A Forcast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression

YouLi Feng , ShanShan Wang
Inner Mongolia university
College of Computer Science and Technology, Inner Mongolia university
Inner Mongolia , China
951092795@qq.com, 2601348947@qq.com

***Abstract*—Bike sharing system is a ways of renting bicycles; bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from a one pick up location and combine with their as-need, customer returns bike to the place, which they would prefer to return. This paper is asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bike share program in Washington, D.C. Firstly, the multiple linear regression model was established by the conventional method, Multiple linear regression equation was obtained by using SPSS software, After comparing the data with the real value, it is indicated that the multiple linear regression model is less accurate. After analysis, we find that the data includes the dummy variables such as the time and the season. Hence this paper proposes a random forest model and a GBM packet to improve the decision tree. The results and the accuracy of multiple regression analysis are greatly improved when use of random forest model to predict the demand for bicycle rental.**

Keywords—Bike sharing system; Multiple linear regression analysis; Random forest; GBM

I. INTRODUCTION

With the rapid development of the global low-carbon movement and the increasing in the number of private cars, city planning to expand the number of vehicles is increasing rapidly, traffic congestion and environmental pollution problems are becoming more and more serious, the public bicycle rental system is as the new public transport system to flourish in the world, the public bicycle can not only in the short distance travel play the flexible and efficient advantages, but also can effectively extend the scope of public transport service. Real time to master the number of bicycle rental bicycles to guide the needs of the public, is conducive to the formation and planning departments to develop rental policy.

At present, the public bicycle rental system is mainly for the rental demand forecast and the scheduling of the lease point [1] From the research in this area is relatively small, the main reasonn because of the complex needs of public bicycles, the impact of public bicycle rental demand forecast has many factors, it is hard to consider all them together. Now, in terms of public bike rental system, there are mainly Southeast University, such as Lu Qiang[2] and Li Yanhong, Beijing Jiaotong University [3], respectively, from the bus and taxi OD data on the residents of the travel rules of the study. According to the highway toll station measured OD matrix data which is based on time slice of the traffic flow for the estimation and prediction, Camus [4] is to obtain the travel law of people. WuYao, the professor of Chang'an University, is based on multinomial logit model to forecast the demand of urban public bicycle rental [5]. However, there are few studies on the impact of weather factors on the number of public bicycle rental until now.

In this paper, we provide a forecast for the rental demand of bicycles in the capital of Washington, which is based on the historical data of bicycle rental data. According to the characteristics of the data, we use the method of multiple linear regression analysis and random forest two methods to forecast the bicycle rental demand.

II. BASED ON MULTIPLE LINEAR REGRESSION MODEL TO FORECAST RENTAL DEMAND

A. Data introduction

Bike sharing system is a ways of renting bicycles, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from a one pick up location and combine with their as-need, customer returns bike to the place, which they would prefer to return. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Therefore, the bike sharing system is functioned as a sensor network, which can be used for studying mobility in a city.

You are provided hourly rental data spanning two years, The data mainly includes the following factors:

Date time - hourly date+ timestamp

season - 1 = spring, 2 = summer, 3 = fall, 4 = winter

978-1-5090-5507-4/17/$31.00 ©2017 IEEE
ICIS 2017, May 24-26, 2017, Wuhan, China

weather
1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
temp - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
count - number of total rentals

weather ,temp ,atemp, humidity ,windspeed and count are numerical variable, Datetime,season and weather are discrete and discontinuous variable

B. Multiple linear regression model

This article is mainly to the bicycle rental demand forecast, for prediction problem, the conventional method is multivariate linear regression analysis model .First of all, we use the multiple linear regression model [6-8]. The basic mathematical model of multiple linear regression analysis.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (1)$$

According to the seven factors which affected the bike rental dem and (of which the weather and season are qualitative variable, its value is discrete and continuous, and they must be in the form of dummy variable into the model, while the others are common nu meric variables). Bicycle rental's demanding forecast model can be established as follows:

$$\text{count} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7 + \varepsilon, \quad (2)$$

Where,β_0 to β_7 is unknown parameters in the model, known as the regression coefficient, ε as the random error term. Count as the dependent variable, table bike rental demand forecast. X_1 To X_7 is the independent variable, which is the factors that influence the bike rental demand respectively called: weather, temp, attempt, humidity, wind speed, date time and the season.The regression model above has been established, and come to the next we will introduced the data to test required by the multiple regression analysis normality and linear relationship between the two premise condition is satisfied .We will import the SPSS work area, calculate the descriptive statistics of the continuous variables as a result, shown as below table 1 and table 2.

TABLE I. DESCRIPTIVE STATISTICS RESULTS OF CONTINUOUS VARIABLE FACTORS

	quantity	*min*	*max*	*Standard Deviation*	*skewness*
temp	5737	0.82	41	8.16	0.113
atemp	5737	0.76	45.46	8.87	0.005
humidity	5737	0	100	20.07	-0.105
windspeed	5737	0	57	8.27	0.576

TABLE II. THE DEPENDENT VARIABLE OF CONTINUOUS VARIABLE DESCRIPTIVE STATISTICAL RESULTS

	Quantity	*min*	*max*	*Standard Deviation*	*skewness*
count	5737	1	968	17.83	0.876

We use the Matlab [9] for continuous variables to the linear relationship between the drawings and get the figure 1、2、3、4. According to the Washington public bicycle data collation. Shown as in Figures 1, 2, 3 and 4, we can find out that these factors have a strong linear relationship with the count.

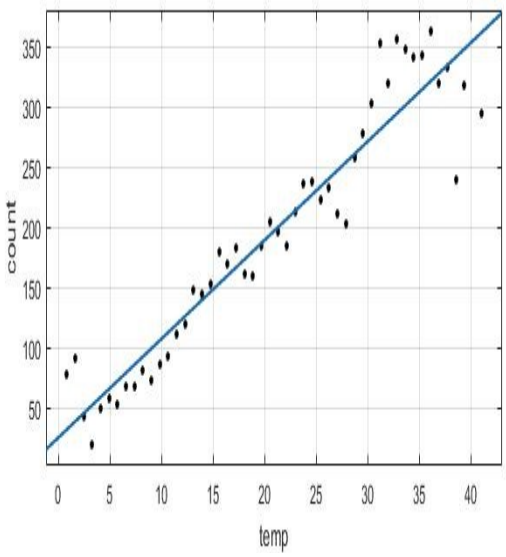

Fig.1 The relationship between the temp and the count

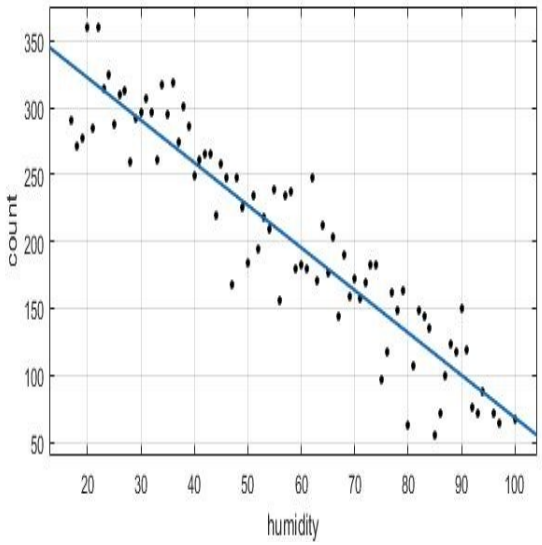

Fig.2 The relationship between the humidity and the count

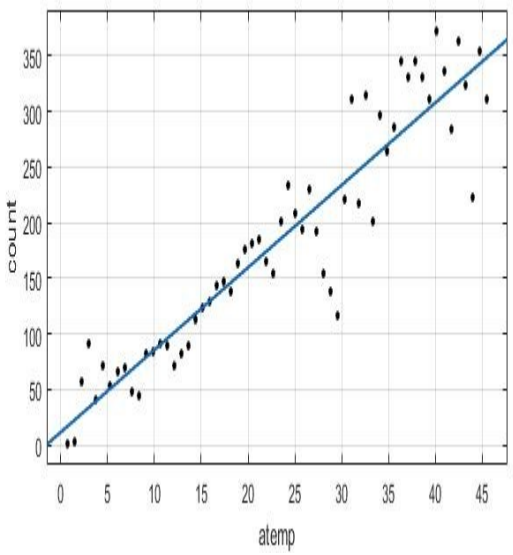

Fig.3 The relationship between the humidity and the count

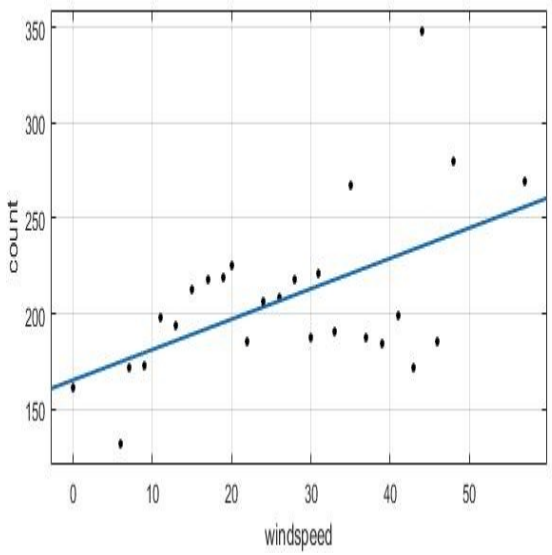

Fig.4 The relationship between the windspeed and the count

According to the above analysis, using SPSS for multivariate linear regression analysis[10], multiple linear regression equation is obtained for,

$$\text{count} = 16.143 + 7.504X_1 + 10.20X_2 - 5.934X_3 - 10.87X_4 + 17.018X_5 - 1.761X_6 + 1.008X_7 \quad (3)$$

Linear regression equation, we see the Model Summary, in the table 3, there is the adjusted R square is 0.327, low value, show the fit of the equation more bad, in the "one-way Anova, satisfy the F test, Sig. 0.00 is less than 0.005 with significant.

TABLE III. THE MODEL SUMMARY

Model	*R*	*R Square*	*Adjusted R Square*	*Sth.Error of the Estimate*
1	$.572^a$	.327	.327	146.32

According to the multivariate linear regression equation, we get through this equation, we forecast test set data, and the prediction accuracy is only 50%. The result indicated that, although it is not bad, the linear relationship between the various factors the fit of the overall is not good, the results of the multivariate linear regression model bike rental demand forecasting is not ideal, largely because of the season, the weather are contained in such dummy variable factors, the regression analysis of the impact on the accuracy of the result of the forecast. Bike rental demand forecasting, therefore, is the conventional method of multivariate linear regression model but is not ideal.

III. BASED ON RANDOM FOREST BICYCLE RENTAL DEMAND FORECASTING MODEL

By the conventional method of multivariate linear regression in front of bicycle rental demand forecasting model, we found that the conventional method is not suitable for bicycle rental demand forecast .Through again to view the data, we found that the factors contained in season, the weather is such a dummy variable factors, such as season, is 1, 2, 3, 4, such expressions, makes linear regression analysis is not accurate, according to the characters of such data, let me think about the method of random forests. So this paper proposes a bicycle rental demand forecasting model based on random forest

A. Random forests model

Random forest classification [11] by random vector grow into "tree", every tree growth without complete pruning .And at the time of spanning tree, each node variables are only a small number of variables in a randomly selected. Namely in the use of a variable (column) and data (rows) conducted on the use of randomization. Through this way of randomly generated a lot of trees were used for classification and regression analysis, which is so called "random forest". Every tree in the forest depend on a random vector, vector in the forest are all independent identically distributed. The final decision tree is based on random vector potential tree "vote" on generated, namely the random forest classification of choice with the most votes .If the purpose is to

return, by averaging the results of these tree mean value of the dependent variable.

B. *The construction of random forest model*

Because of the random forest is not the decision tree pruning, a kind of typical single classifier, the first step of training set is to recursive analysis, generate a shape such as inverted tree structure; The second step analysis of the tree from the root node to leaf node path, produce a series of rules; Finally, according to these rules, classification or projections for new data. The following is the structure of the random forest model process:

(1) n samples were selected from the sample set with random sampling;

(2) k features are randomly selected from all the features, and the decision tree is constructed by using these features;

(3) repeat the above two steps m times, generate m decision tree model, the formation of random forest;

(4) For new data, after each tree decision, finally to make predictions;

The random forest construction and prediction process are as follows.

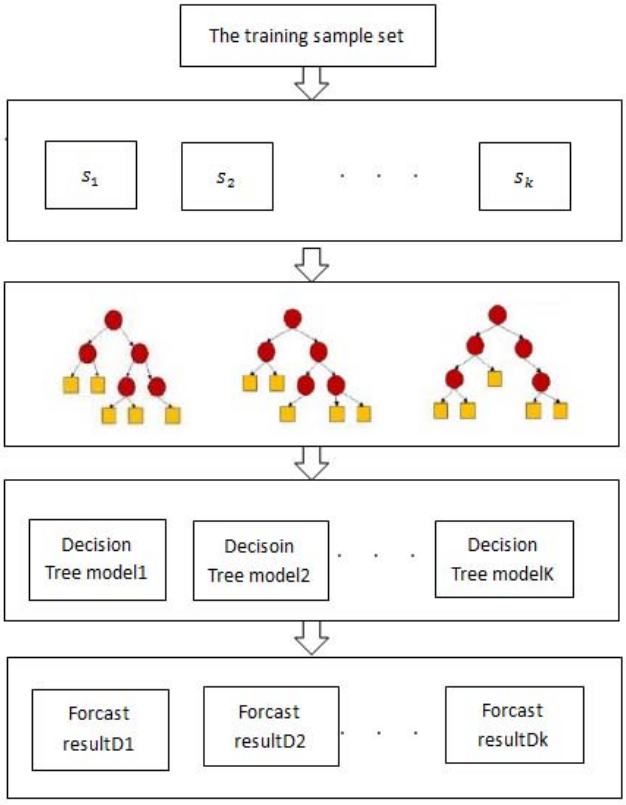

Fig.5 The random forest construction and prediction process

C. *GBM improving the capacity of decision tree in the random forest*

At the time of decision tree structure with random forests, in order to improve the efficiency, we use the GBM package [12] to improving the capacity of decision tree, every loss function model was established in the previous model of gradient descent direction .Loss function describes the unreliable degree of the model, the greater the loss function, explain the easier model error (in fact, there is a variance and deviation balanced problem, but it assumes that the greater the loss function, model, the more error prone).If our model can be reduced to keep the loss function, shows our model constantly improved, and the best way is to make the loss function in the Gradient in the direction of the up and down.

In GBM package, important parameter Settings are as follows:

- distribution
- n.trees
- shrinkage
- bag.fraction
- interaction.depth

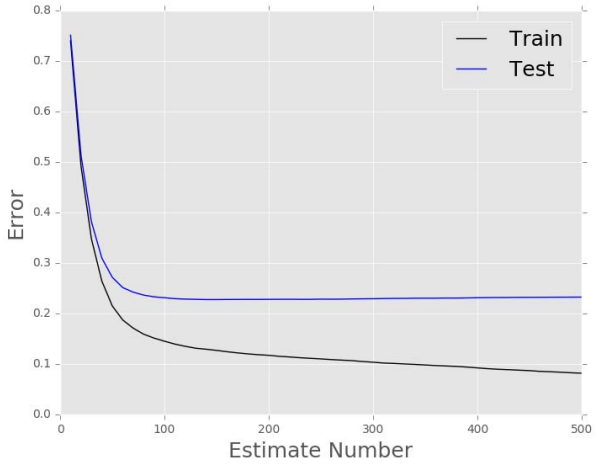

Fig.6 error vs number of estimator

Distribute that we choose the gaussian distribution ,because of forecasting the problem, shrinkage is as small as possible, but the shrinkage is too small, then the number of iterations need to increase in order to achieve the optimal model, the required time and the corresponding increase in computing resources.so We are the shrinkage parameter is 0.005, while n.trees 5000.

Through our random forests to data modeling and GBM package improving the capacity of decision tree, get the model, we forecast bicycle rental demand, get above, as shown in figure 6, we see that the random forest forecast with the increase of the number of iterations, lower error rates at 10%. We found that in

the random forest effectively solved the problems in the multiple linear regression, the random forest algorithm, in this season, the weather, time, the factors can effectively solve the random forests, forests that immediately increases the prediction accuracy, through random forest prediction for the number of car rental, with 80% accuracy rate, compared with multiple linear regression analysis, the accuracy has been greatly improved.

IV. CONCLUSIONS

In this paper, first of all, the above bicycle rental demand forecast is the conventional multiple linear regression model to forecast, the prediction accuracy is too low, although a good linear relationship between factors, also accord with the normal distribution of factors, however the characteristics of some factors, makes the result error is very high. Hence this article's bike rental demand forecast, the conventional multiple linear regression model is not applicable. According to this paper proposes a bike rental demand forecasting model based on random forest, with GBM package to improving the capacity of decision tree in the process of random forests, random decision tree is built into the forest, forest have multiple random forest model generalization ability is strong, and at the time of training, between the tree and the tree were independent of each other, and without losing accuracy. Final result accuracy greatly improved with 82% accuracy rate.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under No.71461023.

REFERENCES

[1] JIAO, Yuntai, LI, Wenquan, FENG, Peiyu, DING, Ran. A Scheduling Demand Model for Public Bicycle Rental Station [J]. Transportation and information security, 2014, 32(4): 8-13

[2] LU, FangQiang, Chen, XueWu, HuXiaoJian. Characteristic Research of Resident's Bus Trip Based on Bus OD Data[N]. Journal of Transportation Engineering and Information, 2010(2).

[3] LI, YanHong, Yuan, ZhenZhou. Analysion Trips Characteristic of Taxi in Suzhou Based on OD Data[N]. Journal of Transportation Engineering and Information, 2007(5).

[4] Camus R,Cantarella G E,Inaudi D,Real-time estimation and prediction of origin-destination matrices per time slince[J].International journal of Forecasting,1997,13(1):13-19

[5] Qian, Jin. Forecast and Analysis of the Demand for the Lease of City Public Bicycle's Rental Station[D]. Chang'an University: , 2015..

[6] Wang, Huiwen, Meng, Jie. Multiple linear regression prediction modeling method[J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, (4):

[7] LinBin. Multiple linear regression analysis and its applicatio[N]. CHINA SCIENCE AND TECHNOLOGY INFORMATION May, 2010(9)

[8] LIJun. Multiple linear regression method based on factor analysis and its application in stock prediction[D]. Nanjing University, 2014.

[9] Xiao-han Guan, Meng-meng Zhang，Yong Zheng，《Matlab Simulation in Signals & Systems——Using Matlab at different levels》，College of Electromechanical Engineering North China University of Technology，Beijing, China，2009

[10] Wei, Zhijing, liu, XiYu, Zhao, QingZhen. The Analysis Based on Statistic Software SPSS and Multiple Linear Regression Analysis[N]. Information technology and information, 2005(1).

[11] Breiman L. Random forests. Machine learning，2001，45 (1) :5-32

[12] https://www.52ml.net/10145.html