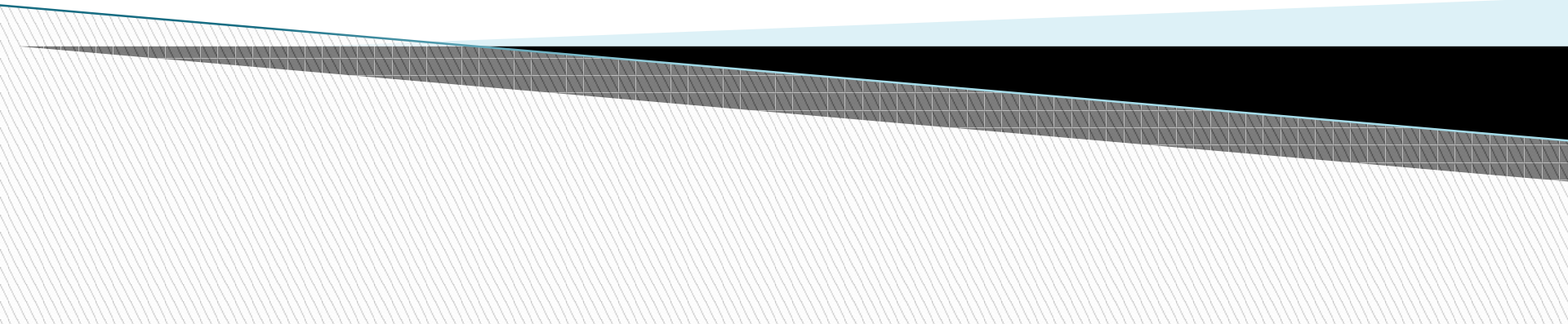


LEAD SCORE CASE STUDY

-Leelavati



Solution Methodology

Data Cleaning and Data manipulation

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns if it contains large volume of missing values
4. Check and handle outliers in the data

Exploratory Data Analysis

Feature Scaling and Creating dummy variables

Classification technique i.e. Logistic Regression is used for model making and prediction.

Validation of the Model.

Model Presentation

Conclusions and recommendations



Problem Statement

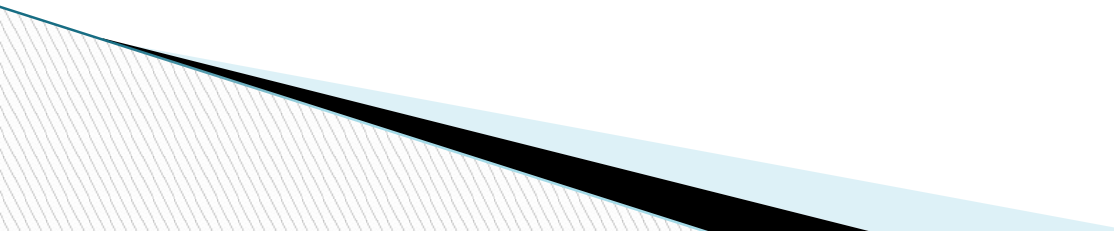
X Education is a company that sells online courses to the professionals and students. It has many Leads visiting their website to buy the courses. And 30 out of 100 would like to buy them. Hence the conversion rate is only 30%.

Hence, to be efficient with their marketing, they want to identify the potential buyers i.e. 'Hot Leads'.

If through this analysis, if they find out potential leads, their sales is expected to grow high.

Business Objective

X Education wants to now promising leads. For which, they want to build a model to find the 'Hot Leads'. And this model can be deployed for future use also.



Data Manipulation

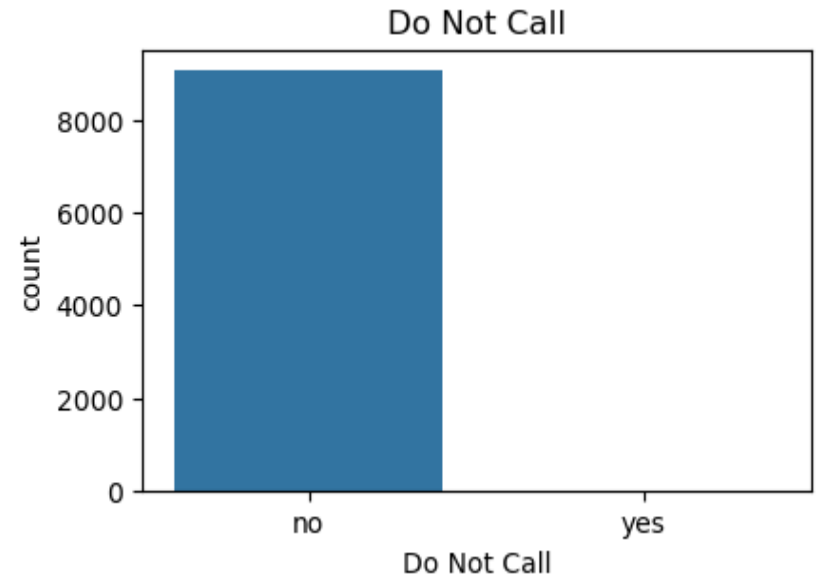
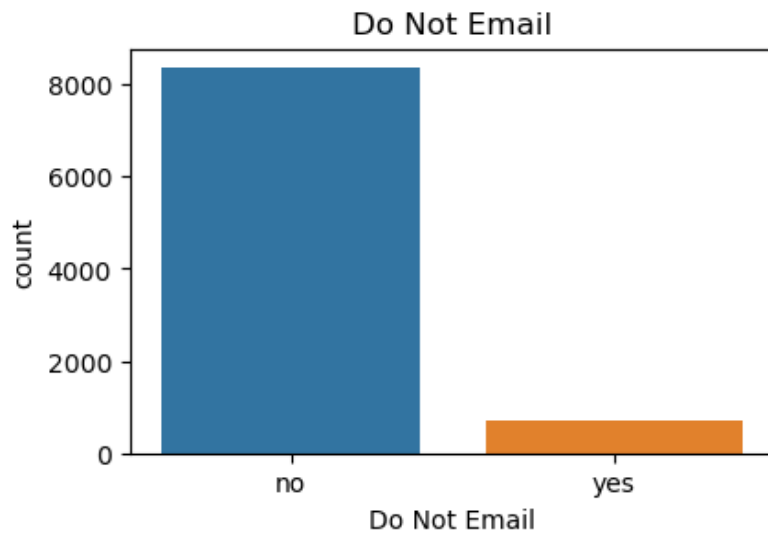
Total number of rows = 9240, Total number of columns = 37.

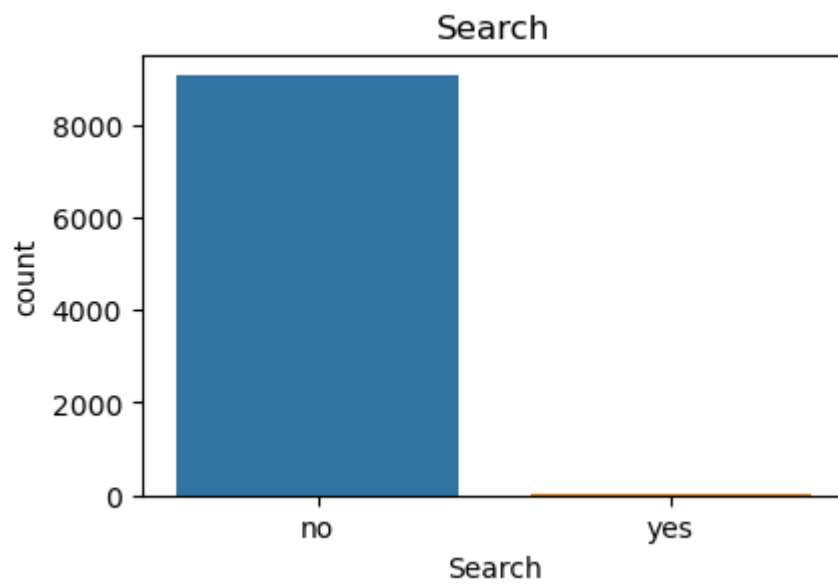
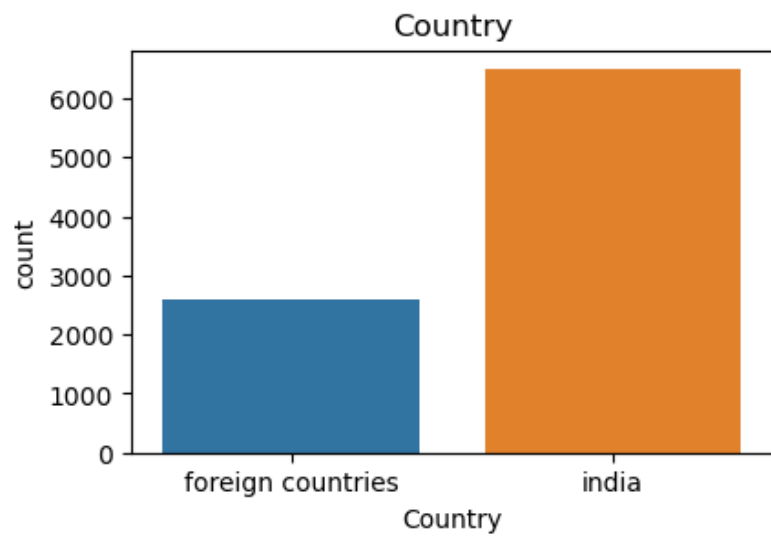
Single value features like “Magazine”, “Receive More Updates about Our Courses”, “Update me on supply”, “Chain Content”, “Get Updates on DM content”, “I agree to pay the amount through Cheque”, etc. have been dropped.

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

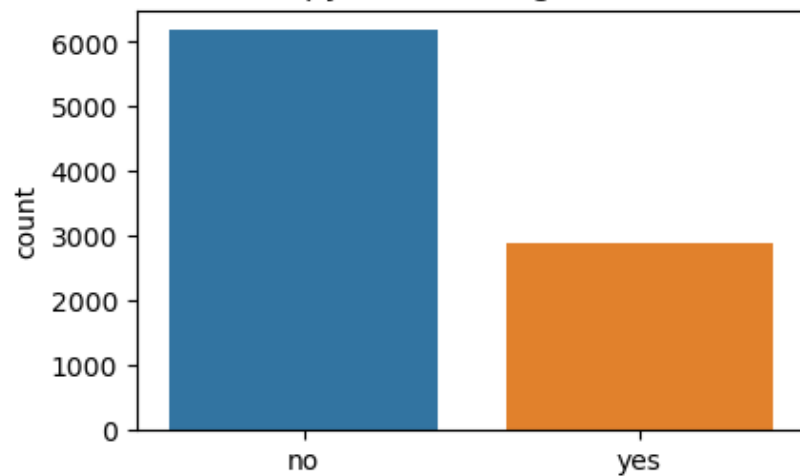
Dropping the columns having more than 35% as missing values such ‘How did you hear about X Education’ and “Lead Profile”.

Exploratory Data Analysis



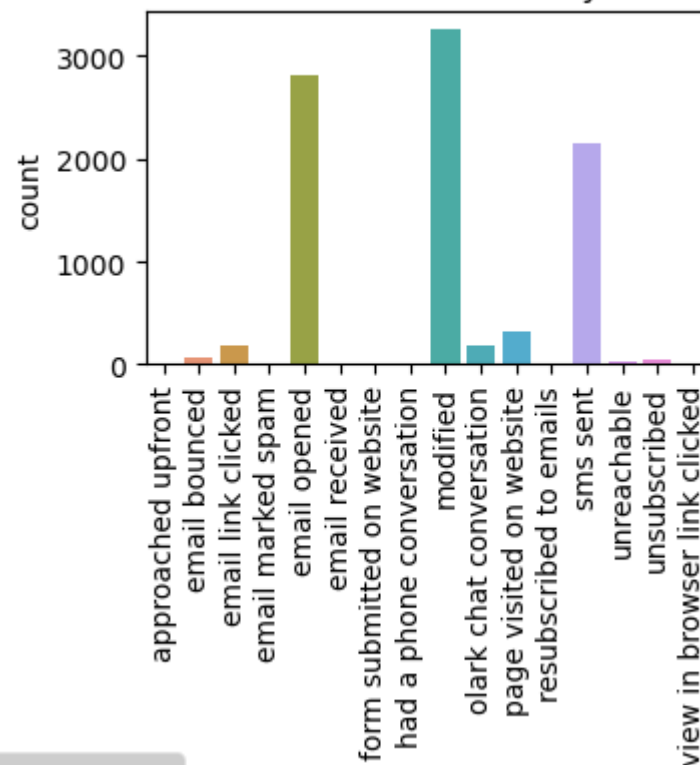


A free copy of Mastering The Interview



A free copy of Mastering The Interview

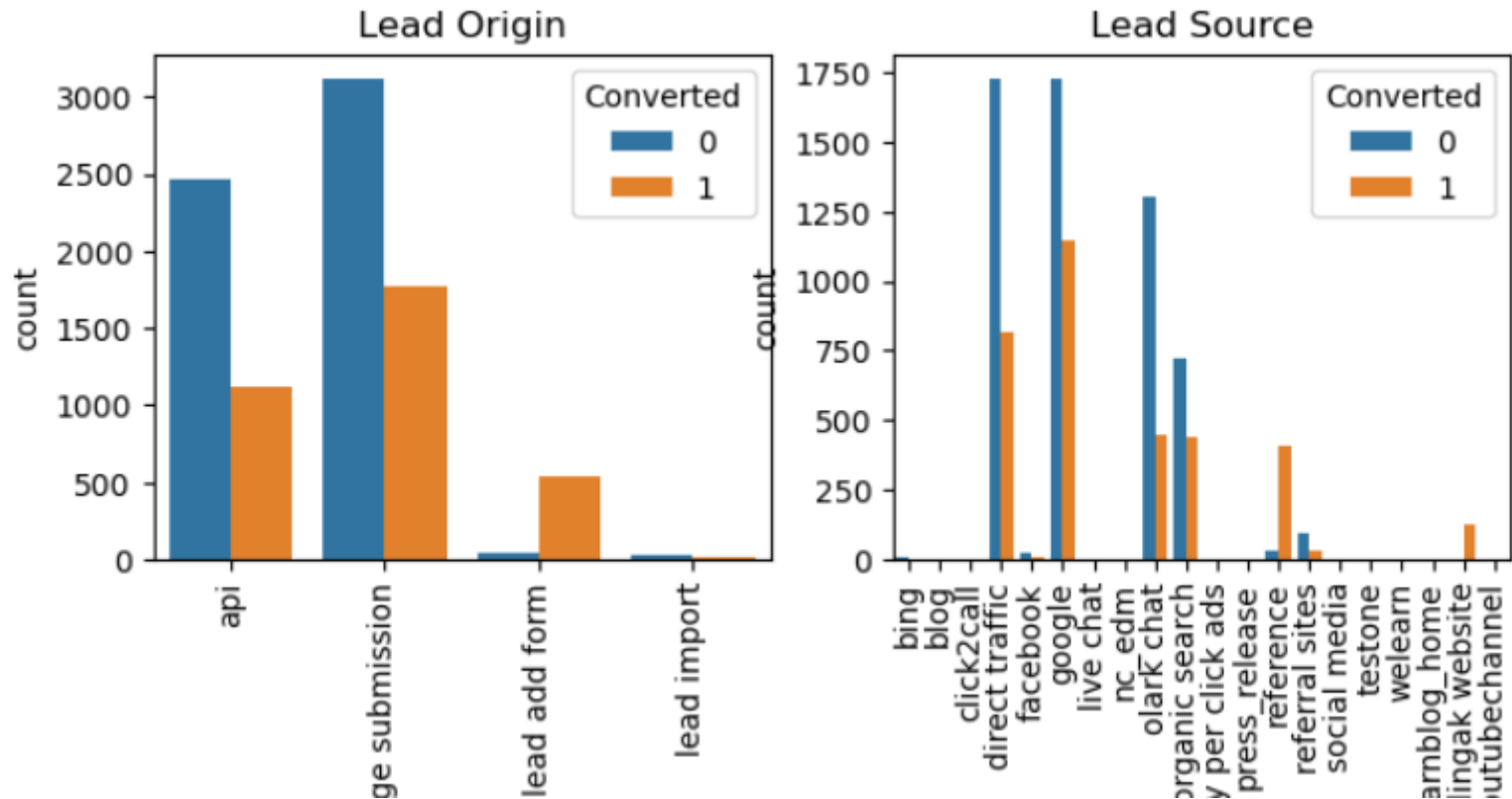
Last Notable Activity

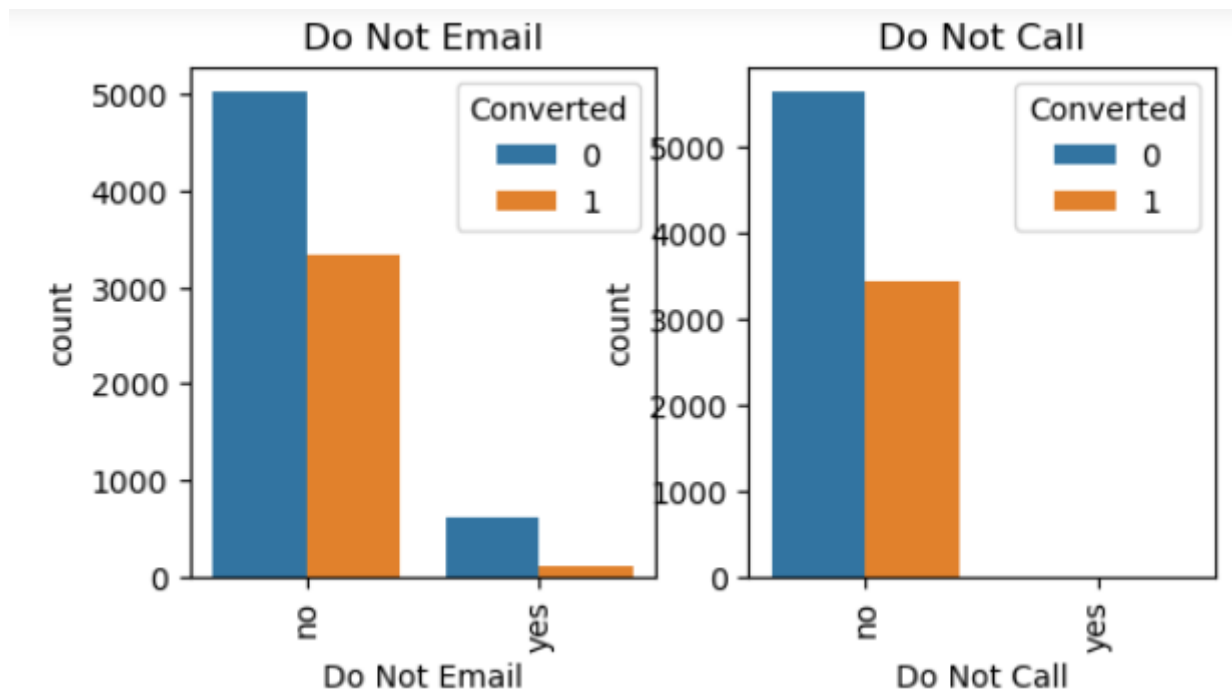


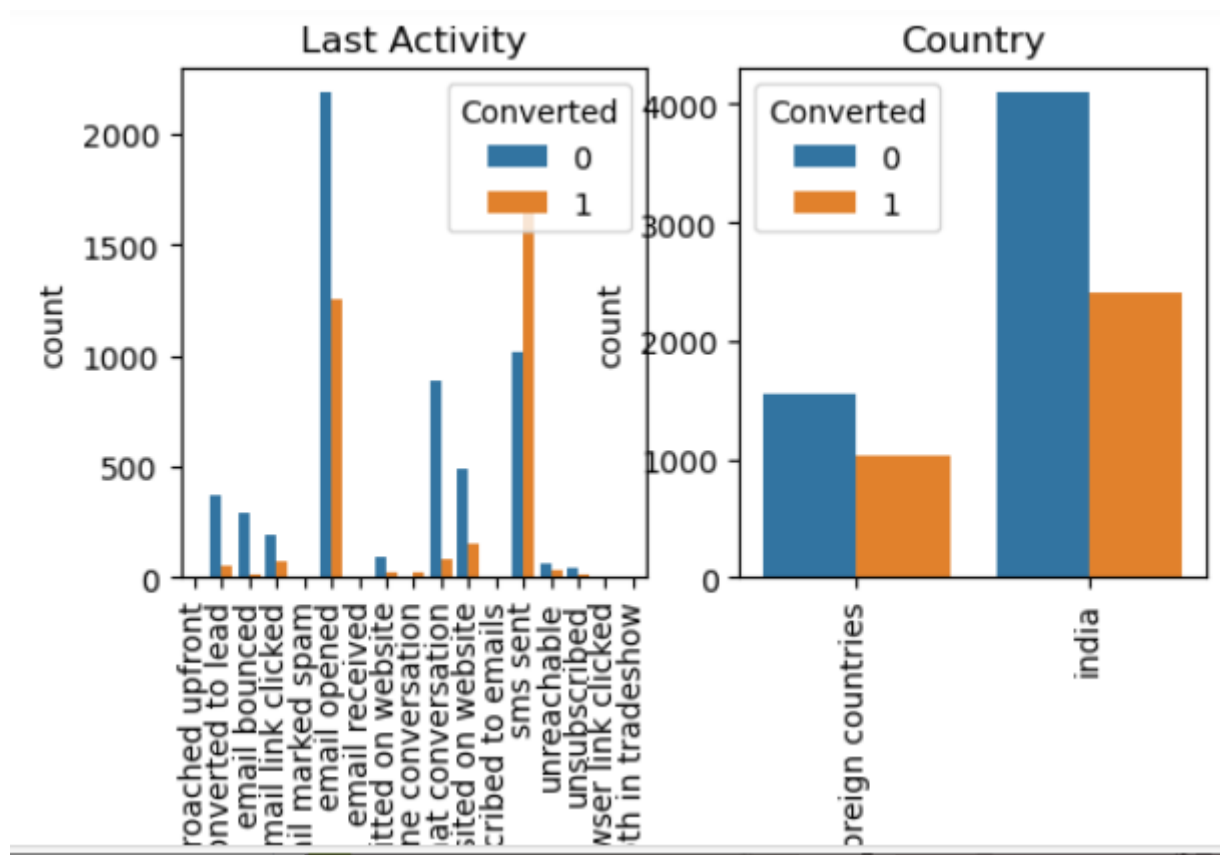
ase study.ipynb#

Last Notable Activity

Categorical Variable Relation



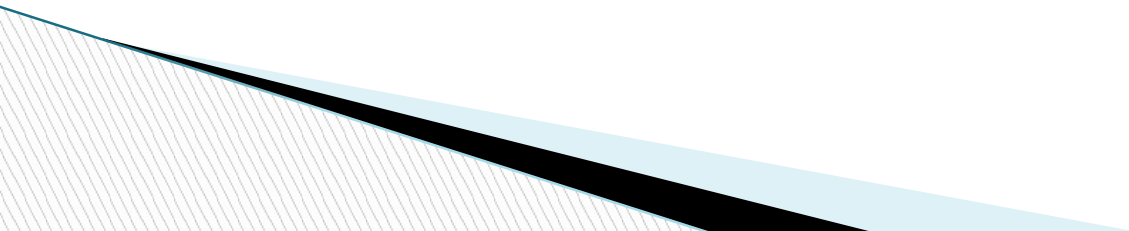




Data Conversion

Numerical Variables are normzalized.

Dummy variables are created for object type variables.



Model Building

Splitting the data into train and Testing sets.

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

Use RFE(Recursive Feature Elemination) for feature selection.

Running RFE with 15 variables as output.

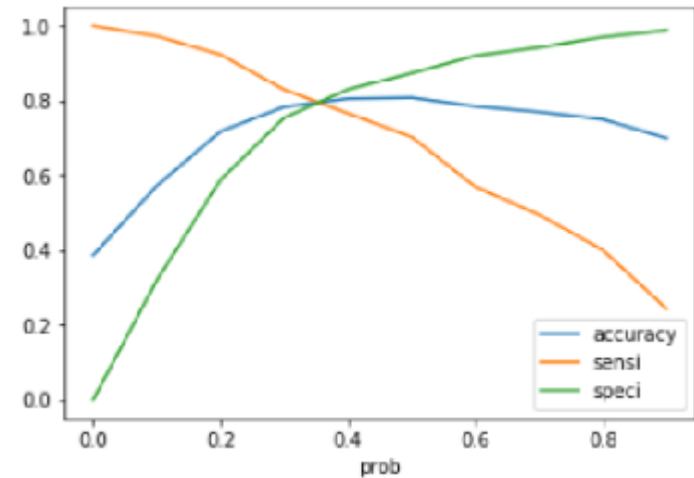
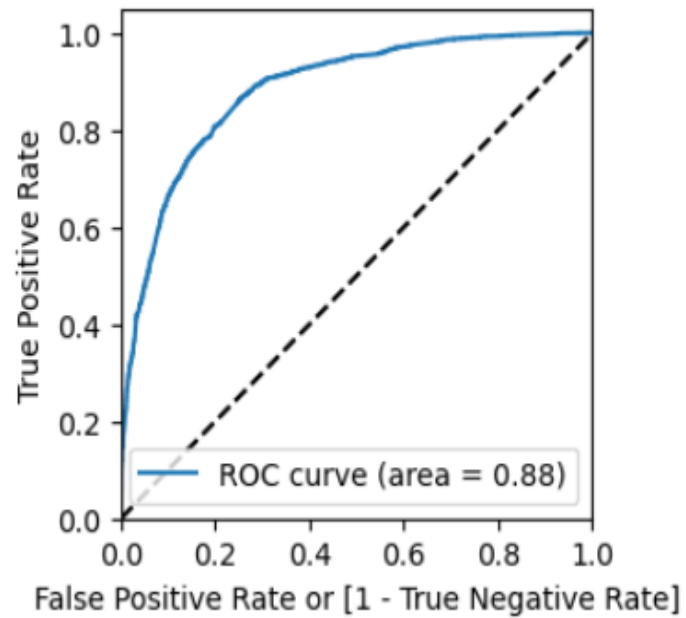
Building model by removing the variable whose p-value is greater than 0.05 and vif value is greater than 5.

Prediction is performed on the dataset.

Overall Accuracy is found to be 81%

ROC curve

Receiver operating characteristic example



Conclusion

1. It was found that the variables that mattered the most in the potential buyers are :
2. The total time spent on the website.
 - 1.Total number of visits.
 - 2.When the lead source was:
 - a) Google
 - b)Direct Traffic
 - c)Organic Search
 - d)Welingak website
3. When the last activity was:
 - a)SMS
 - b)Olark Chat Conversation
4. When the lead origin is lead add format.
5. When the current occupation is if they are working professionals.
6. Jotting these points and working on them will help the X Education to flourish as they have very high probability of getting almost all the 'Hot Leads'.