Lead Scoring Case Study

Submitted By

1. Zeeshan Alam
2. Yatin Sharma
3. Yogeesh N

Agenda

1. Problem statement and Objective

02. Problem Approach

03. Exploratory Data Analysis

04. Correlations

05. Model Evaluation

06. Observation and Conclusion

Problem Statement

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 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. A typical lead conversion process can be represented using the following funnel:

Business Objective

- ✓ Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- ✓ The CEO want to achieve a lead conversion rate of 80%.
- ✓ They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

Problem Approach

- ✓ Importing the data and inspecting the data frame
- ✓ Data preparation
- ✓ EDA
- ✓ Dummy variable creation
- ✓ Test-Train split
- ✓ Feature scaling
- ✓ Correlations
- ✓ Model Building (RFE Rsquared VIF and p- values)
- ✓ Model Evaluation
- ✓ Making predictions on test set

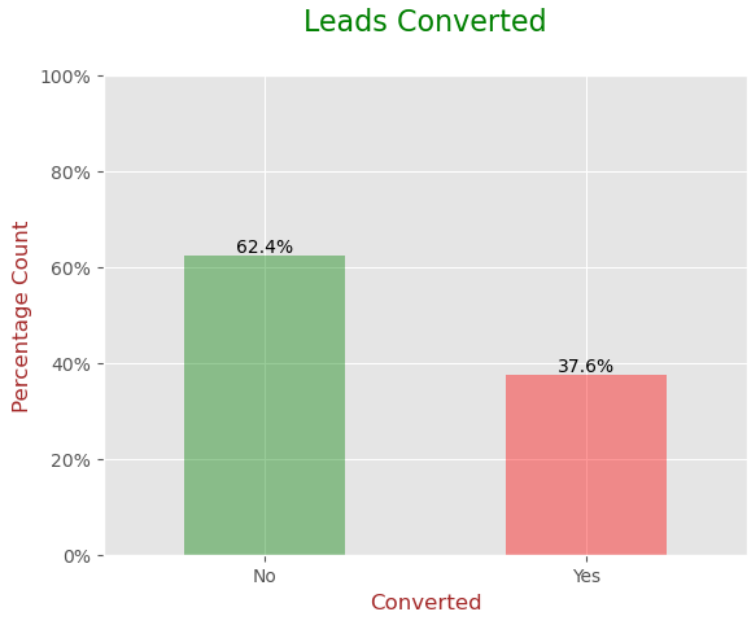
Data Manipulation

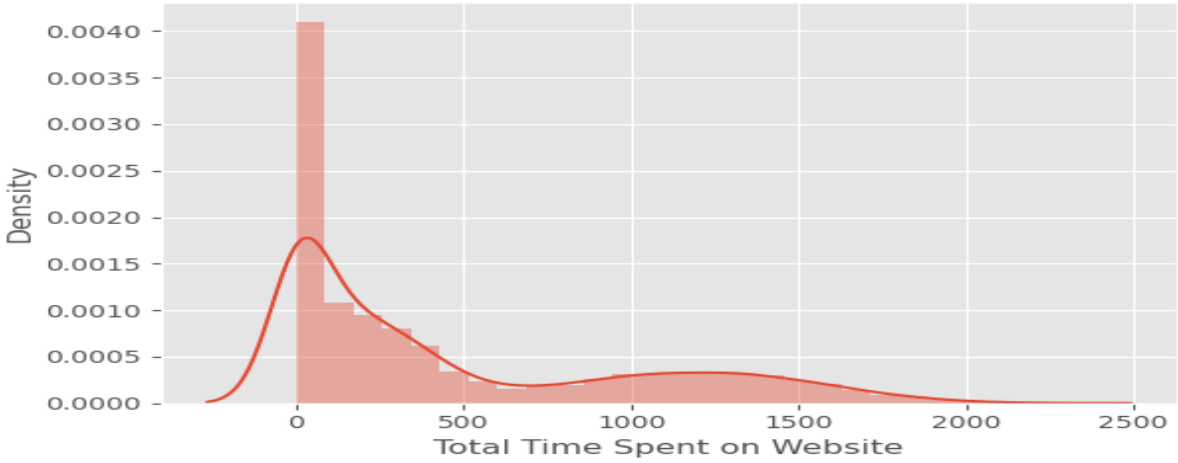
- ✓ Total Number of Rows =37, Total Number of Columns =9240.
- ✓ Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply” ,Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped. Removed the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- ✓ 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.
- ✓ Dropping the columns having more than 30% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.
- ✓ Clubbed certain features like Email, Country to reduce the excess number of features.

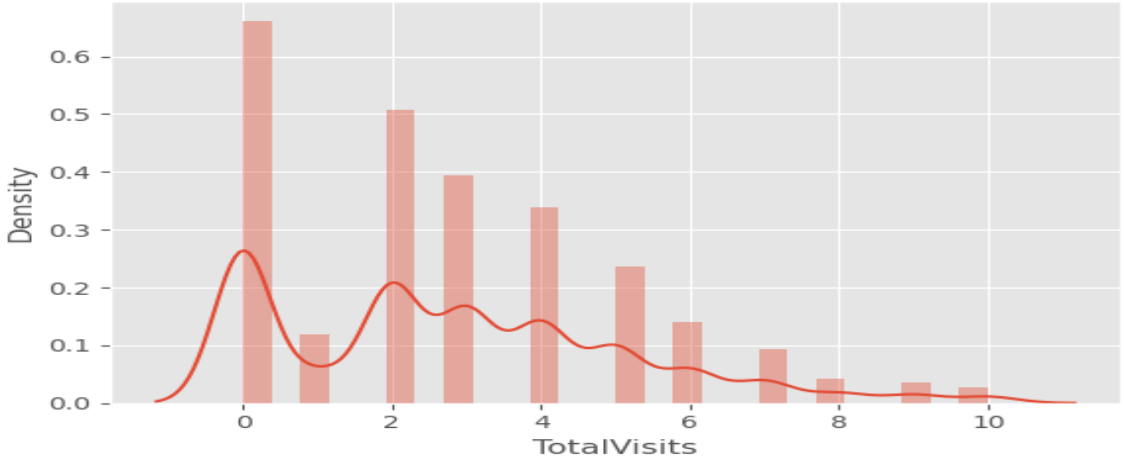
Exploratory Data Analysis (EDA)

Checking if Data is Imbalanced or not

- Data is imbalance when one value is present in majority and other is in minority meaning an uneven distribution of observations in dataset
- Data imbalance is in the context of Target variable only
- Target variable` is `'Converted'` which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted

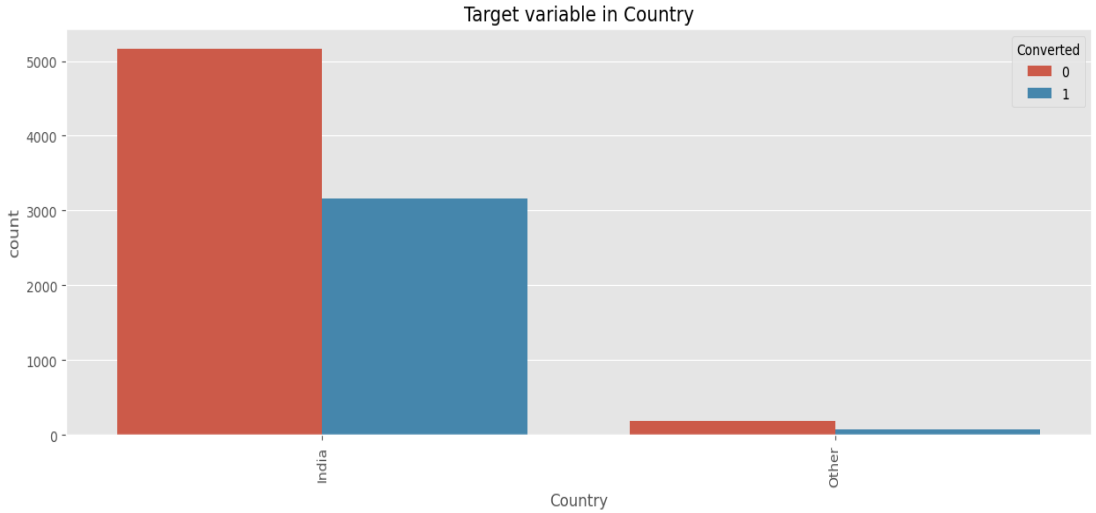


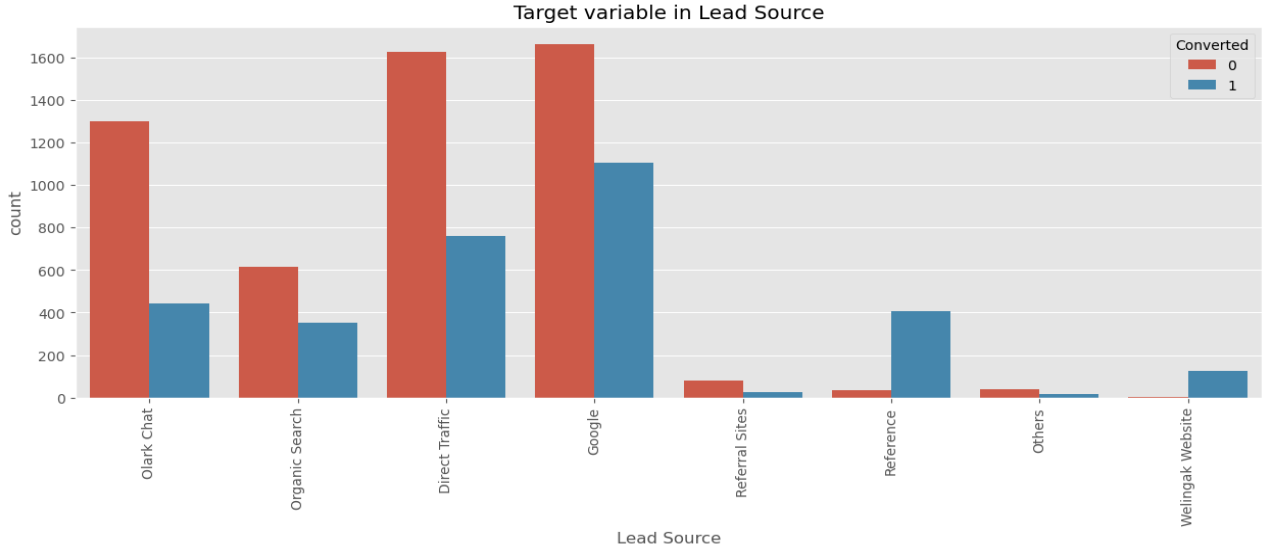


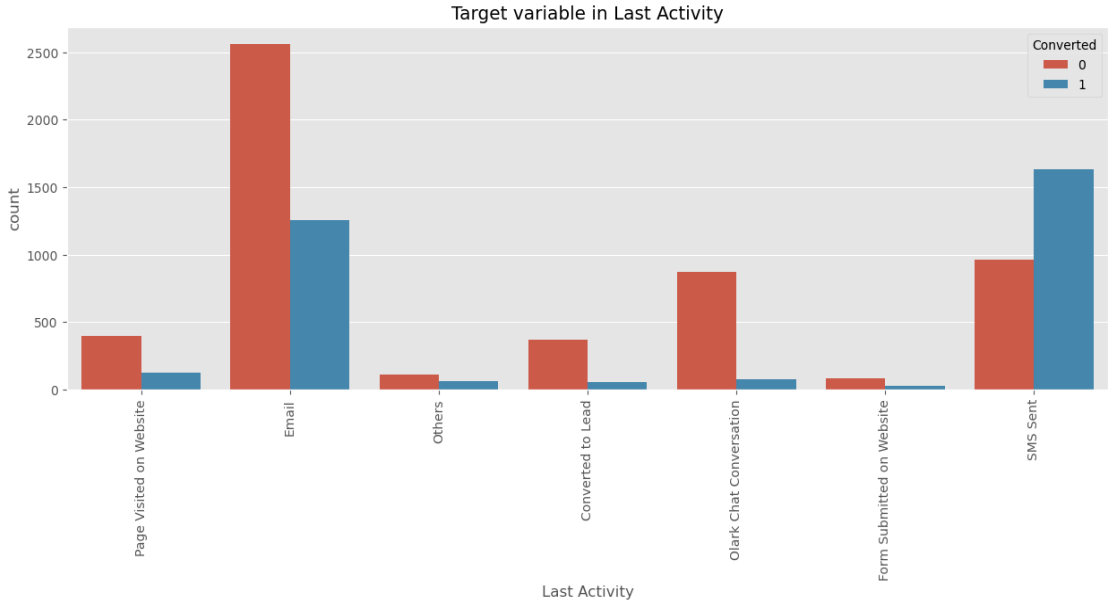


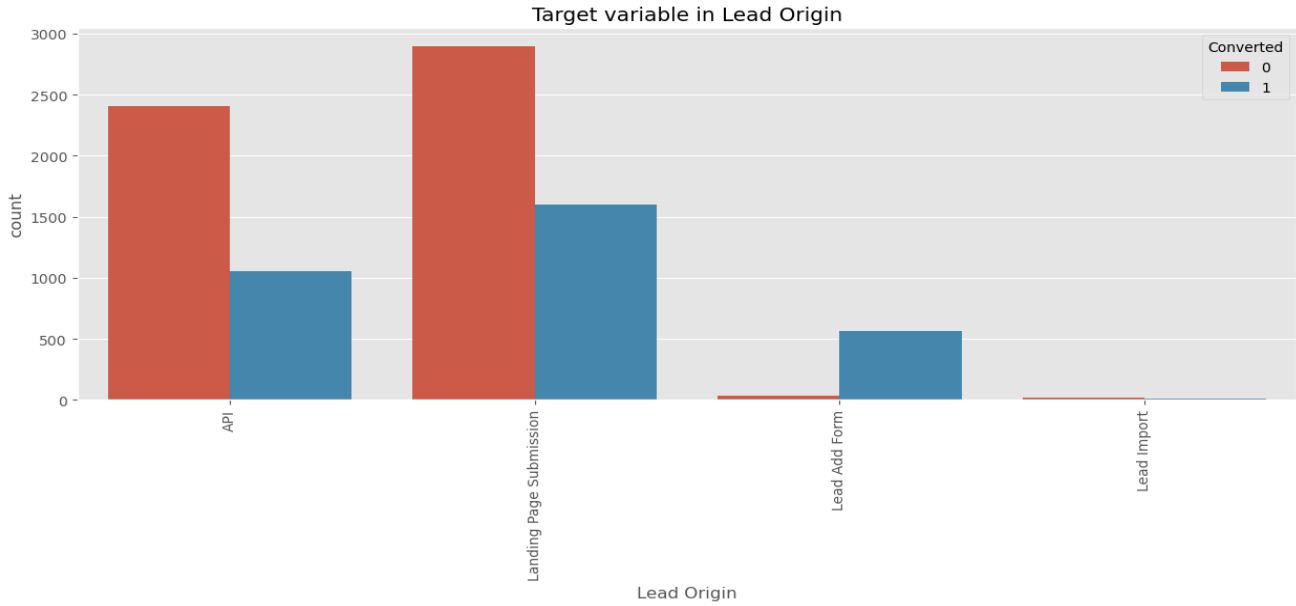
"The graph indicates distribution of Leads based on 'Total Visits,' and 'Total Time Spent on Website ' Higher values in these metrics seem to be positively associated with increased lead conversion rates, suggesting that prospects who visit the website more frequently, spend more time on the site, and view more pages are more likely to convert into leads."

EDA

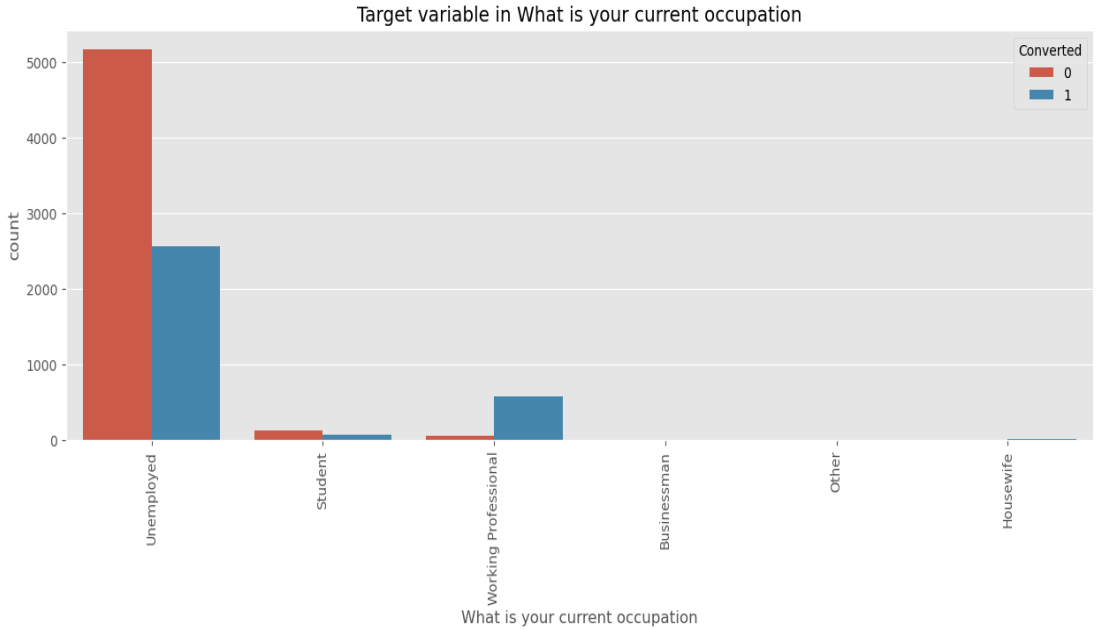

Target variable in Country
Converted
0
1
count
5000
4000
3000
2000
1000
0
India
Other
Country

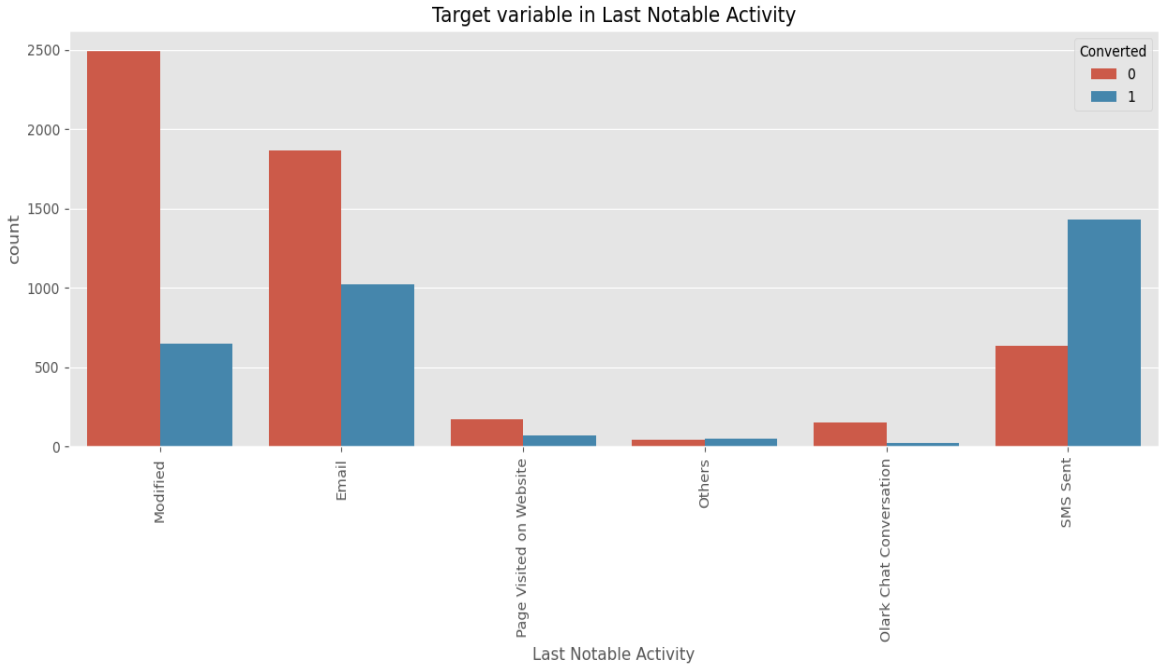

Target variable in Lead Source
Converted
0
1
count
1600
1400
1200
1000
800
600
400
200
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Others
Welingak Website
Lead Source


Target variable in Last Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
Page Visited on Website
Email
Others
Converted to Lead
Olark Chat Conversation
Form Submitted on Website
SMS Sent
Last Activity


Target variable in Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

EDA





Observations from EDA Process -

1.Maximum lead conversion happened from Landing Page Submission followed by API.

2.Major lead conversion in the lead source is from 'Google' followed by direct traffic

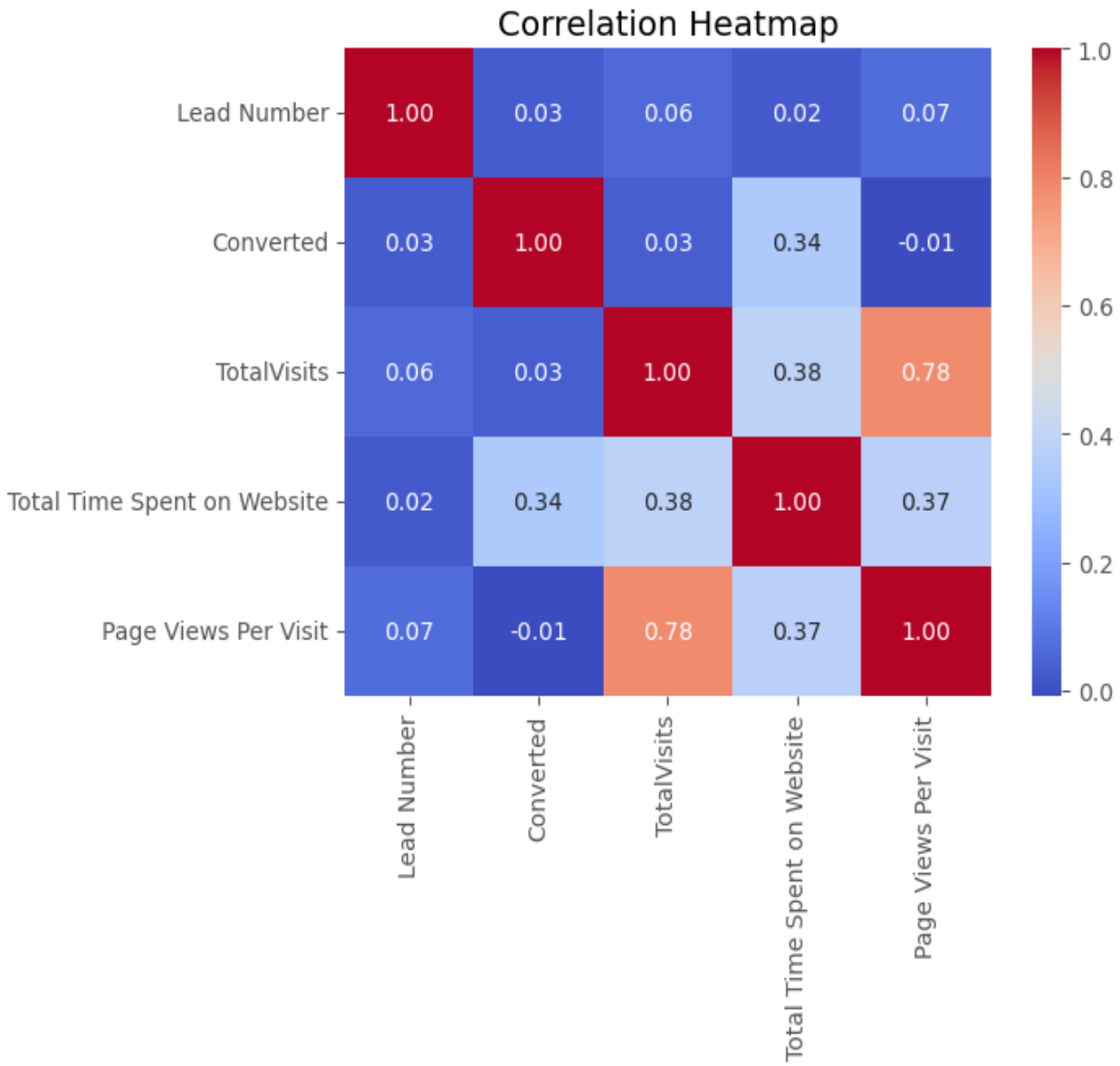
3.Major lead conversion is from the Unemployed Group for Better Career Prospects

4.Major lead conversion from Total Visits, Total Time Spent on Website, Page Views Per Visit

5.Major conversion has happened when the last activity is SMS sent and Opened the Email

6.Working Professionals have high rate of Conversion

Correlation Matrix



1. TotalVisits and Page Views Per Visit:
There is a moderate positive correlation of approximately 0.78 between TotalVisits and Page Views Per Visit. This suggests that, on average, as the total number of visits increases, there tends to be a higher number of pages viewed per visit.

2.Total Time Spent on Website and Converted:
There is a moderate positive correlation of approximately 0.34 between Total Time Spent on Website and the likelihood of conversion (Converted). This implies that prospects who spend more time on the website are more likely to convert.

3.TotalVisits and Converted:
The correlation between TotalVisits and Converted is relatively low (0.03). There is a weak positive correlation, suggesting that there may be a slight tendency for leads with more visits to have a slightly higher likelihood of conversion.

4. Page Views Per Visit and Converted:
The correlation between Page Views Per Visit and Converted is slightly negative (-0.008). This weak negative correlation suggests that, on average, a higher number of page views per visit is not strongly indicative of conversion.

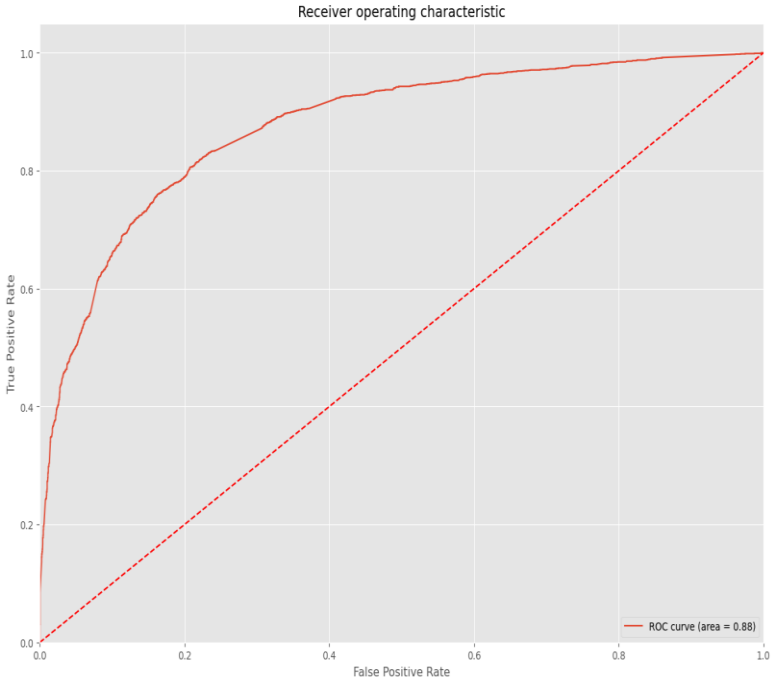
Model Building

- 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%

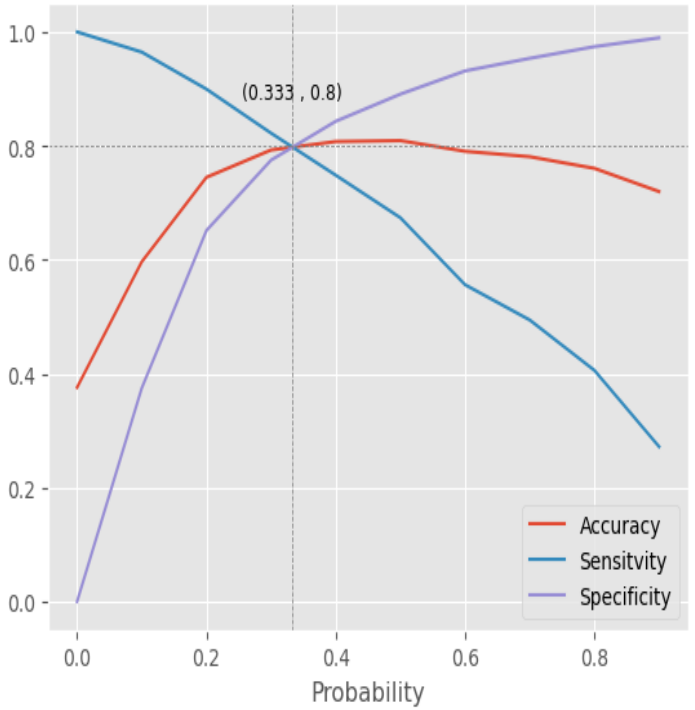
Model Evaluation

0.41 is the tradeoff between Precision and Recall -

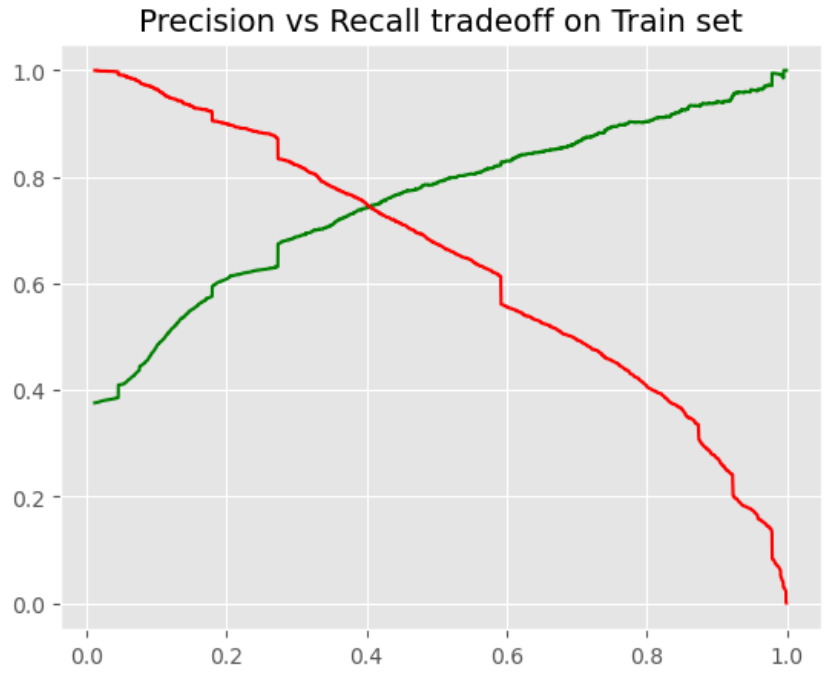
Thus we can safely choose to consider any Prospect Lead with Conversion **Probability higher than 41 % to be a hot Lead**



Area under ROC curve is 0.88 out of 1 which indicates a good predictive model



0.333 is the approx. point where all the curves meet, so 0.333 seems to be our `Optimal cutoff point` for probability threshold .



The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimize the performance of the model based on business requirement, Here our probability threshold is 0.41 aprrox from above curve.

Observations

Lead Origin_Lead Add Form	4.53012
Total Time Spent on Website	4.422485
What is your current occupation_Working Professional	2.753437
Last Notable Activity_SMS Sent	1.34821
Lead Source_Olark Chat	1.077277
Lead Origin_Landing Page Submission	-0.29549
Last Activity_Page Visited on Website	-0.37515
Last Notable Activity_Modified	-0.54183
Last Activity_Olark Chat Conversation	-1.52481
const	-2.05541

A high positive coefficient indicates that a variable has a stronger influence on predicting the probability of leads converting to take up X-Education's course.

For Train Set

```
********************************
********************

Confusion Matrix
[[2989  766]
 [ 462 1800]]

********************************
********************

True Negative          : 2989
True Positive          : 1800
False Negative         : 462
False Positve          : 766
Model Accuracy         : 0.7959
Model Sensitivity      : 0.7958
Model Specificity      : 0.796
Model Precision        : 0.7015
Model Recall           : 0.7958
Model True Positive Rate (TPR)  : 0.7958
Model False Positive Rate (FPR) : 0.204

********************************
********************
```

For Test Set

```
************************************
************

Confusion Matrix
[[1382  225]
 [ 264  709]]

************************************
************

True Negative          : 1382
True Positive          : 709
False Negative         : 264
False Positve          : 225
Model Accuracy         : 0.8105
Model Sensitivity      : 0.7287
Model Specificity      : 0.86
Model Precision        : 0.7591
Model Recall           : 0.7287
Model True Positive Rate (TPR)  : 0.7287
Model False Positive Rate (FPR) : 0.14

************************************
************
```

The evaluation matrices are pretty close to each other so it indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.

Conclusion

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

1.Lead Origin_Lead Add Form (4.530120):
•The high positive score suggests that this particular lead origin has a strong positive impact on predicting lead conversion. Leads from this source are more likely to convert.

2.Total Time Spent on Website (4.422485):
•The high positive score indicates that the more time a user spends on the website, the more likely they are to convert. It suggests that user engagement on the website is a significant predictor of lead conversion.

3.What is your current occupation_Working Professional (2.753437):
•The positive score indicates that leads with a current occupation of "Working Professional" are more likely to convert compared to other occupations.

4.Last Notable Activity_SMS Sent (1.348210):
•The positive score suggests that sending an SMS as the last notable activity is positively correlated with lead conversion. Leads who received an SMS as the last notable activity are more likely to convert.

5. Lead Source_Olark Chat (1.077277):
•The positive score indicates that leads coming from Olark Chat are more likely to convert compared to other lead sources.

6.Lead Source: Certain lead sources significantly contribute to conversions, particularly:

1. Google
2. Direct traffic
3. Organic search
4. Visits through the Welingak website

Conclusion

6. Last Activity: The nature of the last activity performed by the potential buyer plays a crucial role, especially activities like:

1. SMS interactions
2. Olark chat conversations

7. Current Occupation: Individuals identified as working professionals are more likely to convert into buyers.

8. What is your current occupation_Working Professional: working professionals are more likely to convert into buyers.

Keeping these in mind, X Education can flourish as they have a very high chance to persuade potential buyers to change their mind and buy their courses.

Thank you