Seoul Bike Dataset

Andrew Danda, 구민우, 방설화

2024-04-22

Question

Can we predict the number of riders using the Seoul Bike Sharing based on the date and the weather?

Methods

We plan on using Lasso, Ridge, Decision tree and random forest for this dataset.

The Dataset

This is a dataset containing Seoul Bike sharing ridership from December 1, 2017 to November 30, 2018.

Dependent variable	
$Rented_{B}ike_{C}ount$	The number of bikes rented
Independent variables	
1. Time Varibles	
$Date$	The Date (dd/mm/yyyy)
$Hour$	The Hour (integer between 1 and 24)
$Holiday$	Dummy variable if the day is a holiday or not
$Weekend$	Dummy variable if the day is a weekend or not
$FunctionalDay$	Dummy variable if the bikes were functional or not
$Seasons_{S}pring$	Dummy variable if the season is Spring or not
$Seasons_{S}ummer$	Dummy variable if the season is Summer or not
$Seasons_{A}utumn$	Dummy variable if the season is Autumn or not
$Seasons_{W}inter$	Dummy variable if the season is Winter or not
2. Weather varibles	
$Temperature$	Temperature in Celsius
$Humidity$	Humidity (%)
$Wind_{S}peed$	Wind speed in meters per second
$Visibility$	Visibility in Kilometers
$Dew_{P}oint_{T}emperature$	Dew Point Temperature in Celsius
$Solor_{R}adiation$	Solar Radiation in millijoules Per square meter
$Rainfall$	Rainfall in millimeters
$Snowfall$	Snowfall in centimeters

```r
library(tidyverse)
library(dplyr)
library(fastDummies)
```

```r
bikeData <- read.csv("SeoulBikeData.csv", stringsAsFactors=FALSE, fileEncoding="latin1

# Clean dataset

# rename columns
bikeData <- bikeData %>%
  rename("Rented_Bike_Count" = "Rented.Bike.Count",
         "Temperature" = "Temperature..C.",
         "Humidity" = "Humidity...",
         "Wind_Speed"= "Wind.speed..m.s.",
         "Visibility" = "Visibility..10m.",
         "Dew_Point_Temperature" = "Dew.point.temperature..C.",
         "Solar_Radiation" = "Solar.Radiation..MJ.m2.",
         "Rainfall" = "Rainfall.mm.",
         "Snowfall" = "Snowfall..cm.",
         "Functioning_Day" = "Functioning.Day")

#divide the visibility by 100 to change it's units from 10s of meters to kilometers
bikeData$Visibility <- bikeData$Visibility / 100

#add weekend:
bikeData$Weekend <- ifelse(lubridate::wday(as.Date(bikeData$Date,format = "%d/%m/%Y"),
#lubridate::wday(as.Date("21/04/2024",format = "%d/%m/%Y"),label = TRUE, week_start =

# Dummy variables
bikeData$Holiday <- ifelse(bikeData$Holiday == "No Holiday", 0, 1)
bikeData$Functioning_Day <- ifelse(bikeData$Functioning_Day == "Yes", 1, 0)

#Holiday Dummies
bikeData <- bikeData %>% dummy_cols(select_columns = c("Seasons"))



#remove all data where functioning day is false
#We will only use data where the bikes are functioning.
bikeData <- bikeData %>% filter(Functioning_Day == 1)

summary(bikeData)
```

```
    Date            Rented_Bike_Count      Hour        Temperature
 Length:8465        Min.   :   2.0    Min.   : 0.00   Min.   :-17.80
 Class :character   1st Qu.: 214.0    1st Qu.: 6.00   1st Qu.:  3.00
 Mode  :character   Median : 542.0    Median :12.00   Median : 13.50
                    Mean   : 729.2    Mean   :11.51   Mean   : 12.77
                    3rd Qu.:1084.0    3rd Qu.:18.00   3rd Qu.: 22.70
                    Max.   :3556.0    Max.   :23.00   Max.   : 39.40
    Humidity        Wind_Speed        Visibility     Dew_Point_Temperature
 Min.   : 0.00   Min.   :0.000   Min.   : 0.27   Min.   :-30.600
 1st Qu.:42.00   1st Qu.:0.900   1st Qu.: 9.35   1st Qu.: -5.100
 Median :57.00   Median :1.500   Median :16.90   Median :  4.700
 Mean   :58.15   Mean   :1.726   Mean   :14.34   Mean   :  3.945
 3rd Qu.:74.00   3rd Qu.:2.300   3rd Qu.:20.00   3rd Qu.: 15.200
 Max.   :98.00   Max.   :7.400   Max.   :20.00   Max.   : 27.200
 Solar_Radiation     Rainfall           Snowfall          Seasons
 Min.   :0.0000   Min.   : 0.0000   Min.   :0.00000   Length:8465
 1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.00000   Class :character
 Median :0.0100   Median : 0.0000   Median :0.00000   Mode  :character
 Mean   :0.5679   Mean   : 0.1491   Mean   :0.07769
 3rd Qu.:0.9300   3rd Qu.: 0.0000   3rd Qu.:0.00000
 Max.   :3.5200   Max.   :35.0000   Max.   :8.80000
    Holiday        Functioning_Day    Weekend        Seasons_Autumn
 Min.   :0.0000   Min.   :1       Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1       1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :1       Median :0.0000   Median :0.0000
 Mean   :0.0482   Mean   :1       Mean   :0.2884   Mean   :0.2288
 3rd Qu.:0.0000   3rd Qu.:1       3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1       Max.   :1.0000   Max.   :1.0000
 Seasons_Spring   Seasons_Summer   Seasons_Winter
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.2552   Mean   :0.2608   Mean   :0.2552
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```