*Explores the College dataset using a generalized linear model, confusion matrix, and ROC curve to identify model accuracy*

**Assignment**

**3**

A3

ALY6015 Intermediate Analytics

Assignment 3 – GLM and Logistic Regression

**PREPERATION:**

By: John DiSessa

For: Professor Goulding

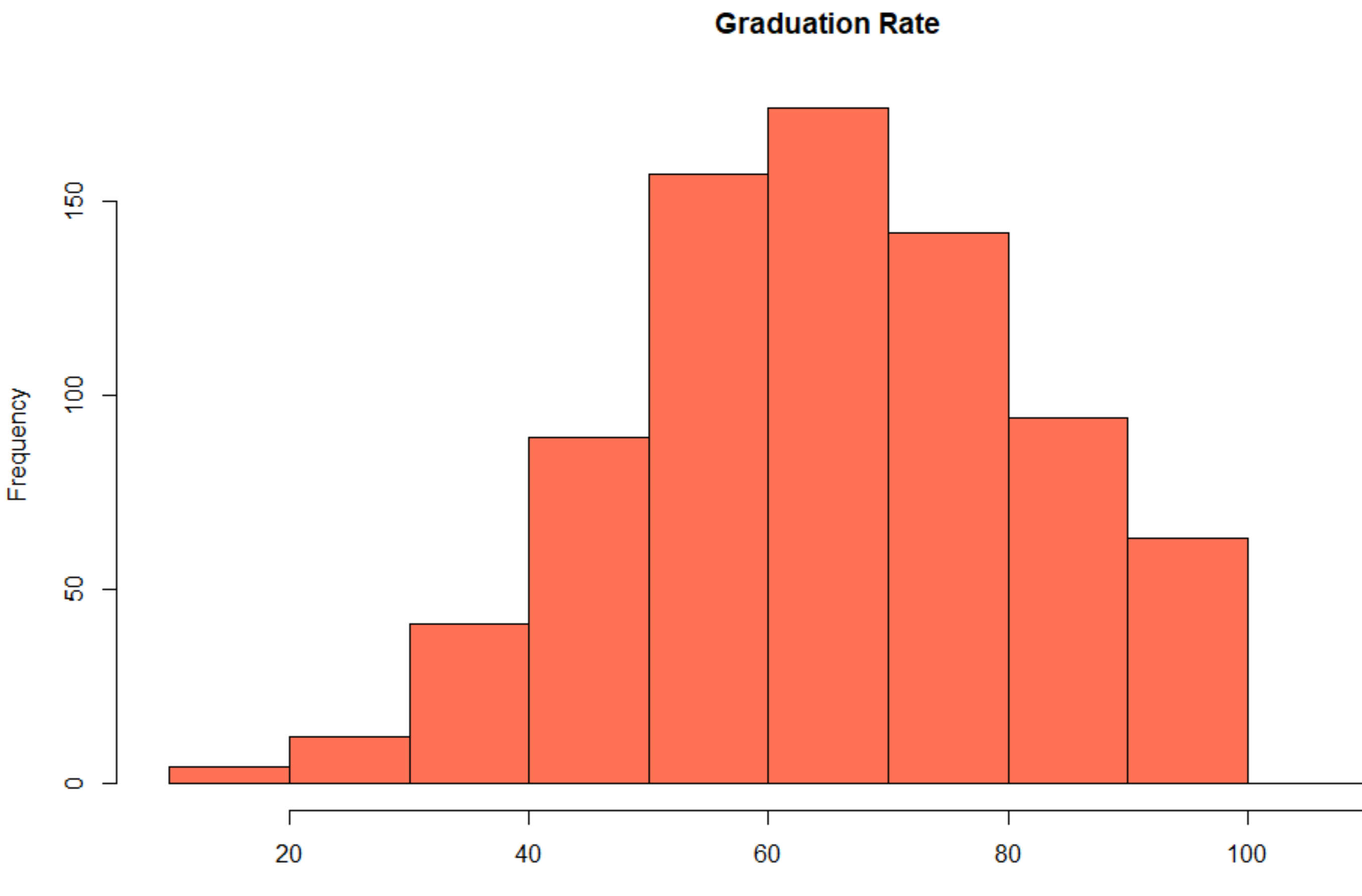
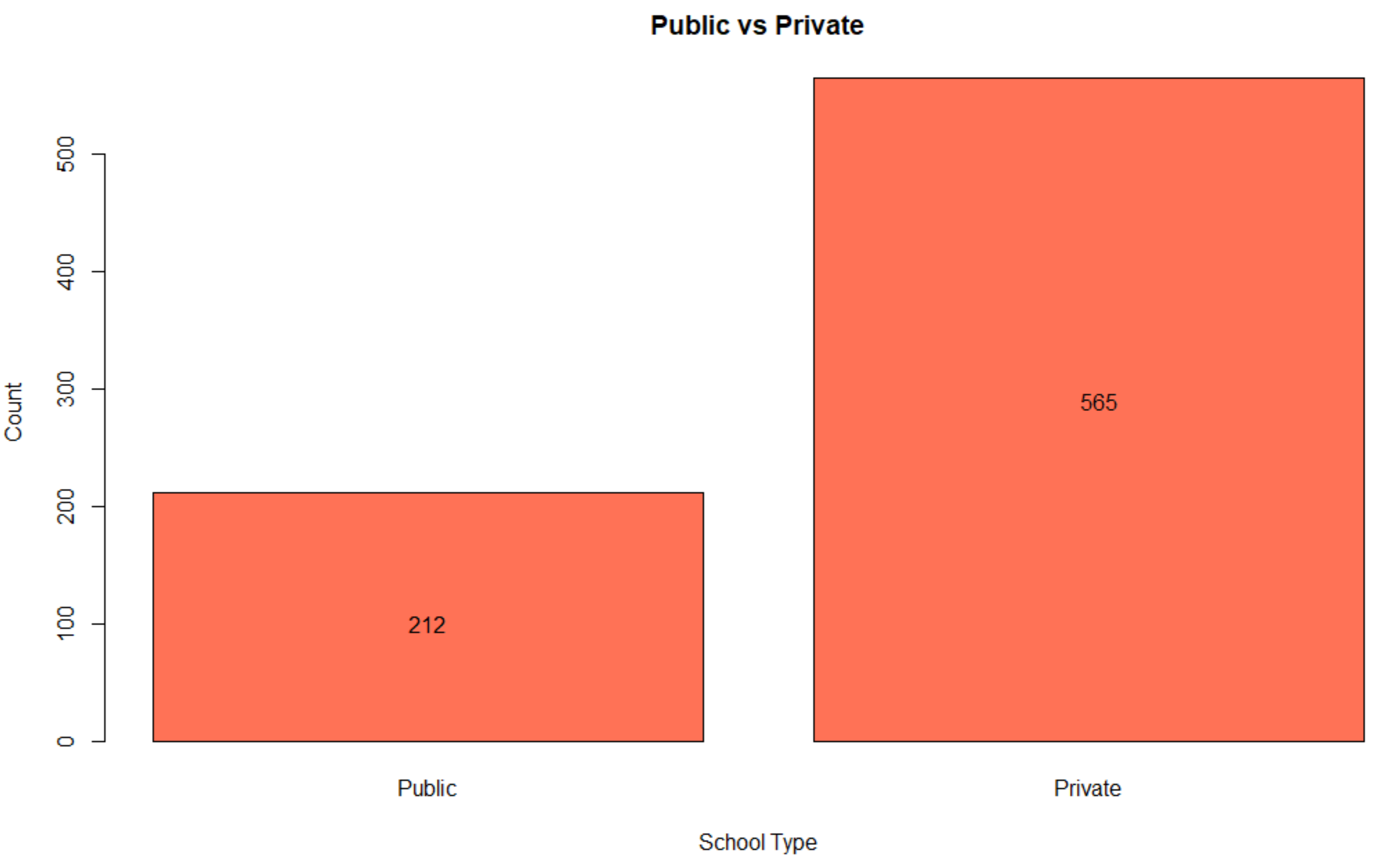
On: October 23rd, 2021

**Introduction**

The College dataset from the ISLR package contains data on 777 colleges and universities with variables such as if the school is private or public, full-time undergraduate population, school expenditure per student, and graduation rate just to name a few. The analysis below summarizes the dataset, creates a model to predict if a school is private or public, and analyzes the accuracy of that model by plotting a confusion matrix and ROC curve.

**Exploratory Data Analysis**

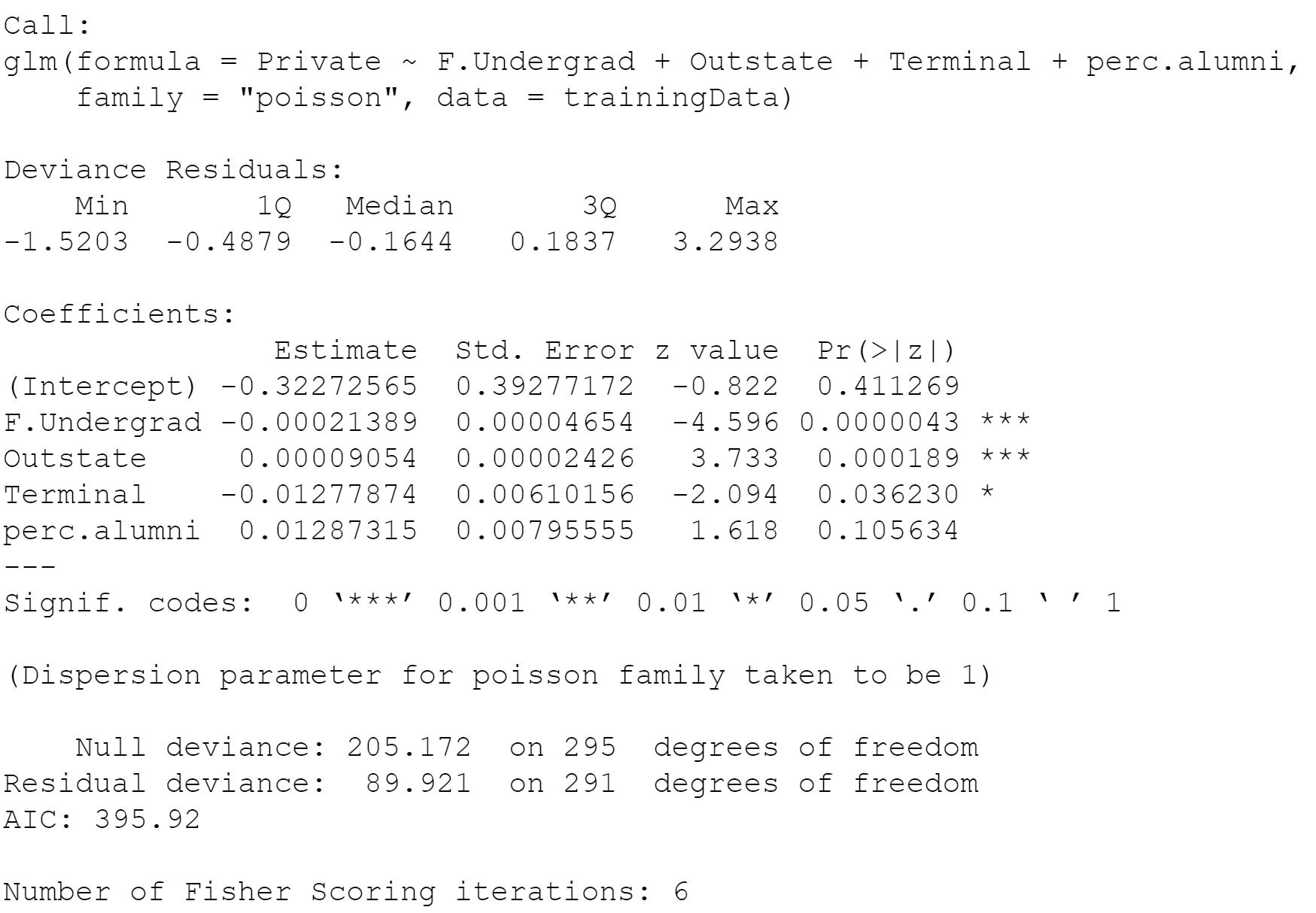
One of the main variables we explored in this assignment was whether a school was public or private. Based on the bar plot below we can see that 27% of our dataset is made up of public universities. Our histogram looks at the graduation rates of our colleges and shows it is basically normally distributed. The average graduation rate is 65%. We also looked at the average number of full-time undergraduates and average expenditure per student. From our preliminary analysis, the average college graduates 65% of their students, has about 3,700 full-time undergraduates, and spends $9,660 per student.



**Generalized Linear Model**

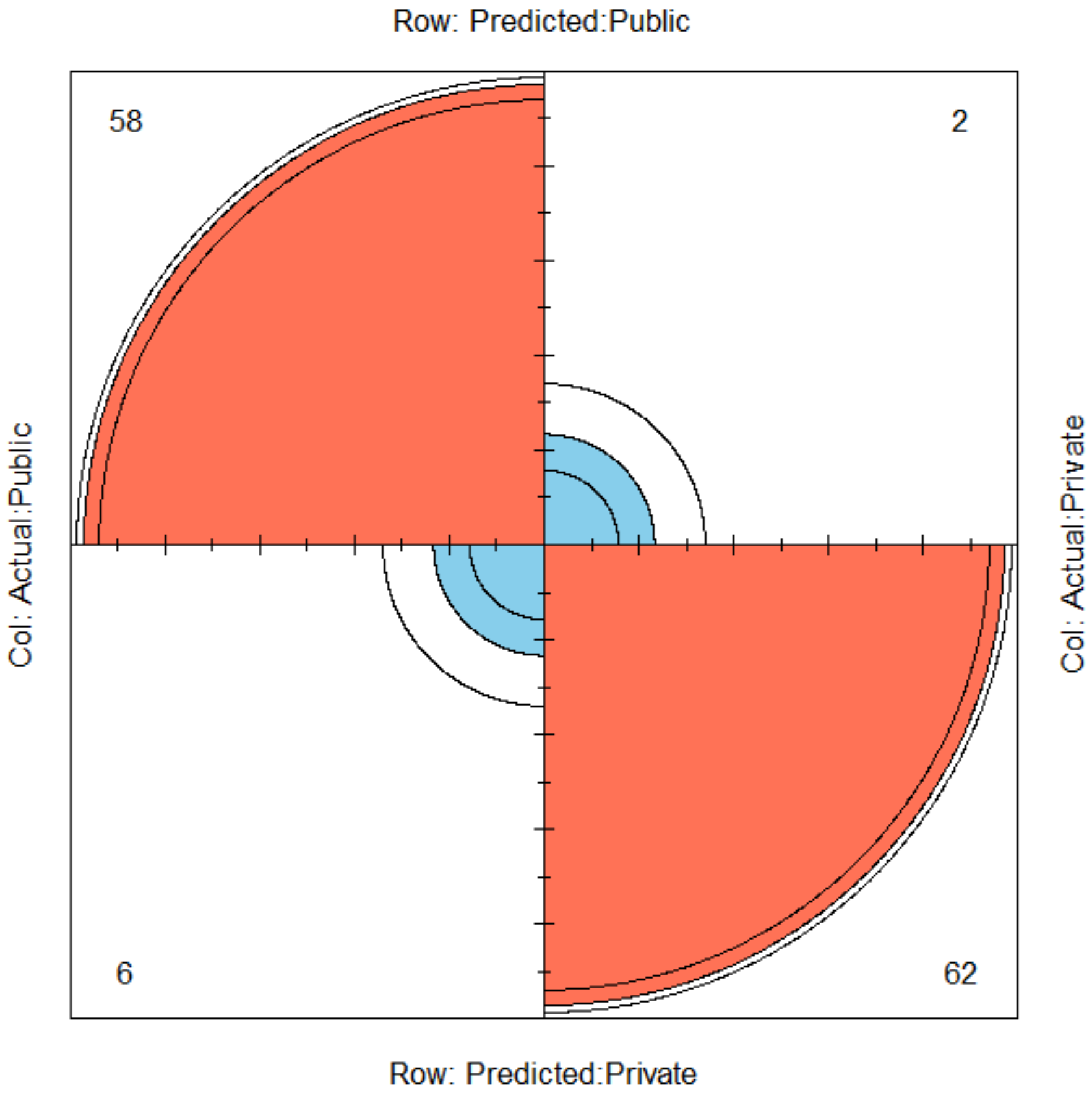
Next, we decided to build a Generalized Linear Model in order to predict if a school is private or public given certain predictors. First, we split the data into training and testing datasets. We built the model using training data and then evaluated it’s accuracy using the test data. Based on the training data, we see how the model reacts compared to the testing data to see how accurate it is. If we built a model based on all of the data, the model would be “overfit”. It would be entirely based on fitting this one particular dataset 100%, so much so, that it really would not be a good model to predict colleges outside of this dataset.

We then ran a model with all variables in the training dataset to find out which predictors were most significant. We picked the top 4 variables and built a GLM with them, as seen below. This second model is certainly better than our first model with all variables since the first model had a higher AIC score (prediction error estimation) and many insignificant variables. Our new model predicts if a school is private or public based on the number of full-time undergraduates, out-of-state tuition, percent of faculty with terminal degrees, and percent of alumni who donate.



**Model Accuracy**

Now that we have a pretty good model, we need to evaluate how accurate it actually is for correctly predicting if a school is private or public. We created the following confusion matrix to easily represent the accuracy, precision, recall, and specificity of our model.



Our testing dataset contained 64 public schools and 64 private schools. The top left shows 58 true positive values (the model predicts public and the schools were actually public). The bottom right shows the 62 true negative values (the model predicts private and the schools were actually private). We have very small Type I (false positive) and Type II (false negative) errors as indicated by the bottom left and type right values. The model had 6 values that it incorrectly predicted as public when really, they were private. The model had 2 values that it incorrectly predicted as private when really, they were public. In our case, Type I and Type II errors are equally damaging since, in both cases, the schools are simply just misclassified with no serious repercussions.

However, there are many other instances where one error is more damaging than the other. In cancer screenings, Type II errors are worse since the test would show cancer negative when really the patient was cancer positive. That means the patient wouldn’t get any treatment. In the criminal justice system, Type I errors have severe consequences such as ruling someone committed a crime when in fact they did not (convicting an innocent person directly contradicts “innocent until proven guilty).

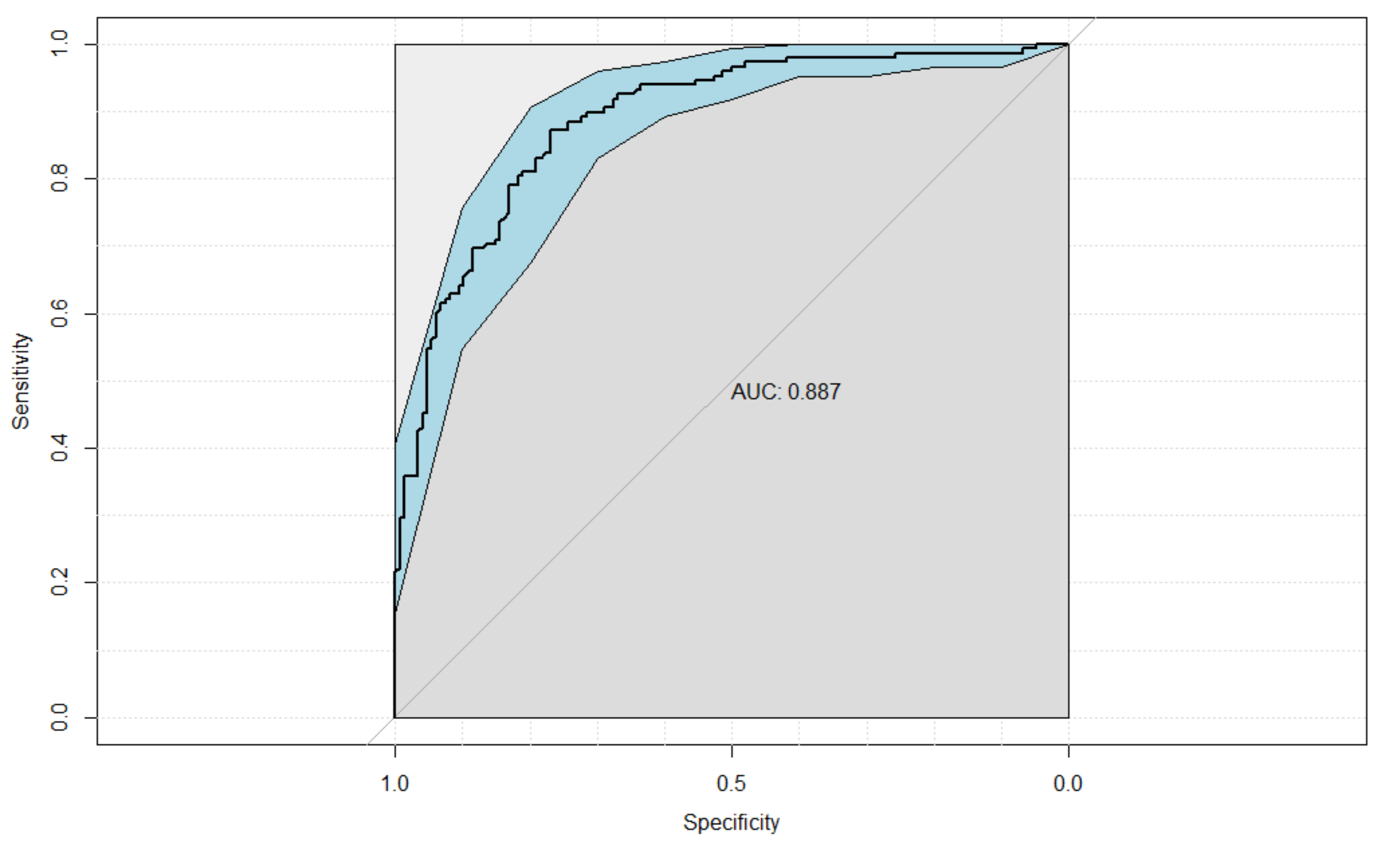
In order to better interpret the confusion matrix, we calculated

* Accuracy (correct public predictions)
* Recall (when the result is public, how often did the model predict public)
* Precision (when the prediction is public, how often was the result actually public)
* Specificity (correct private predictions)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision | Specificity |
| Model Accuracy | 93.75% | 96.67% | 90.63% | 91.18% |

**Receiver Operating Characteristic Curve**

The following ROC curve shows us how likely our model is to distinguish between public and private colleges. Even though we already calculated our model’s accuracy, calculating the area under the ROC curve allows us to determine the model’s ability to distinguish between public and private colleges at different thresholds. Overall, our model has an 88.7% chance of correctly distinguishing between these values. As you can see from the curve, however there are certain cutoff points where the curve jumps up in sensitivity (true positive rate) depending on the false positive rate. Accuracy is still a good measure of how successful our model is at predicting public or private colleges, but the ROC curve gives a more complete picture without assuming just one true or false positive rate.



**Conclusion**

After analyzing the College dataset, we created a generalized linear model that predicted if a college was public or private based on the number of full-time undergraduates, out-of-state tuition, percent of faculty with terminal degrees, and percent of alumni who donate. The confusion matrix indicated high accuracy, recall, precision, and specificity values as well as low type I and type II errors. We avoided overfitting the model at the expense of potentially not find the most accurate model possible. However, that would be a consideration for future analysis. Lastly, we created the ROC curve and found the area under the curve to give us a complete comprehensive view of this model’s ability to predict if a college is public or private.

**Citations**

“Advantages of AUC vs Standard Accuracy.” *Data Science Stack Exchange*, 1 Oct. 1962, https://datascience.stackexchange.com/questions/806/advantages-of-auc-vs-standard-accuracy.

Ajay SinghAjay Singh 26111 gold badge33 silver badges99 bronze badges, et al. “Calculate Accuracy and Precision of Confusion Matrix in R.” *Stack Overflow*, 1 Feb. 1961, https://stackoverflow.com/questions/13548092/calculate-accuracy-and-precision-of-confusion-matrix-in-r.

Carl O'BeirneCarl O'Beirne 18933 silver badges1515 bronze badges, and StupidWolfStupidWolf 36.9k1717 gold badges2424 silver badges5555 bronze badges. “How to Interpret Confusion Matrix in R.” *Stack Overflow*, 1 June 1968, https://stackoverflow.com/questions/60858241/how-to-interpret-confusion-matrix-in-r.

*Choosing Logisitic Regression’s Cutoff ... - Github Pages*. http://ethen8181.github.io/machine-learning/unbalanced/unbalanced.html.

ethen8181. “Machine-Learning/Unbalanced.r at Master · ETHEN8181/Machine-Learning.” *GitHub*, https://github.com/ethen8181/machine-learning/blob/master/unbalanced/unbalanced\_code/unbalanced.R.

“Fourfoldplot: A Prettier Confusion Matrix in Base R.” *Fourfoldplot: A Prettier Confusion Matrix in Base R | Swimming the Data Lake*, http://jvera.rbind.io/post/2017/07/17/fourfoldplot-a-prettier-confusion-matrix-in-base-r/.

Holtz, Yan. “Barchart.” *The R Graph Gallery*, https://www.r-graph-gallery.com/barplot.html.

Kassambara, and Sfd. “Best Subsets Regression Essentials in R.” *STHDA*, 11 Mar. 2018, http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/.

Khintibidze, Lasha. “Visualize Confusion Matrix Using Caret Package in R.” *Delft Stack*, 16 May 2021, https://www.delftstack.com/howto/r/visualize-confusion-matrix-in-r/.

Markham, Kevin. “Simple Guide to Confusion Matrix Terminology.” *Data School*, Data School, 3 Feb. 2020, https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.

Max Kuhn. Contributions from Jed Wing, Steve Weston. “CreateDataPartition: Data Splitting Functions in Caret: Classification and Regression Training.” *CreateDataPartition: Data Splitting Functions in Caret: Classification and Regression Training*, 2 May 2019, https://rdrr.io/rforge/caret/man/createDataPartition.html.

Narkhede, Sarang. “Understanding AUC - Roc Curve.” *Medium*, Towards Data Science, 15 June 2021, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

robk@statmethods.net, Robert Kabacoff -. “Generalized Linear Models.” *Quick-R: Generalized Linear Models*, https://www.statmethods.net/advstats/glm.html.

“Set Seed in R with Set.seed() Function ▷ [with Examples].” *R CODER*, 22 July 2020, https://r-coder.com/set-seed-r/#:~:text=4%20Unset%20seed-,What%20is%20to%20set%20seed%20in%20R%3F,)%20(0%2C1).

Sharma, Prateek. “Decoding the Confusion Matrix.” *Medium*, Towards Data Science, 11 Jan. 2020, https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb.

“Interpretation of R's Output for Binomial Regression.” *Cross Validated*, 1 May 1962, https://stats.stackexchange.com/questions/86351/interpretation-of-rs-output-for-binomial-regression.

Zach. “How to Create a Confusion Matrix in R (Step-by-Step).” *Statology*, 1 Apr. 2021, https://www.statology.org/confusion-matrix-in-r/.