Lead Scoring Assignment

PREPARED BY SWATI & IFFATH

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%.

Business Objective

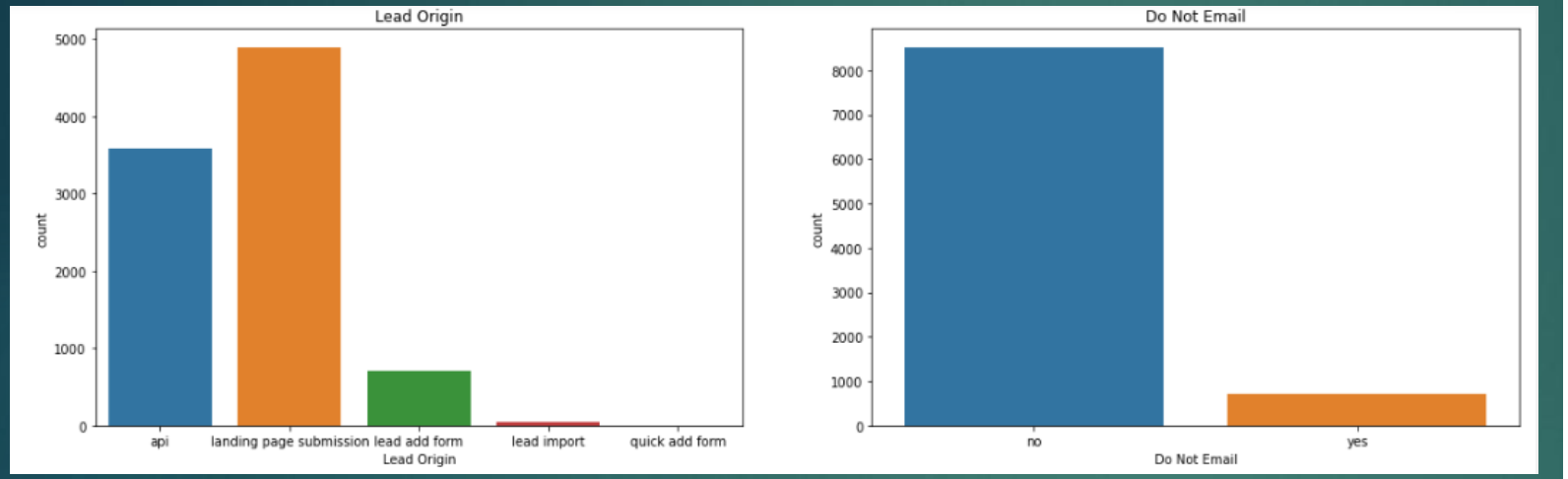
X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

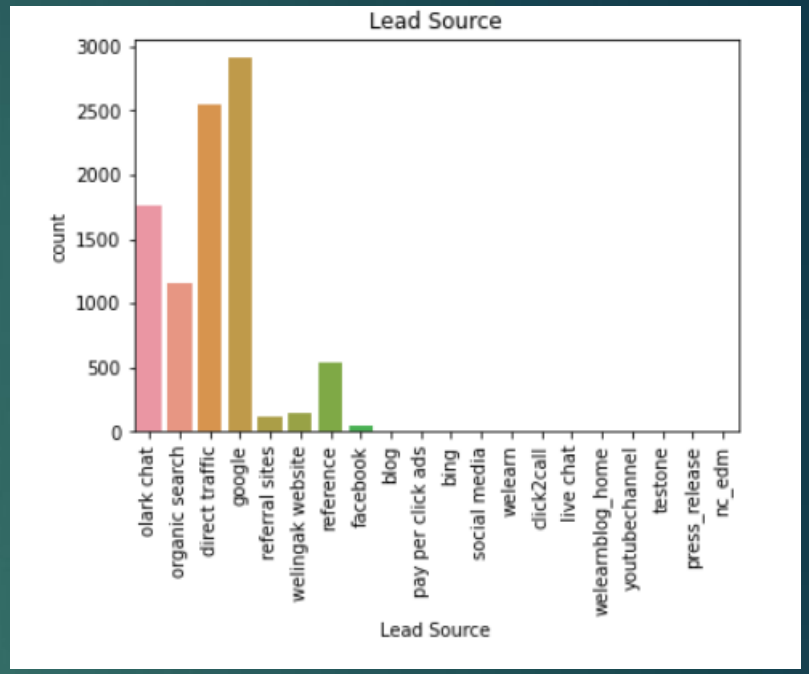
To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Approach

- Import the data
- Perform data cleaning for analysis
- Scaling the data
- Building data model
- Building logistic regression model
- For each leads assign the lead score
- Test the model on train set
- Evaluate model
- Test the model
- Measure the accuracy, specificity and sensitivity of the model

EDA- Leads data analysis

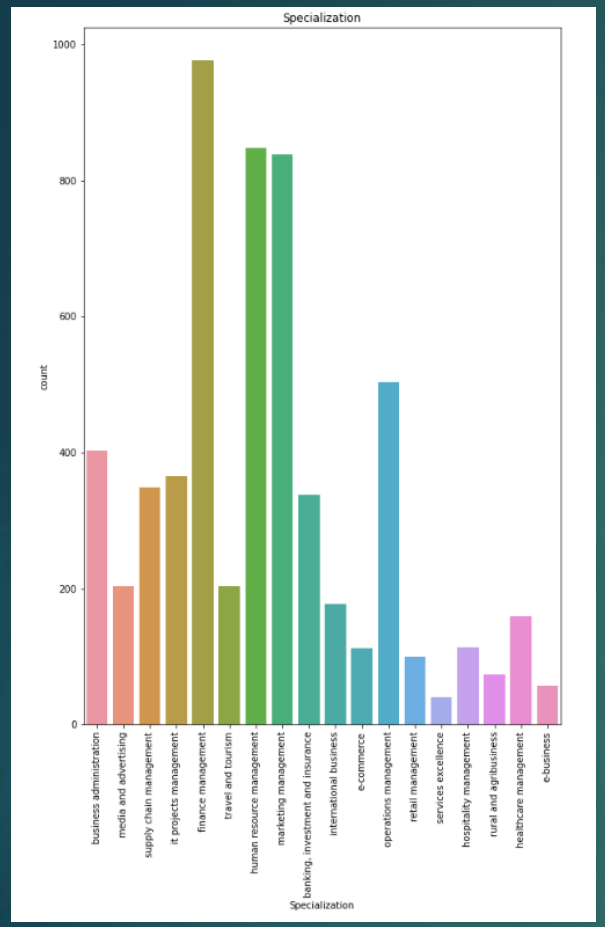


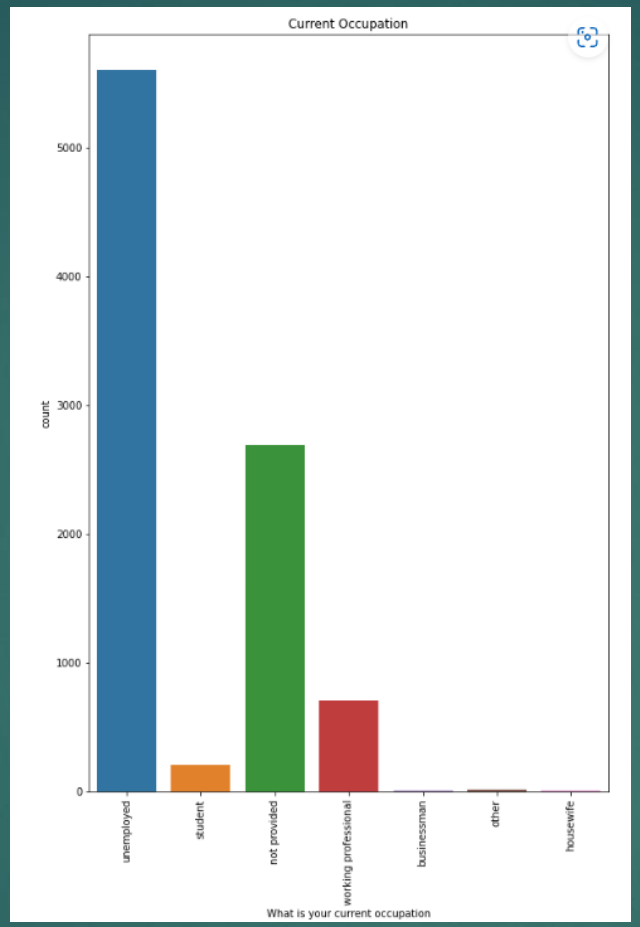


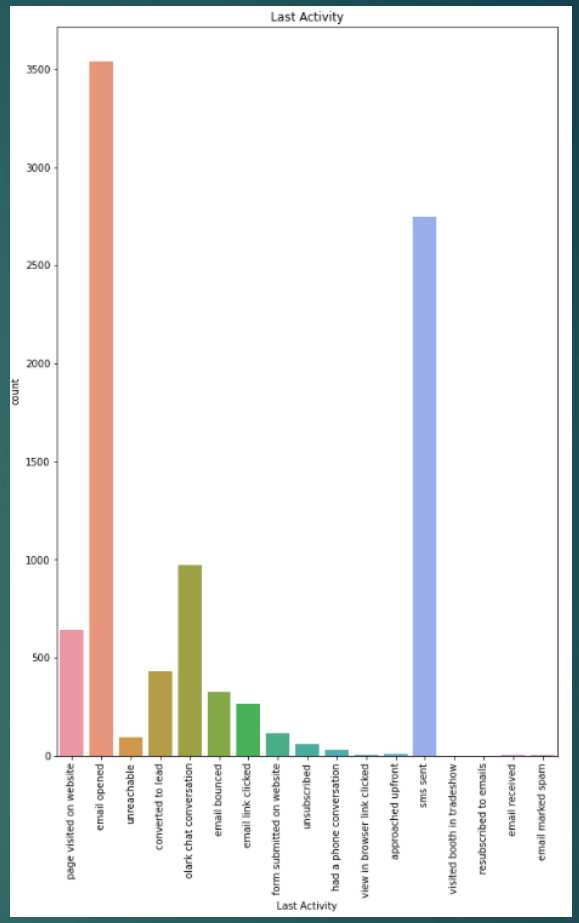
Comments

- 38% customer got converted as per leads data
- For Lead Origin Landing page has max count
- Google has maximum count for Lead Source

EDA- Leads data analysis



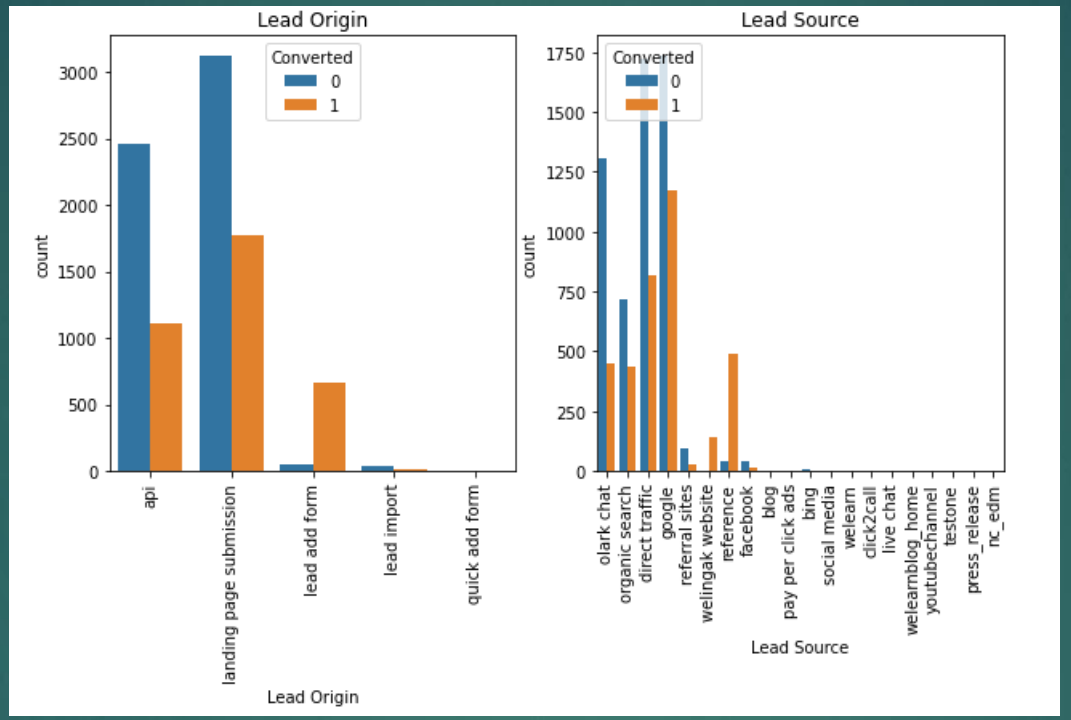




Comments

- Highest leads have came from Finance management for Specialization,
- For Current Occupation it is max for unemployed and
- For Last activity it is max for email opened.

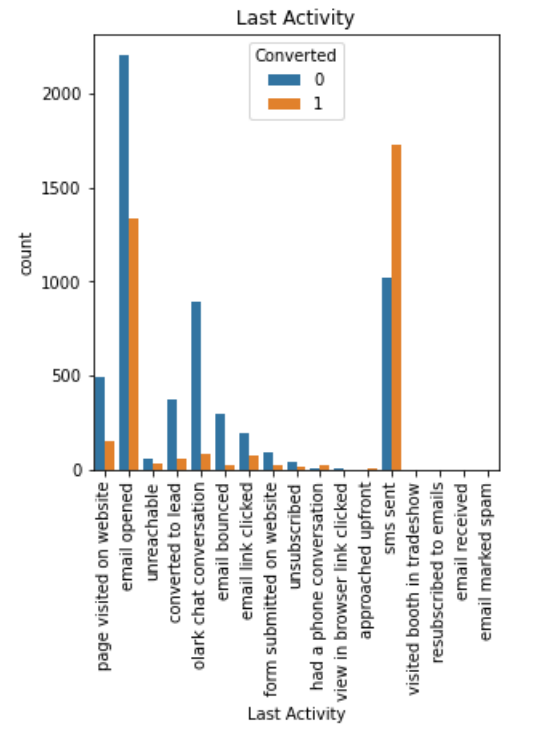
EDA- Leads data analysis

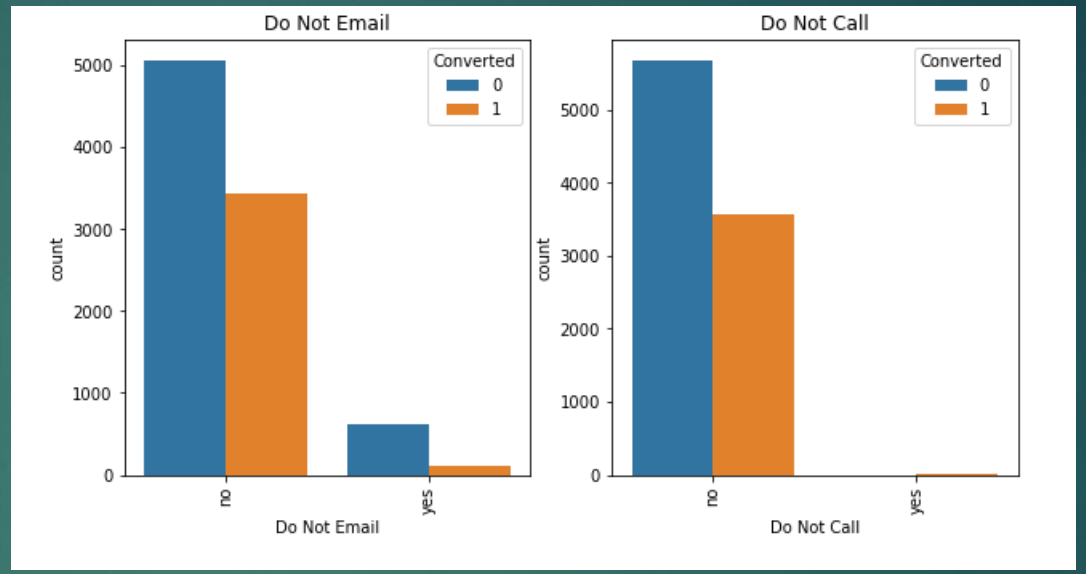


Comments

- Lead Origin with landing page submission has a greater conversion rate
- Where as for Lead Source Google shows maximum conversion rate

EDA- Leads data analysis

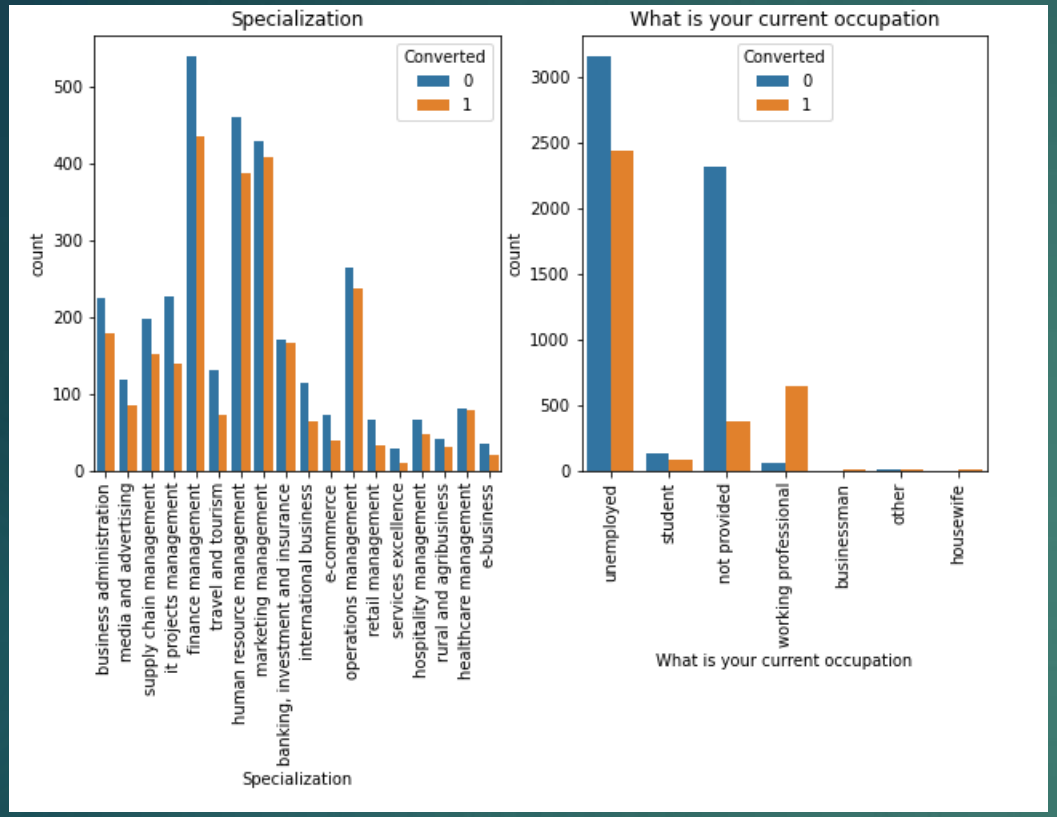


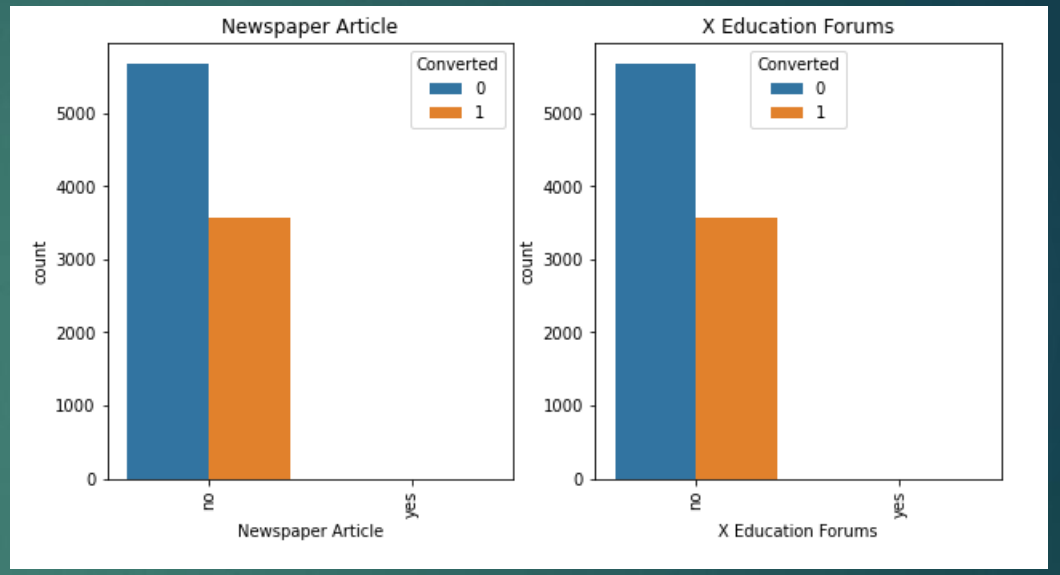


Comments

- Sms sent has highest conversion rate followed by email opened
- Also people who selected do not email/call as no have higher conversion rate

EDA- Leads data analysis





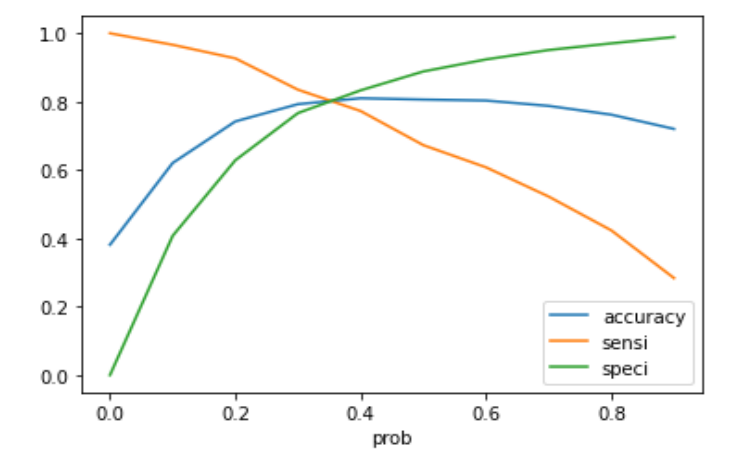
Comments

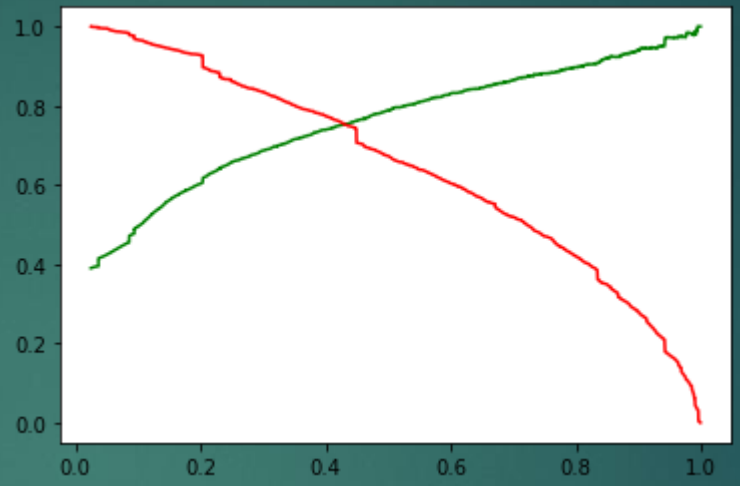
- Conversion rate is higher for unemployed
- And it is seen higher for finance, management, banking specialization

Model Building Approach

- Splitting data into train and test sets.
- Scale variable in train set.
- Build the first model.
- Use RFE to eliminate less relevant variables.
- Build the next model.
- Eliminate variables based on high p- values
- Check VIF value for all the existing columns.
- Predict using train set
- Evaluate accuracy and other metrics.
- Predict using test set.
- Precision and recall analysis on test predictions

Model Evaluation (Train set)

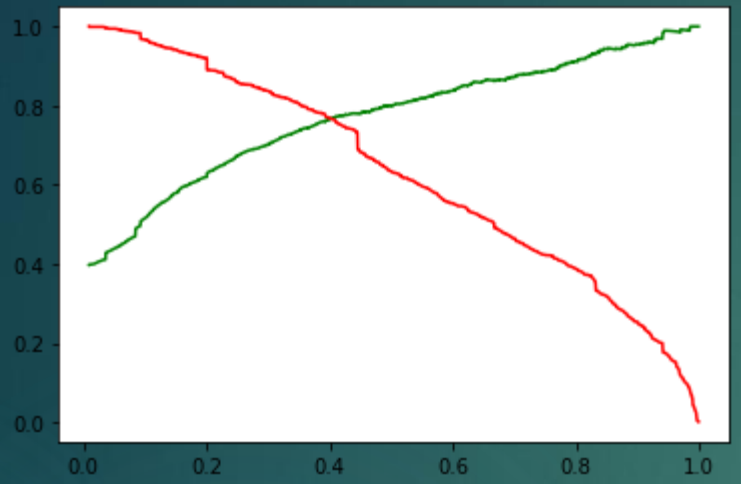




Comments

- Accuracy : 79%
- Sensitivity : 83%
- Specificity : 77%
- Precision: 69%
- Recall: 84%

Model Evaluation (Test set)



Comments

- Accuracy : 80%
- Sensitivity : 83%
- Specificity : 77%
- Precision: 70%
- Recall: 83%

Summary

EDA:

- People spending higher than average time are the promising leads, so targeting them and approach them can be helpful to conversions.
- Finance, Marketing and Human resource management has high conversion rates, people from this specialization can be a hot leads.
- Landing page submission can help find out more leads.

Logistic regression model:

- The model shows high close to 80% accuracy
- The threshold has been selected from Accuracy, Sensitivity, Specificity measures and Precision, Recall curves
- The model shows 84% Sensitivity and 77% Specificity