Original Research

Town trip forecasting based on data mining techniques

Mohammad Fili[1], Majid Khedmati[1]*

Received: 22 October 2020 / Accepted: 16 December 2020 / Published online: 29 December 2020

* Corresponding Author: Khedmati@sharif.edu
[1]Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran.

Abstract
In this paper, a data mining approach is proposed for duration prediction of the town trips (travel time) in New York City. In this regard, at first, two novel approaches, including a mathematical and a statistical approach, are proposed for grouping categorical variables with a huge number of levels. The proposed approaches work based on the cost matrix generated by repetitive post-hoc tests for different pairs. Then, a random forest model is constructed for the prediction of the type of trips, short or long. Finally, based on the trip type and each of the mathematical and statistical approaches, separate artificial neural networks (ANN) are developed to predict the duration time of the trips. According to the results, the mathematical approach performs better and provides more accurate results than the statistical approach. In addition, the proposed methods are compared with some other methods in the literature in which the results show that they perform better than all other methods. The RMSE of mathematical and statistical approaches is, respectively, 4.23 and 4.27 minutes for short trips, and the related value is 9.5 minutes for long trips. In addition, a modified version of the nearest neighborhood approach, entitled modified nearest neighborhood (MNN), is proposed for the prediction of the trip duration. This model resulted in accurate predictions where its RMSE is 4.45 minutes.

Keywords - Artificial neural network (ANN); Forecasting trip duration; Grouping categorical variables; Modified nearest neighborhood (MNN); Random forest

1. INTRODUCTION

Imagine you plan to leave home to go to JFK airport tomorrow at 8 a.m. When should you depart home in order to be there on time? This example is one of many instances by which the importance of trip duration can be indicated. Daily, a vast number of people are commuting via taxi inside cities. Therefore, the prediction of the trip duration is of high significance since it is a key element in scheduling the daily events. In addition, with the advent of application-based taxi services, the importance of duration prediction is now higher than anytime. Accurate prediction is vital for business success, especially in the long term, since fare calculation is heavily dependent on the duration of the trips. That is, underestimating the duration will lead to lower fare and consequently less income for the business. On the other hand, overestimating the duration will result in overcharging the passengers, which would result in their dissatisfaction and accordingly, they will be attracted by the rivals, which are offering more reasonable prices.

Travel time is considered as the amount of time required to traverse a route between two desired points. Link measurement and point measurement are two different methods for calculating the time duration [1]. Wu, Ho, and Lee [2] predicted travel time duration with support vector regression. They used data between February 15 and March 21, 2003, which was collected by loop detectors in Taipei, Taiwan. They proposed a set of parameters for support vector regression that can predict travel time duration quite well. Cho and Kwac [3] tried to predict trip duration with machine learning techniques for University Avenue to San Francisco on 101 N highway. Their goal was to show some

shortcomings of previous studies [4]-[5], and to overcome these problems by proposing some better models. Hence, they proposed an algorithm in which the days of weeks were considered in the model and consequently improved the accuracy of the prediction. Zhan et al. [6] estimated urban link travel time duration using taxi GPS data collected by New York City Taxi and Limousine Commission. They used the k-shortest path algorithm to extract 20 shortest paths as the reasonable path set. Then, the link travel times were estimated by minimizing the error between real and expected travel time for paths. Wang et al. [7] used a space-time delay neural network model for travel time prediction and compared its performance with other models, including Naïve, ARIMA, and STARIMA, and showed that their proposed model outperforms all of them.

In another research, Li and Chen [8] used a data-mining-based approach in order to predict travel time in freeway with non-recurrent congestion. They used three data mining techniques, including k-means, decision tree, and artificial neural networks, for this purpose. Zhang and Haghani [9] used the gradient boosting machine (GBM) in order to improve the travel time prediction. The GBM algorithm combines additional trees in order to correct the errors for previous trees and hence improves the overall accuracy [10]. They showed that the GBM model has advantages in travel time prediction in freeways in comparison to other models.

Antoniades et al. [11] applied different models on yellow taxi data collected with GPS in New York City to predict the trip duration and fare. They used linear regression and random forest models and, comparing these models, showed that the best model was the random forest with the validation RMSE of 5.24 minutes for time and 2.28 dollars for fare prediction. In the random forest model, they used average speed as one of the inputs for their model, while there is no information about the average speed for a future trip. In fact, if we have such information, we could easily make a point estimation for the duration, based on the distance and average speed. Later, Jaiwal et al. [12] improved the prediction accuracy by using clustering and some other models, including ridge regression, random forest, gradient boosting, and ensemble methods. They used data for yellow taxis in New York City from January 2016 to June 2016 and applied k-means with 40 clusters using pick-up and drop-off points. Then, they utilized the output of the k-means as input in their predictive models and obtained the RMSE of 4.87 by gradient boosting method.

There are different approaches to find the most important factors. One way might be to use as many features as might be important to an expert and use the design of experiments to find those factors which are statistically significant. An example of such an approach is incorporated in [13]. In this study, a series of preliminary analyses are used to find the significant variables.

While weather condition is an influential factor in trip duration, it was not considered in previous studies. Thus, in this paper, we propose some data mining techniques for clustering the attributes and then predicting the trip duration in New York City. In this regard, two novel approaches, including a mathematical approach and a statistical method, are proposed first in order to group the categorical variables with a huge number of levels, and then, the type of the trips is predicted by a random forest model.

ANN is a versatile tool which can be used for different forecasting applications, from political affairs [14] to statistical analysis [15]. In this study, artificial neural networks are proposed for the prediction of short and long trips. The performance of the proposed methods is evaluated and compared with other models in the literature.

The rest of the paper is organized as follows. The details of the data used in this study are described in section 2. Section 3 is dedicated to the mathematical and statistical approaches that are proposed for grouping the features, while section 4 contains the proposed ANN and MNN models for the prediction of trip duration. Finally, the concluding remarks are presented in section 5.

2. DATA

The data used in this paper are related to the yellow taxis in New York City gathered from New York City Taxi and Limousine Commission's trip data for April 2016, which are a subset of the larger dataset for years between 2009 and 2016 [16].

The total number of observations for this subset is around 13 million. The most important features in the original dataset, which are used in this paper, are pick-up and drop-off date, time, coordinates, and trip distance. The selected features are pre-processed in order to extract the features of interest, including the trip duration as the difference between drop-off and pick-up time, theoretical distance based on the pick-up and drop-off points, and average speed by dividing the distance by duration. According to the fact that whether a trip is started or ended in JFK or La Guardia airports, two features are also created and added to the dataset. In addition, the data points are explored to investigate in which borough a trip starts and to which ends. This was done by crossing number method [17].

On the other hand, in order to add the weather conditions to the model, the historical data for weather are collected from the Wunderground website [18] via the application programming interface (API). Figure 1 shows the bar plot of weather conditions.

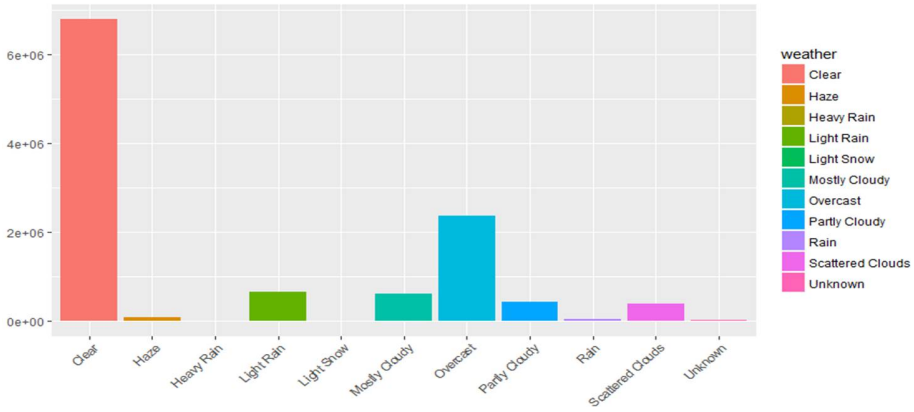

FIGURE 1
Barplot of weather conditions

In order to prepare the data for modeling, a collection of pre-processing methods, including logical relationship inspections, noise elimination, and missing data management tasks are performed. Due to the computational limitations, stratified sampling is used in order to make sure that sufficient observations are gathered from each level of the weekday. In this study, 130,682 records equal to 70% of the data are used in the training phase of the model, and the rest is remained for testing the models.

3. GROUPING CATEGORICAL VARIABLE'S LEVELS

In this study, the attributes such as weather condition and trip hour have a number of levels, and putting them all into a model may not be a good idea where some of them have the same impact on the duration of the trip. Thus, grouping them into a limited number of categories will give better results where, at the same time, it leads to the simplification of the models. Figure 2 shows the average duration of the trip for different levels of hours.

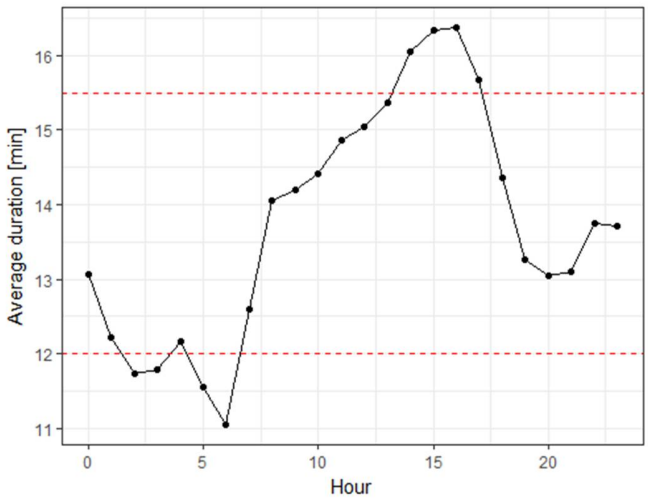

FIGURE 2
Average duration against hours in a day

Before grouping the levels of the mentioned attributes, it should be proved that the features are statistically important. It can be done by showing that at least a pair of the levels for each attribute have significantly different values where ANOVA can be used for this purpose. However, one must, first, make sure that the normality of the data, as well as the homogeneity of the variance assumptions, are satisfied. The normality and homogeneity of the data are represented in Appendix A. Now, those pairs having a significant difference in duration mean should be identified. To do so, the Tukey test is a good choice that uses family error, and it demonstrates the pairs that can be considered as different groups from the rest of the levels. Figure 3 shows one of the outputs from the Tukey test.

The test is replicated 1000 times, each time with different samples, and the results are summarized in the matrix N, where each element n_{ij} of the matrix, shows the number of times that levels i and j are identified as different. This matrix is first transformed to a probability matrix, P, by dividing all the elements to the total number of tests, and then, the matrix P is converted to the cost matrix C by applying an appropriate transformation. The transformation function should transform data in such a way that low probabilities get smaller penalty or cost, while high probabilities take higher costs. A quadratic function of the form $f(pr) = \alpha Pr^2$ is used in this paper. Figure 4 shows the cost matrix C.

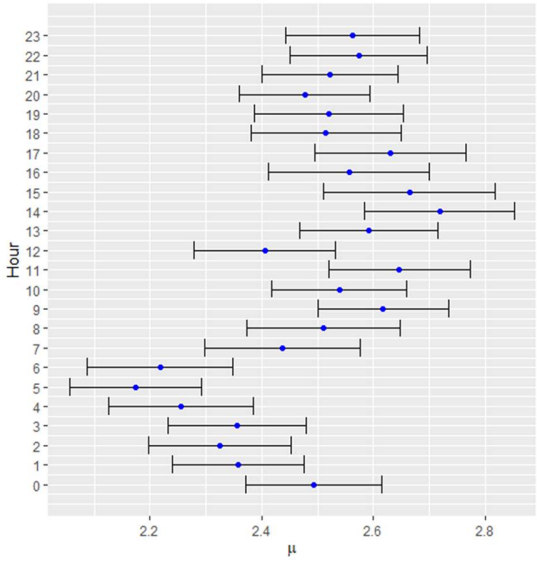

FIGURE 3
Tukey output for pairwise comparison

In matrix C, element c_{ij} represents the cost of putting two levels i and j into one group. Now, a grouping algorithm can be applied to the data where two methods are proposed for this purpose; a mathematical model and a statistical approach. The details of these models are presented in the next two subsections.

I. Mathematical Modeling

In this section, a mathematical approach is proposed in order to find the best grouping of the levels in which the best grouping means one with the minimum cost among all possible combinations. The proposed model is as follows:

$$Min\ Z = \sum_i \sum_j c_{ij} x_{ij} \qquad (1a)$$

$$Subjected\ to:$$

$$\sum_{j \neq i} x_{ij} \geq m - 1\ ; \qquad \forall\ i \qquad (1b)$$

$$x_{ij} + x_{jk} \leq x_{ik} + 1\ ; \qquad \forall\ i, j, k \qquad (1c)$$

$$x_{ij} = \{0,1\} \qquad (1d)$$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	*	0	0	0	0	112	289	0	0	0	0	1	3	8	31	50	12	3	0	0	0	0	0	0
1	0	*	0	0	0	4	54	0	1	4	13	63	88	130	260	272	154	87	12	0	0	0	2	1
2	0	0	*	0	0	0	9	0	13	36	70	191	239	266	392	402	310	232	61	2	0	1	22	18
3	0	0	0	*	0	0	10	0	12	41	63	193	225	283	407	405	310	232	63	1	0	1	22	15
4	0	0	0	0	*	1	26	0	6	16	31	116	150	197	333	337	235	147	28	1	0	1	11	6
5	112	4	0	0	1	*	0	2	267	356	376	448	452	461	467	466	462	452	370	156	100	153	321	312
6	289	54	9	10	26	0	*	35	421	453	450	470	471	470	471	471	470	466	457	331	289	341	435	445
7	0	0	0	0	0	2	35	*	4	12	28	90	133	173	312	318	205	138	22	0	0	0	7	5
8	0	1	13	12	6	267	421	4	*	0	0	0	0	0	5	7	1	0	0	0	0	0	0	0
9	0	4	36	41	16	356	453	12	0	*	0	0	0	0	1	1	0	0	0	0	0	0	0	0
10	0	13	70	63	31	376	450	28	0	0	*	0	0	0	0	1	0	0	0	0	0	0	0	0
11	1	63	191	193	116	448	470	90	0	0	0	*	0	0	0	0	0	0	0	1	2	1	0	0
12	3	88	239	225	150	452	471	133	0	0	0	0	*	0	0	0	0	0	1	2	1	0	0	0
13	8	130	266	283	197	461	470	173	0	0	0	0	0	*	0	0	0	0	0	3	4	2	0	0
14	31	260	392	407	333	467	471	312	5	1	0	0	0	0	*	0	0	0	1	20	39	21	2	2
15	50	272	402	405	337	466	471	318	7	1	1	0	0	0	0	*	0	0	0	31	45	24	3	3
16	12	154	310	310	235	462	470	205	1	0	0	0	0	0	0	0	*	0	0	8	12	6	0	0
17	3	87	232	232	147	452	466	138	0	0	0	0	0	0	0	0	0	*	0	2	4	1	0	0
18	0	12	61	63	28	370	457	22	0	0	0	0	1	0	1	0	0	0	*	0	0	0	0	0
19	0	0	2	1	1	156	331	0	0	0	0	1	2	3	20	31	8	2	0	*	0	0	0	0
20	0	0	0	0	0	100	289	0	0	0	0	2	1	4	39	45	12	4	0	0	*	0	0	0
21	0	0	1	1	1	153	341	0	0	0	0	1	0	2	21	24	6	1	0	0	0	*	0	0
22	0	2	22	22	11	321	435	7	0	0	0	0	0	0	2	3	0	0	0	0	0	0	*	0
23	0	1	18	15	6	312	445	5	0	0	0	0	0	0	2	3	0	0	0	0	0	0	0	*

FIGURE 4
The cost matrix based on the Tukey test for hours of the day

where variable x_{ij} is defined as a Boolean variable, which is 1 if the levels i and j are in the same group and zero if otherwise. Also, c_{ij} is the cost of putting levels i and j in one group. In this model, the objective function (1a) calculates the total cost of grouping; the first constraint ensures that the grouping is balanced; that is, each group or cluster has at least m members. The proposed model needs sensitivity analysis for different values of parameter m. The second constraint states that if i and j are in one group, and j and k are grouped together; consequently, i and k must be within the same group as j. The results of the model for different values of m are summarized in Table 1.

TABLE I
MATHEMATICAL MODELING SUMMARY

m	Number of Groups	Total Cost	Average Cost per Group	Increase Amount
2	11	0	0	-
3	7	18	2.57	2.57
4	5	38	7.6	5.03
5	4	92	23	15.4
6	4	168	42	19
7	3	282	94	52
8	2	2162	1081	987
9	2	2162	1081	0
10	2	3010	1505	424

In addition, the results are demonstrated in Figure 5, where it shows the increasing trend of the cost by increasing the value of m. Based on Figure 5, $m = 7$ is the elbow point where a sudden increase happens, and the slope changes dramatically. Considering the fact that larger values of m lead to more information, but higher costs and smaller values of m lead to lower cost and less information, the point $m = 7$ is selected as the best value which results in the most information and the least cost in comparison to the other values of m.

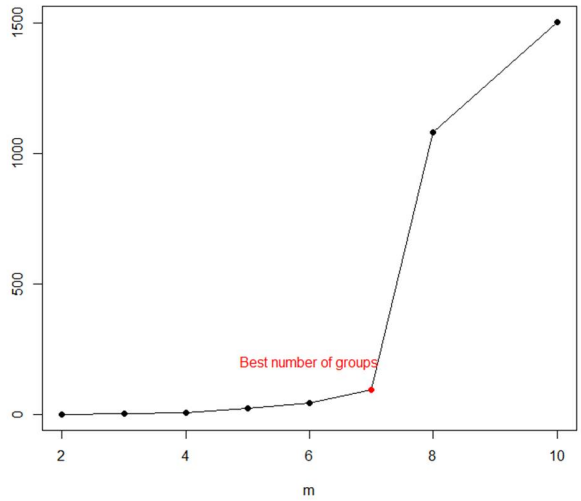

FIGURE 5
The average cost for different values of m

TABLE 2
RECOMMENDED GROUPING BY THE MATHEMATICAL APPROACH

number of groups = 3 m = 7 Cost = 282		
Group 1	**Group 2**	**Group 3**
0	1	**11**
8	2	**12**
9	3	**13**
10	4	**14**
18	5	**15**
19	6	**16**
20	7	**17**
21		
22		
23		

According to $m = 7$, all the levels are grouped in 3 groups in which the results are summarized in Table 2.

II. Statistical Modeling

In the statistical approach, the cost matrix C is treated as a distance matrix. Based on this matrix, the higher the cost, the more distant are the two levels from each other, and as a result, it is less likely to be grouped into one cluster. Consequently, elements with lower costs or penalties will tend to be within the same group as they are near to each other.

For this purpose, the bottom-up hierarchical clustering [19] is used where it puts, initially, each element in the separate clusters and then, at each step, it finds the two nearest clusters according to a specified linkage measure and merges them into one cluster. This procedure will be continued until all the elements are put into a single cluster. The final result can be shown as a tree, which is known as the dendrogram. This dendrogram can be cut at any height to give the best clustering at that level. Some of the linkage measures include the minimum distance, maximum distance, and mean distance. In this paper, the complete and mean linkages are used, and as an example, the dendrogram for complete linkage is demonstrated in Figure 6.

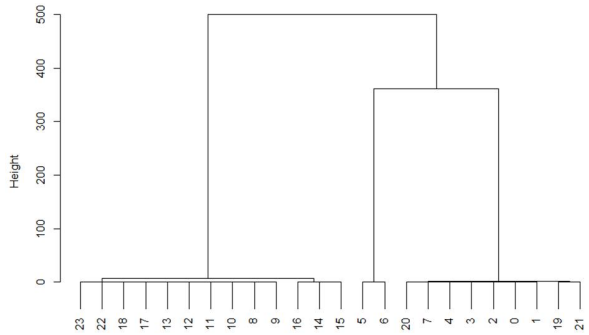

FIGURE 6
Dendrogram for hour with complete linkage

Based on the results (not reported here), the complete linkage leads to better performance, and accordingly, the final result of the statistical approach, based on complete linkage and for different cut-offs at various heights, is summarized in Figure 7.

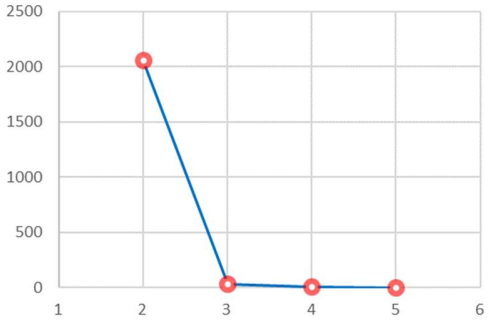

FIGURE 7
Costs of grouping with the statistical approach

Again, based on the statistical approach, three groups are determined as the best result in which it satisfies the least cost and the most information measures, simultaneously. The grouping result is shown in Table 3.

TABLE 3
RECOMMENDED GROUPING BY STATISTICAL APPROACH

number of groups = 3 **Cost = 32**		
Group 1	**Group 2**	**Group 3**
0	5	**8**
1	6	**9**
2		**10**
3		**11**
4		**12**
7		**13**
19		**14**
20		**15**
21		**16**
		17
		18
		22
		23

4. METHOD

In this section, the modeling procedure and the related details of each step are described. In this regard, first, a brief description of the model is provided, and then, different stages of the model are introduced.

In this paper, two different modeling approaches of artificial neural network (ANN) and modified nearest neighborhood (MNN) are proposed. In the first model, in order to predict the trip duration for a new observation, a classifier is used to identify the type of the trip, that is, short or long. Then, a neural network is constructed for each type, and its performance is evaluated and compared with both mathematical and statistical grouping methods. In the second modeling approach, a modified version of the nearest neighborhood is proposed, which is both very straightforward and precise. In the following subsections, the details of each of the approaches are provided.

I. Trip type classification

In order to predict the type of trips, a random forest model is proposed. In this model, features like theoretical distance, pick-up and drop-off longitude and latitude, trip hour, weather condition, and weekdays are used. In the initial dataset, based on the results of some splitting methods such as entropy or information gain, the split point of 40 is determined to categorize the trips with a duration less than or equal to 40 minutes as short trips, and the trips with bigger duration as long trips.

Also, the K-means clustering algorithm is used in order to segment the city into different clusters, which resulted in 4 clusters. These city segments are used as inputs in the random forest as well as other variables defined earlier. All the variables mentioned above are used as input of the random forest model in which the hyper-parameter of the random forest model is tuned via 10-fold cross-validation, and selection of 3 variables has resulted in the best performance of the model.

Figure 8 shows the variable importance for this model. Then, the model has been applied to the test data where the accuracy measure of the model for the test set is obtained as 0.979.

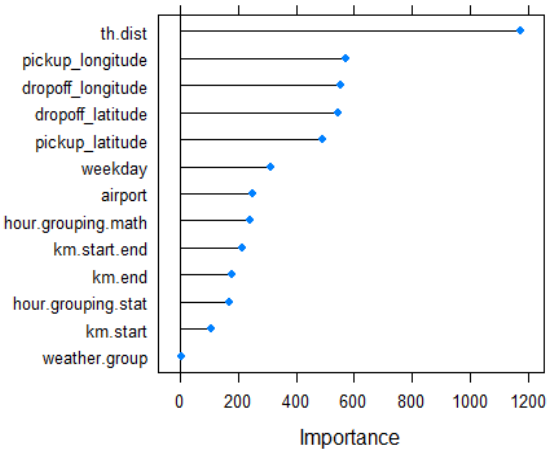

FIGURE 8
Variable importance in random forest model

II. Prediction with ANN

In this section, based on the type of trip, a neural network is proposed for the prediction of the trip duration. Due to the high frequency of short trips compared to long trips, both mathematical and statistical grouping approaches are considered in the development of the neural network for short trips.

The neural networks proposed in this paper are three-layer, fully connected, feed-forward networks in which their hyper-parameters are tuned via 5-fold cross-validation. Based on trial-and-error, the best neural network consists of one input layer with 22 nodes, one hidden layer with 4 nodes, and the output layer with 1 node. For readability, only the shape of the ANN model for long trips is represented in Figure 9. As it is shown in this figure, some dummy variables like those created for the weather, city segments, weekdays, and grouping levels are introduced as inputs to the model. The results for each type of trip are discussed later in section 5.

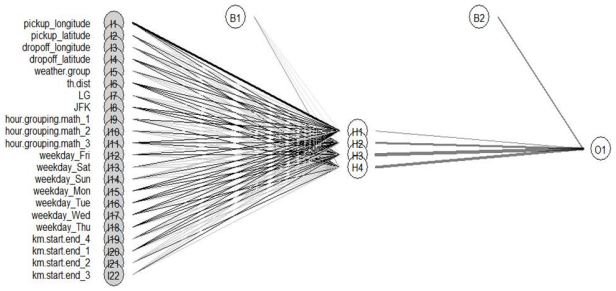

FIGURE 9
ANN model for long trips

III. Modified nearest neighborhood (MNN)

In this section, a new model is introduced specific to the trip duration, but the concept can be generalized to almost any prediction problem. This model is a lazy learner and starts the prediction when a new observation is given to the model. Then, it uses all available observations to find the nearest neighbors and finally finds the predicted value for that new observation based on the average of the values for the target variable of observations in its neighborhood.

In this model, the features are divided into two categories of primary and secondary features. The primary features are those the decision maker has a strict point of view; that is, for finding the nearest neighbors, the decision maker tries to get exactly the same observations based on the value for primary features. In this study, the primary features are weekday, hour, weather condition, La Guardia, and JFK airports. Unlike the k-nearest neighborhood method, the number of neighbors is not defined in MNN by the modeler. Instead, the control parameters are the upper bound of search radius for pick-up and drop-off points entitled UB_{pick} and UB_{drop}, respectively. In the second phase, the model searches, among the filtered observations obtained in the first phase, for the trips that their starting point is within the circle with a predefined radius; i.e., UB_{pick}. Then, the model will do the same for the remaining observations to find the trips for which the difference with their endpoint is smaller than UB_{drop}. Finally, it gives the prediction based on the average duration of the observations, which are the nearest neighbors.

The superiority of MNN over KNN is that it doesn't impose a predefined number of instances to the model; thus, the model searches for really similar observations. This matters especially in the problems where the neighbors may have different behavior based on some key features. In other words, one may find two close observations, one at 6 a.m. and the other at 8 a.m. while they are from two different situations with completely different traffic behavior and hence, they have a large difference in their target value. In this example, although the observations are so close and the time difference is small, but considering them as a neighbor will impact the average, and consequently, the prediction will be no longer accurate. As it was mentioned earlier, the model needs the cut-off value for the start and end radius. The smaller the upper bounds, the more accurate the prediction will be. This is shown in Figure 10. On the other hand, it will increase the chance of encountering an empty list for nearest neighbors. For large datasets, this will be no longer a serious concern, but for a small training set, this model may not be helpful.

Due to the computational limitations, we picked only 1,000 records, and the result was reasonable. The RMSE turned out to be 4.447 minutes, which is acceptable, while the sample size was not very large. Therefore, it is very likely to have more precise predictions by increasing the sample size.

IV. Performance evaluation

In this section, the performance of the proposed methods is evaluated and compared in terms of RMSE measure. In addition, the proposed methods are compared to ridge regression, random forest, and gradient boosting methods [11]-[12]. The results of the prediction are summarized in table 4.

TABLE 4
PREDICTION RESULT FOR TRIP DURATION

Model	RMSE
ANN for short trips with mathematical grouping	**4.23**
ANN for short trips with statistical grouping	**4.27**
ANN for long trips	**9.60**
MNN	**4.45**
Ridge regression	**5.66**
Random Forest	**5.15**
Gradient Boosting	**4.87**

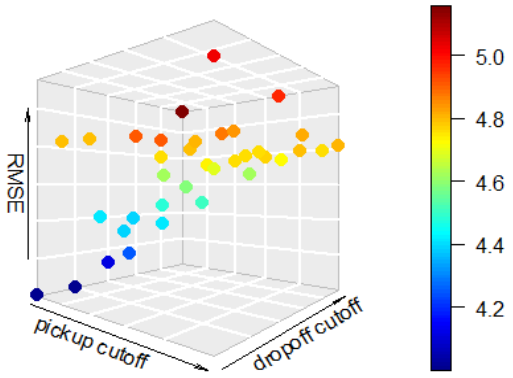

FIGURE 10
RMSE against different amount of control parameters in the MNN model

The first three rows of the table show the ANN-based approaches for the prediction of the trip duration under mathematical and statistical grouping methods. Then, the MNN approach is shown in the fourth row, and other approaches for comparison are represented in rows 5 to 7. Based on the results, the ANN approach for short trips under the mathematical method provides the best performance among all other approaches. Although the RMSE for long trips is high, the number of long trips is too few that it will not affect the whole RMSE of the proposed ANN approach. In addition, the proposed MNN performs better than the three other approaches of ridge regression, random forest, and gradient boosting. Consequently, the approaches proposed in this paper outperform the competing approaches in the literature, while ANN performs better than MNN.

Moreover, the ANN benefits from the ability to classify the type of trips.

5. Concluding Remarks

In this study, two new approaches, including mathematical and statistical methods, were proposed for grouping the categorical features with high number of levels. This will help the model to eliminate unnecessary information, and hence, besides simplification, it can lead to higher accuracy in predictions. Then, a random forest model was proposed for determining the type of trip as a short or long trip. Finally, an ANN was constructed for the prediction of the trip duration based on each of the mathematical and statistical models.

In addition, a new model as a modified version of the nearest neighborhood abbreviated by MNN was proposed for trip-duration prediction. Then the performance of the proposed approaches was evaluated and compared to the competing methods in terms of root mean squared error (RMSE) in which it was concluded that the ANN with the mathematical grouping method for short trips, outperforms all other approaches. Although the RMSE for long trips is high, the number of long trips is too few compared to short trips, and hence, it will not affect the total RMSE. However, since all conditions are kept fixed in the case of long trips, using the mentioned modeling approaches might not be a good idea. Instead, it is recommended to predict the long trips via dynamic models in which the time is also considered, or modify the model such that it takes the changes of key attributes into account as time goes on. In addition, the MNN provided better results than other competing methods.

In order to improve the accuracy of the predictions, one might consider other variables, including paths as well as traffic parameters. Also, if the path can be predicted, the prediction can be made more accurately, which is recommended for future researches.

Competing Interests

The authors declare that they have no competing interests.

References

[1] Turner, S. M., Eisele, W. L., Benz, R. J., & Douglas, J. (1998). Travel time data collection handbook. In *Federal Highway Administration, USA*.

[2] Wu, C. H., Ho, J. M., & Lee, D. T. (2004). Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, 5(4), 276-281. https://doi.org/10.1109/TITS.2004.837813.

[3] Cho, Y., Kwac, J. (2007). *A Travel Time Prediction with Machine Learning Algorithms*. http://cs229.stanford.edu/proj2007.

[4] Kwon, J., & Petty, K. (2005). Travel time prediction algorithm scalable to freeway networks with many nodes with arbitrary travel routes. *Transportation Research Record*. https://doi.org/10.3141/1935-17

[5] Kwon, J., Mauch, M., & Varaiya, P. (2006). Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. *Transportation Research Record*, 1959, 84-91. https://doi.org/10.3141/1959-10

[6] Zhan, X., Hasan, S., Ukkusuri, S. V., & Kamga, C. (2013). Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33, 37-49. https://doi.org/10.1016/j.trc.2013.04.001

[7] Wang, J., Tsapakis, I., & Zhong, C. (2016). A space-time delay neural network model for travel time prediction. *Engineering Applications of Artificial Intelligence*, 52, 145-160. https://doi.org/10.1016/j.engappai.2016.02.012
[8] Li, C. Sen, & Chen, M. C. (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
[9] Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308-324. https://doi.org/10.1016/j.trc.2015.02.019
[10] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189-1232. https://doi.org/10.1214/aos/1013203451
[11] Antoniades, C., Fadavi, D., Amon, A. F. J. (2016). *Fare and Duration Prediction: A Study of New York City Taxi Rides*. http://cs229.stanford.edu/proj2016/report
[12] Jaiwal, H., Bansal, T., Jakate, P., Saxena, T. (2016). NYC Taxi Rides: Fare and Duration Prediction. https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a077.pdf
[13] Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. Journal of Industrial Engineering International, 9, 1-9. https://doi.org/10.1186/2251-712X-9-1
[14] Zolghadr, M., Niaki, S. A. A., & Niaki, S. T. A. (2018). Modeling and forecasting US presidential election using learning algorithms. Journal of Industrial Engineering International, 14, 491-500. https://doi.org/10.1007/s40092-017-0238-2
[15] Maleki, M. R., Amiri, A., & Mousavi, S. M. (2015). Step change point estimation in the multivariate-attribute process variability using artificial neural networks and maximum likelihood estimation. Journal of Industrial Engineering International, 11, 505-515. https://doi.org/10.1007/s40092-015-0117-7
[16] 2016 Yellow Taxi-Trip Data. (n.d.). https://data.cityofnewyork.us/Transportation/2016-Yellow-Taxi-Trip-Data/k67s-dv2t
[17] O'Rourke, J. (1998). Computational Geometry in C. 2nd edition, Cambridge.
[18] weather data. (n.d.). www.wunderground.com
[19] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. 3rd edition, Elsevier.
[20] Montgomery, D. C. (2012). Design and Analysis of Experiments. 8th Edition, John Wiley.

AUTHOR (S) INFORMATION

Mohammad Fili, M.Sc., Department of Industrial Engineering, Sharif University of Technology

Majid Khedmati, Assistant Professor, Department of Industrial Engineering, Sharif University of Technology

APPENDIX A

The first step is to draw a sample out of data for each level. For determining the size of the sample, operational characteristic (OC) curves can be used. Selecting the power and the minimum difference to be identified with at least $(1-\alpha)\%$ confidence will lead to the sample size. In this paper, the sample size of 100 is chosen, which can identify real inequalities between treatments' mean, with the power of 0.9, if at least a difference of 0.5 unit exists with at least 95% confidence.

Figure A.1 shows that the duration distribution is highly skewed; thus, the normality assumption will be violated, and hence a transformation is needed. For this purpose, assume that $\mu_y, {\sigma_y}^2$ are mean and variance of the random variable y, respectively, and their relationship is as follows [20].

$${\sigma_y}^2 = g(\mu_y) \qquad (A.1)$$

where, y^T, which is the transformed data, is defined as $y^T = h(y)$. Then, based on the Tailor series, we can show that:

$$Var(y^T) = \left(h'(\mu_y)\right)^2 g(\mu_y) = a > 0 \qquad (A.2)$$

With a simple replacement, we will have:

$$h'(\mu_y) = \frac{\sqrt{a}}{\sqrt{g(\mu_y)}} \qquad (A.3)$$

Taking an integral from equation (A.3) will result in the following equation in which the parameter a is a constant:

$$h(\mu_y) = \sqrt{a} \int \frac{d\mu_y}{\sqrt{g(\mu_y)}} \qquad (A.4)$$

Now, following the steps mentioned above, one can determine the appropriate transformation. But it is important to consider that the average and variance of the population are unknown, and hence, their point estimates should be used instead. Figure A.2 shows a linear relationship between factor levels mean duration and standard deviation. Since the relationship is linear, then we can write $s_i \propto \bar{y}_i$ and consequently, $\sigma_y = k\mu_y$ where k is constant.

Now raising both sides to the power of two will result in ${\sigma_y}^2 = (k\mu_y)^2 = g(\mu_y)$ and substituting this relationship in equation (A.4) will result in:

$$h(\mu_y) = \sqrt{a} \int \frac{d\mu_y}{\sqrt{(k\mu_y)^2}} = \frac{\sqrt{a}}{\sqrt{k}} \ln(\mu_y) \qquad (A.5)$$

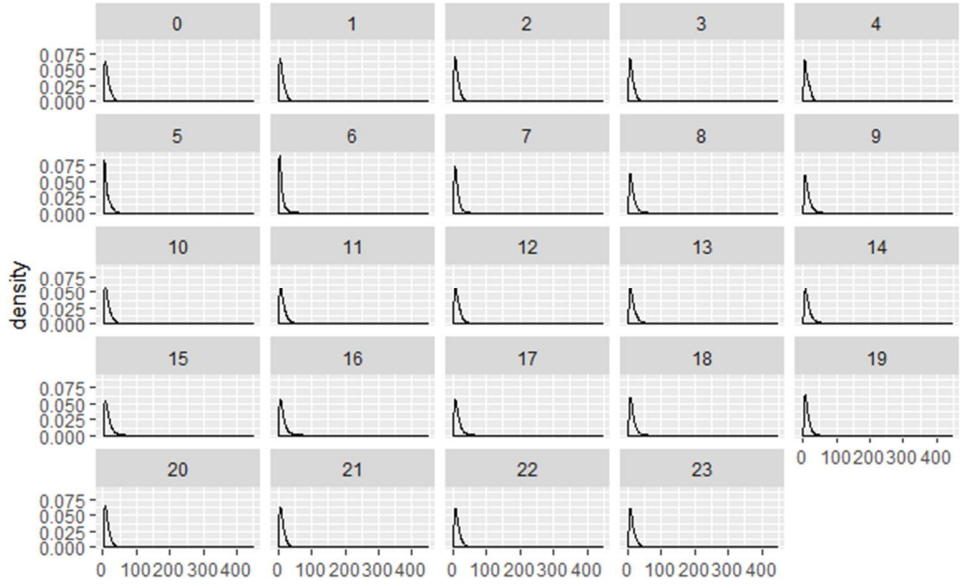

FIGURE A.1.
Duration distribution for different hour levels

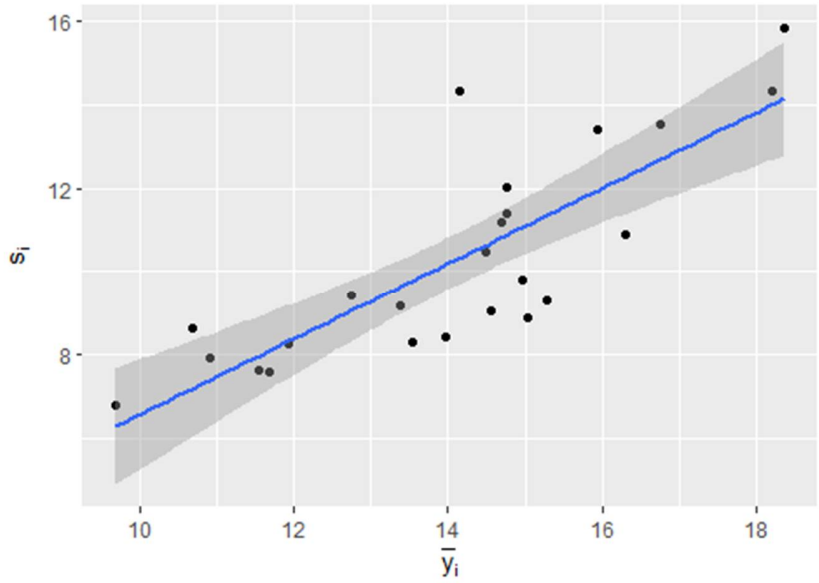

FIGURE A.2.
Scatterplot of mean duration against standard deviation

Transforming data will result in normality, where it is shown in Figure A.3. After making sure about the normality and homoscedasticity assumptions, ANOVA can be used to see whether we can reject the null hypothesis and conclude that the attribute is significant or not. In this study, the p-value for the ANOVA was 1.4e-13; thus, the null hypothesis will be rejected. Checking residuals also did not give any evidence for violation of model assumptions.

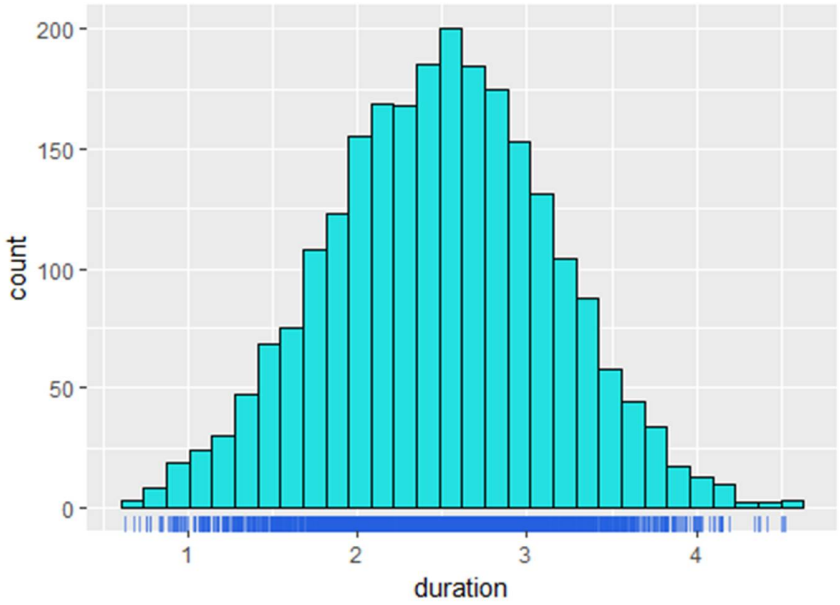

FIGURE A.3.
Histogram of transformed data