**ALY6015 CRN 80403**

**Intermediate Analytics**

**NORTHEASTERN UNIVERSITY**

**SANA YASIN**

Date :05 May 2023

**MODULE 2 -GLM and LOGISTIC REGRESSION**

**Professor: Richard He**

Logo

Description automatically generated

**INTRODUCTION :-**

**Generalised Linear Models :-**

**Generalised liner modeling** technique is the generalisation of linear regression predictive model to various non linear distribution.For GLM we do not need our target variable to be normally distributed. The target variable can be binary or even counts.

GLM is advanced modelling technique and is applied when the relationship between predictor and target variable is non linear .

The various types of GLM are :-

1. **Logistic Regression**-used when target variable is 0 or 1
2. **Possion Regresion**-used when target variable is counts or frequency
3. **Linear Regression**-for normally distributed target variable

There are 3 componenets of GLM

1. **Random component** :- the target variable distribution
2. **System component** :- the linear combination of predictor variables
3. **Link Functions** :- function that specifies the link between random and systematic components

**LOGISTIC REGRESSION** :-

Logistic Regression is a type of genaralised linear model in which the target variable is binary or categorical variable expressed in form of either 0 or 1 . It can be treated as classification model.

In logistic regression we predict the probabilities of occurance of target variable based on given set of predictor variables . The relationship between predictor variables and probabilities is non linear and ranges between 0 and 1 .

Diagram

Description automatically generated

Hence the link function used in logistics regression is logit or probit function which gives the outcome in form of 0 or 1 .



**OBJECTIVE:-**

The objective of this project is use college dataset having list of public and private US colleges from the 1995 with various measures like no of appication , enrollment and acceptance rate , tution fees and no of graduates etc .

1. We need to build a logistic regression model to predict whether a college with given set of measures is public/private .The target variable is binary with 2 level that is “yes /No”.

2.After model Creation we need to test the model with test data and check the model accuracy .

**DATASET:-**

Graphical user interface, text, application

Description automatically generated

The null value in dataset is 0.

The dataset has 777 rows and 18 columns.

**EXPLORATORY DATA ANALYSIS: -**

Table showing distribution of Target variable.

| **Private** | **Freq** |
| --- | --- |
| No | 212 |
| Yes | 565 |

Table showing Descriptive statistics for Column **Apps**



**Descriptive** statistics for Column **Accept**

Graphical user interface, text, application

Description automatically generated

Descriptive statistics for Column **Enroll**

Graphical user interface, text, application, Word

Description automatically generated

Descriptive statistics for Column **Top10perc**

Graphical user interface, text

Description automatically generated

Descriptive statistics for Column **Top25perc**

Graphical user interface, text

Description automatically generated

Descriptive statistics for Column **F.Undergrad**

Graphical user interface, text

Description automatically generated

Descriptive statistics for Column **P.Undergrad**

Text

Description automatically generated

Descriptive statistics for Column **Outstate**

Text

Description automatically generated

Descriptive statistics for Column **Room.Board**

Text

Description automatically generated

Descriptive statistics for Column **Books**

Graphical user interface, text, application

Description automatically generated

Descriptive statistics for Column **Personal**

Text

Description automatically generated

Descriptive statistics for Column **PhD**

Text

Description automatically generated with medium confidence

Descriptive statistics for Column **Terminal**

Text

Description automatically generated

Descriptive statistics for **S.F.Ratio**

Graphical user interface, text

Description automatically generated

Descriptive statistics for **perc.alumni**

Graphical user interface, text

Description automatically generated

Descriptive statistics for **Expend.**

Graphical user interface, text

Description automatically generated

Descriptive statistics for **Grad. Rate**

Graphical user interface, text

Description automatically generated

Descriptive statistics for **Acceptance Rate**

Graphical user interface, text

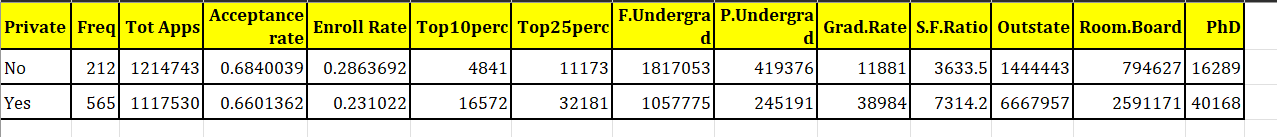
Description automatically generated

Descriptive statistics for **Enrollment Rate**

Graphical user interface, text

Description automatically generated

Table showing sum of variables for each type of university.



**Plots and Graphs: -**

A picture containing timeline

Description automatically generated

A picture containing chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

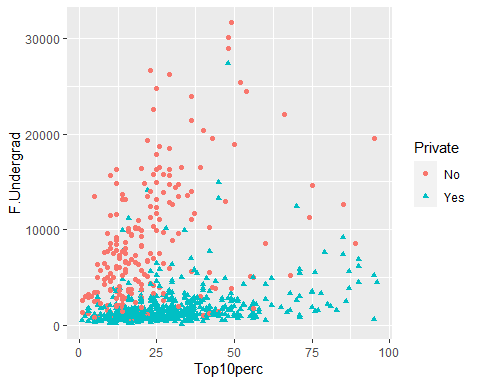
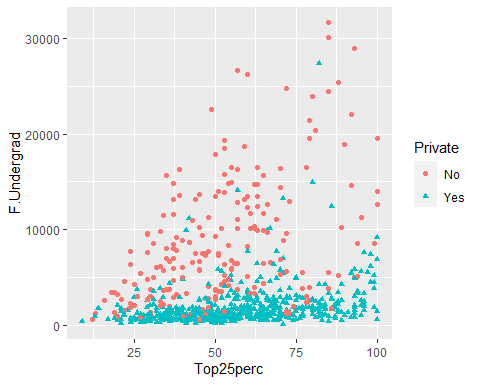
Box and whisker chart

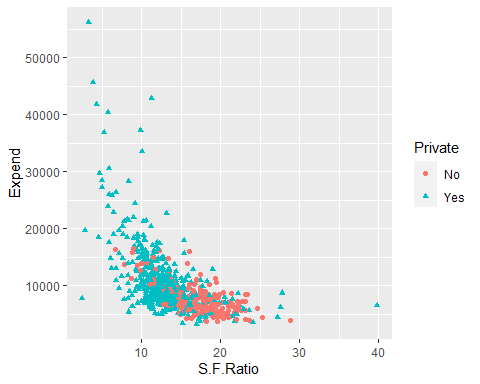
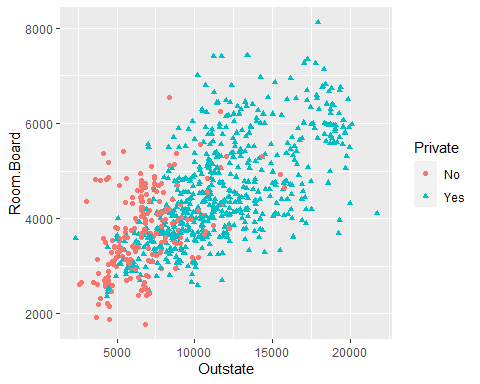
Description automatically generated with low confidence

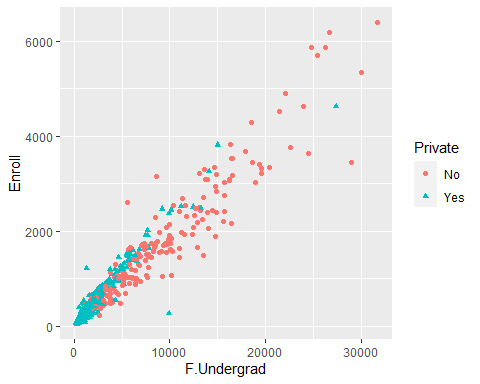
Chart, box and whisker chart

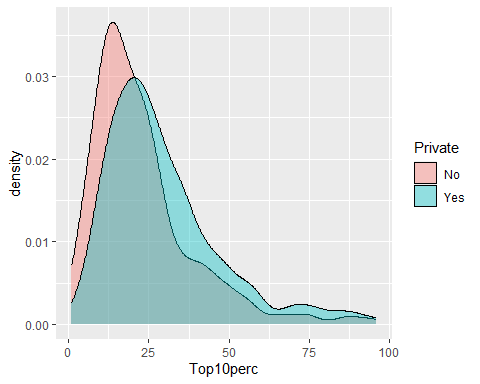
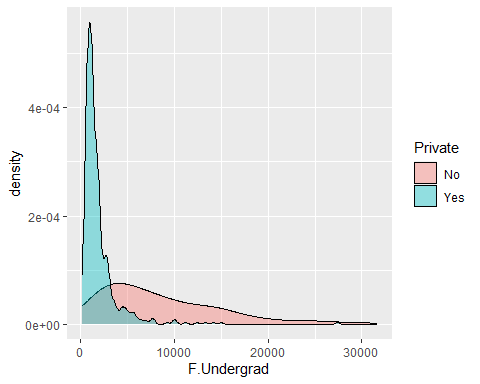
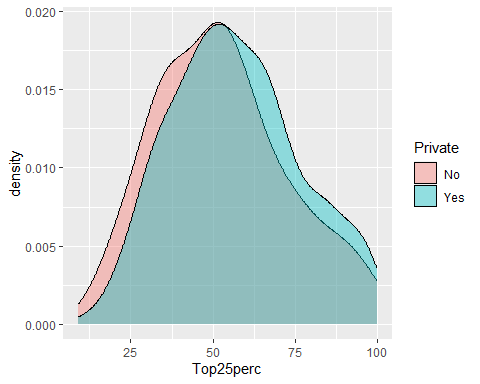
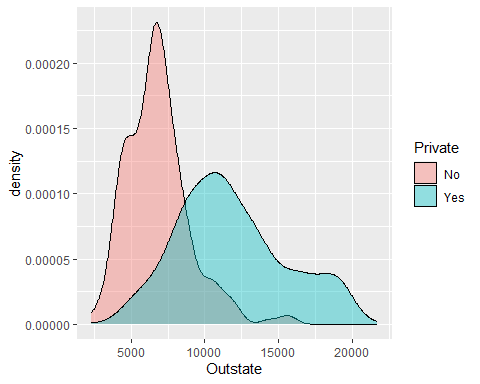
Description automatically generated

Scatterplots showing graphs of variation of various measures with respect to type of universities: -



**Observation** :-

1. From the above boxplots , we can say that the data is skewed and there is considerable number of outliers in the dataset.
2. The huge number of varation in the number of fulltime undergraduates in public universities and private universities. The number is greater in public universities .
3. Most of the full time undergraduates from top 10% of H.S class prefers public universities that private
4. Most of the full time undergraduates from top 25% of H.S class prefers public universities that private
5. The acceptance rate and Enrollment rate is higher in public universities.
6. The out-of-state tuition is very high in private universities than public universities.
7. The full time undergraduate student prefer public universities than private as the out-of-state tuition is less in public than in private.

**Correlation matrix: -**

Chart, scatter chart

Description automatically generated

From the correlation matrix, we can say that our probability that the university in public/private largely depends on the number of fulltime undergraduates and out-of-station tuition fees.

**Splitting the Dataset into Train and Test: -**

Graphical user interface, text

Description automatically generated

The train dataset has 545 rows and 20 columns.

The train dataset has 232 rows and 20 columns.

**Building Logistic Regression Model: -**

We perform logistic regression model by taking all the independent variables and checking for results.

Table

Description automatically generated

From the above model summary, we can say the significant variables are `F.undergrad` and `Outstate` which means lower the number of full time undergraduate students increases the chances the university being labelled as Private .

Similarly, Higher Education increases the cost of out-of-state tuition fees, the chances of university being labelled as private increases.

Hence we will rebuilt the model with only significant variables and interpret the results:-

Text

Description automatically generated

In this model we can see that AIC score has increased from the previous model.

**MODEL METRICS AND PERFORMANCE: -**

Prediction the Train dataset: -

We will first test our model against the train dataset to check the model accuracy.

Text

Description automatically generated

**Confusion matrix: -**

Confusion matrix is the measure of how well our model is performing the predicting the target variable on give set of data. It is matrix of predicted values and actual values

There are 4 types of possible combinations which we derive from confusion matrix

Table

Description automatically generated

1. **True Positive (TP)**: - The predicted value is true.
2. **False Positive (FP)**: - The predicted value is false but in actuality it is true. This is type 1 error.
3. **False Negative (FN)**: - The predicted value is false but in actuality it is true. This is type 2 error.
4. **True Negative (TN): -**The predicted value is false, but it is actually false.

Based on this combination of values we will derive 4 model metrics: -

**Accuracy**: - measure of correct prediction

**(TP+TN)/Total**

**Sensitivity**: - It is the measure the determines that from all the positive classes, how many are predicted correctly.

**TP/(TP+FN)**

**Specificity: -** It is the measure the determines that from all the Negative classes, how many are actually predicted correctly.

**TN/(FP+TN)**

**Precision: -** It is the measure the determines that from all the cases we have predicted as positive, how many are actual positive correctly.

**TP/(TP+FP)**

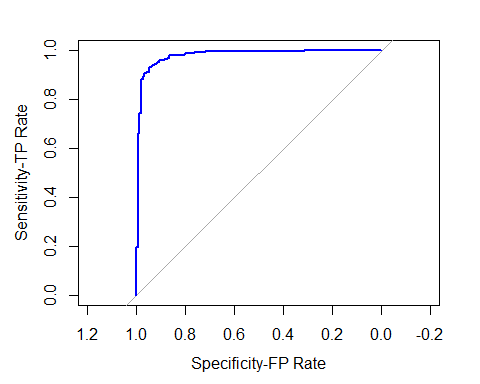
We will compute these measures using the train and test dataset.

**Train Dataset**

Graphical user interface, text, application

Description automatically generated

**ROC curve:** -

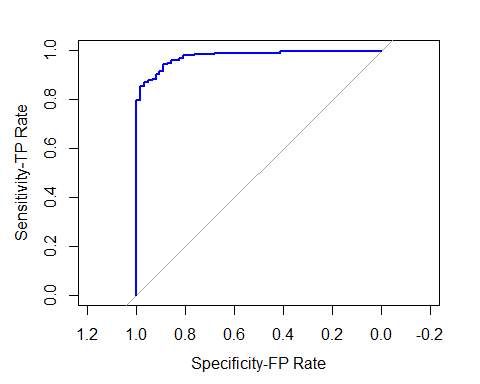


Area under the curve: **0.9777**

Confusion Matrix for Test dataset: -

Graphical user interface, text, application

Description automatically generated



**Area under the curve: 0.9736**

**CONCLUSION:**

The model accuracy with train and test data is 94.13% and 92.67% respectively which means even the model is quiet is accurate in predicting the target variable.

The AUC model with both train and test data is closer to 1 so our model has performed well on both the train and test datasets.

**REFERENCES: -**

**(Alice , atmathew August 17, 2015, Narkhede May 9, 2018**

**, Kida Sep 23, 2019**

**)**

Alice, M. "How to perform a Logistic Regression in R." <https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/> September 13, 2015

atmathew ( August 17, 2015). "Evaluating Logistic Regression Models."

Kida, Y. (Sep 23, 2019

). "Generalized linear models." from <https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab>.

Narkhede, S. (May 9, 2018

). "Understanding Confusion Matrix." from <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>.

**APPENDIX: -**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text, chat or text message

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**