Journal of Universal Computer Science, vol. 27, no. 6 (2021), 564-581
submitted: 20/1/2021, accepted: 20/4/2021, appeared: 28/6/2021

Forecasting Air Travel Demand for Selected Destinations Using Machine Learning Methods

Murat Firat
(Istanbul University-Cerrahpasa, Istanbul, Turkey
https://orcid.org/0000-0002-7480-6600, bmfirat@gmail.com)

Derya Yiltas-Kaplan
(Istanbul University-Cerrahpasa, Istanbul, Turkey
https://orcid.org/0000-0001-8370-8941, deryayiltas@gmail.com)

Ruya Samli
(Istanbul University-Cerrahpasa, Istanbul, Turkey
https://orcid.org/0000-0002-8723-1228, ruyasamli@iuc.edu.tr)

Abstract: Over the past decades, air transportation has expanded and big data for transportation era has emerged. Accurate travel demand information is an important issue for the transportation systems, especially for airline industry. So, "optimal seat capacity problem between origin and destination pairs" which is related to the load factor must be solved. In this study, a method for determining optimal seat capacity that can supply the highest load factor for the flight operation between any two countries has been introduced. The machine learning methods of Artificial Neural Network (ANN), Linear Regression (LR), Gradient Boosting (GB), and Random Forest (RF) have been applied and a software has been developed to solve the problem. The data set generated from The World Bank Database, which consists of thousands of features for all countries, has been used and a case study has been done for the period of 2014-2019 with Turkish Airlines. To the best of our knowledge, this is the first time that 1983 features have been used to forecast air travel demand in the literature within a model that covers all countries while previous studies cover only a few countries using far fewer features. Another valuable point of this study is the usage of the last regular data about the air transportation before COVID-19 pandemic. In other words, since many airline companies have experienced a decline in the air travel operation in 2020 due to COVID-19 pandemic, this study covers the most recent period (2014-2019) when flight operation performed on a regular basis. As a result, it has been observed that the developed model has forecasted the passenger load factor by an average error rate of 6.741% with GB, 6.763% with RF, 8.161% with ANN, and 9.619 % with LR.

Keywords: Air travel, Airline load factor, Artificial Neural Networks, Gradient Boosting, Linear Regression, Random Forest, Travel demand
Categories: I.0, I.2.0, I.2.1, I.2.6, I.m, J.m
DOI: 10.3897/jucs.68185

1 Introduction

One of the main products in the airline sector is the seat inventory of the flights that cannot be stocked as it is only available until the flight operation. Thus, any unbooked seat is a commercial loss after a flight operation. For this reason, a high level of the load factor for the flights is expected due to marketing requirements. Supplying a suitable load factor requires to operate on the markets that have high travel demands.

Determining the markets having high air travel demand is probably one of the most critical decision-making processes in the airline travel sector. This process requires some evaluations of many effective factors. However, evaluating those factors may not be easy as they may vary depending on various numbers of different dynamics. The pricing policy may be an essential factor in this case, but many others affect air travel demand too [Doganis 2010, Smyth and Pearce 2008, Yang et al. 2019]. There are also several cultural and social impacts associated specifically with the personal air travel demand such as average personal income, international business, academical collaborations, study abroad, and foreign holidays [Shaw and Thomas 2006]. Several impacts have been categorized based on the air travel distances by [Schubert et al. 2020] and it has been derived that the gender and life quality have influenced the long-distance travel, whereas personal income, age, geographical context, etc. have had effects on the shorter distance travels.

Initially, this study focuses on the features that are related to air travel demand globally. Totally 318 features were determined accordingly and applied to the model to forecast load factors of the flights operated from Turkey to each country by the Turkish Airlines. This approach also makes it possible to assess the flight operation feasible between any two countries.

The rest of this paper is designed as follows. Section 2 reviews the studies on forecasting air travel demand with different methods and region/route targets. Section 3 presents the data set generation, pre-processing and the proposed approach. Section 4 discusses the experimental results and lastly Section 5 describes the concluding remarks of this study.

2 Literature Review

Forecasting air travel demand has been regarded as a key point in transportation systems since many years. Over the past few decades, several models about forecasting the air travel demand have been developed to manage the transportation system and to improve the efficiency. Artificial Neural Network (ANN) is a popular method in this area and used in many studies. [Blinova 2007] focused on expanding air travel demand between the federal regions of Russia. [Tsekeris 2009] studied on a model to determine the factors that affected air travel demand in the Aegean islands of Greece using ANN. Similarly, [Alekseev and Seixas 2009] used ANN when studying on the domestic air travel demand in Brazil. [Srisaeng 2015] and [Srisaeng et al. 2015] studied on the estimation of air travel demand in Australia. [Srisaeng 2015] used ANN and Adaptive Neuro-Fuzzy Inference System (ANFIS) with considering low cost carriers, while [Srisaeng et al. 2015] used the Genetic Algorithm (GA) for predicting the results of the years between 1992 and 2014. [Ghomi and Forghani 2016] studied on a model to forecast the demand for a flight operated in Turkey using ANN and Box–Jenkins methods with five years of data. [Koc and Arslan 2018] used ANN to forecast a domestic air travel demand in Turkey. [Erjongmanee and Kongsamutr 2018] used Multiple Linear Regression (MLR) and ANN to forecast air travel demand in Thailand. [Sulistyowati et al. 2018] applied ANN and Support Vector Regression (SVR) to forecast the air cargo and the travel demand for three major airports in Indonesia. Apart from ANN, various methods were applied to solve this forecasting problem too. [Ba-Fail et al. 2000] analyzed the determinants of domestic air travel demand in Saudi Arabia and proposed a Linear Regression based equation. [Abed et al. 2001] proposed a model that provided an estimation of international air travel demand in Saudi Arabia. [Brons et al. 2002] examined 37 studies in the literature to investigate the price elasticities of air travel demand globally. [Ashiabor et al. 2007] used Logistic Regression to observe the road and air travel demand in the USA. [Wang et al. 2011] analyzed the determinants of air travel demand and used these evaluated determinants for applying the MLR model to estimate air travel demand for China. [Carson et al. 2011] performed air travel demand estimation in the USA utilizing monthly air traffic data of some major airports from 1990 to 2004 with Ordinary Least Squares Regression (OLS). [Kopsch 2012] studied on the domestic air travel demand about price elasticities for business and leisure travel in Sweden by Prais–Winsten (PW) Regression. [Sivrikaya and Tunc 2013] applied Semi-Logarithmic Regression on each origin-destination pair in Turkey to evaluate potential routes that may provide high profit in the domestic market. [Totamane et al. 2014] predicted the air cargo demand using the Potluck Problem method, [Li et al. 2014] studied historical air travel demand between the origin and destination pairs in the USA using Evolutionary Strategy (ES) method. [Laik et al. 2014] implemented the airport-based passenger load forecasting for multiple airline companies operating at the same airport by using Decision Trees (DT). [Jungmittag 2016] used the seasonal Box–Jenkins model to forecast air travel demand for Frankfurt Airport in Germany with a data set that involves monthly air travel data from 2003 to 2012. [Dantas et al. 2017] combined the Bagging and Holt-Winters (HW) methods to forecast travel demand for 14 countries for the periods of 2007 to 2014. [Chang et al. 2017] applied some methodology for the USA destinations executing the Multi-Objective Programming (MOP). [Deveci et al. 2017] used an alternative interval type-2 fuzzy TOPSIS method to evaluate the best destination pairs between Turkey and North America based on the Multiple Criteria Decision Making (MCDM). [Yang et al. 2017] studied on the demand elasticities of air travel and used the Log-Linear Regression models for the estimations of the China mainland. [Boonekamp et al. 2018] applied the Gravity Model (GM) to analyze the air travel demand among the European airports. They used a data set for 2010 that consists of the route-based air travel data. [Chang et al. 2019] studied on origin and destination pair between Taiwan and Europe that yields maximum total revenue and minimum total cost. They handled some possible destinations in Europe and evaluated the best ones that provide the best profit when executing the Compromise Programming (CP). [Plakandaras et al. 2019] forecasted the demands for the air, road, and train transportation in the USA using the method of SVR. They used monthly data set of transportation between 2000 and 2015.

Table 1 summarizes these literature studies with the methods, the feature numbers, and the region properties. In the literature studies, the numbers of the countries and the data features are very limited. On the other hand, this study covers the implementations for the analysis of all countries at the same time. Another novelty of this study is the usage of a new data set that involves all countries and has not been utilized in any study before.

Table 1 proves that the relevant literature studies in recent years have contained mostly ANN or Regression methods. This current study has been initially originated by implementing Gradient Boosting (GB) and Random Forest (RF) that are mentioned in Section 3.2. After focusing on these boosting and bagging methods, the experimental comparisons have been made with considering the most common methods, namely ANN and Linear Regression (LR).

Studies	Method(s)	Feature Number	Region
[Blinova 2007]	ANN	1	Russia
[Tsekeris 2009]	ANN	8	Greece
[Alekseev and Seixas 2009]	ANN	3	Brazil
[Srisaeng 2015]	ANN, ANFIS	8	Australia
[Ghomi and Forghani 2016]	ANN, Box–Jenkins	1	Turkey
[Koc and Arslan 2018]	ANN	12	Turkey
[Erjongmanee and Kongsamutr 2018]	ANN, MLR	13	Thailand
[Sulistyowati et al. 2018]	ANN, SVR	3	Indonesia
[Ba-Fail et al. 2000]	Stepwise Regression	16	Saudi Arabia
[Abed et al. 2001]	Stepwise Regression	16	Saudi Arabia
[Brons et al. 2002]	Meta-Regression	6	Multiple
[Ashiabor et al. 2007]	Logistic Regression	3	USA
[Wang et al. 2011]	MLR	4	China
[Carson et al. 2011]	OLS	7	USA
[Kopsch 2012]	PW	6	Sweden
[Sivrikaya and Tunc 2013]	Semi-Logarithmic Regression	14	Turkey
[Totamale et al. 2014]	Potluck Problem	13	North America
[Li et al. 2014]	ES	8	USA
[Laik et al. 2014]	DT	5	Asia
[Srisaeng et al. 2015]	GA	11	Australia
[Jungmittag 2016]	Box–Jenkins	1	Germany
[Dantas et al. 2017]	Bagging, HW	1	Multiple
[Chang et al. 2017]	MOP	2	Taiwan-USA
[Deveci et al. 2017]	MCDM	11	Turkey-North America
[Yang et al. 2017]	Log-Linear Regression	7	China
[Boonekamp et al. 2018]	GM	15	Europe
[Chang et al. 2019]	CP	4	Taiwan-Europe
[Plakandaras et al. 2019]	SVR	5	USA

Table 1: Relevant studies

3 Materials and Methods

3.1 Data Sets

The executions in this study were applied to the data publicly presented from The World Bank which contains more than 70 databases, each of which consists of hundreds of features (population, unemployment rate, gross domestic product (GDP), gross national income (GNI), export & import volume, and so on) that belong to each country [The World Bank 2020]. In this study, The World Bank databases are unified in-country, feature, period (year), and feature value level, and missing data has been cleared. The final data set consists of all countries, with each having 1983 features for the years from 2013 to 2019. The sample format of the unified data set is shown in Table 2.

Country	Year	The World Bank Data Set (1983 Features)				Airline Company Operation Data			
		Feature-1	Feature-2	Feature-3 to Feature-N-1	Feature-N	Seat Capacity		Load Factor (Output)	
		Population	GDP (USD)		Interest Rate	Eco Class	Bus Class	Eco Class	Bus Class
Albania	2019	2867000	1.53E+10		5.87%	..	..	..	..
Algeria	2019	43053000	1.70E+11		8.71%	..	..	..	..
Austria	2019	8865000	4.46E+11		NA	..	..	..	..
Azerbaijan	2019	10032000	4.80E+10		17.55%	..	..	..	..
Bahrain	2019	1641000	3.86E+10		NA	..	..	..	..
Bangladesh	2019	163046000	3.03E+11		4.88%	..	..	..	..
Belarus	2019	9467000	6.31E+10		2.30%	..	..	..	..
Belgium	2019	11483000	5.30E+11		NA	..	..	..	..

Table 2: Unified data set format

Some of the features in the data set that are related to the air travel demand are as follows:

- Population, total
- Unemployment, total (% of the total labor force) (modeled ILO estimate)
- GDP per capita (current US$)
- GNI per capita, purchasing power parity (PPP) (current international $)
- International tourism, number of arrivals
- Imports of goods and services (current US$)
- Exports of goods and services (current US$)
- International tourism, receipts (current US$)
- International tourism, expenditures (current US$)
- International tourism, expenditures for travel items (current US$)
- International tourism, receipts for travel items (current US$)
- International tourism, expenditures (% of total imports)

- International tourism, receipts (% of total exports)
- Air transport, passengers carried
- Air transport, registered carrier departures worldwide
- International tourism, expenditures for passenger transport items (current US$)
- Air transport, freight (million ton-km)
- International tourism, receipts for passenger transport items (current US$)
- International tourism, number of departures.

The World Bank databases do not provide all values of features for all countries and years. So, some features in the data set may not contain the values for some countries for a year. This means that there are some missing values for features in the country and year levels. For this reason, the density rate of each feature has been evaluated to prove the accuracy of the model when the features with low-density rates are also applied to the model as the inputs. Therefore, using a feature with a lower density rate may negatively affect the accuracy of the model. In the following sections, the effect of the data density rates on the model's accuracy is reviewed in detail. Table 3 shows the count of features in specific intervals of the density rates. Table 3 indicates that there is only one feature whose value is available for over 95% of the countries.

Density Rate Interval (%)	Number of Features
100 – 95	1
95 – 90	13
90 – 85	159
85 – 80	159
80 – 75	69
75 – 70	84
70 – 65	73
65 – 60	126
60 – 55	161
55 – 50	101
50 – 45	230
45 – 40	153
40 – 35	105
35 – 30	69
30 – 25	96
25 – 20	54
20 – 15	68
15 – 10	57
10 – 5	74
5 – 0	131
Total	1983

Table 3: Number of features between certain density rates

It is obvious that all the features in The World Bank data set are not correlated with the air travel demand. The features like “population” and “GDP per capita” may be relevant to the air travel demand, while a feature like “threatened bird species” may not. At this point in this study, all features are reviewed and assigned to a value between 0 and 10 for demonstrating the importance in the demand evaluation. This operation has been planned and used firstly in this study. If the feature has most relevance, the value is assigned as 10 to show its importance. Likewise, if the feature has no relation with the air travel demand, the value is set to 0. The number of features for each importance value is shown in Table 4.

Importance*	Number of Features
0	224
1	253
2	381
3	303
4	173
5	195
6	216
7	93
8	39
9	87
10	19
Total	1983

* (0 → Least Important, 10 → Most Important)

Table 4: Number of features by the importance

The input parameters in the data set are based on the features of the countries by year, while the output parameters are based on the annual international flight operation data of Turkish Airlines. As of February 2020, Turkish Airlines has the largest country coverage in the worldwide air transportation with operating in 126 countries [Turkish Airlines 2020]. The flight frequencies of each country are different. This means that different seat capacity is available for each country. For some countries, the frequency of the flight operation is too low, for example one time per week. In this study, it is experienced that applying these countries to the model increases the error rate. A flight operated three times per week with 300 seat capacity on the plane makes 1800 (3 x 300 x 2) seat capacity for one week and 93600 (1800 x 52) seat capacity for one year in the model. In this study, it is estimated that the countries (12 for 2019) with under 60000 annual seat capacity should be kept out of the scope to eliminate the higher error rates.

3.2 Methods

3.2.1 Machine Learning

Machine learning is a set of algorithmic methods applied to develop a model created for evaluating on a data set. The process aims to obtain some problematic results to be received by the other algorithmic methods. A model is developed by taking a data set as the inputs and producing the results as the outputs according to a learning type such as supervised, unsupervised, or reinforcement. In this study, the machine learning methods with the supervised learning are implemented to get the passenger load factor according to the applied features; therefore, the regression process is valid. In this section, specifically the methods of ANN, LR, GB, and RF are described because of their usages in the proposed model.

3.2.1.1 Artificial Neural Network

ANN is a modeling tool based on simulating human brain-behavior on different types of systems. In this approach, the basis of the human brain and the nervous system is modeled by the artificial elements called neurons. ANN is used as a powerful tool to investigate the behavior of an existing system to design new processes, to optimize the conditions, or to predict the behavior of a given system with the given parameters. About the prediction goal, the most popular ANN structures are feed-forward networks that are mostly trained from the input data using an error backpropagation algorithm [Rumelhart et al. 1986]. There are usually three layers in this type of ANN: input layer, hidden layer, and output layer. The input and output layers represent the independent and the dependent variables of the system respectively, while the hidden layer is used to perform the transformations. The neuron numbers in the hidden layer are not related to the input or output data and an absolute rule to determine a correct number has not been found yet. Too many hidden layer neurons provide successful training and memorizing, but unsuccessful testing and generalizing. A hidden or output unit in an ANN operates as in Eq. (1).

$$y_j = f\left(\sum_i w_{ji}\, x_i + b_j\right) \tag{1}$$

where $i = 1,2,\dots,n$ and $j = 1,2,\dots,m$. y_j is transformed output by the jth hidden or output node, $f(.)$ is an activation function, w_{ij} is a synaptic weight from node i to node j, x_i is an input node, and b_j is a bias at the node j. The learning algorithm of the ANN structure in this paper was chosen as the backpropagation training algorithm. The backpropagation algorithm looks for the minimum value of the error function in weight space.

The combination of weights, which minimizes the error function, is a solution to the learning problem.

3.2.1.2 Linear Regression

LR is a statistical method that provides new estimations by expressing the relationship between the independent and the dependent variables in a mathematical expression. The mathematical relationship between the variables is expected to be a linear equation. A basic linear regression model can be expressed by Eq. (2).

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \dots a_nx_n + k \tag{2}$$

In Eq. (2), y represents the result, $a_1, a_2, a_3, \ldots, a_n$ and the model coefficients, $x_1, x_2, x_3, \ldots, x_n$, are the dependent variables, and k is the constant noise value. The training process is completed when the coefficients are determined by providing the minimum error rate using various data.

3.2.1.3 Decision Tree

DT is one of the most popular methods used for the regression and classification processes. DT consists of the root, branch, and leaf nodes. In DT, the data reaches from the root to the leaf nodes recursively. While the root and the branch nodes provide a group according to their data properties, the leaf nodes contain the result values of the sample data. One of the most important disadvantages of DT is the possibility of the over-fitting. While it provides excellent training over the training set, it causes highly different results (and errors) when the data is changed little.

Therefore, although the training phase can be completed with over-fitting by a small error rate, the results of the verification phase can be much different. Instead of constituting a single DT for regression or classification, including many DT and combining their output can give more accurate results. This method is generally called an ensemble as seen in Figure 1, and in this method, different techniques can be used in the phases of the constituting subtrees, training, and combining the outputs. Some of these techniques are the boosting and bagging. The basic purpose of this ensemble method is to create a model whose learning ability is high by combining multiple low learning ability trees. For this purpose, GB and RF algorithms are used in this study. Although these two algorithms are similar in some points, they have some differences. In the GB, the subtrees are dependent on each other as applied in the boosting method. When adding a new tree to the structure, there will be some correction over the previous tree structure. Because the trees are constituted sequential, there is no parallel working. At this point, the modeling process can be longer than the RF method. In the RF, small trees are constituted by making random sampling over the public data set. These little trees are trained individually (independent from the others) performing the bagging method. So, there can be a similar working in constituting subtrees and training them. Despite each subtree's low learning ability, the combined tree will have a high learning ability. Also, the over-fitting problem is prevented. But, in general, the performance of the GB is higher than that of the RF method.

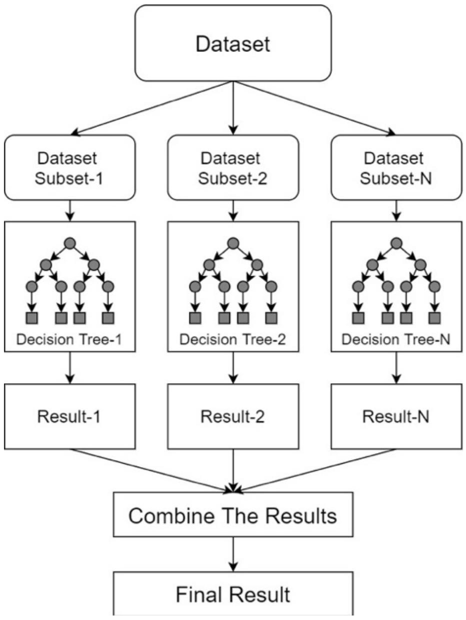

Figure 1: Ensemble method

In this study, an application was developed with Java. The XGBoost [Chen and Guestrin 2016] library was used for the GB method, and WEKA [Weka 2020] Java API is used for the ANN, LR, and RF methods. XGBoost is an open library that contains many supervised machine learning methods and can be executed on Linux, Windows, and macOS platforms. The library supports various programming languages such as Python, Ridava, Scala, Java, and C++. Weka is an open-source application that was developed by Java and consists of many machine learning methods. Different ways can be applied for the data preparation, classification, regression, and grouping over this application. Weka was designed in Waikato University (New Zealand) and is free to the users. It has a graphical user interface as well as a Java API that provides to build the applications.

3.2.2 Software Model

In the data set, each feature has some missing values of which the ratio determines the sparsity and density of the current feature. A data set is expected to have low missing values of the features to get better results. In this study, the density rate of each feature in the data set is different for any relevant period and country. The missing values of some features are over 95%, which makes the density rates of these features lower than 5%. During the computations at first, the density rate of each feature was calculated. Then, the features with a density rate lower than a definite threshold value were neglected from the data set and applied to the software model to obtain the results. The threshold value was 5% at the first iteration of the simulation and then at each step it was increased by 5%. The generated data set was used, and the results were recorded (up to 95%). A similar iteration was also applied for the feature importance factor in

Table 4. At the first step of this iteration, the features with an importance value lower than 0 were neglected from the data set and then they were used in the software model. At each step, the feature importance level was increased by 1, then the generated data set was applied to the software model, and the results were recorded. In this way, the results were registered for all combinations of the feature density rate and the feature importance value. At the end, the feature set with the lowest error rates were evaluated.

4 Results and Discussions

4.1 Experimental Study

The developed model uses machine learning methods and produces the average load factor when the definite seat capacity is provided for the country's flights. For this purpose, the airline travel data is required to be applied as the output parameters of the training phase. The load factor data of the international flights operated by the Turkish Airlines was used in this study. The data was transformed into the origin-destination country level, and for each pair, the total seat capacity and the total count of the booked seat were evaluated on an annual basis. Since the input parameters consist of the data that contains the features of each country from year to year, by generating the output parameters every year, the input and output data are given in pairs of the country and year levels. The data provided by the Turkish Airlines is kept private since it contains sensitive commercial information. In this study, the proposed model covers the data belonging to 7 years (2013-2019). Basically, on average, 2% of the relevant countries was selected for verification while the rest of the countries were applied for training operation in a specific period. After then, the countries that were chosen for the verifications were used for the next period, and the evaluated result was compared with the real value. The absolute percentage error value of these two results was recorded to analyze the error rates with different parameters. The proposed model covers additional input parameters of the total seat capacity, country distance, and business class seat capacity for the relevant country near the World Bank data set values. The country distance is static, but the other two parameters are dynamic. Therefore, the results can be diversified according to the dynamic parameters.

It should be noticed here that the estimation for 2020 was not performed, since many airline companies (including Turkish Airlines) suspended the air travel operations for several months in 2020 when the COVID-19 pandemic emerged. As mentioned by [Gallego and Font 2020], air travel demand covered the flight searches between the months of May and September in 2020 decreased by almost 50% in Asia and 30% in Europe and the America. Thus, the passenger statistics for the year 2020 obviously yield exceptional results when considering the main goal of this study as forecasting the air travel demand.

4.2 Performance Evaluation

An output that gives some results for the specific rules from an entire data of a model has some error percentages. This error rate is the main criterion for determining the performance of the model. There are various metrics to compute this error percentage. In this study, the Mean Absolute Percentage Error (MAPE) is used as in Eq. (3).

$$MAPE = 100 * \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \quad (3)$$

where n represents the sample number, y_i is the result value of sample i, e_i is the difference between the estimation result of sample i and the value of y_i. Equation (3) gives the deviation of the produced result from the demanded result as a percentage rate.

In this study, once the features with the lowest error rates were determined, the estimation was applied again for the years of 2014, 2015, 2016, 2017, 2018, and 2019 by using ANN, LR, GB, and RF. At each estimation process, the training data set was generated from the features of the previous year. This way makes it possible to estimate the passenger load factor (for a country) of the next year using the features of the current year. Each estimation process is repeated 500 times to make the results more robust and homogeneous. In each estimation process for a period, 98% of the countries were randomly chosen for the training, and the remaining 2% were selected for the verification. In a verification step, the real values of the passenger load factors were compared with the values estimated by the software model, and then the mean error rates were recorded. The general results for all methods were listed in Table 5 and Figure 2 respectively.

Train Period	Estimation Period	ANN (%)	LR (%)	GB (%)	RF (%)
2013	2014	6.318	7.480	5.743	5.412
2014	2015	7.597	7.412	6.540	6.646
2015	2016	10.511	10.206	8.900	9.067
2016	2017	9.108	12.208	7.980	7.894
2017	2018	8.148	11.516	6.134	5.761
2018	2019	7.283	8.890	5.148	5.798
Average		8.161	9.619	6.741	6.763

Table 5: General MAPE results

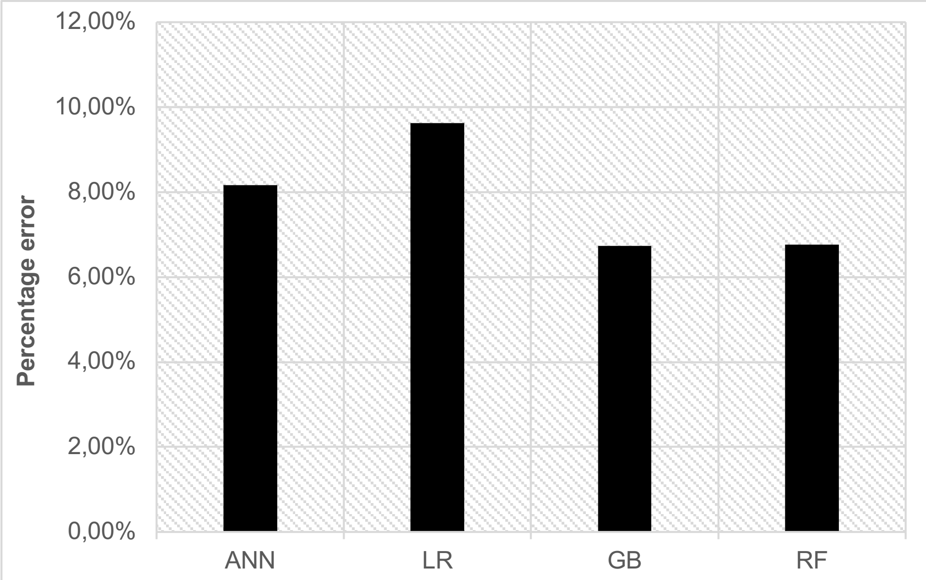

Figure 2: The comparison of average MAPE results

Totally 132000 (20 x 11 x 6 x 100) regression processes were performed in the simulation for all combinations of the feature density rates, the feature importance level, and the estimation periods (2014-2019). Each simulation was run 100 times executing the GB to determine the features with the best results. Different data set combinations were generated and applied to the model at each regression process, and the results were recorded. The simulation results show that the features with a density rate of over 45% and the importance level upper than five should be used to get the lowest error rates. The average error rates over the feature importance levels and the threshold values of the feature data densities are shown in Figure 3 and Figure 4 respectively. In this way, a total number of 318 features that yield the lowest error rates were determined.

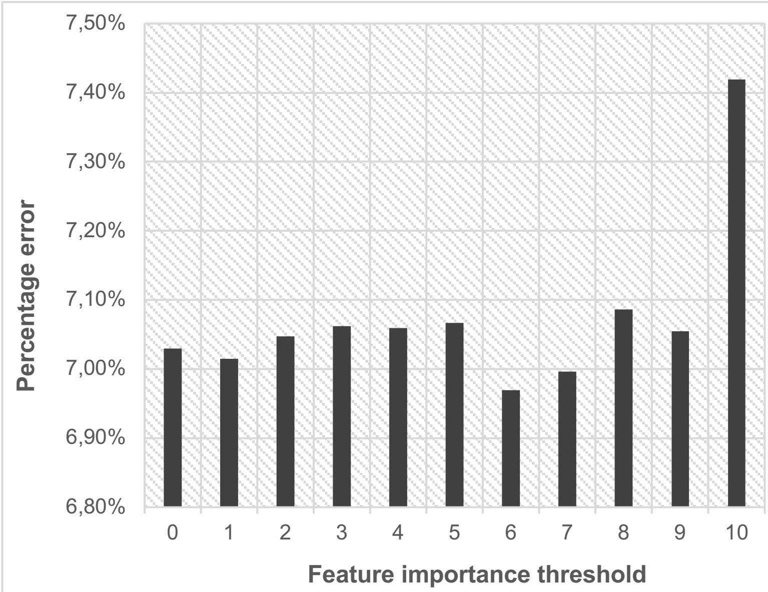

Figure 3: The percentage error rates according to the feature importance levels

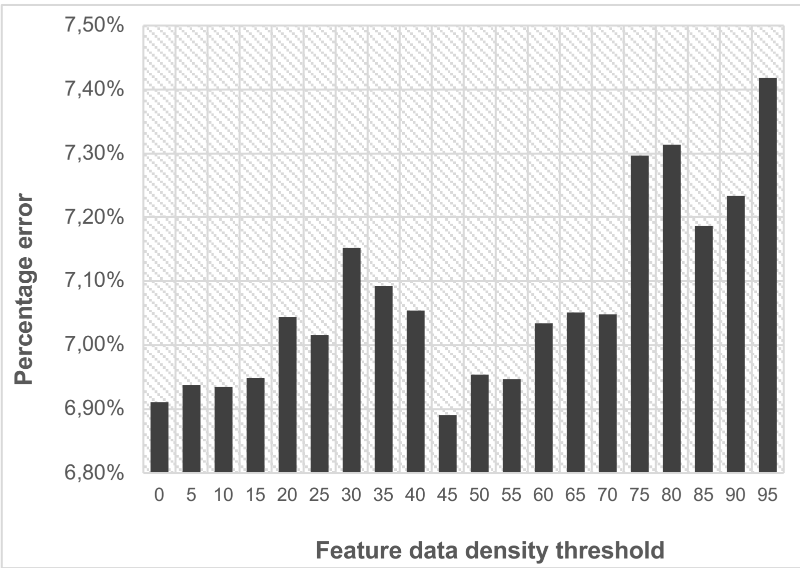

Figure 4: The percentage error rates according to the feature data density threshold

5 Conclusions and Future Work

In this study, a model was developed for forecasting the travel demand corresponding to the flights performing with the predetermined capacity between any two countries. Once the features were determined considering The World Bank data, then the estimation process was applied using various methods to compute the results. The computational evaluations were implemented with including the regression operation as 500 times for each period (each time a different training-validation set was built). The ANN, LR, GB, and RF methods gave the MAPE values of 8.161%, 9.619%, 6.741%, and 6.763%, respectively for the estimation periods of 2014-2019. Consequently, the GB method produced the best results. The period is very important in air travel industry as representing the data analysis just before COVID-19 pandemic. Since many airline companies have experienced a decline in the air travel operation in 2020 due to COVID-19 pandemic.

It may be a future work to study the RF and ANN methods more intensely to get better results. Another future work can be various data analysis on the statistical information of the passengers before and during COVID-19 to present the worldwide effects of this pandemic on any transportation type. Also, the study can be improved with considering different airports and making the estimation for two predetermined airports instead of countries.

The season variables have an important role on affecting the airline passenger demand for specific regions. These season variables were not taken into consideration in the present model, because they were not available for all regions. In another future work, the season variables of the regions can be determined and applied to the current model to get better results. Additionally, the ticket fee was not handled as a parameter in this model. However, this parameter is one of the important factors having effect on the passenger requests. The variation of the passenger requests can be observed through this parameter.

Acknowledgements

We thank the Turkish Airlines for providing the data that consist of the annual international passenger statistics.

References

[Abed et al. 2001] Abed, S. Y., Ba-Fail, A. O., Jasimuddin, S. M.: “An Econometric Analysis of International Air Travel Demand in Saudi Arabia”, Journal of Air Transport Management, 7(3), (2001), 143–148, https://doi.org/10.1016/S0969-6997(00)00043-0.

[Alekseev and Seixas 2009] Alekseev, K. P. G. and Seixas, J. M.: “A Multivariate Neural Forecasting Modeling for Air Transport – Preprocessed by Decomposition: A Brazilian Application”, Journal of Air Transport Management, 15(5), (2009), 212–216, https://doi.org/10.1016/j.jairtraman.2008.08.008.

[Ashiabor et al. 2007] Ashiabor, S., Baik, H., Trani, A.: "Logit Models for Forecasting Nationwide Intercity Travel Demand in the United States", Transportation Research Record: Journal of the Transportation Research Board, 2007(1), (2007), 1–12, https://doi.org/10.3141/2007-01.

[Ba-Fail et al. 2000] Ba-Fail, A. O., Abed, S. Y., Jasimuddin, S. M.: "The Determinants of Domestic Air Travel Demand in the Kingdom of Saudi Arabia", Journal of Air Transportation World Wide, 5(2), (2000), 72–86.

[Blinova 2007] Blinova, T. O.: "Analysis of Possibility of Using Neural Network to Forecast Passenger Traffic Flows in Russia", Aviation, 11(1), (2007), 28–34, https://doi.org/10.1080/16487788.2007.9635952.

[Boonekamp et al. 2018] Boonekamp, T., Zuidberg, J., Burghouwt, G.: "Determinants of Air Travel Demand: The Role of Low-Cost Carriers, Ethnic Links and Aviation-Dependent Employment", Transportation Research Part A: Policy and Practice, 112, (2018), 18–28, https://doi.org/10.1016/j.tra.2018.01.004.

[Brons et al. 2002] Brons, M., Pels, E., Nijkamp, P., Rietveld, P.: "Price Elasticities of Demand for Passenger Air Travel: A Meta-Analysis", Journal of Air Transport Management, 8(3), (2002), 165–175, https://doi.org/10.1016/S0969-6997(01)00050-3.

[Carson et al. 2011] Carson, R. T., Cenesizoglu, T., Parker, R.: "Forecasting (Aggregate) Demand for US Commercial Air Travel", International Journal of Forecasting, 27(3), (2011), 923–941, https://doi.org/10.1016/j.ijforecast.2010.02.010.

[Chang et al. 2019] Chang, Y. -C., Lee, W. -H., Wu, C. -H.: "Airline New Route Selection Using Compromise Programming - The Case of Taiwan-Europe", Journal of Air Transport Management, 76(C), (2019), 10–20, https://doi.org/10.1016/j.jairtraman.2019.01.012.

[Chang et al. 2017] Chang, Y. -C., Woon, H. K., Yen, J. -R., Hsu, C. -J.: "A Destination Selection Model for Long Haul Routes – The Case of Taiwan-US", Journal of Air Transport Management, 64(A), (2017), 60–67, https://doi.org/10.1016/j.jairtraman.2017.06.026.

[Chen and Guestrin 2016] Chen, T. and Guestrin, C.: "Xgboost: A Scalable Tree Boosting System", In Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, (2016), 785–794, https://doi.org/10.1145/2939672.2939785.

[Dantas et al. 2017] Dantas, T. M., Oliveira, F. L. C., Repolho, H. M. V.: "Air Transportation Demand Forecast Through Bagging Holt Winters Methods", Journal of Air Transport Management, 59, (2017), 116–123, https://doi.org/10.1016/j.jairtraman.2016.12.006.

[Deveci et al. 2017] Deveci, M., Cetin Demirel, N., Ahmetoğlu, E.: "Airline New Route Selection Based on Interval Type-2 Fuzzy MCDM: A Case Study of New Route Between Turkey- North American Region Destinations", Journal of Air Transport Management, 59, (2017), 83–99, https://doi.org/10.1016/j.jairtraman.2016.11.013.

[Doganis 2010] Doganis, R.: "Flying Off Course: Airline Economics and Marketing", Routledge.

[Erjongmanee and Kongsamutr 2018] Erjongmanee, S. and Kongsamutr, N.: "Air Passenger Estimation Using Gravity Model and Learning Approaches: Case Study of Thailand", In Proc. 5th Int. Conf. on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), (2018), 36–41, 10.1109/ICAICTA.2018.8541335.

[Gallego and Font 2020] Gallego, I. and Font, X.: "Changes in Air Passenger Demand as a Result of the COVID-19 Crisis: Using Big Data to Inform Tourism Policy", Journal of Sustainable Tourism, (2020), https://doi.org/10.1080/09669582.2020.1773476.

[Ghomi and Forghani 2016] Ghomi, S. M. T. F. and Forghani, K.: "Airline Passenger Forecasting Using Neural Networks and Box-Jenkins", In Proc. 12th Int. Conf. on Industrial Engineering (ICIE), (2016), 10–13, 10.1109/INDUSENG.2016.7519342.

[Jungmittag 2016] Jungmittag, A.: "Combination of Forecasts Across Estimation Windows: An Application to Air Travel Demand", Journal of Forecasting, 35(4), (2016), 373-380, https://doi.org/10.1002/for.2400.

[Koç and Arslan 2018] Koç, İ. and Arslan, E.: "Demand Forecasting for Domestic Air Transportation in Turkey Using Artificial Neural Networks", In Proc. 6th Int. Conf. on Control Engineering & Information Technology (CEIT), (2018), 1–6, 10.1109/CEIT.2018.8751869.

[Kopsch 2012] Kopsch, F.: "A Demand Model for Domestic Air Travel in Sweden", Journal of Air Transport Management, 20, (2012), 46–48, https://doi.org/10.1016/j.jairtraman.2011.11.006.

[Laik et al. 2014] Laik, M. N., Choy, M., Sen, P.: "Predicting Airline Passenger Load: A Case Study", In Proc. IEEE 16th Conference on Business Informatics, (2014), 33–38, 10.1109/CBI.2014.39.

[Li et al. 2014] Li, T., Baik, H., Spencer, T.: "An Optimization Model to Estimate the Air Travel Demand for the United States", In Proc. 2014 Integrated Communications, Navigation and Surveillance Conference (ICNS), (2014), E2-1-E2-11, 10.1109/ICNSurv.2014.6819985.

[Plakandaras et al. 2019] Plakandaras, V., Papadimitriou, T., Gogas, P.: "Forecasting Transportation Demand for the U.S. Market", Transportation Research Part A: Policy and Practice, 126, (2019), 195–214, https://doi.org/10.1016/j.tra.2019.06.008.

[Rumelhart et al. 1986] Rumelhart, D. E., Hinton, G. E., Williams, R. J.: "Learning Representations by Back-Propagating Errors", Nature, 323, (1986), 533–536.

[Schubert et al. 2020] Schubert, I., Sohre, A., Ströbel, M.: "The Role of Lifestyle, Quality of Life Preferences and Geographical Context in Personal Air Travel", Journal of Sustainable Tourism, 28(10), (2020), 1519-1550, https://doi.org/10.1080/09669582.2020.1745214.

[Shaw and Thomas 2006] Shaw, S. and Thomas, C.: "Discussion Note: Social and Cultural Dimensions of Air Travel Demand: Hyper-Mobility in the UK?", Journal of Sustainable Tourism, 14(2), (2006), 209-215, 10.1080/09669580608669053.

[Sivrikaya and Tunc 2013] Sivrikaya, O. and Tunc, E.: "Demand Forecasting for Domestic Air Transportation in Turkey", The Open Transportation Journal, 7(1), (2013), 20–26, 10.2174/1874447820130508001.

[Smyth and Pearce 2008] Smyth, M. and Pearce, B.: "IATA Economics Briefing No 9: Air Travel Demand"; [online] https://www.iata.org/en/iata-repository/publications/economic-reports/air-travel-demand/, (Accessed 20 January 2021).

[Srisaeng 2015] Srisaeng, P.: "Utilizing Advanced Modelling Approaches for Forecasting Air Travel Demand: A Case Study of Australia's Domestic Low Cost Carriers"; RMIT University, Doctoral Dissertation, RMIT University Theses and Dissertations Archive (2015), https://researchrepository.rmit.edu.au/discovery/fulldisplay/alma9921861869601341/61RMIT_INST:ResearchRepository?tags=scholar.

[Srisaeng et al. 2015] Srisaeng, P., Baxter, G., Richardson, S., Wild, G.: "A Forecasting Tool for Predicting Australia's Domestic Airline Passenger Demand Using a Genetic Algorithm", Journal of Aerospace Technology and Management, 7(4), (2015), 476–489.

[Sulistyowati et al. 2018] Sulistyowati, R., Suhartono, Kuswanto, H., Setiawan, Astuti, E. T.: "Hybrid Forecasting Model to Predict Air Passenger and Cargo in Indonesia", In Proc. Int. Conf. on Information and Communications Technology (ICOIACT), (2018), 442–447, 10.1109/ICOIACT.2018.8350816.

[The World Bank 2020] The World Bank: "DataBank"; [online] https://databank.worldbank.org, (Accessed 5 December 2020).

[Totamane et al. 2014] Totamane, R., Dasgupta, A., Rao, S.: "Air Cargo Demand Modeling and Prediction", IEEE Systems Journal, 8(1), (2014), 52–62, 10.1109/JSYST.2012.2218511.

[Tsekeris 2009] Tsekeris, T.: "Dynamic Analysis of Air Travel Demand in Competitive Island Markets", Journal of Air Transport Management, 15(6), (2009), 267–273, https://doi.org/10.1016/j.jairtraman.2008.11.008.

[Turkish Airlines 2020] Turkish Airlines: "Financial and Operational Data", [online] http://investor.turkishairlines.com/en, (Accessed 5 December 2020).

[Wang et al. 2011] Wang, Y., Wang, J., Dang, Y., Wang, Z.: "A Prediction Model of China's Air Passenger Demand", In Proc. 2011 IEEE Int. Conf. on Grey Systems and Intelligent Services, (2011), 347–350, 10.1109/GSIS.2011.6044120.

[Weka 2020] Weka: "Weka 3: Machine Learning Software in Java", [online] https://www.cs.waikato.ac.nz/~ml/weka/, (Accessed 5 December 2020).

[Yang et al. 2019] Yang, J., Shebalov, S., Klabjan, D.: "Semi-Supervised Learning for Discrete Choice Models", IEEE Transactions on Intelligent Transportation Systems, 20(11), (2019), 4145–4159, 10.1109/TITS.2018.2882847.

[Yang et al. 2017] Yang, T., Liu, M., Ye, Y.: "Empirical Estimation of Civil Aviation Demand Elasticity: The Case of China Mainland", In Proc. 2017 4th Int. Conf. on Industrial Economics System and Industrial Security Engineering (IEIS), (2017), 1–6. 10.1109/IEIS.2017.8078649.