* **Interpret the error rate in terms of number of college applications on average we are off with each prediction method**

**Slide 1:**

Introduce ourselves.

**Slide 2:**

Explain that our data set came pre-cleaned, give a few examples of the type of information contained and where the data came from.

This dataset was collected in 1995 from 777 different colleges but only for that one year.

List the predictors?

**Slide 3:**

Methods that we used to analyze the data.

Linear Model

Decision Trees

Bagging

Random Forest

No formulas – but verbally convey what each model does.

**Slide 4:**

Linear Regression Results

Significance

Explain what top10perc means.

State that we removed statistically insignificant predictors in order to improve fit.

**Slide 5:**

Tree Results (Side by side)? -Should we run a larger tree?

**Slide 6:**

Bagging Results

An *ensemble* method is an approach that combines many simple “building block” models to obtain a single and potentially very powerful model. These simple building block models are sometimes known as *weak learners* since they may lead to mediocre predictions on their own. Why would bagging provide a better result than random forest in this context?

**Slide 7:**

Random Forest Results

The number of trees B is not a critical parameter with bagging; using a very large value of B will not lead to overfitting.

**Slide 8:**

Boosting

The number of trees B. Unlike bagging and random forests, boosting can overfit if B is too large, although this overfitting tends to occur slowly if at all. We use cross-validation to select B.

Cross Validation graph to select B

**Slide 9:**

BART?

Why was Bart so bad?

**Slide 10:**

Comparison of methods and results

Should we order the methods in terms of best to worst

**Slide 11:**

Summary Slide!!

**Decision Trees:** Unpruned and pruned decision trees show similar MSE and error rates.

**Ensemble Methods:** Bagged, Random Forest, and Boosting models generally outperform individual decision trees and linear models, as evidenced by lower MSE and error rates.

**Model Complexity:** Bagged and Random Forest models with larger values of B (number of trees) tend to have lower MSE and error rates, indicating the benefit of increasing model complexity.

**Comparative Performance:** Boosting models show competitive performance, offering lower error rates compared to decision trees and linear models.

**Model Interpretability:** Linear models have higher error rates compared to ensemble methods, indicating potential limitations in capturing complex relationships present in the data.

**Slide 12:**

Question Slides