LEAD SCORE CASE STUDY

By,
Shashank Pandey & Ashish Bachuwar

Problem Statement

- An education company i.e., X Education sells online courses to industry professionals gets a lot of leads, but 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 i.e., only 30%
- 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.

Solution Methodology

- Data Cleaning and Manipulation:

  1. Check and handle duplicate data, **2.** Handling null and missing values, **3.** Drop columns with missing/null values > 45% **4.** Imputation remaining NA values, **5.** Check and handle outliers

- EDA:

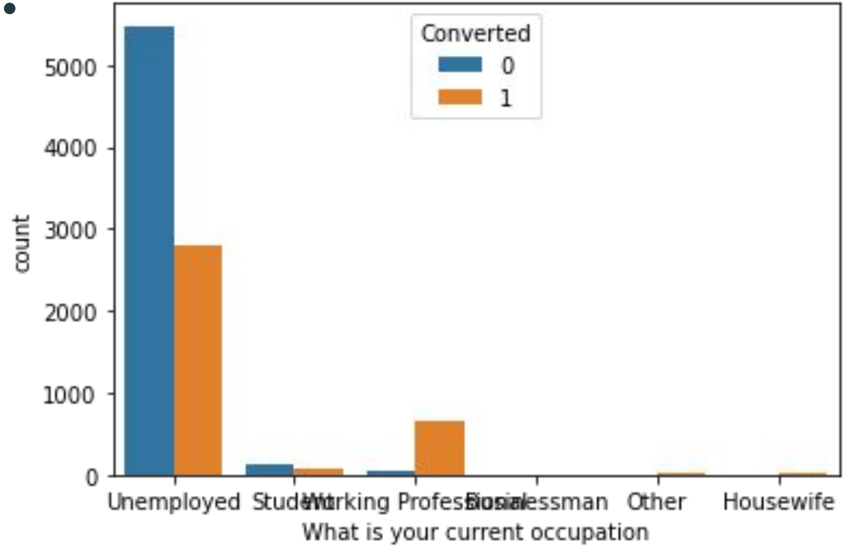
  1. Univariate data analysis on categorical variables: Count plot, **2.** Bivariate data analysis: correlation coefficients and pattern between the variables etc.

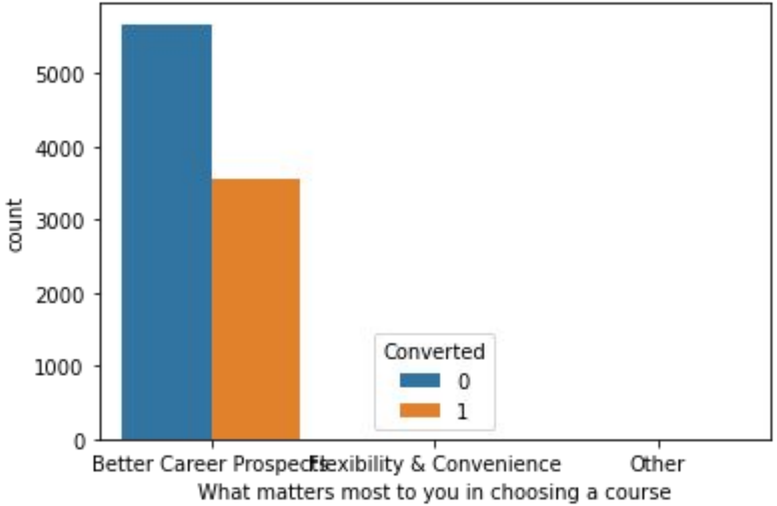
- Feature Scaling & Dummy Variables and encoding of the data
- Classification technique: logistic regression used for the model making and prediction
- Validation of the model
- Model presentation
- Conclusions and recommendations.

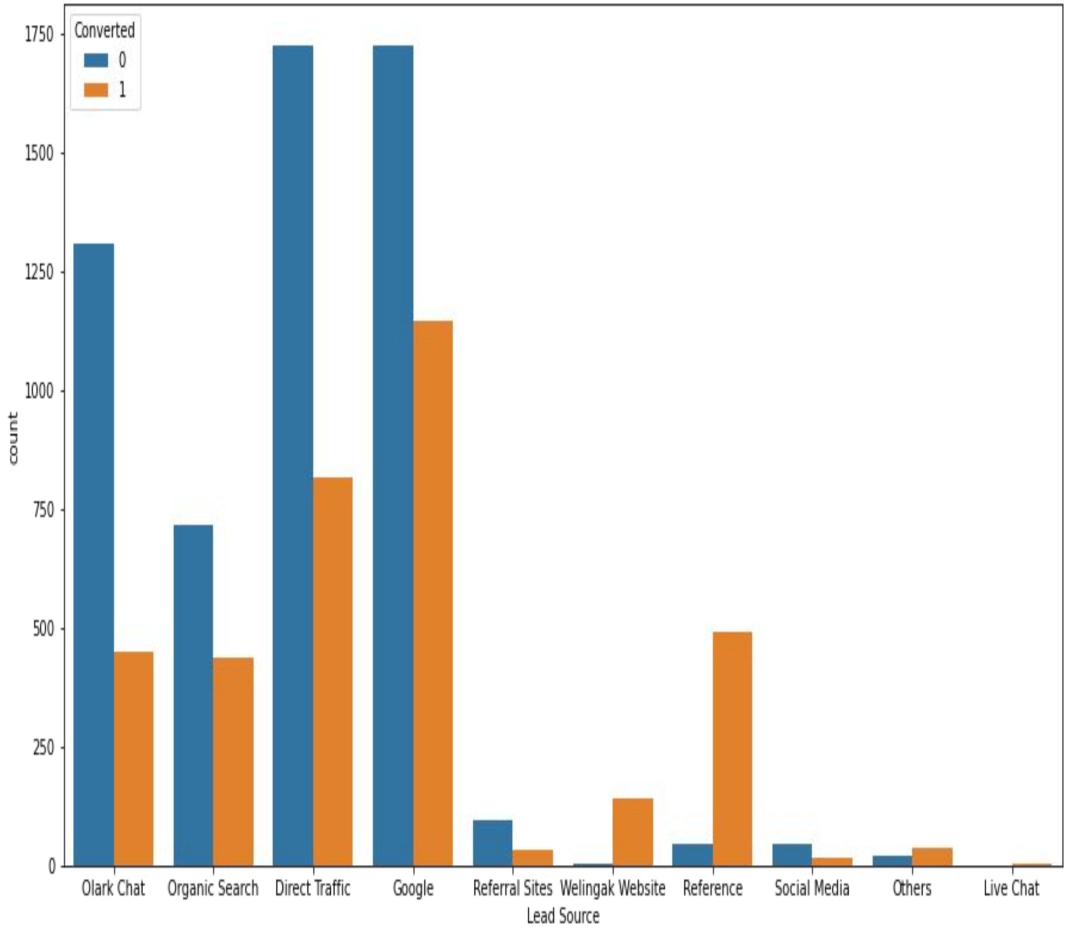
Data Manipulation

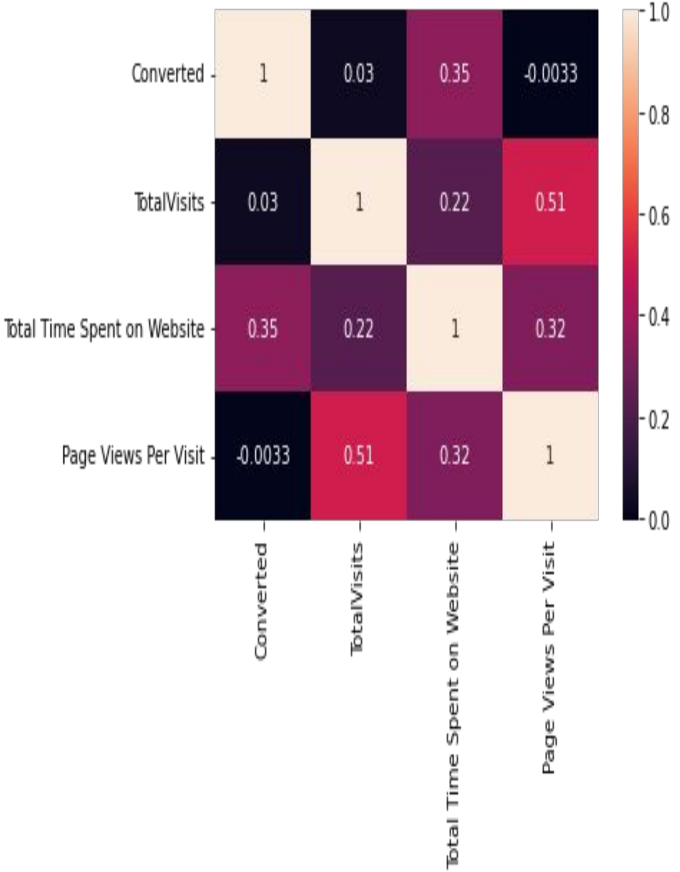
- Total Number of Rows =9240, Total Number of Columns =37
- Dropped ‘Prospect ID’ & ‘Lead Number’, since these columns were of no use
- Dropped all the columns which had null values > 45%
- Handled missing values from the other columns and imputed it with the most repeated one i.e., mode or not specified based on the column
- Created/Combined columns wherever required
- Created dummy variables for the categorical variables i.e., for binary & multi-level

EDA


Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


count
5000
4000
3000
2000
1000
0
Converted
0
1
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course


Converted
0
1
1750
1500
1250
1000
750
500
250
0
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Social Media
Others
Live Chat
Lead Source


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.03
0.35
-0.0033
0.03
1
0.22
0.51
0.35
0.22
1
0.32
-0.0033
0.51
0.32
1
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

Model Building

- Split the data into train & test datasets of 70:30
- Used RFE for Feature Selection i.e., top 15 features to build the model
- Built the model by removing the features whose p- value was > 0.05 and VIF value > 5

Training Dataset Metrics:

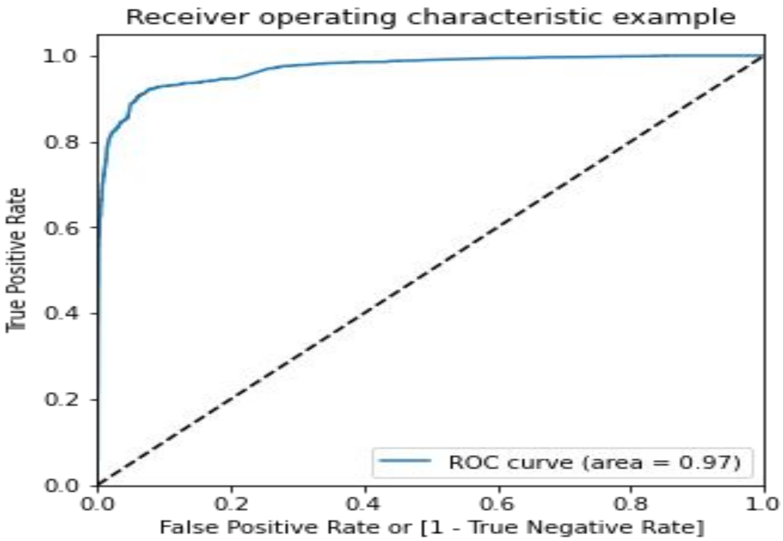
Accuracy : 92.27% Precision : 88.55% Recall : 91.48% Sensitivity : 91.48% Specificity : 92.76%

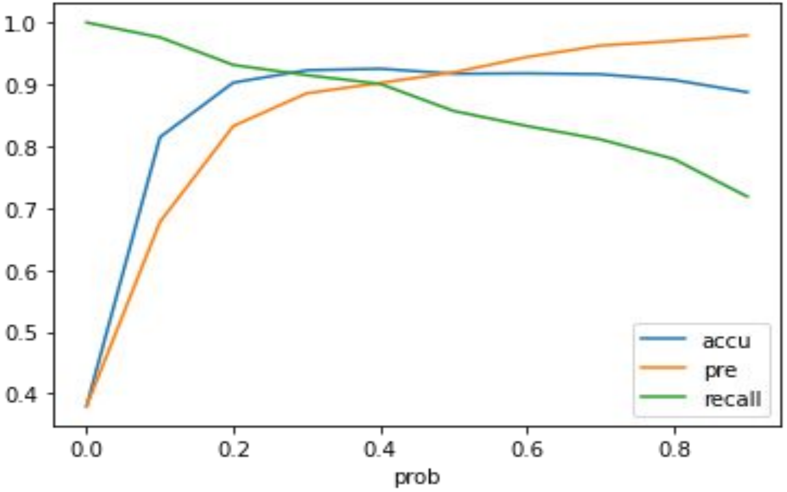
Test Dataset Metrics:

Accuracy : 92.82% Precision : 89.23% Recall : 92.32% Sensitivity : 92.32% Specificity : 93.13%

ROC Curve & Cutoff Probability

- The ROC curve is at 0.97 which is very close to 1
- The cutoff probability from the right side figure can be considered as 0.3, since that is the optimum point





Recommendations

The top features/variables for the company to consider for more lead conversion are :

- The total time spent on Website
- **When the lead source was:**

a. Direct traffic **b.** Welingak website

- Lead Origin_Lead Add Form
- Total visits to the website
- **Tags with:**

a. Will revert after reading the email **b.** Other_Tags **c.** Closed by Horizzon **d.** Ringing **e.** Interested in other courses **f.** Lost to EINS

- **When the last activity was:**

a. SMS **b.** Olark chat conversation

Keeping these in mind, the X Education company can have more leads conversion rate, as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses