

LEAD SCORE CASE STUDY

SUBMITTED BY :Aseem Pasha A

Problem Statement: X Education is an online course provider that specializes in selling courses to professionals in various industries. The company promotes its courses on different websites and search engines like Google. When visitors arrive on the website, they have the option to explore the available courses, complete a form to express interest, or watch informational videos. By providing their email address or phone number, these individuals are considered leads. Additionally, X Education also receives leads through referrals from past customers. Once these leads are obtained, the sales team initiates contact by making phone calls, sending emails, and employing other communication methods. During this process, some leads are successfully converted into paying customers, while the majority do not proceed further. On average, X Education achieves a lead conversion rate of approximately 30%.

Business Goal: X Education is seeking assistance in identifying the most prospective leads, those with the highest probability of becoming paying customers. They require a model that assigns a lead score to each lead, indicating the likelihood of conversion. Customers with higher lead scores are expected to have a greater chance of converting, while those with lower scores are expected to have a lower likelihood of conversion. The CEO has provided a rough target lead conversion rate of approximately 80%.

Steps performed in the case study

► Following steps were performed in building of model and categorizing the leads as “hot leads” and “cold leads”:

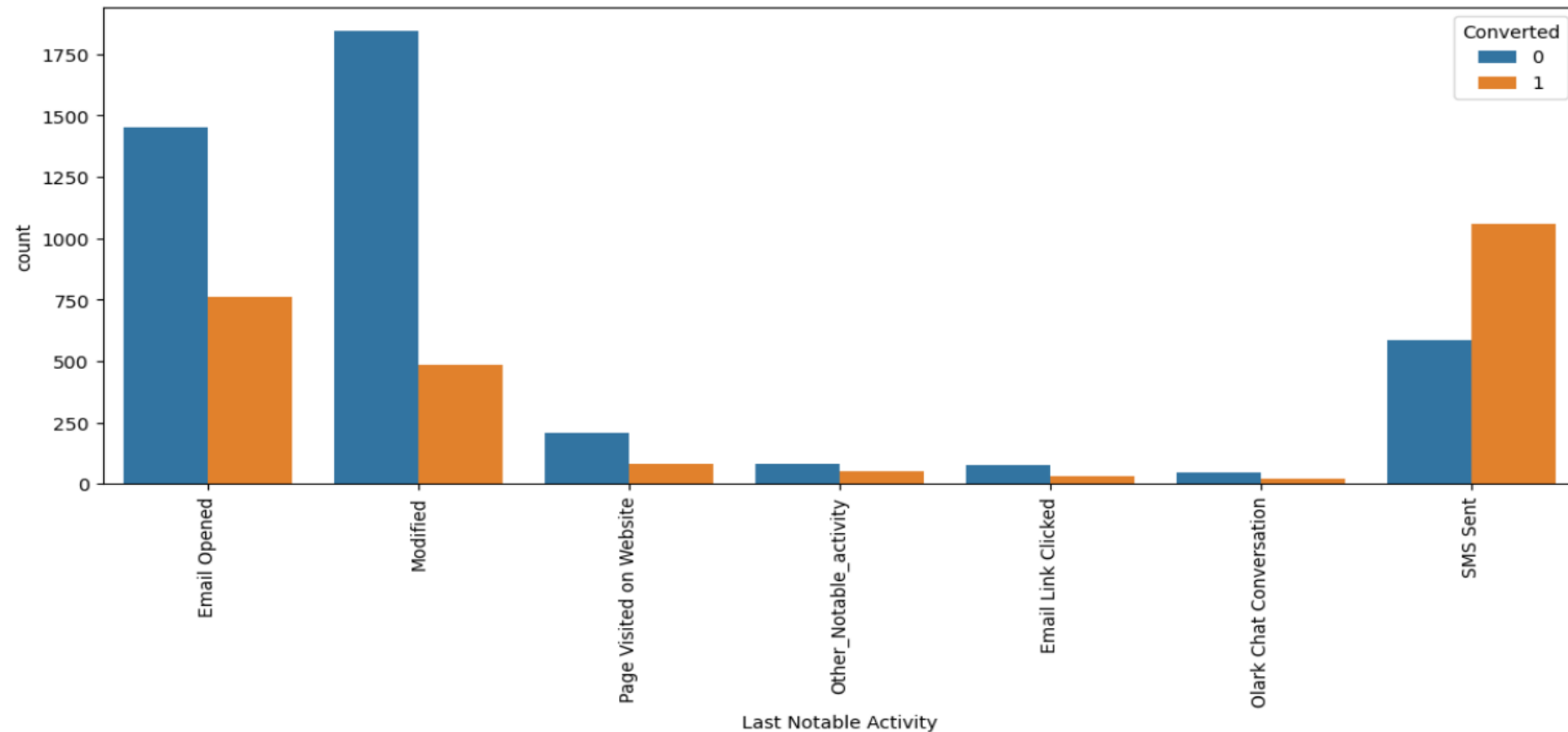
1. Importing of Data
2. Understanding of Data
3. Cleaning of Data
4. Univariate and Bivariate Analysis
5. Multivariate Analysis
6. Data Preparation
7. Train-Test Split
8. Feature Scaling using Standard Scaler
9. Feature Selection using RFE
10. Model Building
11. Model Evaluation
12. Plotting of ROC curve
13. Calculating optimal cut off point
14. Calculating Sensitivity, Specificity, Accuracy
15. Calculating Precision and Recall
16. Prediction of Test set
17. Deciding leads as “hot leads” and “cold leads” based on lead score.

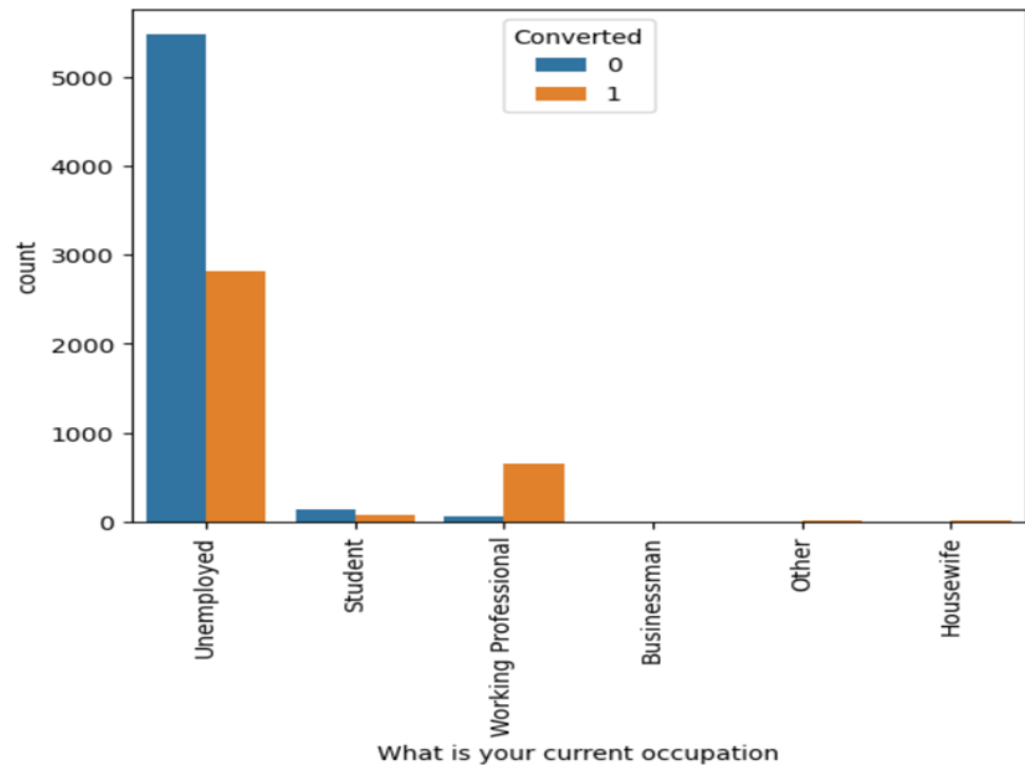
Data Manipulation

- ▶ Total Number of Rows =37, Total Number of Columns =9240.
- ▶ Single value features 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.
- ▶ Removing the “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 has no enough variance, 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.
- ▶ Dropping the columns having more than 30% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Exploratory Data Analysis

Last Activity by customer is either SMS or Email

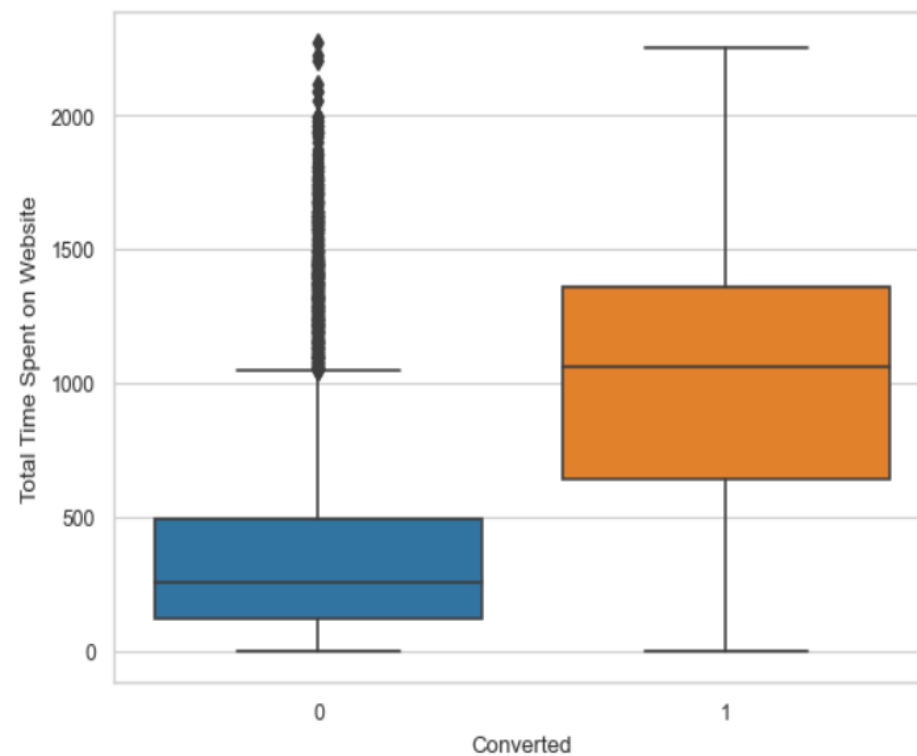




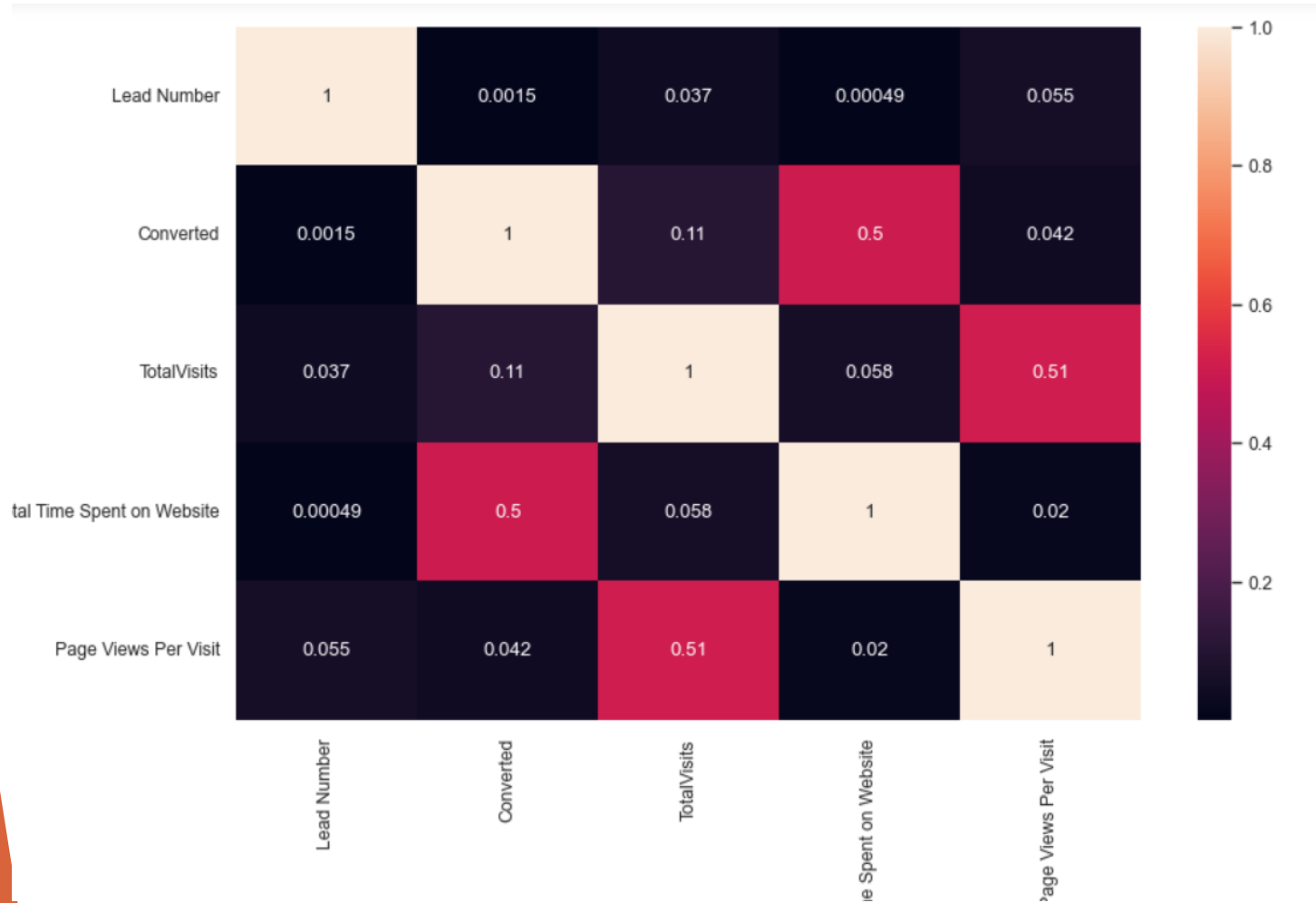
Conversion chances are more if user
Spend more time on Website

Occupation of clients is mostly
Unemployed

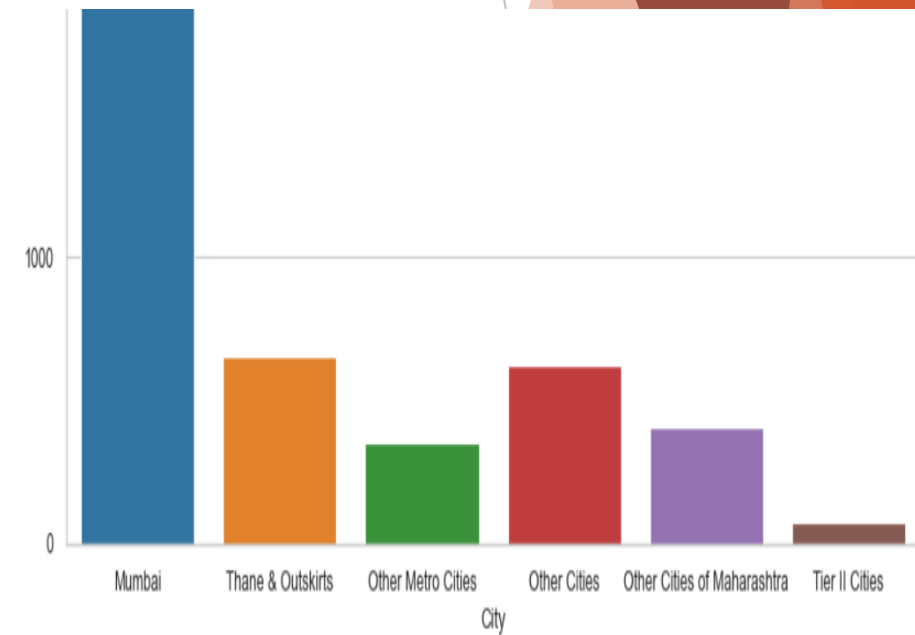
Total Time Spent on Website



Heat Map



Most of them belong from Mumbai

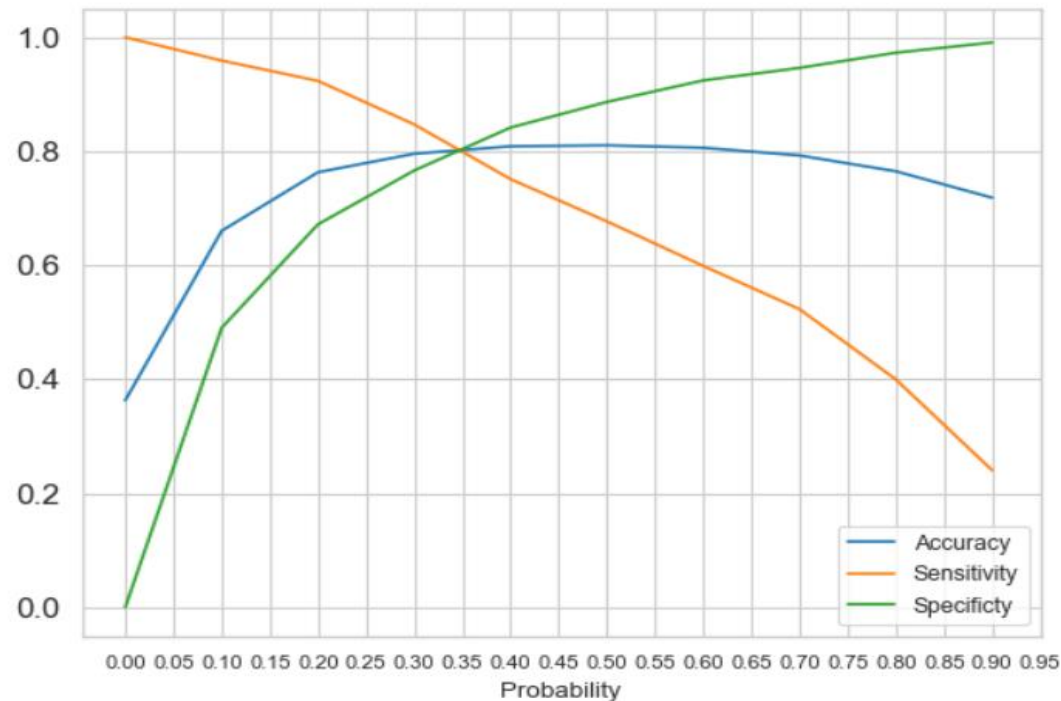


- ▶ Variables Impacting the Conversion Rate • Do Not Email • Total Visits • Total Time Spent On Website
- ▶ Lead Origin - Lead Page Submission
- ▶ Lead Origin - Lead Add Form
- ▶ Lead Source - Olark Chat
- ▶ Last Source - Welingak Website
- ▶ Last Activity - Email Bounced
- ▶ Last Activity - Not Sure
- ▶ Last Activity - Olark Chat Conversation
- ▶ Last Activity - SMS Sent
- ▶ Current Occupation - Working Professional

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 70:30 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 81%

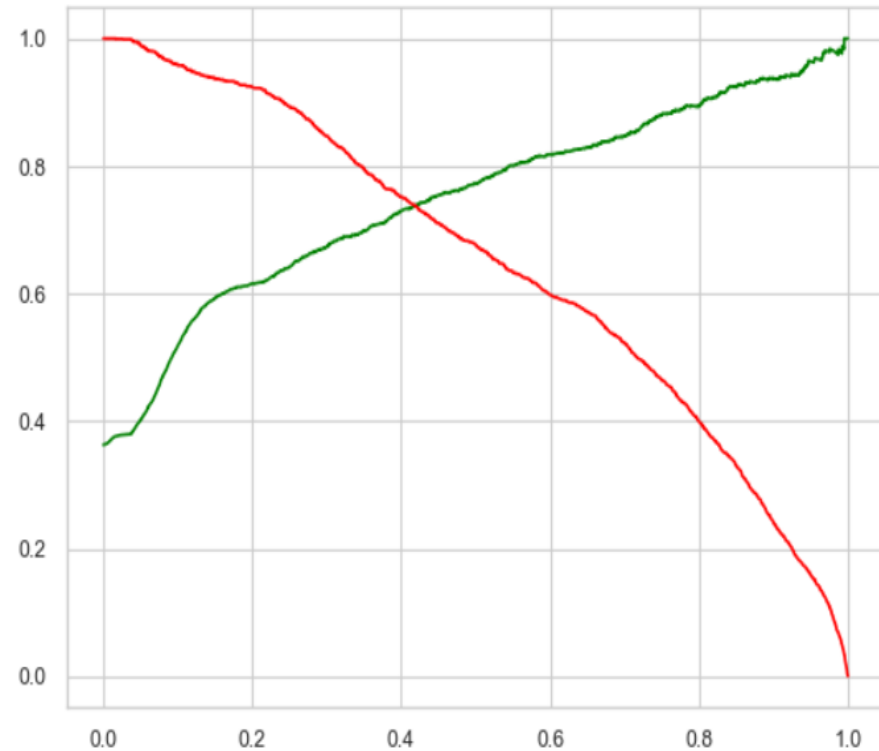
Model Evaluation - Sensitivity and Specificity on Train Data Set



Observation

From the above curve we can see that the optimal cutoff is at 0.35.

Recall & Precision Curve



As seen from above, there is tradeoff between Precision and Recall. Precision and Recall are inversely related means if one increases other will genuinely decrease.

Conclusion

Let us compare the values obtained for Train & Test:

Train Data:

Accuracy : 80.16%

Sensitivity : 79.72%

Specificity : 80.41%

Test Data:

Accuracy : 80.87%

Sensitivity : 79.29%

Specificity : 81.79%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

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

- ▶ 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

Thank You