A Work Project, presented as part of the requirements for the Award of a Master's degree in **Business Analytics** from the Nova School of Business and Economics.

Forecasting Weekly Emergency Department Demand in a Portuguese Private Hospital - Generalising to a Medium-sized Unit

Inês Corvo Marques da Fonseca

Work project carried out under the supervision of:

Professor Iolanda Velho

09-01-2023

Abstract

The international Emergency Department (ED) overcrowding crisis affects both private and public Portuguese hospitals, which can be mitigated by an efficient medium-term operational planning. In this light, a Machine Learning multi-step-ahead predictive tool to forecast weekly ED arrivals in the largest unit of a private Portuguese healthcare provider, CUF, was developed. Linear Regression, SARIMAX and LSTM were evaluated and compared. SARIMAX, which obtained the best results, proved to have adequate predictive accuracy to support ED management. Additionally, the question of whether this model could be generalised to a medium-sized CUF ED unit was studied.

Keywords: Healthcare, Emergency Department, Machine Learning, Time Series, Multi-step-ahead Forecasting, Model Generalisation

Acknowledgments

Firstly, I would like to express my gratitude to our advisor, Professor Iolanda Velho, for her relentless guidance and availability throughout this project. In addition, I would also like to thank to Engineer João Leal and Engineer Mariana Raposo, for supporting our understanding of CUF's business operations. Most important, I would like to emphasise the role my fellow colleagues Beatriz Felisberto, Carolina Cotrim and Miguel Alfaro had in the development of this thesis, which was writen under the Field Lab format. Lastly, I would like to acknowledge the fundamental role of my family and friends in my academic journey, who have always encouraged and supported me in all aspects of my life. Without them, this achievement would not have been possible.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Contents

1. Introduction

1.1. Motivation

Hospitals aim to provide proper and quality care services. To do so, they need to well manage their resources, optimise their processes, and align the services provided with demand (Ferreira et al. 2019). The European Action Plan for Strengthening Public Health Capacities acknowledges the Emergency Department (ED) as the most crucial element to increase efficiency and better manage a hospital (Gille et al. 2020). Although each hospital is unique depending on its surrounding context, generally, EDs are facing an increasing patient demand, affluence of non-urgent patients, scarcity of inpatient beds, large delays between patient arrival and treatment, and staff reductions. Additionally, the shutdown of other local EDs exacerbates the situation. Hence, these departments are suffering from what was declared a global health issue: ED overcrowding; and the situation in Portugal is no exception (Graham et al. 2018; Morley et al. 2018).

To better understand the specificities of the Portuguese case, it is important to further examine the Portuguese health structure. There are three systems that form the Portugal Health System: the public sector designated as National Health System (NHS), the health subsystems, and the private voluntary health insurance (Simões et al. 2017).

Data from the Organization for Economic Co-operation and Development shows that at least one in each five ED visits could be solved through other means (Rocha 2020), overcrowding the ED (Simões et al. 2017). This increased affluence to EDs could be explained by the high volume of citizens that are facing serious obstacles when trying to obtain non-urgent care through the NHS. Ideally, patients would be associated with a General Practitioner (GP), commonly called Family Doctor, to easily access non-critical medical appointments and minor surgeries. However, more than three quarts of a million individuals do not have a GP (Portuguese National Health Services

2017), and even those who do, are facing a lengthy process to obtain treatment (Simões et al. 2017). Adding this to the fact that public urgent care cannot legally deny providing treatment and is open 24 hours, patients began going straight to emergency care, even in non-urgent cases (Oliveira 2020). The impact of this issue is vast and it can result in declining patient satisfaction, long waiting times, adverse patient outcomes, and lower quality of care (Graham et al. 2018).

Furthermore, individuals with higher purchasing power shifted to the private sector, driving this sector to grow and diversify its services. In fact, since 2016, the number of Portuguese private hospitals surpasses the number of public hospitals. (Simões et al. 2017; Oliveira 2020). With so many patients opting for this sector, the situation in private EDs became more inefficient. Thus, both public and private hospitals are suffering from overcrowding and are seeking to implement new strategies to ensure the provision of superior services, by foreseeing critical situations (Sebastião et al. 2021).

By being financed by private entities, private hospitals are subject to large investments and lower budget restrictions, leading to greater efficiency in managing resources and infrastructures (Marques et al. 2021). Additionally, they have more means to employ cutting-edge strategies and technologies to enhance their services.

In parallel, there has been a rapid growth in the use of digital platforms and tools, that enable to gather of a tremendous volume of data. This allows for the creation of valuable decision-making tools that leverage insights from data, and that resort to Artificial Intelligence (AI), and more specifically, Machine Learning (ML) (Sarker 2021).

The use of ML has become popular in healthcare due to the growing use of electronic health records (EHR) in hospitals that generate an unprecedented amount of patient data. Specifically, ML applications in healthcare can be segmented into two areas: (i) the use of robotics in surgeries to increase

positive outcomes and decrease complications that may arise; (ii) and the use of algorithms that create a learning process through experience (Becker 2019).

The latter can have several purposes related to improving diagnostics and treatment efficiency, predicting eventual complications, or assisting managers to improve decision-making. More particularly to the ED, concerns such as patient demand forecasting, probability of admission or readmission, patient deterioration and treatment optimization are common research topics (Bates et al. 2014).

An example of a Portuguese healthcare provider that continuously strives for innovation, incorporating recent technologies in its activities, is CUF. The present study was conducted in collaboration with CUF, the current market leader in private healthcare services in Portugal. In 2021, it registered 2,3 million doctor appointments, 63,000 online appointments, and 282,000 urgent care visits, which resulted in an overall net profit of 34,7 million euros (*Relatório Qualidade e Segurança Clínica* 2021).

CUF delivers its services across 19 units, providing care in almost 50 medical and surgical specialities. It has seven hospitals that offer urgent care 24 hours a day and three hospitals operating on a specific schedule. CUF's ED is divided into three services: General Medicine, which consists of all adult medical specialties, Paediatrics, and Obstetrics (only available in two hospitals) (*Relatório Qualidade e Segurança Clínica* 2021).

All CUF EDs, as most urgent care facilities in Portugal, use the Manchester Triage System (MTS) to adequately prioritize patients by their intensity of pain, clinical severity, and level of emergency, consisting in one of the most reliable methods to use in a hospital urgent care setting (Azeredo et al. 2015). The MTS works by allocating the patient's primary complaint to one of the 52 available flowchart diagrams and then, by using key discriminators, each patient is assigned to one of the

triage categories (Santos et al. 2014). The MTS encodes patients into five categories: ”non-urgent” as blue; ”standard” as green; ”urgent” as yellow; ”very urgent” as orange; and ”immediate” as red. It also establishes the maximum time it should take for a patient to be seen by a doctor, ranging between 10 minutes in most urgent cases and 240 minutes in less urgent patients (Azeredo et al. 2015). By using this system, hospitals can allocate patients to the most adequate service, allowing for faster and improved patient treatment and functioning of the emergency departments (Santos et al. 2014).

In terms of data, CUF has an EHR system widespread throughout all 19 units and that gathers both clinical (*e.g.*, lab results), and operational data (*e.g.*, waiting time in urgent care). This generates an opportunity to produce new business insights and operational tools for hospital management. For instance, in 2020, motivated by the COVID restrictions, CUF launched a ML digital symptom evaluator integrated within their mobile app, guaranteeing that patients have access to safe and reliable information and recommendations about their condition and treatment.

1.2. Objectives

Motivated by CUF's culture of innovating to surpass present challenges, and the global ED overcrowding problem, the present project aims to develop a multi-step-ahead forecasting tool that efficiently predicts the number of weekly ED arrivals for one of CUF's units, enhancing the response to ED demand.

Due to the lack of research on the topic in Portugal, this thesis also presents an overview of the international related work.

To accomplish the primary goal, the following intermediate objectives were defined. The first was to identify which were the best performing models found in literature and to choose the most appropriate to employ. The second was to evaluate the performance metrics and computational time

of the selected models in order to reach the most suitable solution. Lastly, a data-driven analysis was elaborated to provide final business insights and recommendations.

1.3. Thesis Outline

The present thesis is structured into eight chapters. Chapter 1. outlines the thesis' Motivation and Objectives. Chapter 2. presents the State of the Art, followed by the Theoretical Definitions relating to the models and metrics used. In Chapter 3., the Research Approach is described. Afterwards, in Chapter 4. relevant business insights are highlighted through an Exploratory Data Analysis. In Chapter 5., all Modelling steps are described, and in Chapter 6. the Results obtained are discussed. Chapter 7. presents the final Conclusions and Business Recommendations, along with the identification of the Limitations, and several suggestions for possible Future Work. Finally, Chapter 8. comprises the Individual Contribution carried out.

2. Theoretical Background

2.1. State of the Art

The ED overcrowding problem, and its respective causes, consequences, and solutions are explored in multiple papers, namely by Asplin et al. 2003 and by Moskop et al. 2009. Despite the lack of research on the implementation of AI to mitigate ED overcrowding in Portugal, internationally the investigation repository is extensive. As it will be possible to see below, the methods applied are immense, ranging from less complex models, such as Regressions, to more sophisticated algorithms, like Neural Networks (NN) and Ensemble models. The influence of weather, calendar and ambient factors on predictive accuracy is also a prevalent matter of discussion among authors. Furthermore, researchers have considered different time granularities when modelling the data, covering hourly, daily, weekly, and monthly, as well as other less common forecasting intervals. From a patient

acuity perspective, numerous papers constrict their subject of research to only low-acuity patients, due to the randomness of urgent ED presentations.

To combat ED overcrowding, hospitals need to maximize their knowledge about patient affluence levels. To do so, it is necessary to anticipate the seasonality adjacent to ED arrivals, as well as comprehend the existence of some unpredictable presentations (Sun et al. 2009; Asheim et al. 2019). As Asheim et al. 2019 highlights, on the one hand, knowing the sheer volume of daily patients enables to elaborate an operational plan in the medium-term, and assign medical staff rotations. On the other hand, understanding the behaviour of hourly ED arrivals in detail allows to further delegate resources in real-time, and aids with tactical planning – activating reserve medical staff, discharging patients, and freeing additional beds.

As a consequence of the complexity of the ED overcrowding problem, various possible solutions have been researched. Some authors focused on triage optimization, such as Raita et al. 2019, who developed an alternative to the Emergency Severity Index (ESI), the prevalent tool in the United States. Others aimed to predict ED patient waiting times to better allocate resources (Kuo et al. 2020; Pak et al. 2021; Benevento et al. 2021), having Pak et al. 2021 targeted only low acuity patients. In addition, Hong et al. 2018, Araz et al. 2019, and Roquette et al. 2020 resorted to classification models to foresee whether or not an ED patient would be admitted.

Narrowing to the scope of ED arrivals, historically, the research on this topic dates to at least the beginning of the millennium (Gul et al. 2020). Batal et al. 2001, using a Step Wise Linear Regression to forecast daily patient volume, focused primarily on the importance of calendar and weather factors in the accuracy of predicted arrivals, concluding that calendar variables were key when building an accurate model, whereas weather factors were not.

Given the variability and specificity of the data being modelled, the impact of calendar variables,

such as day of the week, month of the year, and holidays, in the improvement of model performance diverges between studies. For instance, Hertzum 2017 and McCarthy et al. 2008 found that the hour of the day has a significant impact on performance, while Xu et al. 2013 observed that holiday variables did not influence ED patient arrivals. Additionally, monthly ED demand seems to peak during Winter (Kadri et al. 2017; Almeida et al. 2020; Vollmer et al. 2021), while Summer appears to be the lowest ED demand period, especially August (Almeida et al. 2020, Rocha et al. 2021, Caldas et al. 2022). Regarding weekly patterns, Monday tends to show a rise in the volume of ED arrivals (Xu et al. 2013; Kadri et al. 2014; Jilani et al. 2019), while the weekends show fewer cases (Sudarshan et al. 2021; Vollmer et al. 2021; Caldas et al. 2022).

In parallel, the influence of climatic and ambient factors on model performance is not consensual, as some research corroborates that they improve accuracy (Jones et al. 2008; Sudarshan et al. 2021), while some state the opposite (Sun et al. 2009; Calegari et al. 2016).

Jones et al. 2008 employed a Time Series Regression, and considering time horizons of 1, 7, 14, 21, and 30 days, attained MAPE scores between 8.91% and 9.04%. Furthermore, generally, larger predictive horizons yielded larger MAPEs, which Calegari et al. 2016, and Tuominen et al. 2022 also support. Contrarily, when approaching predictive horizons shorter than 24 hours, there is evidence that these predictions exhibit higher errors, and that the shorter the time interval, the more substantial the error (Hertzum 2017; Whitt et al. 2019; Rocha et al. 2021). For instance, Hertzum 2017 developed both Regression and ARIMA models to forecast the hourly and daily ED arrivals at five Danish hospitals, obtaining MAPEs between 47%-58% for hourly predictions, and 9%-11% for daily predictions. Although the hourly errors obtained are excessive, rounding 50%, they are in line with other results found in the literature (Jones et al. 2009; Boyle et al. 2011; Asheim et al. 2019).

The Autoregressive Intregated Moving Average (ARIMA) algorithms are a family of Time Series (TS) models significantly present throughout the literature. Sun et al. 2009 predicted the daily volume of ED arrivals in a hospital in Singapore stratified by triage category. The optimal ARIMA models obtained MAPEs between 4.8% and 16.9%, emphasising that the lowest MAPE regards all patients independently of the triage category, while the highest concerns the most critical. Moreover, Kam et al. 2010 compared univariate and multivariate Seasonal ARIMA (SARIMA), concluding that the use of the multivariate model overperformed by 0.416%, yielding a MAPE of 7.372%. Additionally, Kadri et al. 2014 assessed the stationarity of the time series and observed its presence. This justified the employment of a daily Autoregressive Moving Average (ARMA) model achieving a R^2 of 0.99 and a RMSE of 0.141. Tuominen et al. 2022 developed an ARIMA with exogenous variables (ARIMAX), resorting to features inspired by Whitt et al. 2019, attaining a MAPE of 6.6%.

Furthermore, Champion et al. 2007 compared ARIMA with Simple Seasonal Exponential Smoothing (SES), another common approach to modelling Time Series. With an R^2 of 0.71 and an RMSE of 3.3 per day, SES was reckoned the optimal model.

Some authors noted that simpler models can have a better performance when modelling daily ED arrivals – Jones et al. 2008 obtained a MAPE of 8.91%-9.04% using Time Series Regression opposed to 9.68%-9.85% using Artificial NN (ANN); Whitt et al. 2019 attained a MAPE of 8.4% with SARIMAX, obtaining worse results with Multilayer Perceptron (MLP); and, concerning weekly ED arrivals, Aladeemy et al. 2016 achieved better performance (MAPE of 4.91%) using SARIMA, when compared to an ANN model.

Harrou et al. 2020 predicted both hourly and daily ED arrivals using NN. The researchers found Variational AutoEncoder (VAE) to be the best model, with MAEs of 0.295 (hourly) and 2.318

(daily) and R^2 scores of 0.949 (hourly) and 0.925 (daily). A study by Sudarshan et al. 2021 concluded that, compared to Convolutional Neural Network (CNN) and Random Forest Regressor (RFR), Long Short Time Memory (LSTM) performs best for both daily and weekly forecasting, exhibiting MAPE values of 8.04% and 8.91%, respectively.

Jilani et al. 2019, exploring an emerging area of ML, conducted one of the first research using Fuzzy Time Series (FTS) in the prediction of daily and weekly ED arrivals. The author obtained MAPE scores between 3% and 3.6%, compared with the 6% to 7% values obtained using ARIMA and NN.

By definition, there is a vastness of Ensemble model combinations, with some having already shown the potential to achieve extraordinary results. Yu et al. 2017 forecasted ED arrivals with granularities of 1, 2 and 3 months, and observed MAPEs of, approximately, 1.0%, 1.2% and 2%, respectively. These results were attained using an Ensemble model of Wavelet Decomposition (WD) and ANN, and resorting to Simple Addition. When predicting daily ED arrivals with an ARIMA-ANN, Yucesan et al. 2020 achieved their best result with a MAPE of 0.49%, closely followed by the 0.92% obtained with ARIMA-LR. These were the lowest MAPE values found in the research examined for the purpose of this thesis.

As previously mentioned, the literature on the prediction of daily ED arrivals in Portuguese hospitals is limited. Carvalho-Silva et al. 2018 distinguish themselves by building an ARIMA model with fairly good results, obtaining a MAPE of 5.92%. The study was performed with data from a hospital in Braga and detected the presence of distinct patterns when examining data from different triage categories and ED services. Additionally, Almeida et al. 2020 performed the first research on Paediatric Emergency Department (PED) arrivals in a hospital in Lisbon and achieved a MAPE value of 10.7%. The authors also concluded that there are extensive differences between the adult

and paediatric patients arriving to the ED. For example, PED visits generally happen due to acute respiratory diseases and trauma, which are ultimately related to seasonal matters, while adult visits are mostly characterized by non-seasonal matters, such as chronic diseases, ageing, and comorbidities. Thus, it is not viable to generalize patient volume levels from adult ED studies to PED services, and consequent management decisions. More recently, Caldas et al. 2022 predicted non-critical patient arrivals in a 4-week interval. Training a multivariate Temporal Fusion Transformer (TFT) model with calendar variables, a MAPE of 5.90% was obtained. It is important to emphasise that the authors disregarded the data related to the COVID-19 period, as it compromised model performance, also corroborated by Maddigan et al. 2022.

Authors focus predominantly on accuracy to evaluate the models, however the trade-off between the model performance and computational time appears to be crucial when choosing the best model for real-time applications (Benevento et al. 2021; Rocha et al. 2021). In this case, the models may need to be retrained multiple times, in a diminished time interval, which consequently leads to greater computational times, and loss of decision-making value. For instance, both Benevento et al. 2021 and Rocha et al. 2021 opted for the model with better balance between performance and computational time, despite achieving higher performance with different models.

2.2. Theoretical Definitions

The current chapter presents the most relevant theoretical concepts mentioned in the literature which were applied in the development of the the present thesis, namely the models and metrics used.

2.2..1 Time Series Forecasting

A Time Series (TS) can be defined as a series of historical measurements of an observable and quantifiable variable, equally spaced throughout time (Bontempi et al. 2013; Maçaira et al. 2018).

When forecasting TS using ML, the general goal is to predict future behaviour from past samples. To do so, a regressor is fitted to historical data (train set) and predicts on new unseen data, which was intentionally holdout (test set), meaning the predictions are obtained out-of-sample (Tashman 2000). To gauge model performance, the evaluation of out-of-sample errors is preferred for two main reasons: (i) the best fit is achieved by maximizing the train performance, therefore in-sample errors are presumably inferior to out-of-sample errors, and (ii) the optimal in-sample fit may not warrant the best future predictions since past behaviours may not endure and new behaviours may arise (Tashman 2000).

Part of the value in forecasting TS is held in the ability to predict multiple periods ahead, called steps. In one-step TS Forecasting, the model predicts a single step into the future, the period immediately after the last training period. In contrast, in muti-step-ahead TS Forecasting, the model predicts a sequence of future periods of the TS. Multi-step-ahead predictions are valuable in business contexts as they allow to estimate farther time horizons in advance, aiding with planning (Wang et al. 2022).

There are two main strategies when creating multistep models: single-output and multi-output modelling, with the main difference between them laying in the type of returned object (Wang et al. 2022). While single-output models return a single scalar prediction, multi-output models return a vector of scalar predictions each referent to a single time period. Within these two umbrella strategies it is worth mentioning the Recursive Single-output (Ben Taieb et al. 2012). The Recursive Single-output strategy resorts to a single model that is recursively trained to predict one-step-ahead, and then uses the predicted value (or computations of the predicted value) as an input feature for the next step (Ben Taieb et al. 2010).

2.2..2 ARIMA, SARIMA and SARIMAX

Autoregressive Integrated Moving Average (ARIMA) is a supervised algorithm that forecasts a target variable resorting to a linear combination of lags. ARIMA models do not require other inputs besides the target values and the respective dates. ARIMA models are constructed based on Autoregressive, *AR(p)*, Integrated, *I(d)*, and Moving Average, *MA(q)*, terms. The AR component is a linear combination of past data points, while the MA component is a linear combination of current and past residuals. The Integrated component is responsible for transforming the TS into stationary data resorting to differentiation. The p, d, and q parameters respectively concern the number of autoregressive terms, the differentiation order, and the number of moving averages (Siami-Namini et al. 2018).

The Box Jenkins Method is a common approach to the application of ARIMA, and considers three main steps: identification, parameter estimation, and diagnostics (Fattah et al. 2018). The first consists of checking whether the data is stationary. If it is stationary, the parameter d is set to 0, and the model simplifies to ARMA. Otherwise, d is the number of differentiation steps needed to reach stationarity. In the parameter estimation phase, the remaining parameters are selected using methods such as the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF), to minimize the residuals. Lastly, the achieved model is evaluated, assessing the accuracy of the forecasts. If the model performance does not meet the established standards, further parameters could be tested, and the process repeated (Siami-Namini et al. 2018; Fattah et al. 2018).

Arisen from ARIMA, there are SARIMA and SARIMAX. SARIMA comprises a seasonal component, being represented by $ARIMA(p,d,q) \times (P,D,Q)_m$, where the first trio refers to the non-seasonal component and the second to the seasonal component. When the training data con-

templates exogenous variables and a seasonal component, the model is denominated SARIMAX (Nasiru et al. 2013).

2.2..3 Linear Regression

Linear Regression (LR) is a supervised algorithm that takes training data as historical examples, and calculates the statistical relationship between the independent variables and the target one, through a linear combination of the input features. Afterwards, given a particular set of new data points, and based on the found relationship, it computes the value of the target variable (Tranmer et al. 2020).

This algorithm can be classified as Simple or Multiple, where the difference lies in the number of independent variables considered – if in presence of only one variable, the algorithm is designated Simple LR, otherwise it is named Multiple LR (Maulud et al. 2020).

2.2..4 Neural Networks

Artificial Neural Networks (ANN) were introduced by Rosenblatt in 1957 (Mondal et al. 2014). Aiming to surmount other models' limitations, intelligent supervised black-box models with low interpretability were created.

An ANN model is a group of artificial neurons, known as nodes, which exchange information replicating the electric communication exchanged by neurons in the human brain.

This architecture comprises input, hidden, and output layers of connected nodes (Kröse et al. 1993). The relation between each pair of connected nodes is given by the weight coefficient which indicates the level of importance of the given relationship in the network (Svozil et al. 1997). The input layer consists of input nodes that receive data and transmits it to one or more hidden layers (Kröse et al. 1993). A hidden layer is a group of nodes in between the input and output layers. It processes the information acquired by resorting to parallel computations, such as applying an activation func-

tion using the weights assigned, and transmitting it to the next layer culminating in the output layer (Kröse et al. 1993). The output layer is the last layer, where the output nodes conduct the final calculations yielding the predicted value (Kröse et al. 1993).

To achieve the best predictions, throughout this interconnection the ANN models are constantly improving the loss score by optimizing the weights of each node, through a process called backpropagation (Chollet 2018). The goal of backpropagation is to iteratively achieve the minimum loss value by spreading backwards the error computed on the output layer into the network, and by adjusting the parameters (Haykin 2004). Moreover, the activation function and the backpropagation technique allow the ANNs to be flexible and adaptable to non-linear data.

Deep Feedforward Network

A Feedforward Neural Network (FNN) is a model characterized by the sequential flow of data from the input layer to the output layer, without any feedback loop existing during the training process. This means that outputs from the model are not fed back into the training process and that, ultimately, the network does not create a cycle (Goodfellow et al. 2016).

The multilayer perceptron (MLP) is a particular class of FNN where each neuron in each layer is linked to every neuron of the next layer (Svozil et al. 1997).

The development of FNNs was a corner stone for the progress of other types of NNs, such as Convolutional Neural Networks and Recurrent Neural Networks (Goodfellow et al. 2016).

Recurrent Neural Networks

Recurrent Neural Networks (RNN) are based on the FNN architecture, being distinctive by maintaining information over time using memory. Due to their aptness to use the internal state to recall previous inputs, RNN combine memorised and current information to compute the output, and transmit it to the next layer (Chujie et al. 2018). Although theoretically the RNN models can accu-

rately predict long-term horizons, in practice, they do not learn precedent information in long time intervals, as a consequence of vanishing and exploding gradient problems (Chujie et al. 2018).

Long Short-Term Memory

The Long Short-Term Memory (LSTM) model is a complex type of RNN, which is suited for long-term predictions using Time Series historical data (Chujie et al. 2018; Sudarshan et al. 2021). Aiming to surpass the RNN's pitfalls, the LSTM model was developed, which merges the ability to make predictions using short and long-term memory (Hochreiter et al. 1997; Zaytar et al. 2016; Chujie et al. 2018).

The LSTM algorithm is composed of associated subnets, also known as memory blocks, with at least one memory cell and three gates (input gate, output gate, and forget gate) (Sudarshan et al. 2021). The input gate selects the volume of information from the preceding layer that is saved in the cell, whereas the output gate controls the portion of data that goes to the following layer (Hochreiter et al. 1997). The forget gate aims to identify whether the information is crucial and should be reminded. It applies a sigmoid function where the value zero means the knowledge collected should be deleted, while an output of one indicates that it should be memorised (Sudarshan et al. 2021). This complex structure allows the model to keep and later read the information in long-term horizons and, consequently, mitigate the gradient problems (Sudarshan et al. 2021).

When tuning an LSTM, there are three hyperparameters: learning rate, batch size, and the number of epochs. The learning rate refers to how rapidly the model learns and whether it converges. Finding the right learning rate is crucial because an immoderately high value leads to divergence, whereas an excessively small value increases significantly how long the model takes to converge (Fang et al. 2005). Moreover, the batch size regulates the number of samples used in each iteration. Larger batch sizes require more memory and are more computationally expensive, contrary

to smaller batch sizes, as they exploit less memory (Masters et al. 2018). Finally, the number of epochs designates the number of iterations considered in the fitting phase. Deciding on the right number of epochs is fundamental, since it affects underfitting/overfitting. Too high of a value can escalate model complexity and wrongfully capture training noise. On the other hand, too low of a value may prevent the model from learning possible patterns (Rafiq et al. 2001). Nevertheless, it is important to emphasise that the LSTM models generally need memory resources and can have high computational times (Hochreiter et al. 1997; Sudarshan et al. 2021).

2.2..5 Performance Metrics

Mean Bias Error

The Mean Bias Error (MBE) is a metric used to evaluate model bias. The MBE averages the difference between the real and the predicted values, as observed in the following formula:

$$MBE = \frac{1}{n}\sum_{t=1}^{n} y_t - \hat{y}_t = \frac{1}{n}\sum_{t=1}^{n} e_t$$

MBE can take both positive and negative values. A positive MBE means that the model is on average over-predicting, *i.e.*, the algorithm is forecasting higher values than the real values, whereas a negative MBE indicates under-prediction.

Mean Absolute Error

The Mean Absolute Error (MAE) is a metric that evaluates model performance, by calculating the average of the prediction errors. It computes the average of the absolute differences between the real target values and the predicted values:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|$$

Since MAE only considers absolute values, the positive and negative scores do not counterbalance each other. Given MAE definition, the lower the MAE score, the better the model performs, indicating that the model adequately fitted the data, and therefore made accurate predictions.

Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) quantifies the forecasting accuracy, as it measures the average deviation between the forecasted and the real target values, independently if said deviation is positive or negative. Being a percentual metric, it allows to compare the performance of two models which were trained and tested with data of different magnitudes. Bellow, it is presented the formula for MAPE:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t}$$

When in presence of zero or close to zero real target values, MAPE results in infinite or undefined values, which is a strong downfall of this metric.

Root Mean Squared Error

The Root Mean Squared Error (RMSE) is an absolute quadratic scoring measure to assess the performance of the model. It expresses the root average of the squared deviation between the real and the predicted values. The RMSE is calculated by the following expression:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

The lower the RMSE value, the better the model performance, meaning that the difference between the real and the forecasted values are minimal. The RMSE is significantly affected by the outliers' values and penalizes larger errors more than minor errors. Consequently, the use of RMSE is adequate when large errors are unwanted.

Coefficient of Determination

The Coefficient of Determination (R^2), also known as R^2, measures the percentage of variance in the dependent variable that can be explained by the variance of independent variables. R^2 ranges from 0 to 1 and is calculated by the following formula:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \overline{y}_t)^2}$$

Note that, RSS stands for the sum of the squares residuals and TSS stands for the total sum of squares. When the R^2 is equal to 1, the model is predicting with 100% accuracy, whilst when the R^2 is equal to 0, the model is not making correct predictions. Thus, a higher R^2 implies a better fit between the model and the data. Note that, this metric may only be used to evaluate linear models.

3. Research Approach

The comprehensive literature review presented was conducted by first defining keywords (Machine Learning, Time Series, Healthcare, Emergency Department) and by prioritizing the reading of articles published in recent years or which are highly cited.

The review allowed to gauge how extensive the ED overcrowding problem is, with researchers conducting analysis on multiple hospitals in several countries. Moreover, it was also possible to conclude that a wide range of models have already been tested and proven to be successful in predicting patient ED demand. By comparing the results present in the literature, the Linear Regression, ARIMAs and LSTM models were chosen to be tested.

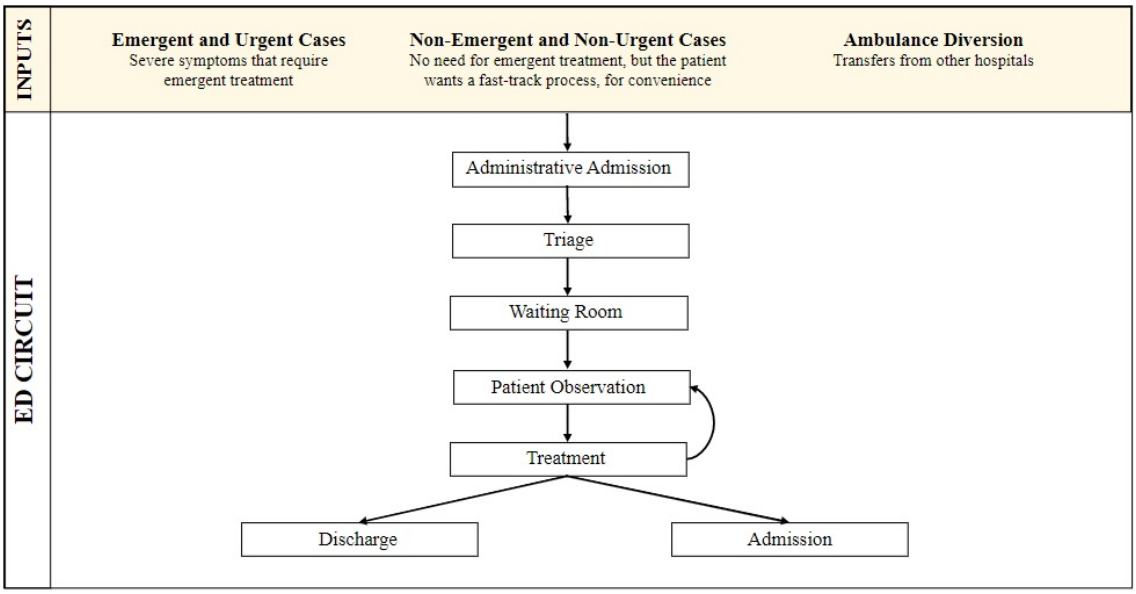


Figure 3..1: ED Patient Flow

After conducting the literature review, and to validate the business problem with CUF's administration, a visit to one of CUF's ED units was carried out. A dialogue with business managers and medical staff allowed to understand the ED day-to-day operations, how the respective resources are allocated, and what are the major challenges experienced in this department. The ED patient flow was comprehended and can be visualised in Figure 3..1. It was in this business understanding phase that the ED unit which would be the focus of this research was chosen.

Afterwards, an Exploratory Data Analysis (EDA) was performed, where the first insights were withdrawn and patterns in the data were identified. Firstly, ED arrivals were characterized, followed by an analysis of patient demand patterns and behaviour.

Subsequently, the Modelling phase was initiated. The approach in this phase began with the addition of external variables regarding weather and COVID. Secondly, the data transformations needed according to each model specificities were identified and performed. The modelling steps were outlined considering the requirements for each model. All models were evaluated based on their fitting time and forecasting ability.

4. Exploratory Data Analysis

4.1. Dataset Introduction

The dataset explored included anonymised information regarding CUF ED arrivals between the 1^{st} of January 2017 and the 30^{th} of September 2022. It was comprised of one main table, where each line represented an arrival to an ED, and 22 tables that contained additional information regarding patients, triage, and medical speciality, among others. For confidentiality reasons, the corresponding relational schema cannot be displayed.

The received dataset comprised adult General Medicine, Paediatrics, and Obstetrics ED services. However, from a business standpoint, the adult General Medicine ED has a higher business relevance and is the only service available in every unit that provides urgent care. Additionally, as previously reviewed, conclusions about a specific ED service functioning may not apply to other services. Thus, the service variable was filtered to only include adult General Medicine.

All pertinent variables were gathered into one single table and analysed in the EDA chapter.

4.2. Choice of ED Unit

The dataset included data recorded in 10 different CUF units. For the model to be as accurate as possible, it was necessary to filter the data to a single unit, because each unit has a particular surrounding context, and consequently may present specific ED patient demand behaviour. From a business perspective, a hospital with a higher volume of ED arrivals translates to a higher business potential. Additionally, a higher affluence of patients may generate significant overcrowding. Therefore, the study focuses on the hospital unit with the highest total number of ED arrivals.

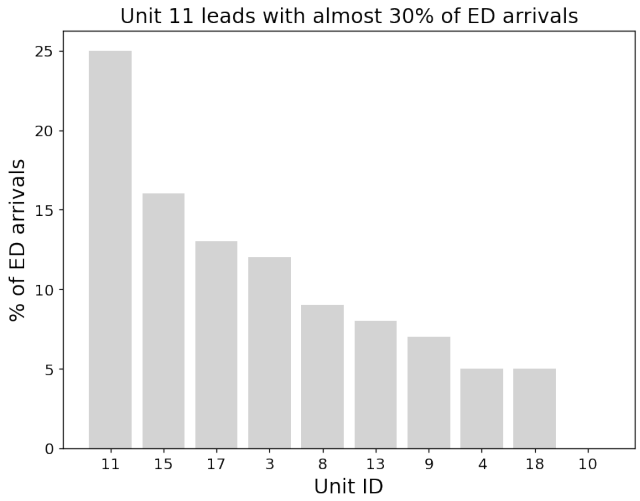


Figure 4..1: Percentage of ED arrivals per CUF unit

It can be observed in Figure 4..1 that unit 11 concentrates over 25% of CUF's ED arrivals, thus being the hospital unit with the highest ED patient volume. Therefore, unit 11 was chosen to be the object of study of this thesis. From here onwards, all analysis will be relative to unit 11.

4.3. Missing Values Analysis

Only 12 out of the 42 variables in the dataset have missing values. Figure A.1 in Appendix shows the number and respective percentage of these missing values. The $especialidade\_medico$ variable, that indicates the medical specialty of the ED doctor, is the one with higher percentage of missing values, 5.79%, followed by $temp\_espera\_max$ with 1.23%. The rest of the variables do not have a significant portion of missing values with all of them presenting less than 1%. Since the EDA is purely informative, no treatment of the missing values was performed.

4.4. Patient Characterization

In this section, the patient demographic will be analysed in detail to fully understand the characteristics of the demand.

Similarly to the distribution of the general population, where 52% of the Portuguese population is female (PORDATA 2022), around 58% of ED arrivals are female patients, which can be noted in

Appendix A.2.

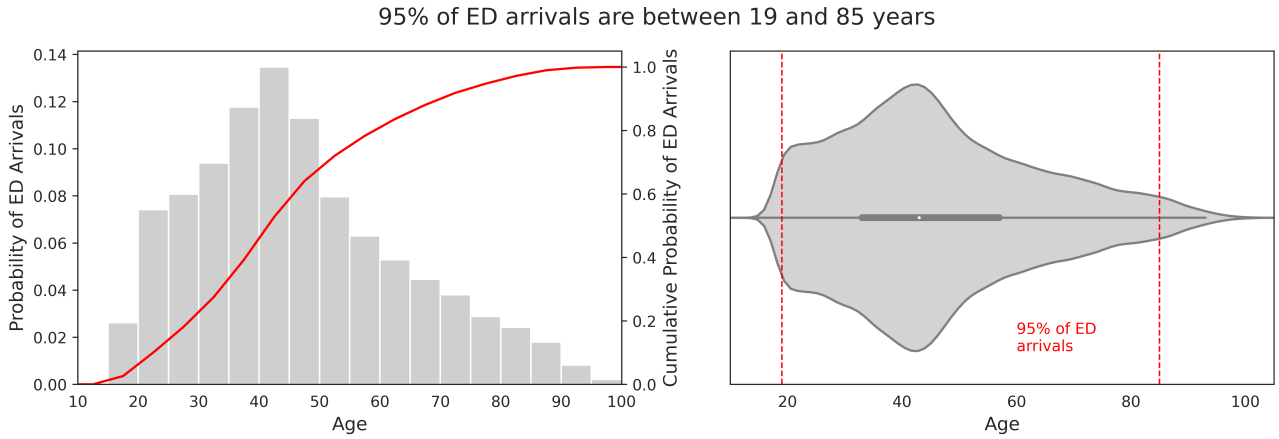


Figure 4..2: Age distribution of ED arrivals

Regarding age, as observed in Figure 4..2, 95% of patients exhibit ages between 19 and 85 years. The mean and median age of arrival are, respectively, 43 and 46 years, and only around 20% of arrivals respect to patients over 60 years old.

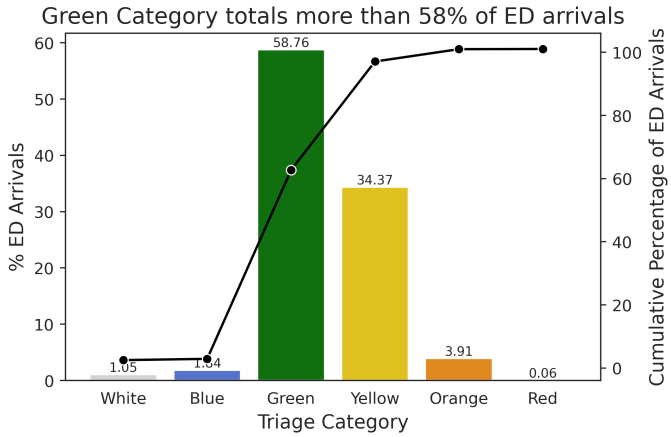


Figure 4..3: Percentage of ED arrivals per triage category

As for the urgency of the patient's condition at arrival, translated by the attributed category at triage (MTS), it can be observed in Figure 4..3 the prevalence of non-urgent cases, represented by the colours *Blue* and *Green*, with these categories summing up to over 60% of arrivals. Note that the *White* category is attributed to patients who need routine assistance with dressing changes and administration of intravenous antibiotics, resorting to the ED only when more adequate services are

not available. Figure A.3 in Appendix shows that *Domicile* is the most common destination after leaving the ED for all levels of urgency. Nevertheless, the *Orange* and *Red* categories have a higher percentage of patients with *Hospitalisation* as destination.

Additionally, close to 92% of patients discharged from the ED have *Domicile* as destination, and around 5% are hospitalized following their ED visit, as seen in Figure A.4 in Appendix. Moreover, as observed in Figure A.5 in Appendix, only a small portion of patients exhibit mobility constraints upon arrival, with less than 2% of patients arriving in a wheelchair, and less than 0.50% arriving on a stretcher.

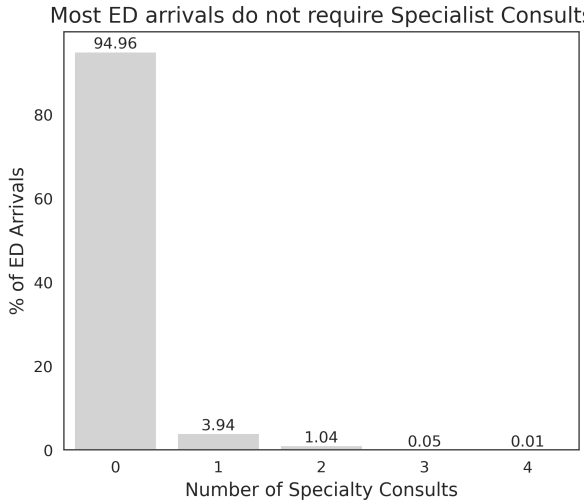


Figure 4..4: Percentage of ED arrivals per number of specialist consults required

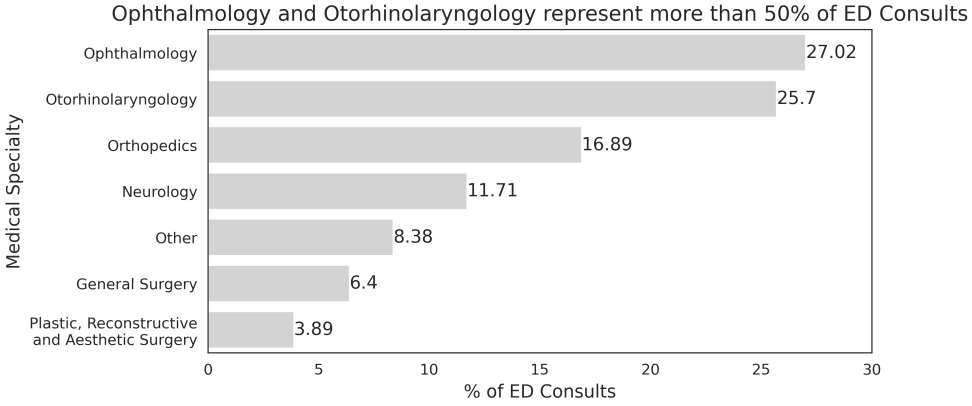
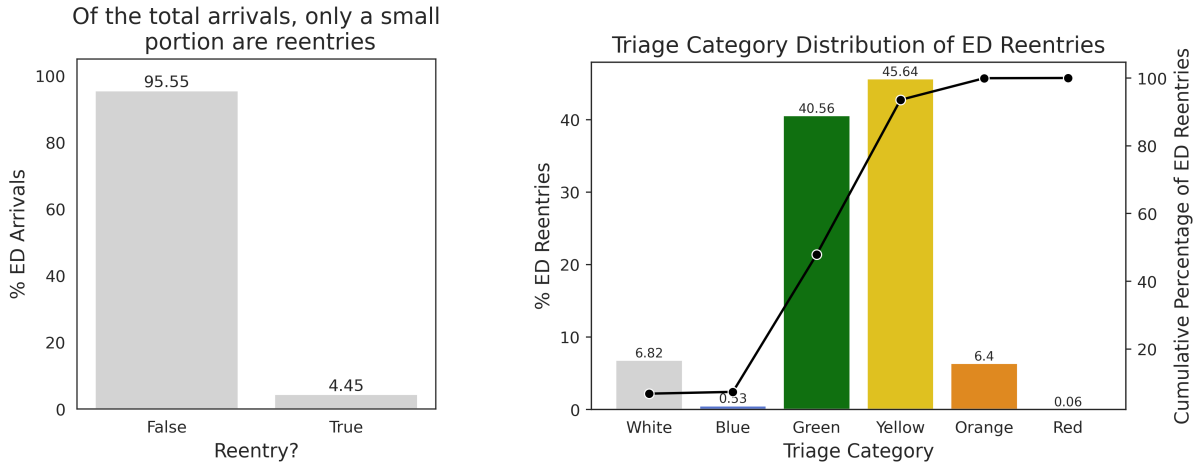


Figure 4..5: Percentage of ED consults per specialty

To provide context, when a patient arrives to the ED, the first observation is in most cases provided by a General and Family Medicine or Internal Medicine ED doctor. For cases that require further specialized medical treatment, one or more specialty consultations may be called for. As observed in Figure 4..4, around 95% of arrivals do not require further specialized urgent care, and close to 4% require only one Specialty Consult. Figure 4..5 evidences the distribution of ED consults per Specialty, where Ophthalmology has the highest number of ED consults with 27%, closely followed by Otorhinolaryngology with approximately 26%.



(a) Percentage of re-entries

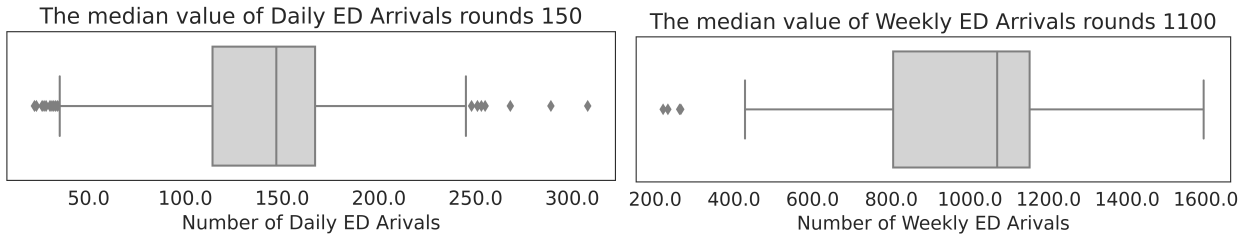
(b) Percentage of ED re-entries per triage category

Figure 4..6: Distribution of re-entries

Following an ED visit or hospitalization, it is not unusual for patients to deteriorate and once again seek medical care, resorting to the ED. Arrivals of patients who return to the ED less than 72 hours after their previous visit are considered ED re-entries. As seen in Figure 4..6, almost 5% of ED arrivals are characterised as re-entries. When studying the triage category distribution of ED re-entries, it is possible to highlight that it is significantly different from the previously observed distribution of ED arrivals. In Figure 4..6, 46% of re-entries are attributed to the *Yellow* category, and there is an increase in the *Orange* category, approximately from 4% to 6%.

4.5. Patient Demand Volume Characterization

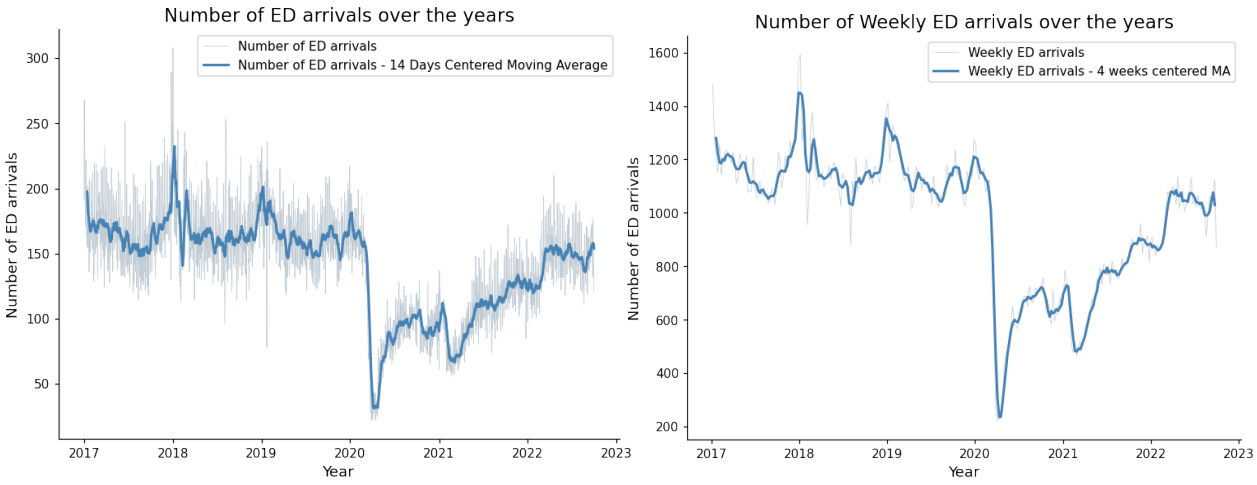
In this section, the patterns in ED patient demand will be identified.



(a) Distribution of daily ED arrivals (b) Distribution of weekly ED arrivals

Figure 4..7: Distribution of ED arrivals

Figure 4..7 illustrates the daily and weekly ED arrivals distributions, both negatively skewed. Regarding daily ED arrivals, 50% of data points lay between 114 and 167, being the median 131 arrivals per day. Additionally, 1.38% of the daily data are considered outliers laying below 35 and above 246. When examining the weekly arrivals, the interquartile range is between 805 and 1152 arrivals, and the median value is registered at 1069 arrivals per week. Outliers are exclusively registered below 428 weekly arrivals, representing 1.67% of the data.



(a) Daily ED arrivals (b) Weekly ED arrivals

Figure 4..8: ED arrivals throughout time

Figure 4..8 demonstrates the number of daily and weekly ED arrivals throughout the years. A spike in ED arrivals can be observed in the beginning of every year, which corresponds to the rougher winter period characterized by the profusion of respiratory diseases (Miller 1973). One exception to this phenomenon is in the year 2020, where a plummet in the ED arrivals was verified due to the spreading of the COVID-19 and subsequent implemented restrictions. Over the years, until 2020, the ED arrivals presented fairly stable values, although a yearly seasonal component can be observed. After 2020, ED arrivals have been increasing, not yet achieving pre-pandemic levels.

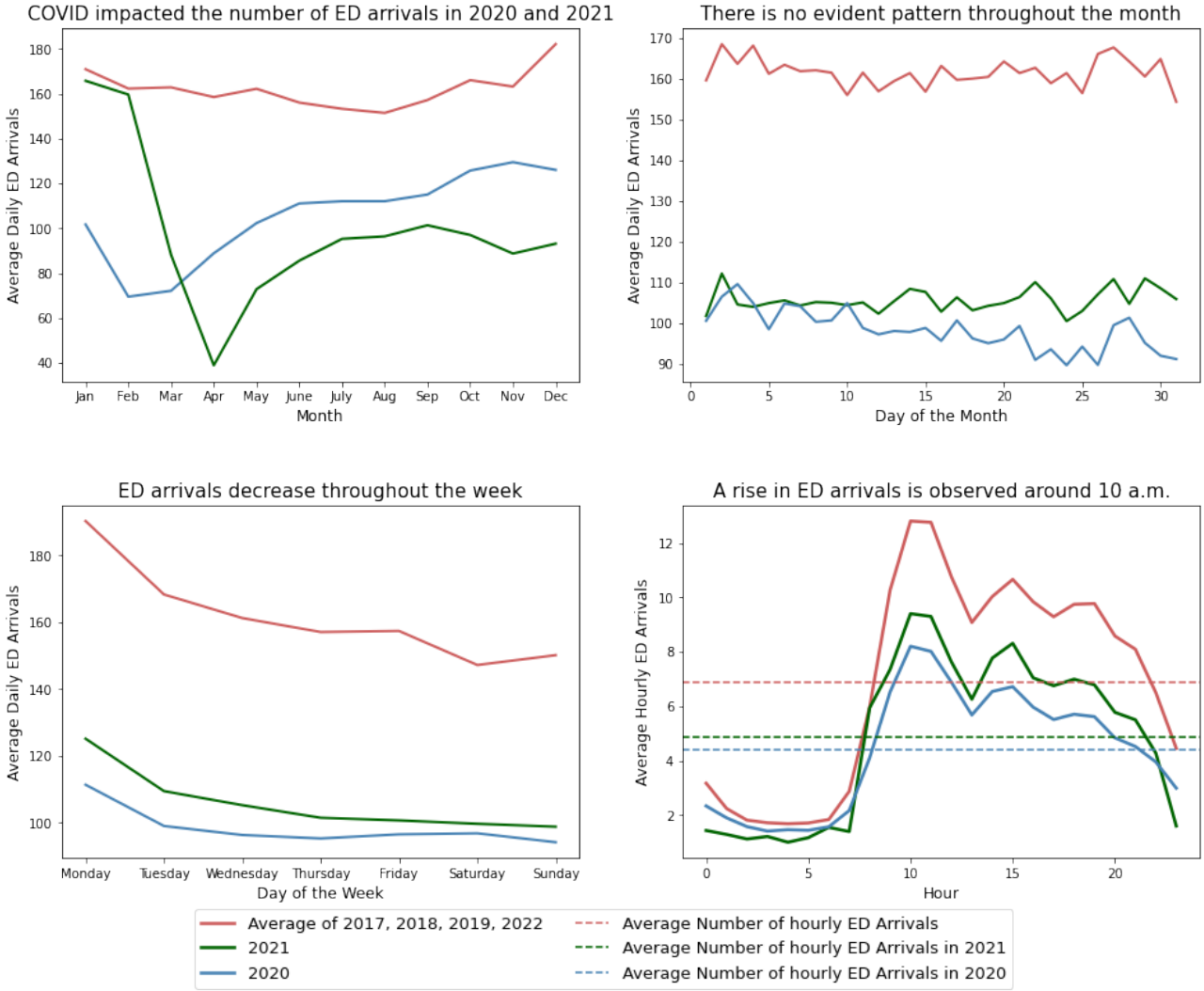


Figure 4..9: Average number of ED arrivals throughout the year, month, week, and day

Considering that the pandemic period, the years 2020 and 2021, might have changed important patterns observed in the ED demand, the historical data from these periods were analysed separately. As Figure 4..9 shows, the main difference between the pandemic and non-pandemic years is in the volume of daily patients recorded, which is considerably lower in 2020 and 2021.

Regarding the average daily ED arrivals per month, there is a drop between March and April 2020. Additionally, in 2021, it is possible to see an increase throughout the year, explained by the easing of COVID-19 restrictions. Moreover, no meaningful patterns stand out when analysing the day of the month.

Focusing on the day of the week, Monday is the day with higher average daily ED arrivals, decreasing for the rest of the week for all the examined time periods. Respecting to the average hourly ED arrivals, there is a noticeable peak in demand between 9 a.m. and 12 p.m., which decreases around midday due to the lunch hour, and increases once again for the afternoon period. Likewise, night periods present the lowest demand throughout the day.

Furthermore, the impact of a holiday on daily ED arrivals can be observed in Figure A.6 in Appendix. From Figure A.6 in Appendix, it is possible to conclude that some national holidays such as *Carnaval*, *Ano Novo*, and *Implantação da República* show a higher number of average daily arrivals than non-holiday days (*NA*). Note that all the mentioned holidays occur in the Fall and Winter periods. These seasons have, in general, higher average daily ED arrivals, which may contribute to the inflation of ED arrivals in these holidays (Figure A.7 in Appendix).

It is also important to analyse if there are any patterns between the severity of the patient illness and other variables, such as time. Consulting Figures A.8, A.9, and A.10 in Appendix it is possible to conclude that there is no pattern between patient severity and season of the year, month of the year and day of the week since the proportions of colours are stable along the given period. For

instance, even though the average daily ED arrivals varies throughout the days of the week, the percentage of each Triage Category does not significantly change. Nevertheless, when conducting the same analysis for the hour of the day, a different conclusion is retrieved.

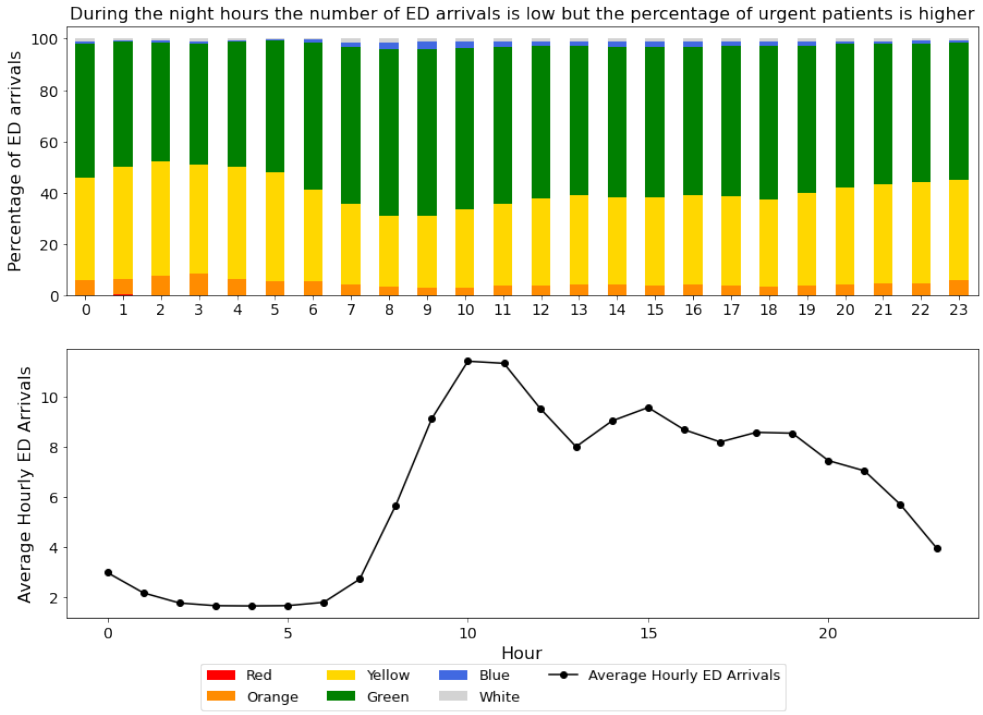
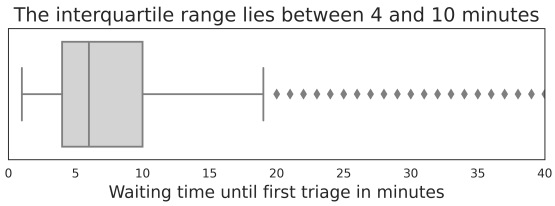


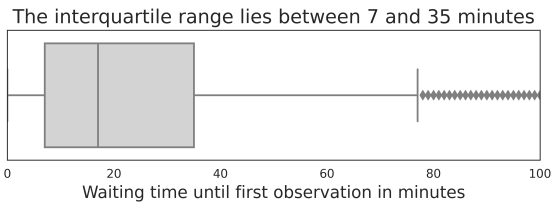
Figure 4..10: Average hourly ED arrivals throughout the day and corresponding triage category distribution

Figure 4..10 shows evidence that the percentage of *Green* arrivals decreases and the percentage of *Yellow* and *Orange* arrivals increases during the night period, when the average hourly ED arrivals reach a minimum. Thus, in less comfortable hours of the day, between 8 p.m. and 6 a.m., at least 40% of patients present urgent conditions.

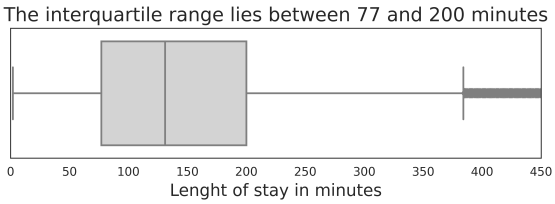
Studying the time patients spend in the ED is also relevant to assess the efficiency of patient flow. The integral distributions of waiting time until first triage, waiting time until first observation, and length of stay can be found in Figures A.11, A.12, A.13 in Appendix.



(a) Distribution of waiting time until first triage - zoomed



(b) Distribution of waiting time until first observation - zoomed



(c) Distribution of length of stay - zoomed

Figure 4..11: Zoomed distributions of waiting times and length of stay

Figures 4..11a, 4..11b, and 4..11c represent the respective zoomed distributions. These distributions have 6.7%, 5.6%, and 4% outliers, respectively, all found in the right tails of the distributions. Moreover, the average waiting time until first triage and first observation are around 8 and 27 minutes, respectively, while the average length of stay in the ED is 166 minutes.

Furthermore, the percentage of patients that wait for more than they should have waited to be seen by a doctor, according to MTS, is 4.6%.

Figure A.14 in Appendix shows the relation between the number of daily ED arrivals and the waiting time until first triage, waiting time until first observation, and length of stay in the ED. Figure A.14a demonstrates a slightly positive correlation between the waiting time until first triage and daily ED arrivals. Additionally, Figure A.14b presents a positive correlation between waiting time until first observation and daily ED arrivals meaning that on a day when the ED receives more patients, those patients will potentially wait longer until being first seen by a doctor. Lastly, Figure A.14c represents a negative correlation between the length of stay in the ED and daily ED arrivals, which implies that when the ED is crowded, patients, on average, stay less time in the ED.

5. Modelling

In this Chapter, the modelling phase is addressed. As an approach to the problem at hand, crucial topics which guided this phase are explored. All computations performed were developed in $Python$. As reflected in the EDA, the presence of abnormal demand behaviour in 2020 and 2021 due to the COVID pandemic was confirmed. Following the ED demand plummet, the number of ED arrivals has been recovering. However, demand remains in a transitional phase, still below pre-COVID levels. This raised the question of whether these periods should be included in the training dataset, since they may have a negative impact on model performance. Thus, three datasets were created – one with the complete time series (*All* dataset), one without the year 2020 (*S20* dataset), and the last without the years 2020 and 2021 (*S20_21* dataset). Since ARIMA models are not able to deal with time gaps, the data of the two latter datasets was shifted forward one and two years, respectively.

Additionally, to develop the multi-step-ahead forecasting models for LR and LSTM, the main challenge was to design a viable strategy that allowed the implementation of the model in a real-world setting. This strategy must respect the timeline in which the input information for future predictions is acquired.

5.1. Choice of Time Granularity Modelled

The feasible planning horizon of ED resource allocation depends not only on the hospital's knowledge about patient affluence levels, but also on its ability to cope with a rapidly changing demand (Asheim et al. 2019). From an operational perspective, planning in medium-term is ideal. On the one hand, planning in short-term reveals to be a difficult task, as there are multiple varying factors susceptible to unpredictability, such as staff availability and unseen punctual events. On the other hand, planning in long-term can be challenging due to the uncertainty associated with estimating

costs and trends for a distant future.

For instance, in CUF, staff rotations are designed on a monthly basis. To aid in scheduling medical staff allocation throughout the month, it is necessary to have an estimate of how ED demand behaves in this period. In parallel, having a prediction of the number of ED arrivals, although aiding in tactical planning, does not facilitate assigning rotas, as adjusting them on such short notice is a difficult task. Thus, having information regarding weekly ED arrivals balances these two scenarios (Choudhury et al. 2020).

5.2. Chain of Operations

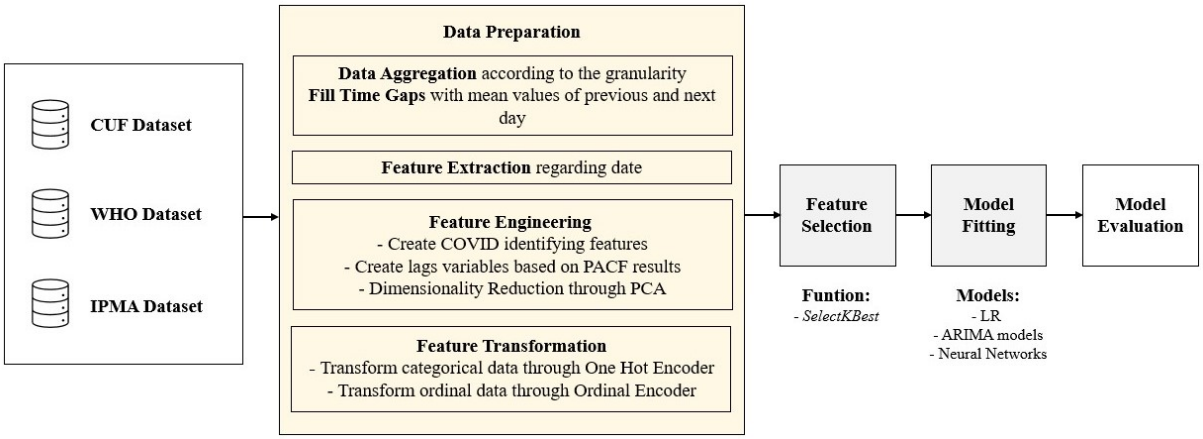


Figure 5..1: Chain of Operations

Building ML pipelines for TS forecasting is a highly complex task, especially for the feature engineering phase (Wang et al. 2022). Moreover, the complexity escalates when modelling multiple steps ahead. After pondering the cost-benefit ratio of developing a pipeline specialised for each model, and since only the best multi-step model will be generalised, the decision was to not contemplate the feature selection and fitting stages in the pipeline. It is worth highlighting that after electing the model with the best multi-step performance, the previous pipeline will be adapted to also fit this model to the train data.

Thus, striving for a smooth model implementation, a series of sequential steps were performed,

creating a workflow which can be easily interpreted and replicated. The diagram of the chain of operations can be consulted in Figure 5..1.

The first stage consisted in loading the three datasets – ED data (CUF), COVID data (WHO), and weather data (IPMA) – to proceed with the Data Preparation stage. This next stage was divided into various steps to prepare and transform the data, so it was suitable to train and test the ML algorithms.

The Data Preparation starts with the aggregation of the data according to the desired granularity, and then it applies feature engineering techniques to fill any existing time gaps. Specifically, the initial dataset, which contained one entry per ED arrival, was aggregated into daily data, by counting the number of ED arrivals per day. In this step, a time gap of two days was found and filled. Note that, the numerical data was created by calculating the mean value of the previous and the next day of each column. Thereafter, for weekly granularity, the data was once again aggregated, by counting the number of ED arrivals per week.

To standardise the number of weeks of each month, it was considered that every month would have 4 weeks, and consequently, each year comprised 48 weeks. This implies that week 1 includes the first 7 days, week 2 includes days from the 8^{th} until the 14^{th} day, week 3 from the 15^{th} until the 21^{st} day, and week 4 the remaining days of the month. Consequently, the number of days present in week 4 is not constant every month. From here onwards, the process applied is analogous independently of the granularity.

In the Feature Extraction step of the pipeline, because LR and LSTM do not handle datetime objects, the relevant date features must be extracted into manageable categorical or numerical columns (Nunno 2014). Also, binary variables that flag whether a certain day is a holiday, business day, or part of a weekend were added.

During the Feature Engineering step, new variables regarding weather and COVID data were created. To encode the impact of COVID restrictions (Assembleia da República 2022), the variables *estado_emergencia* and *estado_calamidade* were built. With data from a World Health Organization (WHO) dataset (World Health Organization 2022) and based on the ratio between the number of COVID deaths and cases, the three waves with higher virus spread were categorically identified (*vagas_covid*). Weather variables regarding past monthly average temperature and precipitation level, gathered from a Portuguese Institute for Sea and Atmosphere (IPMA) dataset, (Portuguese Institute for Sea and Atmosphere 2022) were also included.

Moreover, since LR and LSTM do not intrinsically factor in the behaviour of the dependent variable throughout time, creating other features that contain this information, such as lags of ED arrivals, was the approach employed to attempt to overcome this obstacle (Nunno 2014). These lags were created based on the respective PACF plots, as can be seen in Figure A.15 in Appendix. All lags with a PACF value above 0.05, meaning they are within the 95% confidence interval, were considered influential and therefore added. Furthermore, the mean, median, minimum, and maximum of previous periods were also included.

By definition, the process of lag construction may introduce multicollinearity. To avoid compromising model performance while still retaining the underlying information, a PCA algorithm was employed on the lag features. The number of components was chosen based on the elbow method to capture at least 95% of the explained variance. PCA enables to reduce dimensionality whilst preserving the most significant information, by creating new linearly not correlated components, and discarding the lags. Resorting to PCA, for each one of the datasets All, $S20$, and $S20\_21$, the number of principal components that explained 95% of variance were 13, 16, 21 and were computed based on 47, 55, and 52 lags, respectively (Figures A.16, A.17, A.18 in Appendix).

The data dictionary can be found in Appendix A.19. Here, the lists of variables and their descriptions can be consulted.

After generating the desired features, the categorical and the ordinal variables were encoded through One-Hot and Ordinal Encoders. LR, ARIMAs, and LSTM are intrinsically different models and, consequently, the features that optimize each model are not necessarily the same. Therefore, the Feature Selection stage is specific to each model.

With the data prepared, *TimeSeriesSplit* was used to split the data into train and test sets. The test size consists of 4 weeks, totalling a month, and the training dataset comprises the remaining data. Finally, all models are fitted to the train set, and later predicted out-of-sample. Based on the predictions made by the model, the model performance and Fit Time were evaluated.

5.3. Baseline

Currently, CUF manages its ED resources resorting to a yearly macro analysis. Based on historical data, CUF managers estimate the affluence for the subsequent year. These estimates are sporadically adjusted when medical staff expresses the need to. As 2020 was classified as an atypical year for ED demand, historical data from this period is not accounted for in CUF's estimation.

To compare the performance of the developed ML models with CUF's current method to estimate ED arrivals, a baseline model was established. To replicate the method in place, the baseline prediction averages the mean of the previous month, and the mean of the respective weeks of 2019 and 2021.

5.4. ARIMA, SARIMA and SARIMAX

When modelling ARIMAs, understanding the characteristics of the time series is crucial to find the most suitable model. Accordingly, and based on the Box Jenkins Method, the TS was first assessed in terms of stationarity and seasonality. To visually analyse stationarity, the mean and standard de-

viation of weekly ED arrivals were plotted (Figure A.20), examining whether these two measures remain constant throughout time. Additionally, using the *seasonal_decompose* function, the trend, the seasonal, and the residual elements were plotted (Figure A.22).

To fit the model to the data, the optimal SARIMA hyperparameters must be identified. Since employing a brute force search was not adequate, the *auto_arima* function was chosen to perform this task. *auto_arima* is able to find both non-seasonal and seasonal parameters (*p, d, q, P, D, Q*). However, due to the extensive computational time and memory required to assess all parameters, parameters *d* and *D* were set *a priori* based on the result of the Augmented Dickey Fuller (ADF) test, and the observation of the data differentiation plots with first and second degrees of differentiation (A.21). When not inputted, *auto_arima* sets parameter *m* as 1, meaning there is no seasonal component. Hence, *m* was defined in accordance with the findings obtained in this phase.

Lastly, the *trend* parameter needs to be adjusted (Smith et al. 2017–). To determine the optimal *trend*, an exhaustive cross-validation grid search was applied and the *trend* chosen based on the best model performance. When performing cross validation, for each fold the data was split into train and test sets. SARIMA automatically performs multi-step-ahead forecastings when the predictions are made out-of-sample for multiple time periods.

Using the parameters of the model with the best performance, a SARIMAX was built to study whether adding exogenous variables would improve model performance. Note that the PCA variables were not included as inputs. The purpose behind their construction was to overcome LR's and LSTM's inability to deal with datetime objects, however this is not an obstacle for SARIMA. In fact, SARIMA was designed specifically to model Time Series.

Having recourse to all calendar and weather variables, a multi-step SARIMAX was employed using cross validation. Subsequently, a cross validation grid search was the chosen method to find

the optimal k parameter in the *SelectKBest* function. With only the selected k best features, the SARIMAX was evaluated through cross validation.

5.5. Linear Regression

To assess the impact of COVID data on the quality of the predictions, and choose the most significant features, all three datasets (*All, S20* and *S20_21* datasets), with and without feature selection, were modelled through LR. Later, all six models were evaluated with cross-validation and compared to elect the best model. Onwards, the process of feature selection and fitting the LRs is explained.

It is important to mention that the PCA features were not included in the inputs to avoid leaking future information into preceding one-step predictions. To compute the PCA features it is necessary to have computed all lags beforehand. Thus, the PCAs for a time period t can only be calculated in the t - *1* period, which implies that the use of these features is limited to predicting one-step-ahead. Given the high number of features, two main drawbacks were assessed: whether multicollinearity was being introduced by correlated independent variables, and whether there was overfitting.

Firstly, the Variance Inflated Factor (VIF) was computed to quantify the presence of multicollinearity (Thompson et al. 2017). A high value of VIF indicates the presence of multicollinearity. Regardless, the most commonly used thresholds for variable elimination may not be adequate. It is necessary to contextualise these values based in the different factors that may contribute to instability (O'brien 2007). Therefore, at this point, no variables were removed.

Secondly, a Lasso regularization was contemplated to test whether the model performance would increase. The Lasso method reduces model complexity by shrinking the least important coefficients to zero (Muthukrishnan et al. 2016). Nevertheless, when there are highly correlated features, which was the case, Lasso selects only one variable among them as it does not perform group selection,

discarding possible significant interactions between variables (Zou et al. 2005). Because of this limitation, Lasso revealed not to be the best approach. Instead of Lasso, the *SelectKBest* function was employed to perform feature selection. This function finds the best group of features according to the *k* highest R^2 scores, mitigating Lasso's disadvantage, by taking into consideration the interactions between variables, and dealing with multicollinearity. After filtering out the irrelevant features, all models were evaluated and performances compared.

Based on the one-step model with the best performance, a recursive multi-step-ahead LR was developed. In order to predict multiple steps ahead, it was necessary to ensure that the PCA features did not contain underlying future information. Thus, in each step, the prior prediction substituted the real observed value when recalculating the lags. These updated lags were then transformed into usable PCA features.

5.6. LSTM Model

Proceeding to LSTM, a more complex model, the influence of COVID data in model performance was evaluated using all three datasets (*All, S20* and *S20_21* datasets), with and without feature selection. The PCA variables were not included in the inputs for the same reason as in LR. Afterwards, all six models were compared using cross validation to select the best one-step LSTM. Ahead, the process of data transformation, tuning, feature selection, and fitting is described.

As LSTM has the particularity of requiring the input data to have a three-dimensional shape, several data transformations were done *a priori*. Firstly, the data were scaled using the *StandardScaler* function (Pedregosa et al. 2011), since some variables can assume a high range of values. This transformation allows to enhance the quality of the predictions (D.K. et al. 2019). Secondly, the input was reshaped into three dimensions: (*number of samples, number of time steps, number of features*).

To improve model performance, the learning rate and the batch size were tuned using a cross-validation grid search. The Adam algorithm was the optimizer chosen, since it has been demonstrated to be better and faster when compared to other optimizers, requiring fewer parameters for tuning (Gupta 2021). Moreover, the number of epochs was tuned using *EarlyStopping*, which interrupts model training when the MAE ceases to improve (Chollet et al. 2015). Additionally, the model was fitted to the train set without randomly shuffling the data to maintain chronology.

It is important to note that NN algorithms are not influenced by multicollinearity problems due to their overparameterization (De Veaux et al. 1994). Thus, it was not required to assess this drawback . After filtering out the irrelevant features resorting to the *SelectKBest* function, all models were evaluated, and performances were compared.

During the evaluation of each model, the loss functions of both train and test set were assessed in function of the model complexity, to assess overfitting. The test loss quantifies how well the model predicts new data (generalisation error), whereas the training loss quantifies the train goodness of fit. Ideally, the model should be trained with enough data to minimize the test error. The goal is to balance this trade-off, to neither underfit nor overfit (Ying 2019).

As done in LR, based on the model with the best performance, a recursive multi-step-ahead LSTM was developed. Similarly, to avoid using future information in the inputs and include the PCAs, these values were updated in each step with an analogous process to the one developed for LR. With the introduction of new features, the learning rate hyperparameter was revised. The new value was found through a cross validation grid search.

The main difference between the LR and LSTM multi-step algorithms was the need to scale the numerical variables and reshape the input data into a three-dimensional array before feeding it into the LSTM model. For the predictions to be on the same scale as the real values, the scaling process

was inverted after each step's prediction is obtained, resorting to the *inverse_transform* method (Pedregosa et al. 2011) of *StandardScaler*.

6. Results and Discussion

In this chapter, the results from the modelling phase are analysed and discussed.

6.1. ARIMA, SARIMA and SARIMAX

Firstly, as formerly described in Chapter 5, stationarity and seasonality were analysed. Note that this analysis was replicated for the three used datasets (*All, S20*, and *S20_21*), however, the following analysis concerns the dataset with all years.

Figure A.20 in Appendix shows the evolution of the mean and standard deviation throughout time, with a rolling average of three months. These two statistical measures do not remain constant throughout time, indicating that the data may not be stationary. Moreover, through Figure A.22 in Appendix, it is possible to confirm that the data has a yearly seasonal component, and therefore the model employed must be a SARIMA.

Regarding parameter selection, the results obtained from the ADF test confirm that the data is non-stationary, indicating that parameter d is at least 1. Figure A.21 in Appendix contemplates the results from the All dataset, and suggests that with only one differentiation, the data becomes stationary, which was confirmed by an ADF test. Consequently, d was defined as 1.

With respect to parameter D, the ADF test indicates that the seasonal component of all three datasets was stationary. Hence, D was considered as 0. Figure A.22 showed a yearly seasonality and, since it was defined that there were 4 weeks in a month, totalling 48 weeks in a year, the parameter m was set as 48. The analysis was repeated for all three datasets, yielding identical conclusions regarding parameters d, D, and m.

Data used	Model	***p***	***d***	***q***	***P***	***D***	***Q***
All	SARIMA_{All}	1	1	1	2	0	1
S20	SARIMA_{S20}	0	1	1	2	0	0
S20_21	SARIMA_{S20_21}	5	1	0	1	0	1

Table 6..1: SARIMA optimal parameters

Through *auto_arima*, the remaining hyperparameters that optimise model performance were found, and can be observed in Table 6..1. Furthermore, the optimal parameter $trend$ found for all three datasets was the default value n.

Data used	Model	MAPE	MAE	RMSE	R^2	Std Dev	Fit Time
All	SARIMA_{All}	3.16	33.76	43.62	0.8950	32.66	33.0303
S20	SARIMA_{S20}	3.24	35.58	43.42	0.8814	38.33	7.1748
S20_21	SARIMA_{S20_21}	4.06	44.12	60.76	0.6985	52.54	17.7491

Table 6..2: SARIMA performance

The average predictive performances of these three models, obtained through cross validation, can be found in Table 6..2. In a first analysis, it stands out that $\text{SARIMA}_{\text{S20\_21}}$ performs worse than the other two models. This fact may be explained by the lack of ability of the model to understand the decline of ED arrivals during COVID with the absence of both the years 2020 and 2021.

Both $\text{SARIMA}_{\text{All}}$ and $\text{SARIMA}_{\text{S20}}$ exhibit similar performances, with MAPE scores of 3.16% and 3.24%, respectively. Nevertheless, the first shows a slightly better performance in all metrics used, except for the Fit Time. Since a Fit Time of 33 seconds is not significant, this was considered the best SARIMA model. Therefore, it is possible to conclude that $\text{SARIMA}(1,1,1) \times (2,0,1)_{48}$ is able to forecast weekly ED arrivals with a MAE of 33.76 and a RMSE of 43.62.

Data used	Model	Exogenous features	MAPE	MAE	RMSE	R^2	Std Dev	Fit Time
All	$SARIMA_{All}$	None	3.16	33.76	43.62	0.8950	32.66	33.0303
	$SARIMAX_3$	3	3.03	33.10	40.83	0.8820	35.60	28.9747
	$SARIMAX_4$	4	3.05	33.26	39.70	0.8952	35.68	38.7254
	$SARIMAX_6$	6	3.26	36.92	39.93	0.8796	35.47	36.1489

Table 6..3: Performances of best SARIMA and SARIMAX

Subsequently, the results of the SARIMAX models were evaluated. The results from the three best k values obtained with $SelectKBest$ can be viewed in Table 6..3, alongside with the best SARIMA model. The k value of 3 corresponds to the following features: a binary variable identifying whether the week is the fourth week of the month; the minimum daily temperature registered in the previous month; and the maximum wind speed registered in the previous month, in km/h.

The number of selected features which maximize model performance is considerably low given the total number of columns. When comparing the results displayed in both Tables 6..2 and 6..3, it is possible to conclude that SARIMAX with 3 and 4 exogenous features present slightly higher MAPE, MAE, and RMSE scores. As these metrics alone do not present significant differences, the selection of the best model cannot be made solely based on these performance metrics. Thus, the trade-off between the effort of gathering the weather variables and the increase in performance resulting from the use of these variables needs to be considered.

The test set corresponds to the summer period, so weather variables such as minimum temperature and wind speed might not have as greater impact on the prediction as they might have in more severe winter periods. To safeguard model performance in colder periods, the SARIMAX models with 3 and 4 exogenous variables were considered superior.

When focusing on these two models, although the difference between them is not significant, the SARIMAX with 3 exogenous variables has lower MAPE, MAE, Standard Deviation, and Fit Time.

Thus, SARIMAX$(1,1,1)\times(2,0,1)_{48}$ using the above mentioned 3 exogenous variables, SARIMAX$_3$, was elected as the best model, achieving a MAPE of 3.03%, a MAE of 33.10, and a RMSE of 40.83.

Step	MAPE	MAE	RMSE	Std Dev
t + 1	2.44	25.63	31.63	25.33
t + 2	1.74	17.78	22.23	13.51
t + 3	5.28	53.22	67.16	66.33
t + 4	2.67	35.77	42.78	38.08

Table 6..4: SARIMAX$_3$ average performance of each step

Moreover, in Table 6..4 it is possible to observe the SARIMAX$_3$ average performance per step. As the number of steps increases, it is possible to see differences in model performance, without any pattern being observed. Here, the worst performance is verified in step 3.

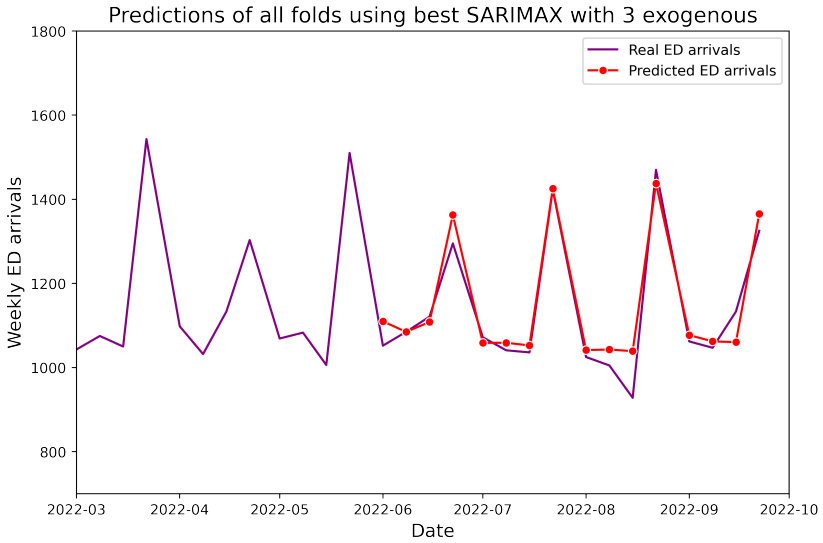
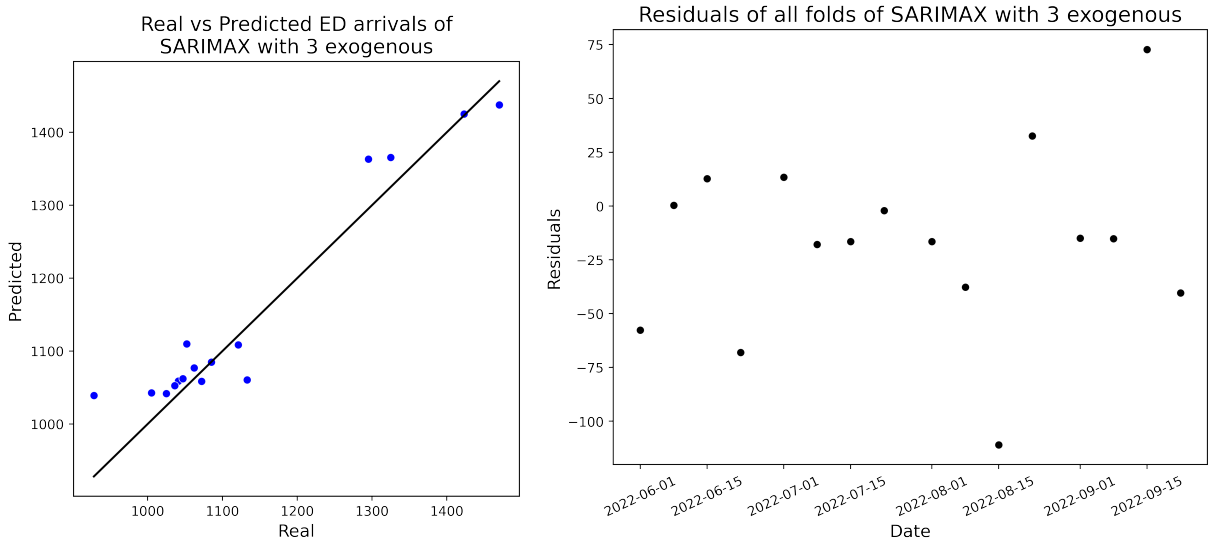


Figure 6..1: SARIMAX$_3$ predictions of all folds



(a) Real vs Predicted ED arrivals of $SARIMAX_3$

(b) Residuals obtained using $SARIMAX_3$

Figure 6..2: $SARIMAX_3$ residuals

The predictions made by this model can be observed in Figure 6..1, alongside the real observations. Moreover, when comparing the real values to the predictions (Figure 6..2), it can be inferred that, generally, prediction values exceed real values. Resorting to the Figure 6..2, no apparent residual pattern can be perceived.

6.2. Linear Regression

Data used	Model	Features selected	MAPE	MAE	RMSE	R^2	Std Dev	Fit Time
All	LR_{All}	All	7.50	85.71	94.42	0.3334	36.42	0.0313
	$LR_{All\text{-}FS}$	15	4.07	48.62	61.46	0.7237	58.59	0.0082
S20	LR_{S20}	All	6.66	74.74	88.68	< 0	48.67	0.0205
	$LR_{S20\text{-}FS}$	12	3.81	45.36	55.99	0.7112	54.30	0.0085
S20_21	LR_{S20_21}	All	9.54	107.50	118.57	< 0	45.83	0.0227
	$LR_{S20\_21\text{-}FS}$	12	7.09	78.14	86.02	0.3390	61.07	0.0088

Table 6..5: Performances of the one-step-ahead LR models

Six variations of LR were modelled, as explained in the previous chapter. In Table 6..5, the one-step-ahead LR results can be observed. The results were obtained through cross validation and

regard the average results of the folds.

As expected, applying feature selection improves model performance for all three datasets. Note that, the Standard Deviation of the errors is the only metric that does not improve with feature selection. All models using the datasets without feature selection exhibit very poor performance, with both LR_{S20} and LR_{S20_21} displaying a negative R^2 score. LR_{All} presents a R^2 of 0.3334, implying that only 33,34% of the variability of the target is explained by the variance of inputs. Despite this score being positive, it is not satisfactory as it falls below the 50% threshold. The R^2 of the $LR_{S20\_21\text{-}FS}$ model is also below 50%, meaning the only two viable candidates for best one-step-ahead LR are $LR_{All\text{-}FS}$ and $LR_{S20\text{-}FS}$ models.

The models using these two datasets, All Data and $S20$, both filtered by the most relevant features, demonstrated similar results. Although the $LR_{S20\text{-}FS}$ has a lower R^2 score of 0.7112, it obtained a lower MAPE of 3.81%, implying that the deficit of less than 0.01 in R^2 score is counterbalanced by the 0.26% MAPE improvement. Moreover, given that all other $LR_{S20\text{-}FS}$ metrics are superior when compared to the metrics of the $LR_{All\text{-}FS}$, $LR_{S20\text{-}FS}$ is elected the best one-step-ahead LR model. The selected features can be observed in Table A.1 in Appendix.

Data used	Model	MAPE	MAE	RMSE	R^2	Std Dev	Fit Time
S20	$LR_{S20\text{-}FS}$	3.81	45.36	55.99	0.7112	54.30	0.0085
	Multi-step LR	3.72	44.07	53.93	0.7326	52.38	2.4361

Table 6..6: Performances of the one-step-ahead and the multi-step-ahead LR models

Step	MAPE	MAE	RMSE	Std Dev
t + 1	1.47	15.39	15.69	29.36
t + 2	1.67	17.69	21.34	35.64
t + 3	5.62	59.92	68.64	85.36
t + 4	6.10	83.28	95.56	76.53

Table 6..7: Multi-step-ahead LR average performance of each step

From the one-step $LR_{S20\text{-}FS}$ model, the multi-step-ahead LR model was constructed. The performance obtained with the multi-step-ahead LR model can be examined in Tables 6..6 and 6..7. Using a multi-step approach improves the predictive accuracy of LR and attains a MAPE of 3.72%, a R^2 of 0.7326 and a Standard Deviation of 52.38. Also, when analysing the average performance of each step, which was obtained through cross validation, as the predictions go further into the future the model performance decreases. This was expected since the recursive approach leads to error accumulation throughout steps. While the first step is predicted with a MAPE of 1.47% and RMSE of 15.69, the fourth step is predicted with a MAPE of 6.10% and RMSE of 95.56. When comparing MAE and RMSE values for each step, it is noticeable that the discrepancy between these values increases in each additional step. RMSE penalises larger error, thus it can be concluded that the model predicts with larger error for steps more distant in the future.

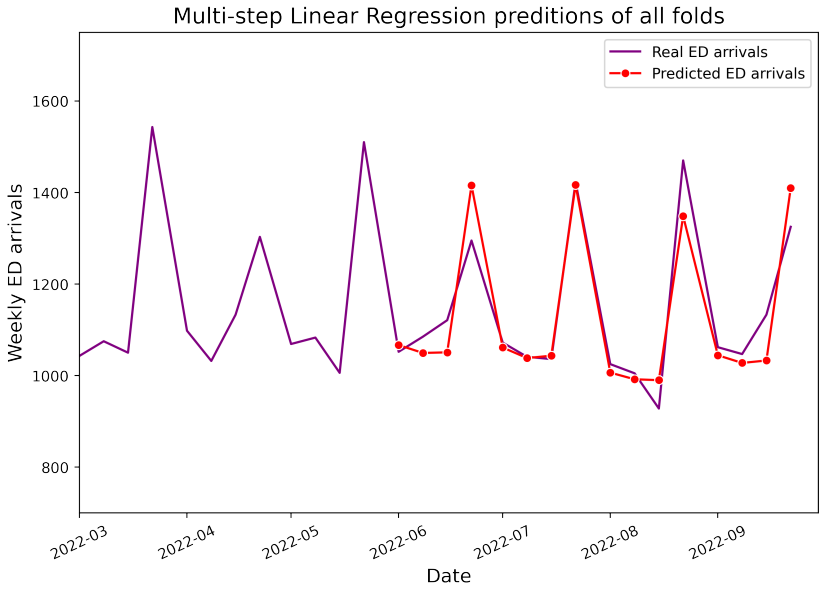
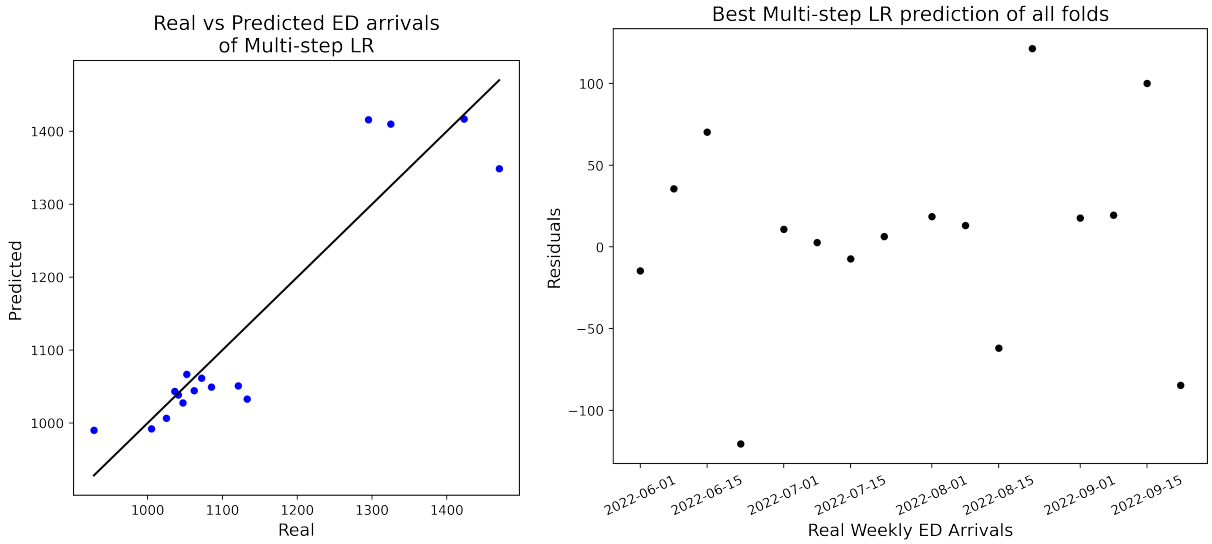


Figure 6..3: Multi-step-ahead LR predictions of all folds



(a) Real vs Predicted ED arrivals of multi-step-ahead LR

(b) Residuals obtained using multi-step-ahead LR

Figure 6..4: Multi-step-ahead LR residuals

The multi-steps-ahead LR predictions throughout time can be observed in Figure 6..3. Lastly, the plots of the multi-steps-ahead predictions and real observations can be found in Figure 6..4. The residual values are also scatter plotted in Figure 6..4. In Figure 6..4 it is possible to identify two distinct collections of data points, the first cluster on the bottom left corner and the second cluster on the right upper corner. Although all points are relatively close to the $y = x$ optimal line, the data points belonging to the second cluster appear to be more dispersed and may be all fourth step predictions.

Additionally, the residuals plot can be observed in Figure 6..4. The residuals seem to be evenly dispersed, which indicates they are close to a normal distribution. Figure 6..4 and the computed MBE value of 7.88 may imply that the model bias is low.

6.3. LSTM

Data used	Model	Features selected	MAPE	MAE	RMSE	Std Dev	Fit Time
All	$LSTM_{All}$	All	4.63	51.22	56.33	49.86	6.3405
	$LSTM_{All\text{-}FS}$	20	4.53	48.70	55.58	46.72	5.4569
S20	$LSTM_{S20\text{-}FS}$	125	4.86	54.63	65.86	62.74	9.9110
S20_21	$LSTM_{S20\_21}$	All	5.76	67.71	81.44	69.53	10.5943
	$LSTM_{S20\_21\text{-}FS}$	65	4.20	46.39	52.01	49.32	12.4465

Table 6..8: Performances of the one-step-ahead LSTM models

As explained in the previous chapter, six one-step-ahead LSTM models were developed. The LSTM one-step-ahead performance results were computed using cross validation, and can be observed in Table 6..8.

The batch size was fine-tuned and set as 32. Moreover, the learning rate was tuned for each dataset. For the All dataset the rate of 0.0040 was optimal, whereas for both $S20$ and $S20\_21$ datasets this parameter took the optimal value of 0.0008.

Regarding whether feature selection was beneficial to LSTM performance, it can be highlighted that, for $S20$, reducing the number of features does not improve the quality of predictions. In this case, the $SelectKBest$ returned 125 as the optimal number of features, which coincides with the total number of columns. Contrarily, for the remaining datasets, All and $S20\_21$, feature selection enhances model performance, by reducing the number of features to 20 and 65, respectively.

Independently of feature selection, all models exhibit a good performance with MAPE values between 4.20% and 5.76%. $LSTM_{All\text{-}FS}$ and $LSTM_{S20\_21\text{-}FS}$ are the two candidates for the best model. These two have a very similar performance with MAPEs of 4.53% and 4.20%, respectively. Besides including different years, the main difference between the models is the number of features selected. Following Occam's Razor, which defends that the simplest explanation is preferable,

$LSTM_{All\text{-}FS}$ was chosen as the best model, due to its lower number of input features.

Data used	Model	MAPE	MAE	RMSE	Std Dev	Fit Time
All	$LSTM_{All\text{-}FS}$	4.53	48.70	55.58	46.72	5.4569
	Multi-step LSTM	3.60	40.73	49.82	183.44	7.2951

Table 6..9: Performances of the one-step-ahead and the multi-step-ahead LSTM models

Step	MAPE	MAE	RMSE	Std Dev
t + 1	3.25	34.02	35.52	25.16
t + 2	1.49	15.15	22.02	33.11
t + 3	5.69	56.46	77.47	84.87
t + 4	4.01	57.27	72.19	75.99

Table 6..10: Multi-step-ahead LSTM average performance of each step

The multi-step-ahead LSTM, built based on the best one-step-ahead model with the 20 best features ($LSTM_{All\text{-}FS}$), slightly improves model performance in all metrics, except for Standard Deviation of errors and Fit Time. Despite the increase in Fit Time being trivial, the discrepancy in the Standard Deviation is not. As for the MAPE, the Multi-step LSTM exhibits 3.60% of forecasting errors. The MAE and RMSE round 40 and 50 weekly ED arrivals, respectively. Regarding the average performance of each step, the results obtained are good, especially of the second step, which acquired a MAPE of 1.49% and RMSE of 22.02. The remaining steps have relatively similar results with MAPEs varying between 3.25% and 5.69%. Additionally, from the third step, the quality of the predictions decreases, achieving MAE and RMSE values above 70 and 55, respectively. When comparing the quality of predictions throughout time, contrarily to the multi-step-ahead LR model, the performance does not decrease as further into the future the steps are.

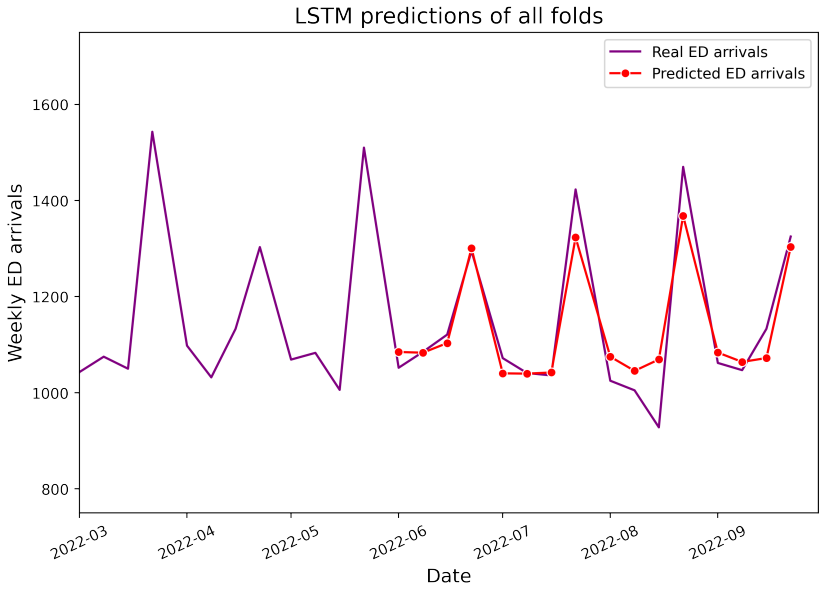
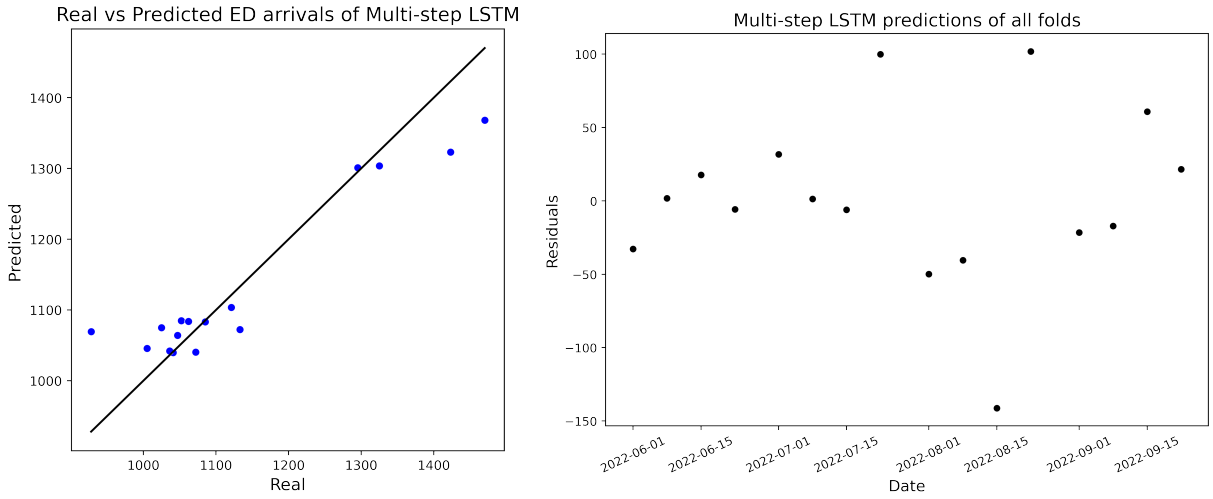


Figure 6..5: Multi-step-ahead LSTM predictions of all folds



(a) Real vs Predicted ED Arrivals of multi-step-ahead LSTM

(b) Residuals obtained using multi-step-ahead LSTM

Figure 6..6: Multi-step-ahead LSTM residuals

Finally, Figure 6..5 exhibits the predicted versus the real number of weekly ED arrivals, while the residual values of all folds are displayed in Figure 6..6. Figure 6..6 demonstrates two well-defined clusters: the first on the bottom left corner and the second on the top right corner. In general, the data points are close to the optimal line demonstrating the good quality of the forecasts. Neverthe-

less, the second cluster presents values distant from the optimum line that may represent the fourth step prediction. Moreover, Figure 6..6 reveals a high dispersion of the residuals, indicating that the model bias is significantly low. In fact, the MBE value of the LSTM multi-step-ahead is 1.31.

6.4. Final Discussion

	MAPE	MAE	RMSE	R^2	Std Dev	Fit Time
Baseline	6.34	73.28	94.59	0.6467	69.21	-
$SARIMAX_3$	3.03	33.10	40.83	0.8820	35.60	28.9747
Multi-step LR	3.72	44.07	53.93	0.7326	52.38	2.4361
Multi-step LSTM	3.60	40.73	49.82	-	183.44	7.2951

Table 6..11: Performances of the best models

Table 6..11 exhibits the results obtained for all best multi-step-ahead models, alongside the baseline values. All developed models show superior performance when compared to the baseline values, reducing MAPE in around 3%.

The performance of the developed models achieved satisfactory results according to the literature review. In fact, the performances registered surpassed some of the performances reported in research on weekly ED arrival forecasting. For example, Calegari et al. 2016 yielded a MAPE of 10.67% using Exponential Smoothing, while Sudarshan et al. 2021 achieved a MAPE value 8.91% through LSTM. Likewise, Aladeemy et al. 2016 resorting to SARIMA achieved 4.91% MAPE.

Additionally, the benchmark MAE and RMSE are substantially higher than the MAE and RMSE of the three developed ML models. Even the worst performing model, the Multi-step LR, significantly improves predictive performance as it outperforms the baseline, on average, in around 30 and 40 weekly ED arrivals, for MAPE and RMSE, respectively.

Although all three models could be employed to accurately predict weekly ED demand, the Multi-

step LR falls behind the other models in terms of performance. Despite 73.26% of the variance in the dependent being explained by the variance of the independent features, the R^2 falls short of expectations.

Comparing $SARIMAX_3$ and Multi-step LSTM, $SARIMAX_3$ outperforms in all metrics. Thus, $SARIMAX_3$ distinguishes itself as the model with the best performance. Nevertheless, even though there was minor fine-tuning, the multi-step LSTM obtained a good performance, showing the potential to yield superior results.

7. Conclusions

7.1. Business Implications and Recommendations

After employing the developed methods and achieving what was considered a good model performance, it is fundamental to understand what knowledge these predictions bring to the business itself, and how they may contribute to ED management.

When using these predictions to support decision-making, there are three scenarios: (i) the prediction corresponds exactly to the observed number of weekly ED arrivals, which is the ideal situation but difficult to obtain; (ii) the prediction is higher than the observed number of weekly ED arrivals or (iii) the predicted value is lower than the observed value of weekly ED arrivals. Each one of these scenarios differently impacts how cost-effective the ED is, needing to be carefully considered to achieve optimal ED management. To understand the impact of the usage of the model as a supportive management tool, the negative and positive consequences of these scenarios are addressed below.

Firstly, the scenario where the model prediction is equal to the real value would be the one with higher benefit, virtually, allowing for a perfectly planning of resources to adequately respond to

patient demand. However, as already concluded, the model is not 100% accurate.

Secondly, the last two scenarios entail that there is a cost associated with managing the ED based on deceiving estimates of weekly ED arrivals. This cost is a function of the ratio between the forecasting error and the real volume of ED demand. As this ratio increases, the associated cost increases as well. For instance, failing to predict 5 ED arrivals on a universe of 100 real ED arrivals has way less impact than failing to predict 50 ED arrivals in that same context. However, the costs of over and underpredicting ED arrivals are rooted in different causes.

In planning, overpredicting may lead to unnecessary higher operational costs as resources are wrongly allocated according to this prediction. For instance, excessive medical staff may be requested to be available, and patients may be moved to other wards to release beds. Maintaining the forecasting error fixed, the overpredicting cost may increase when the observed weekly ED arrivals decrease.

Contrarily, underpredicting may generate ED overcrowding due to the lack of means available to handle the unexpectedly higher volume of patients, hindering the ED flow. Hence, this may result in higher waiting times, patient dissatisfaction and, overall, lower quality of care. Here, the higher the number of weekly ED arrivals, the more costly it is to underpredict by larger margins. This reflects the result of the cumulative effect of ED overcrowding.

To quantify these two costs, it would be necessary to have extensive knowledge on confidential information regarding CUF's business operations. This information was not available for the purpose of this thesis, hence no further analysis on this topic was performed.

To conclude, the main business recommendation suggested is to use the weekly ED arrivals forecast tool to aid in resource planning into three main areas: (i) the ED allocation of sub-contracted doctors, (ii) inpatient bed management and (iii) inclusion of ED predictions in other department's

management.

Having a more fine-grained estimation of the number of ED arrivals per week of the month allows to revise the number of subcontracted doctors needed to fulfil the needs of the ED. Additionally, based on the prediction of weekly ED arrivals and the historical average percentage of patients that require hospitalisation following an ED visit, weekly adjustments to the number of beds available for ED patients can be made. Finally, the ED is usually a gateway to the hospital circuit and consequently to other hospital departments (Choudhury et al. 2020). Therefore, it is wise to weigh the volume of ED demand in the estimation of the number of patients that may be transferred to other departments, such as the Operating Room and the Hospitalisation Ward.

The recommendations for the mentioned areas enable to better manage not only the ED, but the hospital as whole.

7.2. Research Conclusions

Overcrowding is a common problem present in both public and private hospitals, which negatively affects the quality of urgent care. The development of ML tools that aid decision making, may mitigate this global healthcare issue. Thus, using SARIMAX, LR and LSTM models, a predictive tool to forecast ED arrivals was developed.

From a business standpoint, the most pertinent choice was to model weekly ED arrivals, as they allow to plan in medium-term. Furthermore, contrarily to SARIMAX models, which predict multiple steps ahead by default when dealing with out-of-sample predictions, LR and LSTM do not. Hence, a recursive strategy was developed for these two models since there is value in estimating ED demand ahead of time.

Additionally, as ED demand patterns are highly influenced by the context in which the hospital is inserted, the analysis was narrowed to a single ED unit. The unit with the highest ED patient

volume, and consequent business potential, was chosen.

As concluded in the EDA, ED arrival patterns and volume changed after the COVID period. Although the number of ED arrivals in 2022 reflect an increasing tendency, they have not achieved pre-COVID levels. To infer the impact of the pandemic, all models were trained with all years, without 2020, and without 2020 and 2021. The results obtained regarding this topic were contradictory, with two of the models presenting better results when using the entire dataset, and the remaining presenting better performance when excluding data from 2020. These inconsistent results corroborate the hypothesis that 2022 comprises a transitional phase for ED demand.

To possibly improve model performance, additional variables regarding weather and COVID were inputted as external features and proven to be beneficial. Moreover, to mitigate LR and LSTM pitfalls regarding not being able to handle datetime objects, neither capture the variations of the dependent variable throughout time, lags of weekly ED arrivals were created.

Concerning the accuracy of the predictions obtained, the performances of multi-step-ahead SARIMAX, LR and LSTM models exceeded the benchmarks, all attaining MAPEs of around 3%. SARIMAX$(1,1,1) \times (2,0,1)_{48}$ obtained the best results using three exogenous features related to weather and calendar variables. This model achieved a MAPE of 3.03% and an RMSE of 40.83. It is worth highlighting that the multi-step-ahead LSTM obtained a slightly lower performance without extensive fine-tuning of the hyperparameters.

7.3. Limitations

The context of private healthcare is very particular. As this study focused on data from a single hospital unit belonging to a private healthcare provider, one limitation may be the lack of generalisation ability to, not only other units, as well as public hospitals.

Moreover, when constructing the weekly data, the number of ED arrivals in the fourth week are

inflated when compared to the remaining weeks of the month. This may be a limitation as the model may have difficulties in predicting a target variable with disparate values.

Aditionally, the tunning of SARIMA resorted to the function *auto_arima*. This function minimizes the AIC, as *auto_arima* does not accept MAPE as a valid loss function. This may be translated into a limitation since the minimization of AIC does not imply the minimization of MAPE and AIC was not considered when selecting the best model.

Lastly, this research focused on three models, being two of them linear. LSTM, the only non-linear model, requires extensive fine-tuning. Although LSTM hyperparameters were defined without an extensive analysis, it obtained satisfactory results. Hence, the lack of further research on non-linear models may constitute a limitation. Accordingly, some recommendations regarding possible future work are mentioned hereinafter.

7.4. Future Work

Firstly, different feature selection techniques may be studied to assess whether an alternative further improves model performance. For instance, the implementation of an Elastic Net regularization may be considered as it mitigates Lasso regularization's disadvantages.

Furthermore, with the equivalent objective, an extensive analysis of the number of optimal hidden layers, activation function, and dropout rate of LSTM may be conducted. Based on the State of the Art, and concerning the extension of the work developed, it would be interesting to build an Ensemble model combining the two best models found.

This study only focused on General Medicine services, leaving out Paediatrics and Obstetrics. As stated in the literature, conclusions retrieved from a specific service may not be viable to generalise to other services. Thus, further implementation of the referred model may be developed in the future to provide a reliable tool for forecasting ED demand in these services.

In addition, according to CUFs ED managers, exogenous variables, such as public events and football matches, have a high influence on CUF ED arrivals. As it was not possible to include these features in this research, a supplementary investigation could be performed.

To conclude, the present research constructed a tool which predicts weekly ED demand. In order for CUF to be able to benefit from this tool, the next step would be the development of an interface to incorporate this model into CUFs ED operational planning and its deployment.

7.5. Additional Research - Individual Contribution

To complement this research the question of whether the model developed for the large-sized unit may be generalised to medium-sized units was developed and can be found in Chapter 8, which comprises my Individual Contribution for the Work Project.

8. Generalising to a Medium-sized Unit

8.1. Motivation

In more recent years, the research in predictive ML methods has expanded to various health fields (Javaid et al. 2022). Its applications in operational planning as well as in hospital resource management have revealed to be advantageous in the mitigation of the ED overcrowding international crisis and consequent repercussions in patient outcomes (Kadri et al. 2014). With accurate estimates of ED demand, strategic decisions are more informed, and managers may easily make the necessary adjustments so that patient flow in the ED may be optimised (Zhao et al. 2022 and Forero et al. 2011). This motivates the generalisation of some applications to other ED units, in hopes of achieving similar outcomes.

However, since the demand experienced in an ED is modelled from the historical data of that specific ED, the generalisation of these types of ML models may not be guaranteed as site-specific patterns in demand may arise (Ryu et al. 2022). Numerous factors, such as the size of the ED unit, how specialised the services it provides are, which resources it has available, its geographic location, and the encompassing patient demographic, highly influence a unit's ED demand (He et al. 2011).

Moreover, if the generalised model does not accurately predict ED demand, in this case, weekly ED arrivals, the purpose of the generalisation is defeated as it does not provide useful operational insights. Thus, uncertainty of generalisation constitutes a limitation of multiple publications, as researchers usually study a single hospital, or hospitals in a bounded region or country.

In this light and considering that all ten CUF's EDs need to operate efficiently and would benefit from strategic insights, the present Individual Contribution explores the prospect of generalising

the best model obtained in the Group Part (Chapters 1 through 7) to a medium-sized CUF unit.

8.2. ED Unit Choice

When evaluating which unit would be the most appropriate to represent a medium-sized hospital, the average percentages of total ED arrivals registered in each unit were computed.

When disregarding the ED arrivals recorded in units that comprise less than 1% of the total ED arrivals, illustrated by the *other* category, the average percentage was found to be 11.07%.

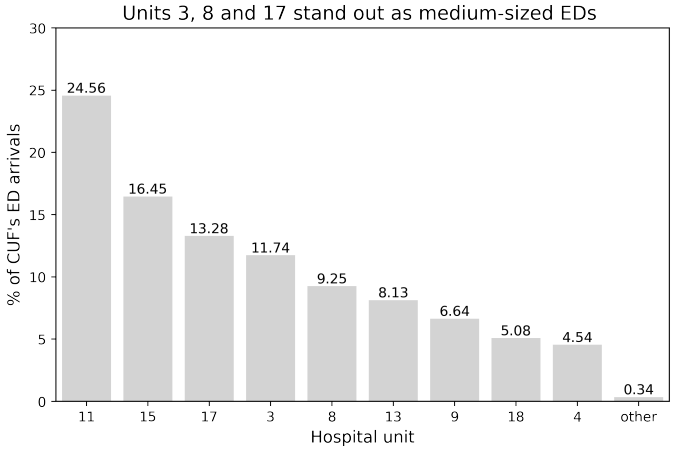


Figure 8..1: Percentage of ED arrivals per ED unit

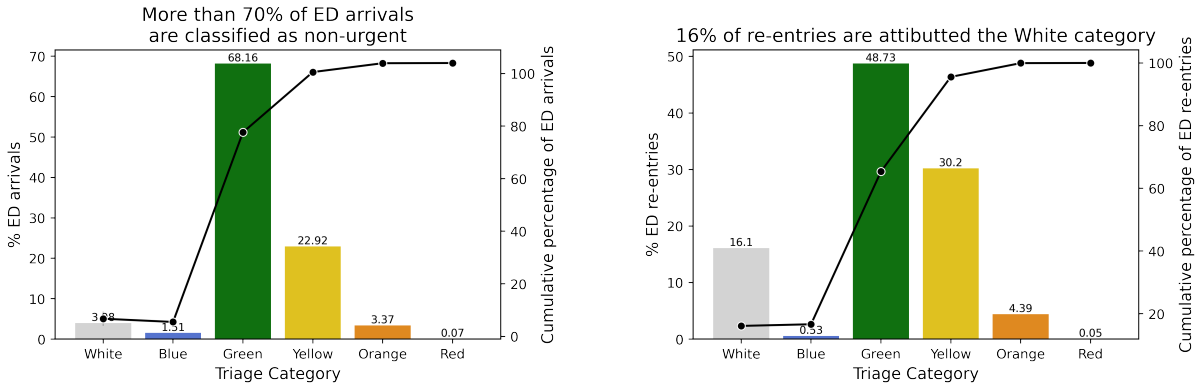
Observing Figure 8..1, units 8, 3 and 17 stand out as the three units with closer percentages to this average exhibiting 9.25%, 11.74% and 13.28% of total ED arrivals, respectively. In terms of geography, unit 3 is situated in the same metropolitan area as unit 11 and, thus, to fathom the isolated effect of ED size in the variations in ED demand, unit 3 was chosen to be the target of the generalisation.

8.3. Exploratory Data Analysis

As mentioned in the introduction, patient characterisation can provide further insight into the ED demand. Thus, the main differences between units 3 and 11, considering both patient and demand volume characterisation, are highlighted below.

8.3..1 Patient Characterisation

Considering age, as observed in Figure A.23 in Appendix, 95% of patients of unit 11 are between the ages of 19 and 88, being the mean and median age recorded 49 and 48 years, respectively. When compared to unit 11, unit 3 exhibits an older patient demographic, with around 20% of arrivals respecting to patient with ages above 68 years, indicating that the population may be slightly aged.



(a) Percentage of ED arrivals per triage category

(b) Percentage of ED re-entries per triage category

Figure 8..2: (a) Unit 3 - Arrivals; (b) Unit 3 - Re-entries

The vast majority of ED arrivals in unit 3 exhibits non-urgent conditions, with around 68% of patient arrivals being attributed the *Green* category, and almost 23% the *Yellow* category (Figure 8..2). In both units the sum of the *Green* and *Yellow* ED arrivals is similar, rounding 91% for unit 3 and 93% for unit 11(Figure 4..3), although the *Yellow* to *Green* ration differs.

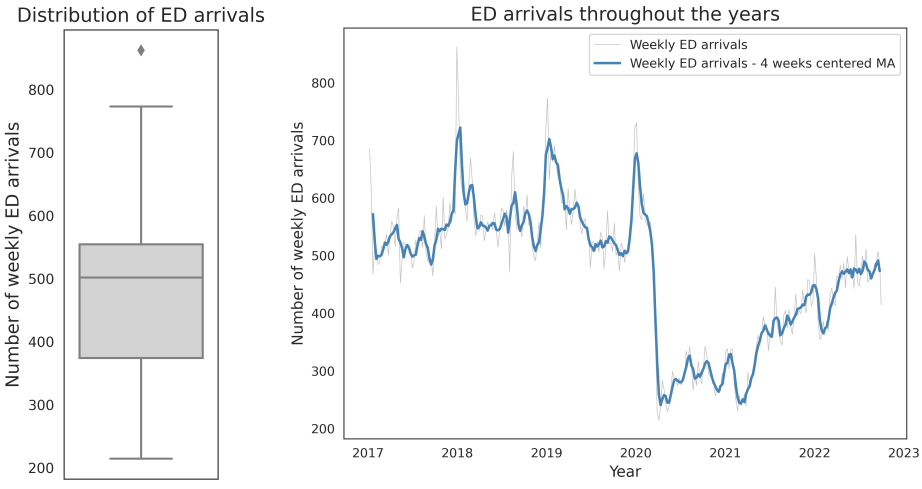
Fig A.24 in Appendix shows that close to 88% of ED arrivals in unit 3 have *Domicile* as destination, being *External Consultation* the second most common destination (around 7%), followed by *Hospitalisation* (almost 3%) and *Transfer* (around 1.50%). When further analysing the destinations per triage categories, in Figure A.25 in Appendix is observable that across all triage categories except for *White*, the percentage of transfers increased compared to unit 11, which may indicate that unit 3 does not have enough resources to respond to the demand, having to transfer some of its patients to

other units. Additionally, across all triage categories, the *Hospitalisation* rate decreases. Moreover, the *External Consultation* discharge destination is unique to unit 3, being the destination of 6.6% *Blue* arrivals, 14% *Red* arrivals, and 4% to 5% *Green*, *Yellow* and *Orange* arrivals.

Lastly, in terms of re-entries (Figure A.26 in Appendix), the percentage of ED arrivals considered as re-entries in unit 3 is 4%. When the distribution of re-entries per triage category in unit 3 (Figure 8..2(b)) is dissected and compared to the same distribution in unit 11 (Figure 4..6(b) in Appendix), major differences in the percentages of the *White* and *Yellow* categories are evident.

While in unit 11, only 6.8% of re-entries were *White*, here in unit 3, this percentage escalates to 16%. Indicating that there is a higher percentage of patients that require systematic care, such as dress changes or the administration of intravenous antibiotics. Contrarily, the percentage of *Yellow* re-entries falls from almost 46% (unit 11) to 30% (unit 3). This may be justified by the smaller dimension of unit 3, as it is possible that patients, when deteriorating and presenting more critical conditions, opt for a larger hospital's ED.

8.3..2 ED Demand Volume Characterisation



(a) Distribution of weekly ED arrivals

(b) Weekly ED arrivals throughout the years

Figure 8..3: Weekly ED arrivals

Figures 8..3(a) and (b) display the univariate distribution of weekly ED arrivals and their evolution throughout the years, respectively. The median value of weekly ED arrivals in unit 3 is 501.5, being the interquartile range between 373 and 554 weekly ED arrivals.

From Fig 8..3(b), the impact of COVID on weekly ED arrivals can be observed. Pre-pandemic, a spike in the weekly ED arrivals at the beginning of each year was evident, however the post-pandemic behaviour has changed, and in recent years a plummet at the beginning of each year can be observed. Comparing to the original data of unit 11 (Figure 4..8), one can establish that the weekly ED demand felt in unit 3 has not recovered to pre-pandemic levels and is increasing at a slower pace than the ED demand of unit 11.

Lastly, to evaluate the patient flow in the ED, the distributions of the waiting times and length of stay were scrutinised. Figures A.27, A.28 and A.29 in Appendix show the zoomed distributions of waiting time until first triage, waiting time until first observation and length of stay, respectively. For all distributions, the medians exhibit lower values than the corresponding medians recorded in unit 11, which may be a sign of overall lower patient volumes, more efficient ED patient flow or lower ED overcrowding.

8.4. Modelling

Three consecutive steps were taken in the modelling stage and are here described: firstly, generalising the previously obtained best model for the data of unit 3, then building a new SARIMA with retuned parameters, and finally including exogenous variables to construct a new SARIMAX.

Firstly, the SARIMAX$(1,1,1) \times (2,0,1)_{48}$ model with 3 exogenous variables previously obtained was fitted to the new data and its performance was evaluated.

Secondly, to understand if there was a more adequate set of hyperparameters to model the weekly

ED arrivals of unit 3, a new SARIMA was tuned. To do so, an analogous process to the one depicted in Section 5.4. was followed. This process began with the visual assessment of the data's stationarity and seasonality through Figures A.30, A.31 and A.32 in Appendix. As the 12-week rolling mean and standard deviation of the time series vary throughout time, one can establish the first evidence to support the non-stationarity hypothesis.

Additionally, through the observation of the multiplicative seasonal decomposition plot (Figure A.31 in Appendix), the existence of a yearly seasonal component was evident, thus *m* was set to 48, as one year of data comprises 48 weeks (12 months with only 4 weeks each).

Optimal parameters for:	*p*	*d*	*q*	*P*	*D*	*Q*	*M*
unit 11	1	1	1	2	0	1	48
unit 3	0	1	1	1	0	1	48

Table 8..1: SARIMA optimal parameters

The suspicions regarding parameters d and D were confirmed resorting to the ADF test and through the observation of the plots of 1st and 2nd order of differentiation (Figure A.32 in Appendix). To find the remaining hyperparameters, alike what was performed for unit 11's data, $auto\_arima$ was employed. The optimal trend parameter remained as *'n'* and all hyperparameters can be consulted in Table 8..1.

The last step was to develop the retuned SARIMAX. Therefore, feature selection was performed through a cross-validation grid search employing $SelectKBest$. Based on the MAPE obtained for each *k* value, the optimal number of exogenous features was found to be *k* = 12. The list of these exogenous features can be found in Table A.2 in Appendix.

8.5. Results and Discussion

Onwards, the results obtained by each model fitted to the data of unit 3 will be presented and discussed.

Model	MAPE	MAE	RMSE	R^2	Std Dev	Fit Time
$SARIMAX_3$	3.03	33.10	40.83	0.8820	35.60	28.9747

Table 8..2: Unit 11 Best Model

To serve as point of comparison, the results achieved previously and described in Chapter 6. are presented in Table 8..2. Note that these results were obtained using the data from unit 11.

Model	MAPE	MAE	RMSE	R^2	Std Dev	Fit Time
$Baseline_{unit3}$	10.56	54.27	61.15	0.3821	42.32	-
$SARIMAX_{generalised}$	4.35	21.64	26.50	0.6421	21.35	61.2319
$SARIMA_{retuned}$	5.43	26.94	31.72	0.5650	27.09	6.2302
$SARIMAX_{retuned}$	4.27	21.47	25.37	0.6801	22.27	16.7721

Table 8..3: Performances of best SARIMA and SARIMAX - Unit 3

Table 8..3 displays the performances obtained for the macro-analysis baseline and for each one of the tested models: the $SARIMAX_3$ generalised to unit 3 ($SARIMAX_{generalised}$), the retuned SARIMA, and the retuned SARIMAX with twelve exogeneous features.

Firstly, when analysing the overall performance, the baseline is clearly inferior, meaning that the SARIMA and SARIMAX predictions yielded more accurate results than the method currently in place in unit 3. Thus, in theory, all three ML models would aid unit 3's managers in their strategic decisions regarding the ED resource allocation.

As for the generalisation approach, when comparing the $SARIMAX_3$ and $SARIMAX_{generalised}$, the

only comparable performance metrics are MAPE and R^2 since the models were trained and tested in two different datasets of unequal scales. Between the two models, $\text{SARIMAX}_{\text{generalised}}$ obtained a worse MAPE, which was expected as the parameters used were optimised resorting to the data of unit 11. Nevertheless, the performance of the $\text{SARIMAX}_{\text{generalised}}$ revealed to be satisfactory, with a MAPE of 4.35%, a RMSE of 26.50 and a R^2 of 0.6421. Note that, although the performance of $\text{SARIMAX}_{\text{generalised}}$ is good and could provide accurate predictions of weekly ED arrivals, there was the possibility of obtaining even better predictive performances with the retuning of the hyper-parameters. Thus, $\text{SARIMA}_{\text{retuned}}$ and $\text{SARIMAX}_{\text{retuned}}$ were constructed.

Moving forward to the performances of the retuned models, $\text{SARIMA}_{\text{retuned}}$ performed slightly worse than $\text{SARIMAX}_{\text{retuned}}$, yielding a MAPE value of 5.43% and R^2 score of 0.5650. The performance of the retuned SARIMA, $\text{SARIMA}_{\text{retuned}}$, is also worse than the performance obtained with the generalised model, $\text{SARIMAX}_{\text{generalised}}$.

When introducing exogenous features, all performance metrics improve besides the fit time. With recourse to the twelve exogenous features, the $\text{SARIMAX}_{\text{retuned}}$ achieves a MAPE value of 4.27%, RMSE of 25.47 and a R^2 score of 0.6880, being the best in modelling unit 3's weekly ED demand.

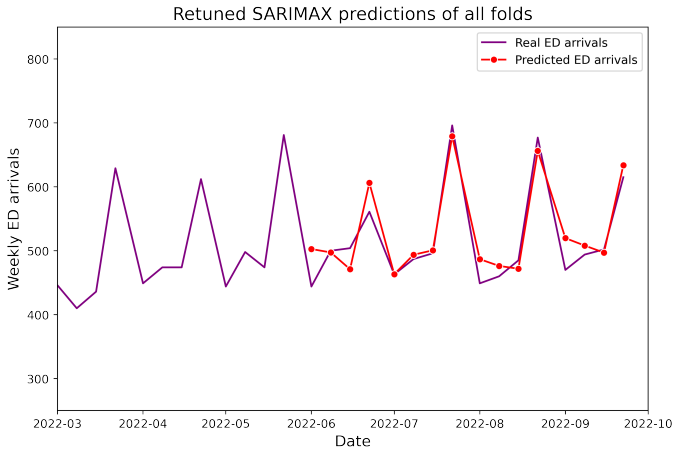
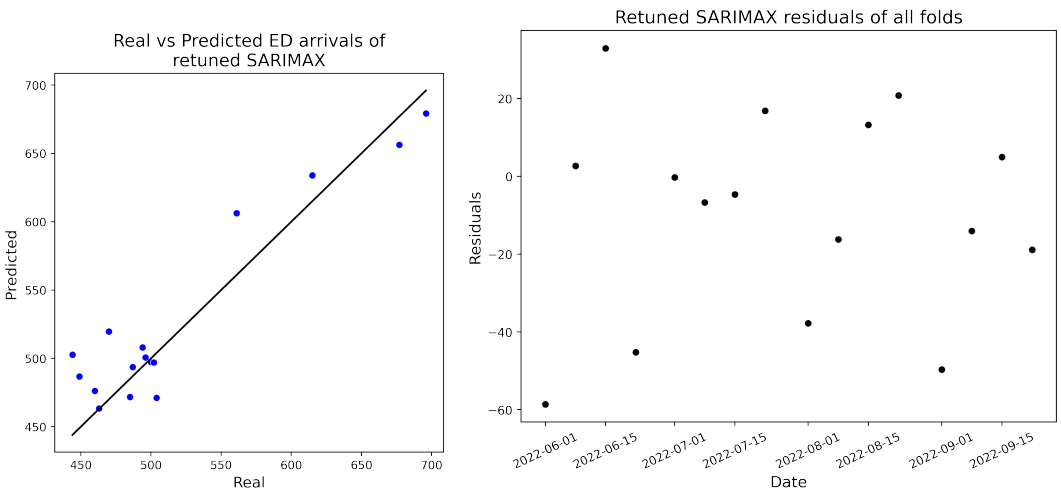


Figure 8..4: Best $\text{SARIMAX}_{\text{retuned}}$ Weekly Predictions of All folds



(a) Real vs Predicted ED Arrivals of $\text{SARIMAX}_{\text{retuned}}$ Best Model

(b) $\text{SARIMAX}_{\text{retuned}}$ obtained residuals

Figure 8..5: $\text{SARIMAX}_{\text{retuned}}$ Best Model

The relevant $\text{SARIMAX}_{\text{retuned}}$ plots can be observed in Figures 8..4 and 8..5. Additionally, Table A.33 in Appendix presents the average performance of each step. In Figure 8..5 the residuals seem normally distributed meaning there is low bias. The MBE obtained rounded -10.06, meaning on average the model overpredicts. Strangely, the first step is the one with highest MAPE, MAE and RMSE, followed by the fourth step. Note that, although the $\text{SARIMAX}_{\text{retuned}}$ attained the best performance the generalisation model $\text{SARIMAX}_{\text{generalised}}$ did not fall behind by a large margin.

8.6. Conclusions

This Individual Contribution was motivated by the question of whether a model developed for a large-sized CUF unit could be applied to a medium-sized CUF unit and still attain good results. To explore this generalisation concept, firstly it was necessary to comprehend the main influencing factors of ED demand. The size of the hospital unit; how specialised the services it provides are; which resources it has available; its geographic location; and the patient demographic it encompasses stood out as determining factors. However, isolating and quantifying the impact of each one

of these factors is a challenging task.

In specific, when comparing units 3 and 11, the patient demographics are similar. The main found discrepancy was the possibility of unit 3's patients being in general slightly older. Additionally, the urgency of the conditions these patients presented was also lower.

To test the generalisation hypothesis formulated, the model which best predicted unit 11's ED demand, SARIMAX$(1,1,1) \times (2,0,1)_{48}$ with 3 exogenous variables, was fitted to unit 3's data and assessed out-of-sample. Although $\text{SARIMAX}_{\text{generalised}}$ obtained a satisfactory MAPE of 4.35%, it did not quite achieve the performance verified for unit 11 (3.03% MAPE). This was expected, as the parameters used were the optimal parameters for the data of unit 11, and not unit 3. Thus, a new SARIMA was tuned to model the weekly ED arrivals of unit 3. By introducing twelve exogenous variables, the retuned SARIMAX$(0,1,1) \times (1,0,1)_{48}$ revealed to be the best model, with a MAPE value of 4.27%.

From a business perspective, both the simple generalisation ($\text{SARIMAX}_{\text{generalised}}$) and retuned model ($\text{SARIMAX}_{\text{retuned}}$) could be employed to accurately predict weekly ED arrivals for unit 3. Both models would aid in equal degrees the operational and strategic planning of the ED, meaning that it would be possible to generalised the best model obtained for the large-sized unit 11, to the medium-sized unit 3 and still achieve accurate results.

However, regarding the more general question of model generalisation, this research does not gather sufficient evidence to state that a model obtained for a large-sized CUF unit can be generalised to any medium-sized CUF unit. To answer this question, additional EDs would have to be analysed and the contributing factors for each unit's ED demand would need to be further studied.

References

Aladeemy, Mohammed, Chun-An Chou, Xiaojun Shan, Mohammad Khasawneh, Srikanth Poranki, and Krishnaswami Srihari. 2016. "Forecasting Daily Patient Arrivals at Emergency Rooms: A Comparative Study." May.

Almeida, Helena, Margarida Sousa, Maria Mascarenhas, Ana Russo, Manuel Barrento, Manuel Mendes, Paulo Nogueira, and Ricardo Trigo. 2020. "The Dynamics of Patient Visits to a Public Hospital Pediatric Emergency Department: A Time-Series Model." *Pediatric emergency care* Publish Ahead of Print (September). https://doi.org/10.1097/PEC.0000000000002235.

Araz, Ozgur M., David Olson, and Adrian Ramirez-Nafarrate. 2019. "Predictive analytics for hospital admissions from the emergency department using triage information." *International Journal of Production Economics* 208:199–207. ISSN: 0925-5273. https://doi.org/https://doi.org/10.1016/j.ijpe.2018.11.024.

Asheim, Andreas, Lars P. Bache-Wiig Bjørnsen, Lars E. Næss-Pleym, Oddvar Uleberg, Jostein Dale, and Sara M. Nilsen. 2019. "Real-time forecasting of emergency department arrivals using prehospital data." *BMC Emergency Medicine* 19, no. 1 (August): 42. ISSN: 1471-227X. https://doi.org/10.1186/s12873-019-0256-z.

Asplin, Brent R., David J. Magid, Karin V. Rhodes, Leif I. Solberg, Nicole Lurie, and Carlos A. Camargo. 2003. "A conceptual model of emergency department crowding." *Annals of Emergency Medicine* 42 (2): 173–180. ISSN: 0196-0644. https://doi.org/https://doi.org/10.1067/mem.2003.302.

Assembleia da República. 2022. *Estado de emergência — Declarações e Relatórios,* December. https://www.parlamento.pt/Paginas/estado-emergencia.aspx.

Azeredo, Thereza Raquel Machado, Helisamara Mota Guedes, Ricardo Alexandre Rebelo de Almeida, Tânia Couto Machado Chianca, and José Carlos Amado Martins. 2015. "Efficacy of the Manchester Triage System: a systematic review." *International Emergency Nursing* 23 (2): 47–52. ISSN: 1755-599X. https://doi.org/https://doi.org/10.1016/j.ienj.2014.06.001.

Batal, Holly, Jeff Tench, Sean McMillan, Jill Adams, and Phillip S. Mehler. 2001. "Predicting Patient Visits to an Urgent Care Clinic Using Calendar Variables." *Academic Emergency Medicine* 8 (1): 48–53. https://doi.org/https://doi.org/10.1111/j.1553-2712.2001.tb00550.x.

Bates, David, Suchi Saria, Lucila Ohno-Machado, and Anand Shah. 2014. "Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients." *Health affairs (Project Hope)* 33 (July): 1123–31. https://doi.org/10.1377/hlthaff.2014.0041.

Becker, Aliza. 2019. "Artificial intelligence in medicine: What is it doing for us today?" *Health Policy and Technology* 8 (2): 198–205. ISSN: 2211-8837. https://doi.org/https://doi.org/10.1016/j.hlpt.2019.03.004.

Ben Taieb, Souhaib, Gianluca Bontempi, Amir F. Atiya, and Antti Sorjamaa. 2012. "A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition." *Expert Systems with Applications* 39 (8): 7067–7083. ISSN: 0957-4174. https://doi.org/https://doi.org/10.1016/j.eswa.2012.01.039.

Ben Taieb, Souhaib, Antti Sorjamaa, and Gianluca Bontempi. 2010. "Multiple-output modeling for multi-step-ahead time series forecasting." Subspace Learning / Selected papers from the European Symposium on Time Series Prediction, *Neurocomputing* 73 (10): 1950–1957. ISSN: 0925-2312. https://doi.org/https://doi.org/10.1016/j.neucom.2009.11.030.

Benevento, Elisabetta, Davide Aloini, and Nunzia Squicciarini. 2021. "Towards a real-time prediction of waiting times in emergency departments: A comparative analysis of machine learning techniques." *International Journal of Forecasting,* ISSN: 0169-2070. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.10.006.

Bontempi, Gianluca, Souhaib Ben Taieb, and Yann-Aël Le Borgne. 2013. "Machine Learning Strategies for Time Series Forecasting." In *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures,* edited by Marie-Aude Aufaure and Esteban Zimányi, 62–77. Berlin, Heidelberg: Springer Berlin Heidelberg.

Boyle, Justin, Melanie Jessup, Julia Crilly, David Green, James Lind, Marianne Wallis, Peter Miller, and Gerard Fitzgerald. 2011. "Predicting emergency department admissions." *Emergency medicine journal : EMJ* 29 (June): 358–65. https://doi.org/10.1136/emj.2010.103531.

Caldas, Francisco M., and Cláudia Soares. 2022. *A Temporal Fusion Transformer for Long-term Explainable Prediction of Emergency Department Overcrowding.* https://doi.org/10.48550/ARXIV.2207.00610.

Calegari, Rafael, Flavio S. Fogliatto, Filipe R. Lucini, Jeruza Neyeloff, Ricardo S. Kuchenbecker, and Beatriz D. Schaan. 2016. "Forecasting Daily Volume and Acuity of Patients in the Emergency Department." *Computational and Mathematical Methods in Medicine* 2016 (September): 3863268. ISSN: 1748-670X. https://doi.org/10.1155/2016/3863268.

Carvalho-Silva, Miguel, M. Teresa T. Monteiro, Filipe de Sá-Soares, and Sónia Dória-Nóbrega. 2018. "Assessment of forecasting models for patients arrival at Emergency Department." EURO 2016—New Advances in Health Care Applications, *Operations Research for Health Care* 18:112–118. ISSN: 2211-6923. https://doi.org/https://doi.org/10.1016/j.orhc.2017.05.001.

Champion, Robert, Leigh D. Kinsman, Geraldine A. Lee, Kevin A. Masman, Elizabeth A. May, Terence M. Mills, Michael D. Taylor, Paulett R. Thomas, and Ruth J. Williams. 2007. "Forecasting emergency department presentations." *Australian Health Review* 31 (1): 83–90. https://doi.org/10.1071/AH070083.

Chollet, François. 2018. *Deep Learning with Python.* Manning Publications Co. ISBN: 9781617294433.

Chollet, Francois, et al. 2015. "Keras." https://github.com/fchollet/keras.

Choudhury, Avishek, and Estefania Urena. 2020. "Forecasting hourly emergency department arrival using time series analysis." *British Journal of Healthcare Management* 26 (1): 34–43. https://doi.org/10.12968/bjhc.2019.0067.

Chujie, Tian, Jian Ma, Chunhong Zhang, and Panpan Zhan. 2018. "A Deep Neural Network Model for Short-Term Load Forecast Based on Long Short-Term Memory Network and Convolutional Neural Network." *Energies* 11 (12). https://doi.org/https://doi.org/10.3390/en11123493.

D.K., Thara, PremaSudha B.G, and Fan Xiong. 2019. "Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques." *Pattern Recognition Letters* 128:544–550. ISSN: 0167-8655. https://doi.org/https://doi.org/10.1016/j.patrec.2019.10.029.

De Veaux, Richard D., and Lyle H. Ungar. 1994. "Multicollinearity: A tale of two nonparametric regressions." In *Selecting Models from Data,* edited by P. Cheeseman and R. W. Oldford, 393–402. New York, NY: Springer New York. ISBN: 978-1-4612-2660-4.

Fang, X., H. Luo, and J. Tang. 2005. "Structural damage detection using neural network with learning rate improvement." *Computers & Structures* 83 (25): 2150–2161. ISSN: 0045-7949. https://doi.org/https://doi.org/10.1016/j.compstruc.2005.02.029.

Fattah, Jamal, Latifa Ezzine, Zineb Aman, Haj El Moussami, and Abdeslam Lachhab. 2018. "Forecasting of demand using ARIMA model." *International Journal of Engineering Business Management* 10:1847979018808673. https://doi.org/10.1177/1847979018808673.

Ferreira, D.C., and R.C. Marques. 2019. "Do quality and access to hospital services impact on their technical efficiency?" *Omega* 86:218–236. ISSN: 0305-0483. https://doi.org/https://doi.org/10.1016/j.omega.2018.07.010.

Forero, R., S. McCarthy, and K. Hillman. 2011. "Access Block and Emergency Department Overcrowding." In *Annual Update in Intensive Care and Emergency Medicine 2011,* edited by Jean-Louis Vincent, 720–728. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-18081-1. https://doi.org/10.1007/978-3-642-18081-1_63.

Gille, Felix, Anna Jobin, and Marcello Ienca. 2020. "What we talk about when we talk about trust: Theory of trust for AI in healthcare." *Intelligence-Based Medicine* 1-2:100001. ISSN: 2666-5212. https://doi.org/https://doi.org/10.1016/j.ibmed.2020.100001.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* Http://www.deeplearningbook.org. MIT Press.

Graham, Byron, Raymond Bond, Michael Quinn, and Maurice Mulvenna. 2018. "Using Data Mining to Predict Hospital Admissions From the Emergency Department." *IEEE Access* 6:10458–10469. https://doi.org/10.1109/ACCESS.2018.2808843.

Gul, Muhammet, and Erkan Celik. 2020. "An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments." *Health Systems* 9 (4): 263–284. https://doi.org/10.1080/20476965.2018.1547348.

Gupta, A. 2021. *A comprehensive guide on deep learning optimizers.* https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/.

Harrou, Fouzi, Abdelkader Dairi, Farid Kadri, and Ying Sun. 2020. "Forecasting emergency department overcrowding: A deep learning framework." *Chaos, Solitons & Fractals* 139:110247. ISSN: 0960-0779. https://doi.org/https://doi.org/10.1016/j.chaos.2020.110247.

Haykin, Simon. 2004. “1 FEEDFORWARD NEURAL NETWORKS : AN INTRODUCTION.”

He, Jun, Xiang-Yu Hou, Ghasem Toloo, Jennifer Patrick, and Gerry Gerald. 2011. “Demand for hospital emergency departments: A conceptual understanding.” *World journal of emergency medicine* 2 (January): 253–61. https://doi.org/10.5847/wjem.j.1920-8642.2011.04.002.

Hertzum, Morten. 2017. “Forecasting Hourly Patient Visits in the Emergency Department to Counteract Crowding.” 10:1–13. https://doi.org/https://doi.org/10.2174/1875934301710010001.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-term Memory.” *Neural computation* 9 (December): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Hong, Woo Suk, Adrian Daniel Haimovich, and R. Andrew Taylor. 2018. “Predicting hospital admission at emergency department triage using machine learning.” *PLOS ONE* 13, no. 7 (July): 1–13. https://doi.org/10.1371/journal.pone.0201016.

Javaid, Mohd, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. 2022. “Significance of machine learning in healthcare: Features, pillars and applications.” *International Journal of Intelligent Networks* 3:58–73. ISSN: 2666-6030. https://doi.org/https://doi.org/10.1016/j.ijin.2022.05.002.

Jilani, Tahseen, Gemma Housley, Grazziela Figueredo, Pui-Shan Tang, Jim Hatton, and Dominick Shaw. 2019. “Short and Long term predictions of Hospital emergency department attendances.” *International Journal of Medical Informatics* 129:167–174. ISSN: 1386-5056. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2019.05.011.

Jones, Spencer S., R. Scott Evans, Todd L. Allen, Alun Thomas, Peter J. Haug, Shari J. Welch, and Gregory L. Snow. 2009. "A multivariate time series approach to modeling and forecasting demand in the emergency department." *Journal of Biomedical Informatics* 42 (1): 123–139. ISSN: 1532-0464. https://doi.org/https://doi.org/10.1016/j.jbi.2008.05.003.

Jones, Spencer S., Alun Thomas, R. Scott Evans, Shari J. Welch, Peter J. Haug, and Gregory L. Snow. 2008. "Forecasting Daily Patient Volumes in the Emergency Department." *Academic Emergency Medicine* 15 (2): 159–170. https://doi.org/https://doi.org/10.1111/j.1553-2712.2007.00032.x.

Kadri, Farid, Fouzi Harrou, Sondès Chaabane, and Christian Tahon. 2014. "Time Series Modelling and Forecasting of Emergency Department Overcrowding." *Journal of Medical Systems* 38, no. 9 (July): 107. https://doi.org/10.1007/s10916-014-0107-0.

Kadri, Farid, Fouzi Harrou, and Sun Ying. 2017. "A Multivariate Time Series Approach to Forecasting Daily Attendances at Hospital Emergency Department." November. https://doi.org/10.1109/SSCI.2017.8280850.

Kam, Hye Jin, Jin Ok Sung, and Rae Woong Park. 2010. "Prediction of Daily Patient Numbers for a Regional Emergency Medical Center using Time Series Analysis." *Healthc Inform Res* 16 (3): 158–165. https://doi.org/10.4258/hir.2010.16.3.158.

Kröse, Ben, B. Krose, Patrick van der Smagt, and Patrick Smagt. 1993. "An introduction to neural networks." *J Comput Sci* 48 (January).

Kuo, Yong-Hong, Nicholas B. Chan, Janny M.Y. Leung, Helen Meng, Anthony Man-Cho So, Kelvin K.F. Tsoi, and Colin A. Graham. 2020. "An Integrated Approach of Machine Learning and Systems Thinking for Waiting Time Prediction in an Emergency Department." *International Journal of Medical Informatics* 139:104143. ISSN: 1386-5056. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2020.104143.

Maçaira, Paula Medina, Antônio Marcio Tavares Thomé, Fernando Luiz Cyrino Oliveira, and Ana Luiza Carvalho Ferrer. 2018. "Time series analysis with explanatory variables: A systematic literature review." *Environmental Modelling & Software* 107:199–209. ISSN: 1364-8152. https://doi.org/https://doi.org/10.1016/j.envsoft.2018.06.004.

Maddigan, Paula, and Teo Susnjak. 2022. *Forecasting Patient Demand at Urgent Care Clinics using Machine Learning*. https://doi.org/10.48550/ARXIV.2205.13067.

Marques, Isabel, Zélia Serrasqueiro, and Fernanda Nogueira. 2021. "Managers' Competences in Private Hospitals for Investment Decisions during the COVID-19 Pandemic." *Sustainability* 13 (4). ISSN: 2071-1050. https://doi.org/10.3390/su13041757.

Masters, Dominic, and Carlo Luschi. 2018. *Revisiting Small Batch Training for Deep Neural Networks*. https://doi.org/10.48550/ARXIV.1804.07612.

Maulud, Dastan, and Adnan M. Abdulazeez. 2020. "A Review on Linear Regression Comprehensive in Machine Learning." *Journal of Applied Science and Technology Trends* 1, no. 4 (December): 140–147. https://doi.org/10.38094/jastt1457.

McCarthy, Melissa L., Scott L. Zeger, Ru Ding, Dominik Aronsky, Nathan R. Hoot, and Gabor D. Kelen. 2008. "The Challenge of Predicting Demand for Emergency Department Services." *Academic Emergency Medicine* 15 (4): 337–346. https://doi.org/https://doi.org/10.1111/j.1553-2712.2008.00083.x.

Miller, D. L. 1973. "Collaborative studies of acute respiratory disease in patients seen in general practice and in children admitted to hospital. Aims, field methods and morbidity rates." *Postgraduate Medical Journal* 49 (577): 749–761. ISSN: 0032-5473. https://doi.org/10.1136/pgmj.49.577.749.

Mondal, Prapanna, Labani Shit, and Saptarsi Goswami. 2014. "Study of Effectiveness of Time Series Modeling (Arima) in Forecasting Stock Prices." *International Journal of Computer Science, Engineering and Applications* 4 (April): 13–29. https://doi.org/10.5121/ijcsea.2014.4202.

Morley, Claire, Maria Unwin, Gregory M. Peterson, Jim Stankovich, and Leigh Kinsman. 2018. "Emergency department crowding: A systematic review of causes, consequences and solutions." *PLOS ONE* 13, no. 8 (August): 1–42. https://doi.org/10.1371/journal.pone.0203316.

Moskop, John C., David P. Sklar, Joel M. Geiderman, Raquel M. Schears, and Kelly J. Bookman. 2009. "Emergency Department Crowding, Part 1—Concept, Causes, and Moral Consequences." *Annals of Emergency Medicine* 53 (5): 605–611. ISSN: 0196-0644. https://doi.org/https://doi.org/10.1016/j.annemergmed.2008.09.019.

Muthukrishnan, R, and R Rohini. 2016. "LASSO: A feature selection technique in predictive modeling for machine learning." In *2016 IEEE International Conference on Advances in Computer Applications (ICACA),* 18–20. https://doi.org/10.1109/ICACA.2016.7887916.

Nasiru, Suleman, Albert Luguterah, and Lea Anzagra. 2013. "The Efficacy of ARIMAX and SARIMA Models in Predicting Monthly Currency in Circulation in Ghana." *Mathematical Theory and Modeling* 3 (September): 2013.

Nunno, Lucas. 2014. "Stock market price prediction using linear and polynomial regression models." *Computer Science Department, University of New Mexico: Albuquerque, NM, USA.*

O'brien, Robert M. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." *Quality & Quantity* 41, no. 5 (October): 673–690. ISSN: 1573-7845. https://doi.org/10.1007/s11135-006-9018-6.

Oliveira, Flávia Cristina Dias. 2020. "Fatores que contribuem para a afluência de casos não urgentes nos Serviços de Urgência: um estudo no Hospital da Senhora da Oliveira, Guimarães, E.P.E." Master's thesis, School of Economics and Management, University of Minho, July.

Pak, Anton, Brenda Gannon, and Andrew Staib. 2021. "Predicting waiting time to treatment for emergency department patients." *International Journal of Medical Informatics* 145:104303. ISSN: 1386-5056. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2020.104303.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

PORDATA. 2022. *População residente, média anual: total e por sexo.* (accessed: 25.11.2022), https://www.pordata.pt/db/portugal/ambiente+de+consulta/tabela/5831968.

Portuguese Institute for Sea and Atmosphere. 2022. *Boletim Climatológico,* December. https://www.ipma.pt/pt/publicacoes/boletins.jsp?cmbDep=cli%5C&cmbTema=pcl%5C&idDep=cli%5C&idTema=pcl%5C&curAno=-1.

Portuguese National Health Services. 2017. *2016 com maior atividade assistencial e com mais portugueses com médico de família.* https://www.acss.min-saude.pt/2017/02/20/2016-com-maior-atividade-assistencial-e-com-mais-portugueses-com-medico-de-familia/. Accessed: 2022-10-31.

Rafiq, M.Y, G Bugmann, and D.J Easterbrook. 2001. "Neural network design for engineering applications." *Computers & Structures* 79 (17): 1541–1552. ISSN: 0045-7949. https://doi.org/https://doi.org/10.1016/S0045-7949(01)00039-6.

Raita, Yoshihiko, Tadahiro Goto, Mohammad Kamal Faridi, David F. M. Brown, Carlos A. Camargo Jr., and Kohei Hasegawa. 2019. "Emergency department triage prediction of clinical outcomes using machine learning models." *Crit Care* 23 (64). https://doi.org/https://doi.org/10.1186/s13054-019-2351-7.

Relatório Qualidade e Segurança Clínica. 2021. Accessed: November 25, 2022. CUF. https://www.cuf.pt/sites/portalcuf/files/documents/2022-05/RelatorioQualidadeSegurancaClinicaCUF2021.pdf.

Rocha, Carlos Narciso, and Fátima Rodrigues. 2021. "Forecasting emergency department admissions." *Journal of Intelligent Information Systems* 56:509–528. https://doi.org/https://doi.org/10.1007/s10844-021-00638-9.

Rocha, Patrícia Alves da. 2020. "A Procura de Cuidados de Saúde Urgentes em Portugal." Master's thesis, NOVA University Lisbon, out.

Roquette, Bruno P., Hitoshi Nagano, Ernesto C. Marujo, and Alexandre C. Maiorano. 2020. "Prediction of admission in pediatric emergency department with deep neural networks and triage textual data." *Neural Networks* 126:170–177. ISSN: 0893-6080. https://doi.org/https://doi.org/10.1016/j.neunet.2020.03.012.

Ryu, Alexander J., Santiago Romero-Brufau, Ray Qian, Heather A. Heaton, David M. Nestler, Shant Ayanian, and Thomas C. Kingsley. 2022. "Assessing the Generalizability of a Clinical Machine Learning Model Across Multiple Emergency Departments." *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 6 (3): 193–199. ISSN: 2542-4548. https://doi.org/https://doi.org/10.1016/j.mayocpiqo.2022.03.003.

Santos, André Peralta, Paulo Freitas, and Henrique Manuel Gil Martins. 2014. "Manchester Triage System version II and resource utilisation in the emergency department" [in en]. *Emerg. Med. J.* 31, no. 2 (February): 148–152.

Sarker, Iqbal H. 2021. "Machine Learning: Algorithms, Real-World Applications and Research Directions." *SN Computer Science* 2. https://doi.org/https://doi.org/10.1007/s42979-021-00592-x.

Sebastião, Sara, and Madalena Cunha. 2021. "O paciente emergente no serviço de urgência: estratificação do risco Clínico." *Servir* 2, no. 01 (December): 65–74. https://doi.org/10.48492/servir0201.25679.

Siami-Namini, Sima, and Akbar Siami Namin. 2018. *Forecasting Economics and Financial Time Series: ARIMA vs. LSTM.* https://doi.org/10.48550/ARXIV.1803.06386.

Simões, Jorge, Gonçalo Augusto, Inês Fronteira, and Cristina Hernández-Quevedo. 2017. "Portugal: Health system review." *Health Systems in Transition* 19 (2): 1–184.

Smith, Taylor G., et al. 2017–. *pmdarima: ARIMA estimators for Python.* [Online; accessed ¡today¿]. http://www.alkaline-ml.com/pmdarima.

Sudarshan, Vidya K., Mikkel Brabrand, Troels Martin Range, and Uffe Kock Wiil. 2021. "Performance evaluation of Emergency Department patient arrivals forecasting models by including meteorological and calendar information: A comparative study." *Computers in Biology and Medicine* 135:104541. ISSN: 0010-4825. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104541.

Sun, Yan, Bee Hoon Heng, Yian Tay Seow, and Eillyne Seow. 2009. "Forecasting daily attendances at an emergency department to aid resource planning." *BMC Emerg Med* 9 (1). https://doi.org/https://doi.org/10.1186/1471-227X-9-1.

Svozil, Daniel, Vladimír Kvasnicka, and Jirí Pospichal. 1997. "Introduction to multi-layer feed-forward neural networks." *Chemometrics and Intelligent Laboratory Systems* 39 (1): 43–62. ISSN: 0169-7439. https://doi.org/https://doi.org/10.1016/S0169-7439(97)00061-0.

Tashman, Len. 2000. “Out-of-sample tests of forecasting accuracy: An analysis and review.” *International Journal of Forecasting* 16 (October): 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0.

Thompson, Christopher Glen, Rae Seon Kim, Ariel M. Aloe, and Betsy Jane Becker. 2017. “Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results.” *Basic and Applied Social Psychology* 39 (2): 81–90. https://doi.org/10.1080/01973533.2016.1277529.

Tranmer, Mark, Jen Murphy, Mark Elliot, and Maria Pampaka. 2020. *Multiple LinearRegression (2nd Edition),* January. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/2020-1-multiple-linear-regression.pdf.

Tuominen, Jalmari, Francesco Lomio, Niku Oksala, Ari Palomäki, Jaakko Peltonen, Heikki Huttunen, and Antti Roine. 2022. “Forecasting daily emergency department arrivals using high-dimensional multivariate data: a feature selection approach.” *BMC Medical Informatics and Decision Making* 22, no. 1 (May): 134. ISSN: 1472-6947. https://doi.org/10.1186/s12911-022-01878-7.

Vollmer, Michaela A.C., Ben Glampson, Thomas Mellan, Swapnil Mishra, Luca Mercuri, Ceire Costello, Robert Klaber, Graham Cooke, Seth Flaxman, and Samir Bhatt. 2021. “A unified machine learning approach to time series forecasting applied to demand at emergency departments.” *BMC Emergency Medicine* 21, no. 9 (January). https://doi.org/https://doi.org/10.1186/s12873-020-00395-y.

Wang, Can, Mitra Baratchi, Thomas Bäck, Holger H. Hoos, Steffen Limmer, and Markus Olhofer. 2022. "Towards Time-Series Feature Engineering in Automated Machine Learning for Multi-Step-Ahead Forecasting." *Engineering Proceedings* 18 (1). ISSN: 2673-4591. https://doi.org/10.3390/engproc2022018017.

Whitt, Ward, and Xiaopei Zhang. 2019. "Forecasting arrivals and occupancy levels in an emergency department." *Operations Research for Health Care* 21:1–18. ISSN: 2211-6923. https://doi.org/https://doi.org/10.1016/j.orhc.2019.01.002.

World Health Organization. 2022. *WHO Coronavirus (COVID-19) Dashboard — Daily cases and deaths by date reported to WHO,* December. https://covid19.who.int/data.

Xu, M., T.C. Wong, and K.S. Chin. 2013. "Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network." *Decision Support Systems* 54 (3): 1488–1498. ISSN: 0167-9236. https://doi.org/https://doi.org/10.1016/j.dss.2012.12.019.

Ying, Xue. 2019. "An Overview of Overfitting and its Solutions." *Journal of Physics: Conference Series* 1168, no. 2 (February): 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.

Yu, Lean, Geye Hang, Ling Tang, Yang Zhao, and Kin Keung Lai. 2017. "Forecasting Patient Visits to Hospitals using a WD&ANN-based Decomposition and Ensemble Model." *Eurasia Journal of Mathematics, Science and Technology Education* 13 (November). https://doi.org/10.12973/ejmste/80308.

Yucesan, Melih, Muhammet Gul, and Erkan Celik. 2020. "A multi-method patient arrival forecasting outline for hospital emergency departments." *International Journal of Healthcare Management* 13 (sup1): 283–295. https://doi.org/10.1080/20479700.2018.1531608.

Zaytar, Mohamed Akram, and Chaker El Amrani. 2016. "Sequence to Sequence Weather Forecasting with Long Short-Term Memory Recurrent Neural Networks." *International Journal of Computer Applications* 143 (June): 7–11. https://doi.org/10.5120/ijca2016910497.

Zhao, Xinxing, Joel Weijia Lai, Andrew Fu Wah Ho, Nan Liu, Marcus Eng Hock Ong, and Kang Hao Cheong. 2022. "Predicting hospital emergency department visits with deep learning approaches." *Biocybernetics and Biomedical Engineering* 42 (3): 1051–1065. ISSN: 0208-5216. https://doi.org/https://doi.org/10.1016/j.bbe.2022.07.008.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–320. https://doi.org/https://doi.org/10.1111/j.1467-9868.2005.00503.x.

A Appendix

	Number of Missing Values	Percentage of Missing Values (%)
especialidade_medico	17101	5.79
tempo_espera_max	3629	1.23
mtr_tempo_ate_primeira_triagem	1647	0.56
mtr_tempo_ate_primeira_observacao	1073	0.36
dsc_destino	741	0.25
mtr_tempo_permanencia_ap	722	0.24
tipo_mobilidade_pk	529	0.18
tipo_triagem_pk	529	0.18
dsc_cor_triagem	529	0.18
dsc_mobilidade	529	0.18
cod_sexo	46	0.02
idade	1	0.00

Figure A.1: Number and percentage of missing values

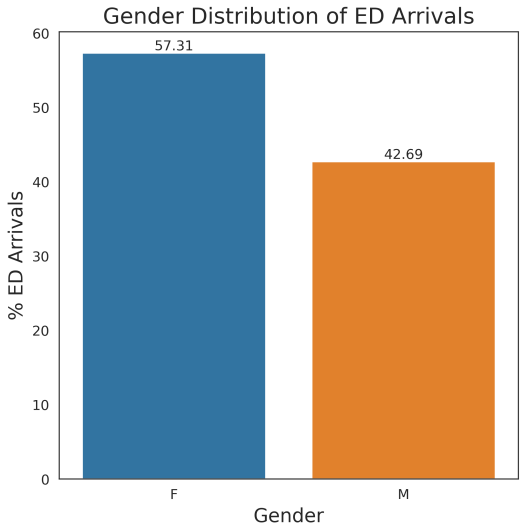


Figure A.2: Gender distribution of ED arrivals

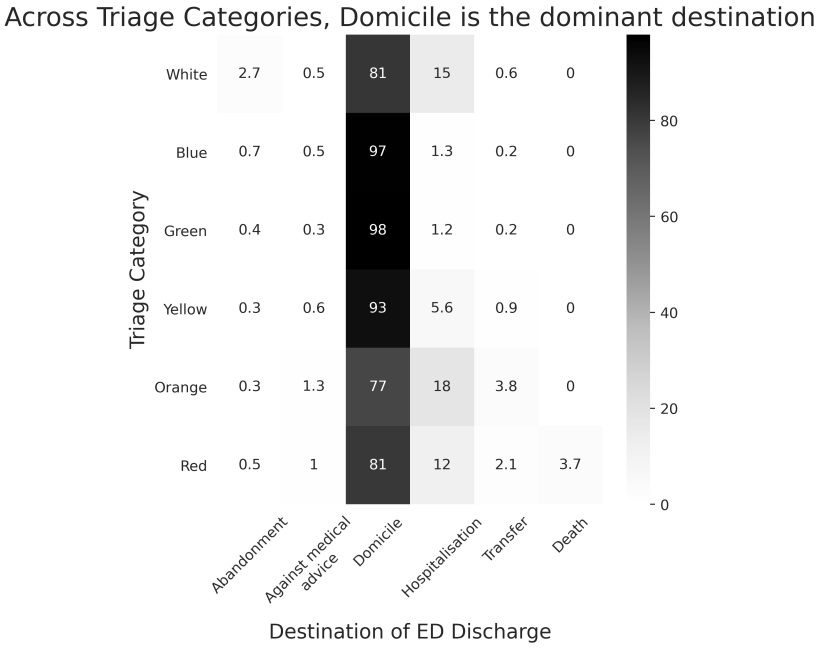


Figure A.3: Percentage of ED discharge destination per triage category

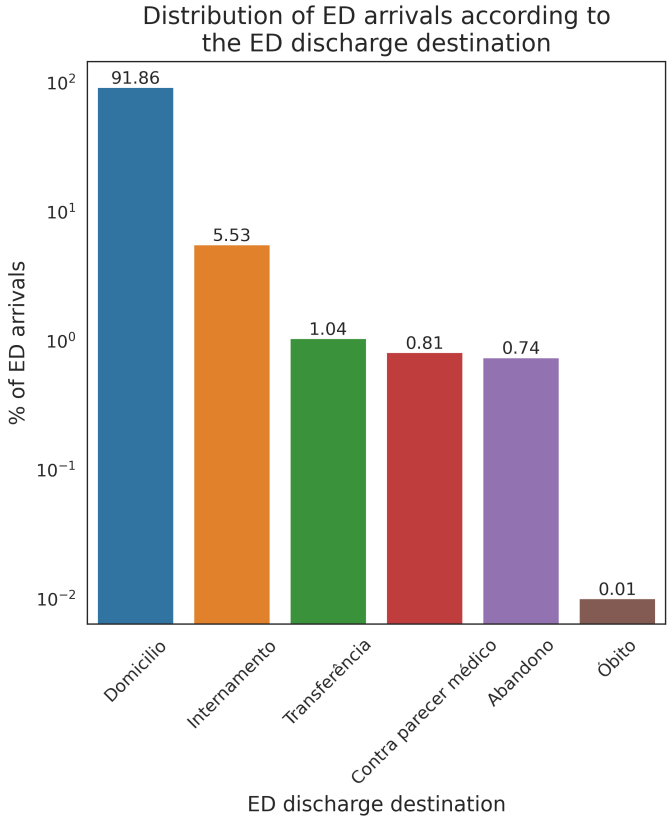


Figure A.4: Distribution of ED arrivals per discharge destiny

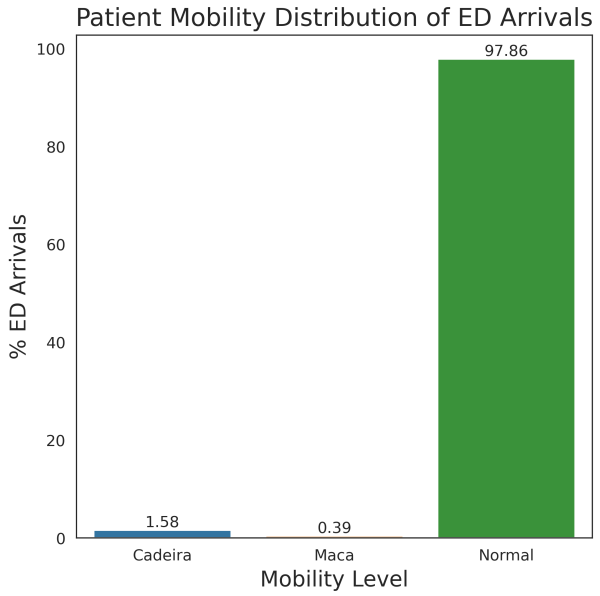


Figure A.5: Distribution of ED arrivals per patient mobility

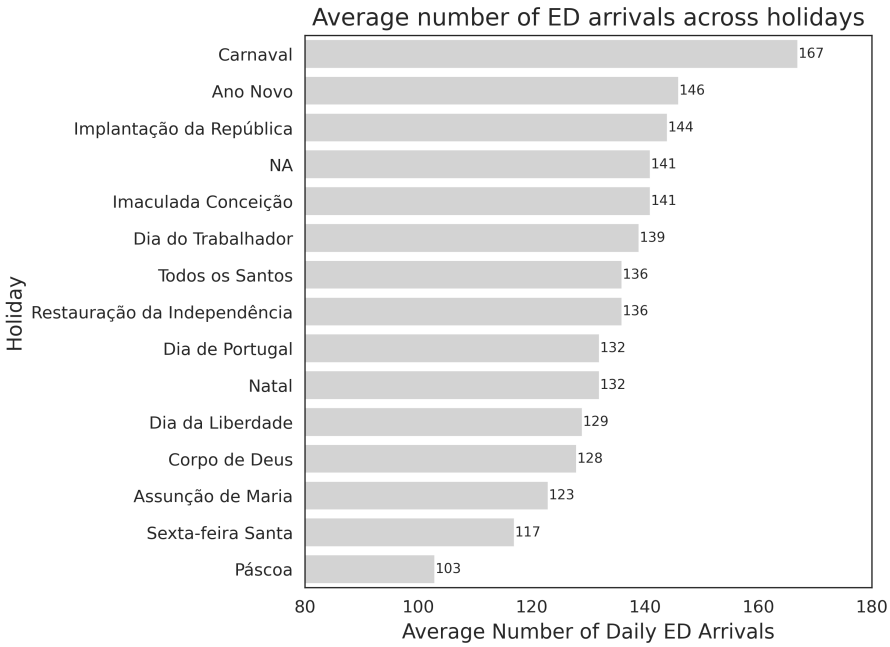


Figure A.6: Average daily ED arrivals on national holidays

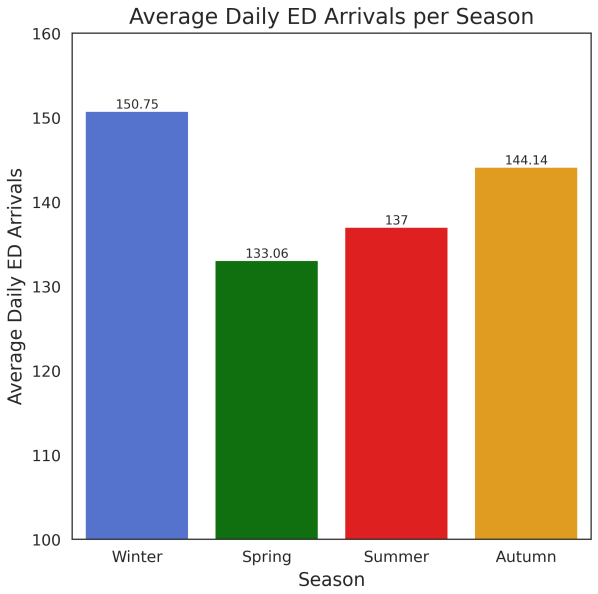


Figure A.7: Average daily ED arrivals per season

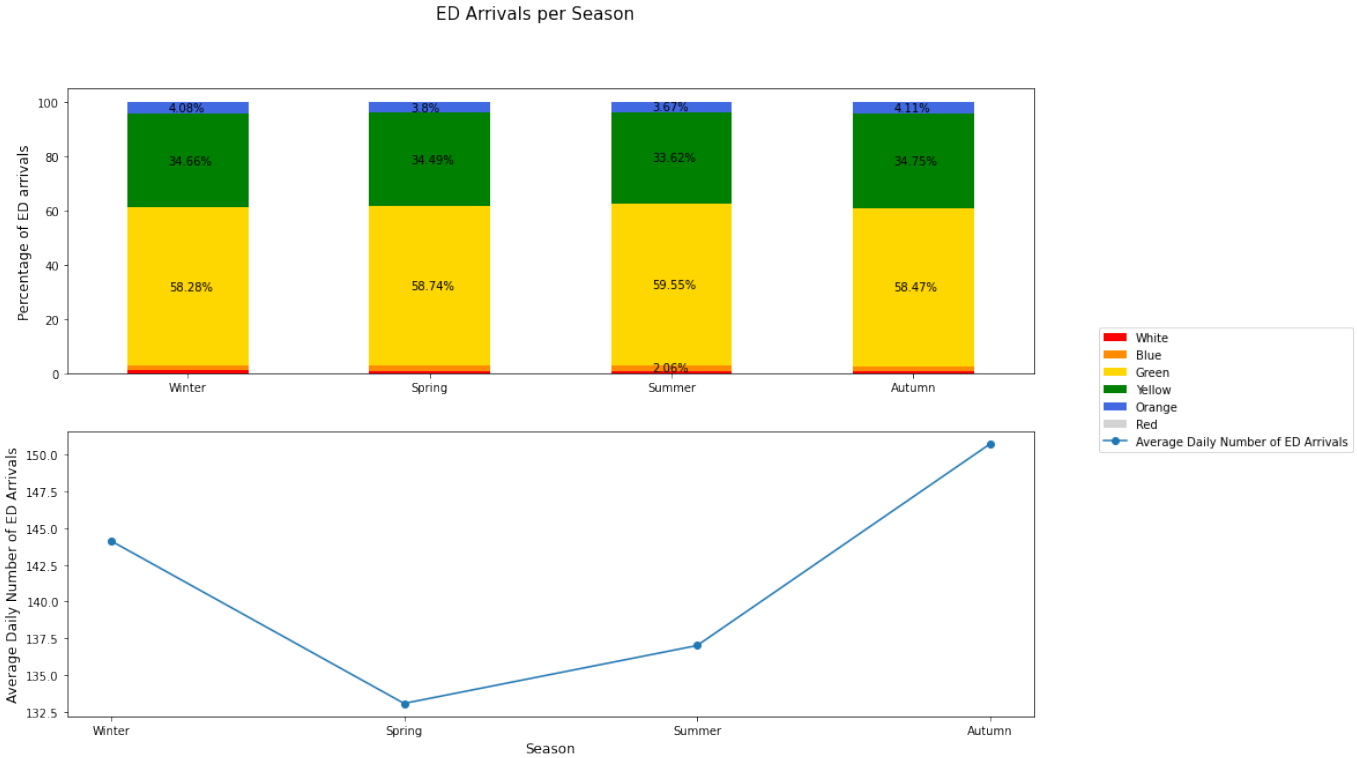


Figure A.8: Average daily ED arrivals per season

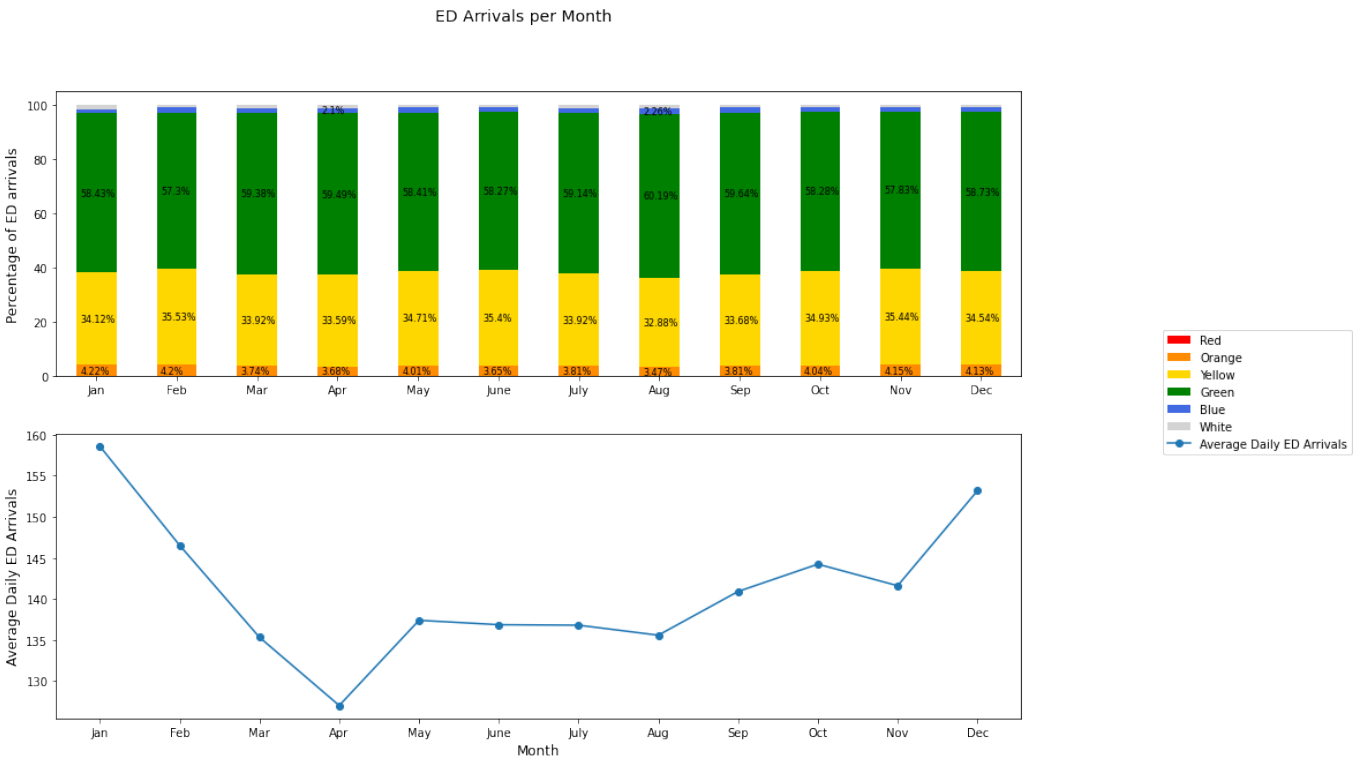


Figure A.9: Average daily ED arrivals per month

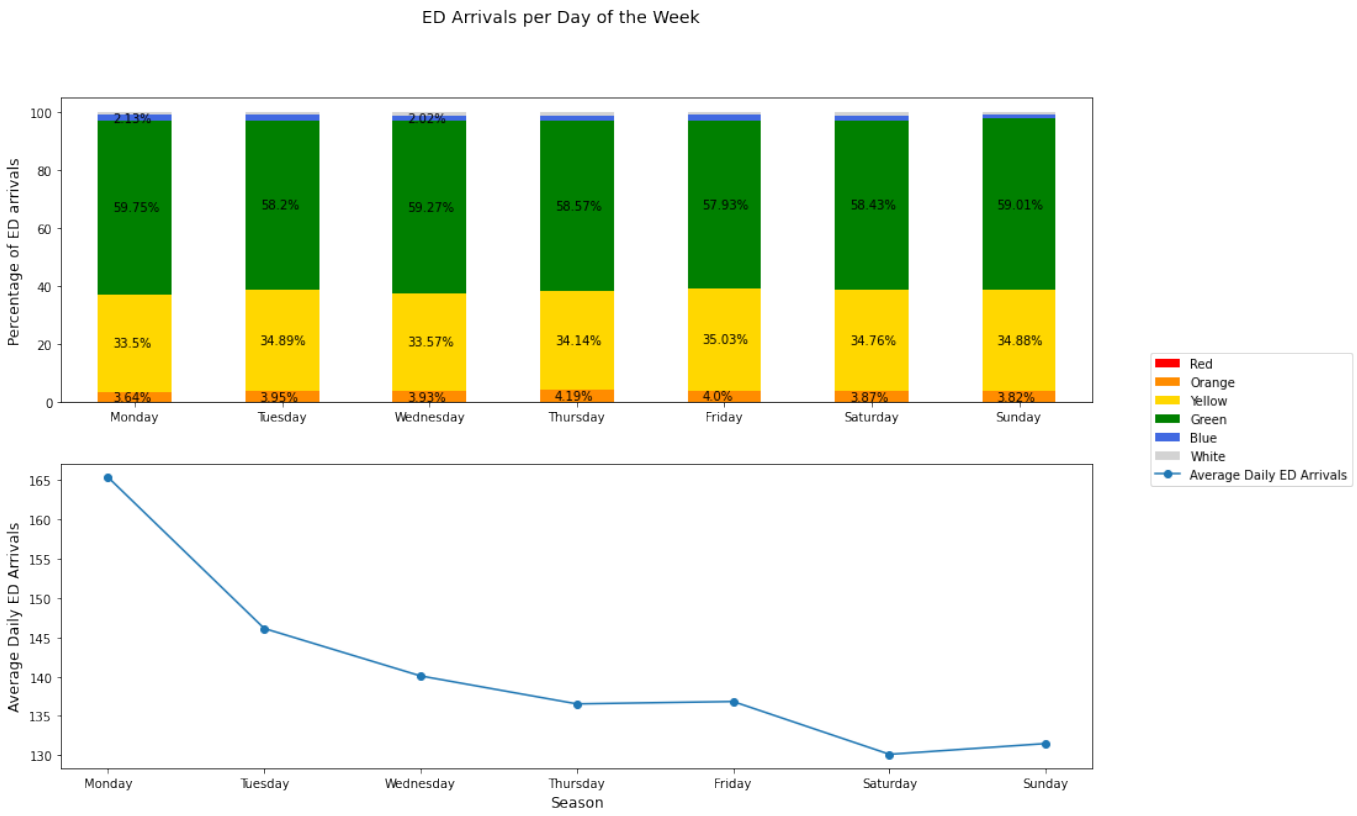


Figure A.10: Average daily ED arrivals per day of the week

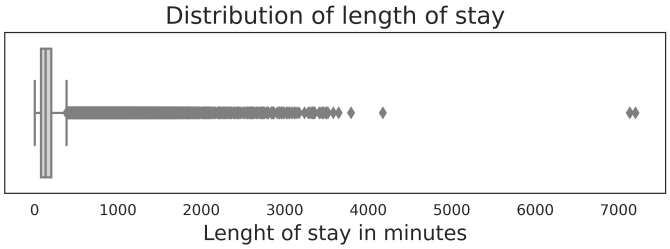


Figure A.11: Distribution of length of stay

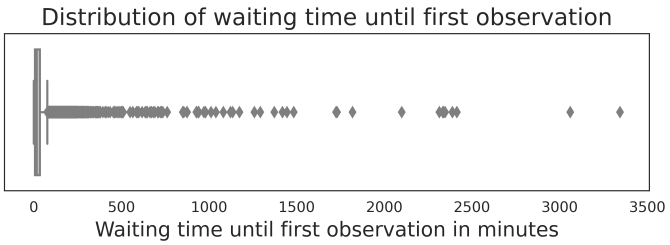


Figure A.12: Distribution of waiting time until first triage

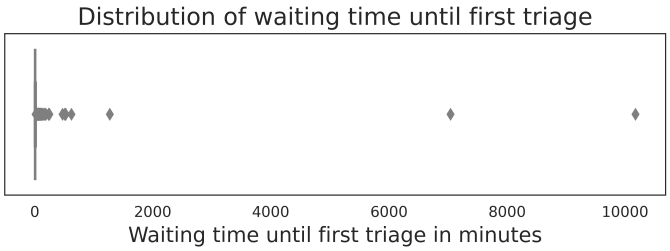
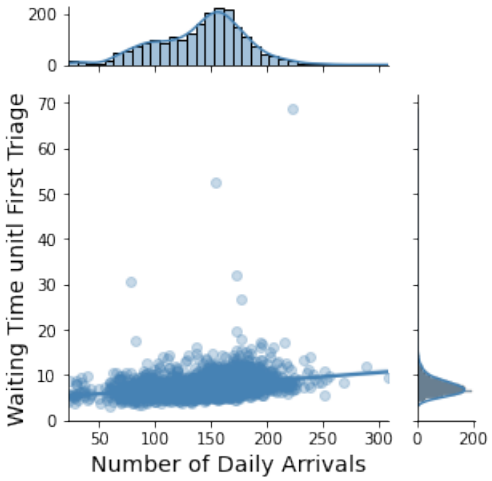
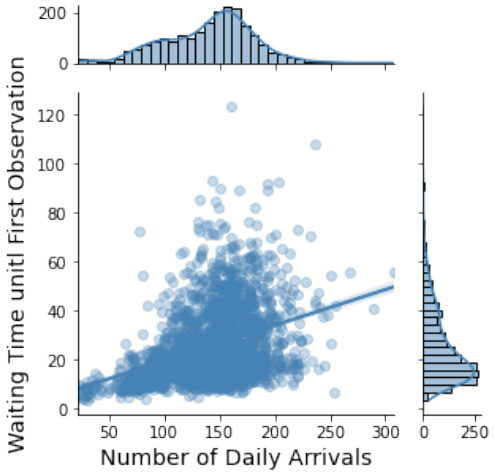


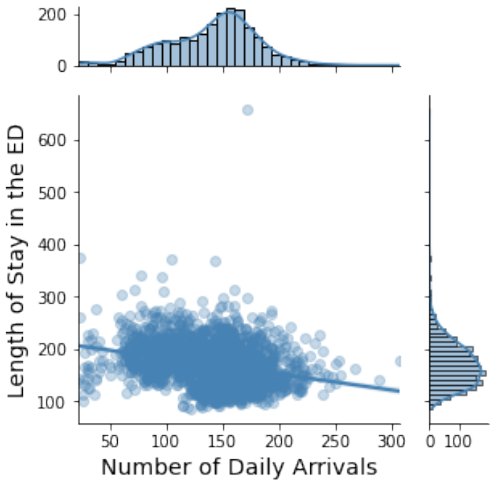
Figure A.13: Distribution of waiting time until first observation



(a) Daily ED arrivals and waiting time until first triage



(b) Daily ED arrivals and waiting time until first observation



(c) Daily ED arrivals and length of stay

Figure A.14: Daily ED arrivals, length of stay and waiting times

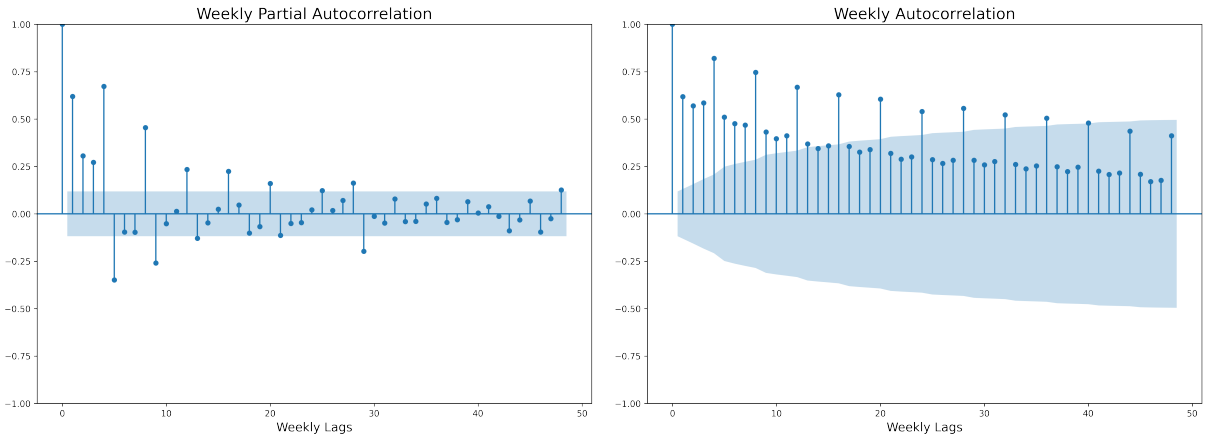


Figure A.15: Partial Autocorrelation and Autocorrelation Functions - *All Data*

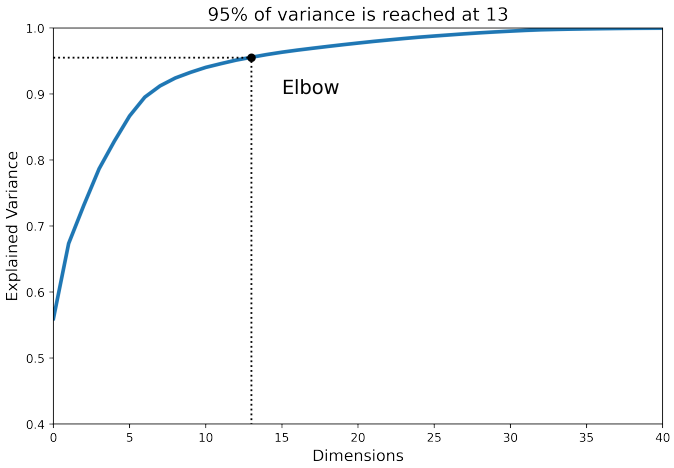


Figure A.16: Explained variance per number of principal components - *All Data*

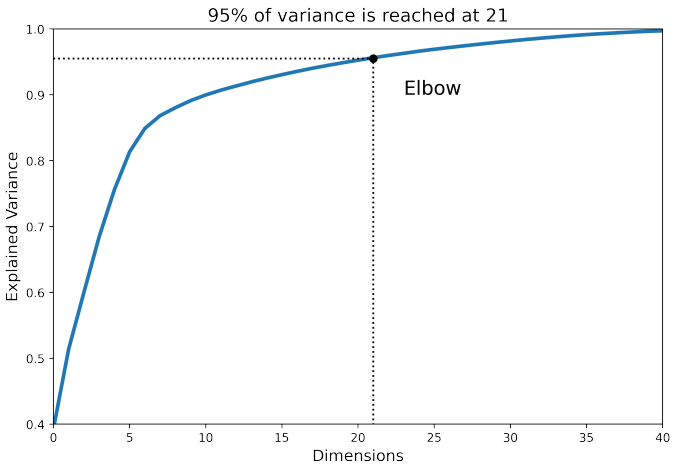


Figure A.17: Explained variance per number of principal components - $S20$

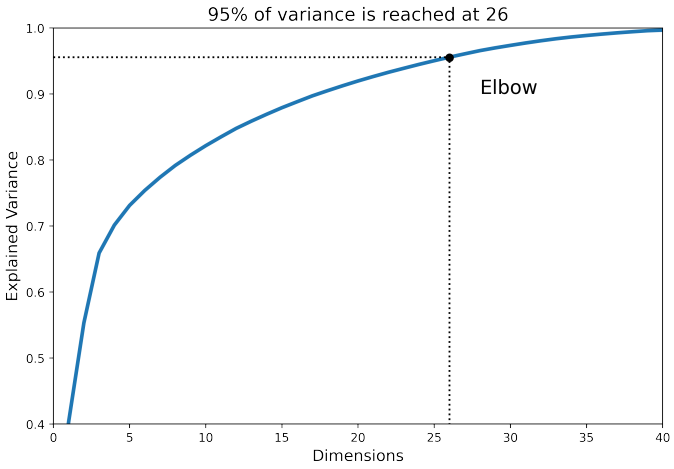


Figure A.18: Explained variance per number of principal components - $S20\_21$

dados_semanais

one entry per week

276 rows and 47 columns

Column Name	Description	Data Type	Count	Null %	Unique	Min	Mean	Max	Example	Confidential
semana_dados	Week number on the dataset	integer	276	0	276	1	-	276	239	No
ano	Year	integer	276	0	6	2017	-	2022	2021	No
semestre	Semester of the year	object	276	0	2	1	-	2	2	No
trimestre	Quarter of the year	integer	276	0	4	1	-	4	4	No
mes	Month of the year	integer	276	0	12	1	-	12	12	No
semana_mes	Week of the month	integer	276	0	4	1	-	4	3	No
total_episodios	Total number of weekly arrivals	integer	276	0	241	212	1071	2137	882	No
estacao	Season of the year	object	276	0	4	-	-	-	Autumn	No
flg_feriado	0 if a national holiday did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_ano_novo	0 if a "Ano Novo" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_dia_portugal	0 if a "Dia de Portugal" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_carnaval	0 if a "Carnaval" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_sexta_feira_santa	0 if a "Sexta-Feira Santa" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_assuncao_de_maria	0 if a "Assunção de Maria" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_corpo_de_deus	0 if a "Corpor de Deus" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_dia_do_trabalhador	0 if a "Dia do Trabalhador" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_implatacao_da_republica	0 if a "Implantação da República" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_dia_da_liberdade	0 if a "Dia da Liberdade" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_restauracao_da_independencia	0 if a "Restauração da Independência" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_todos_os_santos	0 if a "Dia de Todos os Santos" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_imaculada_conceicao	0 if a "Imaculada Conceição" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_natal	0 if a "Natal" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
flg_pascoa	0 if a "Páscoa" did not happened during the week; 1 if it happened	integer	276	0	-	-	-	-	1	No
vagas_covid	Categorical variable that identifies the waves with higher virus spread	integer	276	0	4	0	-	3	2	No
estado_emergencia	Identifies whether the week corresponds to a period when the country was a state of emergency	integer	276	0	2	0	-	1	0	No
estado_calamidade	ifies whether the week corresponds to a period when the country was a state of public cala	integer	276	0	2	0	-	1	0	No
semana_ano	Week number of the year	integer	276	0	48	1	-	48	47	No
pca_i, $i \in \{1, 2, 3, 4, \ldots, 13\}$	Components obtained through the aplication of the PCA algorithm	integer	276	0	-	-	-	-	-	No
lag_temp_min_media_mes_30	Monthly average of minimum daily temperature registered in the previous month	float	276	0	49	6.9	13.09	18,9	10.1	No
lag_temp_max_media_mes_30	Monthly average of maximum daily temperature registered in the previous month	float	276	0	56	13.5	22.14	32.4	17.8	No
lag_temp_min_abs_mes_30	Minimum daily temperature registered in the previous month	float	276	0	56	0.9	9.42	16	6.7	No
lag_temp_max_abs_mes_30	Maximum daily temperature registered in the previous month	float	276	0	63	17.6	31.2	227.6	22.1	No
lag_precip_total_mes_30	Total precipitation registered in the previous month, in mm	float	276	0	62	0	42.54	232.3	16.3	No
lag_precip_max_mes_mm_30	Maximum daily precipitation amount registered in the previous month, in mm	float	276	0	61	0	14.65	64.6	7.4	No
lag_intens_vento_max_mes_kmh_30	Maximum wind speed registered in the previous month, in km/h	float	276	0	48	44.3	62,62	87.5	70.2	No

Figure A.19: Data dictionary - Weekly data

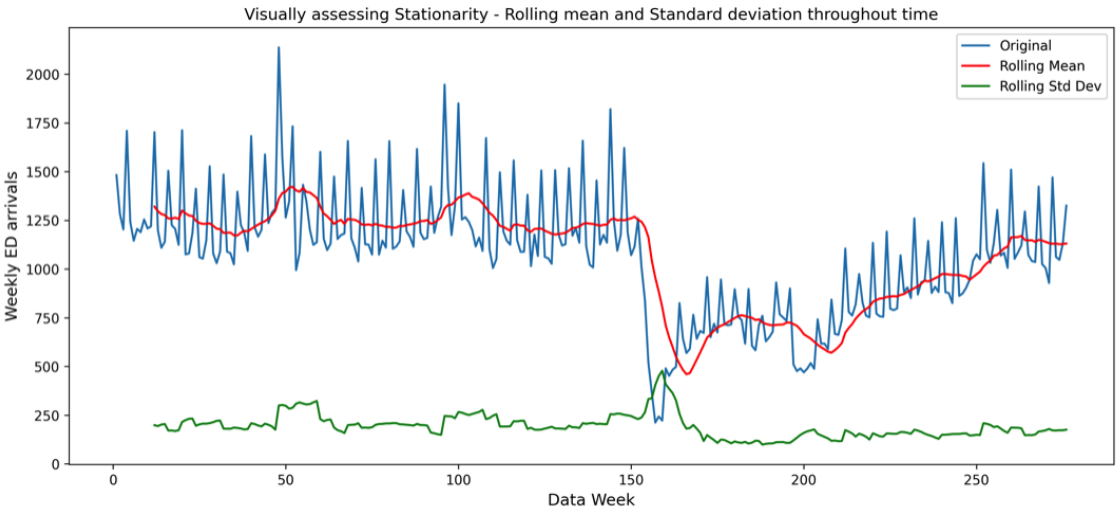


Figure A.20: Stationary - *All Data*

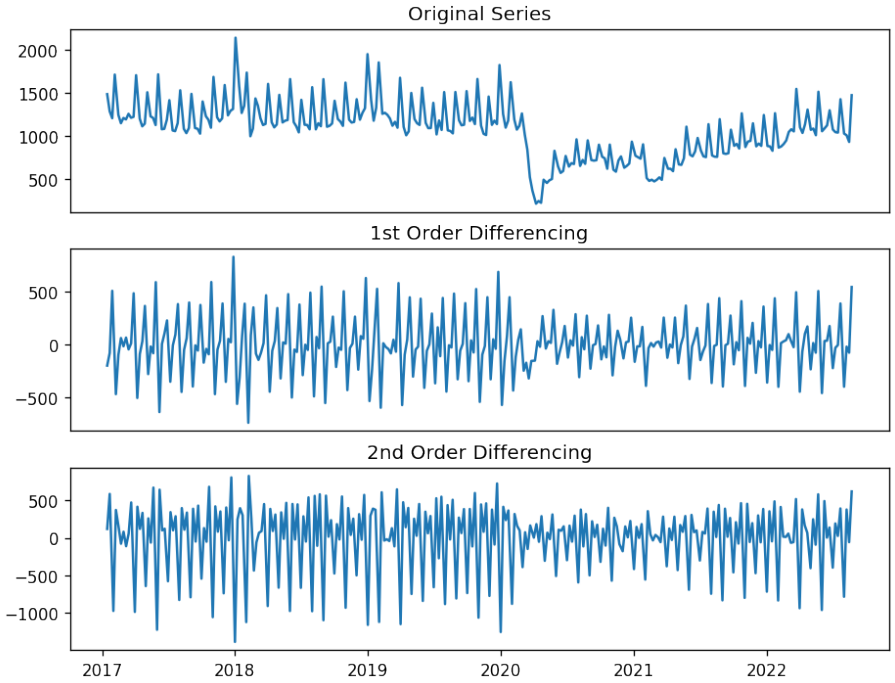


Figure A.21: Differentiation order - *All Data*

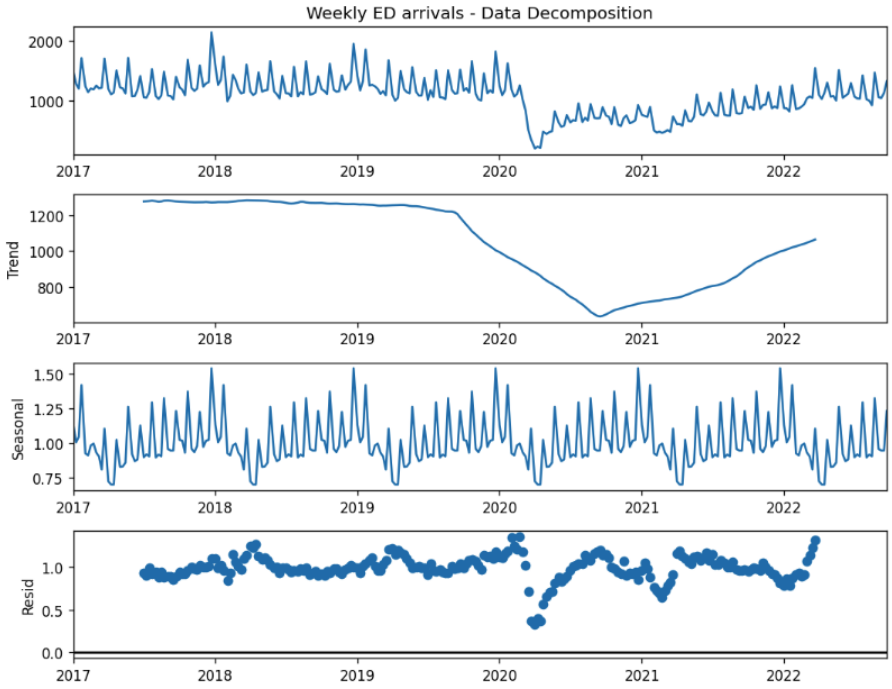


Figure A.22: Decomposition - *All Data*

Features	**Coefficients**
semana_mes_4	383.493
ano_2021	251.937
estado_emergencia	216.177
lag_temp_min_abs_mes_30	9.399
lag_temp_max_media_mes_30	9.047
lag_precip_max_mes_mm_30	0.740
pca_6	0.095
pca_1	0.058
pca_5	0.024
pca_4	0.008
pca_2	0.004
pca_3	0.002

Table A.1: Linear Regression: Feature selection - $S20$ dataset

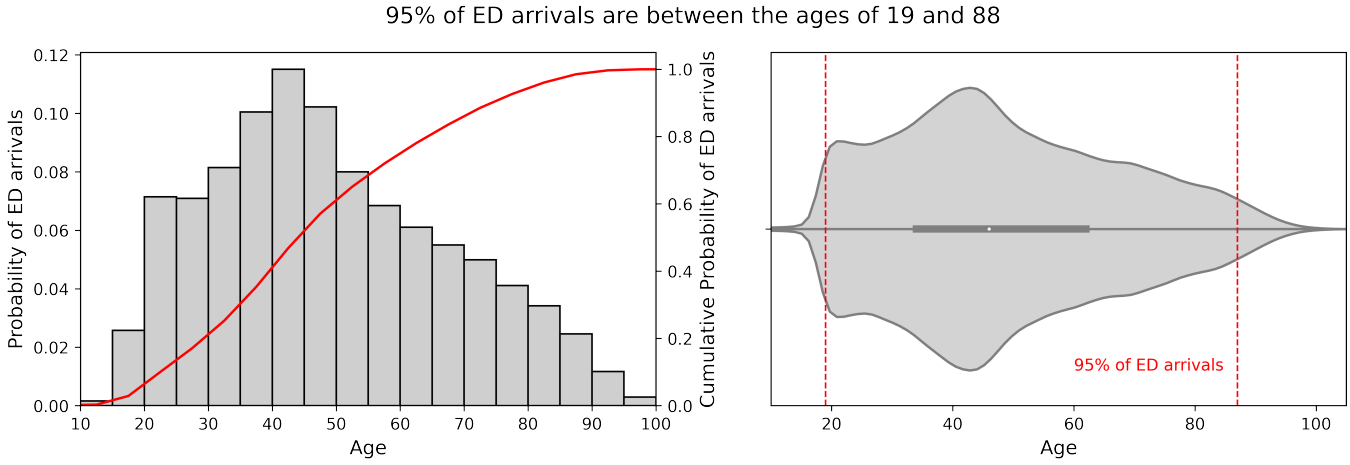


Figure A.23: Unit 3 - Age distribution of ED arrivals

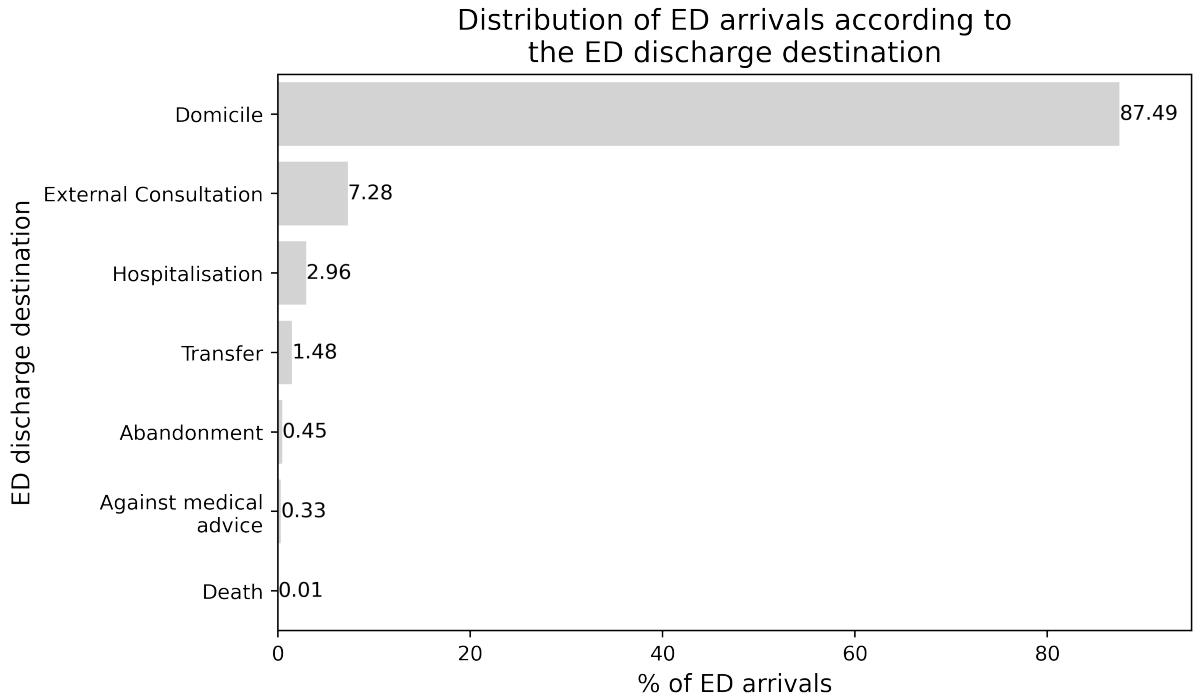


Figure A.24: Unit 3 - ED discharge destination

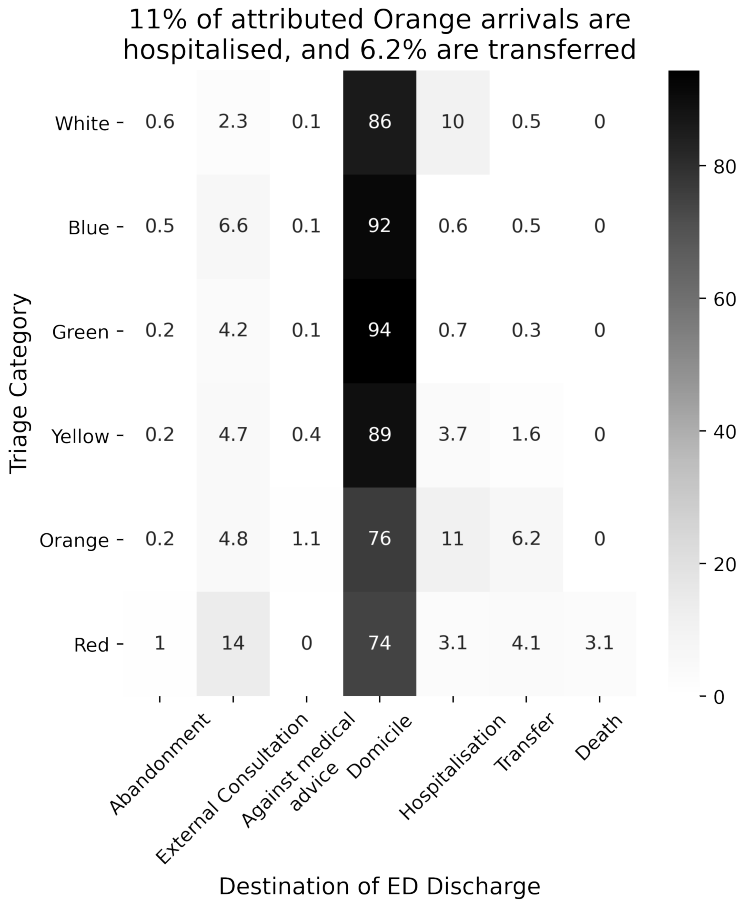


Figure A.25: Unit 3 - Percentage of ED discharge destination per triage category

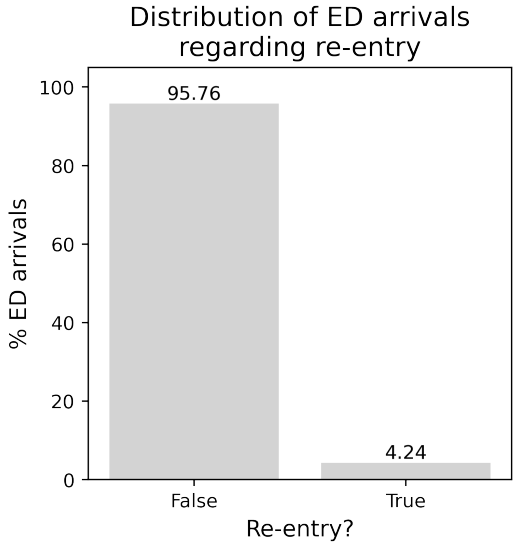


Figure A.26: Unit 3 - Percentage of re-entries

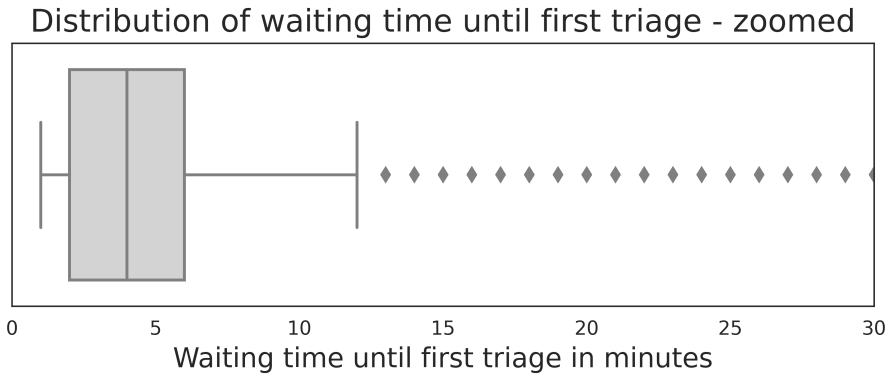


Figure A.27: Unit 3 - Distribution of length of stay

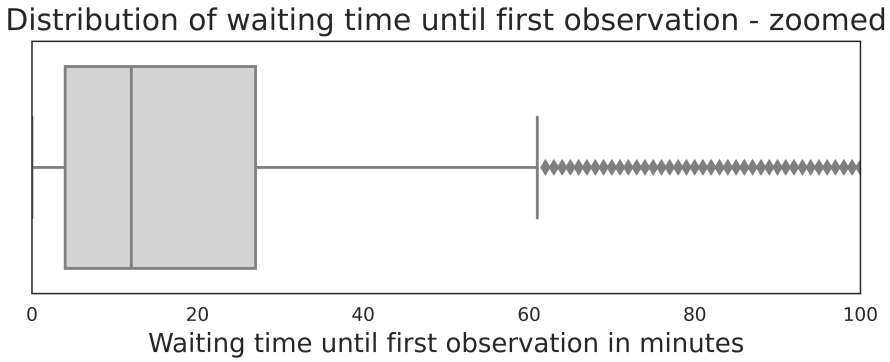


Figure A.28: Unit 3 - Distribution of waiting time until first triage

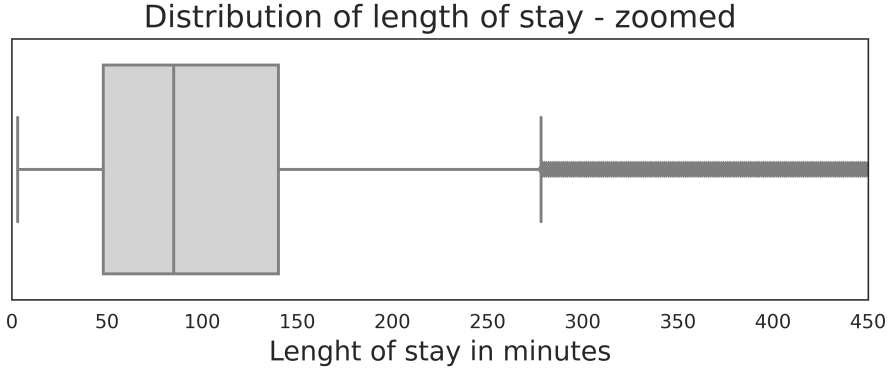


Figure A.29: Unit 3 - Distribution of waiting time until first observation

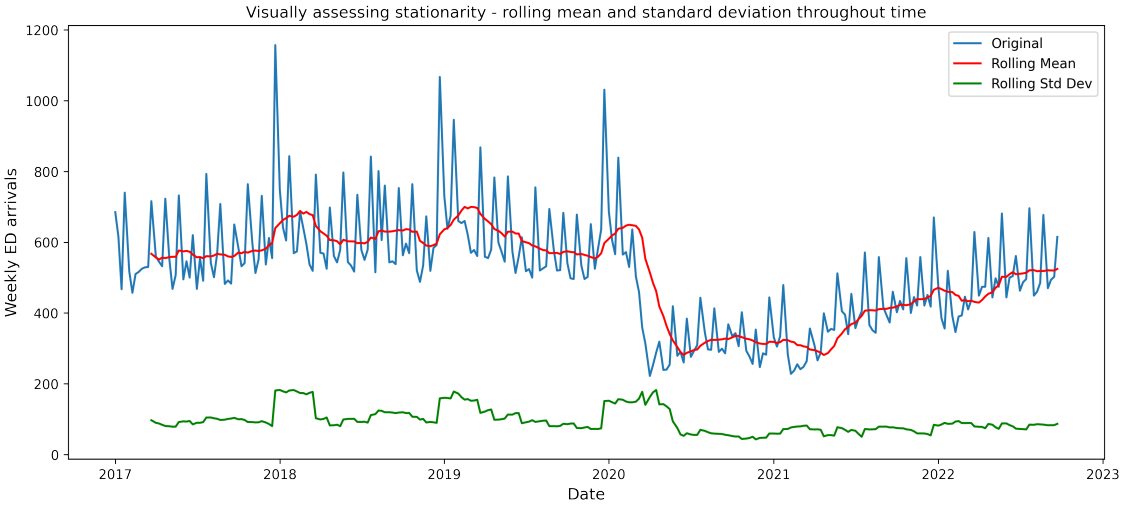


Figure A.30: Unit 3 - Rolling mean and standard deviation throughout time

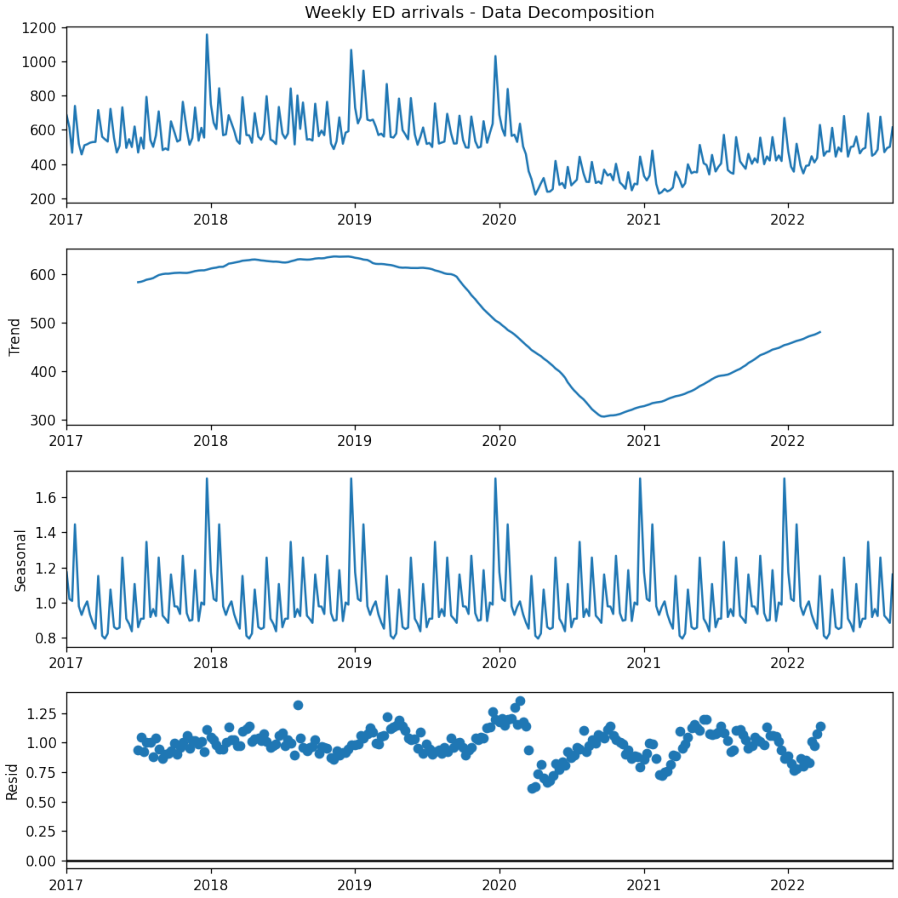


Figure A.31: Unit 3 - Seasonal decompose

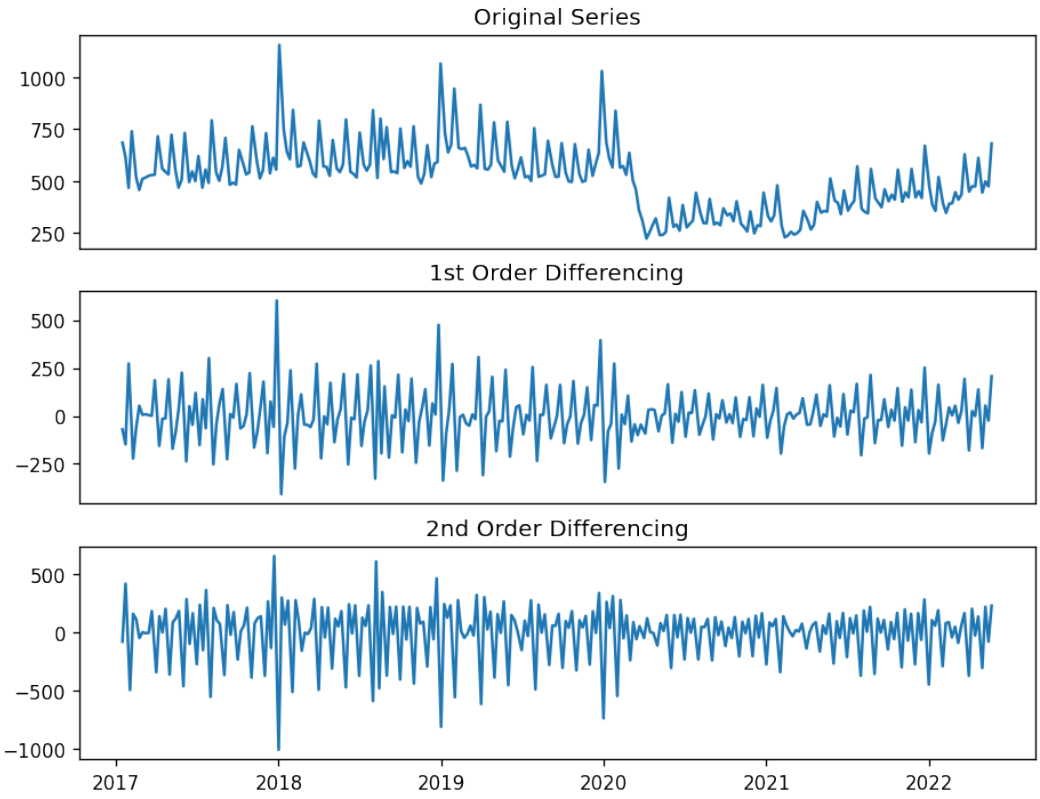


Figure A.32: Unit 3 - First and second order of differentiation

Selected exogenous features	
→ *semana_mes_4*	→ *lag_temp_min_media_mes_30*
→ *vagas_covid_1*	→ *lag_temp_max_media_mes_30*
→ *vagas_covid_2*	→ *lag_temp_min_abs_mes_30*
→ *vagas_covid_3*	→ *lag_temp_max_abs_mes_30*
→ *estacao_winter*	→ *lag_precip_max_mes_mm_30*
→ *estado_emergencia*	→ *lag_intens_vento_max_mes_kmh_30*

Table A.2: Unit 3 - Exogenous variables selected through *SelectKBest*

Step	MAPE	MAE	RMSE	Std Dev
t + 1	8.07	36.61	42.83	22.23
t + 2	2.07	9.91	11.33	7.38
t + 3	2.79	13.92	18.04	13.83
t + 4	4.16	25.43	27.92	27.12

Figure A.33: Unit 3 - $\text{SARIMAX}_{\text{retuned}}$ average performance of each step