DCU

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

CA683i

Forecasting In-Patient & Day Case Waiting Times in Ireland

Master of Computing

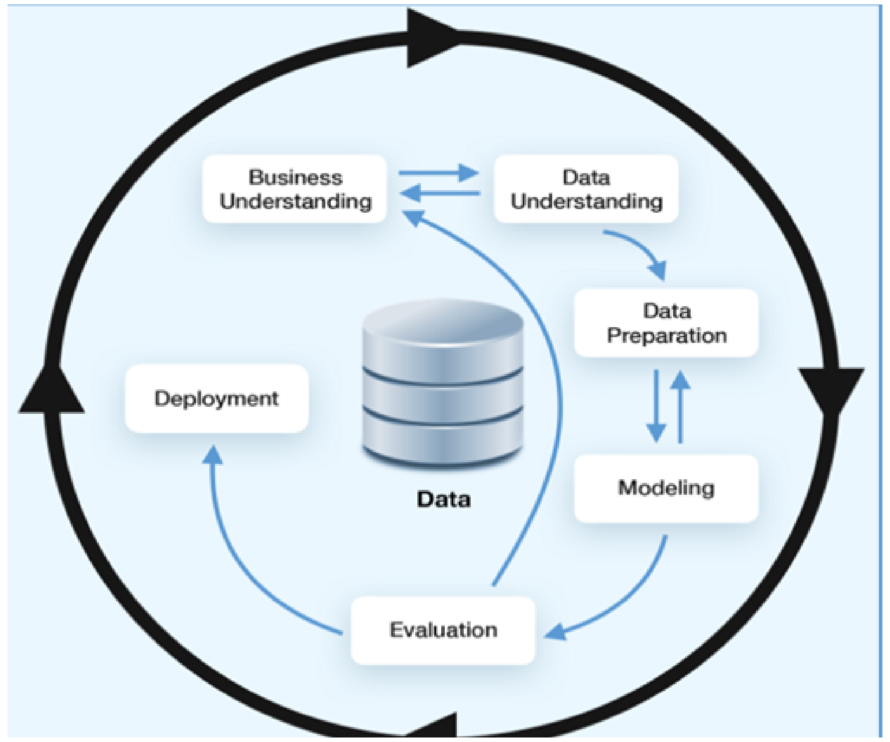
- Hao Io Cheong 21267114
- Ankit Malik 20211119
- Sakthi Somaskandan 14346091

Introduction

- Project objective is to leverage **Time series algorithms** to deliver an effective business strategy for improving Ireland domestic hospital waiting time efficiency
- Research is focusing on two categories only **In-Patient & Day Case**
- Based on HSE data, each year **1 million people** have a planned day case procedure and approximately **94,000 patients** have an elective inpatient procedure
- Data informs that waiting times can range up to **20 months**.
- Analysis is based on historic data and without external supporting datasets.
- Various algorithms are reviewed such as – **ARIMA**, **SARIMA**, **Prophet** and **TBATS**

Methodology

- Cross-Industry Standard Process data mining methodology is used
- Process could between reverse orders



Research Questions

- Which time series forecasting model is most suitable for predicting in-patient & day case waiting time?
- Discover and unleash other potential factors involved within the scope of evaluation in waiting times...

Business Understanding

To establish an effective forecasting model for predicting waiting times based on the provided medical historical data.

Dataset

- Data is provided by **National Treatment Purchase Fund (NTPF)**, in collaboration with HSE & Department of Heath Ireland
- Including various data features, specialties, age categorization and case count etc.
- 8 years of medical insight records

Archive Date	Date when the data was captured
Hospital Group	Group to which the hospital belongs
Hospital HIPE	Hospital In-patient Enquiry number
Hospital	Name of the Hospital
Specialty HIPE	Specialty In-patient Enquiry number
Specialty	Name of Specialty
Case Type	Classification of In-Patient or Day Case
Adult/Child	Classification of Adult and Child
Age Categorization	Age category of patient
Time Bands	Waiting time band for availing service
Count	Number of patients

#	Year	Record Count
1	2014	35049
2	2015	43884
3	2016	49695
4	2017	54303
5	2018	55377
6	2019	52415
7	2020	58457
8	2021	15887

Data Preparation

- Renamed column titles for data inconsistent issue
- Merging datasets into one
- Examine if there are further data quality issues (missing, replacing values)
- Establish function for resolving any case sensitive issues
- Consolidate data rows and unify one row only for per date
- Dropping all unused columns and only keeping “date” and “patients” columns
- Data from 2021 was removed as part of the cleansing activity due to insufficient volume.
- Splitting into training and test data sets.

Data Selection

- Training Data range from 2014 – 2016.
- Test Data range from 2016 – 2018.
- Attributes limited for modelling: Date & Count

	date	patients
0	2014-01-30	24473
1	2014-02-27	24035
2	2014-03-27	23288
3	2014-04-29	24382
4	2014-05-29	23635
...	...	...
82	2020-11-26	26630
83	2020-12-23	27807
84	2021-01-28	29113
85	2021-02-25	25521
86	2021-03-25	22162

Data Exploration

- Data Decomposition was done to identify seasonality, stationarity, autocorrelation in the data.
  - Seasonality refers to data fluctuations over different calendar cycles.
  - Trend component provides the overall trend of time series data.
  - Cycle Component helps to understand the non-seasonal decreasing and increasing pattern
  - Noise component informs random fluctuations in the data
  - Auto-correlation

Related Work / Literature Review

Before initiating the research there was sufficient literature review done on existing material related to in-patient and day case waiting times. The review was divided into three specific areas:

- **Forecasting Technique**
  - Analyzed research papers and blogs for different types of forecasting techniques based on complexity
  - Identified factors to consider while choosing a model such as - Seasonality, Trend, Stationarity
  - Identified evaluation metrics to assess the forecasting model such as - Root mean square error and mean absolute percent error
  - Explored classical forecasting methods like Auto-regression & Moving Average and the moved on to more complex ones like - ARIMA, SARIMA & TBATS

- **Factors affecting waiting time** - Researched on possible factors that can affect waiting times such as:
  - Staff & Capacity
  - Location
  - Recruitment Delays
  - Unplanned circumstances like COVID-19

- **Accurately predicting using time-series model**
  - Assessed if historic data alone is sufficient to predict waiting times accurately
  - Identified additional data that can be used to enhance the accuracy

Modelling

Modelling phase is used to assess and agree on the modelling techniques to be carried out for the research question.

The following models were considered and evaluated:

- Auto-regressive Integrated Moving Average (ARIMA)
- Seasonal Auto-regressive Integrated Moving Average (SARIMA)
- Trigonometric Box Cox ARMA trend seasonality (TBATS)
- Prophet

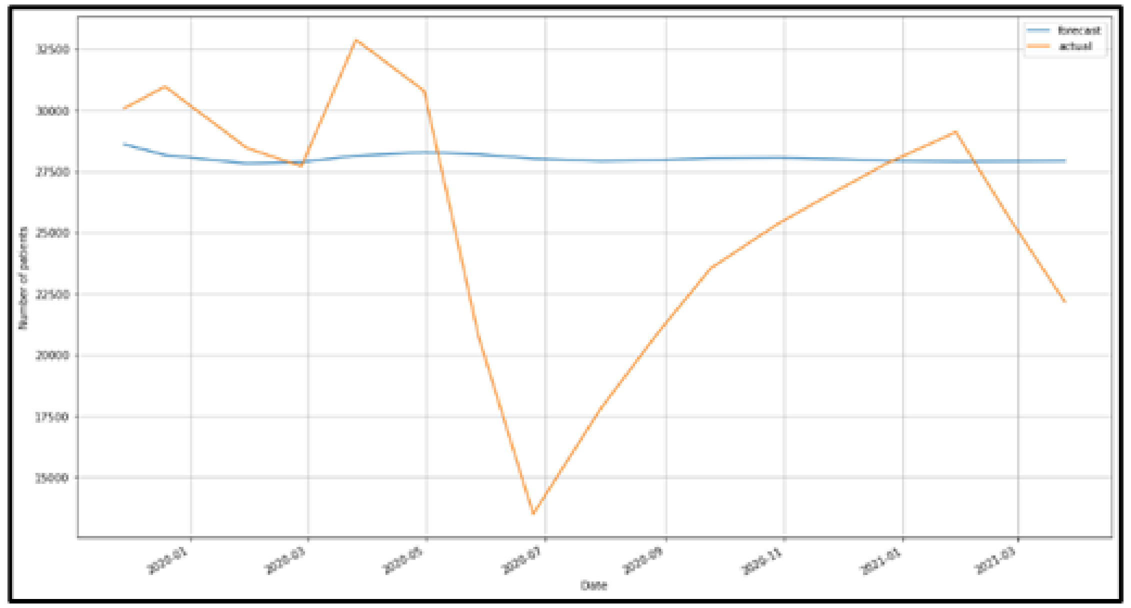
Lowest values of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were finalized as the test criteria for evaluating model performance.

- **Auto-regressive Integrated Moving Average (ARIMA)** consists of three parameters.   These parameters are defined as ARIMA(p,d,q):
  - Lag Order (p): Number of lag observations
  - Degree of differencing (d): Number of times raw observations are differenced
  - Order of Moving Average (q): Size of moving average window

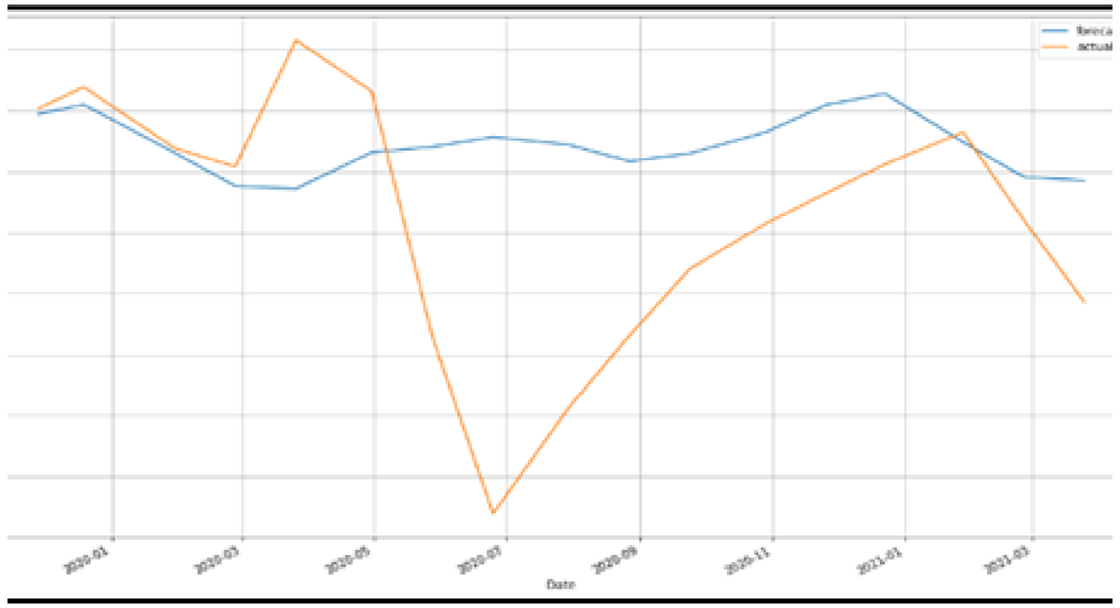
  The methodology used for implementation is Box-Jenkins which is divided in three categories:
  - Identification of best-fir p,d,q values using grid search method
  - Evaluation of the model
  - Diagnosis by running statistical tests
- **Seasonal Auto-regressive Integrated Moving Average (SARIMA)** is and extension of ARIMA for seasonality which has its own P,D,Q parameters along with s known for seasonal periodicity.
- **Trigonometric Box Cox ARMA trend seasonality (TBATS)** is used on seasonal time series where training data is used for model training and fitting to the data by providing season length information.
- **Prophet** is an open source library released by Facebook for time series forecasting. Due to time constraints, this model was not analyzed in detail.

Evaluation

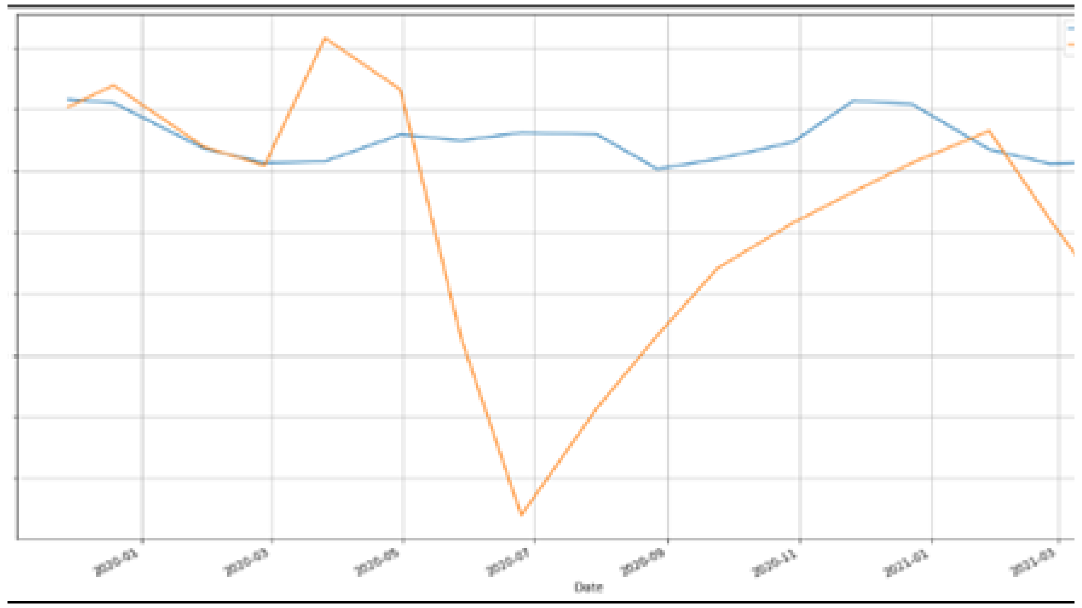
Performance of the three mentioned models is discussed below based on 0-3 months time bands for reference.



ARIMA



SARIMA



TBATS

Observations

- ARIMA: RMSE value of 5580 and MAPE of 0.205.
- ARIMA model has the best overall accuracy in predicting the in-patient & day case waiting time.
- Based on the findings, further tuning of the models is required to ensure an accurate prediction
- However, the ARIMA model can serve as a baseline for future work.

RMSE and MAPE values of all models are predicted below for 0-3 month time band, please refer report for further details:

Model	RMSE Value	MAPE
ARIMA	5580	0.205
SARIMA	5919.37	0.217
TBATS	5905.44	0.215

Conclusion

- ARIMA model has the best overall accuracy in predicting the in-patient & day case waiting time.
- Additional factors were discovered that could be responsible for influencing waiting times. It was understood that historical data alone is not sufficient to accurately predict future waiting times.
  - Average increase in waiting times during pandemic - an unforeseen situation
  - Average increase or decrease in waiting time due to changing staff capacity
  - Average staff to patient ratio

Model	RMSE Value	MAPE
ARIMA	5580	0.205
SARIMA	5919.37	0.217
TBATS	5905.44	0.215

Future Study

- Run more statistical tests and explore more time series models such as NARNN, hybrid SARIMA-NARNN to predict accurately
- Alter the forecasting model for a targeted speciality or age group e.g., elder people are likely to need healthcare access earlier than younger people excluding emergency cases. We can use the data to target the 65+ age group and even extend it to the most impacted or critical specialities such as Neurosurgery, Vascular Surgery and Cardiology.
- Explore external factors - population at different age ranges, time of year, likely infections and other potential factors.