Bike Sharing Demand Analysis

Yisha He (yh2uq) Shuyang Yu (sy2sj) Ren He (rh5ve) Jingying Pan (jp9pbd)

Abstract

This project[1] studies the demand of bike sharing system in Washington D.C.. The aim is to predict bike rental demand using accessible information (time and weather). This project also focuses to find the optimal regression model and classification model that minimizes mean squared prediction error (MSPE) and accuracy respectively. Methods of *Ordinary Least Squares* (*OLS*), *Ridge*, *Lasso*, *Partial Least Squares* (*PLS*), *KNN* and *Decision Tree* are used for the regression analysis. Methods of logistic, *Support Vector Machine* (*SVM*) and *Random Forest* are applied to the classification analysis.

Introduction

Bike sharing systems are a means of renting bicycles based on smartphone applications. People rent and return bikes via kiosk location networks throughout a city. Using these systems, people can rent a bike at one location and return it to a different place as needed. Bike sharing systems make people's city travels a lot easier. There are more than 500 bike-sharing programs around the world and the competition in between is fierce. Several factors may decide the prosperity of a bike sharing system: how to control bike cost, how to maintain bike quality, how to set promotions and advertisement etc.. However, how to forecast the demand of bike rents is the most crucial factor that ensures the sustainability of a bike sharing system. Understanding and successfully predicting the bike rental demand via accessible information (time and weather etc.) will help the system decides its strategy to rebalance bikes, improving efficiency and productivity accordingly.

This project analyses bike rental demand of Capital bikeshare (CaBi) system. Capital bikeshare (CaBi) is a bike sharing system operating in Washington D.C. metropolitan area from September 2010. It is one of the largest bike sharing systems in the area. This dataset was provided by Hadi Fanaee Tork from website of Capital bikeshare[2]. The goal of this project is to predict bike rental demands through weather and time information.

Descriptive Analysis

2.1 Data Description

Hourly data in the span of two years (from Jan 2011 to Dec 2012) is analysed in the analysis. Explanatory variables include datetime, season, holiday, workingday, weather, temp (temperature), atemp ("feels like" temperature), humidity, windspeed. Count (number of bike rents in one hour) serves as the response variable. There are 10,886 observations in the dataset, 8078 observations (80%) as training set and 2178 observations (20%) as testing set.

2.2 Data Visualizations

Exploring the dataset, some important features are discovered. From Figure 1, there is an upward trend of bike rents from 2011 to 2012. It seems that Capital Bikeshare's expanding strategy in the end of 2011 works, causing the shift of bike share demand in level in 2012. Within a year, more bikes are shared from April to September than the rest of the year. This may be due to the temperature effect, that people refrains riding bicycles from October to March as the temperature is lower in the period.

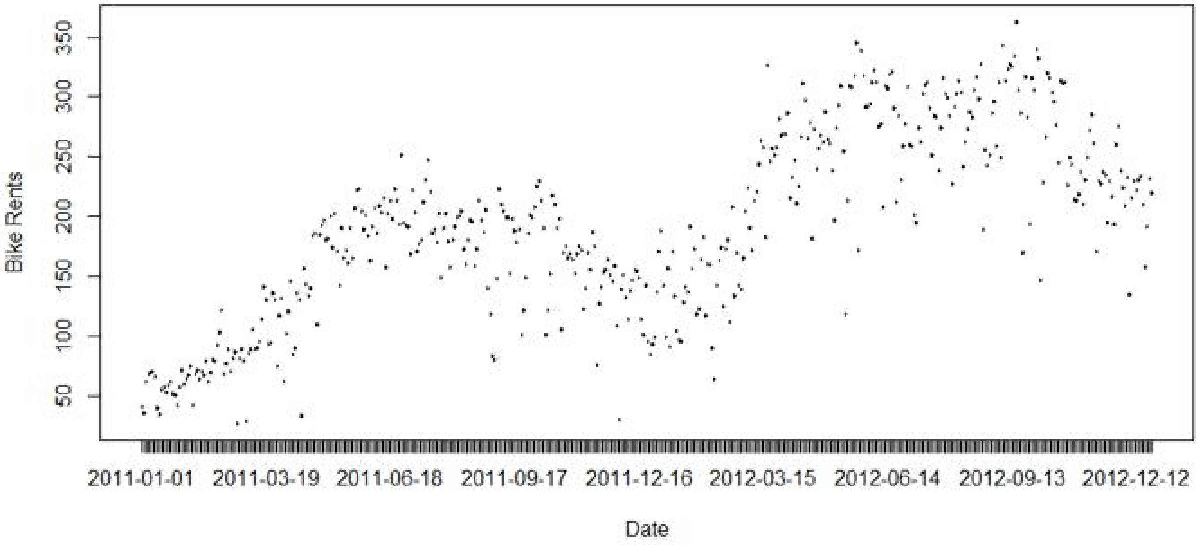


Figure 1. Hourly Bike Rents from 2011 to 2012

According to Figure 2, there exist significant differences of bike rents with hours and seasons. The majority of bikes are rented during the day, and the peak hours of bike demand within a day are around 8 A.M. and 5 P.M.. This is consistent with the commuting hours for most people. Separating the seasonal factor, spring has overall the lowest bike rental demand of four seasons. This may be due to the fact that spring is denoted from 1 January to 31 March in a year, which is the coldest time in Washington D.C. area. Weather can be the hidden truth reason that cause the bike rental demand.

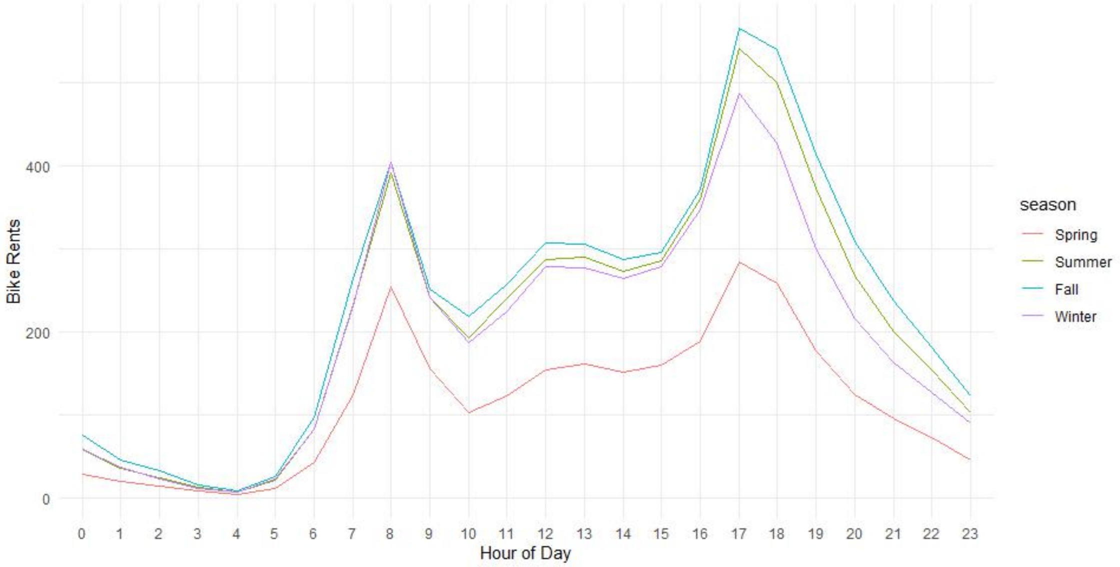


Figure 2. Hourly and Seasonal Differences in Bicycle Demand

From Figure 3, weather dominates the bike demand during daytime. When the weather is good or normal, people are more likely to use a bike rental system than when the weather is bad or very bad.

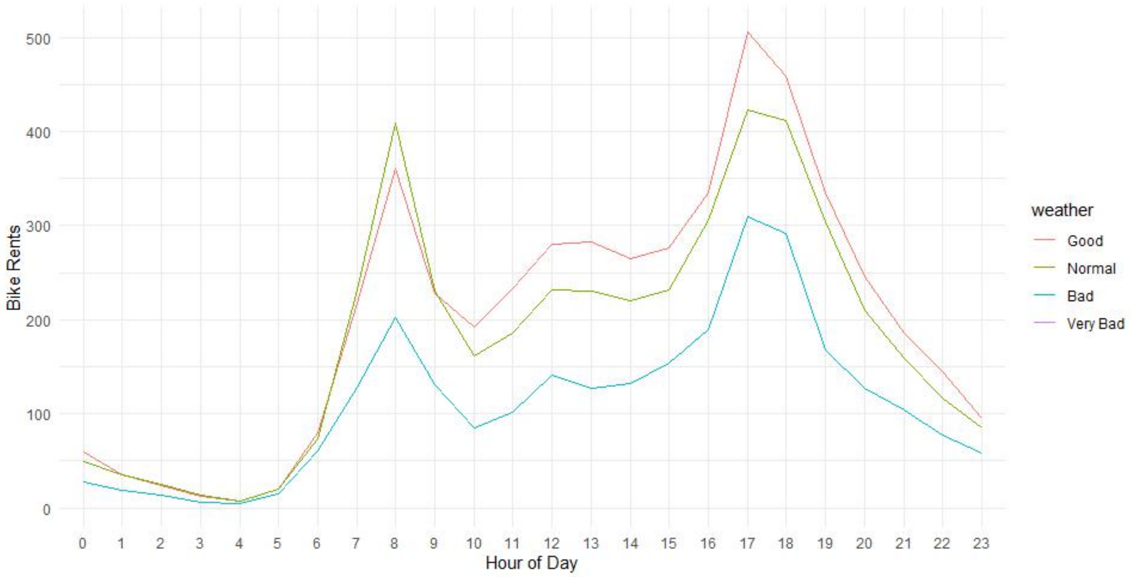


Figure 3. Weather Differences in Bicycle Demand

People have very different demand patterns between weekdays and weekends. From Figure 4, very few people rent bikes in the morning on weekends, and the demand for bikes on weekends is highest from 12 p.m. to 5 p.m.. People are more active at midnight (from 12 a.m. to 2 a.m.) on Saturdays and Sundays.

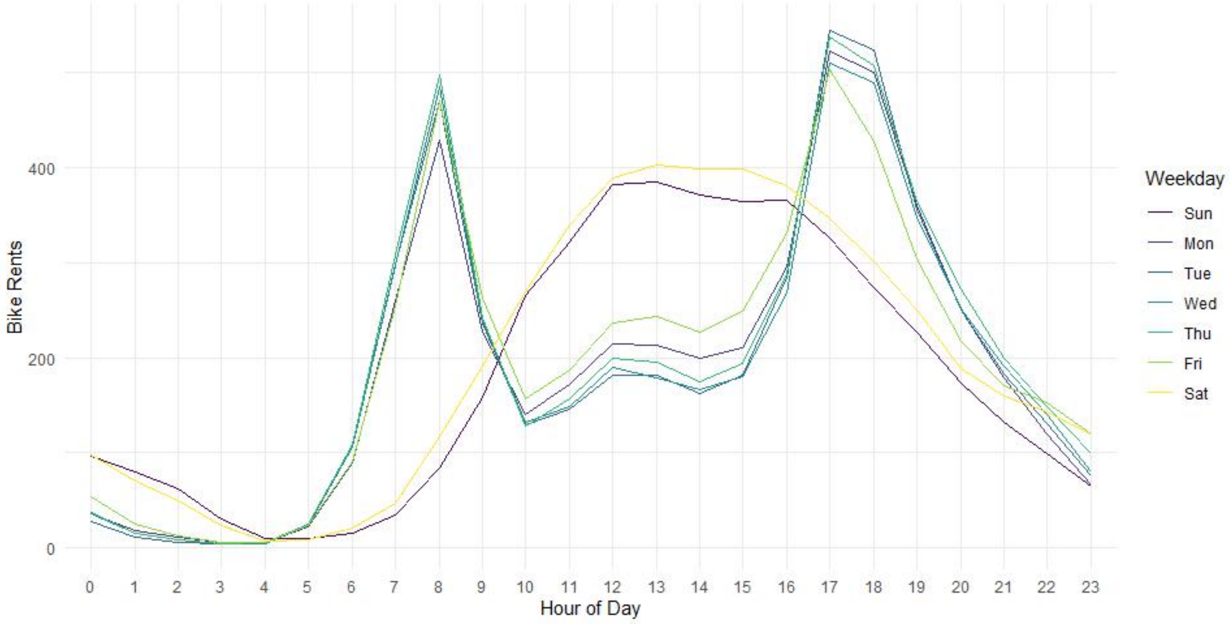


Figure 4. Weekday Differences in Bicycle Demand

3. Continuous Regression

3.1 Ordinary Least Squares (OLS) Model

Ordinary Least Squares (*OLS*) method is conducted on the training set, and the model is as follows:

Model 1: count ~ season + holiday + workingday + weather + temp + atemp + humidity + windspeed + date + hour + month

The mean squared prediction error (MSPE) for this model is 9866.26 and the average residual of the *OLS* model on the testing data is 99.33, which is higher than expected. However, as there is significant difference in bide demand pattern between weekdays and weekends, an interaction term (*time:workingday*) is added into the original *OLS* model as below:

Model 2: count ~ season + holiday + workingday + weather + temp + atemp + humidity + windspeed + date + hour + month + hour:workingday

ANOVA analysis comparing the two models gives F statistic of 214.78 and p-value of 2.2*10^(-16). This indicates the improved model is reasonable and significantly better illustrating the bike demand variable. The MSPE for the new model is 7073.13, much lower than the original OLS model.

3.2 Decision Tree Model

Then a *Decision Tree* model was performed on the training set, with MSPE of 2114.92. From the tree model regression results, we plot the influence of each variable on bike rental demands in Figure 5.

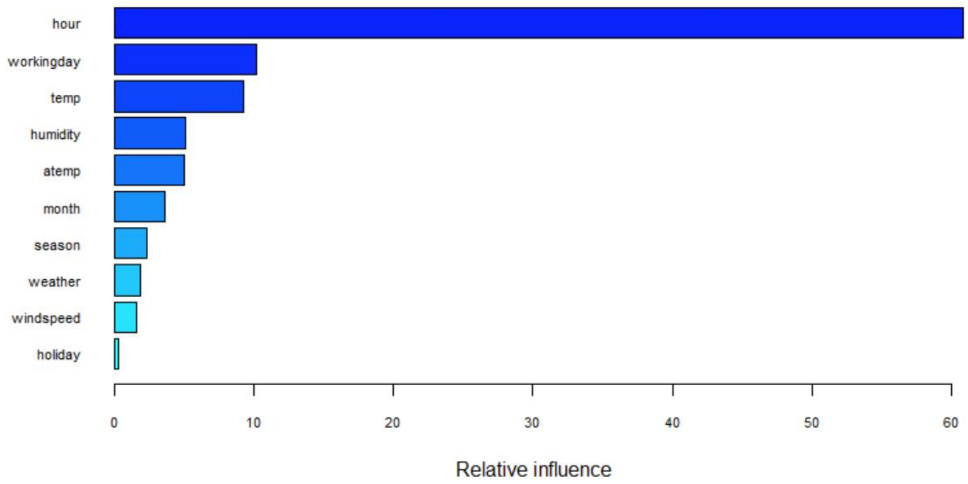


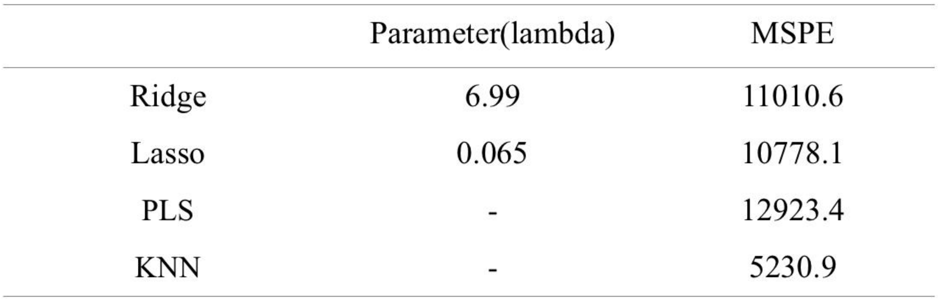
Figure 5. Relative Influence of Each Variable on Count

From Figure 5, *hour* is the most important factor that determines bike demand. *Workingday* and *temperature* also have significant effects on *count*. However, *weather* and *windspeed* have little influence on count. This may indicate that the weather cannot significantly change people's bike rental pattern.

3.3 Other regression models

We also noticed that some of the variables may be correlated, to deal with correlated covariates, we conducted *Ridge* regression; we also tried *Lasso* regression to see if we can do some variables selection, but the result shows none of the coefficients can be shrunken to 0. Models such as *Partial Least Squares* and *KNN* are also conducted in the analysis. However, these models do not overperform the OLS model or the *Decision Tree* model. Table 1 illustrates the results of these models.

Table 1. Parameter and MSPE of Other regression Models



	Parameter(lambda)	MSPE
Ridge	6.99	11010.6
Lasso	0.065	10778.1
PLS	-	12923.4
KNN	-	5230.9

The lambda parameters of *Ridge* and *Lasso* are determined by 5-folder cross validation method. Among the above four regression models, *KNN* performs best with the smallest MSPE on the testing dataset.

3.4 Discussion of Continuous Regression

Linear regression is an example of a parametric approach because it assumes a linear functional form for the regression. Such parametric methods have several advantages. They are often easy to fit, as one only need to estimate a small number of coefficients. In the case of linear regression, the coefficients are simple to interpret, and inference analysis can be performed. However, parametric methods have disadvantages. By construction, strong assumptions are implemented with the regression. If the implemented functional form is far from the true model, we can hardly get accurate predictions, and any conclusions of such models will be misspecified.

In contrast, non-parametric methods (such as *KNN* method) do not explicitly assume a parametric form for the regression, providing an alternative and more flexible approach for regression analysis.

In the bike sharing demand analysis, *KNN* model performs better than *OLS* model, indicating that the true model is far from linear and the other models we conducted (*Ridge*, *Lasso*, *Partial Least Squares*).

Of all the continuous regression models, *Decision Tree* model results in highest predicting accuracy. It implies complex relationship exists between the features and the response. Yet *Decision Tree* model may not be robust. A small change in the training dataset can cause a large change in the estimated tree.

4. Classification Analysis

Trying to simplify the model, we considered transforming the data, making it a classification problem by separating variable count with its median into a new 0-1 variable named as demand. By doing that, we can make our model easier to interpret. As the total amount of bikes available in the city is nearly constant throughout the period of observation. A response variable determining whether the demand is high or low gives the company more an explicit yes-no suggestion whether they need to put more effort into maintenance of the bikes at some specific times.

We have the same training and testing set as we had in our continuous regression analysis, and we continue to set seed 1693 for consistency and reproducibility. In the following analysis, we tried models including logistic classification, support vector machine and random forest. Among all, the random forest model has the best performance.

4.1 Logistic Model

Logistic classification model is similar to ordinary least square regression. It assumes equality of conditional probability and linear combination of independent variables transformed by the logistic function. The use of transformation can help guarantee that linear combination of the independent variables would fall in range (0, 1).

log(count) ~ season + holiday + workingday + weather + temp + humidity + windspeed + hour + month + year + hour:workingday

The obtained model shows that all predictors except *holiday* are significant at 5% level. This result is reasonable, as holiday and working day only differ at national holidays, making them highly collinear. The model has accuracy 79.71%, sensitivity 80.18%, and specificity 79.23% in the testing set. Performance of the *Logistic* model is considerable.

4.2 Support Vector Machine Model

Support Vector Machine (SVM) is a supervised learning model that incorporates associated learning algorithms that optimizes classification problems. It efficiently finds the criteria of assigning each observation into a category as a non-probabilistic binary linear classifier. It can be visualized as separation of observation points by the gap they fall within a mapped space. Generally, the separation is achieved by finding a maximum margin hyperplane that is determined by maximizing its distance from every observation from either category. The plane can be either linear or nonlinear. We have tried both in our analysis.

Variables *season, holiday, workingday, weather, temp, humidity, windspeed, hour, month* and *year* are implemented. The support vector machine model using a linear hyperplane obtains result with accuracy of 79.43%, sensitivity of 79.74%, and specificity of 79.13% in the testing set. Using a nonlinear hyperplane, the support vector machine model obtains accuracy of 84.39%, sensitivity of 84.72%, and specificity of 84.07% in the testing set.

From the result we see that effect of our predictors show signs of nonlinearity. For example, even though a quadratic term for temperature only has coefficient of 0.08 in a linear regression, the correlation between temperature and an individual's desire to ride a bike is not perfectly linear. When the temperature is either too low or too high, the demand for bike rentals diminishes.

4.3 Random Forest Model

Random forest model is an ensemble learning method that can be applied in classification analysis. It utilizes a multitude of *Decision Trees* and outputs mode of the classification result of individual trees. Comparing with a single *Decision Tree*, the random forest is less prone to the problem of overfitting the training set.

Similar to the *SVM* models, variables *season, holiday, workingday, weather, temp, humidity, windspeed, time, month* and *year* are considered. After tuning with cross validation, we find that when the number of trees increases beyond 500, the efficiency of the model no longer increase significantly. We then ran a random forest model with tree number 500 and nodesize 1, the model reaches accuracy of 93.80%, sensitivity of 94.47%, and specificity of 93.14% in the testing set. The random forest model has the best performance of all the models we have tested.

In the following graph we show the importance of each variable in the random forest model, which is similar to what we have obtained in the influence plot in section 2.2.

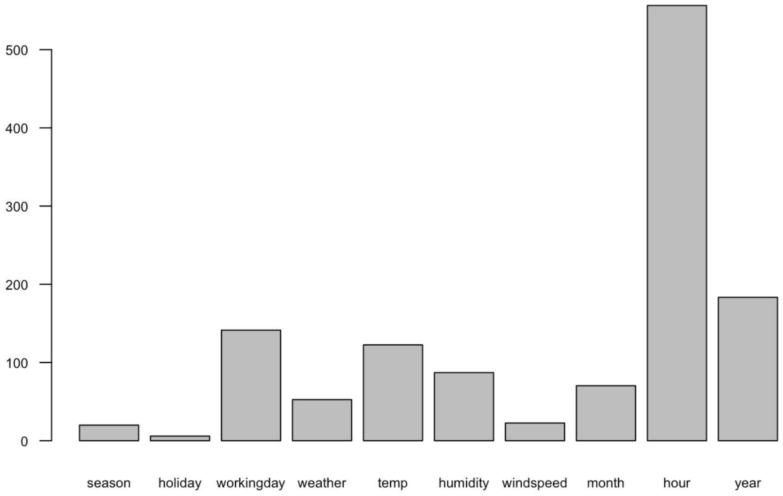


Figure 6. Relative Influence of Each Variable on Bike Demand

4.4 Discussion of Classification Analysis

Solely from the result of running classification models, it is hard to tell which independent variable is more important regarding to the prediction of the response. Meanwhile, from the fact that nonlinear *SVM* performs better than the *Linear SVM*, we can infer that the explanatory variables have nonlinear effects on the response. As the *Random Forest* model performs the best in predictions, we see that the interactions between independent variables are potentially important factors to be considered.

5. Limitations and Conclusions

5.1 Limitations

The dataset is ideal as there are no missing values, and we have dealt with the problem of multicollinearity beforehand. However, the applications of our analysis is limited due to boundary of the dataset. First, the data is outdated. It is questionable whether the analysis result gives consistent prediction for the current situation. Second, we could not obtain location information of each ride. If with departure and arrival locations of each ride, we can test point process models on the dataset, and potentially build heatmaps for Capital Bikeshare that notify the specific locations they should allocate more bikes. Moreover, if the data of usage duration is available, we would be able to analyze purpose of bike rides and provide maintenance suggestions to Capital Bikeshare.

5.2 Conclusions

Overall, *Decision Tree* model is selected with the best prediction performance in regression analysis and *Random Forest* model is with the best prediction result in classification analysis. Using the information supplied in the dataset, we can predict the hourly demand for bike rents using *Decision Tree* model with MSPE of 2114.92. This result should be useful for Capital Bikeshare to rebalance their bike supplies accordingly.

Considering time trends, the demand of bike rents steadily increases over the period of observations. Spring has the overall lowest bike demands compared with other seasons. The analysis also suggests that the usage of bike has different but consistent routine between working days and holidays. The regression models finds that variables *hour, workingday* and their interaction term are very important predictors that decides bike demand in Washington D.C.

between 2011 and 2012. Factors related with weather and temperature impose little influence on an individual's daily usage of bike. Classification analysis performs similar results, with *hour* and *workingday* as the most influential factor that decides bike demand.

Reference:

[1] Project Source: https://www.kaggle.com/c/bike-sharing-demand/overview

[2] Dataset source: https://www.capitalbikeshare.com/system-data