Lead Scoring Case Study

By: Anuj Aggarwal

Problem Statement

- XEducation, an education company selling online courses, needs assistance in recognising and selecting promising leads that will most likely convert to paying customers.
- Our aim is to analyse and identify factors that help result in high customer conversion rate. This is done by assigning lead scores to each of the leads such that customers with high lead score indicate high chance of conversion

GOAL FOR THE CASE STUDY

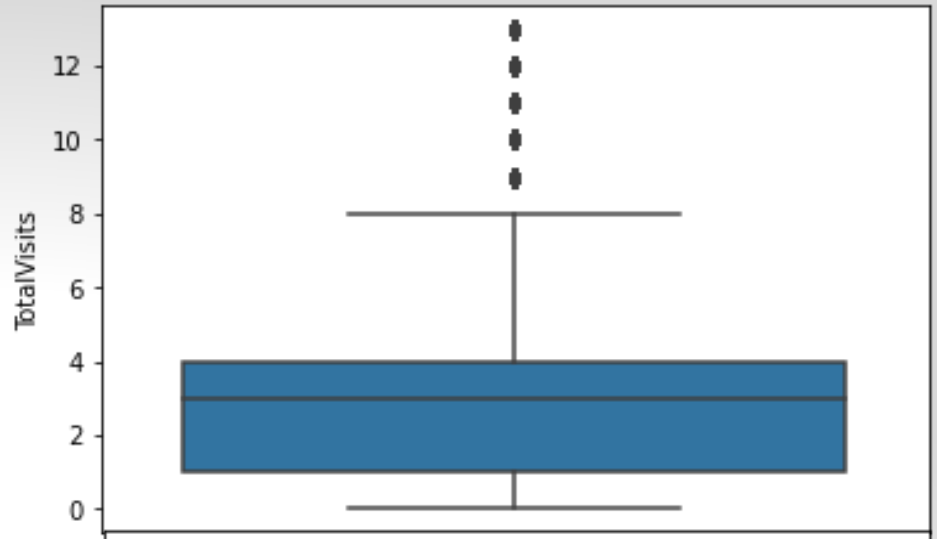
There are quite a few goals for this case study:

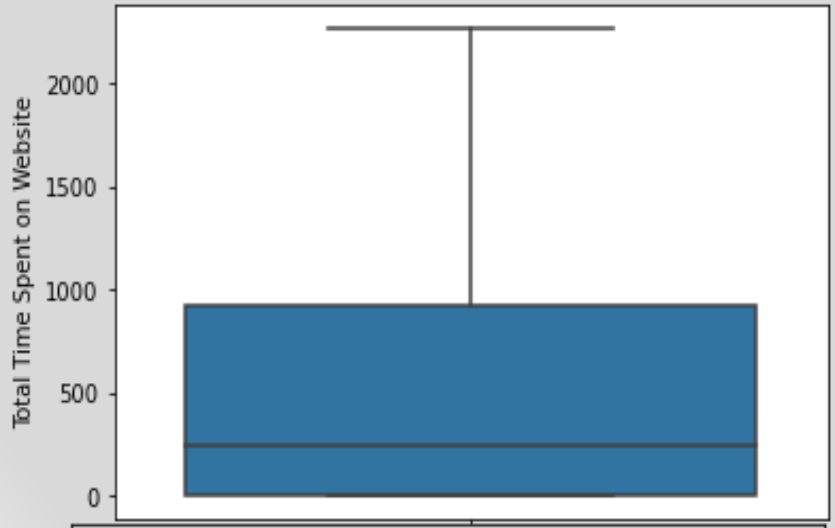
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

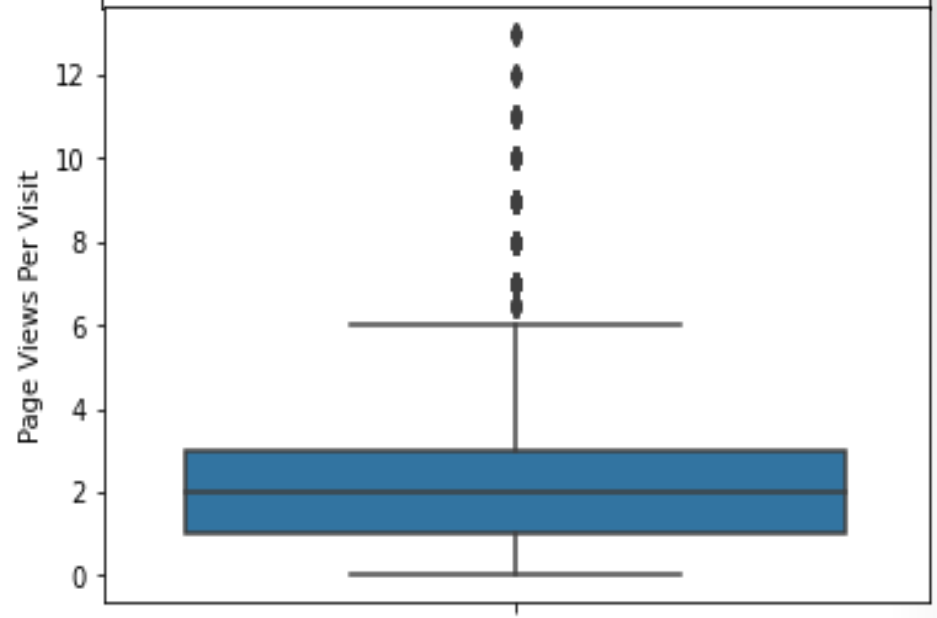
APPROACH

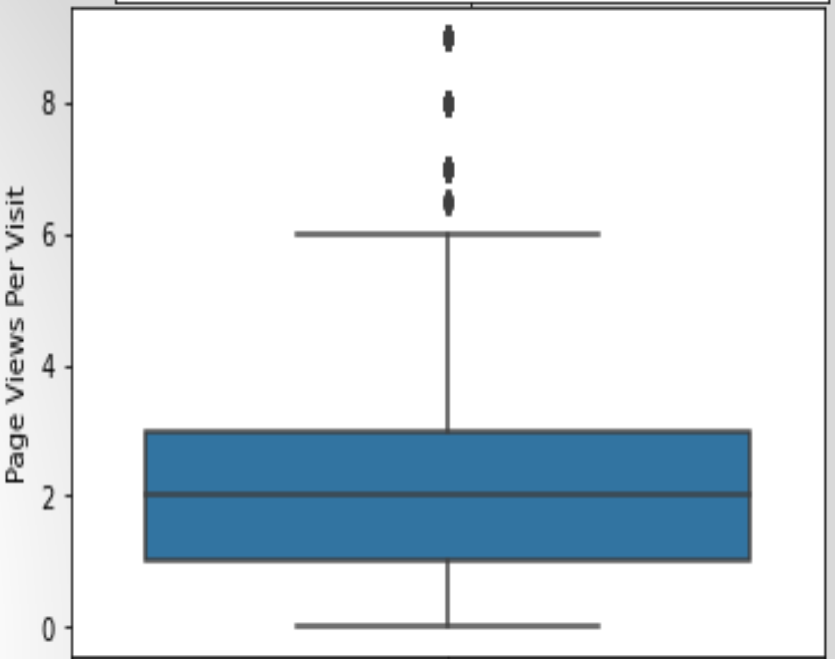
- Importing the Dataset
- Inspecting the DataFrame
- Missing Value Treatment
- Data Preparation
- Univariate Analysis
- Outlier Treatment
- Creating Dummy Variables
- Train - Test Split
- Feature Scaling
- Model Building
- Feature Selection Using RFE
- Plotting the ROC Curve
- Finding Optimal Cut-Off Points
- Precision and Recall
- Making Predictions on the Test Set
- Assigning the Lead Score to the Data

OUTLIERS

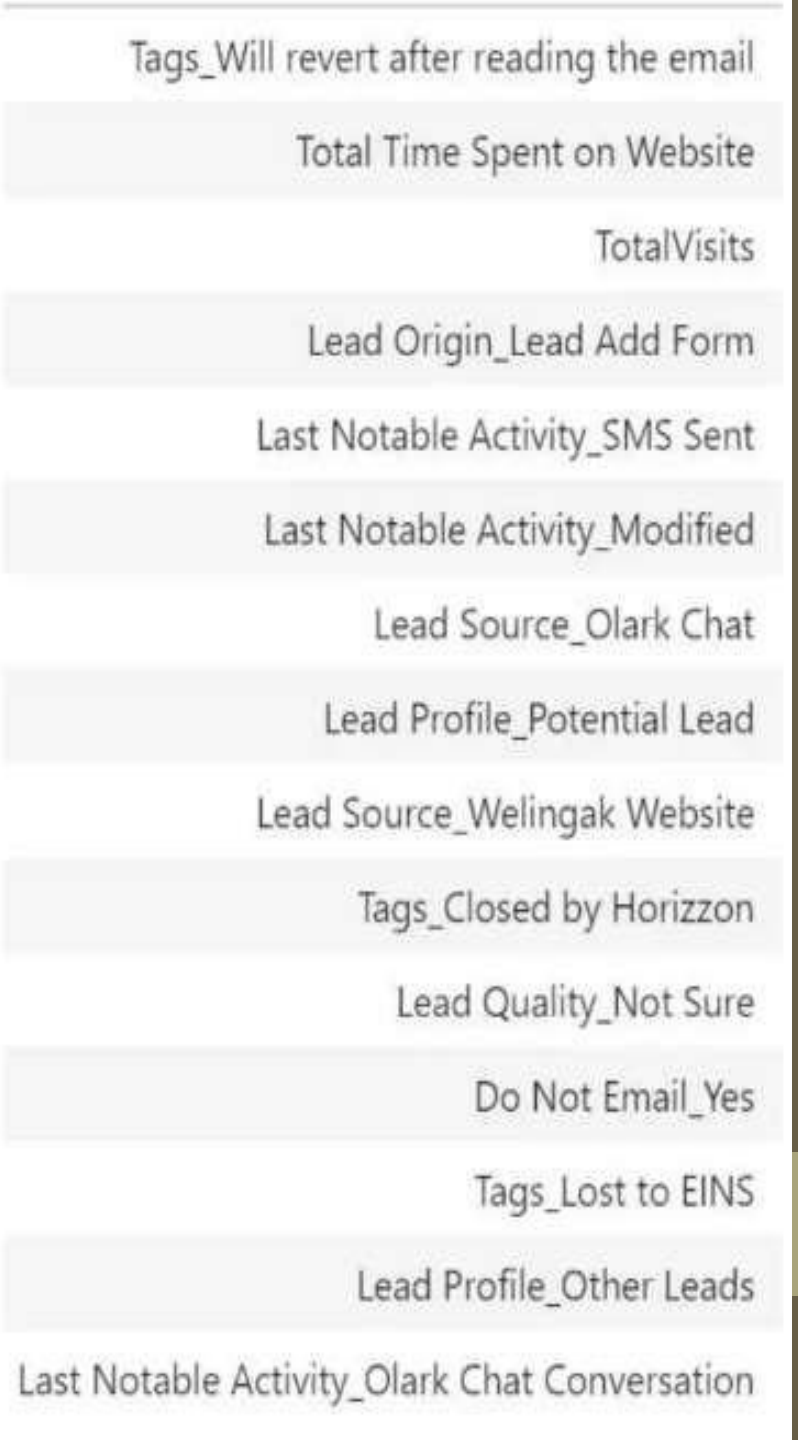

TotalVisits
0
2
4
6
8
10
12


Total Time Spent on Website
0
500
1000
1500
2000

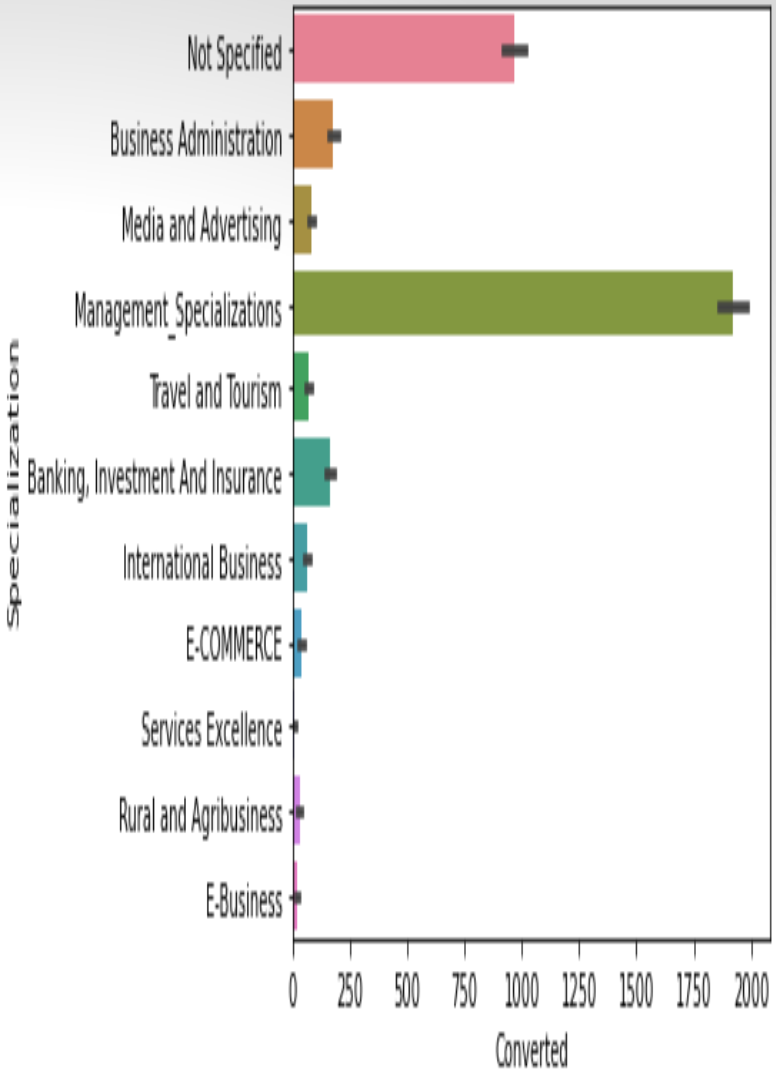

Page Views Per Visit
0
2
4
6
8
10
12

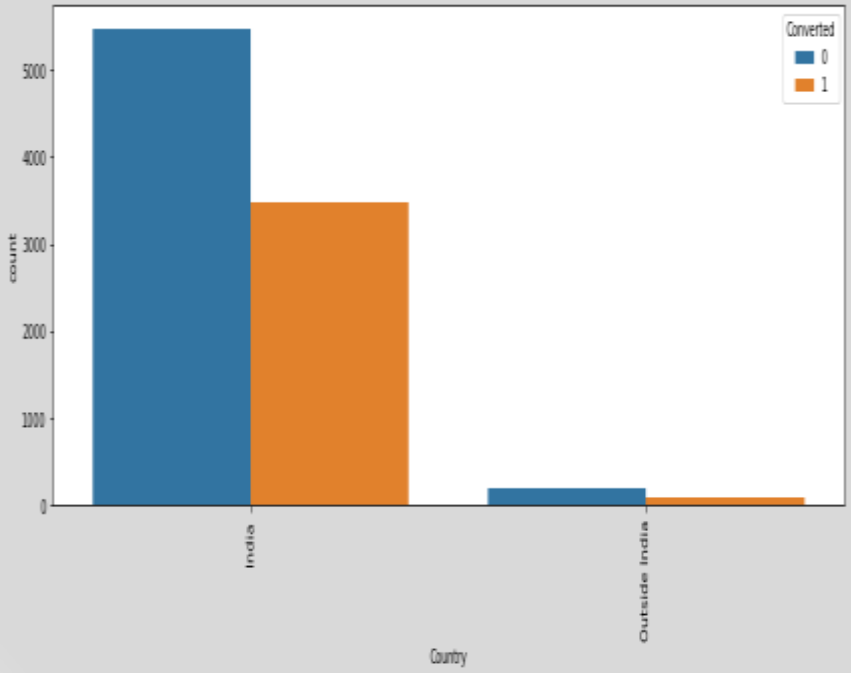

Page Views Per Visit
0
2
4
6
8

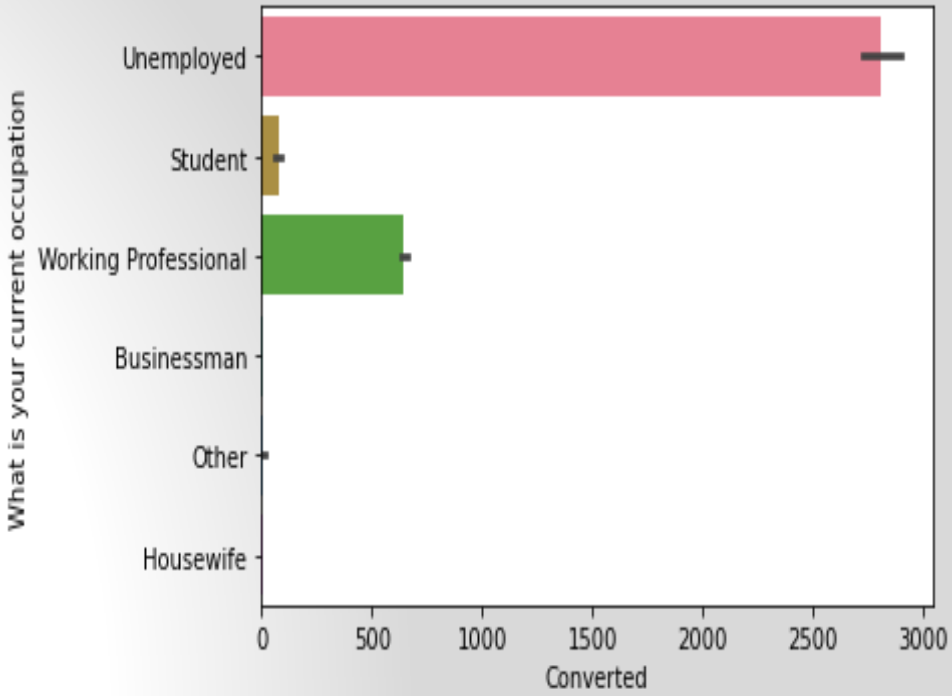
The Top factors that influence the conversion of leads.

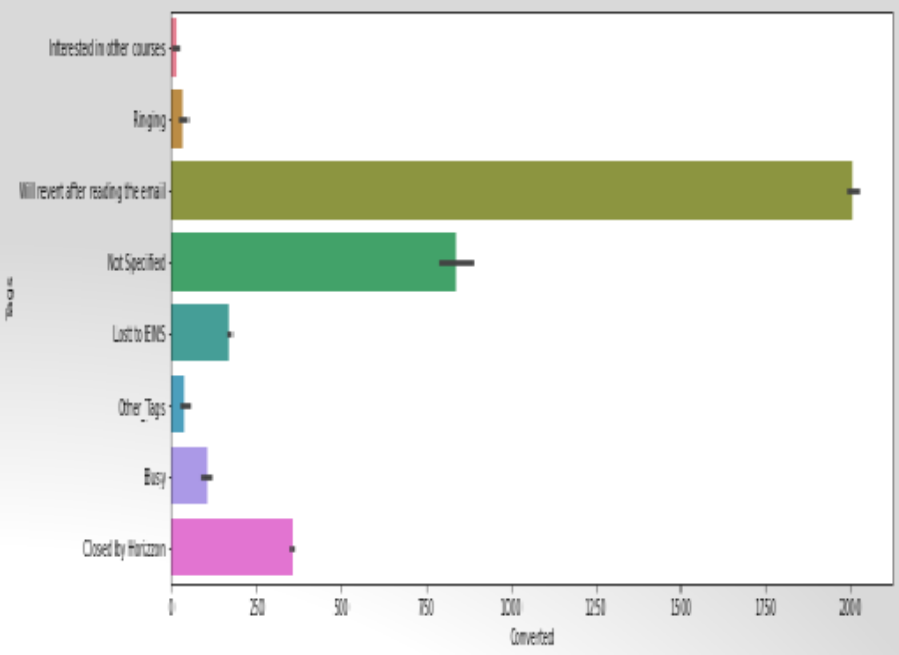


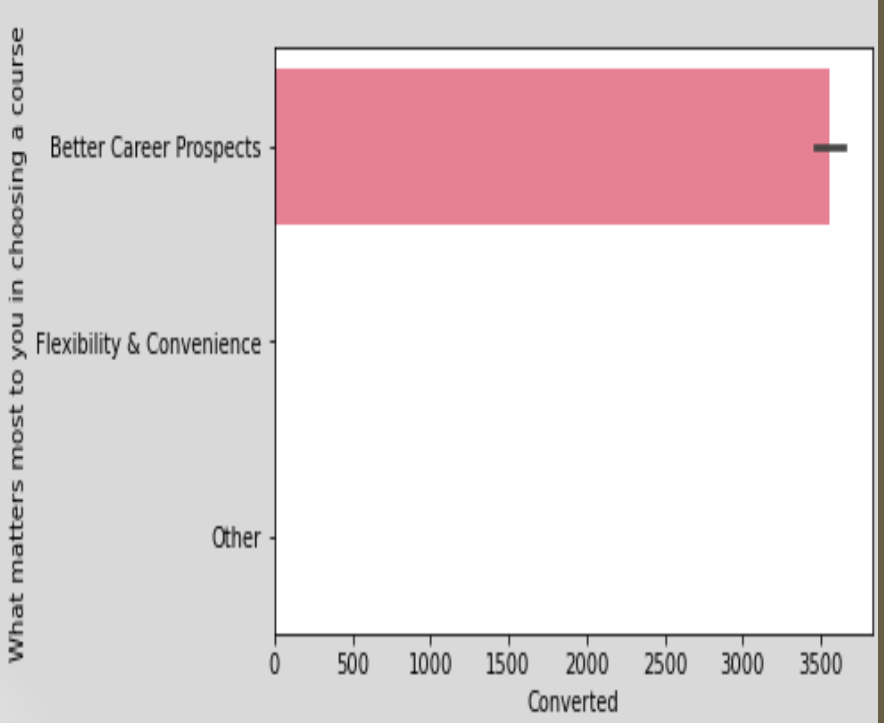
EDA

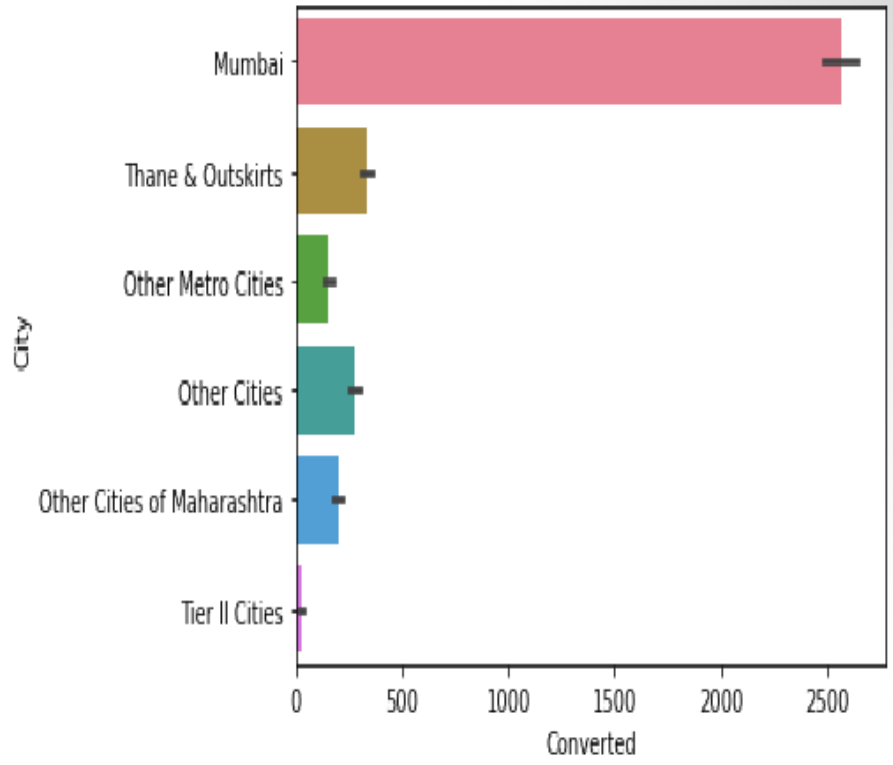

Specialization
Not Specified
Business Administration
Media and Advertising
Management_Specializations
Travel and Tourism
Banking, Investment And Insurance
International Business
E-COMMERCE
Services Excellence
Rural and Agribusiness
E-Business
0
250
500
750
1000
1250
1500
1750
2000
Converted

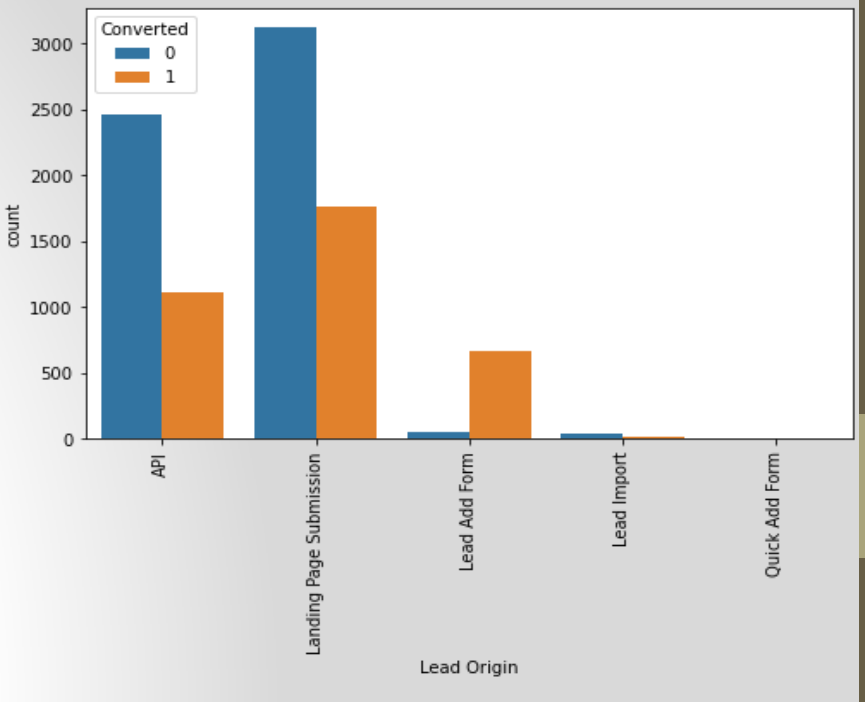

Converted
0
1
count
5000
4000
3000
2000
1000
0
India
Outside India
Country

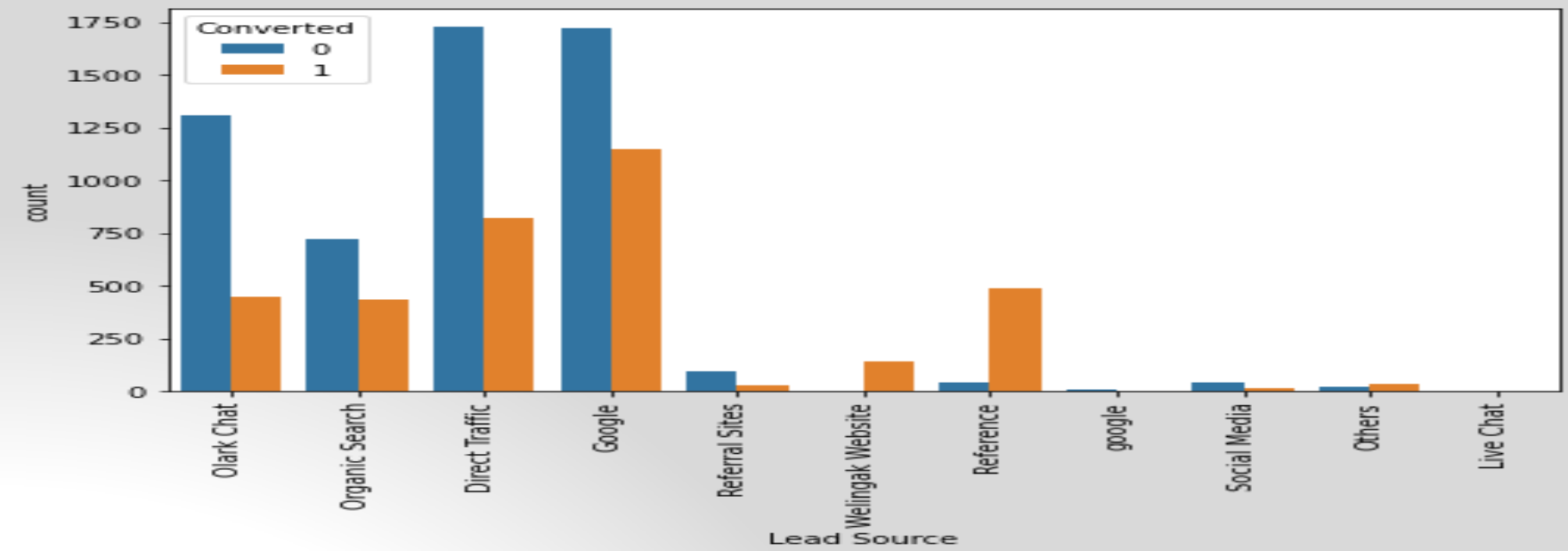

What is your current occupation
Unemployed
Student
Working Professional
Businessman
Other
Housewife
0
500
1000
1500
2000
2500
3000
Converted

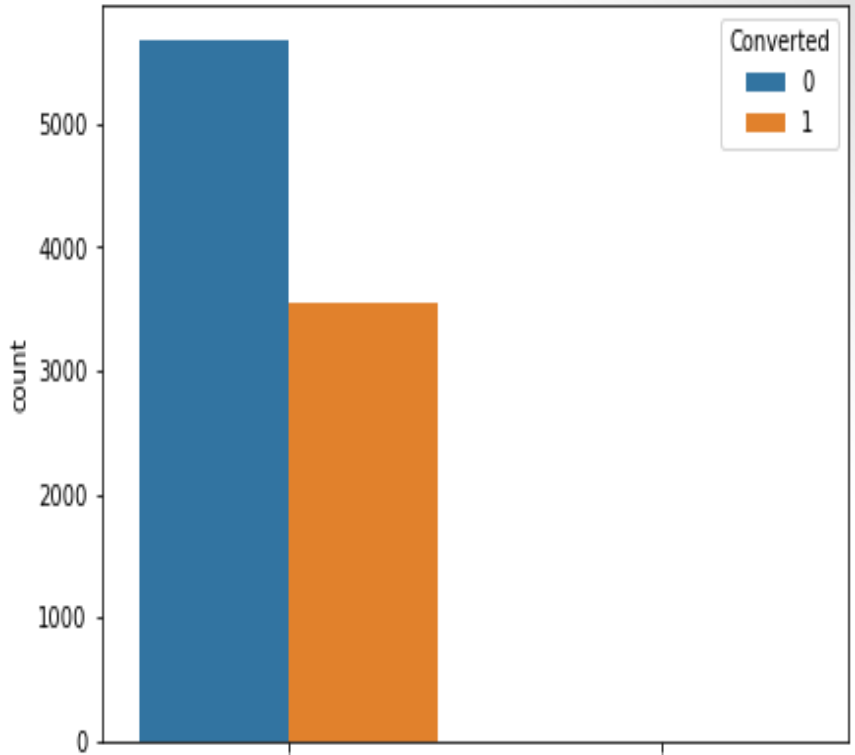

Interested in other courses
Ringing
Will revert after reading the email
Not Specified
Lost to EINS
Other_Tags
Busy
Closed by Horizzon
0
250
500
750
1000
1250
1500
1750
2000
Converted

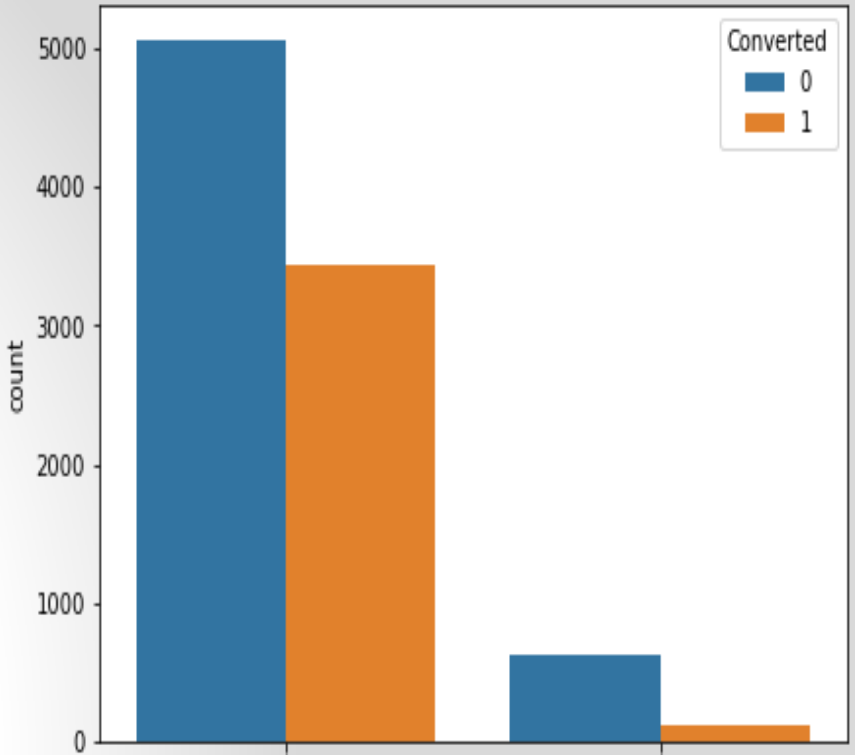

What matters most to you in choosing a course
Better Career Prospects
Flexibility & Convenience
Other
0
500
1000
1500
2000
2500
3000
3500
Converted


City
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
0
500
1000
1500
2000
2500
Converted


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Social Media
Others
Live Chat
Lead Source

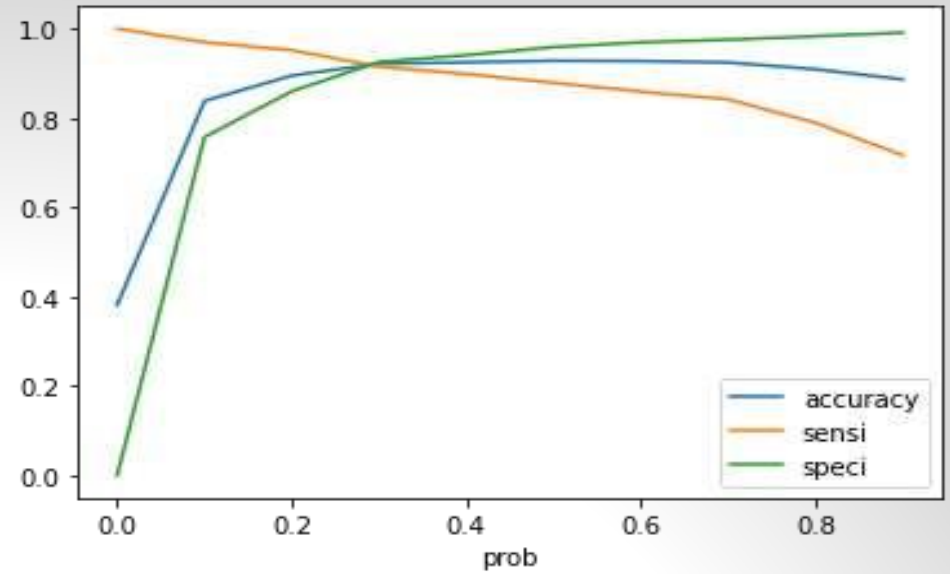

Converted
0
1
count
5000
4000
3000
2000
1000
0

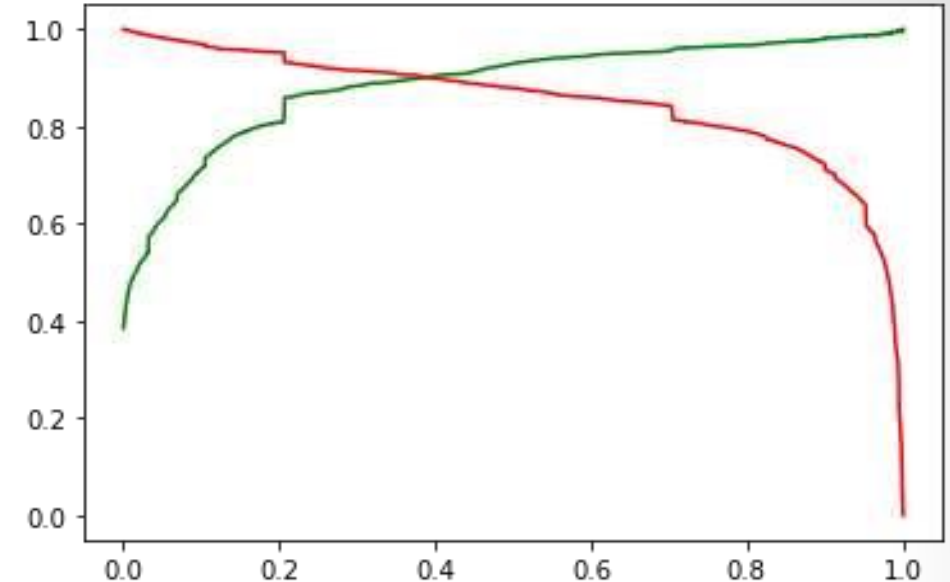

Converted
0
1
count
5000
4000
3000
2000
1000
0

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 p-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)





Final Observation:

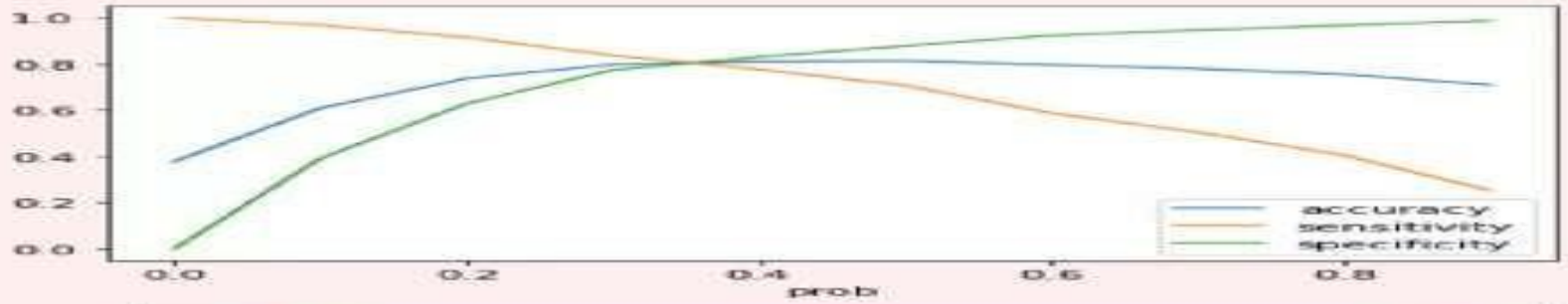
Train Data:

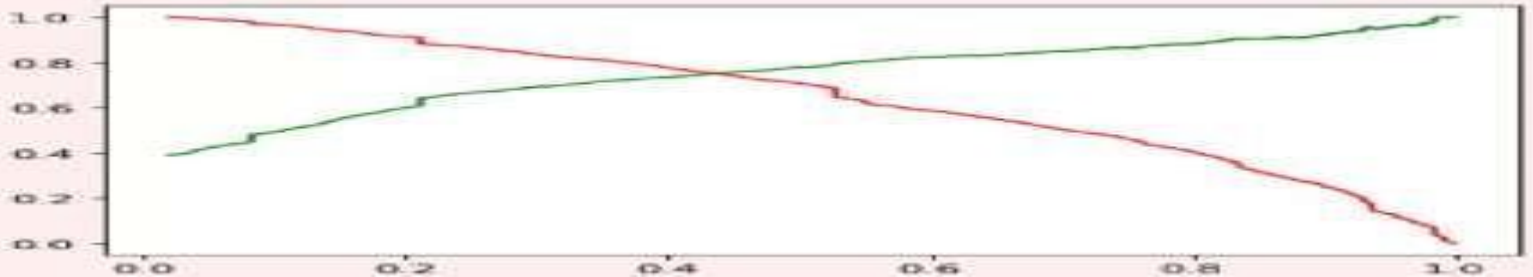
Accuracy : 92.14%

Sensitivity : 91.49%

Specificity : 92.54%

Model Evaluation(Test)





Test Data:

Accuracy : 92.57%

Sensitivity : 91.18%

Specificity : 93.46%

The model seems to be performing well. Can be recommend this model in making good calls based on this model

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

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