Lead Scoring Case Study

By:
Lav Soni
Anuradha Murthi Needi

Problem Statement

INTRODUCTION: An education company named X Education sells online courses to industry professionals. Once 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. The typical lead conversion rate at X education is around 30%.

GOAL: Company wishes to identify the most potential leads, also known as “Hot Leads”

The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. 80%

Overall Approach

- ❖Source the data for analysis
- ❖Reading & Understanding the data
- ❖Data Cleaning
- ❖EDA
- ❖Feature Scaling
- ❖Splitting the data into test & train dataset
- ❖Preparing the data for modelling
- ❖Model building
- ❖Model Evaluation-specificity & sensitivity or precision recall
- ❖Making predictions on the test set.

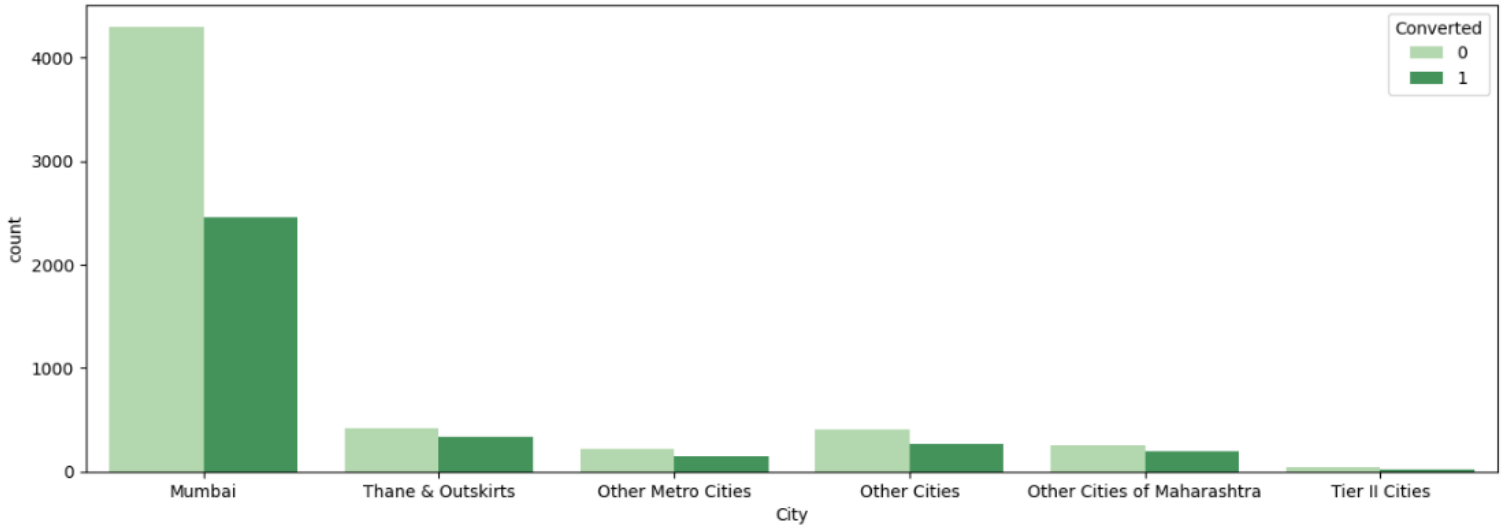
Dataset

Provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Data Sourcing, Cleaning & Preparation

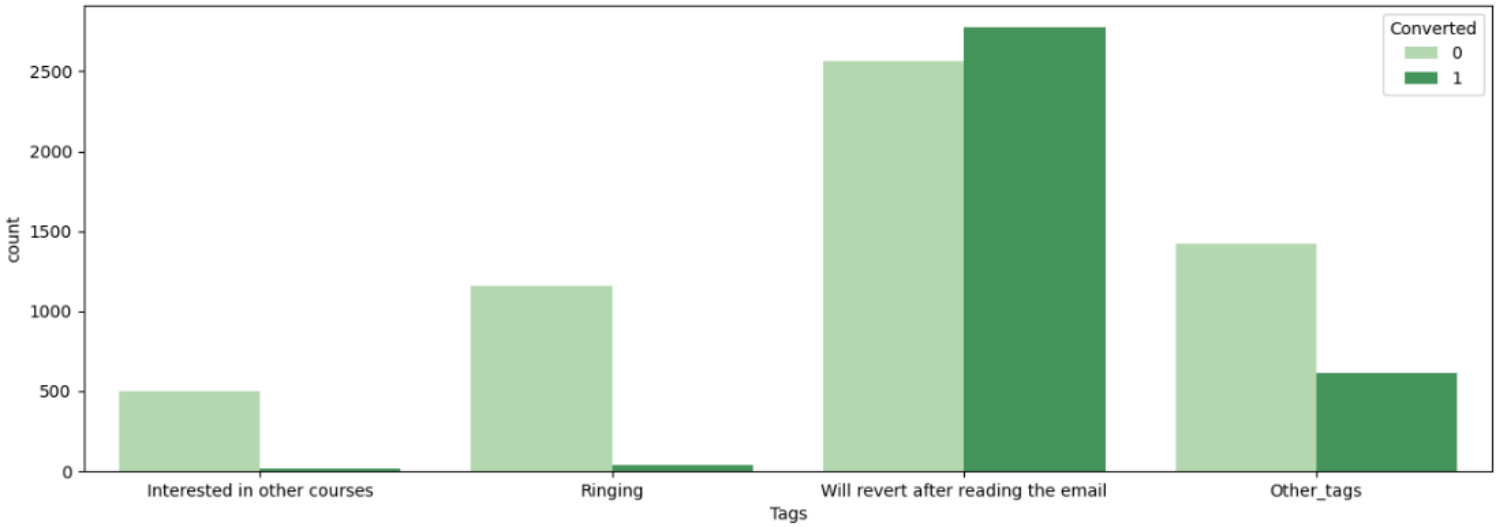
- ❖Read the data from csv file
- ❖Outlier treatment
- ❖Data Cleaning-Handling null values & removing higher Null values data
- ❖Removing redundant column in the data
- ❖Imputing Null values
- ❖EDA
- ❖Feature Standardization

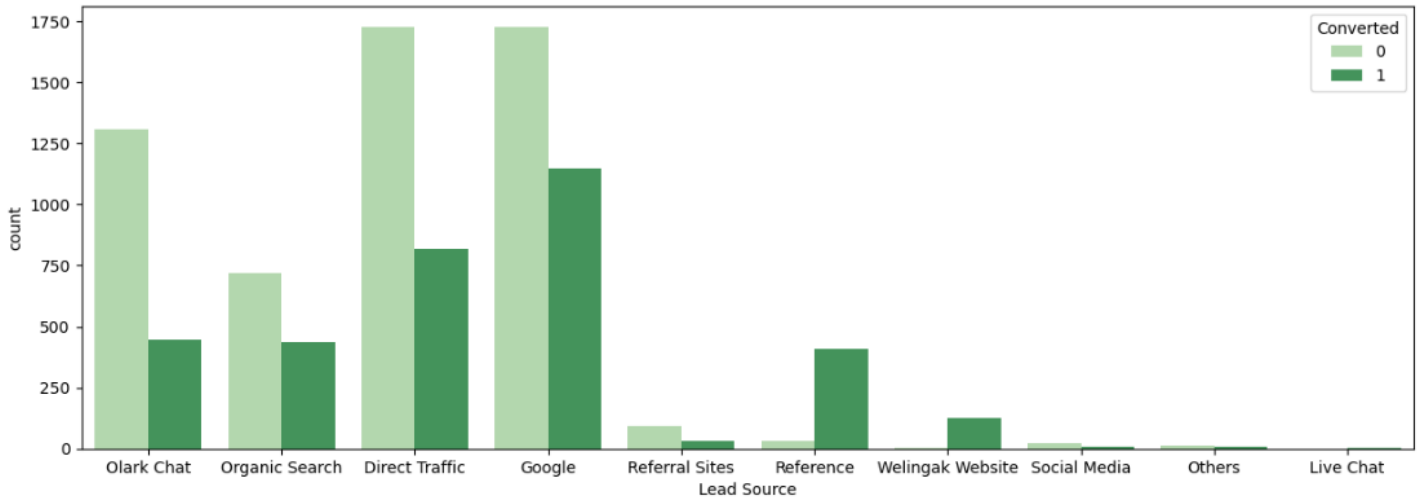
Exploratory Data Analysis



Mumbai city has high lead value

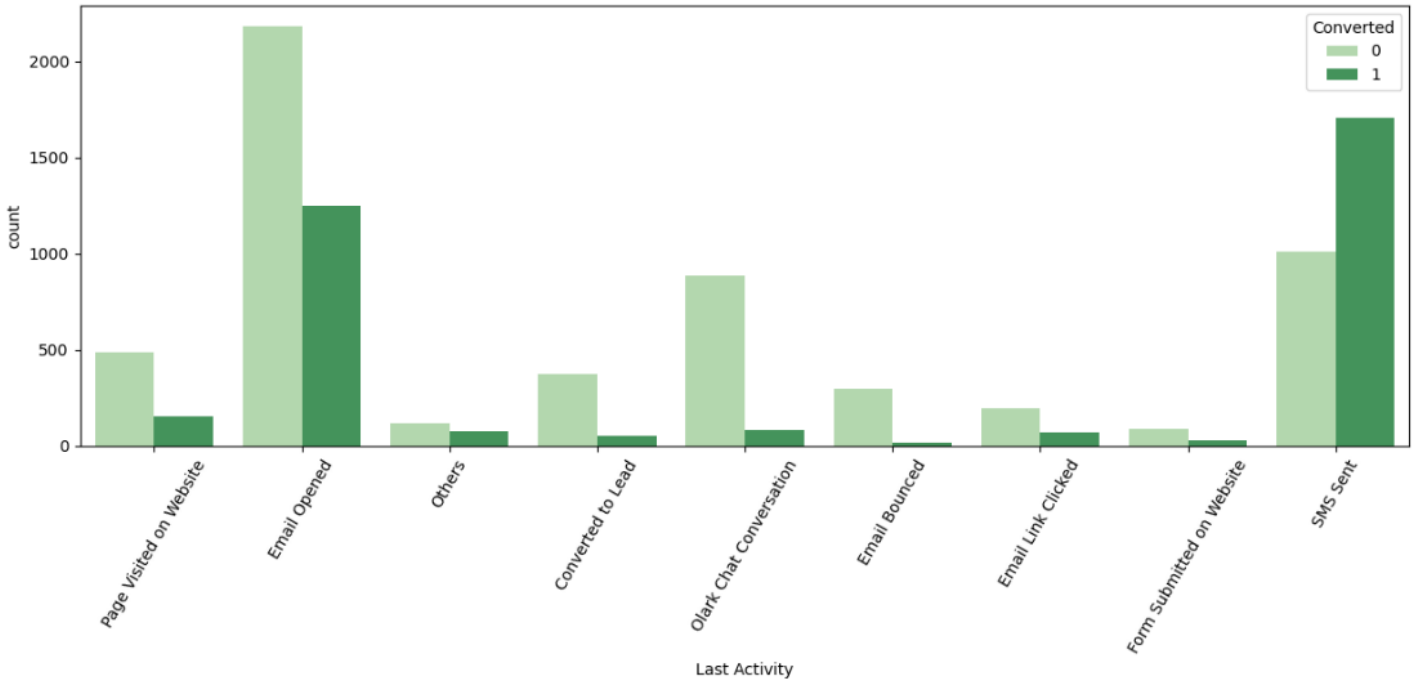
The conversion rate is high for “Will revert after reading the email” Tag

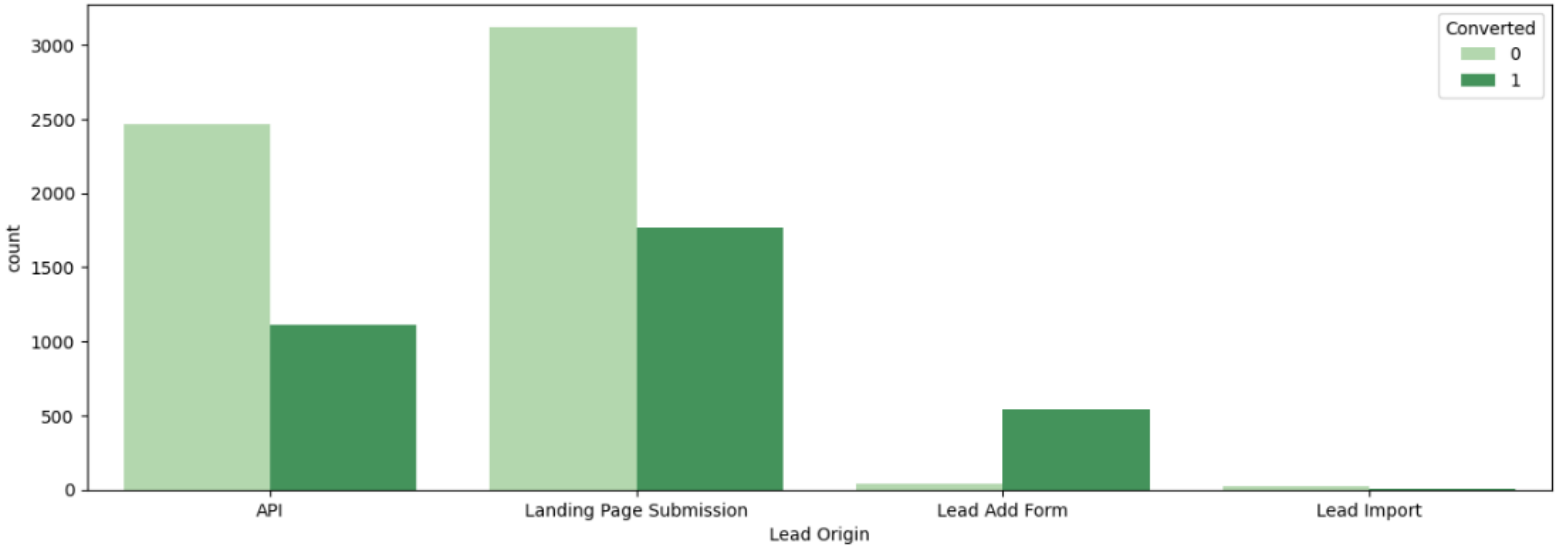




Major conversion in Lead source is from Google

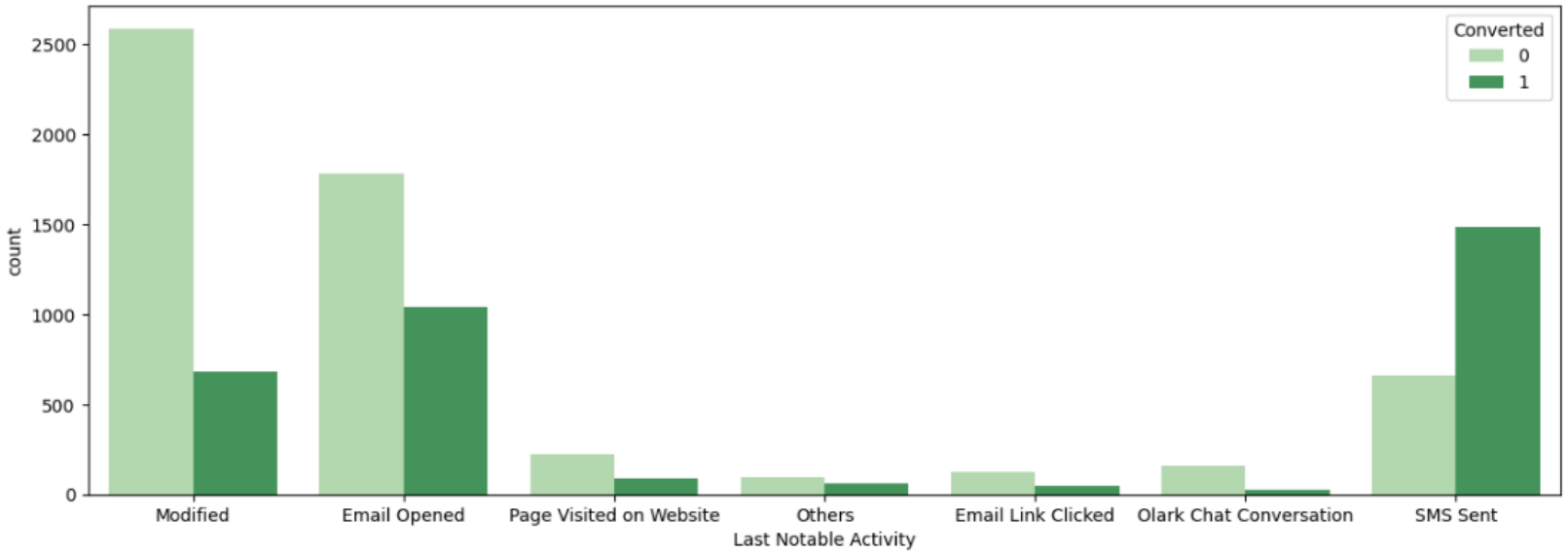
Major conversion in Lead Activity is from “SMS Sent”

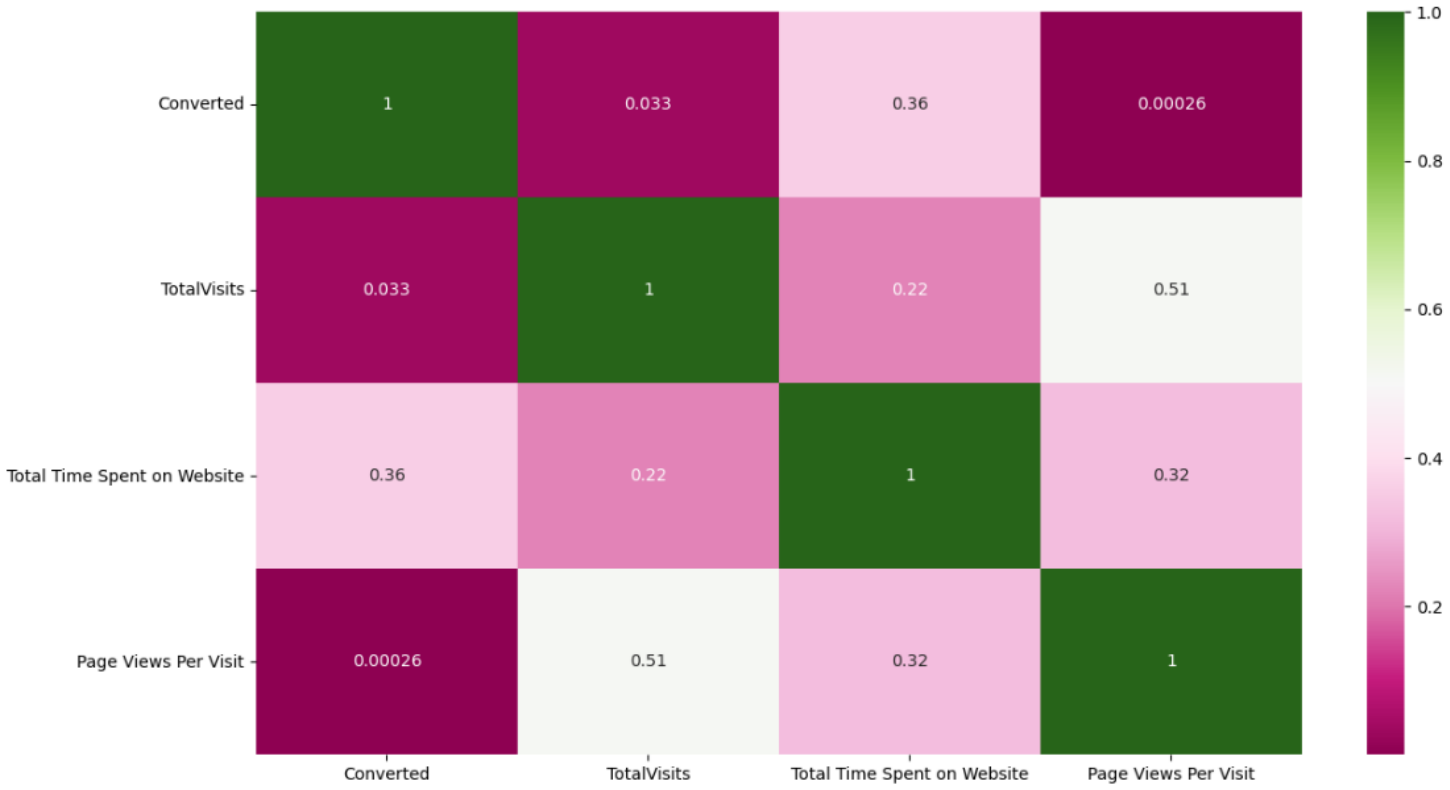




Maximum conversion happened in Lead origin from “Lading Page Submission”

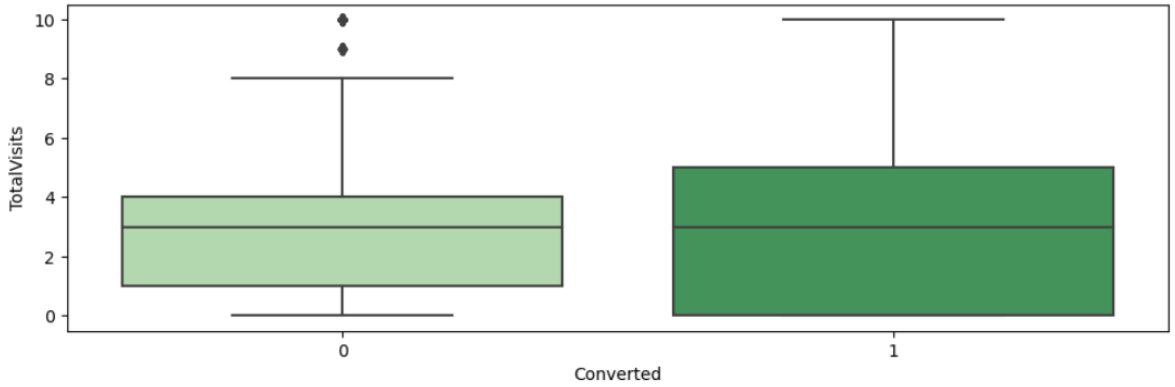
Major conversion in Lead Notable Activity is from “SMS Sent” and “Email Opened”

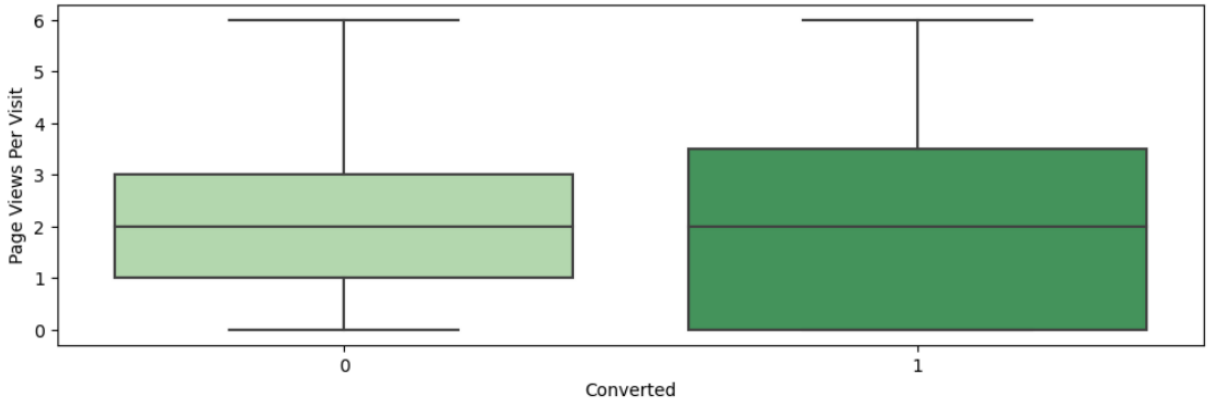


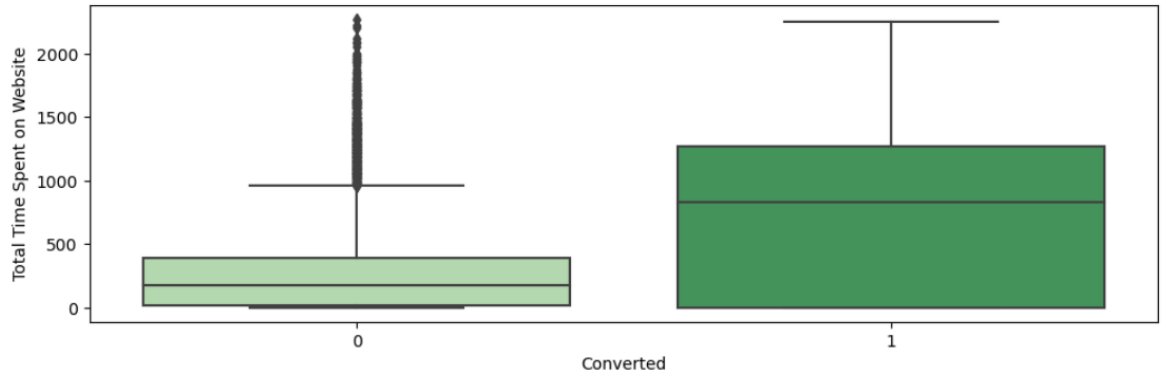


Heat map for checking correlation on continuous variables

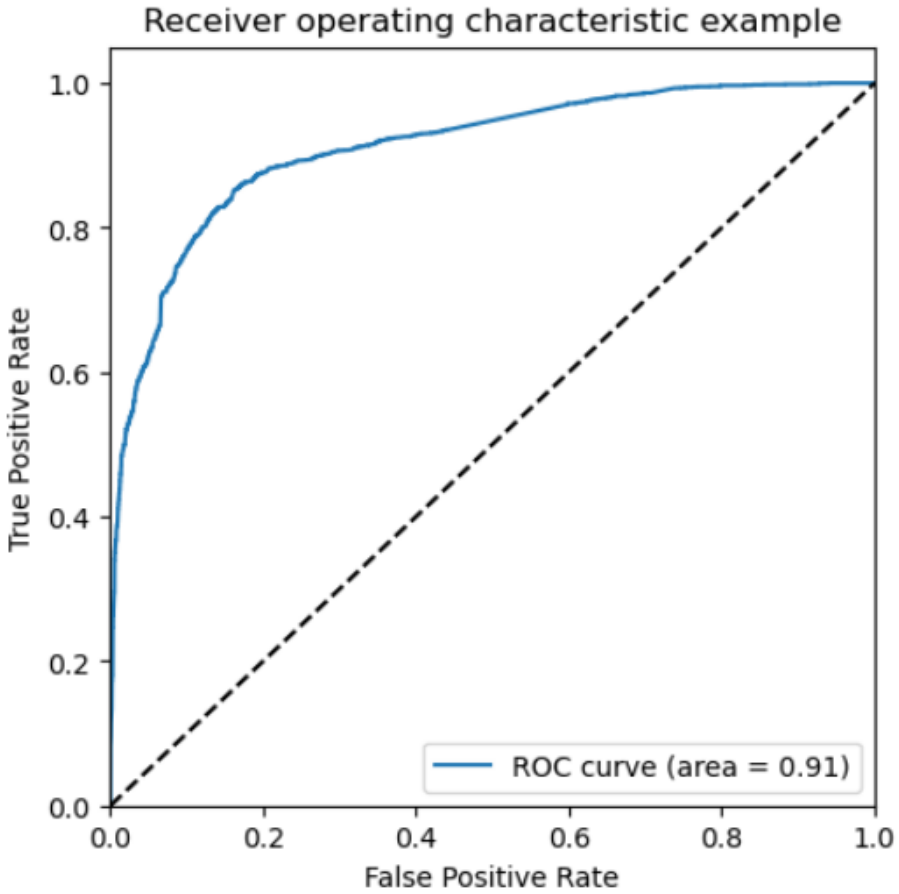
Outlier Analysis







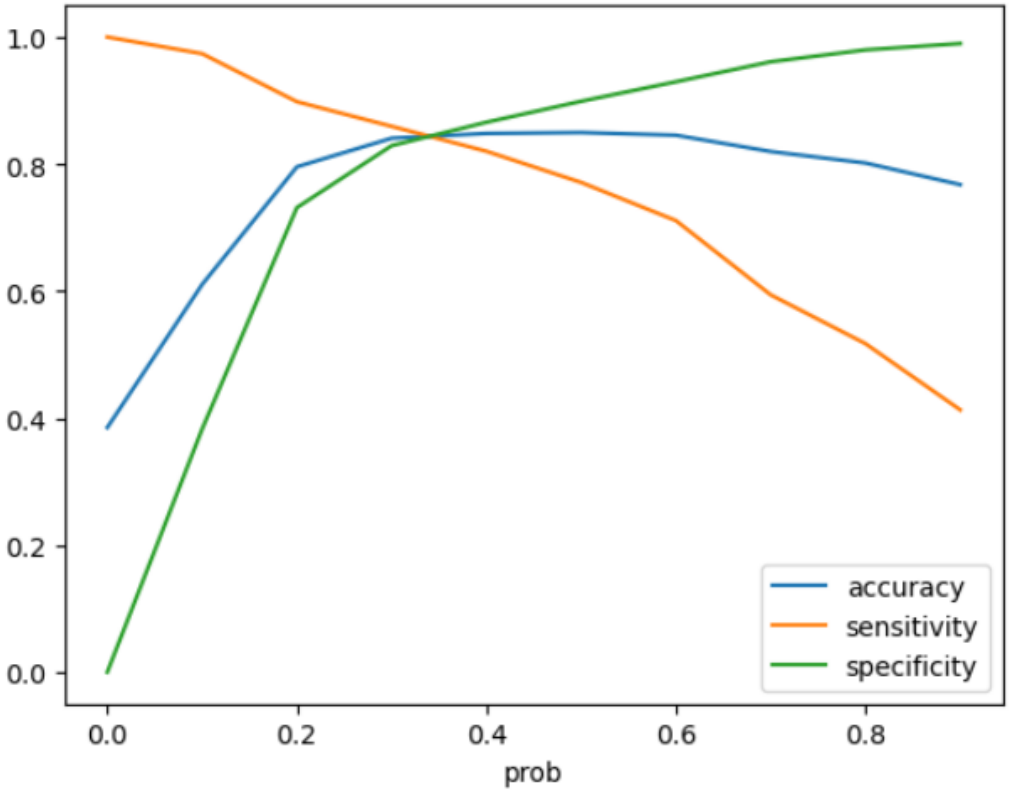
ROC Curve



- ROC curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

- The ROC Curve should be a value close to 1. We are getting a good value of 0.91 indicating a good predictive model

Model Evaluation – Accuracy, Sensitivity, Specificity on Train data set



- From the curve above, 0.35 is the optimum point to take it as a cutoff probability

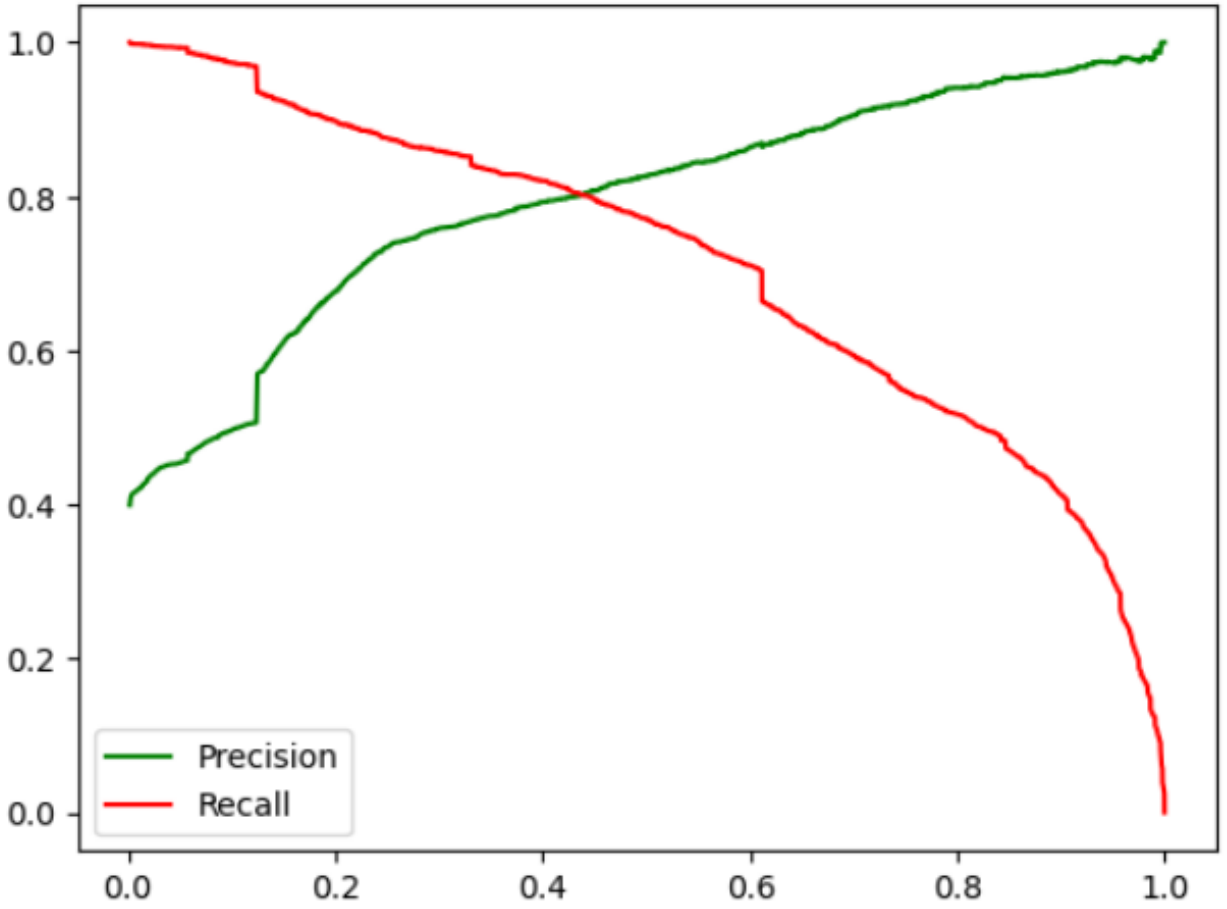
		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Confusion Matrix

3311	594
406	2040

Accuracy : 84.2%
Sensitivity : 83.4%
Specificity : 84.7%
False Positive rate : 15.2%
Positive Predictive Value : 77.4%
Negative Predictive Value : 89.1%

Model Evaluation – Precision and Recall on Train data set



Precision : 82.7%

Recall : 77.1%

Model Evaluation – Accuracy, Sensitivity, Specificity on Test data set

Confusion Matrix

3311	594
406	2040

Accuracy : 83.5%
Sensitivity : 80.6%
Specificity : 85.1%
False Positive rate : 14.9%
Positive Predictive Value : 75.5%
Negative Predictive Value : 88.5%

Variables Impacting Conversion Rate

- ❖ Lead Sources Welingak Websites
- ❖ Lead Sources_Reference
- ❖ Last Notable Activity_SMS Sent
- ❖ What is your current occupation_Working Professional
- ❖ Last Notable Activity_Others
- ❖ Total Time Spent on Website
- ❖ Last Activity_Others
- ❖ Lead Origin_Landing Page Submission
- ❖ Specialization_Others
- ❖ Do Not Email

Conclusion

- We have checked Sensitivity, Specificity, Precision and Recall Metrics. We have considered optimal cut off based on Sensitivity, Specificity for calculating final prediction
- The lead score calculated shows the conversion rate on the final predicted model is 83.4%(Train set) and 80.6%(Test set)
- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted

Thank You