Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi
Yıl: 2022 Cilt: 31 No: 1 Sayfa: 133-144
https://dergipark.org.tr/tr/pub/cusosbil

ASSESMENT OF TICKET PRICE FORECASTING IN TURKEY

Ayşe Tuğba DOSDOĞRU[1], Aslı BORU İPEK[2],
Mustafa GÖÇKEN[3], Mehmet ÖZÇALICI[4]

Article Info

Research Article
DOI: 10.35379/cusosbil.1030398

Article History:
Received 30.11.2021
Revised 28.03.2022
Accepted 14.04.2022

Keywords:
K-means Algorithm,
Linear Regression,
Support Vector Regression,
Genetic Algorithm Based Artificial Neural Network,
Ticket Price Forecasting.

ABSTRACT

Fast, reliable, and comfortable transportation of people increase the level of livability in cities. It also influences people's quality of life. Therefore, research is needed to improve transportation services. Various models are developed to analyze the transportation services but each of them has its own advantages and disadvantages. Today, companies collect large amounts of data to improve their service quality. To survive in a competition environment, they must use the collected data to create value for their customers and employees. Many factors affect the transportation services. Therefore, it is difficult to solve the problems in transportation services using classical methods. The main goal of our study is to determine the bus ticket price accurately. In this study, the k-means algorithm, which is popular because of its simplicity and versatility, is firstly used to discover information that is more meaningful. Then the bus ticket price, which is one of the most important elements of passenger transportation, is forecasted using six different forecasting models including linear regression, support vector regression, regression tree, gaussian process regression, genetic algorithm based artificial neural network, and an ensemble model. The results of this study showed that proposed forecasting models can meet expectations in dynamic environmental conditions.

TÜRKİYE'DE BİLET FİYATI TAHMİNİNİN DEĞERLENDİRİLMESİ

Makale Bilgisi

Araştırma Makalesi
DOI: 10.35379/cusosbil.1030398

Makale Geçmişi:
Geliş 30.11.2021
Düzeltme 28.03.2022
Kabul 14.04.2022

Anahtar Kelimeler:
K-ortalamalar algoritması,
Doğrusal regresyon,
Destek vektör regresyonu,
Genetik algoritma tabanlı yapay sinir ağı,
Bilet fiyatı tahmini.

ÖZ

İnsanların hızlı, güvenilir ve konforlu olarak taşınması şehirlerde yaşanabilirlik düzeyini de arttırmaktadır. Ayrıca insanların yaşam kalitesine de etki etmektedir. Bu nedenle ulaşım hizmetlerinin iyileştirilmesi için çalışmaların yapılması gerekmektedir. Ulaşım hizmetlerini analiz etmek için çeşitli modeller geliştirilmiştir ancak her birinin kendine özgü avantajları ve dezavantajları vardır. Günümüzde işletmeler hizmet kalitesini artırmak için yüklü miktarda veri toplamaktadır. Artan rekabet ortamında ayakta durabilmek için topladıkları verileri, müşterilerine ve çalışanlarına değer yaratacak şekilde kullanmak zorundadırlar. Ulaşım hizmetlerini etkileyen birçok faktör vardır. Bu nedenle ulaşım hizmetlerindeki problemleri klasik yöntemlerle çözmek zordur. Çalışmamızın temel amacı otobüs bileti fiyatını doğru belirlemektir. Çalışmamızda ilk olarak basitliği ve çok yönlülüğü nedeniyle popüler olan k-ortalamalar algoritması, daha anlamlı bilgiler keşfetmek için kullanılır. Daha sonra yolcu taşımacılığının en önemli unsurlarından biri olan fiyat, doğrusal regresyon, destek vektör regresyonu, regresyon ağacı, gauss süreç regresyonu, genetik algoritma tabanlı yapay sinir ağı ve topluluk modeli kullanarak tahmin edilmiştir. Bu çalışmanın sonuçları tasarlanan tahmin modellerinin dinamik çevre koşullarındaki beklentileri karşılayabildiğini göstermiştir.

Yazarlar çalışmanın etik kurallara bağlı olarak hazırlandığını taahhüt eder.

[1] Doç. Dr., Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi, Endüstri Mühendisliği Bölümü, adosdogru@atu.edu.tr, ORCID: 0000-0002-1548-5237
[2] Dr. Öğr. Üyesi, Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi, Endüstri Mühendisliği Bölümü, aboru@atu.edu.tr, ORCID: 0000-0001-6403-5307
[3] Doç. Dr., Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi, Endüstri Mühendisliği Bölümü, mgocken@atu.edu.tr, ORCID: 0000-0002-1256-2305
[4] Doç. Dr, Kilis 7 Aralık Üniversitesi, Uluslararası Ticaret ve Lojistik bölümü, mozcalici@kilis.edu.tr, ORCID: 0000-0003-0384-6872
Alıntılamak için/Cite as: Dosdoğru, A. T., Boru İpek, A., Göçken, M., Özçalıcı, M. (2022), Assesment Of Tıcket Prıce Forecastıng In Turkey, Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 31 (1), 133-144.

INTRODUCTION

With the establishment of the General Directorate of Highways in March 1950, a new era has started for highway history in Turkey. A total of 61,500 kilometers of the road network, including the current state and provincial roads, was formed after this period. However, it is difficult to compare these roads with today's road standards. In the 1970s, the existing network became insufficient, and the network was expanded to improve the standards. By the 1980s, "1983-1993 Transportation Master Plan" was formed to eliminate the drawbacks of the excessive use of roads and to resolve problems related to the increasing number of vehicles. At the beginning of 2000, double way or express way, which is defined as a low cost, has been adopted. Details about the highway history in Turkey can be found in Keçeci (2006). In Turkey, both the passenger and the freight transportation are generally carried out on highway based transportation. For many years, many problems have been encountered due to the highway-weighted transportation policy in Turkey. It is important to use alternative transportation networks in both freight and passenger transportation in terms of proper use of resources in Turkey. In recent years, important projects have been put into practice regarding railways and highways and some of them have been carried out (Çetin, Barış, & Saroğlu, 2011). Details about the transportation system in Turkey can be found in Kapluhan (2014).

Transportation services contribute greatly to the economy. In the process of economic development, transportation activities constitute a driving force. Thus, the decisions, policies and plans to be implemented for transportation directly influence the economic development (Kapluhan, 2014).

The important characteristics of transportation are the starting points, arrival points, the possibility of providing direct transportation, providing flexibility in carrying capacity and route selection. In addition, road transportation is an indispensable element for transportation, because it is a complementary type of railway, seaway and airway transports (Kögmen, 2014). Bus management is of high importance to the economic, social, and cultural impacts of the society because of collective human transportation. The increase in the reliability of the buses due to the developments in today's technology, as well as the shorter transportation time and the increase in the quality of transportation services gave an impetus to the development process of bus management (Kara, 1999). Knowing the competitive behavior and determining the transportation costs are the main functions affecting bus management, but not enough for success. The success of these two variables depends on the implementation. The transport service in bus companies is directly proportional to the size of the geographical area it performs. The organization of the service hours and routes will fundamentally affect the service (Alnıaçık & Özbek, 2009). In addition, bus companies provide many alternatives in the routes and there is no obvious difference between the services they provide. Furthermore, the continuous change and demand of passengers enforce bus companies to provide the best service as speed, comfort, etc.

In a highly competitive environment, the main function of the bus companies is to predict and organize the future. In addition, the internet has become one of the most important channels for businesses (Saran, 2005). Especially, online booking and online ticket sales are two of the most important parts of marketing activities on the internet. For example, Gül and Boz (2012) presented the effect of online booking and online ticket sales on the sales of the bus companies. In the study, the general characteristics of the bus companies that carry out intercity long distance passenger transportation are determined. The results of the study demonstrated that the most important aim of the companies is to make advertising and publicity using Internet site. This aim is followed by communicating with customers, providing information to the customers and creating database about customers. These results show that the bus companies participating in the survey generally use their websites as a means of communication. Many bus companies developed websites related to transportation services. Websites can perform many functions such as reducing the distance between companies and customers, delivering the company's products and services more quickly to customers, increasing the recognition of companies and ensuring customer satisfaction. Hence, companies give importance to the design and content of their website and to keep their websites constantly updated (Karakan, Türkmen, Giritlioğlu, & Kılıç, 2016). Karakan et al. (2016) analyzed the content of the intercity bus company website operating in the transportation sector in Turkey. They developed a questionnaire consisting of forty questions in order. The questionnaire considered the four main activities : online transactions, travel information center, corporate information and the services provided.

In recent years, lots of attentions have also been devoted to analyze and forecast the transportation system (e.g. Abdella, Zaki, Shuaib, & Khan, 2021; Stavinova, Chunaev, & Bochenina, 2021; La, & Heiets, 2021; Truong, 2021; Zhao et al., 2022; Chiu, Chen, & Lee, 2022). Wohlfarth et al. (2011) forecasted the travel price. In the study, the k-means clustering algorithm was employed. In addition, classification tree and random forest were applied to the training dataset. Then, price evolution modeling and price decrease event prediction were

applied to the testing dataset. Tsai, Mulley, and Clifton (2013) employed a univariate model (Autoregressive Integrated Moving Average (ARIMA)) and a multivariate direct demand model (partial adjustment model) to forecast public transport demand. Sevuktekin et al. (2014) used the statistical models including ARIMA, single exponential smoothing and linear trend models to analyze the development in transportation sector in Turkey. Li and Li (2018) employed the random forest algorithm, time series-random forest algorithm, and autoregressive moving average-random forest algorithm to forecast ticket prices. The results demonstrated that the combinations of algorithms has some advantages.

Lin et al. (2013) analyzed the potential effect of bus travel time and selected the candidate variables. Based on historical global positioning system data and automatic fare collection system data, the sub-ANN model was created for each data cluster, and then integrated as a hierarchical ANN model. Li and Chen (2014) utilized the k-means clustering to categorize the traffic characteristics. In the study, the back-propagation ANN was employed. In addition, the classification and regression tree were used to provide an effective solution for travel time prediction. Yu et al. (2016) proposed the ANN method based on urban land use and bus accessibility of each zone to forecast the bus passenger trip flow. In the study, the inputs of ANN are each traffic zone land use, bus accessibility, area, and distance to other zones. The output node is the bus passenger flow from one traffic zone to another traffic zone. In the light of previous studies, we determined that various models are implemented to analyze the transportation system but each of which has its own merits and limitations.

In this study, forecasting models which aim at forecasting ticket prices for intercity passenger buses are developed. The main purpose of our study is to determine the intercity bus ticket price by using linear regression, support vector regression, regression tree, gaussian process regression, genetic algorithm based artificial neural network (GA-ANN), and ensemble model. There is no single method available in the literature to forecast the price of a bus ticket. The novelty of this paper is to improve the accuracy of forecasting and to evaluate the impact of using clustering methods on ticket price forecasting of six different forecasting methods. To the best of our knowledge, no systematic paper is devoted to studying of six clustering based methods simultaneously and comparing the performance of these methods in terms of ticket price forecasting.

Bus companies use different strategies to survive in an increasingly competitive environment. In order to compete, companies must first determine the correct transportation strategy. One of the biggest problems encountered in bus management is the determination of the price of bus tickets. Price is an important part of customer satisfaction. While determining the price, customer satisfaction should be ensured and the profit should be maximized. There are many factors affecting the price of bus tickets. Therefore, it is difficult to determine the bus ticket price by classical methods such as moving average method. The disadvantage of the classical method is that it is unable to identify complex nonlinear interactions that are implicit in ticket price forecasting. The main objective of our study is to determine the bus ticket prices accurately. Firstly, the dataset is collected. Then, clusters with homogeneous properties are formed. Here, the k-means algorithm, which is popular because of its simplicity and versatility, is used to divide the general dataset into homogeneous subsets. Each cluster is divided into training and test sets. Finally, six different forecasting models are used to determine the price of bus tickets using the distance (km) between cities as input.

PROPOSED METHODOLOGY

In this study, we firstly collect dataset and present the descriptive features of dataset. Then, clusters with homogeneous features are created and each cluster is divided as training and testing sets. After that, six different forecasting models are applied. Note that ensemble model is the arithmetic average of all five different forecasting results. Finally, the performance of each cluster is evaluated with graphs and statistical performance measurements.

Dataset Description

An intercity distance and bus ticket price dataset are needed to forecast the bus ticket price between two cities. Distance will be used as the input and the ticket price will be used as the output. Intercity distance matrix which is obtained from General Directorate of Highways (www.kgm.gov.tr.) includes the distances between each city in Turkey.. On the other hand, bus ticket prices between cities are not accessible from a single source. However, there are various online firms where passengers can buy bus tickets. The basic business logic of these firms can be summarized as follows. Users are asked to determine the two cities they want to travel from and the dates of their journeys. Later, these sites list the buses that are available within the respective dates. In this study, required ticket price dataset was scanned with web scraping technique, which can be defined as extracting data

from websites programmatically. One of the sites operating in this way was randomly selected. The name of the site will not be given due to privacy concerns.

Between June 25, 2018 and June 28, 2018, prices were scanned for buses that would take place one day later. There are 81 provinces (cities) in Turkey. The total number of all possible routes between these cities is calculated as 3240 with the help of the combinations formula $\frac{81!}{(81-2)!\times 2!} = 3240$. Therefore, 3240 web inquiries were made on the site and price information was recorded. However, there exist only 1422 trips out of 3240 possible city pairs. There were no bus services on the remaining 1818 routes between the dates examined.

Collected dataset has 1422 rows and four columns. However, there can be outliers in the dataset. In this study, samples that return error beyond three standard deviations from the least square regression line are defined as outlier and removed from the dataset. Fourteen of the observations are detected as outlier and extracted from the dataset. So there remained 1408 (=1422-14) observations in the dataset.

Table 1. Descriptive Statistics of Distance and Price

	Distance	**Number of Trips**	**Minimum Price (krş)**	**Maximum Price (krş)**	**Mode Price (krş)**
Minimum	44	1	1300	1300	1300
Maximum	1752	89	19000	20000	19000
Mean	638.29	5.3	8520.7	9222.59	8826.23
Standard Deviation	330.64	8.26	3470.16	3450.74	3438.67
Median	604	2	8500	9000	9000
Kurtosis	2.83	29	2.74	2.59	2.69
Skewness	0.52	4.34	0.23	0.06	0.14

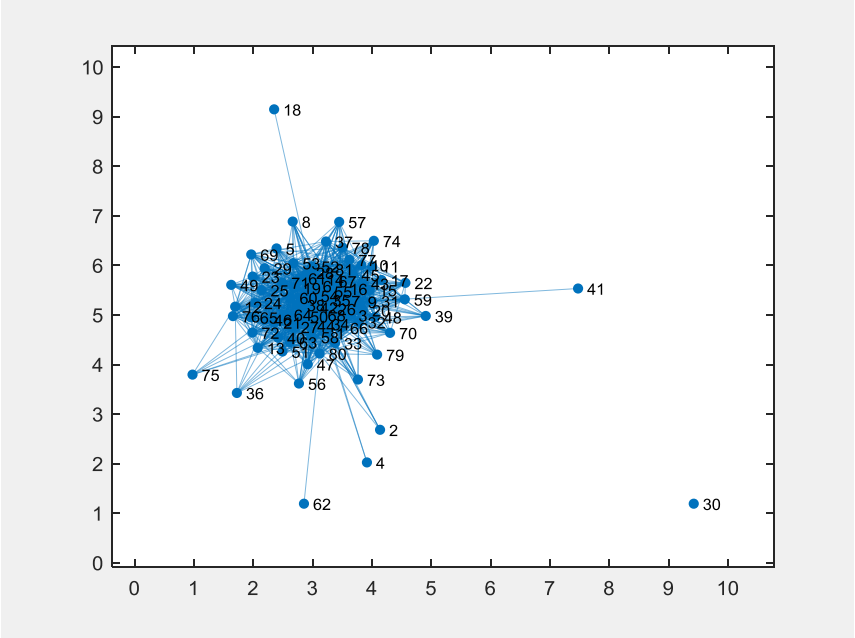

Figure 1. Network Plot of Trips

Minimum and maximum prices may vary for certain reasons. For example, some companies may charge lower prices for certain trips than other companies to gain competitive advantage. Similarly, due to the low demand some trips may be more expensive than they are supposed to be. Descriptive statistics of distance and price are given in Table 1 where distance, number of trips, minimum price, maximum price, and mode price are given. Note that currency unit is Kuruş (krş).

The arrangement of the bus network, which is the connection between the individual cities, is shown in Figure 1. Numbers represents the vehicle registration plates of Turkey. The figure was arranged in such a way

that the cities with more bus trips would be in the center. The cities with fewer bus services are located around the periphery of the figure.

Clustering

Dataset includes bus trips with a wide range of features. Some of the trips are in short distance while others are long distance. Using all of the distance and price information in a single model may not produce successful forecasting results. For this reason, before creating forecasting models, dataset is clustered into homogenous subsets. For each cluster, randomly selected 80 observations are used for out-of-sample performance measurement purposes, while other observations are used for model construction purposes. Because each cluster has a different number of observations, the size of the training and test set varies between clusters.

The aim is to use the distances between the cities as an input variable to predict the maximum bus fare. So there are two variables that will be used in the analysis: (I) distance between cities and (II) maximum price of the bus trip. The distance of the trip and the price (from now on price will be used instead of maximum price) is not homogenous. In other words, behind the pricing strategies lie different dynamics other than the length of the distance between the cities. For this reason, clustering trips in homogenous subsets will help increase the accuracy of the forecasted prices.

K-means clustering algorithm is popular due to its simplicity and versatility (Zhao, Deng, & Ngo, 2018). It is a classification algorithm based on partition. In the first step of the algorithm, a value for k is to be set. Then, n samples are divided into k clustering subsets. The similarity of clustering is evaluated. One of the most commonly used evaluation function is the square sum function of the error. It is calculated using Equation 1:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{c_i} (x_i^{(j)} - m_i)^2 \quad (1)$$

where $x_i^{(j)}$ represents the jth sample of the ith cluster center. m_i denotes the ith cluster center. J represents the sum of the squares of errors. Briefly, k-means clustering algorithm can be summarized using following steps.

Step 1: k samples are selected from the data set.

Step 2: The Euclidean distance is calculated between each sample and the k centers. The sample is divided into its nearest subset.

Step 3: The mean of each subset was calculated.

Step 4: A new cluster center is generated, return to step 2, otherwise the k-means clustering algorithm ends. Details about the k-means clustering algorithm can be found in (Ye, Huang, Teng, & Li, 2018). In this study, k-means clustering algorithm is applied to divide overall dataset to homogenous subsets. It should be noted that an optimal value for k, which represents the number of clusters, should be determined before executing the algorithm. Thus, Silhouette analysis is performed to determine the optimal number of clusters.

Silhouette values are calculated using the Equation 2 (Kaufman & Rousseeuw, 2005):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

where $a(i)$ represents the average distance of i to all other objects of cluster a and $b(i)$ is the minimum average distance of i to other objects in all other clusters.

Different k numbers (from two to ten) are employed and average silhouette values are recorded. Higher silhouette values indicate the quality of clustering. The highest average silhouette values are reached in k = 3. In other words, the optimal number of clusters is determined to be three for this dataset.

After determining the optimal number of clusters, a final k-means algorithm is applied to get clusters. In Figure 2, a grouped scatter diagram is plotted where x-axis represents the distance and y-axis represents the maximum price. Figure 2 indicates that cluster borders are more prominent along y-axis. In other words, price can be used to name the clusters. First group is named as Cluster 1 (low-priced trips), second group is named as Cluster 2 (medium-priced trips) and finally third group is named as Cluster 3 (high-priced trips). In this study, each cluster is considered to be a separate scenario. Thus, a total of three scenarios are used to compare the prediction performance of proposed methods.

To measure the quality of the clusters silhouette analysis is performed and descriptive statistics of silhouette values are presented in Table 2. As can be seen from Table 2, all of the silhouette values are positive which indicates a good clustering performance. To further analyze the quality of the clustering process, one-way ANOVA is also performed. ANOVA results indicate that all of the clusters have statistically different mean values from each other. That means all of the clusters are different from each other. In addition, Tukey-Kramer post-hoc analysis is performed. It is reported that all of the pair comparisons have p value smaller than 0.05.

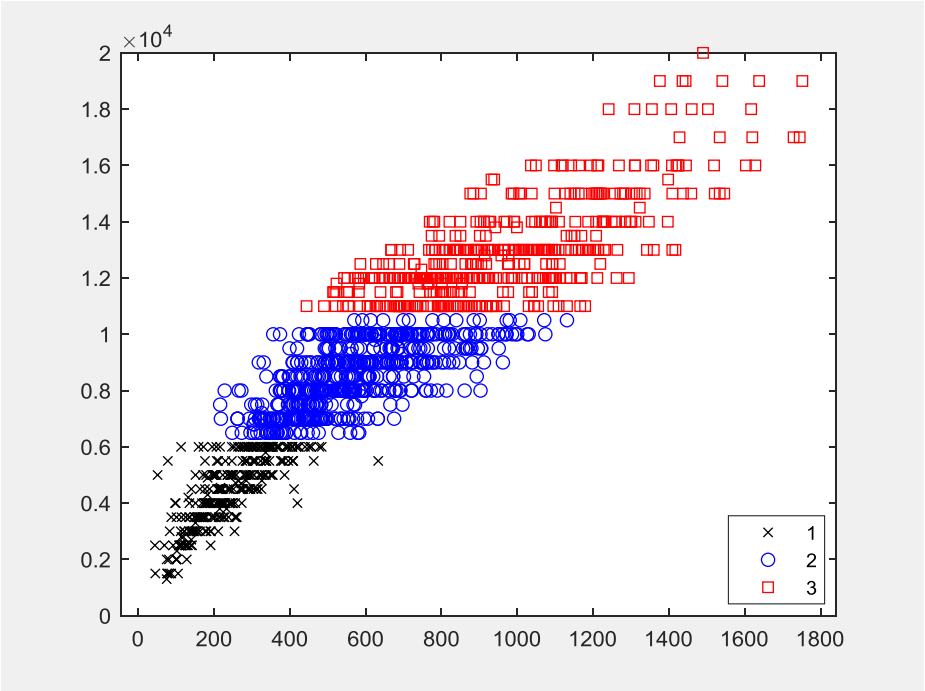

Figure 2. Distance, Price, and Clusters

Table 2. Descriptive Statistics of Silhouette Values

	Cluster 1	Cluster 2	Cluster 3
Minimum	0.4756	0.0384	0.058
Maximum	0.9226	0.9253	0.8738
Mean	0.7959	0.737	0.6683
Standard Deviation	0.1619	0.2176	0.2652
Median	0.8664	0.8382	0.7822
Kurtosis	2.7923	5.6985	3.596
Skewness	-1.2124	-1.6812	-1.4526

Linear Regression

If we represent the response variable by Y and the explanatory variables by $X_1, X_2, \dots, X_K$ then a general model relating these variables is denoted as follows.

$$E[Y|X_1 = x_1, X_2 = x_2, \dots, X_K = x_K] = \Phi(x_1, x_2, \dots, x_K) \quad (3)$$

The linear models are represented as:

$$\Phi(x_1, x_2, \dots, x_K) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K \quad (4)$$

which is linear in the parameters β_j. Details about linear regression can be found in (Seber & Lee, 2012).

The results of the linear regression model are shown in Table 3. All three models have significant F-scores. The results of the linear regression show that increasing the distance by one kilometer increases the price by 4.57 krş in cluster 1. Also, increasing the distance by one kilometer increases the price by 5.08 krş in cluster 2 and increasing the distance one kilometer increases the price by 9.70 krş in cluster 3.

Table 3. Results of The Linear Regression Analysis

Cluster -1				
	Estimate	**SE**	**T stat**	**P value**
Intercept	5968.7	134.09	44.511	1.6724e-180
Distance	4.5716	0.22566	20.259	6.9579e-68
Number of observations: 528, Error degrees of freedom : 526 R-Squared: 0.438, Adjusted R-Squared : 0.437 F-statistic vs. constant model : 410, p-value = 6.96e-68				
Cluster - 2				
	Estimate	**SE**	**T Stat**	**P value**

Intercept	8056.7	233.5	34.504	7.4038e-124
Distance	5.0769	0.23462	21.639	5.1493e-70
Number of observations: 416, Error degrees of freedom: 414 R-squared : 0.531, Adjusted R-squared: 0.53 F-statistics vs. constant model: 468, p-value =5.15e-70				
Cluster - 3				
	Estimate	SE	t Statistic	P Value
Intercept	2139	139.37	15.348	2.1578e-36
Distance	9.6966	0.53882	17.996	8.8802e-45
Number of observations: 217, Error degrees of freedom : 215 R-squared: 0.601, adjusted r-squared : 0.599 F-statistic vs. constant model : 324, p – value = 8.88e-45				

Support Vector Regression

In a regression problem, we are given a set of training patterns $(x_1, y_1), \dots, (x_l, y_l)$,where $x_i \in R^n, i = 1, \dots, l,$ and $y_i \in R$. Each y_i is the desired target, or output, value for the input vector x_i. A regression model is learned from these patterns and used to predict the target values of unseen input vectors. Support vector regression is a nonlinear kernel-based regression method which tries to locate a regression hyperplane with small risk in high-dimensional feature space (Yeh, Huang, & Lee, 2011). The support vector regression problem formulation can be found in Awad and Khanna (2015).

There are different kernel functions for support vector regression analysis. Different support vector regression models with different kernel functions are applied for forecasting and their performances are measured in this study. These kernel functions are linear, radial basis, Gaussian, second order polynomial and third order polynomial kernel functions. It is observed that the best Mean absolute percent error (MAPE) values are obtained with linear kernel function. Instead of presenting the results with all other kernel functions, only linear kernel function results are presented.

Regression Tree

Regression trees have been used by the researchers in many fields. Its accuracy has been generally competitive with linear regression. It can be much more accurate on nonlinear problems but tends to be somewhat less accurate on problems with good linear structure (Breiman, Friedman, Olshen, & Stone, 1998). Assume that the training dataset is denoted as $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Each y_n represents the nth output and $x_n = (x_{i_1}, x_{i_2}, \dots, x_{i_s})$ denotes the corresponding input of the "s" predictors in. The objective function of regression tree is to determine boxes $B_1, B_2, \dots, B_j$ that minimize the residual sum of squares and calculated using Equation 5:

$$\sum_{j=1}^{J} \sum_{i \in B_j} \left(y_i - \hat{y}_{B_j}\right)^2 \quad (5)$$

where $\hat{y}_{B_j}$ is the mean response for the training observations within the jth box. Details about the regression tree can be found in (Ruiz-Abellón, Gabaldón, & Guillamón, 2018).

Parameters of the regression tree is as follow; minimum parent size is ten, select the split predictor that maximizes the split-criterion gain over all possible splits of all predictors, pruning is active and prune criterion is mean squared error, each observation has equal weights.

Gaussian Process Regression

Assume that the function f conforms to a Gaussian process with prior mean μ and kernel k. Suppose that F_x is an observation of $f(x)$ that has been corrupted by zero-mean, i.i.d. Gaussian noise. Hence, $f(x)$is a hidden variable whose posterior distribution we can infer after observing samples of F_x at various locations in the domain. The resulting inference is called Gaussian process regression. Details can be found in (Lizotte, Wang, Bowling, & Schuurmans, 2007).

For Gaussian process regression model, fitting method is exact, basis function is constant, the dataset is not standardized, computation method is QR factorization based approach, kernel function is squared exponential kernel function, distance method is $(x - y)^2$, prediction method is exact, optimization method is quasi-newton approach.

Genetic Algorithm based Artificial Neural Network

Appropriate design of ANN can allow a larger robustness, flexibility, and opportunities for forecasting. However, there is no exact method to determine the parameters of ANN. The design of the ANN is a difficult

task. Therefore, parameters of the ANN are optimized by genetic algorithm (GA) in this study. Details about the GA-ANN can be found in (Göçken, Özçalıcı, Boru, & Dosdoğru, 2016). Hidden layer size is fixed at two. Using higher number of hidden layer will increase the completion time of the analysis. So it is determined as two. GA also determines the hidden layers' activation functions. Both the number of neurons in the hidden layer and activation function types can affect the performance of the neural network model. So these parameters are decided to be determined by GA.

In other words, GA is used to optimize the following items:

- Number of neurons in the first hidden layer (1)
- Number of neurons in the second hidden layer (2)
- Activation function type of the first hidden layer (3)
- Activation function type of the second hidden layer (4)

Table 4. Optimized GA-ANN Parameters For Each Cluster

	1	**2**	**3**	**4**
Cluster 1	28	20	Normalized Radial Basis	Radial Basis
Cluster 2	26	19	Radial Basis	Radial Basis
Cluster 3	27	20	Triangular Basis	Elliot Sigmoid

Optimized GA-ANN parameters for each cluster are given in Table 4. Activation function is selected from the following list: hyperbolic tangent sigmoid, hard limit, triangular basis, radial basis, normalized radial basis, Elliot symmetric sigmoid, log-sigmoid, pure linear, soft max, inverse, saturating linear, positive linear, and symmetric saturating linear. The training algorithm of ANN is Levenberg-Marquardt backpropagation algorithm. By considering the computation power, upper bounds for the first and second hidden layers are determined to be 30 and 20, respectively. Parameters of the GA are as follow; population size is 50, elite count is three, crossover fraction is 0.8, migration fraction is 0.2, and generations are 400.

Ensemble Model

Ensemble methods train multiple learners to solve the same problem. In contrast to ordinary learning approaches, which try to construct one learner from training data, ensemble methods try to construct a set of learners and combine them (Zhou, 2012). A model average ensemble of different forecasting models can be used to overcome the sensitivity of a specific training data and a biased forecast. In this study, ensemble model which is the arithmetic average of the forecasts produced by the considered models is used for not only increasing forecasting accuracy but also getting unbiased forecasts (ensemble).

RESULTS AND DISCUSSION

Descriptive statistics of the results are presented in Table 5 in which distance and price are given for each cluster. Cluster 1 (low-priced trips) are 21.20 % of all trips. Cluster 2 (median-priced trips) are 43.40 % of all trips and Cluster 3 (high-priced trips) are 35.40 % of all trips. Cluster 1 is priced in average 20.39 krş per km. Also, Cluster 2 and Cluster 3 are priced in average 15.98 krş/km and 14.00 krş/km, respectively. For Cluster 1, the unit price which is calculated by dividing price to distance, is much higher (20.39 krş/km) than that of Cluster 3 (14.00 krş/km).

Table 5. Descriptive statistics of clusters

	Cluster 1		**Cluster 2**		**Cluster 3**	
	Distance	**Price**	**Distance**	**Price**	**Distance**	**Price**
Minimum	44	216	444	1300	6500	11000
Maximum	633	1131	1752	6000	10500	20000
Mean	238.54	567.44	964.51	4420.37	8537.99	12937.3
Standard deviation	98.13	166.47	246.55	1231.56	1154.96	1705.38
Median	233	549	924.5	4500	8500	12500
Kurtosis	3.04	3.05	2.99	2.28	1.82	4.94
Skewness	0.43	0.58	0.56	-0.42	-0.12	1.33

There is no performance measure that is valid in all conditions and different performance measures treat different aspects of accuracy (Cameron & Windmeijer, 1997). In Table 6-8 we used various performance metrics

to evaluate the prediction performance for linear regression, support vector regression, regression tree, gaussian process regression, GA-ANN, and ensemble model. In Table 6-8, the computation time (unit is seconds) is also given. According to the computation time, ensemble model gives the best results for all clusters.

The MAPE is a relative measure, which expresses errors as a percentage of the actual data (Göçken & Boru, 2016). The MAPE is calculated using Equation 6:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{e_i}{a_i}\right| \quad (6)$$

where,$e_i = p_i - a_i, p_i$ is the predicted lead time, a_i is the actual lead time.

When MAPE values are considered, gaussian process regression is the best model for Cluster 1 and Cluster 2. On the other hand, regression tree gives the best MAPE value in Cluster 3.

Mean absolute relative error (MARE) can be defined as average absolute value of relative differences between actual value and predicted value. MARE is calculated using Equation 7:

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{a_i}\right| \quad (7)$$

According to the results of MARE in Cluster 1 and Cluster 2, gaussian process regression is the best method. However, regression tree gives the best MARE value in Cluster 3.

Mean squared relative error (MSRE) is Mean square relative error (MSRE) is another measure of performance for assessing the performance of the forecasting model. MSRE can be defined as the square of the mean absolute value of the relative differences between the actual and predicted values and calculated using Equation 8.:

$$MSRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{a_i}\right|^2 \quad (8)$$

When MSRE values are taken into account, support vector regression is the best method for Cluster 1. On the other hand, gaussian process regression gives the best MSRE value in Cluster 2 and regression tree has the best MSRE value in Cluster 3.

Root mean squared relative error (RMSRE) is defined as square root of MSRE and calculated using Equation 9:

$$RMSRE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{a_i}\right|^2} \quad (9)$$

When RMSRE values are considered, support vector regression is the best method for Cluster 1. However, gaussian process regression gives the best RMSRE value in Cluster 2 and regression tree has the best RMSRE value in Cluster 3.

Table 6. Performance Results of Cluster 1

	MAPE	MARE	MSRE	RMSRE	Computation time
Linear Regression	9.4228	0.0942	0.0133	0.1154	0.2200
Support Vector Regression	9.2574	0.0926	0.0129	0.1135	0.3179
Regression Tree	9.3105	0.0931	0.0136	0.1165	20.019
Gaussian Process Regression	9.1529	0.0915	0.0132	0.1150	0.3651
GA-ANN	9.3921	0.0939	0.0141	0.1186	1090.8
Ensemble Model	9.1705	0.0917	0.0131	0.1143	0.0821

Table 7 Performance results of Cluster 2

	MAPE	MARE	MSRE	RMSRE	Computation time
Linear Regression	5.8953	0.0590	0.0054	0.0737	0.1523
Support Vector Regression	5.6142	0.0561	0.0052	0.0721	0.2013
Regression Tree	5.9267	0.0593	0.0056	0.0751	20.096
Gaussian Process Regression	5.5845	0.0558	0.0050	0.0709	0.2916
GA-ANN	7.6229	0.0762	0.0106	0.1028	768.12
Ensemble Model	5.8461	0.0585	0.0055	0.0745	0.0654

Table 8. Performance results of Cluster 3

	MAPE	MARE	MSRE	RMSRE	Computation time
Linear Regression	16.5533	0.1655	0.0543	0.2329	0.0959
Support Vector Regression	15.6594	0.1566	0.0459	0.2142	0.0966
Regression Tree	15.3939	0.1539	0.0445	0.2109	16.993
Gaussian Process Regression	15.7823	0.1578	0.0494	0.2222	0.1034

GA-ANN	21.1338	0.2113	0.0894	0.2989	2688.6
Ensemble Model	16.3296	0.1633	0.0513	0.2265	0.0713

CONCLUSION

In Turkey, a continuous increase has been observed in road transportation. This increase is observed in both freight transportation and passenger transportation. In addition, it is seen that the most investment is made to the highway considering the transportation infrastructure investments. One of the reasons why the highway is significant is that the characteristics of transportation services can positively affect the livability of cities.

It is determined that Turkey's road passenger transportation market consists of three different clusters. These clusters are named as Cluster 1 (low-priced trips), Cluster 2 (medium-priced trips) and Cluster 3 (high-priced trips). These three clusters have different characteristics that affect the pricing strategies of firms. It is determined that, ticket price is the dominant variable in this segmentation. One of the prominent finding is that forecasting models show better performance in Cluster 1. This observation may occur due to the pricing strategies for short distance trips that are adopted by the firms. Two important facts may drive these changes. These facts are demand and competition. They can distort the quality of forecasting on short trips. One of the most significant findings that is reached at the end of the study is that the price per km is much higher for Cluster 1 (low-priced trips) while it is smaller for Cluster 3 (high-priced trips). Companies can use the findings of this study to determine the optimal prices for their customers in the Turkish road passenger transport market. Although each proposed forecasting model provides satisfactory results, it is possible to choose ensemble model among the other methods. According to the results of this paper, it can be said that the performance measurement values are cluster dependent over testing period. In addition, the results can be used to price new trips. If the firms are about to launch a new bus service in a new destination, they can use this model to determine the introductory price. In other words, the proposed price forecasting model can be used as a decision support model to price new services.

REFERENCES

Abdella, JA, Zaki, NM, Shuaib, K., & Khan, F. (2021). Airline ticket price and demand prediction: A survey. *Journal of King Saud University-Computer and Information Sciences*, 33(4), 375-391. https://doi.org/10.1016/j.jksuci.2019.02.001

Alnıaçık, Ü., & Özbek, V. (2009). Otobüs işletmelerinde hizmet kalitesinin ölçümü-Kandıra Gürkan turizm örneği. *International Journal of Economic and Administrative Studies, 1*(3), 125–138.

Awad M., & Khanna R. (2015). *Support Vector Regression*. In Efficient Learning Machines. Apress.

Breiman, L., Friedman, JH, Olshen, RA, & Stone, CJ (1998). *Classification and regression tree*s. Chapman & Hall/CRC.

Cameron, AC, & Windmeijer, FAG (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics, 77*(2), 329–342. https://doi.org/10.1016/S0304-4076(96)01818-0

Çetin, B., Barış, S., & Saroğlu, S. (2011). Türkiye'de karayollarının gelişimine tarihsel bir bakış. *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,1*(1), 123–150.

Chiu, SM, Chen, YC, & Lee, C. (2022). Estate price prediction system based on temporal and spatial features and lightweight deep learning model. *Applied Intelligence*, 52(1), 808-834. https://doi.org/10.1007/s10489-021-02472-6

Göçken, M., & Boru, A. (2016). Integrating metaheuristics and ANFIS for daily mean temperature forecasting. *International Journal of Global Warming*, 9(1), 110-128. https://doi.org/10.1504/IJGW.2016.074326

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, AT (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems with Applications, 44*, 320–331. https://doi.org/10.1016/j.eswa.2015.09.029

Gül, H., & Boz, M. (2012). İnternet ortamında pazarlama, online rezervasyon; şehirlerarası otobüs firmaları üzerine bir araştırma. *İnternet Uygulamaları ve Yönetimi Dergisi, 3*(1), 5–30.

Kapluhan, E. (2014). Historical progress and current state of highway transportation in Turkey with respect to transportation geography. *The Journal of International Social Research, 7*(33), 426–439.

Kara, H. (1999). Otobüs işletmelerinde gelir arttırıcı yönetsel stratejiler. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 2*, 195–205.

Karakan, Hİ, Türkmen, S., Giritlioğlu, İ., & Kılıç, M. (2016). İstanbul Esenler Otogarı'nda faaliyet gösteren otobüs işletmelerinin web site içeriklerinin analizine yönelik bir çalışma. *Journal of Suleyman Demirel University Institute of Social Sciences, 23*(1), 291–310.

Kaufman, L., & Rousseeuw, PJ (2005). *Finding groups in data: an introduction to cluster analysis*. Wiley-Interscience (Vol. 33). Wiley-Interscience. https://doi.org/10.1002/9780470316801

Keçeci, A. (2006). Türkiye'de karayolu taşımacılığı, *Dışişleri Bakanlığı Yayınları Uluslararası Ekonomik Sorunlar Dergisi, 20*, Retrieved from: http://www.mfa.gov.tr/turkiye_de-karayolu-tasimaciligi-.tr.mfa#

Kögmen, Z. (2014). *Karayolu taşımacılığının diğer taşımacılık modlarıyla karşılaştırılması ve sağladığı avantajlar*. Ulaştırma ve Haberleşme Uzmanlığı Tezi, Ulaştırma, Denizcilik ve Haberleşme Bakanlığı, 2014.

La, J., & Heiets, I. (2021). The impact of digitalization and intelligentization on air transportation system. *Aviation*, 25(3), 159-170. https://doi.org/10.3846/aviation.2021.15336

Li, CS & Chen, MC (2014). A data mining based approach for travel time prediction in freeway with non-recurrent congestion. *Neurocomputing, 133*, 74–83. https://doi.org/10.1016/j.neucom.2013.11.029

Li, Y., & Li, Z. (2018). Design and implementation of ticket price forecasting system. In *AIP Conference Proceedings*, 1967, 040009, https://doi.org/10.1063/1.5039083.

Lin, Y., Yang, X., Zou, N., & Jia, L. (2013). Real-time bus arrival time prediction: Case study for Jinan, China. *Journal of Transportation Engineering, 139* (11), 1133–1140. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000589

Lizotte, D., Wang, T., Bowling, M., & Schuurmans, D. (2007). Automatic gait optimization with gaussian process regression. In *IJCAI, 7*, 944–949.

Ruiz-Abellón, MC, Gabaldón, A., & Guillamón, A. (2018). Load forecasting for a campus university using ensemble methods based on regression trees. *Energies, 11*(8), 2038, 1–22. https://doi.org/10.3390/en11082038

Saran, M. (2005). İnternet ve halkla ilişkiler. *Ege Üniversitesi İletişim Fakültesi Yeni Düşünceler Hakemli E-Dergisi*, (1), 61-75.

Seber, GAF & Lee, AJ (2012). *Linear regression analysis*. John Wiley & Sons.

Sevuktekin, M., Keser, HY, Ay, S., & Cetin, I. (2014). Transportation sector in Turkey: future expectations about railway transportation of Turkey. *Kafkas Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 5*(8), 99–116.

Stavinova, E., Chunaev, P., & Bochenina, K. (2021). Forecasting railway ticket dynamic price with Google Trends open data. *Procedia Computer Science*, 193, 333-342. https://doi.org/10.1016/j.procs.2021.10.034

Truong, TMT (2021). A proposal for electronic ticketing based on travel behavior, towards the integrated public transport for smart city in Hanoi, Vietnam. In *AIP Conference Proceedings*, 2428(1), p. 040005. AIP Publishing LLC. https://doi.org/10.1063/5.0070722

Tsai, CH, Mulley, C., & Clifton, G. (2013). Forecasting public transport demand for the Sydney Greater Metropolitan Area: A comparison of univariate and multivariate methods. In *Australasian Transport Research Forum 2013 Proceedings*, Brisbane, Australia.

Wohlfarth, T., Clémençon, S., Roueff, F., & Casellato, X. (2011). A data-mining approach to travel price forecasting. In *10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, Honolulu, United States. https://hal.archives-ouvertes.fr/hal-00665041

www.kgm.gov.tr. (2018, September 5). Retrieved from: http://www.kgm.gov.tr/Sayfalar/KGM/SiteEng/Root/MainPageEnglish.aspx.

Ye, S., Huang, X., Teng, Y., & Li, Y. (2018). K-means clustering algorithm based on improved Cuckoo search algorithm and its application. In *IEEE 3rd International Conference on Big Data Analysis*, 422–426. https://doi.org/10.1109/ICBDA.2018.8367720

Yeh, CY, Huang, CW, & Lee, SJ (2011). A multiple-kernel support vector regression approach for stock market price forecasting. *Expert Systems with Applications*, *38*(3), 2177–2186. https://doi.org/10.1016/j.eswa.2010.08.004

Yu, S., Shang, C., Yu, Y., Zhang, S., & Yu, W. (2016). Prediction of bus passenger trip flow based on artificial neural network. *Advances in Mechanical Engineering*, *8*(10), 1–7. https://doi.org/10.1177/1687814016675999

Zhao, WL, Deng, CH, & Ngo, CW (2018). K-means: A revisit. *Neurocomputing*, *291*, 195–206. https://doi.org/10.1016/j.neucom.2018.02.072

Zhao, Z., You, J., Gan, G., Li, X., & Ding, J. (2022). Civil airline fare prediction with a multi-attribute dual-stage attention mechanism. *Applied Intelligence*, 52(5), 5047-5062. https://doi.org/10.1007/s10489-021-02602-0

Zhou, ZH (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

Author Contributions

All authors contributed equally.

Acknowledgement

This work was supported by Scientific Research Projects Commission of Adana Alparslan Türkeş Science and Technology University. Project Number: 18103024.