LEAD SCORING CASE STUDY

Submitted By :

Sakshee Suryawanshi
Sainadh Channamsetty
Sakshi Srivastav

Lead Scoring Case Study

Introduction:

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Business Objective:

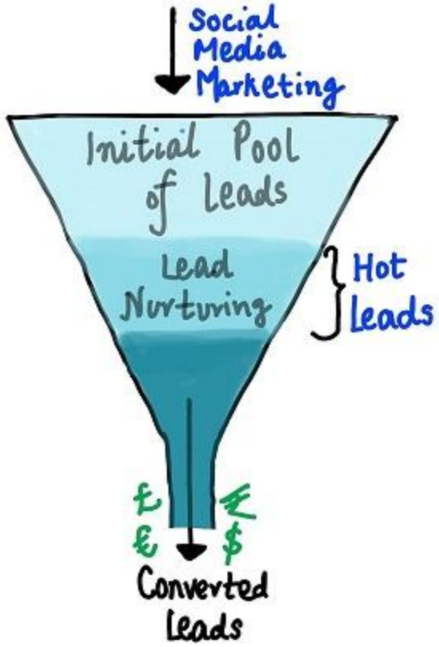
- X Education has to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.
- The goal is to build a Logistic Regression based model for X Education company to help in selecting the most promising leads to figure out the ‘Host Leads’ of company having a higher conversion chance.

Problem Statement

- 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.

- A typical lead conversion process can be represented using the following funnel:



Solution Steps :

1. Understanding and Cleaning Data

2. Exploratory data Analysis

3. Data Preparation : Create Dummy Variables

4. Splitting Data into Train-Test set

5. Model Building

6. Model Evaluation

7. Prediction on Train and Test datasets

8. Prediction using Precision-Recall methods

Understanding and Cleaning Data

- There were few variables with missing/null values which were needed to be take care by dropping unwanted columns with null values having missing values percentage higher than 35% and some with unique values.

- Some categorical variables like “**Specialization**”, ”**How did you hear about X Education**”, “**City**” and “**Lead Profile**” had level called 'Select' which was as good as a null value, therefore, to not lose data they were changed as ‘no data’ value.

- Although, later on all variables with level ‘Select’ were either distributed proportionally or were removed while creating dummies. No major outliers were found.

- Handling the categorical columns with a smaller number of missing values:

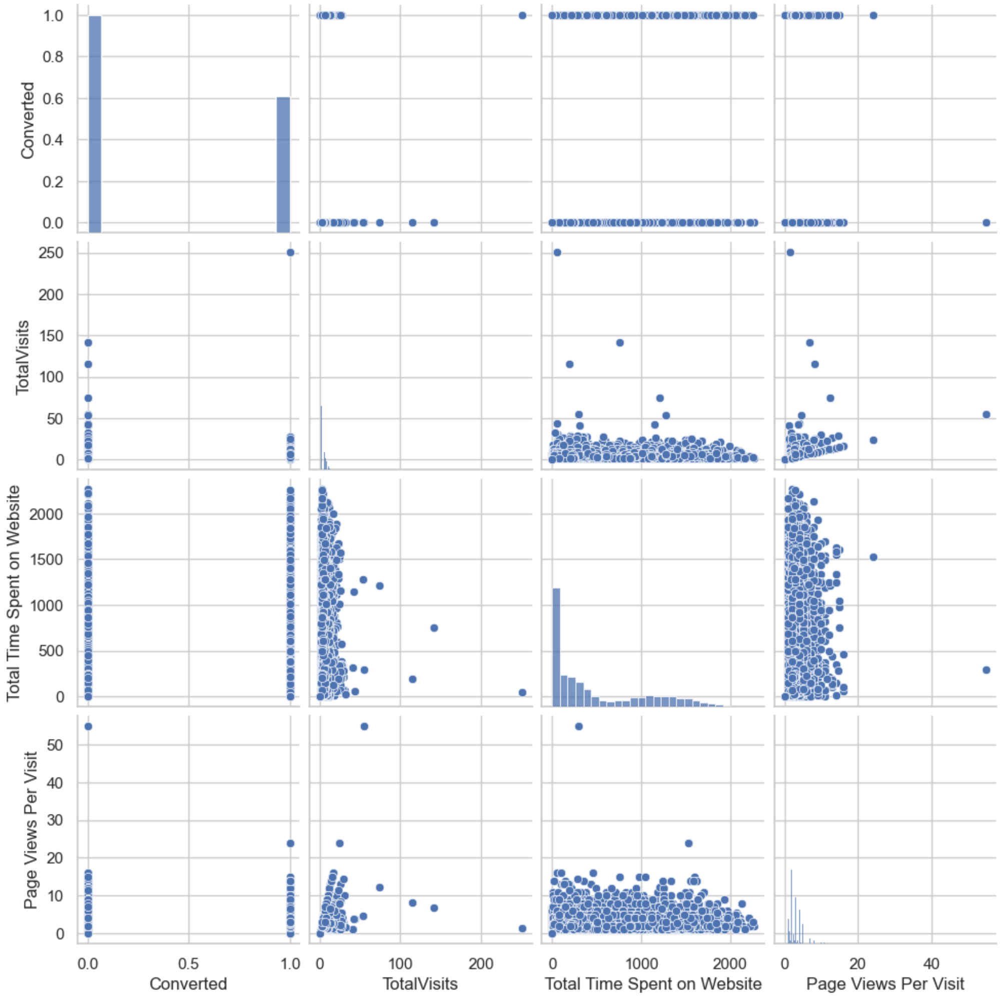
  - Merge categories that have low representation of categories.

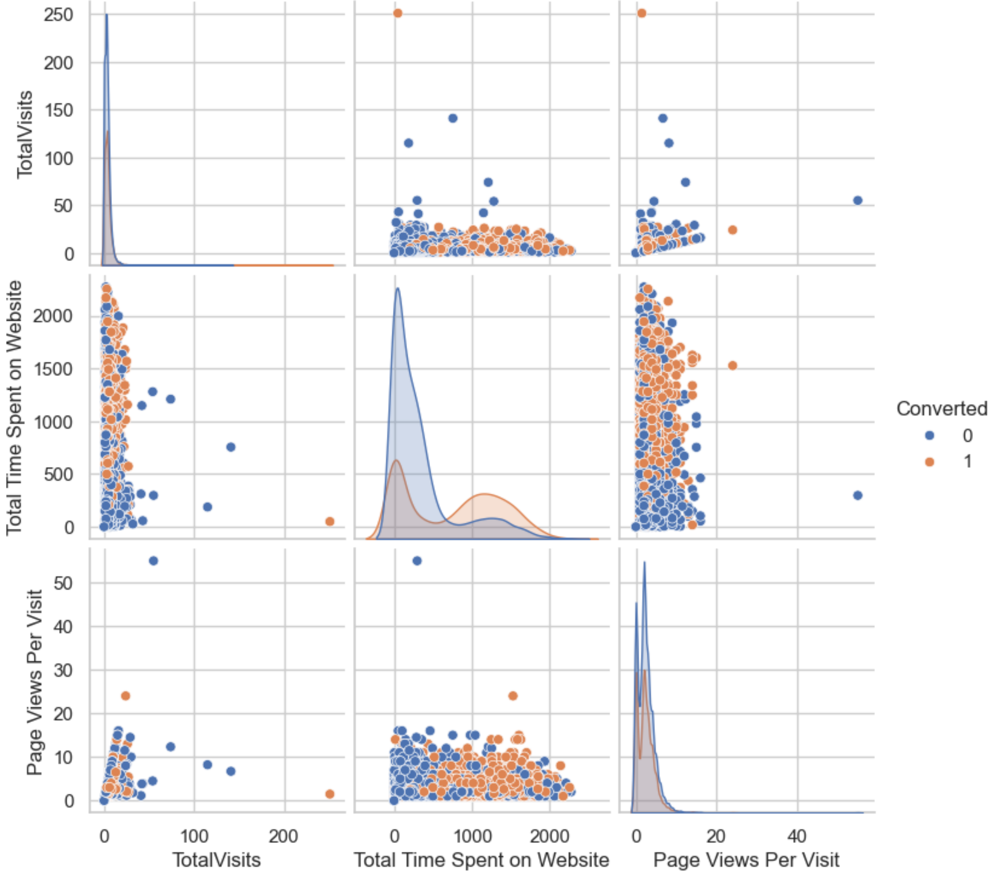
  - Impute the missing values.

- Handled columns with Binary values :
  - Drop those columns with significant data imbalance
  - Drop the columns having only 1 unique entry

Exploratory Data Analysis (EDA):

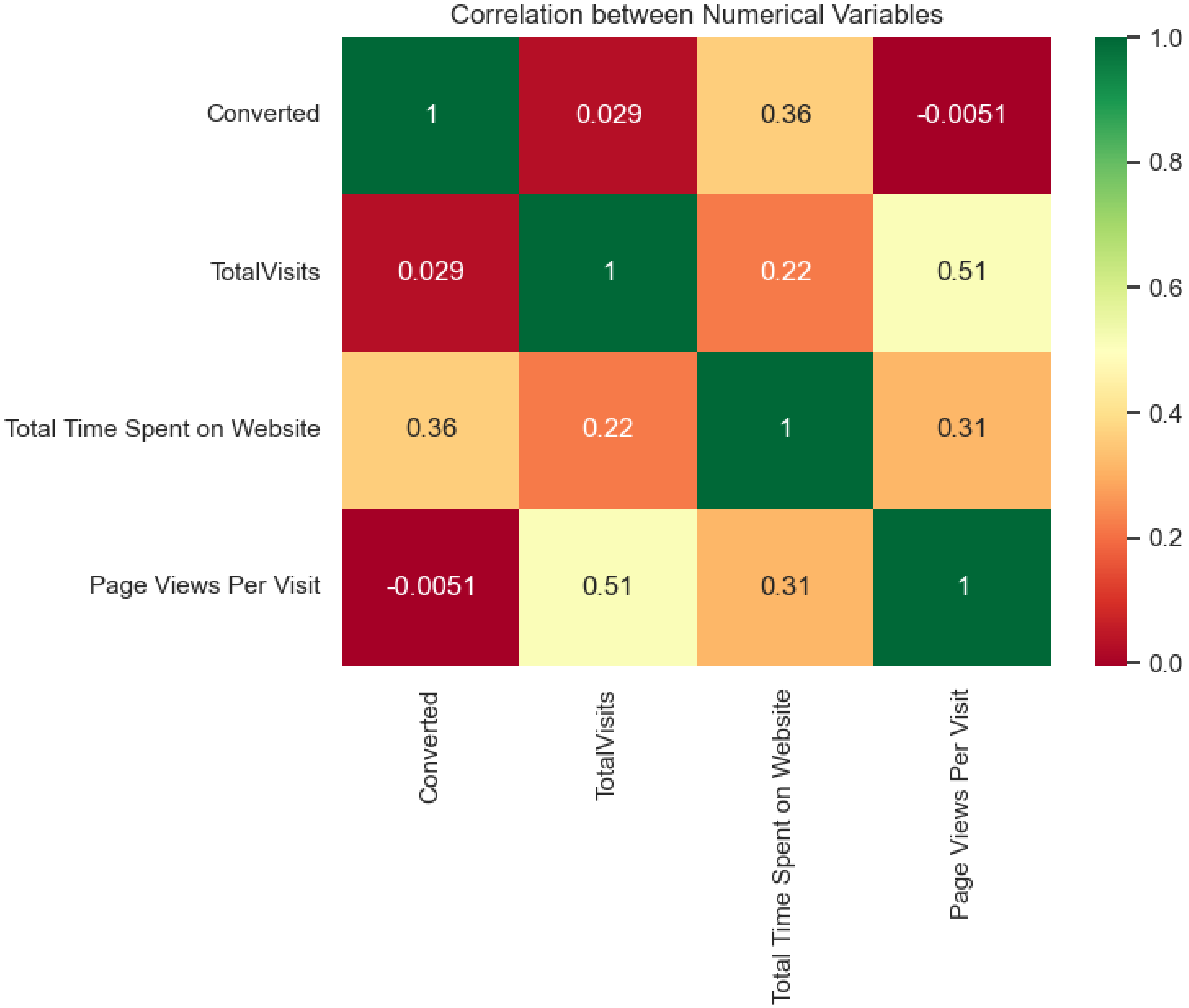
Numerical Analysis :





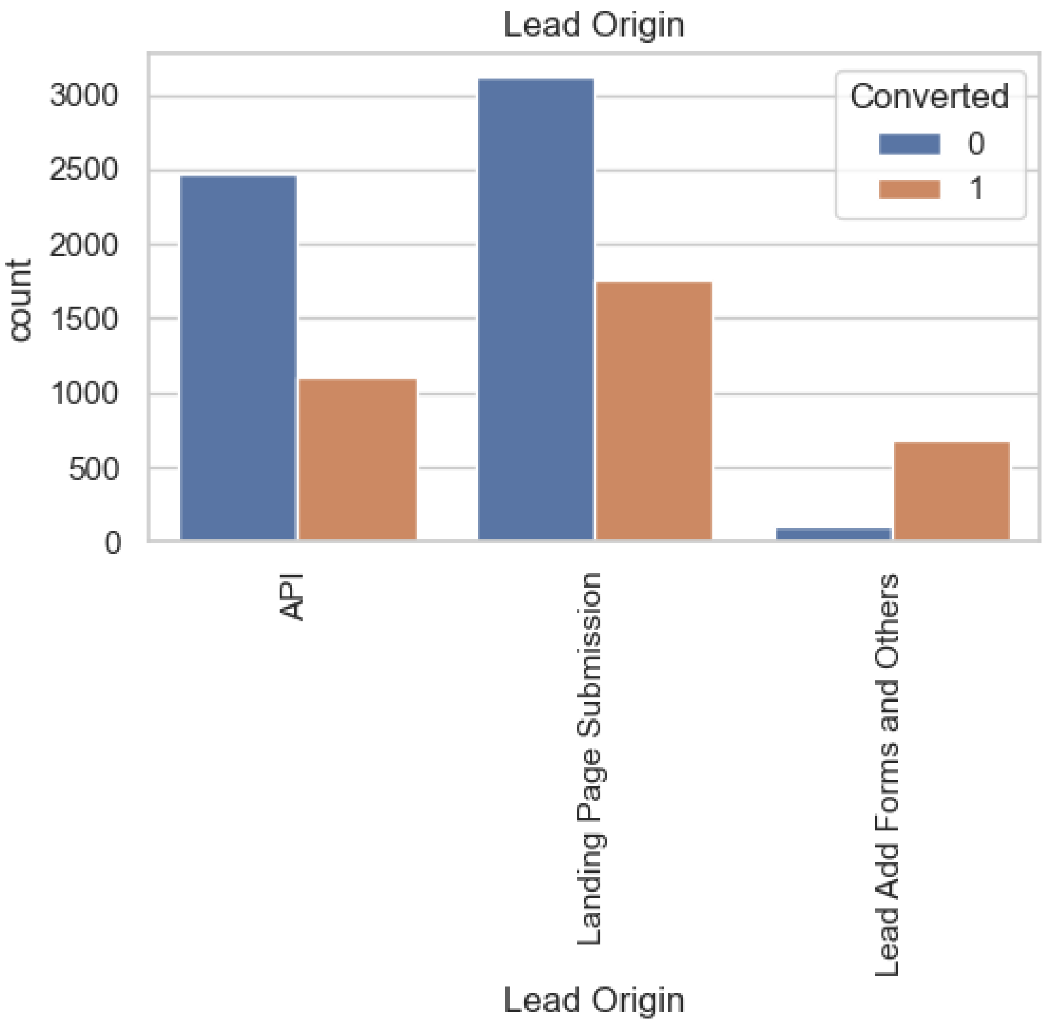
numeric variable analysis w.r.t ‘Converted’

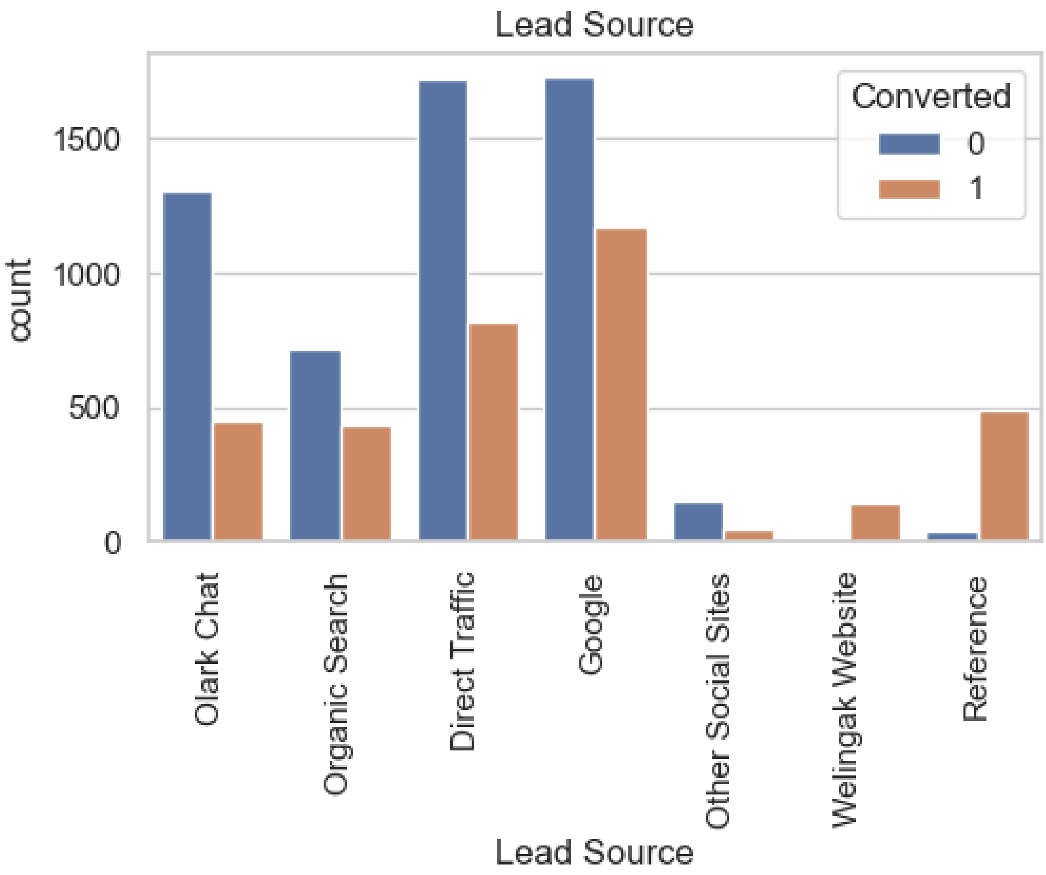
Correlation Matrix between Numeric Variables



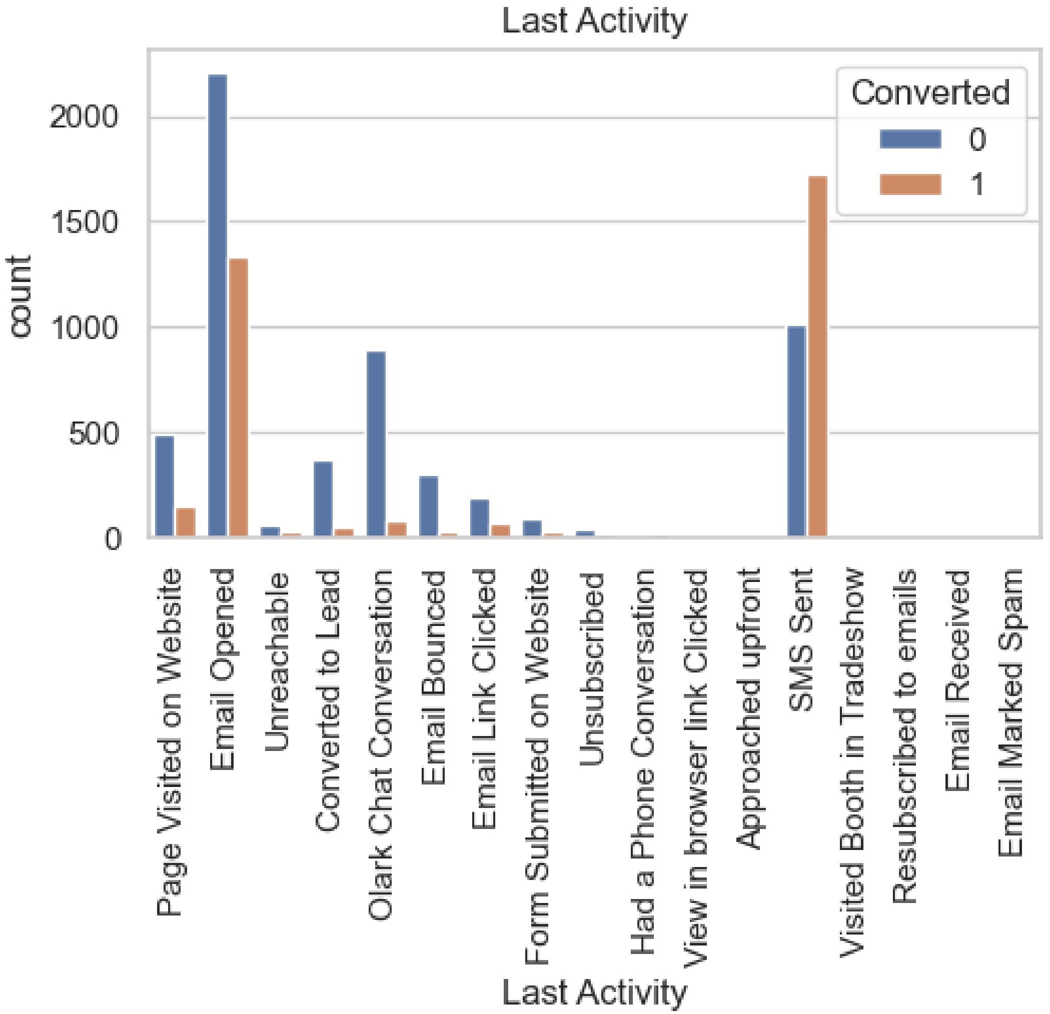
	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1	0.029	0.36	-0.0051
TotalVisits	0.029	1	0.22	0.51
Total Time Spent on Website	0.36	0.22	1	0.31
Page Views Per Visit	-0.0051	0.51	0.31	1

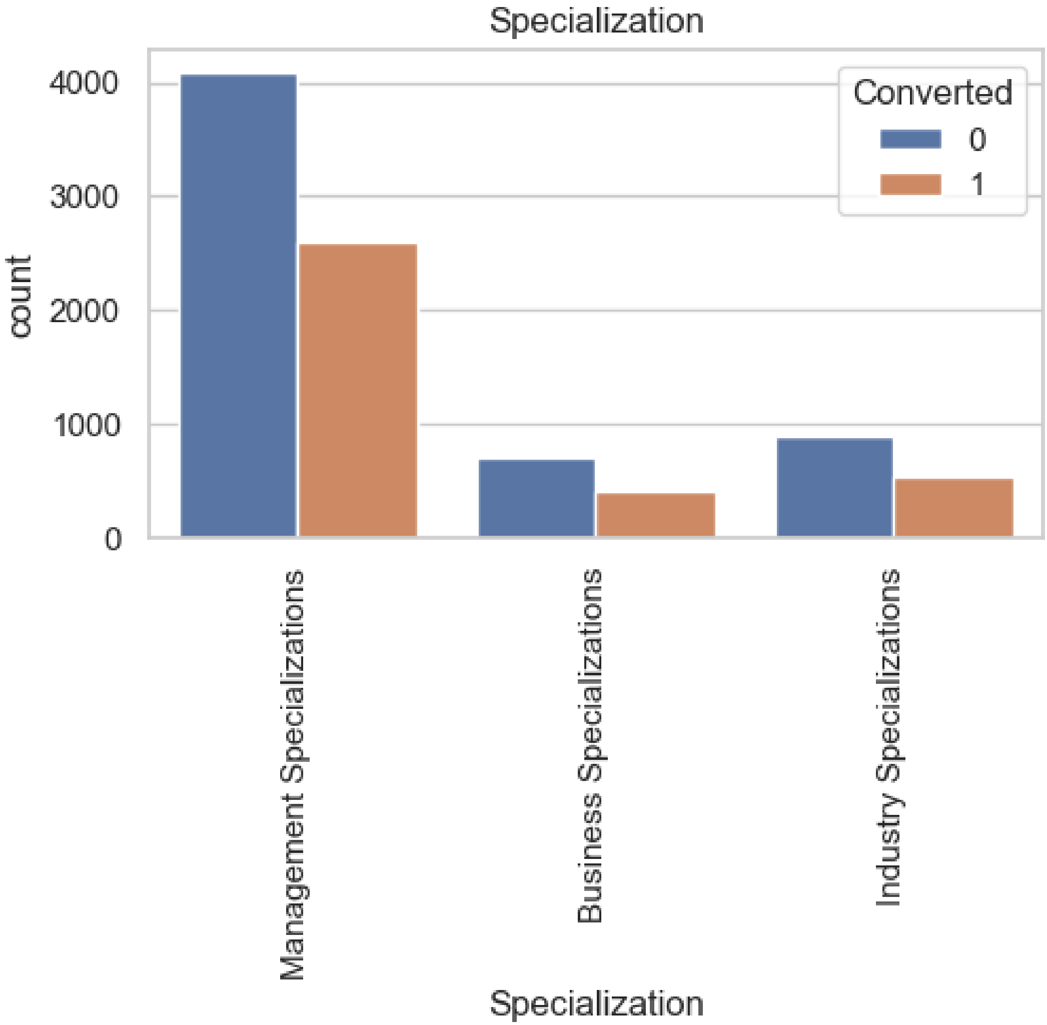
Analysis of Categorical Variables w.r.t. Target variables 'Converted'



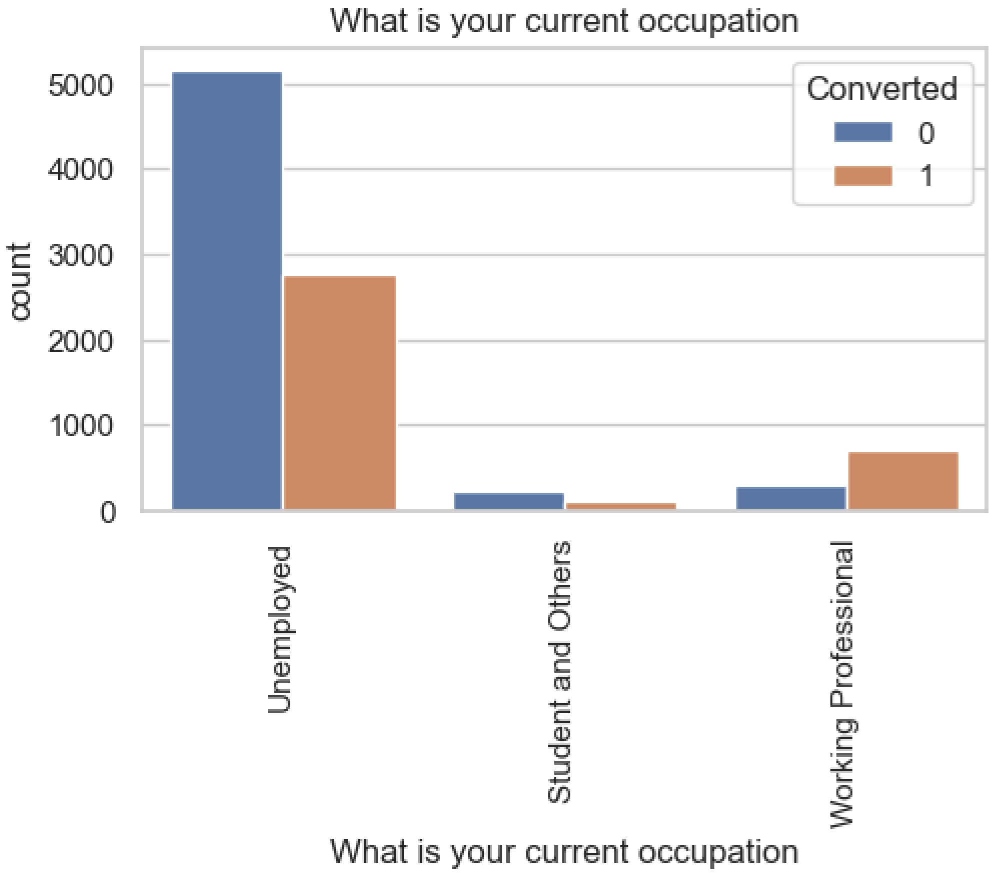


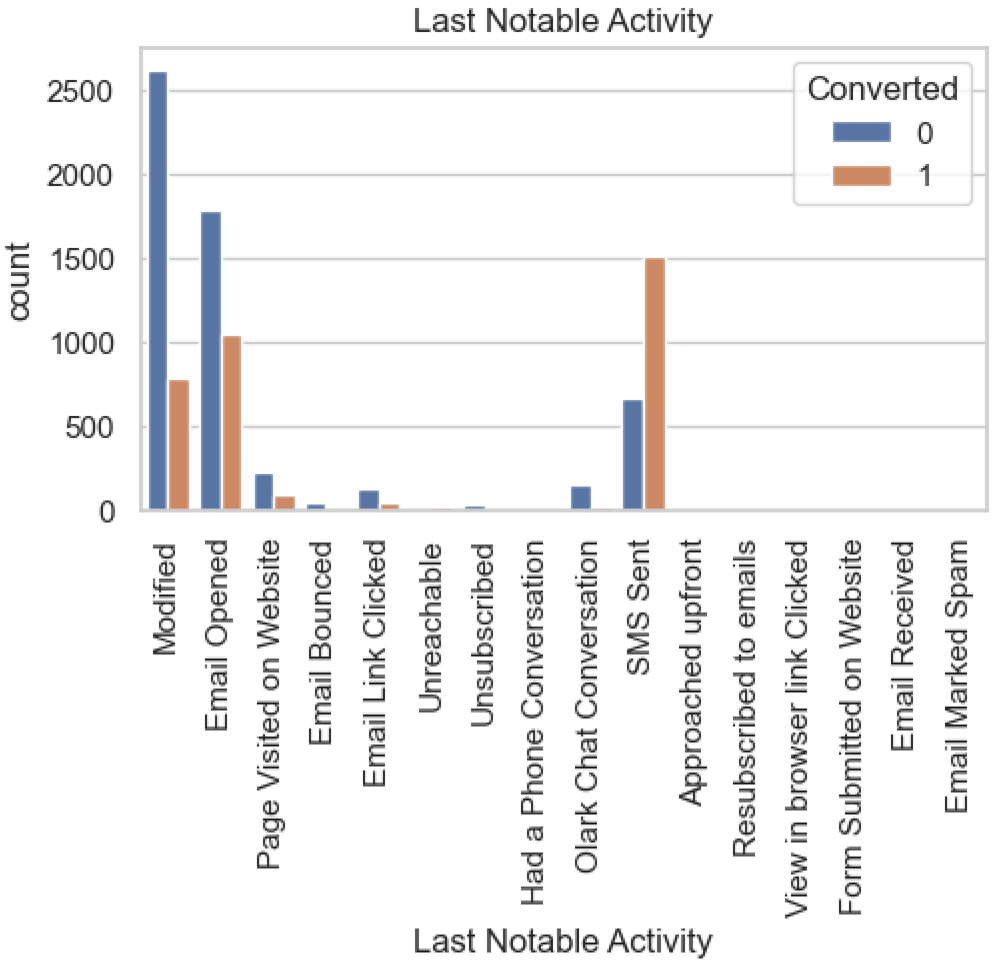
Analysis of Categorical Variables w.r.t. Target variables ‘Converted’



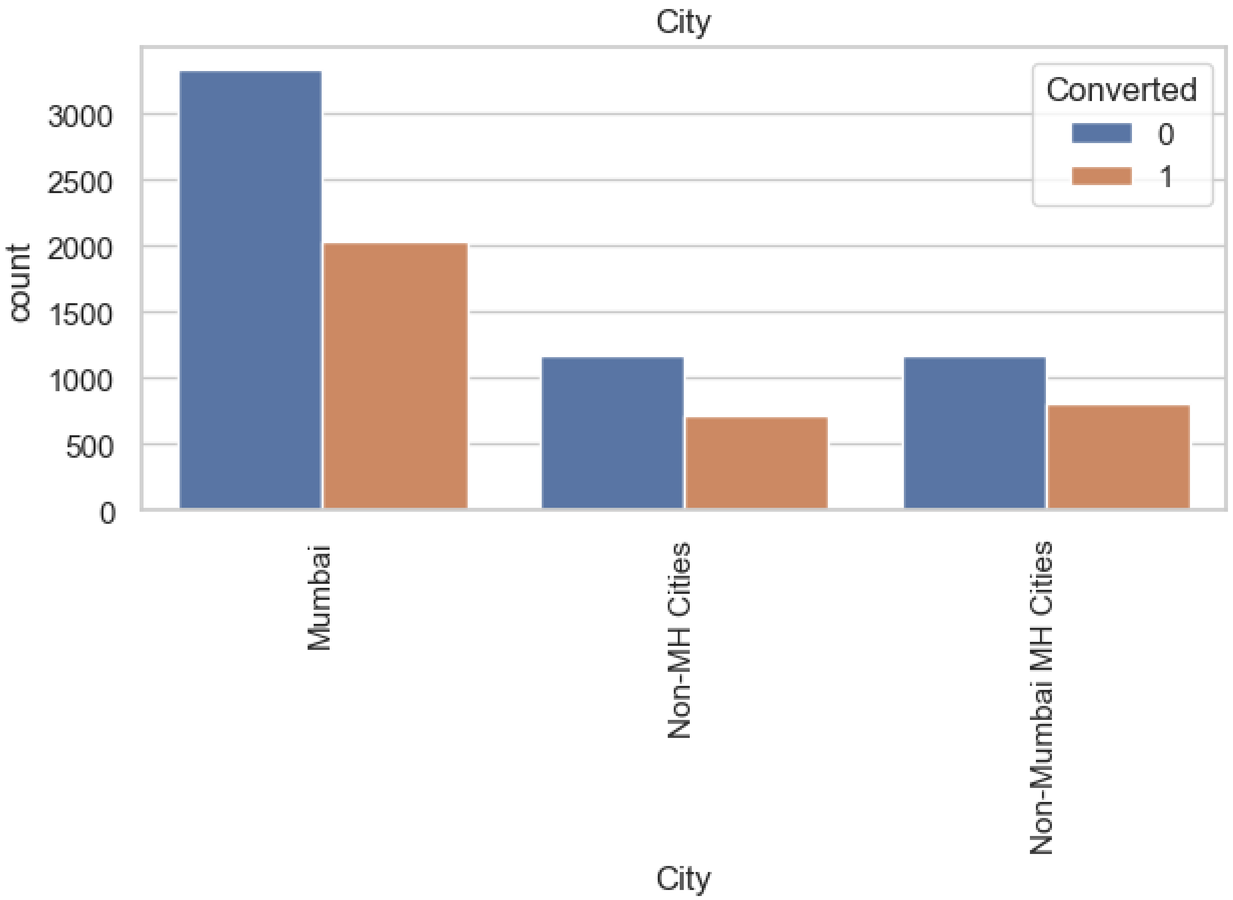


Analysis of Categorical Variables w.r.t. Target variables ‘Converted’





Analysis of Categorical Variables w.r.t. Target variables ‘Converted’



Data Manipulation and Model Building

Dummy Variables:

- Created dummy variables for categorical columns and converted the binary values to numeric elements i.e. 0/1.
- For rescaling numeric variables, we used MinMaxScaler.

Train-Test Data Split :

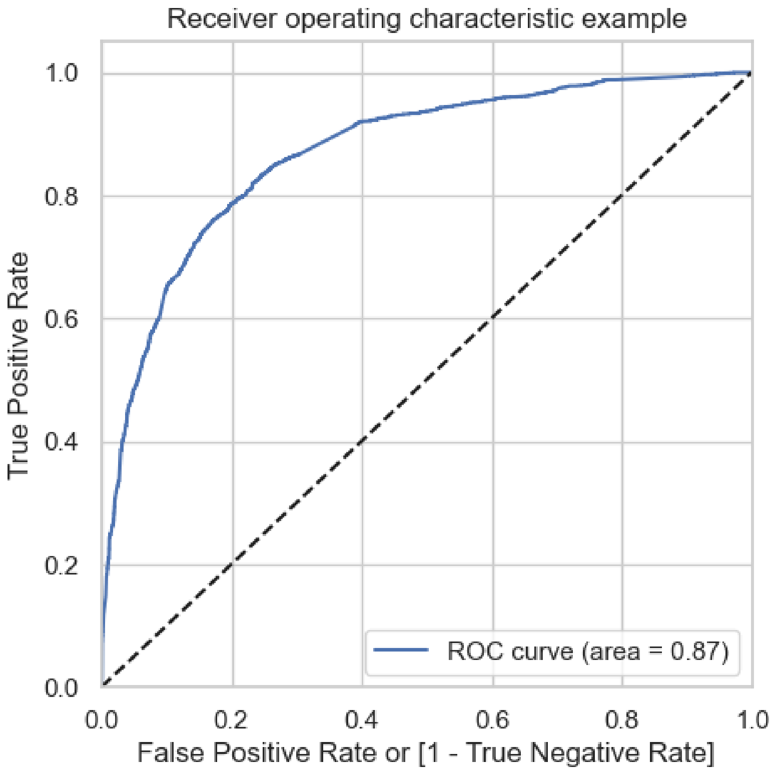
- The dataset was split into 70% for train and 30% for test sets.

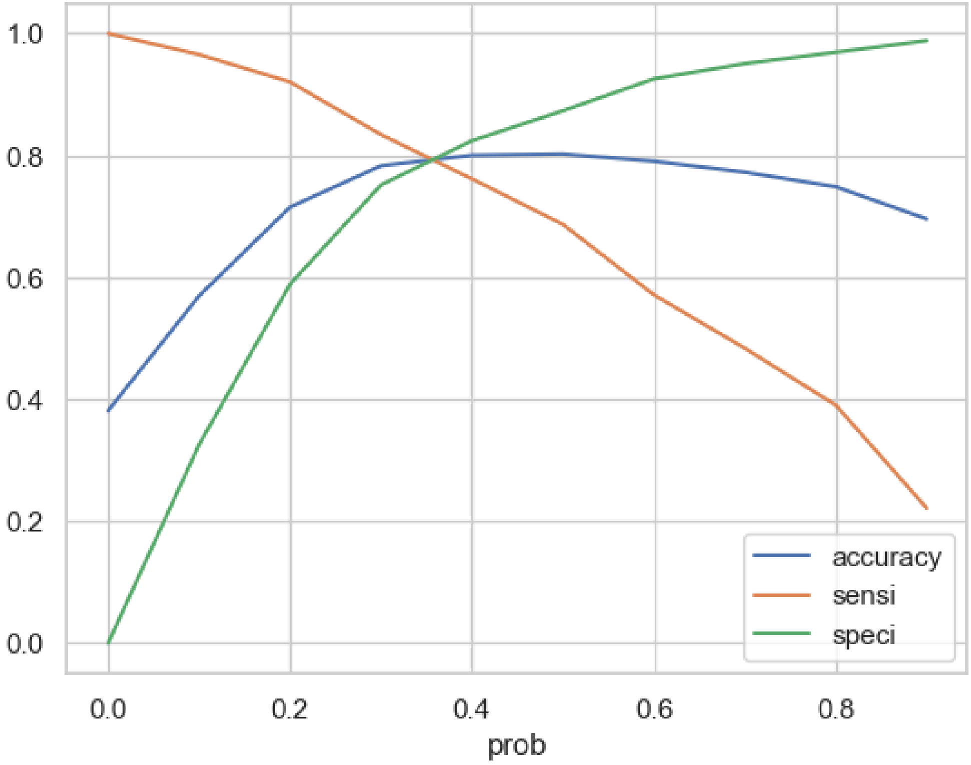
Model Building :

- We used RFE to select the top 15 significant variables to perform model building.
- Later on, the variables with VIF > 5 and p-value > 0.05 values were dropped to build model more significant one.
- At last Model 3 was finalized for performing the Prediction and Model evaluation.

Model Evaluation and Prediction

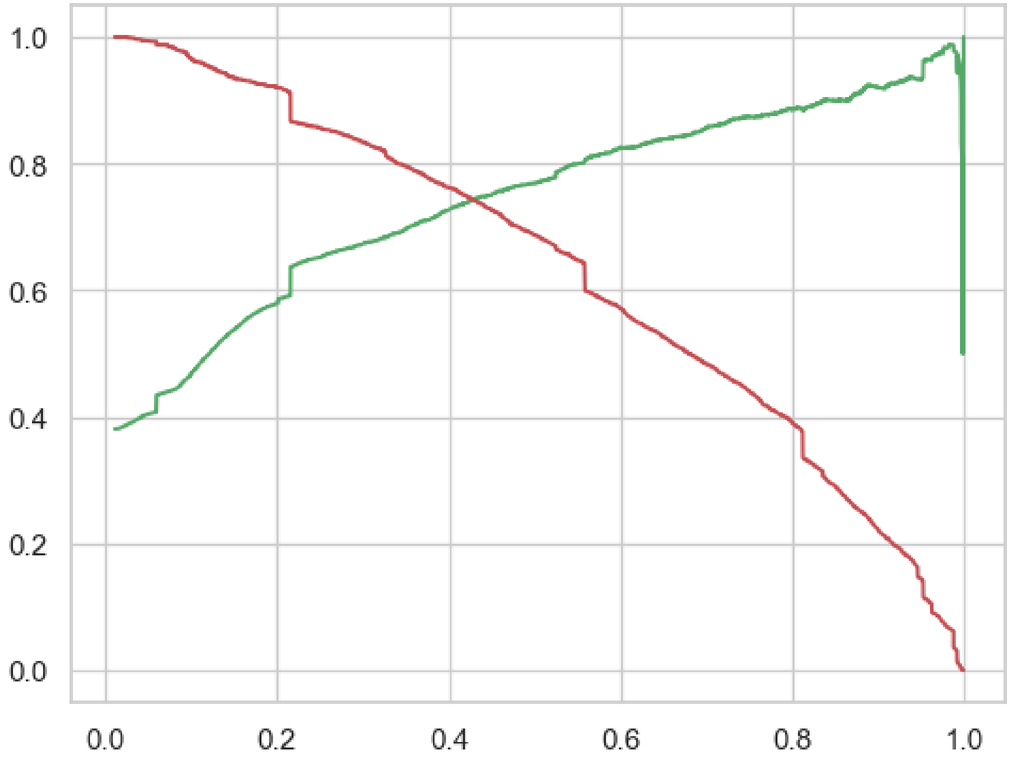
- By creating Optimal cutoff (**ROC curve**) i.e. **0.87** and confusion matrix, the optimal cutoff is around **0.35**.
- Performing prediction on test dataset using cutoffs **0.35**, **Accuracy**, **Sensitivity** and **Specificity** was around **80**%.





Precision and Recall Tradeoff

- This method is used to recheck the prediction made on Train and Test dataset.
- After Performing Precision-Recall of train set and confusion matrix, the optimal cutoff was found around **0.43**.
- The final Precision and Recall was around **77%** and **73%** with overall Accuracy of **80%**.



“Thresholds for Precision and Recall Plot”

Conclusion :

Some significant variables that mattered the most in identifying the promising "Hot Leads" for company:

1. "TotalVisits" : Leads visiting company website often shows potential. Highlight our unique strengths to convince them!
2. "Total Time Spent on Website" : Leads spending more time on websites. This shows their interest in company’s offering.
3. "Page Views Per Visit" : Average number of pages on the website viewed during the visits
4. "Lead Source“ : The source of Lead can be, includes
   - Google
   - Olark Chat
   - Welingak website
   - Reference
5. "Last Activity" : Majority of last activity performed by customer can be
   - Olark Chat Conversation
   - Converted to Lead
   - Email Bounced
   - Had a Phone Conversation
6. "Last Notable Activity“ : The last notable activity performed by the customer can be
   - SMS Sent
7. The Leads having Current Occupation as "Working Professional".

The X Education can flourish their high conversion chance by keeping above factors in mind and can get almost all the potential buyers to become their Converted Leads and buy their courses.