Bike Sharing Demand Predictions

Akhila Masadi

15/06/2022

Summary

In the urban transportation systems, bike-sharing has increasingly become a popular part. However, there has been a challenge in balancing the demand for bike-sharing in different regions. To ensure service efficiency and support timely re-balancing, there needs to be accurate demand prediction. It is necessary to know the number of bikes needed in each station at a given time to help ensure the demand is always met for customer satisfaction. This is because, at times, there are few bikes than the demand, and at other stations, there are more than enough bikes than what is demanded. Therefore, this project intends to predict bike sharing demand and ensure a balance of available bikes and demand for each station. This will be dependent on various factors like precipitation, day of the week, season, the hour of the day, and weather conditions. The project implements the Decision Tree and Random Forest models for the predictions in R language. From this, the most accurate model to provide efficient and effective predictions was identified as Random Forest, which could be used for accurate bike sharing demand prediction. The dataset will be retrieved from the UCI Machine Learning Database.

**Introduction**

It is always difficult for the bike-sharing systems to determine the stations which require many bikes and stations that require a smaller number of bikes. However, in the case where there is bike-sharing demand prediction, the bike-sharing companies can easily predict the number of bikes that are to be required through the utilization of the available data. Therefore, this paper analyzes and visualizes data to clearly illustrate most stations with a higher demand for bikes. With the fast prediction for this project, it will be easier to avoid uneven bike distribution with high accuracy. Factors like climate details from the available data will be useful in analyzing and predicting the number of bikes on demand.

Literature Review

The prediction of the bike-sharing demand has been an interesting topic for many researchers. The prediction of bike-sharing demand includes the creation of predictive models that essentially evaluate the probability of requests for the bikes in each area dependent on historical data and information from previous studies. Tong *et al.* (2017) implements a statistical model used to predict the number of bikes rented every hour by considering climate conditions, time information on the days of the week, and the event of holidays. Dell'Olio *et al.* (2011) implemented a method that was essential in assessing potential demand to help determine the drop-off and pick-up stations by using GIS. In addition, the Poisson model was also created to help consider the effects of climate conditions and calendar events on the number of bikes on demand. As identified in several research studies reviewed in this paper, there is less research on the demand for public sharing of bikes. Therefore, in this paper, analysis and prediction of public bike-sharing demand using data mining techniques and Machine Learning techniques like Decision Tree and Random Forest algorithms will be put in place.

Theory

H1: There is a lower demand for bikes when the temperatures are high compared to cold temperatures.

H2: The demand for bikes varies for different seasons.

#install.packages("xtable")

library(ggplot2)

library(dplyr)

##

## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':

##

## filter, lag

## The following objects are masked from 'package:base':

##

## intersect, setdiff, setequal, union

library(randomForest)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##

## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':

##

## combine

## The following object is masked from 'package:ggplot2':

##

## margin

library(lubridate)

##

## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':

##

## date, intersect, setdiff, union

library(xtable)

Data

The data was retrieved from <https://www.kaggle.com/datasets/bvc5283/Bike_Shared_Data/download>

## datetime season holiday workingday weather temp atemp humidity

## 1 2011-01-01 00:00:00 1 0 0 1 9.84 14.395 81

## 2 2011-01-01 01:00:00 1 0 0 1 9.02 13.635 80

## 3 2011-01-01 02:00:00 1 0 0 1 9.02 13.635 80

## 4 2011-01-01 03:00:00 1 0 0 1 9.84 14.395 75

## 5 2011-01-01 04:00:00 1 0 0 1 9.84 14.395 75

## 6 2011-01-01 05:00:00 1 0 0 2 9.84 12.880 75

## windspeed casual registered count

## 1 0.0000 3 13 16

## 2 0.0000 8 32 40

## 3 0.0000 5 27 32

## 4 0.0000 3 10 13

## 5 0.0000 0 1 1

## 6 6.0032 0 1 1

Check for missing values in the dataset.

sum(is.na(Bike\_Shared\_Data))

## [1] 0

There are no missing values. Next, from the datetime column a column named hour has been created.

Bike\_Shared\_Data$hour <- hour(ymd\_hms(Bike\_Shared\_Data$datetime))

head(Bike\_Shared\_Data)

## datetime season holiday workingday weather temp atemp humidity

## 1 2011-01-01 00:00:00 1 0 0 1 9.84 14.395 81

## 2 2011-01-01 01:00:00 1 0 0 1 9.02 13.635 80

## 3 2011-01-01 02:00:00 1 0 0 1 9.02 13.635 80

## 4 2011-01-01 03:00:00 1 0 0 1 9.84 14.395 75

## 5 2011-01-01 04:00:00 1 0 0 1 9.84 14.395 75

## 6 2011-01-01 05:00:00 1 0 0 2 9.84 12.880 75

## windspeed casual registered count hour

## 1 0.0000 3 13 16 0

## 2 0.0000 8 32 40 1

## 3 0.0000 5 27 32 2

## 4 0.0000 3 10 13 3

## 5 0.0000 0 1 1 4

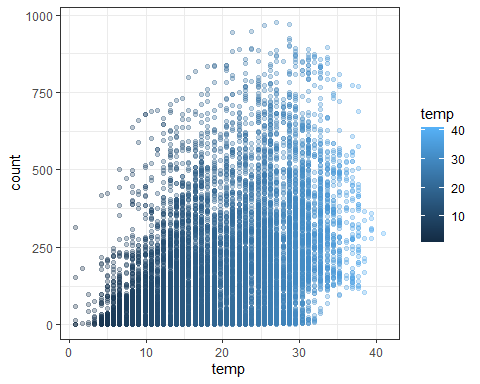
## 6 6.0032 0 1 1 5

Methodology

Since the dataset is clean, we start working on the Bike\_Shared\_Data dataset. The first thing is to perform an Exploratory Data Analysis on the Bike\_Shared\_Data dataset.

The first thing involves checking whether there is a relationship between the number of bikes rented and temperature.

ggplot(data = Bike\_Shared\_Data, aes(temp,count)) + geom\_point(alpha = 0.3, aes(color = temp)) + theme\_bw()



Next was to identify if there is any correlation between temperature and count of bikes rented.

cor(Bike\_Shared\_Data[,c('temp','count')])

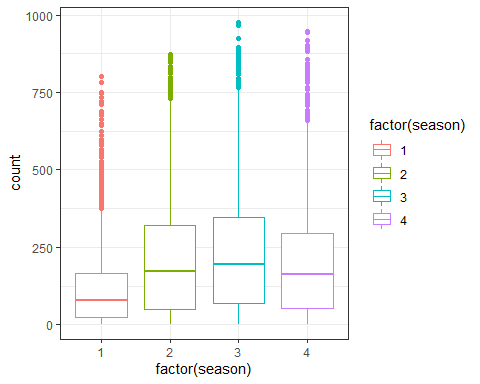
## temp count

## temp 1.0000000 0.3944536

## count 0.3944536 1.0000000

Distribution of bikes count and season

ggplot(Bike\_Shared\_Data,aes(factor(season),count)) + geom\_boxplot(aes(color = factor(season))) + theme\_bw()

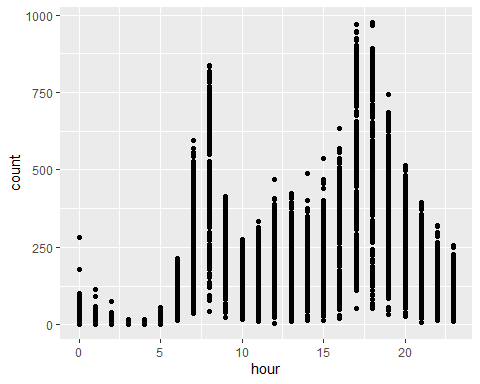


Next is the analysis on the relationship between hour of working day and count of rented bikes.

plot\_1 <- ggplot(filter(Bike\_Shared\_Data,workingday == 1), aes(hour,count))

plot\_1 <- plot\_1+ geom\_point()

print(plot\_1)



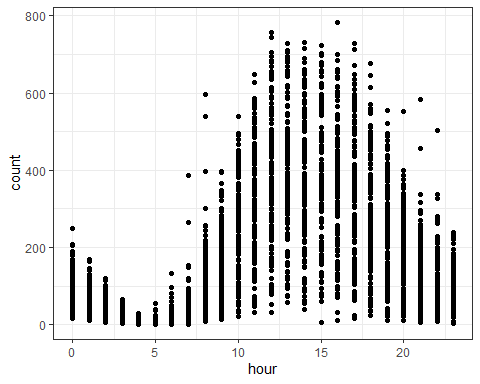
Next, was the analysis for the relationship between hour of non-working day and count of rented bikes.

plot\_2 <- ggplot(filter(Bike\_Shared\_Data,workingday == 0), aes(hour,count))

plot\_2 <- plot\_2+ geom\_point()

#plot\_2 <- plot\_2 + geom\_point(position=position\_jitter(w=1,h=0),aes(color = temp),alpah=0.5)

print(plot\_2 + theme\_bw())



To predict count of rented bikes in regards to the temp variable, a model has been created.

temp\_model <- lm(count ~. - datetime - atemp, Bike\_Shared\_Data)

print(summary(temp\_model))

##

## Call:

## lm(formula = count ~ . - datetime - atemp, data = Bike\_Shared\_Data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4.998e-11 -1.900e-14 -3.000e-15 1.300e-14 7.908e-11

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -5.974e-13 6.423e-14 -9.301e+00 < 2e-16 \*\*\*

## season -2.948e-13 1.028e-14 -2.868e+01 < 2e-16 \*\*\*

## holiday 2.854e-13 6.603e-14 4.323e+00 1.55e-05 \*\*\*

## workingday 2.057e-13 2.717e-14 7.572e+00 3.97e-14 \*\*\*

## weather 5.975e-14 1.874e-14 3.188e+00 0.00144 \*\*

## temp -8.023e-14 1.637e-15 -4.901e+01 < 2e-16 \*\*\*

## humidity 2.082e-14 7.178e-16 2.901e+01 < 2e-16 \*\*\*

## windspeed -4.461e-15 1.397e-15 -3.192e+00 0.00142 \*\*

## casual 1.000e+00 3.176e-16 3.149e+15 < 2e-16 \*\*\*

## registered 1.000e+00 9.050e-17 1.105e+16 < 2e-16 \*\*\*

## hour 7.101e-16 1.715e-15 4.140e-01 0.67891

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

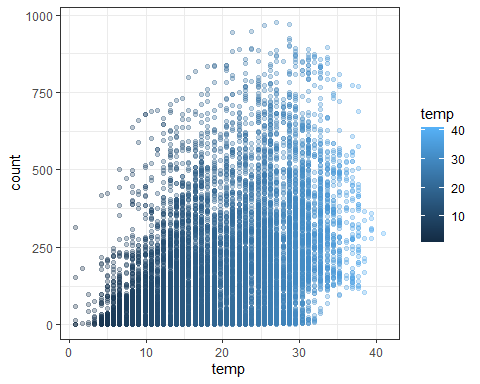
## Residual standard error: 1.108e-12 on 10875 degrees of freedom

## Multiple R-squared: 1, Adjusted R-squared: 1

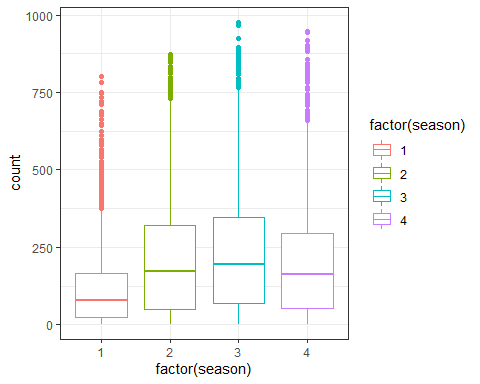
## F-statistic: 2.908e+31 on 10 and 10875 DF, p-value: < 2.2e-16

Results

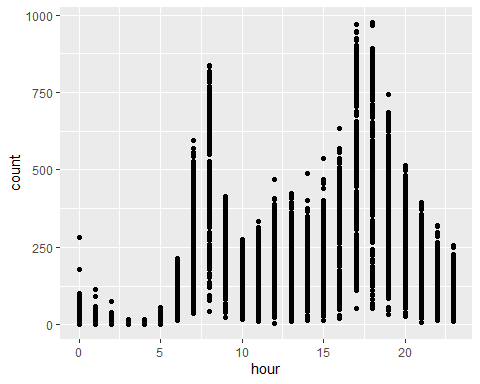
From the analysis above, we were able to identify an increase in the count of rented bikes with an increase in temperature.



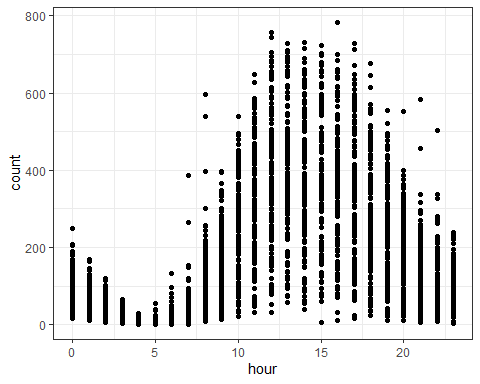
There are more bikes rented in winter compared to spring.



There is an increase in the count of rented bikes in the evening and morning hours, around 5 PM and 8 AM respectively.



During non-working days, fewer bikes are rented in the morning, but the number increases afternoon.



Implications

From the results of this project, the model did not provide a time series analysis, given the nature of the data. Therefore, I would recommend a model that can fully account for this dataset, especially when the dataset might keep growing.

Conclusion

From the analysis and results of this project, the theory that there is a lower demand for bikes when the temperatures are high compared to cold temperatures is not true. There is a higher demand for bikes at the time when temperatures are higher and a lower demand for bikes when temperatures are lower. In addition, the demand for bikes varies with the seasons. This is identified from the project since more bikes are rented in winter compared to spring.

References

Dell’Olio, L., Ibeas, A., & Moura, J. L. (2011, June). Implementing bike-sharing systems. In Proceedings of the Institution of Civil Engineers-Municipal Engineer (Vol. 164, No. 2, pp. 89-101). Thomas Telford Ltd.

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. Sustainable Cities and Society, 54, 101882.

Tong, Y., Chen, Y., Zhou, Z., Chen, L., Wang, J., Yang, Q., … & Lv, W. (2017, August). The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1653-1662).