Data-Driven Business Analytics for the Tourism Industry in the UK: A Machine Learning Experiment Post-COVID

Johnson Uke Obogo
Department of Computing and Informatics
Bournemouth University
Poole, United Kingdom
s5229337@bournemouth.ac.uk

Festus Fatai Adedoyin
Department of Computing and Informatics
Bournemouth University
Poole, United Kingdom
fadedoyin@bournemouth.ac.uk

***Abstract*— The use of data-driven business analytic models has had a significant impact on several sectors of the economy. In the UK, the tourism industry has contributed significantly to the economy. The contribution of tourism to the UK economy is estimated to be £145.9 billion (7.2%) of UK GDP. Regardless of its economic value, tourism is also one of the most vulnerable sectors, as it is susceptible to natural disasters, civil unrest, crisis, and pandemics, all of which can fully shut down the industry. Hence, an accurate and reliable tourism demand forecast is important. Apart from COVID-19, no other occurrence in modern history has had such a broad impact on the economy, industries, everyone and businesses in the world (Galvani et al., 2020). However, with the impact of COVID19 on the industry, it is imperative to reassess potential recovery plans for the UK economy, particularly for local tourism businesses. Macroeconomic data is collected over many source markets for the UK and a machine learning algorithm is tested to assess the future of the industry.**

Keywords— tourism businesses, COVID-19, data-driven business analytics, machine learning

I. INTRODUCTION

Tourism is one of the fastest-growing industries, with many countries relying on it for revenue, development, and growth. These changes are the product of tourist consumption of goods and services, taxes imposed on tourism-related businesses, and potential employment opportunities in industries that provide services to the tourism industry. According to the United Nations World Tourism Organization (UNWTO) (2019), global tourism growth continues to outpace global economic growth, making tourism a global driving force for economic growth and development that has spread globally. Tourism has developed into a significant source of revenue, economic development, jobs, tax revenue, wages, and foreign exchange, as well as a source that contributes to a country's national Gross Domestic Product (GDP).

The contribution of tourism to the UK economy is estimated to have contributed £145.9 billion (7.2%) of UK GDP. Regardless of its economic value, tourism is also one of the most vulnerable sectors, as it is susceptible to natural disasters, civil unrest, crisis, and pandemics, all of which can fully shut down the industry. Hence, accurate and reliable tourism demand forecast is important. Accurate tourism demand forecasts are helpful in making strategic, tactical, and operational decisions. Forecasting in tourism entails predicting the course of potential demand, hence, offering useful knowledge to destination management and tourism providers [1]. A variety of measures are used in tourism market modelling and forecasting. However, tourist arrivals are a widely used metric.

Time series and econometric models have been used in the past to forecast tourism demand [2]. Research by [2] shows autoregressive integrated moving average (ARIMA) models as the most frequent models in forecasting tourism demand. However, econometric models have the advantage of being able to recognize economic factors that affect tourism demand [3-4]. Most notably, a variety of tourism demand forecasting in times of crisis has been conducted by researchers, [5] assess the effects of both the global economic crisis and the swine flu pandemic on demand for the United Kingdom. Furthermore, Artificial neural networks (ANNs) and support vector machines (SVMs) are commonly used machine learning methods used for tourism demand forecasting [6]. These methods do not rely on particular statistical characteristics such as normality and linearity of the distribution of the dataset [2], as well as being more robust against skewed, missing, duplicate, and noisy data [2]. As a result, machine learning–based approaches usually produce better results, not only for noncausal but also for causal approaches [7], [2], [8].

Apart from COVID-19, no other occurrence in modern history has had such a broad impact on the economy, industries, everyone and businesses in the world (Galvani et al., 2020). The COVID-19 pandemic has caused an excessive number of deaths globally and led to severe losses in revenue in the United Kingdom tourism sector. As reported by the (World Health Organization; WHO 2021), 111,102,016 confirmed cases of COVID-19 and 2,462,911 deaths had been reported globally as of February 22, 2020. These deaths rate cuts across countries that are tourism top source market to the UK. The figure 1 shows the leading inbound travel markets for the United Kingdom (UK) in 2018 and 2019, ranked by number of visits.

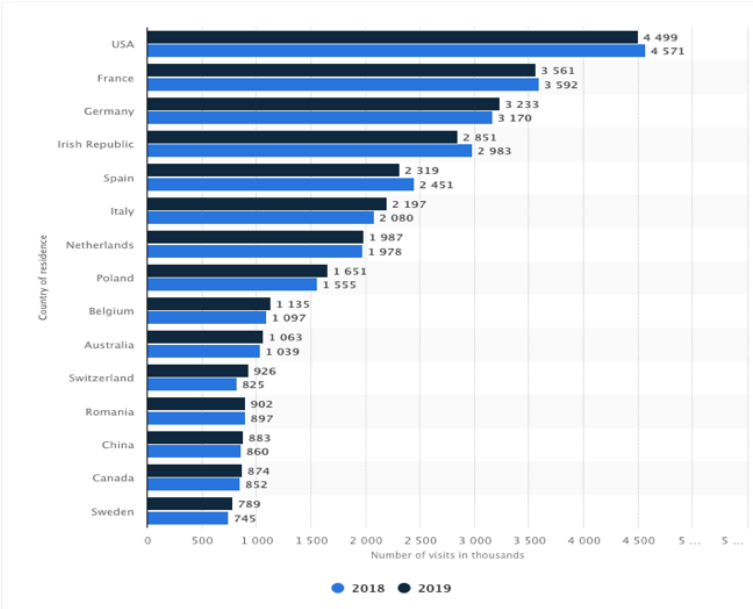

Figure 1: Leading inbound travel markets for the United Kingdom in 2018 and 2019

Travel restrictions were imposed on all flights to the United Kingdom due to the high death rate and rapid spread of the virus. This resulted in a total of 7.0 million visits to the UK from overseas residents in Quarter 1 (Jan to Mar) 2020 (mostly in March), a 16 percent decline from the same timeframe the previous year. As a result of the COVID-19 pandemic, foreign visitors to the UK spent £4.3 billion in Quarter 1 2020, which was 10% less than in Quarter 1 2019, as travel was limited due to the coronavirus (COVID-19) pandemic. The (United Nations specialised origination; UNSO 2021) forecasts that the pandemic may cause a further decrease in tourism of 60 percent to 80 percent in relation to 2019. This has the potential to put many businesses and income streams at risk, as well as halt progress toward achieving sustainable development goals. Pandemics like the COVID-19 pandemic are sporadic and unpredictable. Hence, accurate inbound tourism demand forecasting is critical for the tourism industry in the United Kingdom to effectively address potential challenges, encourage tourism, and allocate adequate resources for operations, marketing, investment, and financial planning. This project utilises inbound tourist arrivals data from 38 countries, the dataset is taken from online reputable source. Different machine learning algorithm such as Random Forest regression, support vector regression, Multiple Linear Regression and Polynomial regression will be used to forecast the inbound tourism demand to the UK while examining COVID-19 deaths from the originating countries.

The tourist destination of United Kingdom serves as a case for this study, using arrival time series data between 1995 and 2018 of 38 sending countries to the UK. Furthermore, the arrival data is reduced to top ten major source market countries to the UK (i.e., Belgium, Canada, Germany, France, Italy Netherlands, Poland, Sweden, Spain, and United State of America) with high COVID-19 deaths and other countries with low covid 19 deaths.

This study aims to see how well machine learning models like multiple linear regression, polynomial regression, support vector regression, and random forest regression predict inbound tourist arrivals to the UK. Also, to take a different approach by applying machine learning methods in forecasting the inbound tourism demand to the UK from 38 countries, 10 source counties with high COVID-19 deaths and other countries with low COVID-19 deaths. The findings will contribute to the tourism industry and the government by providing information needed to make better plans, policy, and decision. Also, it aims to give some references for future research.

To realise the study's aim and attempt to respond to the research question, the follow research objective will be sought after: to implement and build machine learning models for forecasting inbound tourism demand to the UK; evaluation of applied machine learning models; to implement and use the machine learning model with the best performance to forecast inbound tourism demand to the UK from 38 countries; to implement and use the machine learning model with the best performance to forecast inbound tourism demand to the UK from 10 top source market countries with high COVID-19 deaths (minimum of 10,000 deaths) because they are relatively mature tourism source market to the UK; and to implement and use the machine learning model with the best performance to forecast inbound tourism demand to the UK from countries with low COVID-19 deaths (below 10,000 deaths).

II. Literature Review

A. Tourism Demand Forecasting

Forecasting tourism demand is a well-established research area that has been the focus of numerous studies in the tourism and hospitality fields. Studies on tourism demand forecasting can be divided into qualitative and quantitative approaches [2]. With regards to forecasting models, a number of techniques have been used to forecast tourism demand. These approaches can be classified as time-series models, econometric models, or AI models, according to [2]

Time series models use historical data to classify patterns and trends in order to forecast future values of the series based on previous values and to investigate the relationships between different tourism demand variables and tourist arrival volumes. These time series models are widely used in forecasting tourism demand [9-10].

Econometric models have the advantage of integrating relationships between tourism demand (dependent variable) and its influencing factors (explanatory variables), hence, econometric models are replacing time series models in forecasting tourism demand [2]. Tourism price and tourism income are two widely used economic variables correlated with origin and destination countries [11].

Artificial Intelligence (AI) models in tourism demand forecasting is still new and are closely related to machine learning and soft computing methods. AI models have risen in popularity as data volume has increased and data characteristics have become more complex. They can capture complex relations and patterns in a large volume of data. Artificial neural networks (ANN) (Constantino, Fernandes, and Teixeira, 2016), machine learning techniques such as support vector regression (SVR) model [12-13] ; and fuzzy time series models [14] are popular AI models used in tourism demand forecasting [15-16].

Machine learning is a cutting-edge methodology for recognizing, understanding, and analysing highly complex data structures and patterns [17]. With a systematic input of more recent results, it allows for consequential learning and enhances model predictions [18-19]. Machine learning research is a branch of artificial intelligence (AI) that aims to teach computers new information through the input of data such as texts, images, and numerical values, as well as support their interaction with other computer networks. Machine learning techniques have the advantage over statistical methods in that they do not make any assumptions about the data, such as normal distribution, linearity, or noncollinearity [20],[2], [21]. Concretely, the machine learning methods ANN, rough set theory, fuzzy time-series method, genetic algorithms (GAs), SVMs, and, most recently, deep learning approaches [22] are commonly used for tourism demand forecasting [22-25]. Applications of ANNs for tourism demand prediction are presented by [23-24],[26].

B. Role of IoT in Data collection for Tourism demand forecasting

The Internet of Things (IoT) is a global network in which sensors communicate with one another over the internet in order to collect and transmit data. This data is collected over the internet and stored in a database, where it is analysed to assist various business models in resolving various business issues. Data collection can be costly, and accurate data is needed to validate a quantitative model. Observing and capturing complex and multidirectional data may be part of the data collection process. Since manual data collection is vulnerable to bias and recording errors, automated data collection has grown in significance and acceptance. IoT plays an important role in the collection of data in tourism demand forecasting as sensors are placed at the data collection points with unique identifiers to collect data.

In forecasting tourism demand, data is critical. When it comes to Time Series Forecasting, for example, the data obtained in the past assists in predicting future patterns. The majority of the data used in each of the methods above will be primary data. Primary data should be obtained in an impartial and error-free manner. The following benefits can be achieved by using real data produced by IoT devices and can be used in analytical models for forecasting tourism demand.

For many years, the tourism industry has been adapting to emerging technology and has shown several changes in its operations and processes. With the advent of the Internet of Things, real-time data can be collected in an unbiased and error-free manner, and these data can be used to predict inbound tourism demand, airline flight volume, and other transportation access areas. In addition, the Internet of Things has enabled many cities to become smarter by providing smart buildings, smart hospitals, and smart transportation. Hence, making tourist have a wonderful experience.

Real-time data produced by IoT devices aids tourism organizations and the public sector in making decisions based on forecasting performance, preventing further losses. It also aids tourism organizations in making forecasting error corrections, as real-time insights into forecasting errors enable organizations to respond quickly to mitigate the effects of an operational problem.

When designing forecasting models, the alternative model is chosen based on the most recent vintage of historical data available. The outputs are produced, and the resulting forecast errors are compared to evaluate the alternative forecasts. The comparison is also expanded to include the forecast produced using real-time data. This results in two types of benefits from using IoT to collect real-time data: More data can be used to develop the model and the data can be over time revised.

C. Tourism and COVID-19 in the UK Economy

Tourism is a major part of the United Kingdom economy. The Office for National Statistics (ONS) shows that in 2017 tourism directly contributed an estimated GBP 67.8 billion in gross value added (GVA), or around 3.9% of the total United Kingdom economy.

Since late 2019, the corona virus disease 2019 (COVID-19) pandemic has wreaked havoc on the world's health and social structures, as well as wreaking havoc on the global economy. As shown in figure 2, the tourism and hospitality industry is one of the most adversely impacted by the COVID-19 pandemic [27]. Airline staff have been reduced by 90%, 80% of hotel rooms are vacant, and tourist destinations lost money in 2020 [27]. Countries' lockdowns, widespread travel bans, and airport and national border closures decreased international tourist arrivals by 67 million in the first quarter of 2020 (2020Q1), according to UNWTO.

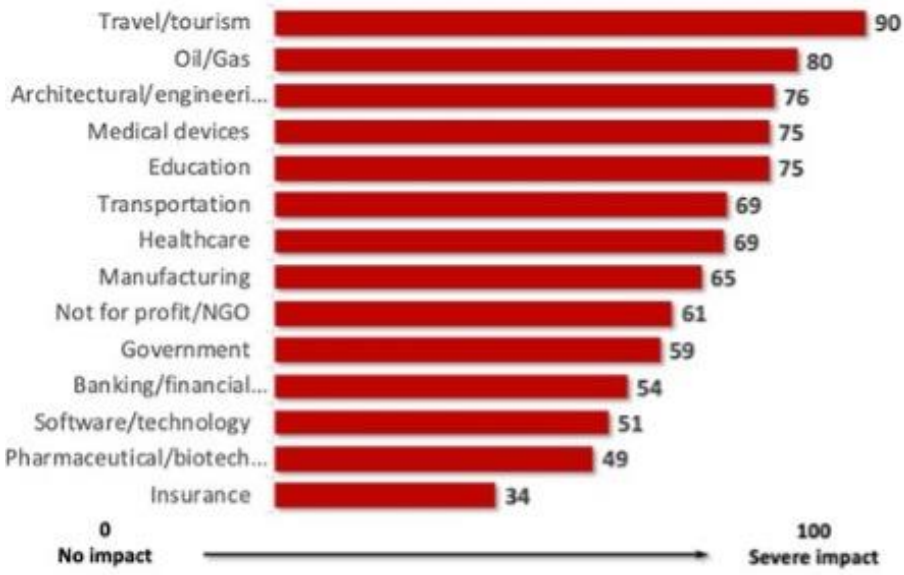

Figure 2: Figure 2: Impact of COVID-19 on Industries [27].

III. Data And Variables

Explanatory variables for tourism demand include macroeconomic variables like Gross domestic product, exchange rate, and consumer price index [28]. These variables were gotten from reliable online open data sources with monthly and annually frequency and were included in the dataset.

In this study, data on the annual inbound tourist arrivals in UK gotten from UNWTO are used. The real Gross domestic product (for the originating countries and destination country), Consumer price index (for the originating countries and destination country), and population data were collected from the International Monetary Fund (IMF). Exchange rate data were collected from the World bank open source. The sample period of the data used in this study is from 1995 to 2018 and the observations of tourist arrivals are at the annual frequency. The inbound tourist arrival dataset used in this study consist of the following columns; relative price; inbound Tourist Arrivals; relative income; and population.

A. Shapiro-Wilk Test

The dataset distribution influences which statistical test should be used to find evidence on the project study issue. The parametric test is used when data are normally distributed; otherwise, the non-parametric test is used [31]. The Shapiro-Wilk test is performed in this study to determine whether the data has a Gaussian (normal) distribution or not. The null hypothesis for the Shapiro-Wilk test is that the data is normally distributed while the alternative hypothesis is that the data is not normally distributed [32]. In Shapiro-Wilk test, if the p value is greater than 0.05, it implies the data is not normally distributed, while if it is less than 0.05, it means the data is normally distributed.

B. Spearman's Correlation Coefficient (R) Test

A statistical measure of the direction and intensity of the relationship between the dependent and independent variables is Spearman's correlation coefficient (R) [33]. In the Spearman's correlation coefficient test, the value of R indicates the strength of the relationship between the dependent and independent variable. R has a value that ranges from +1 to -1. A value of +1 or -1 indicates that the dependent and independent variables are perfectly associated. As R approaches 0, the relationship between the independent and dependent variables will deteriorate. A probability value (p-value) is calculated as part of the statistical test. The p-value is considered statistically significant if its value is less than 0.05 (i.e. p less than 0.005).

C. Data Preparation

In this study, Data is obtained in raw format from various sources, which makes it difficult to train a model and reliably predict the inbound tourist arrivals, making data pre-processing a crucial phase to implement. The steps involved in data preparation are listed below.

D. Data Cleaning

The handling of missing values, as well as the removal of noisy and inconsistent data based on domain information, are all important steps in data cleaning. These steps are essential for converting raw data into a more accurate result. The problem of missing data arises during the compilation process, as a result of a device failure or data entry errors. Missing values are manually or computationally extracted or filled, and noisy data such as outliers or incorrect data is removed based on domain awareness [34]. To account for the role of COVID-19 deaths in forecasting inbound tourism to the UK, the data was further reduced to having just tourism demand data from source market countries to the UK with high and low COVID-19 deaths (top source countries with over 10,000 deaths).

IV. Model and Methodology

in this study, inbound tourist arrivals to the UK as well as some influencing econometric variable like relative price, relative income and population are used to forecast inbound tourism demand to the UK. In this study, actual GDP of the source and destination market countries is chosen to measure the relative income level of the origin country. The relative income is defined as Relative income = GDP_{UK}/GDP_i where GDP_{UK} is the GDP for UK and GDP_i is the GDP for the origin country.

Relative price, CPI, population, GDP, and exchange rates are also taken into account. In this paper, calculating the

relative price variables of tourism in UK is defined as CPI = $\frac{CPI_{UK}}{CPI_i}\ X\ EX_{UK}$

A. Train and Test Split

Train and test split is used to divide the dataset into two known as train and test. The train dataset is used to train and fit the machine learning models. The model enhances its parameter using the train data to forecast output values. The test data is used to assess the machine learning model's performance.

B. Feature Scaling/ Pre-processing

Feature scaling is used for the normalization of the range of separate variables or data set features [35]. Feature scaling is important as machine learning models are more effective when the data has a consistent distribution.

C. K-Means Clustering

Clustering is a grouping technology used to find information and trends about the data structure [36]. K-means clustering is a straightforward and accessible unsupervised machine learning algorithm for combining similar observations/data-points and finding underlying patterns. A cluster refers to a collection of data-points aggregated together because of certain similarities. Inbound tourism to the UK has different tourist arrivals volume among different countries with high and low COVID-19 deaths.

D. Machine Learning Model

In forecasting the inbound tourist arrivals to the UK with the historical data in this study, different machine learning models are explored in this study. This section provides Information on the forecasting models implemented in this study.

1) Support Vection Regression

SVR has recently emerged as an alternative and highly effective means of solving the nonlinear regression problem. SVR has been quite successful in both academic and industrial platforms owing to its many attractive features and promising generalization performance. In this study, SVR is chosen as one of the machine learning models because of some significant features of SVR such as: (i) it can model nonlinear relationships, (ii) the SVR training process is equivalent to solving linearly constrained quadratic programming problems, and the SVR embedded solution meaning is unique, optimal and unlikely to generate local minima, and (iii) it chooses only the necessary data points to solve the regression function, which results in the sparseness of solution.

2) Random Forest Regression (RFR)

For the prediction of future results, many machine teaching approaches can be used, but Random Forest regression machine learning method is employed as one of the methods for forecasting in this study. Random forest regression is a supervised machine learning algorithm that trains and analyses previous data samples through various trees.

3) Multiple Linear Regression (MLR)

Regression is a statistical empirical technique that uses the relation between two or more quantitative variables so that an outcome variable can be predicted from the others. MLR is basically a regression model that examines the relationship between the dependent variable and multiple independent variables. Often the regression models are used to predict future values of the response variable for certain values of the response variables. In this study, the Linear Regression class from the sklearn package is used. The fit () function is used to train the model, adjusting weights according to the data values in order to achieve better accuracy.

4) Polynomial Regression

Polynomial Regression Model is a type of Linear Regression model. Linear Regression calculates the forecast according to the Polynomial Regression forecast method based on an nth degree polynomic, to best fit the historical data. The polynomial degree is indicated by the Degree for Polynomial Regression field. For a best-fit model, a higher-order equation with optimal degree value is required. If the degree is too high or too low, the model can be overfitting or underfitting, respectively [37].

E. Model Evaluation Selection

Mean square error (MSE), mean absolute error (MAE), and root mean squared error were the measurement metrics used to evaluate the in-sample and out-of-sample forecasting output of the different models (RMSE). The correlation coefficient (R^2) was also used to determine how closely the expected values matched the real values.

1) R Squared Score

R squared scare is a statistical measure which represent the proportion of the variance for a dependent variable by an independent variable in regression model, R squared explains to what extend the variance of one variable explain the variance of second variable (Han and Chi, 2016). It could be identified using following formulae:

$$R^2 = \frac{\text{variance explained by model}}{\text{Total Variance}} \quad (1)$$

2) Mean Absolute Error

It is a set of average number of errors in model prediction between model forecasting and real data, it is an average of test results [38]. Its value ranges from 0 to infinity. It is also called negatively oriented score as fewer score value of MAE shows the goodness of learning model. MAE is represented with the formula below

$$\textbf{MAE} = \frac{1}{m}\sum_{i=0}^{m} | x^i - y^i | \quad (2)$$

3) Mean Squared Error (MSE)

Mean square error is a method of monitoring a regression model's efficiency. MSE takes and squares the gap from the regression line to the sample points. Quadrating is important since the sign is negative, which is omitted [38]. The least MSE illustrates closeness of detecting the right match. It could be identified using following formulae:

$$\textbf{MSE} = \frac{1}{m}\sum_{i=1}^{m}(x^i - y^i)^2 \quad (3)$$

4) Root Mean Squared Error (RMSE)

The root mean square error estimates the average magnitude of the forecasting error. RMSE works by squaring the observed and forecast value and then averaged over the observations. The square root of the average is finally taken. Since the errors are squared before being averaged, it gives relatively more weight to larger error differences which means RMSE is very useful [39].

$$\textbf{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x^i - y^i)^2} \quad (4)$$

V. Results and Discussions

The United Kingdom has been a source of tourist destination over the years, and figure 3 shows the number of tourist arrivals to the UK from 1995-2018. Figure 4 shows the number of tourist arrivals from top source market countries to the UK. France has the highest number of inbound tourist arrivals to the UK with USA being the second. Interestingly, as the top source market countries with high tourist arrivals to the UK are also top market countries with high COVID-19 deaths as shown in figure 1.

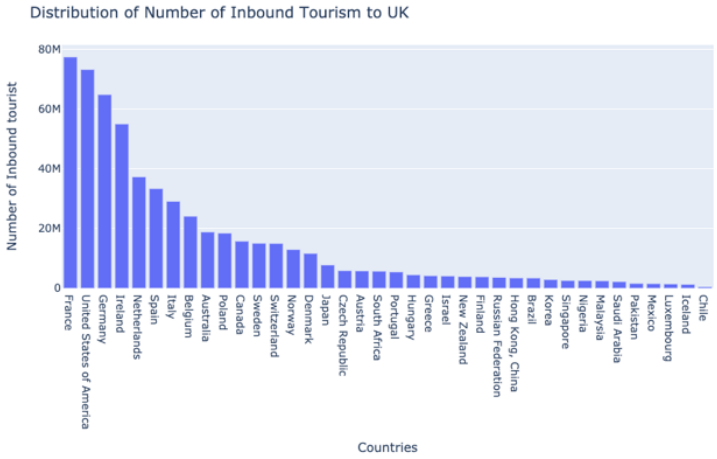

Figure 3: Total Tourist arrival volume per country

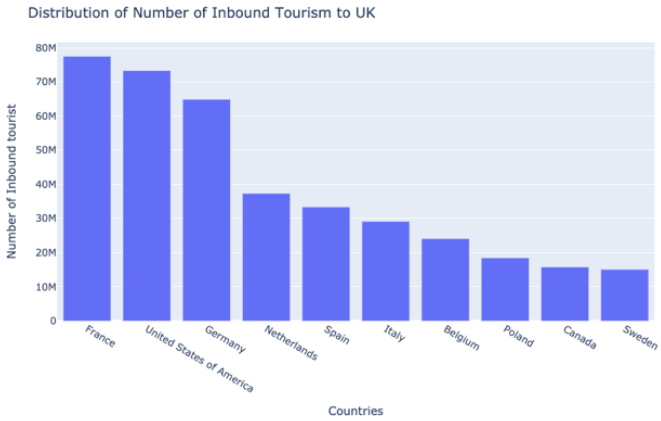

Figure 4: Total Tourist Arrivals volume for top 10 countries

A. Grouping of Countries

The countries are grouped into clusters and each cluster represents the inbound tourist arrival volume to the UK. The deeper the colour the higher the tourist arrivals and vice versa. The deep blue colour represents cluster 1 (countries with high inbound tourist arrival volume to the UK) while the faded colour represent cluster 2 (countries with low tourist arrivals volume). For sample size, K-means clustering was applied on the top 10 source market countries to the UK because K-means clustering on sample size implies that the result extend beyond the selected countries. Figure 5 shows USA, France and Germany being in the same cluster

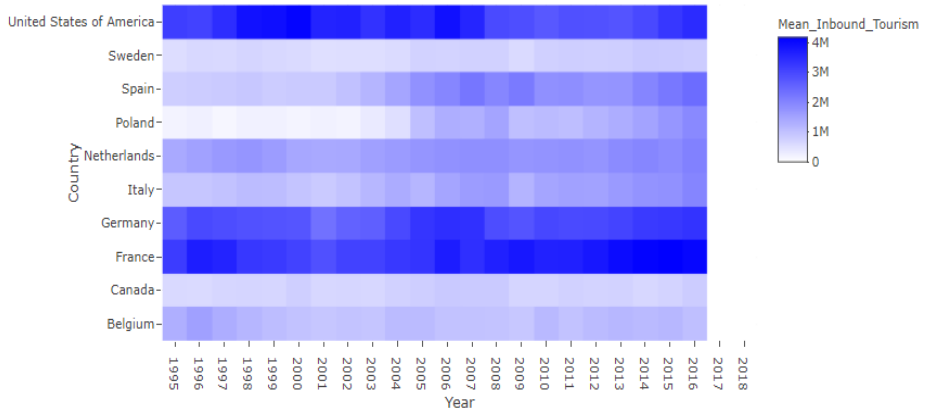

Figure 5: Clustering of Countries with similar trend

B. Machine Learning model performance Evaluation

Table 1,2,3 presents the evaluation result for Random Forest Regression, Support Vector Regression, Polynomial Regression and Multiple Linear Regression model, based on their performance metrics: R^2, MSE, RMSE and MAE, under the three scenarios mentioned above.

	TOURISM FORECAST FOR ALL COUNTRIES			
Metric	R^2	RMSE	MSE	MAE
Random Forest Regression	0.95404	0.21573	0.04654	0.09687
Support Vector Regression	0.59415	0.64110	0.41101	0.40347
Polynomial Regression	0.66266	0.58448	0.34162	0.41388
Multiple Linear Regression	0.22572	0.78335	0.78411	0.63921

Table 1: Evaluation Result for all four models on tourism arrival data from 38 Countries with high and low COVID-19 deaths

	FORECAST FOR SOURCE MARKET TO UK WITH HIGH COVID-19 OF 10,000 AND ABOVE			
Metrics	R^2	RMSE	MSE	MAE
Random Forest Regression	0.99347	0.09429	0.00889	0.05686
Support Vector Regression	0.37360	0.92410	0.85396	0.41214
Polynomial Regression	0.76101	0.57079	0.32580	0.40419
Multiple Linear Regression	0.24601	1.01385	1.02789	0.66024

Table 2: Evaluation result for all four models on tourist arrival data from top source market countries to the UK with high COVID-19 deaths

	COUNTRIES WITH LOW COVID-19 DEATHS			
Metrics	R^2	RMSE	MSE	MAE
Random Forest Regression	0.91798	0.20189	0.04076	0.12129
Support Vector Regression	0.28918	0.59435	0.35325	0.32230
Polynomial Regression	0.35822	0.56475	0.31894	0.40664
Multiple Linear Regression	- 0.23476	0.78335	0.61364	0.60043

Table 3: Evaluation result for all four models on tourist arrival data to the UK from countries with low COVID-19 deaths.

The model predictions performance was examined using $R^{2,}$ RMSE, MSE and MAE. As observed in Table 2, the Random Forest Regression model has the highest R2 value and the lowest RMSE, MSE and MAE as compared to other models in all three scenarios. This shows that the random forest regression model fits the data correctly and is accurate and precise in forecasting inbound tourism demand considering the COVID-19 crises as the prediction error is lower than other models.

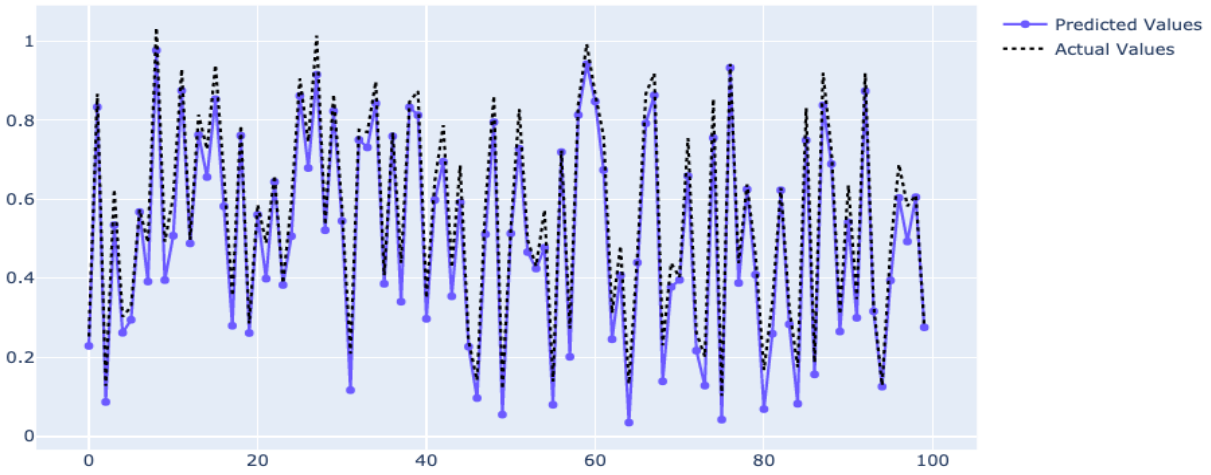

Figure 6: Random Forest Regression model performance on predicting Inbound tourism demand

In figure 6, the Blue colour line indicates what the Random Forest regression model predicted, and the dotted line is the actual value. From the figure, it shows that the predicted values are accurate and very close to the actual. It is basically trying to learn each point and curve.

Table 4: Percentage Relative Difference between predicted and actual value

	Random Forest Regression model performance on predicting Inbound tourism demand for countries with high COVID-19 deaths	*Random Forest Regression model performance on predicting Inbound tourism demand for countries with low COVID-19 deaths*	*Random Forest Regression model performance on predicting Inbound tourism demand for all countries (with high and low COVID-19 deaths)*
Predicted Vaue	**276.7**	**464.52**	**971.11**
Actual Value	**277.16**	**468.98**	**972.81**
Relative difference	**0.20%**	**0.95%**	**0.20%**

Since the forecasting error in Random Forest regression model is lower that other models and since it outperforms other machine learning models, it is used to perform the forecast. The Random Forest regression model predicts a decrease in inbound tourist arrivals in all three scenarios in the study. Table 4 shows that the Random Forest regression model predicts a decrease of 0.2% from all sending countries in the dataset, 0.2% decrease from top source market countries with high COVID-19 deaths and 0.95% decrease from countries with low COVID-19 deaths. Although there is a little decrease in the inbound tourist arrival to the UK, it shows the top source market countries would still have the highest inbound tourist arrivals volume to the UK. However, the little decrease could affect the tourism sector if proper actions are not taken.

The tourism demand forecast in this study would not have been possible if there was not enough data for forecasting, indicating the relevance of Internet of Things in tourism demand forecasting. Since effective tourism forecast depends on accurate real time data, this study shows that Internet of Things can influence the entire tourism sector and tourism demand forecasting by various means including: (i) making available the data about tourism demand available, consistent, and immutable.; (ii) the response of the tourism industry can be improved and operating costs for the hospitality sector can be reduced

VI. CONCLUSION

In this study, R^2 is used to demonstrate how well the machine learning models fit the model and in comparing the model's performance RMSE, MSE, and MAE is used. The forecasting study was conducted for United Kingdom, using arrival data of 38 sending countries with high and low COVID-19 deaths. Data analysis was firstly done to identify and understand the number of tourist arrivals from different countries to the UK. It is discovered that some countries are top tourism source market to the UK and then K-means algorithm was applied to group the top source market countries in clusters. Based on the hypothesis testing, an increase in relative income and population would cause an increase in tourist arrivals to the UK. Lastly, different machine learning models to forecast the inbound tourism demand to the UK was implemented.

Random forest regression model fits the data and performed the best as compared to other models based on the evaluation metrics (RMSE, MSA and MAE) and its capability to being robust to outlier while Polynomial regression performed the worst. With regards to the forecast of the best model (Random Forest Regression), it predicts a decrease in the inbound tourist arrivals to the UK. It shows a 0.2% decrease from all countries in the dataset, 0.2% decrease from top source market countries with high COVID-19 deaths and 0.95% decrease from countries with low COVID-19 deaths. Based on the analysis and forecasting result, it shows that economic factors play an important role in inbound tourism to the UK and UK would continue to receive high volume of inbound tourist from the top source market countries. However, since population has a significant effect on inbound tourism to the UK, and the top source market countries also has huge population, it is recommended that effective and significant restrictions are implemented on tourist arrivals from the top market countries in order to prevent the spread of the COVID-19 virus.

REFERENCES

[1] Vanhove, N. 2011. The Economics of Tourism Destinations.London: Elsevier.

[2] Song, H. and Li, G. (2008) 'Tourism demand modelling and forecasting-A review of recent research', Tourism

Management, 29(2), pp. 203–220. doi: 10.1016/j.tourman.2007.07.016.

[3] Peng, B., H. Song, and G. Crouch. 2014. "A Meta-analysis of International Tourism Demand Forecasting and Implications for Practice." Tourism Management 45:181–93.

[4] Athanasopoulos, G., H. Song, and J. A. Sun. 2018. "Bagging in Tourism Demand Modelling and Forecasting." Journal of Travel Research 57 (1): 52–68.

[5] Page, S., Song, H. and Wu, D. C. (2012) 'Assessing the Impacts of the Global Economic Crisis and Swine Flu on Inbound Tourism Demand in the United Kingdom'. doi: 10.1177/0047287511400754.

[6] Moro, S., and P. Rita. 2016. "Forecasting Tomorrow's Tourist." Worldwide Hospitality and Tourism Themes 8 (6): 643–53.

[7] Kon, S. C., and W. L. Turner. 2005. "Neural Network Forecasting of Tourism Demand." Tourism Economics 11:308–28.

[8] Ricardo, H., I. Gon.avales, and A. C. Costa. 2018. "Forecasting Tourism Demand for Lisbon's Region through a Data Mining Approach." In Proceedings of the 11th IADIS International conference on Information Systems 2018, 58–66. IS2018.

[9] Andrew W.P, Cranage D.A, C.K. Lee (1990) 'Forecasting hotel occupancy rates with time series models: An empirical analysis' Hospitality Research Journal, 14 (2) (1990), pp. 173-182

[10] Frechtling and Frechtling, 2001 D.C. 'Forecasting tourism demand: methods and strategies'. Butterworth-Heinemann, Oxford, Boston (2001).

[11] Ayeh, J. and Lin, V. S. (2011) 'Tourism and Hospitality Research', (March 2016). doi: 10.1177/1467358411415466.

[12] Pai P.-F, W.-C. Hong, P.-T. Chang, C.-T. Chen (2006). 'The application of support vector machines to forecast tourist arrivals in Barbados: An empirical study'. International Journal of Management, 23 (2) (2006), p. 375

[13] Zhang, B., X. Huang, N. Li, and R. Law. (2017). "A Novel Hybrid Model for Tourist Volume Forecasting Incorporating Search Engine Data." Asia Pacific Journal of Tourism Research 22 (3): 245–54.

[14] Chen, K. Y. and Wang, C. H. (2007) 'Support vector regression with genetic algorithms in forecasting tourism demand', Tourism Management, 28(1), pp. 215–226. doi: 10.1016/j.tourman.2005.12.018.

[15] Sun, S. et al. (2020) 'Tourism demand forecasting with tourist attention: An ensemble deep learning approach', arXiv.

[16] Zhang, Y. et al. (2020) 'Tourism Demand Forecasting: A Decomposed Deep Learning Approach', Journal of Travel Research. doi: 10.1177/0047287520919522.

[17] Ngiam, K. Y. and Khor, I. W. (2019) 'Big data and machine learning algorithms for health-care delivery', The Lancet Oncology, 20(5), pp. e262–e273. doi: 10.1016/S1470-2045(19)30149-4.

[18] Harrington, P., (2012). Machine learning in action. Manning Publications Co.

[19] Schölkopf, B. and Smola, A. J. (2002) 'Support Vector Machines and Kernel Algorithms', The Handbook of Brain Theory and Neural Networks, (April 2002), pp. 1119–1125.

[20] Yu, G. and Schwartz, Z. (2006) 'Forecasting short time-series tourism demand with artificial intelligence models', Journal of Travel Research, 45(2), pp. 194–203. doi: 10.1177/0047287506291594.

[21] Song, H., R. T. R. Qiu, and J. Park. (2019). "A Review of Research on Tourism Demand Forecasting Methods." Annals of Tourism Research 75:338–62.

[22] Law, R. et al. (2019) 'Tourism demand forecasting: A deep learning approach', Annals of Tourism Research. Elsevier, 75(October 2018), pp. 410–423. doi: 10.1016/j.annals.2019.01.014.

[23] Law, R. and Au, N. (1999) 'A neural network model to forecast Japanese demand for travel to Hong Kong', Tourism Management, 20(1), pp. 89–97. doi: 10.1016/S0261-5177(98)00094-6.

[24] Claveria, O. and Torra, S. (2014) 'Forecasting tourism demand to Catalonia: Neural networks vs. time series models', Economic Modelling. Elsevier B.V., 36, pp. 220–228. doi: 10.1016/j.econmod.2013.09.024.

[25] Mavrommati, A. and Karakitsiou, A. (2018) 'Machine learning methods in tourism demand forecasting: some evidence from Greece', MIBES Transactions, 11(1), pp. 92–105.

[26] Network, A. N. (2019) 'Tourism demand forecasting – a review on the variables and models Tourism demand forecasting – a review on the variables and models'. doi: 10.1088/1742-6596/1366/1/012111.

[27] Silva, E. S. et al. (2019) 'Forecasting tourism demand with denoised neural networks', Annals of Tourism Research. Elsevier, 74(October 2018), pp. 134–154. doi: 10.1016/j.annals.2018.11.006.

[28] Qiu, R. T. R., Park, J., Li, S. N., and Song, H., 2020. Social costs of tourism during the COVID-19 pandemic. Annals of Tourism Research.

[29] Daniel, A. c. M., & Ramos, F. F. R. (2002). 'Modelling inbound International tourism demand to Portugal'. The International Journal of Tourism Research, 4(3), 193.

[30] Braumoeller, B. F., 2004. Hypothesis testing and multiplicative interaction terms. International Organization.

[31] Angelini, C., 2018. Hypothesis testing. In: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics.

[32] McCrum-Gardner, E., 2008. Which is the correct statistical test to use? British Journal of Oral and Maxillofacial Surgery.

[33] Royston, P., 1992. Approximating the Shapiro-Wilk W-test for non-normality. Statistics and Computing.

[34] Hauke, J. and Kossowski, T., 2011. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae.

[35] Zhang S., C. Zhang, Q. Yang (2003). 'Data preparation for data mining'. Applied Artificial Intelligence, 17 (5–6) (2003), pp. 375-381.

[36] Singhal, S. and Jena, M., 2013. A Study on WEKA Tool for Data Preprocessing , Classification and Clustering. International Journal of Innovative Technology and Exploring Engineering.

[37] Khan, H. R. and Hossain, A., 2020. Countries are Clustered but Number of Tests is not Vital to Predict Global COVID-19 Confirmed Cases: A Machine Learning Approach. medRxiv.

[38] Bruno, D. E., Barca, E., Goncalves, R. M., de Araujo Queiroz, H. A., Berardi, L., and Passarella, G., 2018. Linear and evolutionary polynomial regression models to forecast coastal dynamics: Comparison and reliability assessment. Geomorphology.

[39] Willmott, C.J. and Matsuura, K., (2005). 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance'. Climate research, 30(1), pp.79-82.

[40] [39] Bontempi, G., Ben Taieb, S., and Le Borgne, Y. A., 2013. Machine learning strategies for time series forecasting. In: Lecture Notes in Business Information Processing.