Model Validation Assignment Quiz

2024-09-16

Libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(tidymodels)
```

```
## ── Attaching packages ─────────────────────────────────────── tidymodels
1.2.0 ──
## ✓ broom        1.0.6     ✓ rsample      1.2.1
## ✓ dials        1.3.0     ✓ tune         1.2.1
## ✓ infer        1.0.7     ✓ workflows    1.1.4
## ✓ modeldata    1.4.0     ✓ workflowsets 1.1.0
## ✓ parsnip      1.2.1     ✓ yardstick    1.3.1
## ✓ recipes      1.1.0
## ── Conflicts ────────────────────────────────────────
tidymodels_conflicts() ──
## ✖ scales::discard() masks purrr::discard()
## ✖ dplyr::filter()   masks stats::filter()
## ✖ recipes::fixed()  masks stringr::fixed()
## ✖ dplyr::lag()      masks stats::lag()
## ✖ yardstick::spec() masks readr::spec()
## ✖ recipes::step()   masks stats::step()
## • Use suppressPackageStartupMessages() to eliminate package startup
messages
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(lubridate)
```

Read in Data

```
library(readr)
bike_cleaned_4 <- read_csv("bike_cleaned-4.csv",
    col_types = cols(dteday = col_date(format = "%m/%d/%Y")))
View(bike_cleaned_4)
```

Convert Characters variable to factors

```
bike = bike_cleaned_4 %>%
  mutate_if(is.character, as.factor)
bike = bike %>%
  mutate(hr=as.factor(hr))

summary(bike)
```

```
##     instant          dteday              season          mnth
##  Min.   :    1   Min.   :2011-01-01   Fall  :4232   Jul    :1488
##  1st Qu.: 4346   1st Qu.:2011-07-04   Spring:4409   May    :1488
##  Median : 8690   Median :2012-01-02   Summer:4496   Dec    :1483
##  Mean   : 8690   Mean   :2012-01-02   Winter:4242   Aug    :1475
##  3rd Qu.:13034   3rd Qu.:2012-07-02                 Mar    :1473
##  Max.   :17379   Max.   :2012-12-31                 Oct    :1451
##                                                     (Other):8521
##        hr              holiday           weekday            workingday
##  16     :  730   Holiday   :  500   Friday   :2487   NotWorkingDay: 5514
##  17     :  730   NotHoliday:16879   Monday   :2479   WorkingDay   :11865
##  13     :  729                      Saturday :2512
##  14     :  729                      Sunday   :2502
##  15     :  729                      Thursday :2471
##  12     :  728                      Tuesday  :2453
##  (Other):13004                      Wednesday:2475
##       weathersit         temp            atemp              hum
##  HeavyPrecip:    3   Min.   :0.020   Min.   :0.0000   Min.   :0.0000
##  LightPrecip: 1419   1st Qu.:0.340   1st Qu.:0.3333   1st Qu.:0.4800
##  Misty      : 4544   Median :0.500   Median :0.4848   Median :0.6300
##  NoPrecip   :11413   Mean   :0.497   Mean   :0.4758   Mean   :0.6272
##                      3rd Qu.:0.660   3rd Qu.:0.6212   3rd Qu.:0.7800
##                      Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##
##    windspeed          casual         registered        count
##  Min.   :0.0000   Min.   :  0.00   Min.   :  0.0   Min.   :  1.0
##  1st Qu.:0.1045   1st Qu.:  4.00   1st Qu.: 34.0   1st Qu.: 40.0
##  Median :0.1940   Median : 17.00   Median :115.0   Median :142.0
##  Mean   :0.1901   Mean   : 35.68   Mean   :153.8   Mean   :189.5
##  3rd Qu.:0.2537   3rd Qu.: 48.00   3rd Qu.:220.0   3rd Qu.:281.0
##  Max.   :0.8507   Max.   :367.00   Max.   :886.0   Max.   :977.0
##
```

```r
str(bike)
```

```
## tibble [17,379 × 16] (S3: tbl_df/tbl/data.frame)
##  $ instant   : num [1:17379] 1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : Date[1:17379], format: "2011-01-01" "2011-01-01" ...
##  $ season    : Factor w/ 4 levels "Fall","Spring",..: 4 4 4 4 4 4 4 4 4 4
...
##  $ mnth      : Factor w/ 12 levels "Apr","Aug","Dec",..: 5 5 5 5 5 5 5 5 5
5 ...
##  $ hr        : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9
10 ...
##  $ holiday   : Factor w/ 2 levels "Holiday","NotHoliday": 2 2 2 2 2 2 2 2
2 2 ...
##  $ weekday   : Factor w/ 7 levels "Friday","Monday",..: 3 3 3 3 3 3 3 3 3
3 ...
##  $ workingday: Factor w/ 2 levels "NotWorkingDay",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ weathersit: Factor w/ 4 levels "HeavyPrecip",..: 4 4 4 4 4 3 4 4 4 4
...
##  $ temp      : num [1:17379] 0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24
0.32 ...
##  $ atemp     : num [1:17379] 0.288 0.273 0.273 0.288 0.288 ...
##  $ hum       : num [1:17379] 0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75
0.76 ...
##  $ windspeed : num [1:17379] 0 0 0 0 0 0.0896 0 0 0 0 ...
##  $ casual    : num [1:17379] 3 8 5 3 0 0 2 1 1 8 ...
##  $ registered: num [1:17379] 13 32 27 10 1 1 0 2 7 6 ...
##  $ count     : num [1:17379] 16 40 32 13 1 1 2 3 8 14 ...
```

split the data into training and testing sets. 70% of data to training. use random number (set.seed) of 1234. Stratified by the “count” variable

```r
set.seed(1234)
bike_split = initial_split(bike, prop = 0.70, strata = count)
train = training(bike_split)
test = testing(bike_split)
```

Linear Regression Model

```r
bike_recipe = recipe(count ~ season + mnth + hr + holiday + weekday + temp +
weathersit, train)%>%
  step_dummy(all_nominal())

lm_model = linear_reg()%>%
  set_engine("lm")

lm_wflow = workflow()%>%
  add_model(lm_model)%>%
  add_recipe(bike_recipe)
```

```
lm_fit = fit(lm_wflow, train)
```

Summary

```
summary(lm_fit$fit$fit$fit)
```

```
##
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## 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)             -123.7048    66.1177  -1.871 0.061372 .
## temp                     293.4586    12.1953  24.063  < 2e-16 ***
## season_Spring            -35.0395     7.5737  -4.626 3.76e-06 ***
## season_Summer            -43.7722     6.8705  -6.371 1.95e-10 ***
## season_Winter            -62.5367     6.4533  -9.691  < 2e-16 ***
## mnth_Aug                 -15.1863     8.4226  -1.803 0.071405 .
## mnth_Dec                 -14.9604     8.4415  -1.772 0.076380 .
## mnth_Feb                   0.7133     8.4470   0.084 0.932706
## mnth_Jan                   1.3130     8.6231   0.152 0.878982
## mnth_Jul                 -38.9170     8.5386  -4.558 5.22e-06 ***
## mnth_Jun                 -14.4995     5.9791  -2.425 0.015321 *
## mnth_Mar                   4.3908     6.5373   0.672 0.501819
## mnth_May                  -1.3764     5.1503  -0.267 0.789277
## mnth_Nov                 -13.4502     9.2393  -1.456 0.145485
## mnth_Oct                  -1.7687     9.0406  -0.196 0.844894
## mnth_Sep                   5.2989     7.9195   0.669 0.503449
## hr_X1                    -20.7836     6.9908  -2.973 0.002955 **
## hr_X2                    -29.0673     6.9980  -4.154 3.29e-05 ***
## hr_X3                    -41.4592     7.0968  -5.842 5.29e-09 ***
## hr_X4                    -41.2506     7.0386  -5.861 4.73e-09 ***
## hr_X5                    -27.2665     6.9794  -3.907 9.41e-05 ***
## hr_X6                     31.8318     7.0125   4.539 5.70e-06 ***
## hr_X7                    164.5446     7.0278  23.413  < 2e-16 ***
## hr_X8                    305.3583     6.9782  43.759  < 2e-16 ***
## hr_X9                    163.9524     7.0096  23.390  < 2e-16 ***
## hr_X10                   105.9395     6.9986  15.137  < 2e-16 ***
## hr_X11                   138.1987     6.9861  19.782  < 2e-16 ***
## hr_X12                   179.5246     6.9799  25.720  < 2e-16 ***
## hr_X13                   177.5739     7.0533  25.176  < 2e-16 ***
## hr_X14                   152.0364     7.1106  21.382  < 2e-16 ***
## hr_X15                   170.3496     7.0967  24.004  < 2e-16 ***
## hr_X16                   229.1493     7.1110  32.225  < 2e-16 ***
## hr_X17                   384.6252     7.0221  54.774  < 2e-16 ***
```

```
## hr_X18                   342.3854     7.0387  48.643  < 2e-16 ***
## hr_X19                   236.7980     7.0437  33.618  < 2e-16 ***
## hr_X20                   158.1195     7.0488  22.432  < 2e-16 ***
## hr_X21                   107.9022     6.9453  15.536  < 2e-16 ***
## hr_X22                    72.0674     6.9890  10.312  < 2e-16 ***
## hr_X23                    31.3404     7.0004   4.477 7.64e-06 ***
## holiday_NotHoliday        25.5839     6.3712   4.016 5.97e-05 ***
## weekday_Monday            -9.2322     3.8759  -2.382 0.017238 *
## weekday_Saturday          -0.5683     3.7761  -0.151 0.880363
## weekday_Sunday           -13.4256     3.7705  -3.561 0.000371 ***
## weekday_Thursday          -3.7422     3.8044  -0.984 0.325297
## weekday_Tuesday           -7.3370     3.8298  -1.916 0.055420 .
## weekday_Wednesday         -4.2535     3.8010  -1.119 0.263137
## weathersit_LightPrecip   -13.9008    64.8336  -0.214 0.830233
## weathersit_Misty          58.4528    64.7679   0.902 0.366811
## weathersit_NoPrecip       78.2430    64.7522   1.208 0.226938
## ---
## 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
```

See results on the test set

```
lm_fit %>% predict(test) %>% bind_cols(test) %>% metrics(truth = count,
estimate = .pred)
```

```
## # A tibble: 3 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     110.
## 2 rsq     standard       0.627
## 3 mae     standard      80.1
```

```
predict_train = lm_fit %>% predict(test) %>% bind_cols(test) %>%
metrics(truth = count, estimate = .pred)
```

Develop Histogram

```
library(esquisse)
```