Lead Score Case Study

Group Members :

Sivika Singh

Deepak Jha

Debdisha Banerjee

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.

SOLUTION METHODOLOGY

▶ Data Cleaning and Manipulation:
•Checked and handled duplicate data entries.
•Managed NA values and addressed missing data.
•Dropped columns with large amounts of missing values that were not useful for analysis.
•Imputed missing values where necessary.
•Identified and handled outliers in the data.

▶ Exploratory Data Analysis (EDA):
•Conducted univariate data analysis to assess value counts and variable distributions.
•Performed bivariate data analysis to examine correlation coefficients and identify patterns between variables.

▶ Feature Scaling and Encoding:
•Implemented feature scaling and created dummy variables for categorical data.

▶ Classification Technique:
•Used logistic regression for model building and prediction.

▶ Model Validation:
•Validated the model to ensure its accuracy and reliability.

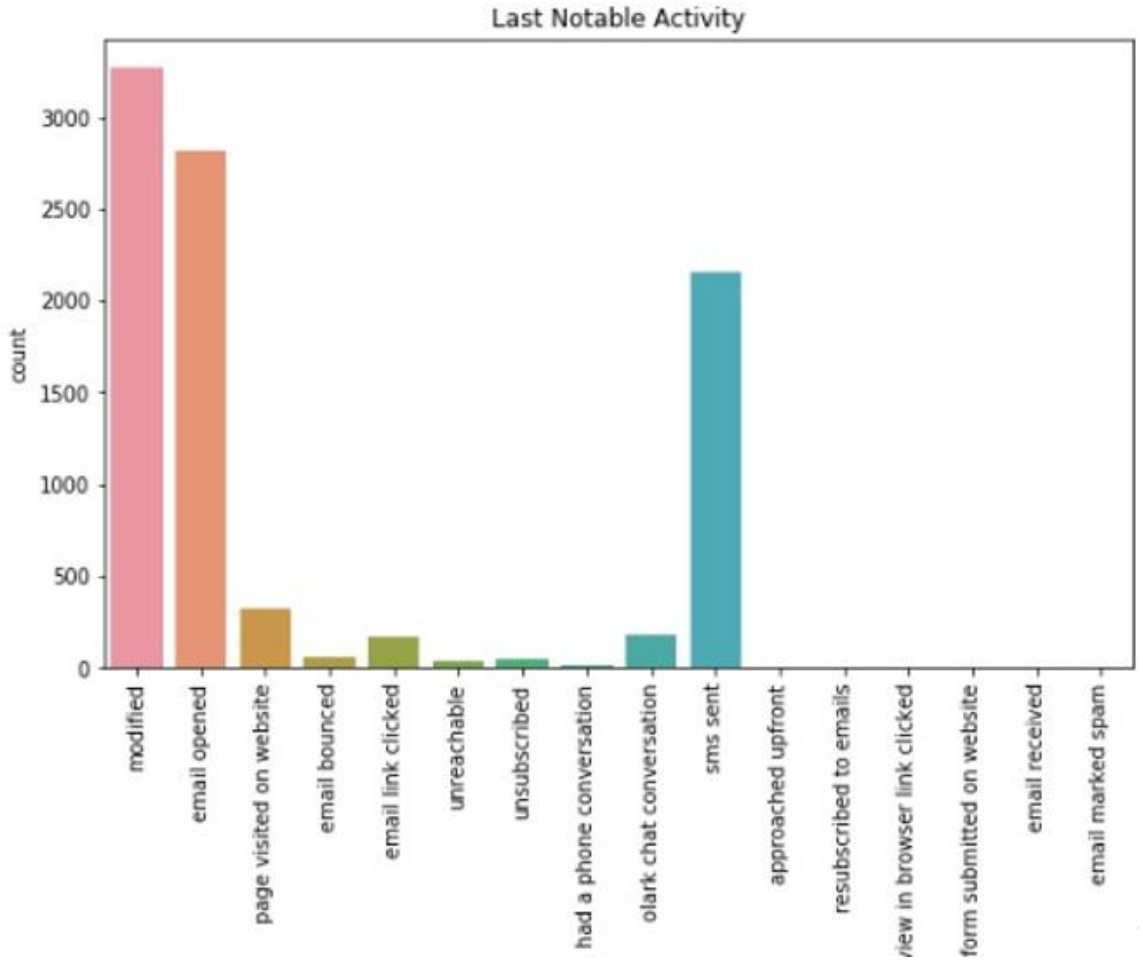
▶ Model Presentation:
•Presented the model, detailing the methodology and findings.

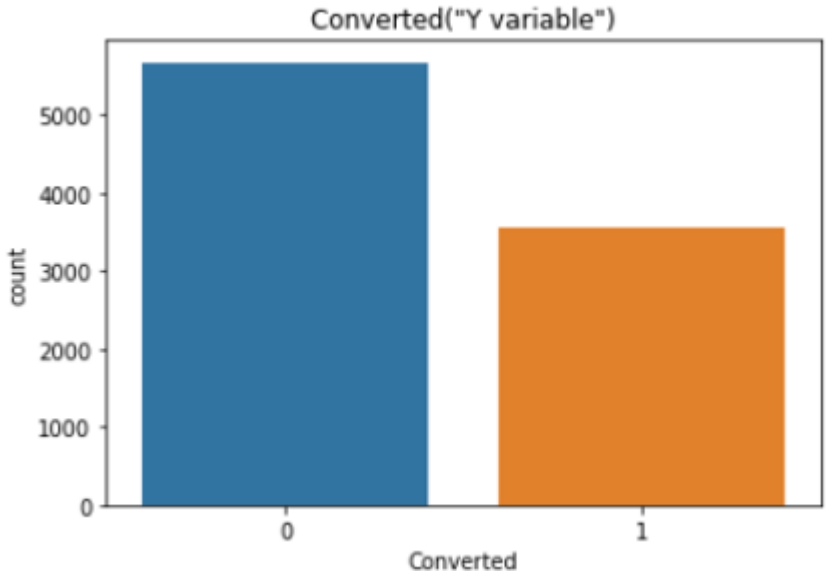
▶ Conclusions and Recommendations:
•Provided conclusions and actionable recommendations based on the analysis.

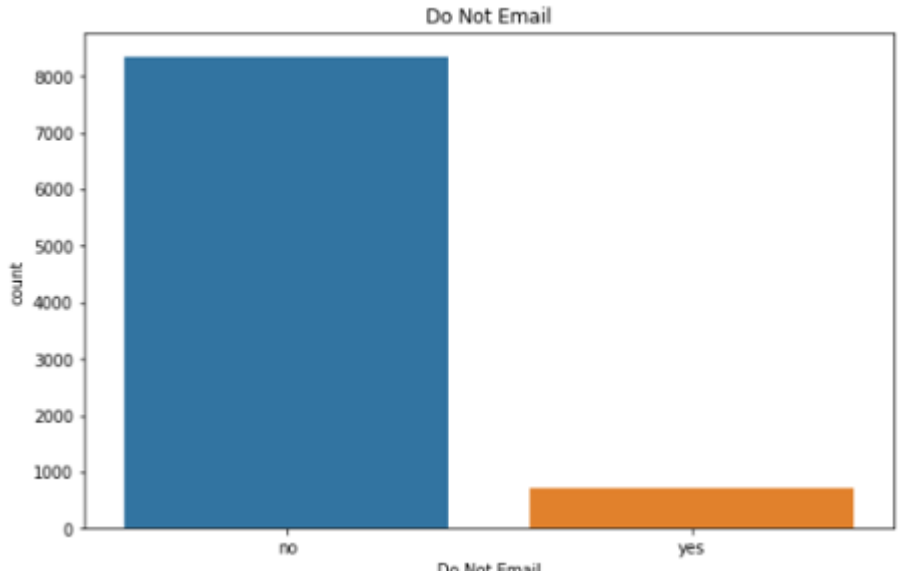
DATA MANIPULATION

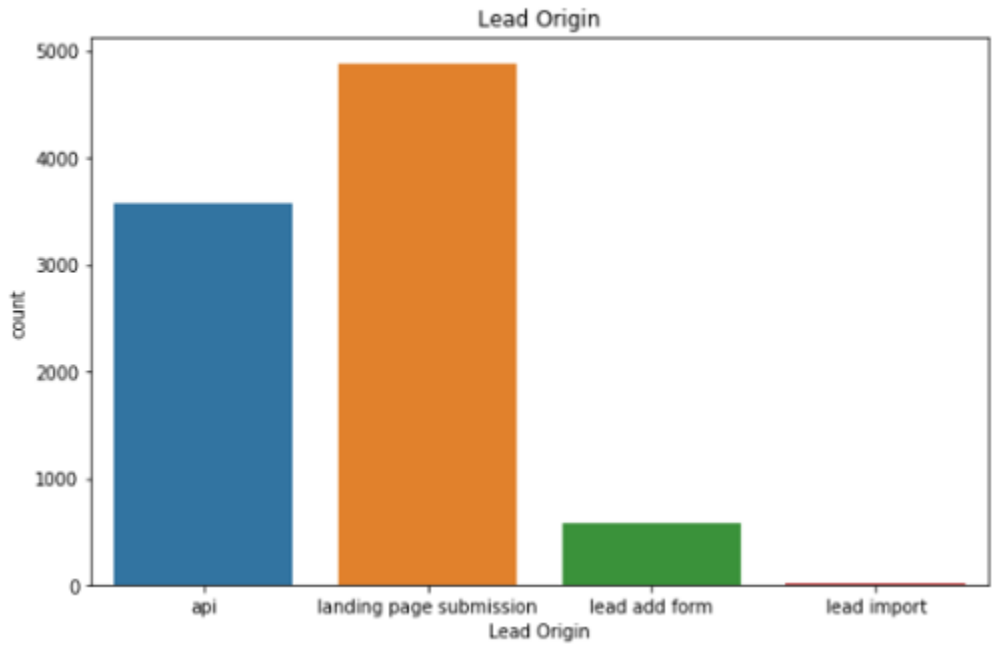
- The dataset contains 37 rows and 9240 columns.
- Features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply Chain Content", "Get updates on DM Content", and "I agree to pay the amount through cheque" have been excluded.
- Columns "Prospect ID" and "Lead Number" have been removed as they are unnecessary for the analysis.
- For some object-type variables, features with insufficient variance have been dropped. These include "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", and "Digital Advertisement".
- Columns with more than 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile', have also been dropped.

EXPLORATORY DATA ANALYSIS

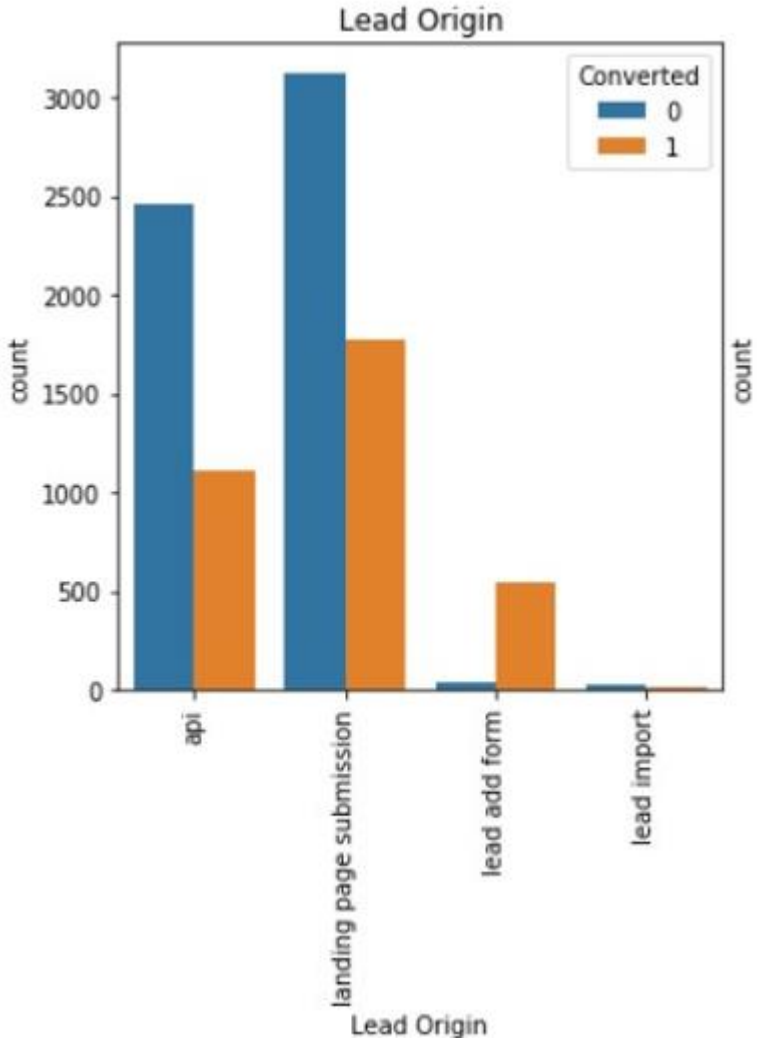


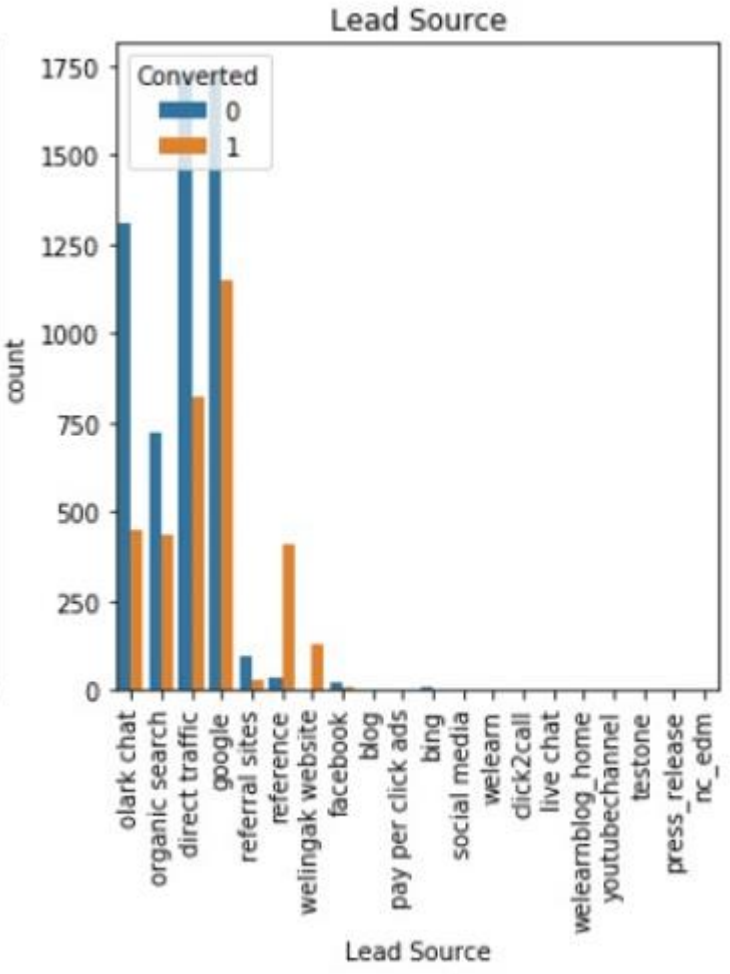

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

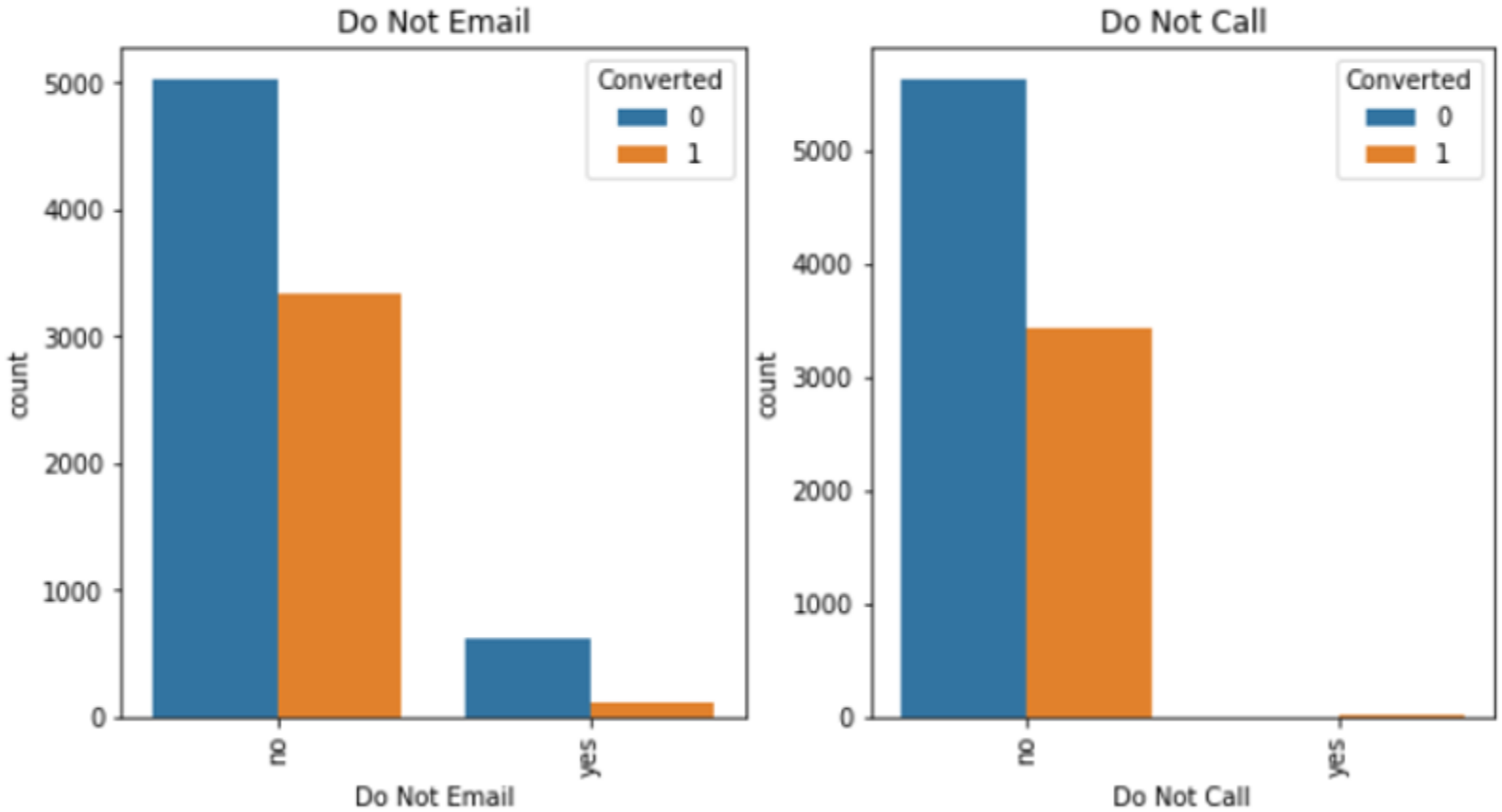

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes

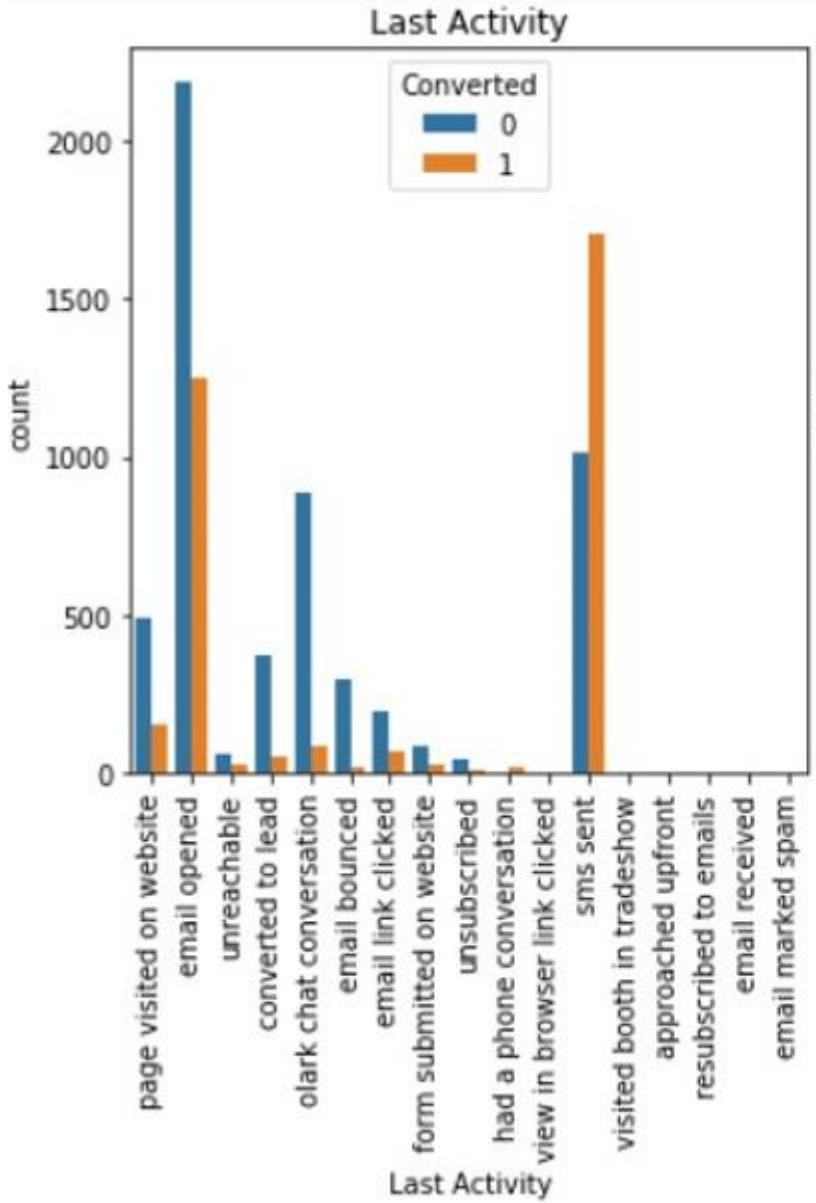

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

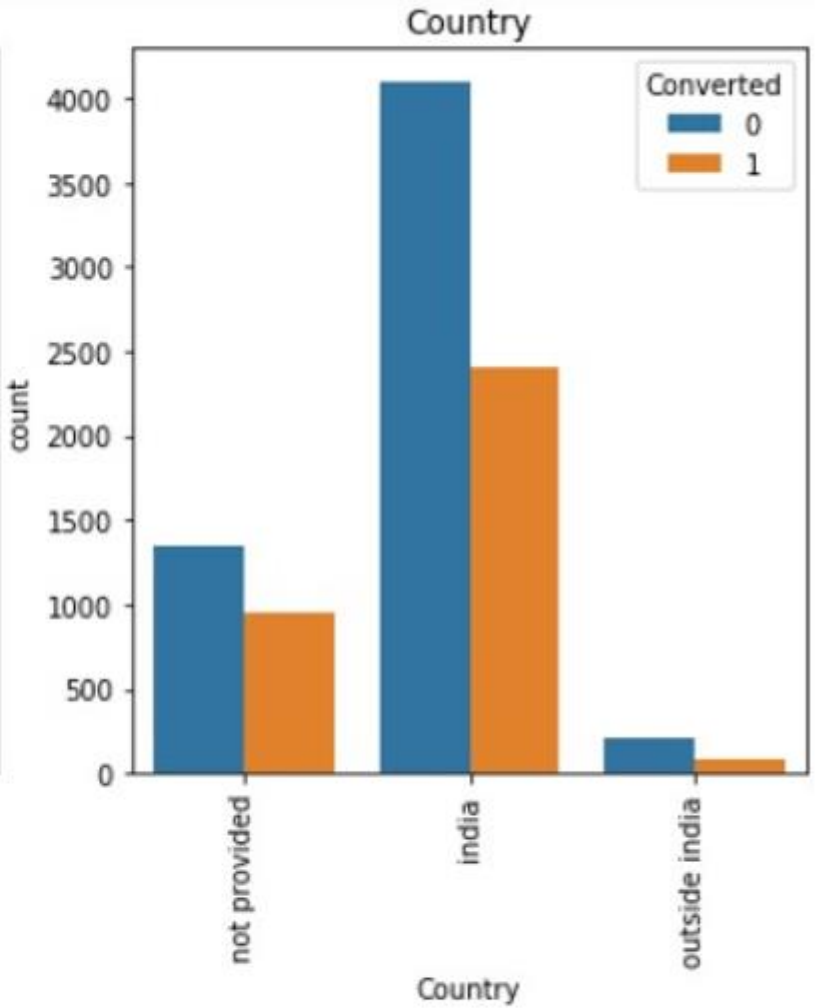
CATEGORICAL VARIABLE RELATION






Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email
Do Not Call
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Call


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity


Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

DATA CONVERSION

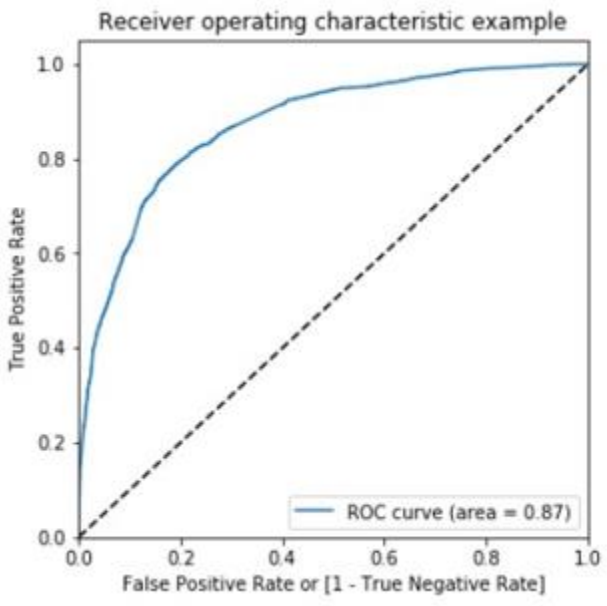
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

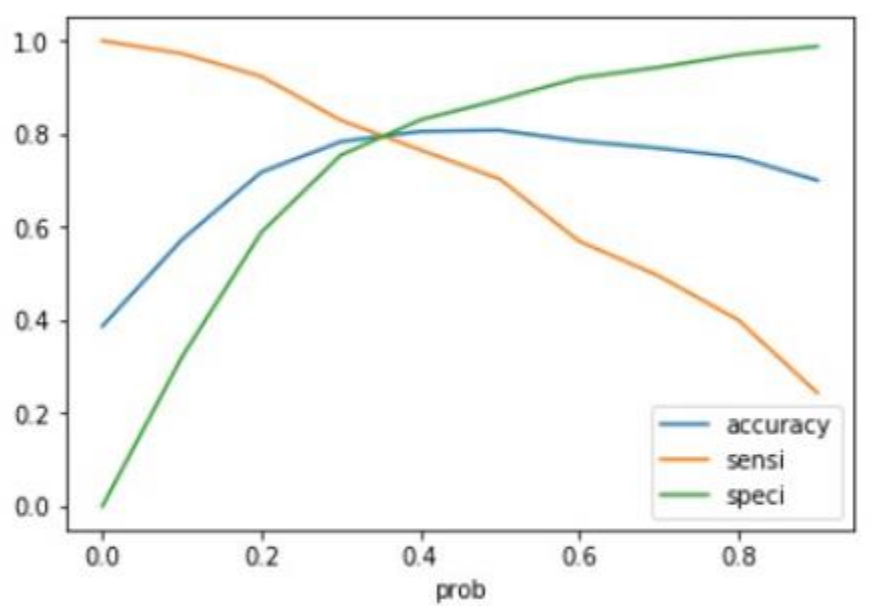
MODEL BUILDING

- 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

- **Finding Optimal Cut off Point**
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.





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 courses.