Lead Scoring Case Study

Pallavi Nishankar | Shekhar More | Shivraj Pujari

DS C47 August Batch

Problem Statement

- An Education company named ‘X Education’ sales online course to industry professionals.
- Now, although X Education gets a lots leads, its lead conversion rate is very poor of about 30%.
- The company wants to increase its lead conversion rate to 80%.

Goal

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Strategy

- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistic regression model

-

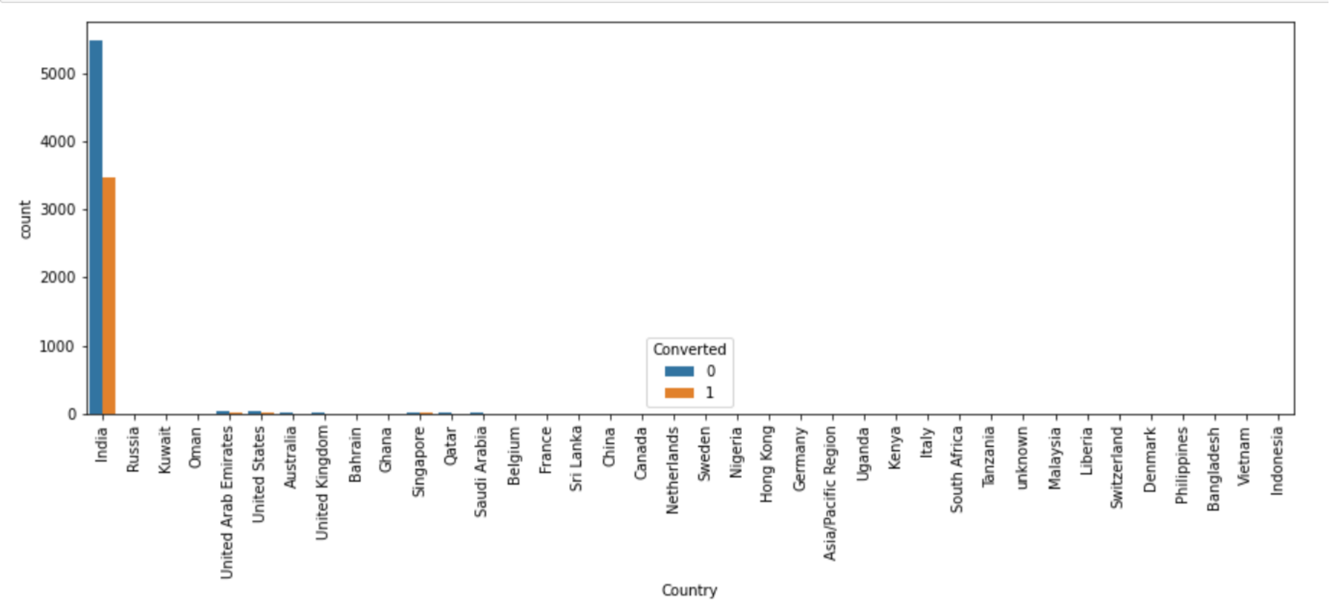
- Test the model on train set
- Evaluate model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics of evolution.

Top factors that impact the conversion of leads

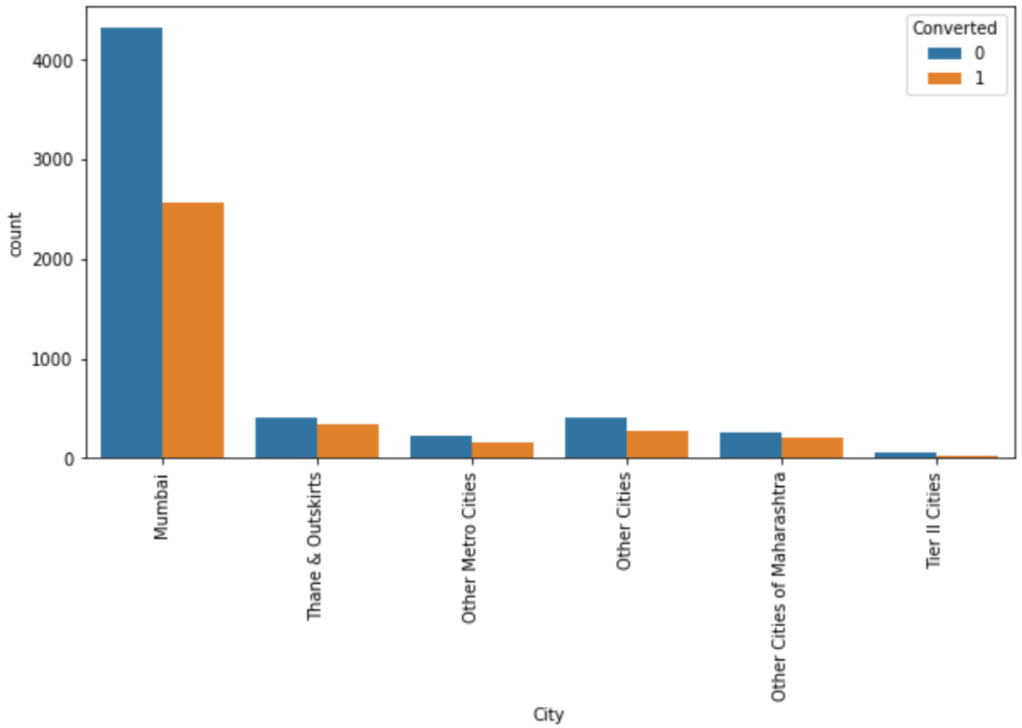
Features
Tags_Will revert after reading the email
Total Time Spent on Website
TotalVisits
Lead Origin Lead Add Form
Last Notable Activity_SMS Sent
Last Notable Activity_Modified
Lead Source_Olark Chat
Lead Profile_Potential Lead
Lead Source_Welingak Website
Tags_Closed by Horizzon
Lead Quality_Not Sure
Do Not Email_Yes
Tags_Lost to EINS
Lead Profile_Other Leads
Last Notable Activity_Olark Chat Conversation

Data Analysis

- Major leads and conversion ratio is high from country ‘India’

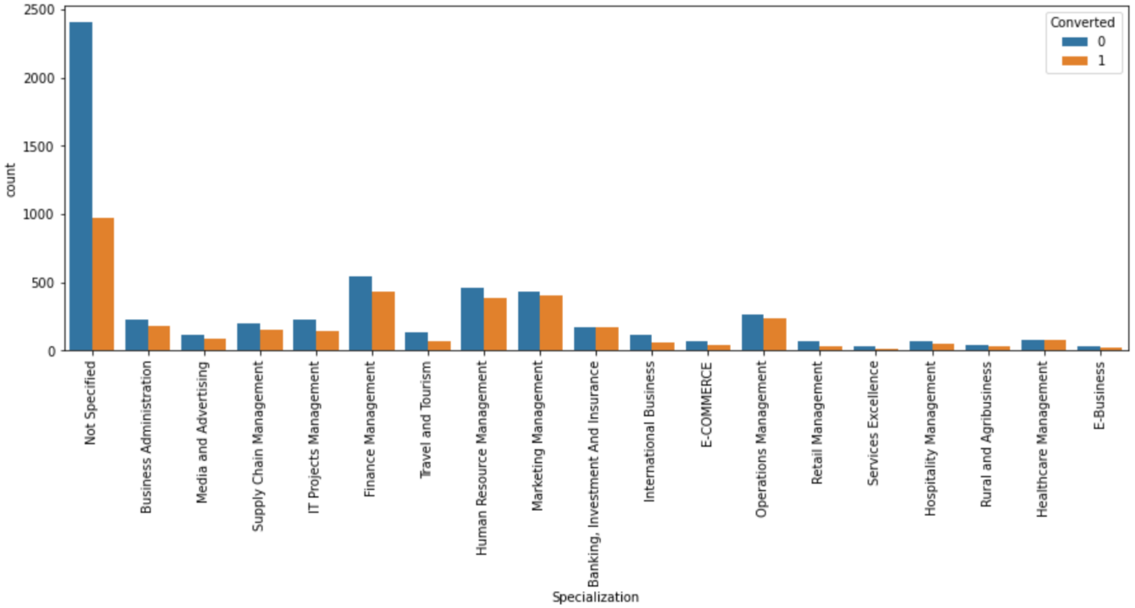


Data Analysis



- Major leads and conversion ration is high from city ‘Mumbai’

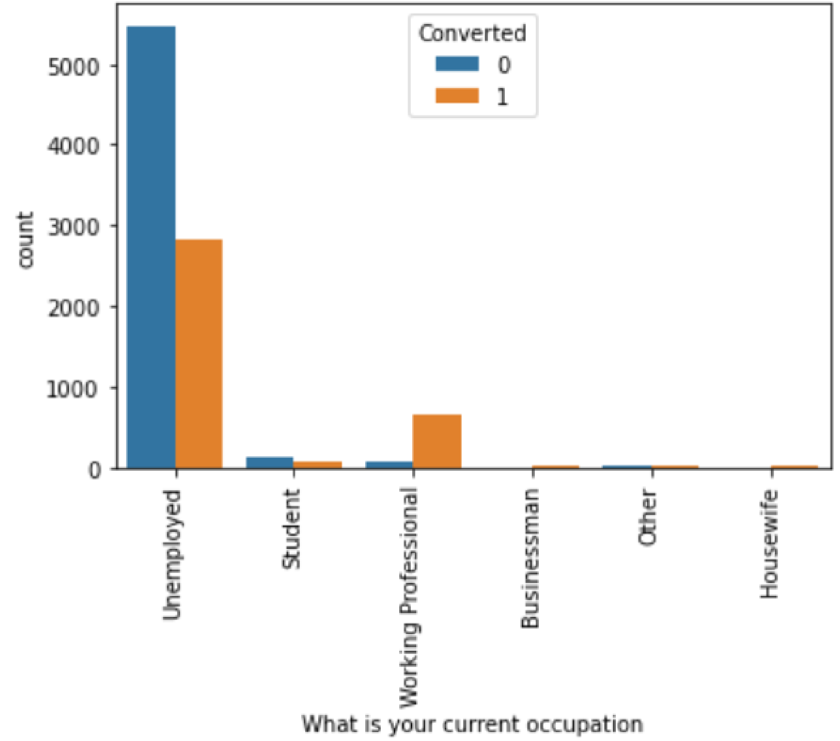
Data Analysis



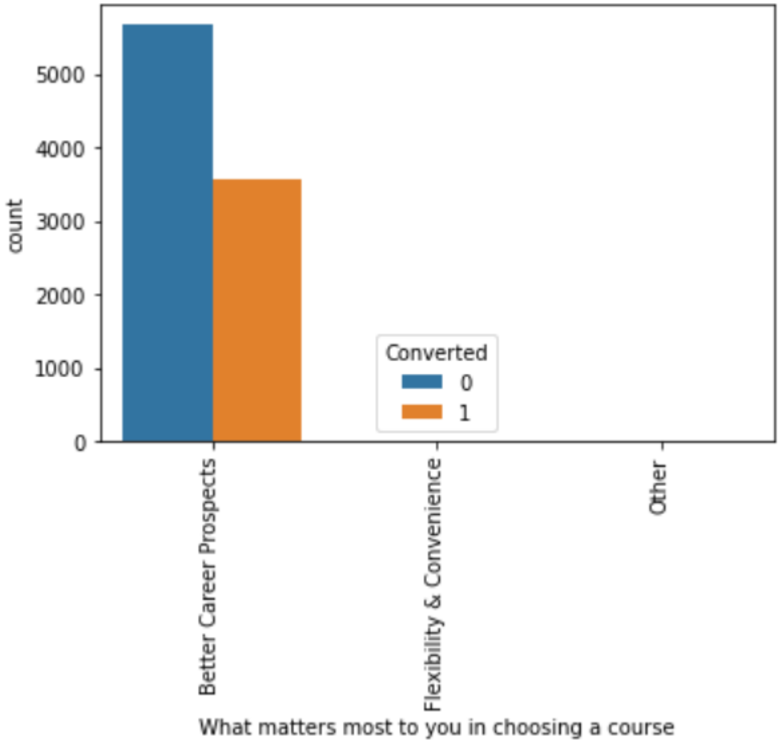
- Major leads and conversion ration is high from specialization category ‘Non-Specialized’

Data Analysis

- Major leads and conversion ration is high from specialization category ‘Un-employed’



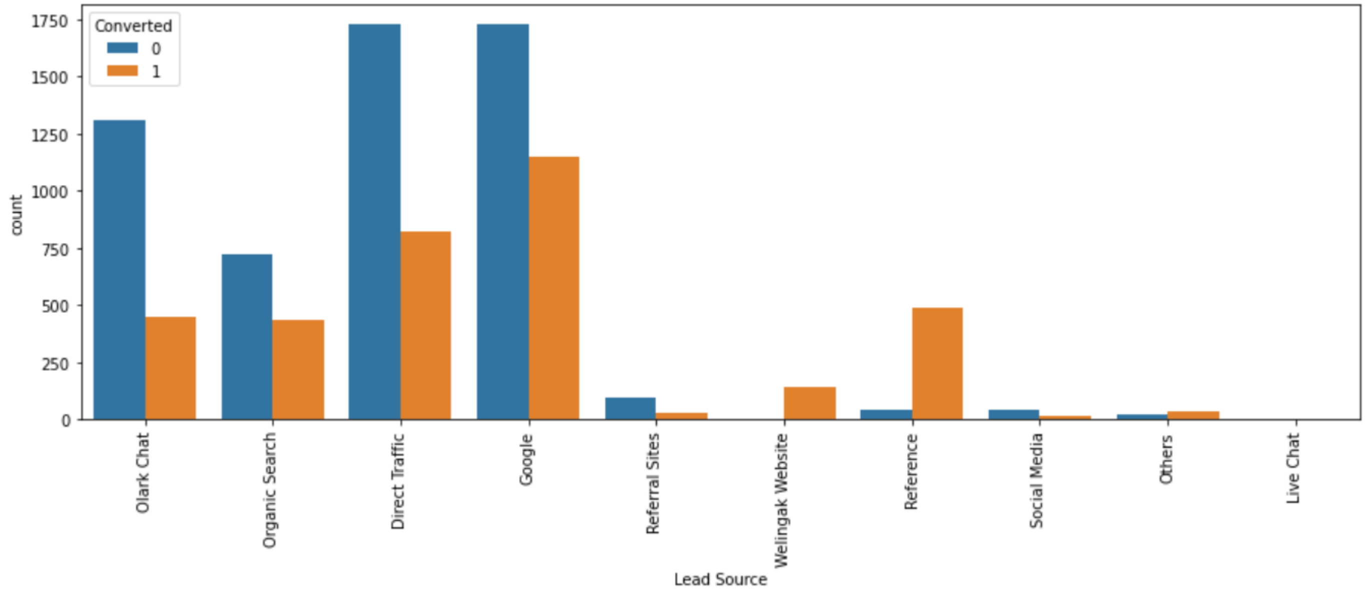
Data Analysis



- Majority converted lead choose course for ‘better career prospects’

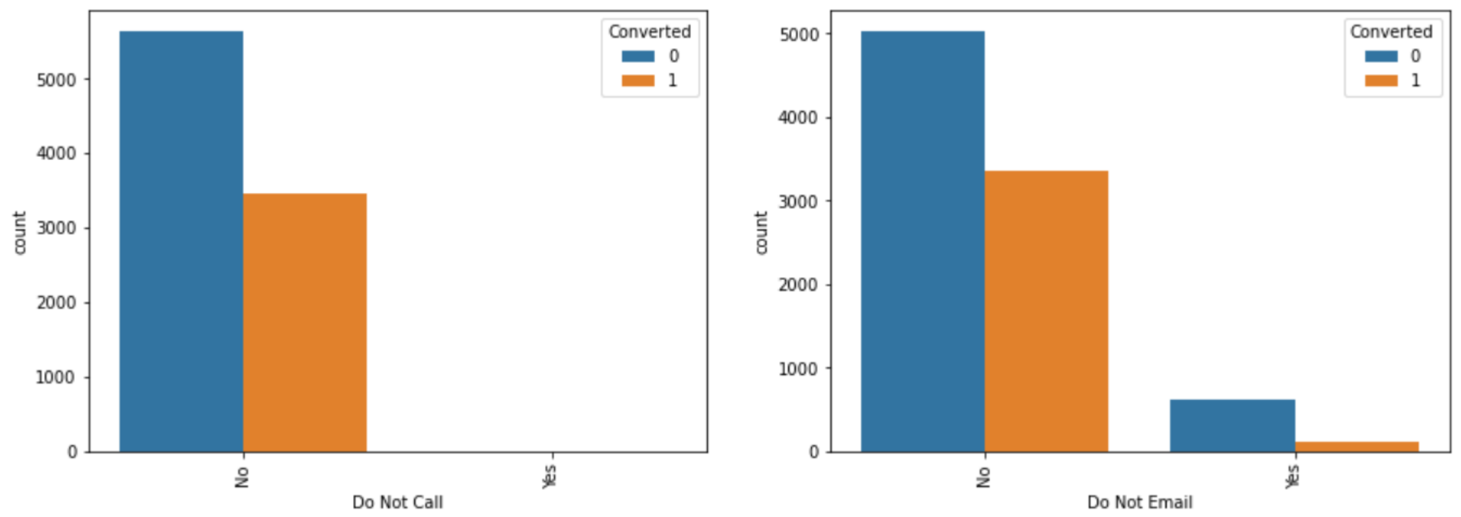
Data Analysis

- Lead Source Analysis for leads Vs conversion ration



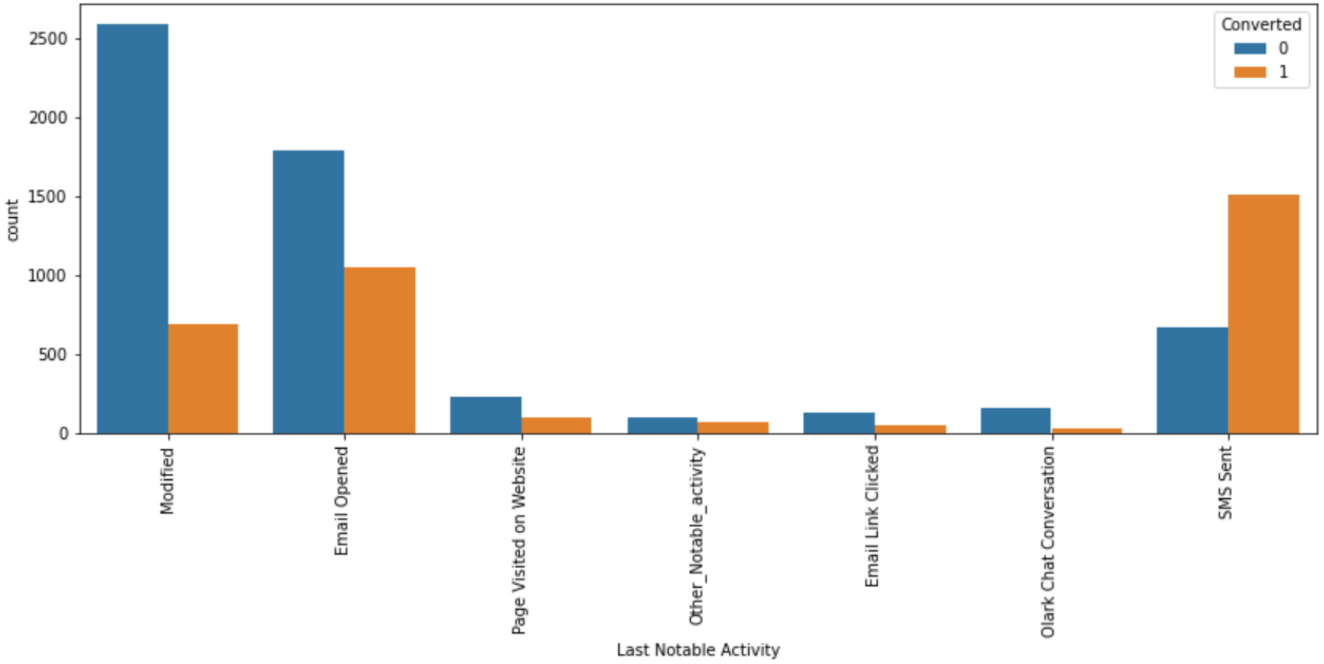
Data Analysis

- Conversion Analysis of ‘Do not call’ and ‘Do not Email’

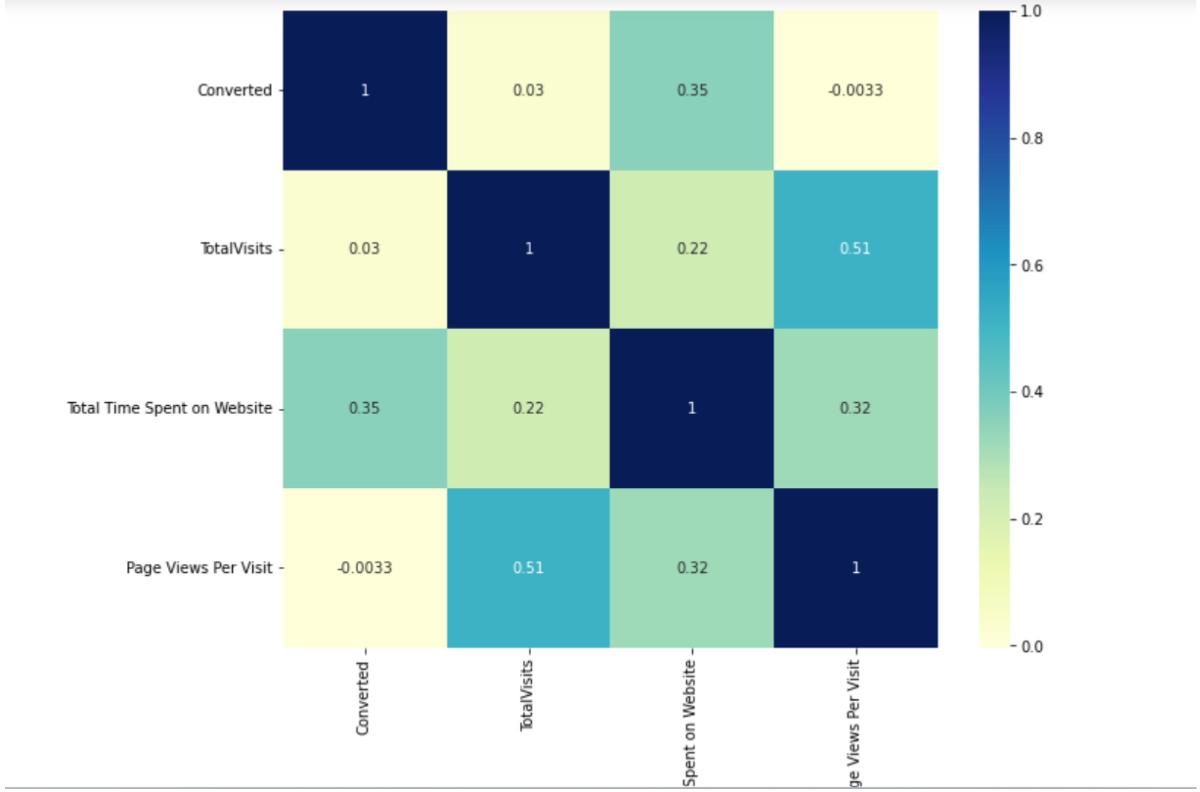


Data Analysis

- Conversion Analysis basis ‘Last Activity’



Data Analysis



- Correlation of numerical values

Data Analysis

- Here are top three variables in derived model which contributed most towards to probability of lead which getting converted.

1) **Total Time Spent on Website:**

- Positive contribution
- Higher the time spent on the website, higher the probability of the lead converting into a customer
- Sales team should focus on such leads

Data Analysis

2) Lead Source_Reference:

- Positive contribution
- If the source of the lead is a Reference, then there is a higher probability that the lead would convert, as the referrals not only provide for cashbacks but also assurances from current users and friends who will mostly be trusted - Sales team should focus on such leads

Data Analysis

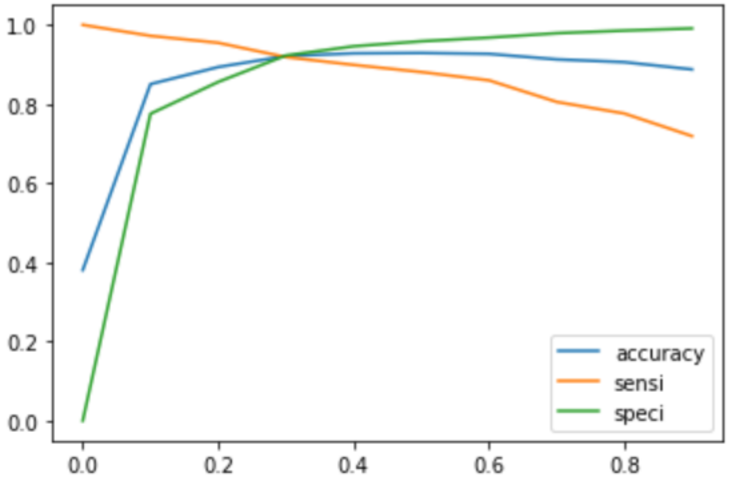
3) What is your current occupation_Student:

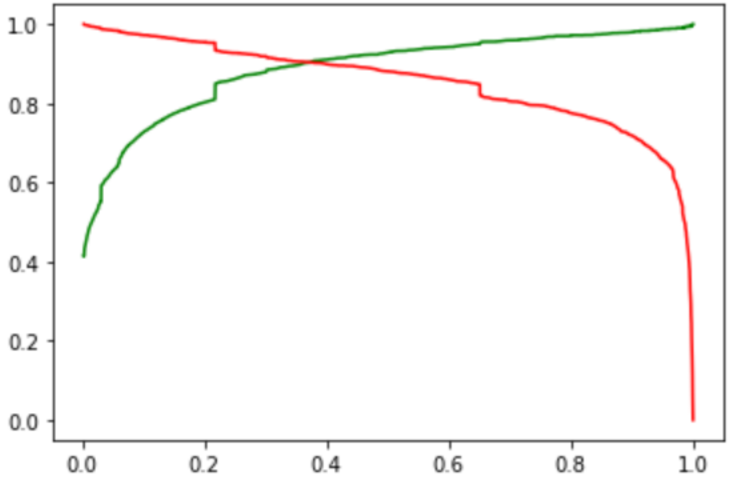
- Negative contribution
- If the lead is already a student, chances are they will not take up another course which is designed for working professionals.
- Sales team should not focus on such leads

Model Building

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

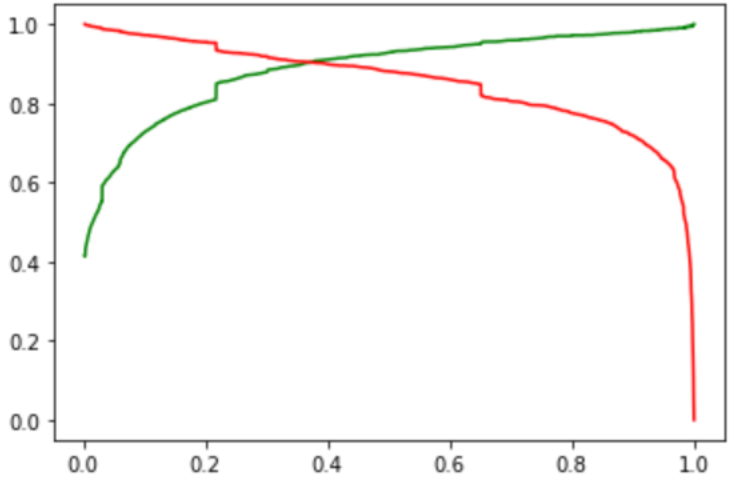
Model Evaluation (TRAIN)





- ACCURACY SENSITVITY AND SPECIFICTY
  - 92.29% Accuracy
  - 91.70 % Sensitivity
  - 92.66% Specificity
- PRECISION AND RECALL.
  - 73.4% Precision
  - 77.6% Recall

Model Evaluation (TEST)



- ACCURACY SENSITVITY AND SPECIFICTY
  - – 92.78% Accuracy
  - – 91.98 % Sensitivity
  - – 93.26% Specificity
- PRECISION AND RECALL.
  - – 74.4% Precision
  - – 75.5% Recall

  Test set threshold has been set as 0.41

Conclusion

- EDA :

- ➢ People spending higher than average time are promising leads, so targeting them and approaching them can be helpful in conversions
- ➢ SMS messages can have a high impact on lead conversion
- ➢ Landing page submissions can help find out more leads
- ➢ Marketing management, human resources management has high conversion rates. People from these specializations can be promising leads
- ➢ References and offers for referring a lead can be good source for higher conversions
- ➢ An alert messages or information has seen to have high lead conversion rate

Conclusion

- Logistic Regression Model:

- The model shows high close to 81% accuracy
- The threshold has been selected from Accuracy, Sensitivity, specificity measures and precision, recall curves.
- The model shows 76% sensitivity and 83% specificity
- The model finds correct promising leads and leads that have less chances of getting converted
- Overall this model proves to be accurate

Thank You!