HarvardX: Ph125.9x Data Science Capstone College Graduation Rates Prediction Project

Chris Davidson

June 1, 2020

Introduction

In the field of Data Sciences and software, like R, provide data scientists the ability to conduct complex analyses of large datasets through machine learning to make predictions or classifications of an outcome. For the second component of the Harvardx Ph125.9x Data Science: Capstone Project, I created a large dataset from data in the Integrated Postsecondary Education Data System (IPEDS) from the US Department of Education's National Center for Education Statistics (NCES).

IPEDS Data

The Department of Education collects data annually through a series of interrelated surveys that all institutions–college, university, technical, and vocational–that participate in federal Title IV financial aid programs established by the Higher Education Act of 1965. Annually, more than 7,500 institutions ranging from research universities, state colleges, private religious and liberal arts institutions, for-profit schools, community and technical colleges, and non-degree-granting institutions complete the surveys for IPEDS.

Data collected from institutions include information in eight areas:

- Institutional Characteristics;
- Institutional Prices;
- Admissions;
- Enrollment;
- Student Financial Aid;
- Degrees and Certificates Conferred;
- Student Persistence and Success; and,
- Academic Libraries, Institutional, and Human Fiscal Resources.

Purpose Statement

As one of the definitive sources for information about postsecondary education in the US, IPEDS is used by Congress, federal and state government agencies, education leaders and researchers, private businesses, media, students, and parents for a variety of purposes. The purpose of this project was to develop and train a series of machine learning algorithms to predict college graduation rates reported in IPEDS and to maximize the accuracy of the algorithm. The Root Mean Square Error (RMSE) served as the value to evaluate the accuracy of each model. The RMSE is a standard way to measure the error between predicted and true values in a model. A smaller RMSE is interpreted as the model being more accurate than a larger RMSE.

This report provides a description of the methods used to download and clean the data, dataset descriptive statistics, the methods to conduct an initial exploratory data analysis of the training subset of data before creating a series of machine learning algorithms. The report will then provide the results and offer a discussion of the analysis before providing concluding remarks.

Methods & Analysis

College Graduation Dataset from IPEDS

The data for this project came from a sample of data from IPEDS. The code to build the college dataset was written to merge eight separate data files to create the College Graduation Dataset that included 10-years of data to account for any anomalies in the data. The ninth file provides the definitions of the variables in the dataset. Once the files were merged, any cases that had at least one null value for any variable were removed from the dataset. After removing cases with null values, eighty percent of the dataset was split into the training dataset used for tuning and 20% split into a test dataset. The code also removed unneeded files from the working directory and environment. Table 1 describes the variables in the dataset. Table 2 provides the first six rows in the dataset.

Table 1: College Graduation Dataset Variables and Definitions

Variable	Definition
UnitID	The identification number given by IPEDS for each institution.
Institution.Name	Official name of the institution.
Sector	Indicator for the control and level of the institution.
Year	Represents the year for which the the data was reported.
GradRate	The ratio of the number of students from the bachelor's degree-seeking cohort, who completed a bachelor's degree within 150 percent of normal time (6-years) divided by the adjusted cohort.
FTE	Derived by calculating the full-time equivalent of the institution's part-time enrollment and then adding it to the full-time enrollment of the institution.
PercAdmit	Ratio of the number of admissions offered divided by the total number of applicants.
PELL	Percentage of full-time, first-time degree/certificate-seeking undergraduate students who were awarded Pell grants.
SF.Ratio	Ratio of the total number of FTE students not in graduate or professional programs divided by the total FTE instructional staff not teaching in graduate or professional programs.
Retention	Percentage of students from the Full-time fall cohort that were enrolled in the prior year.

Table 2: First Six Rows of the College Graduation Dataset

	UnitID	Institution.Name	Sector	Year	GradRate	FTE	PercAdmit	PELL	SF.Ratio	Retention
2	222178	Abilene Christian University	Private not-for-profit, 4-year or above	2018	62	4295	57	30	15	77
3	138558	Abraham Baldwin Agricultural College	Public, 4-year or above	2018	23	3318	69	51	18	64
13	126182	Adams State University	Public, 4-year or above	2018	29	2345	99	54	12	54
14	188429	Adelphi University	Private not-for-profit, 4-year or above	2018	68	7171	74	30	12	80
17	168528	Adrian College	Private not-for-profit, 4-year or above	2018	53	1810	65	36	14	71
18	133872	AdventHealth University	Private not-for-profit, 4-year or above	2018	54	1174	87	62	7	66

The final College Graduation Dataset consisted of 29,103 graduation rates for 1,803 individual institutions across three sector types. An overwhelming majority of colleges are classified as Private not-for-profit, four-year or above; followed by public, four-year colleges; and Private for-private, four-year. Years in the dataset ranged from 2009 to 2018. Graduation Rate in the dataset ranged from 0% to 100% and a mean of 53%. Full-time equivalent enrollment, or FTE, ranged from 6 to 77,707 with a mean of 5,756 students. The low mean for FTEs given the range indicates that many of the institutions in the dataset have much smaller enrollments than larger enrollments. The Percent Admitted, used in this study as a proxy for selectivity in admissions, ranges from 0% to 100% with a mean of 66.62%. The Percent Admitted mean indicates that institutions in the dataset are less selective with a mean above 50%. The percentage of PELL Grant Recipients, a proxy for socioeconomic status, ranged from 0% to 100% with a mean of 40.46%. Given the mean percentage of PELL Grant Recipients, indicates that the first-time full-time degree-seeking students tend to not come from lower socioeconomic statuses. The Student-to-Faculty Ratios range from one student to 107 students for every one faculty member with a mean 14.14. The Retention Rates range from 0 to 100% with a mean of 75.28%, which are relatively high. Table 3 provides the full summary of all of the variables in the College Graduation Dataset.

Table 3: Descriptives of the College Graduation Dataset

UnitID	Institution.Name	Sector	Year	GradRate	FTE	PercAdmit	PELL	SF.Ratio	Retention
Min. :100654	Length:29130	Length:29130	Min. :2009	Min. : 0.00	Min. : 6	Min. : 0.00	Min. : 0.00	Min. : 1.00	Min. : 0.00
1st Qu.:154235	Class :character	Class :character	1st Qu.:2013	1st Qu.: 40.00	1st Qu.: 1084	1st Qu.: 55.00	1st Qu.: 26.00	1st Qu.: 11.00	1st Qu.: 68.00
Median :190567	Mode :character	Mode :character	Median :2018	Median : 53.00	Median : 2348	Median : 69.00	Median : 38.00	Median : 14.00	Median : 76.00
Mean :195760	NA	NA	Mean :2016	Mean : 53.21	Mean : 5756	Mean : 66.62	Mean : 40.46	Mean : 14.14	Mean : 75.28
3rd Qu.:217059	NA	NA	3rd Qu.:2018	3rd Qu.: 67.00	3rd Qu.: 6454	3rd Qu.: 81.00	3rd Qu.: 51.00	3rd Qu.: 17.00	3rd Qu.: 84.00
Max. :490805	NA	NA	Max. :2018	Max. :100.00	Max. :77707	Max. :100.00	Max. :100.00	Max. :107.00	Max. :100.00

Exploratory Analysis of the Training Dataset

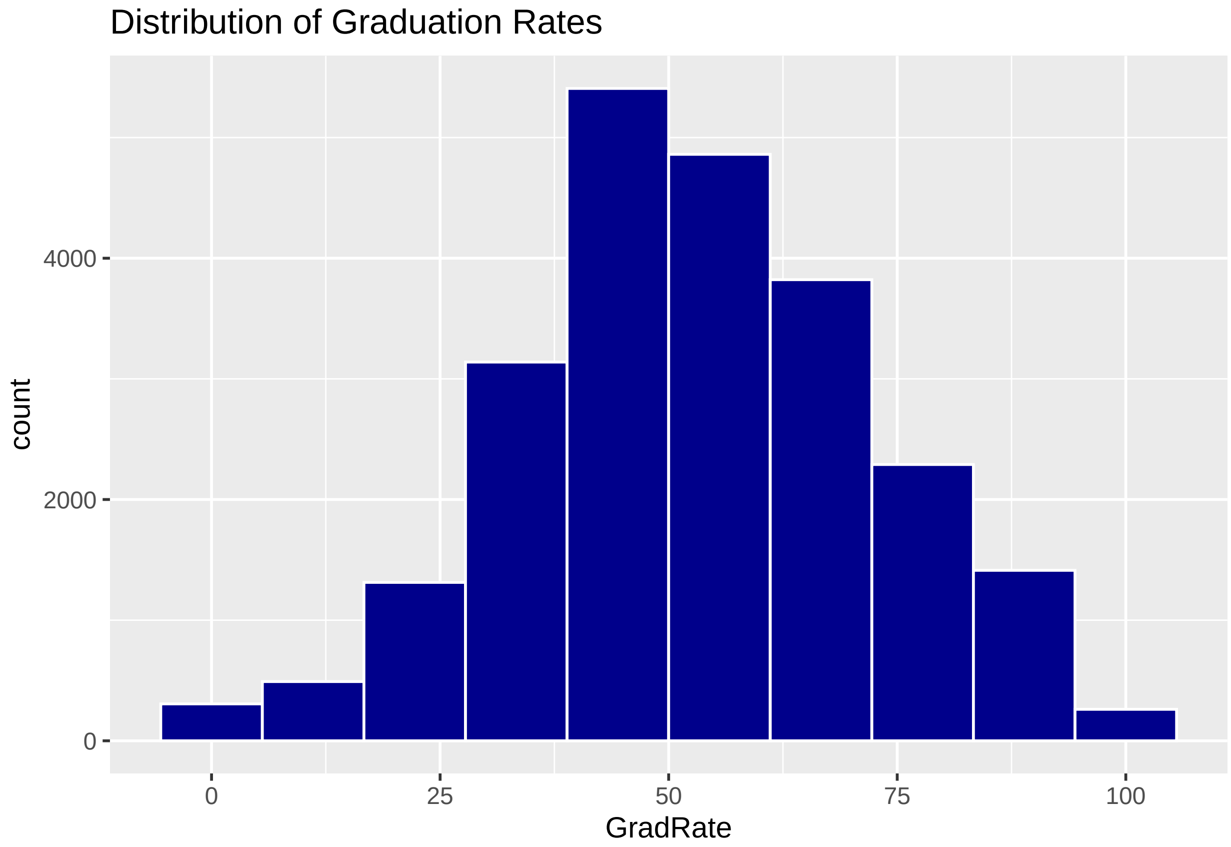
The first step in the analysis was to examine the training dataset to understand its components before using it to create the machine learning algorithms to predict graduation rates. Table 4 summarizes the training set and as we already know, there are no null values in the data because we removed those in the initial cleaning phase.

Table 4: Descriptives of Training Dataset

UnitID	Institution.Name	Sector	Year	GradRate	FTE	PercAdmit	PELL	SF.Ratio	Retention
Min. :100654	Length:23303	Length:23303	Min. :2009	Min. : 0.0	Min. : 6	Min. : 0.00	Min. : 0.00	Min. : 1.00	Min. : 0.00
1st Qu.:154095	Class :character	Class :character	1st Qu.:2013	1st Qu.: 40.0	1st Qu.: 1090	1st Qu.: 55.00	1st Qu.: 26.00	1st Qu.: 11.00	1st Qu.: 68.00
Median :190567	Mode :character	Mode :character	Median :2018	Median : 53.0	Median : 2362	Median : 69.00	Median : 38.00	Median : 14.00	Median : 76.00
Mean :195638	NA	NA	Mean :2016	Mean : 53.2	Mean : 5738	Mean : 66.61	Mean : 40.53	Mean : 14.15	Mean : 75.29
3rd Qu.:217040	NA	NA	3rd Qu.:2018	3rd Qu.: 67.0	3rd Qu.: 6445	3rd Qu.: 81.00	3rd Qu.: 51.00	3rd Qu.: 17.00	3rd Qu.: 84.00
Max. :490805	NA	NA	Max. :2018	Max. :100.0	Max. :77707	Max. :100.00	Max. :100.00	Max. :107.00	Max. :100.00

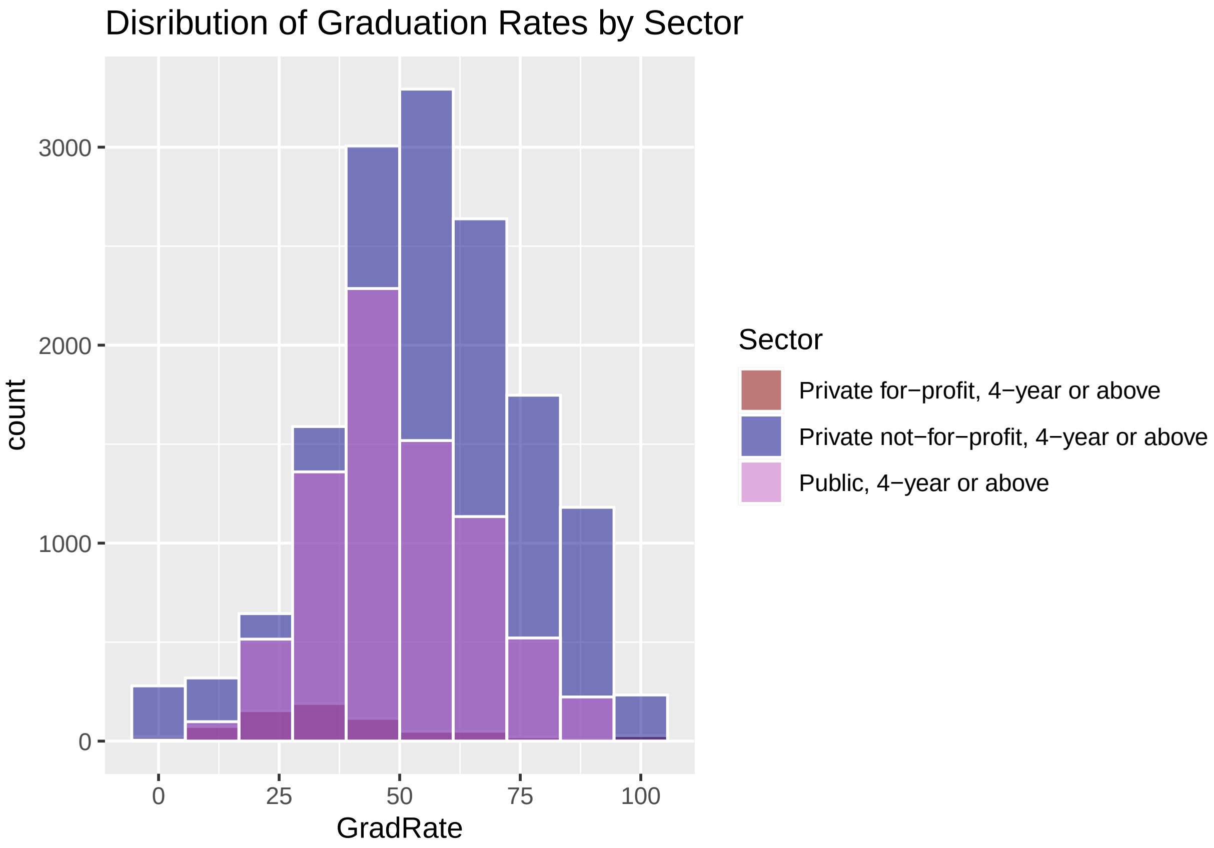
Distribution of Gradaution Rates

Graduation rates in the training dataset ranged from 0% to 100%. The majority of graduation rates were above approximately 40%. Even though some institutions had a 0% graduation rate, the distribution of graduation rates shows that the data is slightly skewed towards higher graduation rates.



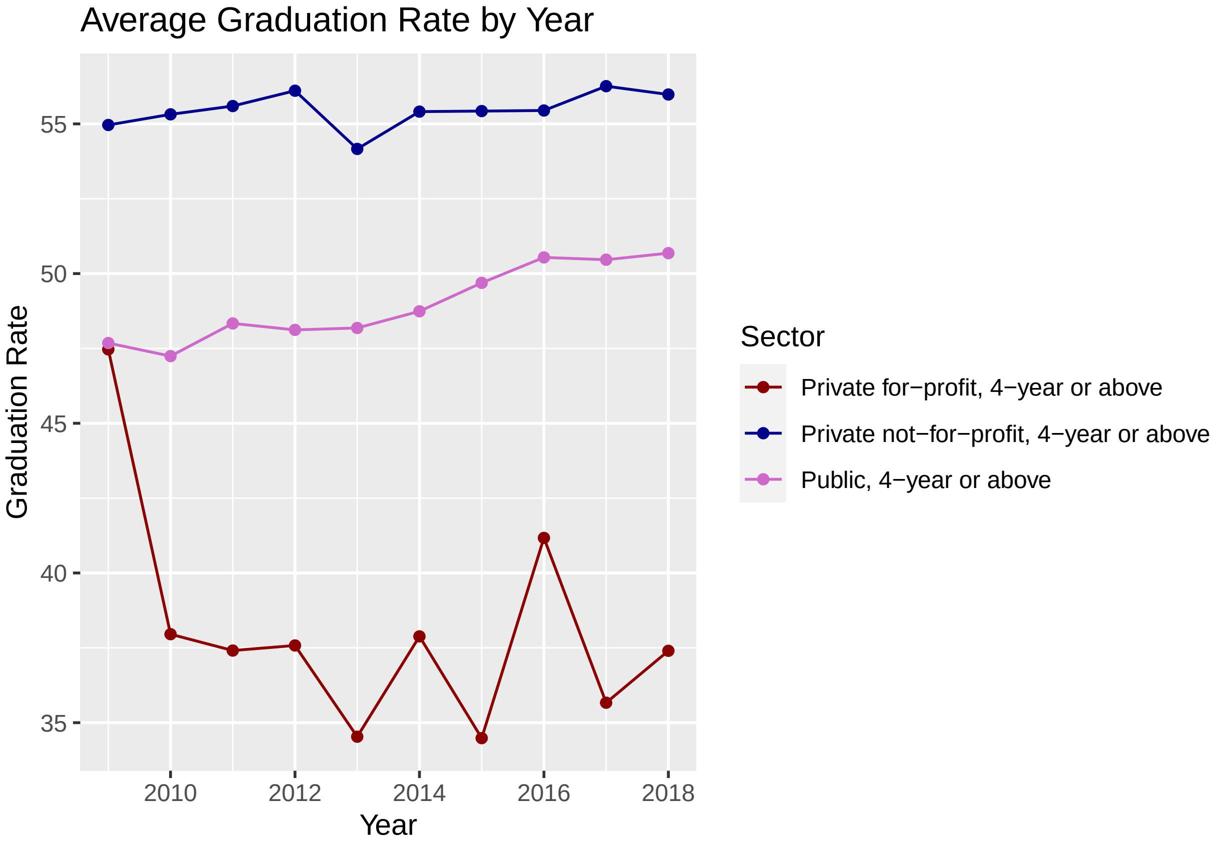
Graduation Rates by Institution Sector

Examining the distribution of graduation rates by institution sector shows the Private not-for-profit institutions had the highest graduation rates of the three sectors, which may partly be because of the number of institutions. Public institutions had the second-highest graduation rates that are centered around in the histogram. Institutions from these two sectors had significantly higher graduation rates than the Private for-profit institutions showing a bias against Private for-profit institutions in the data.



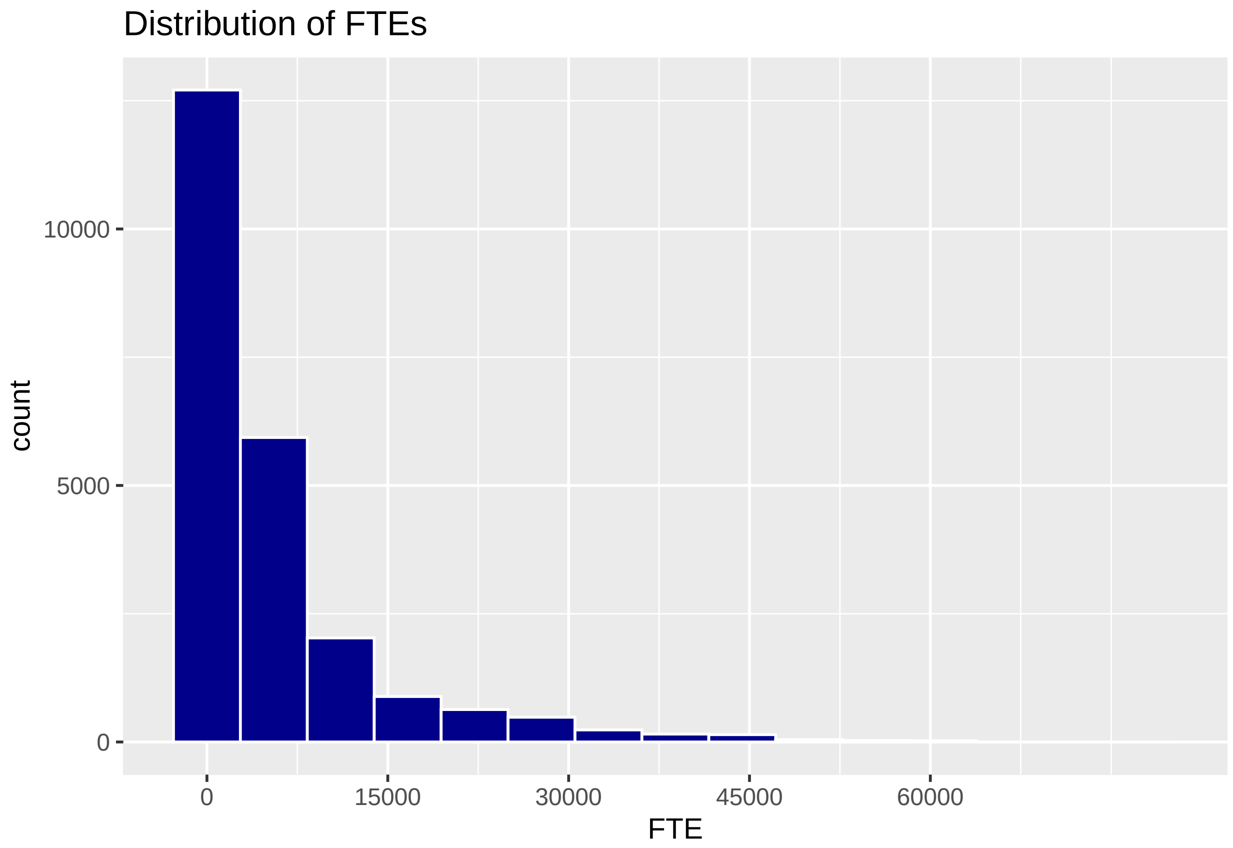
Graduation Rates by Year

As previously mentioned, 10-years' worth of data were used to take into consideration any anomalies or fluctuations in a single year within the data. Examining the average graduation rate by year for each sector, the Private for-profit sector shows the most fluctuations in average graduation rates over the decade. Overall, Public institutions saw a steady increase after a slight decrease in 2010. Private not-for-profit institutions saw a steep decrease from 2012 to 2013 and a steady increase afterward.



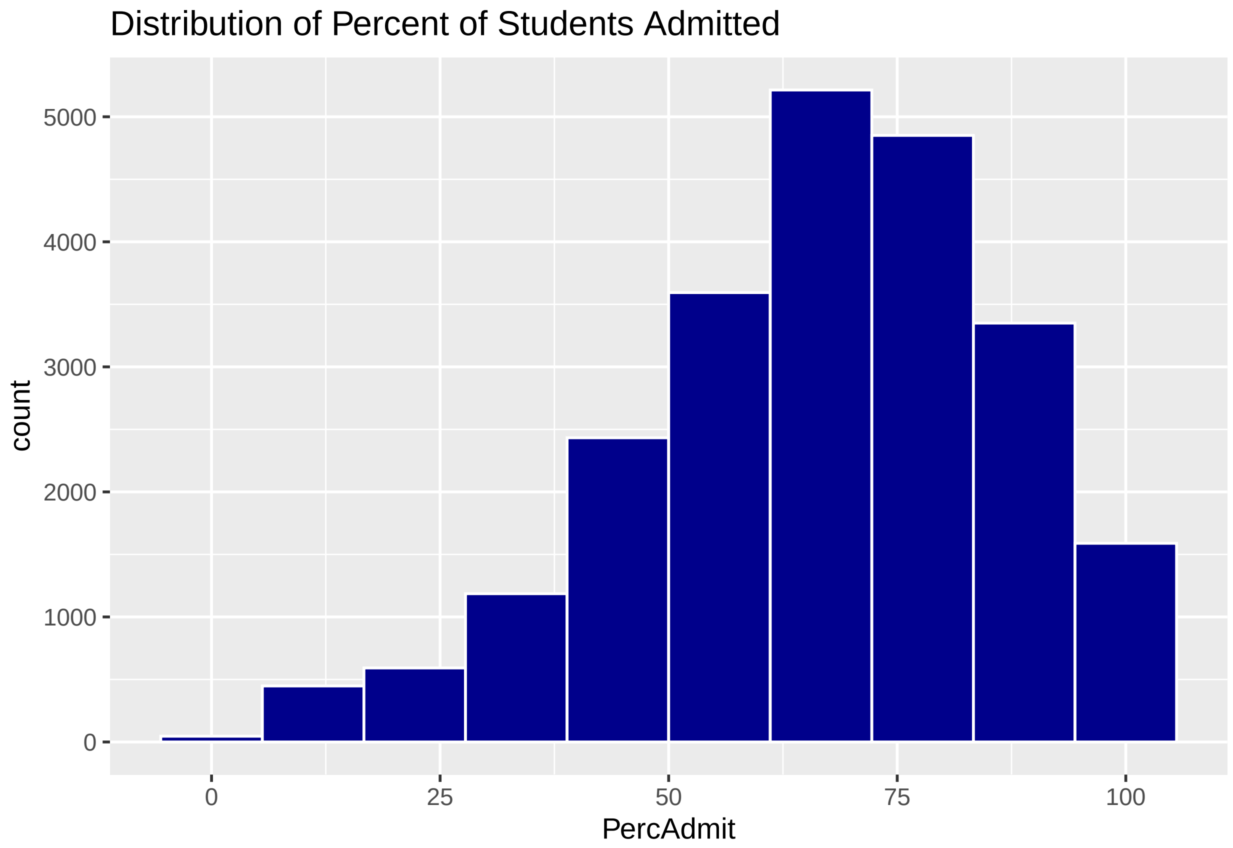
Distribution of FTEs

The distribution of the FTE enrollment shows that FTEs are skewed towards smaller enrollments with fewer than 15,000 total FTEs.



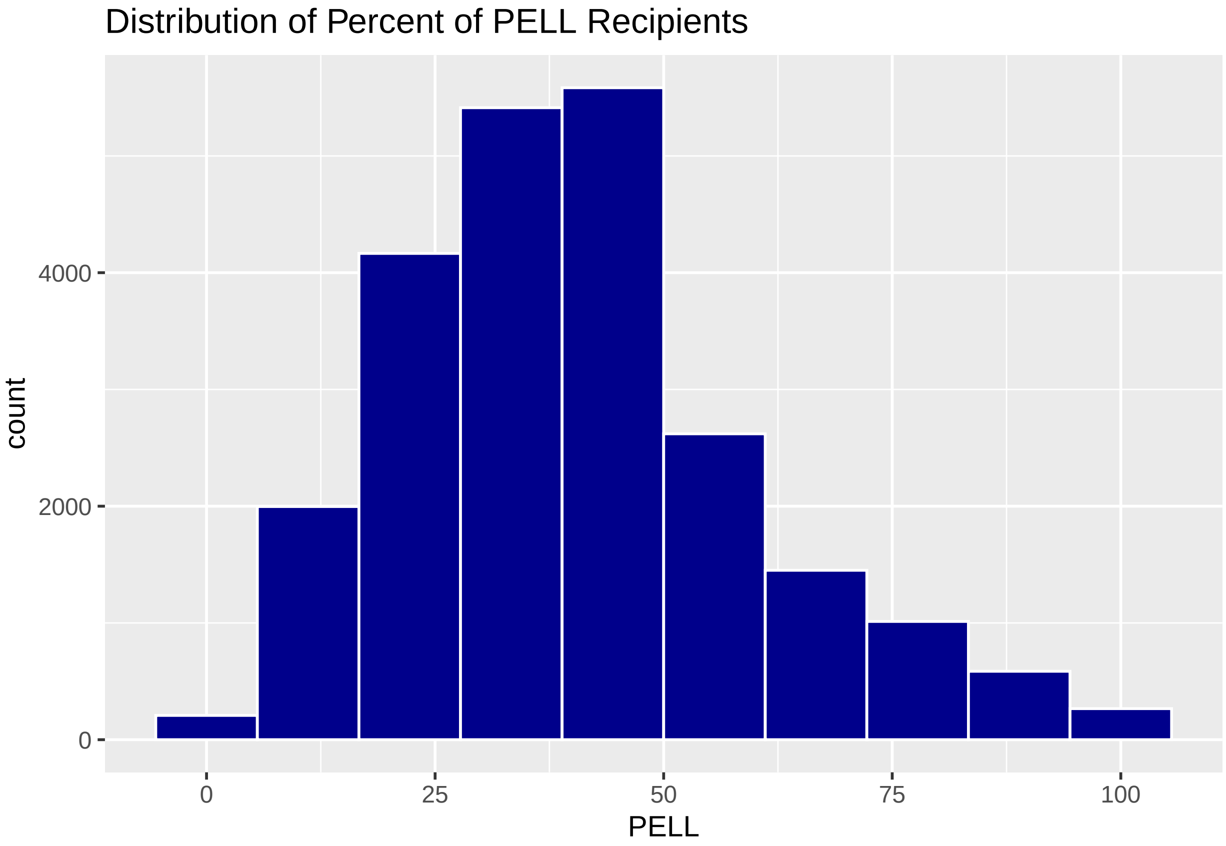
Distribution of Percent Admitted

The distribution of the Percent of Students Admitted, representing the selectivity of institutions, shows that an overwhelming majority of institutions are less selective with the percent admitted well above 50%.



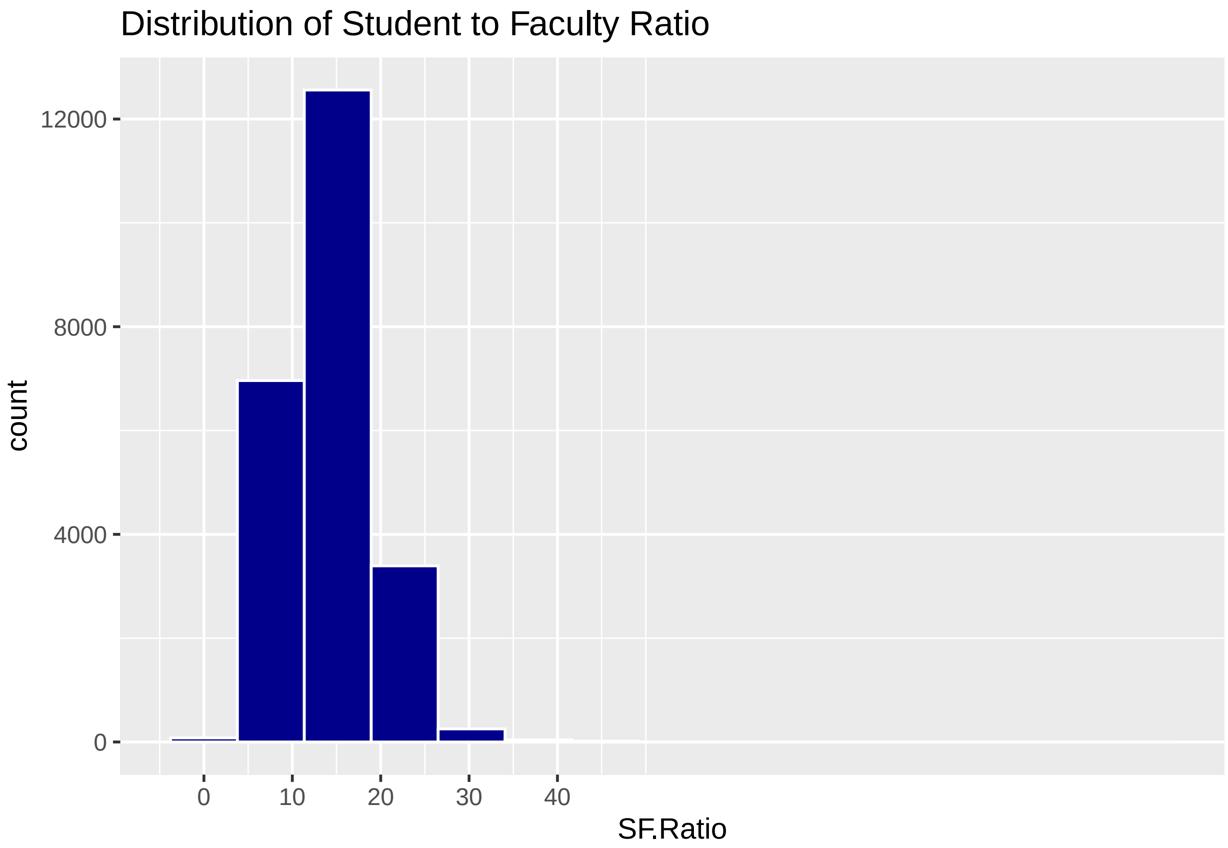
Distribution of Percent of Pell Recipients

The distribution of Pell Grant Recipients shows that an overwhelming majority of institutions do not have a large population of Pell Grant recipients indicating that there are fewer students from lower socioeconomic statuses.



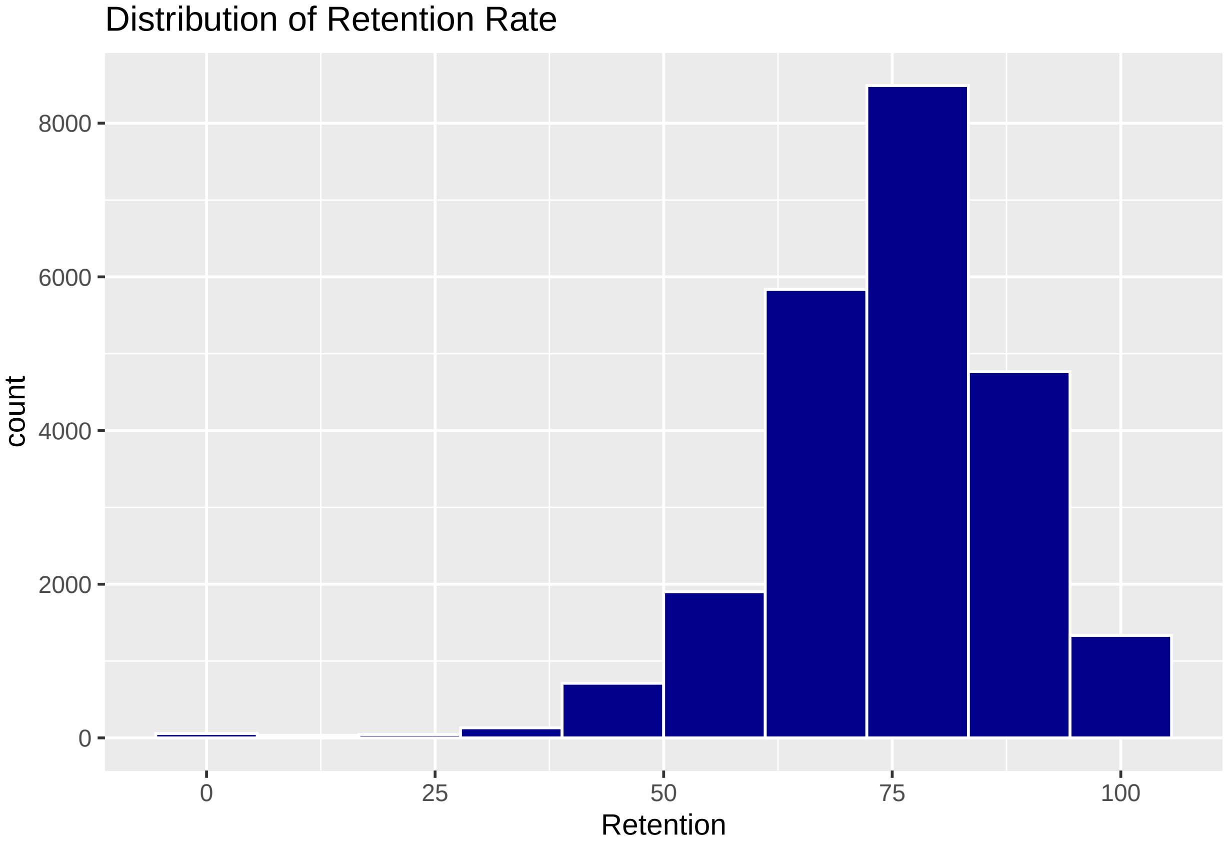
Distribution of the Student to Faculty Ratio

The distribution of student to faculty ration shows that most institutions have a student to faculty ratio of 20 or fewer students for every one faculty member.



Distribution of Retention Rate

The distribution rate shows that a majority of institutions have a 6-year graduation rate of more than 60%, which is a good indicator that many students will be retained and persist to graduation.



Modeling

For the learning algorithms, I used the RMSE to determine the accuracy of the model. As previously stated, the RMSE, or Root Mean Square Error, measures the error between the predicted and observed values. This means that the smaller the RMSE, the more accurate the model. Mathematically, the formula for the RMSE is written as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

For this project, I will create five models to evaluate the predictive power of each model. The five following models: (a) Best Guess, (b) Multiple Linear Regression, (c) Regression Tree, (d) Pruned Regression Tree, and (e) Random Forest models, are defined in subsequent sections.

Model I: Best Guess Model

The first model is a simple best guess for the graduation rate. This model ignores all predictors and all errors are explained as random variation in the model. We know that the mean of the graduation rates is 53.20553, which is close enough to a 50/50 guess. We know that the simplest best guess model is just that, a guess, meaning we can do better and improve the accuracy of the predictions.

Model II: Multiple Linear Regression

Multiple linear regression can be used as a form of machine learning that controls for the various predictors in the model. While simple multiple linear regression can be useful, at times it can be too rigid for predictive power depending on the data. In this case, I will use it as the general baseline along with the Best Guess model as the baseline for more advanced models.

Model III: Regression Tree

Regression trees are a basic yes or no flow chart. Since our outcome variable, graduation rates, is continuous and we have more than a few predictors, a regression tree allows us to predict the outcome by partitioning the predictors. This partitioning occurs and creates nodes, or trees with predictions at the ends. Within the partitioning process, the complexity parameter and the minimum number of observations required before moving on to the next partition.

Model IV: Pruned Regression Tree

Once I create my initial regression tree, I will use the pruning process to apply a higher complexity parameter. This simply means that I will be snipping off partitions that do not meet the complexity parameter criterion.

Model V: Random Forests

The last model I will create will be the random forests model. The goal is to improve the predictive power and reducing any instability by introducing randomness through bootstrapping and by averaging multiple decision trees. Within this process bootstrapping. Since the con to this method is the loss of interpretability, I will examine variable importance to determine which variable(s) are most important in the model.

Results

My purpose for this project was to develop and train a series of machine learning algorithms to predict college graduation rates reported in IPEDS and to maximize the accuracy, measured by the RMSE.

Model I - Best Guess

Without using any other predictors, our best guess for a graduation rate would be 53.20281%. Using the formula and respective function I calculated the RMSE as 19.93759. This RMSE will be used as the baseline to compare all other models moving forward because the model should become more accurate as predictors and algorithms are tuned and tested.

Table 5: Results of RMSEs

Method	RMSE
Model I: Best Guess - Baseline	19.93759

Model II - Multiple Linear Regression

Through the exploratory data analysis, we found that the graduation rates and other variables did not have a normal distribution, which is not surprising given the type of data being used. Examining the linear model, all of the variables were statistically significant with a p-value < .05, except for the Private not-for-profit sector, 4-year and above sector. To accept the linear model, I would need to be examined further because the model only explains about 59.3% in the variance of predicted values, which means the model is not a strong predictor even though the RMSE improved greatly to 12.847343.

Table 6: Multiple Linear Regression

term	estimate	std.error	statistic	p.value
(Intercept)	-527.7927063	55.0835419	-9.5816770	0.0000000
SectorPrivate not-for-profit, 4-year or above	0.0312564	0.4970102	0.0628888	0.9498556
SectorPublic, 4-year or above	-5.1589005	0.5080506	-10.1543037	0.0000000
Year	0.2824828	0.0273447	10.3304309	0.0000000
FTE	0.0003388	0.0000127	26.7014633	0.0000000
PercAdmit	-0.1422368	0.0043089	-33.0096941	0.0000000
PELL	-0.3592566	0.0052122	-68.9256414	0.0000000
SF.Ratio	-0.4658399	0.0220982	-21.0804843	0.0000000
Retention	0.5577021	0.0077751	71.7296882	0.0000000

Table 7: Results of RMSEs

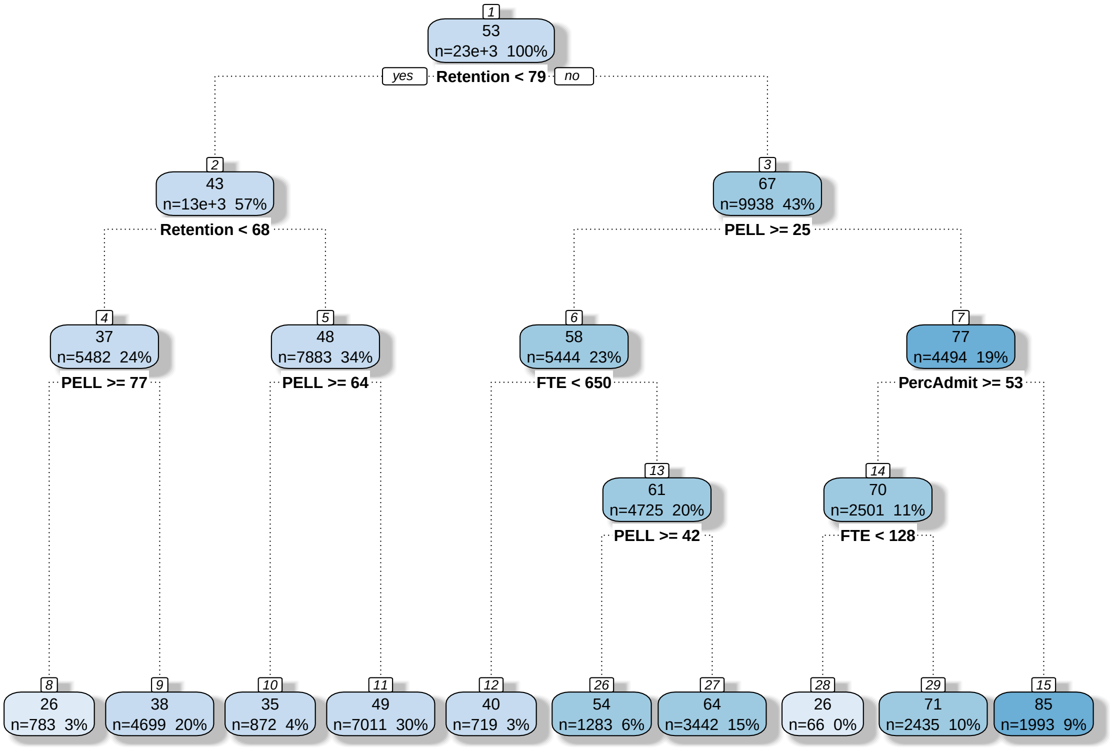
Method	RMSE
Model I: Best Guess - Baseline	19.93759
Model II: Multiple Linear Regression	12.84734

Model III - Regression Tree

Moving beyond simple linear regression, I used Regression Trees as the next way to create a predictive model. In this regression tree, we see a total of nine nodes where the tree branches. For institutions that have a retention rate below 79%, retention rates below 68%, and have 77% or more first-time full-time students with PELL Grants, the average graduation rate is 26%. For the institutions that have below a 77% PELL Grant rate, the average graduation rate is 38%. For colleges that have a retention rate above 68% but under 79%, and have a PELL Grant rate of 64% or above the average graduation rate is 35%. For colleges that have a retention rate above 68% but under 79%, and have a PELL Grant rate of below 64% the average graduation rate is 49%.

For Retention Rates above 79% with a PELL Grant rate of 25% or above and an FTE below 650, the average graduation rate is 40%. For Retention Rates above 79% with a PELL Grant rate of 25% or above and an FTE above 650 and the PELL Grant rate is equal to or above 42%, the average graduation rate is 54%. For Retention Rates above 79% with a PELL Grant rate of 25% or above and an FTE above 650 and the PELL Grant rate is below 42%, the average graduation rate is 64%.

For Retention Rates above 79% with a PELL Grant rate of less than 25%, the Percent Admitted above or equal to 53% and an FTE below 128, the average graduation rate is 26%. For Retention Rates above 79% with a PELL Grant rate of less than 25%, the Percent Admitted above or equal to 53% and an FTE above 128, the average graduation rate is 71%. For Retention Rates above 79% with a PELL Grant rate of less than 25%, the Percent Admitted below 53%, the average graduation rate is 85%. The Regression Tree model improved slightly from the multiple linear regression model to 12.74843.



Rattle 2020−Jun−01 10:25:51 cdavi

Table 8: Results of RMSEs

Method	RMSE
Model I: Best Guess - Baseline	19.93759
Model II: Multiple Linear Regression	12.84734
Model III: Regression Tree	12.74843

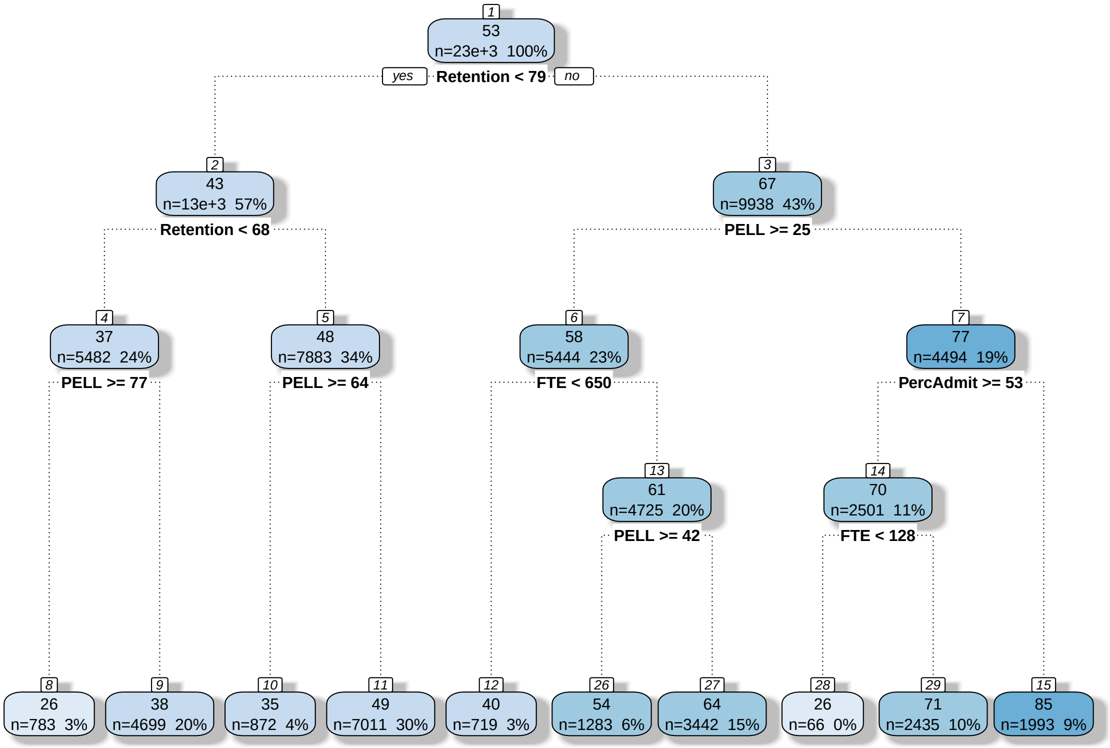
Model IV - Pruned Regression Tree

Taking the previous model, and by using cross-validation, I found the optimal complexity parameter by examining the smallest standard error of the factors and at which split it occurred. By pruning these lower-level decision nodes, I can introduce a little bit of bias in the model that helps stabilize predictions for the future and in turn will be able to better generalize to new data. The lowest complexity parameter occurred at the ninth split; however, after building the Pruned Regression Tree, there is no improvement in the RMSE.

```
##
## Regression tree:
## rpart(formula = GradRate ~ Sector + Year + FTE + PercAdmit +
##     PELL + SF.Ratio + Retention, data = train)
##
## Variables actually used in tree construction:
## [1] FTE       PELL      PercAdmit Retention
##
```

```
## Root node error: 9117120/23303 = 391.24
##
## n= 23303
##
##          CP nsplit rel error  xerror      xstd
## 1  0.347850      0   1.00000 1.00013 0.0088170
## 2  0.089522      1   0.65215 0.65322 0.0080380
## 3  0.044607      2   0.56263 0.56167 0.0074416
## 4  0.029781      3   0.51802 0.51949 0.0075233
## 5  0.028453      4   0.48824 0.49517 0.0071400
## 6  0.016853      5   0.45979 0.46148 0.0067076
## 7  0.014613      6   0.44294 0.44703 0.0066649
## 8  0.011581      7   0.42832 0.42986 0.0063098
## 9  0.011103      8   0.41674 0.41785 0.0062646
## 10 0.010000      9   0.40564 0.41063 0.0062764
```

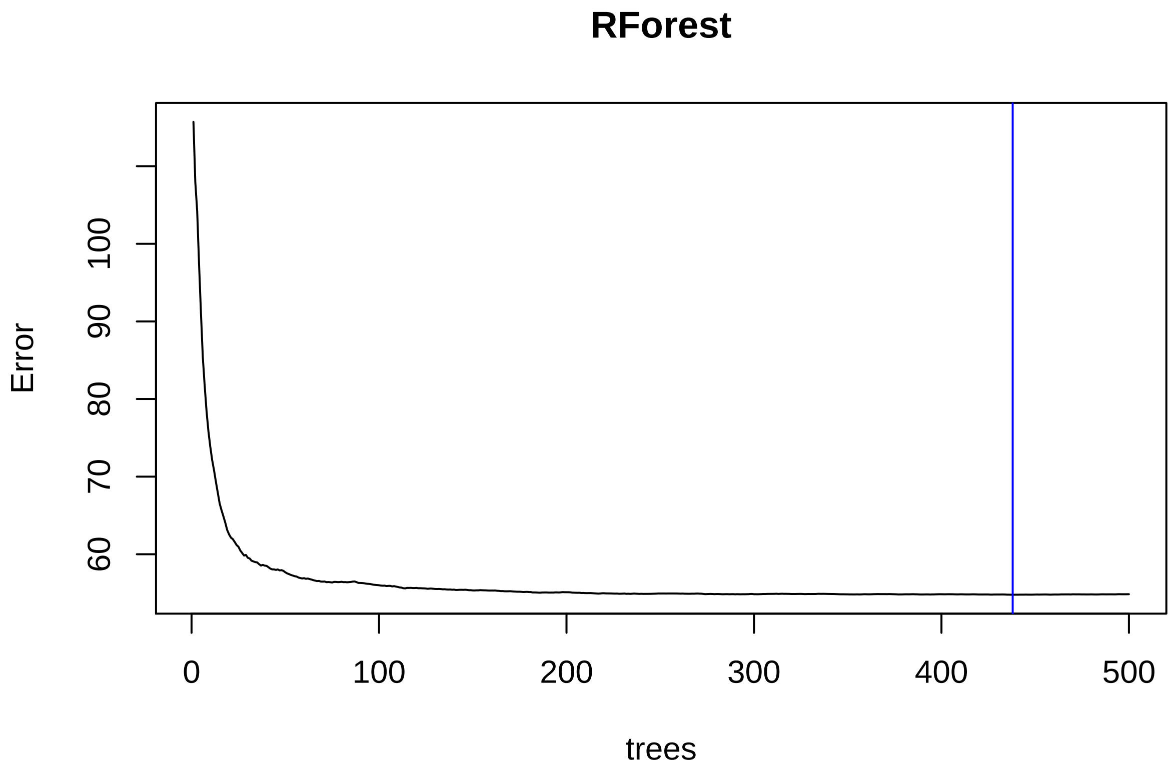
```
## # A tibble: 10 x 5
##        CP nsplit rel.error xerror    xstd
##     <dbl>  <dbl>     <dbl>  <dbl>   <dbl>
##  1 0.348       0     1      1.00  0.00882
##  2 0.0895      1     0.652  0.653 0.00804
##  3 0.0446      2     0.563  0.562 0.00744
##  4 0.0298      3     0.518  0.519 0.00752
##  5 0.0285      4     0.488  0.495 0.00714
##  6 0.0169      5     0.460  0.461 0.00671
##  7 0.0146      6     0.443  0.447 0.00666
##  8 0.0116      7     0.428  0.430 0.00631
##  9 0.0111      8     0.417  0.418 0.00626
## 10 0.01        9     0.406  0.411 0.00628
```



Rattle 2020−Jun−01 10:25:52 cdavi

Model V - Random Forest

I used the Random Forest algorithm to predict college graduation rates as the final model reaches the minimum error estimate occurred is after 438 trees, where the line on the plot levels out. Since the limitation of Random Forests is interpretability, I examined the importance of variables by looking at how many times it appears in the trees. The Retention Rate variable appears to be the variable used most and is driving the model, which is logical because institutions with higher retention rates would have higher graduation rates. After using the model to predict graduation rates, the RMSE dropped significantly to 7.391587 from all other models.



```
## numeric(0)
```

Table 9: Minimum Error Estimate

x
438

Table 10: Variable Importance

	IncNodePurity
Sector	206500.4
Year	261120.3
FTE	1257266.1
PercAdmit	1013661.2
PELL	2407620.1
SF.Ratio	691379.3
Retention	3044081.7

Table 11: Results of RMSEs

Method	RMSE
Model I: Best Guess - Baseline	19.937594
Model II: Multiple Linear Regression	12.847343
Model III: Regression Tree	12.748425
Model IV: Pruned Regresstion Tree	12.748425
Model V: Random Forest	7.391587

Limitations

As with any project, there are limitations. For this project, the limitations include the dataset itself and within some of the models. Within the dataset itself, I did not choose all possible variables from IPEDS that could have been included because of the sheer amount of data collected by the Department of Education. Additionally, some of these variables, including those that I ended up choosing could be correlated with one another and end up skewing some of the results. This is evident in the linear regression model and the percentage of variance, or lack thereof, that is explained.

Conclusion

Using data from IPEDS at the Department of Education, I used a series of machine learning algorithms to predict graduation rates at Public, Private not-for-profit, and Private for-private four-year institutions. After using Multiple Linear Regression, Regression Trees, and Random Forests algorithms, the best model produced came from the Random Forests mode with the smallest RMSE. Moving forward, future research may use these and other algorithms with additional data, including some of the institutions excluded because of missing values. Additionally, other variables could be looked at from IPEDS as well as looking to avoid issues of collinearity and covariance, which affects the final models. With the inclusion of other data and building upon the work of this project, insights gained could provide information for data-informed decision making at institutions of higher education to determine where to invest resources to increase an institution's graduation rate.