Bike Sharing Analysis Report

Brian Boughton and Brian Tulimero

2023-02-04

pacman::p\_load(pacman, dplyr,forecast)

# 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("day.csv")  
pj.df<-read.csv("day.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$holiday<- as.factor(pj.df$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:spring, 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 holiday weekday workingday weathersit  
## 1:181 0:365 1 : 62 0:710 0:105 0:231 1:463   
## 2:184 1:366 3 : 62 1: 21 1:105 1:500 2:247   
## 3:188 5 : 62 2:104 3: 21   
## 4:178 7 : 62 3:104   
## 8 : 62 4:104   
## 10 : 62 5:104   
## (Other):359 6:105   
## temp atemp hum windspeed   
## Min. :0.05913 Min. :0.07907 Min. :0.0000 Min. :0.02239   
## 1st Qu.:0.33708 1st Qu.:0.33784 1st Qu.:0.5200 1st Qu.:0.13495   
## Median :0.49833 Median :0.48673 Median :0.6267 Median :0.18097   
## Mean :0.49538 Mean :0.47435 Mean :0.6279 Mean :0.19049   
## 3rd Qu.:0.65542 3rd Qu.:0.60860 3rd Qu.:0.7302 3rd Qu.:0.23321   
## Max. :0.86167 Max. :0.84090 Max. :0.9725 Max. :0.50746   
##   
## cnt   
## Min. : 22   
## 1st Qu.:3152   
## Median :4548   
## Mean :4504   
## 3rd Qu.:5956   
## Max. :8714   
##

#showing number of Entries and features  
dim(pj)

## [1] 731 12

#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" "holiday" "weekday"   
## [6] "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 holiday weekday workingday weathersit  
## 1:118 0:257 10 : 51 0:497 0:72 0:160 1:316   
## 2:131 1:254 4 : 46 1: 14 1:69 1:351 2:182   
## 3:129 5 : 46 2:72 3: 13   
## 4:133 9 : 44 3:73   
## 11 : 44 4:79   
## 7 : 43 5:72   
## (Other):237 6:74   
## temp atemp hum windspeed   
## Min. :0.05913 Min. :0.07907 Min. :0.1879 Min. :0.02239   
## 1st Qu.:0.34333 1st Qu.:0.34203 1st Qu.:0.5210 1st Qu.:0.13558   
## Median :0.50000 Median :0.48926 Median :0.6317 Median :0.18283   
## Mean :0.49496 Mean :0.47456 Mean :0.6298 Mean :0.19224   
## 3rd Qu.:0.65042 3rd Qu.:0.60481 3rd Qu.:0.7302 3rd Qu.:0.23508   
## Max. :0.86167 Max. :0.84090 Max. :0.9725 Max. :0.50746   
##   
## cnt   
## Min. : 22   
## 1st Qu.:3209   
## Median :4553   
## Mean :4531   
## 3rd Qu.:5912   
## Max. :8714   
##

# dimensions of train data set  
dim(train)

## [1] 511 12

summary(valid)

## season yr mnth holiday weekday workingday weathersit  
## 1:63 0:108 3 :26 0:213 0:33 0: 71 1:147   
## 2:53 1:112 6 :25 1: 7 1:36 1:149 2: 65   
## 3:59 1 :21 2:32 3: 8   
## 4:45 12 :20 3:31   
## 7 :19 4:25   
## 8 :19 5:32   
## (Other):90 6:31   
## temp atemp hum windspeed   
## Min. :0.1383 Min. :0.1162 Min. :0.0000 Min. :0.04665   
## 1st Qu.:0.3217 1st Qu.:0.3196 1st Qu.:0.5177 1st Qu.:0.13308   
## Median :0.4963 Median :0.4779 Median :0.6217 Median :0.17755   
## Mean :0.4964 Mean :0.4739 Mean :0.6235 Mean :0.18641   
## 3rd Qu.:0.6746 3rd Qu.:0.6225 3rd Qu.:0.7231 3rd Qu.:0.23136   
## Max. :0.8383 Max. :0.7904 Max. :0.9625 Max. :0.38807   
##   
## cnt   
## Min. : 441   
## 1st Qu.:2912   
## Median :4540   
## Mean :4443   
## 3rd Qu.:6063   
## Max. :8362   
##

# dimensions of valid data set  
dim(valid)

## [1] 220 12

### Creating a linear regression

# Creating null model  
null<-lm(cnt~1,data=train)  
#Creating Full model  
full<-lm(cnt~.,data=train)  
#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 + holiday + weekday + weathersit +   
## atemp + hum + windspeed, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3410.9 -352.2 70.7 463.7 2750.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1489.66 286.54 5.199 2.97e-07 \*\*\*  
## season2 666.49 232.24 2.870 0.004287 \*\*   
## season3 582.35 274.55 2.121 0.034420 \*   
## season4 1215.35 222.68 5.458 7.72e-08 \*\*\*  
## yr1 2054.11 69.37 29.610 < 2e-16 \*\*\*  
## mnth2 170.85 172.88 0.988 0.323521   
## mnth3 793.18 209.85 3.780 0.000177 \*\*\*  
## mnth4 816.76 308.70 2.646 0.008415 \*\*   
## mnth5 1298.87 330.13 3.934 9.56e-05 \*\*\*  
## mnth6 1045.89 347.22 3.012 0.002730 \*\*   
## mnth7 697.10 383.12 1.820 0.069446 .   
## mnth8 1045.41 370.21 2.824 0.004941 \*\*   
## mnth9 1525.65 328.73 4.641 4.47e-06 \*\*\*  
## mnth10 1106.63 296.16 3.737 0.000209 \*\*\*  
## mnth11 306.22 282.46 1.084 0.278855   
## mnth12 301.78 227.80 1.325 0.185869   
## holiday1 -504.53 219.67 -2.297 0.022059 \*   
## weekday1 230.73 132.84 1.737 0.083046 .   
## weekday2 367.09 128.41 2.859 0.004436 \*\*   
## weekday3 355.90 127.22 2.798 0.005355 \*\*   
## weekday4 384.81 125.75 3.060 0.002336 \*\*   
## weekday5 337.30 127.99 2.635 0.008677 \*\*   
## weekday6 478.46 127.09 3.765 0.000187 \*\*\*  
## weathersit2 -395.55 90.95 -4.349 1.67e-05 \*\*\*  
## weathersit3 -2190.44 251.41 -8.713 < 2e-16 \*\*\*  
## atemp 4258.56 529.01 8.050 6.49e-15 \*\*\*  
## hum -1571.40 357.80 -4.392 1.38e-05 \*\*\*  
## windspeed -2684.40 465.85 -5.762 1.48e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 759.6 on 483 degrees of freedom  
## Multiple R-squared: 0.8469, Adjusted R-squared: 0.8384   
## F-statistic: 98.99 on 27 and 483 DF, p-value: < 2.2e-16

accuracy(opti$fitted.values, train$cnt)

## ME RMSE MAE MPE MAPE  
## Test set -2.228181e-14 738.517 544.0122 -33.39244 45.93015

pred <- predict(opti, newdata = valid)  
accuracy(pred, valid$cnt)

## ME RMSE MAE MPE MAPE  
## Test set -23.02409 821.3445 581.8071 -5.999266 22.421