Hong Kong Metropolitan University
School of Science and Technology

BSc. (Hons) in Data Science and Artificial Intelligence

COMP S461F Data Science Project

2023-2024

Combination Forecasting Using ARIMA and Machine Learning Models

By
So Cheuk Him (Team Leader),
Cheng Caleb

Supervisor: Dr. Tony Chan Moon Tong

Date: 18 April 2024

Acknowledgements

We thank Dr. Tony Chan Moon Tong for his guidance

Abstract

The COVID-19 pandemic has put great pressure on healthcare systems worldwide, therefore, an accurate time series prediction is crucial for hospital admission. However, the existing tools lack complexity, and a single method couldn't handle some complicated time series data. In the project, a case study of a dataset consisting of 365 data in 2021 has been manipulated with the combined approach included ARIMA, Support Vector Regression, Random Forest, Long Short-Term Memory. Through individual and combination forecasting, the effectiveness of various methods can be demonstrated. Each method is modified for the project with the aim of

increasing robustness and accuracy. The result will compare the similarities of the forecasting result which plays an important role in healthcare management and resources distribution.

Table of Content

Chapter 1

Introduction

1.1 Research Background

COVID-19, as known as a global pandemic which infected more than 600 million people in the world ("WHO Coronavirus (COVID-19) Dashboard", 2023), several waves of the pandemic have shown the vitality of COVID19, in early 2020 of United Kingdom, the first wave of the pandemic occurred and is the world's largest outbreak. Until spring 2021, there are around 80% of citizens in the United Kingdom including around 8 million individuals (about half the population of New York) require hospitalization. (Mahase E, 2020)

The dataset to be used is the admission data of UK in 2021, including the second wave of COVID-19 and peak in January 2021. Although Delta variant appeared in July, the rate of hospitalism did not grow gradually. Furthermore, the Omicron variant appeared in early December, and the second peak arose.

The primary objective of this project is to offer support to hospital human resource management by predicting the trend of COVID19. This predictive model will be a critical tool in discovering shortage of medicine or collapse of the hospital system and for health policies including vaccination, quarantine, social distancing, etc. To improve prediction accuracy, forecasts derived from different forecasting methods can be combined by different methods.

Combined forecasting is often considered more accurate than individual forecasting methods. Sumit et al. (2022) proposed a prediction on COVID-19 in India using ARIMA and Prophet model for forecasting who claims that the Prophet model can predict future cases based on the past cases, therefore, particular forecasting method will be performed for research purpose.(Sumit et al., 2022) The consolidation of various forecasting appears beneficial in situations where there is uncertainty, ambiguity about the most accurate method, and need to avoid serious errors.

1.2 Dataset Description

The chosen dataset is related to new admissions of COVID-19 patients to hospitals in the United Kingdom. The dataset contains all the United Kingdom new admissions in 2021, and the number of daily new admissions. We have separated the dataset into 3 parts, consisting 80% of

the data should be placed in the training set, 10% in the validation set, and 10% in the test set for the machine learning model. The purpose of splitting the data is to separate the machine learning part and the test set, ensuring the final result is learnt from the data, but not simply calculation. The attributes of the dataset shown in Table 1.1.

areaName	date	newAdmissions
United Kingdom	1/1/2021	3364
United Kingdom	2/1/2021	3440
United Kingdom	3/1/2021	3650
…	…	…
United Kingdom	29/12/2021	2637
United Kingdom	30/12/2021	2404
United Kingdom	31/12/2021	2108

Table 0.1.1

Table 1.1:

Table 1.1 illustrates the attributes of the dataset. This dataset has three attributes: areaName, date, and newAdmissions. Attributes date and newAdmissions would be primarily focused on the project to forecast the new admissions of COVID-19 because all the data in areaName are inserted as United Kingdom; it is not necessary to take into account different states' new admissions of COVID-19.

The following figure shows the counts of new admissions in 2021.

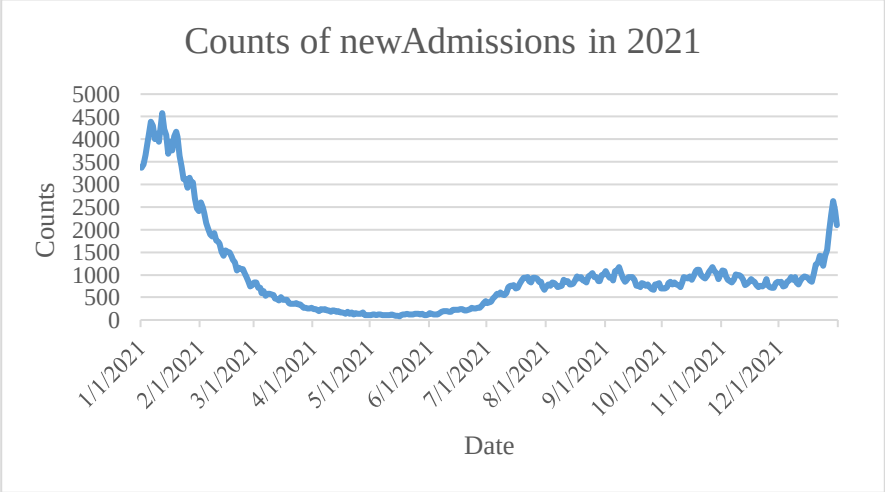


Figure 1.1

Figure 1.1

Figure 1.1 illustrates the number of new hospital admissions for COVID-19 patients in the United Kingdom throughout 2021. The peak of new admissions is observed in January, followed by a decline from January through May. The nadir of new admissions is recorded in

May, after which there is a resurgence from June to December. The graph exhibits no seasonality and demonstrates a declining trend from February to May.

The data description of the dataset is shown in Table 1.2:

Dataset Description	areaName	date	newAdmissions
Smallest number of new admissions in day	United Kingdom	5/16/2021	81
Largest number of new admissions in day	United Kingdom	1/12/2021	4580
Total number of new admissions	United Kingdom	NA	353135
Average daily number of new admissions	United Kingdom	NA	967
Average monthly number of new admissions	United Kingdom	NA	29428
Median of daily number of new admissions	United Kingdom	NA	807
Median of monthly number of new admissions	United Kingdom	NA	26007
Standard deviation of new admissions in 2021	United Kingdom	NA	955

Table 0.2

Table 1.2

Table 1.2 represents an overview of the new hospital admissions for COVID-19 patients in the United Kingdom during the year 2021. To start with, the minimum of new admissions in a single day is 81, recorded on 16th May, while the maximum is 4580 and is recorded on 1st Dec. The total number of the new admission is 353135 while the average daily admission is 967 per day and 29428 per month. The median value for the daily count of new admissions stands at 807. Besides, the median value for the monthly count of new admissions is observed to be 26006.5. Furthermore, the standard deviation for new admissions in the year 2021 is 954.66. The summary statistics concluded the range of the hospital new admission, which can be used to estimate the number of medical personnel that should be on duty during the hard time. Based on the data, forecasting can be developed for estimating the future hospital admissions.

1.3 Values and Significant

The ARIMA has different procedure by comparing to machine learning methods, since ARIMA is not a supervised learning method, the forecast for the ARIMA is effective for a stationary data, on the other hands, machine learning requires a larger dataset for training and testing process, which make machine learning handling complex pattern easily. For forecasting, ARIMA is powerful for short-term prediction, especially the simplicity, the results are easier to interpret the relationship between the data. Furthermore, differencing is needed for ARIMA when the data have no seasonality. After differencing, stationarity appears in the result, which means that mean and variance remains unchanged over time and suitable to imply the ARIMA to the dataset.

Ribeiro at et. (2020) performed a significant study on Short-term forecasting COVID-19 cumulative confirmed cases for Brazil, by the use of ARIMA, cubist regression, random forest, ridge regression, support vector regression, and stacking-ensemble learning to predict cases in a few days and week, the forecast are separated into 1-day, 3 days and 6-days to compare the different prediction result, similar part to our project is the result also essential to the health and resources management.

Chapter 2

Literature Review

2.1 Introduction of combination models

2.1.2 Combination models of Autoregressive Integrated Moving Average and Random Forest (ARIMA-RF)

Noureen at el. (2019) performed a study of comparative forecasting of ARIMA and Random Forest, by using two methods together, a forecasting of short-term and long-term are applied, the founding put in a nutshell of ARIMA are not able to forecast the data precisely, shortcoming of the ARIMA would be a lower accuracy which Random Forest demonstrates a superior performance in terms of accuracy. By using a combination forecasting of these two methods, a larger range of the patterns can be shown, overcome either the overfitting or accuracy of the model individually.

Besides, Kane at el. (2014) provided an article of comparison between ARIMA and RF of the H5N1 outbreaks, to conclude that, a large dataset (including 10 years of data) has been involved in the article, which shows the shortage of ARIMA which is the weakness of forecast the data in a long-term, in other words, the estimation of numerous parameter is poor, in the study, the H5N1 estimation of using ARIMA is less than 1, which is not possible for an probability of the nature of outbreak. The findings of the study presented the importance of comparing various forecasting methods.

2.1.3 Combination models of Autoregressive Integrated Moving Average and Support

Vector Machine (ARIMA-SVM)

One of the most popular linear models for time series forecasting in the past was the autoregressive integrated moving average (ARIMA) model. The nonlinear patterns are difficult for the ARIMA model to capture. Nonlinear regression estimation issues have been effectively solved using support vector machines (SVM), which are a unique neural network technology.

An article written by Pai (2005) who raised a hybrid methodology for forecasting stock price issues that takes advantage of the special strengths of the ARIMA model and the SVM model. The forecasting precision of the suggested approach was tested using actual stock price data sets. Computational testing shows very positive results. The idea for a hybrid model that combines ARIMA and SVM was inspired by evidence suggesting that several forecasting models can complement one another in approximating data sets. In addition, the integrated model is significantly improving the stock price by comparing ARIMA with SVM model.

Besides, a study paper is written by Nie (2012) who offered a hybrid of ARIMA and SVM for short-term load forecasting. This paper represents a combined method; utilize ARIMA and SVM as a hybrid model. To start off, the ARIMA model is employed to predict the linear basic part of the load, where the SVM is doing a forecasting on the non-linear part. SVM, known as its exceptional learning and generalization capabilities, could be used for correcting the deviations from ARIMA forecasting. Upon application of this hybrid model to a large sample prediction, the results demonstrate a high accuracy in forecasting, suggesting a promising potential for practical applications. By comparing to ARIMA and SVR individually, the hybrid ARIMA-SVMs model has effectively capitalized on the respective strengths of ARIMA and SVMs, after the hybrid ARIMA-SVM is performed to a sample data, the result indicated that hybrid form is significantly outperforms by the individual ARIMA and SVMs models.

2.1.4 Combination models of Autoregressive Integrated Moving Average and Long Short-Term Memory (ARIMA-LSTM)

Fan (2021) provided a research paper which is about the well production forecasting based on ARIMA-LSTM model considering manual operations. In the article, a novel combined model is developed that takes the pros and cons of linear and nonlinear. LSTM and ARIMA is combined in this model. Linear trend which is in the production time is filtered by the ARIMA model, and the remaining value is passed to the LSTM model. The combined model ARIMA-LSTM demonstrates a more accurate outcome compared to the individual ARIMA, LSTM models.

2.2 Introduction of Proposed Forecasting Models

2.2.1 Autoregressive integrated moving average (ARIMA) Model

In the examination of this dataset, ARIMA models, a prevalent statistical method for time series forecasting, are employed. The ARIMA model consists of three data-independent processes: auto regression, integration, and moving average. These processes are utilized for parameter e stimation, which is a linear function for historical data and random error (Box et al., 2015). Th e fundamental process of the time series form of the ARIMA model is represented as follows:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_1 - \theta_2 \varepsilon_2 - \cdots - \theta_q \varepsilon_{t-q} \qquad (1)$$

In (1), the terms y_t and ε_t stand for the initial value and random error respectively. ϕ_i $(i = 1, 2, \cdots, p)$ and θ_i $(i = 0, 1, 2, \cdots, q)$ stands for the model's parameter. ε_t is represented as random error which is assumed to be a mean of 0 and a standard deviation of 2. (1) presents the ARIMA models mathematically. They are applied to solve a variety of problems in different situations. When q in (1) is set to 0, AR model with order p is operated. When p in (1) is set to 0, AR model with order q is operated. Therefore, (p, q) are two crucial parameters in ARIMA model.

ARIMA model takes its advantages on time series, in view of the recognition of data pattern, including trends, cycles, and seasonality. In contrast, ARIMA is handy for forecasting, programming languages such as SAS, R and Python can be imply the ARIMA packages and library to satisfied the requirement. Besides, these languages could provide confidence intervals and error metrics including standard error, mean squared error by involving command in Python and R, for SAS, data metrics would be provided to user when utilize the ARIMA procedure, shows that SAS is valuable due to the ability to display multiple data metrics.

Limitation and challenges were found during time series prediction, specifically, complex dynamics and nonlinear interaction is crucial to manipulate. Additionally, ARIMA is a parametric model that aims to capture the prediction with a finite set of parameters together with data assumption. Besides, when applying ARIMA model, outliers and missing values should be eliminate throughout the data preprocessing. Moreover, ARIMA model are not suitable for some extremely short or long time series due to instability.

2.3 Introduction of Machine Learning Models

Machine learning models are also applied to forecast based on the data, including, Random

Forest (RF), Support Vector Machines (SVM), and Long Short-Term Memory (LSTM). These models could manage complicated data patterns and perform effectively in forecasting.

2.3.2 Support Vector Machines (SVM)

The Support Vector Machine (SVM) is a supervised machine learning algorithm predominantly utilized for classification and regression tasks. It operates on the principle of decision planes, which delineate decision boundaries. A decision plane, in this context, is a hyperplane that segregates different sets of objects based on their class memberships. (Boser, 1992)

Fundamentally, SVM is a boundary that optimally separates two classes. The primary objective of SVM is to identify the maximum marginal hyperplane (MMH) that most effectively partitions the dataset into classes. (Asri et al., 2016) research on estimating and identifying breast cancer's effects and risk, they claim that SVM has the highest accuracy among Naïve Bayes, k Nearest Neighbour, and Decision Tree, which has 97.13% accuracy. Last but not least, the SVM algorithm functions by mapping input vectors into a high-dimensional feature space using a kernel function. Subsequently, it identifies a hyperplane within this space that separates the input vectors into their respective classes. The hyperplane is selected to maximize the distance (or margin) between itself and the nearest vectors from each class, known as support vectors. This characteristic gives the algorithm its name - Support Vector Machine. The equation of SVM is:

$$y = f(X) = \sum_{i=1}^{M} W_i X_i + b \tag{2}$$

In (2), X_i stands for the values of input features, while W_i represents the weights assigned to these inputs. The term b refers to bias, and y is utilized for actual values. M represents the total number of data samples.

The Support Vector Machine (SVM) algorithm demonstrates effective performance with smaller datasets and high-dimensional spaces. However, the training process becomes time-consuming when SVM is applied to larger datasets, (Zhang & Yang, 2005) claims that SVM will consume a large memory requirement and computation time, therefore, SVM should be apply parallel training to shorten the training time, which means splitting the data into a smaller dataset and train it at the same time.

Since its initial release, support vector machines (SVM) have been the subject of extensive studies and have been applied to a wide range of tasks, including text categorization,

handwritten character recognition, and pattern recognition (Joachims, 1997; Schölkopf and Burges, 1996; Schmidt, 1996). The SVM concept has been applied to regression issues because of its successful performance in solving real-world classification problems (Smola and Schölkopf, 2004).

The SVM algorithm used for classification problems is defined as Support Vector Classification (SVC), and for regression problems is known as Support Vector Regression (SVR).

In SVR, input data are placed into a higher-dimensional feature space. The data points that are nearest to the hyperplane are defined as support vectors. The hyperplane's orientation and position are significantly affected by support vectors, the hyperplane that fits the data could assist in forecasting.

2.3.3 Long Short-Term Memory (LSTM)

Recurrent neural networks (RNNs) are suitable for time forecasting given that they are analyzing sequential data excellently. However, vanishing and exploding gradients problems and RNNs inability to model over an extended period of time, the training process would be difficult. (Goodfellow, 2016) To tackle these issues, the Long Short-Term Memory Network (LSTM) is implemented. (Hochreiter, 1997).

The following describes the LSTM's learning process, which is depends on these factors:

- x_t:
  The input vector at a given time step, t.
- $b = \{b_i, b_o, b_f, b_c\}$:
  The bias vectors for the input (i), output (o), forget (f), and memory cell (c).
- $W = \{W_i, W_o, W_f, W_c\}$:
  The weight matrices for the input (i), output (o), forget (f), and memory cell (c).
- $U = \{U_i, U_o, U_f, U_c\}$:
  The recurrent weights for the input (i), output (o), forget (f), and memory cell (c).

Four parameters for LSTM unit: an input gate (i_t), an output gate (o_t), a forget gate (f_t), and a memory cell (c_t). These components are essential in the LSTM's operation.

Hyperbolic tangent ($\tan h$) and sigmoid (σ) activation functions are used in LSTM. Below is the mathematical representation of the gates given a specified number of hidden units (H):

$$input\ gate: i_t = \sigma(x_t W_{xi} + H_{t-1} W_{hi} + b_i)$$

Input gate selects which data should be used, while the output gate generates the new long-term memory.

$$output\ gate: o_t = \sigma(x_t W_{xo} + H_{t-1} W_{ho} + b_o)$$

The data that is output from the LSTM unit is controlled by the output gate.

$$forget\ gate: f_t = \sigma(x_t W_{xf} + H_{t-1} W_{hf} + b_f)$$

Which data from previous memory units can be transmitted or forgotten is determined by the forget gate.

$$intermediate\ cell\ state: \tilde{c}_t = \tan h\,(x_t W_{xc} + H_{t-1} W_{hc} + b_c)$$

The term $\tilde{c}_t$ refers to the newly computed memory within the system.

$$cell\ state: c_t = f_t c_{t-1} + i_t \tilde{c}_t$$

The memory cell (c_t) is a combination of the previous memory.

$$new\ state: h_t = o_t \tan h\,(c_t)$$

The output of the LSTM unit is represented as h_t.

LSTM is superior to RNN in serval ways. Initially, it could recognize complex patterns and long-term relationships in sequential data. Furthermore, LSTM is the most suitable method for producing the most accurate model when compared to other approaches like ARIMA, according to (Kırbaş et al., 2020). In addition, LSTM could learn from extended sequences without losing or oversimplifying information because it could avoid the problems that are caused by expanding or vanishing gradients. Moreover, compared to a basic RNN, LSTM is better to tolerate missing or noisy input.

Apart from LSTM's advantages, LSTM exhibits several drawbacks when compared to RNN. Given that LSTM has more parameters and operations in forming the model, it requires more processing resources, it takes more memory and requires longer training and execution times. Additionally, it is simpler to overfit than RNN, the use of regularization techniques like dropout, weight decay, or early termination is necessary to avoid overfitting. Finally, since they have

more hidden layers and states than RNN, they present more difficulties for interpretation and explanation.

A study paper is written by (Ma, Q., 2020) who offer LSTM to the stock market claims that LSTM perform a prediction result that can be used to distinguish between market and accidental fluctuation, demonstrate the decline and long-term trend respectively. Hence, the disadvantage of it would be other factors of the environment may affect the final accuracy.

Chapter 3

Statistical Methodology

3.1 Research Objectives

The objective of this study is to comprehend the benefits and drawbacks of the ARIMA model and machine learning models (Random Forest, SVM, LSTM). The purpose of this study is to survey the respective forecast values. The research employs combined forecasting method with the comparison of using forecasting method individually with the aims of find out particular model. A comparative analysis will be conducted between individual forecasting and combination forecasting to identify the more desirable model by model evaluation and will be selected for the implementation in this project. Therefore, the research objectives are as follows:

1. Can cross-validation or bootstrapping provide more insight in the forecast model?
2. Using ARIMA model and machine learning models to evaluate the predictive performance and comprehend the pros and cons respectively.
3. Combine different model results by comparing the MSE and MAE separately to find out the precise model.
4. Diagnose whether individual forecasting or combination forecasting is more desirable.

3.2 Model Building and Forecasting

This project will be using Autoregressive Integrated Moving Average model (ARIMA) and 3 different machine learning models, first is Bootstrap aggregating (Bagging), next will be Support Vector Machine (SVM), and lastly Long short-term memory (LSTM) network, is a kind of kNN method. Furthermore, this project will consider either combination forecasting, or the individual can obtain a more accurate forecast result. Thus, the project will be using these models to obtain a forecasting the following trend of the admission data of hospitality in UK 2021, the prediction aims to lower the pressure of the UK hospital effort and let the citizens get ready for the next wave of the pandemic.

3.2.1 Autoregressive and Moving Average Models

To start off, we discover that the data are formed by three categories, we will mainly focus on the column “New admission”, by using Autoregressive and Moving Average Models (ARIMA)

model, to analyze which model suits our dataset. Moreover, we will discover the difference between machine learning methods and choose the most appropriate one for our specific dataset.

Table 3.1

The ARIMA Procedure

Name of Variable = newAdmissions	
Period(s) of Differencing	1
Mean of Working Series	-3.45055
Standard Deviation	99.82091
Number of Observations	364
Observation(s) eliminated by differencing	1

Table 3.1. Shows the result of performing SAS to generate an ARIMA (Autoregressive Integrated Moving Average model) procedure, it illustrates the mean, standard deviation and observation elimination by differencing.

Figure 3.1

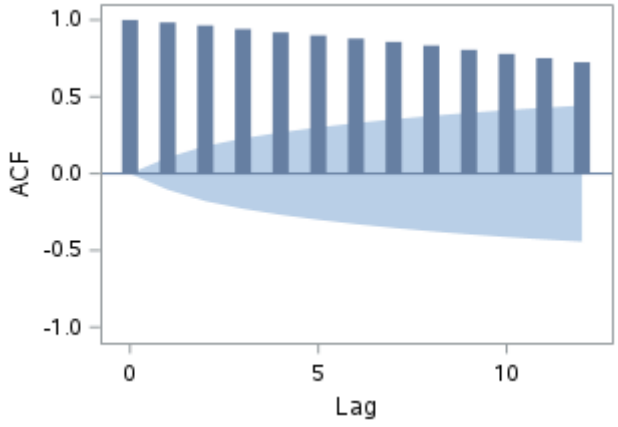


Figure 3.2

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-2.8578	0.2453	-2.04	0.0397		
	1	-3.6167	0.1909	-2.12	0.0329		
	2	-4.6369	0.1385	-2.45	0.0139		
	3	-4.0113	0.1684	-3.02	0.0027		
	4	-3.9404	0.1722	-3.49	0.0005		
	5	-4.7776	0.1326	-3.98	<.0001		
	6	-5.7434	0.0987	-3.68	0.0003		
	7	-7.5885	0.0564	-2.96	0.0032		
	8	-10.7230	0.0224	-3.63	0.0003		
	9	-9.6443	0.0306	-2.64	0.0082		
	10	-10.5358	0.0236	-2.94	0.0034		
	11	-9.2537	0.0343	-3.30	0.0010		
	12	-6.4132	0.0805	-1.39	0.1540		

Figure 3.1. From the above solution, the ACF shows how the consumption is correlated to the previous 12 lags and next 12 lags. The shaded area is statistically irrelevant, together with Figure3.2 an ADF test is performed, the p-value of the ADF test with a drift term and zero mean, it is 0.2453, greater than the significant level of 0.05 indicated the data is non-stationary, differencing is required for further assumption.

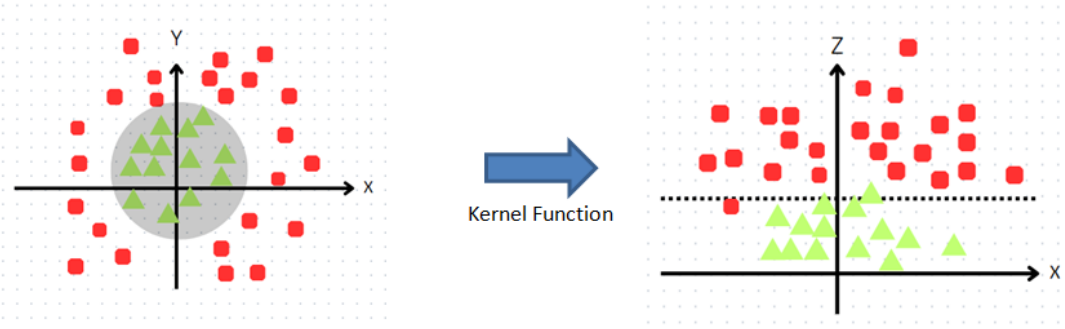
3.2.2 Support vector regression (SVR)

Kernel Method

To convert the incoming data points into a higher-dimensional space, SVR employs kernel functions. It is possible to use a variety of kernel types, including sigmoid, linear, polynomial, and Gaussian (RBF).

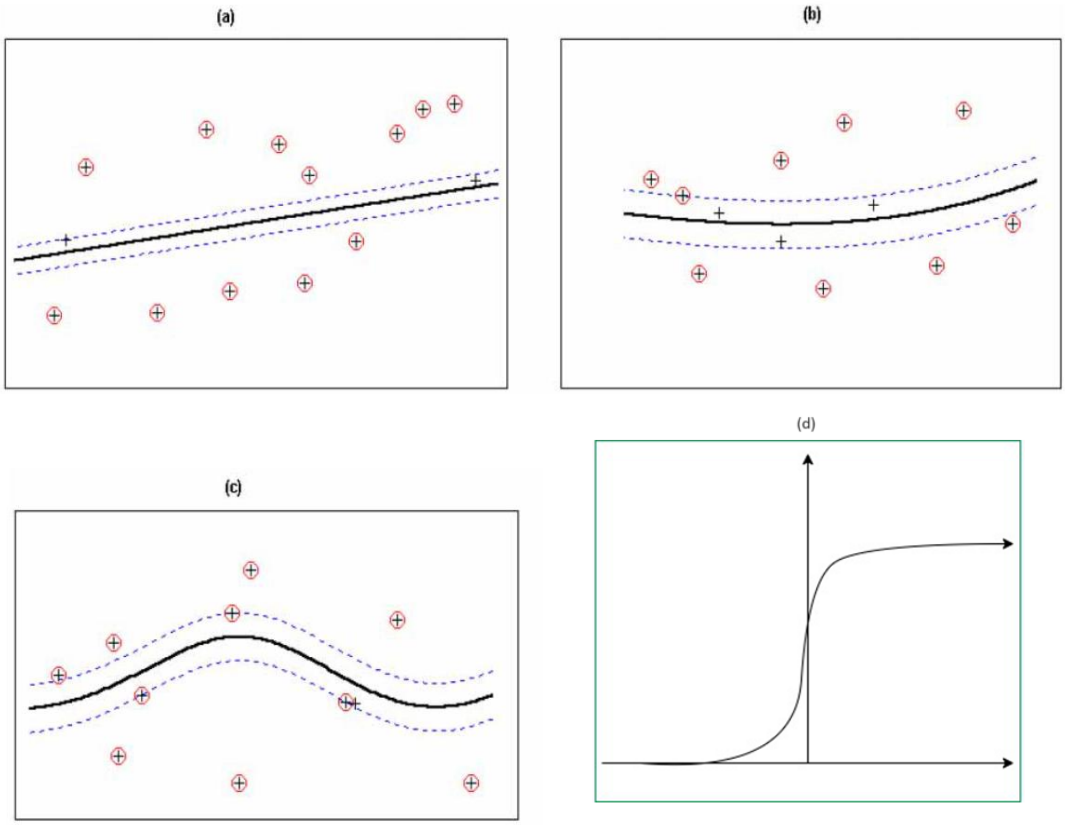
Figure 3.3 shows how the input data will change from 2D to 3D when the kernel method is applied, indicating that SVR could handling the non-linear data in the hyperplane.

Figure 3.3



For instance, when a dataset contains two classes that are excessively comparable to one another, the kernel can be used to effectively separate them but not only viewing the data in linear but also from various perspectives.

Figure 3.4



SVR using various kernels is shown in Figure 3.4. (a)Linear, (b)polynomial, (c)radial basis functions (RBF), (d)and sigmoid are the four types of functions.

Linear kernel has a decision boundary that is either a straight line or a flat plane. It is primarily used when the data is clearly divided by a line or plane. It doesn't make any complex changes to the input data.

On the other hand, polynomial kernel, the data is converted into a higher dimension, therefore a line or plane could be used to split it. To be more accurately represent the complex relationship between data, it incorporates additional features like squares and interaction terms in the kernel.

Radial Basis Function (RBF) is a popular kernel in SVR. The input data could be mapped into a multidimensional space. It generates complicated decision boundaries to execute SVR model therefore it could evaluate for different kinds of data. The RBF kernel provides more flexibility since it generates non-linear feature combinations.

Sigmoid kernel adjusts the data through sigmoid function. It is useful when probability is required since it produces values between -1 and 1. It is not frequently applied in SVR because the limited applicability.

Margin

SVR is separated into two categories: soft margin and hard margin.

When a model strictly prohibits any data point falling inside the error tolerance or margin, a hard margin in SVR is created. Overfitting could result from this, especially in noisy data.

In situations in which the margin is soft, some data points might fall within it. This can help to avoid overfitting in cases where the data is noisy. The degree of softness of the margin can be adjusted using hyperparameters.

More regularized margins are produced with smaller hyperparameters, and less regularized margins are produced with larger hyperparameters.
Hard Margin Support Vector Regression is defined as:

$$\text{Minimize } \frac{1}{2}\left|\left|w\right|\right|^2 \text{ subject to } y_i - w^T x_i - w_0 \leq \epsilon\,,\ i = 1,2,\dots,N \tag{3}$$

In (3) $\frac{1}{2}\left|\left|w\right|\right|^2$ represents the margin of the hyperplane and the nearest data point. The weight vector's Euclidean norm is $\left|\left|w\right|\right|^2$. The input sample is denoted by x_i. Continuous output is denoted by y_i. The margin is denoted by ϵ .

Soft Margin Support Vector Regression is defined as:

$$\text{Minimize } \frac{1}{2}\left|\left|w\right|\right|_0^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4}$$
$$\text{subject to } y_i - w^T x_i - w_0 \leq \epsilon + \xi_i\,, w^T x_i + w_0 - y_i \leq \epsilon + \xi_i^*\,,\ \xi_i, \xi_i^* \geq 0$$

In (4) C represents the regularization parameter used to maximize the margin, and ξ_i , ξ_i^* are slack variables used to calculate for misclassification.

Furthermore, this dataset is better suited for the hard margin SVR. Hard margin is appropriate for small data given that soft margin SVR allows for some error to occur; this dataset only includes data from one year. Regularization parameter C can be used to improve the problem when using hard margin SVR.

Additionally, C and Gamma controls the shape of the decision boundary, hyperparameters tuning is necessary for Support Vector Regression to get an optimal result, after we consumed the grid search, the tuned hyperparameter gamma is 8, C is 4 while the epsilon is 0.01. First, in the grid search process, 5-fold cross validation is used to split the dataset into training and validation set which means the dataset in then divided into 5 equal size fold and perform the

training for 5 times, then, some parameters are assigned some value to Python to explore the optimal value.

Tuning

Tuning is the process of adjusting the parameters to maximize the predictive accuracy of the model. This requires carefully modifying the parameters and evaluating the model's performance to determine the set of parameters that provides the most accurate predictions. By modifying these parameters, the SVR model could perform better and have higher prediction accuracy and reliability.

By changing a few of its parameters, SVR can operate more effectively.
These include the kernel selection, the regularization parameter (C), and kernel-specific parameters for example gamma for the RBF kernel.

These parameters have a significant impact on the model's performance. A single training example's impact on the system is determined by the gamma parameter, the cost parameter balances the allowance of training errors and the imposition of strict margins, and the epsilon parameter shows the epsilon tube within which the loss function in training process is associated with no penalty.

3.3 Model evaluation

The mean absolute error (MAE), root mean square error (RMSE), regression coefficient (R^2), and mean square error (MSE) are the four-performance metrics used to assess the proposed model's performance.
represents the actual value, the mean value, and the estimated value. The number of data inputs is denoted by .
Better model performance is indicated by a lower MAE or RMSE value, which is expressed mathematically as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(y - \hat{y})| \tag{5}$$

RMSE is defined as follows in (6):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y - \hat{y})^2} \tag{6}$$

A model that tends toward 1 indicates a higher level of accuracy in forecasting. R2 is displayed as follows in (7) to show the variance within the independent and dependent parameters:

$$R^2 = 1 - \frac{\sum |y - \hat{y}|}{\sum |y - \bar{y}|} \tag{7}$$

The Mean Squared Error (MSE) will primarily be employed in this project. Five distinct MSEs will be assessed post-modeling, one of which is a combination forecast, while the remaining four are individual forecasts, namely Autoregressive Integrated Moving Average (ARIMA), Bagging, Long Short-Term Memory (LSTM), and Support Vector Machine (SVM). The purpose of the MSE is to quantify the disparity between the predicted and actual values. The MSE is expressed mathematically as follows in (8):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 \tag{8}$$

After the comparison of the 5 MSEs, the solution indicated that the greater the accuracy, the smaller the MSE will be, in other words, the result is close to the actual value, will be implemented in the project.

Chapter 4

Model

4.1 ARIMA

Figure 4.1

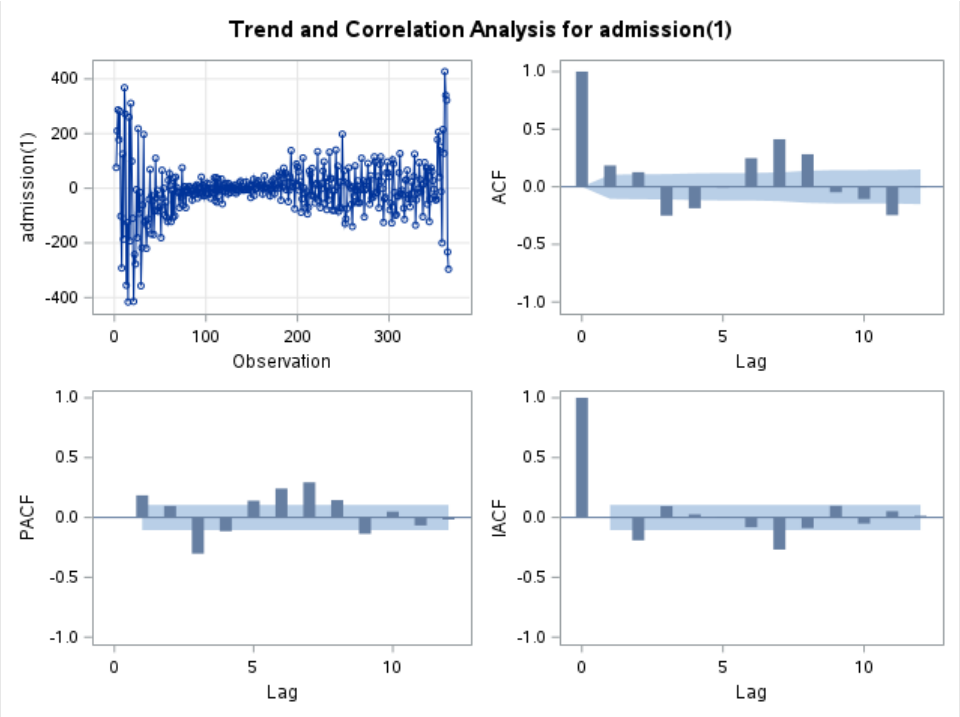


Table 4.1

Constant Estimate	5.102048
Variance Estimate	6369.416
Std Error Estimate	79.80862
AIC	4228.291
SBC	4255.571
Number of Residuals	364

Figure 4.1 We are using differencing methods to check if stationarity exists. From the result, we observed that ACF dies down in a sine wave as k increases while k is the number of lags. From the partial autocorrelation graph above, we observed that an area with light blue color is formed by 2 standard error, therefore, certain data is standing outside the designated area while the others do not. Besides, in the PACF, the result shows that the lag after lag 9 is significant, but some of the data before lag 9 is uncertainly significant, so, perform various of testing is necessary, after testing different p-value, specified lag of 2, 3, 6, 7, 8 are selected according to the result of the PACF graph after the first differencing, however, according to the ACF graph there are 2 lags cuts off but does not selected in the study, including lag 1 and lag 5, the reason why ignoring the lag will be implement in lately this session. Table 4.1 indicates the lowest AIC and BIC among the other lag individually, the AIC of 4228 and BIC of 4255 are slightly smaller than the others, for lag 7, it has a sharply cut off appears, when lag 7 are selected individually, the AIC is 4294, BIC is 4302, both are larger than our selected lag.

Application on Auto ARIMA in R:

Figure 4.2

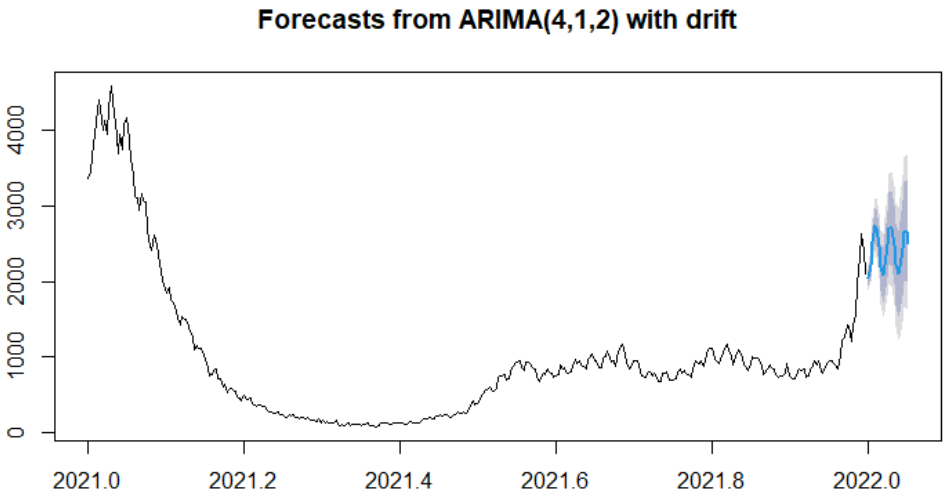


Figure 4.2 Illustrates the Auto ARIMA function in R program, in the graph, an ARIMA(4,1,2) is selected by the program, 4 means the order of autoregressive component, 1 is the order of differencing, while 2 is the moving average component. Blue line is the forecasting result of the function, a slightly downward trend appears, while the grey area is the possible trend in all scenarios, afterwards, more tests are needed to ensure the correctness of the selected model.

T-test for the ARIMA (4,1,2):

Table 4.2

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	-3.67437	5.37850	-0.68	0.4949	0
MA1,1	-0.93403	0.12558	-7.44	<.0001	1
MA1,2	0.06279	0.17899	0.35	0.7259	2
AR1,1	-0.74154	0.12357	-6.00	<.0001	1
AR1,2	0.41131	0.16178	2.54	0.0114	2
AR1,3	-0.11155	0.07475	-1.49	0.1365	3
AR1,4	-0.26441	0.05979	-4.42	<.0001	4

Table 4.2 Represent the t-test result of the ARIMA(4,1,2), the column "Approx Pr > $|t|$" means the p-value associated with the value in the t-value column, therefore, p-value is normally alpha=0.05, since t-test is used for evaluating whether a group of data differs from a known

value, alpha <= 0.05 means passing the t-test. In the result, half of the result is not satisfied with the requirement, in other words, it has no evidence to support the alternative hypothesis.

Second approach for Auto ARIMA in R:

Figure 4.3

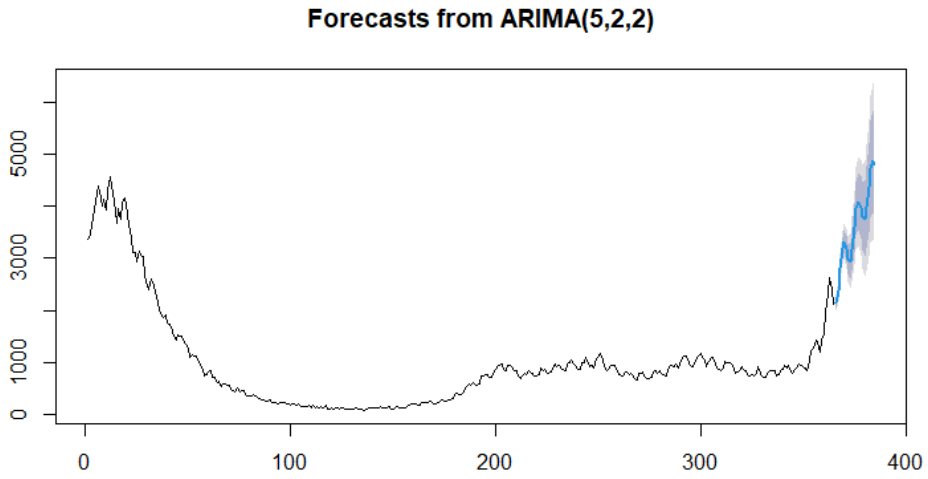


Figure 4.3 Represent the result of Auto ARIMA by inserting data manually, the graph illustrates an ARIMA(5,2,2) with an upward trend occurred, the model has been differencing for twice, also, more tests are needed to ensure the correctness.

T-test for the ARIMA (5,2,2):

Table 4.3

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	1.06563	0.13469	7.91	<.0001	1
MA1,2	-0.43040	0.10957	-3.93	0.0001	2
AR1,1	0.04965	0.13666	0.36	0.7166	1
AR1,2	-0.17351	0.05880	-2.95	0.0034	2
AR1,3	-0.51859	0.04421	-11.73	<.0001	3
AR1,4	-0.38168	0.07759	-4.92	<.0001	4
AR1,5	-0.21602	0.09002	-2.40	0.0169	5

Table 4.3 shows the results of the t-test, in the column "Approx Pr > $|t|$ " shows row3 AR1,1 is not passing the t-test, the result is 0.7166 which is larger than alpha 0.05, but it still have a better result that the ARIMA (4,1,2), next, q-test should be performed for further assumption.

Q-test for the ARIMA (5,2,2):

Table 4.4

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.013	-0.005	0.002	0.009	-0.014	-0.125
12	16.69	5	0.0051	0.046	-0.008	-0.102	0.119	-0.019	0.038
18	21.21	11	0.0313	0.060	0.006	0.029	0.035	-0.035	-0.070
24	30.10	17	0.0256	-0.129	0.006	-0.025	-0.052	0.050	-0.024
30	32.14	23	0.0972	0.046	0.041	0.009	-0.013	-0.032	-0.011
36	34.23	29	0.2311	-0.022	-0.041	0.032	-0.026	-0.031	-0.019
42	40.05	35	0.2560	-0.035	0.048	0.002	-0.044	0.093	-0.013
48	42.96	41	0.3872	0.021	-0.008	0.012	-0.025	-0.051	0.055

Table 4.4 represent the result of the Chi-square test, the column “Pr>Chisq” is the associated p-value, the p-value should be greater than 0.05 to pass the q-test, hence the null hypothesis is true, in the above graph, lag6 –lag 24 are smaller than 0.05 and should be rejected. Thus, more testing is needed in order to find the best fit ARIMA model.

Computing the ARIMA manually:

Table 4.5

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	109.03692	30.19089	3.61	0.0003	0
MA1,1	-0.17235	0.05915	-2.91	0.0038	10
AR1,1	0.12396	0.04459	2.78	0.0057	2
AR1,2	-0.16283	0.03895	-4.18	<.0001	3
AR1,3	0.25580	0.04026	6.35	<.0001	6
AR1,4	0.47901	0.04326	11.07	<.0001	7
AR1,5	0.25727	0.04821	5.34	<.0001	8

Table 4.5 represent the reason of choosing the those selected lags in the above, after various of approaches, some of the lag cannot passes through the t-test, some of the lag needs to be rejected, besides, removing some specific lag will cause some changes to the q-value, some lags need to be computing to the ARIMA manually. Furthermore, AR model with selected lag (2,3,6,7,8) together with MA model with selected lag 10 could get the best result among all the ARIMA model with specific lag, in the figure, p-values of different parameters including MU, AR and MA model are satisfied with the condition of smaller than 0.05, which means proceed the t-test.

In summary, the result indicates that the model is statistically significant, and there is evidence to show the null hypothesis is incorrect and to be rejected, the following study can be performed.

Table 4.6

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.028	0.048	0.010	-0.040	0.100	-0.037
12	6.79	6	0.3408	-0.004	0.002	-0.031	0.003	-0.000	0.034
18	10.41	12	0.5799	0.059	-0.011	0.018	0.060	-0.018	-0.040
24	17.77	18	0.4706	-0.111	0.016	-0.016	-0.061	0.036	-0.035
30	19.48	24	0.7260	0.046	0.038	0.014	-0.009	-0.023	-0.004
36	21.88	30	0.8585	-0.025	-0.013	0.034	-0.043	-0.045	-0.011
42	28.43	36	0.8114	-0.052	0.052	0.020	-0.050	0.087	-0.010
48	30.06	42	0.9157	0.032	0.002	0.021	-0.007	-0.027	0.040

Table 4.6 Shows the Chi Squared test of the selected ARIMA (2, 3, 6, 7, 8), 1, (10)), in the third column of the autocorrelation check of residuals, the p-value are all over 0.05, thus, the ARIMA model passed the Chi Squared test, which determined that the null hypothesis needed to be rejected, in other words, no evidence to show there is significant or difference between the variables. Afterwards, since the ARIMA model can pass two tests, it can be used for the further study of combination forecasting.

Figure 4.3

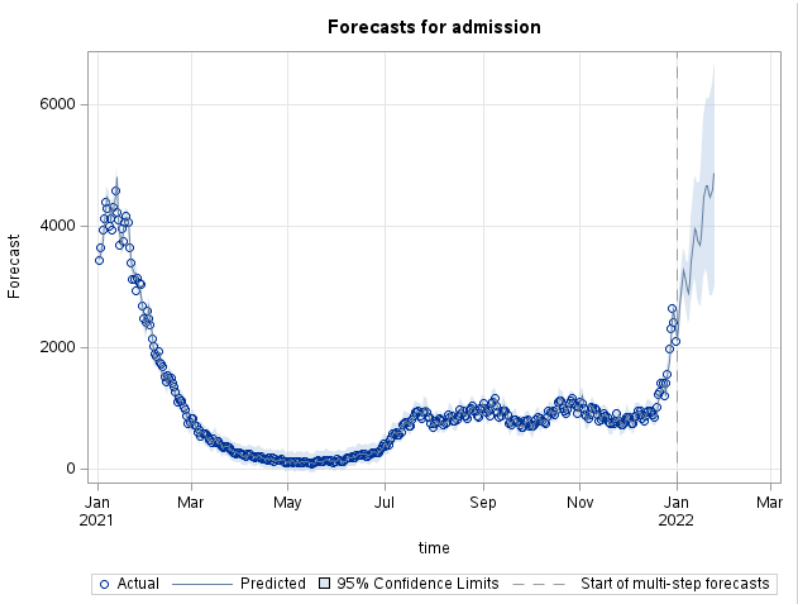


Figure 4.3 illustrates the forecasting result of the AR model with selected lag (2,3,6,7,8) together with MA model with selected lag 10, the trend shows that there would be another peak held in January next year, furthermore, the model needs to be separate into 80% training and 20% of testing for combination.

Result of ARIMA:

In conclusion, the selected ARIMA with an autoregressive order of (2, 3, 6, 7, 8) disparate of using 5 directly, 5 means the selection is AR (1) to AR (5), in the selection, AR (7) and AR (8) is vital because it cuts off gradually, therefore, these 5 lagged observations are used as the predictor in the model. Secondly, like the autoregressive order, the moving average of lag 10 is selected, but not from MA (1) to MA (10), since most of the MA model cannot pass the Chi-squared test, the null hypothesis cannot be rejected.

Figure 4.4

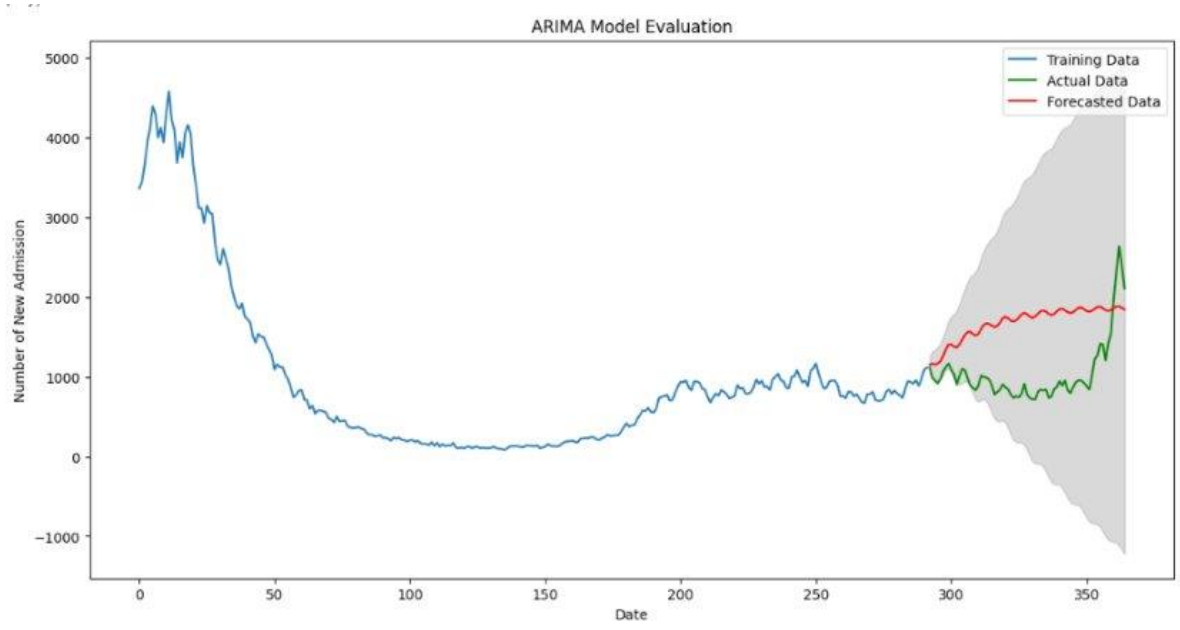


Figure 4.4 illustrates the entire time series with the prediction by ARIMA, the blue line is the original data, representing the training set, the green line is the testing data and the red line is the forecasted data, from the result above, the forecasted data is slightly going upward with seasonality, the prediction is not a straight line but a curve with fluctuation, forms a seasonality

from data starting at 200. Additionally, the shaded area represents the possible range the trend would deviate if any unexpected factors appeared.

Machine learning models

In SVR, RF and LSTM, using the same data set spilt. The data set has 365 data,60% of data (219) used as training set, 20% data (73) used as validation set and 20% data (73) used as testing set.

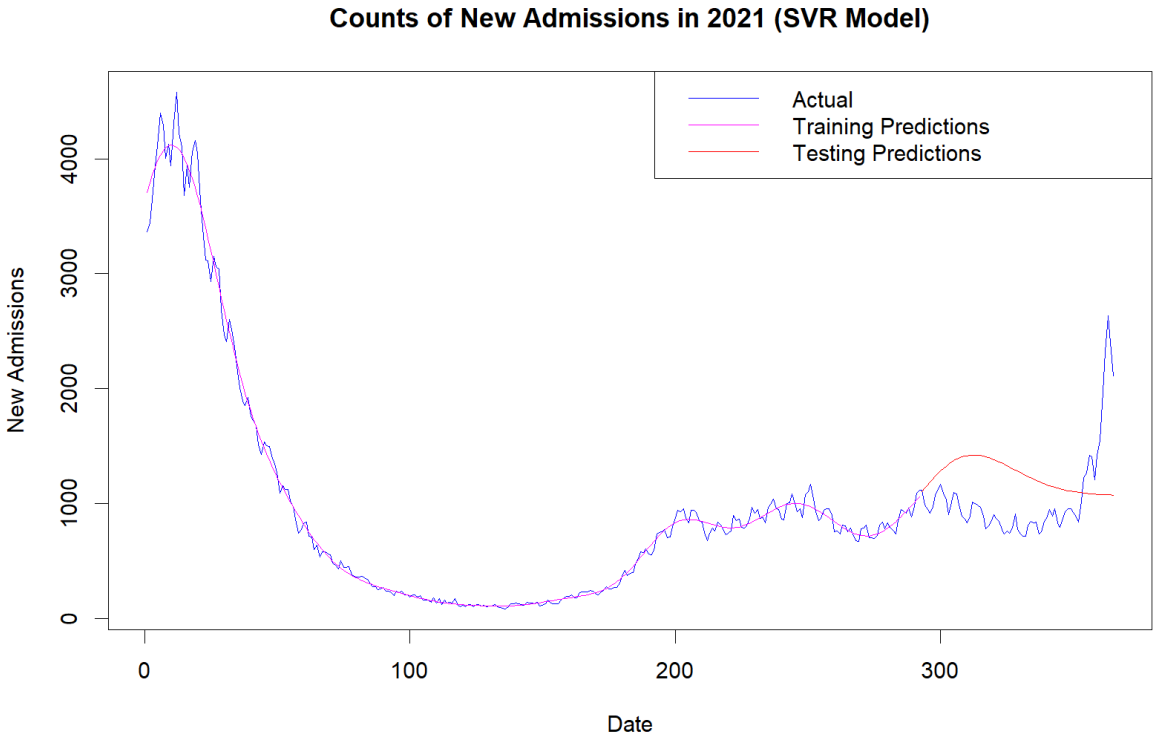
4.2 Support Vector Regression (SVR)

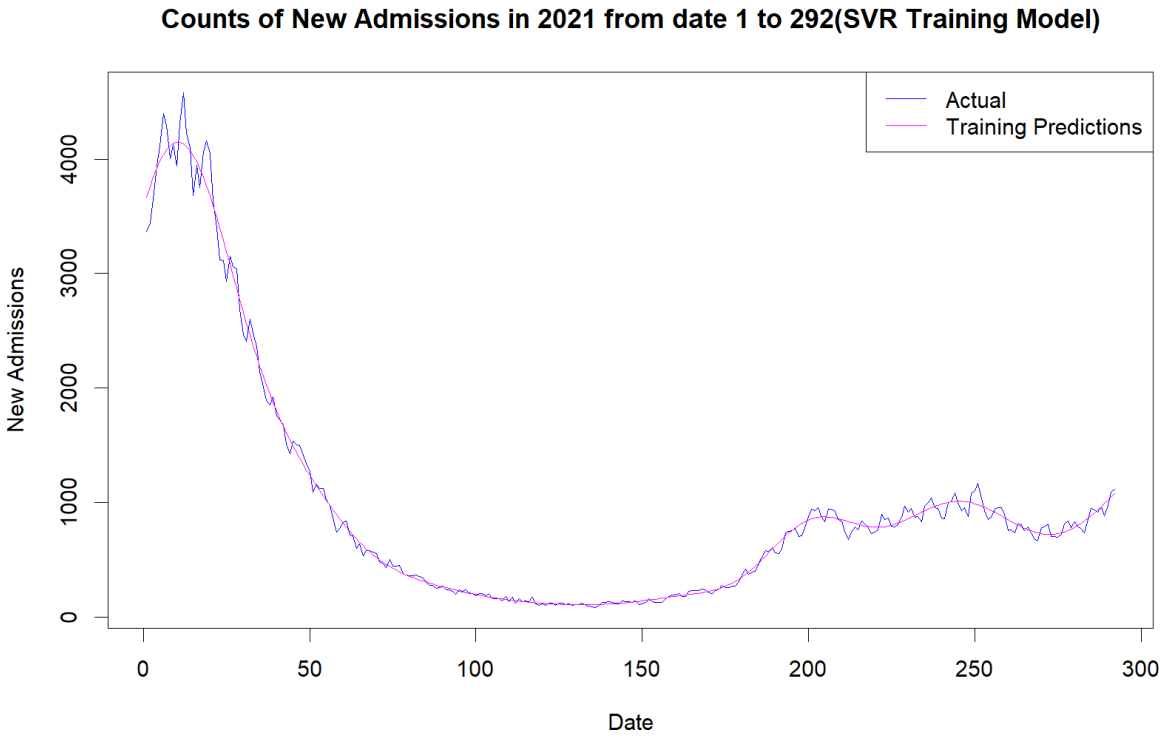
In SVR, cross-validation and grid search is used through data 1-292(training data + validation data=219+73=292).
In cross-validation, 5-fold cross-validation is used for model evaluation.
In grid search, firstly defining the range of the value of the hyperparameter which are gamma, cost and epsilon. For gamma, the range of value set as from 2 to 8 step by 0.1. For cost, the range of value set as from 2 to 4 step by 0.1. For epsilon, the range of value set as from 0.001 to 0.1. Choosing 2 to 8 for gamma and 2 to 4 for cost step by 0.1, for a higher gamma and cost would lead the model more overfitting, therefore using a small value in both gamma and cost can avoid overfitting. Besides, using a small step size as 0.1 could identify the optimal and best parameter for the model in a wide range of selection to enhance the model complexity. Furthermore, according to (Smets, et al., 2007) recommended that applying epsilon from 0.001 to 0.1 is suitable for assemble a SVR model. Secondly, using the tune.svm function to use RBF kernel execute cross-validation and grid search.

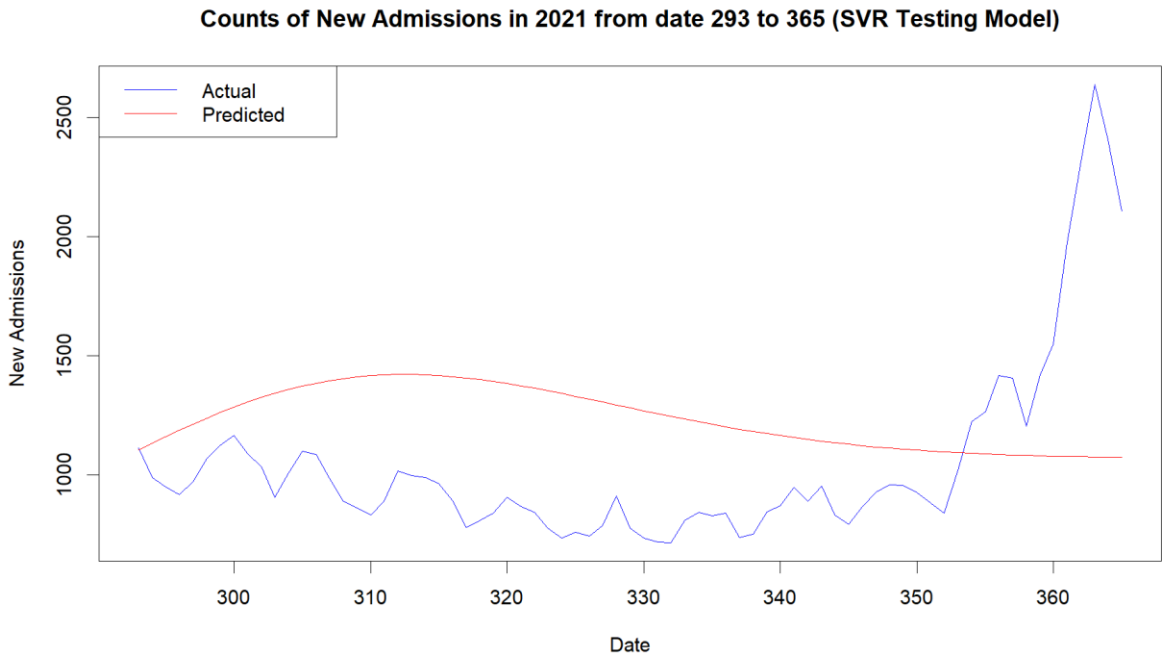
Lastly, after tuning the parameter, outputting the best model with the turned parameter, which are 7.9 for gamma, 4 for cost and 0.01 for epsilon. The training and validation model is finished.

In evolving the testing model, it is not directly used the best model in the cross-validation training model. Since it is a time series model, the sequential data is significantly affecting the performance of the testing model. For loop is used to form a testing model. In the for loop (from 293 to 365), in each iteration, using the best tuned SVR model by reading the original data from date 1 to the current date minus 1 date. For instance, the iteration is in

data 300 now, the best SVR training model would read the data from date 1 to 299 to forecast the date 300 data by employing the best tuned model.
The following figures are the counts of new admission in 2021 by using SVR model.







In training model part, SVR performs significantly excellent in showing the curve of the data. The arc is well-fit the original data. However, SVR have poor performance in testing model, the curve does not match the actual data. It is because that the SVR testing model did not study the whole data in testing set and did not tune the parameter for testing data, it could not efficiently learn the nature of the sequential time series data.

The following table is the model evaluation of SVR in testing model comparing with the actual data from date 293-365.

Model evaluation	Result
MSE	246510.7847
MAE	416.0344365
R-Squared	-0.663150362

In both MSE and MAE showed that the result of SVR training model forecasting output has a large difference between with the original data. For R- Squared, the output is negative number, it means that the SVR testing model is not a suitable model to represent the original data.

Random Forest (RF)

4.3 Data Analysis Software Used

To fulfill the requirement of ARIMA and machine learning model, the project will be conducted by R and SAS programming language for data analysis.

SAS

The SAS software has a variety of functions and packages for different purposes. Furthermore, SAS has a clear graphical user interface that provides a user-friendly interface for debugging and easier to use. This will significantly shorten the analysis process.

While comparing SAS to R, SAS has a better graphical support, which is able to form statistical tables and graph instantly together with more specific statistics. Rather than the option is fewer than R, the data visualization is also important for this project.

R

R is an open-source programming language. In addition, R has a higher learning curve when comparing to SAS, since it is more complicated to user, but it can deal with data mining, data cleaning or some machine learning algorithms. In this project, R was chosen because of its wide variety of statistical methods and ability to handle mathematical notation and formulae.

In a nutshell, SAS would be used in building the ARIMA model while R will mainly focus on machine learning in view of SAS didn't include bagging, LSTM, SVM packages. Either one language may not satisfy the requirements of the project since both languages have their respective strengths and limitations, therefore, we are going to utilize both for the data analysis, in order to ensure the correctness and completeness of the project.

Python

Python is a powerful tool for machine learning, imply the library to python can handle various of tasks, Pandas for data preprocessing and managing, Scikit-learn provide cross-validation procedure, apply libraries Tenserflow and Keras can provide deep learning to train the model. Python is mainly used in the project due to its general-purpose and machine learning libraries.

Chapter 5

Discussion

5.1 Research questions revisited

Revisiting the research questions in following

1. Can cross-validation or bootstrapping provide more insight in the forecast model?
2. Using ARIMA model and machine learning models to evaluate the predictive performance and comprehend the pros and cons respectively.
3. How does the new admission data forecast affect real-life decision-making?
4. Whether individual forecasting or combination forecasting is more desirable.

5.2 Insight on the forecast model

The first question, cross-validation is one of the common techniques for carrying out machine learning, by dividing data to multiple fold, it can assess the correctness of the machine learning model perform on unseen data, furthermore, simple evaluating the model may cause overfitting or underfitting, the insight of the cross-validation including lower the forecasting error, during training procedure, the actual and estimated values must not be totally the same, the difference between two values are the mean squared error, after the use of cross validation, it ensure that model would be trained on the most relevant data int order to get a more precise result. In other words, by employing cross-validation, the result turns out to be more realistic. This method aims to resample the dataset, for a small dataset like the one in our study, it could recognize more reliable assessment of the model on unseen data.

Bootstrapping is important for uncertainty of sampling, it is like cross-validation, it is known as one kind of resampling method, apart from using k-fold, bootstrapping generates samples with replacement, some observation could be appeared more than once. All in all, cross-validation aims on model selection and tuning and bootstrapping aims on understanding the result and estimating, for example, confidence interval together with hypothesis testing, according to Joseph (2020), bootstrapping and traditional method are appropriate to create samples, it provides a robust opportunity for forecasting the dataset, moreover, resampling is more cost-effective than traditional methods, for a larger dataset, more data processing and repeating is needed. The resampling method provides accurate estimates on statistics, which is essential to make future decisions.

5.3 Pros and con of using ARIMA and machine learning methods

In section 4, the modelling part has addressed the inquiry, ARIMA is good at understand the past data and model parameters, while machine learning model can handle more complex pattern, start with the advantage of ARIMA, the method is simpler, steps for performs including check for stationarity, conduct differencing and parameter selection, the cross-validation is not necessary of ARIMA, the performance goals in to compare to the other machine learning methods to check the presence of cross-validation can perform a precise predictive accuracy, moreover, the cross-validation is on the overall performance but not for the parameter selection. In the above result, the optimal machine learning model in the project is Random Forest, the ability of capture a nonlinear relationship in the data, for instance, non-regular and non-linear pattern and trends. Machine learning, with its advantages, is suited for solving non-straightforward relationships.

The cons on ARIMA consist of sensitive to outliers, since the model is followed by a stationary process, outliers that influence the value will be considered as the general trend and capture by the ARIMA, an incorrect forecasting can be shown. On the other hand, a robust machine learning model would have a higher chance of overfitting, the noises in the training dataset could be learn by the model and forecast the identical structure, for further approach, regularization is necessary for adding penalty terms to the loss function in order to avoid memorizing too many details in the training process.

5.4 Decision-making on the result of combination forecasting

During the result session, most of the forecasting methods indicates that there is upward trend in hospital demand, moreover, here are some decisions can be made for the future planning, to start off, the resource sharing and collaboration between different hospital institutions is vital, for example capacity planning and patient transfer, for example asymptomatic infection or mild symptoms cases, by transfer patient to temporary quarantine center for lowering the pressure of the hospital, since hospital couldn't focus only on the COVID-19, injuries are one of the urgent cases require hospital attention, the advantage of the quality of care will be show up and improve the patient flow.

In addition, the focus of our project is beyond the pandemic of COVID-19, but also for unpredicted cases, according to population and rate of using the medical services, some of the going up trend would be capture, the forecasting is accordingly predicting the expansion of the capacity, and to prevent the risk of rapid tends to economic loss. Unpredicted situations including disaster preparation, sufficient capacity is vital for emergencies, including adjusting the operation, mobilizing resources to ensure for handling numerous patients. Patient flow is one of the considerations, through forecasting result, predicting the number of patients arrival, schedule appointment, aims to lower the waiting time of patients, the faster the patient flow also means the vacancy rate of the hospital. Data-driven decision making needs to be monitored in real-time, for example variant of the pandemic, the transmissibility of the virus, all in all, the patient and front-line medical workers' feedback needs to be considered in further decision making.

5.5 Whether individual or combination forecasting performs better

Chapter 6

Conclusion

6.1 Overall summary

This project planned to build a combination model that predicts future new admissions of COVID-19 for improving the resource management. After reviewing admissible literature on forecasting recommended that ARIMA times series modeling and others machine learning model, they are SVR, RF and LSTM could operate as components of combing the model. Three distinct functions of this concept were perceived: i) forecasting by using specific models, ii) combining the models, iii) comparing the models. The technique is typically viable, and the combined model outperforms the single model for forecasting future data, according to a case study including 20 sessions. The three applications that the findings address were the main topic of debate that followed.

6.2 Limitations and future direction

This project is narrow by i) the range of the dataset and ii) methodological limitations.

In scale of the dataset, ARIMA, SVR, RF and LSTM models are built respectively. The data set have 365 data,80% of data (292) used as training set and 20% data (73) used as test set. After building 4 different model, each model has 73 outputs as the future predictions. By comparing to the original data, then selecting the lowest error in each output, and forming the combine model through elected value. It ought to be mentioned that for from a useful perspective, the dataset should be larger for building more accurate model. Exclusively for scientific investigation, further research could employ more representative datasets and incorporate a dimension for comparison; for example, analyzing more than 1 year dataset, collecting not only United Kingdom new admissions data but also using France's.

In methodological restrictions, apart from ARIMA, SVR RF and LSTM, one model for representing sequential time series data is the Time Series Transformer. It makes use of the architecture known as the Encoder-Decoder Transformer, which is ideal for predicting applications. It enables autoregressive generation, which is akin to text generation in that it makes future predictions based on past data.

Rather than importing the complete series at once, transformers manage massive amounts of time series data by picking context and prediction windows for training batches. Imputation approaches can be avoided by utilizing masks to include missing data into the model.

Transformers do have certain restrictions, though. Quadratic computation and memory needs impose limits on the sizes of the context and prediction windows. Transformers are also strong models that readily overfit or pick up erroneous correlations, therefore appropriate regularization and validation are essential.

To sum up, the Time Series Transformer is an effective tool for modeling sequential data that can handle issues like missing values and enable precise predictions. It is important to be mindful of its limitations and the possibility of overfitting, though.

Appendix

areaName	date	newAdmissions	date	newAdmissions	date	newAdmissions
United Kingdom	1/1/2021	3364	1/3/2021	825	1/5/2021	102
United Kingdom	2/1/2021	3440	2/3/2021	835	2/5/2021	120
United Kingdom	3/1/2021	3650	3/3/2021	713	3/5/2021	123
United Kingdom	4/1/2021	3937	4/3/2021	706	4/5/2021	103
United Kingdom	5/1/2021	4115	5/3/2021	599	5/5/2021	120
United Kingdom	6/1/2021	4396	6/3/2021	636	6/5/2021	120
United Kingdom	7/1/2021	4294	7/3/2021	535	7/5/2021	104
United Kingdom	8/1/2021	4002	8/3/2021	576	8/5/2021	111
United Kingdom	9/1/2021	4127	9/3/2021	579	9/5/2021	103
United Kingdom	10/1/2021	3941	10/3/2021	567	10/5/2021	105
United Kingdom	11/1/2021	4309	11/3/2021	551	11/5/2021	109
United Kingdom	12/1/2021	4580	12/3/2021	480	12/5/2021	122
United Kingdom	13/1/2021	4225	13/3/2021	463	13/5/2021	103
United Kingdom	14/1/2021	4100	14/3/2021	428	14/5/2021	97
United Kingdom	15/1/2021	3684	15/3/2021	503	15/5/2021	91
United Kingdom	16/1/2021	3944	16/3/2021	441	16/5/2021	81
United Kingdom	17/1/2021	3751	17/3/2021	445	17/5/2021	102
United Kingdom	18/1/2021	4062	18/3/2021	451	18/5/2021	128
United Kingdom	19/1/2021	4161	19/3/2021	380	19/5/2021	127
United Kingdom	20/1/2021	4051	20/3/2021	364	20/5/2021	132
United Kingdom	21/1/2021	3637	21/3/2021	358	21/5/2021	125
United Kingdom	22/1/2021	3395	22/3/2021	361	22/5/2021	119
United Kingdom	23/1/2021	3118	23/3/2021	368	23/5/2021	116
United Kingdom	24/1/2021	3113	24/3/2021	350	24/5/2021	139
United Kingdom	25/1/2021	2932	25/3/2021	341	25/5/2021	133
United Kingdom	26/1/2021	3150	26/3/2021	298	26/5/2021	132
United Kingdom	27/1/2021	3059	27/3/2021	273	27/5/2021	127
United Kingdom	28/1/2021	3044	28/3/2021	273	28/5/2021	139
United Kingdom	29/1/2021	2687	29/3/2021	251	29/5/2021	105
United Kingdom	30/1/2021	2469	30/3/2021	257	30/5/2021	112
United Kingdom	31/1/2021	2408	31/3/2021	270	31/5/2021	124
United Kingdom	1/2/2021	2605	1/4/2021	234	1/6/2021	157
United Kingdom	2/2/2021	2481	2/4/2021	237	2/6/2021	131
United Kingdom	3/2/2021	2362	3/4/2021	222	3/6/2021	129
United Kingdom	4/2/2021	2142	4/4/2021	197	4/6/2021	126
United Kingdom	5/2/2021	2005	5/4/2021	238	5/6/2021	128
United Kingdom	6/2/2021	1893	6/4/2021	223	6/6/2021	156
United Kingdom	7/2/2021	1852	7/4/2021	235	7/6/2021	177
United Kingdom	8/2/2021	1921	8/4/2021	207	8/6/2021	189
United Kingdom	9/2/2021	1754	9/4/2021	206	9/6/2021	191
United Kingdom	10/2/2021	1726	10/4/2021	186	10/6/2021	201
United Kingdom	11/2/2021	1680	11/4/2021	205	11/6/2021	180
United Kingdom	12/2/2021	1511	12/4/2021	205	12/6/2021	178
United Kingdom	13/2/2021	1428	13/4/2021	182	13/6/2021	223
United Kingdom	14/2/2021	1538	14/4/2021	197	14/6/2021	227
United Kingdom	15/2/2021	1503	15/4/2021	161	15/6/2021	231
United Kingdom	16/2/2021	1495	16/4/2021	160	16/6/2021	230
United Kingdom	17/2/2021	1411	17/4/2021	159	17/6/2021	245
United Kingdom	18/2/2021	1339	18/4/2021	141	18/6/2021	239
United Kingdom	19/2/2021	1272	19/4/2021	179	19/6/2021	211
United Kingdom	20/2/2021	1091	20/4/2021	134	20/6/2021	207
United Kingdom	21/2/2021	1157	21/4/2021	173	21/6/2021	233
United Kingdom	22/2/2021	1120	22/4/2021	122	22/6/2021	242
United Kingdom	23/2/2021	1120	23/4/2021	157	23/6/2021	275
United Kingdom	24/2/2021	1022	24/4/2021	130	24/6/2021	255
United Kingdom	25/2/2021	971	25/4/2021	138	25/6/2021	258
United Kingdom	26/2/2021	864	26/4/2021	135	26/6/2021	268
United Kingdom	27/2/2021	742	27/4/2021	169	27/6/2021	267
United Kingdom	28/2/2021	771	28/4/2021	112	28/6/2021	310
United Kingdom			29/4/2021	103	29/6/2021	368
United Kingdom			30/4/2021	114	30/6/2021	417

areaName	date	newAdmissions	date	newAdmissions	date	newAdmissions
United Kingdom	1/7/2021	373	1/9/2021	1084	1/11/2021	1099
United Kingdom	2/7/2021	392	2/9/2021	999	2/11/2021	1084
United Kingdom	3/7/2021	396	3/9/2021	928	3/11/2021	985
United Kingdom	4/7/2021	470	4/9/2021	952	4/11/2021	889
United Kingdom	5/7/2021	521	5/9/2021	879	5/11/2021	861
United Kingdom	6/7/2021	577	6/9/2021	1077	6/11/2021	831
United Kingdom	7/7/2021	571	7/9/2021	1103	7/11/2021	888
United Kingdom	8/7/2021	608	8/9/2021	1167	8/11/2021	1015
United Kingdom	9/7/2021	559	9/9/2021	1038	9/11/2021	996
United Kingdom	10/7/2021	552	10/9/2021	926	10/11/2021	989
United Kingdom	11/7/2021	594	11/9/2021	851	11/11/2021	965
United Kingdom	12/7/2021	732	12/9/2021	868	12/11/2021	892
United Kingdom	13/7/2021	750	13/9/2021	943	13/11/2021	778
United Kingdom	14/7/2021	755	14/9/2021	952	14/11/2021	808
United Kingdom	15/7/2021	775	15/9/2021	953	15/11/2021	839
United Kingdom	16/7/2021	700	16/9/2021	894	16/11/2021	905
United Kingdom	17/7/2021	710	17/9/2021	754	17/11/2021	867
United Kingdom	18/7/2021	799	18/9/2021	759	18/11/2021	843
United Kingdom	19/7/2021	877	19/9/2021	731	19/11/2021	775
United Kingdom	20/7/2021	938	20/9/2021	813	20/11/2021	735
United Kingdom	21/7/2021	930	21/9/2021	808	21/11/2021	758
United Kingdom	22/7/2021	955	22/9/2021	756	22/11/2021	743
United Kingdom	23/7/2021	867	23/9/2021	783	23/11/2021	786
United Kingdom	24/7/2021	828	24/9/2021	731	24/11/2021	911
United Kingdom	25/7/2021	939	25/9/2021	678	25/11/2021	776
United Kingdom	26/7/2021	941	26/9/2021	672	26/11/2021	734
United Kingdom	27/7/2021	922	27/9/2021	782	27/11/2021	717
United Kingdom	28/7/2021	852	28/9/2021	783	28/11/2021	715
United Kingdom	29/7/2021	840	29/9/2021	810	29/11/2021	809
United Kingdom	30/7/2021	746	30/9/2021	705	30/11/2021	841
United Kingdom	31/7/2021	676	1/10/2021	699	1/12/2021	828
United Kingdom	1/8/2021	739	2/10/2021	698	2/12/2021	840
United Kingdom	2/8/2021	789	3/10/2021	723	3/12/2021	736
United Kingdom	3/8/2021	763	4/10/2021	814	4/12/2021	751
United Kingdom	4/8/2021	836	5/10/2021	840	5/12/2021	845
United Kingdom	5/8/2021	812	6/10/2021	782	6/12/2021	871
United Kingdom	6/8/2021	782	7/10/2021	828	7/12/2021	946
United Kingdom	7/8/2021	729	8/10/2021	791	8/12/2021	888
United Kingdom	8/8/2021	742	9/10/2021	767	9/12/2021	953
United Kingdom	9/8/2021	760	10/10/2021	735	10/12/2021	831
United Kingdom	10/8/2021	894	11/10/2021	850	11/12/2021	793
United Kingdom	11/8/2021	850	12/10/2021	946	12/12/2021	868
United Kingdom	12/8/2021	861	13/10/2021	932	13/12/2021	929
United Kingdom	13/8/2021	790	14/10/2021	915	14/12/2021	957
United Kingdom	14/8/2021	788	15/10/2021	958	15/12/2021	954
United Kingdom	15/8/2021	807	16/10/2021	886	16/12/2021	924
United Kingdom	16/8/2021	870	17/10/2021	974	17/12/2021	881
United Kingdom	17/8/2021	967	18/10/2021	1089	18/12/2021	840
United Kingdom	18/8/2021	913	19/10/2021	1114	19/12/2021	1018
United Kingdom	19/8/2021	950	20/10/2021	1113	20/12/2021	1224
United Kingdom	20/8/2021	872	21/10/2021	988	21/12/2021	1265
United Kingdom	21/8/2021	877	22/10/2021	949	22/12/2021	1417
United Kingdom	22/8/2021	834	23/10/2021	916	23/12/2021	1405
United Kingdom	23/8/2021	966	24/10/2021	972	24/12/2021	1205
United Kingdom	24/8/2021	1000	25/10/2021	1068	25/12/2021	1420
United Kingdom	25/8/2021	1036	26/10/2021	1124	26/12/2021	1548
United Kingdom	26/8/2021	956	27/10/2021	1166	27/12/2021	1975
United Kingdom	27/8/2021	945	28/10/2021	1087	28/12/2021	2315
United Kingdom	28/8/2021	864	29/10/2021	1032	29/12/2021	2637
United Kingdom	29/8/2021	857	30/10/2021	905	30/12/2021	2404
United Kingdom	30/8/2021	997	31/10/2021	1009	31/12/2021	2108
United Kingdom	31/8/2021	1003				

SVR code

```
data <- read.csv("Book1v2.csv", header = TRUE)

set.seed(123)


y <- data$newAdmissions

x <- data$date



library(e1071)


k <- 5


#  k-fold cross-validation and grid search to find the best SVR model

gamma_values <- seq(2, 8, by = 0.1)

cost_values <- seq(2, 4, by = 0.1)

epsilon_values <- 10^(-3:-1)


tuned_parameters <- tune.svm(x = x[1:292],

                             y = y[1:292],

                             kernel = "radial",

                             gamma = gamma_values,

                             cost = cost_values,

                             epsilon = epsilon_values,

                             tunecontrol = tune.control(sampling = "cross", cross = k)

)
```

```
best_model <- svm(x = x[1:292],
                  y = y[1:292],
                  kernel = "radial",
                  gamma = tuned_parameters$best.model$gamma,
                  cost = tuned_parameters$best.model$cost,
                  epsilon = tuned_parameters$best.model$epsilon
)



# use the training model for testing using the for loop
predictions <- numeric(length(y))
for (i in 293:365) {
  partial_data <- data[1:i, ]

  predictions[i] <- predict(best_model, newdata = partial_data[i, "date"])
}

print(predictions)
print(tuned_parameters$best.model$gamma)
print(tuned_parameters$best.model$cost)
print(tuned_parameters$best.model$epsilon)
```

```
plot(data$date[293:365], data$newAdmissions[293:365], type = "l", col = "blue", xlab = "Date", ylab = "New Admissions", main = "Counts of New Admissions in 2021 from date 293 to 365 (SVR Testing Model)")

lines(data$date[293:365], predictions[293:365], col = "red")

legend("topleft", legend = c("Actual", "Predicted"), col = c("blue", "red"), lty = 1)


library(openxlsx)

# Create a data frame with the date and predictions

prediction_data <- data.frame(date = data$date[293:365], predictions = predictions[293:365])


# Create new data points

new_data <- data.frame(date = (max(data$date) + 1):(max(data$date) + 73))


# Use the best model to make predictions

new_predictions <- predict(best_model, newdata = new_data)


# Print the predictions

print(new_predictions)


plot(data$date[1:365], data$newAdmissions[1:365], type = "l", col = "blue", xlab = "Date", ylab = "New Admissions", main = "Counts of New Admissions in 2021 (SVR Model)")

lines(data$date[293:365], predictions[293:365], col = "red")

# Use the best model to make predictions on training data

train_predictions <- predict(best_model, newdata = data[1:292, "date"])


# Plot the training data and predictions
```

```
lines(data$date[1:292], train_predictions, col = "magenta")

legend("topright", legend = c("Actual", "Training Predictions","Testing Predictions"), col =
c("blue", "magenta","red"), lty = 1)


plot(data$date[1:292], data$newAdmissions[1:292], type = "l", col = "blue", xlab = "Date",
ylab = "New Admissions", main = "Counts of New Admissions in 2021 from date 1 to 292(SVR
Training Model)")


# Use the best model to make predictions on training data

train_predictions <- predict(best_model, newdata = data[1:292, "date"])


# Plot the training data and predictions

lines(data$date[1:292], train_predictions, col = "magenta")

legend("topright", legend = c("Actual", "Training Predictions"), col = c("blue", "magenta"), lty
= 1)
```

References

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos, Solitons & Fractals, 135, 109853.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152).

Burges, C. J., & Schölkopf, B. (1996). Improving the accuracy and speed of support vector machines. *Advances in neural information processing systems*, *9*.

DataFlair Team. (n.d.). Advantages of SAS | Disadvantages of SAS Programming. Retrieved from
https://data-flair.training/blogs/disadvantages-and-advantages-of-sas/

Fan, D., Sun, H., Yao, J., Zhang, K., Yan, X., & Sun, Z. (2021). Well production forecasting based on ARIMA-LSTM model considering manual operations. *Energy*, *220*, 119708.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Smola, T. (1997, July). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *ICML* (Vol. 97, pp. 143-151).

Kırbaş, İ., Sözen, A., Tuncer, A. D., & Kazancıoğlu, F. Ş. (2020). Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos, solitons & fractals, 138, 110015.

Kung, F. (2017, January 27). CRAN now has 10,000 R packages. Here's how to find the ones

you need. Revolutions. Retrieved from https://blog.revolutionanalytics.com/2017/01/cran-10000.html

Liu, X., Liu, A., Chen, J. L., & Li, G. (2023). Impact of decomposition on time series bagging forecasting performance. *Tourism Management*, *97*, 104725.

Noureen, S., Atique, S., Roy, V., & Bayne, S. (2019). A comparative forecasting analysis of arima model vs random forest algorithm for a case study of small-scale industrial load. International Research Journal of Engineering and Technology, 6(09), 1812-1821.

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC bioinformatics, 15, 1-9.

Ma, Q. (2020). Comparison of ARIMA, ANN and LSTM for stock price prediction. In *E3S Web of Conferences* (Vol. 218, p. 01026). EDP Sciences.

Mahase, E. (2020). Covid-19: outbreak could last until spring 2021 and see 7.9 million hospitalised in the UK.

Mohan, S., Solanki, A. K., Taluja, H. K., & Singh, A. (2022). Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. Computers in Biology and Medicine, 144, 105354.

Nie, H., Liu, G., Liu, X., & Wang, Y. (2012). Hybrid of ARIMA and SVM for short-term load forecasting. *Energy Procedia*, *16*, 1455-1460.

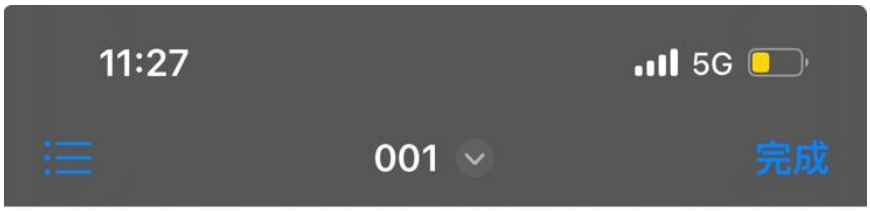
Pai, P. F., & Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, *33*(6), 497-505.

Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*, *140*, 110212.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*, 199-222.

WHO Coronavirus (COVID-19) Dashboard. (n.d.). WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. Retrieved from https://covid19.who.int/

Joseph, T. (2020, August 7). Bootstrapping Statistics. What it is and why it's used. https://www.linkedin.com/pulse/bootstrapping-statistics-what-why-its-used-trist-n-joseph/



Cover page
Abstract
Content table
Acknowledgement
Intro
- Background
- Basic concept and knowledge (1-5 sentence)
- Problem statement (what to solve in the project)
- Value and significant
  - Values
  - Significant (why doing this, any existing work, what you do to overcome the problem)

Literature review (not summary of paper, critically analyzing existing work, identify the weakness of the existing work)
Method
- Methodology
- Design
  - Data collection
  - Model
  - Method to analysis
  - Evaluate metrics
- Implementation details

Experiment and analysis
- Experimental result
- Analysis

Discussion
Conclusion
References

```
library(forecast)
ts_data <-
c(3364,3440,3650,3937,4115,4396,4294,4002,4127,3941,430
9,4580,

4225,4100,3684,3944,3751,4062,4161,4051,3637,3395,3118,
3113,

2932,3150,3059,3044,2687,2469,2408,2605,2481,2362,2142
```

Smets, K., Verdonk, B., & Jordaan, E. M. (2007, August). Evaluation of performance measures for SVR hyperparameter selection. In *2007 International Joint Conference on Neural Networks* (pp. 637-642). IEEE.

Chén zhāomíng. (2022). Shēndù xuéxí: Zuìqiáng rùmén mài xiàng AI zhuāntí shízhàn. [Deep learning: the strongest introduction to AI topic practice.] Shēn zhì shùwèi gǔfèn yǒuxiàn gōngsī.

Tyralis, H., & Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. *Algorithms*, *10*(4), 114.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

Probst, P., & Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*(181), 1-18.