UPGRAD LEAD SCORE CASE STUDY

GROUP MEMBERS

BIBIN JACOB

DIVYA

SUMAN

PROBLEM STATEMENT

- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
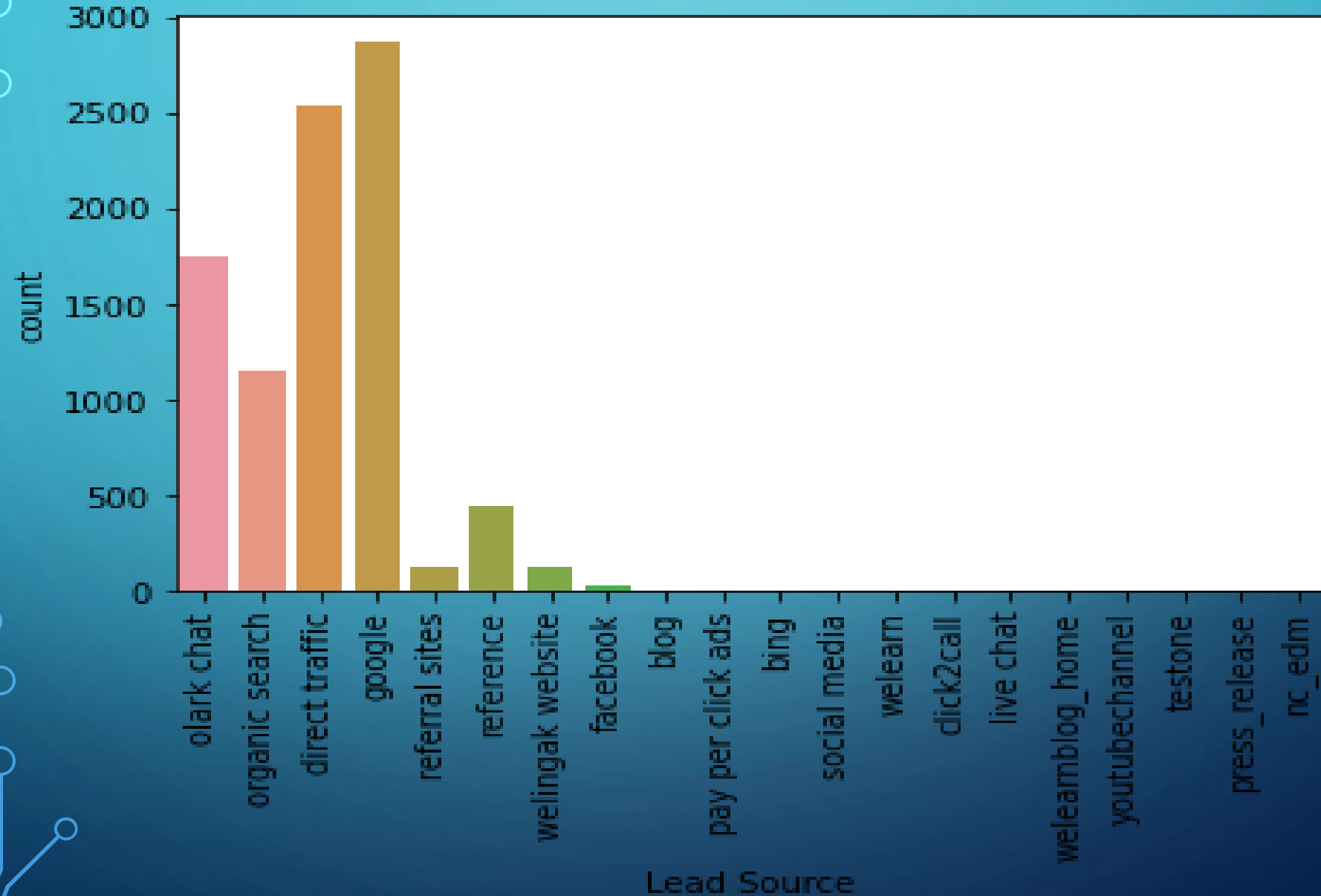
STEPS FOLLOWED

- Reading Data

- Cleaning Data

- EDA

- Creating Dummy

- Splitting data into train and test set

- Building Model

- Making Predictions

- Model Evaluation

- ROC Curve

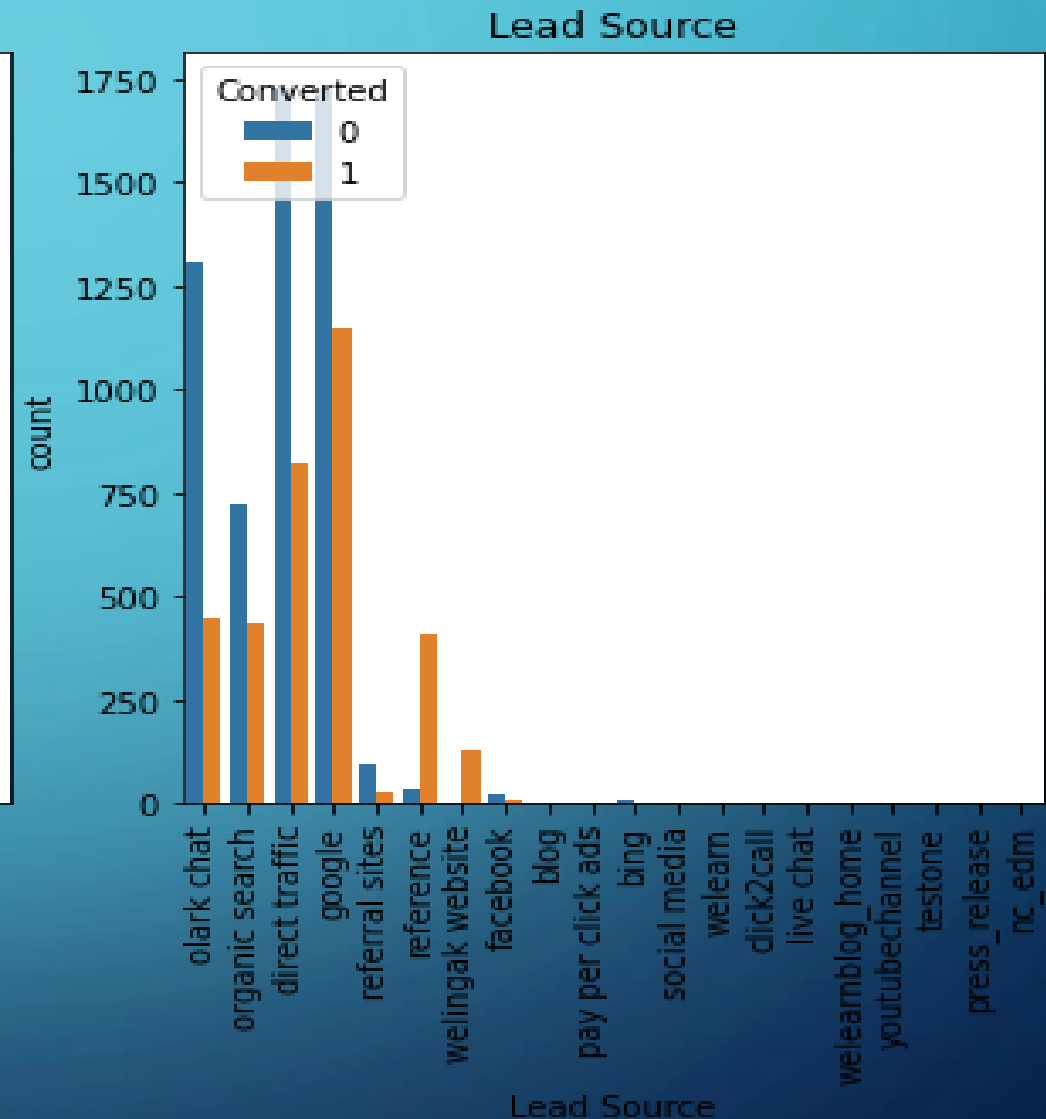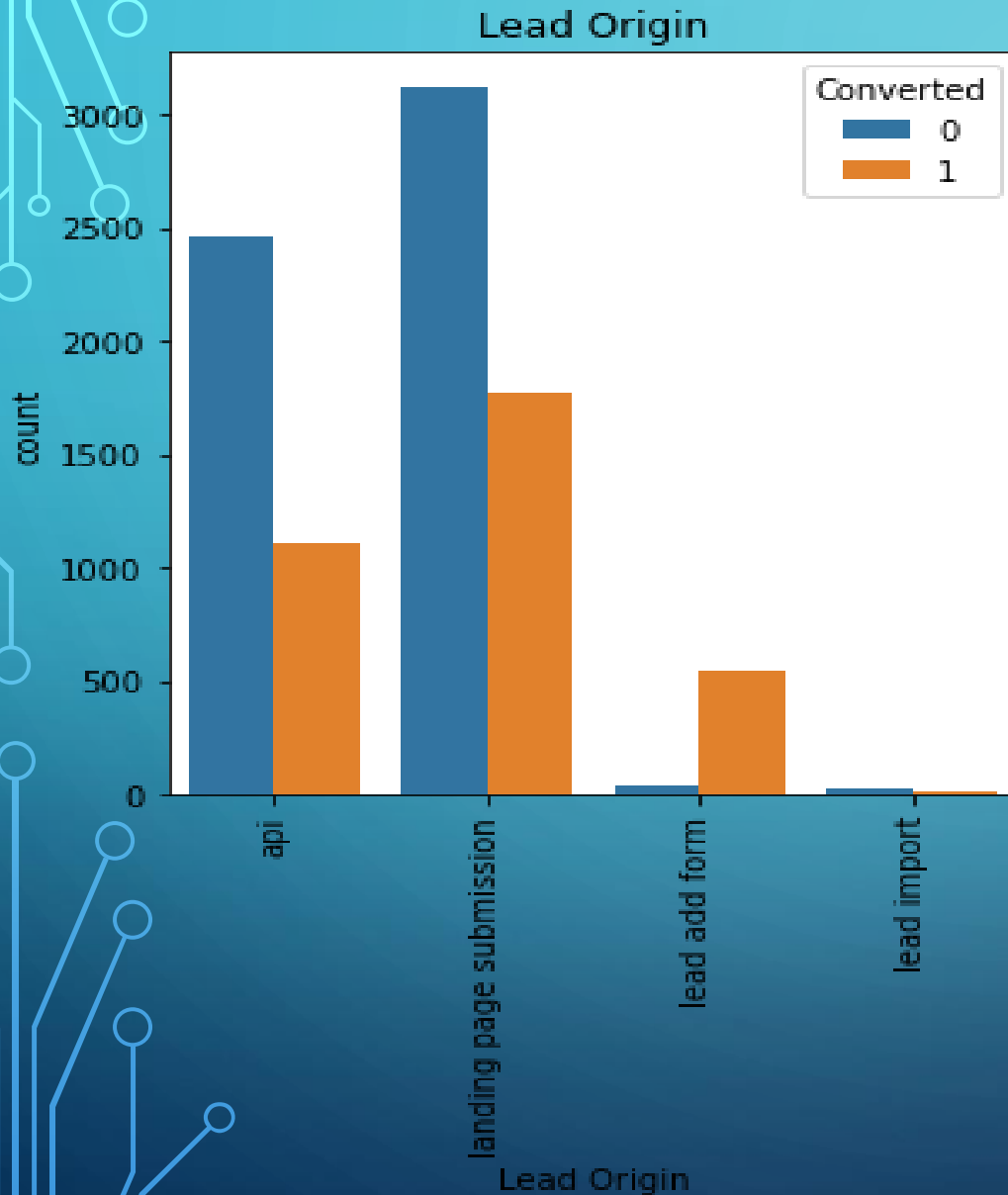- Prediction on test set

- Precision- Recall

DATA MANIPULATION

- Total Number of Rows =37, Total Number of Columns =9240.

- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply"

- Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.

- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.

- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.

- Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.
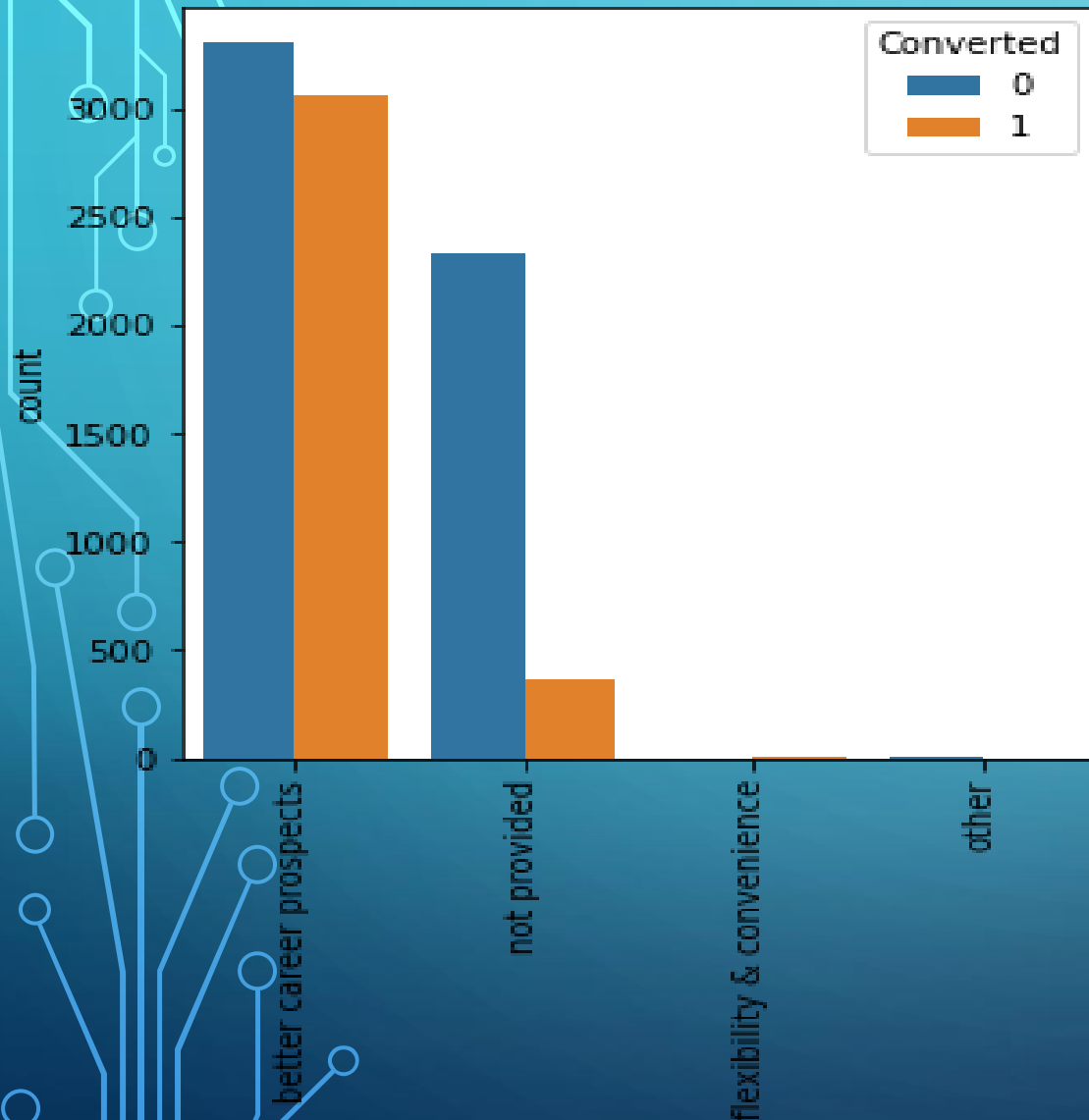
EDA

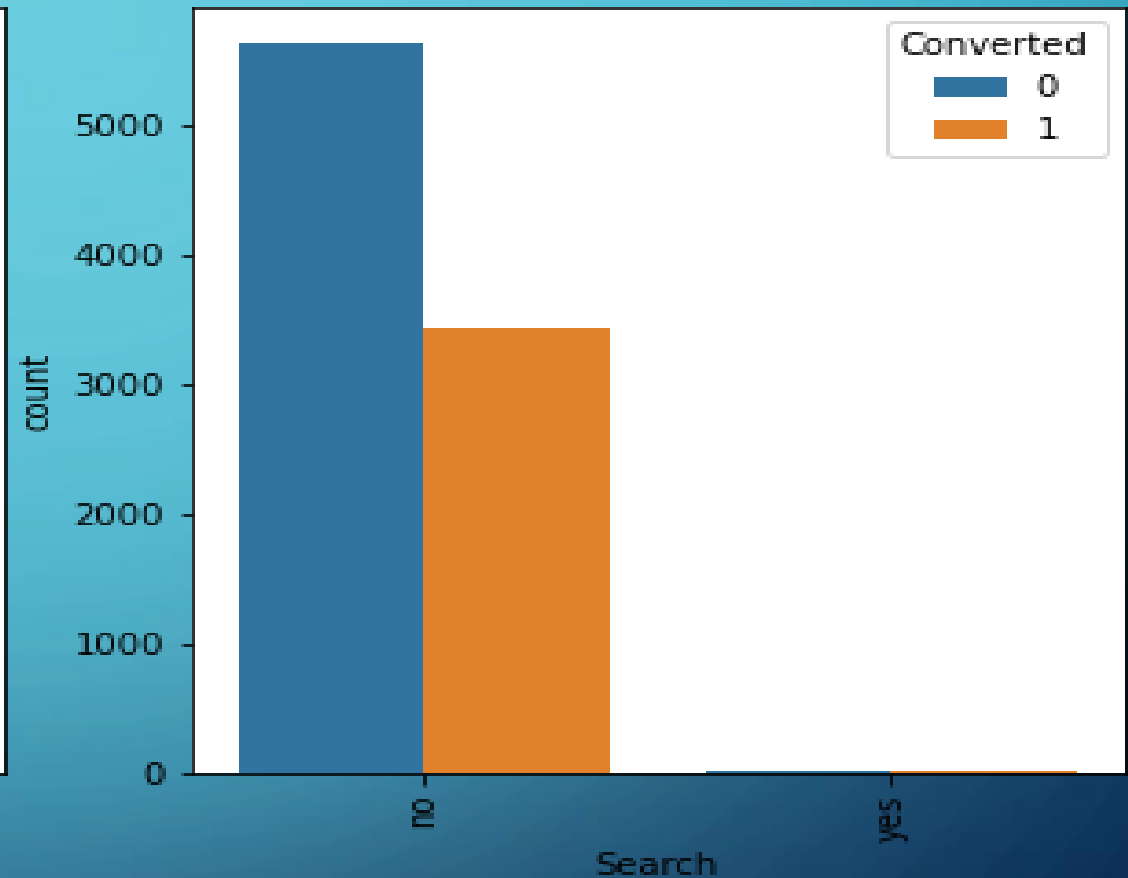
Lead Source

CATEGORICAL VARIABLE

CATEGORICAL VARIABLE

CATEGORICAL VARIABLE
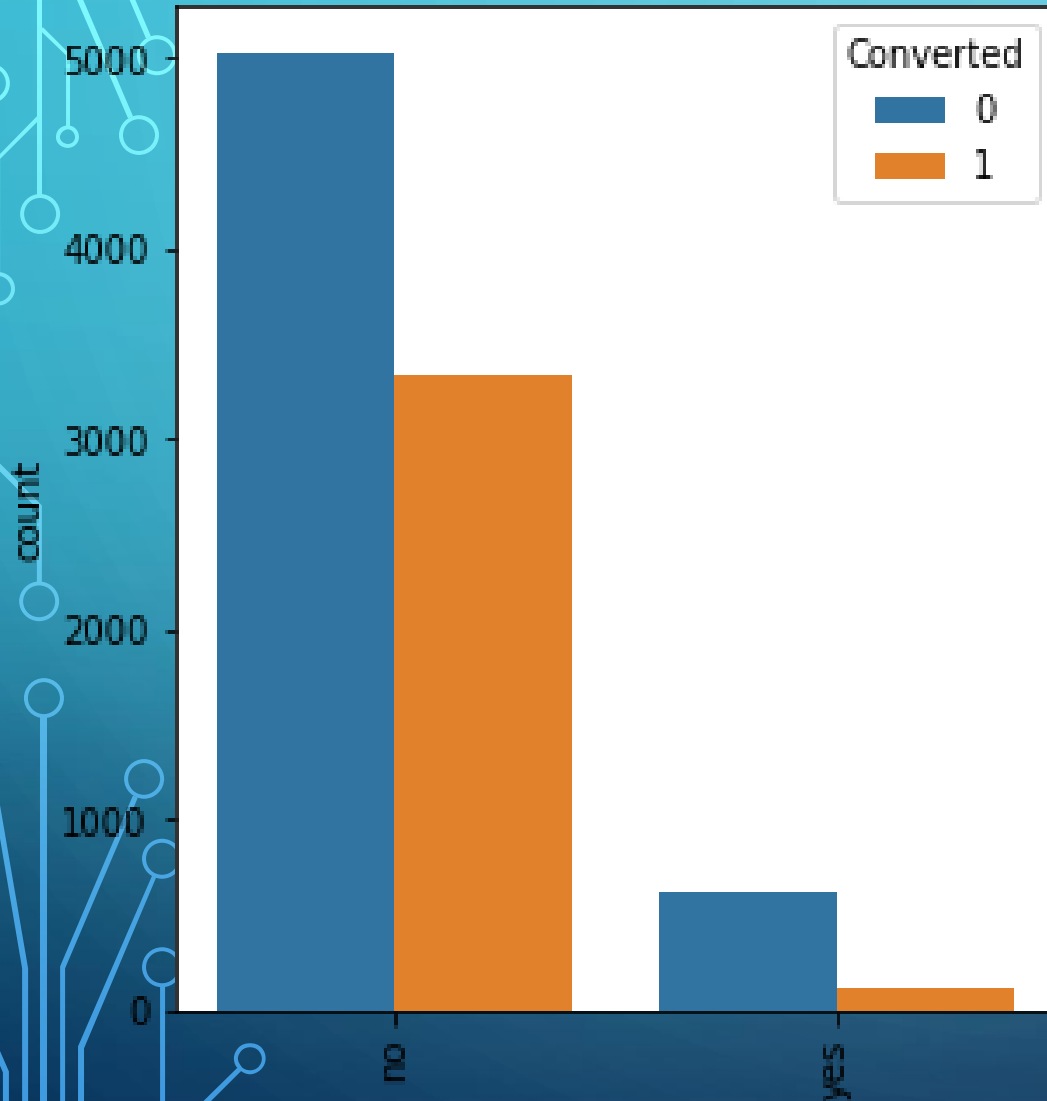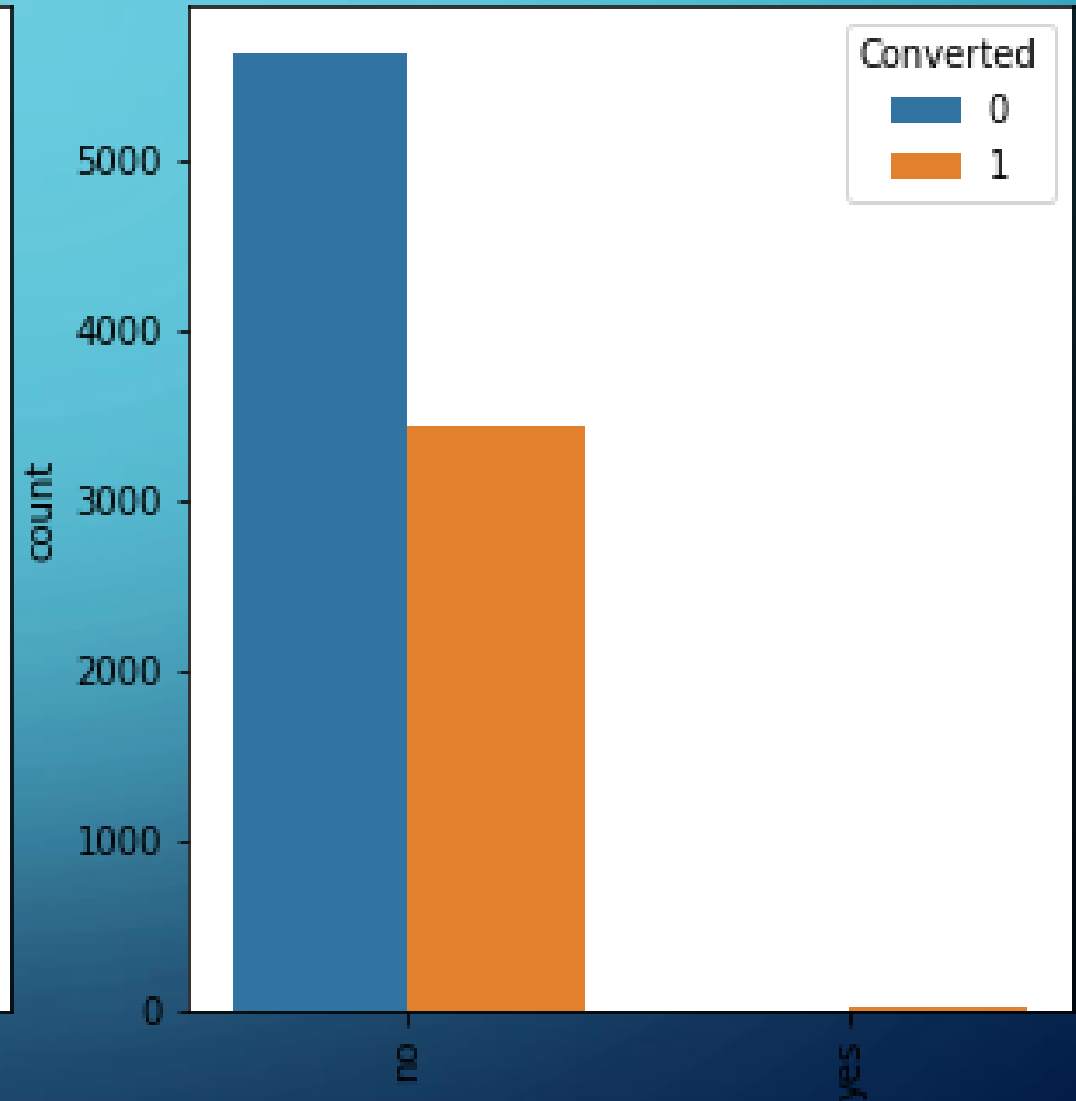
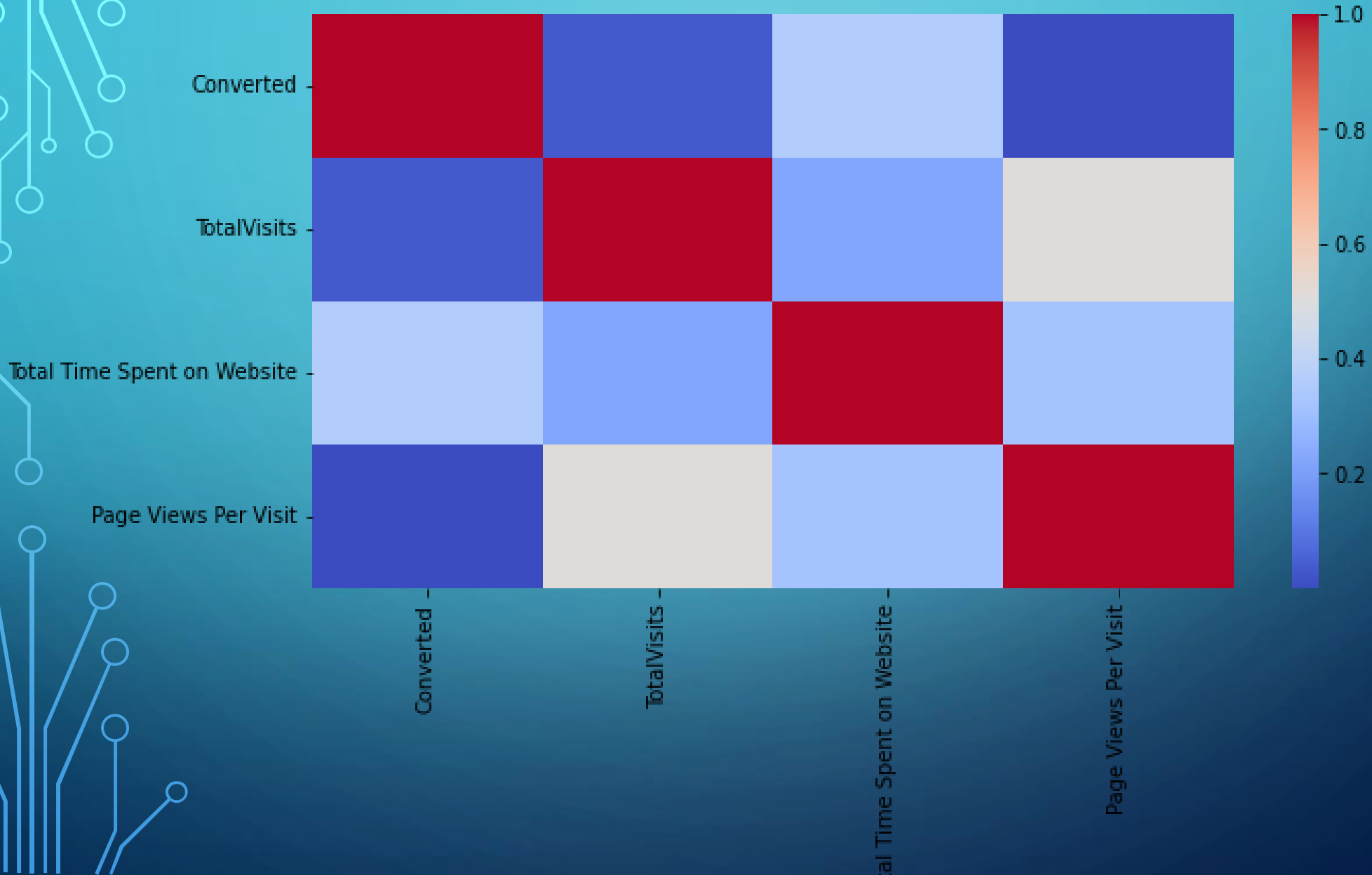# CHECKING FOR CORELATION

MODEL BUILDING

- Splitting the Data into Training and Testing Sets

- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.

- Use RFE for Feature Selection

- Running RFE with 15 variables as output

- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5

- Predictions on test data set

- Overall accuracy 81%

ROC CURVE

Receiver operating characteristic example
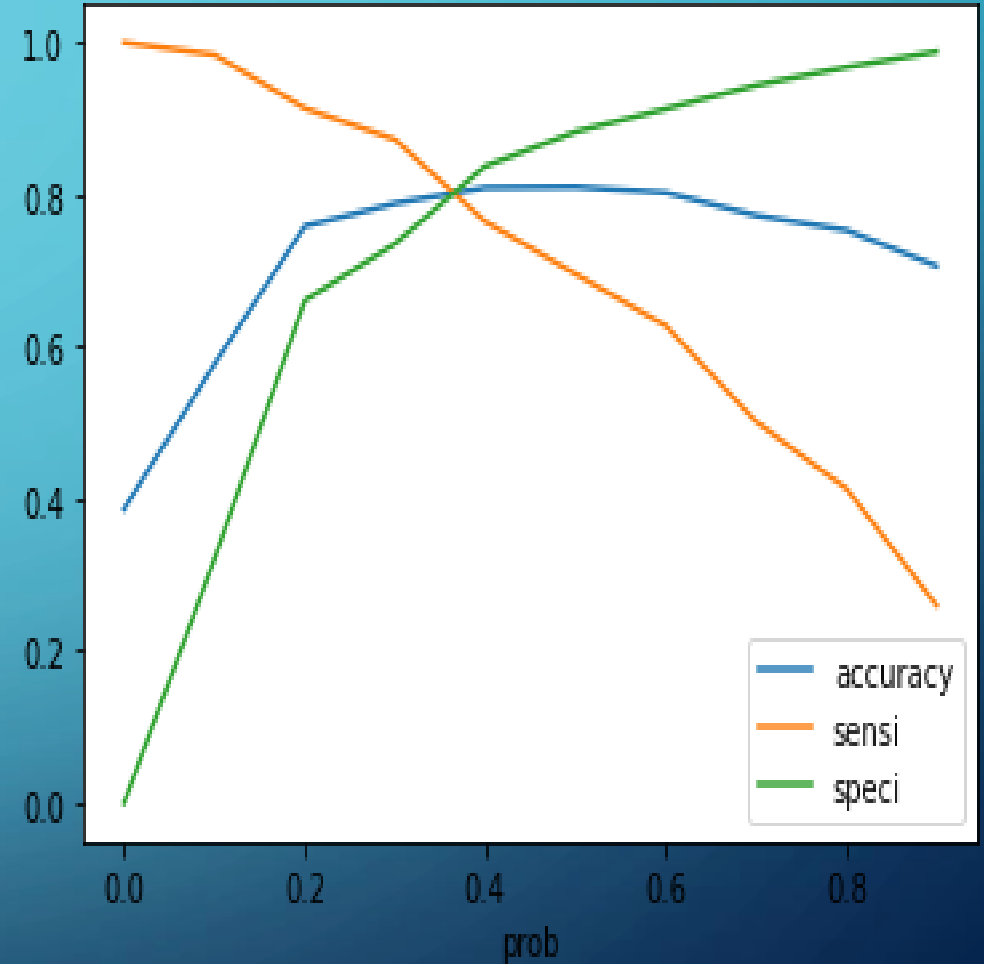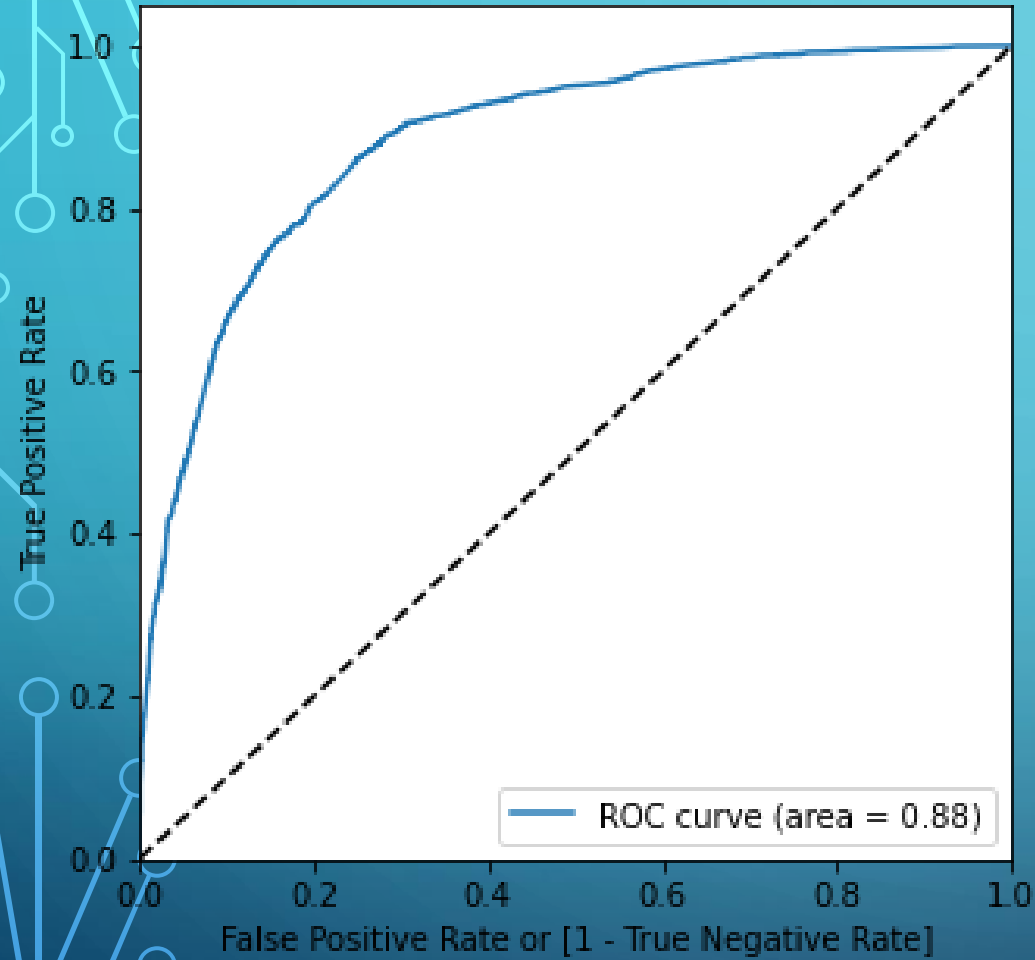
ROC IS 0.87 & OPTIMAL CUTOFF IS 0.35

CONCLUSION

- The total time spend on the Website.

- Total number of visits.

- When the lead source was:

- a. Google

- b. Direct traffic

- c. Organic search

- d. Welingak website

- When the last activity was:

- a. SMS

- b. Olark chat conversation

- When the lead origin is Lead add format.

- When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses