Exercise 2

2023-02-04

#1) Saratoga house prices

Pricing Strategy

Main Focus: More preciously prediction for price

For the tax manager who want to know the precious prediction for price, we made more precious model from the data and suggested the points what elemets affect on how much price is.

Data

- lotSize
- Age
- landValue
- livingArea
- pctCollege
- bedrooms
- fireplaces
- bathrooms
- heating(electronic/)
- fuel
- sewer
- waterfront
- newConstruction
- centralAir

Model

We used the following steps to make the precious model.

- 1. Change qualitative variables into dummy variables in the data
- 2. Split data train/test dataset —- we repeated the followign procedures ten times and take an average of rmse
- 3. Linear regression with all variables(*)
-

4. Knn regression with all variables

- 5. Compared the average of rmse of Linear and Knn model to find better fit model
- 6. Summarized the better model and interpreted its meaning

(*) In the character of the linear model, when we used all variables to estimate, the rmse is lower than when we didn't use them.

Results

The liner model of RMSE is 0.3167590 and The Knn model of RMSE is 0.3235792.

Discussion: Comparison between Linear and LNN model

In this estimation, from the result that rmse of the linear model is smaller than that of knn model, the fitting of the linear model is better than that of the best linear model. We can think this reason is what the liner model that is set up close to the true model.

Conclusion for Tax authority

From the result of the estimation of the linear model(Appendix 1), we can say that elements that increases house prices are "waterfront", more "lotSize", more "landValue", more "livingArea", more "bathrooms." On the other hands, the more age house got or the house is a new constructed one, then the less price house get.

Appendix

1. Result of the model

```
Call:
lm(formula = log(price) ~ ., data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7686 -0.1521  0.0096  0.1700  1.3665

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.120e+01  1.366e-01  82.007  < 2e-16 ***
centralAir_Yes             3.774e-02  1.978e-02   1.908 0.056559 .
newConstruction_Yes       -1.415e-01  4.188e-02  -3.378 0.000751 ***
waterfront_Yes             4.612e-01  9.759e-02   4.726 2.53e-06 ***
sewer_septic               3.404e-02  1.016e-01   0.335 0.737567
sewer_public_commercial    3.612e-02  1.009e-01   0.358 0.720302
fuel_gas                   3.689e-02  2.940e-02   1.255 0.209850
fuel_electric             -6.285e-02  7.263e-02  -0.865 0.386975
`heating_hot air`         -2.197e-03  6.902e-02  -0.032 0.974616
`heating_hot water/steam` -2.589e-02  7.233e-02  -0.358 0.720413
lotSize                    5.007e-02  1.290e-02   3.880 0.000109 ***
age                       -1.402e-03  3.416e-04  -4.103 4.32e-05 ***
landValue                  3.316e-06  2.672e-07  12.413  < 2e-16 ***
livingArea                 2.904e-04  2.621e-05  11.084  < 2e-16 ***
pctCollege                 8.405e-04  8.855e-04   0.949 0.342710
bedrooms                   4.599e-03  1.472e-02   0.312 0.754758
fireplaces                -3.040e-04  1.718e-02  -0.018 0.985882
bathrooms                  9.622e-02  1.956e-02   4.919 9.77e-07 ***
rooms                      7.540e-03  5.498e-03   1.371 0.170447
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2981 on 1363 degrees of freedom
Multiple R-squared:  0.5807,    Adjusted R-squared:  0.5751
F-statistic: 104.9 on 18 and 1363 DF,  p-value: < 2.2e-16
```

2. Reference

Documentation of the Saratago House dataset https://r-data.pmagunia.com/dataset/r-dataset-package-mosaicdata-saratogahouses