National College of Ireland

Forecasting of Hospital Outpatient Waiting Lists in Ireland using Time-Series and Machine Learning

MSc Research Project
Data Analytics

Sonal Srinath
Student ID: 19207638

School of Computing
National College of Ireland

Supervisor: Dr Martin Alain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sonal Srinath
Student ID:	19207638
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr Martin Alain
Submission Due Date:	16/12/2021
Project Title:	Forecasting of Hospital Outpatient Waiting Lists in Ireland using Time-Series and Machine Learning
Word Count:	7084
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	**Sonal Srinath**
Date:	27th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting of Hospital Outpatient Waiting Lists in Ireland using Time-Series and Machine Learning

Sonal Srinath
19207638

Abstract

The Health Services Executives (HSE) endorses health and well-being throughout the country. It manages resources in a way that delivers the best health outcomes in Ireland. The Acute Hospitals Division works meticulously with the Hospital Groups to deliver the hospital services like Inpatient and Outpatient care throughout Ireland. This division not only improves the waiting time for the public hospitals, but also ensures quality, safety, and the financial operations of the Hospital Groups. This research project aims to deliver the resolution for one of the crucial matters that the HSE prioritizes - long patient waiting lists. This research focuses only on outpatient waiting lists. The key role that HSE is responsible for is to guarantee timely access for patients to receive treatment and care because currently, over 3.2 million patients wait for outpatient care and treatment. The National Treatment Purchase Fund (NTPF) works with the HSE and delivers the data for the outpatient services annually. The proposed framework combines time series analysis using seasonal ARIMA along with supervised ML algorithms like ANN, Random Forest, and Linear Regression to study the data from 2014-2021. Results are presented for the four models based on RMSE AND MAE. The random forest model outperformed the other models with an RMSE of 57.97. This research shows promise for seasonal ARIMA and random forest models to help the hospital administrators make appropriate changes to enhance the hospital systems.

1 Introduction

Patients who do not need any emergency treatments and cannot be given any immediate medical treatment are usually listed as a formal record in all the hospitals. In an outpatient clinic, the time period between a patient requesting treatment and the actual appointment is months together. When the consultants give an opinion to the patient, either the appointment date is given immediately or the patient's name is put on the waiting list. The list consists of patients in medical priority-based order to receive the appointment date. The fact that some patients will have to keep needing medical service throughout their lives only makes the matters worse (Wang et al.; 2015). The main motivation for this research is that there is a significant rise in the need for medical services. There is a rise of over 8 percent since last year. The Irish Times published on the 17th of November, 2021 that around 900,000 people are on the waiting lists in the hospitals across the Republic. The consultants have warned that this crisis will only worsen.

The National Treatment Purchase Fund (NTPF) is an independent corporate legal body

with certain responsibilities which was established by the Minister for Health in Ireland. The primary function of the NTPF is to collect and validate the data on all the patients waiting for treatment from public hospitals across Ireland. They also advise the Minister for Health and perform any other task assigned to them by the Minister. Hence the NTPF works meticulously with the Department of Health, the HSE, the public hospitals, and the private nursing homes across Ireland. The NTPF currently collects data for Inpatient, Outpatient, Day Cases, and Planned Procedures.

The Outpatient Waiting List consists of both patients waiting for an appointment date and patients who have been scheduled an appointment date for their treatment. The data report that NTPF presents every month represents the total number of people waiting for a first-time appointment with a consultant. The acute hospital system focuses on this area of service, i.e., delivering outpatient treatment and care. They set up the Outpatient Services Performance Improvement Programme 2016-2020. They handle the standardization in the delivery of the service provided, and they also focus on how the outpatient services could be delivered in places like the Primary Care Centres and GP Practices which are outside the hospitals. The Waiting List Action Plan has been developed by the Department of Health along with the Health Service Executive (HSE) and the NTPF. Hospital Groups and specialties will assess the waiting lists in each of the hospitals to reduce the list and also remove any duplications and make the list more effective by ensuring that the consultants schedule the patients appropriately. To improve the patient care in healthcare settings, forecasting methods are used. Time Series forecasting is found to be vital to manage a healthcare organization (Claudio et al.; 2014). Scientific predictions can be made using these forecasting methods on historical time-stamped data.

Statistical techniques are required to plan the admissions of the patients (Wargon et al.; 2009). This is because of the reason that there are about 112,513 people waiting for planned procedures or pre-admissions[1]. These pre-admissions are for the patients who have received treatment and require additional care for a future date. The main objective of this research is to conduct a thorough analysis of the data presented by the NTPF. The major contribution of this research is to collect the data from NTPF and utilize it to perform a few novel analysis using forecasting techniques like seasonal ARIMA to build a forecasting model and also study the existing correlations in the data using three machine learning algorithms – Artificial Neural Network, Random Forest, and Linear Regression. This research aims to demonstrate different forecasting strategies to predict the number of patients and compare these results of various methods in order to emphasize the influence of forecasting techniques on the healthcare system in Ireland. This improved understanding of the outcome on this study will help hospitals to handle the overcrowding and inefficient systems by making suitable operational adjustments. Waiting lists have been established mainly due to the poor ability of the consultants to control their workload. Also, the administration has to make the treatment plan for all the patients on the waiting lists with making sure the appropriate medical resources like equipment, bed, and staffing are available.

The research question posed in this study investigates the following:
To what extent can time series analysis help in accurately forecasting the outpatients waiting to receive treatment for anywhere between 0-12 months in hospitals across Ireland? Also, can machine learning algorithms like Artificial Neural Network, Random Forest,

[1]https://www.irishtimes.com/news/health

and Linear Regression be used to predict the number of outpatients in the waiting list and outperform the time series model?

The rest of the paper is structured as follows:
Section 2 presents the related work and its impact on this project;
Section 3 describes the research methodology;
Section 4 explains the design specification in detail;
Section 5 shows the Implementation of this study;
Section 6 displays the Evaluation and discussion;
Finally, Section 7 concludes the paper.

2 Related Work

This section of literature review is mainly to demonstrate the evidence of the selected research topic by critically analysing the potential of this study and highlight if any the research gaps from previous studies. This central section conveys any kind of conflicts or reciprocated questions towards this study. This research has three main subsections of the methods utilized in the implementation and are categorized as follows:

- Statistical Analysis
- Machine Learning Algorithm
- Importance of Forecasting and Patient Satisfaction

2.1 Statistical Analysis

A time series is a sequence of observations on a variable measured at successive points in time or over successive periods of time in statistics. The data can be stationary or non-stationary. This paper by Elgohari et al. (2019) utilized to models, the non-seasonal ARIMA(p,d,q) and an exponential smoothing models. For exponential smoothing model, isolation of the seasonality and trend was done from the irregular variations with 10 observations for the moving average. For ARIMA, the Box-Jenkins methodology was used. ARIMA (3,1,3) was the model with the best results with the lowest RMSE value of 1100.832. A Box-Ljung test was used to check the model for unfitting which showed a significant difference in the parameters used for moving average. This predicted an increase in the visits for the following months in the tertiary hospital.

Similarly, the best model was the ARIMA model of the fourth order i.e., ARIMA (4,2,0) at a clinic in Saudi Arabia shown by the study of Abdel-Aal and Mangoud (1998). The MAPE value was found to be 4.23 percent but gives a better value of 1.17 percent in a polynomial fit by extrapolating the growth curve. It was a monthly dataset over a period of 9 years. Since this was a stationary time series, Univariate UBJ-ARIMA model was fitted.

It is difficult to predict the overcrowding in hospitals. This paper by Schweigler et al. (2009) proved how time series perform in emergency department visits by building three models. The first was the average for very hour, second was an ARIMA model and third was a sinusoidal AR model. These models were tested with the accuracy for 4- and 12-hour predictions for a year for the data collected from three tertiary hospitals. The main

limitation of this study was that a higher order ARIMA would have performed better with a lower error rate. The model predicting the occupancy was the result but did not show any factors that was responsible for the overcrowding. The historical average performed slightly less than the AR models with ANOVA p value less than 0.01.

Zhou et al. (2018) went further ahead to combine an ARIMA model with a nonlinear autoregressive neural network (NARNN). This hybrid model was meant to forecast the new admission patients. Along with this model, they also developed an SARIMA and compared the models using RMSE and MAE. The hybrid model did not necessarily outperform the SARIMA although the RMSE value was low. ANN has enhanced the accuracy because of the intrinsic property. ARIMA-ANN hybrid was also used in the research by Yucesan et al. (2018). They also built the ARIMA-LR model. The MAPE values were 0.49 percent and 0.92 percent for both the models respectively. The main limitation in this paper was however, not all the important variables were not taken into consierdation for modelling of the ARIMA. ARIMA works better with ANN because of the limitation of linearity.

However SARIMA performed well, and to back this result, the research study by Zhu et al. (2015) developed three models. The SARIMA, multiplicative seasonal ARIMA (MSARIMA) and combination of MSARIMA with Markov Chain model. MAPE, normalized MSE were used to evaluate the model. The combinational model performed the best to forecast the bed availability in hospitals as well. It was noticed that the combination method works better even in the study by Li et al. (2014). With ARIMA, a time series modeler in Predictive Analytics Software (PAWS) was used by setting the required parameters to build the forecasting model for the next two years. The SARIMA is sensitive to the parameters used. Postma et al. (1995) showed in his study that this model was not suitable for all the conditions and all the group parameters.

2.2 Machine Learning Algorithm

With large data, supervised machine learning is the most powerful tool. This was put to test in a paediatric ophthalmology outpatient clinic. Many models were built including random forest, multiple linear regression and support vector machine and elastic net. K-fold cross validation was used to validate the data to avoid overfitting. RMSE and R square, along with ROC were evaluated. Out of all the models, the random forest model displayed the highest accuracy with an RMSE value of 24.22. The study by Lin et al. (2019) was able to predict all the factors which contribute to the long waiting lists by using machine learning and also ranked the features which were important depending on the MSE values. To further prove the effectiveness of random forest, the research by Patil and Thakur (2019) used a minimized path-awareness method they proposed an improved random forest method called as Incremental Patient Treatment Time Prediction algorithm (IPTTP). This also suggested a Queuing-Recommendation system for the patients with a treatment action plan to reduce the waiting time in India with a low-latency reaction. Machine learning usually is a suitable option for sophisticated and noisy data like waiting list data.

The Journal by Sukmak et al. (2015) worked on time series forecasting using Data Mining. They first constructed two ANN models, Radial Basis Function (RBF) and Multi-Layer Perceptron networks (MLP). RBF was selected depending on the Mean Absolute Percentage Error values. The study pointed out that the traditional time series methods are

not that efficient with linguistic data. This data mining approach is more flexible and powerful hence ANN has been utilized to solve this complex problem with less error rate in a psychiatric hospital. A feed-forward neural network called as RBFregressor has three layers which was used in this study. MAPE was the main evaluation method used and got an average of 35.32 percent MAPE value for all the models.

Similar to this, ANN was used to predict the healthcare visits in Silobela District Hospital NYONI and NYONI (n.d.). They needed a more efficient resource planning because a high number of visits were forecasted. Multi-Layer Perceptron Neural Network under the Feedforward Neural Networks was used to study the data of all age groups in 2019. With about 96 observations, it was predicted that there would be a rise in the visits in the next 24 months, at least 343 visits per month and with a highest of 927 visits and this study was warning sign to the department of health.

2.3 Importance of Forecasting and Patient Satisfaction

This particular problem having a negative effect has increased the attention of the public. Patient waiting time, the service time and patient satisfaction with the service are all correlated and it is important to obtain a better understanding on this relation. Xie and Or (2017) conducted a study on the outpatients in a teaching hospital in China. The study proved that the amount of time the patients were waiting to receive treatment and care was not acceptable and that the patient dissatisfaction could be overcome by increasing the resources of the hospital. They also stated that the hospital has to be more transparent to the patients about the treatment plan and about the healthcare professional. In order to reduce the dissatisfaction of the patients and make it easier on them while they wait, the care providers have to be more empathetic and courteous. It was concluded that patients spent less time in getting the service whereas spent hours in waiting for the prescription slip.

Although there are many studies conducted to increase outpatient satisfaction, they were not backed by robust methods. Hence, Author Sun et al. (2017) performed analysis on a longitudinal time series data to notice the trends using a segmented linear regression model. The strength of the relationship between waiting times and patient satisfaction was determined using Pearson correlation analysis. The outcome displayed a well-designed improvement plan for the management system. It is therefore important to use appropriate forecasting methods to reduce patient dissatisfaction since they are correlated to each other.

2.4 Conclusion to Related Work

The ARIMA(p,d,q) model was widely used in many of the research papers to forecast the outpatient visits. The flexibility and accuracy of the model was validated. Many of the papers used a combinational model of ARIMA and neural network to obtain the respective results. Currently this kind of problem has been implemented by both time series models and general linear models (Guan and Engelhardt; 2019). Out of which most common one is Random Forest in the supervised Machine Learning algorithm which also performed well hence these are the models implemented in this research study. Table 1 shows the summary of the related work.

AUTHOR/YEAR	COUNTRY	METHOD	MAIN FINDINGS
Lin et al., (2020)	Portland	Random Forest, Elastic net, Gradient Boosting Machine, Support Vector Machine, and Multiple Linear Regression	Random forest outperformed all models.
Nyoni et al., (2020)	Zimbabwe	ANN	Suitable for the non-linear dataset and the MSE was found to be 738.3.
Elgohari et al., (2019)	Egypt	ARIMA and Exponential (single, holt's, and dampened trend) models	ARIMA (3,1,3) was the best fitting model and efficient.
Sukmak et al., (2015)	Thailand	Two ANN models: Multi-Layer Perceptron networks (MLP) and Radial Basis Function (RBF)	RBF proved a superior model with MAPE <20% and was selected as the final model.
Abdel-Aal and Mangoud (1998)	Saudi Arabia	ARIMA	ARIMA (4,2,0) resulted in greater accuracy.
Patil and Thakur (2019)	India	Improved Random Forest model called Incremental Patient Treatment Time Prediction (IPTTP) algorithm	minimized path-awareness, acquiescence effectiveness, and low-latency reaction.
Zhou et al., (2018)	China	ASRIMA NARNN SARIMA-NARNN	Hybrid model was the better model.
Zhu et al., (2015)	China	Three models: model combining seasonal regression and ARIMA, multiplicative seasonal ARIMA (MSARIMA) model, and combinatorial model based on MSARIMA and weighted Markov Chain.	The combinational model gave improved results.

Table 1: Summary of the Related Work

3 Methodology

A methodology of any research is meant to deliver a holistic view of the entire process of discovering knowledge. This methodology trails the Cross Industry Process for Data Mining (CRISP-DM) approach but has a unique workflow from analyzing the raw data to providing detailed information on the knowledge that has been discovered during the research. Slightly contrasting with the typical CRISP-DM (Huber et al.; 2019), this methodology translates the current business problem into a Data Mining task by examining the data with exploratory data analysis. The research methodology consists of six namely Business Value, Data Preparation, Exploratory Data Analysis, Modelling, Evaluation, and Deployment as shown in Figure 1.

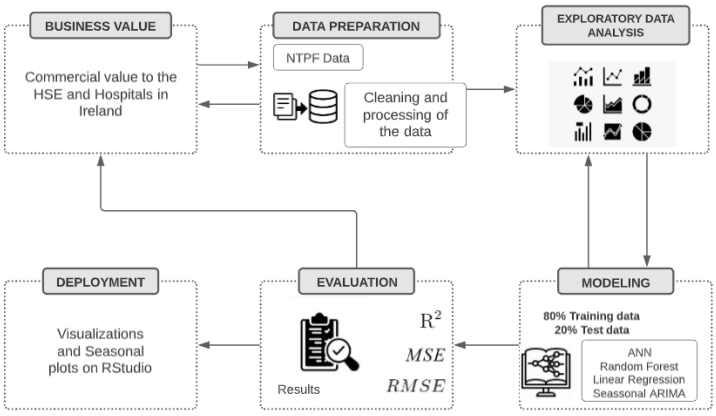


Figure 1: Outpatient Waiting List Methodology

The first step in this methodology is **Business Value**. The fact that in Ireland, the waiting time between a referral to the actual appointment date to see the doctor is months together. The situation can get worse when the patients need to revisit for the treatment, sometime throughout their lives (Wang et al.; 2015). It is crucial for the department of health to keep a close watch on the numbers to make sure it is well managed throughout the country. This is one of the reasons why the HSE works closely with the NTPF to collect the data from all the public hospitals since 2014 till date. They keep this data available to the public so companies can benefit from this and work towards a better health care system for the country. Analyzing this data helps the executives to work and plan accordingly to the needs and resources for the hospitals. Since Ireland is known for the health services for its residents, it is important for the administrations to strategize for a better efficient system to reduce these long waiting lists.

The second step is **Data Preparation**. The data was obtained from the official website of The National Treatment Purchase Fund[2]. It is approved by the department of health in Ireland. This data consists of 8 CSV files for each year from 2014-2021. The data shows the total number of people waiting for a first-time appointment in any consultant-led Outpatient clinic across Ireland in different time bands. The NTPF complies with the Regulations of the Re-use of Public Sector Information, and also encourages the re-use of the data that they produce for free of charge. Figure 2 below displays the description of the dataset.

Data	Description
Date	The date of the data recorded
Group	The hospitals in Ireland are organized into seven hospital groups. This displays in which group the hospital belongs to.
Hospital HIPE	This is the Hospital Inpatient Enquiry number which denoted the principal source of data information from the hospitals
Hospital	This denotes which hospital in Ireland the appointment is taken for.
Specialty HIPE	This number denoted the HIPE for a specific specialty.
Specialty	This is the defined specialty of the consultant.
Adult/Child	This is a binary variable that indicates if the patient is an adult or a child.
Age Profile	This categorizes the age of the patient.
Time Bands	This column depicts the various time frames the patients have to wait for the treatment plan, for example, 3-6 months.
Count	This is the total number of patients on the waiting list.

Figure 2: Data Description

Once the required packages were imported, the 8 CSV files were merged into one file and the dataset was checked for any null or missing values and duplicate values, and it was handled accordingly since they were few missing values. The dataset now contains 582729 rows and 10 columns after it is cleaned and ready for the exploratory analysis. Figure 3 illustrates a snippet of the dataset.

```
In [5]: #Printing the first 5 rows of the dataset
        df.head()
```

Out[5]:

	Archive Date	Group	Hospital HIPE	Hospital	Specialty HIPE	Specialty	Adult/Child	Age Categorisation	Time Bands	Count
0	2014-01-31	Children's Hospital Group	940	Childrens University Hospital Temple Street	NaN	Other	Child	0-15	0-3 Months	4
1	2014-01-31	Children's Hospital Group	940	Childrens University Hospital Temple Street	NaN	Other	Child	0-15	3-6 Months	1
2	2014-01-31	Children's Hospital Group	940	Childrens University Hospital Temple Street	NaN	Other	Child	0-15	6-9 Months	1
3	2014-01-31	Children's Hospital Group	940	Childrens University Hospital Temple Street	NaN	Other	Child	16-64	0-3 Months	1
4	2014-01-31	Children's Hospital Group	940	Childrens University Hospital Temple Street	100.0	Cardiology	Child	0-15	0-3 Months	193

Figure 3: Snippet of the dataset

[2]https://www.ntpf.ie/home/outpatient$_g roup.htm$

Figure 4 displays the summary of the dataset as in, the column name and type of the data.

```
#Printing the summary of the dataset

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 582729 entries, 0 to 22520
Data columns (total 10 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Archive Date        582729 non-null  object
 1   Group               582729 non-null  object
 2   Hospital HIPE       582729 non-null  int64
 3   Hospital            582729 non-null  object
 4   Specialty HIPE      581507 non-null  float64
 5   Specialty           582729 non-null  object
 6   Adult/Child         582729 non-null  object
 7   Age Categorisation  582126 non-null  object
 8   Time Bands          582727 non-null  object
 9   Count               582729 non-null  int64
dtypes: float64(1), int64(2), object(7)
memory usage: 48.9+ MB
```

Figure 4: Summary of the dataset

The third step is **Exploratory Data Analysis**. It is critical to study and analysis the data and its structure before building any models. A few plots were displayed, and a thorough analysis was done to understand the different patterns in the data. This examination of the data is plotted with a few bar charts, box plots and pie charts to maximise the insights from the data.

The fourth step is **Modelling**. One hot encoding was implemented to a few columns and joined back to the main data frame. The data was then split into 80 percent training data and 20 percent test data. Three supervised machine learning models were applied, i.e. Artificial Neural Network, Random Forest, and Linear Regression. After this, the data was then formatted and converted into four separate time series objects to perform a forecasting analysis on the data and fit a seasonal ARIMA model for the four time bands 0-3, 3-6, 6-9, and 9-12 months waiting time. ANN and Random Forest algorithms were implemented because from the related work, it was observed they were reliable and gave good results. Linear regression was selected since it has a noticeably lower time complexity than most machine learning algorithms. For the reason that most of the literature papers suggested a combinational model of ARIMA with either NARNN (nonlinear Autoregression Neural Network) or MSARIMA (Multiplicative seasonal ARIMA), the aim was to try and implement only a simple seasonal ARIMA and observe if it is possible to get a best fit for this data.

The fifth step is **Evaluation**. The main evaluation methods used were Mean Absolute Error, Mean Square Error, Root Mean Square Error and R square value. Table 2 shown below demonstrates the calculation of these evaluation metrics.

- MAE – The error of the predictions of every instance of the dataset is the absolute values, and the mean of this is referred to as the Mean Absolute Error. Lower values of MAE give a better model.

- MSE - This is evaluated by taking the average of the squared differences between the predicted and expected target values in the NTPF dataset. If the MSE value is closest to 0, then it is a good model.

- RMSE – This is used to report the performance of the model and is an important metric to represent the best model since it is an absolute measure of the model. The RMSE value of anywhere between 0.2 to 0.5 is said to accurately predict the data.
- R2 Score – It is the coefficient of determination. This statistical measure shows that the predictions fit the data well.

$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}.$
$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$	$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i \rvert$

Table 2: Calculation of Evaluation Metrics

Evaluation of the seasonal ARIMA can be mainly done by reducing the AIC (Akaike Information Criterion) value of the model. Although RMSE value will be the metric of evaluation, the plots considered are the Normal Q-Q plot and the Residual plots. The Normal Q-Q plot show if the data is normally distributed. In the residual plot for the model, ACF (Auto Correlation Function) is specifically plotted for the residuals. It is important for the lags to be within the blue dotted significant lines which indicates that there are no autocorrelation patterns in the residuals. The residuals also should be normally distributed. If there is a spike in the ACF plot of the residuals it means that there is some part of the time series that has not been captured by the specified model.

The final sixth step is **Deployment**. This last step is important to deploy the project and present the information gathered to help the healthcare service in Ireland. This can be deployed to various hospitals across the country so that they are informed about the number of appointments they can expect and be prepared with appropriate resources and treatments. This will widely benefit the government to manage more systematically.

4 Design Specification

The architectural framework of this research is represented in Figure 5 which consists of a combination of three supervised machine learning models i.e. ANN, Random Forest and Linear Regression along with time series seasonal ARIMA model.

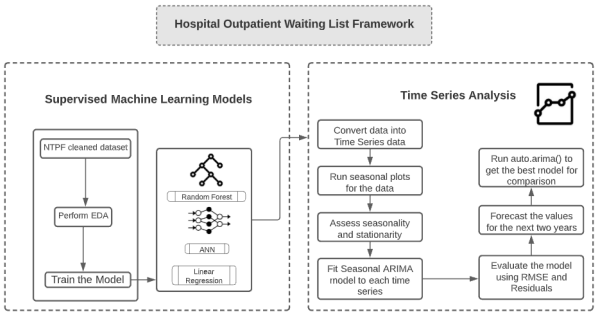


Figure 5: Architectural Framework of the Outpatient Waiting List

Artificial Neural Network: ANN is a machine learning algorithm which works similar to a human neuron. This algorithm is organised into layers. It usually can consist of an input layer, one or more hidden layers and an output layer. The input layer takes in the input values and communicates with the hidden layers. In the hidden layer, the information is transferred to different nodes and multiplied by the weights. They then link to the output layers and produce the results. The ANN model implemented is a simple neural network which consists of four layers. Since the data is not very complex and also has less dimensions, only two hidden layers were implemented. The first layer contains 128 nodes, the second layer consists of 64 nodes, third consists of 16 nodes and the last layer contains only 1 mode which is the output layer. This is demonstrated in the Figure 6 shown below.

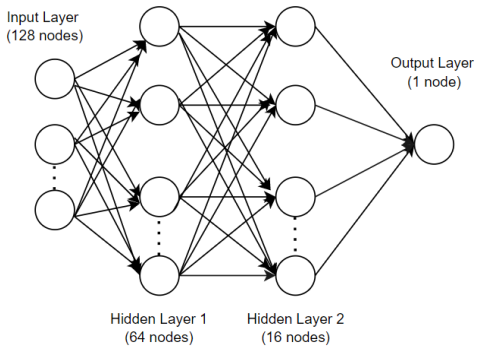


Figure 6: Implemented ANN Model

Random Forest: It is a supervised ensemble machine learning algorithm that consists of many individual decision trees. It builds a forest of uncorrelated trees by building each tree and gives out a higher accuracy. It is especially very useful while handling large datasets. While growing the 100 trees in the model, the regressor adds more randomness to the model. As in, while splitting a node, it looks for the best feature from any random subset of the features rather than focusing on the most essential feature. The implemented random forest model is demonstrated in Figure 7.

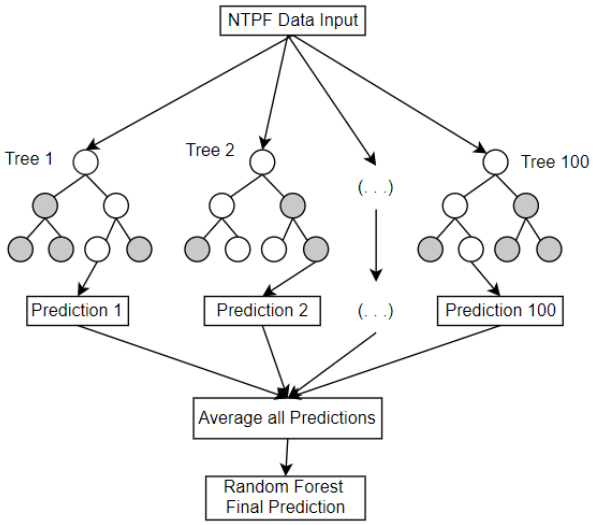


Figure 7: Implemented Random Forest Model

Linear Regression: This is a machine learning technique that provides the best fit for a linear data. This linear relationship is between the dependent and independent variable. However, this model is very sensitive to outliers . It did not provide a good result as there wasn't much linearity between the dependent and independent variable.

Seasonal ARIMA: Autoregressive integrated moving average (ARIMA) is a forecasting model in which the predicted values are a linear function of the actual values and errors of prediction that are called as residuals. The forecast errors consist of lags called as moving average. The Augmented Dickey-Fuller test (ADF) was used to evaluate the assumption of stationary. The Ljung-Box test was used to check if the autocorrelations are close to zero because that means the model has fit the data well. The ARIMA model used in this implementation is a seasonal ARIMA (p,d,q) (P,D,Q) m model where m is the number of observations per year which in this case is 12, p is the order of the autoregressive part, d is the degree of differencing and q is the order of moving average part. The evaluation plots for seasonal ARIMA are Normal Q-Q plots along with the forecasting plot which consists of the predicted values for the next 24 months.

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part of} \\ \text{of the model}}}$$

5 Implementation

5.1 Implementation of Machine learning

The implementation for this section has been done using Python programming language using the pandas python package which is very intuitive and designed to be flexible. This has been implemented on Jupyter Notebook. The dataset was loaded in 8 separate CSV files and the first step was to merge these files into a single dataframe. A few rows were dropped because of missing values. The graphical representation was done on Jupyter Notebook using the Matplotlib and Seaborn libraries. Out of the seven hospital groups in Ireland –(Ireland East Hospital Group, RCSI Hospitals Group, Dublin Midlands Hospital Group, University Limerick Hospitals Group, South/South West Hospital Group, Saolta University Healthcare Group and Children's Health Ireland)- the exploratory data analysis showed that the highest number of patients waiting were under the South/South West Hospital Group as shown below in Figure 8.

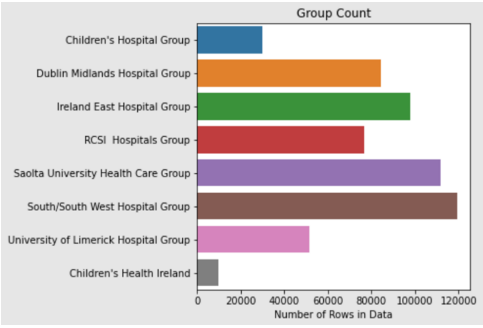
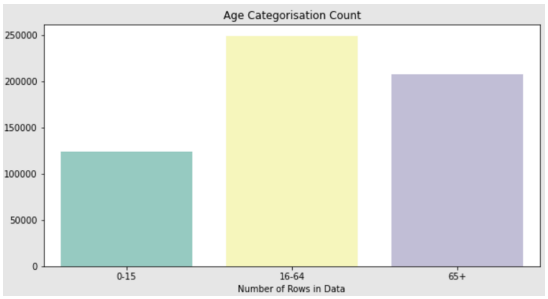
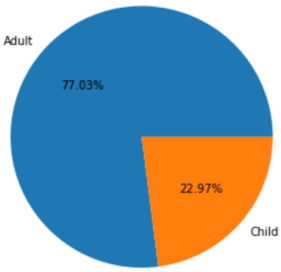


Figure 8: Count in each Hospital Groups

The majority of the patients waiting for treatment are in the age groups between 16 and 64 as represented below in Figure 9a. Out of which 77.03 percent are adults and 22.97 percent are children as shown in the pie chart below in Figure 9b.



(a) Patients in different age bands



(b) Division of Adults and Children

Figure 9: Age Categorization Plots

The below graph in Figure 10 shows that only a few hospitals have more patients waiting for an appointment. The data analysis shows that most number of people are on the waiting list for 0-3 months and the demanded specialties are ENT, Ophthalmology and General Surgery.

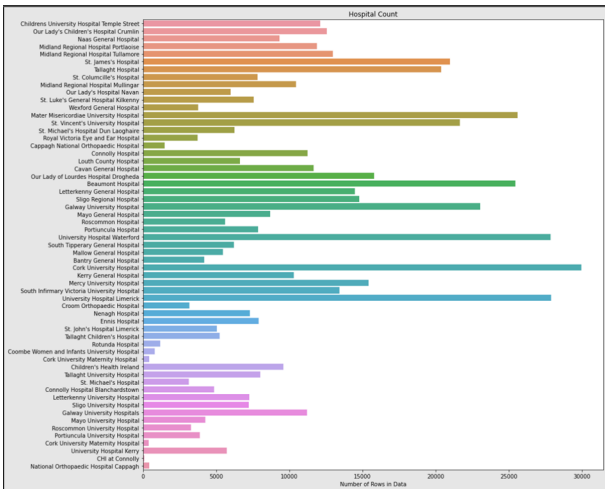


Figure 10: Count in each Hospital

Before the model building, four of the categorical columns –(Group, Adult/Child, Age Categorisation, and Time Bands)- were converted into 20 extra new columns of binary values of 0 or 1 in which these different categories have a value. The data is then converted into 80 percent train and 20 percent test.

For ANN, Keras is mainly a neural network library and the sequential model was implemented by passing a stack of 4 layers to the sequential function. The layers were incrementally stacked by using the add() method and the summary() method was called to display the contents of the model. It is usually a good practice to include the input shape of the model. This helps the model to continuously have weights along with a defined output shape because the model does not display the weights generally. This model was made to run for 100 epochs. In Artificial Neural Networks, one iteration over the entire 80 percent of training data usually makes for 1 epoch and for ANN a few more

than 1 epoch is required to evaluate the model. The pass counts both forward pass and backward pass. Figure 11 shows the training of the ANN model.

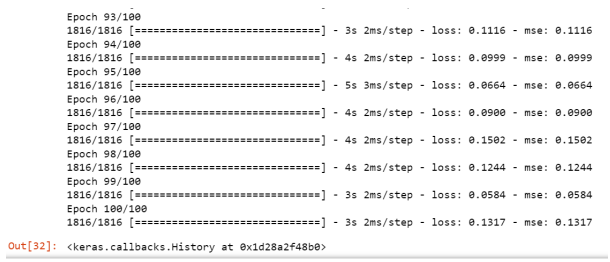


Figure 11: Training ANN Model

To compile a keras model, out of the two arguments, an optimizer is required. It can be passed in by the string identifier or instantiated before passing in model.compile(). For this implementation, the optimizer used is called Adam. The first-order and second-order moments in the adaptive estimation is from the stochastic gradient method which is the Adam optimization. The loss function was set to 'mean_squared_error' to calculate the quantity that the ANN model should try to minimize during the training of the model. The next model built is Random Forest by importing the random forest regressor from the sklearn ensemble module. While growing the 100 trees in the model, the regressor adds more randomness to the model. This process leads to a better model due to the wide diversity. Figure 12 shows the training of the random forest model.

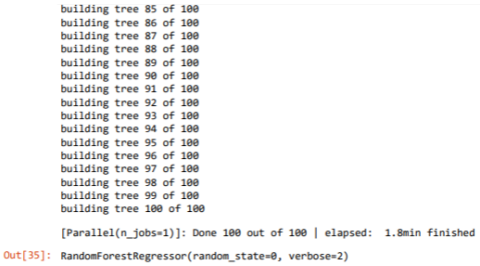


Figure 12: Training Random Forest Model

The final model is Linear Regression. Linear Regression was imported from sklearn.linear_model. This is a linear model and the data is fit to the supervised linear regression algorithm.

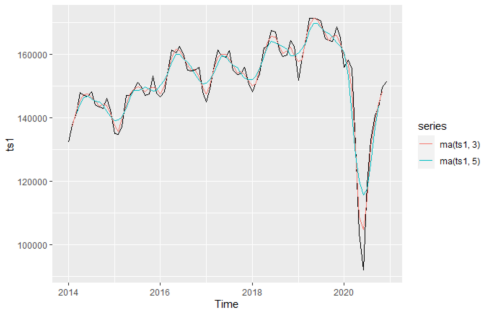
5.2 Implementation of Time Series

This part of the implementation is conducted using the R programming language on an IDE called as RStudio. All the required libraries like haven, tseries, and dplyr were imported. The tseries library is required for a time-series data analysis. The dplyr library helps to solve many challenges regarding data manipulation. After reading the data in the CSV format, it is necessary to understand the data. The date needed to be rearranged and formatted. Two missing values were deleted from the data. The data for the year 2021 was not considered for this section of the study due to a criminal cyber-attack on the HSE that took place in May 2021. The public hospitals were unable to provide and deliver the data for a few months hence the study was conducted from 2014 till 2020.

This time series has been experimented for four different time bands, i.e., patients waiting

from 0-3, 3-6, 6-9, and 9-12 months. Four separate data-frames were formed for each of the time bands and a time series object was created. In R, it is essential to create a time series object by declaring the observations, starting time of the series, and its periodicity where frequency will be 1 for annual, 12 for monthly and so on. The pattern otherwise known as trend of the time series is important to understand the behavior of the data.

One of the simplest methods of smoothing a time series was used – simple moving averages. This is to smooth a curve that damps down these fluctuations. As k increases, the plot becomes increasingly smoothed. Given below in Figure 13 are the simple moving average plots for each of the time series when k=3 and k=5.



(a) Plot for time series 1

(b) Plot for time series 2

(c) Plot for time series 3

(d) Plot for time series 4

Figure 13: Simple Moving Average Plots

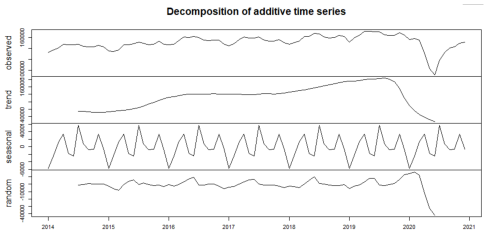
The models depicted a additive model because the seasonal fluctuations did not depend on the level of time series. A seasonal ARIMA model was fit after checking the order of differencing and the Augmented Dickey Fuller (ADF) test to test stationarity. The checkresiduals() function was used to obtain the residuals plot. The forecasting values were plotted and the accuracy was calculated.

5.2.1 Seasonality

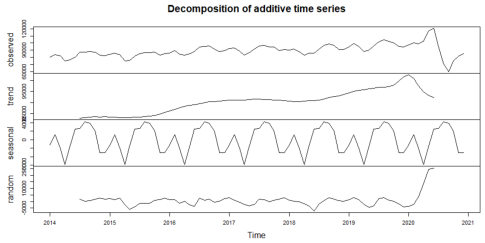
In times series, there are usually three main parameters – Trend, Seasonality, and Cyclic. The random fluctuations and the gradual shifts in the plot are said to exhibit a trend pattern. If this pattern is affected by seasonal factors, the plot is said to be seasonal. There are various seasonal plots like the ggseasonplot() and the ggsubseriesplot(). If there is a rise and fall in the pattern without a fixed frequency, it is said to be cyclic.

Since the dataset consists of a seasonal component, the seasonal decomposition can be

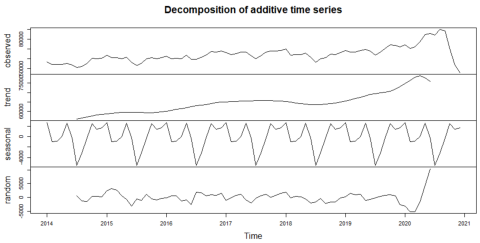
either additive or multiplicative. For the additive decomposition, there is no increase in the amplitude of the plot, whereas, the multiplicative decomposition consists of the increase in the amplitude of the plot. Hence, additive decomposition was necessary for the four time series. These plots in Figure 14 show the time series, seasonal, trend, and irregular components for each of the four time bands.



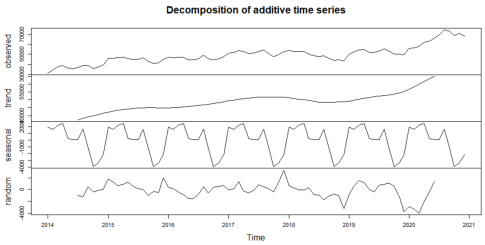
(a) Plot for time series 1



(b) Plot for time series 2



(c) Plot for time series 3

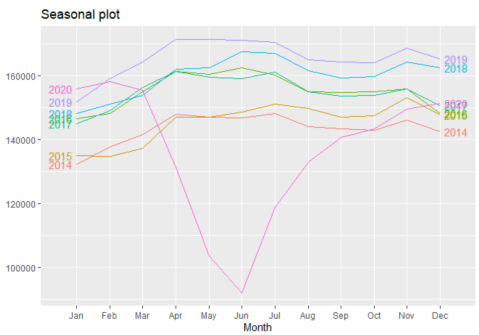


(d) Plot for time series 4

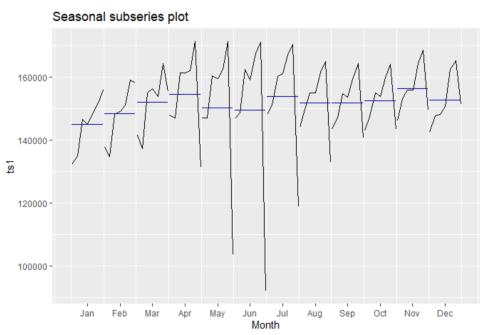
Figure 14: Decomposition for Additive Time Series

A few seasonal plots were displayed for each of the four time series in order to observe the fluctuations in the data. The ggseasonplot() is similar to a time plot but here the data is plotted against the individual seasons for a separate time in years. The seasonal subseries plots are plotted using ggsubseriesplot(). This emphasizes the seasonal patterns where the data for every season are collected together in separate mini time plots. The blue horizontal lines indicate the means for each month. This form of plot enables the underlying seasonal pattern to be seen clearly, and also shows the changes in seasonality over time. It is especially useful in identifying changes within particular seasons.

Demonstrated below in Figures 15-18 are the seasonal plots and seasonal subseries plots for each time series.

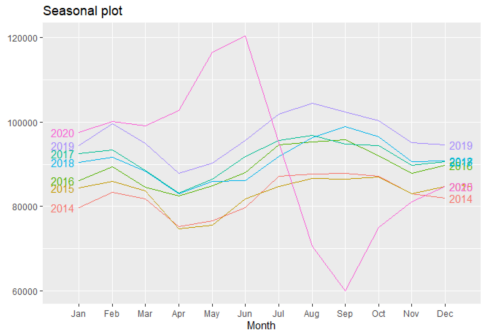


(a) Seasonal Plot

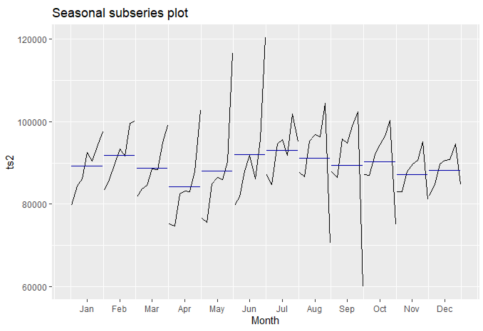


(b) Seasonal Subseries Plot

Figure 15: Plots for Time Series 1

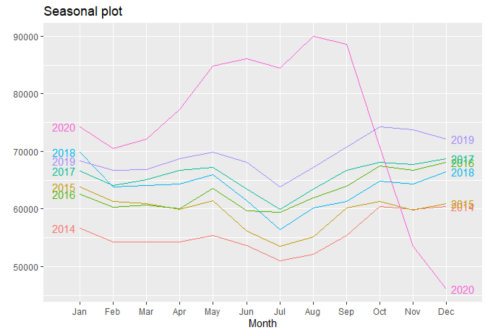


(a) Seasonal Plot

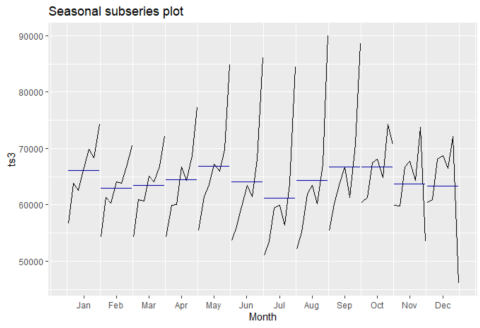


(b) Seasonal Subseries Plot

Figure 16: Plots for Time Series 2

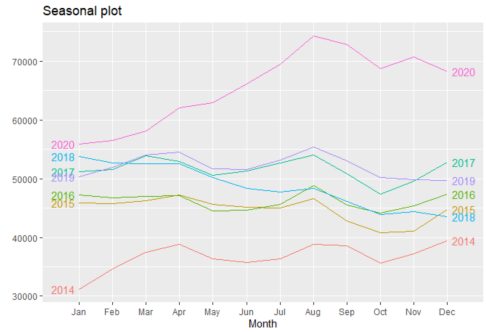


(a) Seasonal Plot

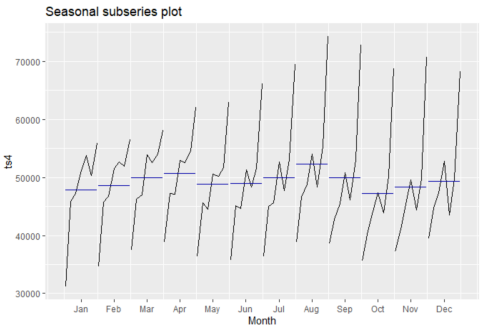


(b) Seasonal Subseries Plot

Figure 17: Plots for Time Series 3



(a) Seasonal Plot



(b) Seasonal Subseries Plot

Figure 18: Plots for Time Series 4

Figure 15a shows that in 2020, there was a sudden decrease in the number of patients waiting for treatment in June. Apart from that it has been constant in the number of people waiting for 0-3 months. However, for people waiting in the time period of 3-6 months (Figure 16a), in 2020 there was a sudden increase seen in Jun and then a sudden decrease in September. All the years are quite consistent on how long the patients wait for the other time periods. The number of people on the waiting list for treatment for over 9-12 months was less in 2014 but has seen a steady increase and the highest in 2020.

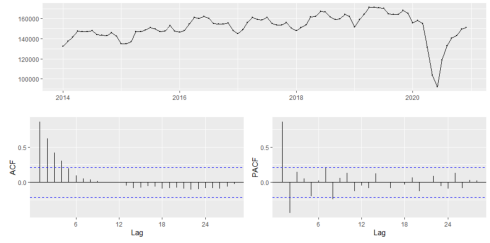
5.2.2 Stationarity

To check whether the time series is stationary, it can be made sure that they have the same mean, variance, and covariance in every point. To make future predictions, it is important to make sure the series is stationary. One of the most common tests for stationarity is the Augmented Dickey Fuller Test (ADF). This Unit Root test consists two hypothesis, the null hypothesis that the time series is non-stationary and the alternate hypothesis that the time series is stationary. Since a few values are greater than 0.05, the Null hypothesis is not rejected and the series is said to be non stationary.

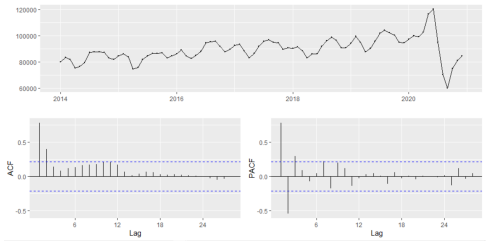
ADF p-value for time band 0-3 months: 0.123.
ADF p-value for time band 3-6 months: 0.02224.
ADF p-value for time band 6-9 months: 0.01.
ADF p-value for time band 9-12 months: 0.5317.

5.2.3 Autocorrelations

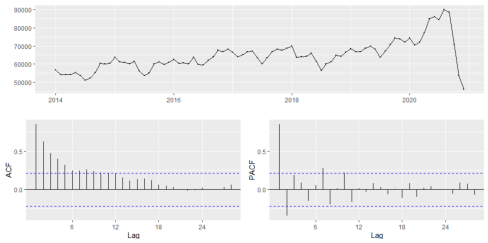
The relations of the time series with the past values are called as autocorrelations. The autocorrelations are plotted according to the period lags. The correlogram or the ACF plot can be plotted using the acf() function. For a shorter lag length, the pacf() is plotted which is the partial autocorrelation function. From these plots, the order of differencing can be calculated manually and be used to construct the seasonal ARIMA(p,d,q)(P,D,Q)m model. Figure 19 displays the ACF and PACF plots by using the ggtsdisplay() function.



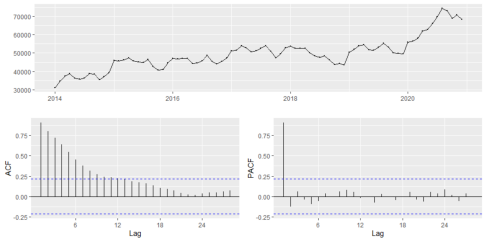
(a) Plot for Time Series 1



(b) Plot for Time Series 2



(c) Plot for Time Series 3



(d) Plot for Time Series 4

Figure 19: ACF and PACF Plots

The auto.arima() function was also applied which is a shorter model that reduces the search time. It uses the Alkaike Information Criterion and the Bayesian Information Criterion to try various values of p,d,q, and find the best model. These steps were implemented for all the four time-bands in order to forecast the values for patients waiting for the time span of 0-3, 3-6, 6-9, and 9-12 months for treatment.

6 Evaluation

6.1 Results for Machine Learning

For regression problems, accuracy cannot be calculated directly. It has to be in terms of error for the predictions. It is important to know how close the predictions were to the expected values. The evaluation metrics used are Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R2 Score. The aim of this study being able to forecast the values for the next two years and also checking if the supervised machine learning algorithm can outperform the seasonal ARIMA time series model.

ANN and Random Forest exhibited a better MAE value compared to Linear Regression. From the table 3 given below, it is observed that both ANN and Random Forest performed well in terms of MSE. However, Linear regression having a value of 14579.12 did not perform well. The residuals are not much spread out for ANN and Random Forest thus performing well but for linear regression, the deviation of the residuals is high with a value of 120.74.

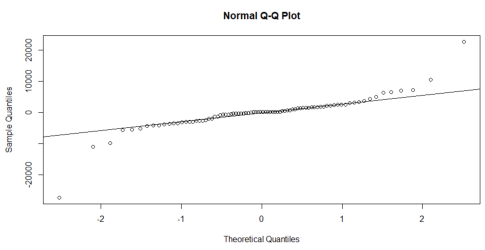
MODEL IMPLEMENTED	MEAN ABSOLUTE ERROR – (MAE)	MEAN SQUARE ERROR – (MSE)	ROOT MEAN SQUARE ERROR – (RMSE)	R^2 SCORE
ANN	50.92	9515.46	97.54	0.42
RANDOM FOREST	22.79	3361.18	57.97	0.79
LINEAR REGRESSION	70.31	14579.12	120.74	0.11

Table 3: Evaluation metrics for ML Models

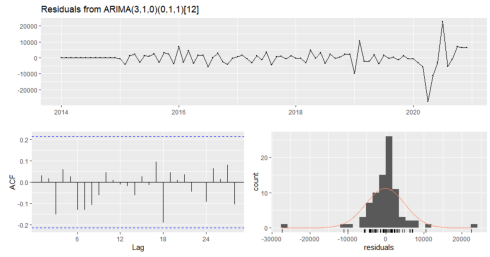
6.2 Results for Time-Series

The final stage after fitting the seasonal ARIMA model is to evaluate and check the accuracy of the model. The normal Q-Q plot along with the residual plot was displayed to evaluate each of the time series objects as shown in the representations below from Figures 20-23.

For the first time band 0-3 months, the Normal Q-Q plot shows that the data is normally distributed. The Ljung-Box shows a p-value of 0.772 which is a little significant than necessary. The RMSE for the model ARIMA(3,1,0)(0,1,1)[12] is 5107.884. The auto.arima() displayed the best model as ARIMA(3,0,0)(0,0,1)[12] with an RMSE of 5002.967 and AIC of 1689.67.



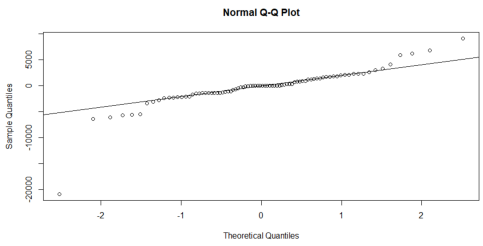
(a) Normal Q-Q plot



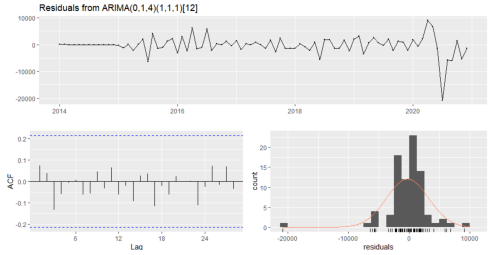
(b) Residuals Plot

Figure 20: Plots for Time Series 1

For the time band 3-6 months waiting time, the Ljung-Box test showed a p-value of 0.4892 which is not significant and shows the model is a good fit. The RMSE of ARIMA(0,1,4)(1,1,1)[12] is 3453.427. The auto.arima() gave an RMSE of 3558.941 and an AIC of 1611.3 for the model ARIMA(0,1,4)(1,0,0)[12]. However, the manual ARIMA model gave a lower RMSE value hence is a better model.



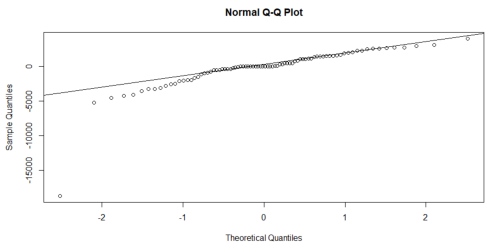
(a) Normal Q-Q plot



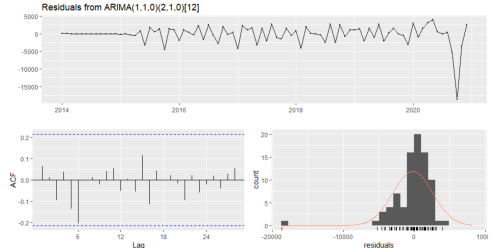
(b) Residuals Plot

Figure 21: Plots for Time Series 2

The data in the figure below is normally distributed and the AIC value is 1351.26 with an RMSE of 2775.273. The Ljung-Box test displayed a p-value of 0.5586. The model that was fit is ARIMA(1,1,0)(2,1,0)[12]. To compare with the auto.arima() function, the model best selected was ARIMA(1,1,0)(2,0,0)[12] with an RMSE value of 2908.719 and an AIC value of 1576.27. The manual model showed better results for the time band 6-9 months.



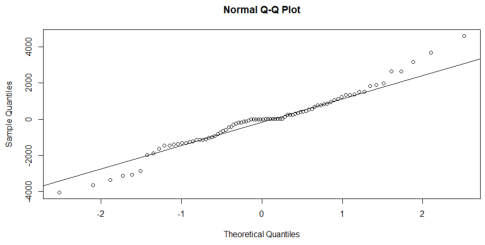
(a) Normal Q-Q plot



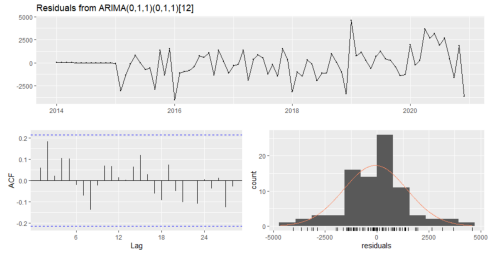
(b) Residuals Plot

Figure 22: Plots for Time Series 3

For the last time band, patients waiting for 9-12 months gave an RMSE value of 1515.037 for the model ARIMA(0,1,1)(0,1,1)[12]. The Ljung-Box test displayed a p-value of 0.5748 which is a well within the range of a good fit model. However, the auto.arima() selected ARIMA(0,2,1)(0,1,1)[12] with an RMSE of 1552.153.



(a) Normal Q-Q plot



(b) Residuals Plot

Figure 23: Plots for Time Series 4

6.3 Discussion

The supervised machine learning algorithms performed considerably well and further proved that ANN and Random Forest models performed well from the literature review for this study. It is observed that the predicted values are very close to the actual values for the model as expected and shown in Figure 24. However, the Linear Regression model did not perform well with an RMSE of 120.74. One of the main disadvantages of Linear Regression is that it always assumes that the relationship between the dependent and independent variables is linear and also the mean values. This model is also very sensitive to the outliers present in the data.

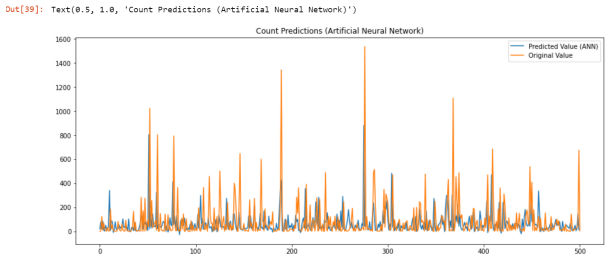
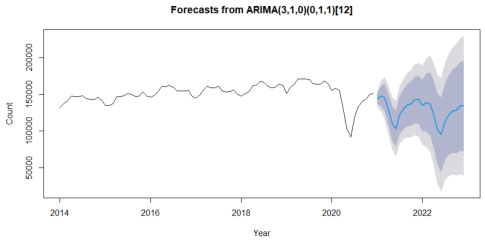
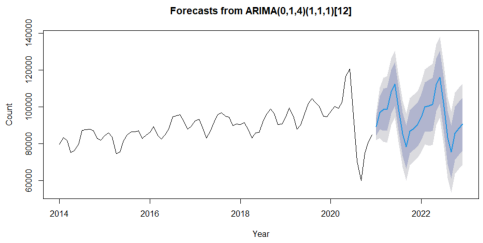


Figure 24: Predicted Values and Actual Values

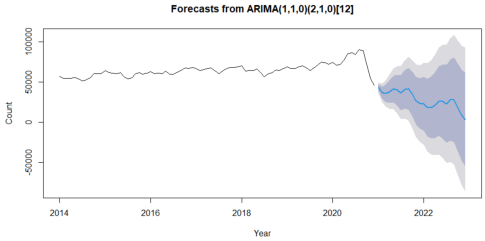
Figure 25 displays the forecasts from the four time-series models that were plotted. The values for the next two years were forecasted keeping in mind the values predicted for 2021 will be redundant and it was plotted to see how many people will still be waiting for that much longer to recieve treatment. Seasonal ARIMA models are very reliable however do not work very well with large datasets. ANN on the other hand work very well for large datasets. These results should be able to warn and make all the hospitals in Ireland more aware of what is to expect in 2022. From the forecasting plots it seems that the number of people waiting for treatment for 6-9 months decreases. In contrast to that, the number increase for people waiting for 9-12 months.



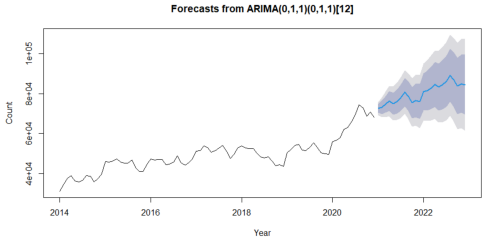
(a) Plot for Time Series 1



(b) Plot for Time Series 2



(c) Plot for Time Series 3



(d) Plot for Time Series 4

Figure 25: Forecasting Plots

7 Conclusion and Future Work

This research proposes to forecast the hospital outpatient waiting lists from the data provided by NTPF working with the HSE in Ireland. The main aim was to implement this study using time-series analysis with a seasonal ARIMA and three supervised machine learning algorithms which are ANN, Random Forest, and Linear Regression. Results demonstrate that time series performed exceptionally well and is more reliable however ANN and Random Forest also presented good results. The model with the best accuracy and lowest RMSE value was the random forest model. This research can potentially enhance the hospital systems and give them a heads up to prepare for the future years by studying the forecasted values. The treatment and the resources can be arranged prior to help reduce the waiting time for many patients and make the system more organized and efficient. In analysing the exploratory data analysis it is clearly evident which hospitals experience more patients on the list, what age group they belong to and which specialization is mostly demanded. This should benefit the HSE in order to recruit the required consultants to the required hospitals throughout Ireland. Time series analysis not only provides a satisfying model with a low RMSE value, but is also very insightful to study such patterns in order to forecast the number for the following years which is extremely beneficial and if made use can yield excellent results in Ireland. Nevertheless, it is vital to point out the limitations to allow further researchers to improve this work in the future to obtain even better results.

The limitations and future work of this research are:

- The time series forecasting could not be performed for the year 2021 due to the missing data from the NTPF because of the cyber-attack. Hence for future work, this data must be extracted in order to get more accurate forecasting results.
- The time series modelling has been done for time bands for patients waiting up to just 12 months. For future work, the data for patients waiting for more than 18+ months can be considered.
- More in-depth study on this data using other combinational neural networks might provide better results since ANN performed well.
- For future work, the study should only consider the data for the children waiting on the list since about 85,000 children are currently on the list and the NTPF only recently started putting out this data. This can be a novel and beneficial study area.

Acknowledgement

The author of this Research Project would like to thank National College of Ireland for providing excellent resources. Thanks also to Dr Martin Alain for the constant feedback and guidance. Finally, I would like to thank my sister Dr Priyanka for being the primary inspiration behind this research topic and my parents for making me reach my highest potential. This is my best effort to showcase everything I have learnt in this course and promote credible information during the study.

References

Abdel-Aal, R. and Mangoud, A. (1998). Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis, *Computer methods and programs in biomedicine* **56**(3): 235–247.

Claudio, D., Miller, A. and Huggins, A. (2014). Time series forecasting in an outpatient cancer clinic using common-day clustering, *IIE Transactions on Healthcare Systems Engineering* **4**(1): 16–26.

Elgohari, H., Bakr, A. and Majeed, M. (2019). Forecasting the number of outpatient visits in tertiary hospital using time series based on arima and es models, *Australian Journal of Basic and Applied Sciences* **13**(7): 70–77.

Guan, G. and Engelhardt, B. E. (2019). Predicting sick patient volume in a pediatric outpatient setting using time series analysis, *Machine Learning for Healthcare Conference*, PMLR, pp. 271–287.

Huber, S., Wiemer, H., Schneider, D. and Ihlenfeldt, S. (2019). Dmme: Data mining methodology for engineering applications–a holistic extension to the crisp-dm model, *Procedia Cirp* **79**: 403–408.

Li, Y., Wu, F., Zheng, C., Hou, K., Wang, K., Sun, N., Xu, B., Zhao, J. and Li, Y. (2014). Predictive analysis of outpatient volumes of a first-class grade a general hospital through arima models, *Chinese Medical Record English Edition* **2**(8): 364–367.

Lin, W.-C., Goldstein, I. H., Hribar, M. R., Sanders, D. S. and Chiang, M. F. (2019). Predicting wait times in pediatric ophthalmology outpatient clinic using machine learning, *AMIA Annual Symposium Proceedings*, Vol. 2019, American Medical Informatics Association, p. 1121.

NYONI, S. P. and NYONI, M. T. (n.d.). Forecasting the number of outpatient visits at silobela district hospital in zimbabwe using artificial neural networks, *EPRA International Journal of Multidisciplinary Research (IJMR)* .

Patil, P. and Thakur, S. (2019). Patient waiting time prediction in hospital queuing system using improved random forest in big data, *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, Vol. 1, IEEE, pp. 1–10.

Postma, M. J., Ruwaard, D., Jager, H. J. C. and Dekkers, A. L. (1995). Projecting utilization of hospital in-patient days in the netherlands: A time-series analysis, *Mathematical Medicine and Biology: A Journal of the IMA* **12**(3-4): 185–202.

Schweigler, L. M., Desmond, J. S., McCarthy, M. L., Bukowski, K. J., Ionides, E. L. and Younger, J. G. (2009). Forecasting models of emergency department crowding, *Academic Emergency Medicine* **16**(4): 301–308.

Sukmak, V., Thongkam, J. and Leejongpermpoon, J. (2015). Time series forecasting in anxiety disorders of outpatient visits using data mining, *Asia-Pacific Journal of Science and Technology* **20**(2): 241–253.

Sun, J., Lin, Q., Zhao, P., Zhang, Q., Xu, K., Chen, H., Hu, C. J., Stuntz, M., Li, H. and Liu, Y. (2017). Reducing waiting time and raising outpatient satisfaction in a chinese public tertiary general hospital-an interrupted time series study, *BMC Public Health* **17**(1): 1–11.

Wang, K., Taneja, A., Zeng, I. and Wong, D. (2015). Improving services to hospital outpatient clinics.

Wargon, M., Guidet, B., Hoang, T. and Hejblum, G. (2009). A systematic review of models for forecasting the number of emergency department visits, *Emergency Medicine Journal* **26**(6): 395–399.

Xie, Z. and Or, C. (2017). Associations between waiting times, service times, and patient satisfaction in an endocrinology outpatient department: a time study and questionnaire survey, *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* **54**: 0046958017739527.

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

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

Zhu, T., Luo, L., Zhang, X., Shi, Y. and Shen, W. (2015). Time-series approaches for forecasting the number of hospital daily discharged inpatients, *IEEE journal of biomedical and health informatics* **21**(2): 515–526.