FINAL PROJECT -- SPRING 2018

For this project we will be exploring the use of tree and logistic regression methods to classify schools as Private or Public based off their features. And we will use linear regression to analyze the data.

Let's start by getting the data which is included in the ISLR library, the College data frame (https://cran.r-project.org/web/packages/ISLR/ISLR.pdf).

A data frame with 777 observations on the following 18 variables.

- Private A factor with levels No and Yes indicating private or public university
- Apps Number of applications received
- Accept Number of applications accepted
- Enroll Number of new students enrolled
- Top10perc Pct. new students from top 10% of H.S. class
- Top25perc Pct. new students from top 25% of H.S. class
- F.Undergrad Number of fulltime undergraduates
- P.Undergrad Number of parttime undergraduates
- Outstate Out-of-state tuition
- Room.Board Room and board costs
- Books Estimated book costs
- Personal Estimated personal spending
- PhD Pct. of faculty with Ph.D.'s
- Terminal Pct. of faculty with terminal degree
- S.F.Ratio Student/faculty ratio
- perc.alumni Pct. alumni who donate
- Expend Instructional expenditure per student
- Grad.Rate Graduation rate

Get the Data

Call the ISLR library and check the head of College (a built-in data frame with ISLR, use data() to check this.) Then reassign College to a dataframe called df.

Check the head, structure and summary of the df.

Data Visualization

Let's explore the data!

Create a scatterplot of Grad.Rate versus Room.Board, colored by the Private column. Please interpret the scatterplot.

Create a histogram of full time undergrad students, color by Private. Please interpret the histogram.

Create a histogram of Grad.Rate colored by Private. You should see something odd here.

What college had a Graduation Rate of above 100% ? Change that college's grad rate to 100%.

Train Test Split

Split your data into training and testing sets 70/30. Set seed as 101.

Decision Tree

Use the rpart library to build a decision tree to predict whether or not a school is Private. Remember to only build your tree off the training data.

Use predict() to predict the Private label on the test data.

Check the Head of the predicted values. You should notice that you actually have two columns (Yes/No) with the probabilities.

Turn these two columns into one column to match the original Yes/No Label for a Private column.

Create a confusion matrix of your tree model. What's the model accuracy?

Use the rpart.plot library and the prp() function to plot out your tree model.

Random Forest

Now let's build out a random forest model!

Call the randomForest package library

Now use randomForest() to build out a model to predict Private class. Use n=500 trees. Add importance=TRUE as a parameter in the model. (Use help(randomForest) to find out what this does.

What was your model's confusion matrix on its own training set?

Predictions

Now use your random forest model to predict on your test set.

What's the confusion matrix on the test set? What's the model accuracy?

Logistic Regression

Use the caTools library to build a logistic regression model to predict whether or not a school is Private. Remember to only build your model off the training data.

Use predict() to predict the Private label on the test data.

Check the Head of the predicted values. You should notice that you actually have two columns (Yes/No) with the probabilities.

Turn these two columns into one column to match the original Yes/No Label for a Private column.

Use table() to create a confusion matrix of your logistic regression model. What's the model accuracy?

Now compare the model accuracy for the above tree model, random forest model and logistic regression model. Which one is the best for this data set?

Multiple Linear Regression

Use the caTools library to build a multiple linear regression model to predict the enrollment. Remember to only build your model off the training data.

Show the summary information of your model. What's the R-squared value of the training data?

Predict the enrollments on test data. Show the head of the prediction dataframe.

Whats are the MSE and R-squared value of the prediction? Is this a good model?

Great job! Thank you for your hard work of this semester!