Lead Scoring Assignment By:

Partha Sarathi Sahoo
Kaustav Bhattacharjee

Problem Statement and Business Goal

Problem Statement : X Education sells online courses to industry professionals. 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. 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 lowly 30%.

Business Goal: X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

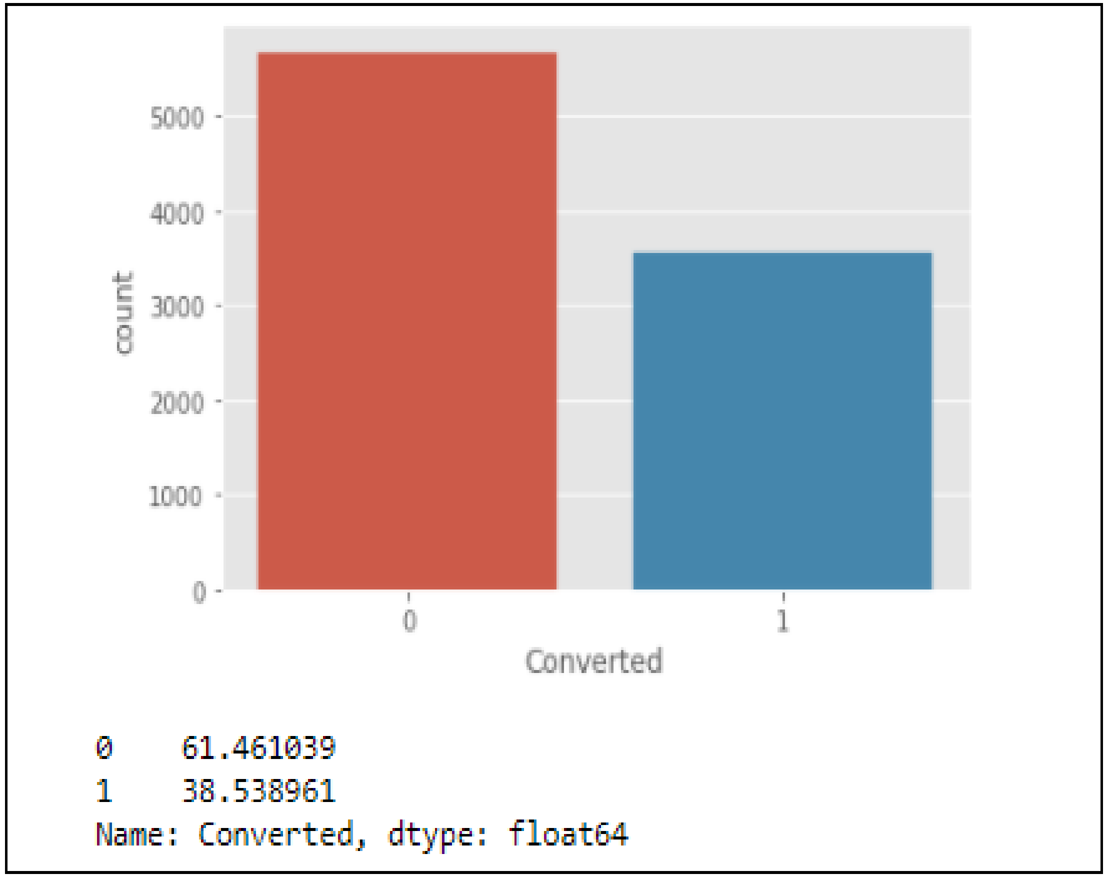
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Approach

- Import the data and difference libraries
- Perform EDA on the dataset and clean up the dataset
- Split the data into test and train dataset
- Create dummy columns for categorical variables
- Scale the train dataset and perform RFE to get top 15 relevant variables
- Building a logistic Regression model and calculate Lead Score (0 to 1).
- Find the optimal cut-off for values
- Evaluating the model by using different metrics – Specificity, Sensitivity and Accuracy.
- Applying the best model on scaled test data based on the Sensitivity and Specificity Metrics.

Current Insights From Data - I

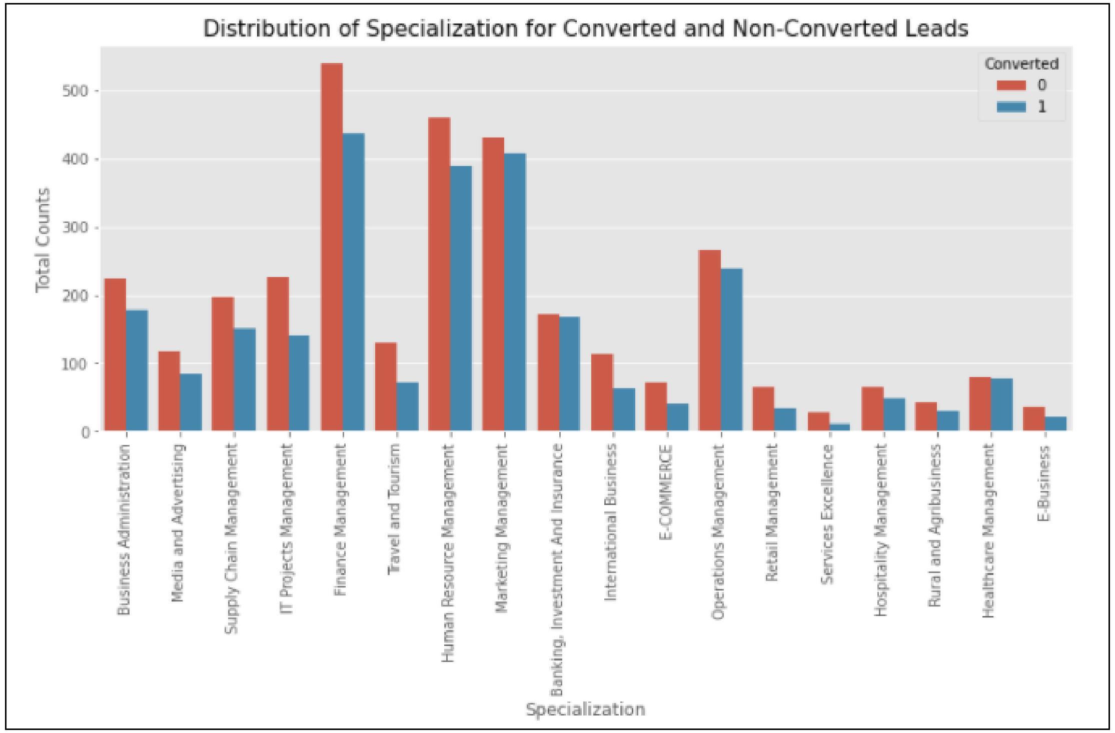
We have 38.5% of overall conversion rate which is very low. Our goal is to improve this number.



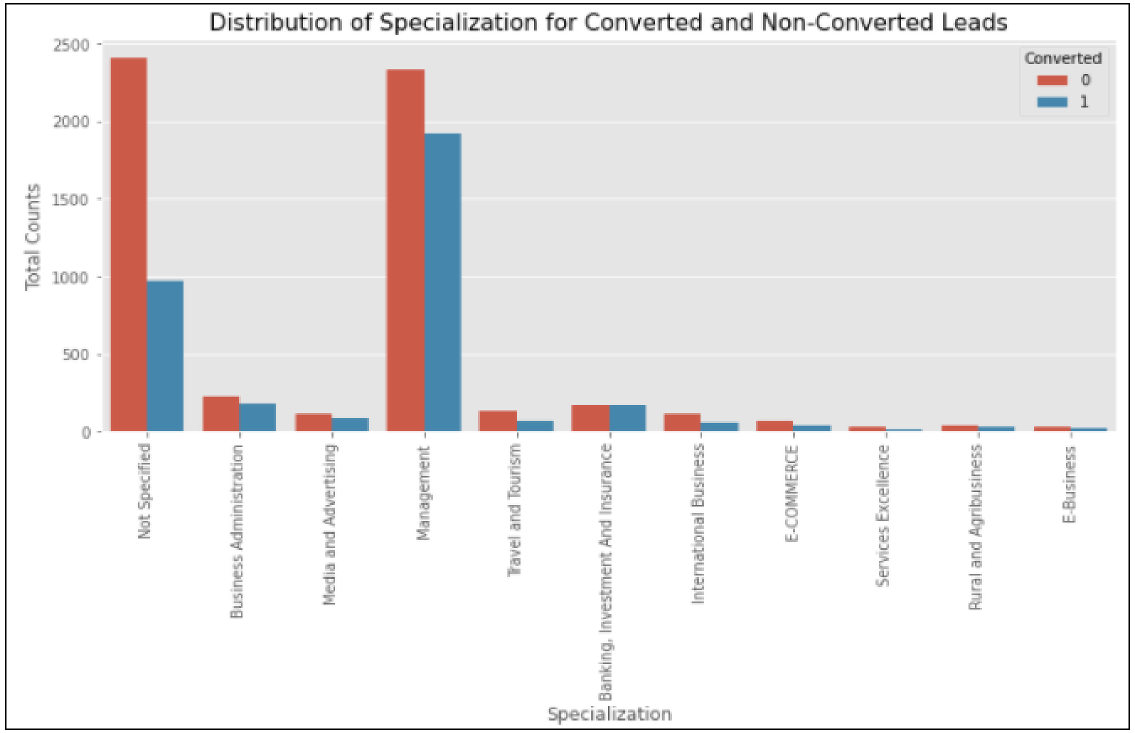
```
0    61.461039
1    38.538961
Name: Converted, dtype: float64
```

Current Insights From Data- II

We can see that 'Finance Management', 'HRM' and 'Marketing Management' has a very high contribution to the over all converted leads. But all the management specializations seems to have similar conversion rate. So we decided to club them together.



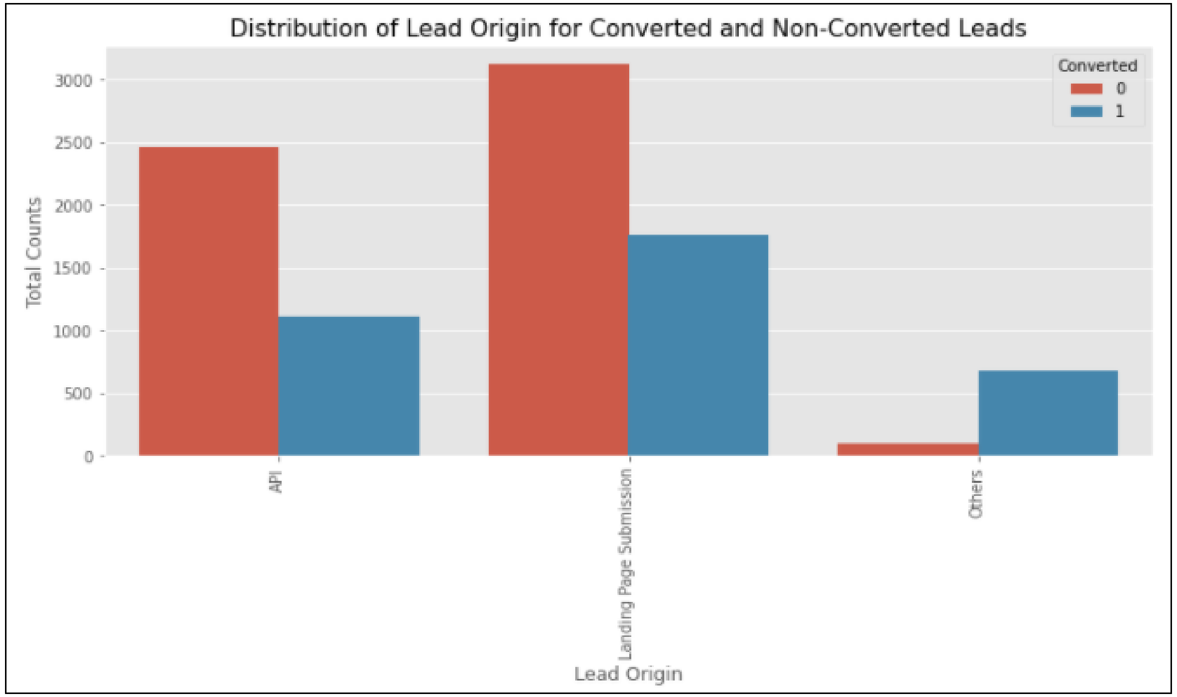
Trends before grouping management specializations grouping



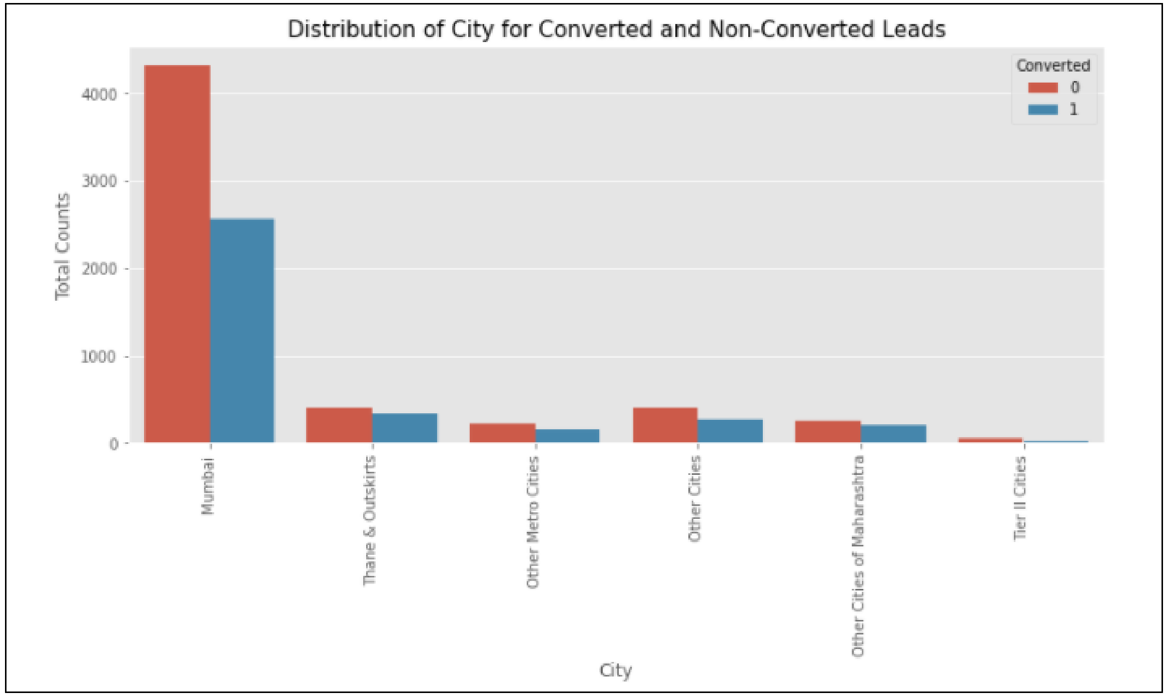
Trends after grouping management specializations grouping

Current Insights From Data - III

Majority of the lead originated from landing page submission. However we see that the other section, albeit small in number has a very good conversion rate.



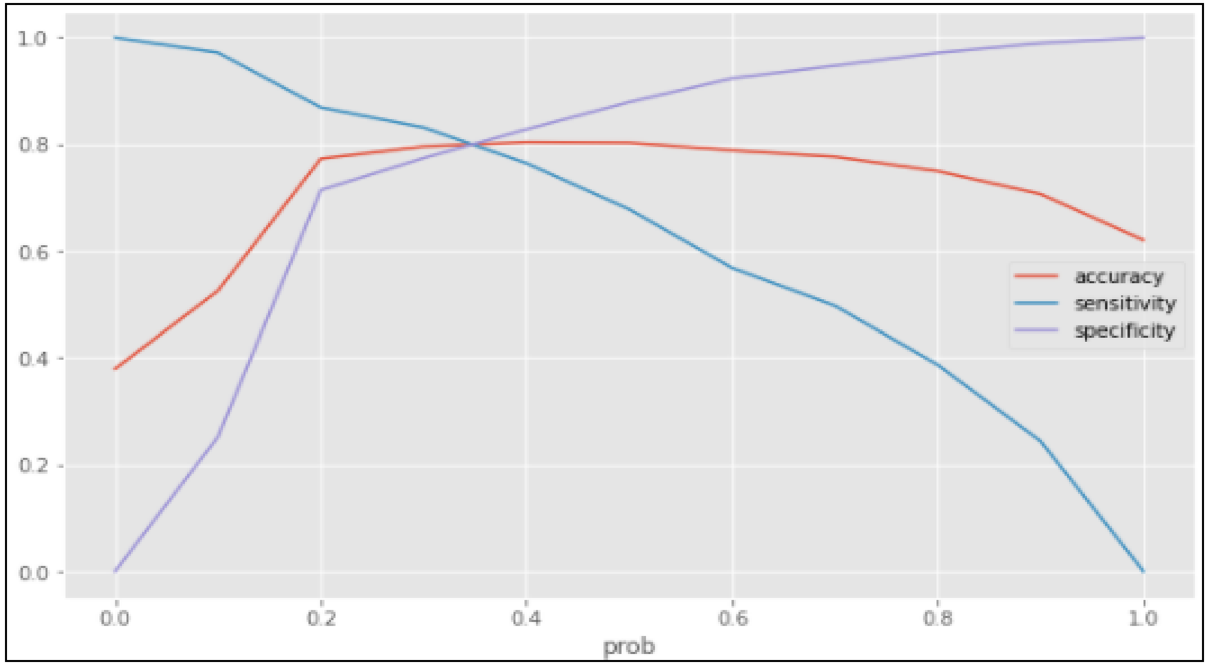
We get most of the leads and most of the converted leads from Mumbai.



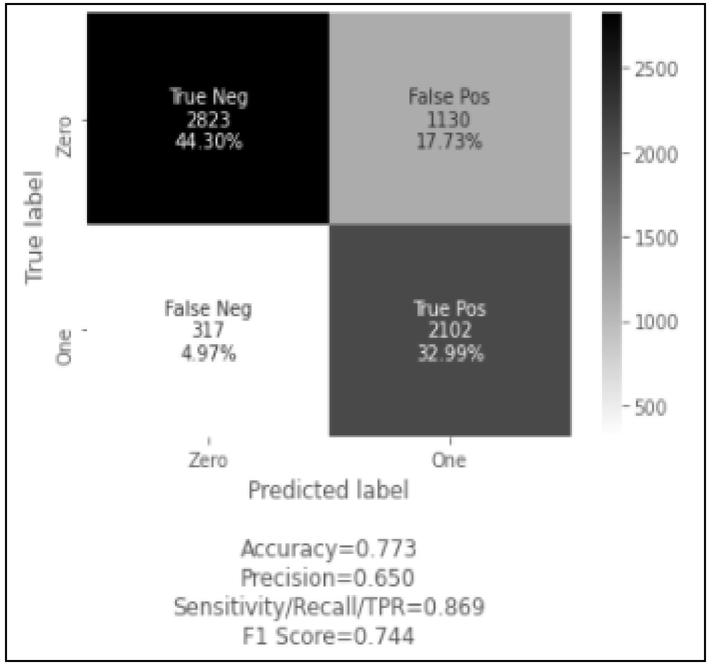
Finding the variables impacting Conversion most

1. We scaled the continuous variables using standard scalar
2. We created dummy variables for all categorical variables.
3. To get to the prediction, we built a model using Logistic Regression.
4. After all the analysis, we found that the most impactful variables are
   a) current occupation_Working Professional
   b) current occupation_Unemployed
   c) Lead Source_Olark Chat
   d) Total Time Spent on Website
   e) Specialization_Management
   f) Lead Origin_Others
   g) Specialization_Banking, Investment And Insurance
   h) Last Notable Activity_SMS Sent
   i) Do Not Email

Model Evaluation on Train Data Set



Plot of sensitivity, specificity and accuracy over different cut-offs

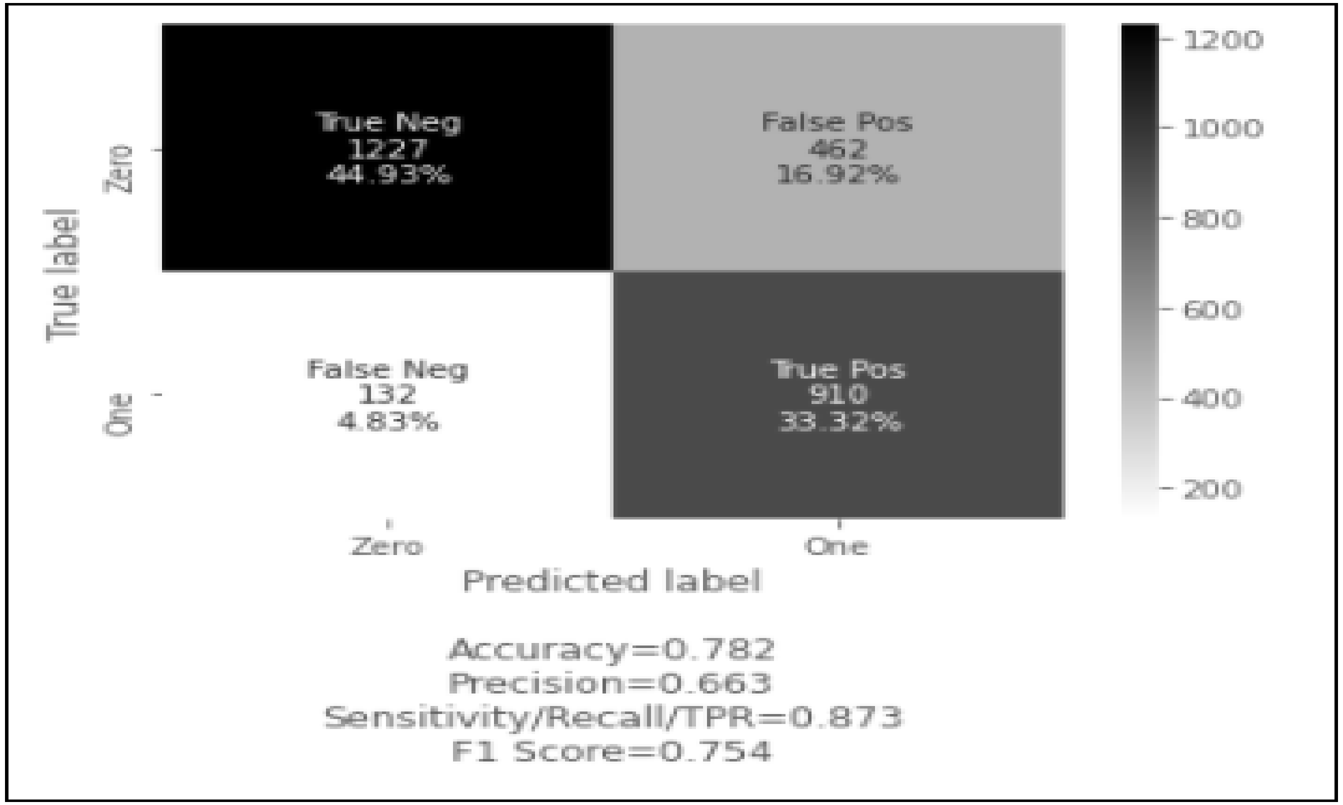


Accuracy=0.773
Precision=0.650
Sensitivity/Recall/TPR=0.869
F1 Score=0.744

Confusion Matrix with 0.2 as cut-off

1. Here we don’t want to miss out on potential converting leads, contacting people who are not going to convert to lead is still okay. So we are going to choose a higher sensitivity in a trade off with lower Specificity.
2. Since any number between 0.2 to 0.8 doesn’t improve accuracy, we are going to set the cut-off to 0.2

Model Evaluation on Test Data Set



Accuracy=0.782
Precision=0.663
Sensitivity/Recall/TPR=0.873
F1 Score=0.754

Scenarios

- If we have more resources to contact leads and we want to contact as many leads as possible, then we should choose a higher lower cut-off in the lead score.
- If we have lower number of resources to contact leads, then we should focus more towards high probability leads. Then we should choose a higher cut-off in the lead score.

Final Points

- ➢ We have tried to maximize the sensitivity, that is True Positive Rate, while keeping accuracy in check
- ➢ Accuracy and Sensitivity values of Train Data are 77.3% , 86.9% where as that of test data are 78.2% and 87.3%.
- ➢ The variables that contribute most for lead getting converted as per the model are
  - a. current occupation_Working Professional
  - b. current occupation_Unemployed
  - c. Lead Source_Olark Chat
  - d. Total Time Spent on Website
  - e. Specialization_Management
  - f. Lead Origin_Others
  - g. Specialization_Banking, Investment And Insurance
  - h. Last Notable Activity_SMS Sent
  - i. Do Not Email