Lead Score Case study

Group Members

Rahul patil

Nithin Vijayabaskaran

Pratik Joshi

Problem Statement

- X Education sells online courses to industry professionals.
- 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.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

- **Data cleaning and data manipulation.**

1. Handle duplicate data and check the whole set.
2. NA values and missing values need to be handled.
3. If a column contains large number of missing values and not useful for analysis those columns to be dropped.
4. Outliers in the data need to be verified and handled.

- **EDA**

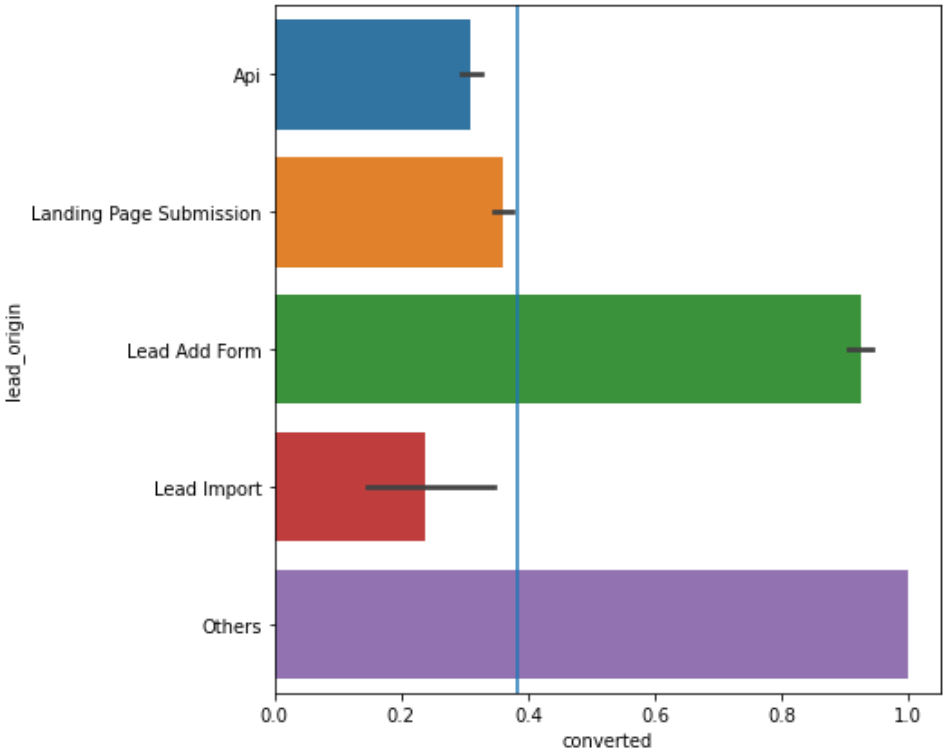
1. Univariate data analysis
2. Bivariate data analysis

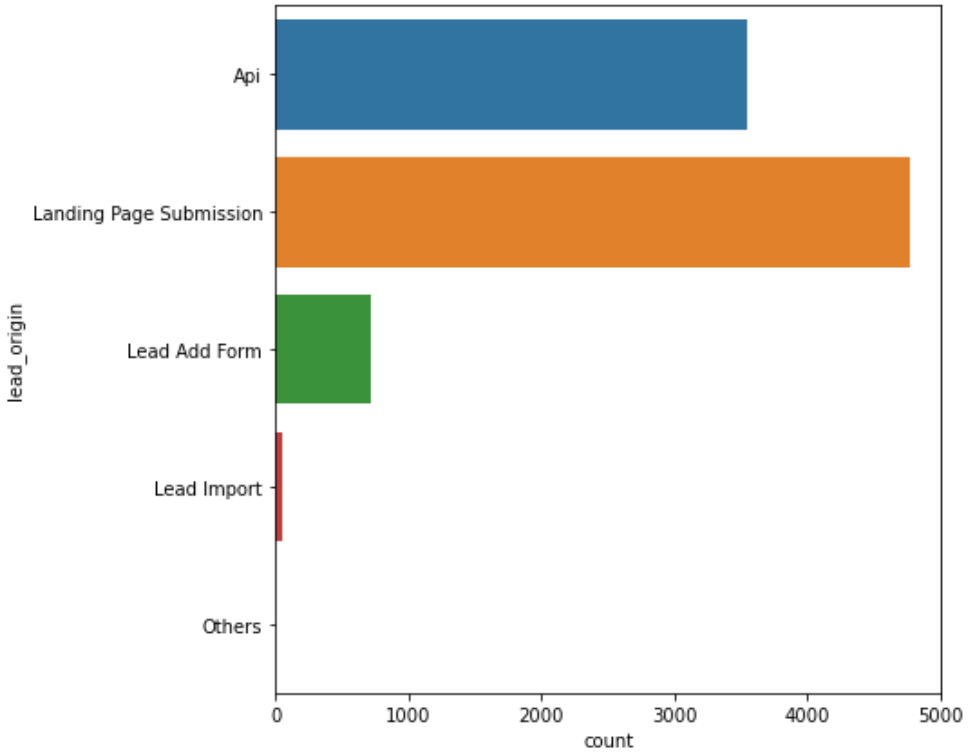
- **Dummy Variables and encoding of the data.**
- **Classification technique: logistic regression used for the model making and prediction.**
- **Model building and validation.**
- **Conclusions and recommendations.**

Data Manipulation

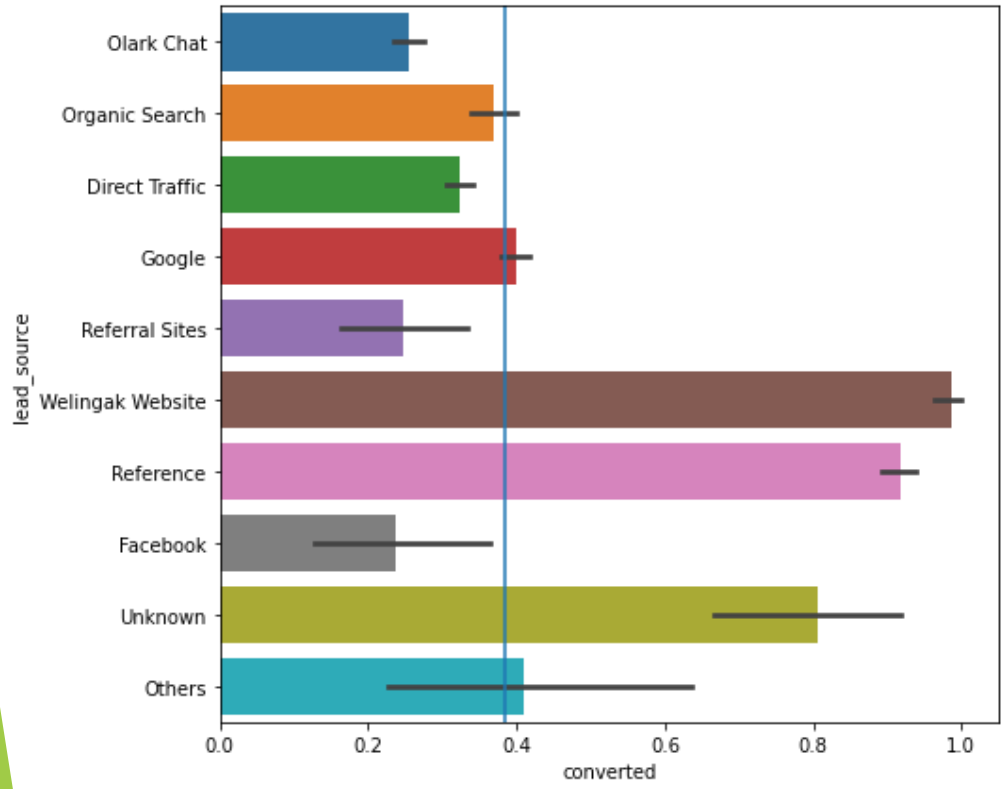
- Total Number of Rows =32, Total Number of Columns =9240.
- Dropping the columns having more than 20% as missing value.

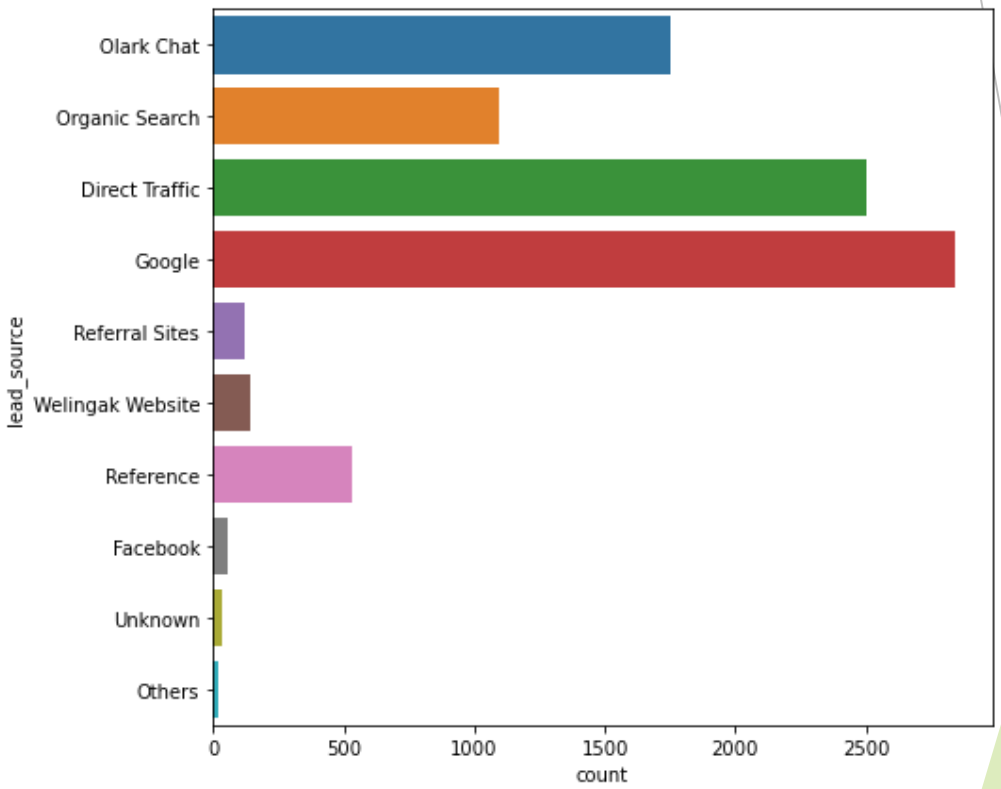
EDA

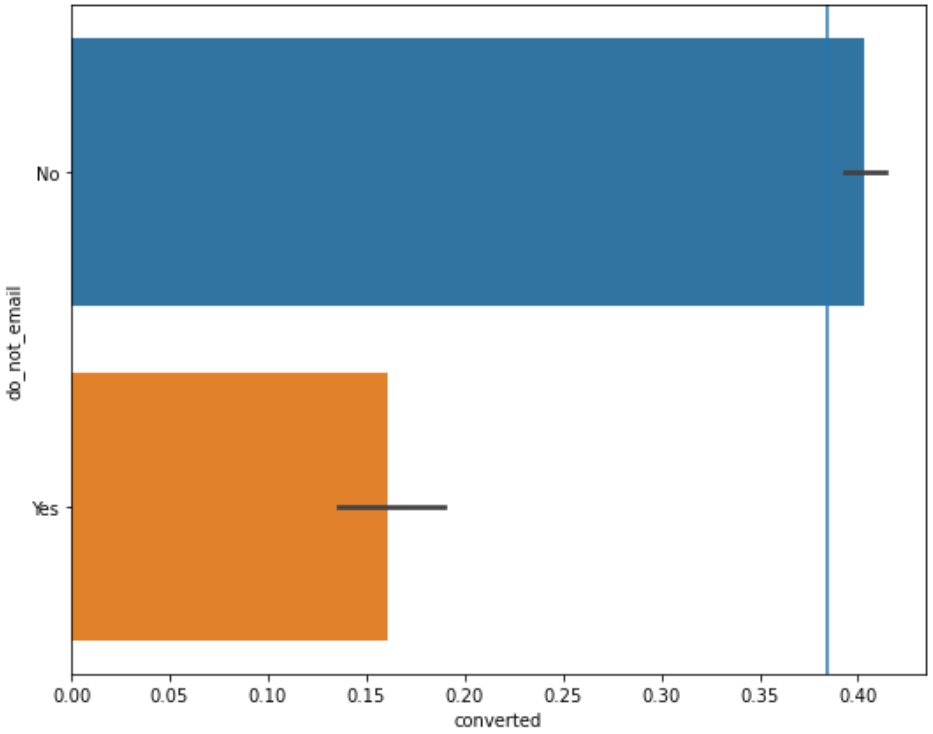

Api
Landing Page Submission
Lead Add Form
Lead Import
Others
lead_origin
0.0
0.2
0.4
0.6
0.8
1.0
converted

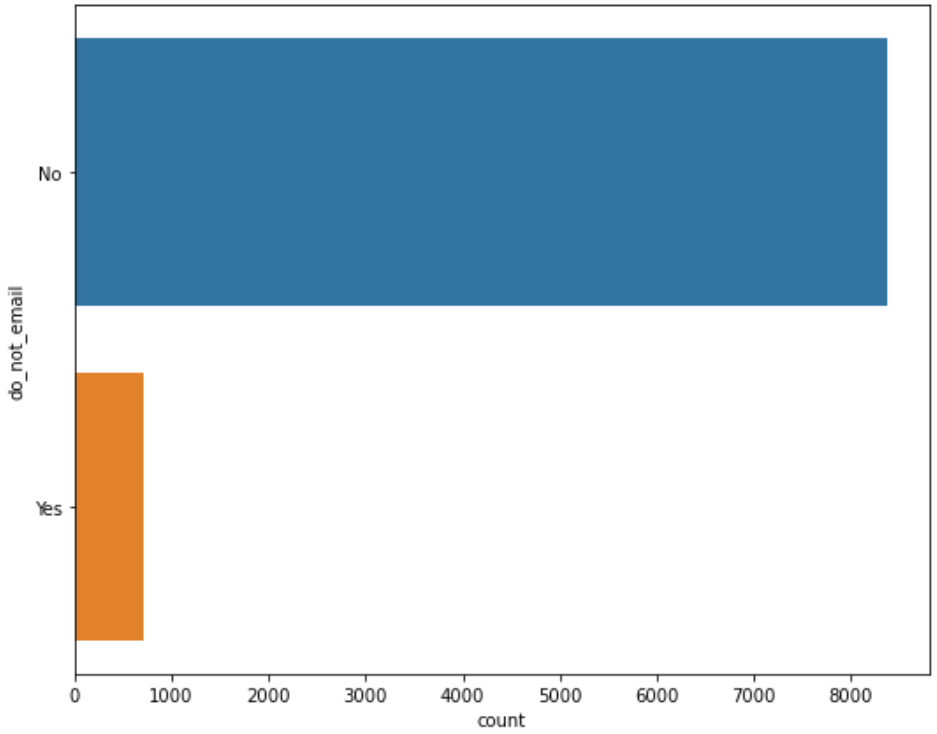

Api
Landing Page Submission
Lead Add Form
Lead Import
Others
lead_origin
0
1000
2000
3000
4000
5000
count

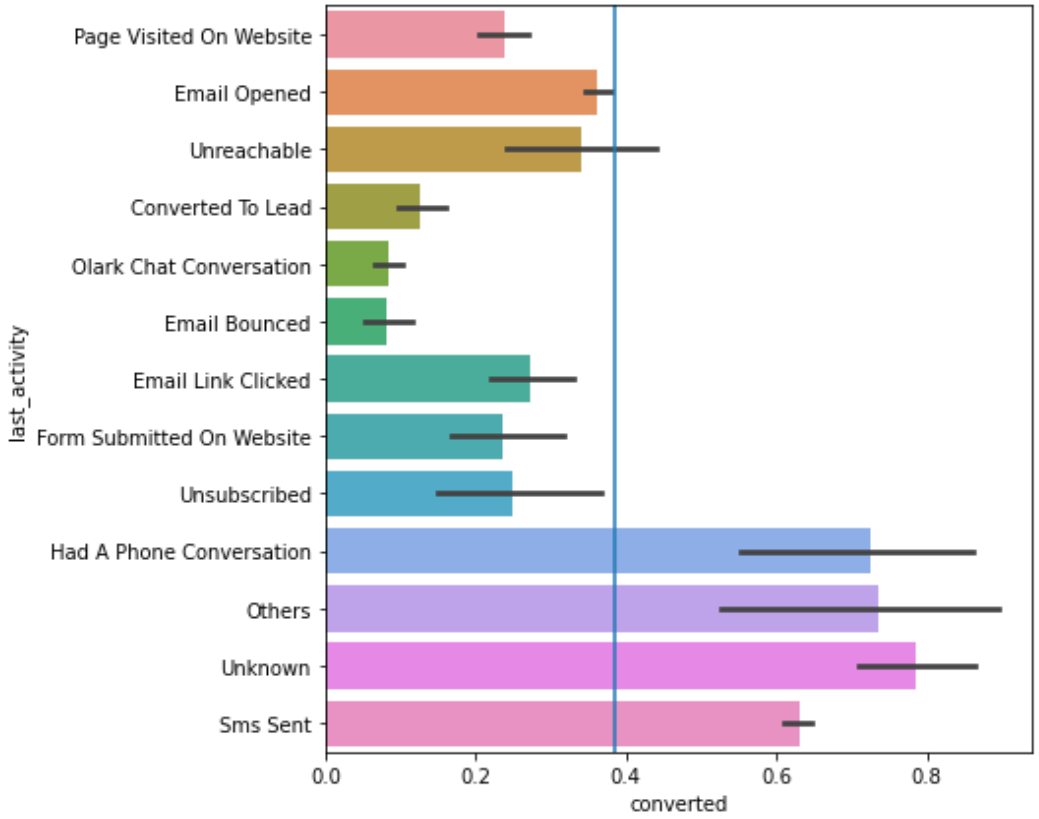
Categorical variable Relation

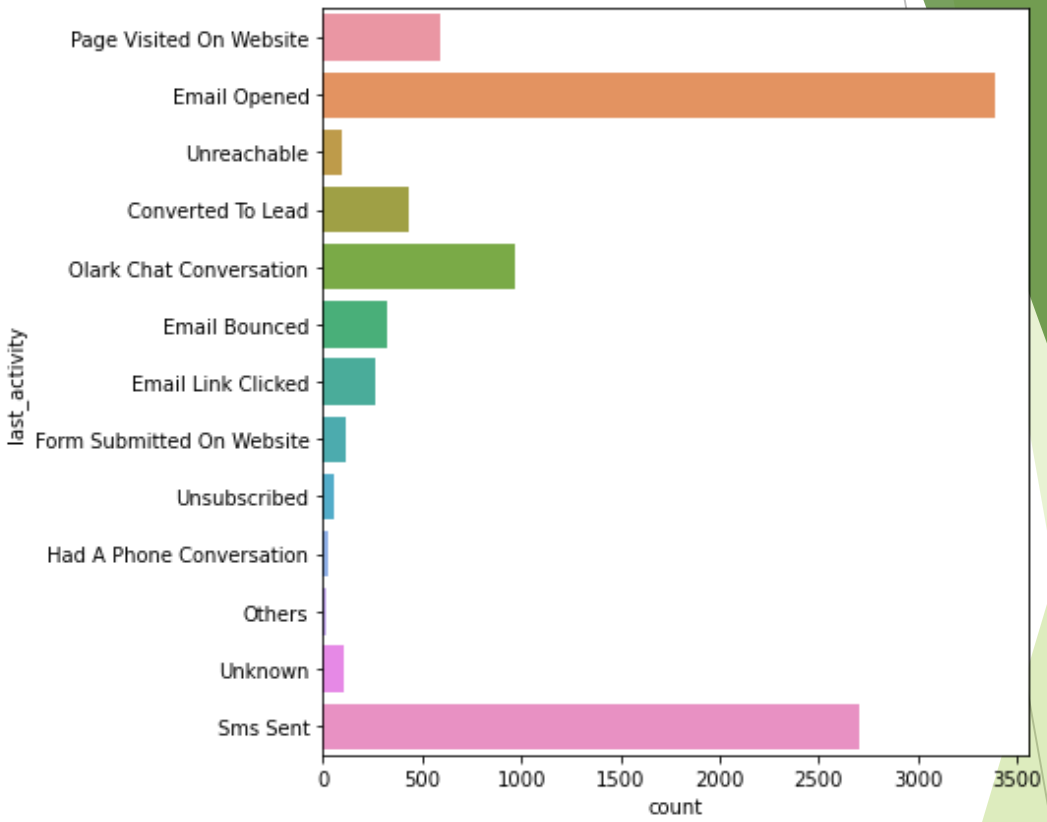




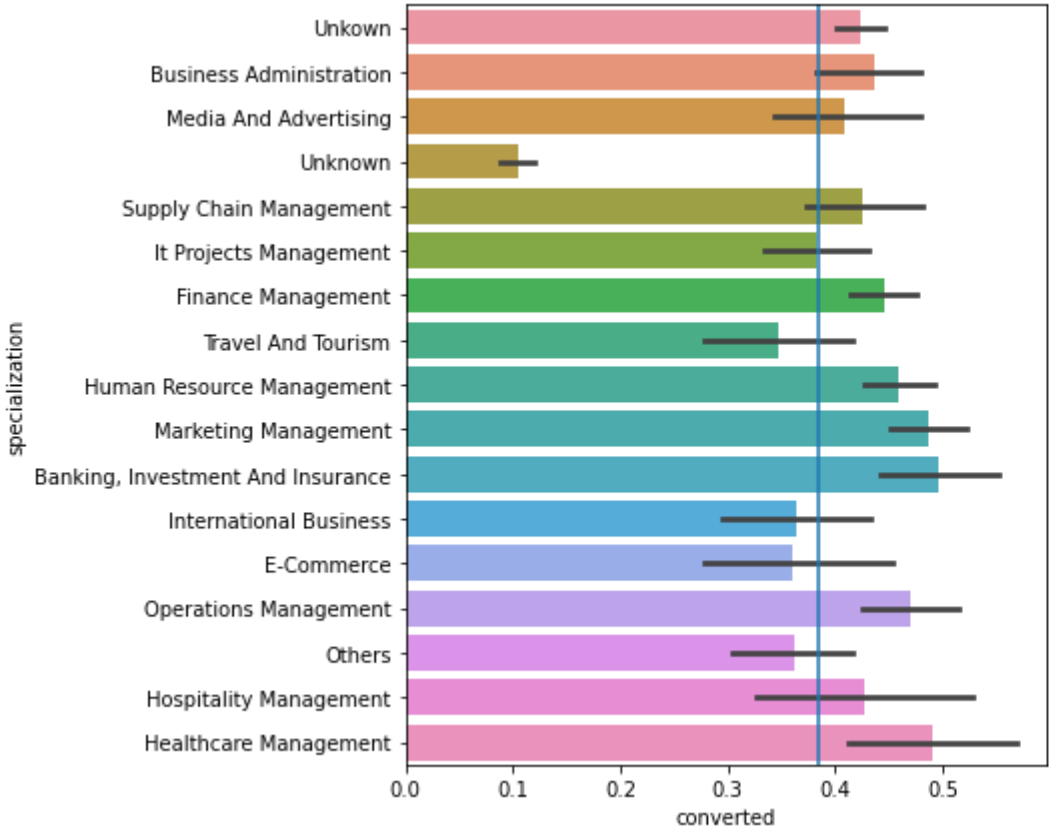

No
Yes
do_not_email
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
converted


No
Yes
do_not_email
0
1000
2000
3000
4000
5000
6000
7000
8000
count

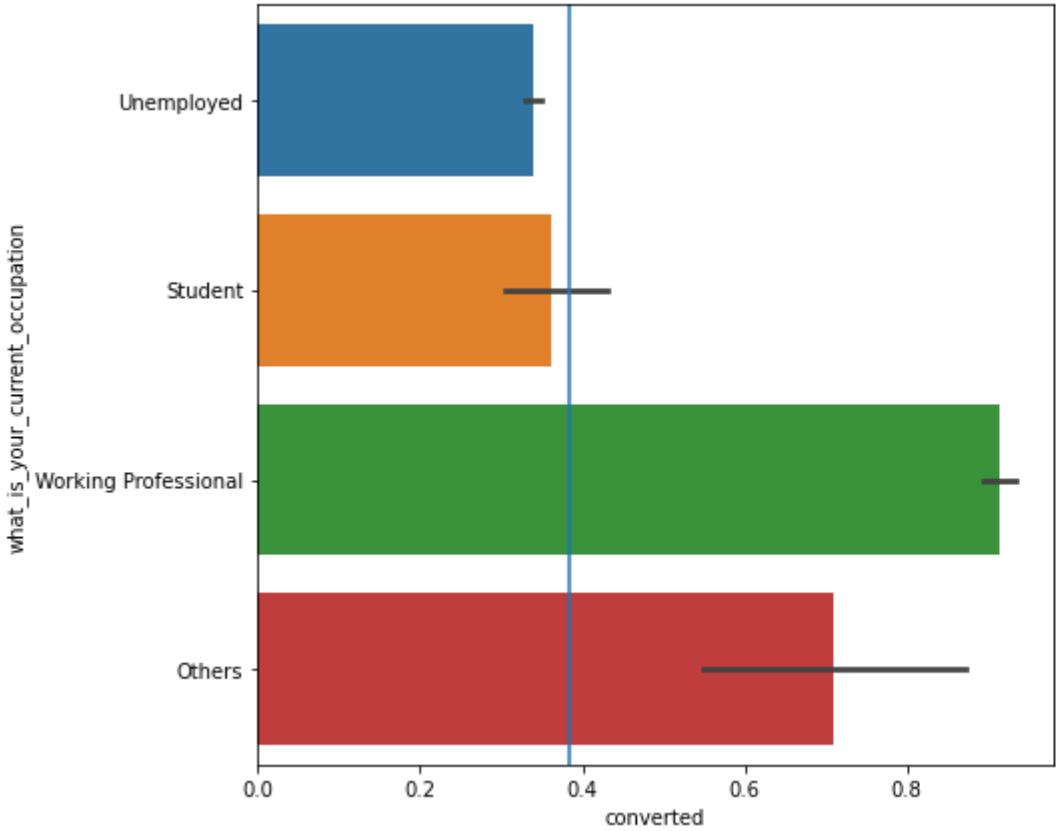


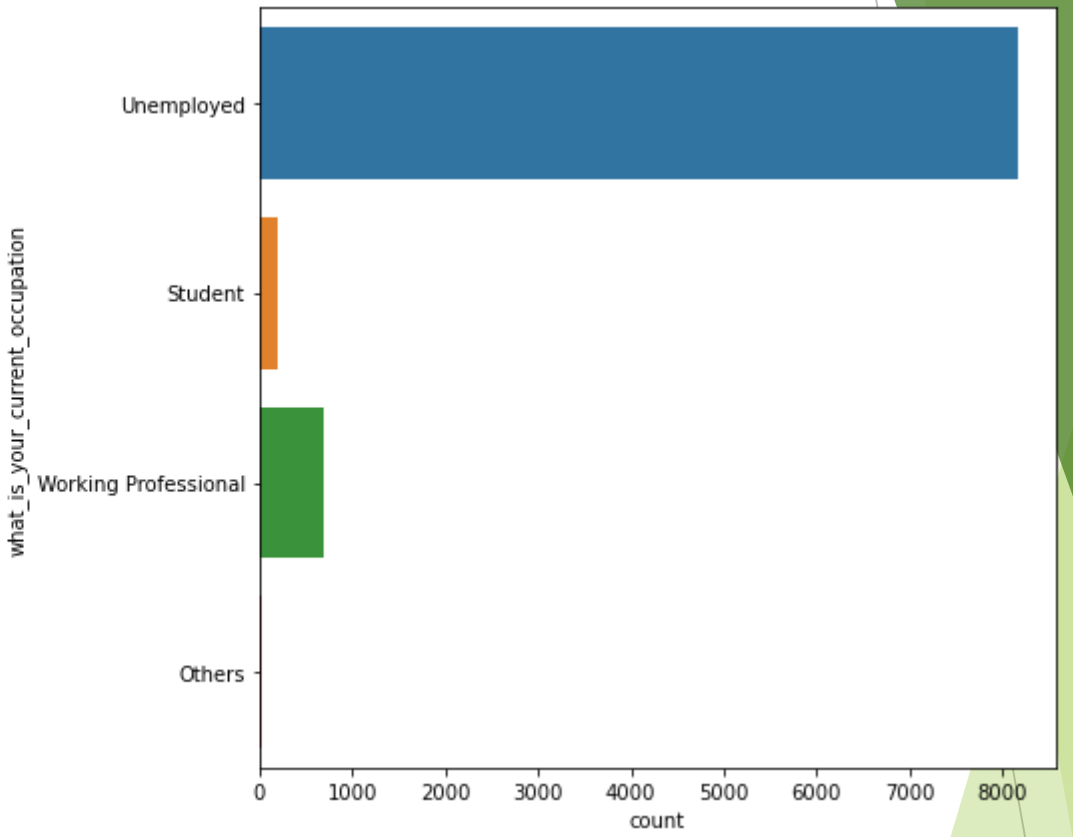


- If the last activity is an sms sent, then there is high chance of conversion

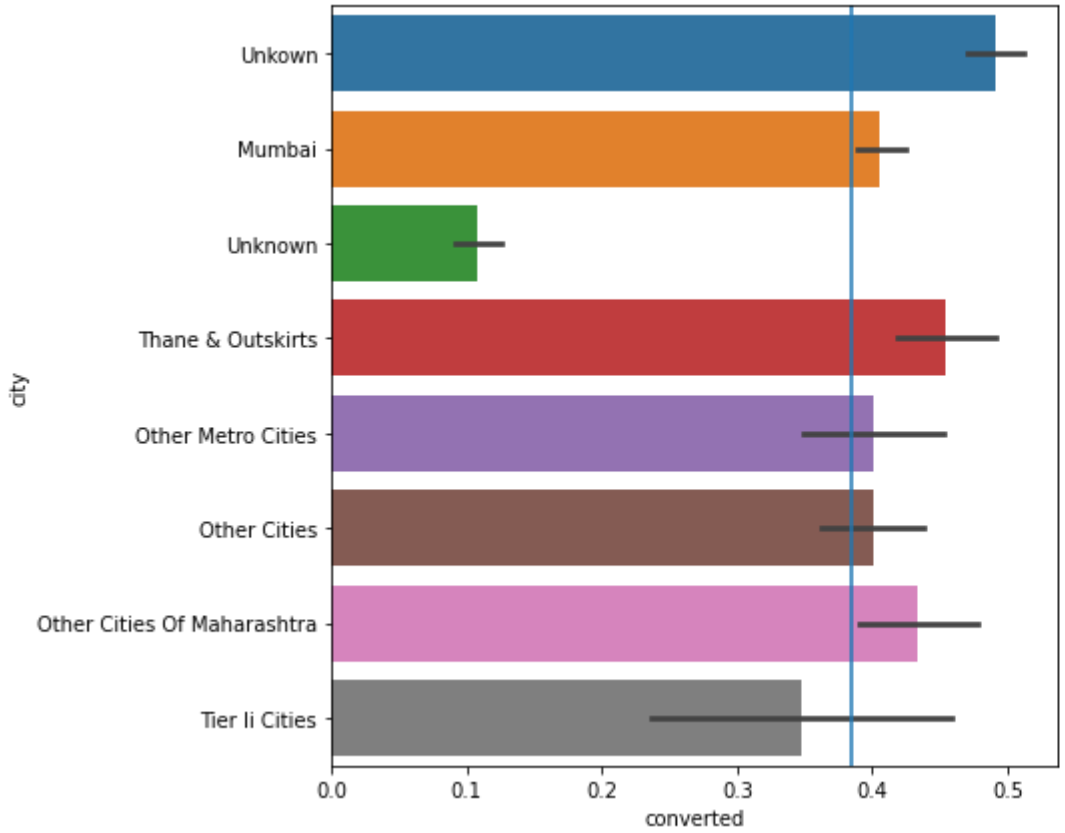


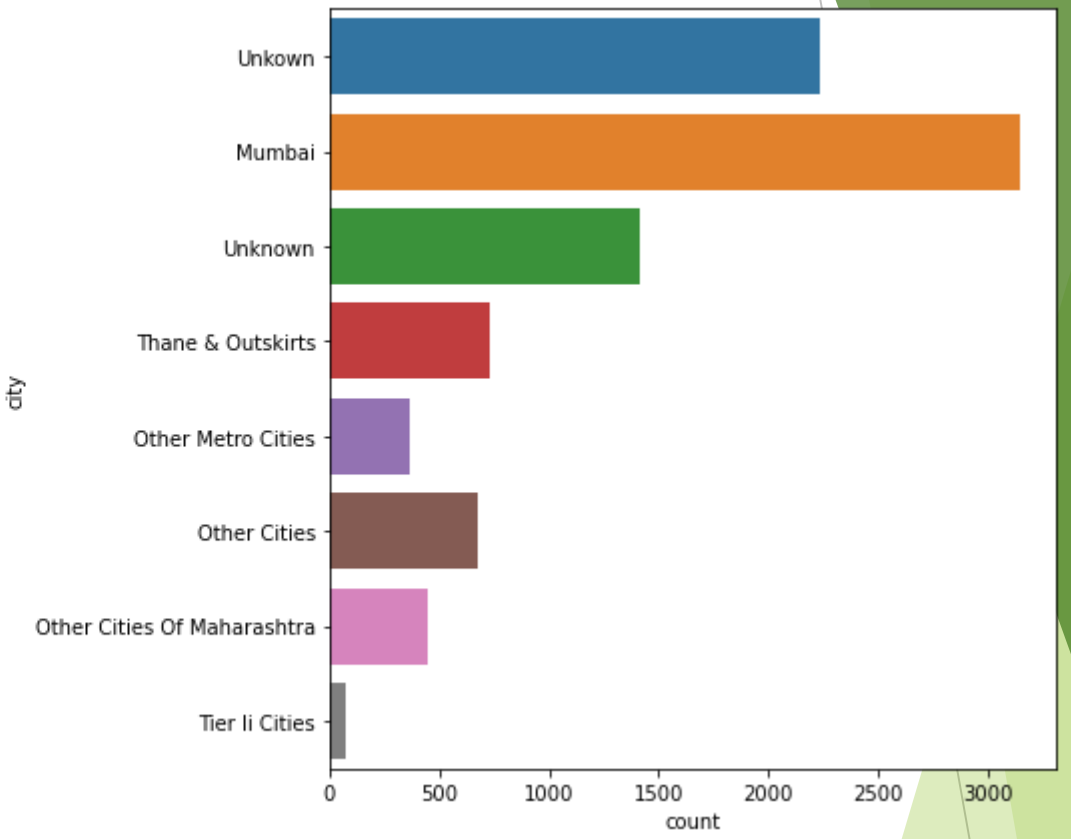
1. If there is no specialization i.e. unknown the conversion rate drops suddenly
2. All Management studies have a good conversion rate



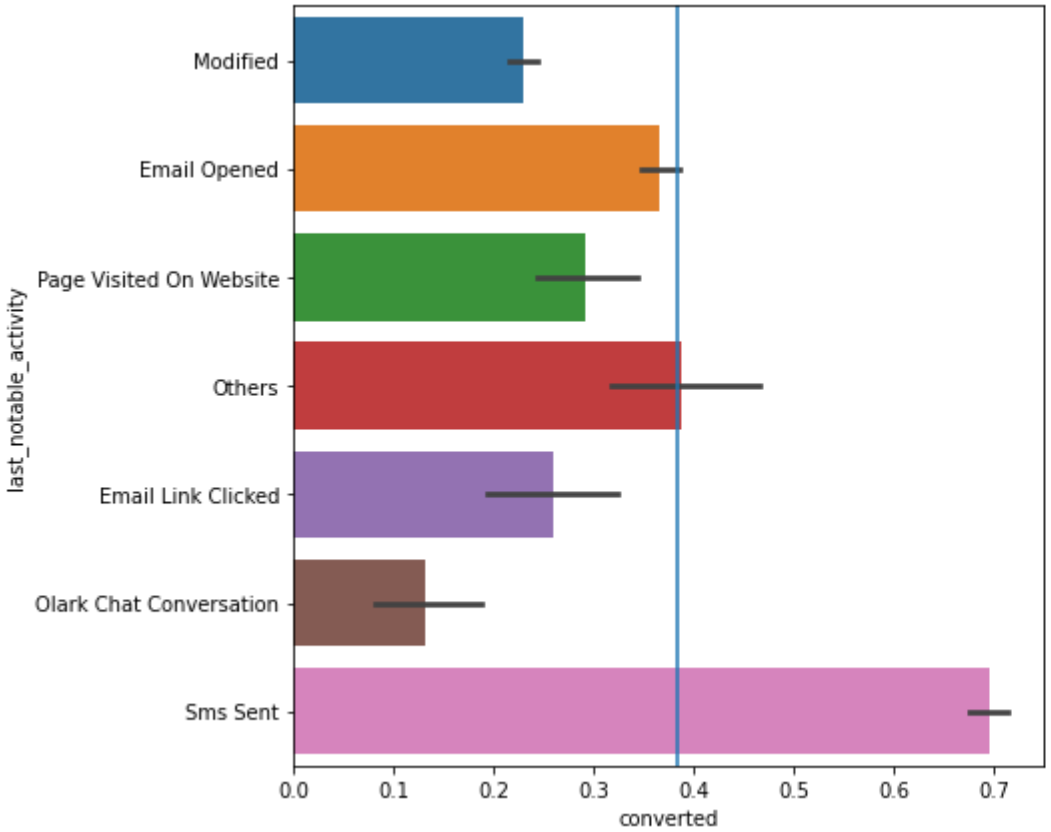


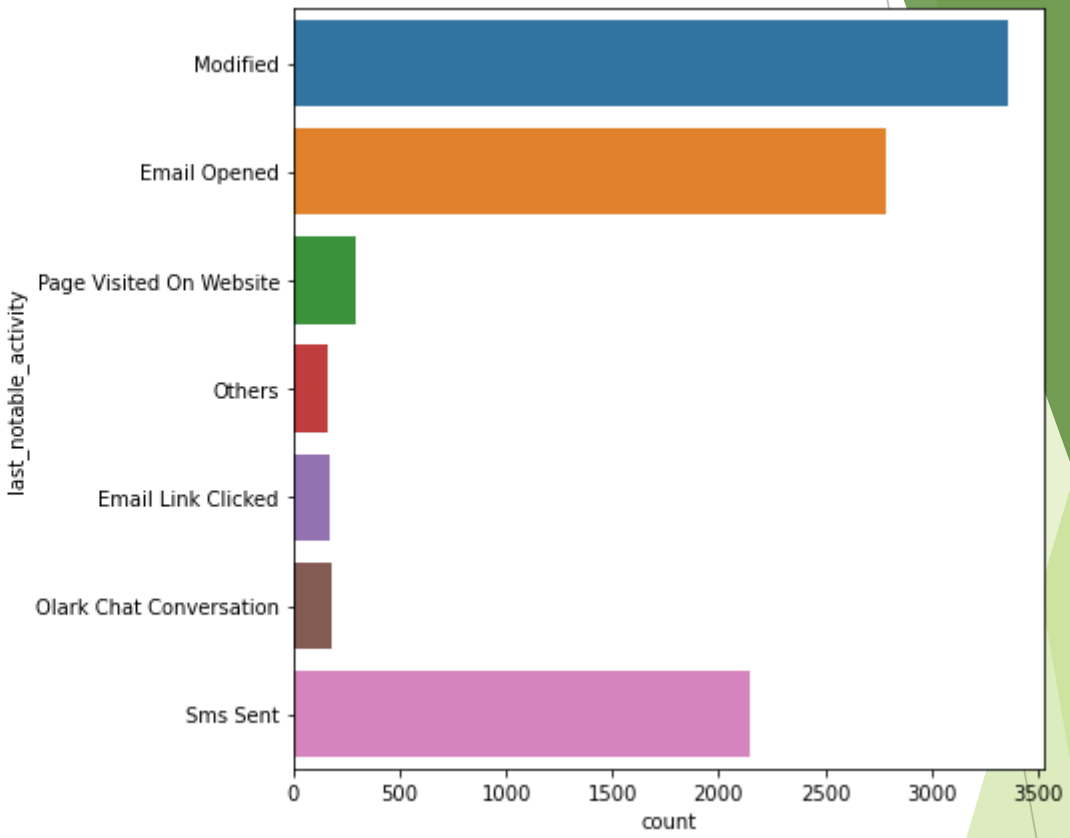
1. Work Professionals are a good bracket to target





1. When city is not mentioned the conversion rate drops



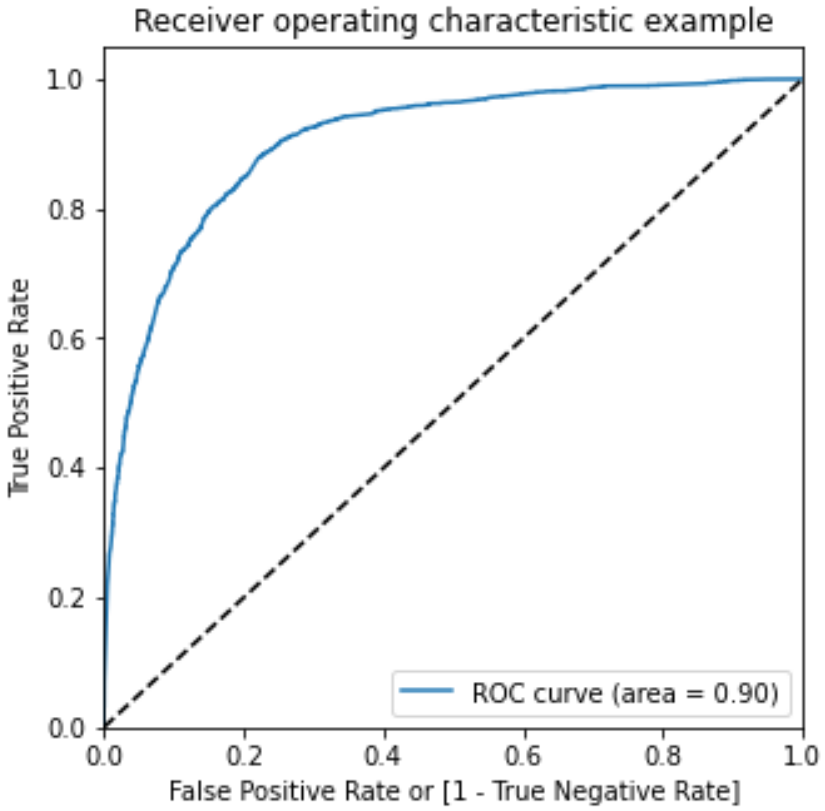


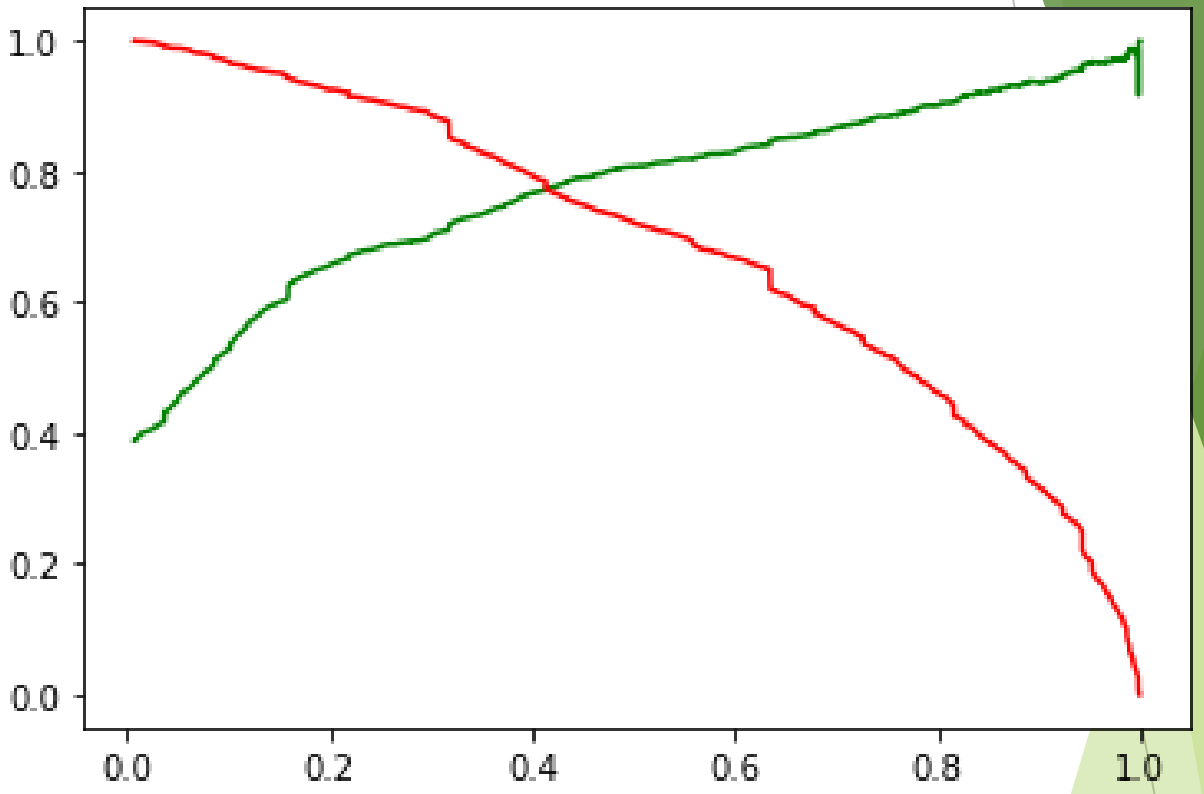
1. Sms Sent is a good indicator

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 80:20 ratio.
- Use RFE for Feature Selection
- Running RFE with 20 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 82%

ROC Curve





- Finding 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 0.40.

Conclusion

Train Data Score:

Accuracy : 82%
Sensitivity (Out of all the converted users, how many we were able to identify) : 81%
Specificity (Out of all the non converted users, how many we were able to identify) : 90%

Test Data Score:

Accuracy : 83%
Sensitivity (Out of all the converted users, how many we were able to identify) : 81%
Specificity (Out of all the non converted users, how many we were able to identify) :85%

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
- The total time spend on the Website.
- Total number of visits.
- When the lead source was: a. Google b. Direct traffic c. Organic search d. Welingak website
- 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.