Seoul Bike Sharing Demand Analysis and Prediction

Ahmad Sadeed (asadeed2), Deepa Nemmili Veeravalli (deepan2), Rui Zou (ruizou4)

2022-07-17

A Model for Prediction Goal

- Family,Form,Fit of Model
- Elements of a good model for Prediction

What Family, Fit and Form to use ?

Family : Parametric Model

Form : Linear Models

Fit : SLR, MLR and GLR Models

What is a good Prediction Model ?

Model assumptions applicability

```
- LINE - Not important for Prediction
- Data Analysis
    - Unusual Observations in Observed Data ?  Guard against over-fitting
        - Leverage, Outliers, Influence
        - Variable Selection
        - Transformations needed ?
```

Model Building and Diagnostics

Maximize R2, Adjusted R2, Multiple R2

```
- Compare Bigger Vs Smaller models
- Compare Models with predictor Interactions, higher order predictors
-  Variable Selection Precedures:
    AIC, BIC , Step and exhaustive
```

Minimize RMSE, LOOCV RMSE

```
- Train, test split
- Select 2-3 Models and compare and contrast
```

Description of the data file

This data file contains count of public bikes rented at each hour in Seoul Bike Sharing System with the corresponding weather data and holidays information. It has 14 variables and 8760 observations. We are interested in using Rented.Bike.Count (a numeric variable) as our response variable and explore how other factors (3 categorical variables and several continuous numeric variables) affect the count of bikes rented at each hour. Among the other 13 variables which we plan to use as potential predictors, we know from intuition that some may have more importance than others, like temperature, humidity, wind speed, visibility, seasons, and holiday, etc.

Background information on the data set

The original data comes from http://data.seoul.go.kr. The holiday information comes from SOUTH KOREA PUBLIC HOLIDAYS. A clean version can be found at UCI Machine Learning Repository.

Attribute Information:

- Date : month/day/year
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature - Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday, No holiday
- Functional Day - Functional or Non-functional days of rental bike system

Our Interest

This data set is interesting to us both personally and business-wise. Recently we have seen a rise in the delivery, accessibility, and usage of regular and electric rental bikes. There are clear environmental, health, and economical benefits associated with the usage of bikes as a mode of transportation. We would like to find out what factors lead to an increase in number of bikes rented and what factors have inverse effect on using rental bikes. Learning about such factors can help a bike rental business manage its inventory and supply without any hindrance. It can also help cities plan accordingly due to an increase of bikers, e.g. opening up more bike lanes during certain days or seasons. Environmentally, we will have a better understanding of the feasibility of turning a city into a “bike city” or looking at alternative options if a city is not friendly to bikers due to harsh weather conditions.

Data in R

The data file can be successfully loaded into R. We have printed out the structure and first few rows of the data file below.

The column names in the `csv` file contains measurement units (like `Wind speed (m/s)`, `Solar Radiation (MJ/m2)`) and characters such as ° and %. We load the data using cleaned up column names.

```
columns = c("Date","Rented.Bike.Count","Hour","Temperature","Humidity",
            "Wind.Speed","Visibility","Dew.point.temperature",
            "Solar.Radiation","Rainfall","Snowfall","Seasons","Holiday",
            "Functioning.Day")
bike = read.csv("../data/SeoulBikeData.csv", col.names = columns)
str(bike)
```

```
## 'data.frame':    8760 obs. of  14 variables:
##  $ Date                 : chr  "01/12/2017" "01/12/2017" "01/12/2017" "01/12/2017" ...
##  $ Rented.Bike.Count    : int  254 204 173 107 78 100 181 460 930 490 ...
##  $ Hour                 : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Temperature          : num  -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
##  $ Humidity             : int  37 38 39 40 36 37 35 38 37 27 ...
##  $ Wind.Speed           : num  2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
##  $ Visibility           : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 ...
##  $ Dew.point.temperature: num  -17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8 -22.4 ...
##  $ Solar.Radiation      : num  0 0 0 0 0 0 0 0 0.01 0.23 ...
##  $ Rainfall             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Snowfall             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Seasons              : chr  "Winter" "Winter" "Winter" "Winter" ...
##  $ Holiday              : chr  "No Holiday" "No Holiday" "No Holiday" "No Holiday" ...
##  $ Functioning.Day      : chr  "Yes" "Yes" "Yes" "Yes" ...
```

```
head(bike)
```

```
##         Date Rented.Bike.Count Hour Temperature Humidity Wind.Speed Visibility
## 1 01/12/2017               254    0        -5.2       37        2.2       2000
## 2 01/12/2017               204    1        -5.5       38        0.8       2000
## 3 01/12/2017               173    2        -6.0       39        1.0       2000
## 4 01/12/2017               107    3        -6.2       40        0.9       2000
## 5 01/12/2017                78    4        -6.0       36        2.3       2000
## 6 01/12/2017               100    5        -6.4       37        1.5       2000
##   Dew.point.temperature Solar.Radiation Rainfall Snowfall Seasons    Holiday
## 1                 -17.6               0        0        0  Winter No Holiday
## 2                 -17.6               0        0        0  Winter No Holiday
## 3                 -17.7               0        0        0  Winter No Holiday
## 4                 -17.6               0        0        0  Winter No Holiday
## 5                 -18.6               0        0        0  Winter No Holiday
## 6                 -18.7               0        0        0  Winter No Holiday
##   Functioning.Day
## 1             Yes
## 2             Yes
## 3             Yes
## 4             Yes
## 5             Yes
## 6             Yes
```

```
library(lubridate)
```

```
## 
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
## 
##     date, intersect, setdiff, union
```

```
bike$Date = as.Date(bike$Date,'%d/%m/%Y')
bike$year = as.numeric(format(bike$Date,'%Y'))
bike$month = as.numeric(format(bike$Date,'%m'))
bike$wday = wday(bike$Date)
```

Data Review

```
# brief review data
length(bike$Rented.Bike.Count)
```

```
## [1] 8760
```

```
bike[(which.max(bike$Rented.Bike.Count)),"Rented.Bike.Count"]
```

```
## [1] 3556
```

```
bike[(which.min(bike$Rented.Bike.Count)),"Rented.Bike.Count"]
```

```
## [1] 0
```

```
length(bike[(which(bike$Rented.Bike.Count >= mean(bike$Rented.Bike.Count))),"Rented.Bike.Count"])
```

```
## [1] 3522
```

```
length(bike[(which(bike$Rented.Bike.Count < mean(bike$Rented.Bike.Count))),"Rented.Bike.Count"])
```

```
## [1] 5238
```

```
median(bike$Rented.Bike.Count)
```

```
## [1] 504.5
```

```
# update with factors
bike$seasons = as.factor(bike$Seasons)
bike$holiday = as.factor(bike$Holiday)
bike$functioning.day = as.factor(bike$Functioning.Day)

bike_data = cbind(rented.bike.count = bike$Rented.Bike.Count,
            hour = bike$Hour,
            temp = bike$Temperature,
            wday = bike$wday,
            humdity = bike$Humidity,
            wind.speed = bike$Wind.Speed,
            visibility = bike$Visibility,
            dew.point.temp = bike$Dew.point.temperature,
            solar.radiation = bike$Solar.Radiation,
            rain  = bike$Rainfall,
            snow = bike$Snowfall,
            season = bike$seasons,
            holiday = bike$holiday,
            functioning.day = bike$functioning.day
            )
```

```
#pairs(bike_data, col = "dodgerblue")
```

```
# cor numeric
round(cor(bike_data), 2)
```

```
##                    rented.bike.count  hour  temp  wday humdity wind.speed
## rented.bike.count               1.00  0.41  0.54  0.03   -0.20       0.12
## hour                            0.41  1.00  0.12  0.00   -0.24       0.29
## temp                            0.54  0.12  1.00 -0.01    0.16      -0.04
## wday                            0.03  0.00 -0.01  1.00   -0.01       0.04
## humdity                        -0.20 -0.24  0.16 -0.01    1.00      -0.34
## wind.speed                      0.12  0.29 -0.04  0.04   -0.34       1.00
## visibility                      0.20  0.10  0.03  0.00   -0.54       0.17
## dew.point.temp                  0.38  0.00  0.91 -0.02    0.54      -0.18
## solar.radiation                 0.26  0.15  0.35  0.03   -0.46       0.33
## rain                           -0.12  0.01  0.05 -0.01    0.24      -0.02
## snow                           -0.14 -0.02 -0.22 -0.01    0.11       0.00
## season                          -0.25  0.00 -0.34  0.00   -0.12       0.11
## holiday                          0.07  0.00  0.06  0.06    0.05      -0.02
## functioning.day                  0.20  0.01 -0.05 -0.02   -0.02       0.01
##                   visibility dew.point.temp solar.radiation  rain  snow season
## rented.bike.count       0.20           0.38            0.26 -0.12 -0.14  -0.25
## hour                    0.10           0.00            0.15  0.01 -0.02   0.00
## temp                    0.03           0.91            0.35  0.05 -0.22  -0.34
## wday                    0.00          -0.02            0.03 -0.01 -0.01   0.00
## humdity                -0.54           0.54           -0.46  0.24  0.11  -0.12
## wind.speed              0.17          -0.18            0.33 -0.02  0.00   0.11
## visibility              1.00          -0.18            0.15 -0.17 -0.12  -0.01
## dew.point.temp         -0.18           1.00            0.09  0.13 -0.15  -0.33
## solar.radiation         0.15           0.09            1.00 -0.07 -0.07  -0.08
## rain                   -0.17           0.13           -0.07  1.00  0.01  -0.02
## snow                   -0.12          -0.15           -0.07  0.01  1.00   0.15
## season                 -0.01          -0.33           -0.08 -0.02  0.15   1.00
## holiday                -0.03           0.07            0.01  0.01  0.01  -0.05
## functioning.day        -0.03          -0.05           -0.01  0.00  0.03   0.22
##                   holiday functioning.day
## rented.bike.count    0.07            0.20
## hour                 0.00            0.01
## temp                 0.06           -0.05
## wday                 0.06           -0.02
## humdity              0.05           -0.02
## wind.speed          -0.02            0.01
## visibility          -0.03           -0.03
## dew.point.temp       0.07           -0.05
## solar.radiation      0.01           -0.01
## rain                 0.01            0.00
## snow                 0.01            0.03
## season              -0.05            0.22
## holiday              1.00            0.03
## functioning.day      0.03            1.00
```

```
# most collinearity of rented.bike.count with:
# temp, hour,dew.point.temp >= 0.38
# next with radiation,season >0.25,
# next visibility ,functioning.day >=0.2
```

Model Building for Prediction Goal

- Family,Form,Fit of Model
- Elements of a good model for Prediction

What Family, Fit and Form to use ?

Family : Parametric Model

Form : Linear Models

Fit : SLR, MLR and GLR Models

What is a good Prediction Model ?

Model assumptions applicability

```
- LINE ? Not important for Prediction
- Unusual Observations in Observed Data ?  Guard against over-fitting
    - Leverage, Outliers, Influence
    - Variable Selection
    - Transformations needed ?
```

Model Building and Diagnostics

Maximize R2, Adjusted R2, Multiple R2

```
- Compare Bigger Vs Smaller models
- Compare Models with predictor Interactions, higher order predictors
-  Variable Selection Precedures:
    AIC, BIC , Step and exhaustive
```

Minimize RMSE, LOOCV RMSE

```
- Train, test split
- Select 2-3 Models and compare and contrast
```

```r
# Goal:to find model for prediction for bike rental, we would use selection criteria that
# implicitly penalize larger models, such as LOOCV  RMSE
calc_loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}
# lm fit with all predictors: additive model
fit_all_additive = lm(rented.bike.count ~ . ,data = as.data.frame(bike_data))
summary(fit_all_additive)
```

```
##
## Call:
## lm(formula = rented.bike.count ~ ., data = as.data.frame(bike_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1247.1  -274.6   -56.1   207.1  2327.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.11e+03   1.13e+02   -9.87  < 2e-16 ***
## hour              2.72e+01   7.27e-01   37.45  < 2e-16 ***
## temp              1.68e+01   3.63e+00    4.62  3.9e-06 ***
## wday              1.21e+01   2.32e+00    5.22  1.8e-07 ***
## humdity          -1.11e+01   1.03e+00  -10.81  < 2e-16 ***
## wind.speed        1.67e+01   5.06e+00    3.29  0.00100 ***
## visibility        2.14e-02   9.53e-03    2.24  0.02506 *
## dew.point.temp   1.35e+01   3.82e+00    3.53  0.00041 ***
## solar.radiation -8.05e+01   7.57e+00  -10.64  < 2e-16 ***
## rain            -5.77e+01   4.27e+00  -13.52  < 2e-16 ***
## snow             3.51e+01   1.11e+01    3.16  0.00161 **
## season          -1.01e+02   4.58e+00  -22.05  < 2e-16 ***
## holiday          1.18e+02   2.15e+01    5.49  4.1e-08 ***
## functioning.day  9.39e+02   2.64e+01   35.64  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 433 on 8746 degrees of freedom
## Multiple R-squared:  0.551,  Adjusted R-squared:  0.55
## F-statistic:  825 on 13 and 8746 DF,  p-value: <2e-16
```

```
calc_loocv_rmse(fit_all_additive)
```

```
## [1] 433.1
```

```
# lm fit order 2 for collinear predictors noted in prior chunk
fit_order_2 =  lm(
   rented.bike.count ~ . ^ 2 + I(temp ^ 2) + I(hour ^ 2) + I(dew.point.temp)
   + I(solar.radiation ^ 2) + I(season ^ 2 + I(visibility ^ 2) + I(functioning.day ^2)),
  data = as.data.frame(bike_data))
summary(fit_all_additive)
```

```
##
## Call:
## lm(formula = rented.bike.count ~ ., data = as.data.frame(bike_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1247.1  -274.6   -56.1   207.1  2327.6
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.11e+03   1.13e+02   -9.87  < 2e-16 ***
## hour             2.72e+01   7.27e-01   37.45  < 2e-16 ***
## temp             1.68e+01   3.63e+00    4.62  3.9e-06 ***
## wday             1.21e+01   2.32e+00    5.22  1.8e-07 ***
## humdity         -1.11e+01   1.03e+00  -10.81  < 2e-16 ***
## wind.speed       1.67e+01   5.06e+00    3.29  0.00100 ***
## visibility       2.14e-02   9.53e-03    2.24  0.02506 *
## dew.point.temp   1.35e+01   3.82e+00    3.53  0.00041 ***
## solar.radiation -8.05e+01   7.57e+00  -10.64  < 2e-16 ***
## rain            -5.77e+01   4.27e+00  -13.52  < 2e-16 ***
## snow             3.51e+01   1.11e+01    3.16  0.00161 **
## season          -1.01e+02   4.58e+00  -22.05  < 2e-16 ***
## holiday          1.18e+02   2.15e+01    5.49  4.1e-08 ***
## functioning.day  9.39e+02   2.64e+01   35.64  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 433 on 8746 degrees of freedom
## Multiple R-squared:  0.551,  Adjusted R-squared:  0.55
## F-statistic:  825 on 13 and 8746 DF,  p-value: <2e-16
```

```
calc_loocv_rmse(fit_order_2)
```

```
## [1] 371.6
```

```
anova(fit_all_additive, fit_order_2)
```

```
## Analysis of Variance Table
##
## Model 1: rented.bike.count ~ hour + temp + wday + humdity + wind.speed +
##     visibility + dew.point.temp + solar.radiation + rain + snow +
##     season + holiday + functioning.day
## Model 2: rented.bike.count ~ (hour + temp + wday + humdity + wind.speed +
##     visibility + dew.point.temp + solar.radiation + rain + snow +
##     season + holiday + functioning.day)^2 + I(temp^2) + I(hour^2) +
##     I(dew.point.temp) + I(solar.radiation^2) + I(season^2 + I(visibility^2) +
##     I(functioning.day^2))
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1   8746 1.64e+09
## 2   8665 1.18e+09 81 456015183 41.3 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```