Predictive Modeling in Higher Education

Haziel Garcia Sanchez

Vix Talbot

Intro to Statistical Learning

Data Description

The college data set:

- 18 variables
- 777 different universities and colleges in the US.
- Data were collected in 1995
- Predicting the number of applications received

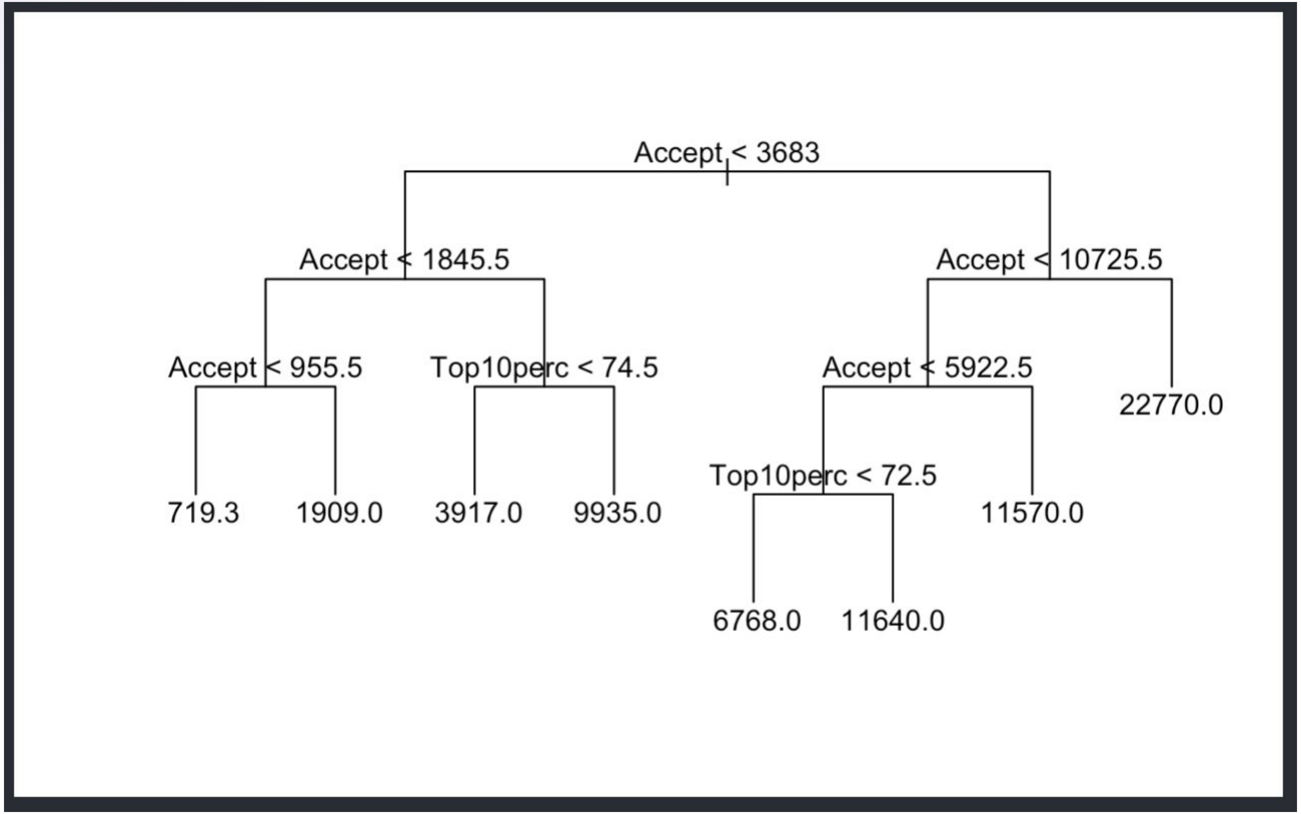
Linear Regression Analysis

```
Call:
lm(formula = Apps ~ ., data = College)

Residuals:
    Min      1Q  Median      3Q     Max
-4908.8  -430.2   -29.5   322.3  7852.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -445.08413  408.32855  -1.090 0.276053
PrivateYes  -494.14897  137.81191  -3.586 0.000358 ***
Accept         1.58581    0.04074  38.924  < 2e-16 ***
Enroll        -0.88069    0.18596  -4.736 2.60e-06 ***
Top10perc     49.92628    5.57824   8.950  < 2e-16 ***
Top25perc    -14.23448    4.47914  -3.178 0.001543 **
F.Undergrad    0.05739    0.03271   1.754 0.079785 .
P.Undergrad    0.04445    0.03214   1.383 0.167114
Outstate      -0.08587    0.01906  -4.506 7.64e-06 ***
Room.Board     0.15103    0.04829   3.127 0.001832 **
Books          0.02090    0.23841   0.088 0.930175
Personal       0.03110    0.06308   0.493 0.622060
PhD           -8.67850    4.63814  -1.871 0.061714 .
Terminal      -3.33066    5.09494  -0.654 0.513492
S.F.Ratio     15.38961   13.00622   1.183 0.237081
perc.alumni    0.17867    4.10230   0.044 0.965273
Expend         0.07790    0.01235   6.308 4.79e-10 ***
Grad.Rate      8.66763    2.94893   2.939 0.003390 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Significant Predictors

- **Accept**
- **Top 10%**
- Private
- Enroll
- Top 25%
- F. Undergrad
- Outstate
- Room.Board
- PhD
- Expend
- Grad. Rate
-

Decision Tree

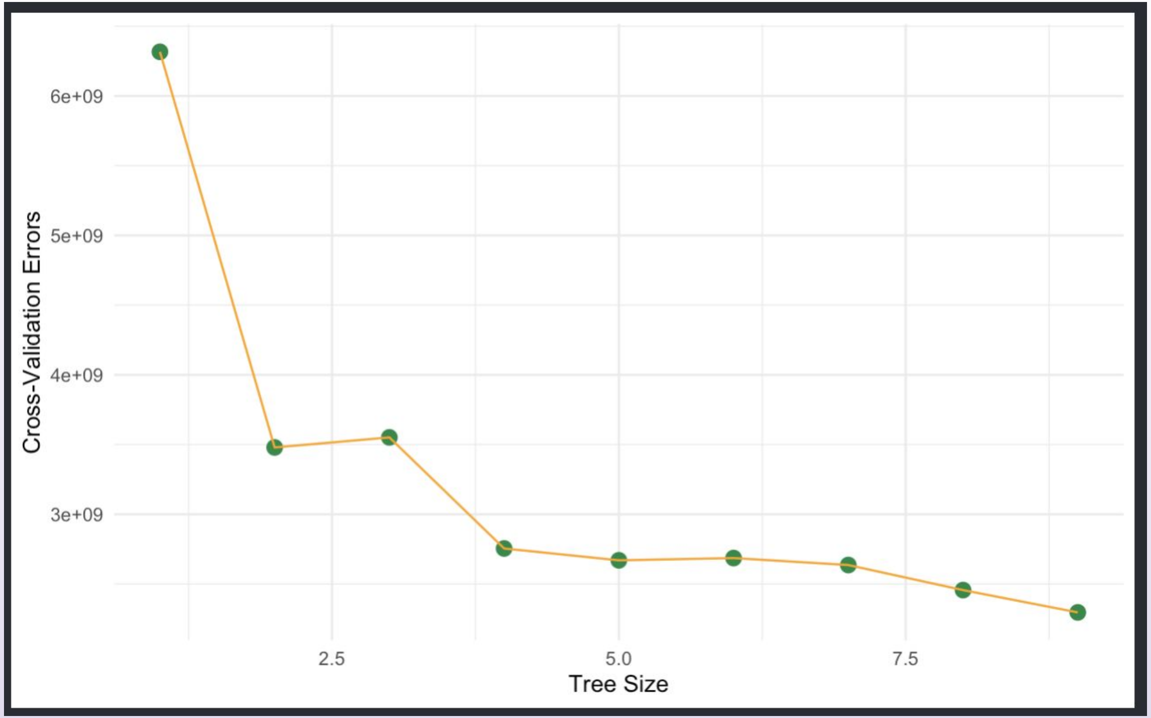


Partitioned Tree

8 Terminal Nodes:

*R*1={*yi* | Accept (*yi*) < 955.5}
*R*2={*yi* | Accept (*yi*) > 955.5 and Accept (*yi*) < 1845.5}
*R*3={*yi* | Accept (*yi*) > 1845.5 and Accept (*yi*) < 3683 and Top10 < 74.5}
*R*4={*yi* | Accept (*yi*) > 1845.5 and Accept (*yi*) < 3683 and Top10 > 74.5}
*R*5={*yi* | Accept (*yi*) > 3683 and Accept (*yi*) < 5922.5 and Top10 < 72.5}
*R*6={*yi* | Accept (*yi*) > 3683 and Accept (*yi*) < 5922.5 and Top10 > 72.5}
*R*7={*yi* | Accept (*yi*) > 5922.5 and Accept (*yi*) < 107525.5}
*R*8={*yi* | Accept (*yi*) > 107525.5}

Is pruning the decision tree helpful?



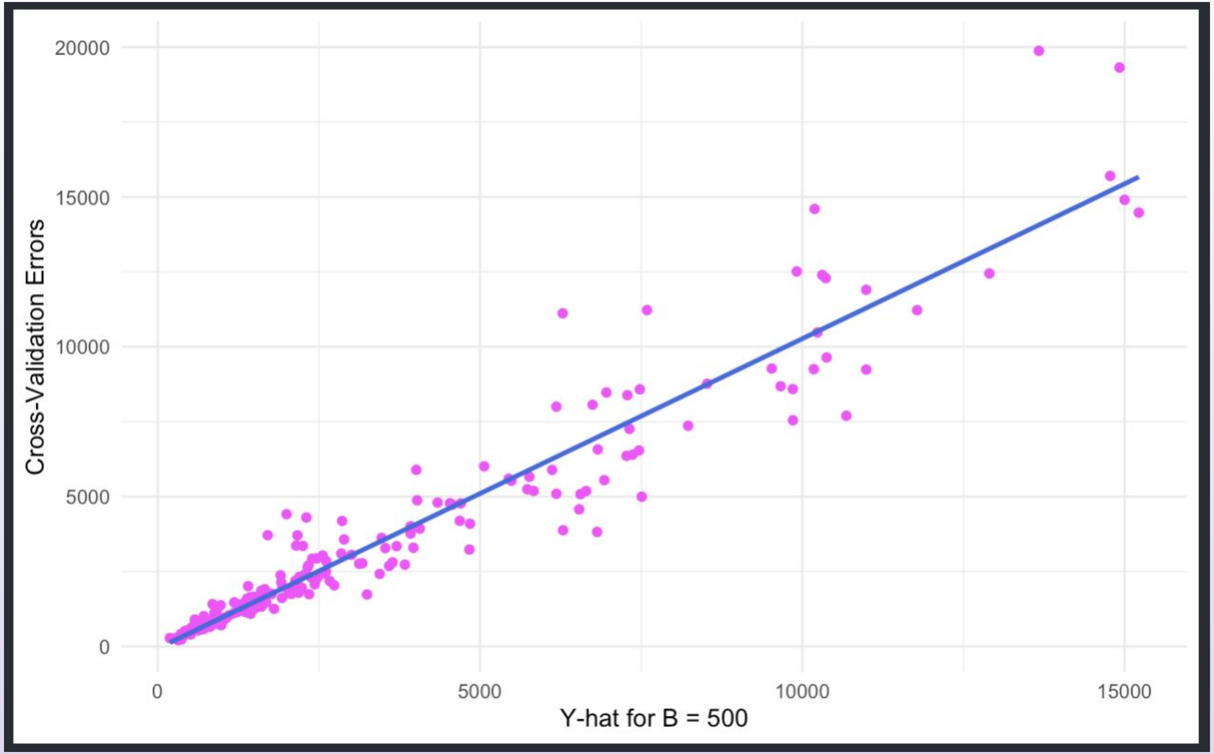
LOOCV (Leave-One-Out-Cross-Validation)

Cross Validation Curve for our decision tree

MSE: 2,277,347.9

RMSE: 1,509.09 college applications

Bagged Model



B = 500

MSE = 1,047,360.81

RMSE = 1,028.76

college applications

B = 1000

MSE = 1,053,600.79

RMSE = 1,011.62

college applications

Random Forest

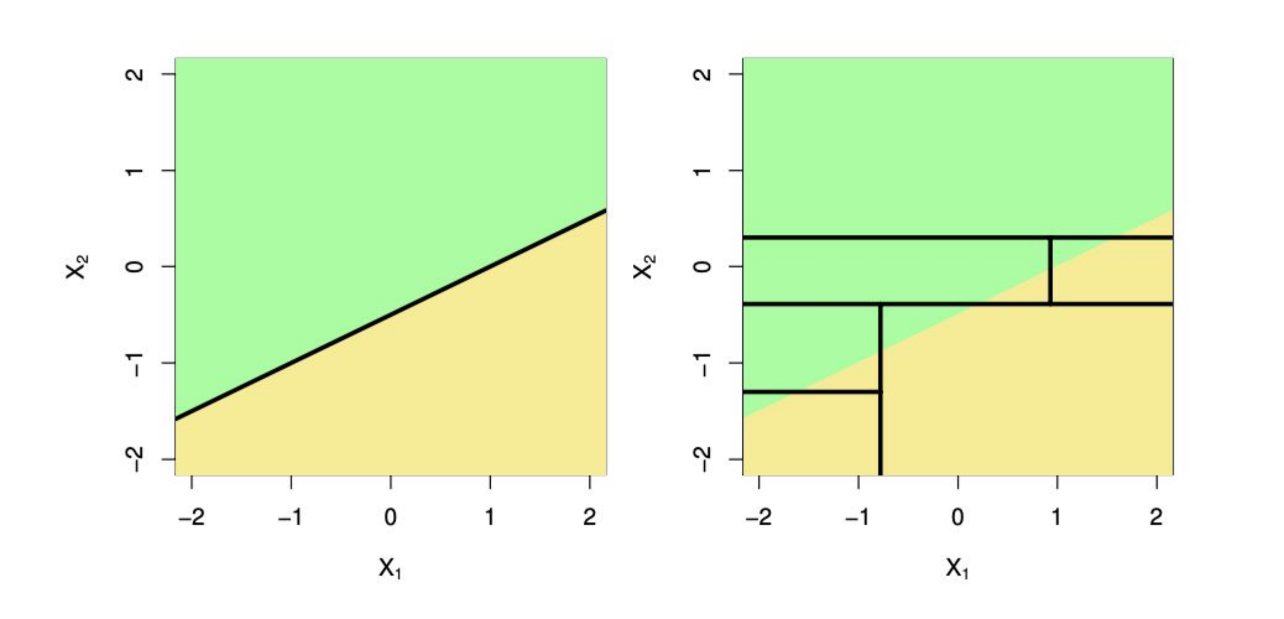


Image Source: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer. Pg 339

B = 500

MSE = 1,039,048.14

RMSE = 1,019.34

college applications

B = 1000

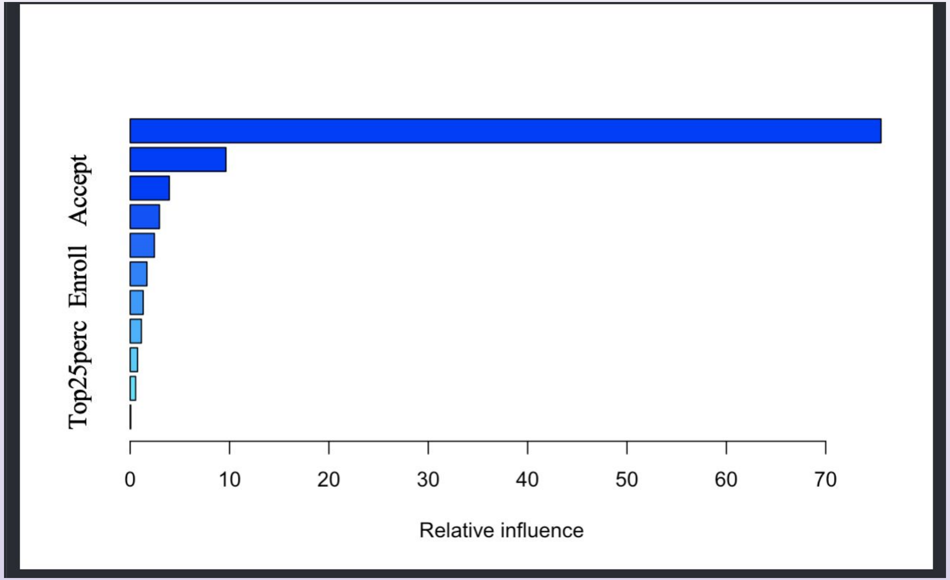
MSE = 1,061,330.47

RMSE = 1,030.21

college applications

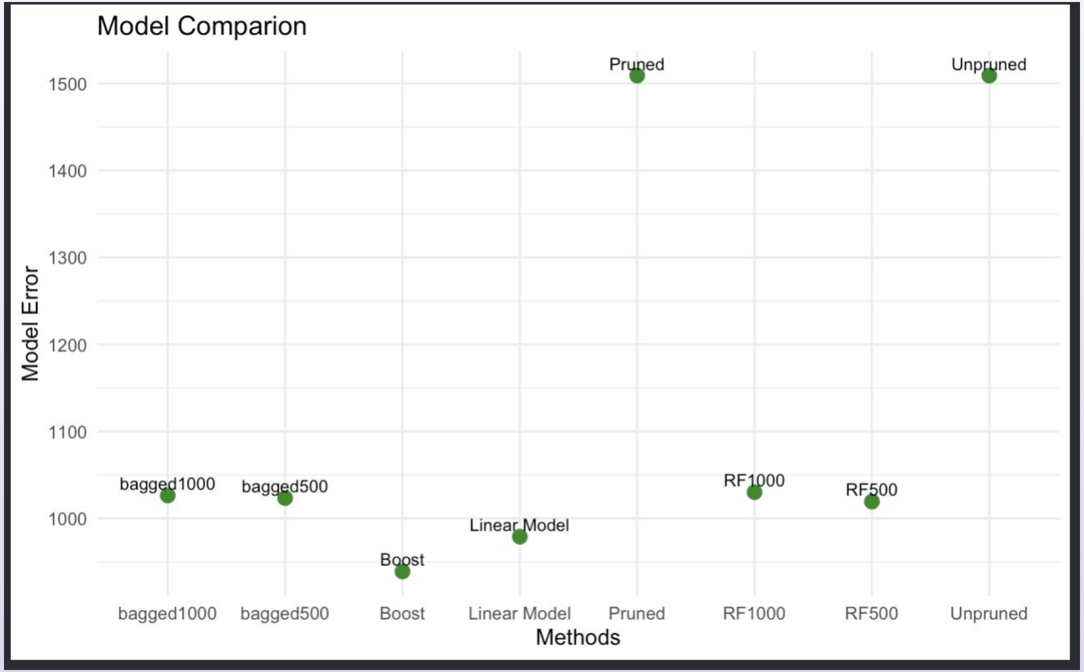
Boosting

	var <chr>	rel.inf <dbl>
Accept	Accept	74.920154805
Enroll	Enroll	7.892554785
F.Undergrad	F.Undergrad	7.652619921
Top10perc	Top10perc	3.174224259
Top25perc	Top25perc	2.479617338
Expend	Expend	1.304734574
Grad.Rate	Grad.Rate	0.931114986
Outstate	Outstate	0.758077700
Room.Board	Room.Board	0.605081796
PhD	PhD	0.274591851



- Boosting is an ensemble learning technique that combines multiple weak learners to create a strong learner.
- In a boosting model is a slow learner
- The final prediction is a weighted sum.

Final Method Comparison



Range of total applications received in original date set

Minimum: 81

Maximum: 48,094

Decision Tree RMSE: 1,509.09 College Apps

Bagged (500 trees) RMSE: 1,028.76 College Apps

Bagged (1000 trees) RMSE: 1,011.62 College Apps

Random Forest (500 trees RMSE: 1,019.34 College Apps

Random Forest (1000 trees) RMSE: 1,030.21 College Apps

Boosting RMSE: 933.43 College Apps

Linear Model RMSE: 979.09 College Apps

Summary

Our Best Model for this data:

Boosting Model where B = 50

- MSE = 555615.65
- RMSE = 939.09 college applications

Key Insights:

- **Decision Trees:** Flexible resulting in higher variability
- **Ensemble Methods:** Generally outperform individual decision trees and linear models, as evidenced by lower MSE and error rates.
- **Final Performance:** Our data ended up having a fairly linear relationship.
- **Comparative Performance:** Boosting models show competitive performance.

Thank You

Q&A