high-r-sq-model.rmd

Avinash Tiwari

2022-07-30

Analysis Workbook

Proposal

Define Helper Functions

EDA

```
# set seed
#set.seed(11)
set.seed(420072022)
# Load libraries
library(ggplot2)
library(faraway)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
ca_housing_data = read.csv('../000_Data/california-housing-prices/housing.csv')
ca_housing_data = subset(ca_housing_data,
                         subset = ca_housing_data$ocean_proximity != "ISLAND")
#head(ca_housing_data)
#View(ca_housing_data)
#str(ca_housing_data)
```

```
# delete rows with missing values - as only 207 rows have missing value for only one variable "total_be
nrow(ca_housing_data)
```

```
## [1] 20635
```

```
cah_data = na.omit(ca_housing_data)
nrow(ca_housing_data) - nrow(cah_data)
```

```
## [1] 207
```

Data cleaning

```
cah_trn_idx  = sample(nrow(cah_data), size = trunc(0.70 * nrow(cah_data)))
cah_trn_data = cah_data[cah_trn_idx, ]
cah_tst_data = cah_data[-cah_trn_idx, ]
```

Variable Creation

```
cah_trn_data$ocean_proximity = as.factor(cah_trn_data$ocean_proximity)
cah_trn_data$pop_per_hh = cah_trn_data$population / cah_trn_data$households
cah_trn_data$rooms_per_hh = cah_trn_data$total_rooms / cah_trn_data$households
cah_trn_data$bedrooms_per_hh = cah_trn_data$total_bedrooms / cah_trn_data$households
```

Model Building

```
mod  = lm(log(median_house_value) ~ longitude + housing_median_age + median_income + pop_per_hh + bedroo
#summary(mod)
#vif(mod)
```

```
mod_fix_int = lm (log(median_house_value) ~ (longitude + housing_median_age + median_income + pop_per_h
                  data = cah_trn_data,
                  subset = (cooks.distance(mod) <= 4 /length(cooks.distance(mod))))
summary(mod_fix_int)
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ (longitude + housing_median_age +
##     median_income + pop_per_hh + bedrooms_per_hh + ocean_proximity)^2,
##     data = cah_trn_data, subset = (cooks.distance(mod) <= 4/length(cooks.distance(mod))))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2475 -0.1783 -0.0117  0.1657  1.8371
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                   4.90e+00   1.24e+00    3.95
## longitude                                    -5.80e-02   1.03e-02   -5.61
## housing_median_age                            7.43e-02   1.54e-02    4.84
## median_income                                 1.20e-01   1.08e-01    1.11
## pop_per_hh                                    2.18e-01   1.84e-01    1.19
## bedrooms_per_hh                               3.23e+00   6.79e-01    4.76
## ocean_proximityINLAND                        -8.55e-01   4.14e-01   -2.07
## ocean_proximityNEAR BAY                      -8.93e+01   6.55e+00  -13.64
## ocean_proximityNEAR OCEAN                    -5.51e+00   4.80e-01  -11.48
## longitude:housing_median_age                  6.01e-04   1.28e-04    4.70
## longitude:median_income                       3.59e-04   9.00e-04    0.40
## longitude:pop_per_hh                          3.32e-03   1.54e-03    2.16
## longitude:bedrooms_per_hh                     2.47e-02   5.63e-03    4.39
## longitude:ocean_proximityINLAND              -1.81e-03   3.43e-03   -0.53
## longitude:ocean_proximityNEAR BAY            -7.30e-01   5.36e-02  -13.62
## longitude:ocean_proximityNEAR OCEAN          -4.62e-02   4.07e-03  -11.34
## housing_median_age:median_income              8.70e-04   1.17e-04    7.46
## housing_median_age:pop_per_hh                -8.35e-04   2.26e-04   -3.69
## housing_median_age:bedrooms_per_hh           -2.75e-04   1.16e-03   -0.24
## housing_median_age:ocean_proximityINLAND     -4.09e-03   5.26e-04   -7.78
## housing_median_age:ocean_proximityNEAR BAY   -2.68e-05   8.17e-04   -0.03
## housing_median_age:ocean_proximityNEAR OCEAN  5.07e-04   6.96e-04    0.73
## median_income:pop_per_hh                      1.54e-02   1.17e-03   13.20
## median_income:bedrooms_per_hh                 9.98e-03   6.30e-03    1.58
## median_income:ocean_proximityINLAND           7.85e-02   3.95e-03   19.87
## median_income:ocean_proximityNEAR BAY         2.69e-02   5.61e-03    4.79
## median_income:ocean_proximityNEAR OCEAN       1.99e-02   4.57e-03    4.36
## pop_per_hh:bedrooms_per_hh                    2.16e-02   8.52e-03    2.54
## pop_per_hh:ocean_proximityINLAND              9.09e-02   5.66e-03   16.04
## pop_per_hh:ocean_proximityNEAR BAY            3.53e-02   1.30e-02    2.71
## pop_per_hh:ocean_proximityNEAR OCEAN          2.21e-02   8.93e-03    2.47
## bedrooms_per_hh:ocean_proximityINLAND        -3.24e-01   4.63e-02   -6.98
## bedrooms_per_hh:ocean_proximityNEAR BAY      -1.83e-01   8.55e-02   -2.14
## bedrooms_per_hh:ocean_proximityNEAR OCEAN    -1.23e-01   6.83e-02   -1.79
##                                              Pr(>|t|)    
## (Intercept)                                   7.7e-05 ***
## longitude                                     2.1e-08 ***
## housing_median_age                            1.3e-06 ***
## median_income                                 0.26631    
## pop_per_hh                                    0.23515    
## bedrooms_per_hh                               2.0e-06 ***
## ocean_proximityINLAND                         0.03882 *  
## ocean_proximityNEAR BAY                       < 2e-16 ***
## ocean_proximityNEAR OCEAN                     < 2e-16 ***
## longitude:housing_median_age                  2.7e-06 ***
## longitude:median_income                       0.69009    
## longitude:pop_per_hh                          0.03064 *  
## longitude:bedrooms_per_hh                     1.2e-05 ***
## longitude:ocean_proximityINLAND               0.59721    
## longitude:ocean_proximityNEAR BAY             < 2e-16 ***
## longitude:ocean_proximityNEAR OCEAN           < 2e-16 ***
## housing_median_age:median_income              9.3e-14 ***
## housing_median_age:pop_per_hh                 0.00022 ***
## housing_median_age:bedrooms_per_hh            0.81305    
## housing_median_age:ocean_proximityINLAND      8.1e-15 ***
## housing_median_age:ocean_proximityNEAR BAY    0.97381    
## housing_median_age:ocean_proximityNEAR OCEAN  0.46681    
## median_income:pop_per_hh                      < 2e-16 ***
## median_income:bedrooms_per_hh                 0.11337    
## median_income:ocean_proximityINLAND           < 2e-16 ***
## median_income:ocean_proximityNEAR BAY         1.7e-06 ***
## median_income:ocean_proximityNEAR OCEAN       1.3e-05 ***
## pop_per_hh:bedrooms_per_hh                    0.01115 *
## pop_per_hh:ocean_proximityINLAND              < 2e-16 ***
## pop_per_hh:ocean_proximityNEAR BAY            0.00666 **
## pop_per_hh:ocean_proximityNEAR OCEAN          0.01339 *
## bedrooms_per_hh:ocean_proximityINLAND         3.0e-12 ***
## bedrooms_per_hh:ocean_proximityNEAR BAY       0.03256 *
## bedrooms_per_hh:ocean_proximityNEAR OCEAN     0.07283 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.276 on 13610 degrees of freedom
## Multiple R-squared:  0.747,  Adjusted R-squared:  0.747
## F-statistic: 1.22e+03 on 33 and 13610 DF,  p-value: <2e-16
```

```r
get_RMSE = function(fitted_y, actual_y){
  n = length(fitted_y)                                      #length of data
  rmse = sqrt((sum(((actual_y - fitted_y) ^ 2)) / n))
}
```

```r
cah_tst_data$ocean_proximity = as.factor(cah_tst_data$ocean_proximity)
cah_tst_data$pop_per_hh = cah_tst_data$population / cah_tst_data$households
cah_tst_data$rooms_per_hh = cah_tst_data$total_rooms / cah_tst_data$households
cah_tst_data$bedrooms_per_hh = cah_tst_data$total_bedrooms / cah_tst_data$households

(get_RMSE(predict(mod_fix_int, cah_trn_data), cah_trn_data$median_house_value)) # 237303
```

```
## [1] 236367
```

```r
(get_RMSE(predict(mod_fix_int, cah_tst_data), cah_tst_data$median_house_value)) # 235732
```

```
## [1] 237914
```

Model Selection

Graphs and Tables

QA: How do we know what we did makes sense?.

Move to final report