Exercise 2

2023-02-12

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)

<Dependent variables(numerical)> - 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

<Dependent variables(non-numerical)> - 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 Split data train/test dataset
- 2 Create squared variables and interaction variables of the numerical data in the SaratogaHouses

we repeated the followign procedures ten times and take an average of rmse

The estimation of the model is

$$log(Price) = \beta_0 + \beta_{\not\models}[numerical\ variables]^2 + \beta_{\not\models}[interaction\ terms\ by\ each\ numerical\ variables]$$
$$+ \beta_{\not\models}[non - numerical\ variables(dummytemrs)]$$

- 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

Results

The liner model of RMSE is 0.2822 and The Knn model of RMSE is 0.3061. 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 "fireplaces", more "newConstructionNo" at the statistically siginificance. However, more "age", "heatinghot water/steam", "waterfrontNo" make its price decrease at the statistically siginificance.

Appendix

1. Result of the model

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.7497 -0.1405  0.0100  0.1576  1.1371

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.149e+01  2.594e-01  44.320  < 2e-16 ***
lotSize                  1.400e-01  9.121e-02   1.535 0.125100
age                     -6.528e-03  1.985e-03  -3.288 0.001037 **
landValue                2.991e-06  2.378e-06   1.258 0.208710
livingArea               1.373e-04  1.658e-04   0.828 0.407822
pctCollege              -2.385e-03  6.711e-03  -0.355 0.722334
bedrooms                 1.206e-01  9.930e-02   1.215 0.224662
fireplaces               4.066e-01  1.081e-01   3.761 0.000176 ***
bathrooms                2.879e-01  1.326e-01   2.171 0.030119 *
rooms                    1.462e-02  3.464e-02   0.422 0.673072
lotSize.sq               1.467e-03  4.970e-03   0.295 0.767818
lotSize._.age           -1.063e-03  4.809e-04  -2.211 0.027197 *
lotSize._.landValue     -7.900e-07  4.299e-07  -1.838 0.066306 .
lotSize._.livingArea    -2.544e-05  4.040e-05  -0.630 0.528950
lotSize._.pctCollege     8.655e-04  1.488e-03   0.582 0.560879
lotSize._.bedrooms       1.467e-02  2.213e-02   0.663 0.507411
lotSize._.fireplaces    -8.360e-03  3.149e-02  -0.265 0.790681
lotSize._.bathrooms     -5.440e-02  3.014e-02  -1.805 0.071331 .
lotSize._.rooms          5.359e-03  9.768e-03   0.549 0.583339
age.sq                   1.744e-05  5.910e-06   2.951 0.003222 **
age._.landValue          1.959e-08  7.786e-09   2.516 0.011983 *
```

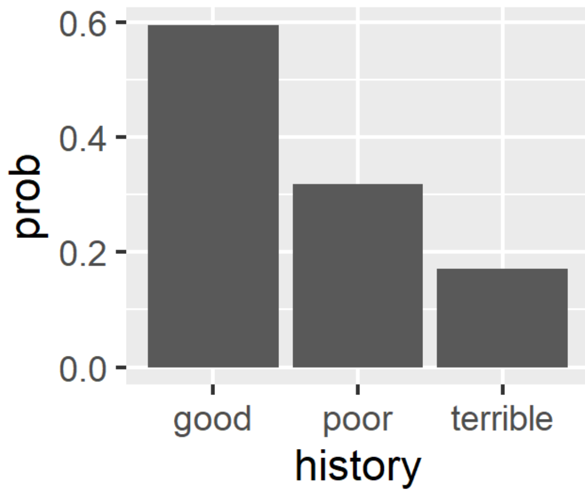
```
age._.livingArea        -4.245e-07  8.106e-07  -0.524 0.600590    
age._.pctCollege         7.604e-05  2.970e-05   2.560 0.010585 *  
age._.bedrooms          -1.111e-04  4.943e-04  -0.225 0.822186    
age._.fireplaces         3.312e-04  6.099e-04   0.543 0.587238    
age._.bathrooms          7.385e-04  6.041e-04   1.222 0.221757    
age._.rooms             -2.342e-04  1.869e-04  -1.253 0.210419    
landValue.sq            -9.147e-12  2.599e-12  -3.520 0.000447 ***
landValue._.livingArea  -9.431e-10  7.145e-10  -1.320 0.187081    
landValue._.pctCollege   9.317e-08  3.619e-08   2.575 0.010144 *  
landValue._.bedrooms    -7.029e-07  4.219e-07  -1.666 0.095931 .  
landValue._.fireplaces  -1.186e-06  5.235e-07  -2.267 0.023574 *  
landValue._.bathrooms    7.389e-07  5.630e-07   1.312 0.189637    
landValue._.rooms        2.372e-08  1.764e-07   0.134 0.893069    
livingArea.sq           -3.616e-08  5.054e-08  -0.715 0.474500    
livingArea._.pctCollege  3.326e-06  2.451e-06   1.357 0.174924    
livingArea._.bedrooms    3.422e-05  4.263e-05   0.803 0.422323    
livingArea._.fireplaces -1.989e-05  5.200e-05  -0.382 0.702190    
livingArea._.bathrooms   9.046e-05  5.673e-05   1.595 0.111041    
livingArea._.rooms      -1.321e-05  2.035e-05  -0.649 0.516440    
pctCollege.sq           -4.168e-05  6.088e-05  -0.685 0.493686    
pctCollege._.bedrooms    3.026e-04  1.391e-03   0.218 0.827838    
pctCollege._.fireplaces -4.274e-03  1.552e-03  -2.754 0.005966 ** 
pctCollege._.bathrooms  -1.383e-03  1.852e-03  -0.747 0.455277    
pctCollege._.rooms      -2.517e-05  5.227e-04  -0.048 0.961603    
bedrooms.sq             -3.858e-03  1.592e-02  -0.242 0.808578    
bedrooms._.fireplaces   -5.325e-02  2.805e-02  -1.899 0.057841 .  
bedrooms._.bathrooms    -6.534e-02  3.109e-02  -2.102 0.035765 *  
bedrooms._.rooms        -1.430e-04  1.105e-02  -0.013 0.989677    
fireplaces.sq            2.558e-02  2.486e-02   1.029 0.303794    
fireplaces._.bathrooms   7.453e-03  3.599e-02   0.207 0.835996    
fireplaces._.rooms       6.778e-03  1.080e-02   0.628 0.530341    
bathrooms.sq            -3.959e-02  2.735e-02  -1.448 0.147982    
bathrooms._.rooms        9.407e-03  1.204e-02   0.781 0.434895    
rooms.sq                 2.086e-04  3.692e-03   0.057 0.954952    
heatinghot water/steam  -4.657e-02  2.266e-02  -2.055 0.040033 *  
heatingelectric          3.426e-02  6.399e-02   0.535 0.592505    
fuelelectric            -5.308e-02  6.323e-02  -0.839 0.401376    
fueloil                 -7.454e-03  2.733e-02  -0.273 0.785087    
sewerpublic/commercial   1.201e-02  2.076e-02   0.579 0.562873    
sewernone               -1.179e-01  8.472e-02  -1.392 0.164108    
waterfrontNo            -5.851e-01  8.477e-02  -6.903  7.9e-12 ***
newConstructionNo        1.391e-01  4.102e-02   3.390 0.000719 ***
centralAirNo            -1.832e-02  1.831e-02  -1.000 0.317270    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2678 on 1318 degrees of freedom
Multiple R-squared:  0.645, Adjusted R-squared:  0.6281 
F-statistic: 38.02 on 63 and 1318 DF,  p-value: < 2.2e-16
```

2) Classification and retrospective sampling

Results

The graph of the default probability by credit history is



The result of the logit model that we built is

```
The coefficients of the logit model
        (Intercept)            duration              amount         installment                 age
              -0.66                0.02                0.00                0.19               -0.02
    historyterrible          purposeedu purposegoods/repair       purposenewcar      purposeusedcar
              -2.02                0.91                0.13                0.94               -0.67

confusion matrix
   yhat
y     0   1
  0 131  11
  1  44  14

out-of-sample accuracy rate
0.725

the result of the null model
  0   1
142  58

the null model accuracy rate
0.71
```

Disucussion

What do you notice about the history variable vis-a-vis predicting defaults?

From the coefficient of the logit model, the poor and terrible of the history made the probability of default decrease.

What do you think is going on here?

Intuitively, the poor and terrible of the history made the probability of default increase. So there is something with the bad estimation. We can think this reason is caused by what the default is rare, and so we cannot collect data randomly(the data is not collected through random sampling) that is biased.

As this evidence, the bar graph has shows that people of the good credit history has the higher default probability. However, this is different from the intuitive result and is not reality.

Do you think this data set is appropriate for building a predictive model of defaults

We don't think so. Because the out-of-sample accuracy rate is 0.725 while the null model accuracy rate is 0.71. Therefore, the improvement of the estimation is so low.

Would you recommend any changes to the bank's sampling scheme?

As we said above, the data should be collected randomly that will make biased decrease.

3) Children and hotel reservations

Model Building

Models

We shows the models that we used in this problems. First, the baseline 1 is

$$children = \beta_0 + \boldsymbol{\beta}\mathbf{X}_{market\ segment,\ adults,\ customer_type,\ is\ repeated\ guest}$$

The baseline 2 is

$$children = \beta_0 + \boldsymbol{\beta}\mathbf{X}_{all\ variables\ excpet\ arriving\ date}$$

The our model is

$$\begin{aligned} children = \beta_0 + \boldsymbol{\beta}\mathbf{X}_{all\ variables\ excpet\ arriving\ date} &+ arriving\ year + arriving\ month \\ &+ average\ daily\ rate \times adults \\ &+ \ days\ in\ waiting_list \times adults \\ &+ stays\ in\ weekend_nights \times adults \\ &+ total\ of\ special\ requests \times adults \\ &+ booking\ changes \times average\ daily\ rate \\ &+ booking\ changes \times days\ in\ waiting_list \\ &+ lead\ time \times booking\ changes \\ &+ (lead\ time)^2 \end{aligned}$$

Check

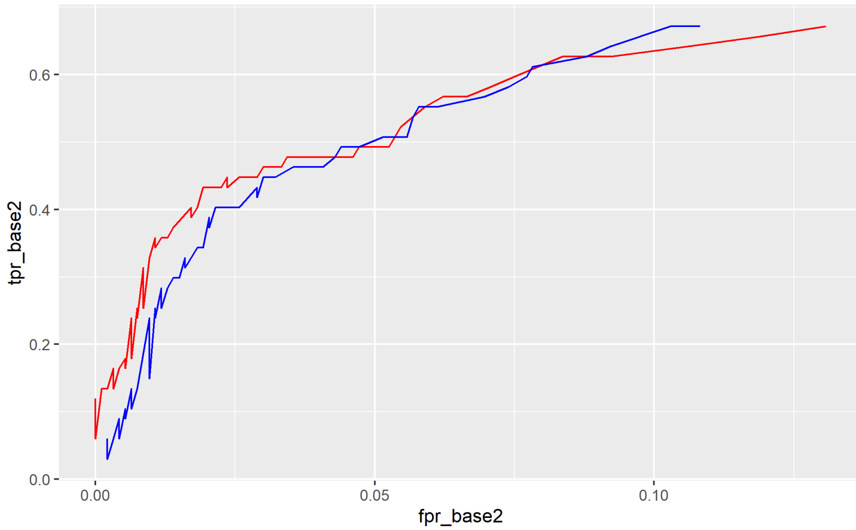
Out-of-sample accuracy rate by each model is

baseline1	baseline2	mymodel
0.9202222	0.9368889	0.9375556

Therefore, the model accuracy of my model is higher than the baseline2 by 0.1% and thn the baseline 1 by 1.7%.

Model validation: step 1

The ROC curve of baseline 2 and my model is



red line: baseline 2, blue line: my model

From the graph, if FPR=0.05 my model has higher tpr than baseline 2, and so my model is better than baseline 2 in this case.

However, in the low FPR, the TPR of baseline 2 is higher than that of my model, and so my model is worse than baseline 2. Also, in the high FPR, the TPR of baseline 2 is lower than that of my model, and so my model is better than baseline 2.

Model validation: step 2

In this case, we assumed a threshold is 50%, and our results is in the following.

predict_base2	predict_model	actual
8	7	14
6	11	21
11	11	14
10	10	19
10	9	19
15	17	21
12	10	26
8	7	26
5	6	19
12	12	24
8	8	18
8	8	17
7	8	17
11	11	17
11	13	24
13	15	21
12	11	20
7	11	12
17	17	25
16	19	28

sum_base2	sum_predict	sum_actual
207	221	402

From the result, the predicting the total number of bookings with children by baseline 2 is 207, that by my model is 221, and that by actual data is 402. The accurancy of the prediction of the our model is around 50%, which is so lower than we expected. However, our model's accurancy of the prediction is higher than the baseline 2's one.