Lead Scoring Case Study

- **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. X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein we need to assign a lead score 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%

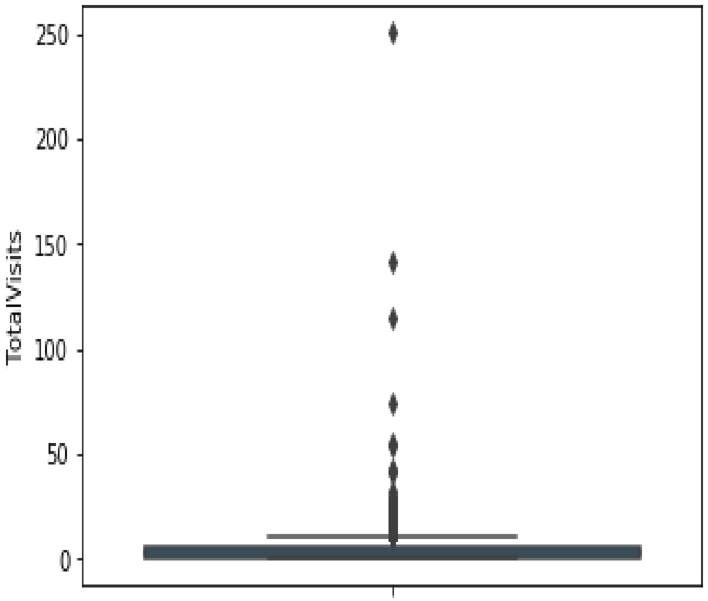
Created by Rudrakanta Ghosh & Dhritiman Banerjee

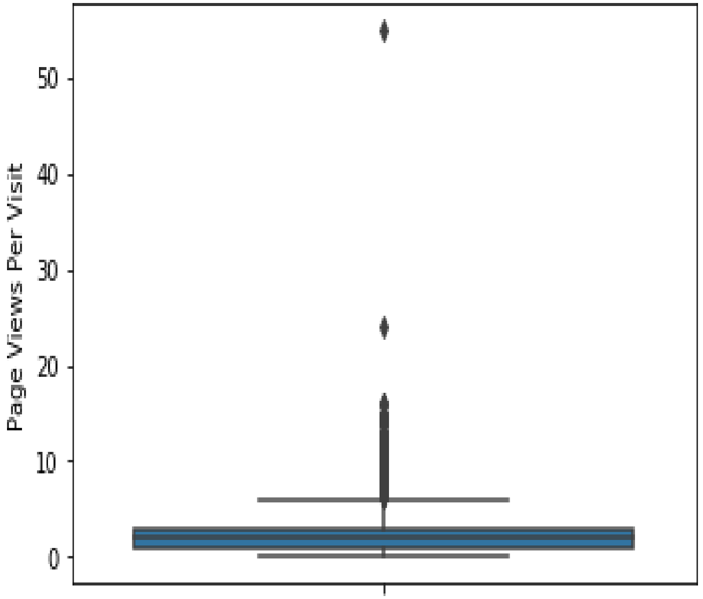
Steps for Data Understanding and Cleaning :

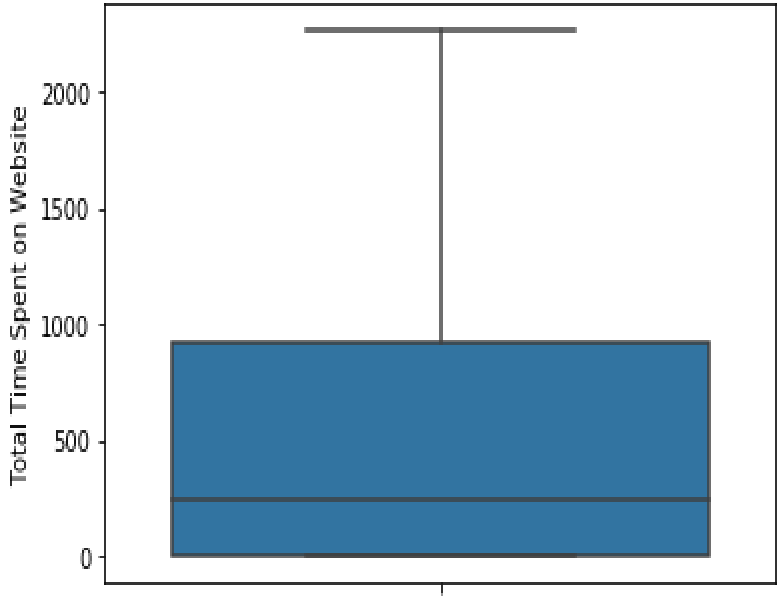
- We are removing few of the columns which doesn't have proper existence in their values. e.g. either Yes or No or negligible count for the secondary value
- Dropped further columns which has 46% of data NULL
- Dropping Country as 95.77% is India from the Non-NULL values
- Columns having values count of less than 200 we are going to merge all of those to a single value as *##Others e.g. LA*Others, LAN_Others and rest wherever applicable.
- For Specialization column we can see NULL & Select values adds up to 3380 which is approx. 30% of the total population. Moreover please note NULL and Select can be considered as identical. Hence merging all those to a new value as Unknown_Specialization
- More than 60% for "How did you hear about X Education" is Select and we can't manipulate this with any other ways like random variable or anything else. Hence dropping this too
- Now for Occupation we will club Student/Other/Housewife/Businessman in one group due to the low count and the NULL to Unknown to maintain the difference
- Again for "What matters most to you in choosing a course" the variance of the values for Other AND Flexibility & Convenience is negligible and this column will not make and sense or difference in our analysis. Hence dropping this too.
- 40% of the City is Unknown hence dropped that too
- For the columns TotalVisits & Page Views Per Visit are having around 137 rows with NULL values which is very less in comparison to the whole dataset. Hence, we are dropping those NULL records

Created by Rudrakanta Ghosh & Dhritiman Banerjee

Outliers Analysis :



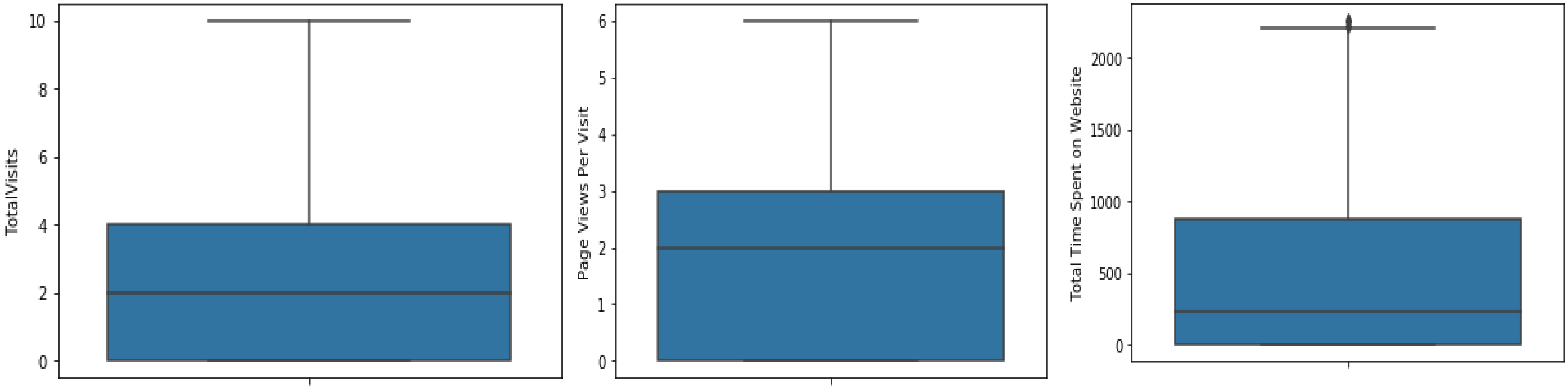




Looking into the boxplots we are considering to remove the outliers for TotalVisits & Page views Per Visit with 0.05 %

Created by Rudrakanta Ghosh & Dhritiman Banerjee

Post Outliers Treatment :



We observed Prospect ID & Lean Number is UNIQUE and can be the used as Identity in future purpose. Hence, preserving these columns for future use.

Current Lead Conversion rate post Outliers treatment is 38%

Created by Rudrakanta Ghosh & Dhritiman Banerjee

Steps for Data Preparation:

- Conversion of column data with binary values
- Dummy variable creation
  - Initially we have manipulated the data on few columns with "Others/Unknown" values hence, deleting "Others/Unknown" dummy column to be clean and simple
- Started with Training and Test Data Set Split
  - Feature Scaling
  - Model Building
  - Running 1st Training Model
  - Feature Selection using RFE
  - Model assesment with Statsmodel
  - Running 2nd Training Model
  - Insignificant feature: LeadProfile_Lateral Student ➔ **p value – 0.999**
  - Running 3rd Training Model
  - Insignificant feature: LeadQuality_High in Relevance ➔ **p value – 0.067**

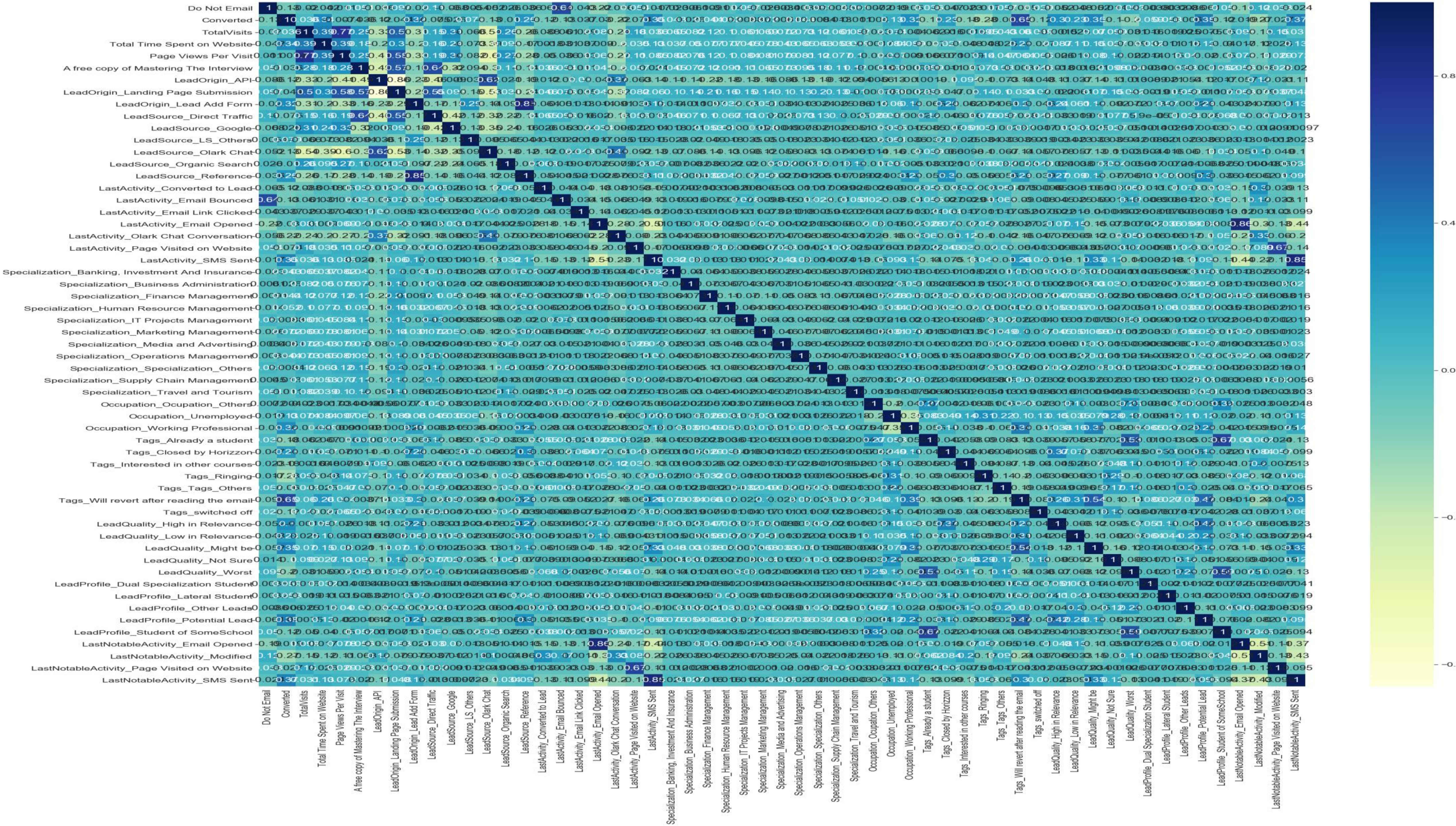
Created by Rudrakanta Ghosh & Dhritiman Banerjee

Continued…

Steps for Data Preparation Cont…:

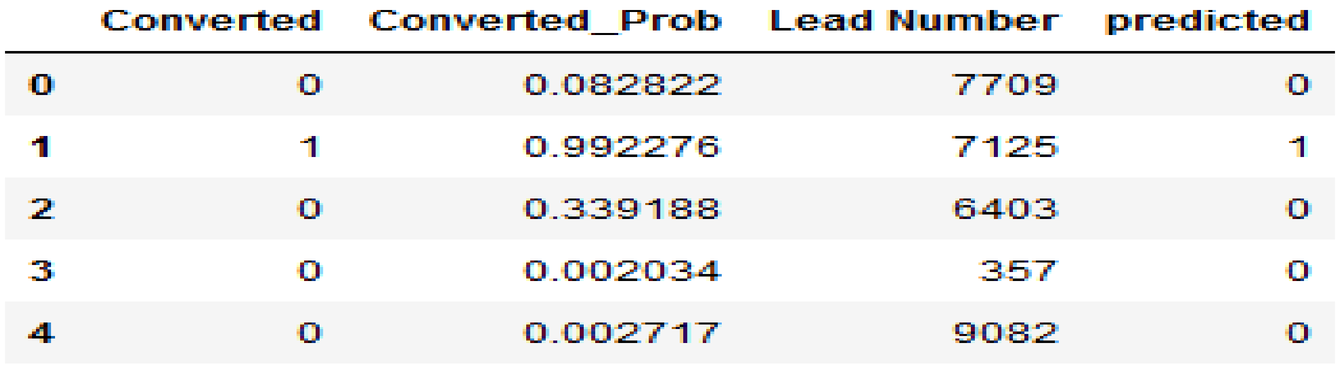
- Started with Training and Test Data Set Split
  - Running 4th Training Model
  - Insignificant feature: Specialization_Travel and Tourism ➔ **p value – 0.052**
  - Running 5th Training Model
  - Create a dataframe that will contain the names of all the feature variables and their respective VIFs
  - We have few highly correlated feature like - LastNotableActivity_SMS Sent & LastActivity_SMS Sent
  - Running 6th Training Model by removing the first one
  - Create a dataframe that will contain the names of all the feature variables and their respective VIFs
  - All VIFs are below 5
  - Creating Dataframe with the actual converted flag and predicted probabilities from 6th model
  - Accuracy, Sensitivity and Specificity -- Train Data
  - **Accuracy – 91%   Sensitivity – 85%   Specificity – 95%**
- **Correlation coefficients among the variables presented in next slide.**
  - **There are many variable which are highly correlated to each other.**

Created by Rudrakanta Ghosh & Dhritiman Banerjee


Do Not Email
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
A free copy of Mastering The Interview
LeadOrigin_API
LeadOrigin_Landing Page Submission
LeadOrigin_Lead Add Form
LeadSource_Direct Traffic
LeadSource_Google
LeadSource_LS_Others
LeadSource_Olark Chat
LeadSource_Organic Search
LeadSource_Reference
LastActivity_Converted to Lead
LastActivity_Email Bounced
LastActivity_Email Link Clicked
LastActivity_Email Opened
LastActivity_Olark Chat Conversation
LastActivity_Page Visited on Website
LastActivity_SMS Sent
Specialization_Banking, Investment And Insurance
Specialization_Business Administration
Specialization_Finance Management
Specialization_Human Resource Management
Specialization_IT Projects Management
Specialization_Marketing Management
Specialization_Media and Advertising
Specialization_Operations Management
Specialization_Specialization_Others
Specialization_Supply Chain Management
Specialization_Travel and Tourism
Occupation_Ocupation_Others
Occupation_Unemployed
Occupation_Working Professional
Tags_Already a student
Tags_Closed by Horizzon
Tags_Interested in other courses
Tags_Ringing
Tags_Tags_Others
Tags_Will revert after reading the email
Tags_switched off
LeadQuality_High in Relevance
LeadQuality_Low in Relevance
LeadQuality_Might be
LeadQuality_Not Sure
LeadQuality_Worst
LeadProfile_Dual Specialization Student
LeadProfile_Lateral Student
LeadProfile_Other Leads
LeadProfile_Potential Lead
LeadProfile_Student of SomeSchool
LastNotableActivity_Email Opened
LastNotableActivity_Modified
LastNotableActivity_Page Visited on Website
LastNotableActivity_SMS Sent
Do Not Email
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
A free copy of Mastering The Interview
LeadOrigin_API
LeadOrigin_Landing Page Submission
LeadOrigin_Lead Add Form
LeadSource_Direct Traffic
LeadSource_Google
LeadSource_LS_Others
LeadSource_Olark Chat
LeadSource_Organic Search
LeadSource_Reference
LastActivity_Converted to Lead
LastActivity_Email Bounced
LastActivity_Email Link Clicked
LastActivity_Email Opened
LastActivity_Olark Chat Conversation
LastActivity_Page Visited on Website
LastActivity_SMS Sent
Specialization_Banking, Investment And Insurance
Specialization_Business Administration
Specialization_Finance Management
Specialization_Human Resource Management
Specialization_IT Projects Management
Specialization_Marketing Management
Specialization_Media and Advertising
Specialization_Operations Management
Specialization_Specialization_Others
Specialization_Supply Chain Management
Specialization_Travel and Tourism
Occupation_Ocupation_Others
Occupation_Unemployed
Occupation_Working Professional
Tags_Already a student
Tags_Closed by Horizzon
Tags_Interested in other courses
Tags_Ringing
Tags_Tags_Others
Tags_Will revert after reading the email
Tags_switched off
LeadQuality_High in Relevance
LeadQuality_Low in Relevance
LeadQuality_Might be
LeadQuality_Not Sure
LeadQuality_Worst
LeadProfile_Dual Specialization Student
LeadProfile_Lateral Student
LeadProfile_Other Leads
LeadProfile_Potential Lead
LeadProfile_Student of SomeSchool
LastNotableActivity_Email Opened
LastNotableActivity_Modified
LastNotableActivity_Page Visited on Website
LastNotableActivity_SMS Sent
0.8
0.4
0.0
−0.4
−0.8

Testing model on Test Data :

- Scaling of Test Data
- Prediction on the Test Data
- Predicted Dataset head ➔➔

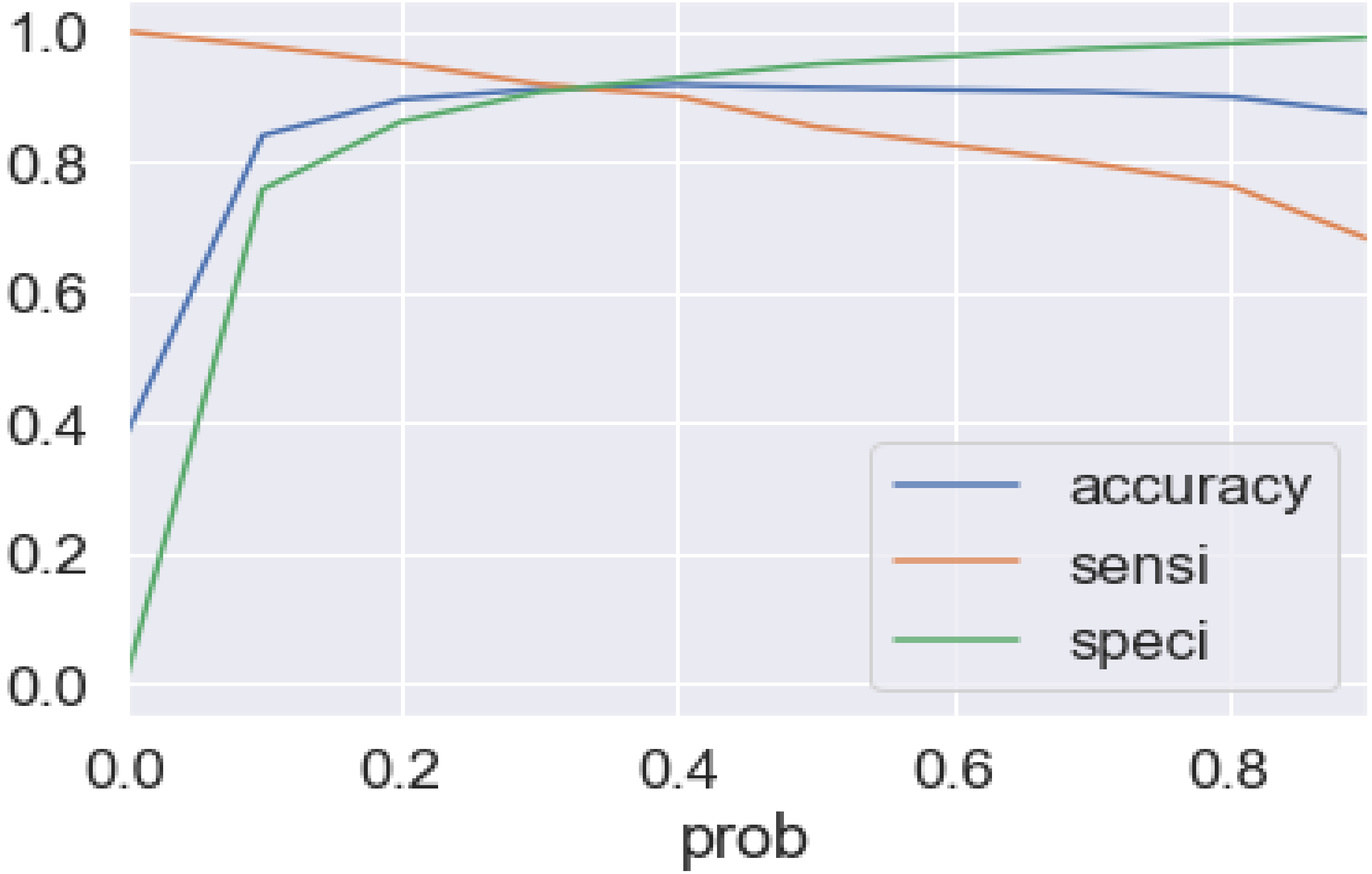


	Converted	Converted_Prob	Lead Number	predicted
0	0	0.082822	7709	0
1	1	0.992276	7125	1
2	0	0.339188	6403	0
3	0	0.002034	357	0
4	0	0.002717	9082	0

- Accuracy, Sensitivity and Specificity -- Test Data
- **Accuracy – 91%   Sensitivity – 84%   Specificity – 96%**

Created by Rudrakanta Ghosh & Dhritiman Banerjee

Finding Optimal Cutoff Point:



Above curve suggests the optimum point to take it as a cutoff probability.

Created by Rudrakanta Ghosh & Dhritiman Banerjee