Bike Sharing Analysis Report

Brian Boughton and Brian Tulimero

2023-02-04

pacman::p\_load(pacman, dplyr)

# Milestone 1

## Introduction

The data set used was collected by Hadi fanaee from the laboratory of Artificial Intelligence and Decision Support at the University of Porto in Portugal. The contains information around bike sharing. It consist of an entry number, date, year, season, month, day of the week, hour, and whether or not it was a weekend. It also contains information around weather like temperature, wind speed and, forecast. Lastly, it gives the amount of casual and registered bike borrows along the total number.

Our goal with this data is to make predictions of future rentals based on day information and predicted forecasts. There are a few unnecessary variables in the dataset that have been removed like the instant number and date along with the casual and registered users because we are currently only interested in the overall users.

## Data Wrangling

### Importing data and setting needed categories to factors and removing the uneeded variables.

pj.dt<-read.csv("hour.csv")  
pj.df<-read.csv("hour.csv")  
pj.df$season<- as.factor(pj.df$season)  
pj.df$yr<- as.factor(pj.df$yr)  
pj.df$mnth<- as.factor(pj.df$mnth)  
pj.df$hr<- as.factor(pj.df$hr)  
pj.df$holiday<- as.factor(pj.df$holiday)  
pj.df$weekday<- as.factor(pj.df$weekday)  
pj.df$workingday<- as.factor(pj.df$workingday)  
pj.df$weathersit<- as.factor(pj.df$weathersit)  
pj<-subset(pj.df, select = -c(instant, dteday,casual,registered))

The final data frame is called simply pj for project and there are other version of the dataset if needed. Description of each feature: \* instant: record index \* dteday : date \* season : season (1:springer, 2:summer, 3:fall, 4:winter) \* yr : year (0: 2011, 1:2012) \* mnth : month ( 1 to 12) \* hr : hour (0 to 23) \* holiday : weather day is holiday or not (extracted from <http://dchr.dc.gov/page/holiday-schedule>) \* weekday : day of the week \* workingday : if day is neither weekend nor holiday is 1, otherwise is 0. \* weathersit : \* - 1: Clear, Few clouds, Partly cloudy, Partly cloudy \* - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist \* - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds \* - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog \* temp : Normalized temperature in Celsius. The values are divided to 41 (max) \* atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max) \* hum: Normalized humidity. The values are divided to 100 (max) \* windspeed: Normalized wind speed. The values are divided to 67 (max) \* casual: count of casual users \* registered: count of registered users \* cnt: count of total rental bikes including both casual and registered

### Getting to understand the data.

summary(pj)

## season yr mnth hr holiday weekday workingday  
## 1:4242 0:8645 5 :1488 16 : 730 0:16879 0:2502 0: 5514   
## 2:4409 1:8734 7 :1488 17 : 730 1: 500 1:2479 1:11865   
## 3:4496 12 :1483 13 : 729 2:2453   
## 4:4232 8 :1475 14 : 729 3:2475   
## 3 :1473 15 : 729 4:2471   
## 10 :1451 12 : 728 5:2487   
## (Other):8521 (Other):13004 6:2512   
## weathersit temp atemp hum windspeed   
## 1:11413 Min. :0.020 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 2: 4544 1st Qu.:0.340 1st Qu.:0.3333 1st Qu.:0.4800 1st Qu.:0.1045   
## 3: 1419 Median :0.500 Median :0.4848 Median :0.6300 Median :0.1940   
## 4: 3 Mean :0.497 Mean :0.4758 Mean :0.6272 Mean :0.1901   
## 3rd Qu.:0.660 3rd Qu.:0.6212 3rd Qu.:0.7800 3rd Qu.:0.2537   
## Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :0.8507   
##   
## cnt   
## Min. : 1.0   
## 1st Qu.: 40.0   
## Median :142.0   
## Mean :189.5   
## 3rd Qu.:281.0   
## Max. :977.0   
##

#showing number of Entries and features  
dim(pj)

## [1] 17379 13

#displaying the number of missing entries  
sum(is.na(pj))

## [1] 0

#showing numeric data features  
names(select\_if(pj, is.numeric))

## [1] "temp" "atemp" "hum" "windspeed" "cnt"

#showing categorical data features  
names(select\_if(pj, is.factor))

## [1] "season" "yr" "mnth" "hr" "holiday"   
## [6] "weekday" "workingday" "weathersit"

We can see that the data consist of 13 different variables and 17,379 observations. There is not missing data entries found within the data. The numeric features are temp, atemp, hum, windspeed, and cnt. the categorical features are season, yr, mnth, hr, holiday, weekday, workingday, and weathersit.

## Model development

### Creating training and test data frames

set.seed(1318)  
#Randomly choosing 70% of the data set  
train.rows<- sample(rownames(pj), dim(pj)[1]\*0.7)  
#Create the training set  
train<-pj[train.rows,]  
#Repeat for Test set  
valid.rows<-setdiff(rownames(pj),train.rows)  
valid<-pj[valid.rows,]

**Summaries**

summary(train)

## season yr mnth hr holiday weekday workingday  
## 1:2962 0:6030 7 :1040 15 : 531 0:11817 0:1733 0:3816   
## 2:3093 1:6135 3 :1033 17 : 528 1: 348 1:1716 1:8349   
## 3:3170 6 :1033 9 : 524 2:1734   
## 4:2940 8 :1033 13 : 520 3:1736   
## 5 :1031 11 : 517 4:1770   
## 12 :1024 21 : 517 5:1741   
## (Other):5971 (Other):9028 6:1735   
## weathersit temp atemp hum windspeed   
## 1:7962 Min. :0.0200 Min. :0.0000 Min. :0.000 Min. :0.0000   
## 2:3211 1st Qu.:0.3400 1st Qu.:0.3333 1st Qu.:0.480 1st Qu.:0.1045   
## 3: 989 Median :0.5000 Median :0.4848 Median :0.630 Median :0.1940   
## 4: 3 Mean :0.4974 Mean :0.4759 Mean :0.627 Mean :0.1897   
## 3rd Qu.:0.6600 3rd Qu.:0.6212 3rd Qu.:0.780 3rd Qu.:0.2537   
## Max. :1.0000 Max. :1.0000 Max. :1.000 Max. :0.8060   
##   
## cnt   
## Min. : 1.0   
## 1st Qu.: 40.0   
## Median :143.0   
## Mean :189.7   
## 3rd Qu.:281.0   
## Max. :977.0   
##

# dimensions of train data set  
dim(train)

## [1] 12165 13

summary(valid)

## season yr mnth hr holiday weekday workingday  
## 1:1280 0:2615 12 : 459 10 : 245 0:5062 0:769 0:1698   
## 2:1316 1:2599 5 : 457 7 : 232 1: 152 1:763 1:3516   
## 3:1326 7 : 448 12 : 231 2:719   
## 4:1292 10 : 444 8 : 230 3:739   
## 8 : 442 20 : 228 4:701   
## 9 : 441 0 : 225 5:746   
## (Other):2523 (Other):3823 6:777   
## weathersit temp atemp hum windspeed   
## 1:3451 Min. :0.020 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 2:1333 1st Qu.:0.340 1st Qu.:0.3333 1st Qu.:0.4800 1st Qu.:0.1045   
## 3: 430 Median :0.500 Median :0.4848 Median :0.6300 Median :0.1940   
## 4: 0 Mean :0.496 Mean :0.4754 Mean :0.6278 Mean :0.1910   
## 3rd Qu.:0.640 3rd Qu.:0.6212 3rd Qu.:0.7800 3rd Qu.:0.2537   
## Max. :0.960 Max. :0.9242 Max. :1.0000 Max. :0.8507   
##   
## cnt   
## Min. : 1.0   
## 1st Qu.: 39.0   
## Median :140.0   
## Mean :188.9   
## 3rd Qu.:280.0   
## Max. :976.0   
##

# dimensions of valid data set  
dim(valid)

## [1] 5214 13

### Creating a linear regression

# Creating null model  
null<-lm(cnt~1,data=pj)  
#Creating Full model  
full<-lm(cnt~.,data=pj)  
#Using stepwise function to find the optimal variable configuration for the model  
opti<-step(  
 full,  
 scope = list(upper = full, lower = null),  
 direction = "both",  
 trace = FALSE  
 )  
summary(opti)

##   
## Call:  
## lm(formula = cnt ~ season + yr + mnth + hr + holiday + weekday +   
## weathersit + temp + atemp + hum + windspeed, data = pj)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -393.87 -60.66 -7.96 51.31 439.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -83.630 6.633 -12.608 < 2e-16 \*\*\*  
## season2 38.178 4.856 7.862 4.00e-15 \*\*\*  
## season3 32.055 5.749 5.575 2.51e-08 \*\*\*  
## season4 67.994 4.882 13.928 < 2e-16 \*\*\*  
## yr1 85.431 1.563 54.658 < 2e-16 \*\*\*  
## mnth2 3.426 3.920 0.874 0.382185   
## mnth3 14.299 4.407 3.244 0.001179 \*\*   
## mnth4 6.230 6.548 0.951 0.341438   
## mnth5 20.657 7.007 2.948 0.003201 \*\*   
## mnth6 6.238 7.205 0.866 0.386617   
## mnth7 -13.269 8.082 -1.642 0.100645   
## mnth8 7.897 7.879 1.002 0.316222   
## mnth9 32.269 7.001 4.609 4.07e-06 \*\*\*  
## mnth10 15.843 6.483 2.444 0.014549 \*   
## mnth11 -9.840 6.238 -1.577 0.114744   
## mnth12 -6.256 4.954 -1.263 0.206718   
## hr1 -17.294 5.345 -3.236 0.001216 \*\*   
## hr2 -26.369 5.364 -4.916 8.91e-07 \*\*\*  
## hr3 -37.112 5.403 -6.869 6.67e-12 \*\*\*  
## hr4 -40.263 5.408 -7.445 1.01e-13 \*\*\*  
## hr5 -23.501 5.373 -4.374 1.23e-05 \*\*\*  
## hr6 35.393 5.359 6.605 4.10e-11 \*\*\*  
## hr7 170.418 5.348 31.864 < 2e-16 \*\*\*  
## hr8 310.801 5.342 58.183 < 2e-16 \*\*\*  
## hr9 163.101 5.347 30.501 < 2e-16 \*\*\*  
## hr10 108.444 5.370 20.196 < 2e-16 \*\*\*  
## hr11 133.843 5.409 24.742 < 2e-16 \*\*\*  
## hr12 173.142 5.456 31.735 < 2e-16 \*\*\*  
## hr13 168.102 5.494 30.600 < 2e-16 \*\*\*  
## hr14 152.249 5.525 27.558 < 2e-16 \*\*\*  
## hr15 161.707 5.535 29.213 < 2e-16 \*\*\*  
## hr16 223.834 5.524 40.522 < 2e-16 \*\*\*  
## hr17 377.535 5.491 68.750 < 2e-16 \*\*\*  
## hr18 345.587 5.455 63.350 < 2e-16 \*\*\*  
## hr19 236.919 5.404 43.841 < 2e-16 \*\*\*  
## hr20 157.293 5.375 29.266 < 2e-16 \*\*\*  
## hr21 107.840 5.353 20.147 < 2e-16 \*\*\*  
## hr22 70.907 5.343 13.272 < 2e-16 \*\*\*  
## hr23 32.112 5.338 6.015 1.83e-09 \*\*\*  
## holiday1 -26.228 4.881 -5.374 7.81e-08 \*\*\*  
## weekday1 9.275 2.973 3.120 0.001812 \*\*   
## weekday2 10.849 2.904 3.736 0.000187 \*\*\*  
## weekday3 13.625 2.900 4.698 2.64e-06 \*\*\*  
## weekday4 13.149 2.901 4.532 5.87e-06 \*\*\*  
## weekday5 17.445 2.892 6.032 1.65e-09 \*\*\*  
## weekday6 16.089 2.878 5.591 2.30e-08 \*\*\*  
## weathersit2 -10.409 1.920 -5.421 6.00e-08 \*\*\*  
## weathersit3 -65.189 3.236 -20.145 < 2e-16 \*\*\*  
## weathersit4 -62.580 58.893 -1.063 0.287970   
## temp 116.384 29.513 3.943 8.06e-05 \*\*\*  
## atemp 127.975 30.624 4.179 2.94e-05 \*\*\*  
## hum -82.802 5.554 -14.909 < 2e-16 \*\*\*  
## windspeed -29.167 7.052 -4.136 3.55e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 101.7 on 17326 degrees of freedom  
## Multiple R-squared: 0.6864, Adjusted R-squared: 0.6854   
## F-statistic: 729.1 on 52 and 17326 DF, p-value: < 2.2e-16