LEAD SCORE CASE STUDY

PRESENTATION BY:

- *HAYAGREEVA SUDARSHAN SUNDARESWARAN*
- *SANSKRUTI BABAR*
- *SANGAMITRA SENTHIL KUMAR*

Objective

To help X Education to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

To build a logistic regression model to assign a lead score value between 0 to 100 to each of the leads which can be used by the company to target potential leads.

To be able to adjust to the company's requirement if changes in the future are demmed necessary

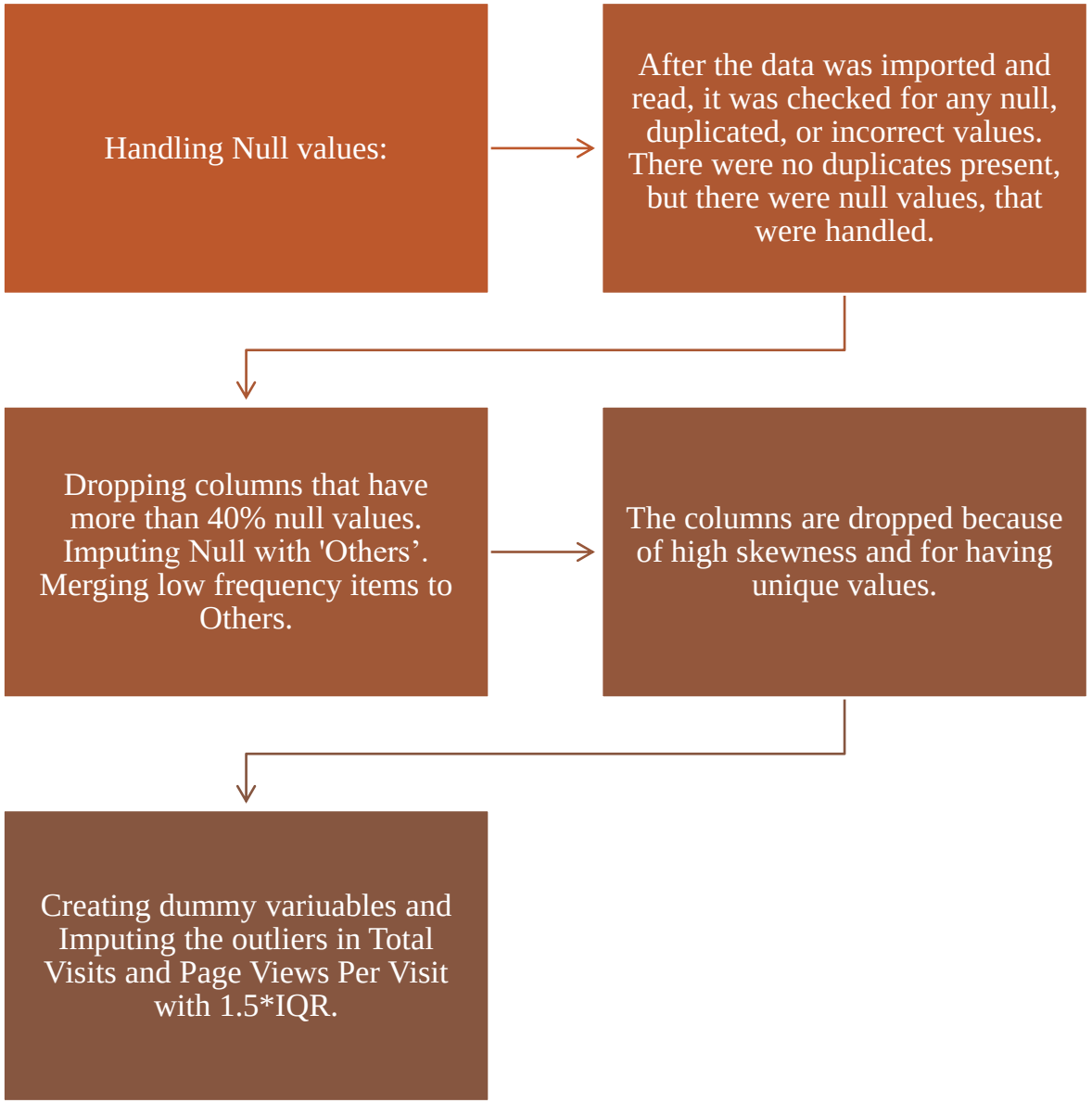
Problem Statement

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.

The company does marketing of its courses on various websites and search engines such as Google. Once the visitors land on the website, they browse through different courses or fill up a form for the course or watch related videos. When the visitors fill up a form and provide their contact details, they are classified as leads.

There are a lot of leads generated in the initial stage but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well in order to get a higher lead conversion. X Education has appointed you to help them select the most promising leads.

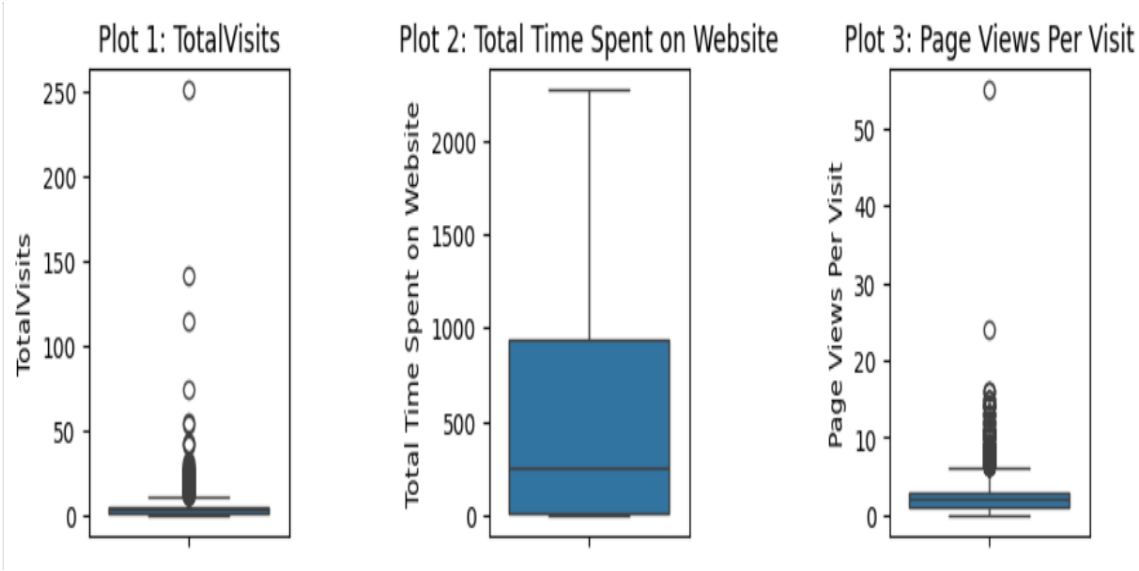
Data Cleaning



Exploratory Data Analysis

Univariate Analysis

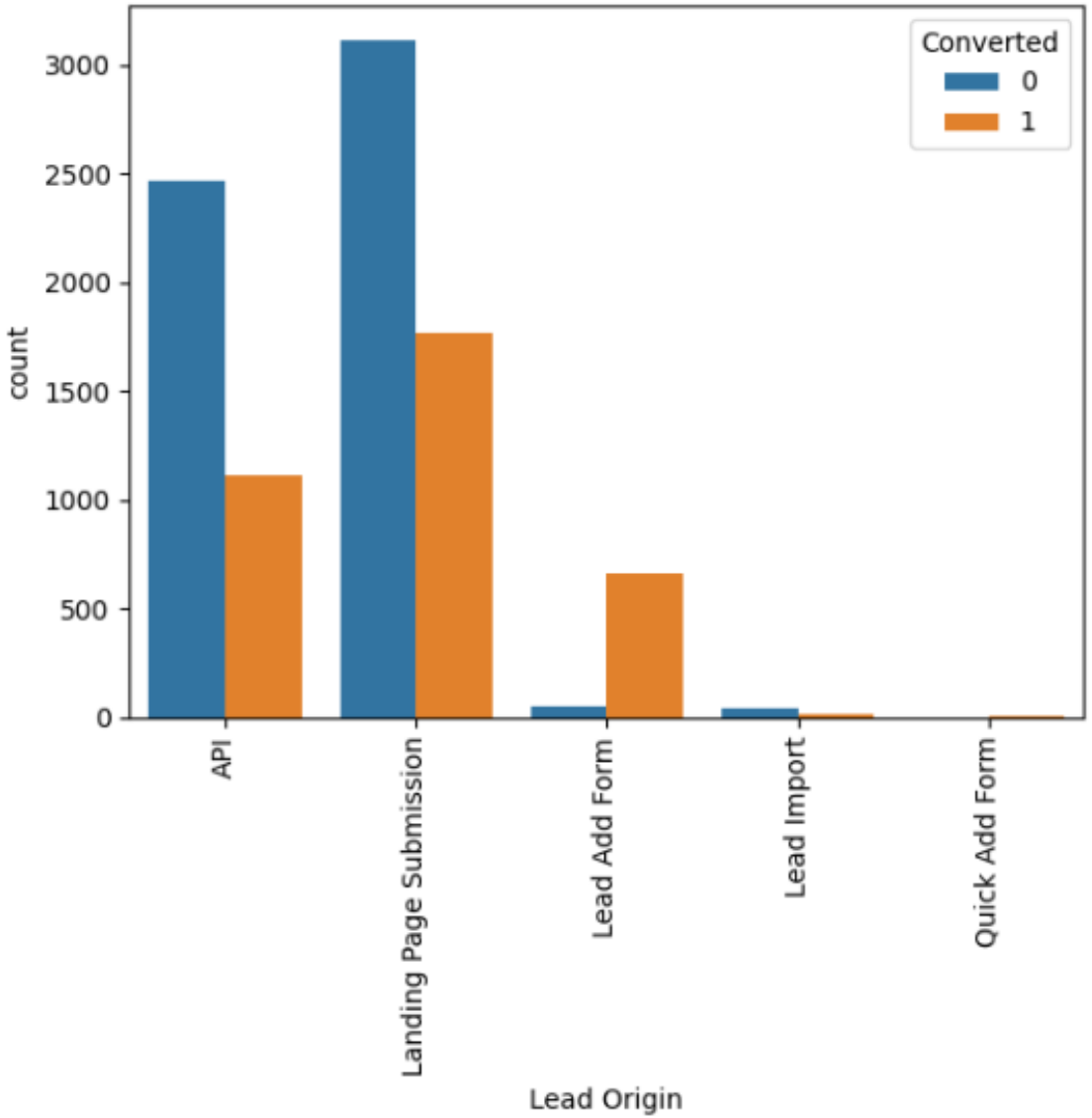
We run the univariate analysis to find the outliers and imputing the outliers in TotalVisits and Page Views Per Visit with 1.5*IQR



Lead origin

Inference:

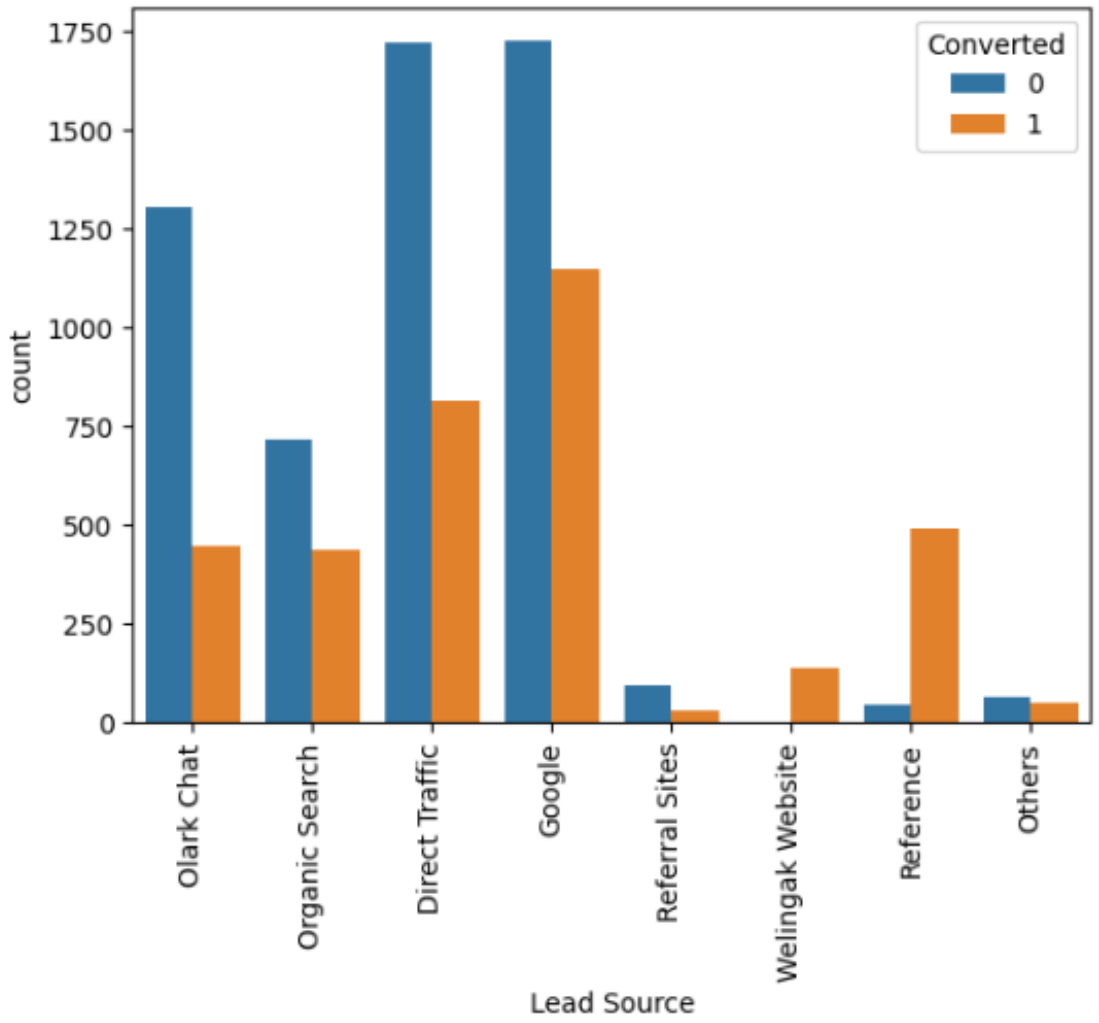
API and Landing Page Submission have around 35-40% conversion rate but count originated is of a considerable amount. Lead Add Form has more than 90% conversion rate but count is not very high. Lead Import and Quick Add From are very less in count.



Lead Source

Inference:

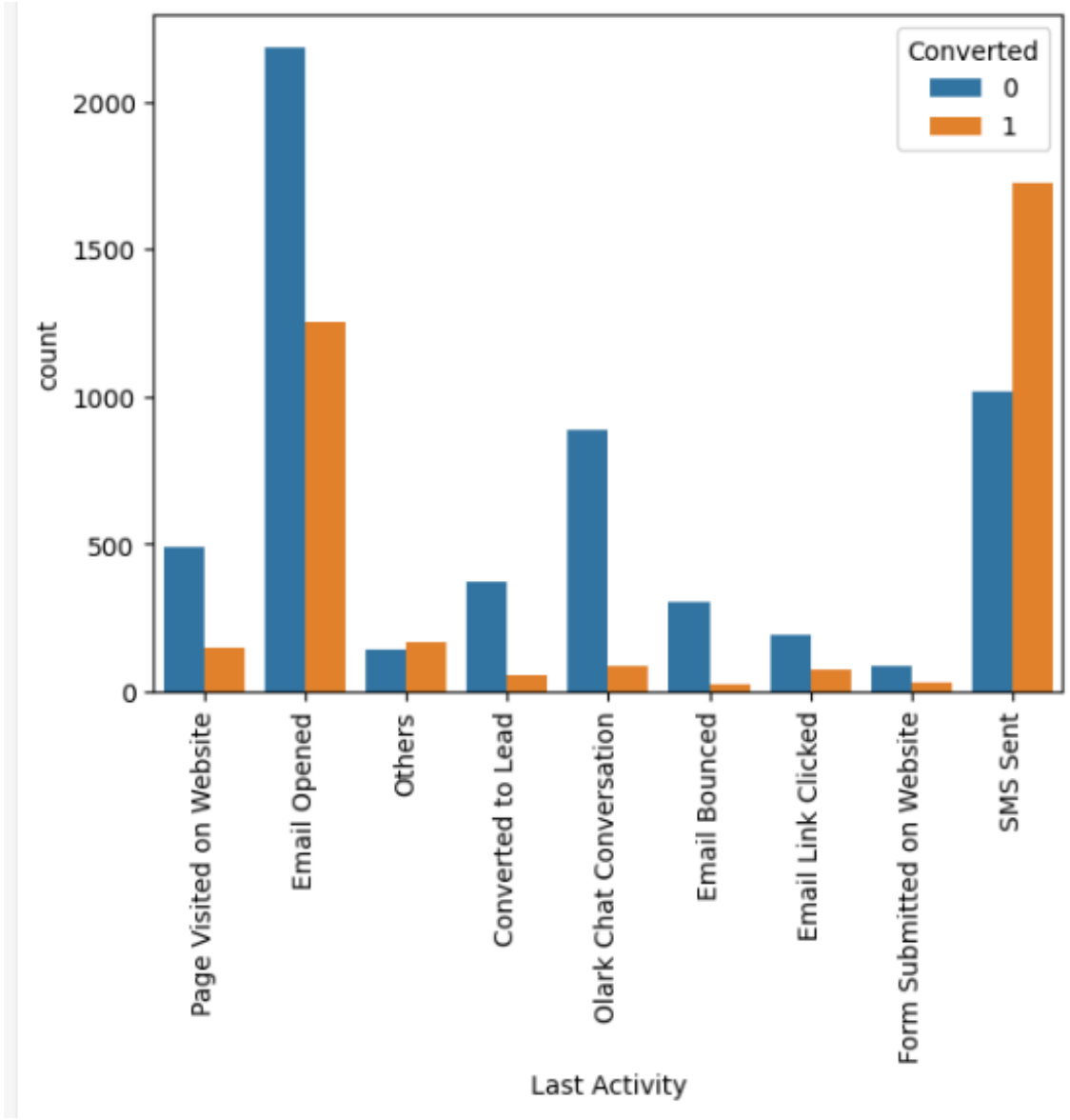
Direct Traffic and Google are generating high number of Lead_data; should be focused for their conversion Welingak and Reference are having more conversion rate hence should be focused to generate more Lead_data Lead_data from Olark Chat and Organic Search should be focused for conversion



Last Activity

Inference:

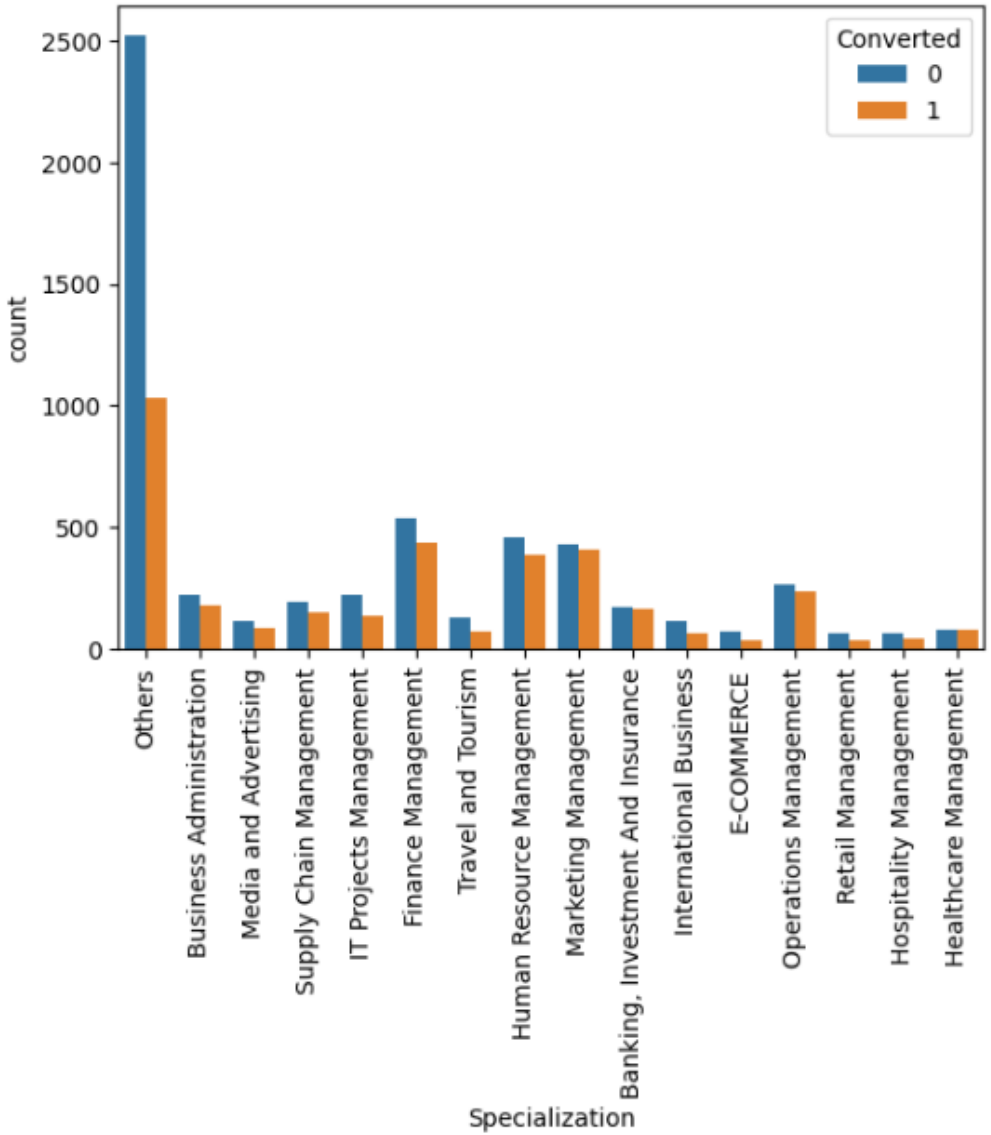
Most of the Lead_data have Email Opened and SMS Sent. More focus should be on Lead_data opening email. And conersion rate for SMS sent is very high and hence they should be targeted for lead conversion



Specialisation

Inference:

The focus should be on those various specializations with less Lead_data but high conversion rate.Most of the Lead_data are from 'Others' specialization but the conversion rate is poor

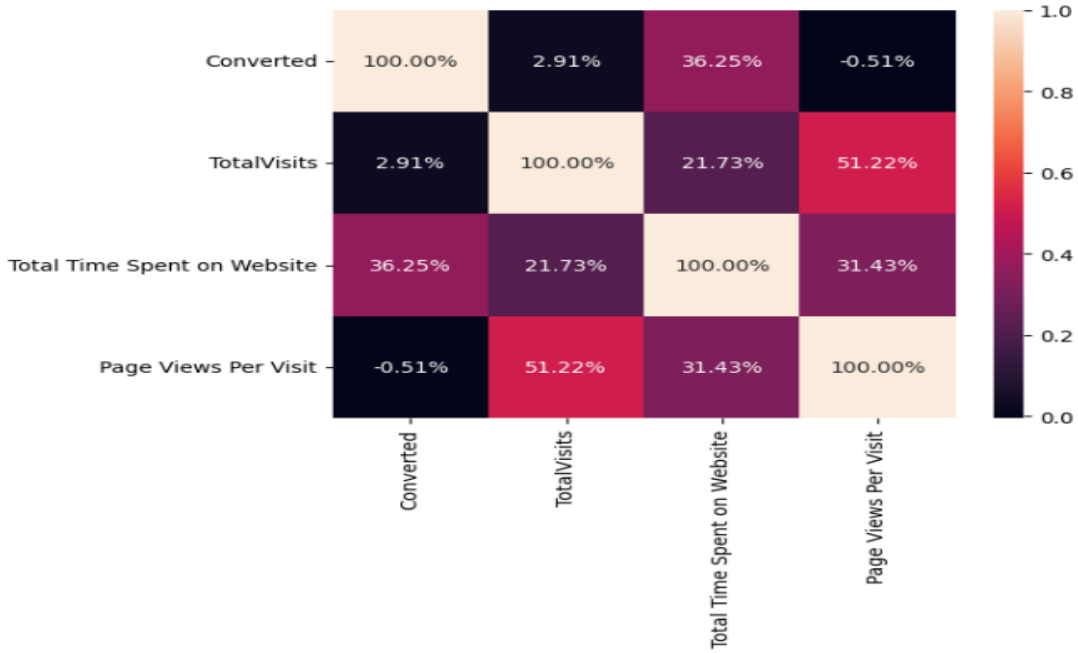


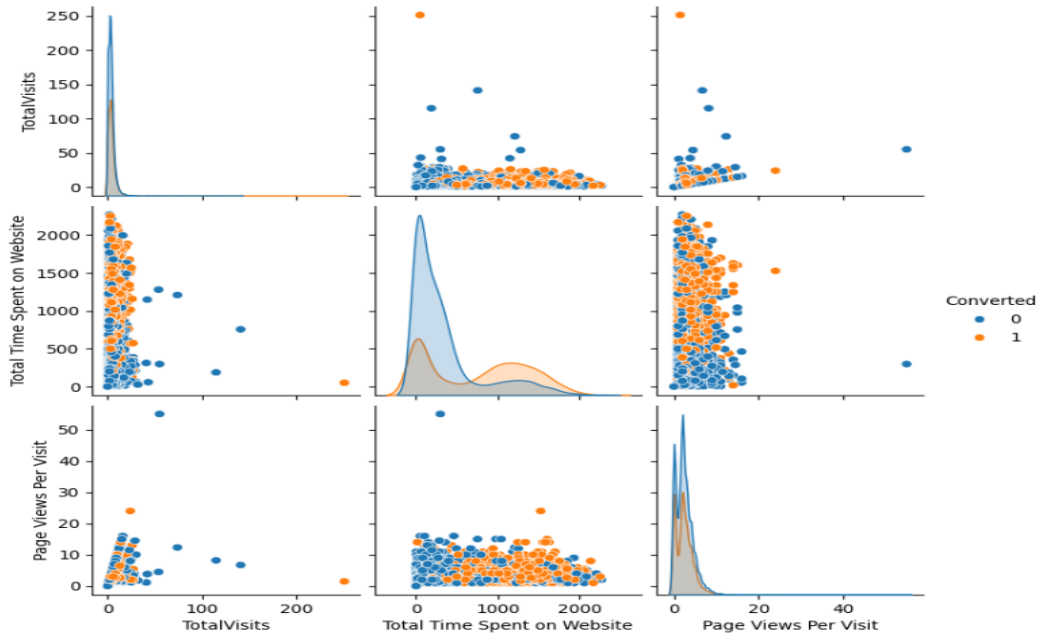
Bi variate analysis

Inference from above heat map

Strong correlation between Total Visits and Page Views Per Visits Converted has good correlation with Total time spent on Website.

More Lead_data are from those who have modified their account or opened email. focus should be on their conversion.SMS sent have high converison

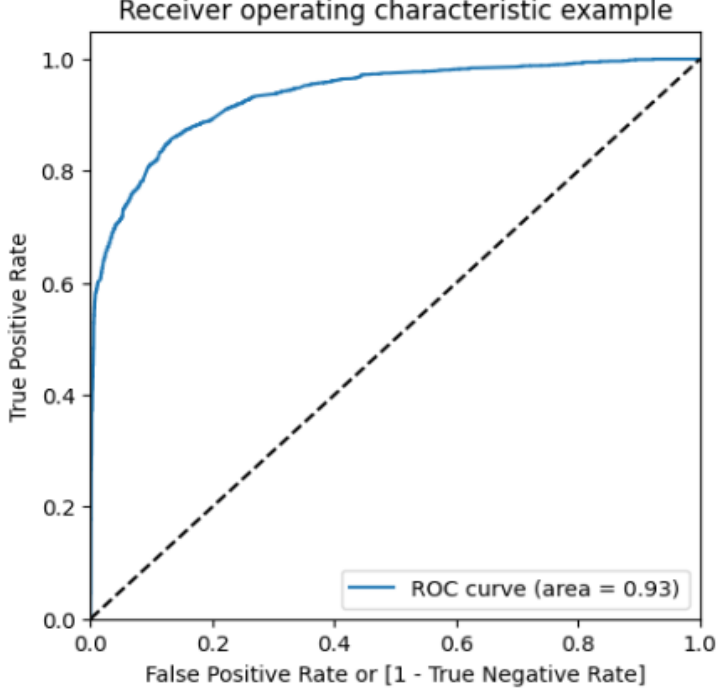


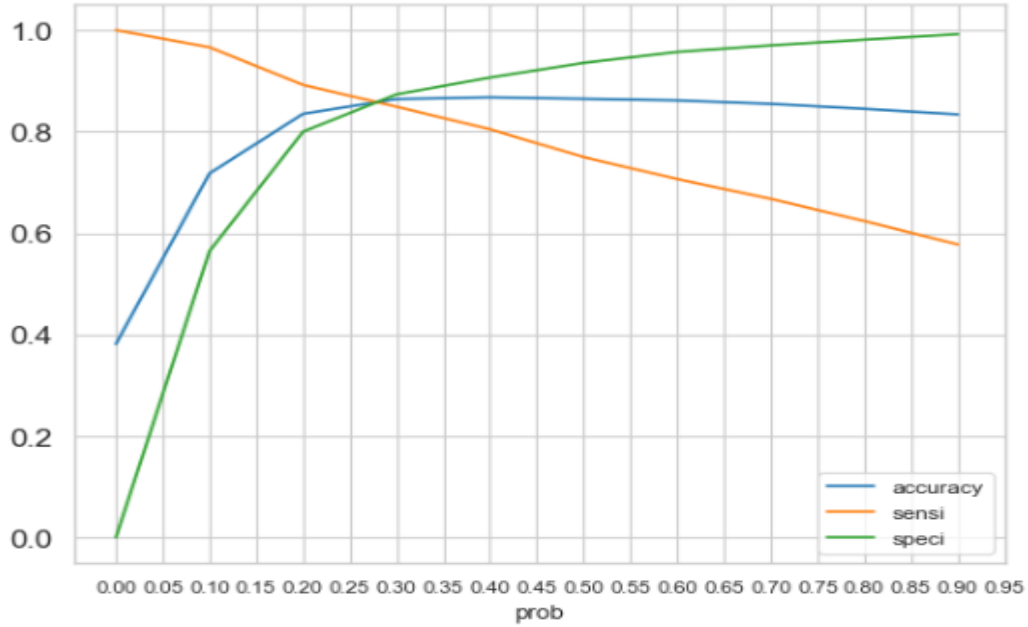


Model Details

Finding Optimal Cut-Off pointOptimal cutoff probability is that prob where we get balanced sensitivity and specificity

From the curve above, 0.28 is the optimum point to take it as a cutoff probability.





Comparing the model and metrics on training set

Metrics on training set	Model
Confusion matrix -1481(TP), 196(TN), 178(FP), 917(FN).	Confusion matrix -3460(TP), 542(TN), 347(FP), 2119(FN).
overall accuracy-86.5%	overall accuracy-86.2%
Sensitivity-83.7%	Sensitivity-85.5%
Specificity-88.3%	Specificity-86.4%
Precision-82.3%	Precision-79.6%
Recall-83.7%	Recall-85.28%

Conclusion

The Leads that have a higher ‘Lead Score’ should be focused on more for a better conversion rate.

Encouraging existing converted leads for referrals by providing some kind of reward for doing so.

The unemployed category should be focused on more.

Students should be focused less, as they have a lower conversion rate