**Capstone Project – Final Report**

**What is the problem you want to solve?**

Analyze the data of colleges and universities and classify them as either public or private. Based on this classification government funding can be given to schools who are public or scholarship to top students so as to improve public schools and bring them at par with private schools. This analysis may be used by the Education department of the Government They may use the analysis to decide on which areas to increase the spending e.g. creating more board rooms for students as they may study in more concentrated way and provide more privacy. This may impact the graduation rates of students.

**Deeper Dive into dataset**

1. The data is taken from the ISLR package. The data frame used is the ‘College’ data frame. Fields to be studied are

* 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

1. We will compare certain characteristics like Graduation Rate , Single Board Room occupancy , number of under grads enrolled in the universities to determine whether the school/college is a private or public one
2. There are some data outliers which were normalized. There was no N/As.

**Scope and Limitations**

This analysis depends on the College data set from the ISLR library. Some of the fields from this dataset are being used in the analysis to determine which colleges qualify as private and which as public. Some of variables analyzed are

* Room.Board Room
* F.Undergrad Number of fulltime undergraduates
* Grad.Rate Graduation rate

There are many other variables can be analyzed like number of PhDs or cost of books or ratio of student to faculty these may not qualify to classify a school as public or private.

**Further Research**

Further research can be done from the dataset. For example questions such as which colleges produce most of the research, whose alumni associations are most likely to donate and how much, application acceptance rate, tuition rate comparison etc can be researched.

===================================================================

**Complete Solution**

**Getting 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 dcf. (We are also loading the other required libraries in the same place)**

**# Load the required libraries used here**

**library(ISLR)**

**library(ggplot2)**

**library(caTools)**

**library(rpart)**

**library(rpart.plot)**

**library(randomForest)**

**#check the data**

**head(College)**

**#assiging College to dcf dataframe**

**dcf <- College**

**Exploratory Data analysis**

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

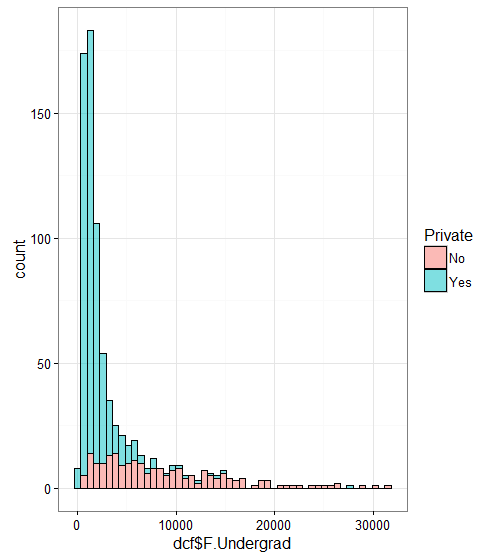
***ggplot(dcf, aes(dcf$Room.Board, dcf$Grad.Rate)) + geom\_point(aes(col=Private), size=2)***



Initial exploratory data analysis shows that Public schools have less graduation rate compared to private schools when compared against the number of board room occupancy. It seems that in private schools, students are provided with private board room where they can concentrate in their studies in silence and without disturbance.

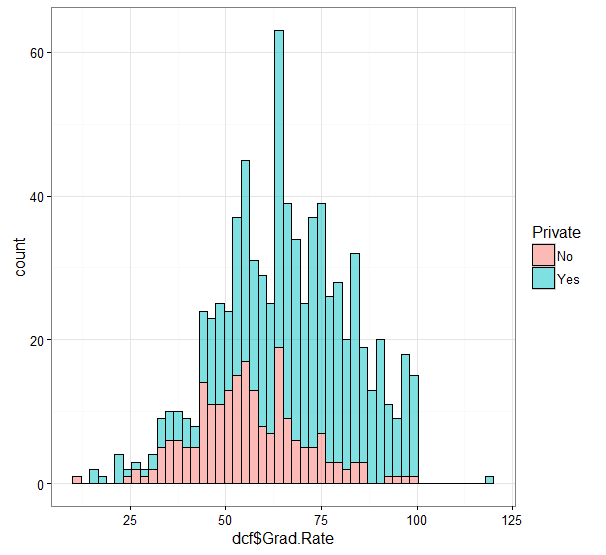
**Create a histogram of full time undergrad students, color by Private.**

**ggplot(dcf, aes(dcf$F.Undergrad)) + geom\_histogram(aes(fill=Private), col='black', bins=50, alpha=0.5) + theme\_bw()**



Another characteristic of private schools is the higher graduation rates

***ggplot(dcf, aes(dcf$Grad.Rate)) + geom\_histogram(aes(fill=Private), col='black', bins=50, alpha=0.5) + theme\_bw()***



**#Outlier data, 125% is more than 100**

**# Cazenovia College has an outliar with graduation rate more than 125%**

**subset(dcf, dcf$Grad.Rate > 100)**

**dcf['Cazenovia College', 'Grad.Rate'] <- 100**

**subset(dcf, dcf$Grad.Rate > 100)**

## Train Test Split

**Split the data into training and testing sets 70/30. Using the caTools library to do this.**

**#Split data into train, test**

**set.seed(101)**

**sample <- sample.split(dcf$Private, SplitRatio = 0.70)**

**train <- subset(dcf, sample == T)**

**test <- subset(dcf, sample == F)**

## Decision Tree

**Use the rpart library to build a decision tree to predict whether or not a school is Private.**

**tree <- rpart(Private ~ ., method = 'class', data = train)**

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

**#Intuitive Prediction**

**tree.preds <- predict(tree,test)**

**Check the Head of the predicted values. We have two columns with the probabilities**

**head(tree.preds)**

No Yes Private

Adelphi University 0.006116208 0.99388379 Yes

Albertus Magnus College 0.006116208 0.99388379 Yes

Alma College 0.006116208 0.99388379 Yes

American International College 0.006116208 0.99388379 Yes

Amherst College 0.006116208 0.99388379 Yes

Arizona State University Main campus 0.941666667 0.05833333 No

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

**tree.preds <- as.data.frame(tree.preds)**

**joiner <- function(x){**

**if (x >= 0.5){**

**return('Yes')**

**} else {**

**return('No')**

**}**

**}**

**tree.preds$Private <- sapply(tree.preds$Yes, joiner)**

**print(head(tree.preds))**

No Yes Private

Adelphi University 0.006116208 0.99388379 Yes

Albertus Magnus College 0.006116208 0.99388379 Yes

Alma College 0.006116208 0.99388379 Yes

American International College 0.006116208 0.99388379 Yes

Amherst College 0.006116208 0.99388379 Yes

Arizona State University Main campus 0.941666667 0.05833333 No

**using table() to create a confusion matrix of your tree model.**

**#creating a confusion matrix**

**table(tree.preds$Private, test$Private)**

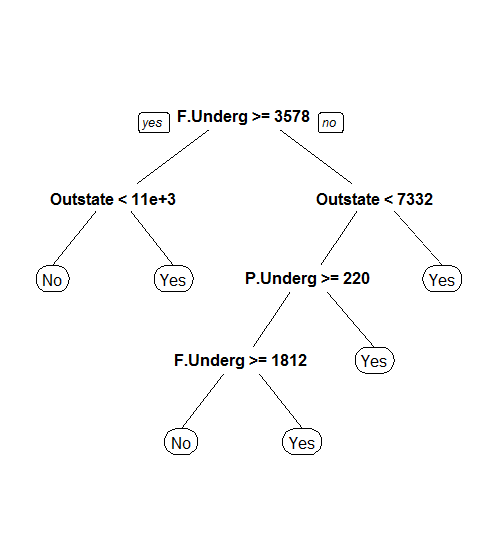
No Yes

No 46 3

Yes 18 166

**#plot the decision**

**prp(tree)**



**## Using randomForest**

**#decision Tree**

**rf.model <- randomForest(Private ~ ., data = train, importance = TRUE)**

**#Checking the importance of Confusion and importance**

**Our model's confusion matrix for the training set**

**rf.model$confusion**

No Yes class.error

No 128 20 0.13513514

Yes 12 384 0.03030303

**# Prediction for randomForest**

**rf.preds <- predict(rf.model, test)**

**table(rf.preds, test$Private)**

rf.preds No Yes

No 52 3

Yes 12 166

**Conclusions**

Private and public schools have certain characteristics which can be used to classify them as either public or private. Using the available dataset in the ISLR library we were able to analyze relationship between boardrooms and graduation rates, number of undergrads enrolled and number of graduates and based on this analysis of relationship use decision tree and randomForest methods to classify schools as public or private. This classification can be used by Government etc to analyze and grant support to required schools.