LEAD SCORE CASE STUDY

ABHISHEK PRADHAN
ADRISH RAY
ABHISHEK SIKERWAR

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 30of 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 Data Manipulation

i. First step to clean the dataset we chose was to drop the variables having unique values.

ii. Then, there were few columns with value ‘Select’ which means the leads did not choose any given option. We changed those values to Null values.

iii. We dropped the columns having NULL values greater than 30%.

iv. Next, we removed the imbalanced and redundant variables.

v. All sales team generated variables were removed to avoid any ambiguity in final solution.

Solution Methodology

- ➢ Data Transformation:

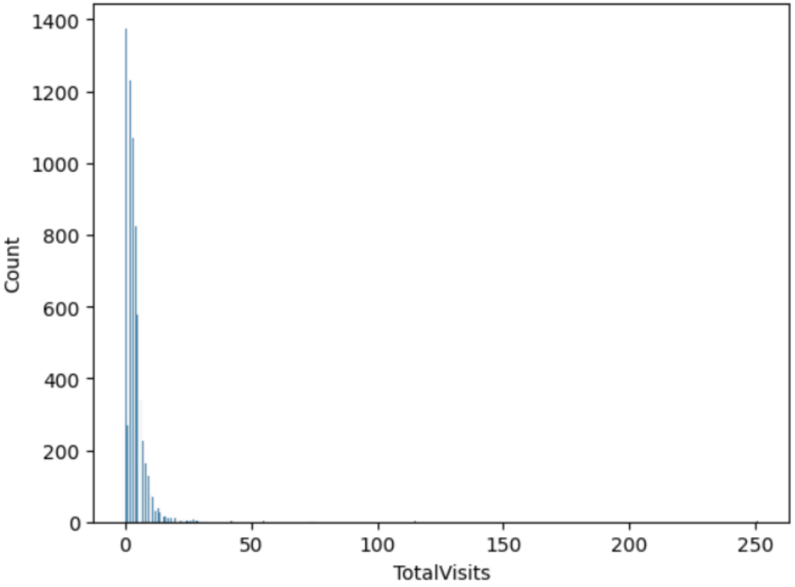
i. Changed the binary variables into ‘0’ and ‘1’.

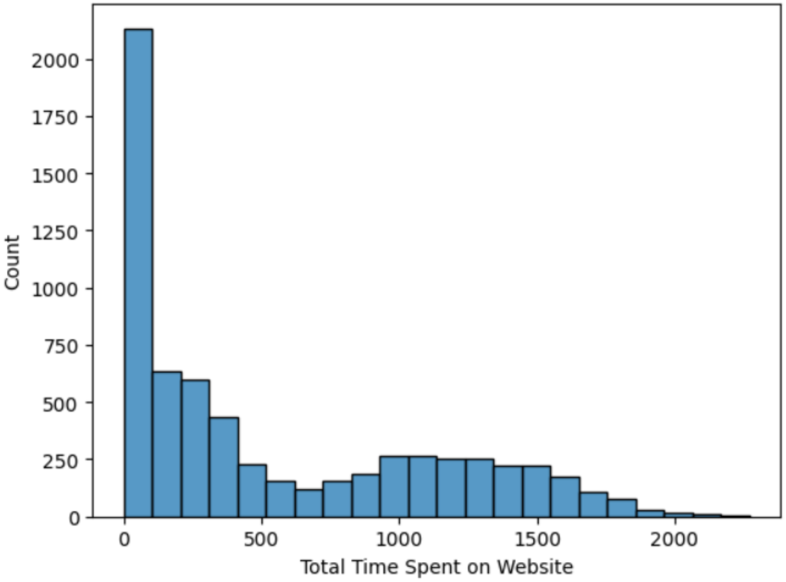
- ➢ Data Transformation:

i. We created dummy variables for the categorical variables.

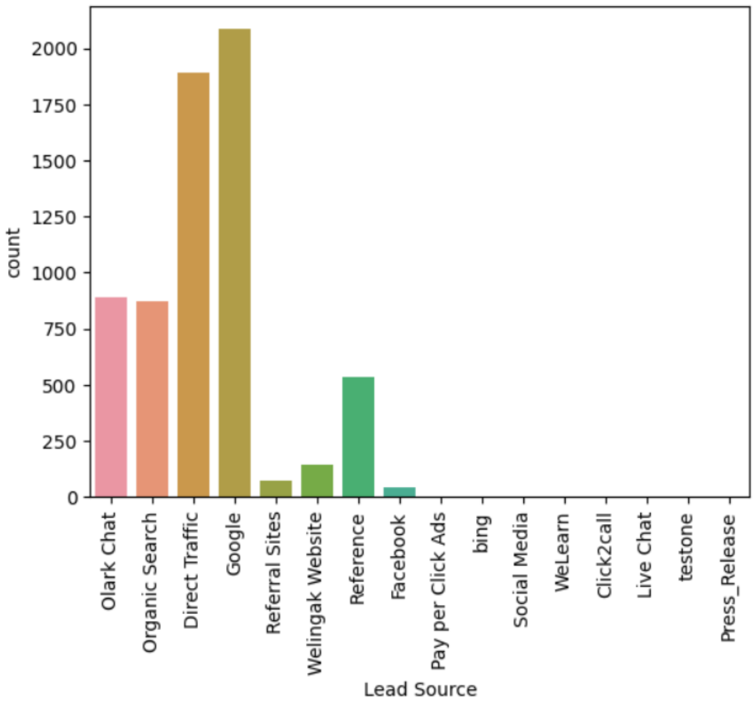
ii. Removed all the repeated and redundant variables

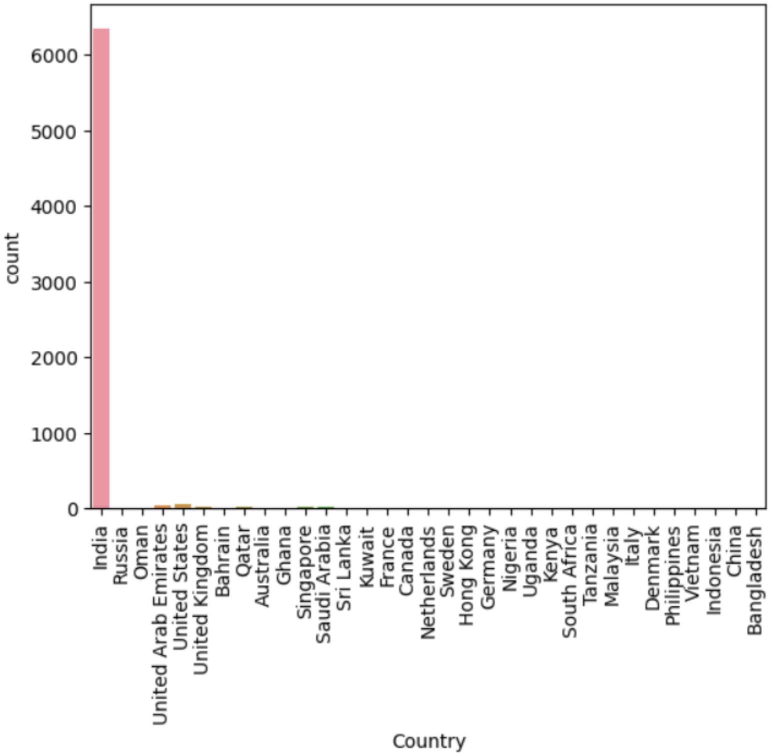
EDA – Univariate Analysis



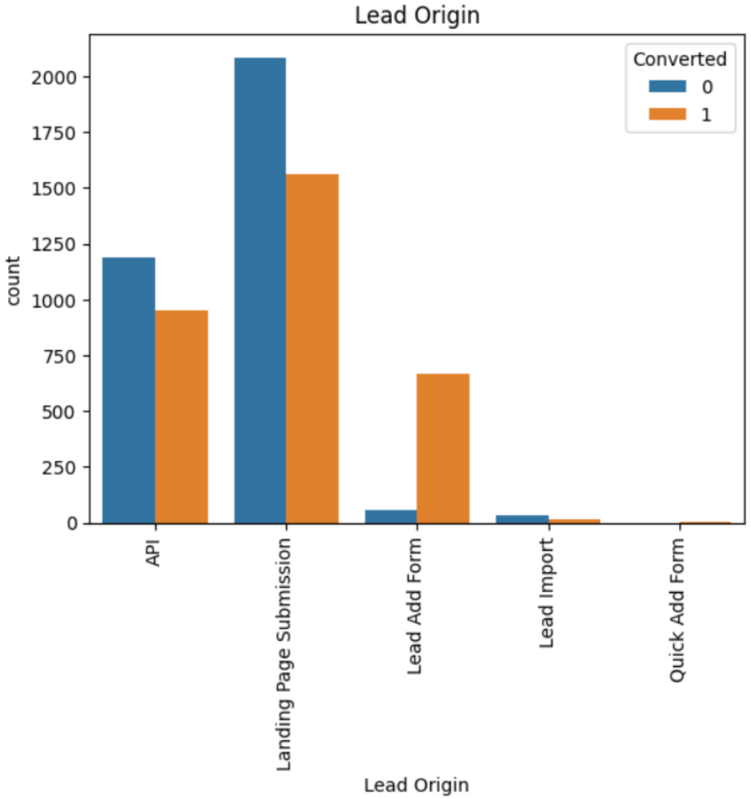


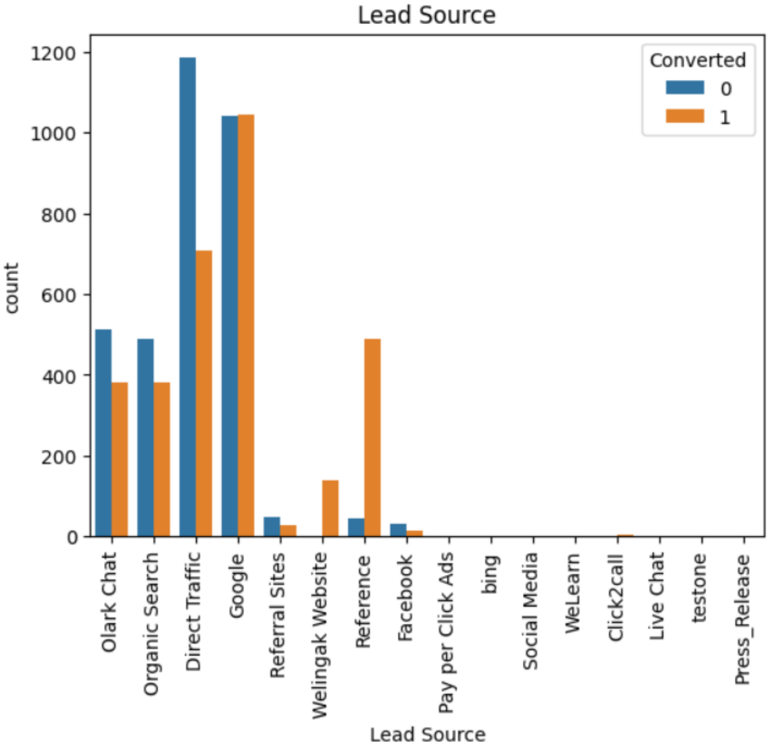
- **'Google' is the lead source for most of the users and Maximum users are from ‘India’.**



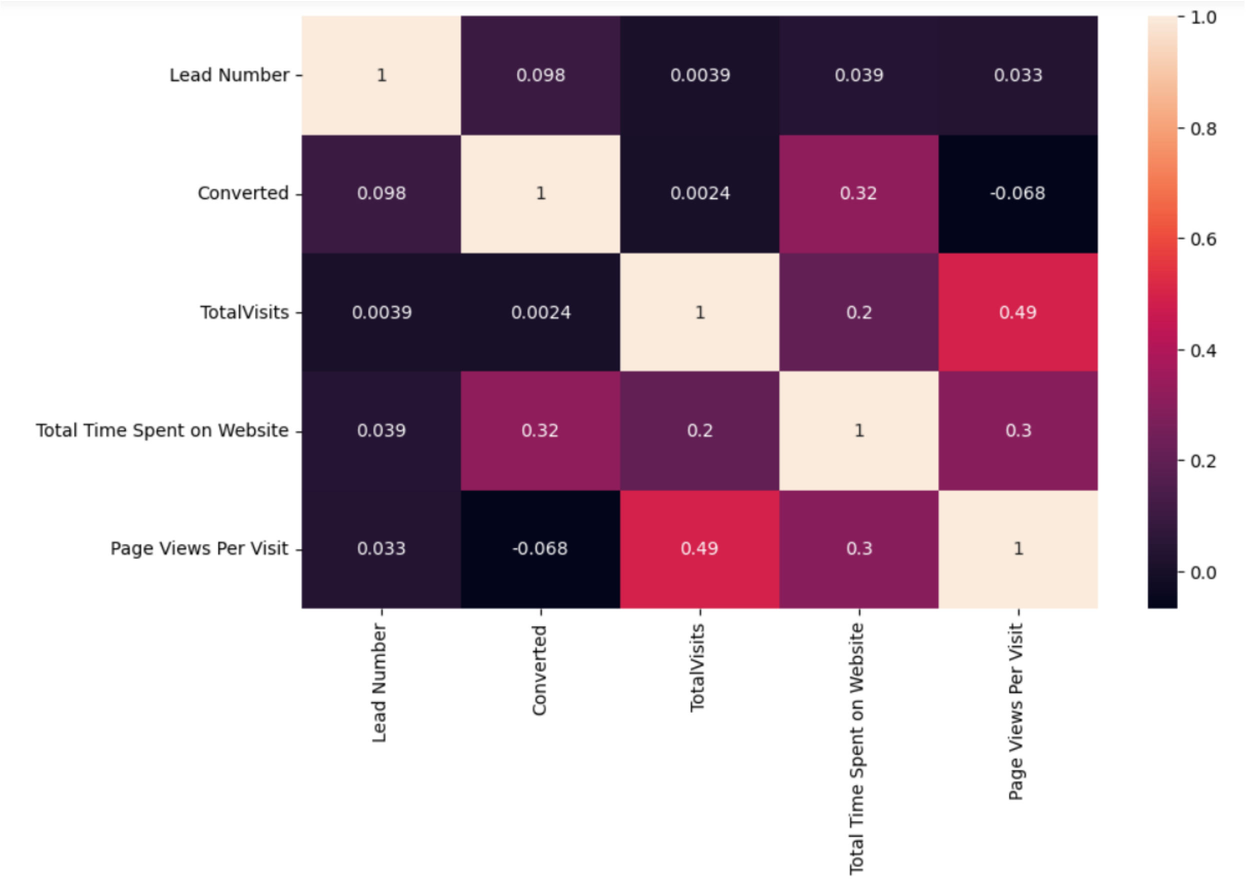


Bivariate Analysis





Multivariate Analysis



Model Building

➢ Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.

➢ Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

➢ Finally, we arrived at the 11 most significant variables. The VIF's for these variables were also found to be good.

➢ For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.

Model Building

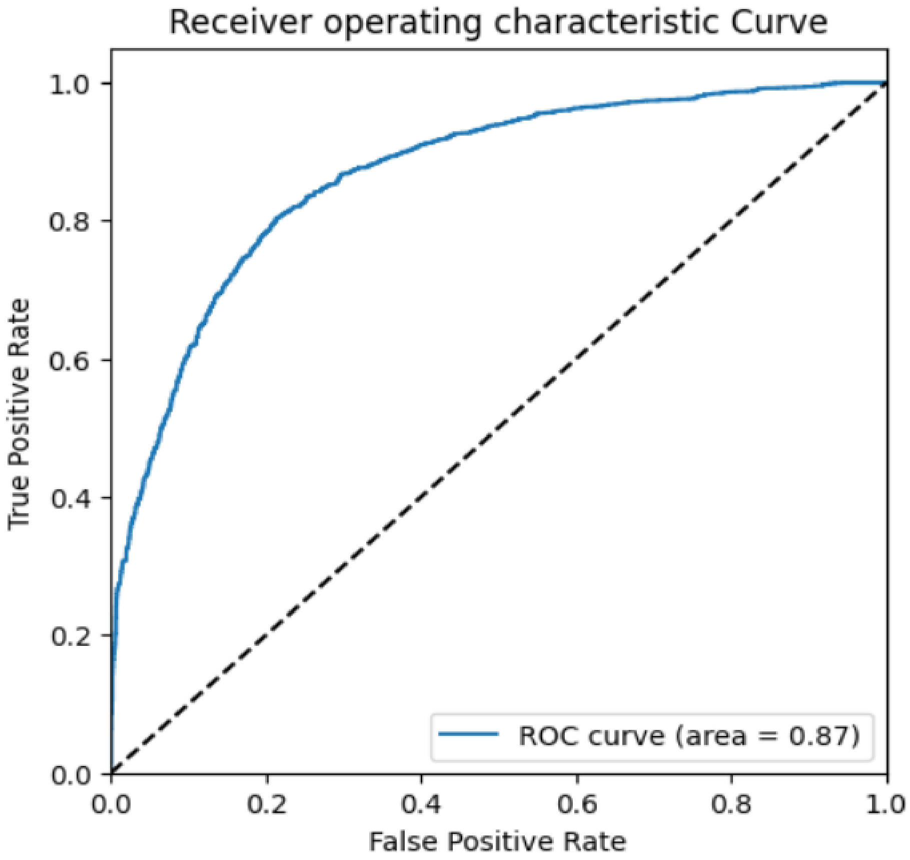
➢ We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 87% which further solidified the of the model.

➢ We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.

➢ Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.43.

➢ Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 78.6%; Recall= 77.13%; Precision= 79.64%.

ROC Curve


Receiver operating characteristic Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.87)

INSIGHTS & CONCLUSION

- Features which contribute more towards the probability of a lead getting converted are:
  1. Do not Email
  2. Total Visits
  3. Total Time Spent on Website

- Most of the users are 'Unemployed’

- 'Email opened' is the last activity for most of the users.

- Most of the users are choosing the course

- 'Modified' is the last notable activity for most of the users.