Lab 10: Predicting house prices

MinJae Jo

2025-11-13

Lab report

```r
str(train)
```

```
## spc_tbl_ [16,512 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ longitude         : num [1:16512] -118 -121 -118 -122 -121 ...
##  $ latitude          : num [1:16512] 34.1 38 33.9 37.4 37.7 ...
##  $ housing_median_age: num [1:16512] 24 46 41 25 19 31 28 17 26 30 ...
##  $ total_rooms       : num [1:16512] 5745 2001 2048 1750 1690 ...
##  $ total_bedrooms    : num [1:16512] 735 428 439 341 327 264 574 261 568 164 ...
##  $ population        : num [1:16512] 2061 1384 1191 999 855 ...
##  $ households        : num [1:16512] 679 401 429 319 296 236 595 269 585 143 ...
##  $ median_income     : num [1:16512] 8.28 1.94 3.8 5.81 3.25 ...
##  $ median_house_value: num [1:16512] 451400 62200 222500 308700 176700 ...
##  $ ocean_proximity   : chr [1:16512] "INLAND" "INLAND" "<1H OCEAN" "NEAR BAY" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   longitude = col_double(),
##   ..   latitude = col_double(),
##   ..   housing_median_age = col_double(),
##   ..   total_rooms = col_double(),
##   ..   total_bedrooms = col_double(),
##   ..   population = col_double(),
##   ..   households = col_double(),
##   ..   median_income = col_double(),
##   ..   median_house_value = col_double(),
##   ..   ocean_proximity = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
str(test)
```

```
## spc_tbl_ [4,128 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ longitude         : num [1:4128] -122 -122 -122 -122 -122 ...
##  $ latitude          : num [1:4128] 37.9 37.8 37.9 37.8 37.8 ...
##  $ housing_median_age: num [1:4128] 52 50 42 50 52 49 48 49 52 52 ...
##  $ total_rooms       : num [1:4128] 1228 2239 1639 2082 729 ...
##  $ total_bedrooms    : num [1:4128] 293 455 367 492 160 447 409 366 390 419 ...
##  $ population        : num [1:4128] 648 990 929 1131 395 ...
##  $ households        : num [1:4128] 303 419 366 473 155 378 335 329 403 395 ...
##  $ median_income     : num [1:4128] 2.12 1.99 1.71 1.64 1.69 ...
##  $ median_house_value: num [1:4128] 155500 158700 159800 108900 132000 ...
##  $ ocean_proximity   : chr [1:4128] "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   longitude = col_double(),
##   ..   latitude = col_double(),
##   ..   housing_median_age = col_double(),
##   ..   total_rooms = col_double(),
##   ..   total_bedrooms = col_double(),
##   ..   population = col_double(),
##   ..   households = col_double(),
##   ..   median_income = col_double(),
##   ..   median_house_value = col_double(),
##   ..   ocean_proximity = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Exercises

Exercise 1 -The test loss was slightly higher than the training and validation losses, as the model had never encountered the test data before. The difference between the losses is small, so the model does not appear to be severely overfit.

Exercise 2 -Areas with high home prices are mostly concentrated in coastal areas, particularly Southern California and the Bay Area.

Exercise 3

i. Median_income showed the most pronounced relationship with house prices.

ii. House prices seem up to about half a million, but the value is cut, so the accuracy of linear models can decrease when predicting high values.

Exericse 4 -The category most concentrated in the lowest house price was INLAND.

Exericse 5 -The 5-fold validation RMSE of this model is about 84,000. In other words, when predicting a house price with just the median_income alone, there is an error of about $80,000 on average.

Exericse 6 -The cross-validation RMSE was much lower in the model created by adding all variables. Therefore, the prediction performance was clearly better than when using only media_income.

Exericse 7 -The test RMSE was slightly higher than the cross-validation RMSE when the model with the best performance was applied to the test data. It is natural for the error to be slightly larger because it is the first time predicting new data. Still, the difference does not widen significantly, so it can be seen that the model tends to predict stably and not overfitting.