Lead Scoring Case Study using Logistic Regression

Submitted by:

Aahna Ranjan

Monish Patle

Sharad Kumar Rai

DS C50

Problem Statement

1. An Education company named X Education provides online courses to the industry professionals.
2. It markets itself on Google and various other websites where viewers can browse about the online courses that they provide.
3. When the viewers provide their email address and phone number, then they are considered as leads.
4. The lead conversion rate for the company is around 30%, which is considered to be very low by the company.
5. Company wants to increase its conversion rate and they want to classify their leads as “hot leads” which are most potential and “cold leads” which are less potential.
6. Company aims at target lead conversion rate to be 80%.

Data

1. The file consists of 9000 data points of past leads.
2. “Converted” is target variable with values as 0 for “not converted” and 1 for “converted”.

Aim

1. To build a linear regression model to predict for the hot leads, on the basis of the lead score between 0 and 100.
2. It is considered that the leads with the higher lead score are “hot leads” and with low lead score are “cold leads”.
3. Hot leads are the leads that are likely to be converted.

Approach

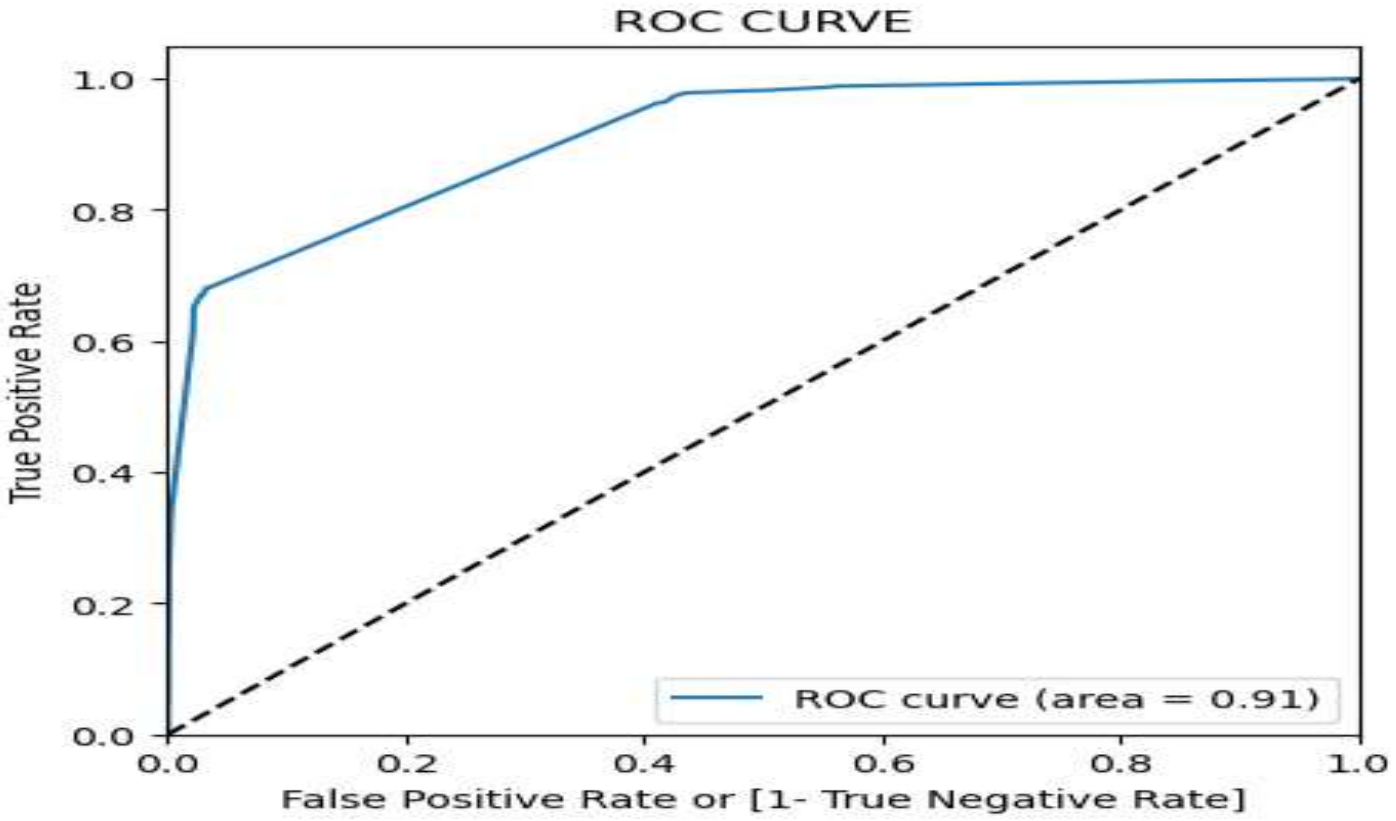
Following steps were performed in building of model and categorizing the leads as “hot leads” and “cold leads”:

1. Reading 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

Approach

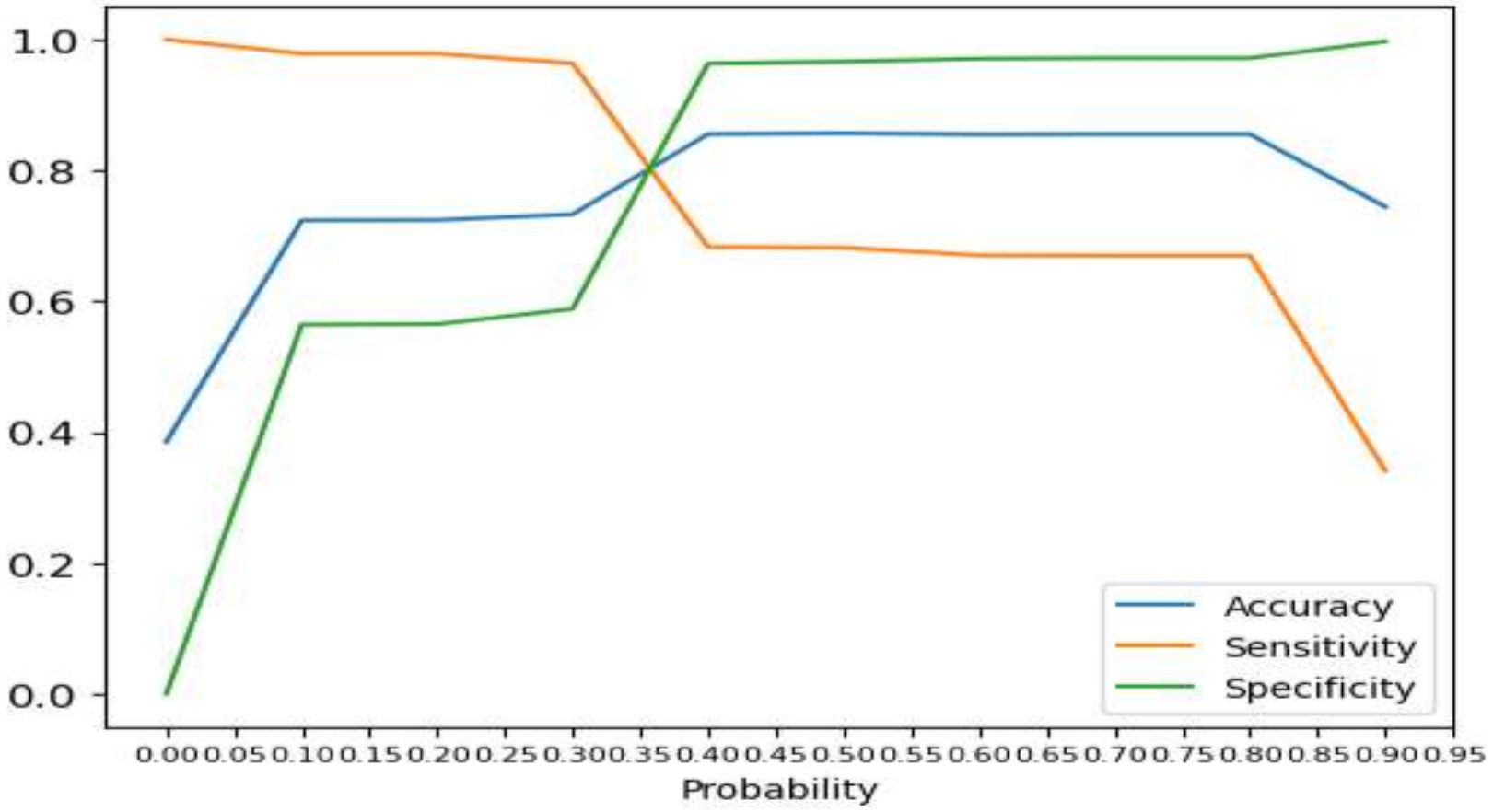
10. Model Building
11. Model Evaluation
12. Plotting of ROC curve
13. Calculating optimal cut off point
14. Calculating Sensitivity, Specificity, Accuracy
15. Calculating Precision and Recall
16. Prediction of Test set
17. Deciding leads as “hot leads” and “cold leads” based on lead score.

Technical Analysis



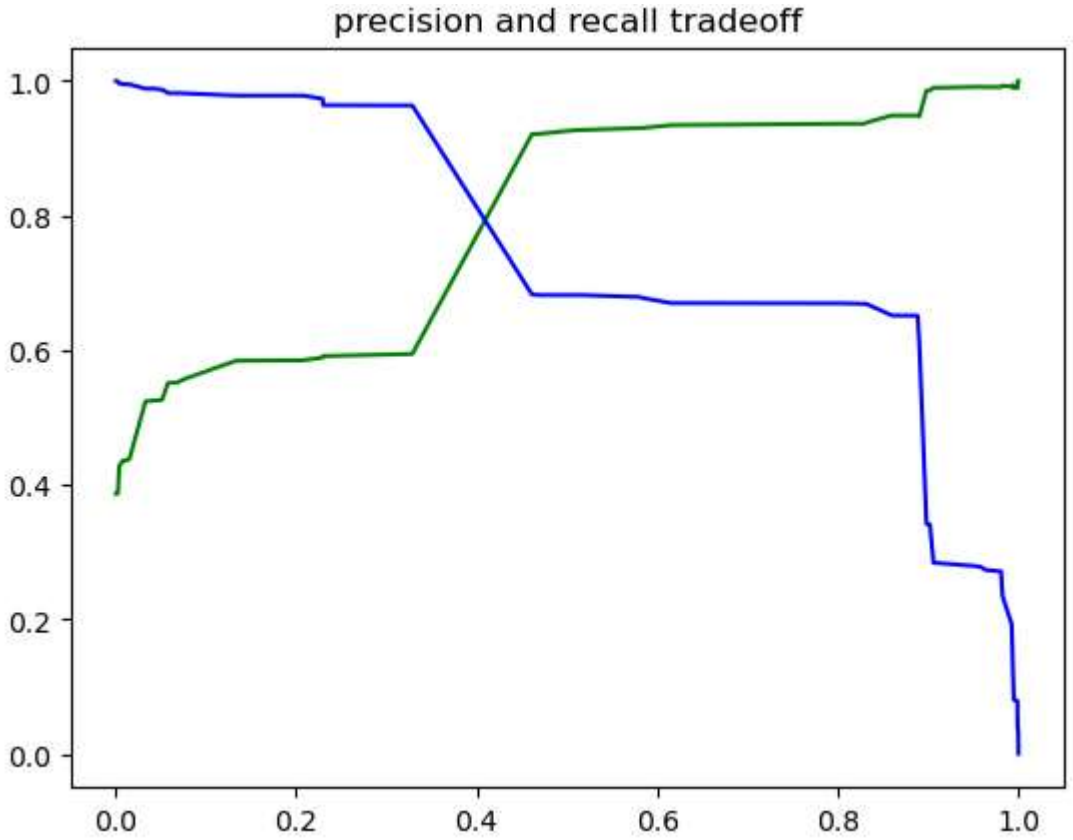
ROC curve has area as 0.91 which indicates that our model is 91% correct.

Technical Analysis



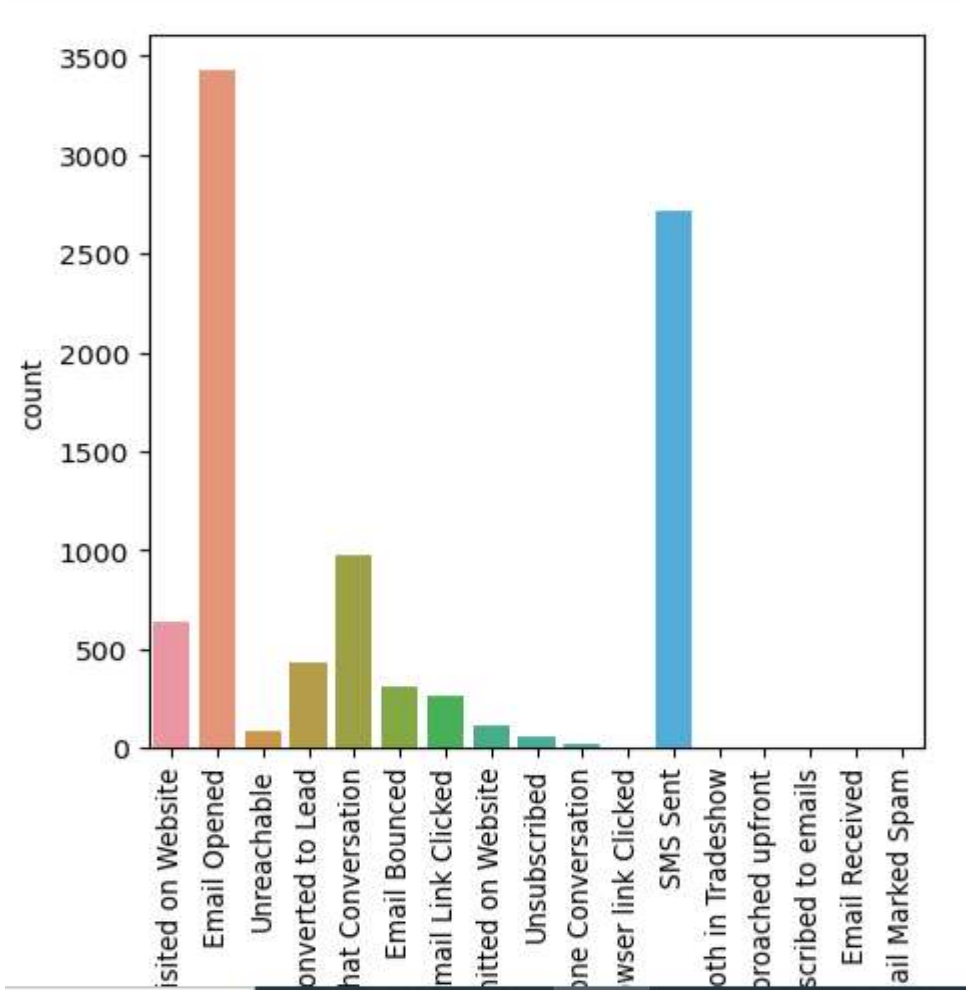
The Accuracy, Specificity, Sensitivity curve meets at 35, which shows that optimal cutoff point is 35%

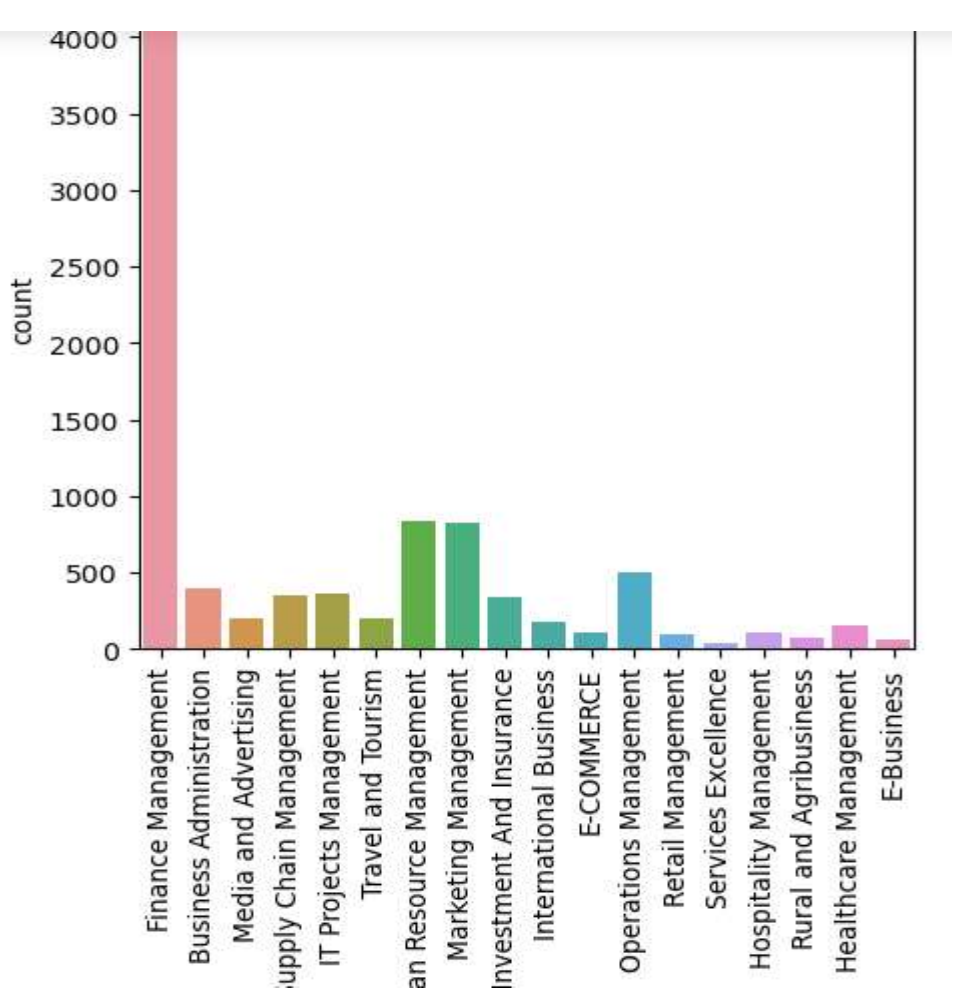
Technical Analysis



This trade off graph shows that precision and recall are inversely related.

Business Analysis





* Email opened and sms sent are the recent activities that are mostly done by leads, so company should focus on these two for contacting viewers..

* Most of the leads are from finance management specialization so company must focus on that.

Suggestions.

1. Current occupation of students, time spent on website, lead origin are 3 top variables which contributed towards probability of leads getting converted into customers.
2. Company should less focus on unemployed leads.
3. Company has to continuously monitor the leads conversion rates and adjust the strategy based on the effectiveness of various channels and approaches.
4. Company should prioritize the leads which middle lead score to convert them to high lead score.

Thank You