Analysis on the Growth of Shared Bike Users Based on Random Forest Model

Kewei Jiang [1, a], Chuanjin Jiang [1, b], Wenjun Hou [1, c] and Yuheng Mo[1, d]

[1]Shanghai Business School, Shanghai, China;

[a] jcj@vip.163.com, [b] jiang@sbs.edu.cn, [c] hwjsgcg22@outlook.com, [d] mo13773376962@163.com

Abstract. By analyzing the data set of hourly rental of shared bikes in Washington, D. C., this paper explores how to achieve the growth of shared bike users based on the methods of data mining and visual exploration. In this paper, machine learning models such as ridge regression, lasso regression, support vector machine regression and random forest regression are mainly selected to predict the needs of shared bike users, and then the random forest regression is verified as the optimal model. The result of this article explores the reasonable scheduling of auxiliary resources in the shared bike industry, improves the utilization rate of bicycle resources.

Keywords: Correlation Analysis; Visualization; Machine Learning; Random Forest Regression; Demand forecast of shared bike users.

1. Introduction

Shared bike is regarded as a low-carbon, environmentally friendly, healthy and economical way to travel. In 2007, a French company named JCDecaux launched the city bicycle rental service, which marked the embryonic form of shared bike. In 2008, Washington, D.C. launched the first urban bike-sharing system in the United States-Capital Bikeshare. Developed by Alta Bicycle Share, an American company, the system originally had 1,000 vehicles and 100 rental points, and now it has grown to 4,400 vehicles and 580 rental points, making it one of the largest bike-sharing systems in the United States. The successful promotion and operation of Capital Bikeshare provides a template and experience for the development of shared bike in other American cities. Therefore, in our research, we focus on the eastern part of the United States with Washington, D.C. as a reference.

There are also some problems in the development of shared bike, which hinder the healthy development of urban traffic and affect the appearance of the city:

- Parked chaos. According to the survey, many users did not park their vehicles according to the regulations after using the shared bike, which seriously affected the normal traffic and the beauty of the city.
- Theft and destruction. Shared bike is faced with the problems of being stolen and destroyed at will, which has brought considerable economic losses to shared bike enterprises.
- Overdelivery. Due to the fierce competition among shared bike enterprises in the market, some enterprises put a mass of vehicles into cities in order to expand their market share, resulting in urban traffic congestion and environmental pollution.
- Tidal effect of shared bike: There is a large demand for bicycles during the peak period, and there is a shortage of bicycles; During the trough, the supply of bicycles was large, and there was a surplus of bicycles.

In addition to the fierce market competition, the above problems in the development process of shared bike are also related to the inaccurate forecast of bicycle demand by shared bike Company and unreasonable resource allocation. Currently, the disorderly parking in shared bike is still a passive response. Therefore, it is of great significance to actively explore, analyze and predict the needs of users in shared bike for optimizing the operation and management of shared bike.

2. Research Status

2.1 Demand of shared bike users

Ryerson M, Cherry C et al. pointed out that the needs of users in shared bike are mainly affected by factors such as distance, temperature, precipitation and air quality, while the demographic characteristics of users themselves (including income, gender, occupation, etc.) will not significantly affect the use needs of public bicycles [1].

Mattson J et al. pointed out that climatic conditions such as temperature, wind power and precipitation are the main factors affecting the demand of shared bike [2]; Faghih-Imani A and Eluru N et al. believed that time variable is also an important factor affecting the demand of shared bike, that is, different time periods of the day, whether it is a working day, whether it is in peak hours, etc[3].

2.2 Shared bike demand forecasting method

Einav L and Levin J pointed out that the rapid development of information technology has gradually improved the availability of large-scale operation management data and private sector data, which has brought new opportunities and challenges to economic research [4]. Therefore, we need to explore some new methods.

Regression models are widely used in prediction-related problems. Its main method is to construct independent variables and dependent variables first, and then use the least square method to find the correlation between them, and then carry out hypothesis test on correlation coefficient and get regression results. Jiang Yueyao et al. thinks that the traditional regression model is simple and easy to explain, but the regression model may not be able to accurately measure the correlation between variables due to multiple collinearity[6]. In addition, the regression model can only explain the correlation but not the causal relationship between variables. Mullainathan S and Spiess J found that the estimation effect of machine learning methods such as random forest is obviously better than that of traditional linear least squares regression (OLS) [5].

2.3 Random forest regression model

Random forest regression is a regression model based on random forest algorithm. It adopts decision tree as the basic classifier, establishes multiple decision trees by randomly selecting features and samples, and then synthesizes the prediction results of multiple decision trees by ensemble learning method to get the final regression result.

Random forest regression can be applied in many fields, such as finance, medical treatment, industry, etc. It can be used to predict stock price, disease risk, production quality and so on. Random forest regression has the following characteristics:

- It can deal with high-dimensional data and a mass of samples, avoiding the problem of over-fitting;
- It can automatically select important features and reduce the workload of feature selection;
- It can deal with nonlinear relations and is suitable for nonlinear regression problems;
- It can detect outliers and noise data, and improve the stability of the model.

3. Empirical Analysis

3.1 Introduction of data sets

This empirical analysis uses the hourly rental data of bike-sharing program in Washington, D.C. from 2019 to 2020, and selects 17,379 samples and some fields for analysis.

Table 1. Description of Field Meaning

Variable name	Explanation of specific meaning
instant	Record index
dteday	Date
season	Season (1: Spring; 2: Summer; 3: Autumn; 4: Winter)
yr	Year (0: 2019; 1: 2020)
mnth	Month (values 1 ~ 12)
hr	Hours (values 0 ~ 23)
holiday	Whether it is a holiday or not
weekday	Day of the week (values 0 ~ 6)
workingday	Working days (1: neither weekend nor holiday; Otherwise 0)
weathersit	1: Sunny and cloudy 2: Fog and cloudy days 3: Light snow and light rain 4: Heavy rain, hail, heavy snow and heavy fog
temp	Standardized temperature, in degrees Celsius; The values are (t-t_min)/(t_max-t_min), t_min=-8, t_max=39 (in the hourly range only)
atemp	Normal somatosensory temperature, in degrees Celsius; The values are (t-t_min)/(t_max-t_min), t_min=-16, t_max=50 (in the hourly range only)
hum	Standardized humidity; Max. 100
windspeed	Normalized wind speed data; Max. 67
casual	Number of unregistered users
registered	Number of registered users
cnt	Total number of users in shared bike (cnt=casual+registered

3.2 Data preprocessing

By looking at the data dimensions, it is found that there are 17379 samples and 17 variables. By looking at the missing data, it is found that there is no missing value in the data. Convert the corresponding numbers in season field and weather field into corresponding English for subsequent visual exploration.

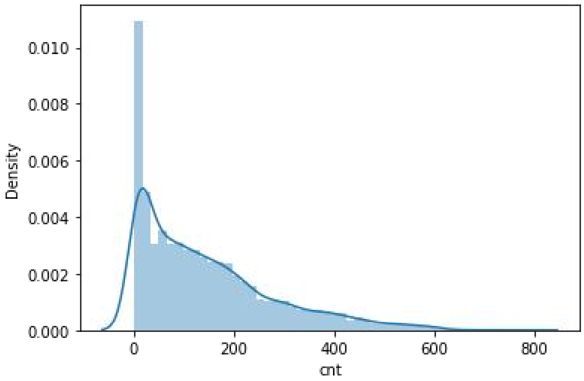

Figure1. Distribution of users

3.3 Data analysis

3.3.1 Correlation analysis

Calculation of correlation coefficient between two characteristic pairs.

Table 2. Correlation coefficients

	cnt	temp	atemp	casual	registered	hum	windspeed	holiday	workingday
cnt	1.000	0.400	0.400	0.690	0.970	-0.320	0.093	-0.031	0.030
temp	0.400	1.000	0.990	0.460	0.340	-0.070	-0.023	-0.170	0.055
atemp	0.400	0.990	1.000	0.450	0.330	-0.052	-0.062	-0.031	0.055
casual	0.690	0.460	0.450	1.000	0.510	-0.350	0.090	0.032	-0.300
registered	0.970	0.340	0.330	0.510	1.000	-0.270	0.082	-0.047	0.130
hum	-0.320	-0.070	-0.052	-0.350	-0.270	1.000	-0.290	-0.011	0.016
windspeed	0.093	-0.023	-0.062	0.090	0.082	-0.290	1.000	0.004	-0.012
holiday	-0.031	-0.170	-0.031	0.032	-0.047	-0.011	0.004	1.000	-0.250
workingday	0.030	0.055	0.055	-0.300	0.130	0.016	-0.012	-0.250	1.000

3.3.2 Matplotlib visualizes the distribution of variables

The box diagram shows the distribution of variables:

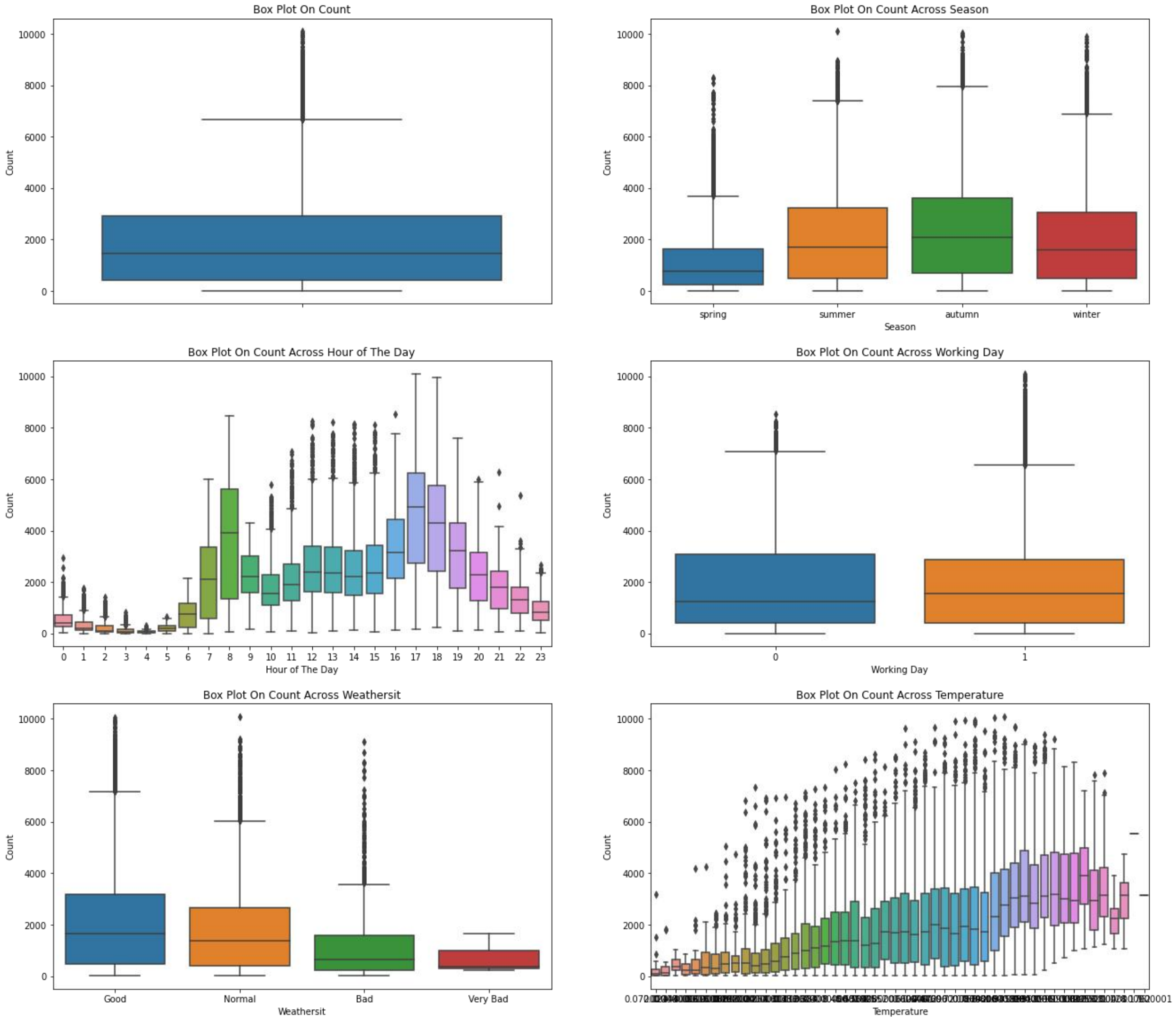

Figure 2. Box diagram shows variable distribution

From the picture "Hour of The Day" in the first column of the second row, there are two peaks: 7-8 o'clock and 17-18 o'clock. It happens to be the morning and evening peaks of commuting on weekdays.

From the last picture "Temperature", the number of bicycle users increases with the increase of Temperature; However, after the Temperature passes a certain critical point, the number of users in shared bike decreases with the increase of Temperature. The visualization results in the picture are consistent with our expectations.

3.4 Random forest regression

3.4.1 Model comparison

The correlation coefficient of variables "temp" and "atemp" is 0.99, which is further verified by visual exploration. Here, in order to reduce the model dimension, we delete the variable "atemp";

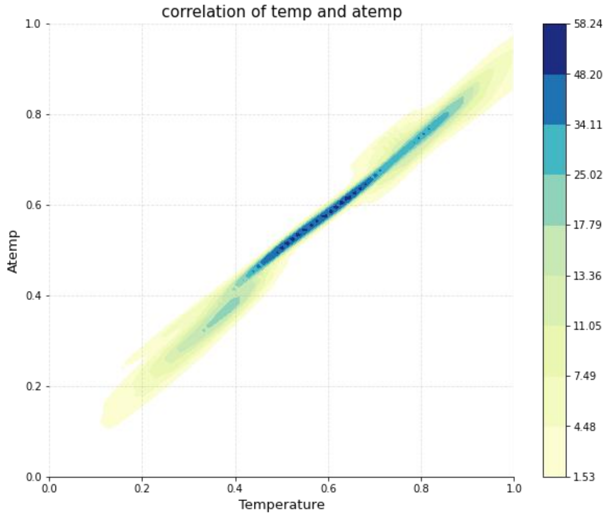

Figure 3. Temp versus atemp correlation display

It is observed that the target variable cnt is a group of continuous values, so the Ridge regression (Ridge), Lasso regression (Lasso), support vector machine regression (SVR), Random Forest Regressor and bagging method are preliminarily selected to build the model, and the fitting situation of the model is compared and analyzed.

The random number seed is set, and the data is divided into training set, verification set and test set according to the ratio of 60: 20:20; The parameter alpha = 0.1 for Lasso, alpha =. 5 for Ridge, gamma and kernel for SVR are set to default values; Mean Squared Error and R^2 score of different models are output.

Table 3. Comparison of MSE and R2 score of different models

Model	Mean Squared Error	R^2 score
Lasso	43103.36	0.07
Ridge	42963.88	0.07
SVR	41659.68	0.10
BaggingRegressor	55698.19	-0.20
RandomForestRegressor	18949.93	0.59

By comparison, the R^2 score of random forest regression model is the largest, which is 0.59, and its mean square error is the smallest, which is 18949.93, so the fitting effect of random forest model is the best.

3.4.2 Training of random forest regression model

By adjusting the parameters of stochastic forest model, further training and optimizing the model, whether it can achieve better fitting effect is observed.The evaluation parameters of each index of the model training set and the verification set are compared.

Table 4. Comparison of Index Parameters of Random Forest Model

Model	Dataset	MSE	MAE	RMSLE	R^2 score
RandomForestRegressor	training	299.54	10.94	0.21	0.98
RandomForestRegressor	validation	19020.41	96.49	0.47	0.59

After debugging, the model effect has been significantly improved, and the R score of the training set has reached 0.98, close to 1, which is 0.39 higher than the previous verification set; The mean square error is 299.54, which is 18650.39 lower than before. The parameters of the validation set have not changed significantly.

3.4.3 Feature importance analysis

The value of feature importance of each variable is output and sorted in descending order.

Table 5. Descending Order of Feature Importance

Feature	Value
hr	0.634691
temp	0.159124
hum	0.050063
workingday	0.046783
windspeed	0.026515
weathersit	0.026307
weekday	0.020375
mnth	0.019887
season	0.013120
holiday	0.003135

It can be intuitively seen from the table that the three factors that have the greatest impact on the number of users in shared bike are hour (hour), temp (temperature) and hum (humidity), and the importance values are 0.63, 0.16 and 0.05 respectively. Combined with tables and visual images, the factor that has the greatest impact on the number of users in shared bike is hour.

Therefore, we should pay special attention to the time of day when launching shared bike. Theoretically speaking, we should increase the number of bicycles in key areas during peak hours and correspondingly reduce the number of idle bicycles during low peak hours. However, increasing the number of shared bike during peak hours of the day will not effectively increase the number of bike users, but may aggravate the tidal phenomenon of shared bike.

4. Summary

The number of users in shared bike has a strong correlation with hour, temp and hum. According to the ranking of feature importance, hour has the greatest influence on the number of users in shared bike.The correlation coefficient between cnt and registered is the strongest, and the correlation coefficient reaches 0.97; Combined with visual exploration, the needs of registered users in shared bike are much higher than those of unregistered users. Therefore, we can significantly increase the total number of users in shared bike by increasing the number of registered users:

1) It is necessary to provide quarterly package service for registered users: 5 free uses +85 commuting uses +10 free uses. Registered users in shared bike have 5 opportunities to use bicycles free of charge. When using bicycles during commuting hours on non-working days, they can choose to deduct the number of free uses, totaling 10 opportunities.

2) It is necessary to provide student discounts to registered students. Registered users who have completed student certification in relevant apps in shared bike can enjoy student discounts.

3) Registered users can accumulate points by riding, and then exchange the points for coupons or goods. For example, sunscreen caps, ice sleeves, wearable fans, warm gloves, etc. These sunscreen and warm-keeping items can be used not only when riding bicycles, but also in other scenes of daily life, so users will be willing to exchange them.

4) In a year, May to October is the peak season for cycling, while the demand for bicycles in summer and autumn is high, and the two time periods are basically consistent. Therefore, we can increase the number of bicycles from May to October to meet the needs of users; Recycle some bicycles for maintenance from December to February of the following year (off-season of cycling).

5) During the peak hours of commuting, enterprises cooperate with urban management and law enforcement departments to dispatch vehicles to meet the travel needs of citizens during peak hours. Peak hours are 7-8 am, 17-18 pm and around 12 noon.

References

[1] CAMPBELLA,CHERRY C,RYERSON M,et al. Factors influencing the choice of shared bicycles and shared electric bikes in Beijing[J]．Transportation Research Part C，2016，67(6) : 399－414．

[2] MATTSON J,GODAVARTHY R. Bike share in Fargo,North Dakota:keys to success and factors affecting ridership[J].Sustainable Cities and Society,2017,34(10):174－182.

[3] FAGHIH-IMANI A,ELURU N,EL-GENEIDY A,et al. How landuse and urban form impact bicycle flows:evidence from the bicycle-sharing system(BIXI) in Montreal[J].Journal of Transport Geography,2014,41(12):306－314.

[4] EINAV L,LEVIN J. Economics in the age of big data[J].Science,2014,346(11):715－721.

[5] MULLAINATHAN S,SPIESS J. Machine learning:an applied econometric approach[J].Journal of Economic Perspectives,2017,31(2):87－106.

[6] Jiang Yueyao. Study on Supply and Demand Forecast and Dispatching of Tidal Points in Shared bike--A Case Study of Wushipu Metro Station in Xiamen [D]. Dongbei University of Finance and Economics, 2023 (02)