X Education – Lead Scoring Case Study

Team Members: Dikshant Singh, Deepali Pathak & Dhanashri Amar Patil

Table Of Content:-

Introduction:

Problem Statement: An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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.

Business Goal:

The company aims to optimize lead conversion by implementing a lead scoring model. This model will assign each lead a score reflecting its likelihood of converting into a paying customer. Leads with higher scores are anticipated to have a greater probability of conversion, while those with lower scores are expected to have a lower probability. The CEO has set a target lead conversion rate of approximately 80%, guiding the development of the lead scoring model.

Analysis Approach

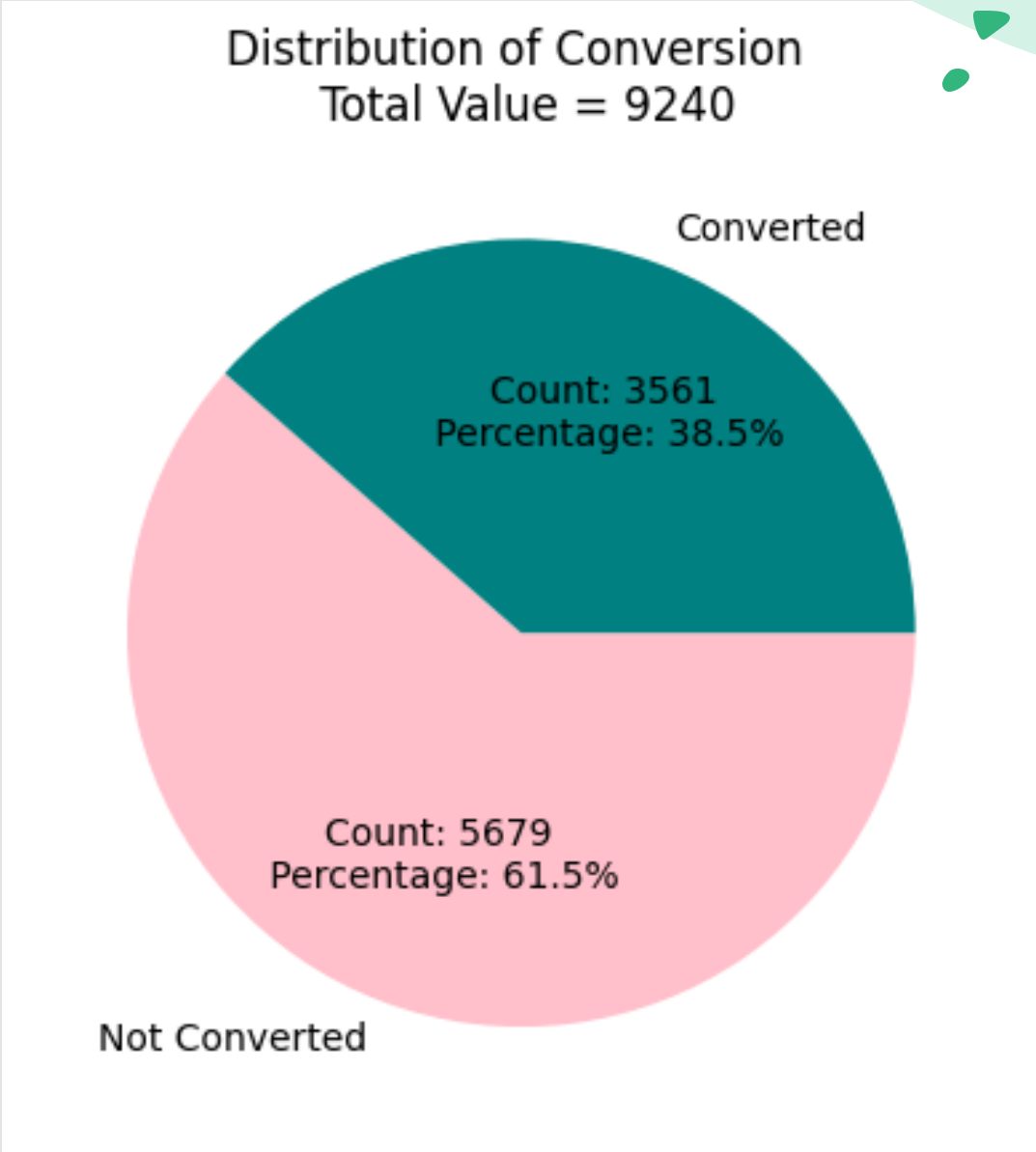
- Data Cleaning
- EDA
- Data Preparation
- Model Building
- Model Evaluation
- Prediction on Test/Unseen Data
- Conclusion

Data Cleaning:

- Handled columns with the value 'Select'.
- Removed all the columns which has more than 25% missing values in them.
- Handled skewed columns like 'Do Not Call', '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', 'I agree to pay the amount through cheque'.
- Categorized categories in Lead Source columns.

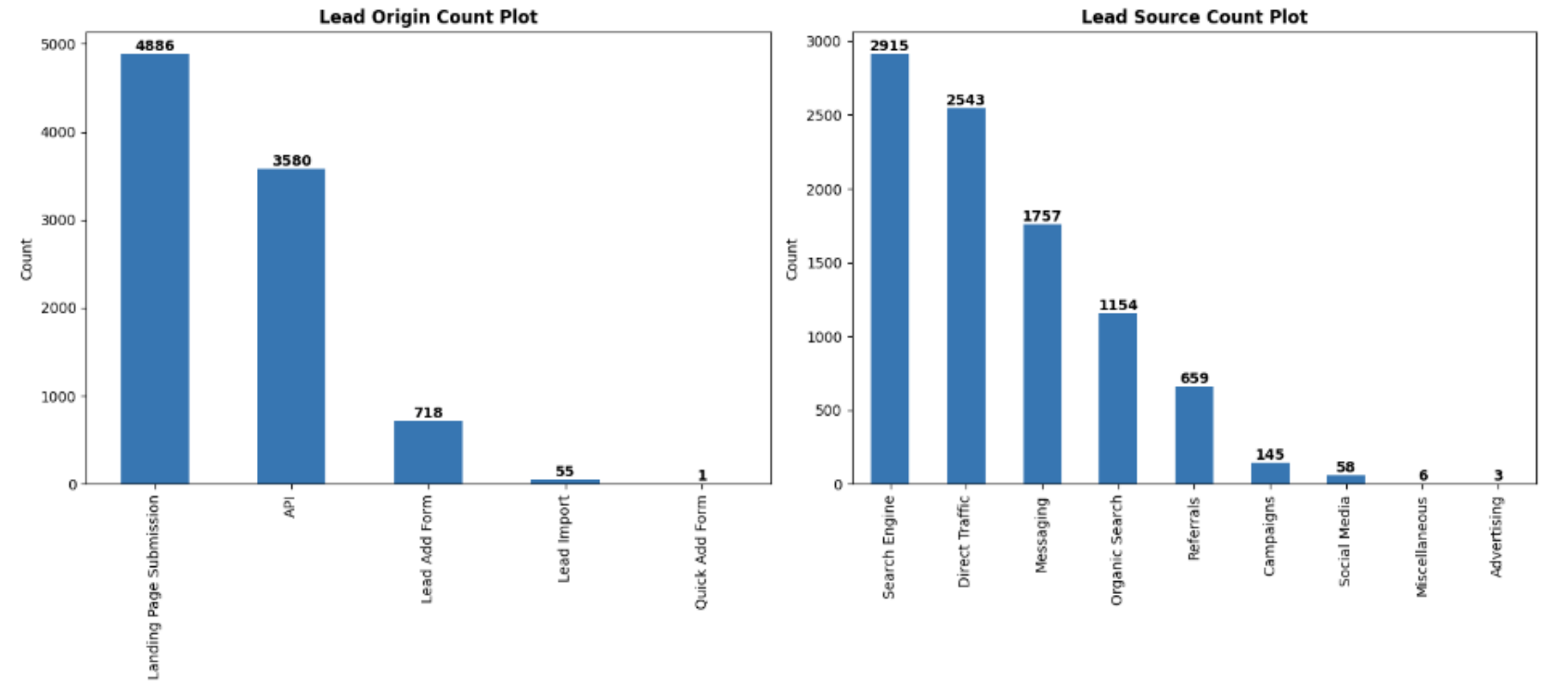
Exploratory Data Analysis:

Distribution of Target Column: As we can see the data is Imbalanced.



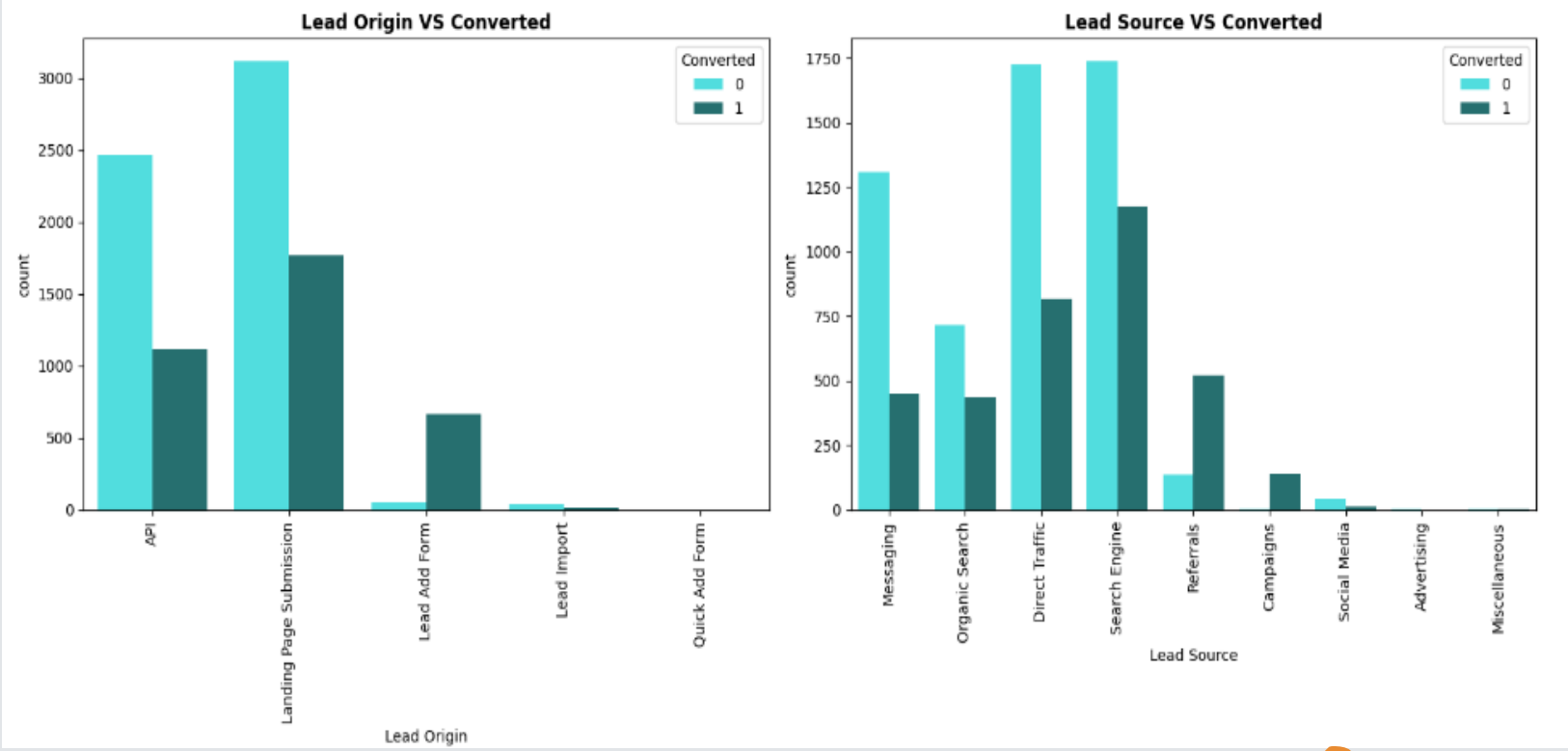
Exploratory Data Analysis:

Company is receiving a greater number of leads from Landing Page Submission and Search Engine



Exploratory Data Analysis:

- Rate of conversion is maximum if lead is generated from Add forms.
- Higher rate of conversion is from referrals and campaigns, but the greatest number of users traffic is from Search Engines.



Data Preperation:

- Encoded categorical data using get_dummies.
- Split the data into 70:30 ratio 70% in train data 30% in test data
- Scaled Numerical columns using MinMaxScaler because numerical data in this case is not uniformly distributed.

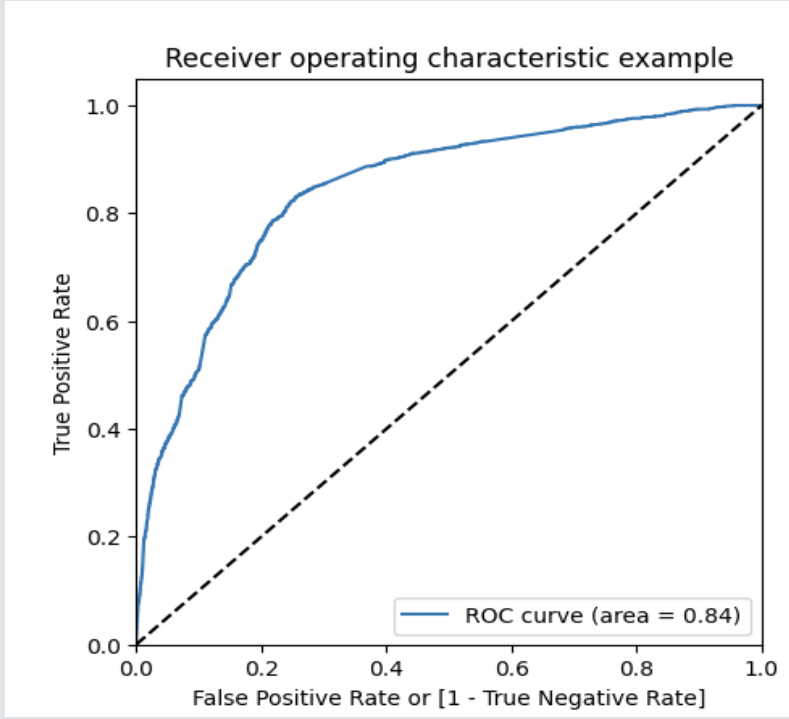
Model Building:

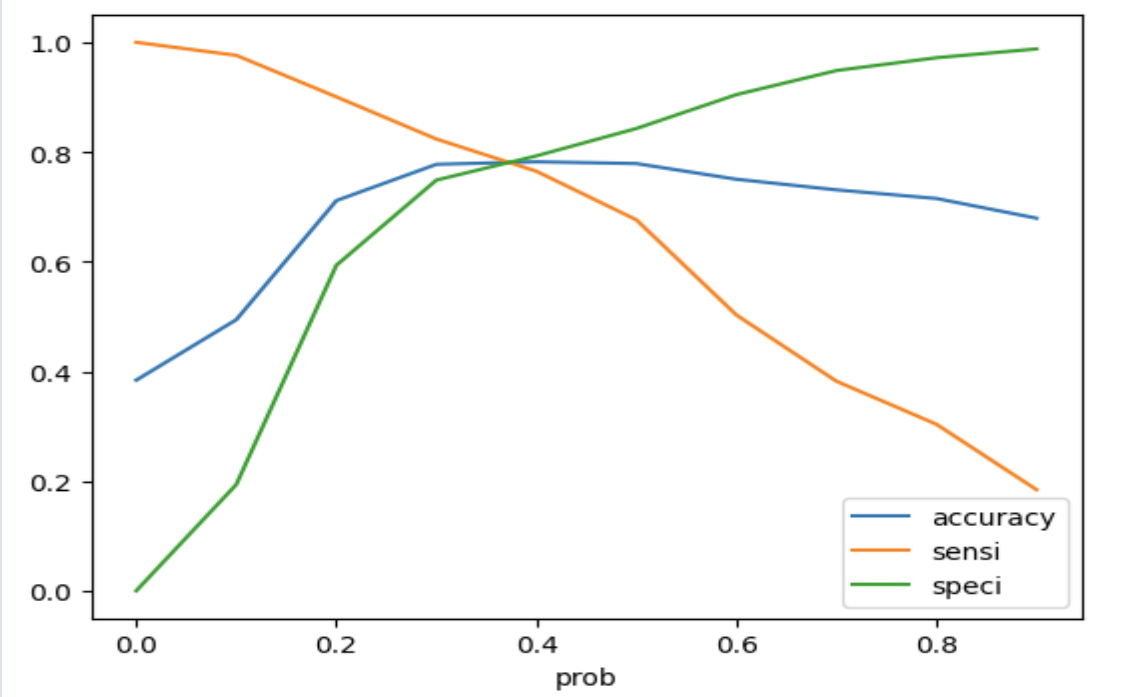
- There were 26 columns before elimination of the columns.
- Using RFE we kept only 17 columns.
- After that we removed the columns by checking P_value less than 0.05 and VIF less than 5.

Model Evaluation:

For Training Dataset: -

- In Model evaluation we got cutoff probability as 0.37.
- Accuracy: 78.18%
- Sensitivity: 78.78%
- Specificity: 77.81%
- And area under ROC is 0.84

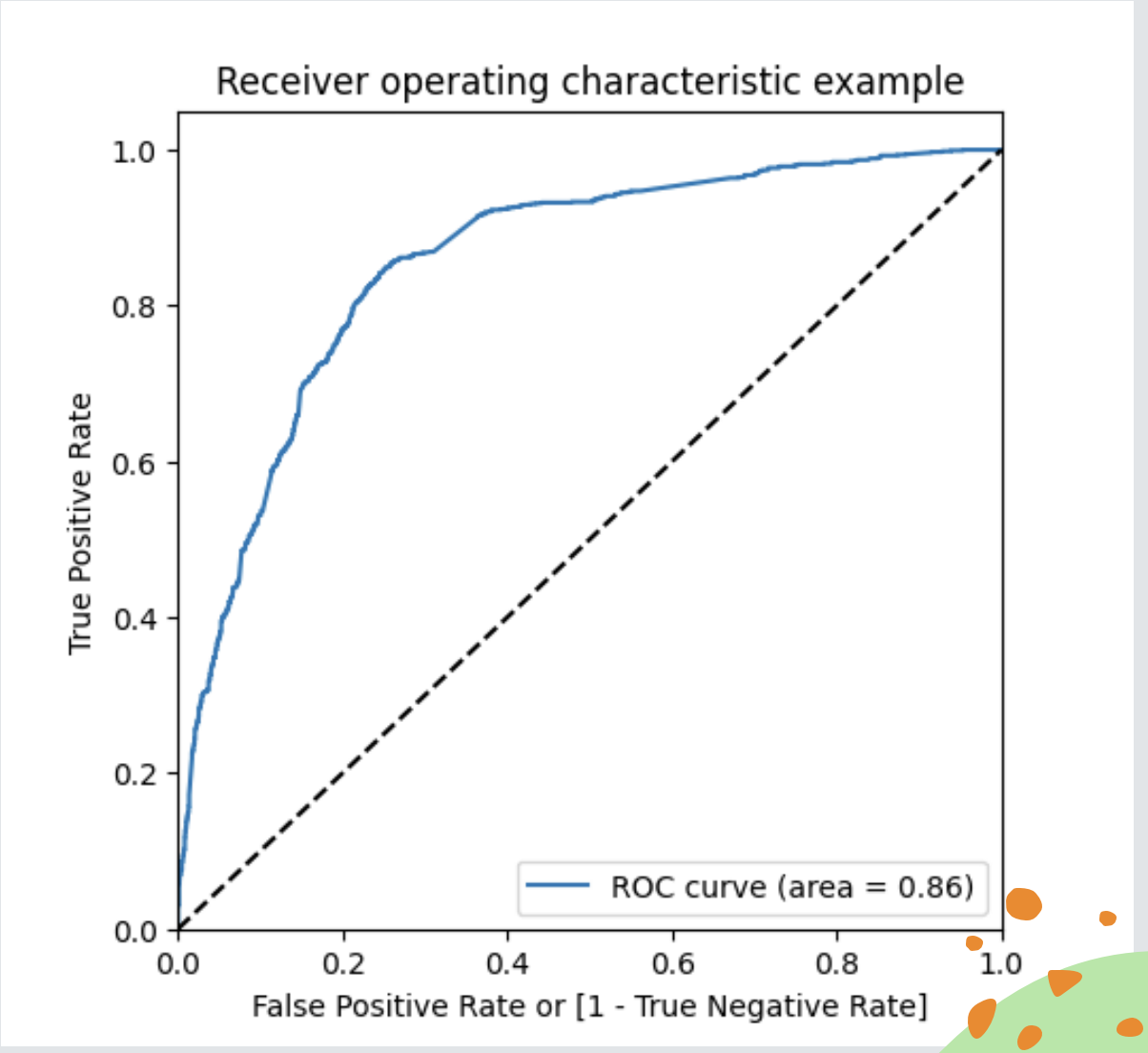




Model Evaluation:

For Test Dataset: -

- Accuracy: 79.04%
- Sensitivity: 81.89%
- Specificity: 77.23%
- Area under ROC curve is 0.86



Recommendation:

- In this business scenario, sensitivity (recall) is the most crucial evaluation metric. It measures the model's ability to correctly identify positive instances, ensuring that we capture as many true positives as possible, which is vital for the success and effectiveness of our strategy.
- In terms of sensitivity, our model has achieved an 81.89% success rate in predicting users who can be converted, effectively identifying the majority of potential conversions.
- Below Mentioned colums are more important for predicting lead conversions: -

Total Time Spent on Website	4.321250
Lead Source_Campaigns	3.469814
Last Notable Activity_Had a Phone Conversation	3.385652
Last Notable Activity_Unreachable	2.041369
Last Notable Activity_SMS Sent	1.572506