Lead Score Case Study

By

Gaurav Sharma

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. Moreover, the company also gets leads through past referrals. 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%.
- Now, although 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.
- X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Objective of this case study

- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- So, we build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Data Description

- We have been provided with a leads dataset from the past with 9240 data points and 37 attributes.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- The target variable is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted..
- Another thing that need to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

Methodology

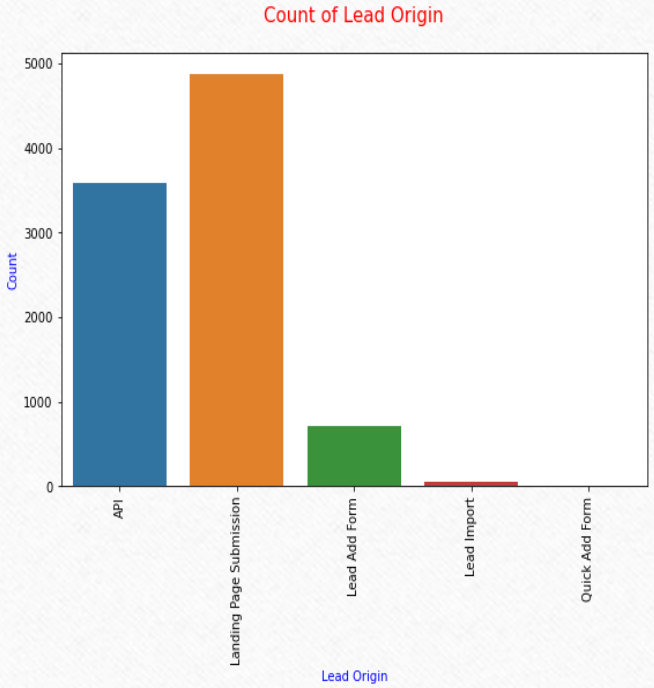
1. Reading and Understanding the Leads Dataset.
2. Data Cleaning -
   a) Handling the Missing Values, Data Imbalance and Outliers,
   b) Dropping the unnecessary data
3. Exploratory Data Analysis aka EDA.
4. Model Preparation –
   a) Dummy Variables
   b) Hot encoding of the data
5. Feature scaling
6. Model Building
7. Feature Selection Using RFE
8. Plotting the ROC Curve
9. Making the Prediction on the Test Data

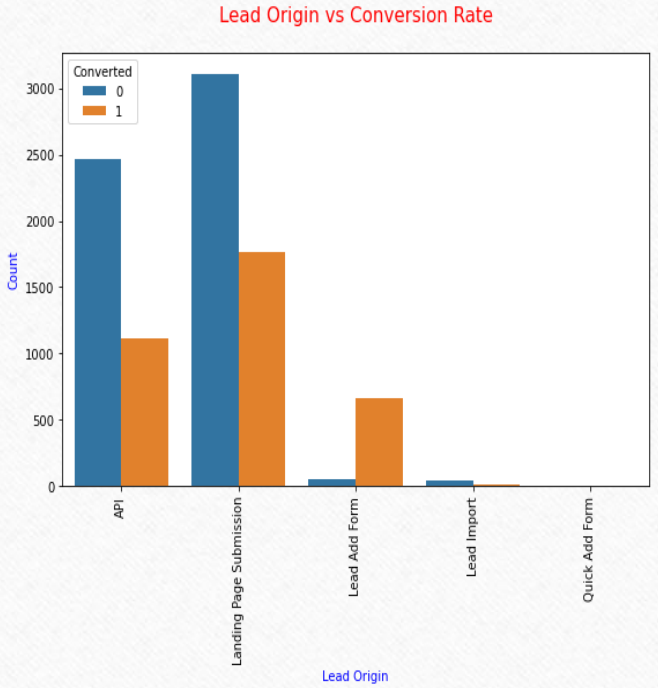
EDA

Exploratory Data Analysis

1. Univariate Analysis
2. Bivariate Analysis

1. Lead Origin

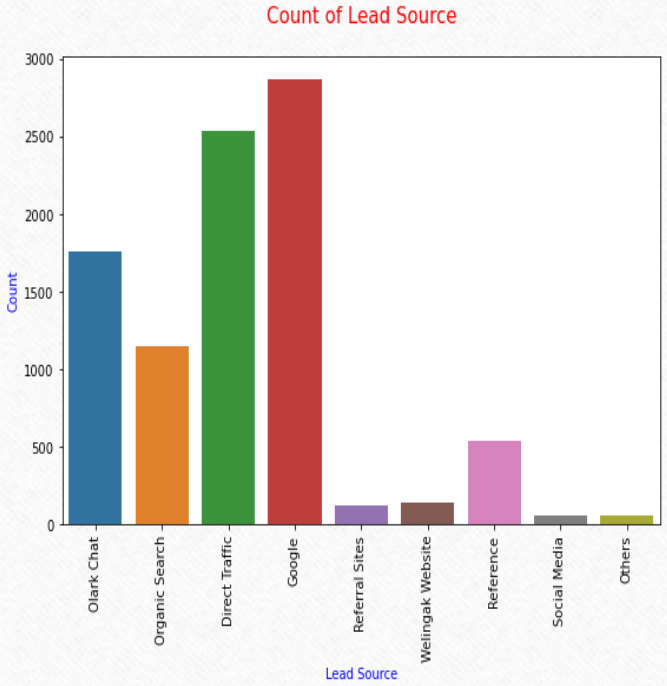


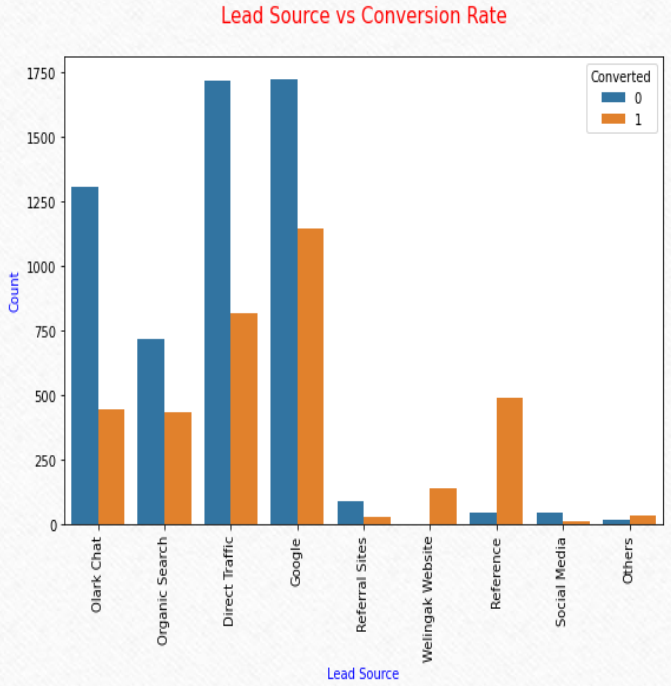


INFERENCE:

- The Maximum number of leads originates by the `Landing Page Submission` origin.
- Almost 70-80% Leads originate by `Landing Page Submission` and `API`.
- Clearly infer from above plots that The Lead Conversion rate of `Lead Origin` is higher than other origins. But Maximum leads identifies by `Landing Page Submission` probably due to it originates maximum number of leads.

2. Lead Source

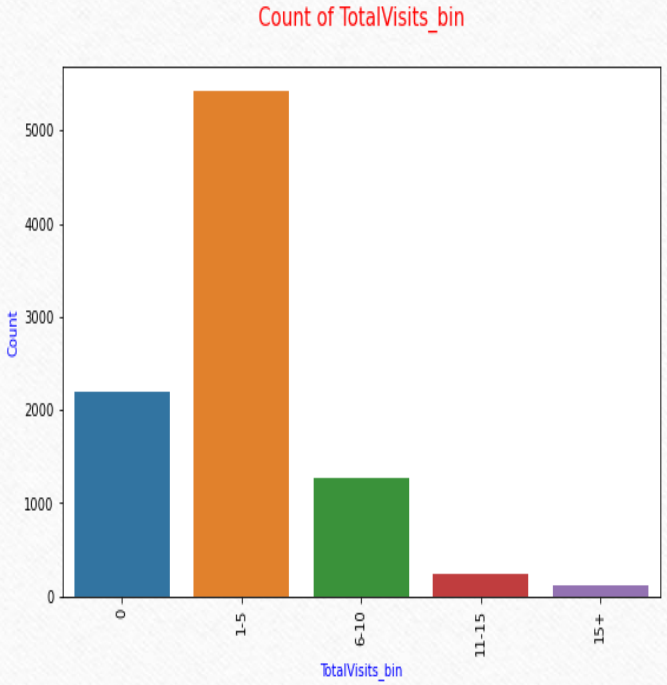


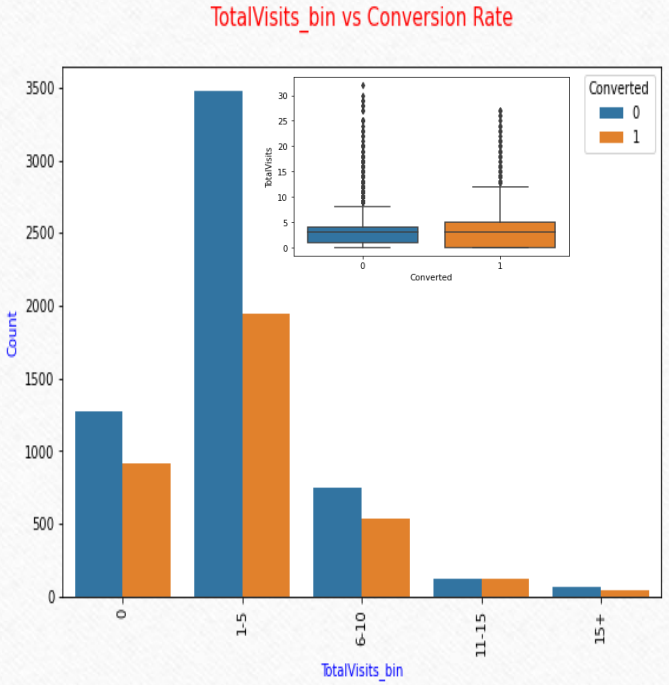


INFERENCE:

- `Google` is the Primary source of leads generated for X Education.
- Apart from `Google` source, Almost 50% Leads are from `Direct Traffic` and `Olark Chat`.
- Clearly infers from above plots that The Conversion rate of `Reference` and `Welingak Website` is better than other origins. But Number of Lead Conversions are from `Google` and `Direct Traffic` Lead sources.

3. Total Visits made by Customers

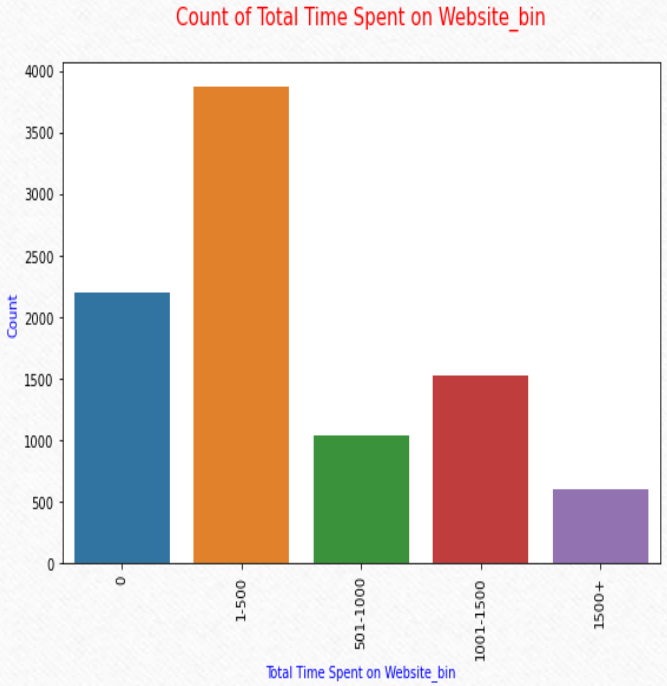


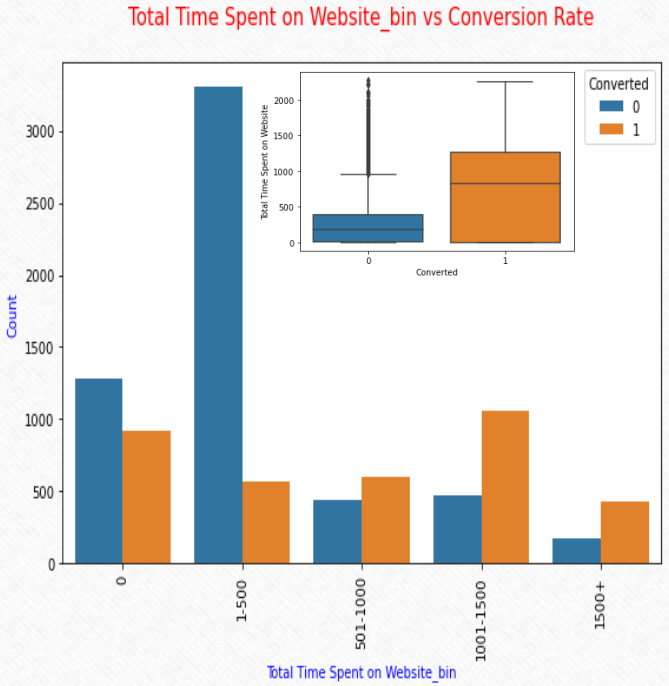


INFERENCE:

- The high numbers of the Customers visit the website by `1-5` only.
- As we clearly infer from above plots that The Maximum Conversions of customer likely visits the website only `1-5` times.
- Around 85-90% of customers visits `1-10` times before making their decision.

4. Total Time Spent on Website

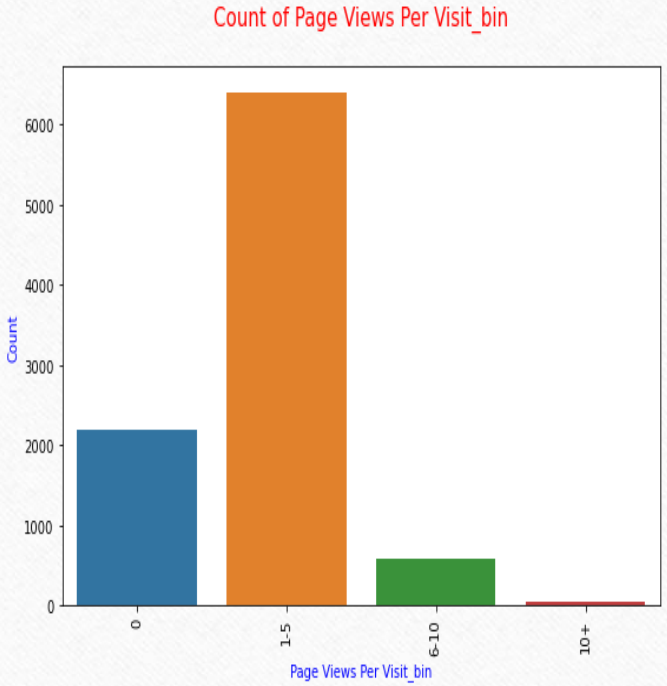


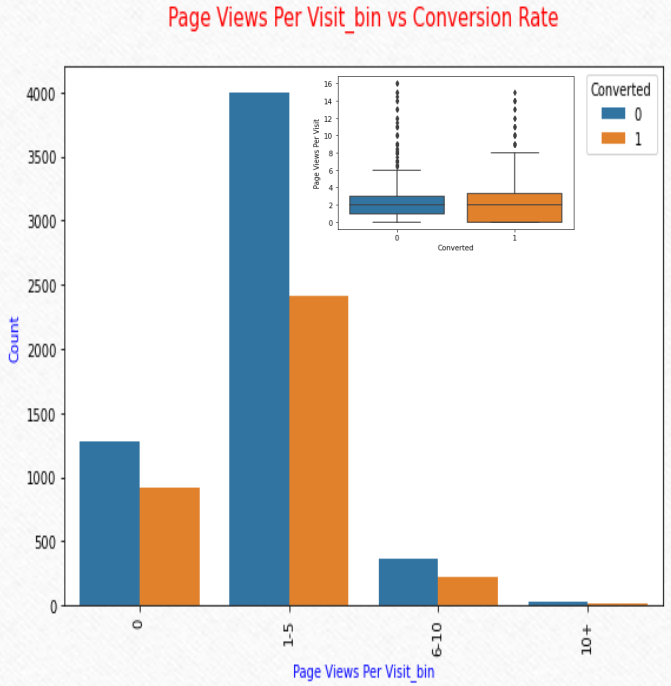


INFERENCE:

- As clearly infer from box plot that Maximum customers has tendency of spending the long time on website.
- The most of customers spent time till `1-500` on the website.
- We can also that the longer the customer spent on time on website, the higher the chances of lead conversion.

5. Avg. Page Views Per Visit by the Customers

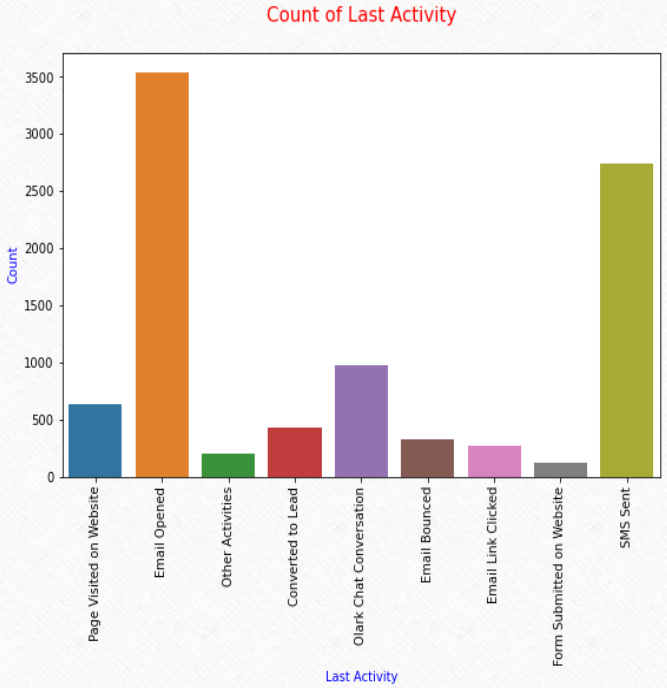


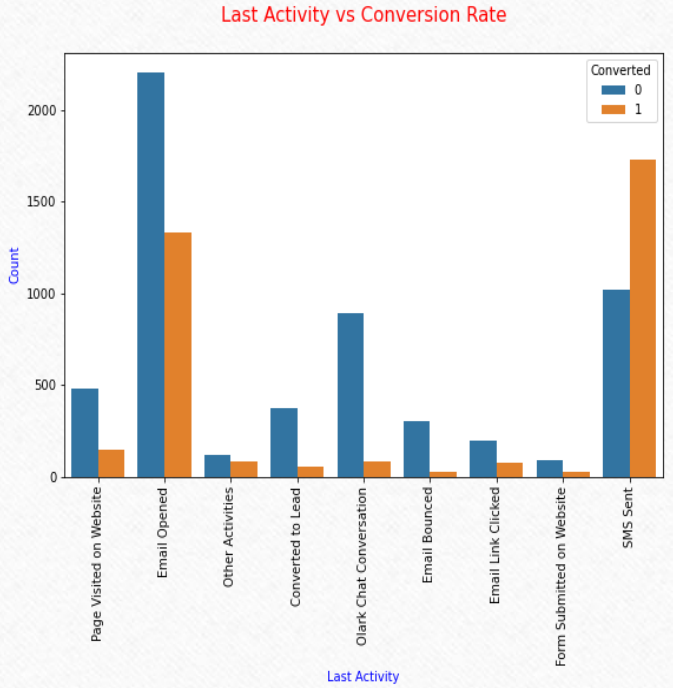


INFERENCE:

- We clearly infer from the plots that the most of customers like to go to 2-4 pages to view.
- Most of the leads converted to customers shows that they view at least 1-5 pages and spend time on the website.

6. Last Activity performed by Customers

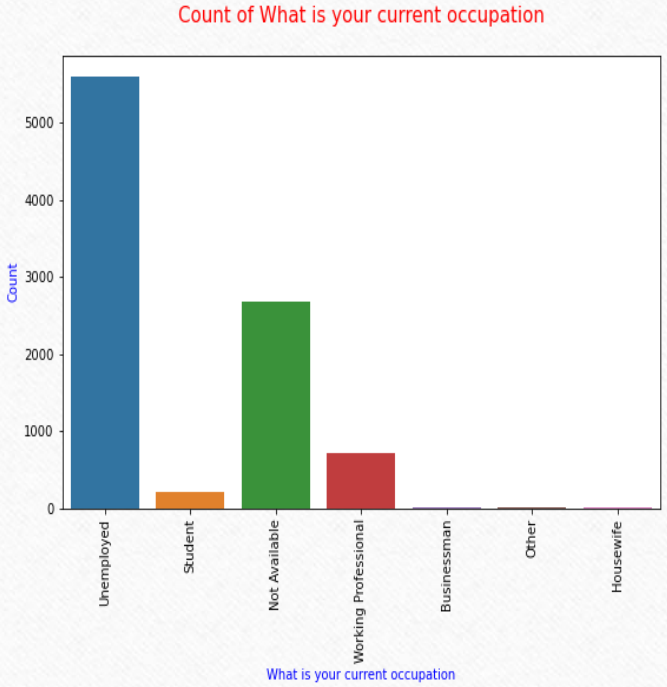


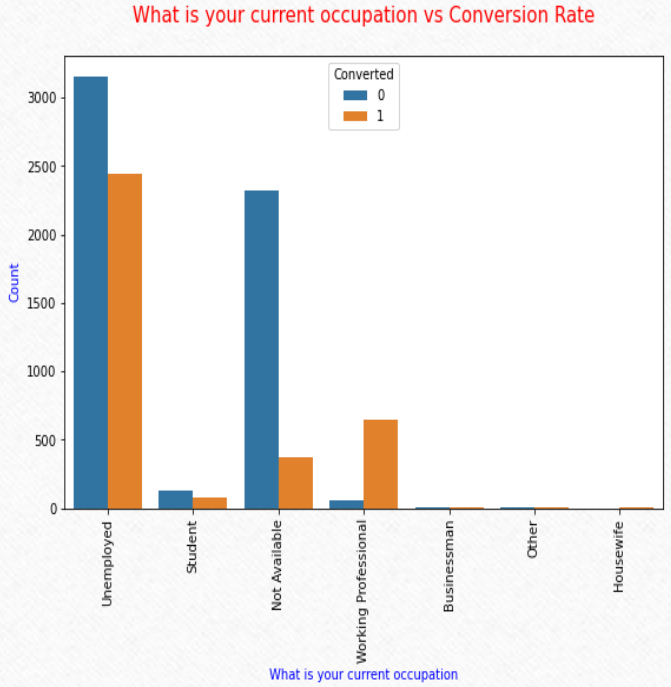


INFERENCE:

- The most of Customer Leads tends to check their `open their email` before making the decision.
- As it can clearly infer from above plots that The Lead conversion rate of `SMS Sent` activity is higher than other activities.

7. Current Occupation of the Customers

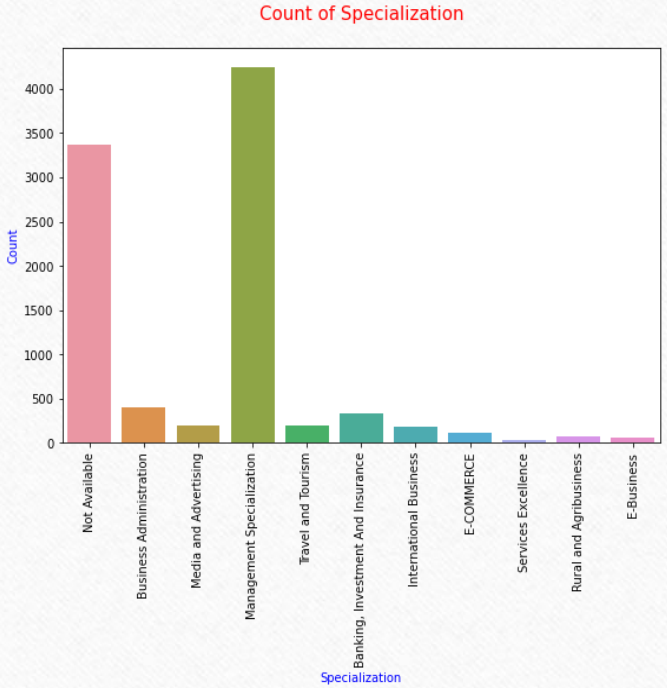


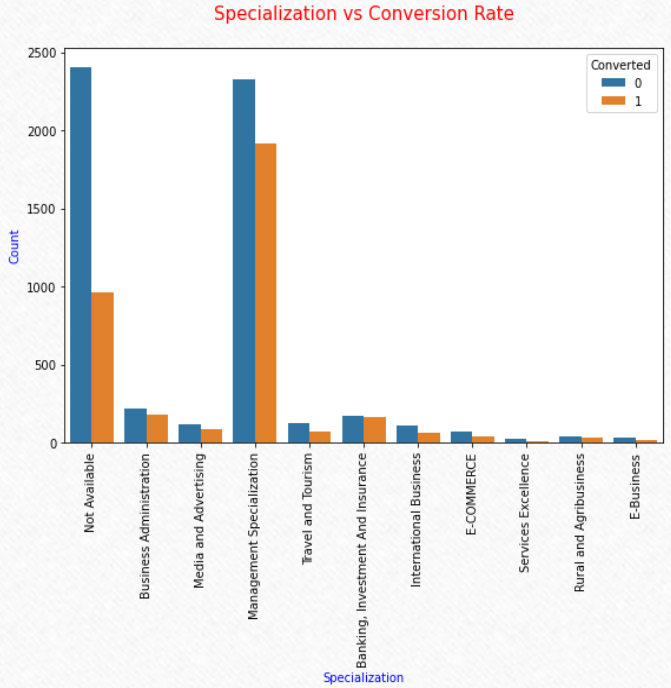


INFERENCE:

- From above figure, we can infer that most of leads of customers are from the `Unemployed` category.
- 2. We can also infer from above plots that The `Working Professional` likes the X education's courses better than customer with other occupations and has higher conversion rate.

8. Specialization of the Customers

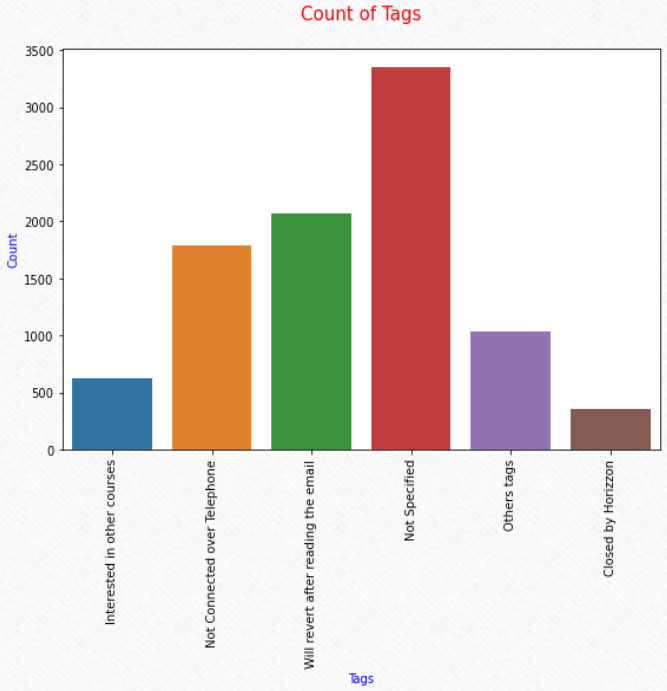


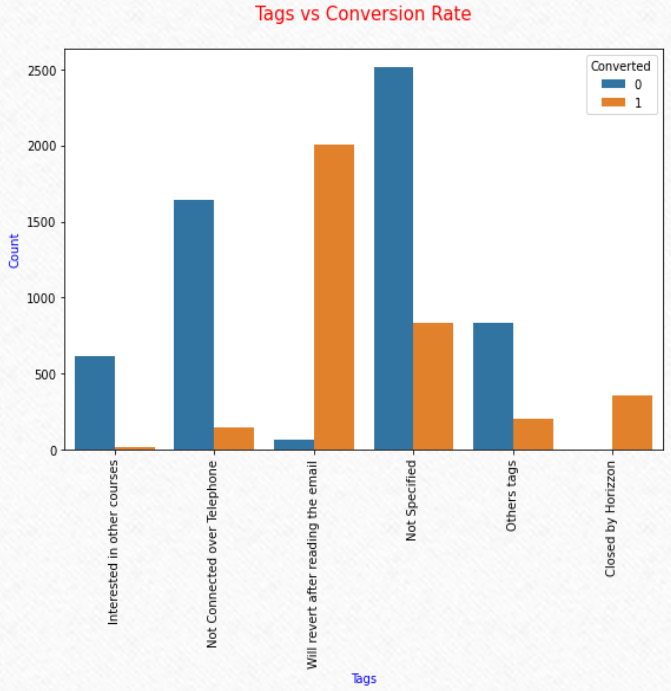


INFERENCE:

- From above figure, we can infer that most of leads of customers are from the `Management Specialization` who shows interest in X Education's courses.
- We can also infer from above plots that The `Management Specialization` has good conversion rate as well.
- The Other Specializations such as `Business Administration` and `Banking, Investment and Insurance` also have good conversion rate from other specializations.

9. Tags Assigned – Lead Status

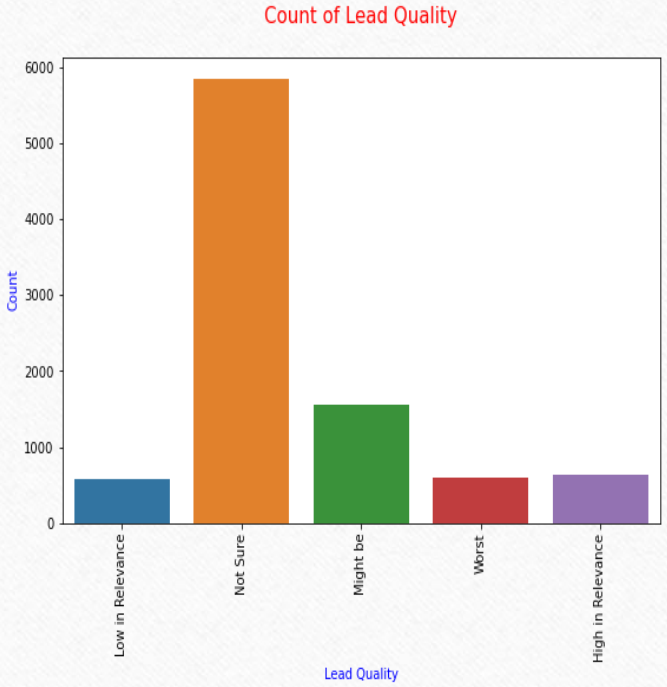


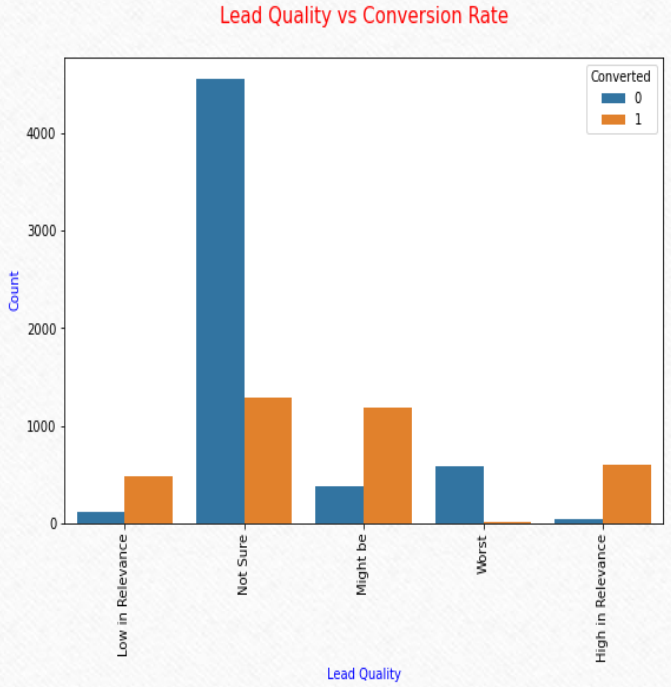


INFERENCE:

- From the above plots, we can infers that the most of customers tends to `revert after reading email`. So, X education Should more focus on such leads who says that they revert after the reading the email.

10. Lead Quality - Remark assigned by Representatives

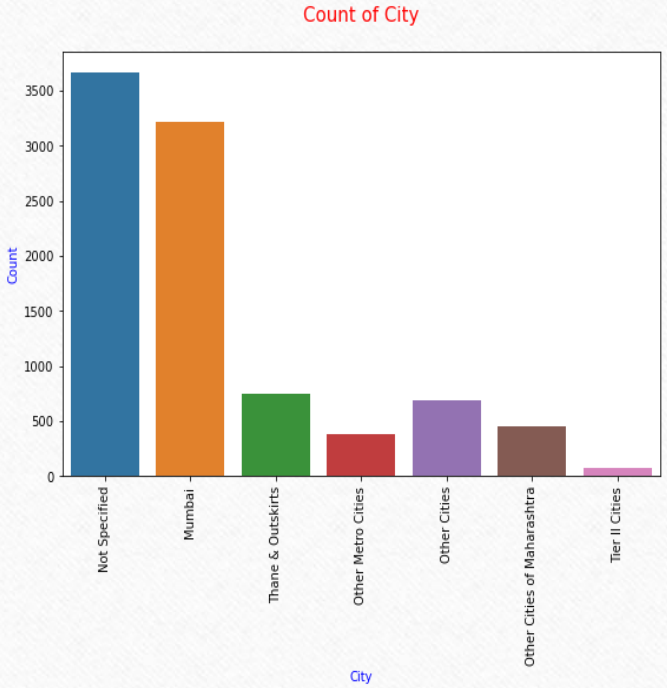


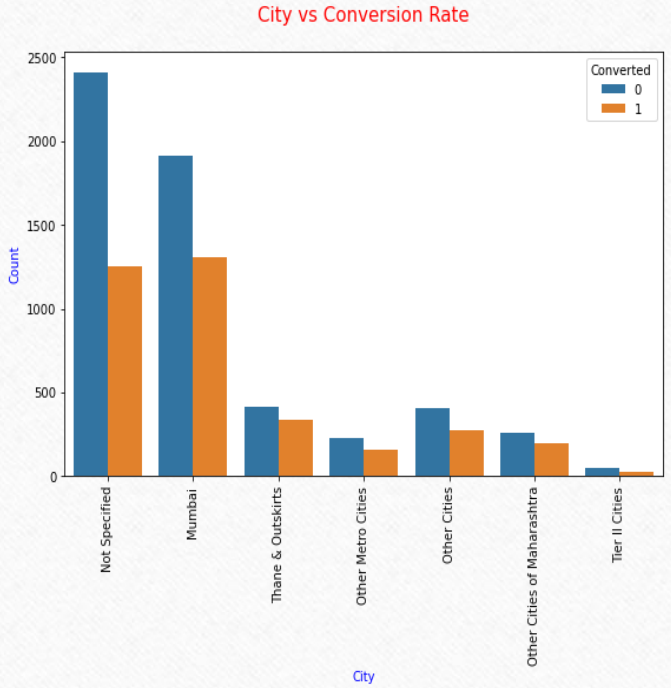


INFERENCE:

- The most of the customer leads are in `Not Sure` categories. They could be in confused minds to go for the course or not.
- Secondly, The `High in Relevance` category's leads have highest conversion rate and the `Worst` has worst lead conversion rate.

11. City of the Customers

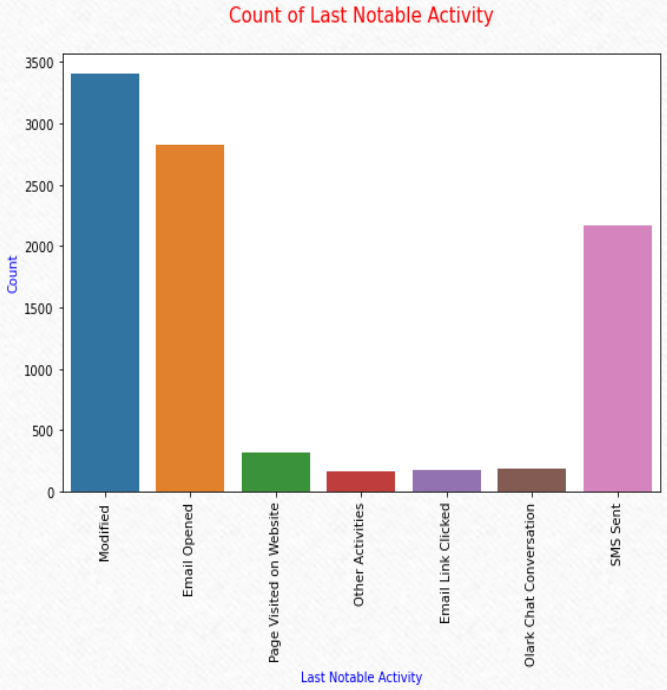


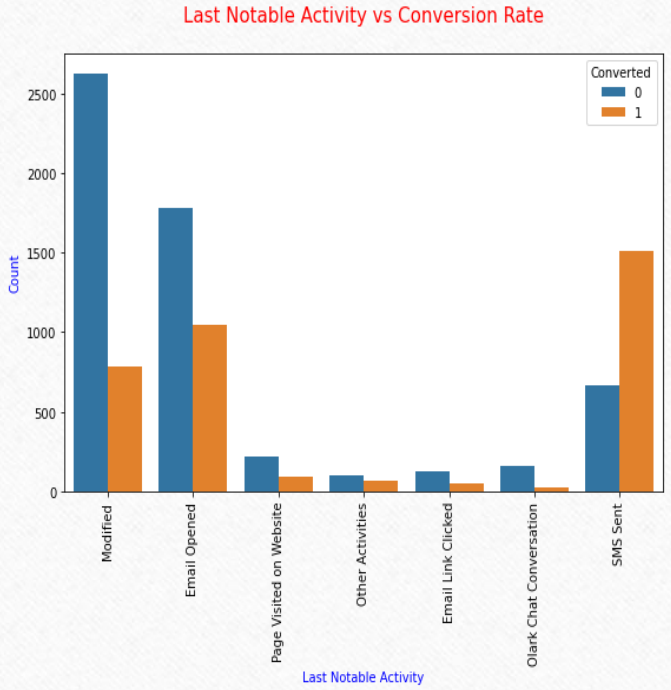


INFERENCE:

- As we can see that the most of customer's leads are from `Mumbai` city and have good lead conversion rate. so, X Education should focus more on the leads of customers who based in Mumbai city.

12. Last Notable Activity performed by Students





INFERENCE:

- The most of leads of Students are from `Modified` category as their `Last Notable Activity`. However, it can clearly observed that highest conversion comes out from the students who's `Last Notable Activity` are recorded as `SMS Sent`.

Logistic Regression Model

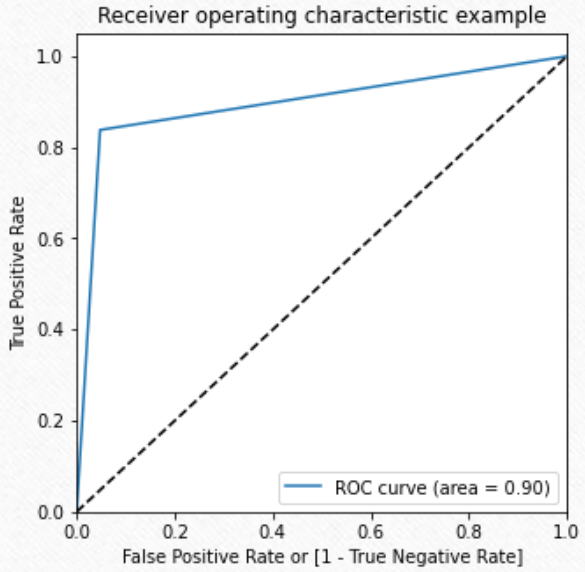
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6461
Model:	GLM	Df Residuals:	6444
Model Family:	Binomial	Df Model:	16
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1431.1
Date:	Mon, 11 Oct 2021	Deviance:	2862.2
Time:	15:08:50	Pearson chi2:	6.52e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

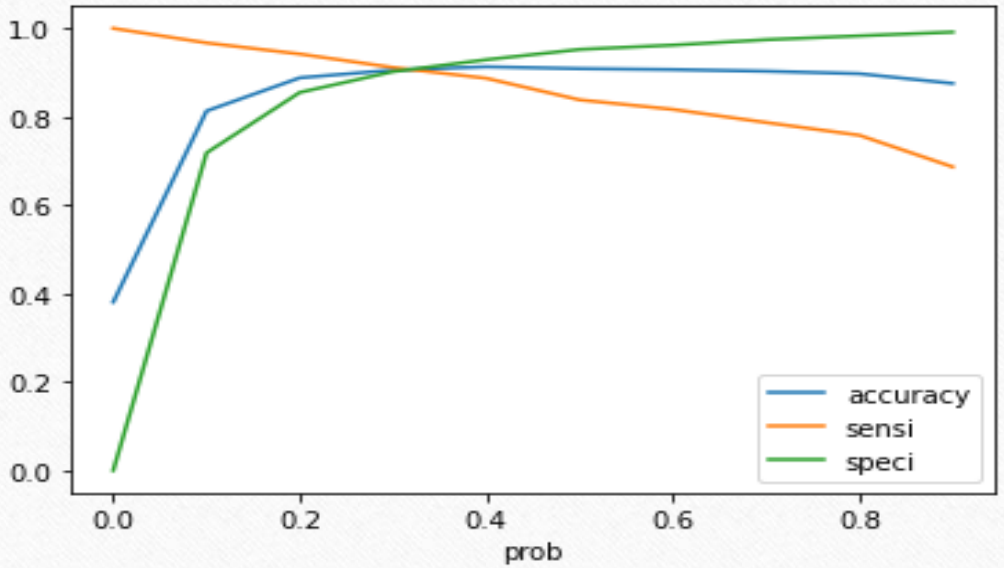
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7723	0.069	-11.116	0.000	-0.908	-0.636
Total Time Spent on Website	1.0948	0.055	19.925	0.000	0.987	1.203
Lead_Origin_Lead Add Form	0.2809	0.076	3.694	0.000	0.132	0.430
Lead_Source_Olark Chat	0.4244	0.056	7.642	0.000	0.316	0.533
Lead_Source_Welingak Website	0.3743	0.094	3.962	0.000	0.189	0.559
Last_Activity_Email Link Clicked	0.2764	0.085	3.252	0.001	0.110	0.443
Last_Activity_Email Opened	0.1852	0.081	2.280	0.023	0.026	0.344
Last_Activity_Olark Chat Conversation	-0.1616	0.077	-2.102	0.036	-0.312	-0.011
Last_Activity_SMS Sent	0.9737	0.077	12.577	0.000	0.822	1.125
current_occupation_Not Available	-0.8586	0.056	-15.404	0.000	-0.968	-0.749
Tags_Closed by Horizzon	1.0012	0.139	7.205	0.000	0.729	1.274
Tags_Interested in other courses	-0.8251	0.090	-9.132	0.000	-1.002	-0.648
Tags_Not Connected over Telephone	-1.2776	0.064	-19.828	0.000	-1.404	-1.151
Tags_Will revert after reading the email	1.3198	0.076	17.321	0.000	1.170	1.469
Lead_Quality_Worst	-0.8805	0.110	-8.027	0.000	-1.096	-0.666
Notable_Activity_Email Link Clicked	-0.2625	0.085	-3.075	0.002	-0.430	-0.095
Notable_Activity_Modified	-0.4837	0.059	-8.197	0.000	-0.599	-0.368

Model Validation

ROC Curve (area = 0.90)



Optimal Cut-off Point at 0.35



Final Observations

	Metrics	Trains_Values	%age	TestSet_Values	%age_test
0	Accuracy	0.909612	90.961	0.908631	90.863
1	Sensitivity	0.895528	89.553	0.901729	90.173
2	Specificity	0.918270	91.827	0.913174	91.317
3	False Positive Rate	0.081730	8.173	0.086826	8.683
4	Positive Predictive Value	0.870751	87.075	0.872359	87.236
5	Negative Predictive Value	0.934622	93.462	0.933864	93.386

Recommendations

X Education should put more focus on following to increase the Lead conversion rate

- When Lead identifies by the "Lead Add Form" as the Origin.
- When Lead Source are "Google", "Direct Traffic", "Welingak website" and "Cloak Chat"
- The Potential buyers like to visit the website frequently and spend time on the website.
- When the last activity performed by customers is "SMS Sent" and "Email Clicked/Open".
- When the Status of the Leads is "Will revert after reading the email" and "Closed by horizon".
- When the leads are of "Working Professional" in the "Management Specialization" or in the "Banking, investment and Insurance" field as their industry domain.
- For any Student leads, if "SMS Sent" performed by student as their last notable activity are the potential buyers of online course offered by X educations.