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

The description of the dataset in the salatago house;

- price: price (1000s of US dollars)
- lotSize: size of lot (square feet)
- Age: age of house (years)
- landValue: value of land (1000s of US dollars)
- livingArea: living are (square feet)
- pctCollege: percent of neighborhood that graduated college
- bedrooms: number of bedrooms
- fireplaces: number of fireplaces
- bathrooms: number of bathrooms (half bathrooms have no shower or tub)
- rooms: number of rooms
- heating: type of heating system
- fuel: fuel used for heating
- sewer: type of sewer system
- waterfront: whether property includes waterfront
- newConstruction: whether the property is a new construction
- centralAir: whether the house has central air

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

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

The estimation of the model is

$$log(Price) = \beta_0 + \beta[All\ variables\ in\ the\ above]$$

* There are no interaction terms and squared terms in the model. * In the knn regression, all variables is scaled for the standardization * In the regression, sometimes the low ranked estimation happens because of the multicolinearlity.

- 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 0.2864057 and The Knn model of RMSE is 0.2907227. Please see the detail of the linear reagression in the appendix.

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 more "lotSize", more "landValue", more "livingArea", more "pctCollege", more "bathrooms", the existence of the central Air and the waterfront. 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
-2.50061 -0.16178  0.00813  0.16657  1.36032

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.116e+01  1.238e-01  90.084  < 2e-16 ***
centralAir_Yes            4.166e-02  1.905e-02   2.187 0.028946 *
newConstruction_Yes      -1.877e-01  4.016e-02  -4.675 3.23e-06 ***
waterfront_Yes            5.074e-01  8.503e-02   5.967 3.07e-09 ***
sewer_septic             -4.991e-03  9.075e-02  -0.055 0.956148
sewer_public_commercial   1.607e-02  9.047e-02   0.178 0.859057
fuel_gas                  3.147e-02  2.756e-02   1.142 0.253618
fuel_electric            -3.192e-02  6.863e-02  -0.465 0.641865
`heating_hot air`         2.526e-02  6.582e-02   0.384 0.701202
`heating_hot water/steam` -7.811e-03  6.871e-02  -0.114 0.909510
lotSize                   4.044e-02  1.169e-02   3.459 0.000558 ***
age                      -1.412e-03  3.120e-04  -4.526 6.52e-06 ***
landValue                 3.224e-06  2.529e-07  12.747  < 2e-16 ***
livingArea                2.739e-04  2.468e-05  11.098  < 2e-16 ***
```

```
pctCollege                1.647e-03  8.216e-04   2.005 0.045212 *
bedrooms                  1.259e-02  1.382e-02   0.911 0.362468
fireplaces                7.317e-03  1.613e-02   0.454 0.650135
bathrooms                 1.111e-01  1.816e-02   6.120 1.22e-09 ***
rooms                     5.441e-03  5.209e-03   1.044 0.296496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2815 on 1363 degrees of freedom
Multiple R-squared:  0.6109,    Adjusted R-squared:  0.6058
F-statistic: 118.9 on 18 and 1363 DF,  p-value: < 2.2e-16
```