Lead Case Study

1. Ashish Gupta
2. Gaurav Sehgal
3. Nithin Skanda

The problem Statement

Problem statement

- An education company X sells online courses to industry professionals.
- X gets lods of leads, but lead conversion rate is poor. About 30 out 100 leads in a day are only converted.
- To make this process more efficient, X wants to identify the most potential leads ‘Hot Leads’.
- The lead conversion rate would go up as sales team will now be focusing more on this potential leads rather than making calls to everyone if they successfully identify this set of Hot Leads.

Business Objective

- X wants to know most promising leads.
- For that they want to build a model which identifies the hot leads.
- Deployment of that model for the future use.

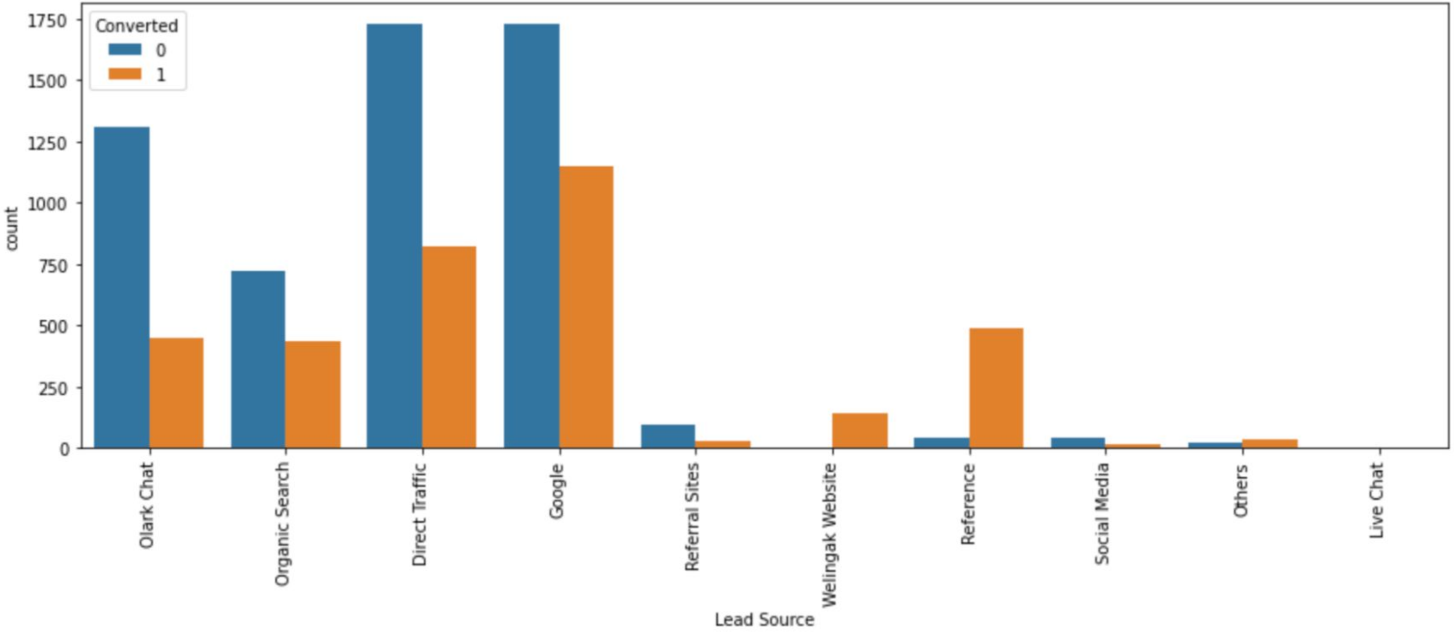
Solution Methodology

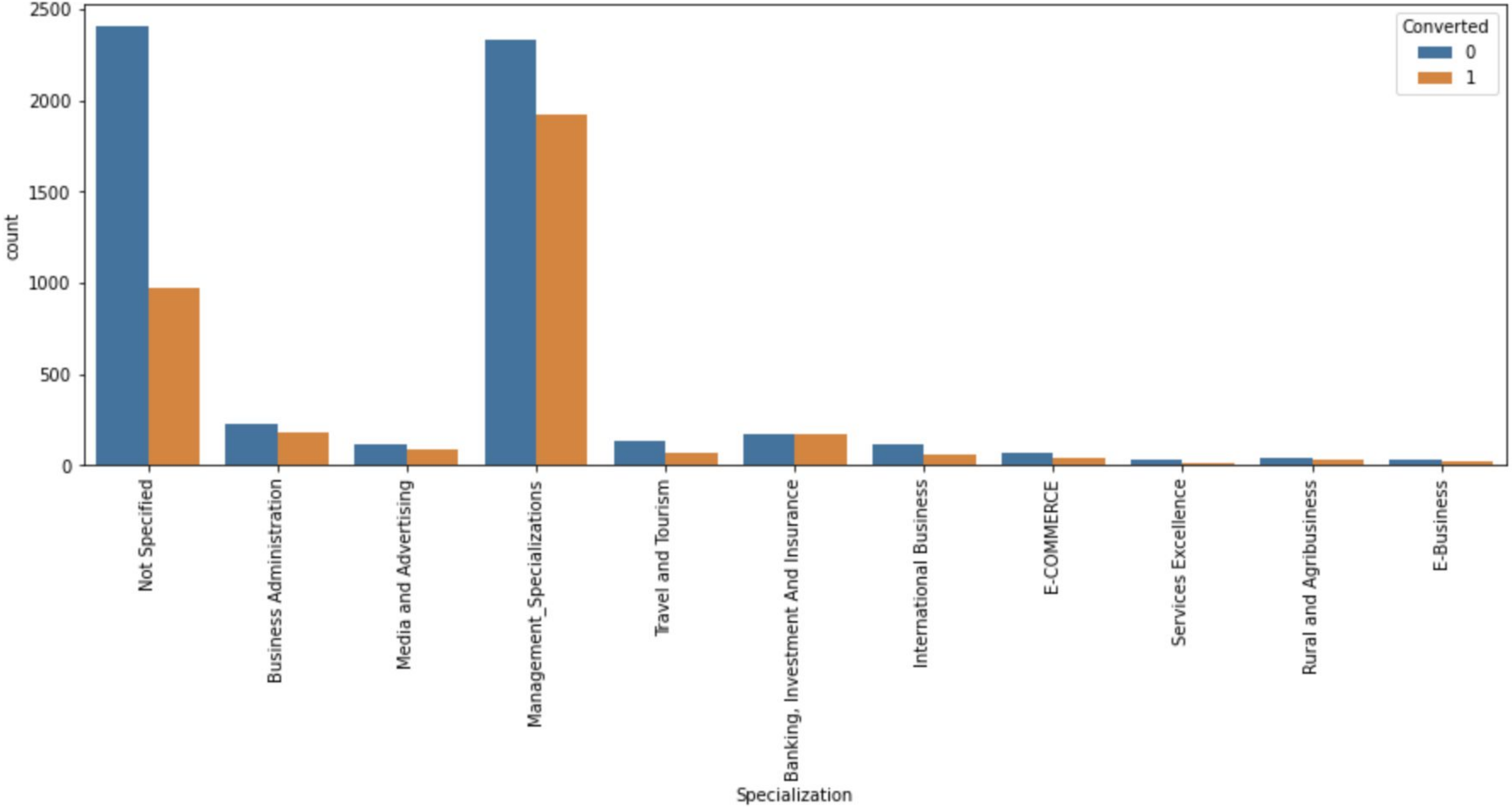
- Data Cleaning and Treatment.
  1. Check and handle duplicate data.
  2. Handle NaN values.
  3. Drop columns if it contains large amount of missing values.
- EDA
  1. Univariate Analysis
  2. Bivariate Analysis
- Dummy Variables and encoding of the data.
- Logistic regression used for the model making and prediction.
- Validation of the model.
- Model Presentation.
- Conclusion and Recommendations.

EDA

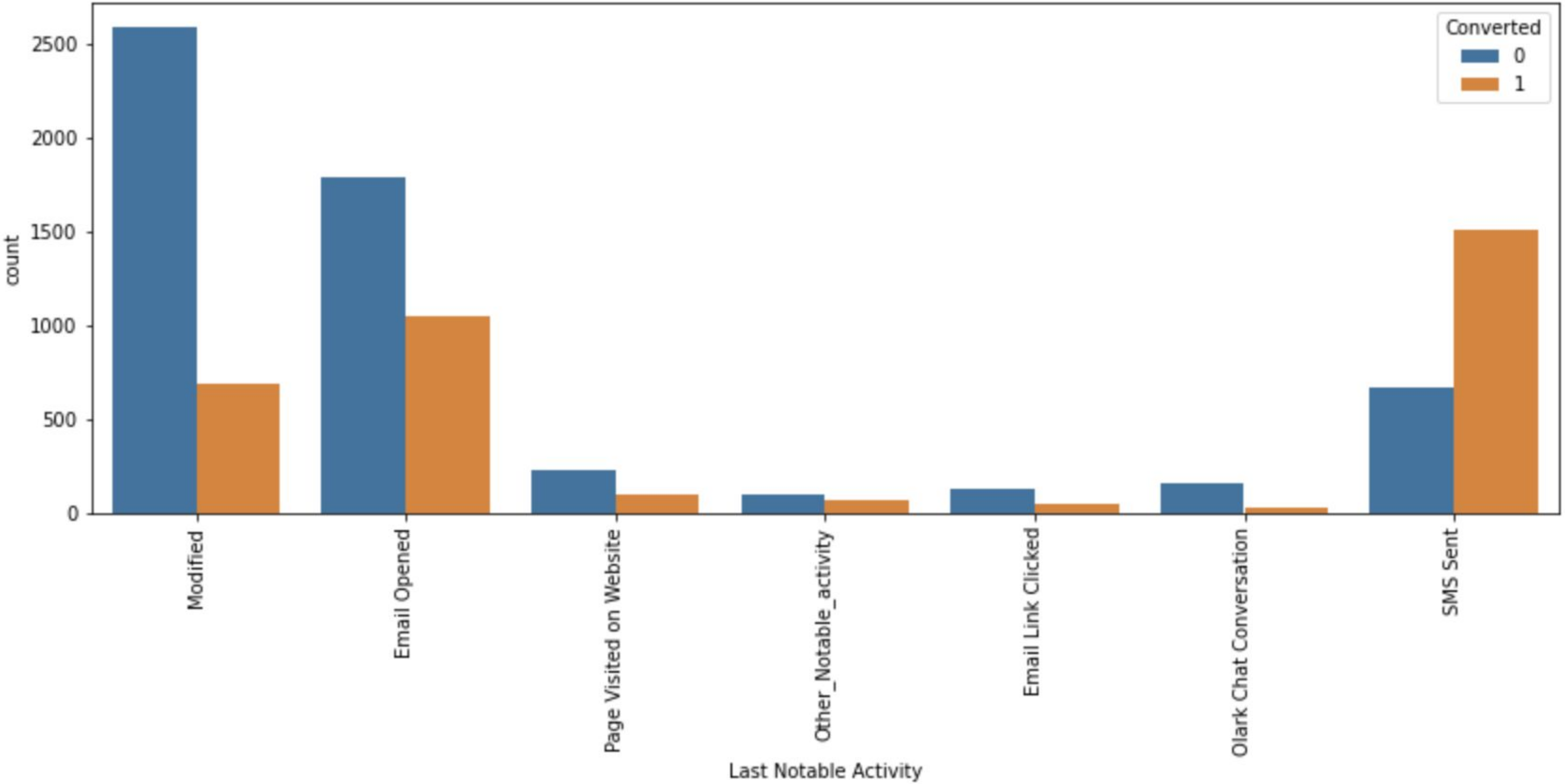
- Total Number of Rows are 37, Columns are 9240.
- Single value features like ‘Magazine’, ‘Receive More Updates About Our Courses’, ‘Update me on Supply’.
- ‘Prospect ID’ and ‘Lead Number’ are removed since it’s not necessary for the analysis.
- Dropped the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

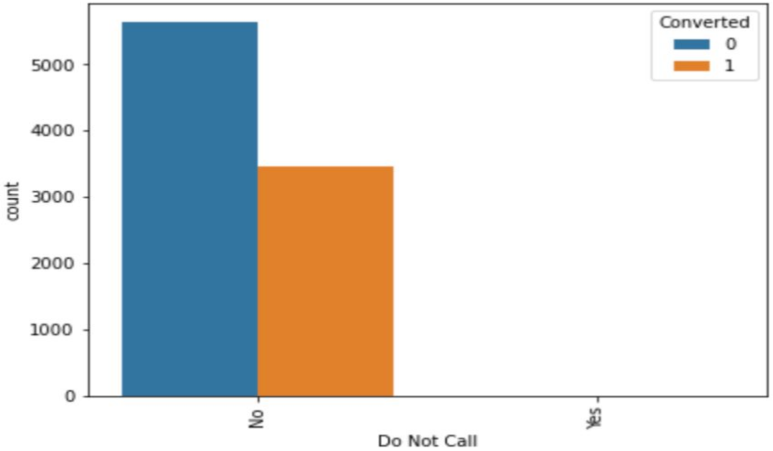
Categorical Attributes Analysis

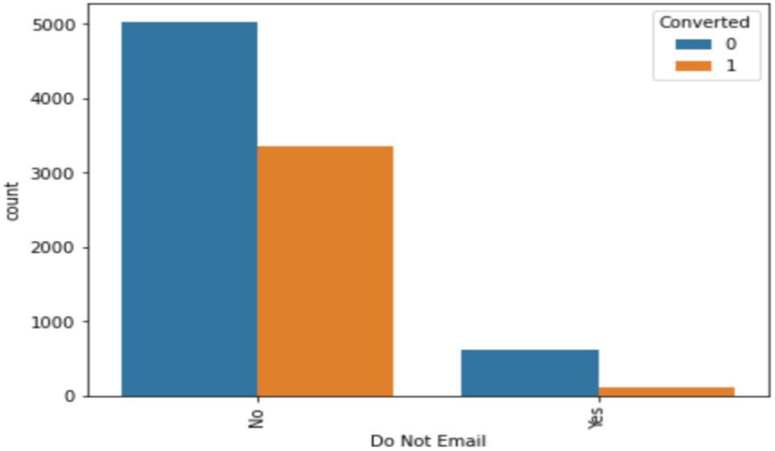


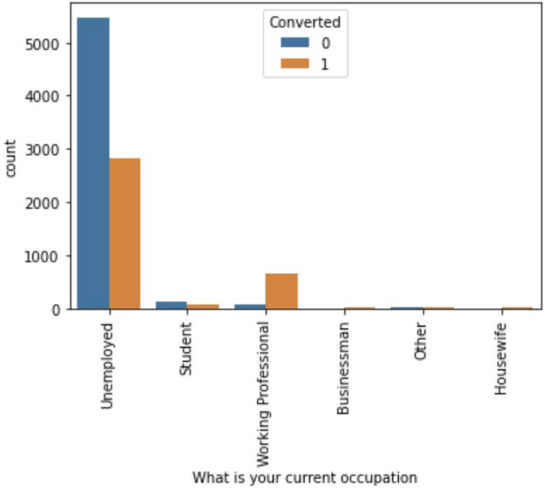


2500
2000
1500
1000
500
0
count
Converted
0
1
Not Specified
Business Administration
Media and Advertising
Management_Specializations
Travel and Tourism
Banking, Investment And Insurance
International Business
E-COMMERCE
Services Excellence
Rural and Agribusiness
E-Business
Specialization

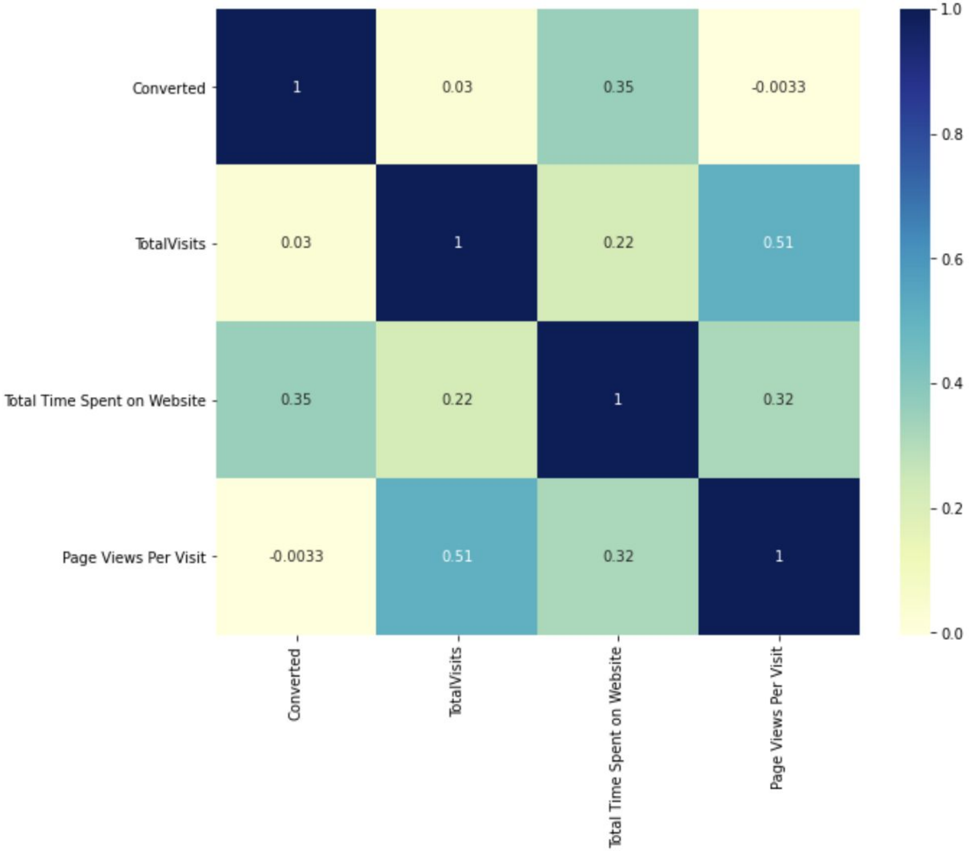

Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Other_Notable_activity
Email Link Clicked
Olark Chat Conversation
SMS Sent
Last Notable Activity

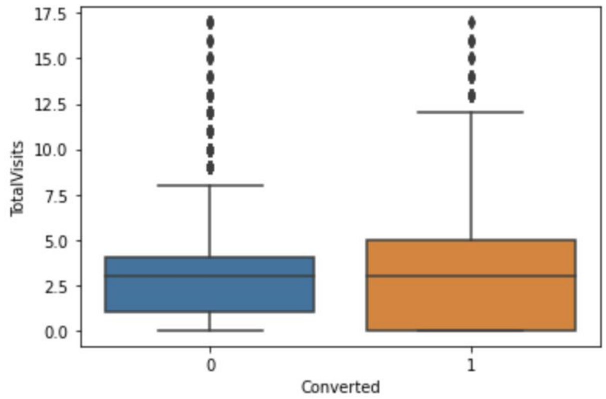

Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
Do Not Call

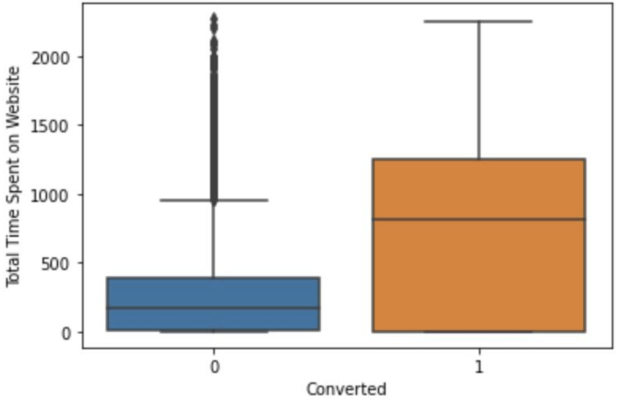

Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
Do Not Email

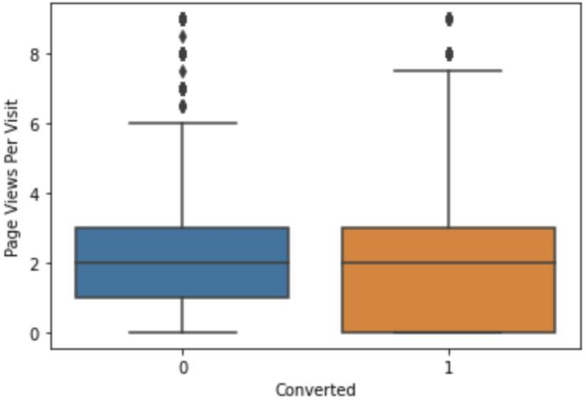

Converted
0
1
count
0
1000
2000
3000
4000
5000
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

Numerical Attributes Analysis




TotalVisits
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
1
Converted

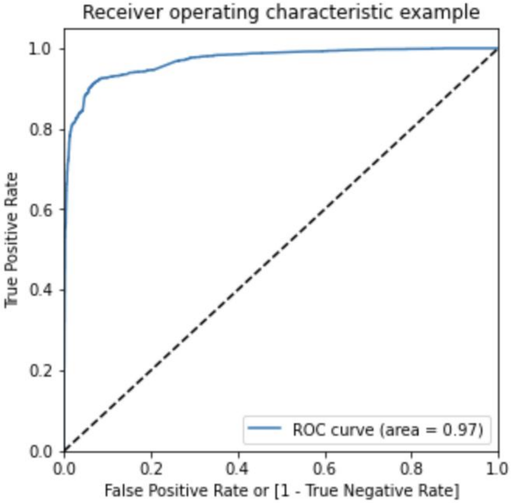

Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted

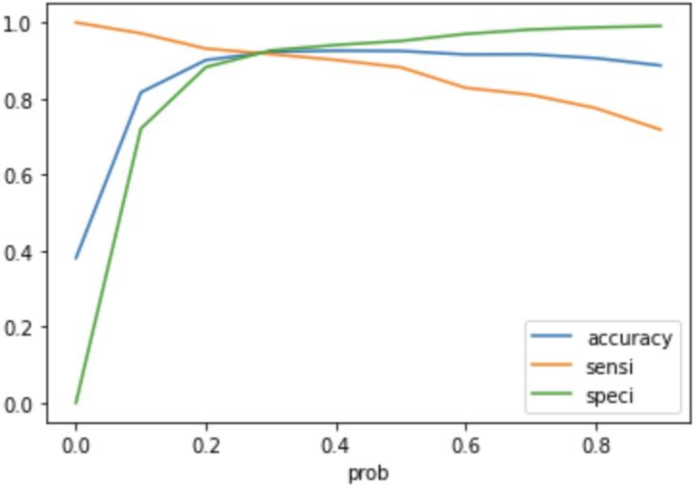

Page Views Per Visit
8
6
4
2
0
0
1
Converted

Model Building

- Splitting the Data into Training and Testing sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Running RFE with 15 variables as output.
- Building Model by removing the variables whole p- value is greater than 0.05 and vif value with the highest value.
- Overall accuracy 92.29%

ROC Curve





- Finding optimal cut off point.
- Probability where we get the balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.3.

Conclusion

- Train Data:
    - Accuracy: 92.29%
    - Sensitivity: 91.70%
    - Specificity: 92.66%
- Test Data:
    - Accuracy: 92.78%
    - Sensitivity: 91.98%
    - Specificity: 92.26%

- It was found out that the variables that mattered the most in the potential buyers are:
    - Total time spent on the website.
    - When lead sources are:
        - Google
        - Direct Traffic
    - When the lead origin is lead add format
    - When their current occupation is as a working professional.
- Keeping these in mind, X can have a very high chance to get almost all the potential buyers to change their mind and buy their courses.