**Slide 2:**

Give an overview of the data.

n = 777

Response = college applications

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

The goal of our project is to predict the number of college applications a university can expect to receive based off of this data.

**Slide 3:**

We initially ran our entire project based off of the entire data set.

After viewing the results we wanted to explore working with a reduced data set.

We performed Linear Regression.

Explain what top10perc means.

**Positive Impacts of removing statistically insignificant predictors when using ensemble methods:**

1. **You may reduce overfitting** in your ensemble models. This is because fewer irrelevant predictors mean the models are less likely to capture noise in the data.
2. Removing insignificant predictors can **simplify the** **models**, making them easier to interpret and potentially improving generalization to unseen data.
3. **Improved Computational Efficiency**

**Negative Impacts:**

1. **Information Loss**: Ensemble methods are often capable of capturing complex interactions between predictors that may not be evident in linear regression.

**However, our data turned out to have a natural linear relationship.**

1. Removingpredictors solely based on linear regression's significance tests may bias feature selection, potentially excluding important predictors that have nonlinear relationships with the target variable.

**Slide 4:**

Decision trees work by considering a split on each feature, and then calculating the node purity produced from each recursive binary split.

The tree then selects the decision splits that result in the most useful information.

We then create a tree of a certain depth based on the most important splits.

**Slide 5:**

Partition space graph.

As expected, Accept and Top10Perc were the split that yielded the most relevant information.

Now with this additional understanding we are able to more accurately predict the number of applications a university may receive, based on which region it belongs in smaller colleges with less prestige are going to receive less interest from prospective students than larger colleges that tend to attract top performing students.

**Slide 6:**

Is pruning our tree helpful?

Sometimes decision trees suffer from model complexity, we’ve already done some work to correct for this issue when we removed less significant predictors from our data set, this helps ensure that our model isn’t picking up noise in the data.

It turns out that out tree actually improved as we increased the depth of the tree.

We didn’t fit a larger tree, because it didn’t occur to us! We will certainly be doing that and reporting those results in our project report.

**Slide 7:**

Bagging –

Bootstrapping and Aggregating

With a normal decision tree we feed the entire date set to the tree and allow it to make it’s prediction

When we bootstrap though, we’re taking random subsets of the data (with replacement) meaning each tree gets a different input in order to make it’s prediction this leads to uncorrelating our trees

We then add all the models together with a weighted average to help us come to one conclusion

Overall, the "aggregating" part of bagging is crucial for leveraging the diversity of individual models created through bootstrap sampling and harnessing their collective predictive power to build a more accurate and robust ensemble model.

**Slide 8:**

Random Forests

They are extremely similar to the bagging model just discussed, they use bootstrapping but when each individual tree goes to make its decisions, it’s no longer allowed to consider all of the features. Instead, it’s forced to only consider a small subset of predictors.

In this case m = 3 for our random forest model.

**Slide 9:**

Boosting!

Still an ensemble method, meaning that we are building multiple decision trees and combining them at the end.

However, boosting differs from our previous methods in how the individual trees are built – and how they’re combined at the end.

“Weak Learners” or “Decision stumps”

Uses least squares loss function as a default for regression problems.

We stop and recalibrate at each progressive step making this model a slow learner.

**Slide 10:**

Method Comparisons

Recall the bias variance trade off.

Our data naturally has a pretty strong linear relationship, however it has a high variability in it’s data. Look at the range of our response variable.

As such it turns out that decision trees, bagging and random forests don’t perform as well with this data, because these model have low bias (they don’t make a lot of assumptions about the relationship in the data) and able to handle a lot of variability.

As we see a linear model which is high bias low variance actually did really well with this data set because the assumption of linearity plays well into this data set and the variability of our data doesn’t hurt this model as much as it does the trees.

Boosting delivers excellent performance because it learns and adjusts as it goes, we can see it did beat out the linear model and well out performed the other ensemble methods.

Our results varied wildly with different seeds, speaking to the large amount of variance in our data set. Our prediction can be viewed as good within the context of the larger colleges – but for small colleges our predictions are virtually useless.

We plan to explore this further, and attempt to normalize our response variable in order to get a better prediction that scales for the smaller schools. We will include the findings from this process in our final report.

**Slide 11:**

Questions?