IET Intelligent Transport Systems

Special Issue: Intelligent Transportation Systems in Smart Cities for Sustainable Environments

Seoul bike trip duration prediction using data mining techniques

ISSN 1751-956X
Received on 29th November 2019
Revised 5th March 2020
Accepted on 17th April 2020
E-First on 14th July 2020
doi: 10.1049/iet-its.2019.0796
www.ietdl.org

Sathishkumar V E[1], Jangwoo Park[1], Yongyun Cho[1] ✉

[1]*Department of Information and Communication Engineering, Sunchon National University, Suncheon, Republic of Korea*
✉ *E-mail: yycho@scnu.ac.kr*

Abstract: Trip duration is the most fundamental measure in all modes of transportation. Hence, it is crucial to predict the trip-time precisely for the advancement of Intelligent Transport Systems and traveller information systems. To predict the trip duration, data mining techniques are employed in this study to predict the trip duration of rental bikes in Seoul Bike sharing system. The prediction is carried out with the combination of Seoul Bike data and weather data. The data used include trip duration, trip distance, pickup and dropoff latitude and longitude, temperature, precipitation, wind speed, humidity, solar radiation, snowfall, ground temperature and 1-hour average dust concentration. Feature engineering is done to extract additional features from the data. Four statistical models are used to predict the trip duration. (a) Linear regression, (b) Gradient boosting machines, (c) *k* nearest neighbour and (d) Random Forest (RF). Four performance metrics root mean squared error, coefficient of variance, mean absolute error and median absolute error is used to determine the efficiency of the models. In comparison with the other models, the best model RF can explain the variance of 93% in the testing set and 98% (R^2) in the training set. The outcome proves that RF is effective to be employed for the prediction of trip duration.

1 Introduction

According to recent studies, it is expected that more than 60% of the population in the world tends to dwell in cities, which is higher than 50% of the present scenario [1]. Some countries around the world are practising righteous scenarios, renderings mobility at a fair cost and reduced carbon discharge. On the contrary other cities are far behind in the track [2]. Urban mobility usually fills 64% of the entire kilometres travelled in the world. It ought to be modelled and taken over by inter-modality and networked self-driving vehicles which also provides a sustainable means of mobility. Systems called Mobility on Demand (MOD) has a vital part in raising the vehicles' supply, increasing its idle time and numbers. Bike-sharing MOD systems are already firmly holding the effective part in short commuting and as 'last mile' mobility resources on inter-modal trips in several cities. Certain issues prevail in the maintenance, design, and management of bike-sharing systems: layout of the station design; fleet size and capacity of the station; detecting broken, lost, or theft bikes; pricing; monitoring of traffic and customer activities to promote behaviour virtuously; and marketing using campaigns etc. System balancing is the hardest endeavour: In the daytime, some stations are likely to be crowded with bike flow, while leaving other stations empty, which hampers pick-up and drop-off, respectively. So, to restore the balance, several manual techniques, like shifting bikes through trucks, cars and even by volunteers are employed. Data analysis techniques and studies focus on dynamic systems and optimisation methods are utilised for complementing the knowledge base of employing optimum rebalancing policies [3].

Today, bike-sharing systems are blooming across more than 1000 cities around the world [4], particularly in big or large cities like New York City, Paris, Washington DC, London, Beijing and Barcelona. To complete a short trip renting a bike is a faster way when compared to walking. Moreover, it is eco-friendly and comfortable too compared to driving. This paved the way for its raising popularity [5]. As in other countries, South Korea has its bike-sharing system called *Ddareungi*, set up in 2015. In English, it is given the name Seoul Bike. The system is designed to tackle traffic congestion, high oil price, and air pollution in Seoul, and to create a healthier society thus improving the quality of life for Seoul residents. *Ddareungi* was first implemented in Seoul in October 2015 in selected areas on the right bank of the Han River, Seoul. As months passed, the station's count rose above 150 and 1500 bikes are made available. Since 2016, the number of stations keeps elevating and covers districts that were not included earlier. According to July 2016 data, there is an availability of 300 stations functioning with 3000 bikes. Furthermore, Seoul mayor Park Won-soon affirms his intention of raising the availability of bikes to 20,000. As of now, there are more than 1500 Bike Rental stations in Seoul running 24 h with advanced technology.

As Advanced Travellers Information Systems and Intelligent Transportation Systems shows rapid growth, the data raised by these systems can be valuable for knowing and enhancing procedures in transport companies and also other transportation dealing organisations, i.e. public and private transportation companies, logistics companies and the local government. Trip duration in public rental bikes is a notable example of a travel analytics problem which profits from data analysis. Knowing the estimated trip duration in advance helps the government organisations and also the travellers as in case of picking the right choice on route planning and timing. For this reason, information on bike trips (more likely GPS data) obtained by rental bikes can be utilised.

Data mining methods can be employed to predict the duration of the trip. By exploring the data acquired by Seoul Bike, data mining methods link trip duration to certain variables stating the trip as the source, destination, weather, time information of the day, day of the week, month and minute. Several algorithms are developed and employed to predict the trip duration. Yet the performance of prediction varies, which results in certain difficulties. Data mining performs the key task of classifying the algorithm that shows an optimum function for a particular issue. However, it is already known that there is no best algorithm for a domain of large issue [6]. Picking an algorithm for a specific issue is done by trial and error method or based on an expert's advice. However, neither of the approaches is satisfactory to the end-user who wish to access the technology in a cost-efficient manner [7]. The Perfect learning algorithm to predict trip duration usually vary for different rental bike sharing system, due to differences in use, and driving habits. Hence, the choice of the algorithm should be rendered at a lower range like the rental bike itself instead of the global level. Whereas, for systems of multiple data sources in

© The Institution of Engineering and Technology 2020

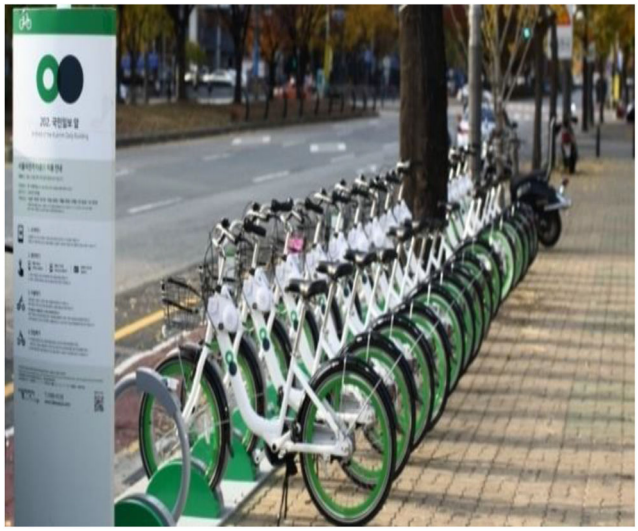

Fig. 1 *Docking stand in Seoul*

which the data structure is the same, raising the reliability of the model is plausible for a specific source. It can be achieved by utilising other data sources for training.

In this analysis, data mining methodology is used to predict the duration of trip of each trip using weather information. Given that weather information plays an important role in transportation, weather patterns information is considered a primary determiner in predicting the duration of the bike rental trip. The data is pre-processed and cleansed and then combined with Seoul weather data. Four regression algorithms are used to predict the duration of each rental bike trip, and the best performing algorithm is picked. Fig. 1 presents a view of the docking station of Seoul rental bikes. The knowledge extracted from this pattern of trip duration can be used to provide convenient public bike sharing and development of tourism services. Accurate travel-time prediction is also critical to the development of intelligent transport systems, route planning, navigation applications and advanced information systems for travellers.

The structure of the rest of the paper is organised as follows. Section 2 reviews the previous studies on bike-sharing systems and prediction approaches. Section 3 describes the algorithms used in detail. Section 4 deals with the preparation of data and exploratory analysis. Section 5 describes the evaluation indices. Section 6 deals with the discussion of results. Finally, Section 7 concludes the paper.

2 Literature survey

A vast range of researches is conducted in the prediction of trip duration. Travel time is the required time for traversing a path or a link between any two points of interest. There are two approaches to the estimation of travel times: point measurement and link measurement [8]. By using actual traffic data, it is shown that simple prediction methods can provide a significant estimate of the duration of trips for beginning shortly (up to 20 min). On the other side, better predictions can be produced with historical data of trips beginning more than 20 min away.

Research by Li *et al.* [9] deal with the issue to predict path travel times when only a low number of GPS floating cars are accessible. An algorithm is developed for learning local congestion patterns of the compact set of commonly shared paths through historical data. In the view of the travel time prediction question, the current trends of congestion around the query path through recent trajectories are established, accompanied by inferring its travel time shortly. Mridha and NiloyGanguly [10] establish a connection (Road Segment) Travel Time Estimation Algorithm, which is named as Least Square Estimation with Constraint, that calculates travel time by 20% more accurately than existing algorithms. The primary concept is the augmentation of a subset of trips along the specific paths utilising logged distance information, rather than fitting the *adhoc* 'route-choice' model. An approach to predict Kriging travel time by Miura [11] used Kriging method, a spatial prediction method as a predictive measure for the travel time of car in a notional four-dimensional space. Every point in the space signifies a particular trip and the co-ordinates of a particular point stand for the start and terminal on the plane. And so the duration of travel is rendered as a function over the four-dimensional space. The system prediction depends on the aspect that the adjacent point (four-dimensional space) holds nearly the same travel time. Here, a breaking down of travel time from source to destination into link travel time is not necessary. The data from 'probe vehicles' are necessarily used in this method for predicting the time of travel in imminence. One case study in London reveals this method's feasibility.

For predicting the duration of the trip on a freeway, Kwon *et al.* [12] used the occupancy details and the flow data, which is from the single loop detectors and past trip duration data. The very method is utilised by Chien and Kuchipudi [13]. Zhang andRice [14] proposed using a linear-based model for predicting the lasting period of a short-term freeway trip. Duration of a trip is considered as a function of the departure time. The outcomes reveal that the error drifts between 5 and 10%, in case of a low dataset, whereas for a larger dataset, the error varies between 8 and 13%. Wu *et al.* [15] used support vector regression (SVR) for the prediction of trip duration. For their examination, they exploited a real highway traffic data. Furthermore, a set of trial and error SVR parameter is proposed which in turn directs to a model that surpasses a baseline model. Balan *et al.* [16] put forward a system which renders period information, including the estimated price and trip duration for the travellers. A historical data of paid taxi trips that comprises almost 250 million entries are taken into consideration for this study.

Given the fast-behavioural change in the network of vehicles, application of the learning algorithms for the prediction of the travel time for various vehicles for a prolonged duration certainly leads to incorrect predictions. So it is important to figure out the optimum algorithm for each context. Using a trial and error method is the other possibility. It intends to find the best fitting algorithm for the particular dataset (i.e. for a specific period and also for a specific vehicle). And the best algorithm is picked out by the process of comparison with other algorithms [17]. The method consumes a lot of time, given several available alternatives. Meta-learning deals with the problem concerning the selection of the algorithm which leads to an optimum model that gives a precise prediction for every trip [18]. The methodology is tried out on the data gathered from a Drive-In project. The results assert meta-learning's capability of picking out the algorithm with optimum precision. Handley *et al.* [19] investigate the application of machine learning and data mining to boost the prediction of travel times in an automobile. Data collected from the San Diego freeway system and *k*-nearest neighbour combined with a wrapper are used to choose useful features and parameters for normalisation. The results suggest that three nearest neighbours greatly outperform

© The Institution of Engineering and Technology 2020

predictions available from existing digital maps when using information from freeway sensors. Analyses often show some surprises about the utility of other features such as day and time of the trip. Hailu and Gau [20] present models that predict two parameters of the fishing trip that is recreational: the length and trip timing in a year. A discrete choice model called logit connects the choosing of trip timing with scheduling events and the demographic characteristics of anglers and the trip's nature are calculated econometrically. A Tobit model is implemented for assessment of the effects of the trip and personal characteristics on fishing trip length. The results indicate that the choice of timing and trip duration can be reported well in terms of personal variables and observable trip. Knowing the connections is a valuable independent variable into the management of tourism/ recreational fishing and the development of models for the simulation of tourism/fishing activities.

Lee *et al.* [21] proposed an algorithm for travel time prediction, which utilises rule-based map reduce grouping of the huge scale of trajectory data. Firstly, the algorithm sets rules for grouping based on real traffic stat. And ascertains the effective classes of velocity for each part of the road. Secondly, it generates a distributed index, which is done by using a grid-based map partitioning method. Also, it possesses the capacity to deduce the query processing cost as the grid cells which includes the question region is retrieved, rather than retrieving the whole network of the road. Also, the time for processing the query can be minimised by calculating the time taken for travelling given queries for each segment in a parallel way.

The above studies provide information about the researches carried out so far in the duration of trip and travel time prediction in various modes of transport. Such studies reflect on the need to predict the duration of the trip for the development of various applications. Many techniques are used to predict the duration of the trip but the use of data mining techniques could be an efficient tool to provide satisfactory results in prediction.

3 Methodology

3.1 Linear regression

Linear regression (LR) is the simplest statistical regression method for identifying the linear link between the independent and the dependent variables. It is done by fitting a linear equation line to the observed data [22]. For fitting the model, it is utmost important to check, whether there is a connection between the variables or features of interest, which is supposed to use the numerical variable, that is the correlation coefficient. The following equation defines an LR line:

$$Y = a + bX, \tag{1}$$

where X is the independent variable whereas Y is a dependent variable. b is the slope of the line and a is the intercept (the value of y when $x = 0$). For figuring out the best fitting line, the least square errors are commonly used, that is done by deducing the addition of squares of the vertical deviation of each point from the line or the addition of squares of the residuals.

3.2 Gradient boosting machine

The gradient boosting machine (GBM) model has more benefits than the LR model that is commonly found in the existing works. It is capable of managing various independent variables (categorical, continuous etc.). Also, it involves a minimum duration for making data. Complex non-linear trip duration time relationships with independent variables that fit, as trip duration time is not required following a certain distribution. Complex non-linear driving time links with the independent variables, as the driving time does not need to obey a particular distribution. In the case of decision trees, some other independent variable values at the higher trees level decide the reaction for an independent variable. So automatically GBM models a mutual notion between independent variables [23]. Besides, it seizes sharp or subtle variation in the duration of the trip and improves predictive accuracy by boosting. The rest of the section enumerates the GBM algorithm in terms of mathematics.

Consider X as a feature set of explanatory variables and approximation function of the response variable y as $F(x)$. This method computes the function $F(x)$ as an additive expansion $F(x)$ depending on the basic function $h(x; a_m)$ [24, 25]. Equation (2) represents $F(x)$:

$$F(x) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \beta_m h(x; a_m) \tag{2}$$

where a_m is the mean of split locations and the terminal node for each splitting variable in the individual decision tree $h(x; a_m)$, β_m is determined for reducing a specified loss function $L(y, F(x)) = (y - F(x))^2$. For effective estimation, gradient boosting approach has been proposed [26]. Its algorithm can be summed up as follows [27]:

Step 1: Initialise $F_0(x)$ to be a constant, $F_0(x) = \arg\min_{\delta} \sum_{i=1}^{N} L(yi, \beta)$*Step 2:* For $m = 1$ to M

For $i = 1, 2, \ldots, N$ compute the negative gradient

$$\tilde{y}_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right] F(x) = Fm - 1(x)$$

Fit a regression tree $h(x; a_m)$ to the target $\tilde{y}_{im}$ Compute a gradient descent step size as $\beta_m = \arg\min_{\delta} \sum_{i-1}^{N} L(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m))$ Upgrade the model as follows: $F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m)$ *Step 3:* Reveal the results of the final model $F(x) = F_M(x)$ To get over the over-fitting issue [4], learning rate (or shrinkage), is used for scaling each base tree model's contribution by presenting a factor of ξ $(0 < \xi \leq 1)$ as given in the following equation:

$$F_{m(x)} = F_{m-1}(x) + \zeta \,.\, \beta_m h(x; a_m), \text{where} \quad 0 < \zeta \leq 1 \tag{3}$$

The lesser the shrinkage value, the lower is the loss function. Nonetheless, it involves the addition of many trees to the model. So, it reveals the existence of an interchange between the rate of learning and the number of trees. The other significant factor for the GBM method is the complexity of the tree which brings up to the number of splits that fits every decision tree. For seizing the complex interactions between variables, the complexity of the tree must be increased. Overall, the best function of the model relies on a collective choice of a number of trees, learning rate, and complexity of the tree.

3.3 K nearest neighbour

k-nearest neighbours (KNNs), a non-parametric learning algorithm is employed for regression or classification [28]. The input in both cases holds *k* closest examples of training in the feature space. The result varies by *k*-NN depending on its utilisation for regression or classification.

In the case of *k*NN *classification*, the result is a member of the class. An object is categorised by a relative majority of its neighbours, the object attributed to a class which is usual among its nearest *k* neighbours (*k* is a positive integer).

In the case of *k*NN *regression*, the result is assumed to be the property value of the object. The output value is computed by taking average values of *k* nearest neighbours.

It is considered as non-parametric because of the absence of explicit mapping relationship between independent and dependent variables. The proximity of neighbouring independent variables and the respective dependent variables are used to render the ultimate scores of the test data. *K* the parameter which specifies the considered number of neighbouring observations must be observed prior to the score. *k*NN is considered to be a simple learning algorithm and when it is practically applied, the performance is considered to be acceptable.

© The Institution of Engineering and Technology 2020

3.4 Random forest

The Random forests (RFs) suggested by Breiman [29] are non-parametric and tree-based ensemble techniques [30, 31]. Unlike traditional statistical methods, RFs contain most easy-to-interpret decision trees models instead of parametric models. A more comprehensive prediction model can be concluded by integrating the analysis results of decision trees models. The main objective of this research is to predict the Seoul bike trip duration by addressing regression mode. The RFs regression is a non-parametric regression method consisting of a set of K trees $\{T_1(X), T_2(X), \ldots, T_K(X)\}$, where $\boldsymbol{X} = \{x_1, x_2, \ldots, x\_\}$, is a dimension independent vector that forms a forest. The ensemble produces P dependent variables corresponding to each tree Y_p $(p = 1, 2, \ldots, P)$. The ultimate result is achieved by computing the mean of all tree predictions. The training process is adopted as follows:

(a) Extract a sample bootstrap from the accessible dataset, i.e. a sample chosen randomly with replacement.
(b) Employ the following adjustments to the bootstrap sample and derive trees at each node. Select the best split between a randomly chosen subset of *mtry* (number of variant predictors which are tested at each node) descriptors. *mtry* serves a vital tuning factor in the RFs algorithm. The tree is grown to maximise size without pruning back.
(c) Step (b) is repeated until the number of trees (*ntree*) defined by the user are grown which are based on a bootstrap observations sample.

For regression, RFs build the number of regression trees K and average the results. Final predicted values are obtained by aggregating of the outcomes of each tree. The following equation defines the RF regression predictor, after k trees $\{T_k(x)\}$ has been grown:

$$f(x) = \left[\sum_{k=1}^{K} T_K(x)\right] / K \tag{4}$$

For each RFs regression tree construction, a fresh training set (bootstrap samples) is established to a replacement from the original training set. So, every time a regression tree is constructed using the randomised sample training from the original dataset. The out-of-bag sample is utilised for examining its exactness and given in (5)

$$\mathrm{GI}\,(t_{X(xi)}) = 1 - \sum_{j=1}^{m} (t_{X(xi),}, j)^2 \tag{5}$$

The inherent validation features improve the tree robustness of the RFs while utilising independent test results. RFs have also been shown to be a viable method of regression and classification and so utilised in this study.

4 Data preparation and exploratory analysis

The objective of this research is to find the trip duration as much as accurate for each of the rental bikes with various predictors considered. Moreover to discuss the performance of different regression models linear regression (LR), K nearest neighbours (kNN), gradient boosting machine (GBM), and random forest (RF) to predict the trip duration.

4.1 Data creation

One year data (January 2018 to December 2018) is downloaded from South Korean website Seoul Public Data Park (Open Data), where data about trips made using Seoul Bike throughout the year is available [32]. The time-span of the dataset is 365 days (12 months). The one-year data consists of 99,87,224 entries, which means nearly 10 million trips are made in one year. The fields in the dataset include rental bike number, pickup date and time, pickup station number and address, dropoff date and time, dropoff station number and address, return dock, trip duration in minutes and total distance in metres. This data does not include latitude and longitude details of pickup and dropoff stations. The data about latitude and longitude details of the rental bike stations are downloaded from the same website and merged using the station number and address. Since details about the rental stations are not updated, the trips without station details are dropped. After this step, the total number of entries is 99,74,018.

Since weather information is the most influential data contributing to the trip duration and was used in previous research studies, the weather information is also added to increase the performance of the prediction models. The weather information is downloaded from the Korean Meteorological Society [33]. An hour wise weather information is used and the weather variables used are Temperature, Precipitation, Windspeed, Humidity, Solar Radiation, Snowfall, Ground Temperature. South Korea has the condition of fine dust which affects the environment a lot. So this can also be used as an influencing variable in trip duration prediction. One hour average fine dust concentration data is also added to the data. Fine dust data contains some missing values and the missing entries are replaced with 0. Fig. 2 shows the whole process involved in data preparation and Fig. 3 presents the whole system flow followed in this research.

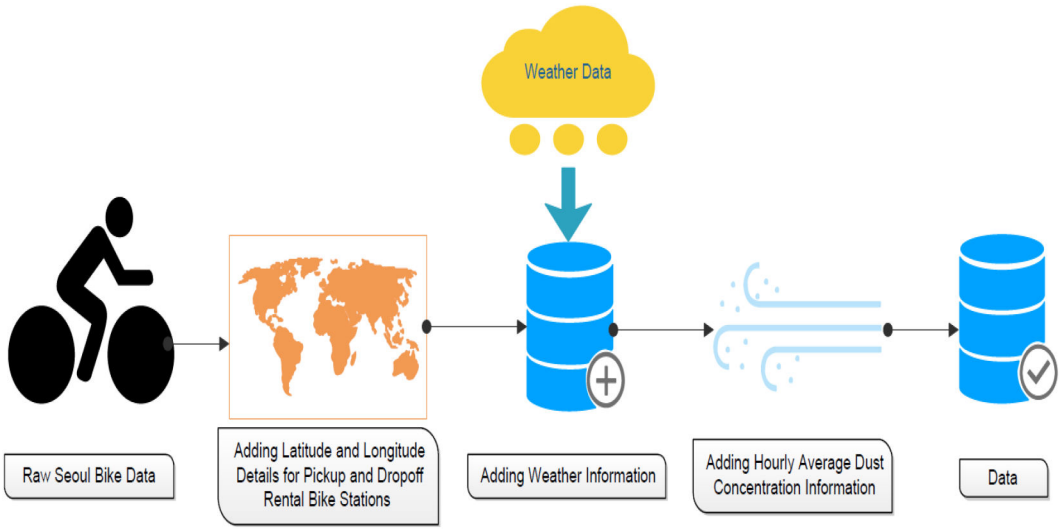

Fig. 2 *Data creation process*

© The Institution of Engineering and Technology 2020

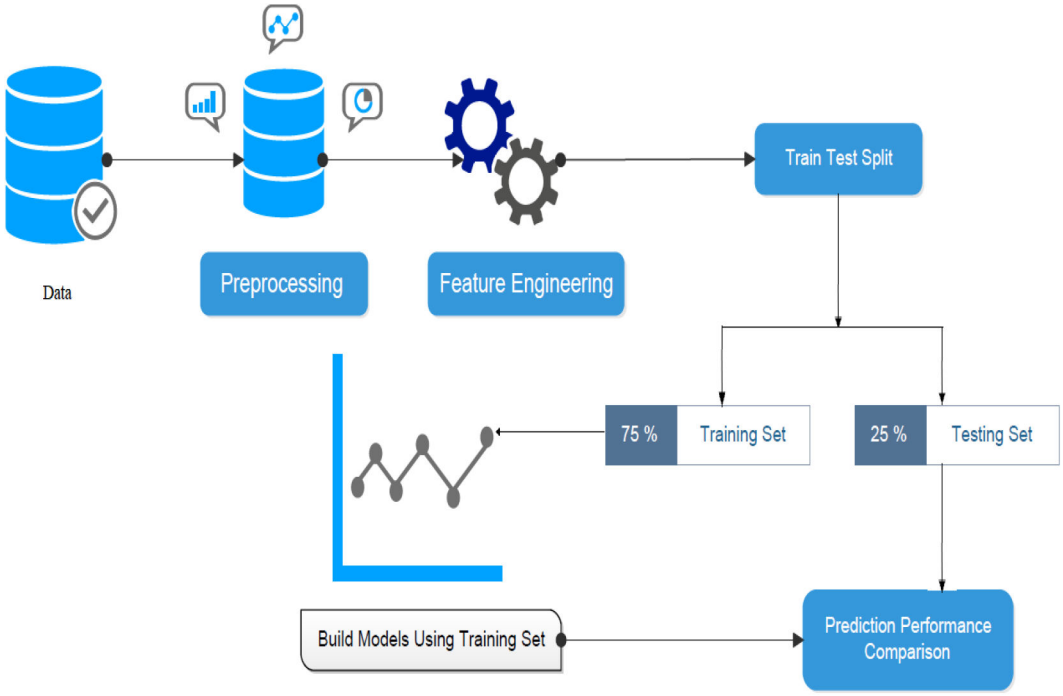

Fig. 3 *System flow*

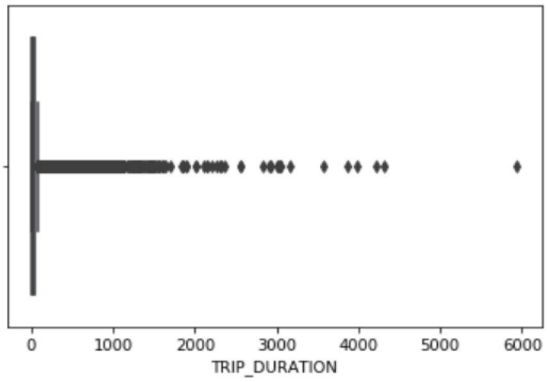

Fig. 4 *Boxplot of trip duration before outliers removal*

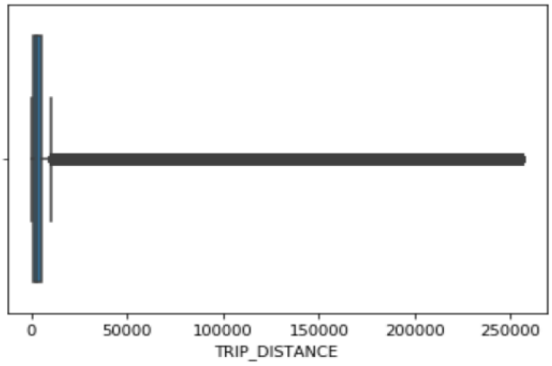

Fig. 5 *Boxplot of trip distance before outliers removal*

4.2 Data preprocessing

To remove noise in the data and to make the prediction algorithms perform better some of the basic pre-processing steps in data mining techniques are done. Dropping 0 entries in Trip duration and Trip distance is the first step. After this step number of entries came down to 98,30,314. Next step is to remove outliers in Trip duration and Trip distance field. Fig. 4 presents the boxplot of trip duration and Fig. 5 shows the boxplot of the distance field. It can be noted from the figures that there are a lot of outliers in both fields including a maximum of 5940 in trip duration and a maximum of 2,55,990 in the trip distance. Removing these outliers

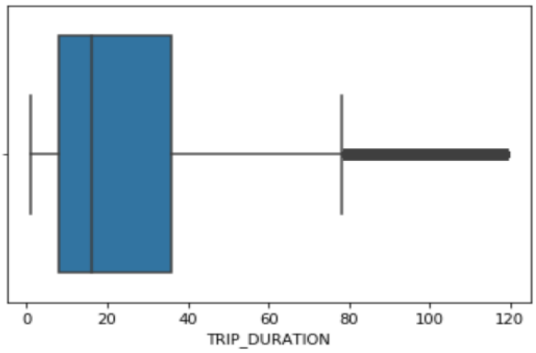

Fig. 6 *Boxplot of trip duration after outliers removal*

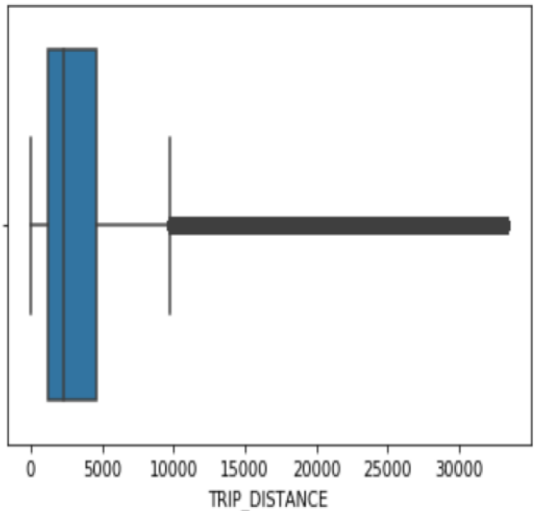

Fig. 7 *Boxplot of trip distance after outliers removal*

improves prediction performance. So data lies outside 3 standard deviations from the mean value is excluded in both trip duration and trip distance.

Fig. 6 shows the boxplot of trip duration after outliers removal and Fig. 7 shows the box plot of trip distance after outliers removal. In the box plot, the median is represented with a black line inside the blue rectangle. The thick line above the upper

© The Institution of Engineering and Technology 2020

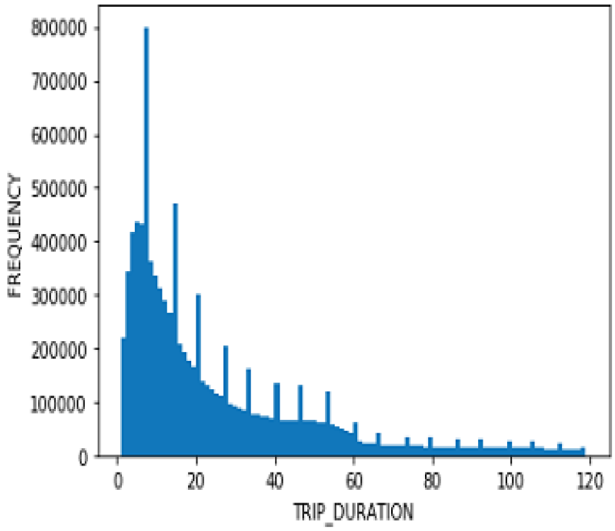

Fig. 8 *Histogram plot for trip duration*

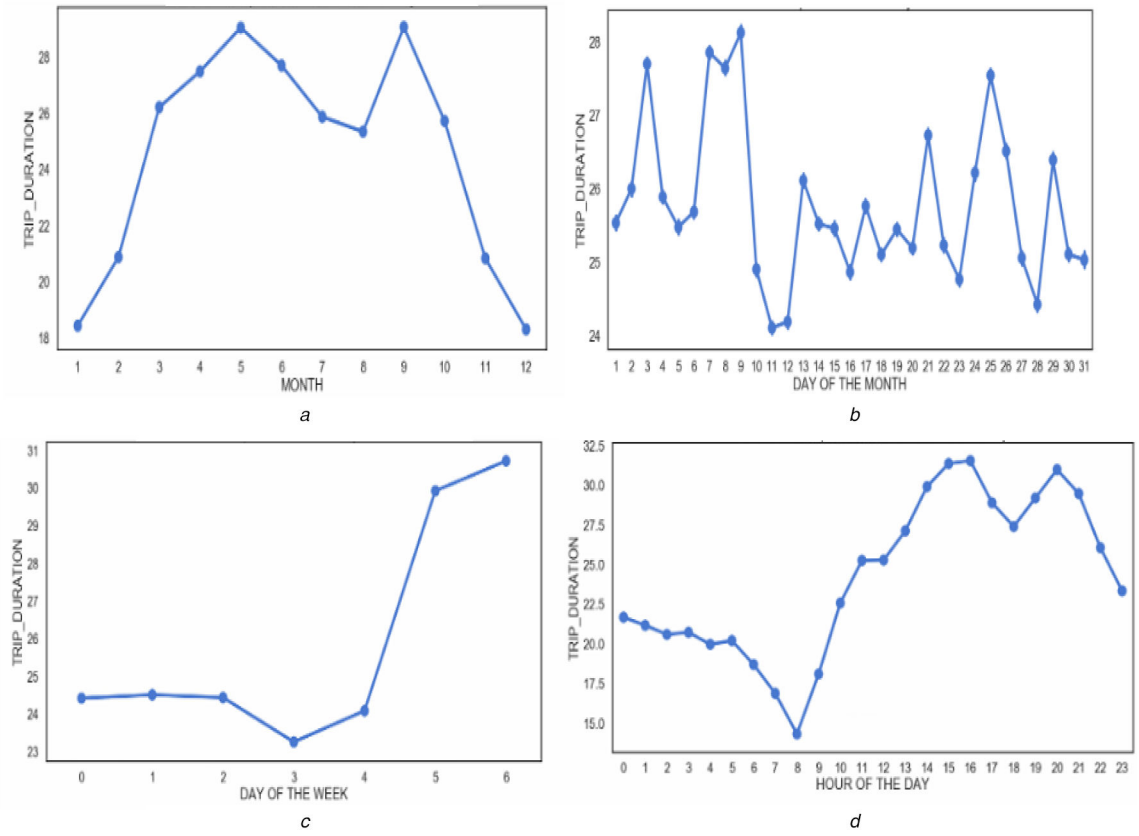

Fig. 9 *Average trip duration across*
(a) Months, ***(b)*** Day of the month, ***(c)*** Day of the week, ***(d)*** Hour of the day

whisker represents the outliers and after the removal of the outliers, the quantity of outliers is reduced compared to Figs. 4 and 5. After outliers removal, the dataset reduced to 96,01,139 entries.

4.3 Exploratory data analysis

Exploratory data analysis (EDA) is a most common approach, which helps in summarising the main characteristics of a dataset by analysing the characteristics, commonly with visual methods. Using a statistical model is optional. However, mainly EDA helps to seek for pieces of information from the data which are beyond the hypothesis testing task or formal modelling. Fig. 8 displays a histogram and the boxplot of the data. It also shows the data distribution that possesses a long tail. It can be understood that most of the trip is between 0 and 20.

Fig. 8 shows the histogram plot for trip duration. This show that the Trip duration data is skewed. Fig. 9 presents the Average duration of the trip across months, day of the week, day of the month, and hour of the day. From the plots, it is clear that trip duration has strong time component and these dependencies are used to predict the trip duration more accurately. From Fig. 9*a* it is clear that average trip duration is very less during January, February, November and December which is winter season in South Korea. This proves that temperature affects the trip duration. Fig. 9*b* presents a plot of average trip duration across the day of the month. It can be seen from the figure that the trip duration across the day of the month is not stable and there is no clear pattern across the day of the month. Fig. 9*c* presents average trip duration across the day of the week and the trip duration is high during the weekends. From Fig. 9*d*, average trip duration is high during the hour 15, 16 and 20 which represents leisure hour with less traffic

© The Institution of Engineering and Technology 2020

Fig. 10 *Latitude and Longitude distribution*
(a) Pickup Latitude and Longitude distribution, ***(b)*** Dropoff Latitude and Longitude distribution

and rush and less during the hours 8 and 18 which represent the morning and evening peak hours in Seoul city, respectively.

Fig. 10*a* shows the Pickup latitude and longitude distribution and Fig. 10*b* shows Dropoff latitude and longitude distribution. It can be seen that Pickup and dropoff latitude is in the range of (126.60–127.80) and pickup and dropoff longitude is in the range of (37.45–37.70). This proves that all the trips are executed within the Seoul city range and no potential outliers associated with latitude and longitude.

4.4 Feature engineering

Next step is to create some additional features from the date/time variable to make the machine learning algorithms work more efficiently. This process of creating additional features from the existing data by using domain knowledge is known as feature engineering. From the Pickup date and time variable, variables such as pickup month, day, hour, minute and day of the week are computed. From Dropoff date and time variable, variables such as dropoff month, day, hour, minute and day of the week are extracted. Although the trip distance is already present in the data, the distance between pickup station and drop off station is computed using haversine function, by using pickup latitude and longitude details and drop off latitude and longitude details. Table 1 presents the list of all the features or variables or parameters and its corresponding Abbreviation, Type (Continuous or categorical) and Measurement.

After creating the final dataset, the next step is to check whether the considered variables used for predicting the trip duration is correlated with the dependent variable. So a correlation plot is created for finding the relationship among the variables. Fig. 11 shows the pairs plot and displays the correlation values of Trip duration with Distance, PLong, PLatd, DLong, DLatd, Haversine, Temp, Precip, Wind, Humid, Snow, GroundTemp and Dust. As can be seen from the plot that the dependent variable Duration has at least least correlation value with the independent variables. This shows the trip duration variable is associated with other variables considered in this study. Positive values represent a positive correlation between the variables and negative value represents a

© The Institution of Engineering and Technology 2020

Table 1 Data variables and description

Parameters/ Features	Abbreviation	Type	Measurement
date	Date	year-month-day hour:minute:second	-Time
trip duration	Duration	Continuous	1, 2, 3, … 5940
trip distance	Distance	Continuous	1, 2, 3, … 33,290
pickup date and time	PD time	year-month-day hour:minute:second	Time
dropoff date and time	DDtime	year-month-day hour:minute:second	Time
pickup longitude	PLong	continuous	Radians
pickup latitude	PLatd	continuous	Radians
dropoff longitude	DLong	continuous	Radians
dropoff latitude	DLatd	continuous	Radians
haversine distance	Haversine	continuous	Kilometres
pickup month	Pmonth	categorical	January, February, March, … December
pickup day	Pday	categorical	1,2,3, … 31
pickup hour	Phour	categorical	0,1,2, … 23
pickup minute	Pmin	continuous	1,2,3, … 60
pickup day of the week	PDweek	categorical	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
dropoff month	Dmonth	categorical	January, February, March, … December
dropoff day	Dday	categorical	1,2,3, … 31
dropoff hour	Dhour	categorical	0,1,2, … 23
dropoff minute	Dmin	continuous	1,2,3, … 60
dropoff day of the week	DDweek	categorical	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
temperature	Temp	continuous	°C
precipitation	Precip	continuous	Mm
windspeed	Wind	continuous	m/s
humidity	Humd	continuous	%
solar radiation	Solar	continuous	MJ/m^2
snow fall	Snow	continuous	cm
ground temperature	GroundTemp	continuous	°C
1 hour average fine dust concentration	Dust	continuous	$\mu g/m^3$

negative correlation between the variables. The highest correlation value exists between Duration and Distance variables.

The combined data set is split into training and test set using train_test_split function [34]. A total of 75% of data is employed for training the models while the remaining is for testing (Table 2).

5 Evaluation indices

Multiple evaluating criteria are used for comparing the performance of regression models. The performance evaluation indices used here are: Root Mean Squared Error (RMSE), Rsquared (R^2), Median Absolute Error (MedAE) and Mean Absolute Error (MAE).

RMSE stands for the sample standard deviation of the residuals between the observed and the predicted values. Large errors can be identified using this measure and the fluctuation of model response regarding variance can be evaluated. RMSE, a scale-dependent metric outputs values having the same units of the measurement. On the other hand, R^2 is the coefficient of determination that ranges from 0 to 1, which reflects the fitting quality. A high R^2 value signifies the predicted values which perfectly fits with the observed values. Formula to compute RMSE and R^2 values are given in (6) and (7) respectively:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2} \tag{7}$$

MAE is to assess the prediction acuteness. MAE, a scale-dependent metric, effectively reveals the error in prediction by preventing the offset between negative and positive errors. The MAE can be calculated by using the followingequation:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left[Y_i - \hat{Y}_i\right]}{n} \tag{8}$$

MedAE is specifically stimulating as it is resistant to deviations. Loss is determined by calculating the median of all the absolute differences between the target and the prediction. In case of $\hat{y}$ being the value predicted out of the ith sample and y_i, being the respective true value, the MedAE figured over n samples is given as follows:

$$\mathrm{MedAE} = \mathrm{median}\left|\left(Y_i - \hat{Y}_i\right), \ldots, \left(Y_i - \hat{Y}_i\right)\right| \tag{9}$$

where Y_i is considered the actual values, $\hat{Y}_i$ is the value predicted from the models and n is the sample size and $\bar{y}$ is the sample average.

6 Results and discussion

Four regression algorithms such as LR, GBM, KNN and RF are used to predict the trip duration. Each of the regression algorithms requires the selection of the best hyperparameters to make them perform best. So it is crucial to select the optimum hyperparameters.

Since the data is large (Nearly 10 million), finding optimal hyperparameters for each of the models is a time consuming and computationally expensive. So to find the optimal hyperparameters a random search was done. For the LR model, intercept value is kept true and the model fit was done. For GBM mode, the optimal set of hyperparameters include $\alpha = 0.9$, learning rate = 0.1, maximum depth is 3, the minimum samples split is 2, and the number of estimators is 100. For KNN the number of neighbours was set to be 5. The number of estimators or number of trees is 10, minimum samples leaf is 1 and minimum samples leaf is 2 for RF.

After training each of the models with its best hyperparameters, the performance of each of the models is tested with testing set and evaluated using four metrics RMSE, R^2, MedAE and MAE. The model's performance is summarised in Table 3. RMSE, R^2, MedAE and MAE values of LR model in testing phase is 16.48, 0.56, 6.90 and 10.12, respectively. RMSE, R^2, MedAE and MAE values of GBM model in the testing phase are 12.58, 0.74, 3.75 and 7.37, respectively. RMSE, R^2, MedAE and MAE values of KNN model in testing phase is 13.93, 0.69, 2.59 and 6.83, respectively. RMSE, R^2, MedAE and MAE values of RF model in testing phase is 6.25, 0.93, 1.20 and 2.92, respectively. The model with the lowest RMSE, MedAE and MAE and highest R^2 is considered the

© The Institution of Engineering and Technology 2020

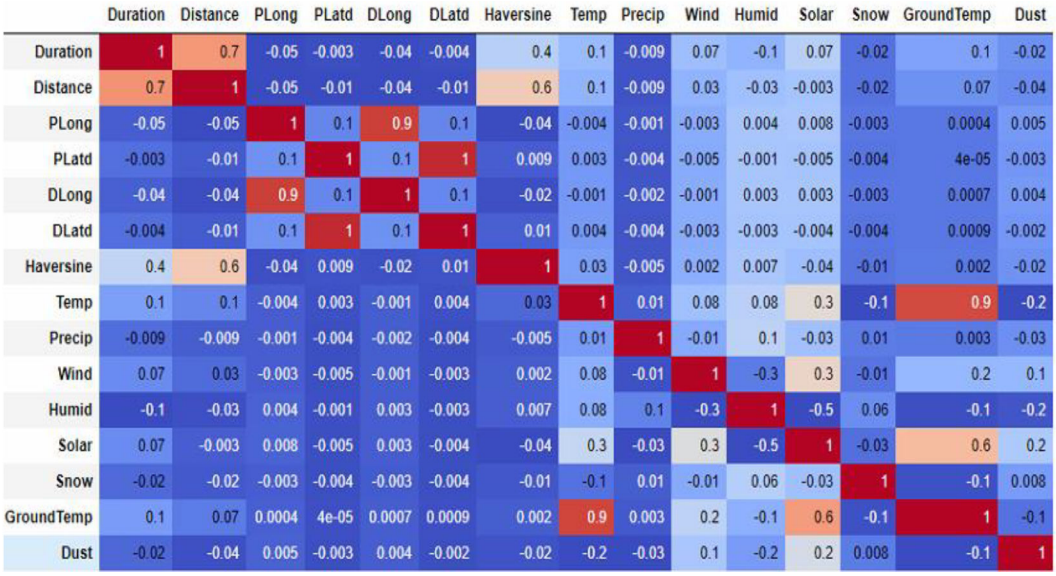

Fig. 11 *Correlation plot*

Table 2 Training and testing dataset

Dataset	Number of observations
training	72,00,854 and 24 variables
testing	24,00,285 and 24 variables

Table 3 Models performance

Training models	RMSE	R^2	MedAE	MAE	Testing RMSE	R^2	MedAE	MAE
LR	16.45	0.56	6.90	10.11	16.48	0.56	6.90	10.12
GBM	12.55	0.74	3.74	7.35	12.58	0.74	3.75	7.37
KNN	11.29	0.79	2.0	5.53	13.93	0.69	2.59	6.83
RF	2.76	0.98	0.40	1.21	6.25	0.93	1.20	2.92

optimum performing model. RF model has the best performance and LR produced the worst performance. RF yields the highest value for R^2 and lowest values for RMSE, MedAE and MAE. Performance of LR is the worst. This is because the trip duration is not linearly dependent on the independent variables. The performance of the GBM and KNN models is almost similar. There is only 1-point difference between values of RMSE, MedAE and MAE values. The performance of RF is two times higher than the performance of GBM and KNN. This shows that RF could be used as an effective tool to predict the trip duration.

7 Conclusion

A data mining approach is used to predict the drip duration using data recorded in the rental bikes and weather information is proposed. The analysis is done with Seoul Bike data. Four regression techniques LR, GBM, KNN and RF is used to predict the trip duration. This statistical data analysis shows interesting outcomes in prediction methods and also in an exploratory analysis. The experimental results prove that the RF model predicts best the trip duration with the highest R^2 and with less error rate compared to LR, GBM and KNN. RF model exhibits its proficiency in the analysis of time-series and statistical learning. However, a few works are done for the prediction of trip duration. The results indicate that RF predictor significantly outperforms other baseline predictors. It proves the applicability of RF for the prediction of trip duration. This trip duration prediction with a combination of weather data and feature engineered variables are used as an effective tool to develop future Artificial intelligence-based Transportation. The efficient trip duration prediction can also provide multiple advantages by providing various applications to users. Future work includes the use of Deep learning techniques for trip duration prediction. Since deep learning techniques are more effective in learning big data with automatic feature extraction and deliver high-quality results, deep learning could be used for trip duration prediction.

8 References

[1] Conservancy, Ocean: 'Stemming the tide: land based strategies for a plastic-free ocean', Ocean Conservancy and McKinsey Center for Business and Environment, 2015

[2] Audenhove, V., François-Joseph, O., Dauby, L., *et al.*: 'The future of urban mobility 2.0: imperatives to shape extended mobility ecosystems of tomorrow', 2014

[3] Calafiore, G.C., Portigliotti, F., Rizzo, A.: 'A network model for an urban bike-sharing system', *IFAC-PapersOnLine*, 2017, **50**, (1), pp. 15633–15638

[4] Wikipedia: 'List of bicycle-sharing systems', 2017

[5] Shaheen, S., Guzman, S., Zhang, H.: 'Bikesharing in Europe, the americas, and Asia: past, present, and future', *Transp. Res. Rec., J. Transp. Res. Board*, 2010, **2143**, p. 159167

[6] Wolpert, D., Macready, W.: 'No free lunch theorems for optimization', *IEEE Trans. Evol. Comput.*, 1997, **1**, (1), pp. 67–82

[7] Giraud-Carrier, C., Vilalta, R., Brazdil, P.: 'Introduction to the special issue on meta-learning', *Mach. Learn.*, 2004, **54**, (3), pp. 187–193

[8] Turner, S., Eisele, W., Benz, R., *et al.*: 'Travel Time Data Collection Handbook', Federal Highway Administration, Report FHWA-PL-98-035, 1998

[9] Li, Y., DimitriosGunopulos, C.L., Guibas, L.: 'Urban travel time prediction using a small number of GPS floating cars'. Proc. of the 25th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems, USA, 2017, p. 3

[10] Mridha, S., NiloyGanguly, S.B.: 'Link travel time prediction from large scale endpoint data'. Proc. of the 25th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems, USA, 2017, p. 71

[11] Miura, H.: 'A study of travel time prediction using universal kriging', *Top*, 2010, **18**, (1), pp. 257–270

[12] Kwon, J., Coifman, B., Bickel, P.: 'Day-to-day travel-time trends and travel-time prediction from loop-detector data', *Transp. Res. Rec.: J. Transp. Res. Board*, 2000, **1717**, (1), pp. 120–129

© The Institution of Engineering and Technology 2020

[13] Chien, S.I.J., Kuchipudi, C.M.: 'Dynamic travel time prediction with real-time and historic data', *J. Transp. Eng.*, 2003, **129**, (6), pp. 608–616
[14] Zhang, X., Rice, J.A.: 'Short-term travel time prediction', *Transp. Res. C: Emerg. Technol.*, 2003, **11**, (3), pp. 187–210
[15] Wu, C.H., Ho, J.M., Lee, D.T.: 'Travel-time prediction with support vector regression', *IEEE Trans. Intell. Transp. Syst.*, 2004, **5**, (4), pp. 276–281
[16] Balan, R.K., Nguyen, K.X., Jiang, L.: 'Real-time trip information service for a large taxi fleet'. Proc. of the 9th Int. Conf. on Mobile Systems, Applications, and Services, MobiSys, ACM, New York, 2011, pp. 99–112
[17] Brazdil, P., Soares, C., Costa, J.D.: 'Ranking learning algorithms: using IBL and meta-learning on accuracy and time results', *Mach. Learn.*, 2003, **50**, pp. 251–277
[18] Zarmehri, M.N., Soares, C.: 'Using metalearning for prediction of taxi trip duration using different granularity levels'. Int. Symp. on Intelligent Data Analysis, Cham, 2015, pp. 205–216
[19] Handley, S., Langley, P., Rauscher, F.A.: 'Learning to predict the duration of an automobile trip'. KDD, New York, NY, USA, 1998, pp. 219–223
[20] Hailu, A., Gao, L.: 'Research note: recreational trip timing and duration prediction', *Tour. Econ.*, 2012, **18**, (1), pp. 243–251
[21] Lee, H., Hong, S., Kim, H.J., *et al.*: 'A travel time prediction algorithm using rule-based classification on MapReduce'. Database and Expert Systems Applications, Cham, 2015, pp. 440–452
[22] Neter, J., Isserman, W., Kutner, M.H.: 'Applied linear regression models', 1989
[23] Elith, J., Leathwick, J.R., Hastie, T.: 'A working guide to boosted regression trees', *J. Anim. Ecol.*, 2008, **77**, (4), pp. 802–813
[24] De'Ath, G.: 'Boosted trees for ecological modeling and prediction', *Ecology*, 2007, **88**, (1), pp. 243–251
[25] Saha, D., Alluri, P., Gan, A.: 'Prioritizing highway safety manual's crash prediction variables using boosted regression trees', *Accident Anal. Prev.*, 2015, **79**, pp. 133–144
[26] Friedman, J.H.: 'Greedy function approximation: a gradient boosting machine', *Ann. Stat.*, 2001, **29**, pp. 1189–1232
[27] Ding, C., Wu, X., Yu, G., *et al.*: 'A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data', *Transp. Res. C, Emerg. Technol.*, 2016, **72**, pp. 225–238
[28] Altman, N.S.: 'An introduction to kernel and nearest-neighbor nonparametric regression', *Am. Stat.*, 1992, **46**, (3), pp. 175–185
[29] Breiman, L.: 'Random forests', *Mach. Learn.*, 2001, **45**, pp. 5–32
[30] Adusumilli, S., Bhatt, D., Wang, H., *et al.*: 'A low-cost INS/GPS integrationmethodology based on random forest regression', *Expert Syst. Appl.*, 2013, **40**, pp. 4653–4659
[31] Zhou, J., Shi, X.Z., Du, K., *et al.*: 'Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel', *Int. J. Geomech.*, 2017, **17**, p. 04016129
[32] 'SEOUL OPEN DATA', http://data.seoul.go.kr/
[33] 'KOREA METEOROLOGICAL ADMINISTRATION', https://www.kma.go.kr/eng/index.jsp
[34] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.*: 'Scikit-learn: machine learning in python', *J. Mach. Learn. Res.*, 2011, **12**, pp. 2825–2830

© The Institution of Engineering and Technology 2020