Lead Scoring Case Study

Batch : DS_C72

Team :

Abhishek Guleria

Lalita Rathod

Mansa Madhusoodanan

Summary

- Problem Statement
- Business Objective
- Approach (Steps Taken)
- Results

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 Objectives

- **X Education** aims to identify the most promising leads.
- To achieve this, they seek to develop a model that can classify **hot leads** effectively.
- The model will be deployed for **future use** to streamline lead prioritization and decision-making

Approach

Data Cleaning and Data Manipulation

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains a large number of missing values and are not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

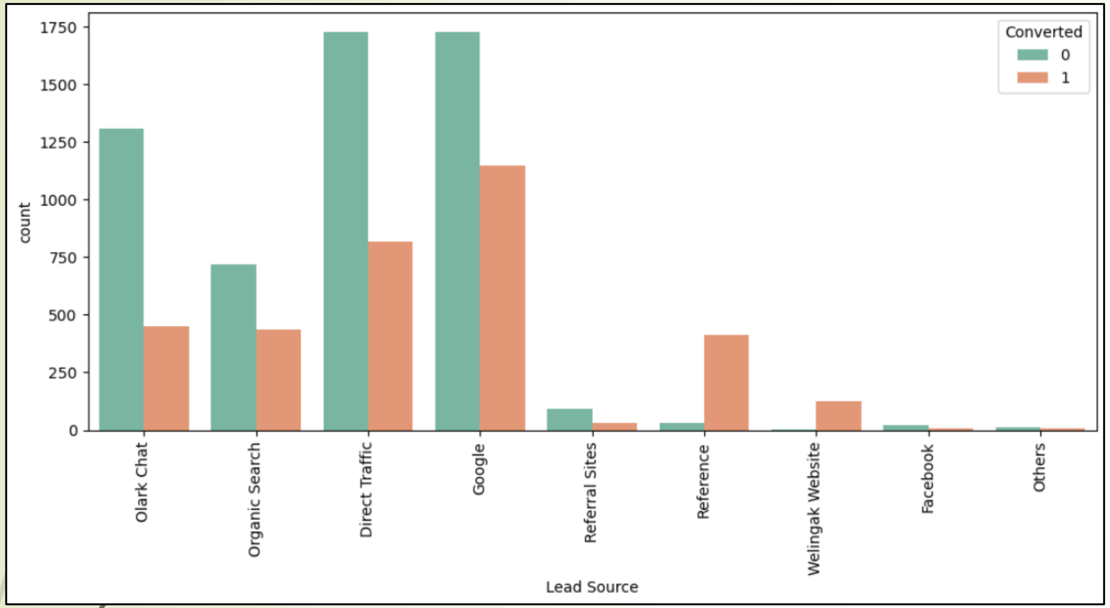
Exploratory Data Analysis (EDA)

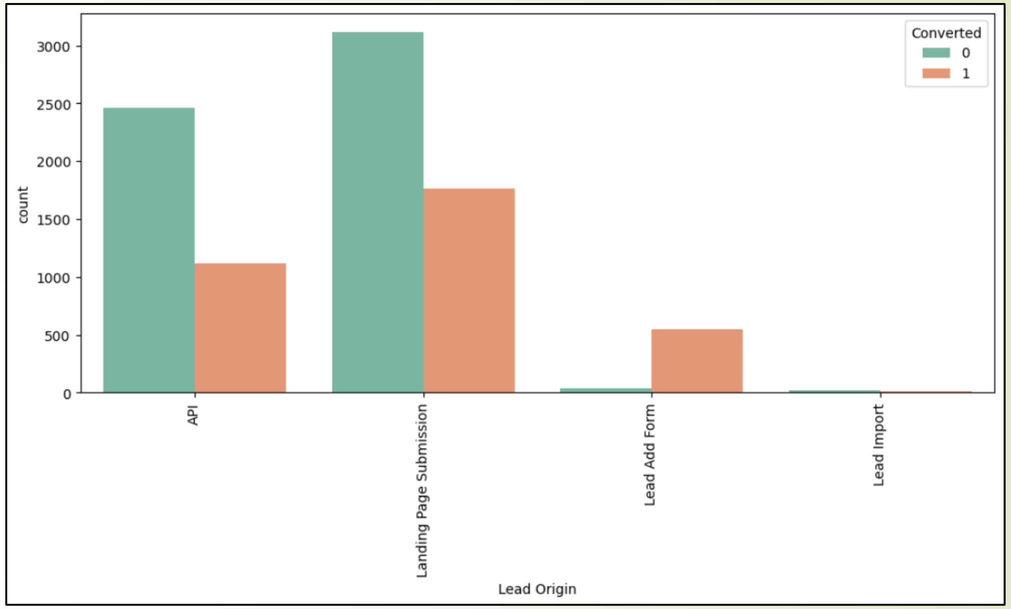
- Univariate data analysis: value count, distribution of variables, etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy variables and encoding of the data.
- Classification technique: logistic regression is used for model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

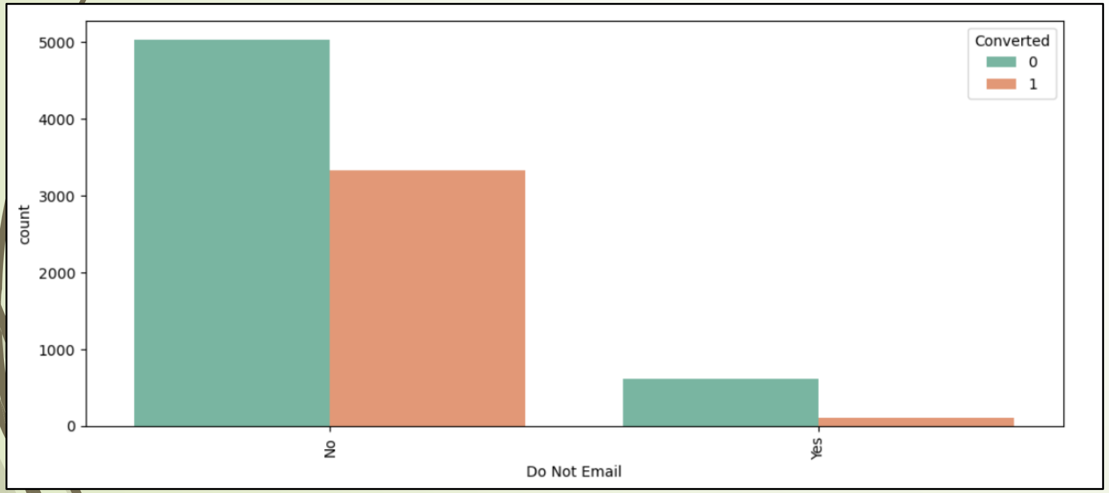
Data Manipulation & Data Conversion

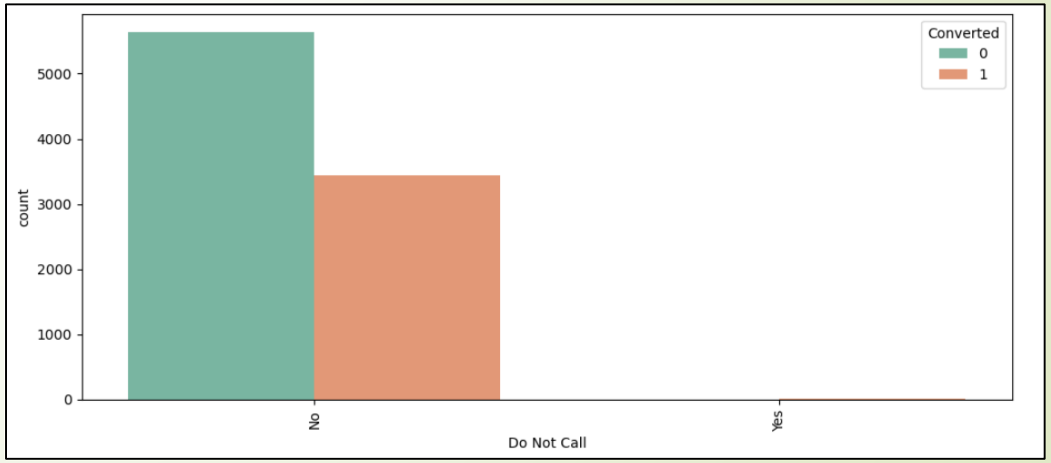
- Total Number of Rows=37,Total Number of Columns =9240.
- Single value features like“Magazine”, “ReceiveMoreUpdates About Our Courses”, “Update my supply”
- After checking for the value counts for some of the object type variables, we find some of the feature which have enough variance, which are dropped, the features are: “What matters most to you in choosing course”
- Dropping the columns having more than 40 % as missing values such as ‘How did you hear about X Education’, ‘Lead Quality’ ‘Lead Profile’ etc.
- Numerical Variables
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240 & Total Columns for Analysis: 37

Exploratory Data Analysis (EDA)

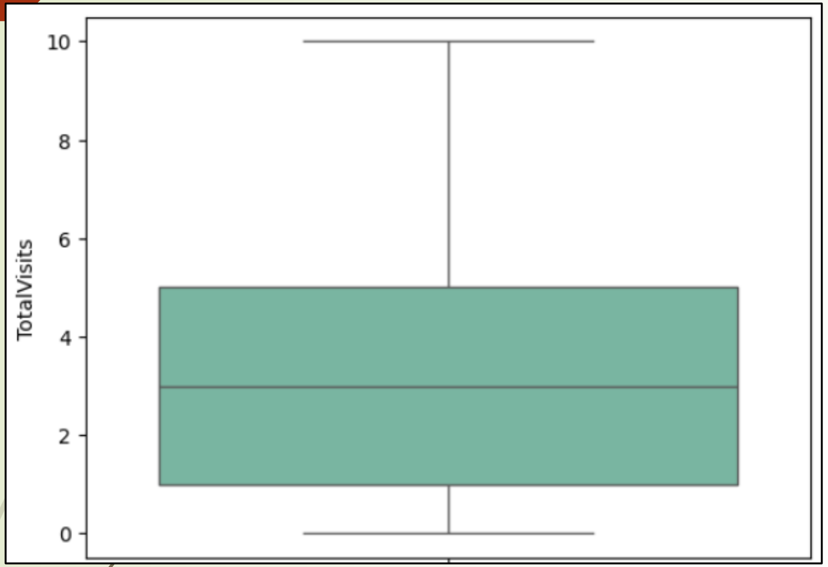


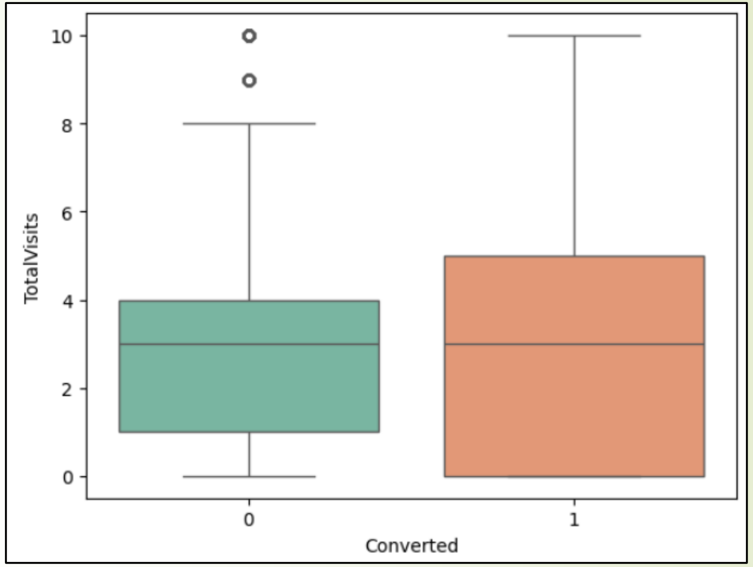


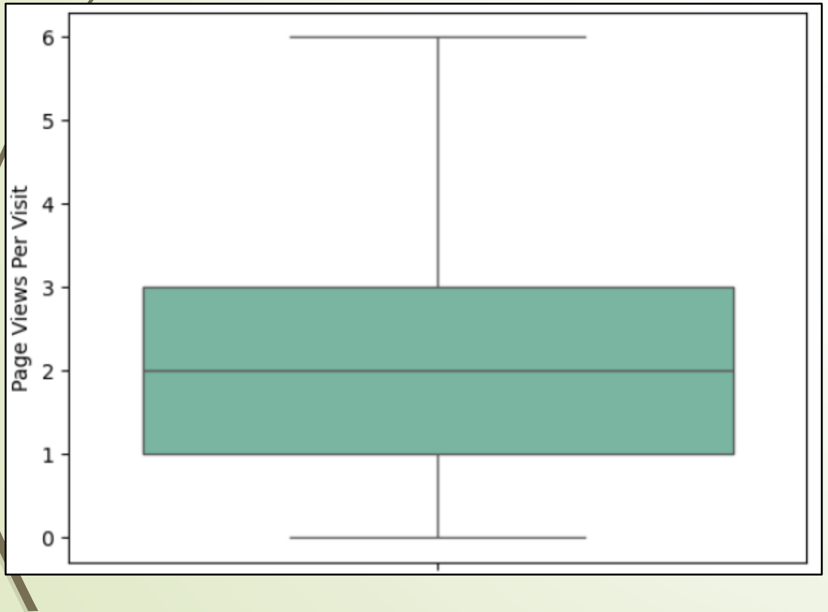


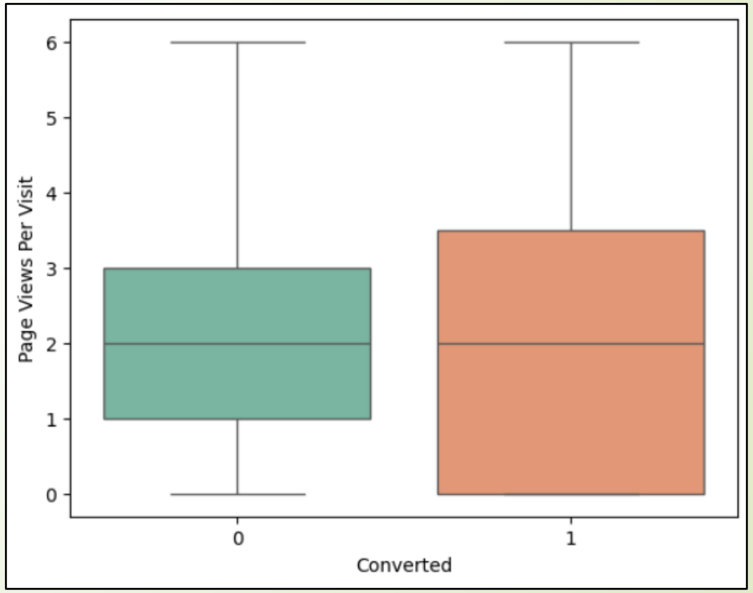


Box Plot

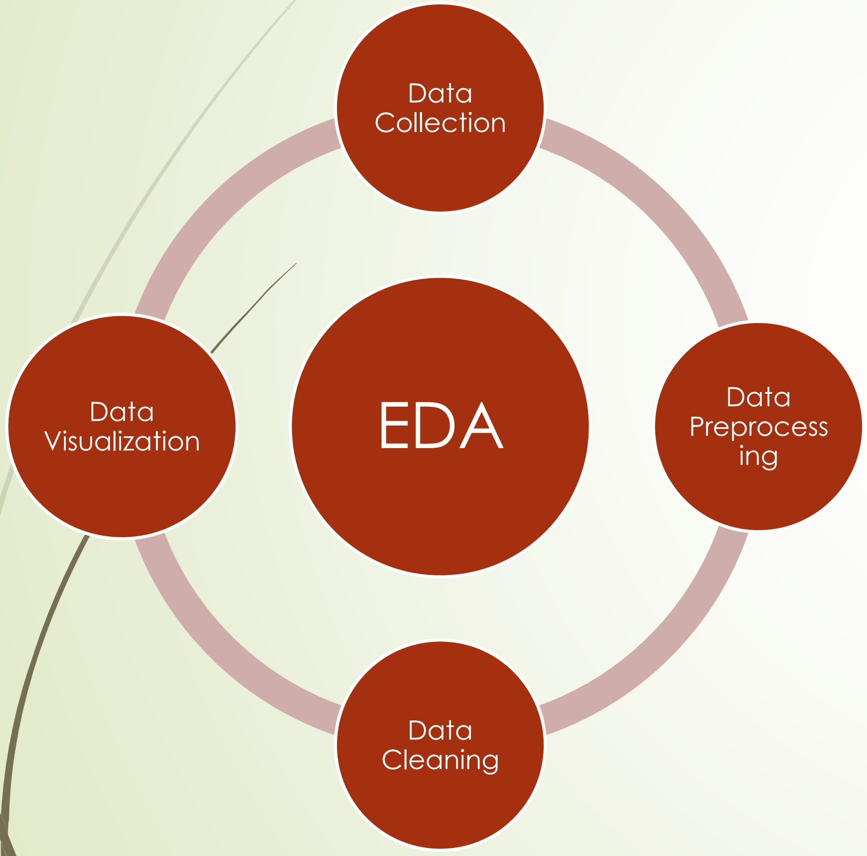

10
8
6
4
2
0
TotalVisits


10
8
6
4
2
0
TotalVisits
0
1
Converted


6
5
4
3
2
1
0
Page Views Per Visit

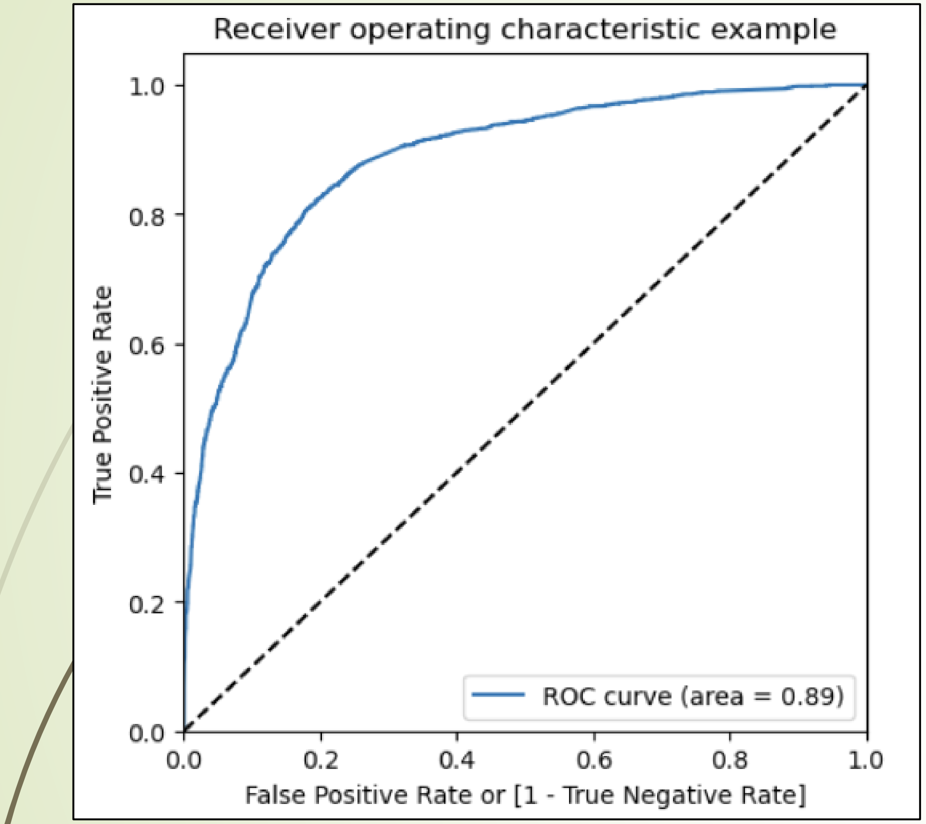

6
5
4
3
2
1
0
Page Views Per Visit
0
1
Converted

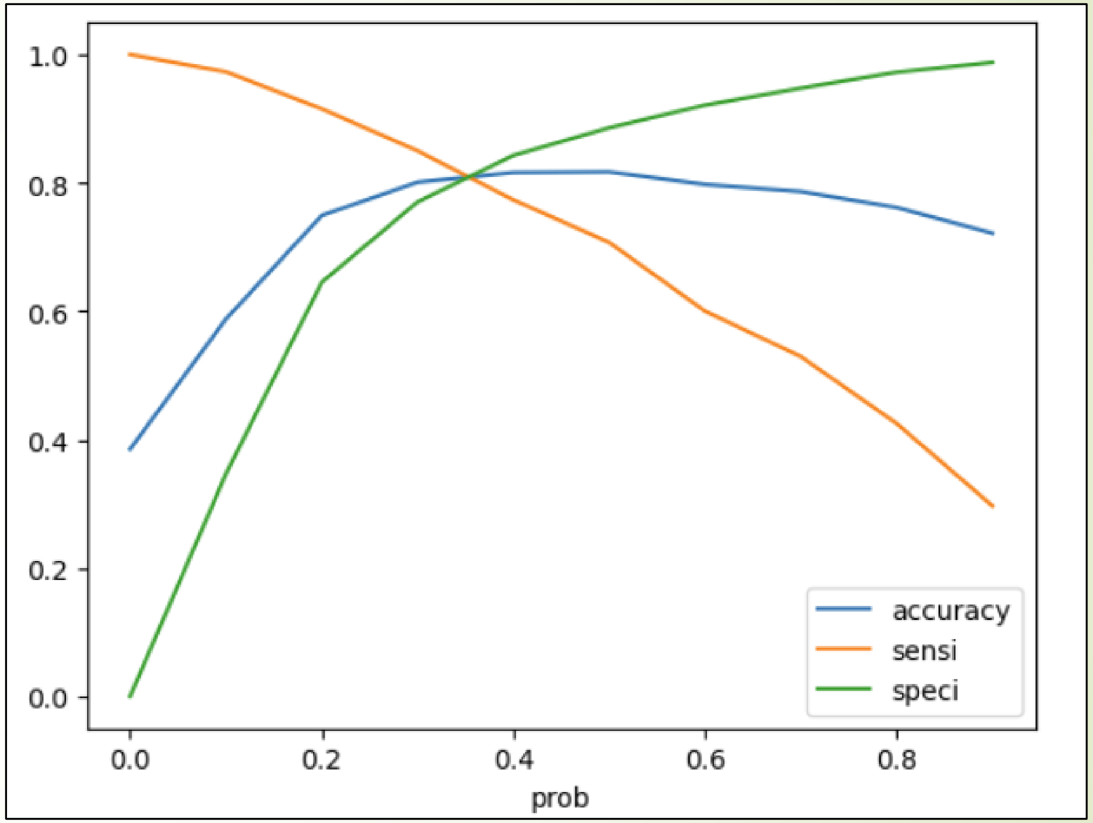
Model Building



- 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 15 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%

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 0.34

Prediction on Test Set

- Before predicting on the test set, we need to standardize the test set
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this we did model evaluation i.e. finding the Accuracy, Sensitivity, and Specificity.
- The Accuracy score on test data was 0.80, Sensitivity 0.81, and Specificity 0.80 approximately.
- This shows that our test prediction is having Accuracy, Sensitivity, and Specificity scores in an acceptable range.
- This also shows that our model is stable with good Accuracy and Sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

Recommendations

It was found that the variables that mattered the most in the potential buyers are as follows :

- lead sources "Welingak Websites" and "Reference".
- "working professionals"
- "more time on the websites"
- "Olark Chat"
- “last activity was SMS Sent”

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

Thank You !