Lead Scoring Case Study Presentation

DS C32

Pratyush Pran

Siddhee Washimkar

Agenda

- Problem Statement
- Analysis Approach
- Visualizations
- Business Terms Understanding

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- While the company markets its courses on several websites and search engines like Google, they receive traffic on their website and get sales leads for further conversion.
- The problem is that their lead conversion rate is poor – which is 30%.
- To make the leads (potential customers) convert into buyers, the conversion process needs to be more efficient, for which 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 and aims to convert the lead percentage rate to 80%. For that they want to build a Model which identifies the hot leads and refer to the model for the future use.

Analysis Approach

- Data Cleaning and Data manipulation
- Check and handle duplicate data
- Check and handle NA values and missing values
- Drop columns, if it contains large amount of missing values that are not useful for the analysis
- Imputation of the values, if necessary
- Check and handle outliers in data

Exploratory Data Analysis (EDA)

- 1. Univariate data analysis for categorical and numerical variables
- 2. Bivariate data analysis: correlation coefficients and pattern between the variables
- Creation of Dummy Variables and encoding of the data
- perform Test Train and split process
- Classification Technique: Logistic regression used for the model making and prediction
- Evaluation of the model

Predictions on Test Set

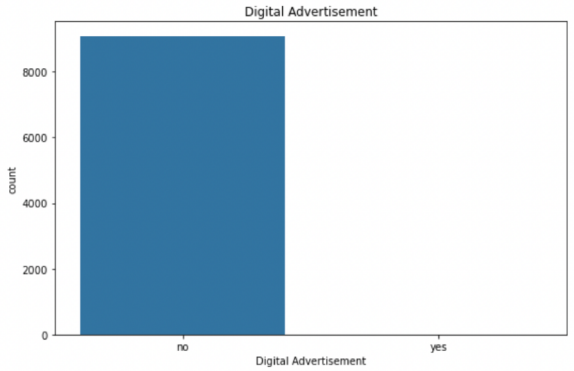
Model Presentation
Conclusions and Recommendations

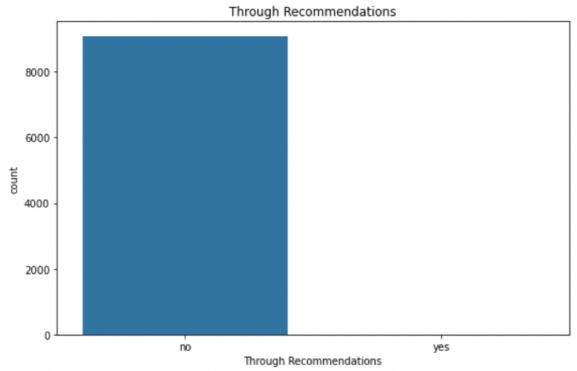
Data Information & Manipulation

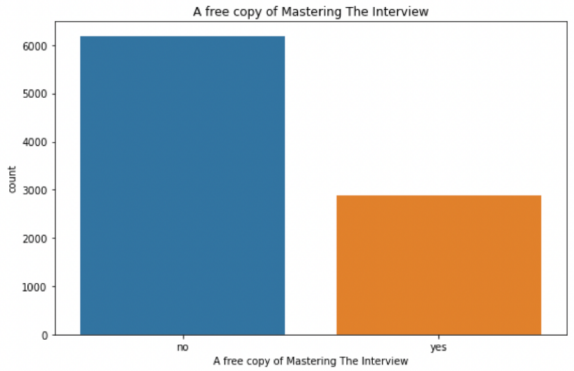
- Data Inspection shows a total Number of 37 row and 9240 columns
- Single value features found in the Data like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removal of “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 do not have enough variance and 35% null values, 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.
- We will also drop the columns having more than 35% of missing values such as ‘How did you hear about X Education’ and ‘Lead Profile’.

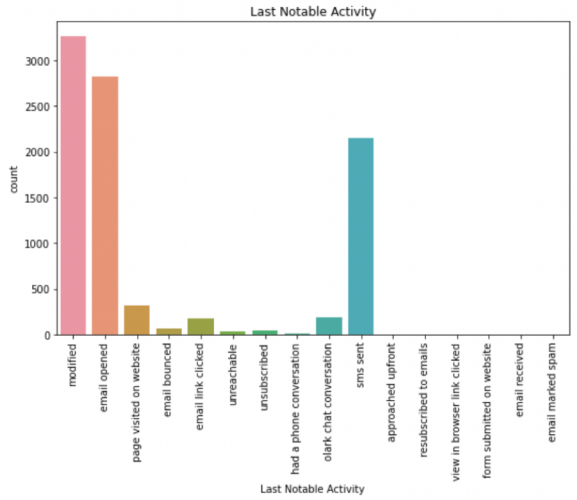
Exploratory Data Analysis

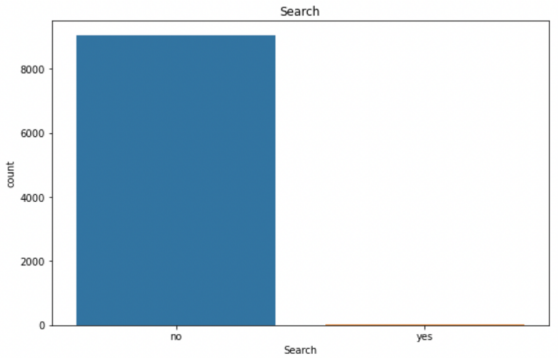
Univariate Analysis for Categorical Variables

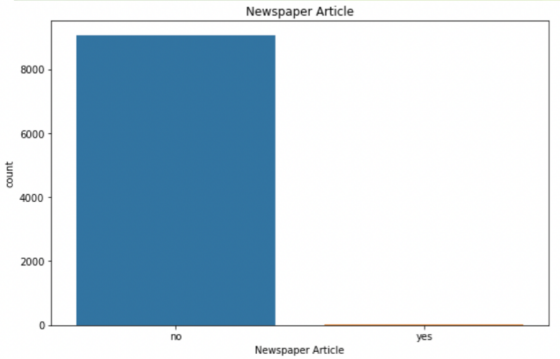


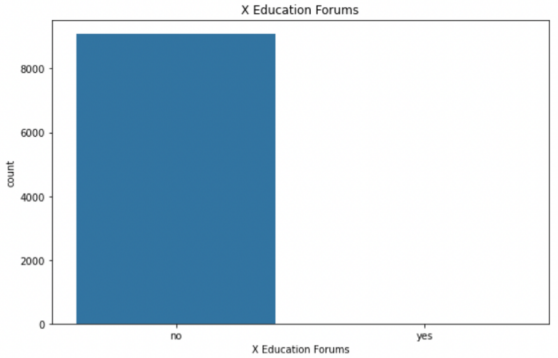


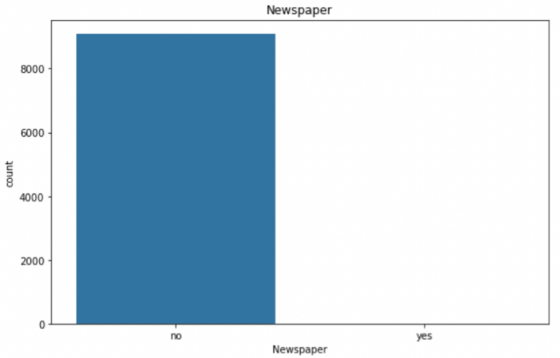


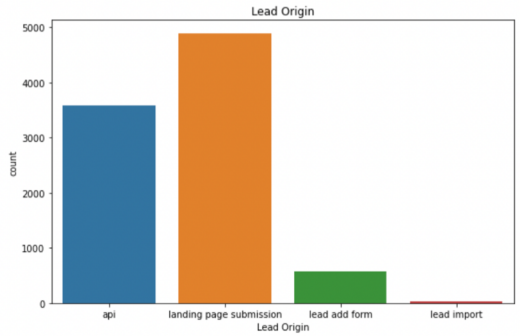


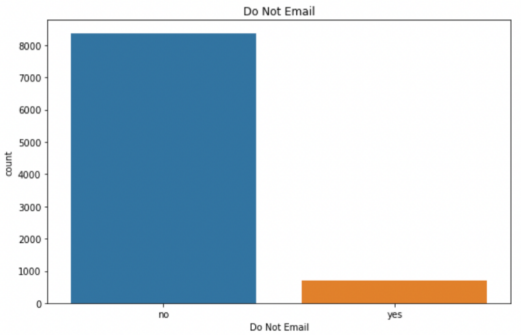


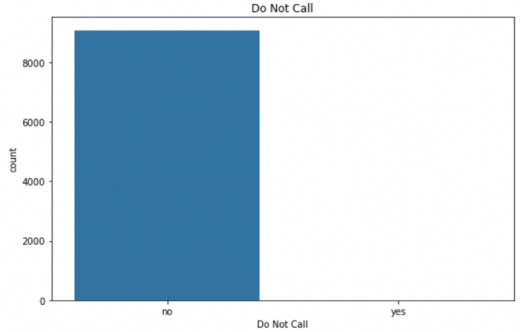


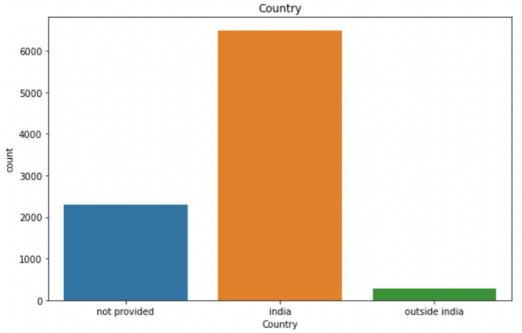




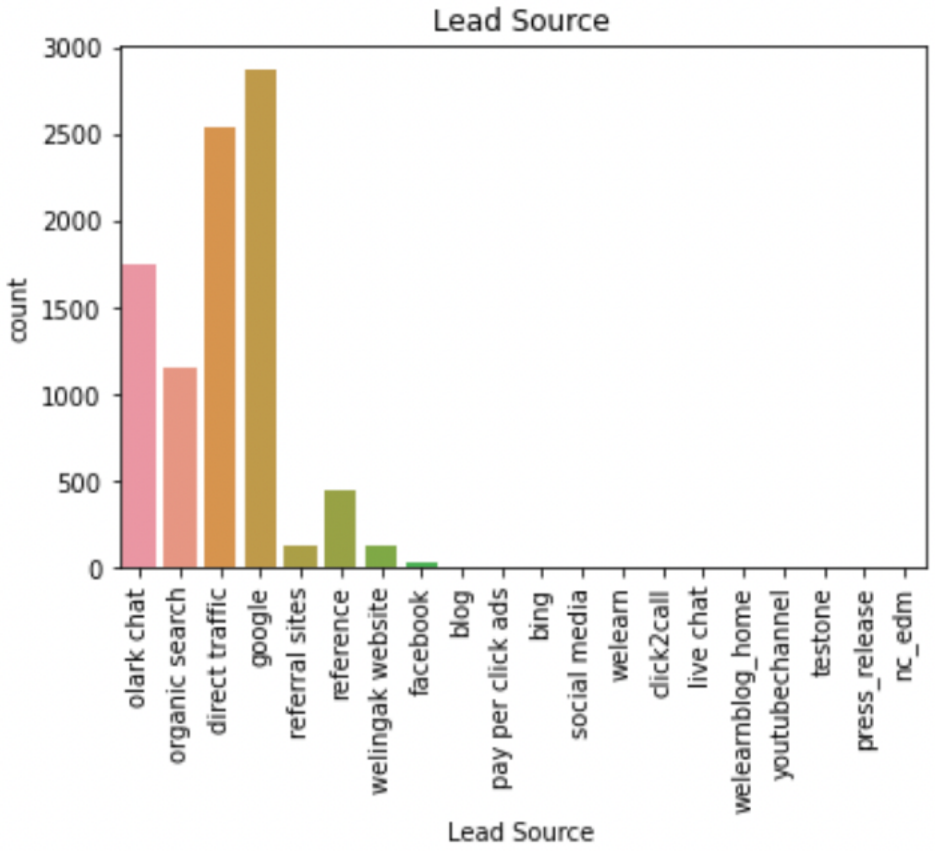








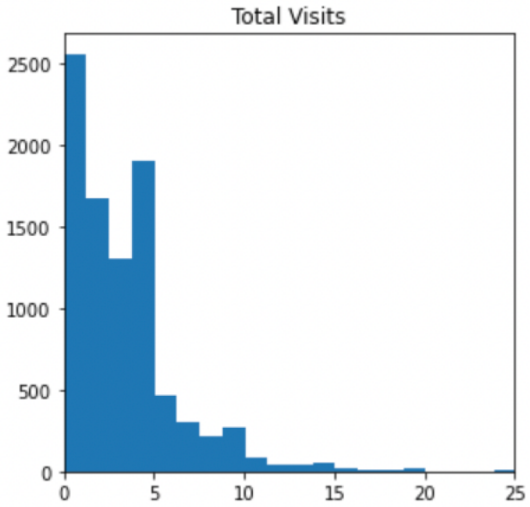
Lead Source

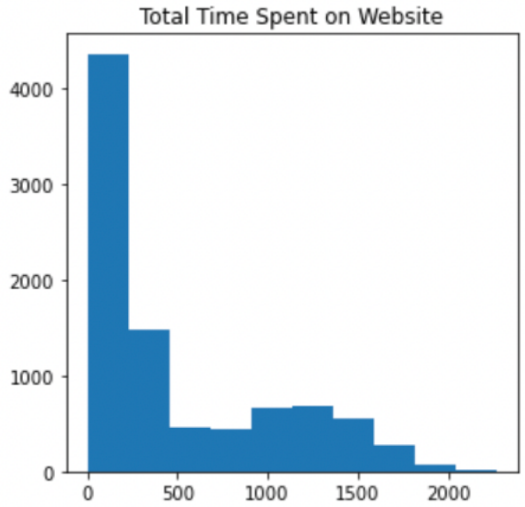


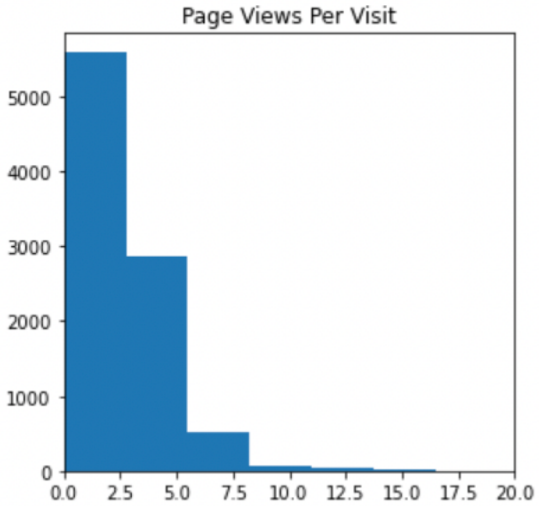
This figure shows us the platforms from where the website garners leads on X Education.

The platform sees highest traction from google followed by a direct traffic to the education portal and Olark chat.

Numerical Variables

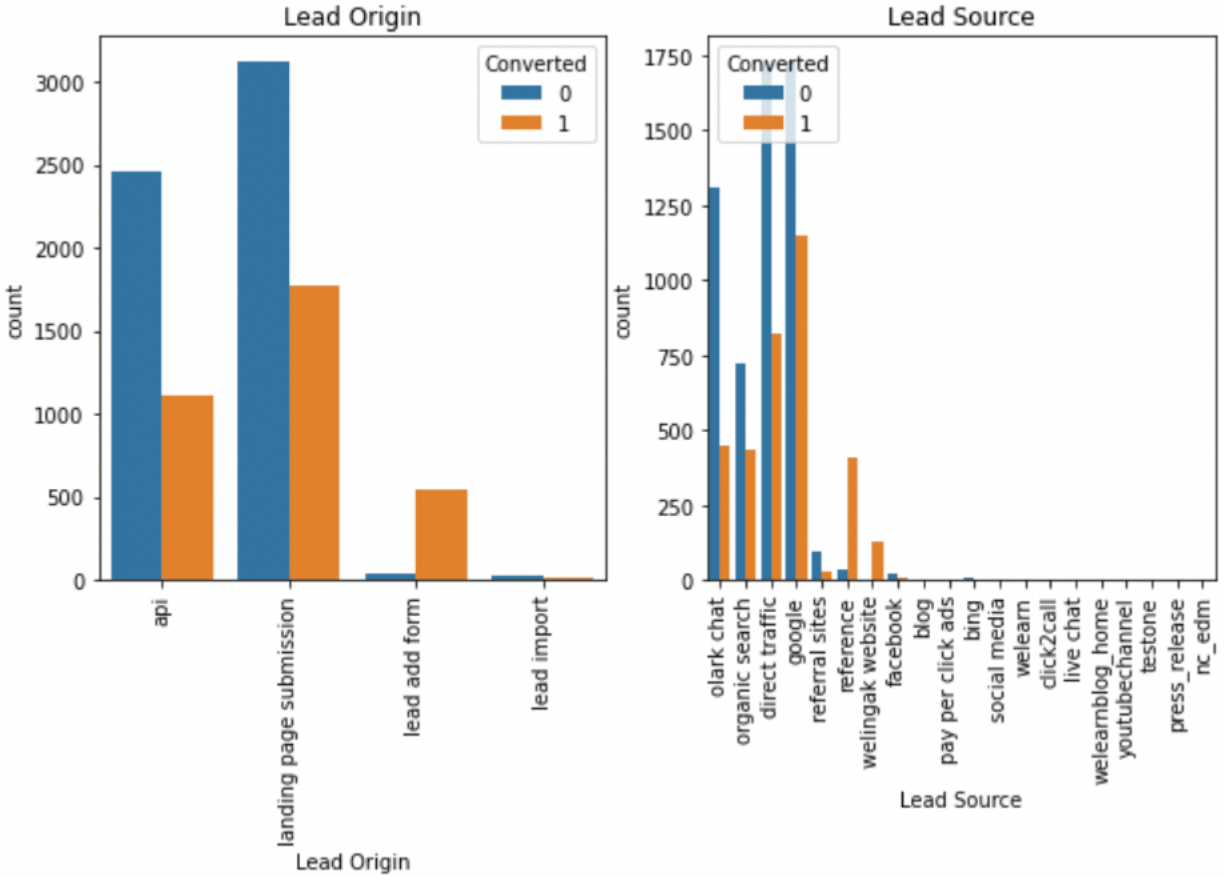






On the left are the numerical variables in the dataset which show us the time spent by the leads on the education platform's website, total number of visits and the page view per visit garnered.

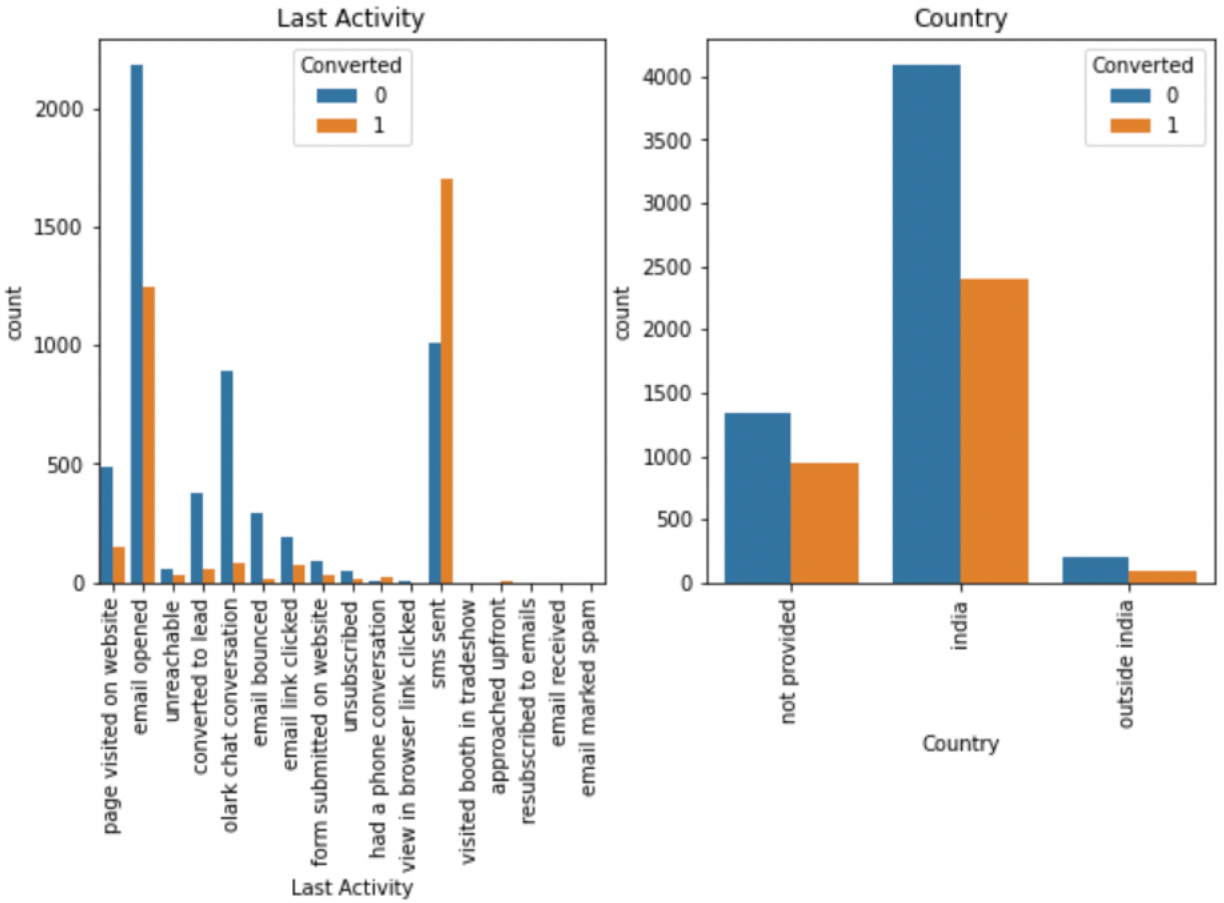
Converted leads



These graphs show us the lead origin and the lead source, and hence the converted leads on the basis of which of the leads are getting converted and which are not, depending on the origin and source.

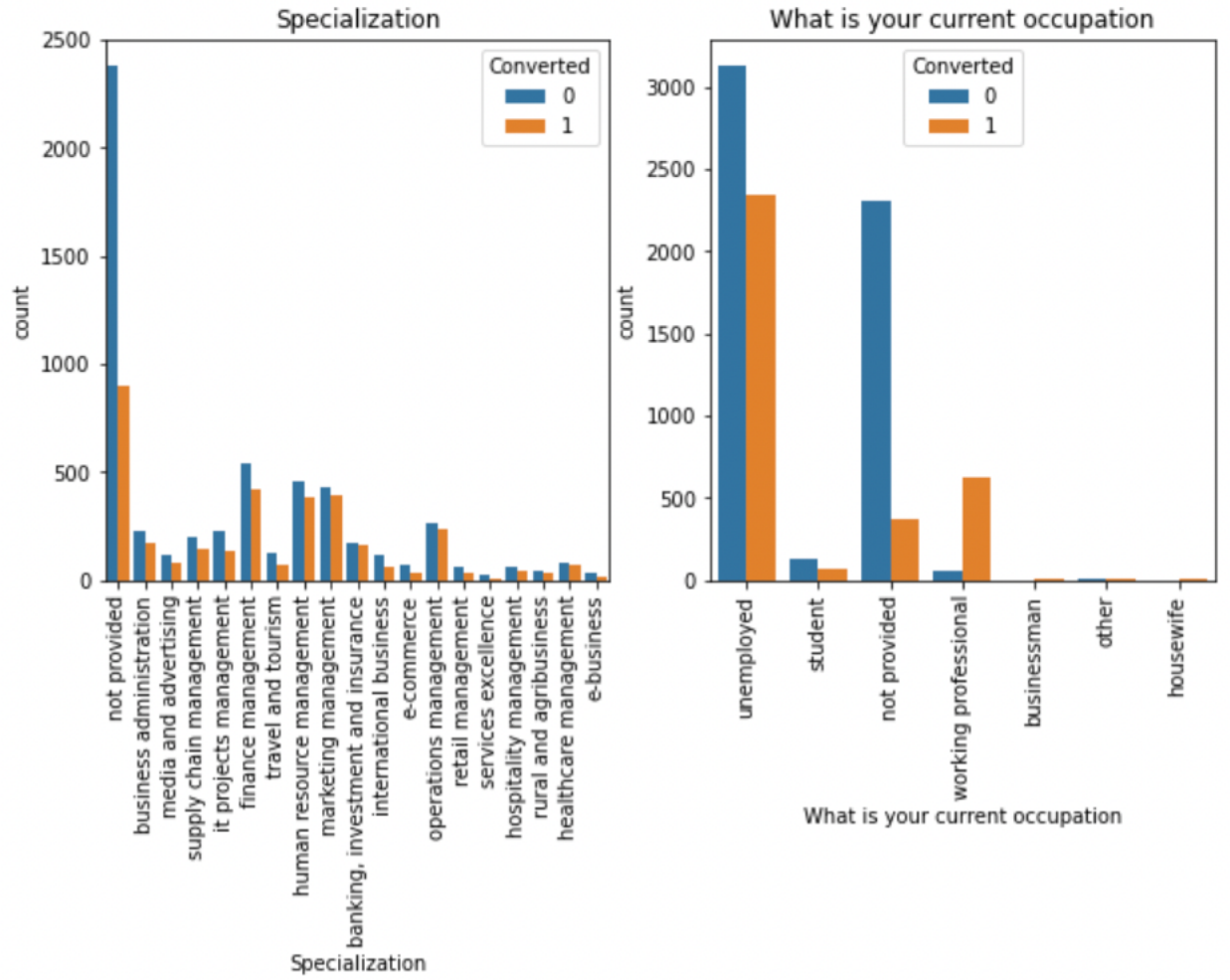
We see that the highest converted leads are garnered from google visits, followed by the direct traffic on the website.

Country of leads



We observe that the leads which are being converted mostly are from India. The last activity by all the leads has been highlighted in the first graph.

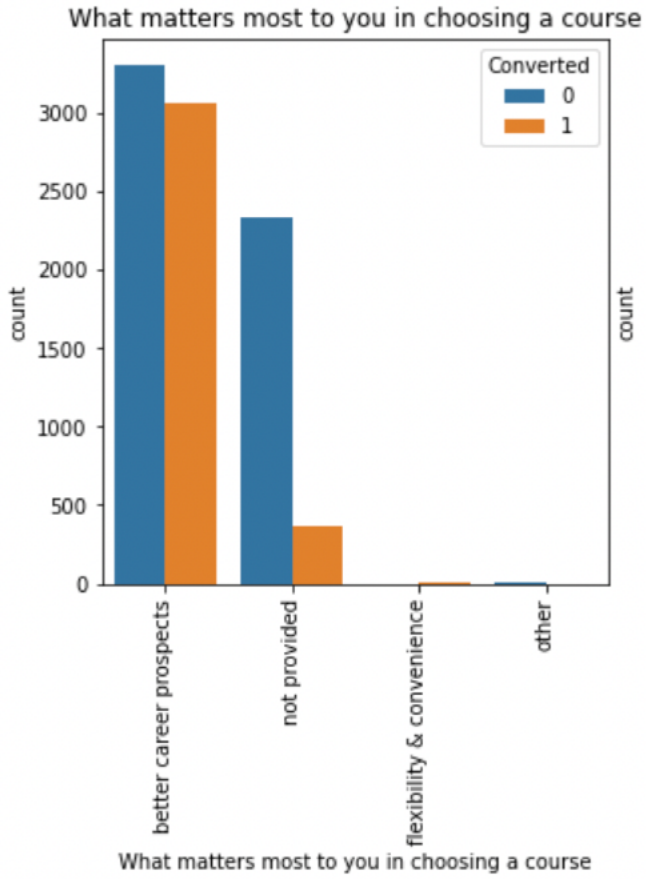
Occupation and Lead Specialization



Most of the leads subscribing to the online courses on X Education are Unemployed people who are pursuing a course. There are some of the working professionals as well who are interested in pursuing a course online while working side by side. This tells us that Unemployed leads are looking for upskilling themselves for better employment opportunities.

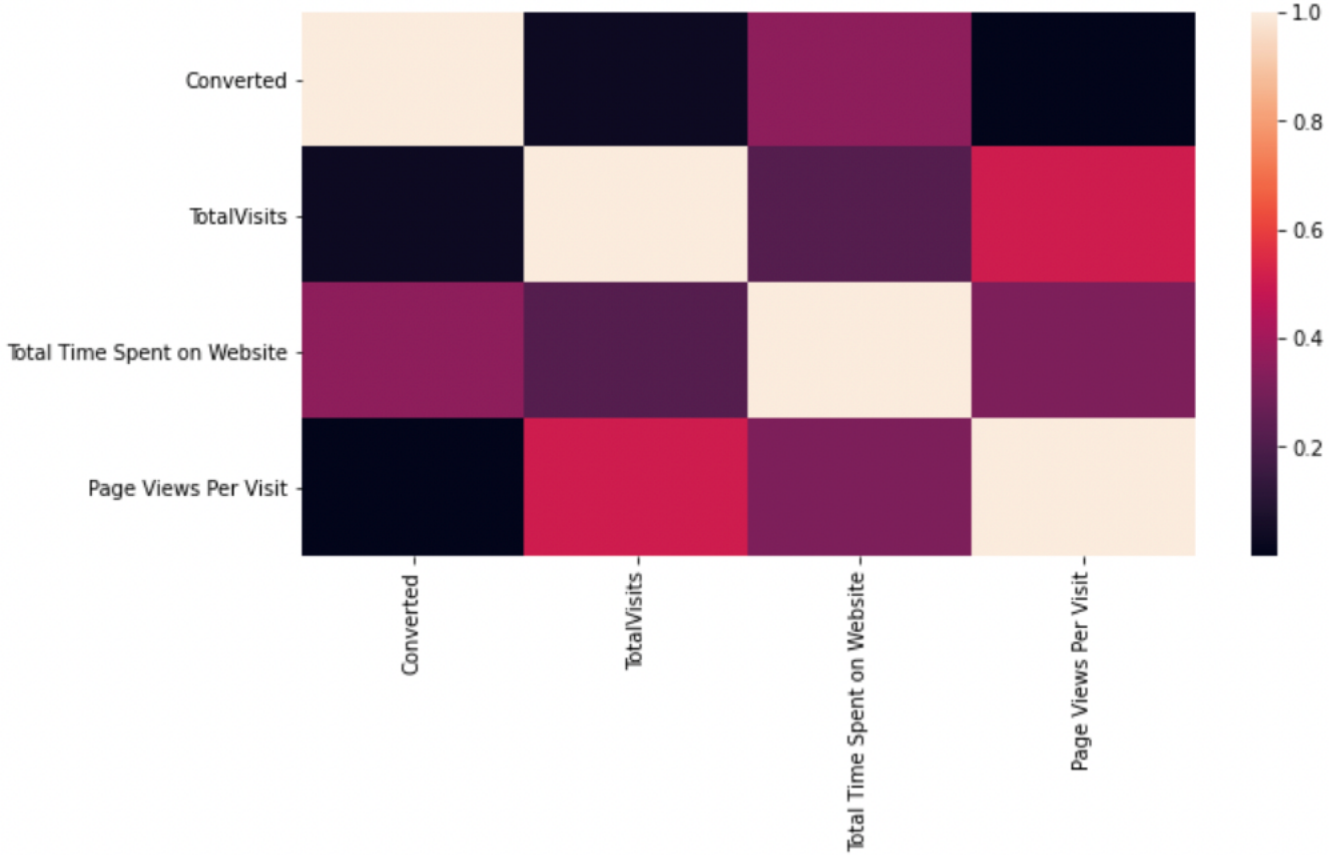
The first graph shows us the industries the leads belong to. The companies sales team can focus on the target audience of finance management and marketing for higher traction according to the data.

Course preference



We see that what matters the most to a lead who is getting converted, is the career prospects that are related to the course that they are pursuing, as the leads are pursuing the course for better career options.

Correlation among the variables



The leads which have a higher visit rate on the website of X Education are getting converted.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6335
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2635.0
Date:	Wed, 10 Nov 2021	**Deviance:**	5270.1
Time:	11:55:42	**Pearson chi2:**	6.48e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

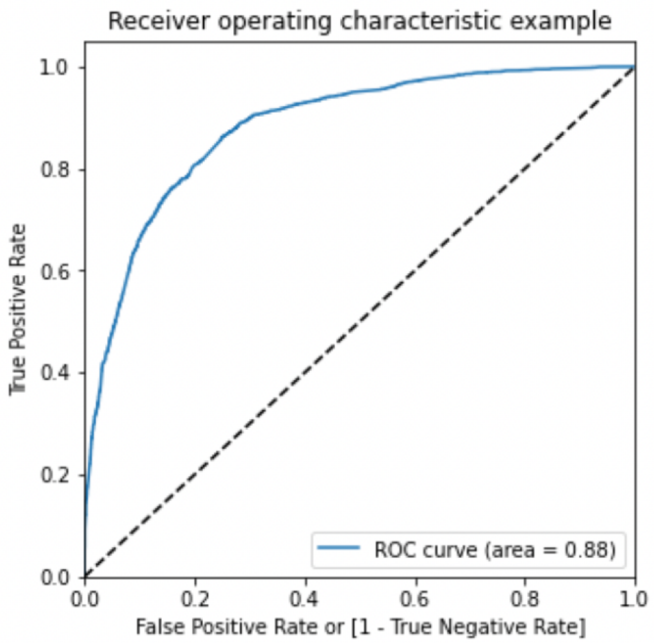
Data Conversion

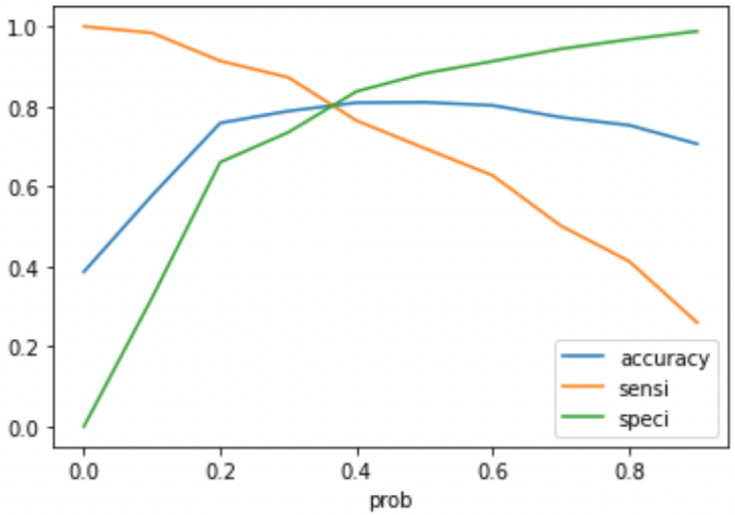
- In the process of Data conversion, we have normalised the Numerical Variables
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 37

Model Building

- During the model building, we have split the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, where we have chosen 70:30 ratio
- We have used RFE for Feature Selection
- We ran the RFE with 15 variables as output
- We are building the 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 of the model is 81%

ROC Curve





- **Finding the Optimal Cut off Point**
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35

Conclusion

- It was found that the variables that mattered the most in the potential buyers according to the descending order are;

- The total time spent on the Website

- Total number of visits

- When the lead source was:

a. Google

b. Direct traffic
c. Organic search
d. Welingak website

- And when the last activity was:

a. SMS

b. Olark chat conversation

- When the lead origin is Lead add format
- When their current occupation is as a working professional

- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.