Case 4

Group 8: Caitlyn Blair, Shruti Hardasani, Rainna Sena
5 November 2024

Boston Housing Data

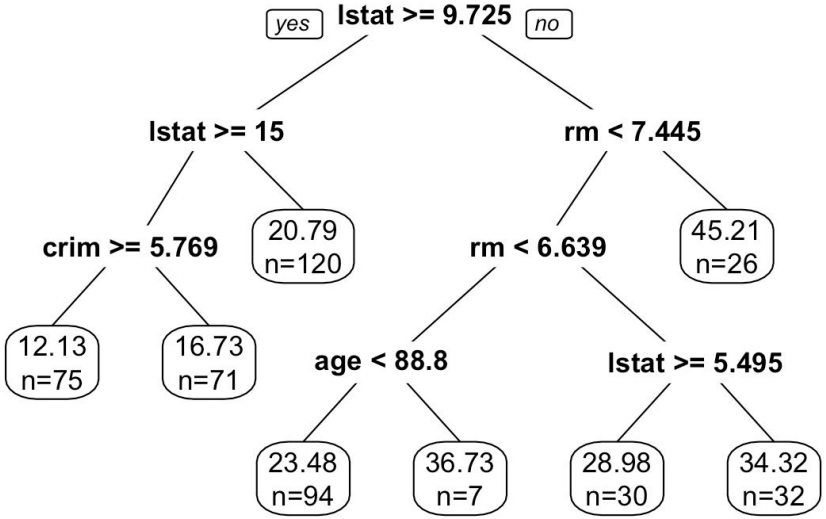
```
n= 455

node), split, n, deviance, yval
      * denotes terminal node

 1) root 455 38564.5500 22.41582
   2) lstat>=9.725 266  6486.3510 17.26429
     4) lstat>=15 146  2606.5860 14.36644
       8) crim>=5.76921 75  1040.8830 12.13200 *
       9) crim< 5.76921 71   795.6992 16.72676 *
     5) lstat< 15 120  1162.0480 20.79000 *
   3) lstat< 9.725 189 15083.8200 29.66614
     6) rm< 7.445 163  6836.9830 27.18712
      12) rm< 6.6385 101  3004.9980 24.39604
        24) age< 88.8 94   885.3631 23.47766 *
        25) age>=88.8 7   975.7143 36.72857 *
      13) rm>=6.6385 62  1763.4590 31.73387
        26) lstat>=5.495 30   488.6680 28.98000 *
        27) lstat< 5.495 32   833.9822 34.31562 *
     7) rm>=7.445 26   965.0785 45.20769 *
```

- Using Boston Housing data, we randomly sampled 90% of the data as the training set and the remaining 10% as the test set
- This shows the outcome of running a regression tree on the data
- Hard to interpret

Regression Tree Model



In Sample MSE

```
> MSE.tree
[1] 15.70865
```

Out of Sample MSE

```
> MSPE.tree
[1] 15.92386
```

The regression tree visualization is much easier to interpret and given the values, we can have a predicted value for any given input.

Linear Regression Model

```
Coefficients:
(Intercept)        crim          zn        chas         nox          rm
  39.559371   -0.114504    0.044809    3.132874  -19.657220    3.524630
        dis         rad         tax     ptratio       black       lstat
  -1.539422    0.287534   -0.009829   -0.963041    0.008518   -0.539330
```

In Sample MSE

```
> mean((pi - boston_train$medv)^2)
[1] 22.64818
```

Out of Sample MSE

```
> mean((pi2 - boston_test$medv)^2)
[1] 15.91855
```

Comparison of CART to Linear Regression Model

CART	Linear Regression
In-Sample MSE: 15.71	In-Sample MSE: 22.65
Out-of-Sample MSE: 15.92	Out-of-Sample MSE: 15.92

While the out-of-sample MSE values were almost identical before rounding, we can conclude that the CART model is a better fit than the linear regression model because the MSE values are overall smaller.

German Data set

This is our classification tree. We did a random sample of 80% for our training set and used the other 20% as the test set.

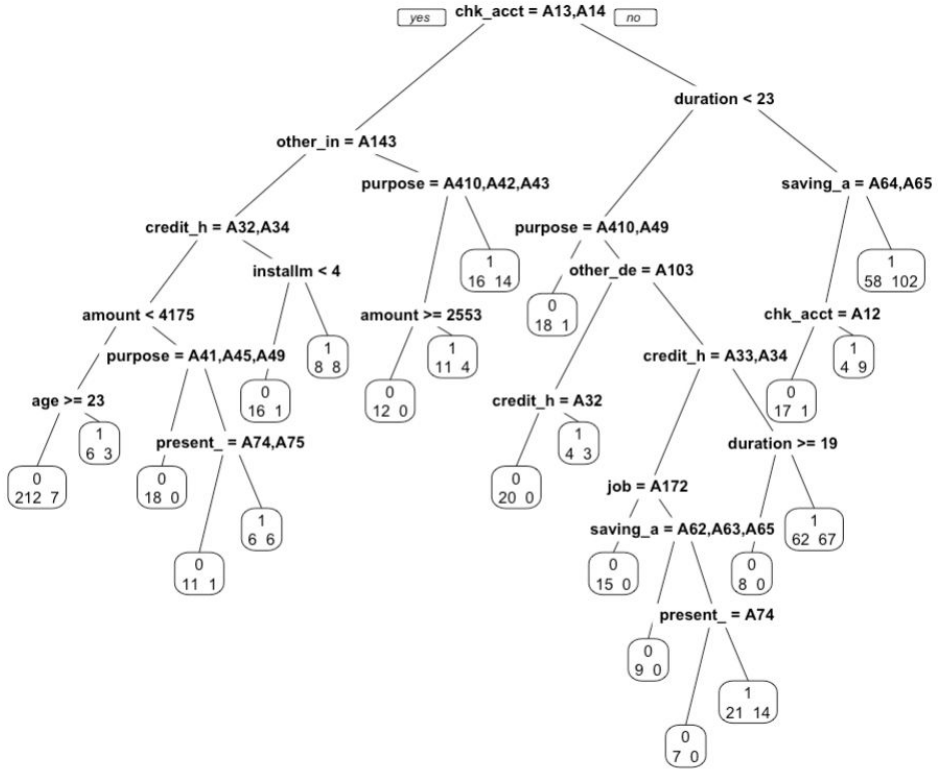
```
n= 800

node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 800 241 0 (0.69875000 0.30125000)
     2) chk_acct=A13,A14 360  44 0 (0.87777778 0.12222222) *
     3) chk_acct=A11,A12 440 197 0 (0.55227273 0.44772727)
       6) duration< 22.5 249  85 0 (0.65863454 0.34136546)
        12) credit_his=A32,A33,A34 228  69 0 (0.69736842 0.30263158)
          24) purpose=A41,A410,A43,A45,A49 96  19 0 (0.80208333 0.19791667) *
          25) purpose=A40,A42,A44,A46,A48 132  50 0 (0.62121212 0.37878788)
            50) other_install=A142,A143 118  40 0 (0.66101695 0.33898305)
             100) credit_his=A34 40   7 0 (0.82500000 0.17500000) *
             101) credit_his=A32,A33 78  33 0 (0.57692308 0.42307692)
               202) amount>=1485.5 40  10 0 (0.75000000 0.25000000) *
               203) amount< 1485.5 38  15 1 (0.39473684 0.60526316)
                 406) sex=A93,A94 17   6 0 (0.64705882 0.35294118) *
                 407) sex=A91,A92 21   4 1 (0.19047619 0.80952381) *
            51) other_install=A141 14   4 1 (0.28571429 0.71428571) *
        13) credit_his=A30,A31 21   5 1 (0.23809524 0.76190476) *
       7) duration>=22.5 191  79 1 (0.41361257 0.58638743)
        14) saving_acct=A64,A65 31  10 0 (0.67741935 0.32258065)
          28) chk_acct=A12 18   1 0 (0.94444444 0.05555556) *
          29) chk_acct=A11 13   4 1 (0.30769231 0.69230769) *
        15) saving_acct=A61,A62,A63 160  58 1 (0.36250000 0.63750000)
          30) duration< 47.5 131  55 1 (0.41984733 0.58015267)
            60) amount>=1549.5 120  55 1 (0.45833333 0.54166667)
             120) purpose=A41 18   5 0 (0.72222222 0.27777778) *
             121) purpose=A40,A410,A42,A43,A45,A46,A49 102  42 1 (0.41176471 0.58823529)
               242) other_debtor=A103 8   2 0 (0.75000000 0.25000000) *
               243) other_debtor=A101,A102 94  36 1 (0.38297872 0.61702128)
                 486) amount< 4231 64  29 1 (0.45312500 0.54687500)
                   972) amount>=2313 43  20 0 (0.53488372 0.46511628)
                    1944) sex=A93,A94 28  10 0 (0.64285714 0.35714286) *
                    1945) sex=A91,A92 15   5 1 (0.33333333 0.66666667) *
                   973) amount< 2313 21   6 1 (0.28571429 0.71428571) *
                 487) amount>=4231 30   7 1 (0.23333333 0.76666667) *
            61) amount< 1549.5 11   0 1 (0.00000000 1.00000000) *
          31) duration>=47.5 29   3 1 (0.10344828 0.89655172) *
```

Classification Tree

This image is easier to interpret for our classification tree. To the left shows the criteria is true while going to the right shows the criteria is false.



In Sample Misclassification Table

Our misclassification rate is computed by
(196+11) / (363+196+11+230) =0.259
25.9% of our predictions are incorrect

Truth	Predicted 0	1
0	363	196
1	11	230

Out of Sample Misclassification Table

Our misclassification rate is computed by
(68+14) / (73+68+14+45)=0.41
41% of our predictions are incorrect

Truth	Predicted 0	1
0	73	68
1	14	45

In Sample Misclassification Cost (Asymmetric)

This model incurs a cost of 0.506 per instance, which is still a high number.

```
[1] 0.50625
```

Out of Sample Misclassification Cost (Asymmetric)

There is a higher cost on the testing data set, indicating that there are more errors on this model.

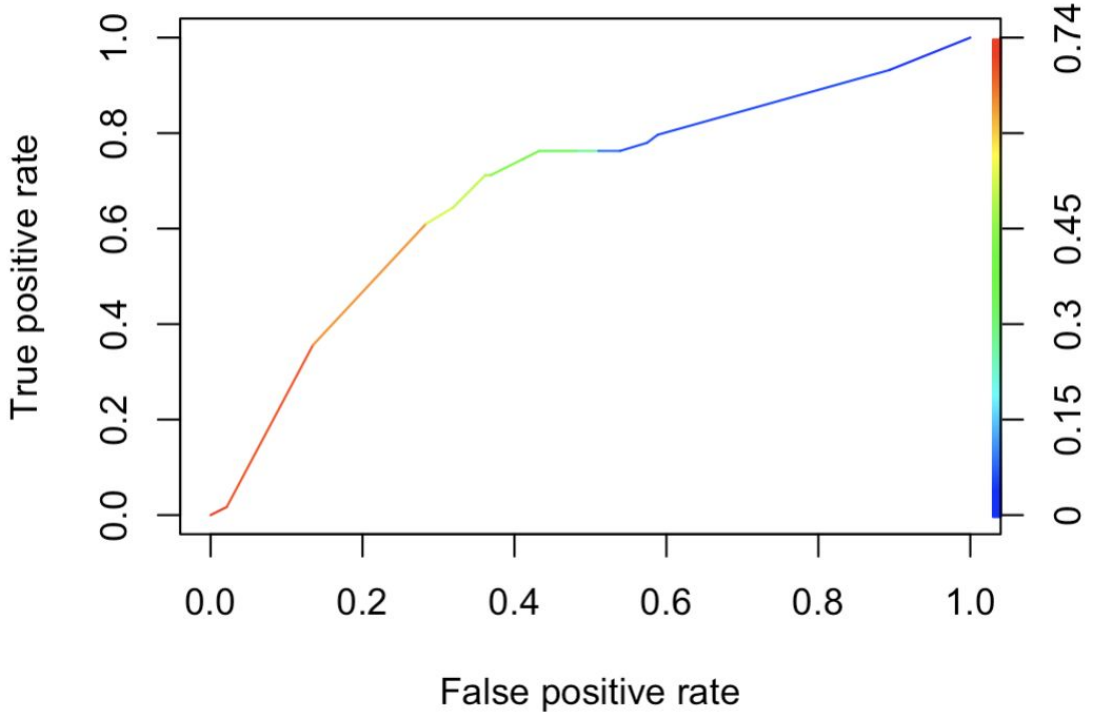
```
[1] 0.6975
```

ROC Curve

Out of Sample AUC

```
[1] 0.682534
```

We can evaluate the model's performance in context of ROC curve. It is an adequate model but can be better.



Thank you