Lead Score Case Study

Group Members | Ganesh Krishnan | Preetha Adiseshan | Somika Shahi

Table Of Contents

Problem statement & Business Objectives

Problem Statement

- X Education sells online courses to industry professionals. X Education seeks to improve its lead conversion rate from the current 30% to a target of 80% by identifying the most promising leads, or "Hot Leads."
- The goal is to develop a predictive model that assigns a lead score to each potential customer, enabling the sales team to focus their efforts on leads with the highest likelihood of conversion, thereby increasing overall efficiency and conversion rates

Business Objectives

- **Increase Lead Conversion Rate**: Improve the lead conversion rate from the current 30% to a target of 80% by focusing sales efforts on the most promising leads.
- **Identify "Hot Leads"**: Develop a model to identify and score leads based on their likelihood of conversion, allowing the sales team to prioritize high-potential customers.
- **Optimize Sales Efforts**: Enhance the efficiency of the sales team by reducing time spent on low-potential leads and concentrating on leads that are more likely to convert.
- **Improve Revenue Generation**: By increasing the lead conversion rate, ultimately boost the company's revenue and growth by converting more website visitors into paying customers.
- **Adaptability for Future Requirements**: Ensure that the model is flexible and can be adjusted to meet future business needs or changes in strategy.
- **Data-Driven Decision Making**: Utilize insights from the model to inform marketing and sales strategies, making more informed and data-driven decisions.

Solution Approach

- Step 1 : Importing Libraries and Loading Data
- Step 2: Inspecting the DataFrame
- Step 3: Data Cleaning
- Step 4: Exploratory Data Analysis (EDA)
- Step 5: Dummy Variables Creation
- Step 6: Test Train Split
- Step 7 : Model Building
- Step 8: Making Predictions & Model Evaluation
- Step 9 : Plotting the ROC Curve
- Step 10 : Prediction on test set & Precision and Recall

Loading Data and Cleaning

Key Observations and Action Items:

- Imported the required Libraries and loaded the data file **'Leads.csv'**
- Total Number of Rows -> **9240**, total number of Columns ->**37**

Handling the Missing Values:

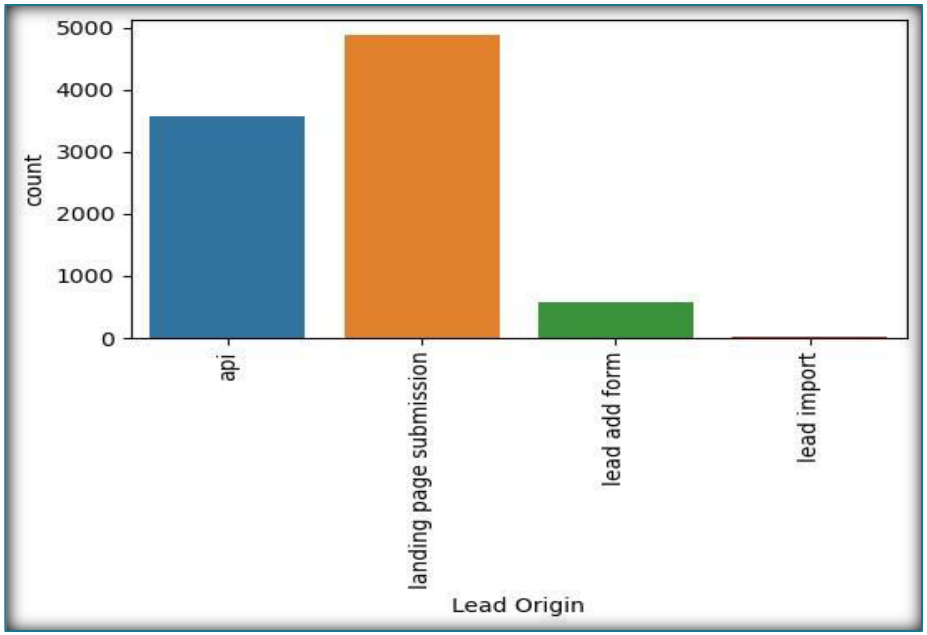
- The "Select" level in categorical variables is treated as a null value and removed, Columns **greater than 40% of nulls values are removed** – "How did you hear about X Education, Lead Profile, Lead Quality, Asymmetrique Profile Score, Asymmetrique Activity Score, Asymmetrique Activity Index, Asymmetrique Profile Index,Tags, City"

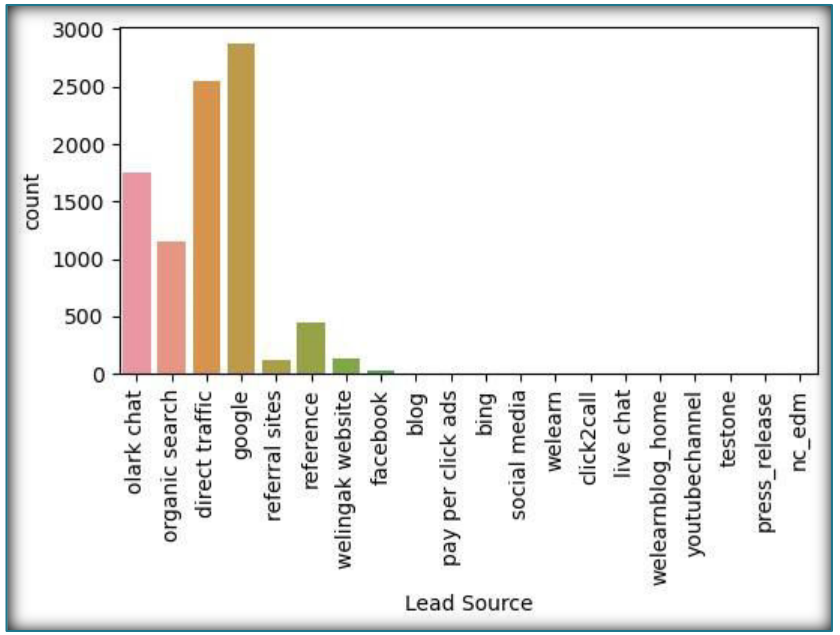
- Columns with more null values are **imputed with "Not Available" values** – "Specialization, What matters most to you in choosing a course, What is your current occupation, Country"

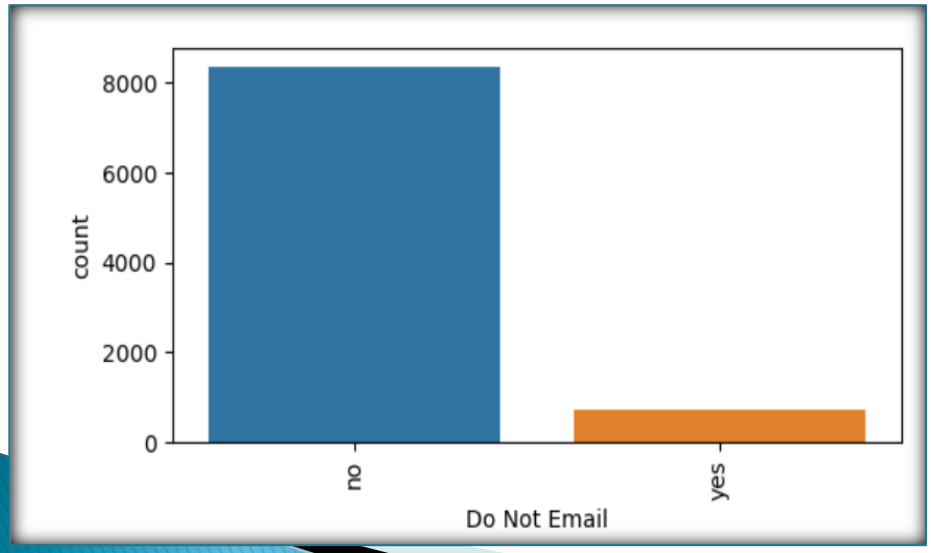
- Columns containing **only one unique value are identifed and removed** "'Magazine','Receive More Updates About Our Courses','Update me on Supply Chain Content','Get updates on DM Content','I agree to pay the amount through cheque' "

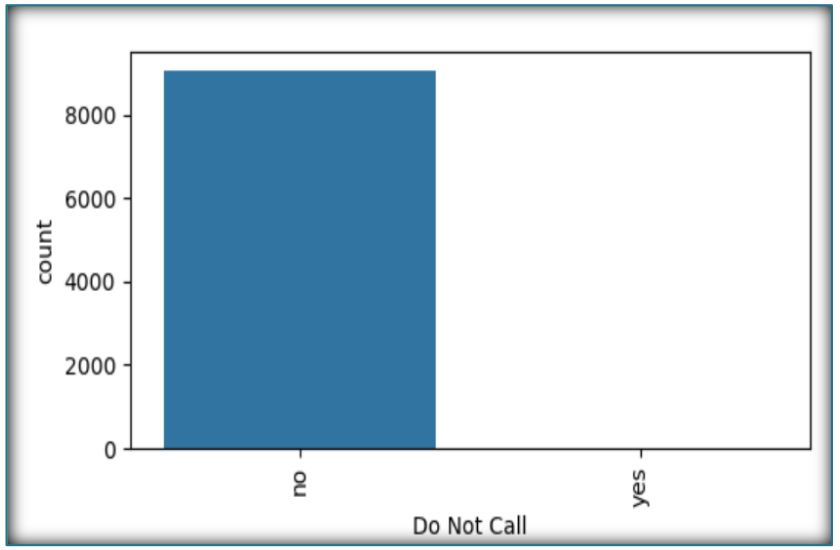
- **ID columns** 'Prospect ID','Lead Number' are dropped as it is not signification for our model analysis.

EDA – Univariate Analysis

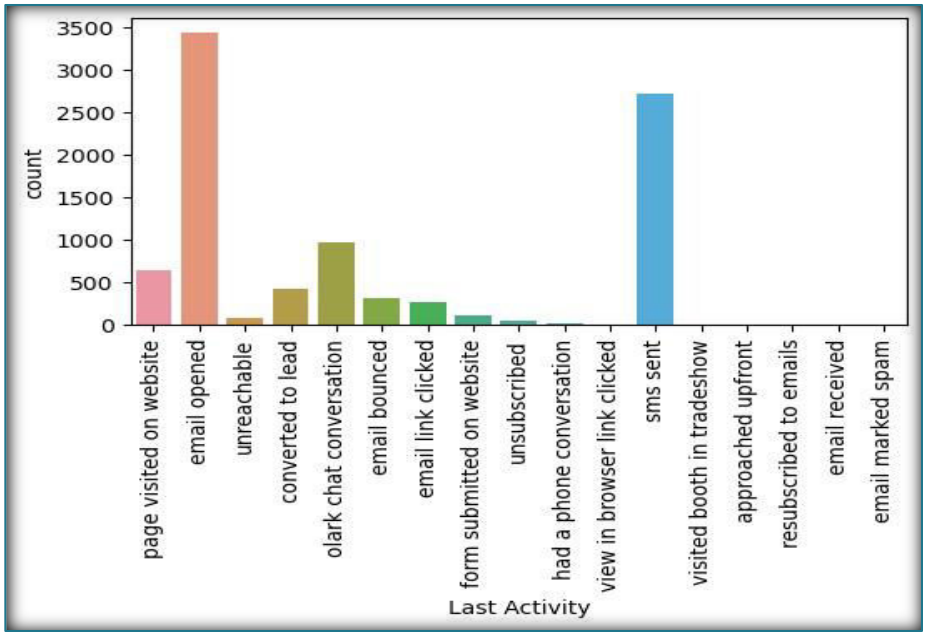


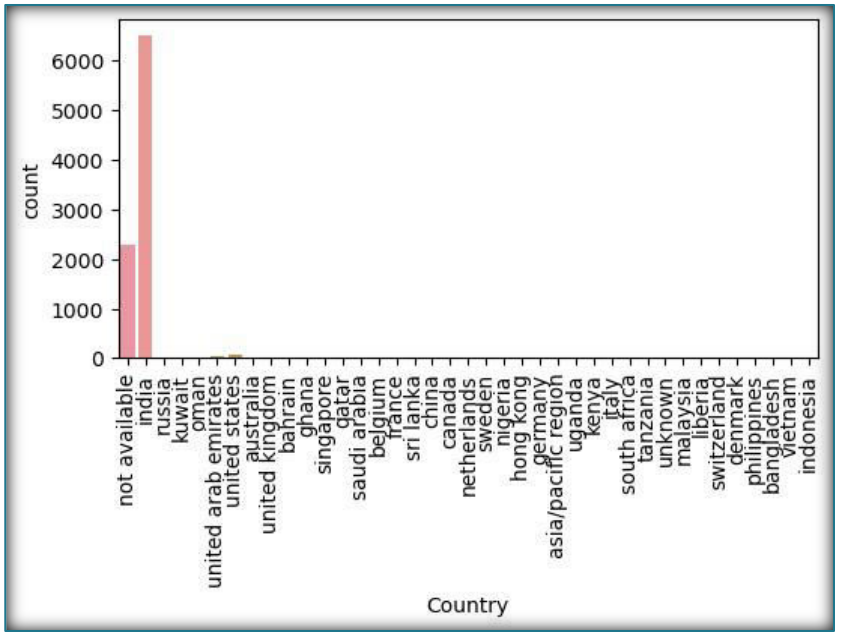


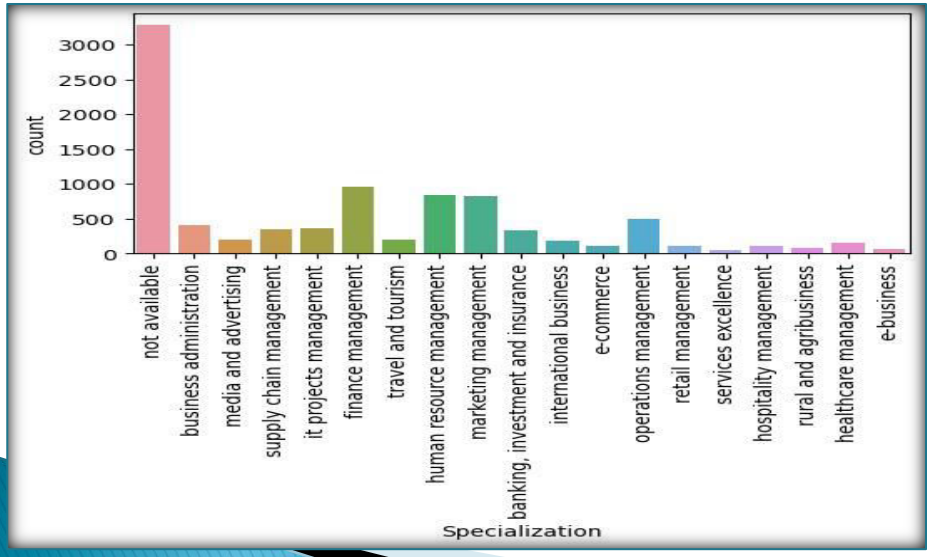


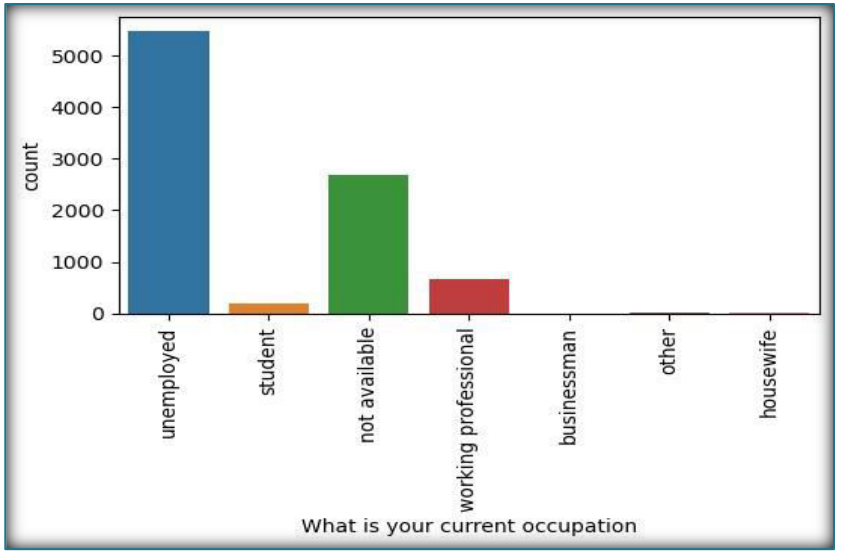


EDA – Univariate Analysis (Cont…)

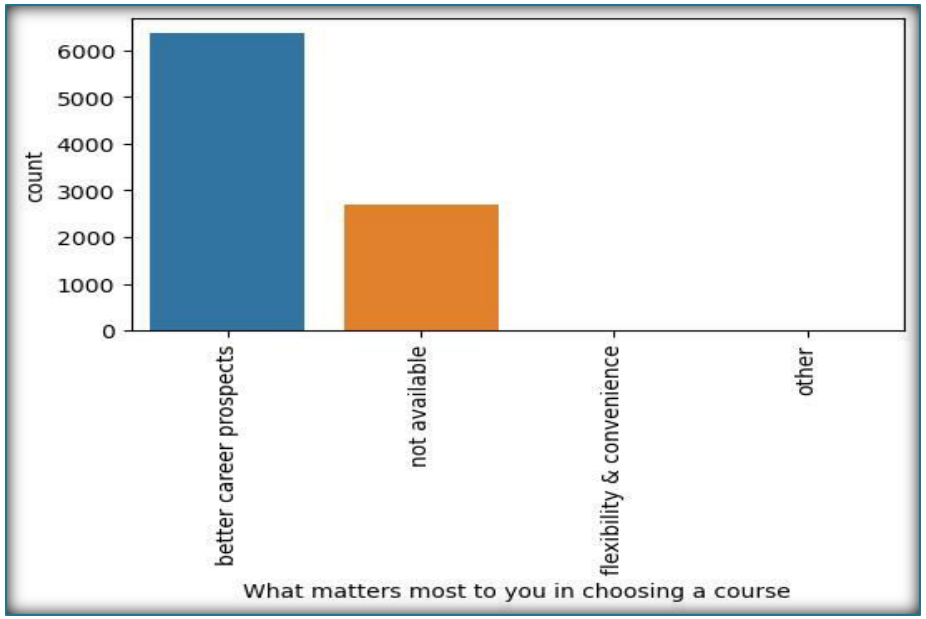


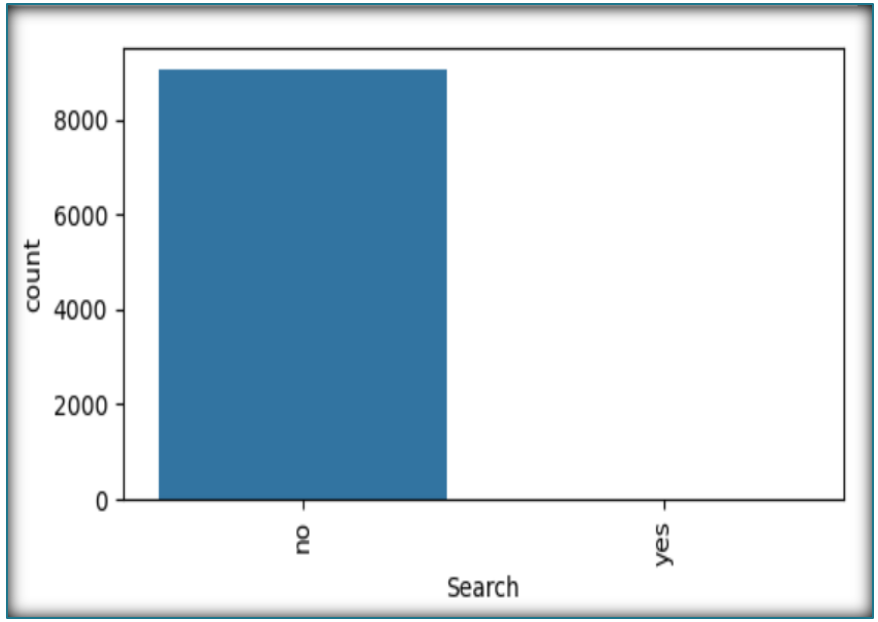


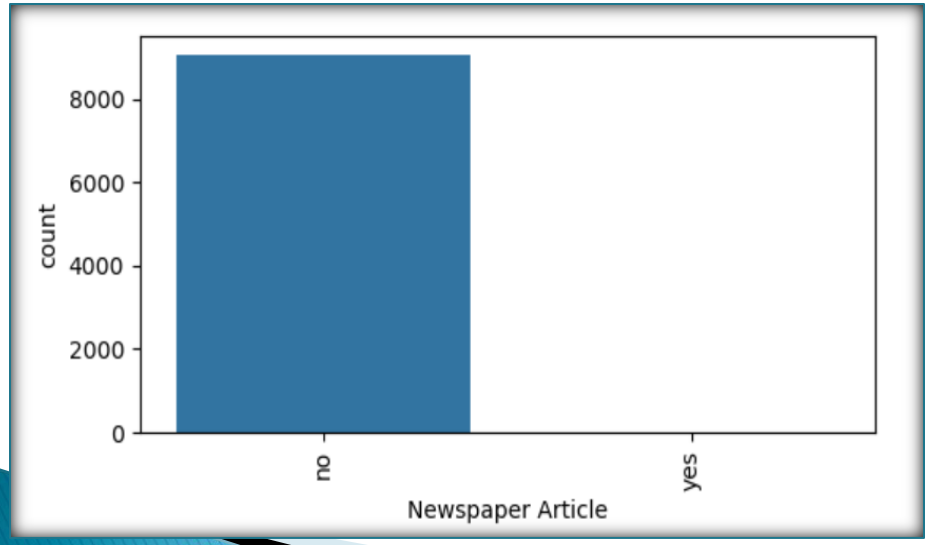


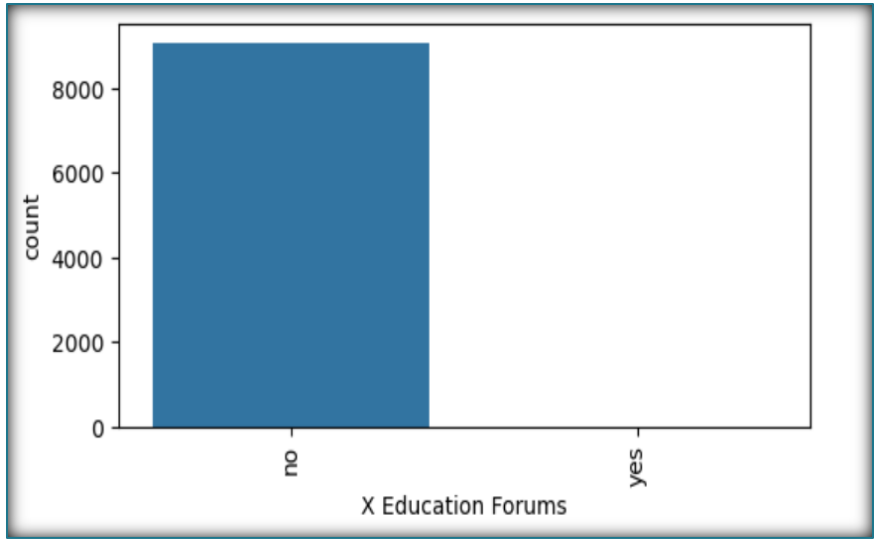


EDA – Univariate Analysis (Cont…)

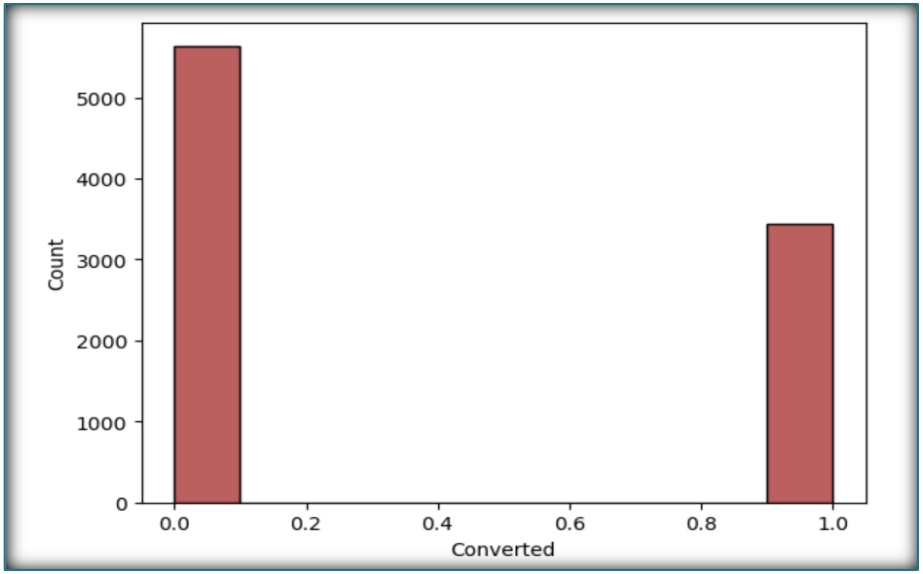


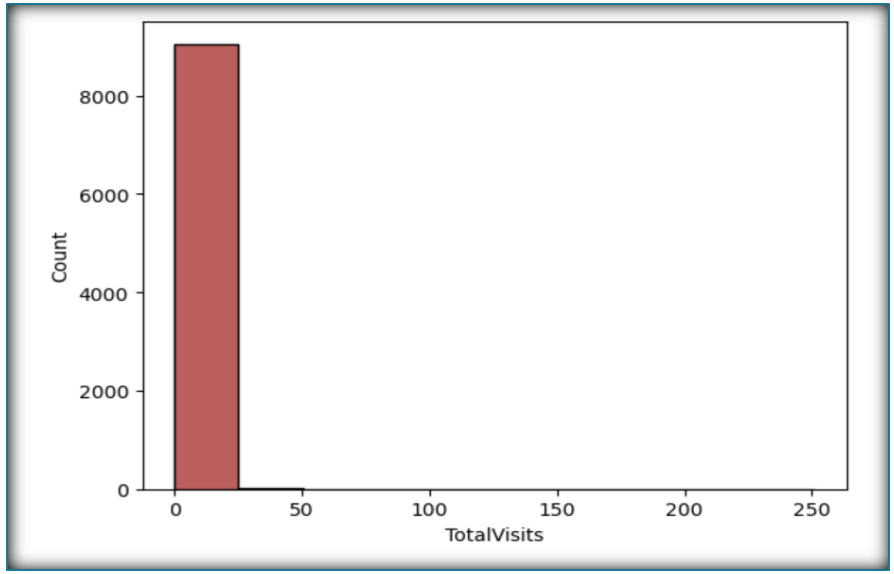


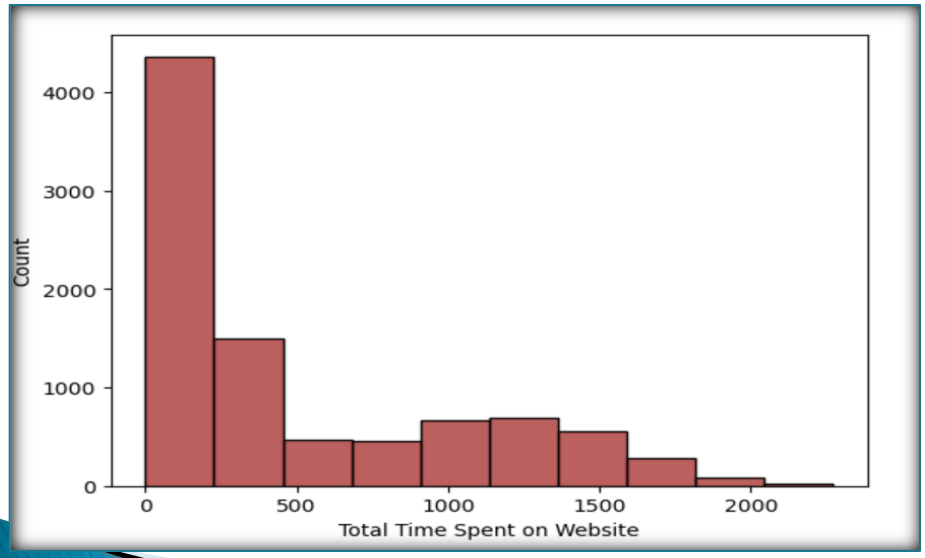


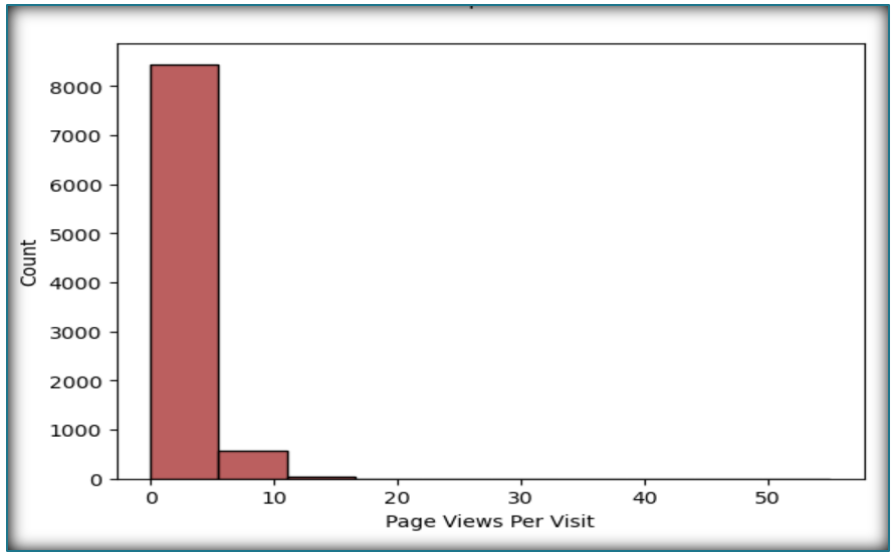


EDA – Univariate Analysis (Cont…)

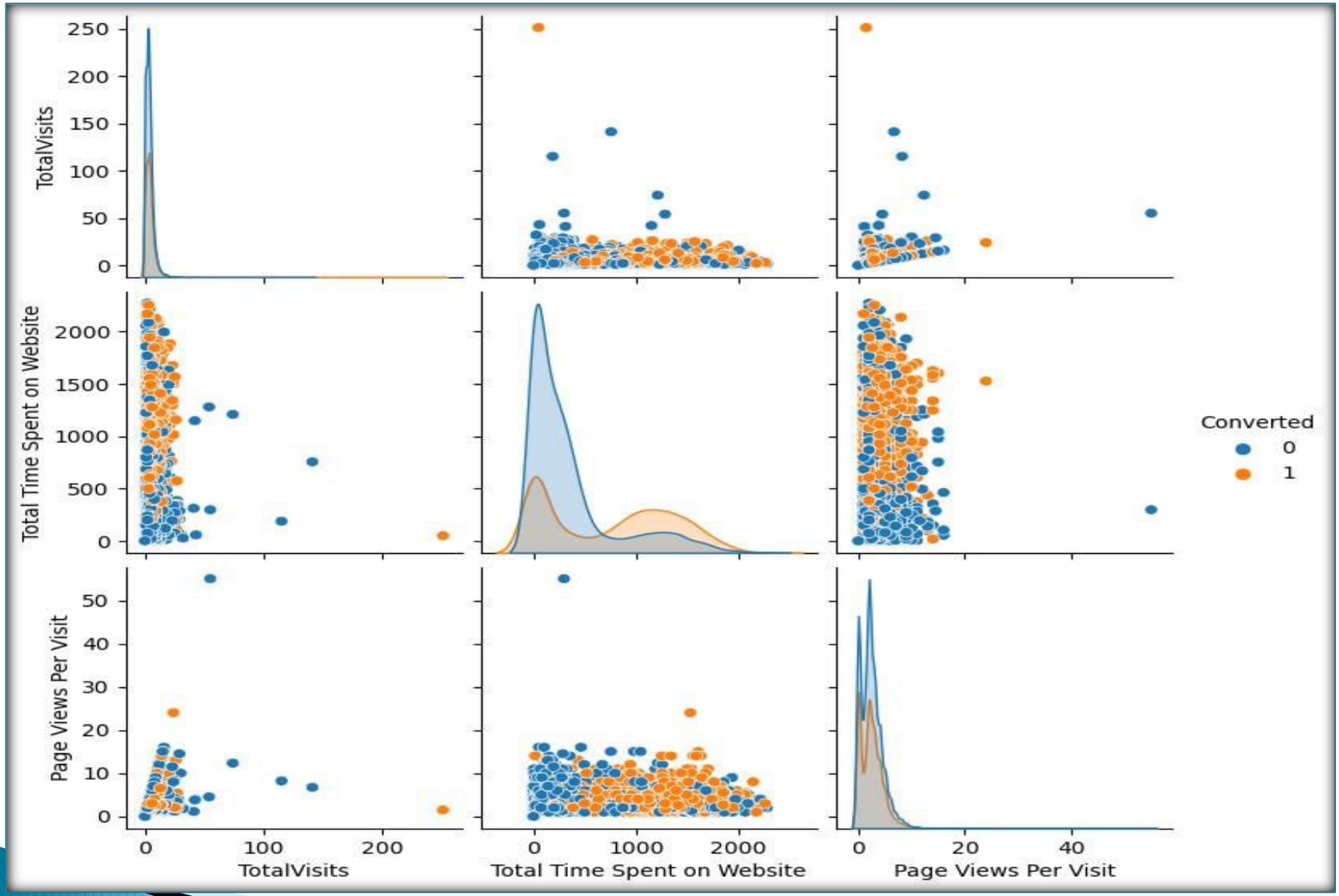








EDA – Bivariate Analysis



EDA Analysis – Inferences

- The majority of the leads are from India
- Better career prospects is what the leads says that matters most to them
- Google followed Direct traffic has produced the most leads to the X education
- Most of the unemployed folks are looking to upskill.
- There is no major outlier found with this dataset
- There are many columns that has minimum data and will be less relevent to our model analysis

Dummy Variable Creation & Test Train Split

- Numerical Variables are scaled

```
# Scale the 3 numeric variables using Minmax Scaler
scaler=MinMaxScaler()
X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']]=scaler.fit_transform(X_train[['TotalVisits','To
X_train.head()
```

- Dummy Variables are created for object type variables (Categorical variables)
- After the dummy variable creation, the data frame holds total Rows for Analysis: 9074 , total Columns for Analysis: 81
- 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.

```
In [48]: # split the training and testing dataset in 70 to 30 ratio
X_train,X_test,y_train,y_test=train_test_split(X,y,train_size=0.7,test_size=0.3,random_state=10)
```

Model Building

- Used the Recursive feature building to remove the weakest features that are not needed for our model build
- Best 15 variables of our model

```
Index(['TotalVisits', 'Total Time Spent on Website',
       'Lead Origin_lead add form', 'Lead Source_olark chat',
       'Lead Source_welingak website', 'Do Not Email_yes',
       'Last Activity_olark chat conversation', 'Last Activity_sms sent',
       'What is your current occupation_housewife',
       'What is your current occupation_other',
       'What is your current occupation_student',
       'What is your current occupation_unemployed',
       'What is your current occupation_working professional',
       'Last Notable Activity_had a phone conversation',
       'Last Notable Activity_unreachable'],
      dtype='object')
```

- Assessed the model with StatsModels
- The final model depicted below has the features with P values < 0.05 and VIF value < 5

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4533	0.113	-30.579	0.000	-3.675	-3.232
TotalVisits	5.5427	1.444	3.838	0.000	2.712	8.373
Total Time Spent on Website	4.6048	0.166	27.690	0.000	4.279	4.931
Lead Origin_lead add form	3.7501	0.225	16.651	0.000	3.309	4.192
Lead Source_olark chat	1.5802	0.111	14.187	0.000	1.362	1.798
Lead Source_welingak website	2.5821	1.033	2.500	0.012	0.558	4.607
Do Not Email_yes	-1.4360	0.170	-8.437	0.000	-1.770	-1.102
Last Activity_olark chat conversation	-1.3974	0.167	-8.348	0.000	-1.725	-1.069
Last Activity_sms sent	1.2672	0.074	17.164	0.000	1.123	1.412
What is your current occupation_other	2.1567	0.755	2.857	0.004	0.677	3.636
What is your current occupation_student	1.2456	0.226	5.502	0.000	0.802	1.689
What is your current occupation_unemployed	1.1632	0.086	13.582	0.000	0.995	1.331
What is your current occupation_working professional	3.6797	0.204	18.008	0.000	3.279	4.080
Last Notable Activity_unreachable	1.8153	0.601	3.022	0.003	0.638	2.993

	Features	VIF
10	What is your current occupation_unemployed	2.30
1	Total Time Spent on Website	2.06
0	TotalVisits	1.85
2	Lead Origin_lead add form	1.58
7	Last Activity_sms sent	1.53
3	Lead Source_olark chat	1.51
6	Last Activity_olark chat conversation	1.37
11	What is your current occupation_working profes...	1.32
4	Lead Source_welingak website	1.31
5	Do Not Email_yes	1.06
9	What is your current occupation_student	1.05
8	What is your current occupation_other	1.01
12	Last Notable Activity_unreachable	1.01

Predictions & Model Evaluation

Predictions on training data set

In [202]:
```
# Data frame with already available convertion rate and probablity of predicted values
y_train_pred_final = pd.DataFrame({'Converted':y_train.values, 'Conversion_Prob':y_train_pred})
y_train_pred_final.head()
```

Out[202]:

	Converted	Conversion_Prob
0	1	0.647883
1	0	0.133180
2	0	0.232946
3	0	0.133180
4	0	0.495090

In [203]:
```
# Creating new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0
y_train_pred_final['Predicted'] = y_train_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.5 else 0)
y_train_pred_final.head()
```

Out[203]:

	Converted	Conversion_Prob	Predicted
0	1	0.647883	1
1	0	0.133180	0
2	0	0.232946	0
3	0	0.133180	0
4	0	0.495090	0

Accuracy:

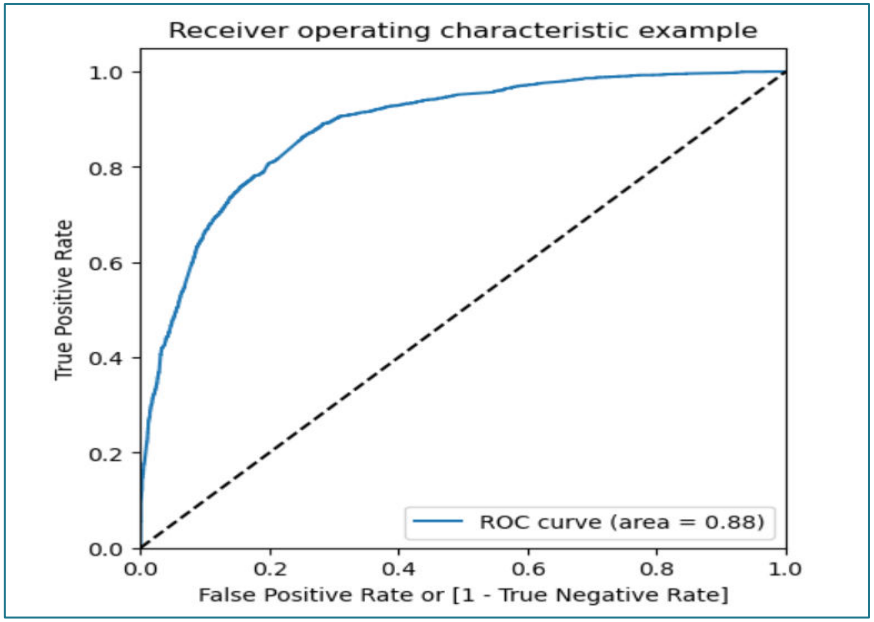
Overall Accuracy of the model is 81%

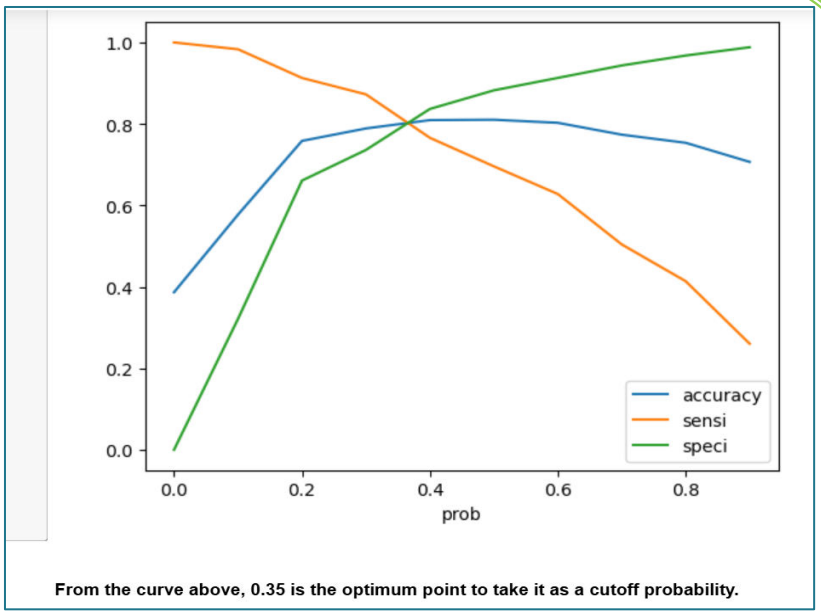
In [208]:
```
# Check the overall accuracy
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted)
```

Out[208]: 0.810266099826799

We have got an accuracy of 81% from this model

ROC Curve

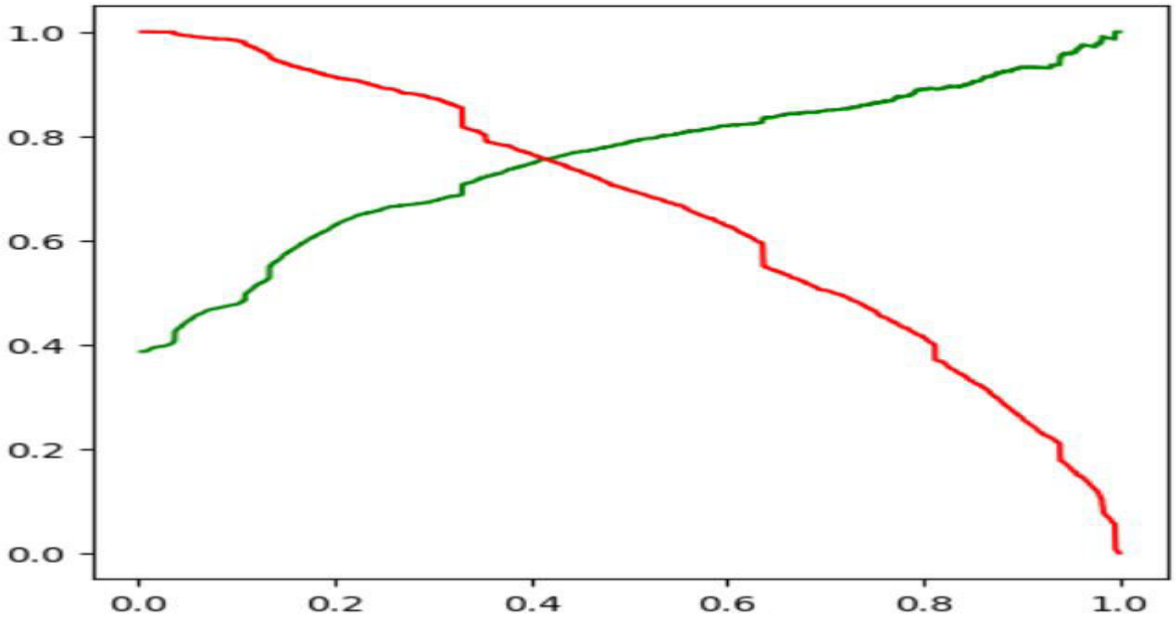




From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

- The area under ROC curve is 0.88 and it looks good
- The previous cutoff of 0.5 is randomly selected and we will find the optimum cut off
- 0.35 is the optimum point identified as a cutoff probability.
- With the current cut off of Conversion_Prob as 0.35 we get accuracy, sensitivity and specificity as just over 80% for train dataset.

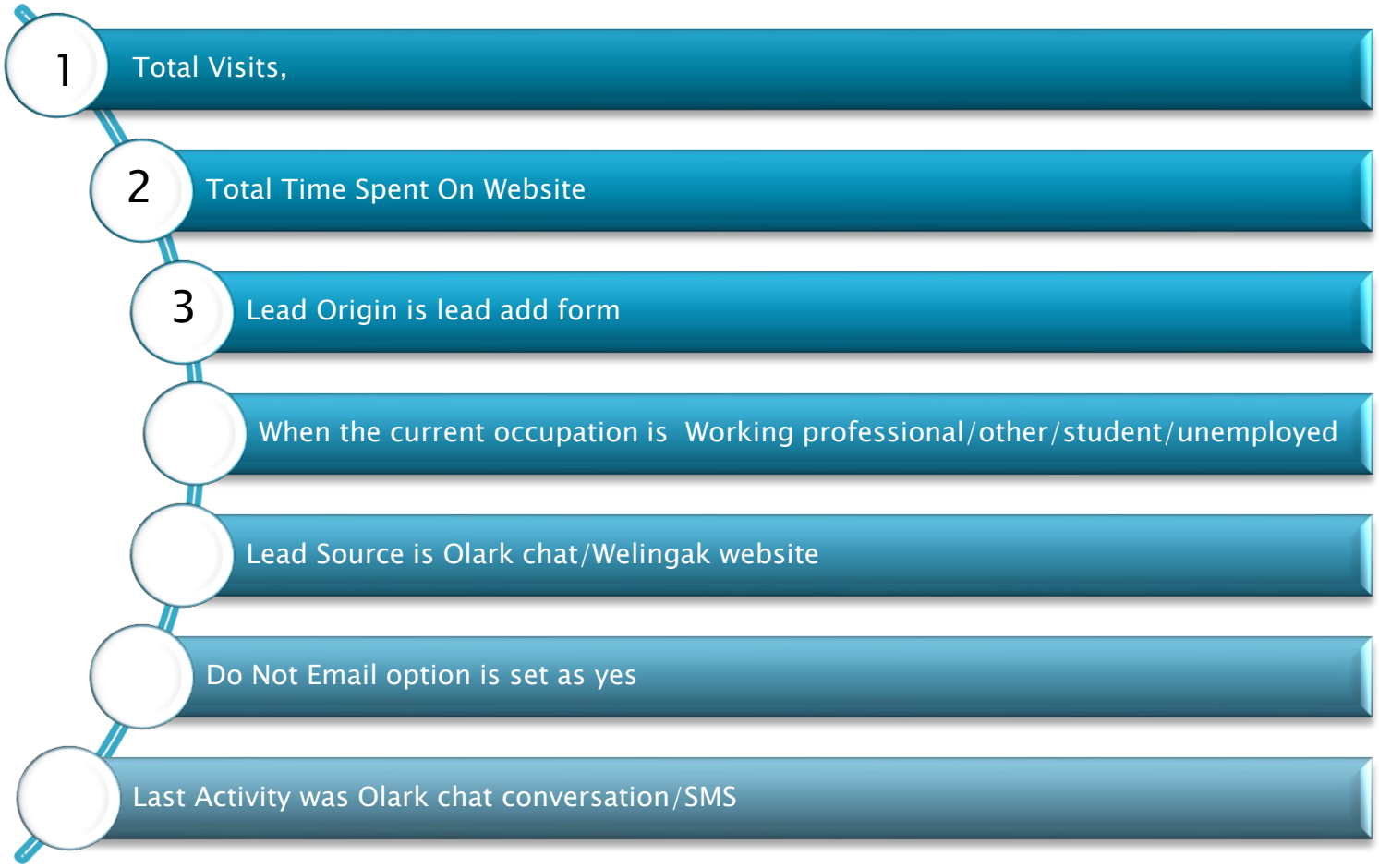
Prediction on test set & Precision and Recall



- With the current cut off of Conversion_Prob as 0.35 we get accuracy, sensitivity and specificity as ~ 81% for test dataset
- With the current cut off of Conversion_Prob as 0.35 we get Precision as 79% and Recall as 70%
- From the precision recall curve, we identified the new cut of as 0.42.
- With the identified cut off as 0.42 we get the accuracy as 81%, Precision as 76% and Recall as 75% for the training dataset
- With the identified cut off as 0.42 we get accuracy as 81%, Precision as 74% and Recall as 76% for the test dataset

Observations

Below are the features that influence the Conversation rate of leads in X education and hence they need to focus on these features:



Thank You!