Mod3Assign1

Adwoa Kessie

2/5/2022

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.2

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(tidymodels)

## Warning: package 'tidymodels' was built under R version 4.1.2

## Registered S3 method overwritten by 'tune':  
## method from   
## required\_pkgs.model\_spec parsnip

## -- Attaching packages -------------------------------------- tidymodels 0.1.4 --

## v broom 0.7.9 v rsample 0.1.1   
## v dials 0.0.10 v tune 0.1.6   
## v infer 1.0.0 v workflows 0.2.4   
## v modeldata 0.1.1 v workflowsets 0.1.0   
## v parsnip 0.1.7 v yardstick 0.0.9   
## v recipes 0.1.17

## Warning: package 'dials' was built under R version 4.1.2

## Warning: package 'infer' was built under R version 4.1.2

## Warning: package 'modeldata' was built under R version 4.1.2

## Warning: package 'parsnip' was built under R version 4.1.2

## Warning: package 'recipes' was built under R version 4.1.2

## Warning: package 'rsample' was built under R version 4.1.2

## Warning: package 'tune' was built under R version 4.1.2

## Warning: package 'workflows' was built under R version 4.1.2

## Warning: package 'workflowsets' was built under R version 4.1.2

## Warning: package 'yardstick' was built under R version 4.1.2

## -- Conflicts ----------------------------------------- tidymodels\_conflicts() --  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()  
## \* Search for functions across packages at https://www.tidymodels.org/find/

bike = read.csv('bike\_cleaned-2.csv')

bike = bike %>% mutate(dteday = mdy(dteday)) #mdy is a lubridate package function

bike = bike %>% mutate\_if(is.character, as\_factor)

bike = bike %>% mutate(hr = as\_factor(hr))

set.seed(1234)  
bike\_split = initial\_split(bike, prop = 0.70, strata = count)  
train = training(bike\_split)  
test = testing(bike\_split)

## Training set has 12163 rows, test set has 5216 rows

model = lm(count~season+mnth+hr+holiday+weekday+temp+weathersit, data=train)  
summary(model)

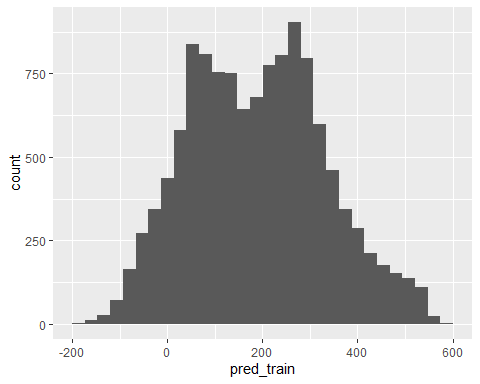
##   
## Call:  
## lm(formula = count ~ season + mnth + hr + holiday + weekday +   
## temp + weathersit, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -427.33 -62.08 -9.82 51.84 503.54   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -81.6699 6.9466 -11.757 < 2e-16 \*\*\*  
## seasonSpring 27.4972 6.3951 4.300 1.72e-05 \*\*\*  
## seasonSummer 18.7645 7.5881 2.473 0.01342 \*   
## seasonFall 62.5367 6.4533 9.691 < 2e-16 \*\*\*  
## mnthFeb -0.5997 5.1373 -0.117 0.90707   
## mnthMar 3.0778 5.7904 0.532 0.59506   
## mnthApr -1.3130 8.6231 -0.152 0.87898   
## mnthMay -2.6894 9.2230 -0.292 0.77060   
## mnthJun -15.8125 9.4879 -1.667 0.09562 .   
## mnthJul -40.2300 10.6077 -3.793 0.00015 \*\*\*  
## mnthAug -16.4993 10.3574 -1.593 0.11119   
## mnthSep 3.9859 9.2187 0.432 0.66548   
## mnthOct -3.0817 8.5334 -0.361 0.71800   
## mnthNov -14.7632 8.2403 -1.792 0.07322 .   
## mnthDec -16.2734 6.5606 -2.480 0.01313 \*   
## hr1 -20.7836 6.9908 -2.973 0.00295 \*\*   
## hr2 -29.0673 6.9980 -4.154 3.29e-05 \*\*\*  
## hr3 -41.4592 7.0968 -5.842 5.29e-09 \*\*\*  
## hr4 -41.2506 7.0386 -5.861 4.73e-09 \*\*\*  
## hr5 -27.2665 6.9794 -3.907 9.41e-05 \*\*\*  
## hr6 31.8318 7.0125 4.539 5.70e-06 \*\*\*  
## hr7 164.5446 7.0278 23.413 < 2e-16 \*\*\*  
## hr8 305.3583 6.9782 43.759 < 2e-16 \*\*\*  
## hr9 163.9524 7.0096 23.390 < 2e-16 \*\*\*  
## hr10 105.9395 6.9986 15.137 < 2e-16 \*\*\*  
## hr11 138.1987 6.9861 19.782 < 2e-16 \*\*\*  
## hr12 179.5246 6.9799 25.720 < 2e-16 \*\*\*  
## hr13 177.5739 7.0533 25.176 < 2e-16 \*\*\*  
## hr14 152.0364 7.1106 21.382 < 2e-16 \*\*\*  
## hr15 170.3496 7.0967 24.004 < 2e-16 \*\*\*  
## hr16 229.1493 7.1110 32.225 < 2e-16 \*\*\*  
## hr17 384.6252 7.0221 54.774 < 2e-16 \*\*\*  
## hr18 342.3854 7.0387 48.643 < 2e-16 \*\*\*  
## hr19 236.7980 7.0437 33.618 < 2e-16 \*\*\*  
## hr20 158.1195 7.0488 22.432 < 2e-16 \*\*\*  
## hr21 107.9022 6.9453 15.536 < 2e-16 \*\*\*  
## hr22 72.0674 6.9890 10.312 < 2e-16 \*\*\*  
## hr23 31.3404 7.0004 4.477 7.64e-06 \*\*\*  
## holidayHoliday -25.5839 6.3712 -4.016 5.97e-05 \*\*\*  
## weekdaySunday -12.8572 3.7603 -3.419 0.00063 \*\*\*  
## weekdayMonday -8.6638 3.8974 -2.223 0.02623 \*   
## weekdayTuesday -6.7687 3.8295 -1.768 0.07716 .   
## weekdayWednesday -3.6852 3.8010 -0.970 0.33231   
## weekdayThursday -3.1739 3.8047 -0.834 0.40418   
## weekdayFriday 0.5683 3.7761 0.151 0.88036   
## temp 293.4586 12.1953 24.063 < 2e-16 \*\*\*  
## weathersitMisty -19.7902 2.3715 -8.345 < 2e-16 \*\*\*  
## weathersitLightPrecip -92.1438 3.8276 -24.073 < 2e-16 \*\*\*  
## weathersitHeavyPrecip -78.2430 64.7522 -1.208 0.22694   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 111.8 on 12114 degrees of freedom  
## Multiple R-squared: 0.6224, Adjusted R-squared: 0.6209   
## F-statistic: 416.1 on 48 and 12114 DF, p-value: < 2.2e-16

## The model has Adjusted R2 of 0.6209. hr is a significant variable because all the hour values are significant. For month nd weekday, only a few values are significant.

pred\_train = predict(model, train)  
pred\_train\_df = data.frame(pred\_train)

ggplot(pred\_train\_df, aes(x=pred\_train)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## The predicted values look to be normally distributed around 200 However, this is not perfectly normal distribution as there is a dip in the histogram right at the center.

Rsquared = function (x, y) cor(x, y)^2  
RsquaredTest = Rsquared(predict(model, test), test$count)

## Training model has R2 of 0.6209 and test has the R2 of 0.627 which are pretty close to each other. However, since in general, if the R2 of test > R2 of training, it might imply overfitting. In this case though, the difference is minimal.