Spring Break Assignment

Cadee Pinkerton, James Owens, Molly Wu

4/4/2022

Problem 1: In the Fullerton Housing data set, let PRICE be the response and consider BEDS, BATHS, SQUARE_FEET, and YEAR_BUILT to be your predictors. First, fit a multiple linear regression model to this data will all four predictors in the model.

```
a=seq(1,195,1)
b=sample(a,175,replace = F)


F.train=Fullerton[b,]
F.test=Fullerton[-b,]


fullerton.price=lm(PRICE~YEAR_BUILT+SQUARE_FEET+BATHS+BEDS,data=F.train)
fit.price=predict(fullerton.price,newdata=F.test)
SSE.1=sum((fit.price-PRICE[-b])^2)
SSE.1
```

```
## [1] 161216811941
```

```
log(SSE.1)
```

```
## [1] 25.80602
```

Problem 2: Consider the four possible models that have only one predictor. Using the 5-fold cross validation technique, compare RMSE of the four models. Which one is a better model?

```
Housing=data.frame(FullertonHousing)


ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~BEDS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   1 predictor
##
```

```
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 156, 156, 157, 155, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   202982.6  0.5217897  151775.2
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~BATHS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 157, 157, 155, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   215390.7  0.4875721  167835.8
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~SQUARE_FEET,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 156, 156, 156, 156, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   133279.7  0.8040331  101547.8
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~YEAR_BUILT,
```

```
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 156, 158, 155, 156
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   291294.3  0.02048746  228738.5
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Out of the four models, the model using square feet as the predictor is the better model.

Problem 3: Consider all six models that have only two predictors. For example lm(PRICE~BEDS+BATHS). Using the 5-fold cross validation technique, compare the RMSE of the six models. Which one is a better model?

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~BEDS+BATHS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 156, 155, 157, 157
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   195232.2  0.5866509  146843
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~BATHS+SQUARE_FEET,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
```

```
##   2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 157, 155, 156, 157, 155
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   131633.1  0.8010372  99692.95
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~SQUARE_FEET+YEAR_BUILT,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 157, 155, 155, 158
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   127512.8  0.8127179  95198.7
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~YEAR_BUILT+BEDS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 157, 156, 156, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   204856.8  0.5300517  153732.6
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~YEAR_BUILT+BATHS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 158, 156, 156, 155
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   212246.5  0.4616824  161443.2
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~SQUARE_FEET+BEDS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 156, 156, 155, 157, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   135198.9  0.8133362  100633.6
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Out of the six models, the model using square feet and the year the house was built as the predictors is the better model.

Problem 4: Consider all four possible models that have three predictors. For example lm(PRICE~BEDS+BATHS+ SQUARE_FEET). Using the 5-fold cross validation technique compare the RMSE of the four models. Which one is a better model?

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~BEDS+BATHS+YEAR_BUILT,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 156, 158, 155, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   196481.5  0.5980961  148301
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~BEDS+BATHS+SQUARE_FEET,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 155, 155, 157, 157, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   129969.8  0.8004078  96453.73
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~BATHS+SQUARE_FEET+YEAR_BUILT,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 156, 155, 156, 156, 157
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   130204.3  0.8197282  96089.4
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~SQUARE_FEET+YEAR_BUILT+BEDS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 157, 155, 155, 157, 156
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   127687.7  0.8208338  93252.23
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Out of the four models, the model using square feet, the number of bedrooms, and the number of bathrooms as the predictors is the better model.

Problem 5: Consider the only model that has four predictors. Using the 5-fold cross validation technique, calculate the RMSE of that model.

```
ctrl <- trainControl(method = "cv", number = 5)
model <- train(PRICE~SQUARE_FEET+YEAR_BUILT+BEDS+BATHS,
               data = Housing, method = "lm", trControl = ctrl)
print(model)
```

```
## Linear Regression
##
## 195 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 157, 157, 156, 155, 155
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   126680.2  0.8170926  93913.24
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Problem 6: Between all the models you fit to this data in parts 2-5 which one has the lowest RMSE or the best goodness of fit? Is that surprising?

Comparing the RMSE of all the models from problems 2-5, the model with the best goodness of fit is the model with the 3 predictors square feet, the number of bedrooms, and the number of bathrooms. This is surprising because in the past we saw that the better model was usually the one with the largest number of predictors. This assignment showed us that is not always true because our model with four predictors had a larger amount of error.