Lead Scoring Case Study For X Education

•••

Boosting Lead Conversion Rates through Predictive Lead Scoring to detect Hot Leads

Table of Contents

- Background of X Education Company
- Problem Statement & Objective of the Study
- Suggested Ideas for Lead Conversion
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Recommendations

Background of X Education Company

Company Overview: X Education is an online education company that specializes in offering courses to industry professionals. They provide a wide range of online courses catering to various domains, including business, technology, marketing, and more.

Target Audience: The company's target audience consists of working professionals and individuals seeking to enhance their knowledge and skills in specific areas. X Education aims to provide high-quality educational content and support to help individuals advance in their careers.

Marketing and Lead Generation: X Education employs various marketing strategies to attract potential customers. They promote their courses through digital marketing channels, such as search engines, social media platforms, and online advertising. These efforts drive traffic to their website and generate leads.

Target Conversion Rate: The CEO of X Education has set a target conversion rate of around 80%, significantly higher than the current conversion rate. The lead scoring model will play a crucial role in achieving this ambitious goal by enabling the identification of high-potential leads for targeted engagement.

X Education is committed to leveraging data-driven insights and predictive modeling techniques to enhance their lead conversion rates. By implementing an effective lead scoring system and focusing on high-potential leads, the company aims to maximize its sales efforts and drive significant growth in customer acquisitions.

Problem Statement:

X Education faces challenges in converting a significant number of leads into paying customers, resulting in a low lead conversion rate. The existing lead nurturing and targeting strategies are not effectively identifying and prioritizing potential leads with a higher likelihood of conversion. This leads to suboptimal resource allocation and hampers the overall sales performance of the company.

Objective of the Study:

The objective of this study is to develop a predictive lead scoring model for X Education that assigns a lead score to each potential lead, indicating their probability of conversion. The model aims to identify and prioritize the most promising leads, referred to as "Hot Leads," for targeted engagement by the sales team. By implementing an effective lead scoring system, the study aims to improve the lead conversion rate and optimize the allocation of sales resources. The ultimate goal is to increase the overall sales performance and achieve a higher conversion rate, aligning with the company's objective of enhancing lead conversion efficiency.

Suggested Ideas for Lead Conversion

Personalized Communication

*By tailoring the messaging and content to address individual pain points and aspirations,

*Establish a stronger connection with leads and increase the likelihood of conversion.

Lead Scoring and Segmentation

*Identify and segment leads into different categories

*Assign Predictive Lead Scores

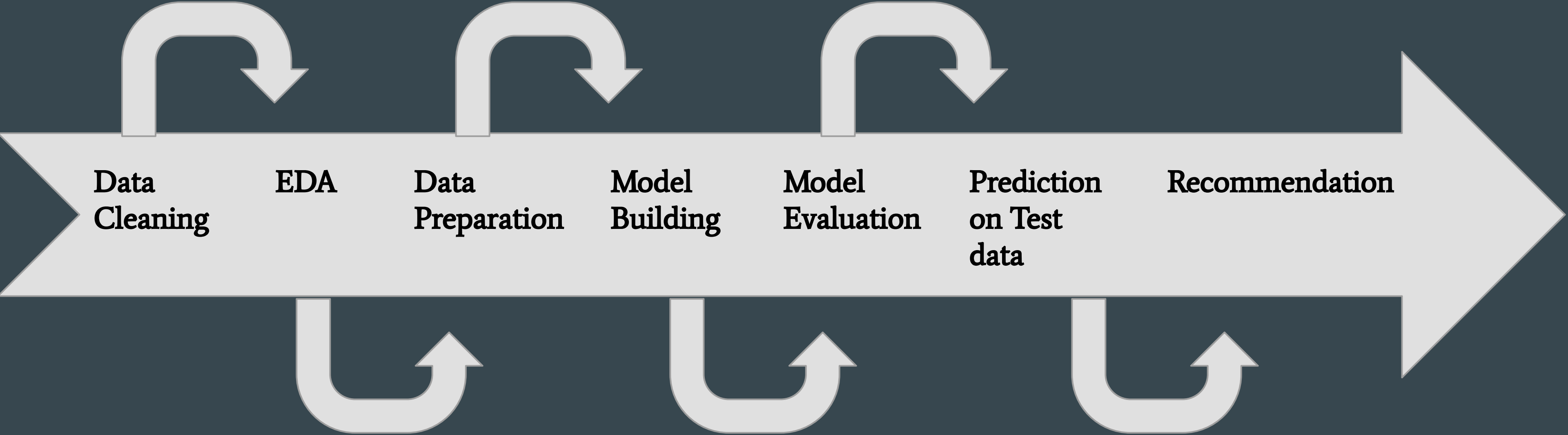
* Hands on experience

Personalized Demos and Trials

*Encourage satisfied customers and partners to refer leads, incentivizing them through referral programs

Strategic Partnerships and Referral Programs

Analysis Approach



Data Cleaning

EDA

Data Preparation

Model Building

Model Evaluation

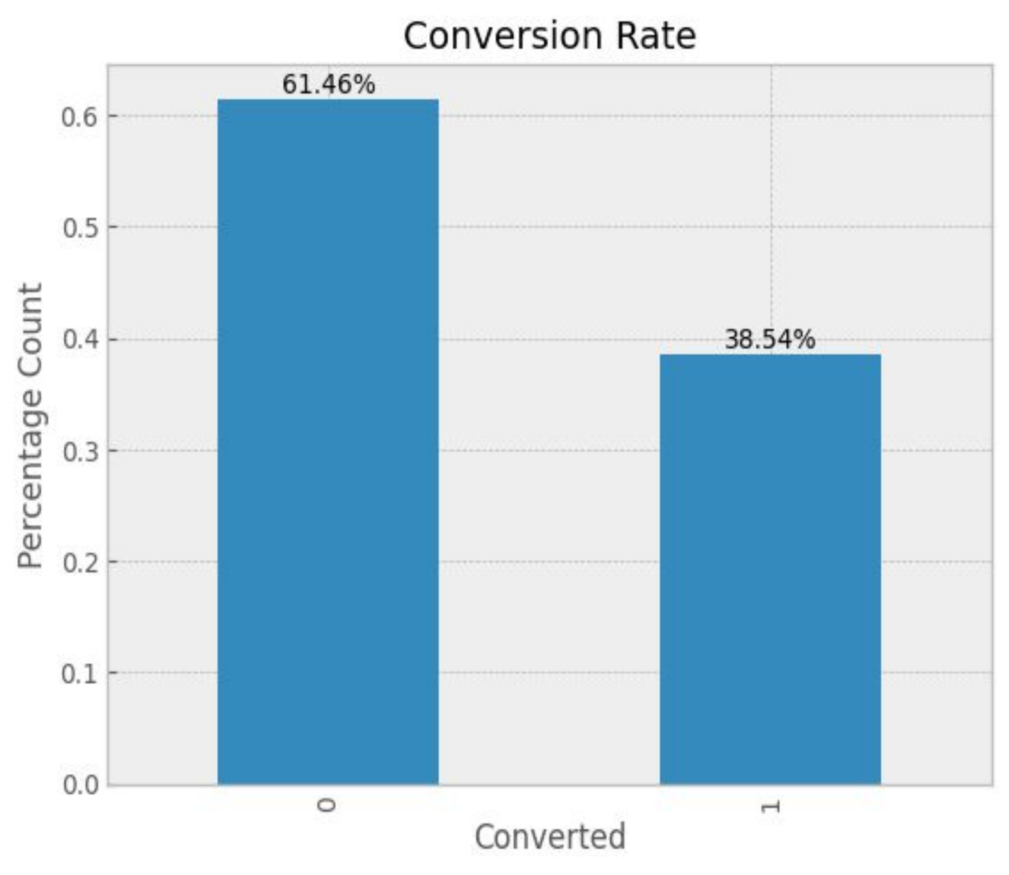
Prediction on Test data

Recommendation

Data Cleaning

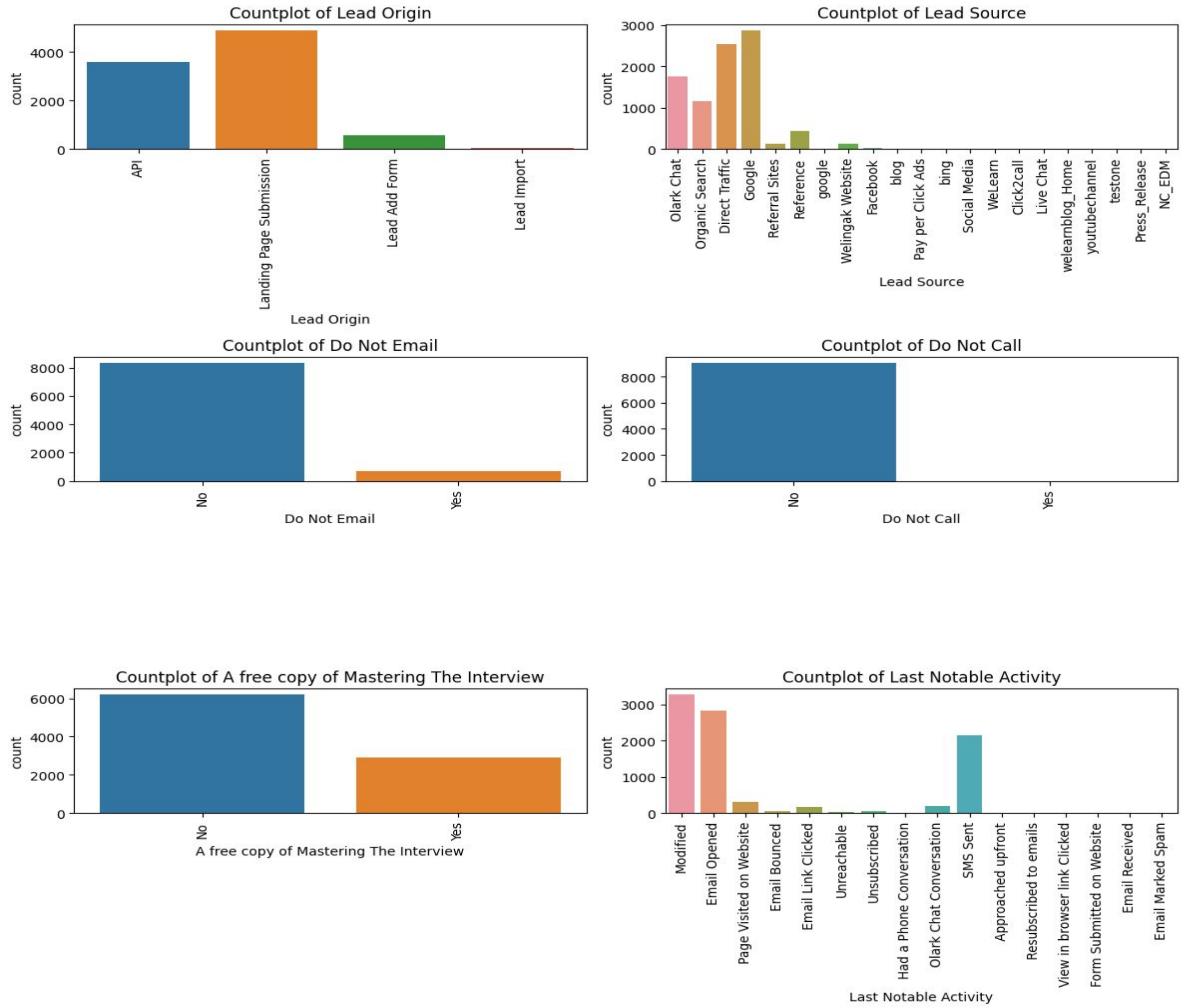
1. Replacing 'select' values: The 'select' values in the dataset are replaced with NaN to indicate missing or unknown values.
2. Removing columns with more than 10% null values: Columns that have more than 10% missing values are dropped from the dataset. This helps to eliminate variables with insufficient data, ensuring the analysis is based on more complete information.
3. Removing columns with single answer throughout: Columns that have only one unique value throughout the dataset are removed as they do not provide any useful information for analysis. The columns 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', and 'I agree to pay the amount through cheque' are removed.
4. Removing prospect id and lead number: The columns 'Prospect ID' and 'Lead Number' are removed from the dataset as they do not contribute to the analysis.
5. Selecting relevant columns: The remaining columns considered important for lead conversion analysis are retained. These columns include 'Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call', 'Converted', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity', 'A free copy of Mastering The Interview', and 'Last Notable Activity'.

EDA



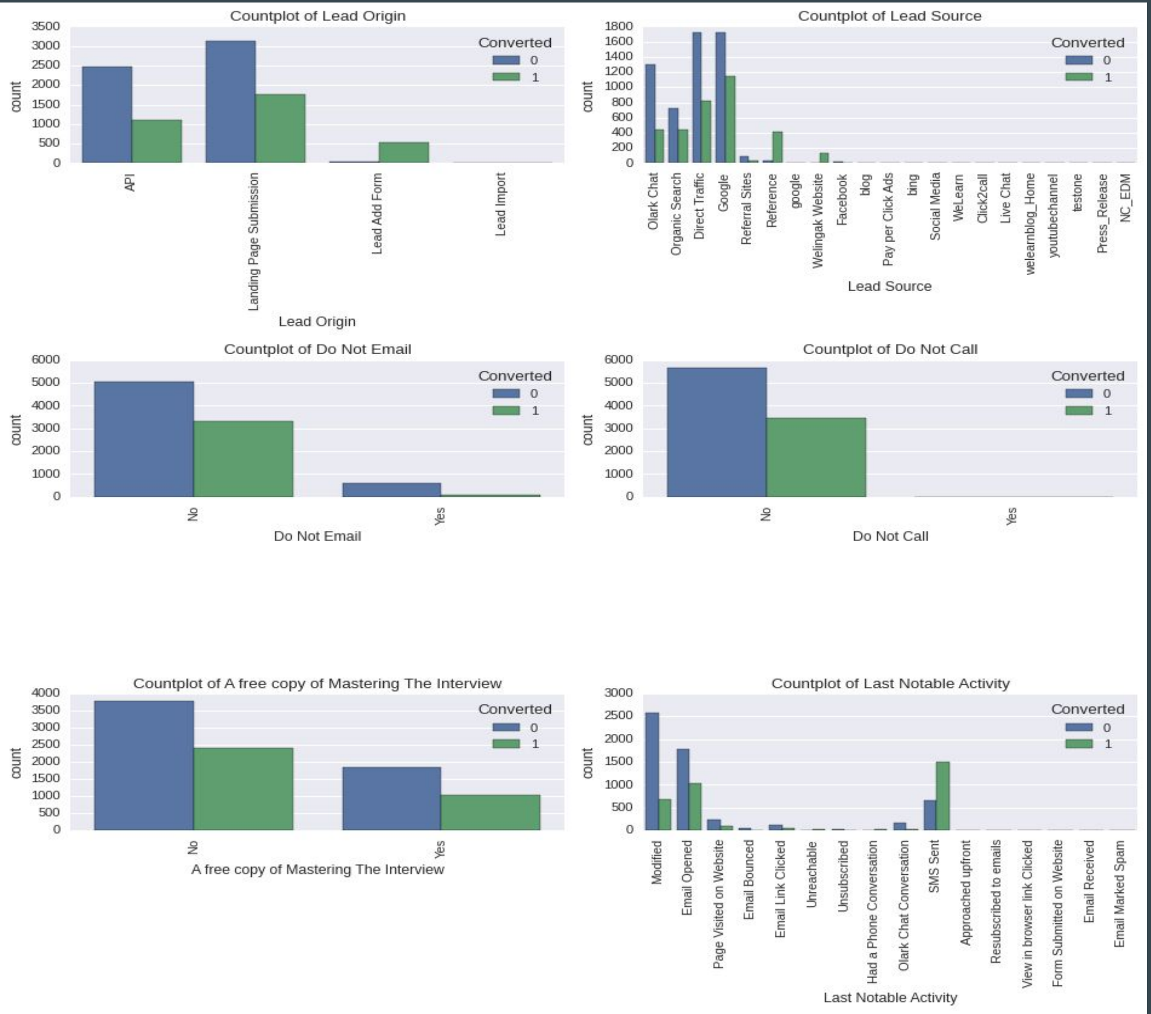
- Data is Imbalanced
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert

Univariate Analysis of Categorical Variables



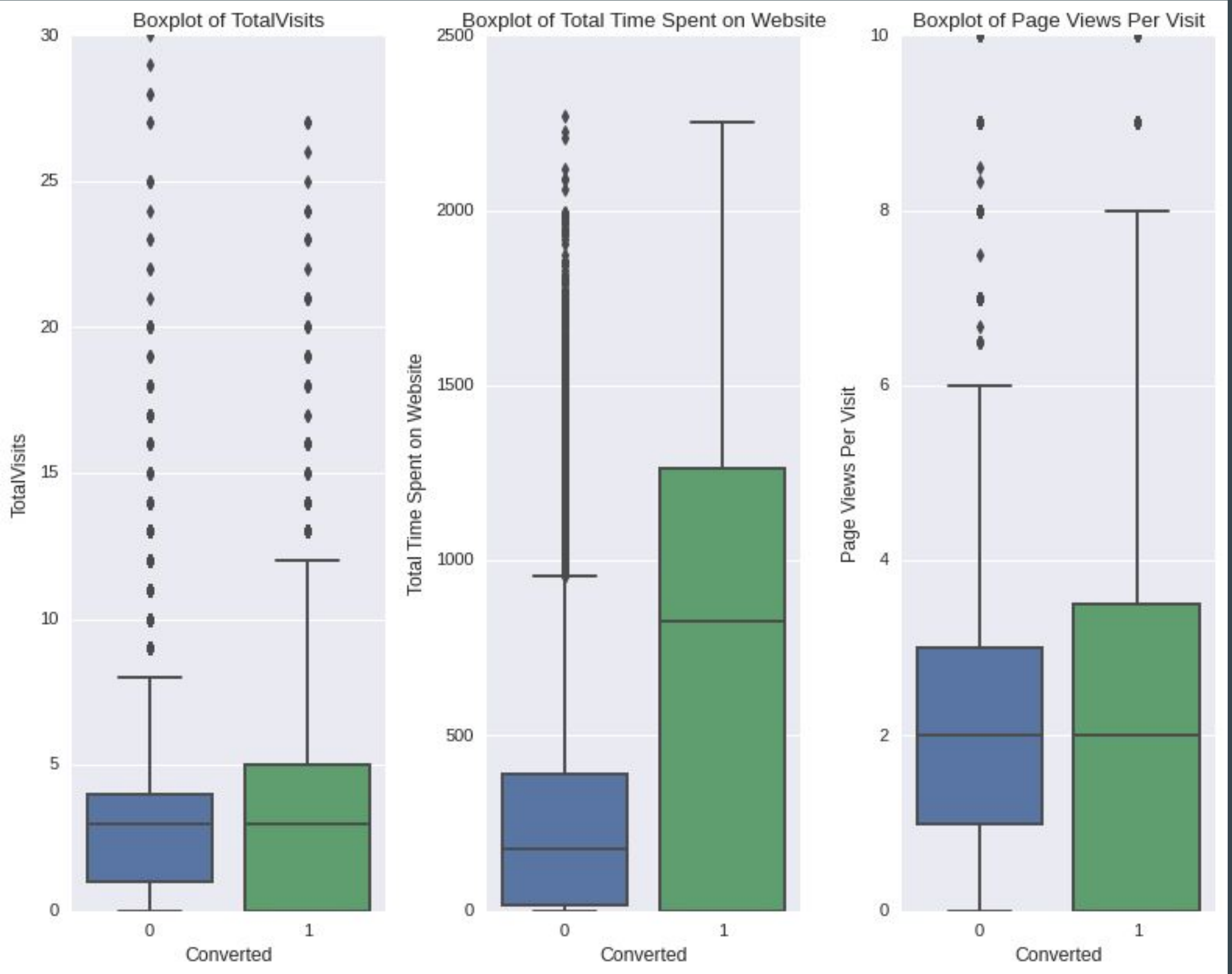
- Lead origin- More than 4000 Lead origin is from Landing page submission
- Lead Source: 58% Lead source is from Google & Direct Traffic combined
- More than 8000 are Do Not Email, Do Not Call
- Modified, Email opened, SMS sent are last notable activity

Bivariate Analysis

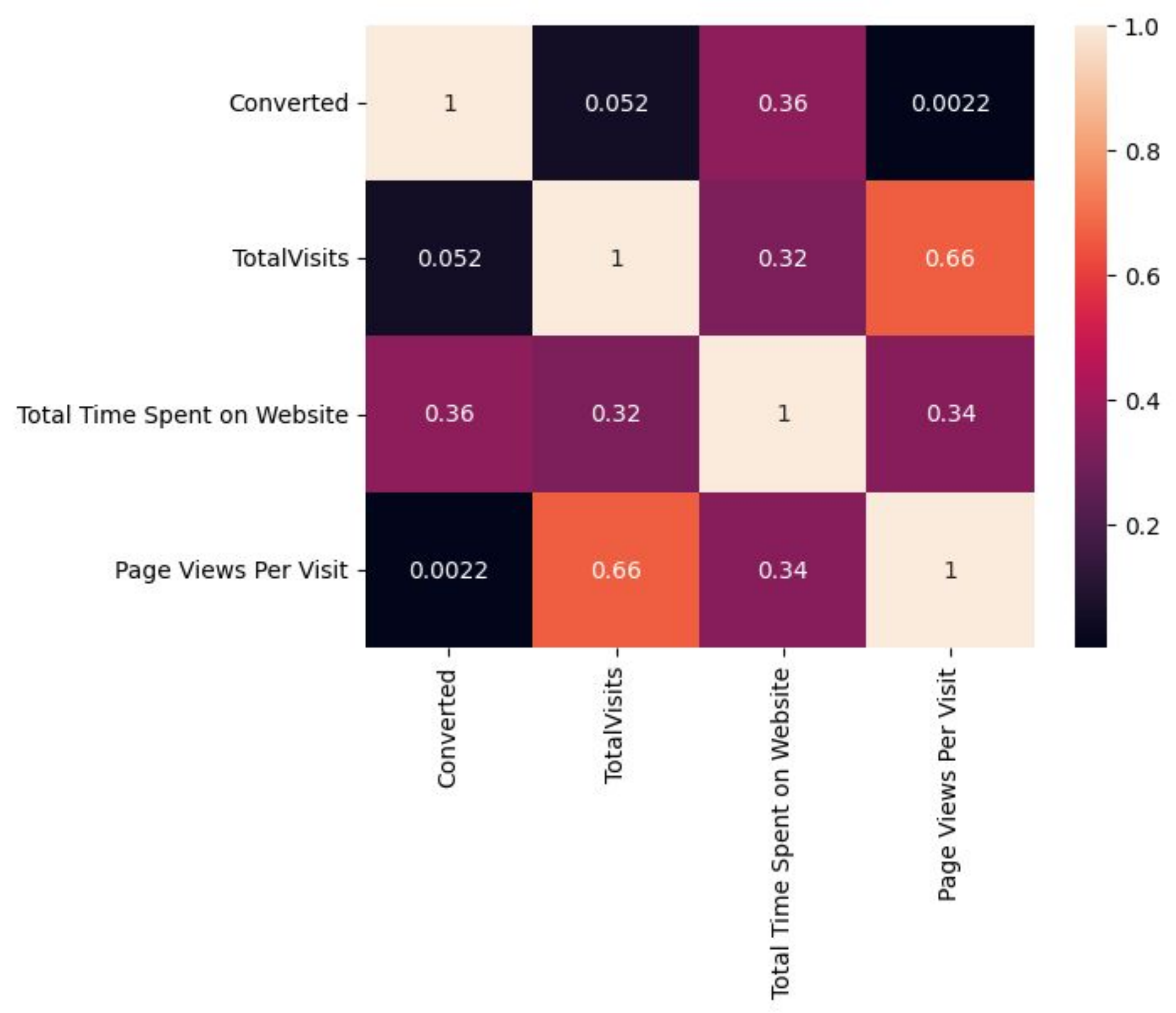


- Around 2000 of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 2000. The "API" identified approximately 2500 of customers with a lead conversion rate (LCR) of 1000
- 90%+ of the people has opted that they don't want to be emailed or Called about the course & 40% of them are converted to leads.
- 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities,
  'Email Opened' activity contributed 38% of last activities performed by the customers, with 37% lead conversion rate.

Bivariate Analysis of Continuous variables



- Past Leads who spends more time on the Website have higher chance of getting successfully converted than those who spends less time as seen in the box-plot



- Heatmap clearly shows ‘Total time time spent on website Has highest Correlation with Converted variable

Data Preparation before Model building

Binary level categorical columns were already mapped to 1 / 0 in previous steps

Created dummy features (one-hot encoded) for categorical variables ● Splitting Train & Test Sets

70:30 % ratio was chosen for the split

Feature scaling- Standardization method was used to scale the features

Checking the correlations

○ Predictor variables which were highly correlated with each other were dropped

Model Building

Feature Selection

The data set has lots of dimension and large number of features.

This will reduce model performance and might take high computation time.

Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.

Then we can manually fine tune the model.

RFE outcome

Pre RFE – 46 columns & Post RFE – 15 columns

Model Building

Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.

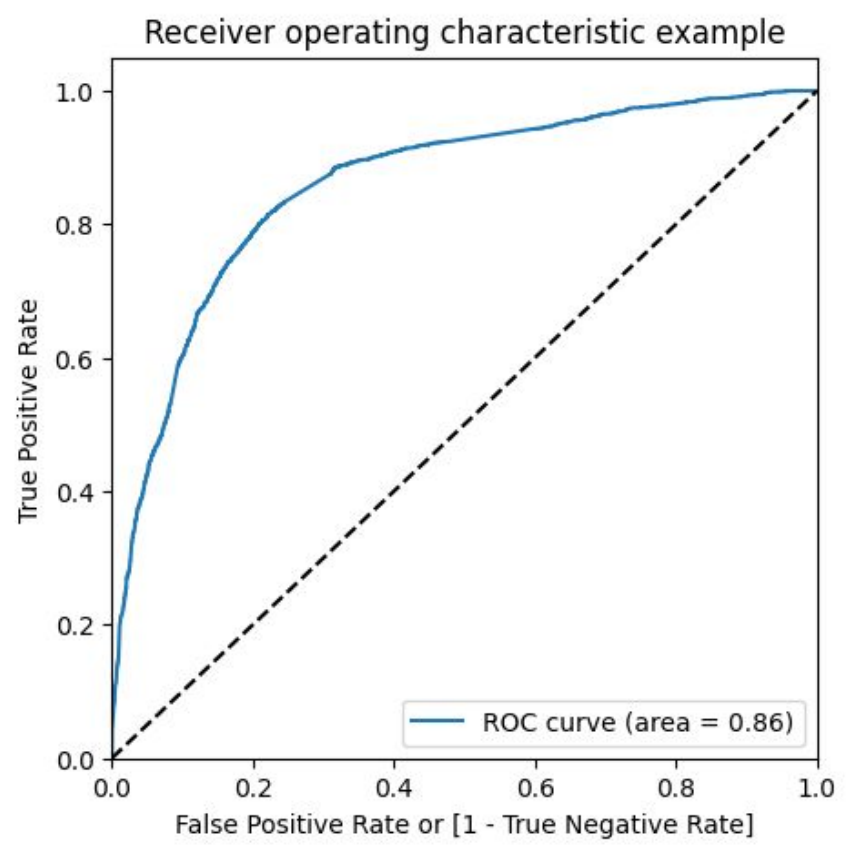
Model 2 looks stable after four iteration with:

significant p-values within the threshold (p-values < 0.05) and

No sign of multicollinearity with VIFs less than 5

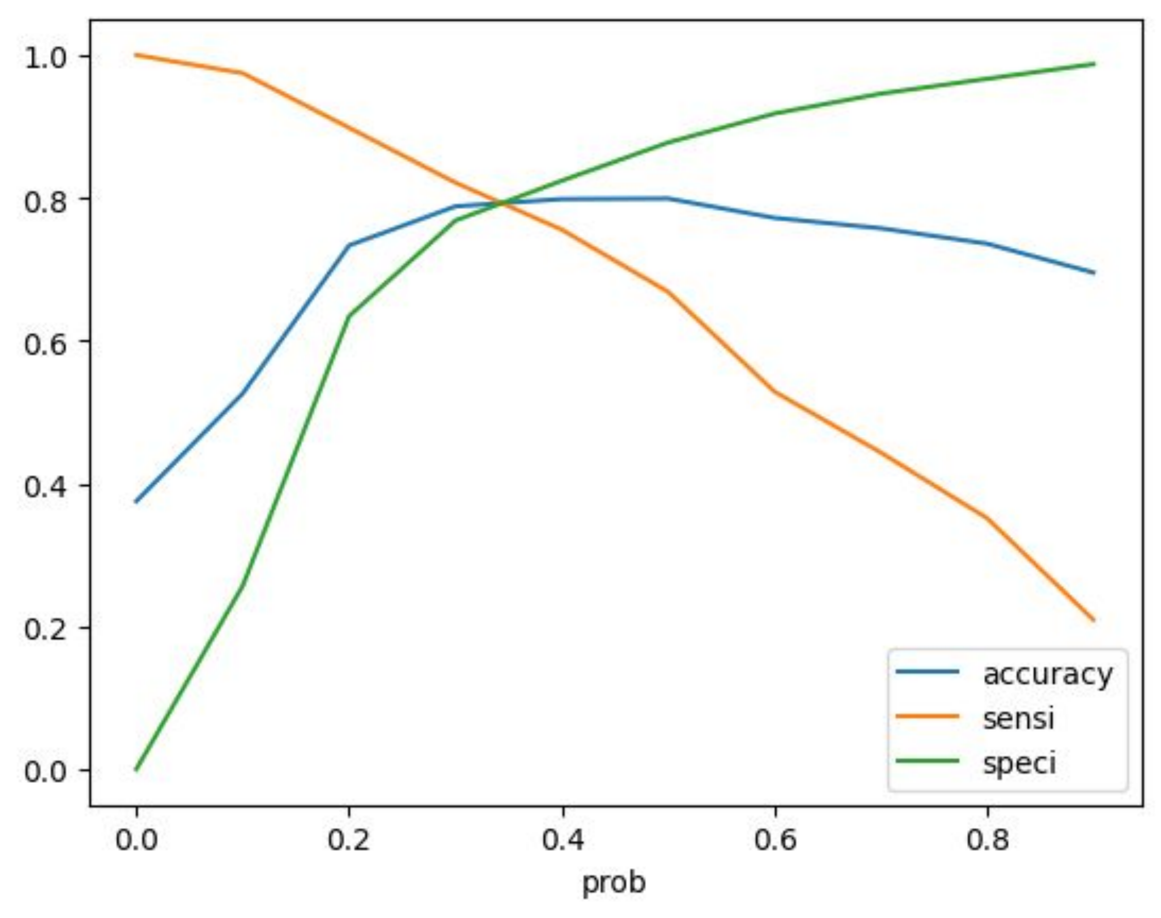
Hence, logm2 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

Model Evaluation - Train dataset



Area under ROC curve is 86% hence it is good predictive model

Model Evaluation - Train dataset



Cut off value-
It was decided to go ahead with 0.35 as cutoff after checking evaluation metrics coming from both plots

Model Evaluation - Test Dataset

Observation: So as we can see above the model seems to be performing well. The ROC curve has a value of 0.87, which is very good. We have the following values for the Train Data:

Accuracy : 79.48%

Sensitivity : 79.07%

Specificity : 79.73%

Test Data:

Accuracy : 79.89%

Sensitivity : 78.60%

Specificity : 80.69%

Recommendations Based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
  - Lead Source_Welingak Website: 4.93
  - Lead Source_Reference: 3.33
  - Last Notable Activity_Had a Phone Conversation: 2.85
  - Last Notable Activity_Unreachable : 1.67
  - Last Notable Activity_SMS Sent : 1.388
  - Total Time Spent on Website: 1.14
  - Last Notable Activity_Unsubscribed : 1.06

  We have also identified features with negative coefficients that may indicate potential areas for
- improvement. These include:
  - Last Notable Activity_Olark Chat Conversation: -1.61
  - Do Not Email_Yes: -1.45
  - Lead Source_Direct Traffic: -1.26

Conclusion

To improve the lead conversion rate, X Education should consider the following actions:

Strengthen the impact of positive features: Give priority to leads coming from 'Lead Source_Welingak Website' and 'Lead Source_Reference' as they have the highest positive coefficients. Focus on nurturing leads who have had a phone conversation or are marked as unreachable. Implement effective strategies for lead engagement through SMS communication. Additionally, continue to emphasize the importance of website engagement and encourage potential leads to spend more time on the website.

Address negative factors: Take steps to reduce the occurrence of 'Last Notable Activity_Olark Chat Conversation' and 'Do Not Email_Yes'. These factors may indicate potential barriers or concerns that hinder lead conversion. Explore ways to improve the effectiveness of communication through chat conversations and consider alternative approaches to email communication. Evaluate the impact of 'Lead Source_Direct Traffic' and identify opportunities to optimize lead generation from this source.

By focusing on the identified positive factors and addressing the negative factors, X Education can enhance its lead conversion rate. Continuous monitoring, evaluation, and refinement of marketing and sales strategies based on these insights will contribute to the overall growth and success of the company.

Thank you