LEAD SCORING CASE STUDY

PRESENTED BY:

ANANTA KUMAAR V R & SAMBRIT SAHA

IIIT-B

Problem Statement

An educational company named X sells online courses to industry professionals

Now although X education gets a lot of Leads, but it's Lead conversion rate is very poor around 30%

The Company wants to increase it 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

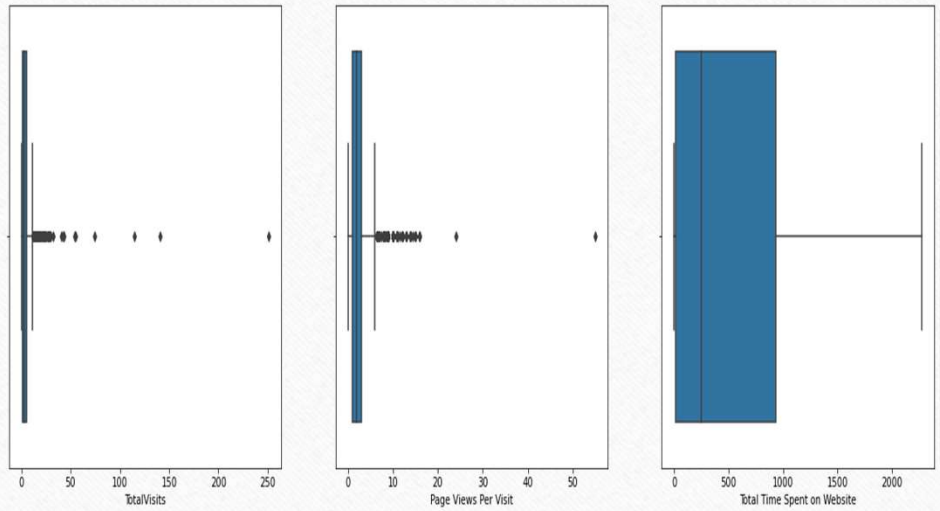
Methods

- Data Reading and Understanding
- Data Cleaning
- Analysing Outliers
- Exploratory Data Analysis – Univariate and Bivariate analysis
- Data Preparation
- Model Building
- Model Evaluation
- Precision-Recall Curve
- Measure of Accuracy, Sensitivity and Specificity

Outliers in the Data

Inference

There was presence of Outliers in the data. By comparing all the three data, we can clearly see that Total Visits and Page Views Per Visit has many Outliers.
When comparing to Total Time Spent in website the outlier value shall be processed by taking median function.

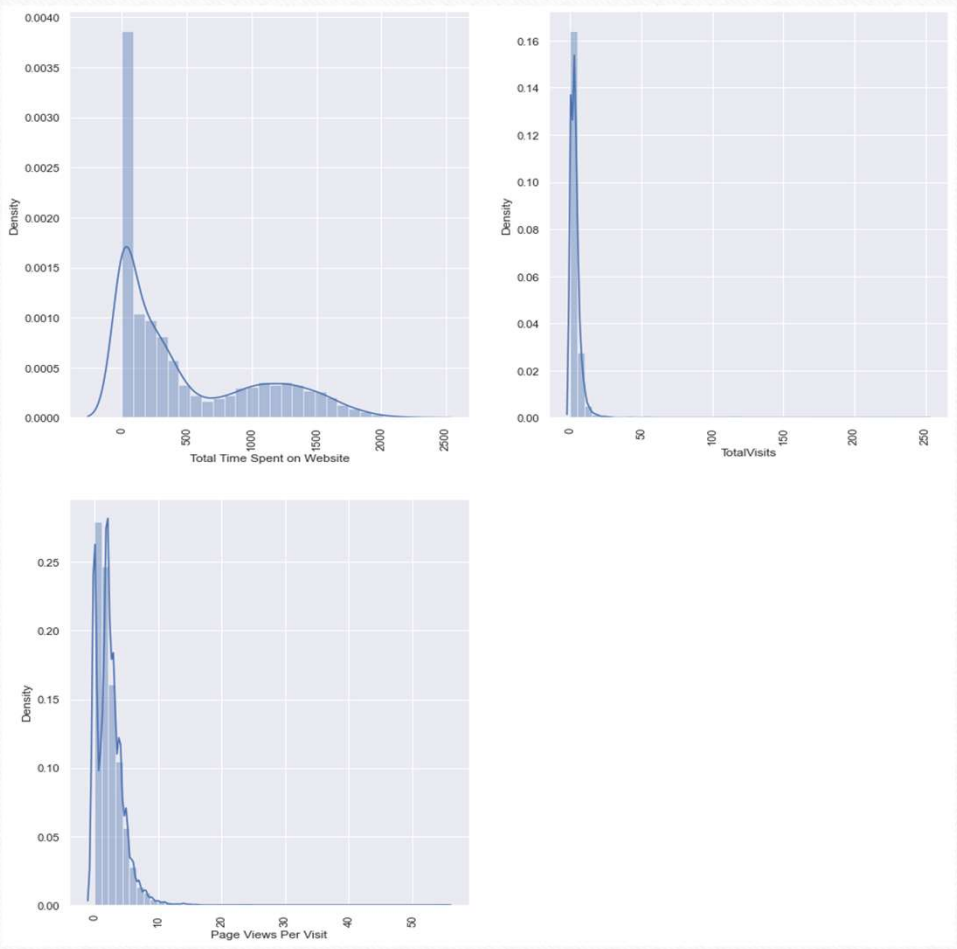


Univariate Analysis

Inference

As we can clearly see, by conducting a univariate analysis for continuous variables, none of the data is not normally distributed.

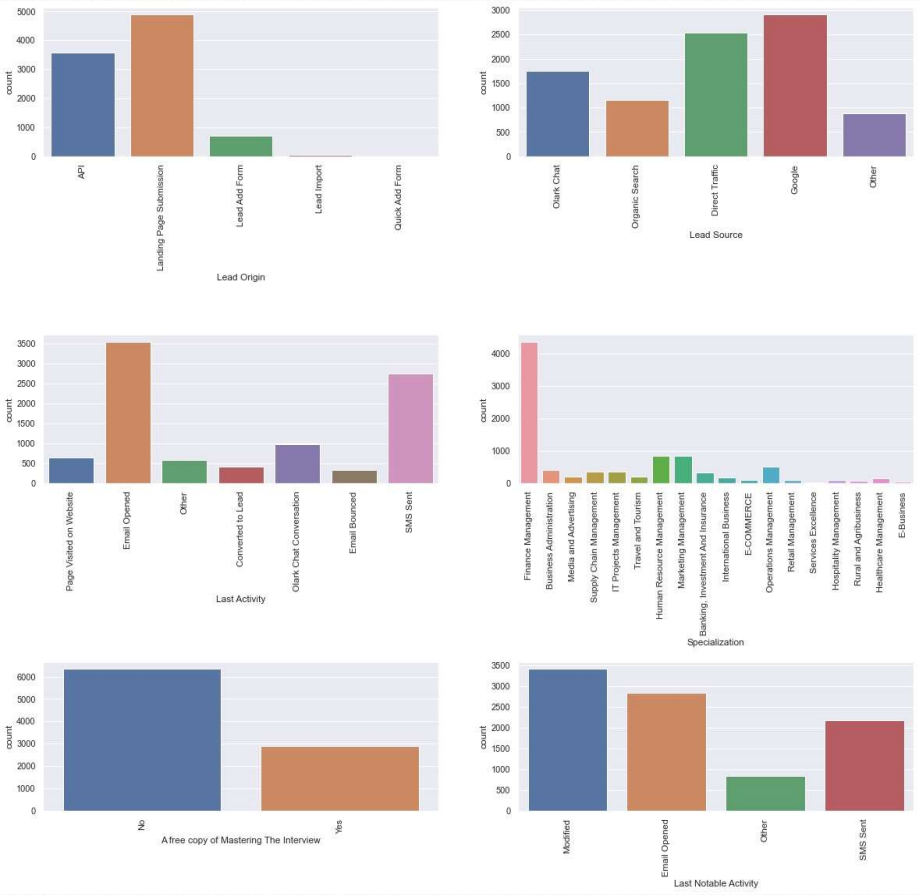
Because the numerical data contains outliers which should be taken out respectively.



Count Plot For Objective Data

Inference

From the data we can see that in column 'Lead Source', 'Google' & 'Direct Traffic' contains the highest values which clearly states that these two categories are of the main points of lead source.
As for the 'Last Activity' column, the categories such as 'Email Sent' & 'SMS Sent' have higher values
For the 'Specialization' column, most people opted for 'Finance Management'



Bivariate Analysis

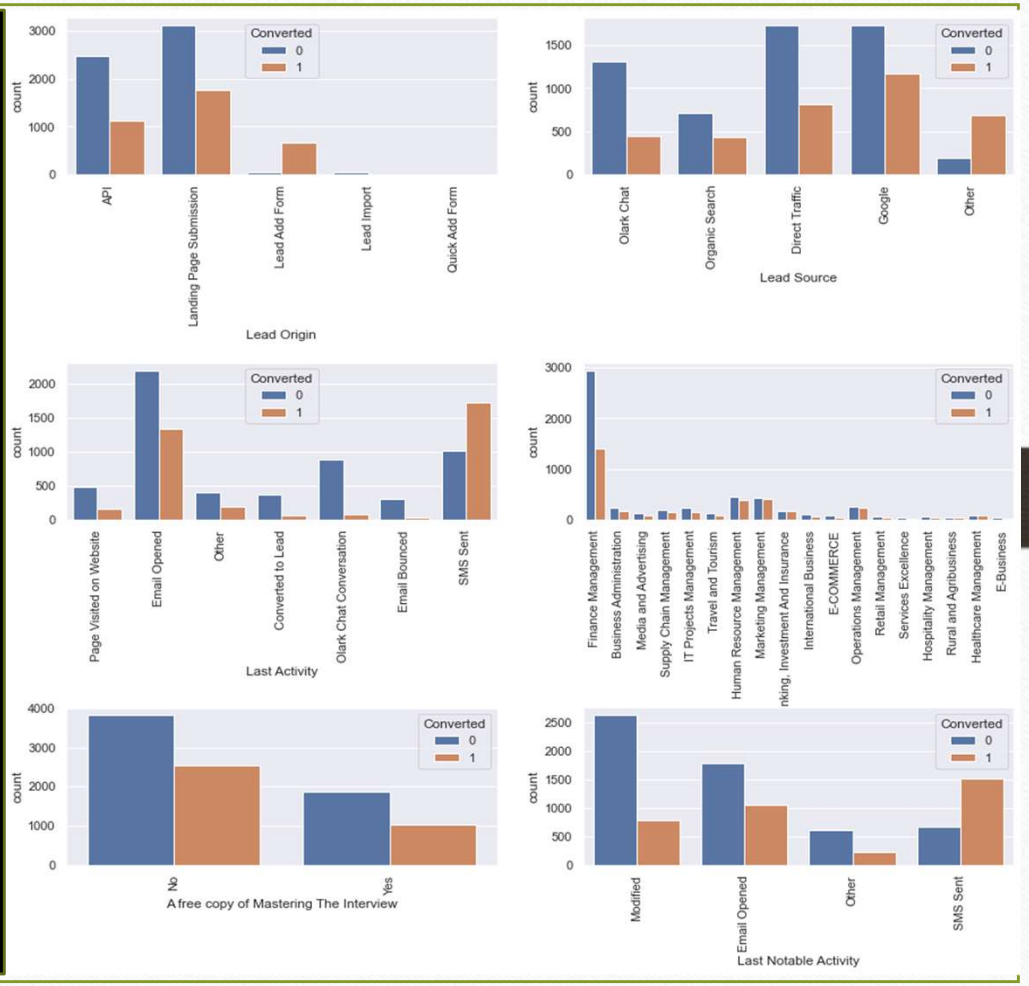
Inference

As we can see in column Lead Source, Direct Traffic and Google is higher whereas Other is lower

In Last Activity Email Opened category has higher value in blue region and SMS Sent category orange region is higher

Last Notable Activity is almost similar to Last Activity for Email Opened and SMS Sent

In Specialization, Finance Management is higher in the blue region



TOP FACTORS THAT IMPACT THE CONVERSION OF LEADS

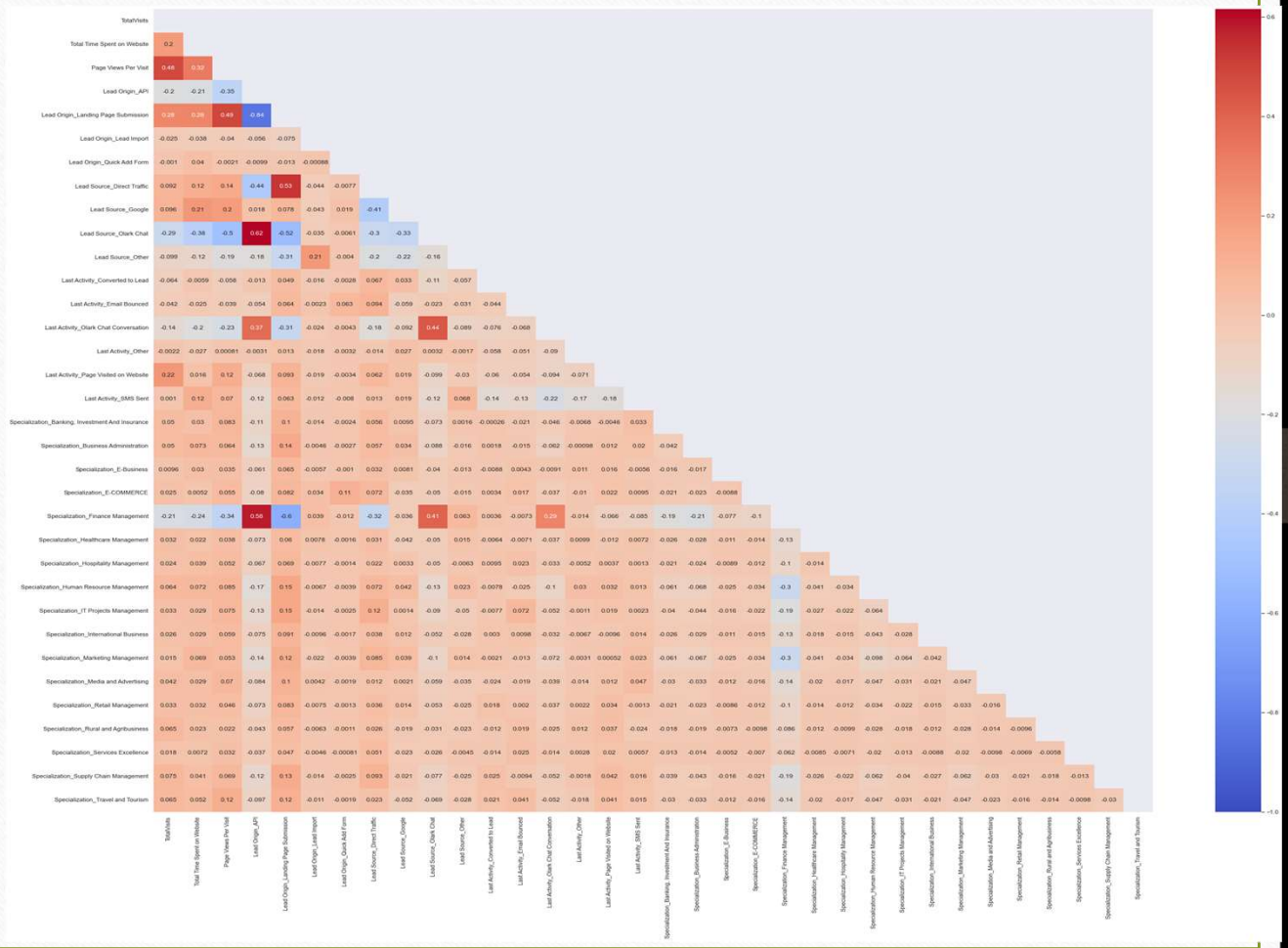
Features

Tags will revert after reading the email
Total time spent on Website
Total Visits
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 Horizon
Lead Quality Not Sure
Do Not Email_Yes
Tags_Lost to EINS
Lead Profile_Other Leads
Last Notable Activity_Olark Chat Conversion

Correlation

Inference

We can see that some values are highly correlated
By using Recursive Feature Elimination we can determine whether to drop columns or not.

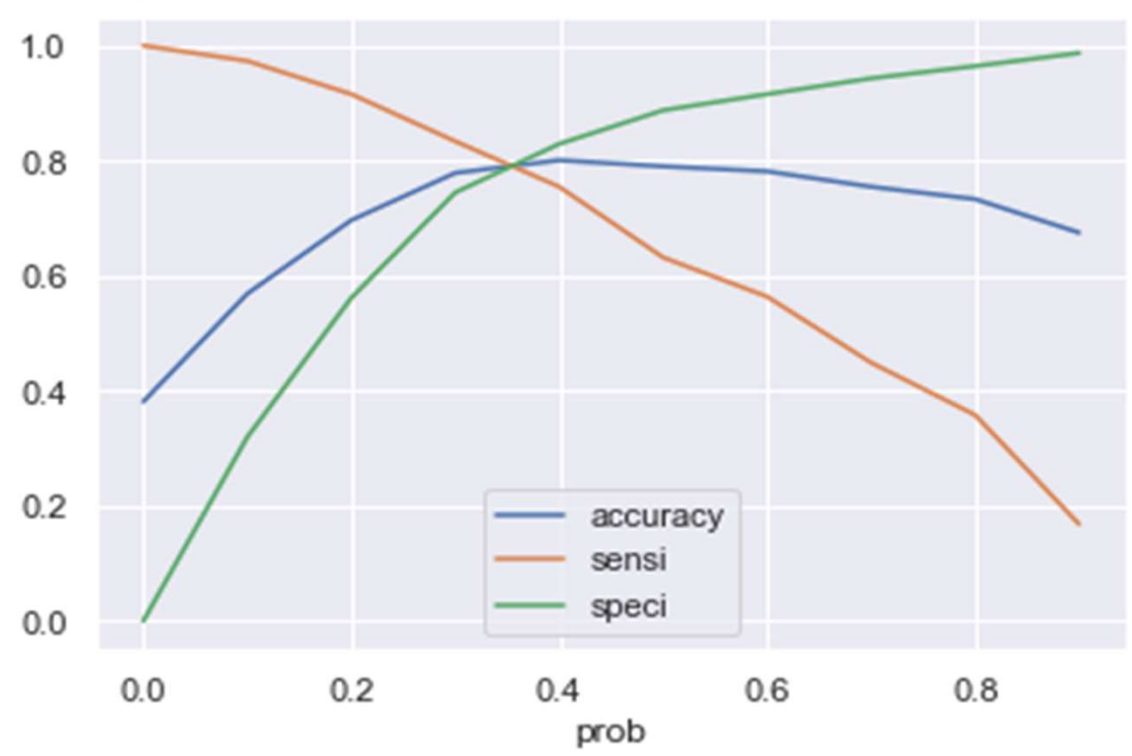


The count plot for accuracy, sensitivity and specificity was plotted to determine the final cut-off for the model.

We can clearly see that, three points coincide each other at 0.8 respectively for the trained data model set.

Hence by using this model we achieve a Right trade off point of having 0.35

That is 35% probability of a prospect is good enough to target as Hot lead and work more on the strategies of the Lead than the Cold ones

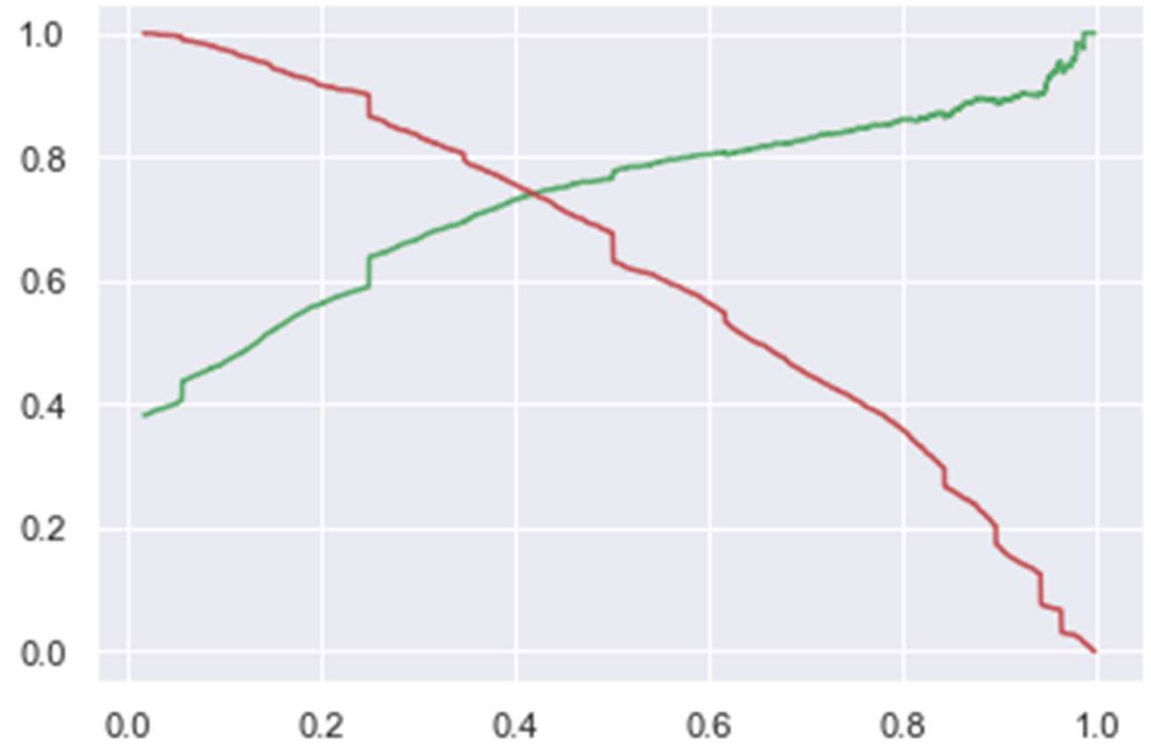


By using the count plot, we can clearly understand that the trained data set contains 79% accuracy, 80% specificity and sensitivity is 79% respectively.

To Achieve a Conversion rate of around 80%
Need to not miss those prospects which can turn to hot leads which is Recall
Should Not overestimate a cold lead which is Precision

A right choice of Probability at which the lead should be considered as potential and can turn into a Hot Lead is the Major need

Hence by using this model we achieve a Right trade off point of having 0.35
That is 35% probability of a prospect is good enough to target as Hot lead and work more on the strategies of the Lead than the Cold ones



The trade-off between Precision and Recall-
Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than 35% to be a hot Lead.

Conclusion

From the dataset using regression models, calculating VIF, splitting of train and test data & visualizing the data, we are finally able to determine the most important variables in potential buyers
Total Time Spent on the Website.
The Total Number of Visits.
The Lead Source of having google, direct traffic and Olark Chart as higher values.
The Last Activity containing higher values were SMS and Olark Chat Conversation.
Lead Origin is Lead Import.

	Lead Number	Lead Score
0	609431	11.97
1	631817	5.52
2	596164	6.22
3	646570	14.09
4	643974	3.80

Lead Number is determined and presented in a data frame format