ALY\_6015 – Intermediate Analytics  
Module 3

Logo, company name

Description automatically generated

Abhishek Mehra

**Introduction**

The College data is taken from the website <https://rdrr.io/cran/ISLR/man/College.html>. It contains the information regarding private and public schools in the USA. The colleges have attributes such as number of applications college received , enrollment of students, number of fulltime and part time undergraduate students and many more such attributes. The shape of the dataset is 777 rows and 19 columns.

**Exploratory data analysis**

The mean number of Apps- application, Accept – acceptance and Enroll- Enrollment of students in all the school are described as below

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Apps** | **Accept** | **Enroll** |
| **count** | 777 | 777 | 777 |
| **mean** | 3001.638 | 2018.804 | 779.973 |
| **std** | 3870.201 | 2451.114 | 929.1762 |
| **min** | 81 | 72 | 35 |
| **25%** | 776 | 604 | 242 |
| **50%** | 1558 | 1110 | 434 |
| **75%** | 3624 | 2424 | 902 |
| **max** | 48094 | 26330 | 6392 |

Figure 1

The figure below shows the distribution of private and public schools in USA. There are 565 private schools and 212 public schools in the given dataset.

Chart, bar chart

Description automatically generated

Figure 2

The figure 3 shows top 4 and bottom 4 colleges with the maximum and minimum number of applications, acceptance and enrollment.

|  |  |  |  |
| --- | --- | --- | --- |
| **college** | **Apps** | **Accept** | **Enroll** |
| **Rutgers at New Brunswick** | 48094 | 26330 | 4520 |
| **Purdue University at West Lafayette** | 21804 | 18744 | 5874 |
| **Boston University** | 20192 | 13007 | 3810 |
| **University of California at Berkeley** | 19873 | 8252 | 3215 |
| **Saint Mary-of-the-Woods College** | 150 | 130 | 88 |
| **College of St. Joseph** | 141 | 118 | 55 |
| **Capitol College** | 100 | 90 | 35 |
| **Christendom College** | 81 | 72 | 51 |

Figure 3

I utilized three features namely ‘Expend', 'Apps',' Grad.Rate' for the training of Logistic Regression model. The dataset was divided into 80% training data and 20% test set. After training of the logistic regression model, I utilized test set for evaluation. The figure below shows confusion matrix results. The zero represents the college as not private and one represents college as private. A confusion matrix is a table that is used to define the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It is also known as an error matrix.

* The number of false positives, false negatives, true positives, and true negatives are listed in a table with two rows and two columns called the confusion matrix. The columns of the matrix match the actual class, and the rows match the anticipated class.
* True Positives (TP) are observations that are both expected and found to be positive.
* The observations that are genuinely negative despite being projected to be negative are known as true negatives (TN).
* Observations that are projected to be positive but turn out to be negative are known as false positives (FP).
* Observations that are projected to be negative but are really positive are referred to as false negatives (FN).

Accuracy, precision, recall, and F1-score are a few classification metrics that may be calculated using a confusion matrix. Further, the observations that are predicted as negative but are actually positive.

Different classification measures, including accuracy, precision, recall, and F1-score, may be calculated using a confusion matrix. It is helpful for figuring out what kind of error the classifier is producing as well.

The more damaging interpretation would depend on the problem statement. False negatives would mean the colleges which are private are classified as public. False positives would mean that the colleges that are public termed as private. In general both are mis representation of the actual classes.

Chart, treemap chart

Description automatically generated

Figure 4

Receiver Operating Characteristic (ROC) is a graphical display that shows how a binary classifier system's diagnostic capacity changes with its discriminating threshold. Plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels yields the ROC curve. The total performance of the classifier is measured by the area under the ROC curve (AUC), which has a value between 0 and 1, with a greater value signifying better performance. A perfect classifier has an AUC of 1, whereas a random classifier has an AUC of 0.5. The ROC curve is a valuable tool for assessing a classifier's performance and contrasting several classifiers.

The figure 5 below shows the ROC area under the curve. The area under the curve stands at 0.849 . ROC AUC describes how efficiently our model was able to distinguish between true negative and true positives. A score near to is considered ideal. Our model was able to distinguish with 84% efficiency.

Chart, line chart

Description automatically generated

Figure 5

**Conclusion**

Our model with three predictors produced a ROC AUC score of 0.849 which shows Expenditure, Application, and Graduation Rate are good predictors in determining if the colleges are private or not. Further we can experiment with more features to determine the label class, keeping in mind overfitting

**References**

* *Sklearn.linear\_model.linearregression*. scikit. (n.d.). Retrieved January 28, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html
* *Sklearn.metrics.confusion\_matrix*. scikit. (n.d.). Retrieved January 28, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion\_matrix.html
* *Sklearn.model\_selection.train\_test\_split*. scikit. (n.d.). Retrieved January 28, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html

**Appendix**

**import numpy as np**

**import pandas as pd**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.linear\_model import LogisticRegression**

**from sklearn.metrics import confusion\_matrix,classification\_report,roc\_auc\_score, ConfusionMatrixDisplay, roc\_curve**

**from matplotlib import pyplot as plt**

**import seaborn as sns**

**import openpyxl**

**#### EDA**

**college\_dataset = pd.read\_csv('../us\_college\_data/College.csv') #reading the college file**

**college\_dataset.rename(columns={'Unnamed: 0':'college'},inplace=True) #renaming unmaned column as college**

**college\_dataset.head() #firt look at the dataframe**

**college\_dataset.shape #looking at the shape of the dataset**

**college\_dataset.describe()**

**# distribution of private and public schools**

**college\_dataset['Private'].describe()**

**plt.figure(figsize=(5,5))**

**sns.countplot(x=college\_dataset['Private'])**

**plt.savefig('countplot\_private')**

**a = college\_dataset[['Apps','Accept','Enroll']].describe()**

**college\_dataset[['Apps','Accept','Enroll']].describe().T**

**table = a.style.set\_table\_attributes("style='display:inline'").set\_caption('Summary Statistics')**

**table.to\_excel("save.xlsx")**

**# a table to put in report to show summary characteristics.**

**college\_wise\_accept = college\_dataset.groupby(['college'])[['Apps','Accept','Enroll',]].mean().sort\_values(by='Apps',ascending=False)**

**college\_wise\_accept = college\_wise\_accept.head(4).append(college\_wise\_accept.tail(4))**

**table = college\_wise\_accept.style.set\_table\_attributes("style='display:inline'").set\_caption('Summary Statistics')**

**table.to\_excel("college\_wise\_accept.xlsx")**

**#### Data preparation**

**#### divinding into train and test set**

**X = college\_dataset.drop(['Private','college'],axis =1) #dropping dependent variable and name of college**

**y = college\_dataset['Private'] #setting the dependent variable**

**y = y.map({'Yes':1, 'No':0}) #converting to zeros and ones for the machine to understand**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y) #80:20 split between train and test**

**X\_train.shape, X\_test.shape, y\_train.shape, y\_test.shape # checking the shape of the train and test**

**lr = LogisticRegression() #intialising object for logistic regression**

**lr.fit(X\_train[['Expend','Apps','Grad.Rate']],y\_train) # fitting**

**predictions = lr.predict(X\_test[['Expend','Apps','Grad.Rate',]])**

**confusion\_matrix(y\_test,predictions)**

**ConfusionMatrixDisplay(confusion\_matrix(y\_test,predictions)).plot()**

**plt.savefig("confusion matrix")**

**plt.show()**

**print(classification\_report(y\_test,predictions))**

**roc\_auc\_score(y\_test,predictions)**

**predictions\_probability = lr.predict\_proba(X\_test[['Expend','Apps','Grad.Rate']])[::,1]**

**fpr, tpr, \_ = roc\_curve(y\_test, predictions\_probability)**

**auc\_score = roc\_auc\_score(y\_test,predictions)**

**#create ROC curve**

**plt.plot(fpr,tpr, label="AUC"+str(auc\_score))**

**plt.ylabel('True Positive Rate')**

**plt.xlabel('False Positive Rate')**

**plt.legend(loc=4)**

**plt.savefig('roc\_curve')**

**plt.show()**

**Our model with three features was able to produce and AUC of 0.84 which is an above average score**