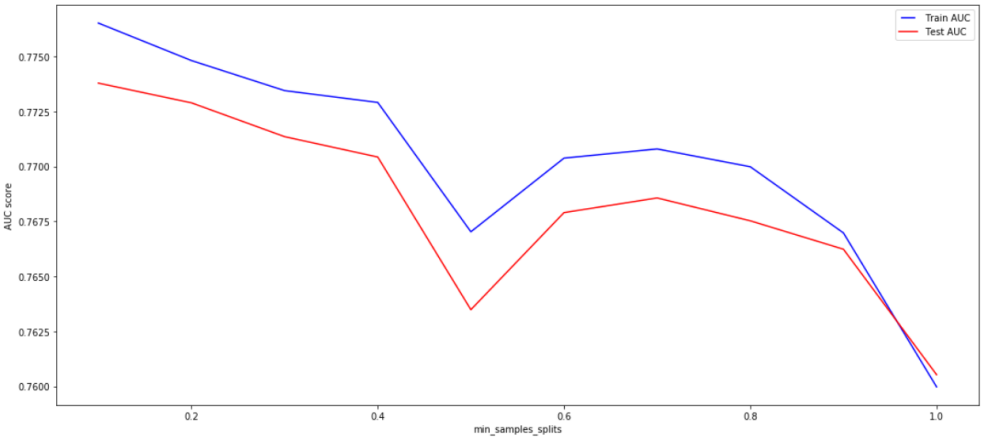


Figure 94: min_sample_splits

Inpatient Admission to the Hospital: Hyper Parameter Analysis

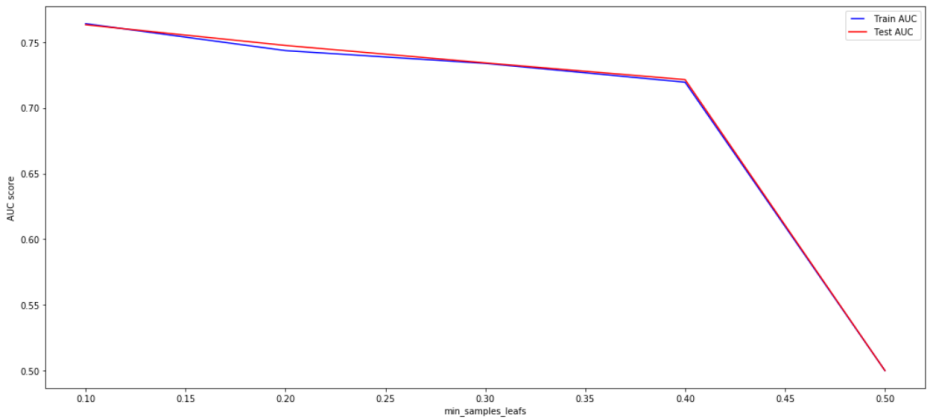


Figure 95: max_sample_leafs

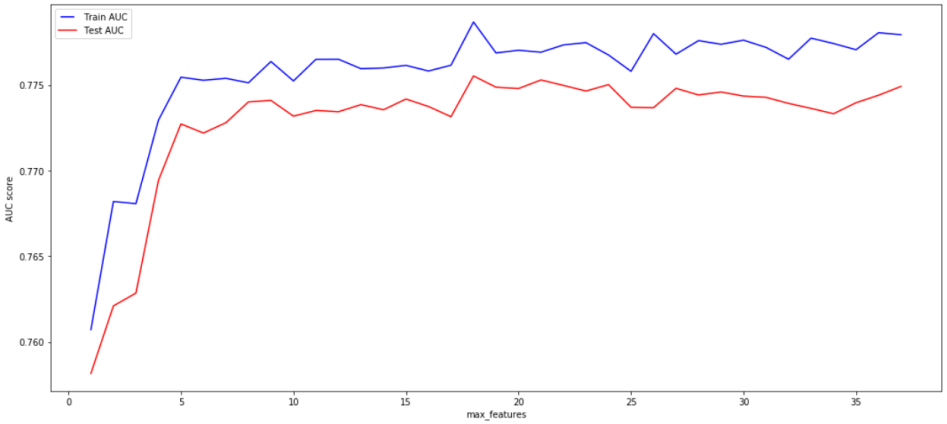


Figure 96: max_features

Inpatient Admission to the Hospital: ROC curve and Precision-Recall Curve

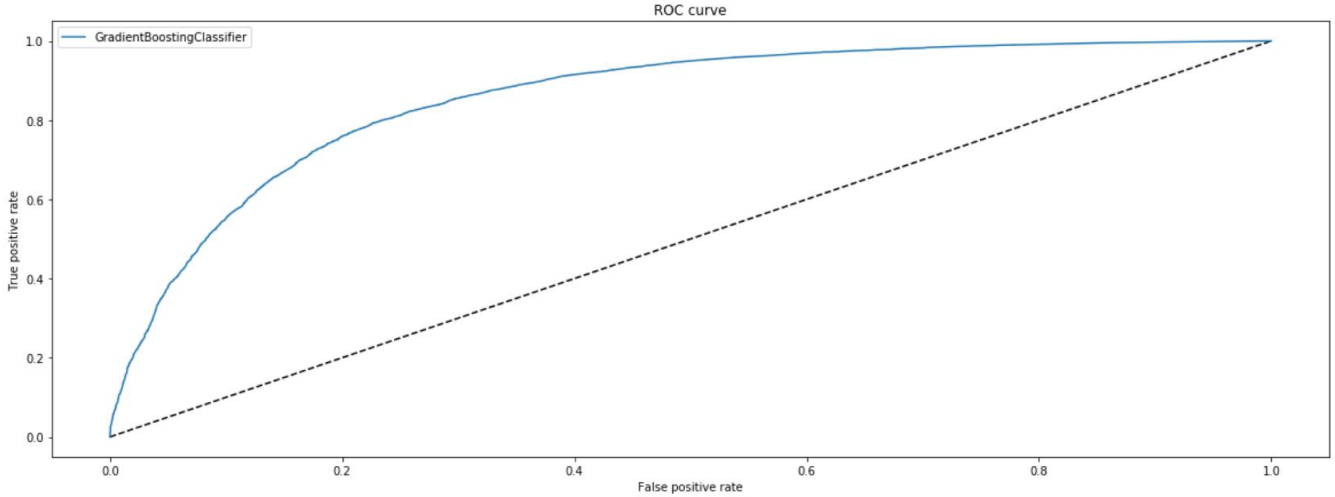


Figure 97: ROC Curve

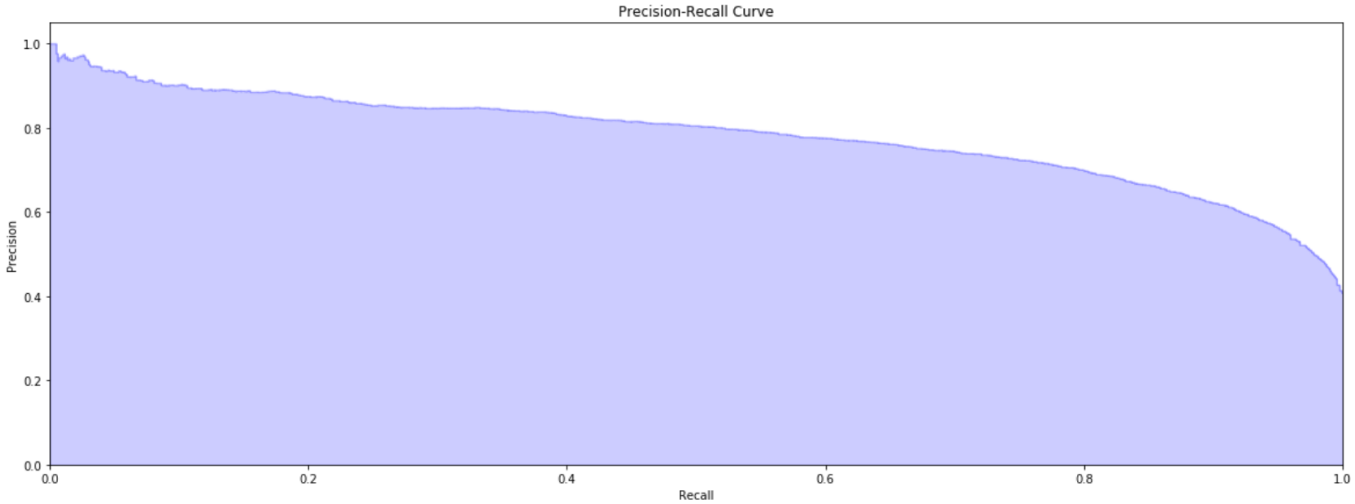


Figure 98: Precision-Recall Curve

Inpatient Admission to the Hospital: Feature Importance

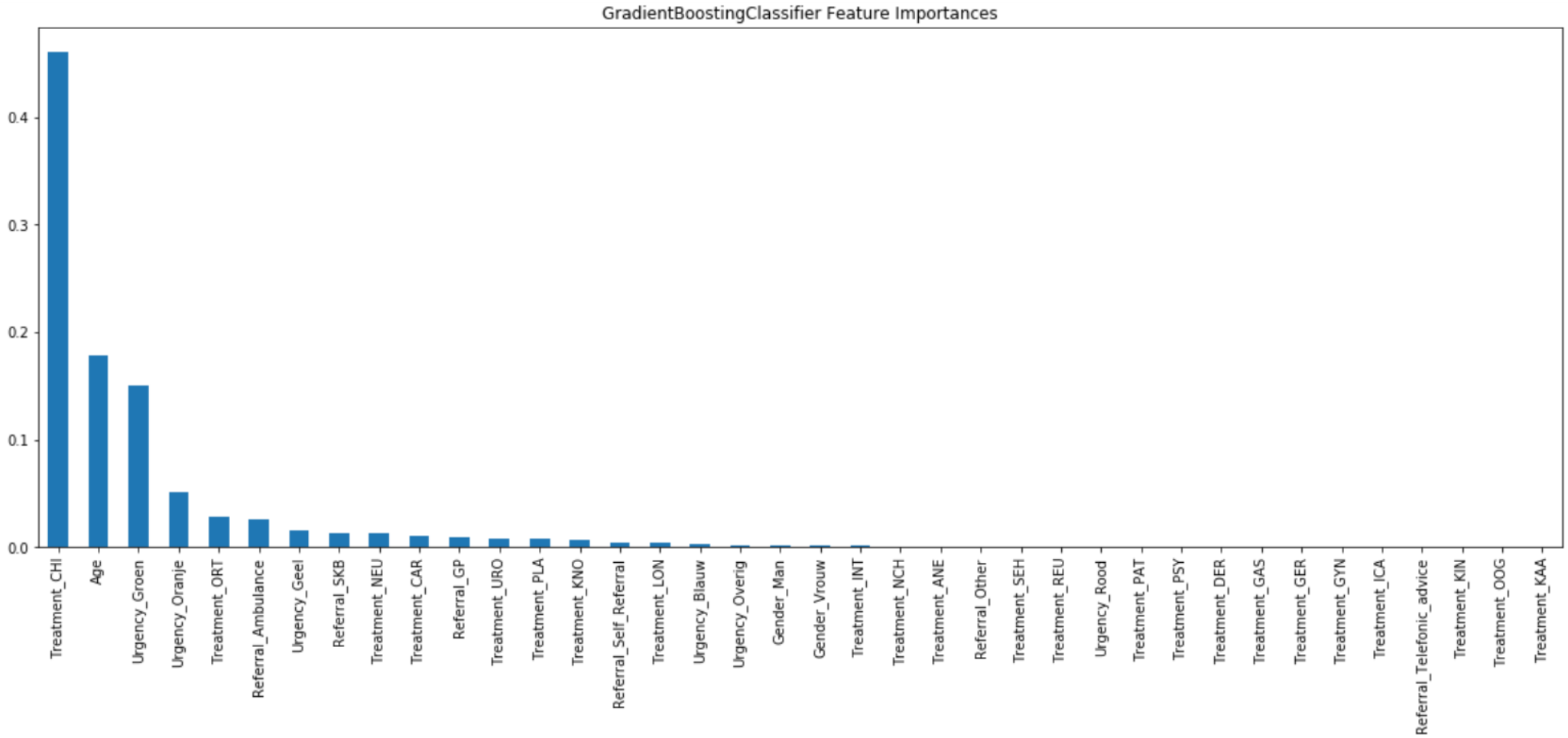


Figure 99: Feature Importance diagram of GradientBoostingClassifier

Inpatient Admission to the Hospital: Feature Importance

Table 39: Feature Importance scores

No	features	feature_importances_	No	features	feature_importances_
1	Treatment_CHI	0.460757042	22	Urgency_Rood	0
2	Age	0.178184457	23	Treatment_ANE	0
3	Urgency_Groen	0.149987949	24	Treatment_DER	0
4	Urgency_Oranje	0.051678985	25	Treatment_GAS	0
5	Treatment_ORT	0.028431097	26	Treatment_GER	0
6	Referral_Ambulance	0.026211205	27	Treatment_GYN	0
7	Urgency_Geel	0.015661859	28	Treatment_ICA	0
8	Referral_SKB	0.01328728	29	Treatment_KAA	0
9	Treatment_NEU	0.013042191	30	Treatment_KIN	0
10	Treatment_CAR	0.010899778	31	Treatment_NCH	0
11	Referral_GP	0.009029394	32	Treatment_OOG	0
12	Treatment_URO	0.008468911	33	Treatment_PAT	0
13	Treatment_PLA	0.007886702	34	Treatment_PSY	0
14	Treatment_KNO	0.006384699	35	Treatment_REU	0
15	Referral_Self_Referral	0.004609943	36	Treatment_SEH	0
16	Treatment_LON	0.004437236	37	Referral_Other	0
17	Urgency_Blauw	0.003269031	38	Referral_Telefonic_advice	0
18	Urgency_Overig	0.002382694			
19	Gender_Man	0.002098072			
20	Gender_Vrouw	0.002062631			
21	Treatment_INT	0.001228846			