# Lead Scoring Case Study

By:
Nandish Jani
Nayanshi Sahu

# PROBLEM STATEMENT

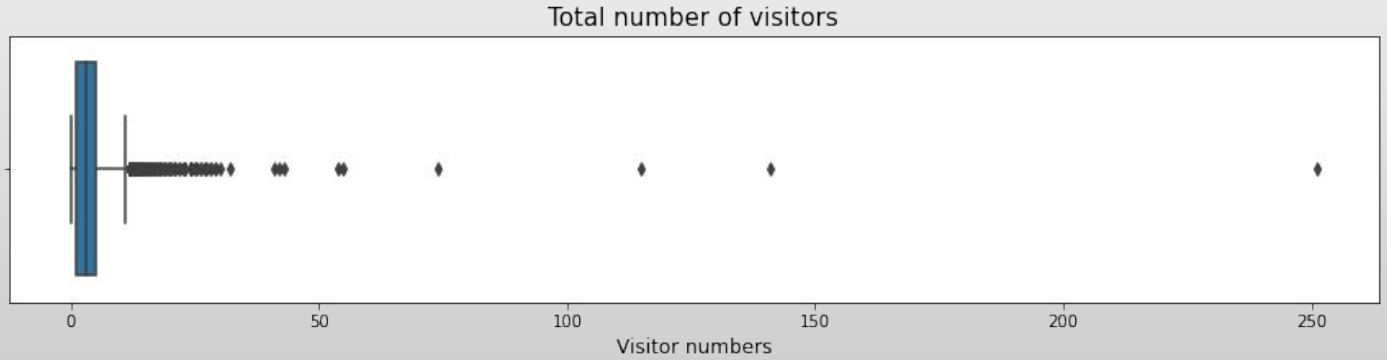
An education company named X Education sells online courses to industry professionals. Dataset is provided for different variables tracking behaviour of both converted and unconverted leads. We have to build a model to meet the CEO's 80% conversion rate expectation.

The goal of the case study is to 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. The target lead conversion rate is expected to be around 80%.

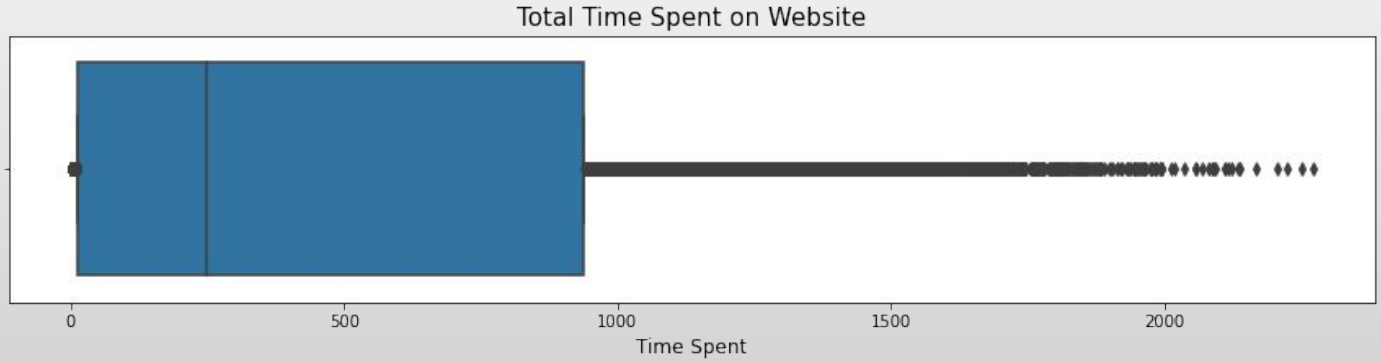
# STEPS INVOLVED IN MODEL BUILDING

1. Data preparation of numerical and categorical variables, cleaning of null values and EDA to understand outliers and various comparative distributions.
2. Data Train Test Split and Feature Scaling using MinMaxScaler
3. Correlation Analysis the numerical data.
4. Logistic Regression Model Building.
5. Different metrics, ROC curve, Optimal Cutoff evaluation.
6. Precision and Recall calculation and analysis.
7. Test set prediction, analysis, and comparison of its metrics with train data.

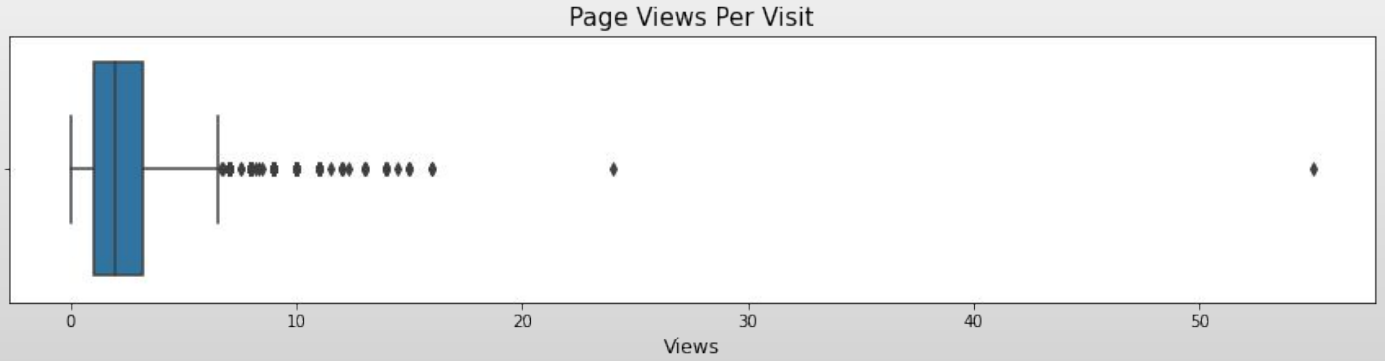
# Outlier Analysis with Boxplots



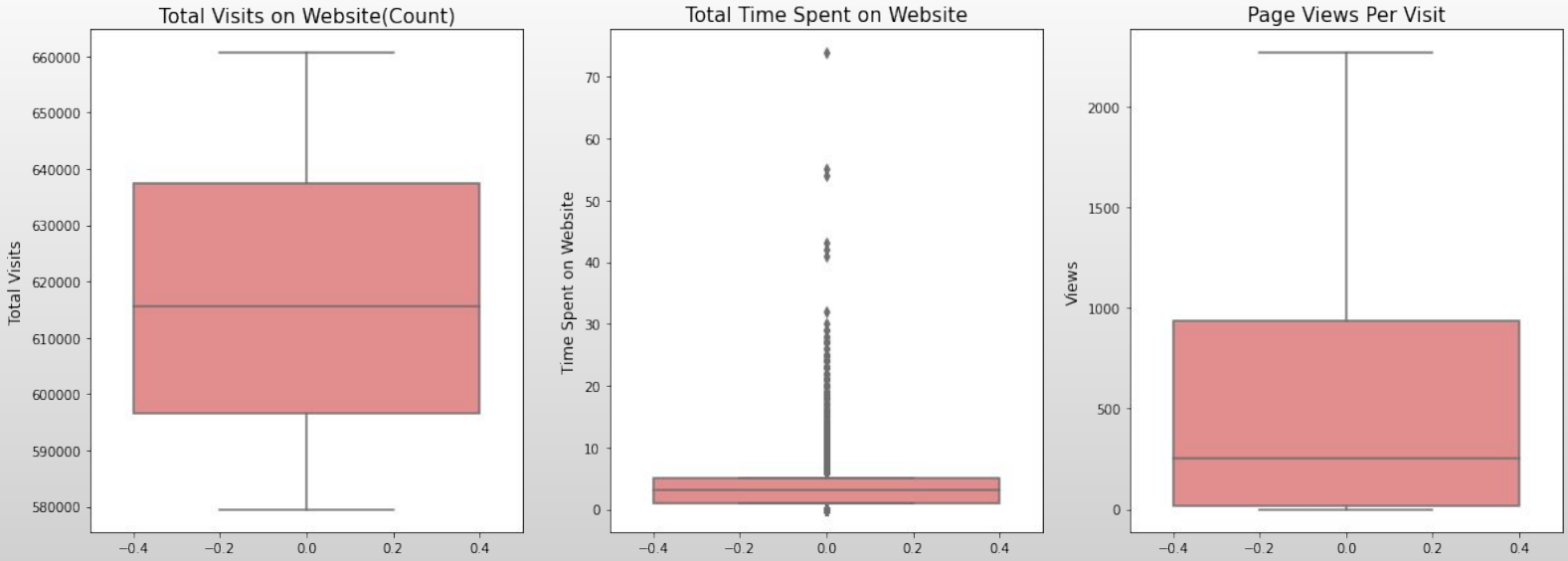

Cases of extreme outliers are clearly visible beyond 100.




No extreme or isolated outliers in "Total Time Spent on Website" variable

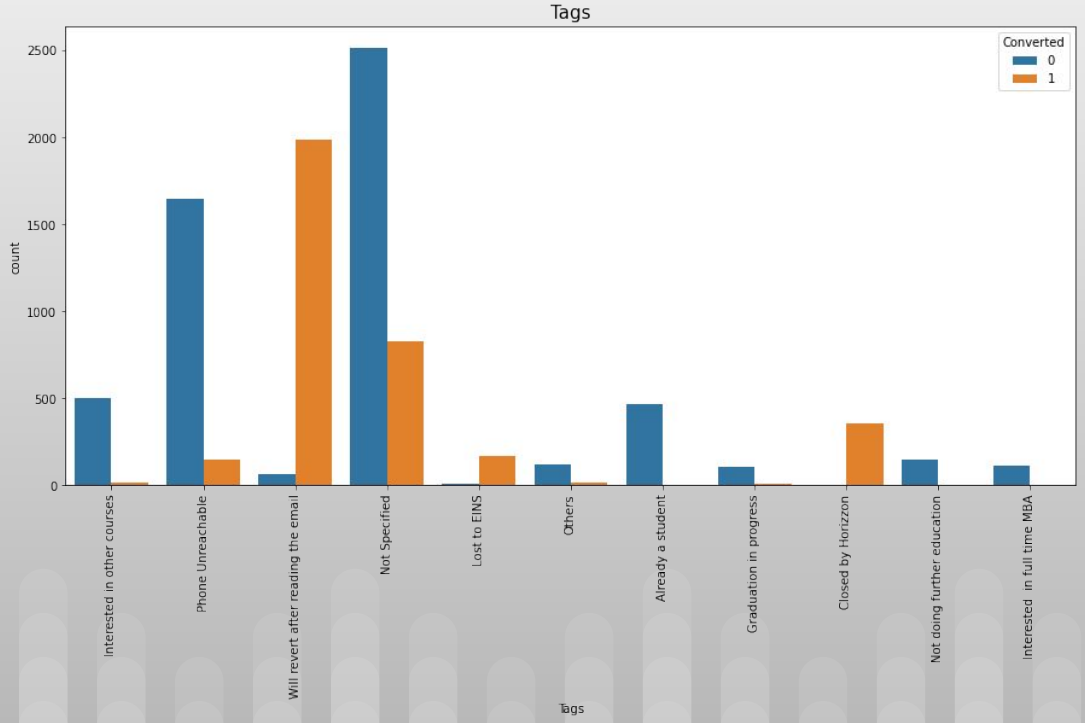



Cases of extreme outliers are clearly visible beyond 20.

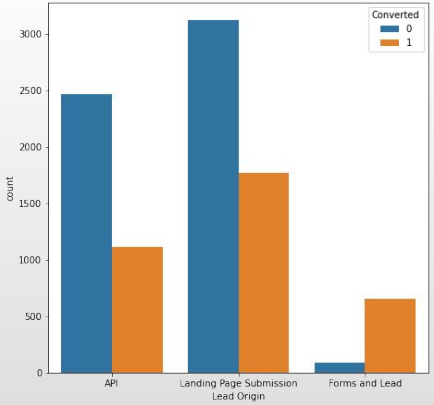


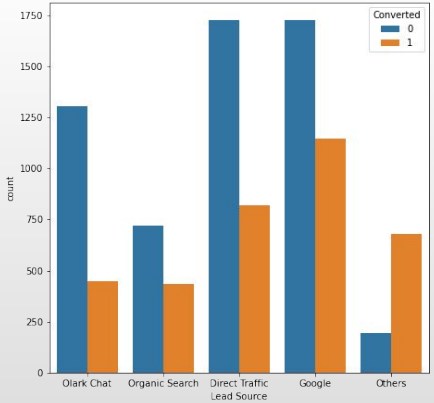

All Outliers are taken care of as we can see on graph. In the variable 'Total Time Spent on Website' we have kept the max at 70 because this value is close to values 50 and 60..

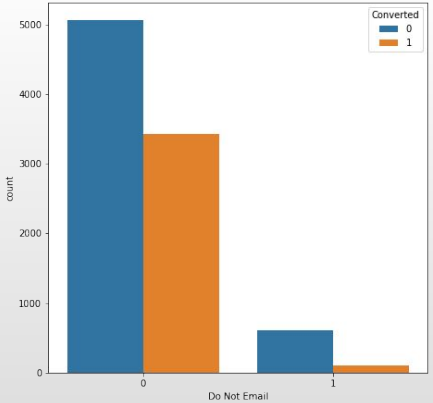
# Exploratory Data Analysis

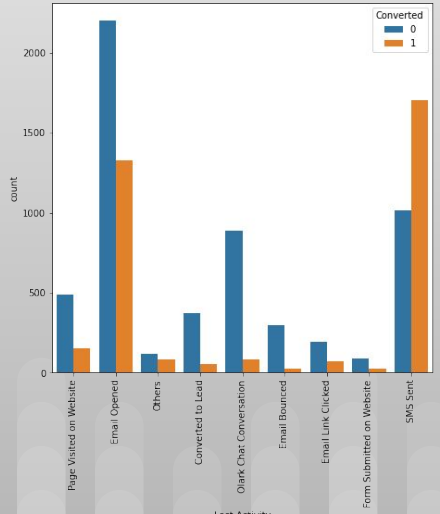


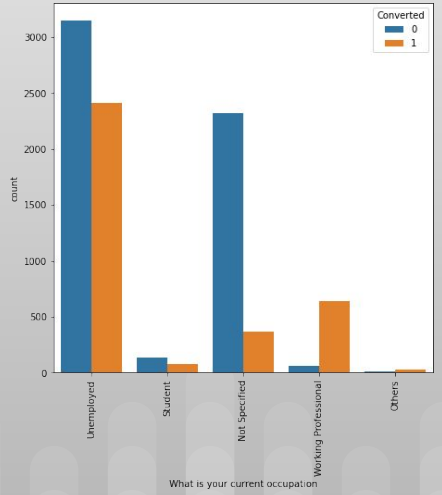

- 'Revert after reading email' and 'Closed by Horizzon' are among the best sources of lead conversion.
- 'Not specified' is also the second highest in count.

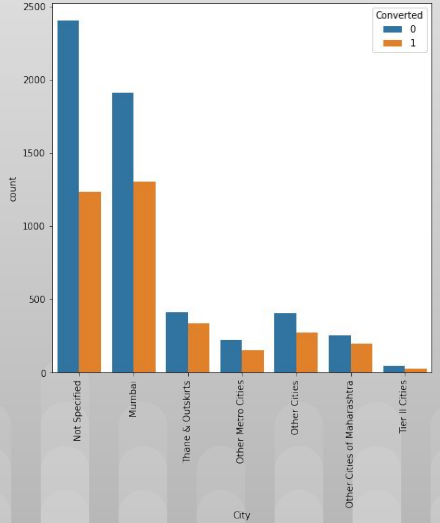


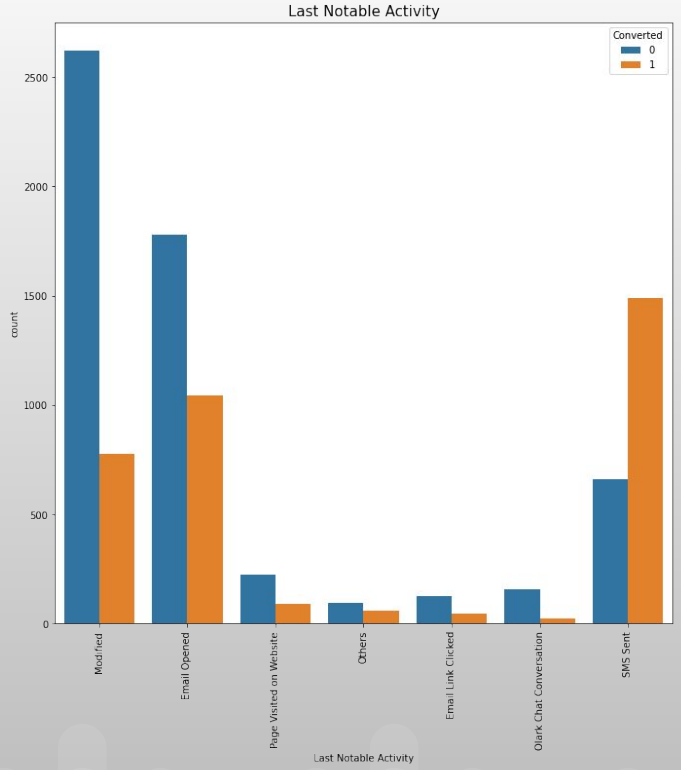



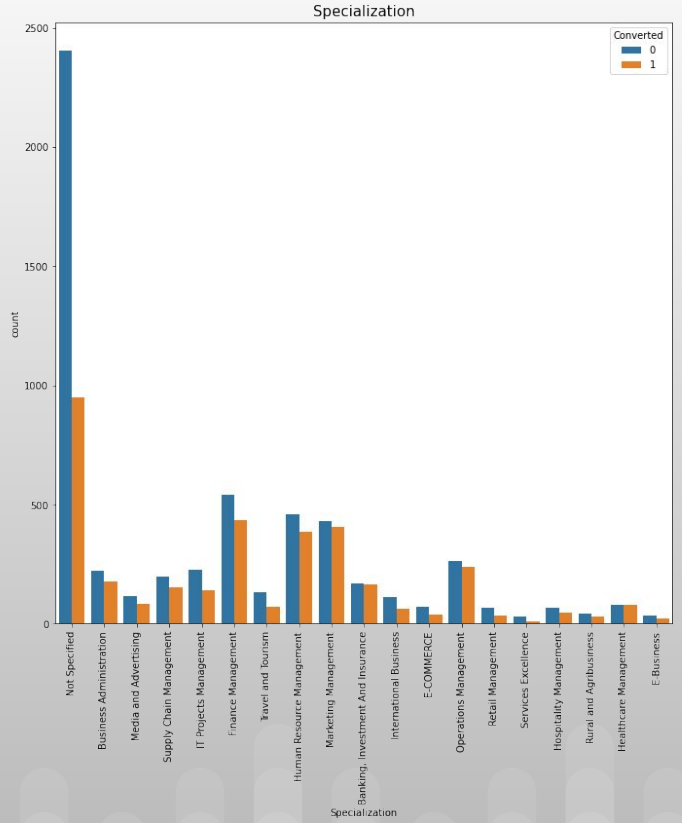




Most leads are converted from:

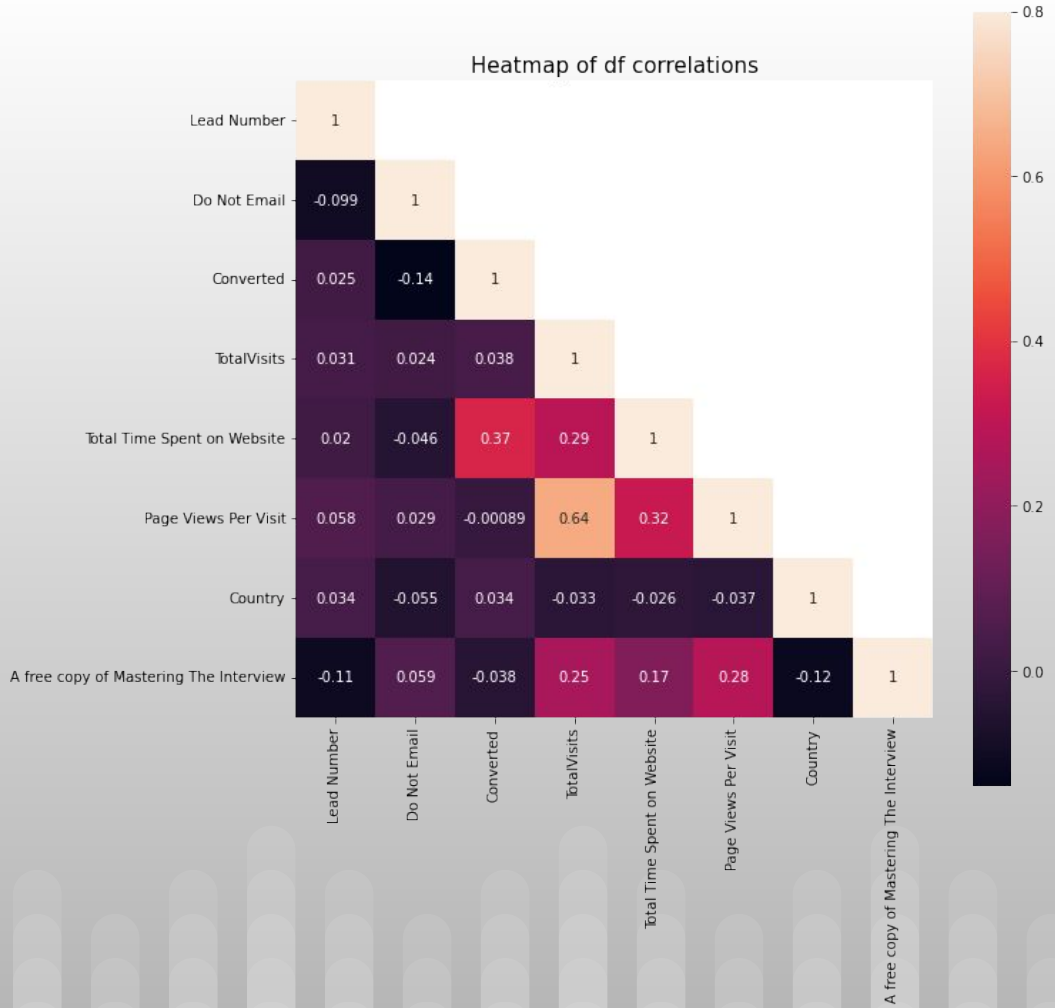
- Leads who are currently unemployed or are working professionals.
- Leads who live in Mumbai and other cities as well..
- When leads took some actions like; 'SMS sent' to any company executive or 'Email Opened', they lead to higher conversion.
- 'Landing page submission' and 'Google' search have shown the maximum lead conversion count.



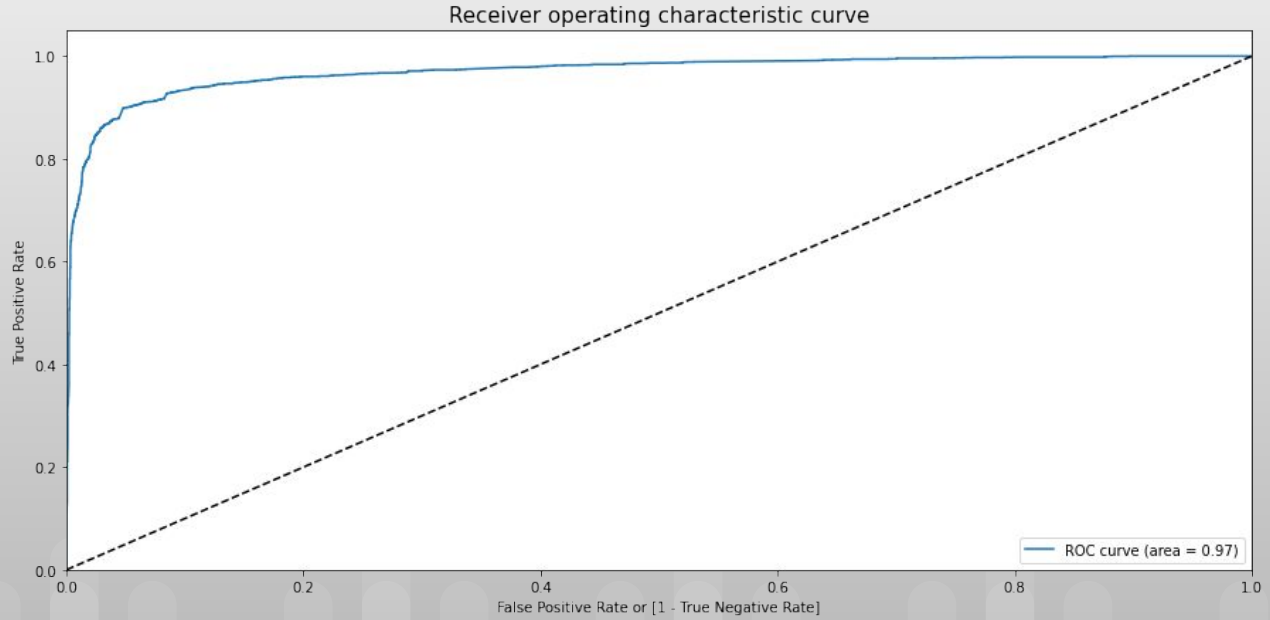

Most leads are converted from the following:

- 'SMS Sent' has the best converted to not-converted ratio.
- Management Specializations like Finance, HR, Marketing, and Operations, have very high lead conversion counts, they can be our focus area.
- Leads who had not mentioned their specializations have the highest conversion count.



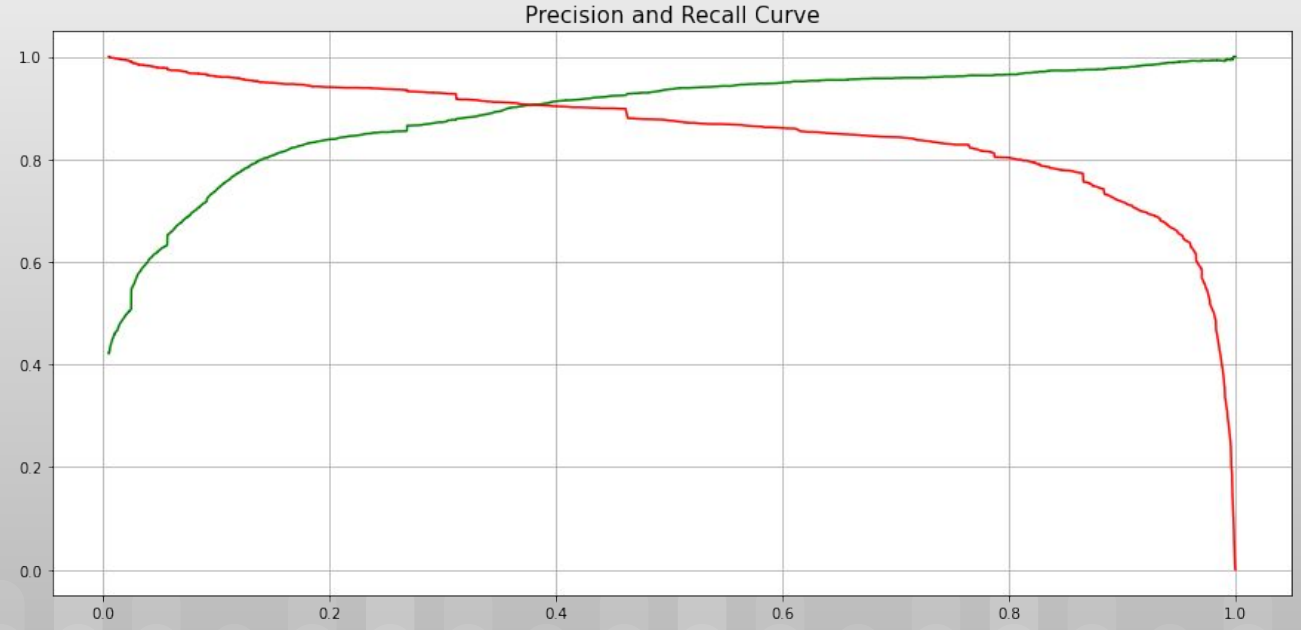

After looking into scatter plot and heatmap we have seen that the variables "TotalVisits","Total time spent on website" and "Page Views per visit" are highly correlated. However we will be dropping on the basis of correlation among dummies later on and not here.

# Receiver Operating Characteristic Curve




The area under the curve is 0.97 which is near to 1 so we can say that the model is working fine.

# Precision and Recall Curves




As per the graph, the intersection point between the two is at 0.38 probability.

# CONCLUSION

| | TRAIN | TEST |
|---|---|---|
| Accuracy | 0.92 | 0.92 |
| Sensitivity | 0.93 | 0.90 |
| Specificity | 0.95 | 0.93 |

- All the metric are fairly high with very little variance between train and test performance. This indicates that the model is working well for the dataset and is not overfitted.
- X Education can achieve its CEO's target of 80% lead conversion rate with this model.
- Top 5 features which matter most for lead conversion as per the model along with their coefficient factors are:
  1. Tags_Lost to EINS 5.963161
  2. Tags_Closed by Horizzon 5.907348
  3. Tags_Will revert after reading the email 3.989566
  4. Total Time Spent on Website 3.651781
  5. Last Notable Activity_SMS Sent 3.019506