



LEAD SCORE CASE STUDY

SUBMITTED BY:

- SANCHIT KUMAR
- PRAGATHI N
- DERIL DANY
DSOUZA

Problem Statement:

- ❖ X Education sells online courses to industry professionals.
- ❖ 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.

Business Objective:

- ❖ X education wants to know most promising leads.
- ❖ For that they want to build a Model which identifies the hot leads.
- ❖ Deployment of the model for the future use.

METHODOLOGY USED

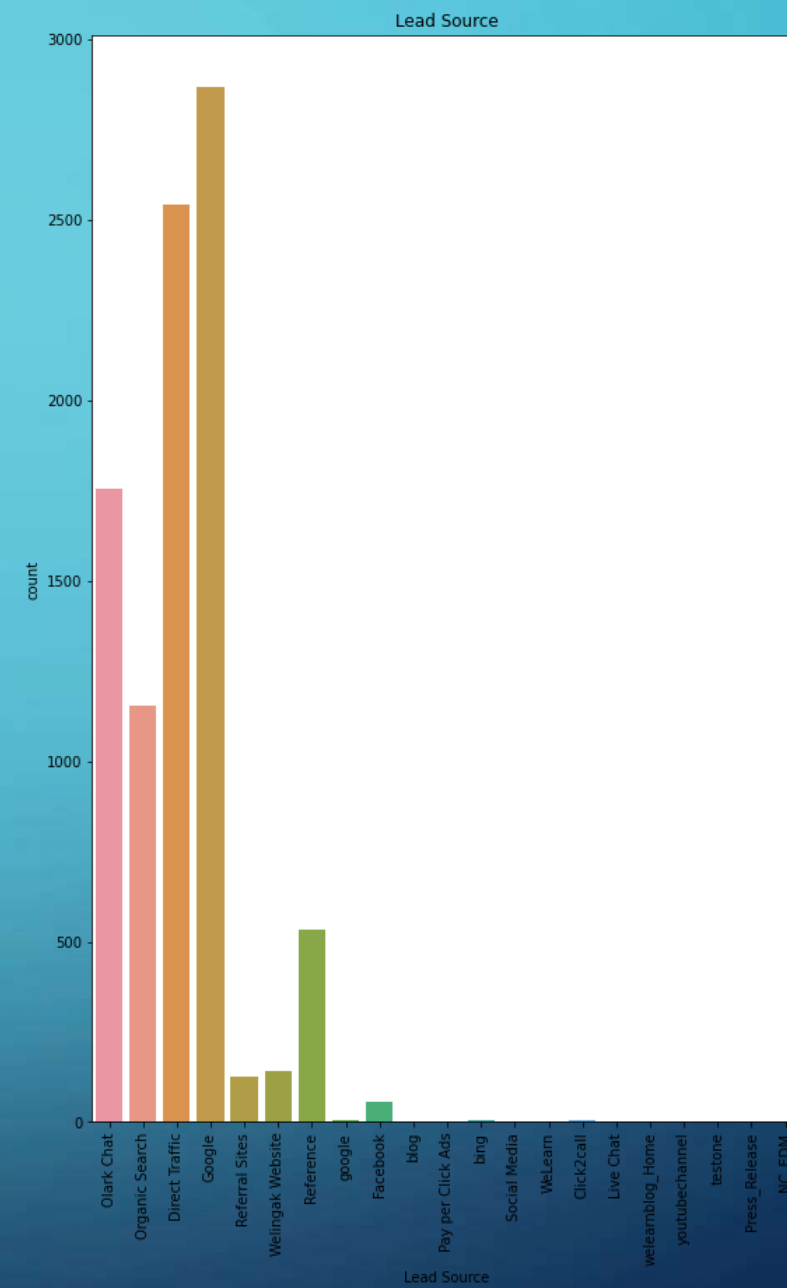
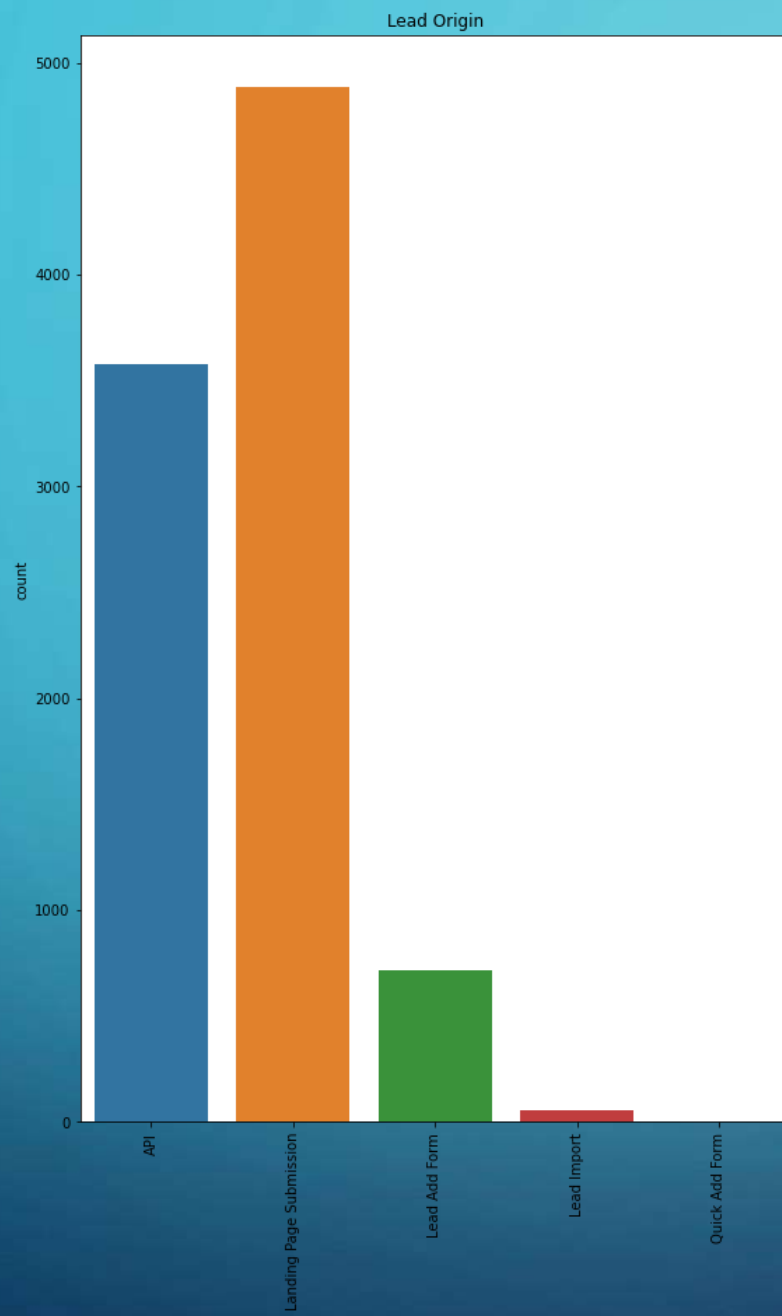
1. Data cleaning and data manipulation.
2. EDA
3. Feature Scaling & Dummy Variables and encoding of the data.
4. Classification technique: logistic regression used for the model making and prediction.
5. Validation of the model.
6. Model presentation.
7. Conclusions and recommendations

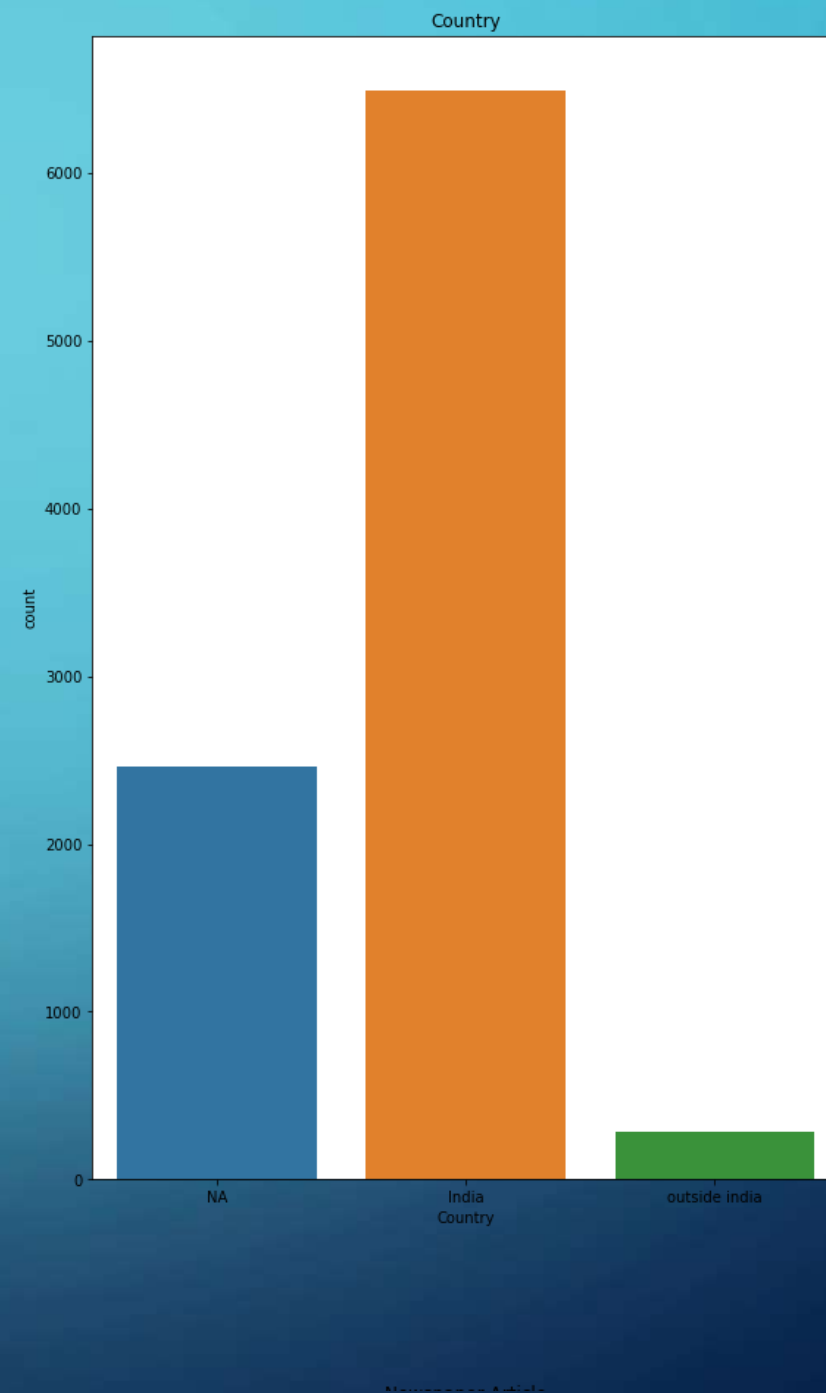
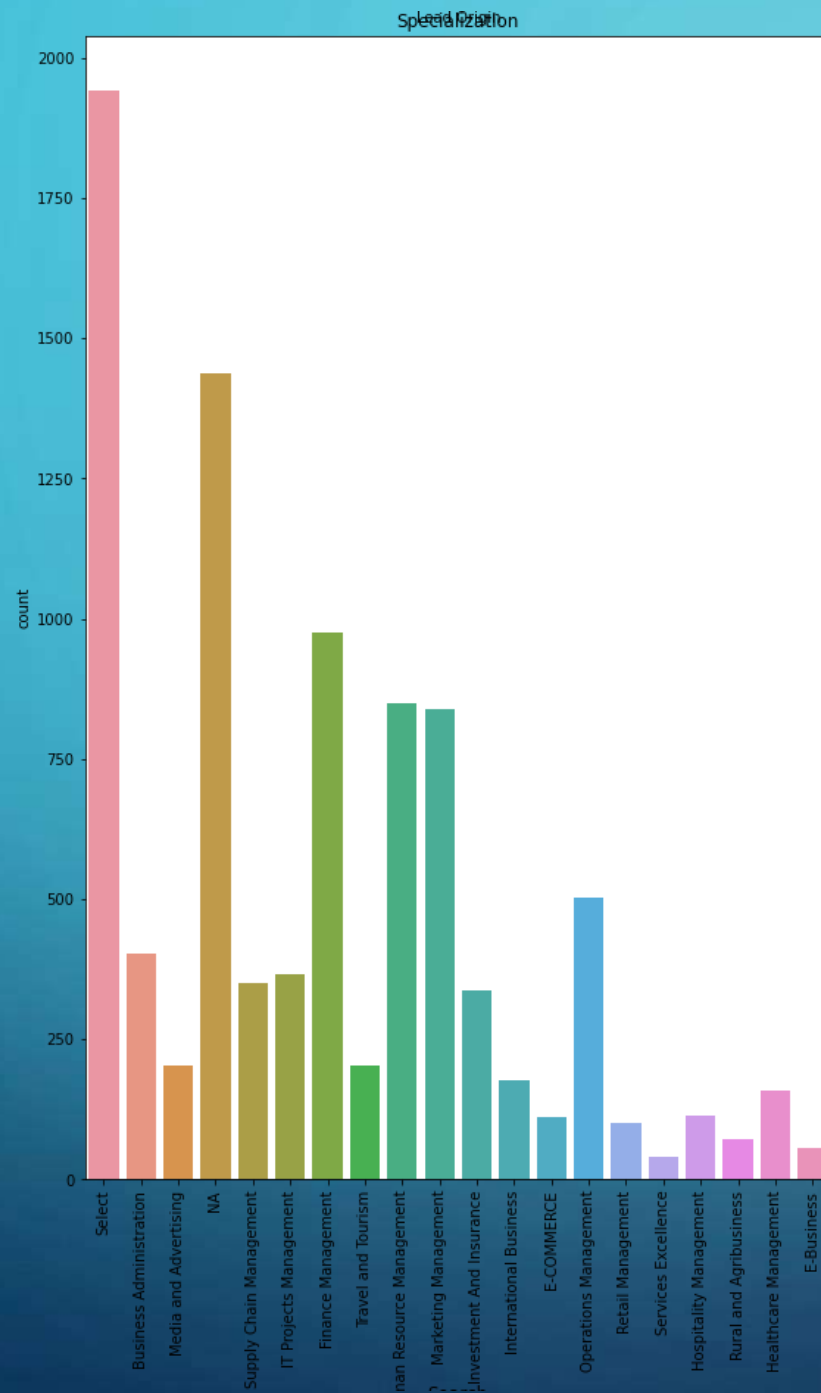
DATA MANIPULATION

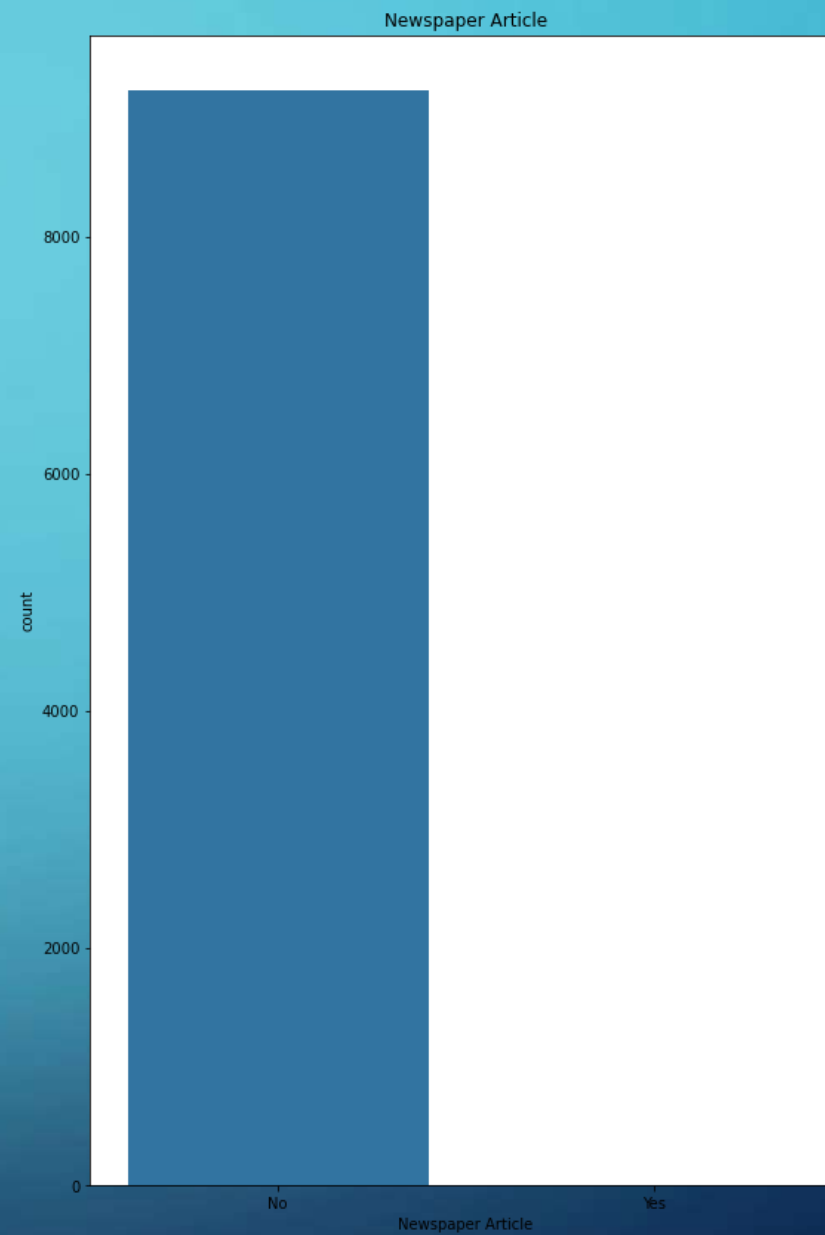
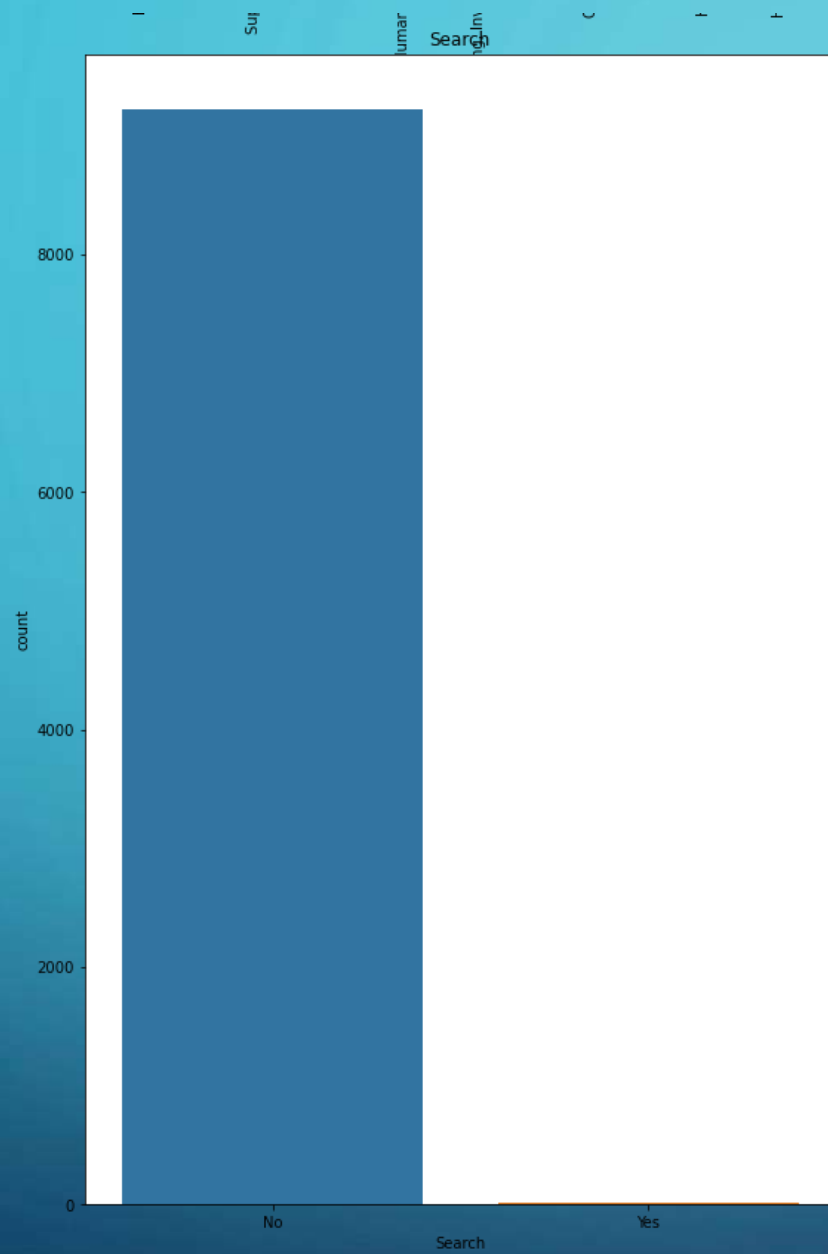
1. Total Number of Rows =37, Total Number of Columns =9240. Single value features like “Do Not Email”, “Last Activity”, “Tags”, “Update me on Supply Chain Content” etc. have been dropped.
2. Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.
3. After checking for the 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.
4. Dropping the columns having more than 40% as missing value such as “Asymmetrique Activity Score”, “Asymmetrique Profile Score”, “Lead Quality”.

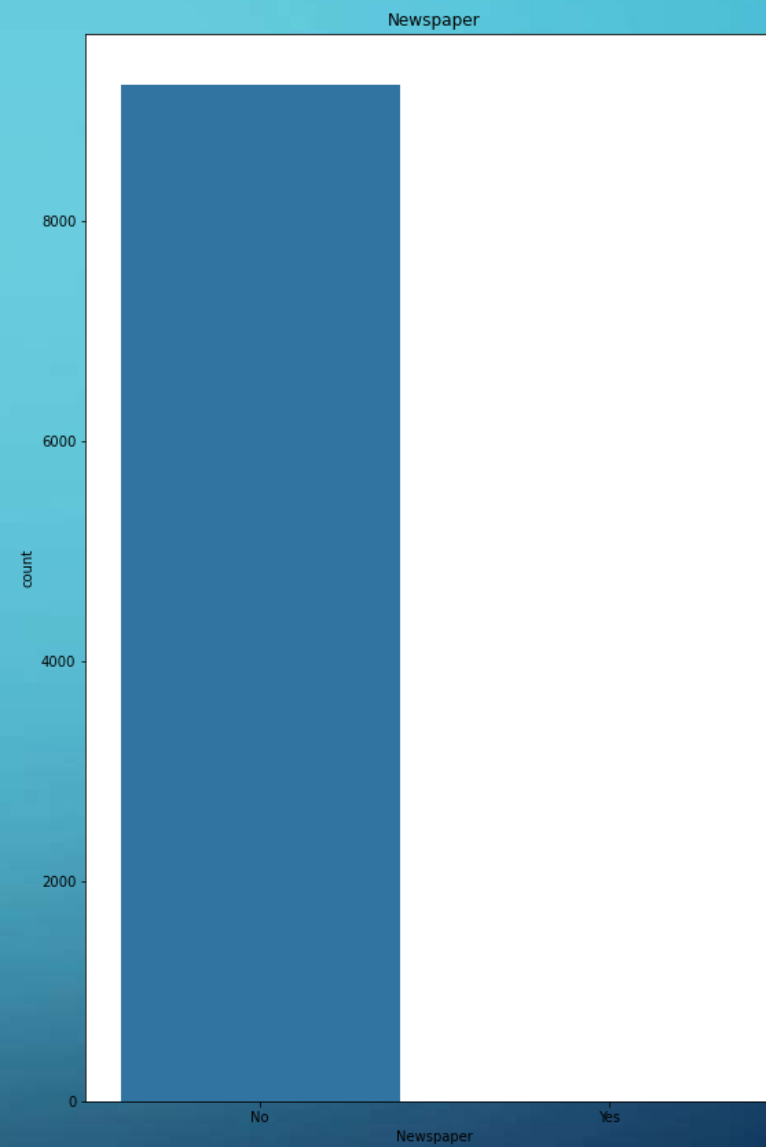
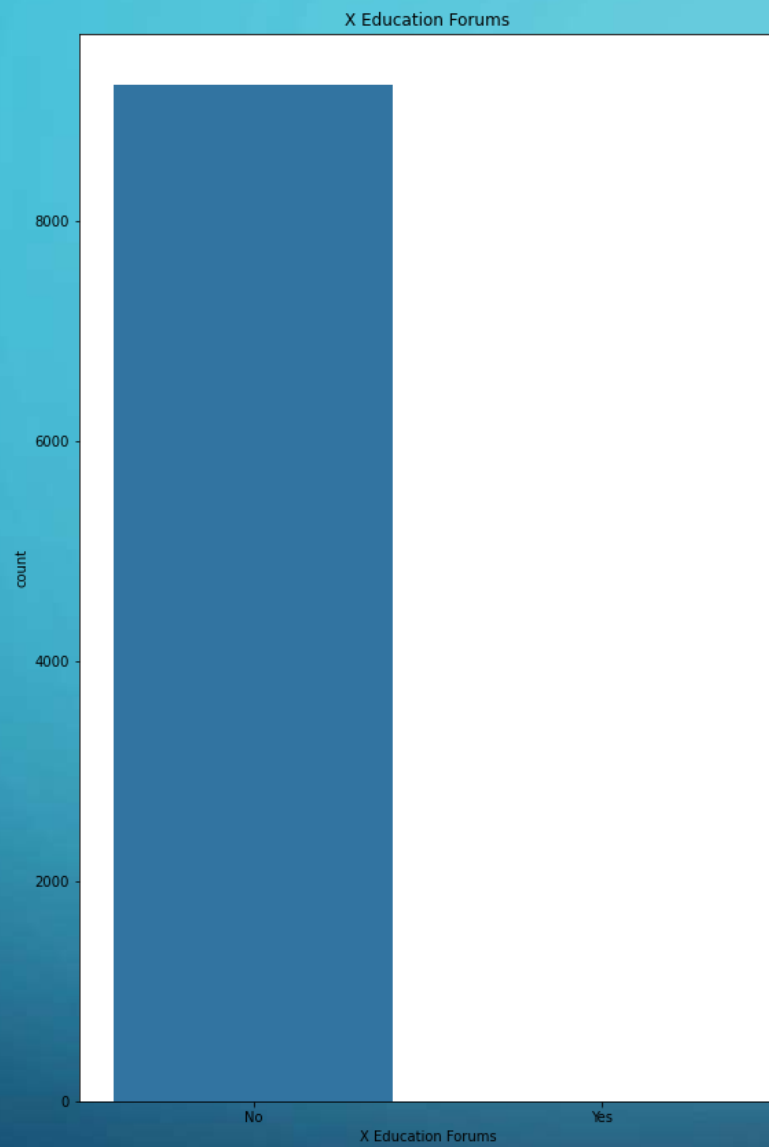


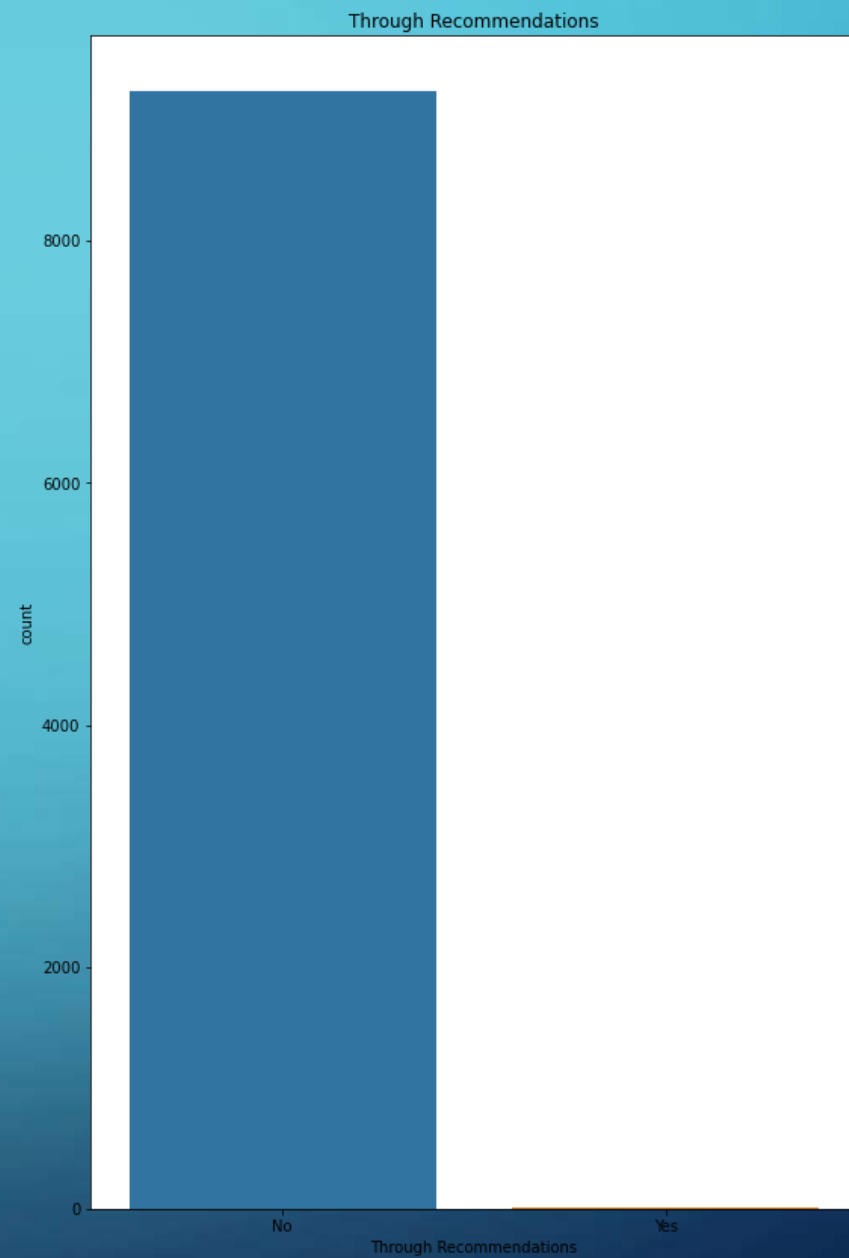
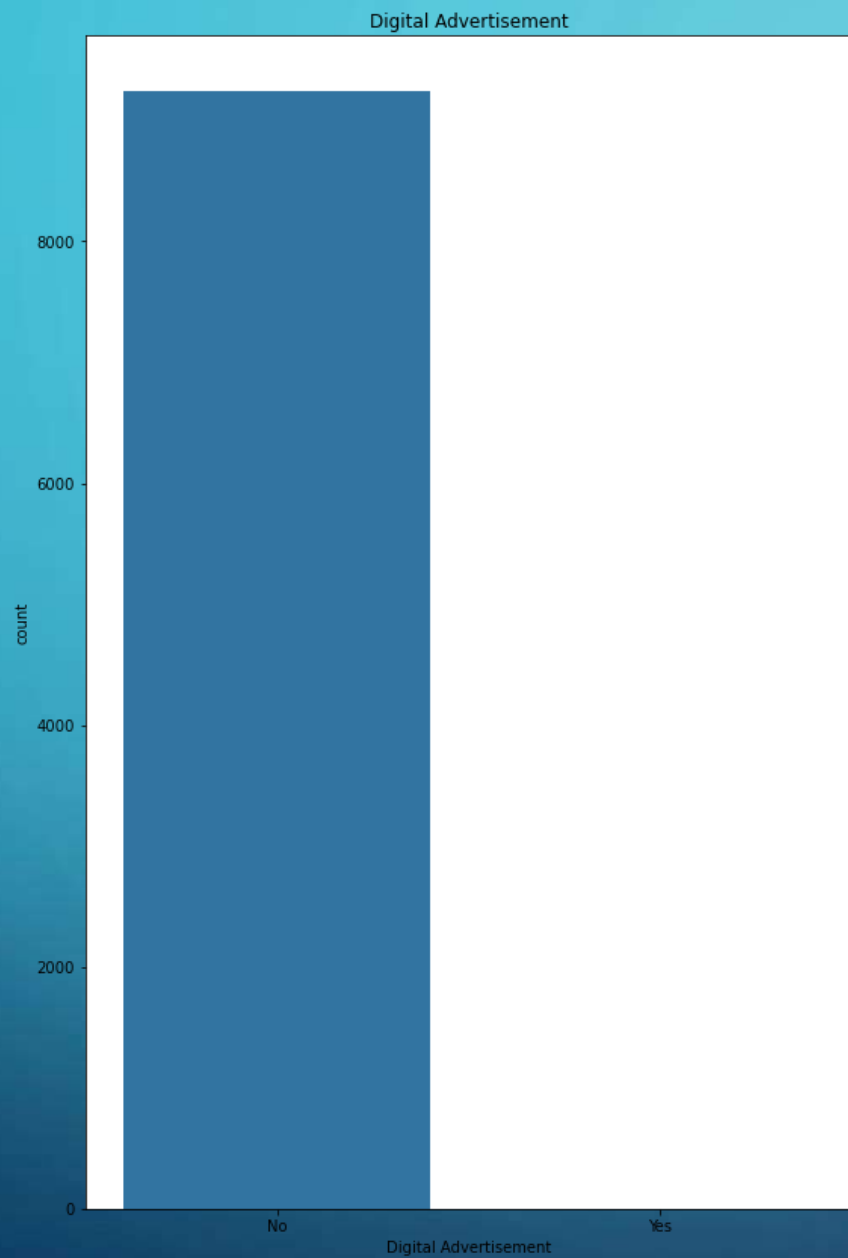
Categorical Variable Relation & Correlation Heatmap

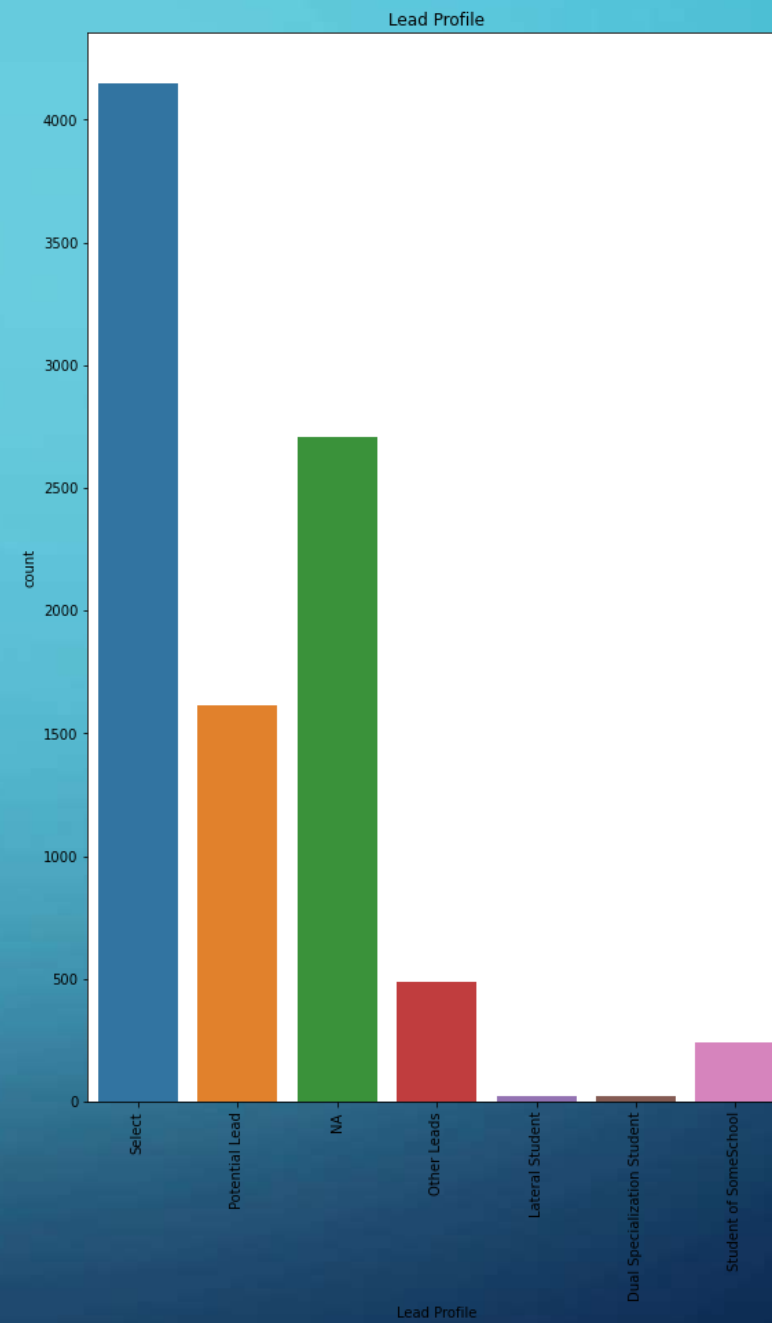
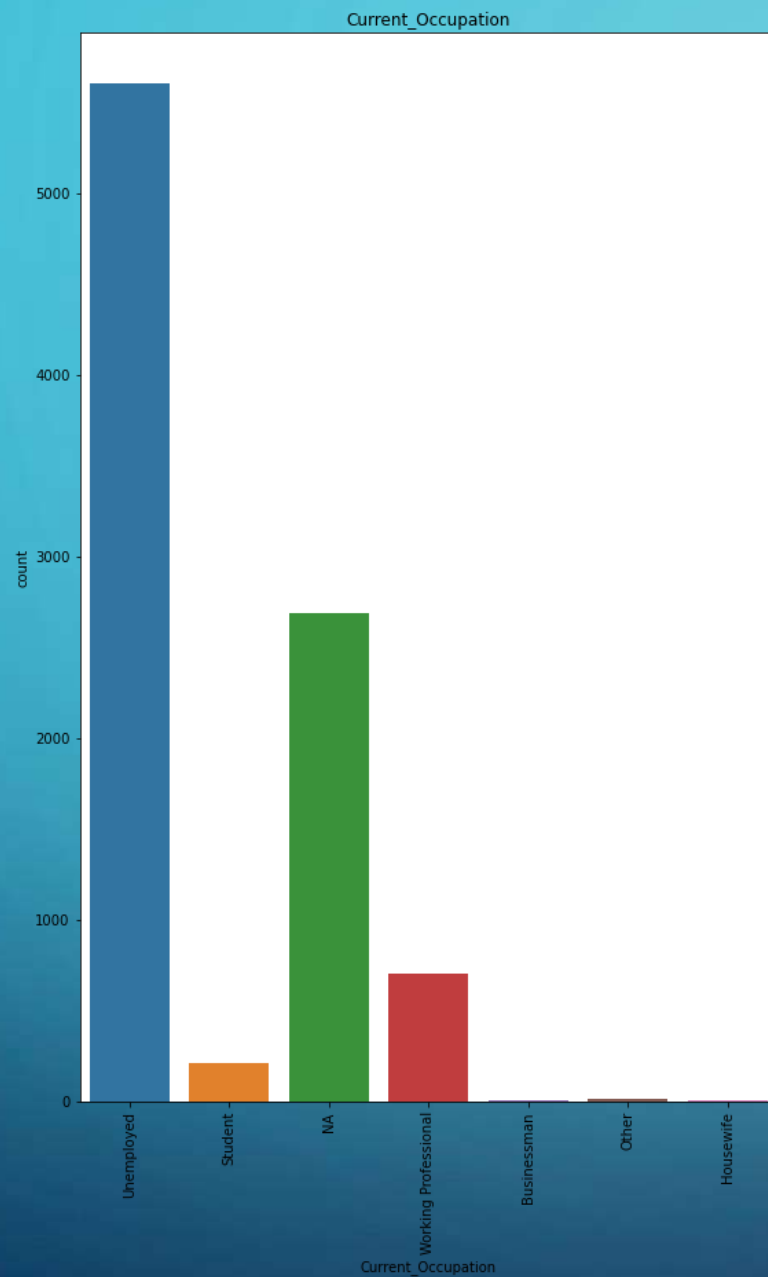










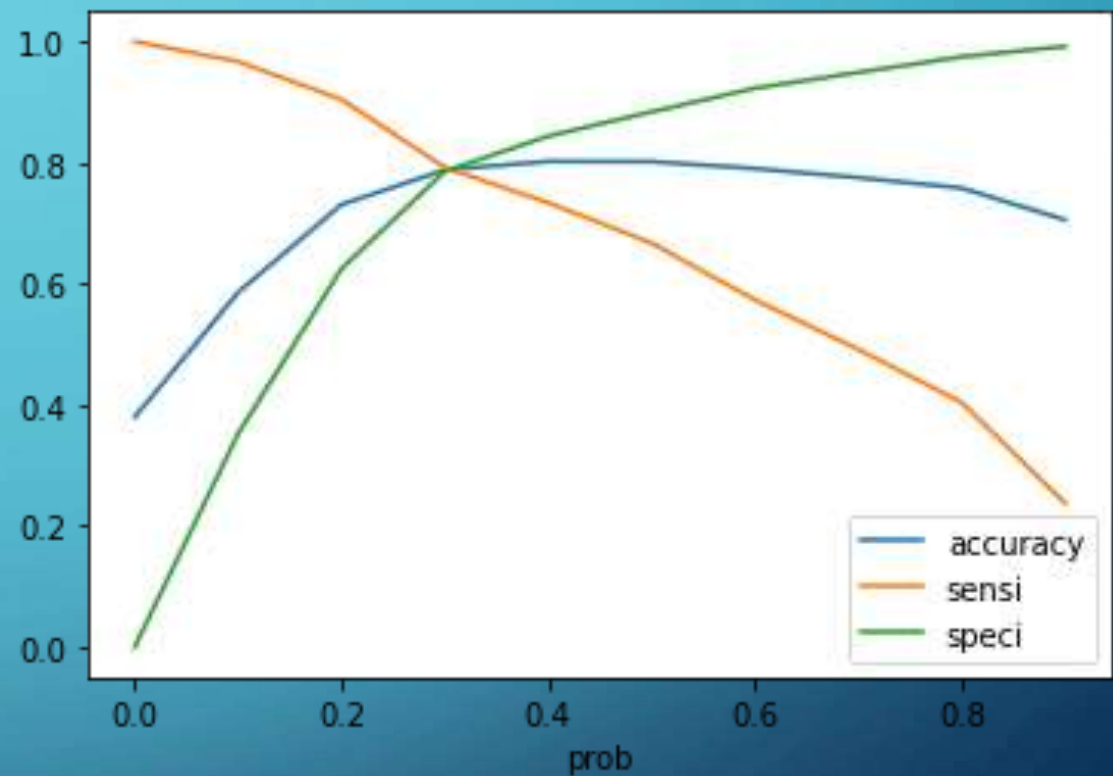
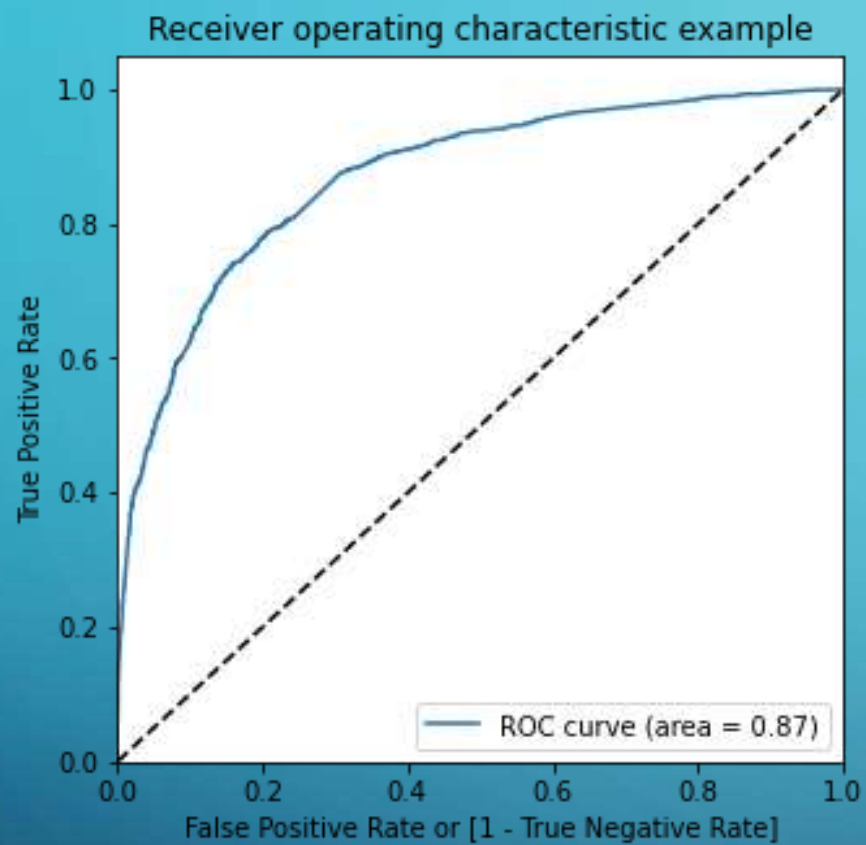




Model Building

1. Splitting the Data into Training and Testing Sets.
2. The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
3. Use RFE for Feature Selection.
4. Running RFE with 15 variables as output.
5. Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
6. Predictions on test data set.
7. Overall accuracy 81%

ROC CURVE



Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- ❖ 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 course.