Lead Scoring Case Study

Submitted by: Pranav Khanna

Problem Statement

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- The company solely relies on leads generated through these platforms for their business.
- Even though a lot of leads are generated in the initial phase, only about 30 percent of them convert into customers.

Business Objective

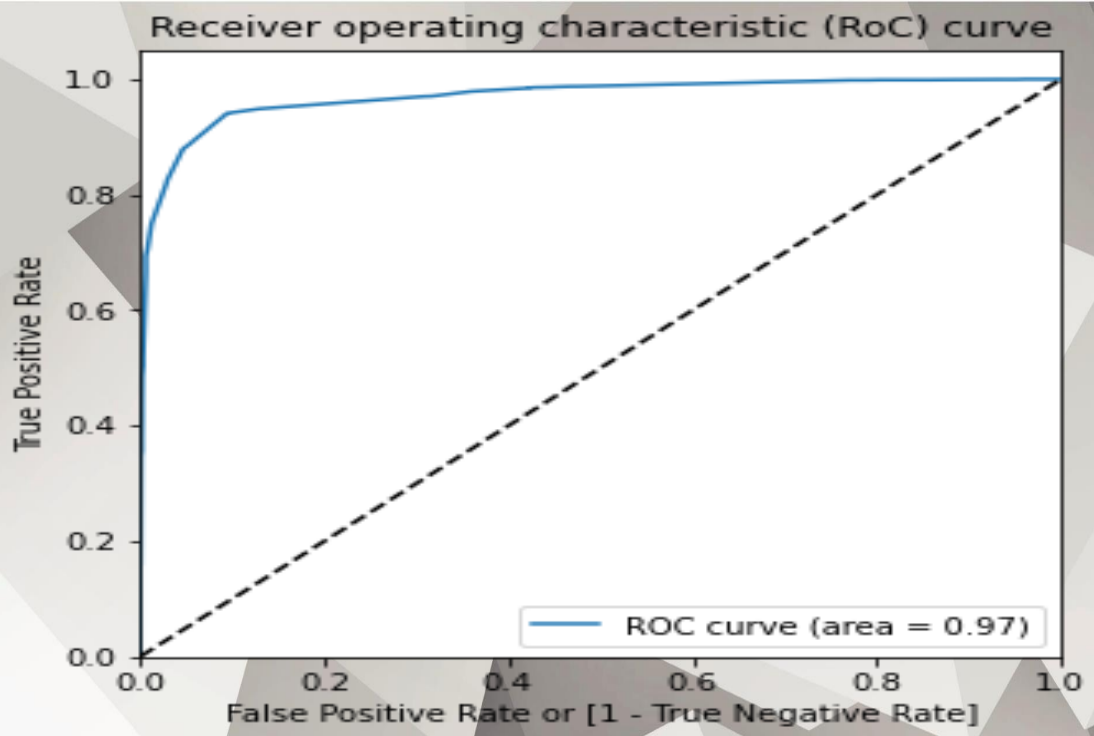
- Company wants make this process more efficient and wishes to identify the most potential leads, also known as ‘Hot Leads’.
- The company wants to analyze and get this conversion rate up as much as possible through data driven solutioning.

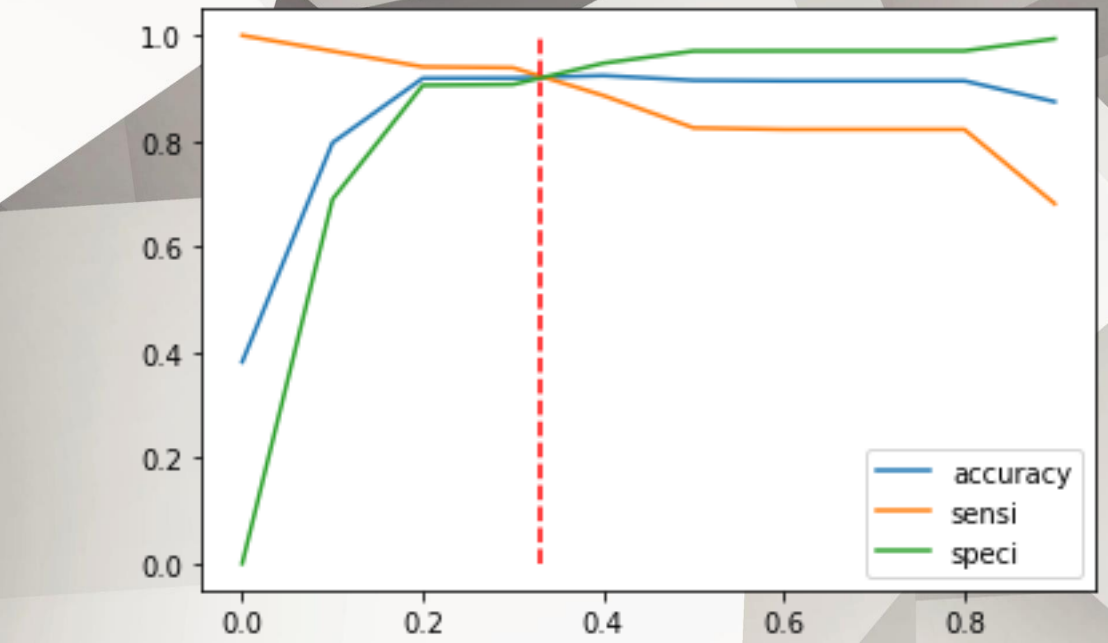
Solution Strategy

- Compiling and analyzing the data provided by the company.
- Cleansing the data and preparing it for modelling.
- Creating and Evaluating a logistic regression model.
- Finding out the features/variables that affect the conversion rate of generated leads.

Data Analysis

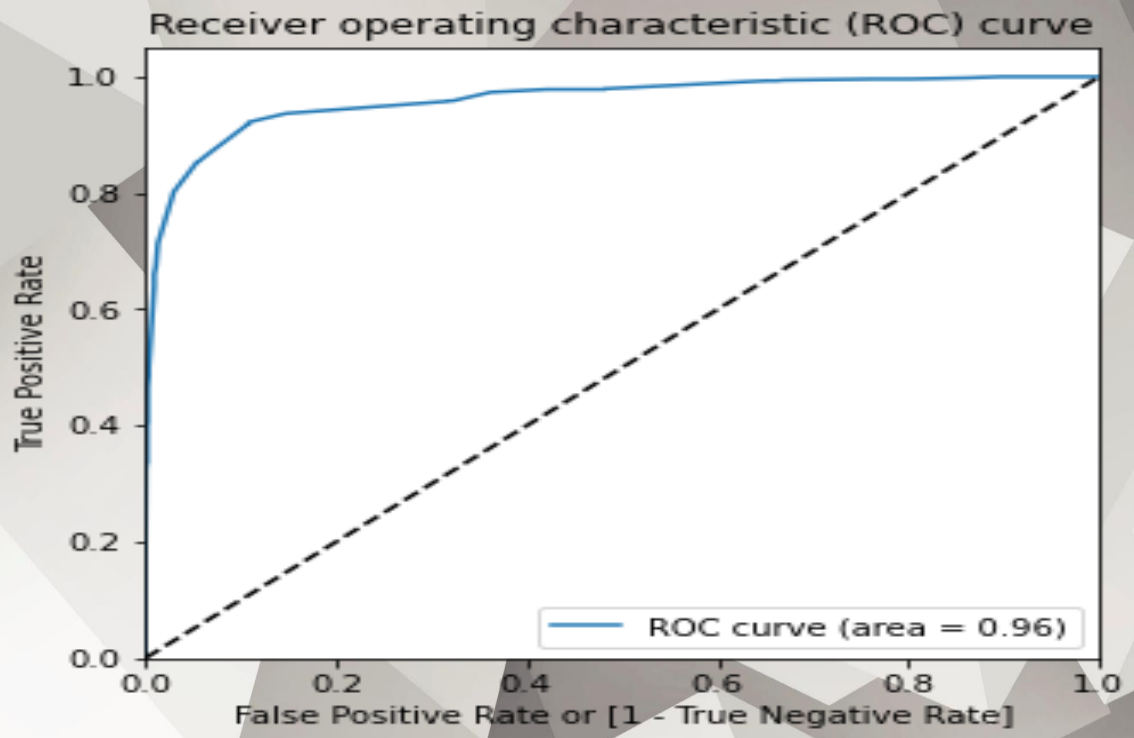
- It was found during analysis that out of 37 features in the dataset, 17 features contained Null Values which could have affected our Analysis.
- Out of these 17 features, 10 features had more than 40 percent null values. These features were analyzed individually and null values were imputed however necessary.
- The numerical continuous null values were imputed with Mean and Median, wherever necessary.
- The categorical null values were substituted with 'Unknown', wherever necessary.
- Outliers were handled for all the numerical variables.
- After the data cleaning was done, variables with only 1 unique value were dropped since they weren't adding any value to the analysis because they did not have any other values to be compared with as in the case of most other categorical variables.

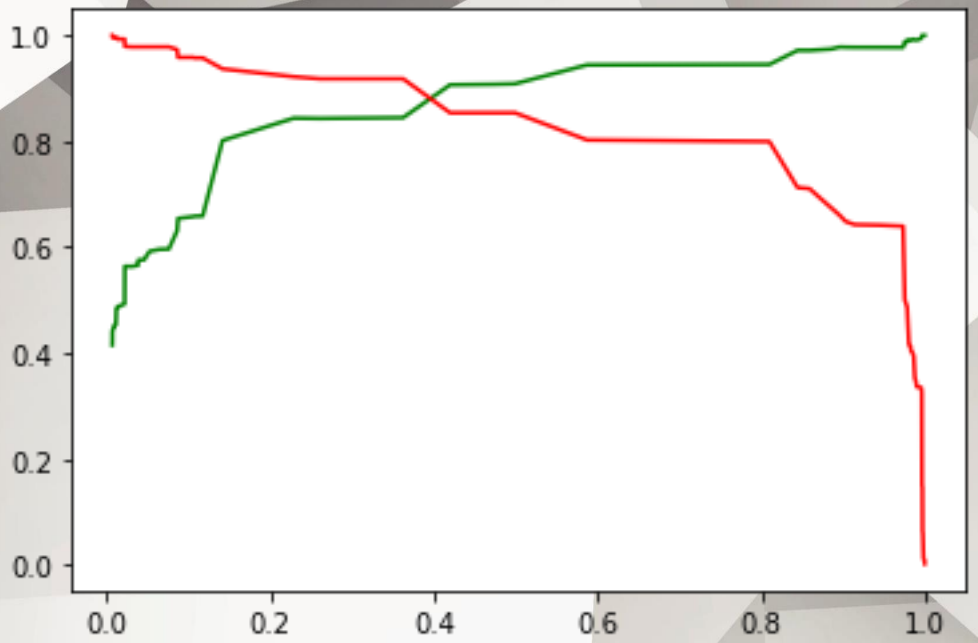




Model Evaluation on Training Dataset

- With the final Model, we get ROC curve area=0.97 which is very good and with cut-off value at 0.33 we got the below data:
  - Accuracy : 91.9 %
  - Sensitivity : 93.9 %
  - Specificity : 90.7 %





Model Prediction on Test Dataset

- With the final Model, we get ROC curve area=0.96 which is very good and with cut-off value at 0.33 we got the below data:
  - Accuracy : 90.2 %
  - Sensitivity : 91.8 %
  - Specificity : 89.2 %

Features that Positively Affect the Model

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Lead Source_Welingak Website
- Tags_Will revert after reading the email
- Last Activity_SMS Sent
- What is your current occupation_Working Professional
- What is your current occupation_Unemployed

Features that Negatively Affect the Model

- Tags_Graduation in progress
- Last Notable Activity_Modified
- Tags_Interested in full time MBA
- Tags_opp hangup
- Tags_Interested in other courses
- Tags_Diploma holder (Not Eligible)
- Tags_Ringing
- Tags_switched off

Conclusion:

- The overall model is able to predict the conversion rate up to 91 % which was expected from it on both Train and Test data with cut-off 0.33.
- Company should focus more on Working Professionals, and Unemployed people who are looking to upskill themselves to be more relevant in the job market.
- Most leads generating source online for the company turns out to be Welingak Website.
- The company should not spend much resources on people interested in other courses and full time MBA, people who hung up the call and did not answer.