ECO395M_Exercise2

Youngseok Yim (EID: yy9739)

2023-02-22

1. Saratoga house prices

Mean RMSE for medium model

```
## [1] 66469.99
```

Mean RMSE for main linear model

```
## [1] 60363.98
```

Mean RMSE for k nearest neighbors with various k values

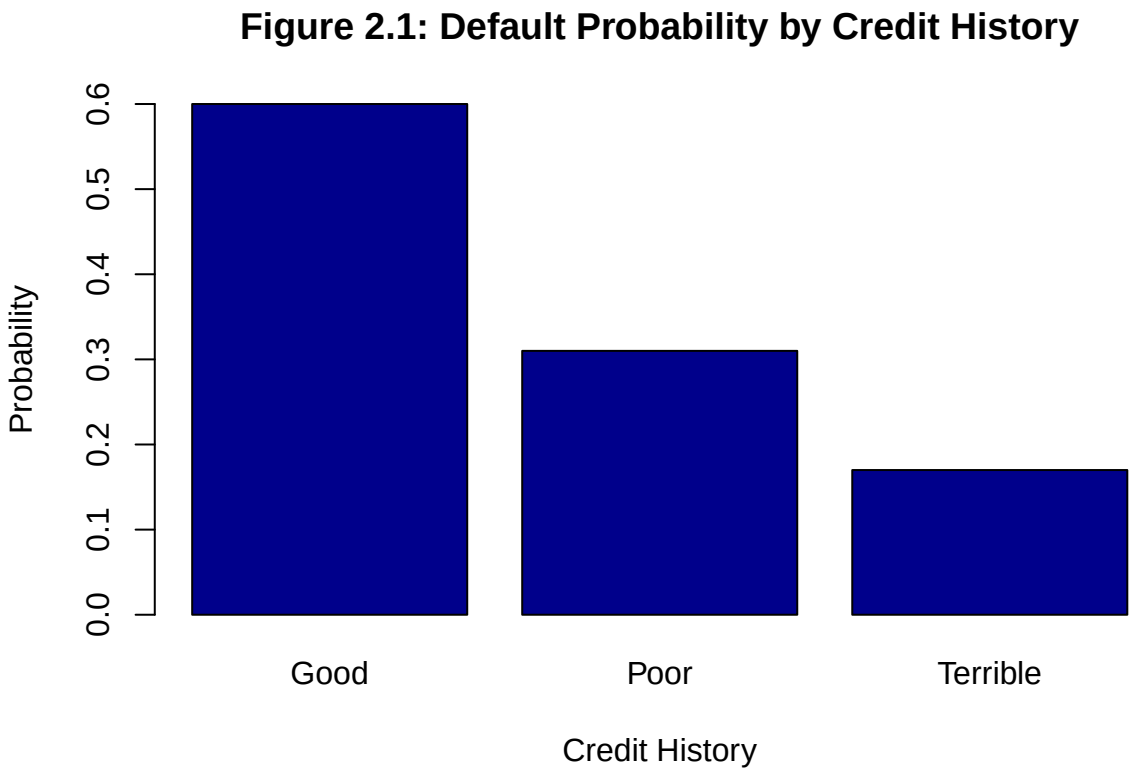
```
##              k     RMSE
## result.1     2 68424.98
## result.2     5 62548.11
## result.3    10 62261.95
## result.4    20 62531.16
## result.5    50 63663.94
## result.6    75 64506.43
## result.7   100 65377.54
## result.8   200 67987.53
## result.9   300 70486.54
## result.10  400 72264.01
```

I created linear and k-nearest neighbors regression models to predict house prices, using selected features that influence the price. For the linear model, I considered lot size, age, land value, living area, number of bedrooms, number of bathrooms, number of fireplaces, number of rooms, type of heating system, type of fuel used, availability of central air, interaction between age and central air, interaction between lot size and land value, interaction between number of bedrooms and rooms, and interaction between number fireplaces and type of heating system for predicting house price.

In the k-nearest neighbors model, I used lot size, age, land value, living area, number of bedrooms, fireplaces, number of bathrooms, number of rooms, type of heating system, type of fuel, and availability of central air as feature variables. Since the k- nearest neighbors model is adaptable to find interactions and nonlinearities, I omitted interaction between the feature variables from the model.

After evaluating both models using out-of-sample RMSE, I found that the k-nearest neighbors model with k=20 had a lower average RMSE and performed better.

2. Classification and retrospective sampling




```
##         (Intercept)            duration              amount         installment
##               -0.71                0.03                0.00                0.22
##                 age         historypoor     historyterrible          purposeedu
##               -0.02               -1.11               -1.88                0.72
## purposegoods/repair       purposenewcar      purposeusedcar       foreigngerman
##                0.10                0.85               -0.80               -1.26
```

Figure 2.1 displays a bar plot of default probabilities based on credit histories (good, bad, and terrible). Surprisingly, the probability of default is higher for individuals with good credit history (0.6) compared to those with bad (0.3) and terrible (0.17) credit history. The logit model coefficients also support this finding, with poor history having a coefficient of -1.11 and terrible history having a coefficient of -1.88, indicating a decrease in the odds of default as credit history worsens.

However, the data set used in the study may not be suitable for predicting default probabilities. The bank selected defaulted loans for inclusion, resulting in an oversampling of defaults compared to a random sample of loans in the bank's portfolio. Therefore, if the purpose is to classify prospective borrowers into high or low default risk, the bank should use a random sampling method to avoid oversampling of defaults.

3. Children and hotel reservations

Baseline 1: Mean RMSE

```
## [1] 0.2682445
```

Baseline 2: Mean RMSE

```
## [1] 0.2332847
```

Main linear model: Mean RMSE

```
## [1] 0.2314623
```

Figure 3.1: ROC curve

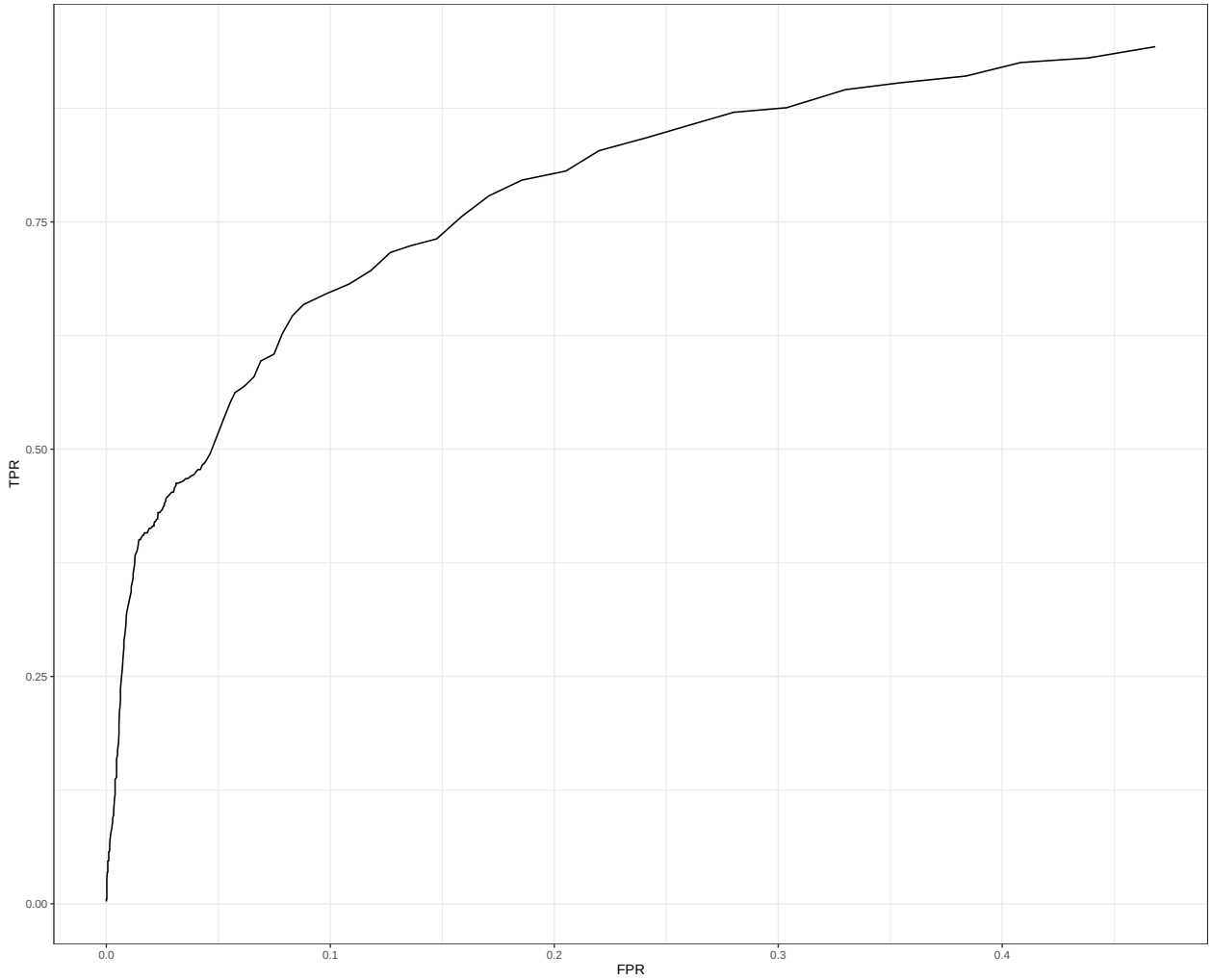


Table 1: Actual v Expected number of bookings with childen

Fold_id	Actual	Expected
1	19	19.41525
2	11	16.10791
3	14	18.95748
4	20	21.30603
5	21	23.52836
6	27	22.76993
7	26	21.07458
8	21	21.83524
9	28	24.47166
10	23	22.75855
11	17	20.62713
12	19	21.54587
13	19	20.18545
14	23	20.23307
15	14	21.17384
16	23	20.74142
17	22	21.71953

Fold_id	Actual	Expected
18	12	20.23556
19	21	22.25859
20	22	23.76595

Table 1 reveals that, among 20 folds each with 250 observations, the model has overpredicted the number of bookings with children in 50% of the folds, while it has underpredicted in 40%. In 10% of the folds, the model has accurately predicted the number of bookings with children. The greatest discrepancy between actual and predicted bookings is 8. Despite the fluctuations, the model's overall performance in predicting the number of bookings with children is considered satisfactory.