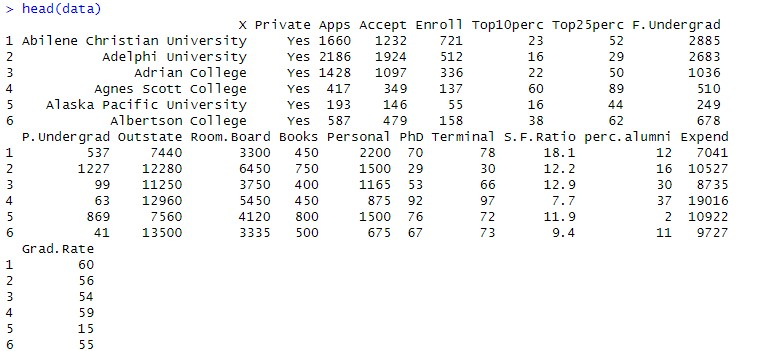
Introduction

This study describes an analysis of the College dataset using logistic regression. The goal is to figure out if a university is private or not based on the facts about it. Exploratory data analysis (EDA), making the logistic regression model, evaluating the model's performance, and figuring out what the results mean are all parts of the report.

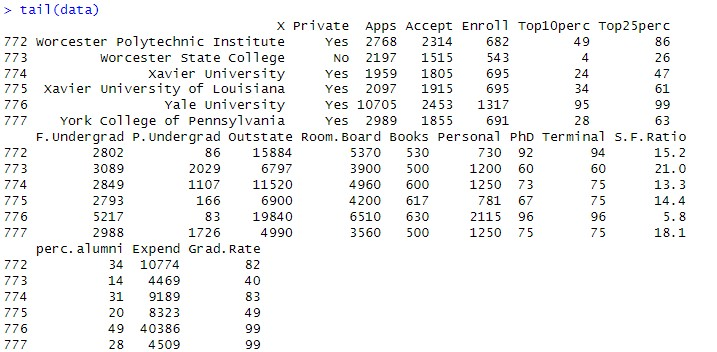
Exploratory Data Analysis

During the EDA step, the goal is to learn about the most important parts of the dataset. We begin by loading the information and looking at the first and last few rows. There were also summaries of the data for each variable. There were no lost data points in the set. A graph of the Out-of-State Tuition (Outstate) variable was made to see how it was spread out. This graph showed how often universities charged different fee rates.

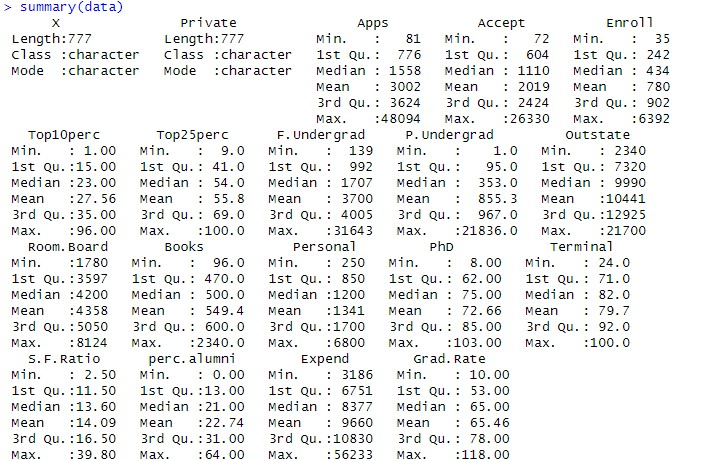
Visualizing the First 5 Values using the head() function.



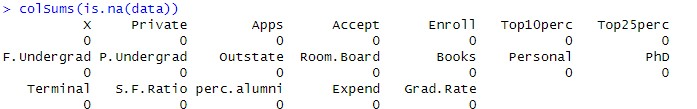
Visualizing the Last 5 Values using the tail() function.



Analyzing the dataset and finding the descriptive statistics of the dataset we use used summary () function.



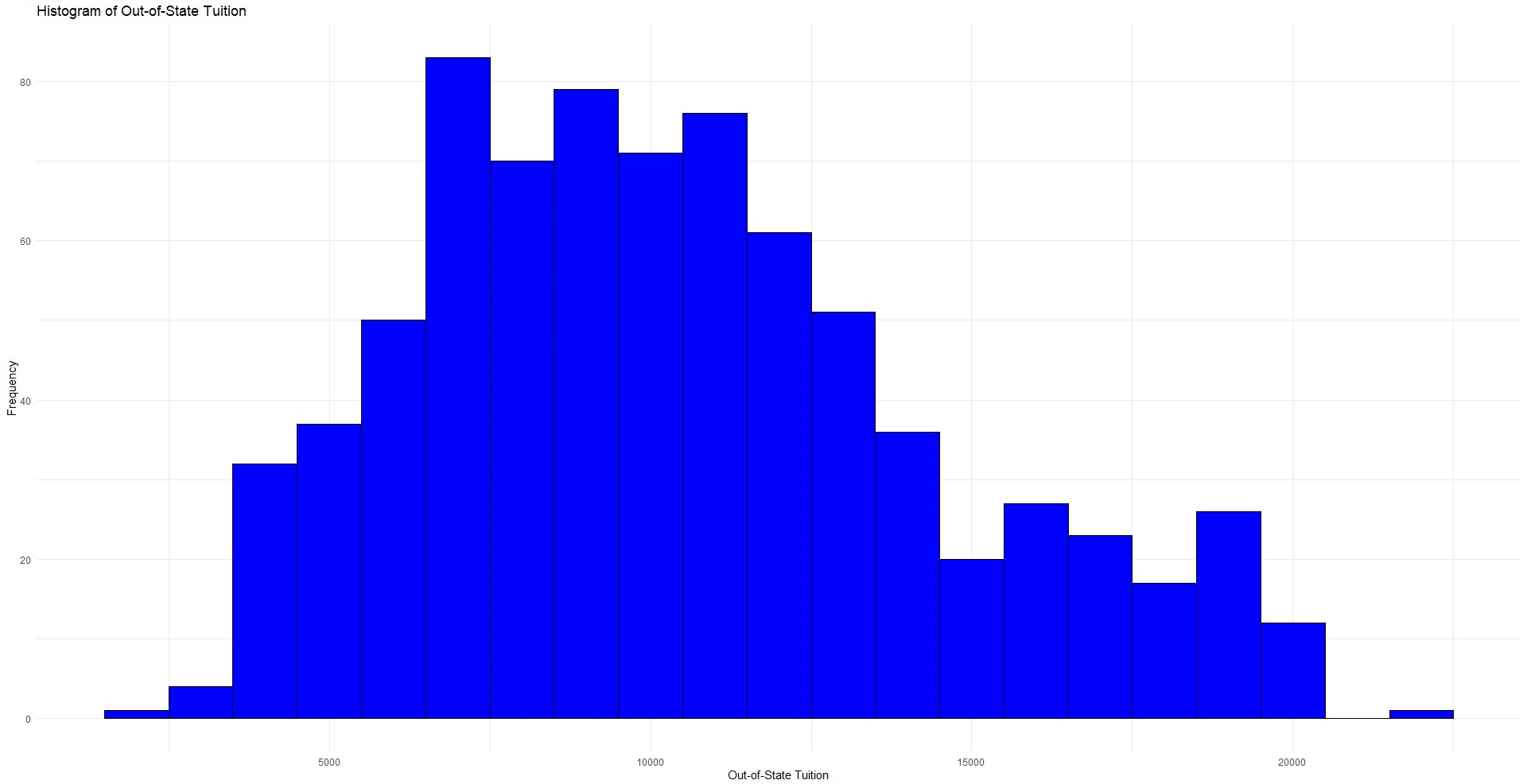
To check if there are any null values we used is.na() function



Histogram for Out-Of-State Tuition

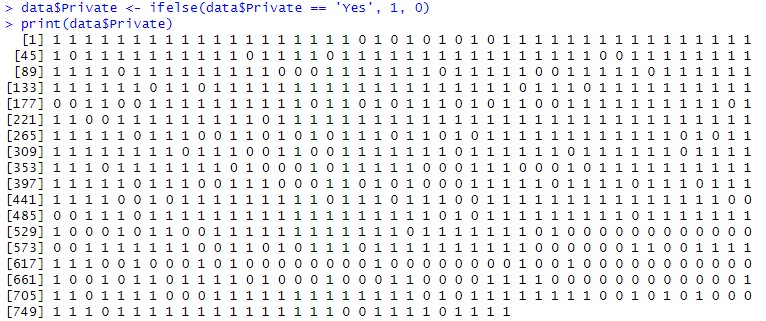
This code's goal is to generate a histogram that shows how the cost of out-of-state tuition varies depending on the dataset. With the x-axis showing out-of-state tuition costs and the y-axis showing frequency of occurrence, the histogram illustrates the frequency of various tuition ranges.

The ggplot2 package, a popular package for data visualization in R, is utilized in the code. By defining the dataset and the variable to be plotted, it initially sets up the plot. The geom\_histogram() function is then used to produce the histogram, which splits the range of out-of-state tuition prices into bins and illustrates the frequency of data points falling into each bin by varying the height of the histogram bars.



Data Preprocessing

The Private column, which says whether a university is private or not, was turned into binary signs (1 for Yes and 0 for No) and then into a factor variable. This is important for logistic regression, which predicts events that can be either yes or no.



Split Train and Test Data

Randomly, the data set was split into a training set of 70% and a test set of 30%. This lets models be tested on data that hasn't been seen yet. So that it could be done again, the random number was set to 123. Also, the predictor factors were normalized by scaling the data. This is helpful for logistic regression because it makes the model less sensitive to how big or small the features are.

Confusion Matrix:

The confusion matrix, which is a table that shows how well the classification model did, is shown in the result. The rows show the real classes in the test set, while the columns show the classes that were forecast. In this case, the rows correspond to the "Private" variable, which probably shows whether an observation is from a private school (0) or not (1). The columns show the classes that were expected ("No" and "Yes").

From the confusion matrix, we can see the following:

True negatives (TN): Three of the observations are correctly labeled as "No" (they belong to schools that are not private).

False positives (FP): There are 60 observations that are wrongly labeled as "Yes" (belonging to private institutions) when they actually belong to non-private institutions.

False negatives (FN): Three findings are wrongly labeled as "No" (belonging to non-private institutions) when they actually belong to private institutions.

True positives (TP): 166 observations are correctly labeled as "Yes" (belonging to private organizations).

In conclusion, the code uses a model for classifying things to guess the names for a test set. The confusion matrix shows how well the model works by showing that it is more accurate at identifying institutions that are not private ("No") but less accurate at identifying institutions that are private ("Yes"). More analysis and evaluation metrics could be used to figure out how well and how well the model works generally.  
  


Model Building

The glm() method in R was used to fit a logistic regression model. The model tried to figure out if a college is public or private by looking at how many students were accepted from the top 10% and top 25% of their high school classes (Top10perc and Top25perc).

Model Evaluation on Test Data

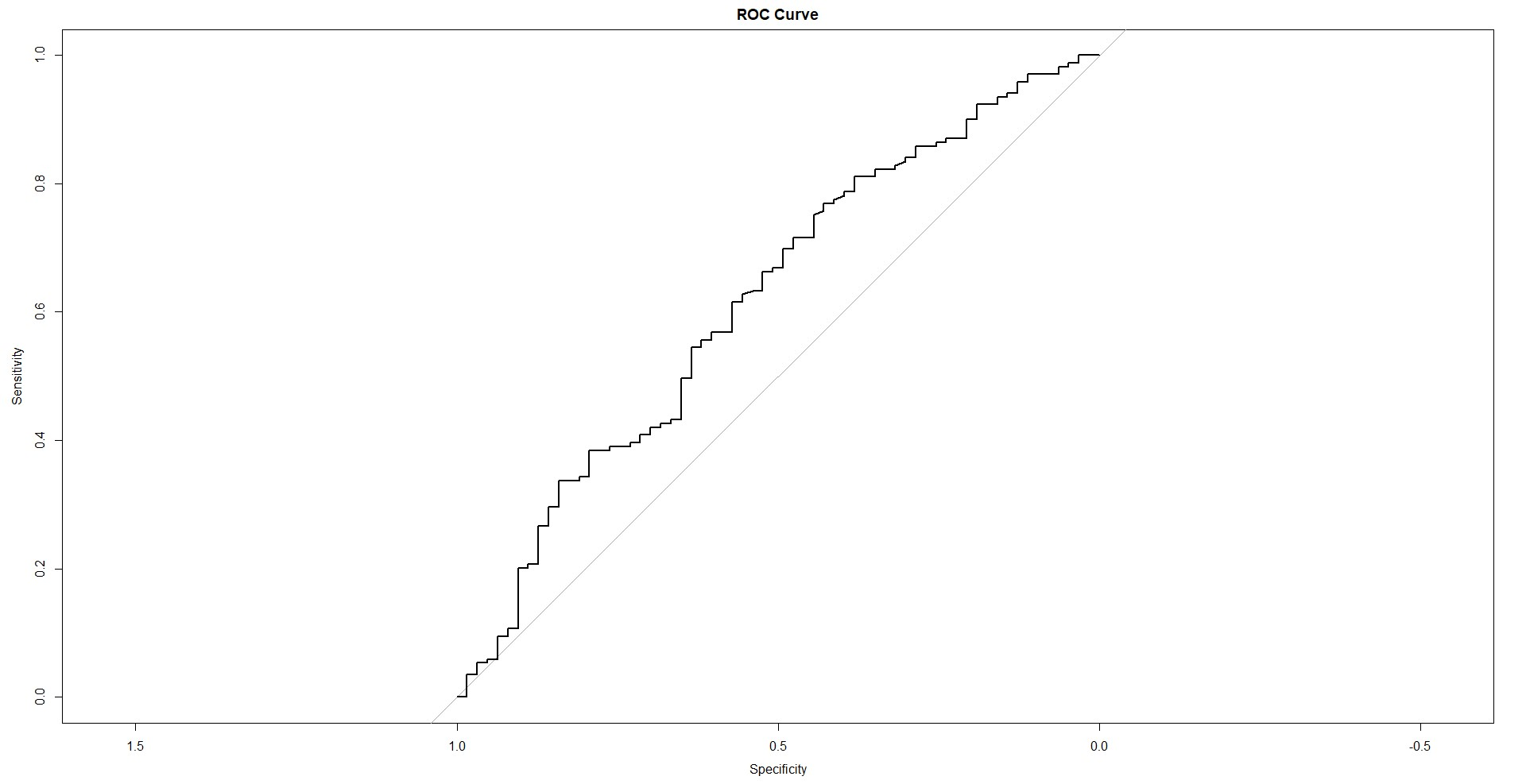
To see how well the model worked with the test data, an uncertainty matrix was made. The confusion matrix showed how many forecasts were true positives, true negatives, false positives, and false negatives.

Both false positives and false negatives can be bad, but in different ways. Depending on the situation and how much each type of mistake costs, one might be seen as more harmful than the other. For example, if the goal is to find private universities for a program where only private universities are qualified, then false negatives would be more harmful because they would mean losing out on potential candidates.

AUC and ROC Curve

The Receiver Operating Characteristic (ROC) curve is a graph that shows how the true positive rate (sensitivity) compares to the fake positive rate (1-specificity) for different threshold values. It's a good way to figure out how well a binary predictor works. The Area Under the Curve (AUC) shows how well the model can classify things. The AUC of a perfect model is 1, but the AUC of a random predictor is only 0.5.

ROC Curve Plot:



In this study, the ROC curve and the AUC for the logistic regression model were both made.

Output for Code:

# Calculate the Area Under the Curve (AUC)

auc(roc\_obj)



The given Receiver Operating Characteristic (ROC) shape has an Area Under the shape (AUC) of 0.6167. This shows that the classification model isn't very good at telling different things apart.

Conclusion

Several metrics were used to build and test the logistic regression model. The ROC curve and AUC number show how well the model classifies universities as private or not based on the percentage of accepted students from the top 10% and top 25% of their high school class. Based on the situation,