Bike Sharing Demand

COMP9417 Machine Learning Project

z5141683 LiJin Fan
z5244467 Chen Wu
z5225227 Mo Wang

Introduction

The bicycle sharing system is a way to rent bicycles through a network of kiosks in the city, automatically obtain membership, and rent and return bicycles. These systems allow people to rent bicycles from one location and return them to another location when needed. Currently, there are more than 500 bike sharing programs worldwide.

In this assignment, we selected the topic about shared bicycles. This is because the global warming trend is becoming more and more serious. People need to travel in a more environmentally friendly way. Shared bicycles are one of the ways. If the use of bicycles can be predicted, which will better serve customers and reduce the space occupied by too many bicycles.

The method we use is, processed the data first, found the correlation between the data. Use relevant data as a training set, different algorithms can be utilized for training, by comparing the results of different algorithms and finally got the best solution.

Dataset and simulation environment

The bike sharing demand simulation software used in this study is self-coded using Python 3.6. The source code has been provided along with this study, and external modules required to run the program are matplotlib, numpy, pandas and sklearn. The data used in this study was downloaded and extracted from Kaggle. The bicycle sharing rent data only includes entries for the two years from January 1, 2011 to December 19, 2012. The data set also contains weather statistics for the corresponding date and time. Since it is a competition data set, the complete data is divided into a training set and a test set. The training set only contains entries from the 1st to the 19th of each month, and the test set contains entries from the 20th to the end of the month but

excluding some important Predictor variables. In data exploration and analysis, I will use the training set to get the complete function and predictor variables. A data set containing 10886 observations and 12 variables will be used in this study.

Evaluation model

According to several discussions in our group, based on considerable analyses, We think that because we want to output discrete values, this is a regression problem in supervised learning. Finally, we use the Root Mean Squared Logarithmic Error (RMSLE) to evaluate our forecast. The RMSLE is calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(P_i+1)-log(a_i+1))^2}$$

Where n is the total number of samples, is the predicted value, and is the actual value. log(x) is the natural logarithm algorithm.

We have analyzed the reasons for selecting the accuracy judgment algorithm, and we believe that the prediction accuracy is our main consideration. The root mean square logarithmic error criterion is generally compared with the standard error criterion. If the range of predicted values is large, the standard error method will be dominated by some large values. In this way, even if many small values are predicted accurately, but a very large value is not accurate, the standard error will be large. Correspondingly, if another relatively poor algorithm is more accurate for this large value, but many small values have deviations, the mean square error may be smaller than the previous one. Therefore, after taking the logarithm and comparing, this problem can be solved to a certain extent.

Data analysis and feature selection

In order to make the data easy to understand and further analyze, the identifiable statistical trends and patterns of the data set are analyzed. By observing the data in the training set and selecting features, it can be found that some features are numerical data, and some are discrete data. Therefore, for numerical data, we choose to use regression curves to observe the relationship between each feature variable. For discrete data, we use scatter plots and heat maps to observe the relationship before each feature variable. After preliminary analysis, the following steps were taken to convert the data into an operational data set of the system:

a. change date time to timestamp

b. Divide the timestamp into days, months, years, and days of the week.

c. Convert seasons, holidays, working days and weather into categorical variables or factors.

d. Convert hours to a factor.

We began our exploration with searching correlations of each feature, getting the following result:

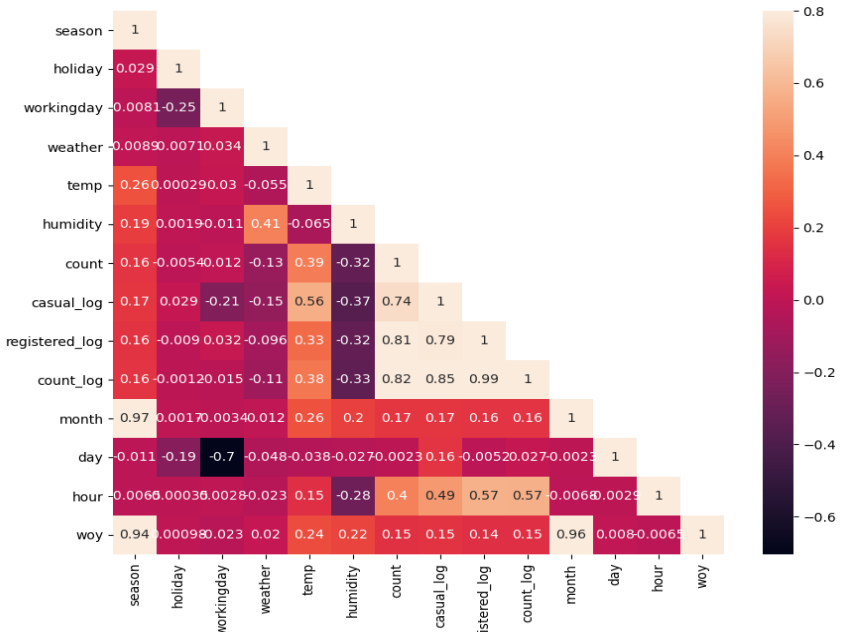

Figure1. Features Correlation

It is clear that the intensity of the color shows the strength of the correlation. There are positive and negative correlations between temperature and humidity and the number of rentals, and the wind speed has almost negligible influence on the number of rentals.

Then it is of significance to analyze the impact of various variables on count, acquiring the following result:

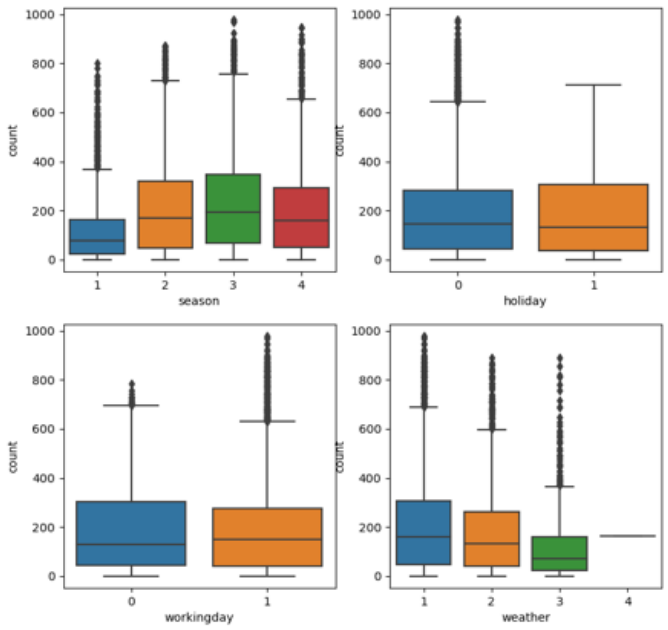

Figure2. The relation between count and features

In terms of seasons, sharing bikes are frequently used in summer and autumn. As can

be seen in the holiday's figure, users tend to use more frequently than non-holidays, which is consistent with our expectations. Simultaneously, work exerts negative impact on using sharing bikes. Also, weather conditions affect the number of rentals significantly, the use of shared bicycles on sunny days is the most.
According to the number of rentals per month, draw the following figure.

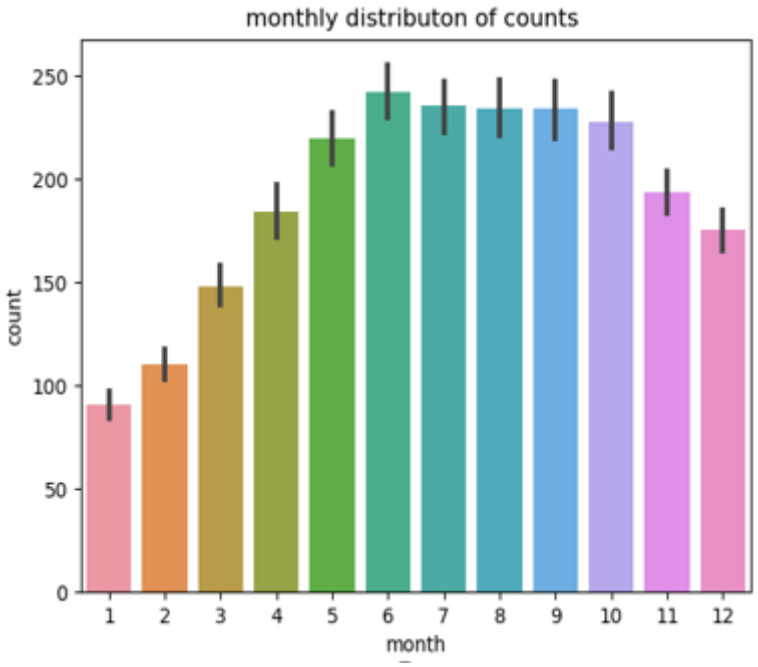

Figure3. Monthly distribution of counts

From the graphic point of view, the rental volume is mainly concentrated in the spring, summer and autumn. The winter months are relatively less frequent. It may not be suitable for outdoor riding because of the low temperature, while the temperature in spring, summer and autumn is suitable for riding.

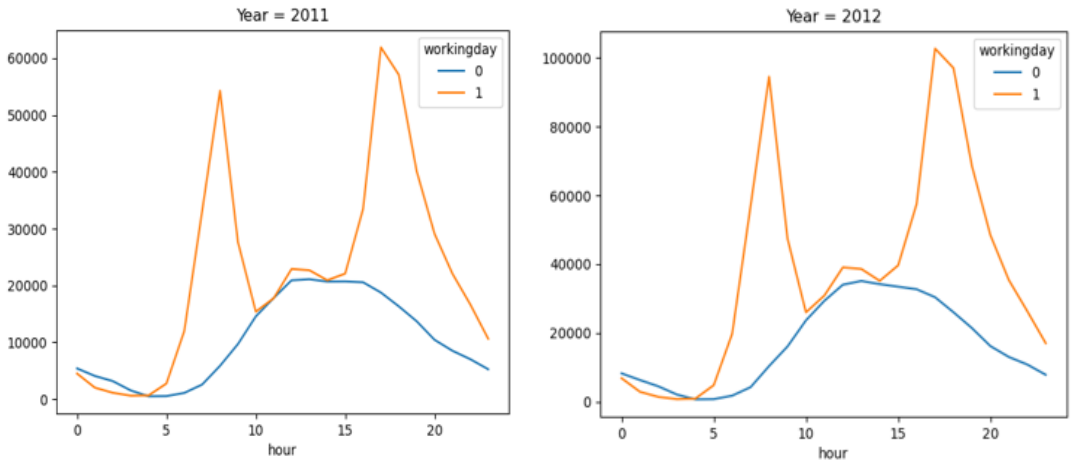

Figure4. Rental status per hour per year

From the picture above, we find that the trend of working days is the same, but the trend of non-working days is the same. This is also consistent with our expectations. The peak consumption of non-working days is concentrated in various time periods during the day.

When analyzing the 'casual', ' registered', the target variables registered and casual are converted into log(a+1) form, making the correlation better, as followed:

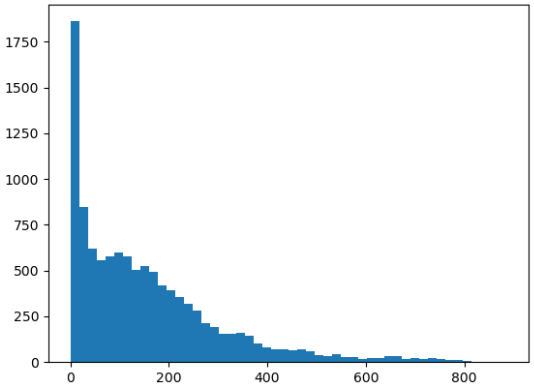

Figure5.1 registered (before log)

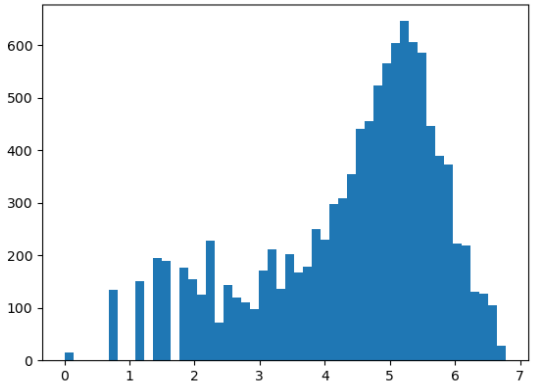

Figure5.2 registered(after log)

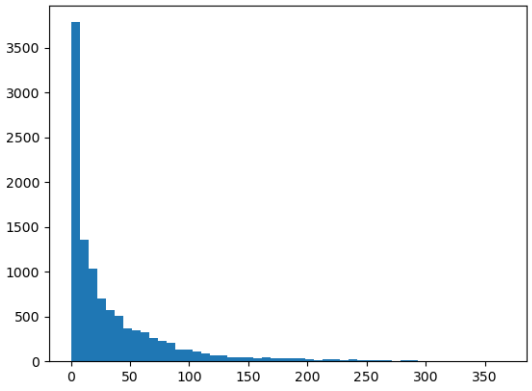

Figure5.3 casual (before log)

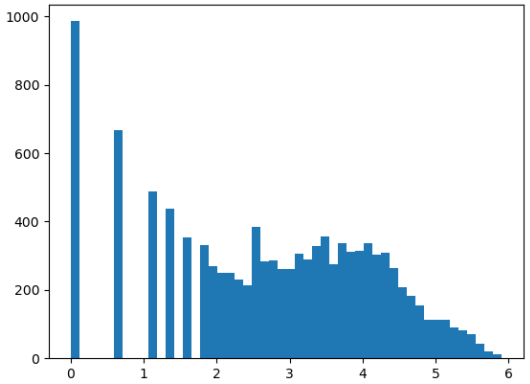

Figure5.4 casual (after log)

Then compare the impact of different humidity on the number of rentals. According to Humidity, draw the following figure:

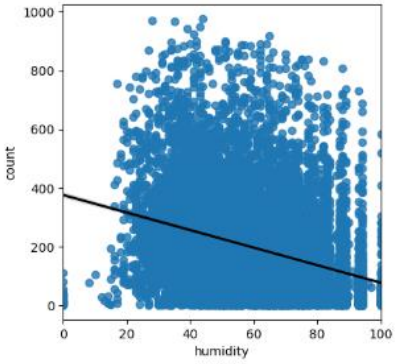

Figure6.1 humidity with count

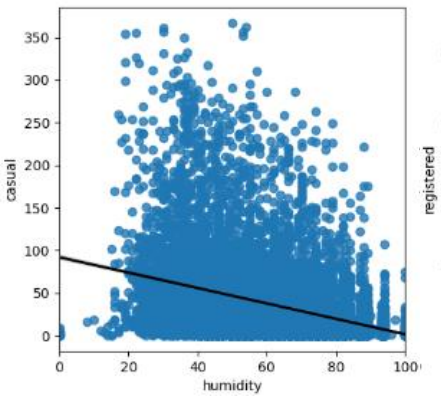

Figure6.2 casual with count

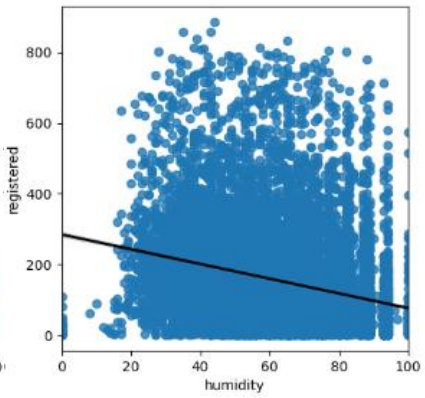

Figure6.3 humidity with registered

According to the scatter plot, we can observe that as the humidity increases, the number of users decreases. This means that when the weather is high and the humidity is high, the number of people who choose to use bikes for go outside becomes low.

The number of rentals, temp and atemp have a positive trend, this can be shown in the following graph:

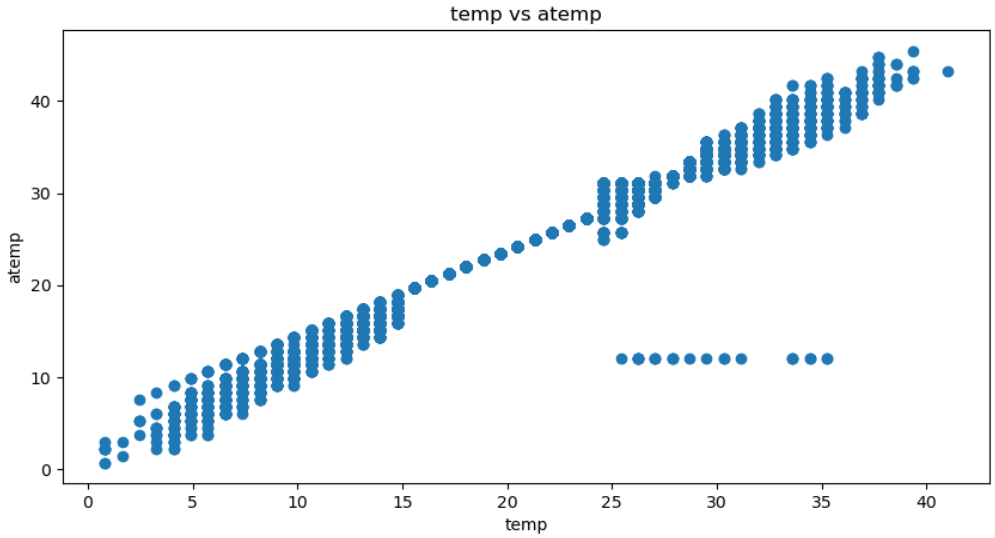

Figure7. Temp vs atemp

Finally, analyze the relationship between wind speed and the number of users, and the chart is as follows:

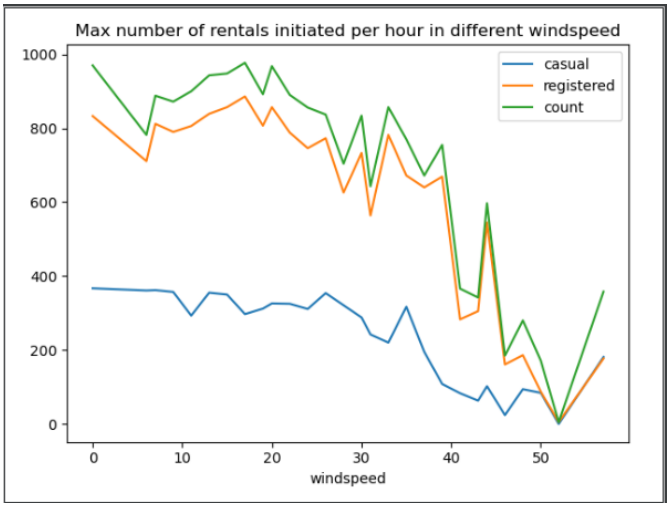

Figure8. Max number of rentals initiated per hour in different windspeed

Observing the chart data, we can see that when the wind speed is greater than 50, the number of users still rises, and the overall does not show a linear correlation. Therefore, the feature of wind speed does not have much impact on the number of users. This feature is discarded when the subsequent model is established.

Model

In order to select the best algorithm for this task, we need to know some characteristics of the machine learning algorithm. The linear regression algorithm cannot fit the nonlinear data well. The KNN algorithm has a huge amount of calculation and the decision tree is prone to overfitting. Therefore, when selecting the model, the above algorithms are preferentially excluded. We selected the following three models for training

SVR

Support vector machine (SVM) itself is proposed for dichotomy, and SVR (Support vector regression) is an important branch of SVM. The difference between SVR regression and SVM classification is that the sample points of SVR end up with only one category. The optimal hyperplane it seeks is not the "most open" of two or more sample points as SVM does, but the minimum of the total deviation of all sample points from the hyperplane.

When we used SVR as a model for training, the results were not very good. So we trying to use other algorithms.

RF

Random forest [1] is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset. It works by creating multiple weak learners for different subsets of the training set, and then combines these weak learners to form a strong learner. The basic idea of random forest is that every time the decision tree is repeatedly trained on the data set, a new decision tree will be generated every time,

and multiple such trees will reduce the overall error of the model.
The experiment is performed in scikit-learn's integrated random forest classifier using Python. Find the best value for each parameter. The parameters that are being considered for tuning are as follows:
Estimated number: number of trees in the forest
Maximum number of elements: the number of elements to be considered for each previous element.
To improve the predictive accuracy and control over-fitting, we trained our model using different values of n_estimators = [100,200,300,400,500,600,700,800,900,1000] and max_depth = [5,10,15]. We found the lowest RMSLE is 0.1867. [3]

GBR
Gradient boosting regressor [2] like random forest is a holistic learning method. Similar to the latter, it uses multiple weak learners, which are combined into one strong learner. However, unlike its random forest, as the name suggests, gradient boosting uses boosting.
Use scikit-learn's Gradient Boosting integrated learning method to analyze in python. To adjust the parameters of the model, we extracted feathers from the original Gradient Boosting paper. The suggestion is to maintain a high estimator first, and then use these best parameters to obtain the estimator to tune other parameters. We tried the Gradient Boosting tree provided by Sklearn for different values of n_estimator (number of trees) and max_depth .
In order to boost the performance, we trained our model using different values of n_estimators =[100,200,300,400,500,600,700,800,900,1000], max_depth = [5,10,15]. The lowest RMSLE is 0.1988 and was obtained with n_estimators = 1000, max_depth = 5. [3]

Result

In order to avoid overfitting, we used our regressors to test the dataset and drawn a conclusion of performance. After a series of tuning parameters, we found that the Random forest model and Gradient boosting regressor, the values obtained are relatively close. Therefore, we analyzed the two models and simulated them according to a certain ratio. When the ratio of the Random forest model to the Gradient boosting regressor is 6:4, the RMLSE obtained is the smallest.

Name	Rmsle_score	Name	Rmsle_score
RF_rmsle	0.18674	0.2 * rf + 0.8 * gbr	0.18939
GBR_rmsle	0.19885	0.3 * rf + 0.7 * gbr	0.18594
SVR_rmsle	0.29883	0.4 * rf + 0.6 * gbr	0.18335
		0.5 * rf + 0.5 * gbr	0.18166
		0.6 * rf + 0.4 * gbr	0.18087
		0.7 * rf + 0.3 * gbr	0.18099
		0.8 * rf + 0.2 * gbr	0.18201

Conclusion

The purpose of this research is to analyze and predict the impact of the characteristics on the demand for shared bicycles by observing the characteristics of each data in the trainset. In the project, we considered support vector regression model, gradient boosting regression model and random forest model. After a series of data analysis, the gradient ascent model and the random forest model were selected as the methods for predicting the demand for shared bicycles in this experiment.

In our study, multiple characteristics were considered, and it is not clear which characteristic most affects the demand for bike sharing.[4] However, it is clear that weather factors have an adverse effect on the demand for shared bicycles. In most studies, unfavorable weather conditions such as rain and hail are negative for all travel needs of registered and casual users on weekends and working day.

With future work, we can use more algorithms to test, such as XGBoost, and we can also group some data, which may make the prediction more accurate.

Reference

[1] Breiman, L., 2001. Random forests, machine learning 45. J. Clin. Microbiol, 2(30), pp.199-228.
[2] Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis, 38(4), pp.367-378.
[3] Sachdeva, P. and Sarvanan, K.N., 2017. Prediction of Bike Sharing Demand. Oriental Journal of Computer Science and Technology, 10(1), pp.219-226.
[4] Eren, Ezgi and Uz, Volkan Emre, 'A Review on Bike-Sharing: The Factors Affecting Bike-Sharing Demand' (2020) 54 Sustainable cities and society.