Lead Scoring Case Study

GROUP MEMBERS:

SUMIT KOSHTI

VAISHNAVI PANDE

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- When people fill up their forms/enrol with email address or phone number, they are classified as leads.
- The typical lead conversion rate at X education is around 30%.
- The company wishes to identify the most potential leads i.e ‘Hot Leads’ by making the lead conversion rate go up by focusing more on communicating with the potential leads rather than making calls to everyone.

Objective

1. To build a logistic regression model to assign a lead score between 0 to 100 to each of the leads which can be used by the company to target potential leads.
2. There are some more problems presented by the company which model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Steps Followed in Case Study

1. Read and Understand the Data
2. Cleaning of the data
3. Data Visualization
4. Preparation of the data
5. Model Building
6. Model Evaluation
7. Model Predictions on test data set
8. Summary

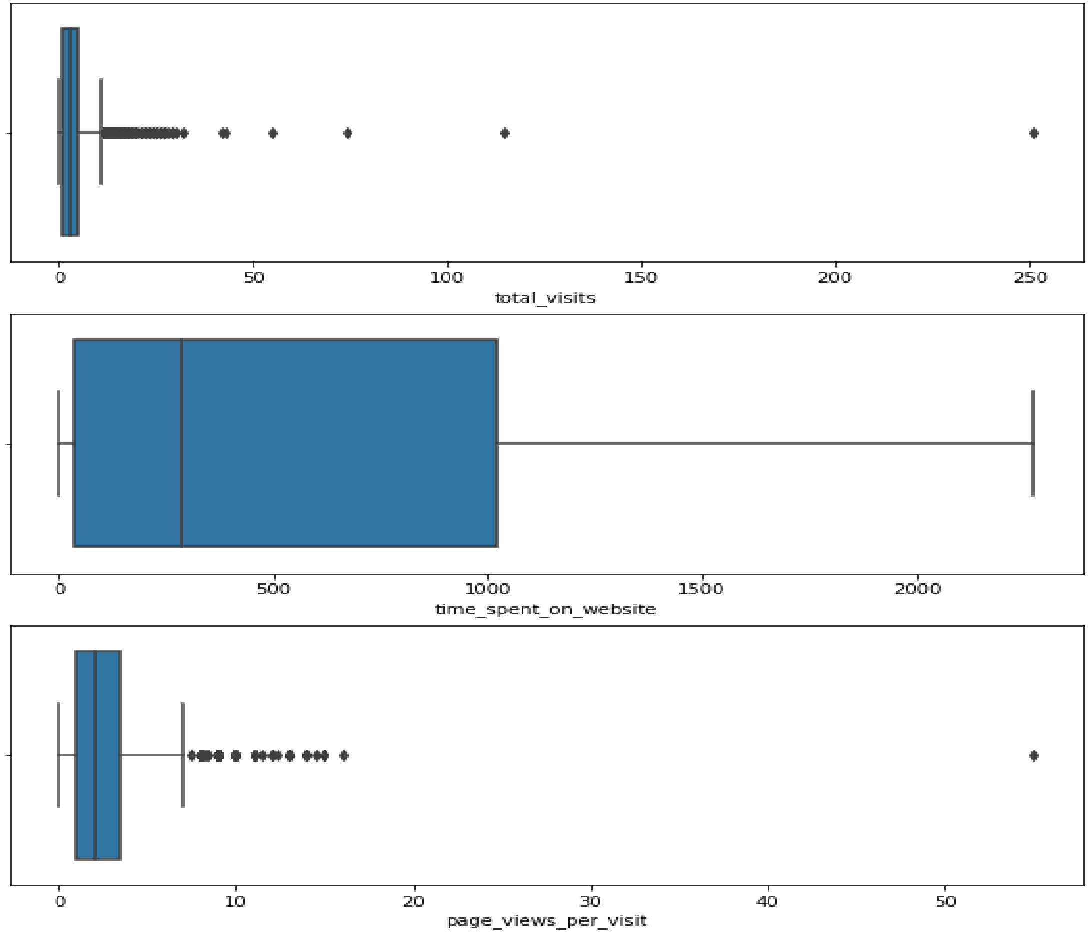
Read and Understand The Data

- Insights obtained by importing the data set and by doing some sanity checks:
  - Data set contains total **37** columns.
  - Data set has **no duplicate** values.
  - Unique Identifiers are Prospect ID and Lead Number.
  - Many columns **has null values**.
- We have dropped the unwanted columns like City and Country.
- Total three Columns from the data set contains a "Select" level.
- Variables "Lead Profile" & "How did you here about X Education" have a lot of rows which have a "Select" value which is no use to analysis. So these variables are dropped.

DATA CLEANING

1. We have dropped several variables which are having more than 3000 null values.
2. We have dropped those variables which are having one value majorly.
3. We have Dropped null rows from the couple of variables.
4. We have renamed some of the columns for better understanding.

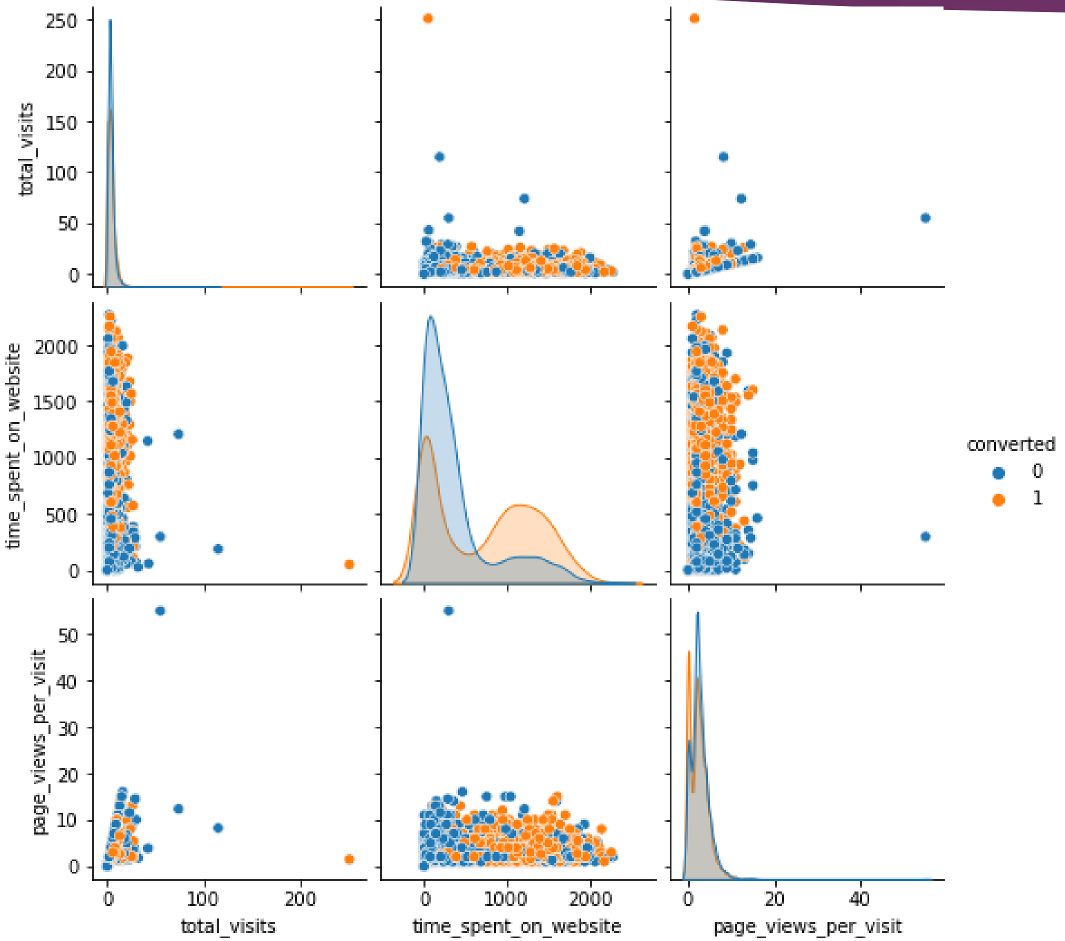
Checking The Outliers



<u>Observation</u>:

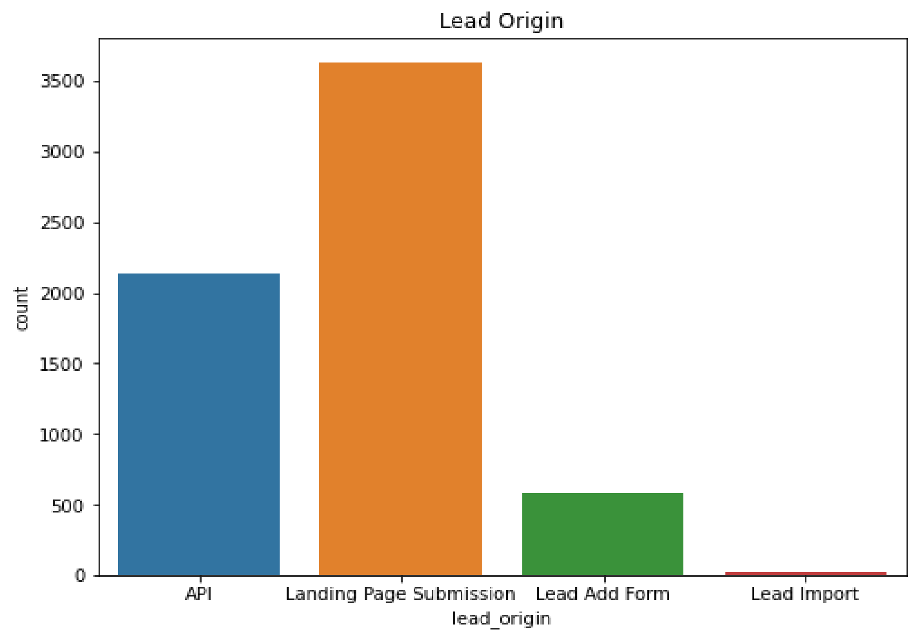
The above boxplots are showing that, columns "total_visits" and "page_views_per_visit" are have upper bound outliers and the data can be capped at 99 percentile.

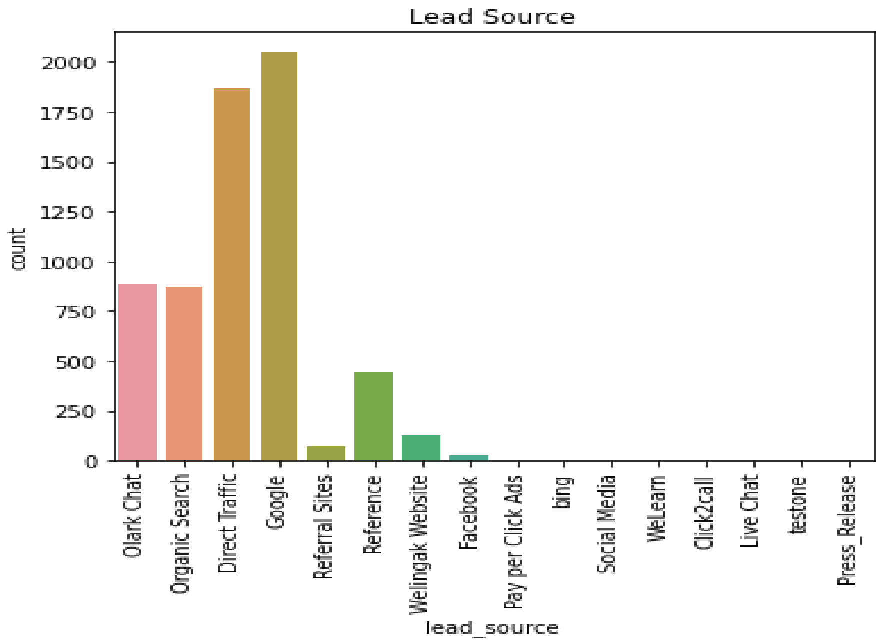
DATA VISUALIZATION



Observation for the continuous variables:

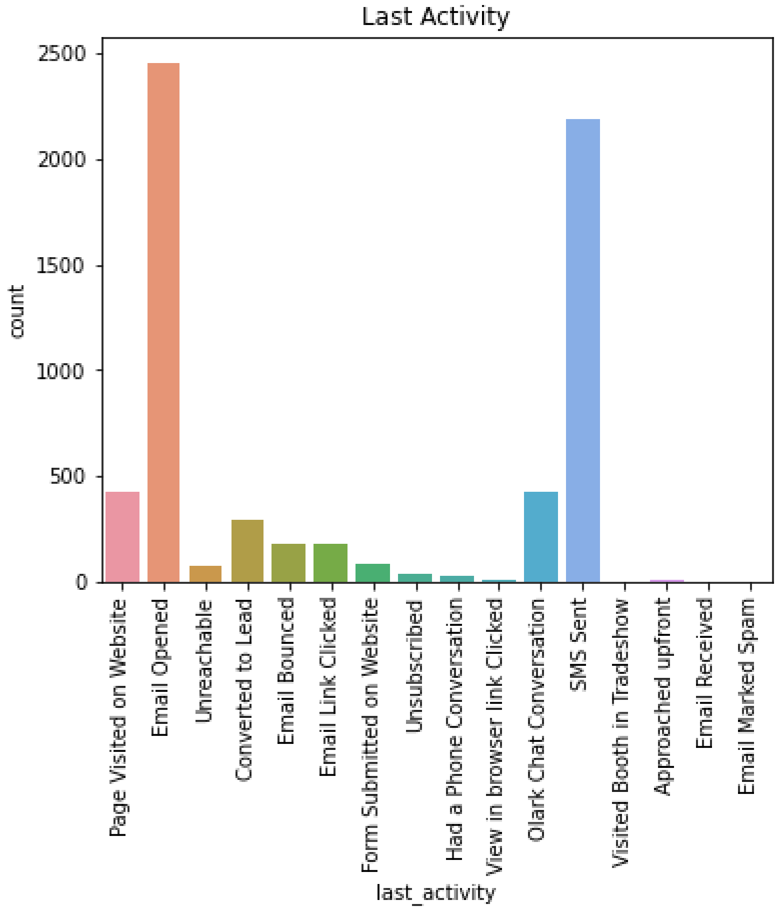
At Some extent there is linear relationship between total_visits & page_views_per_visit variables.

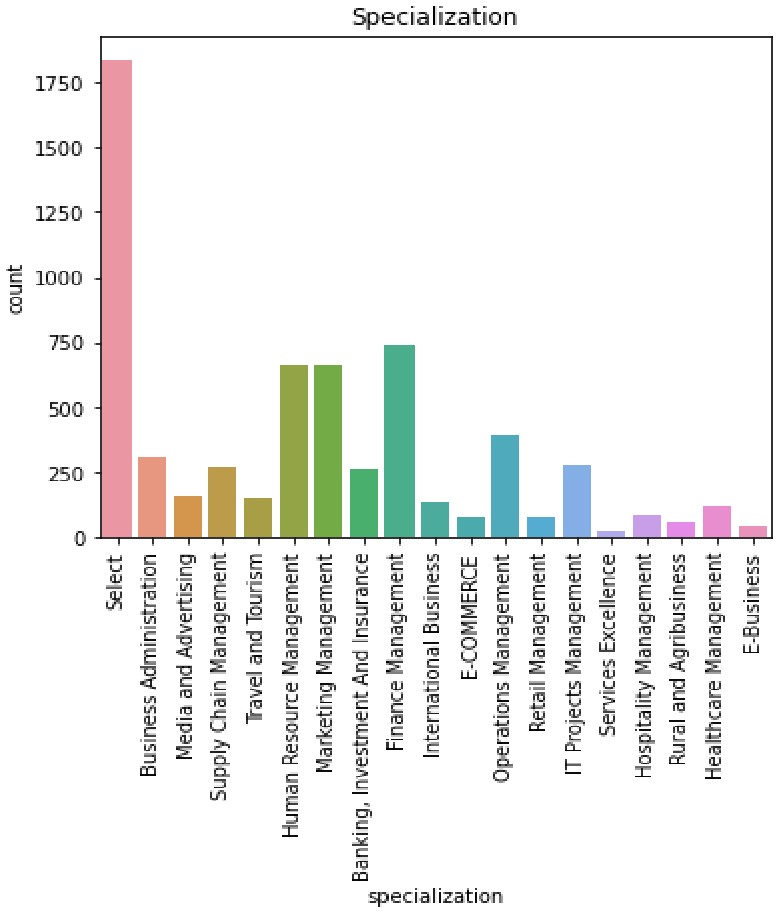




<u>Observation for categorical variables</u>:

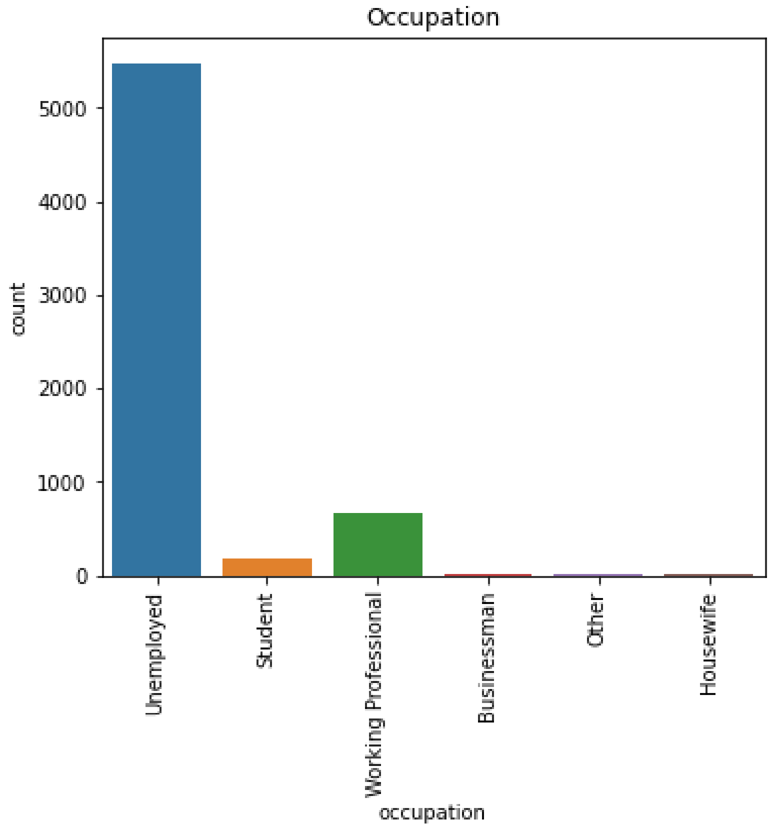
1. Highest lead generated from "Landing Page Submission" Origin.
2. Highest leads generated from sources like "Google" and "Direct Traffic".



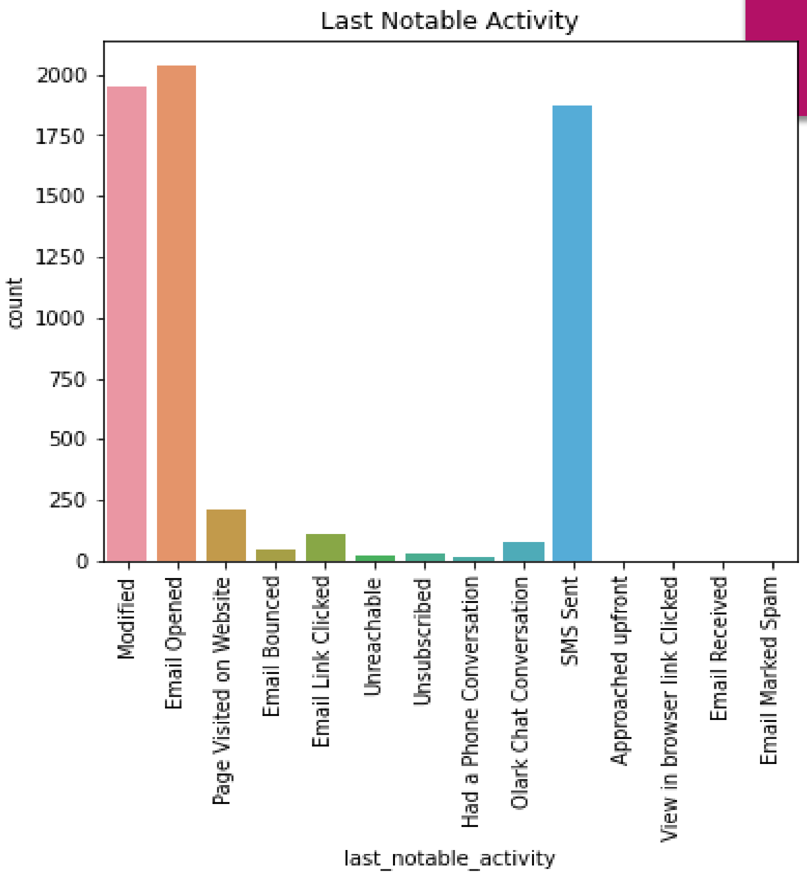


1. Highest leads generated from "Email Opened" and "SMS Sent" activity.

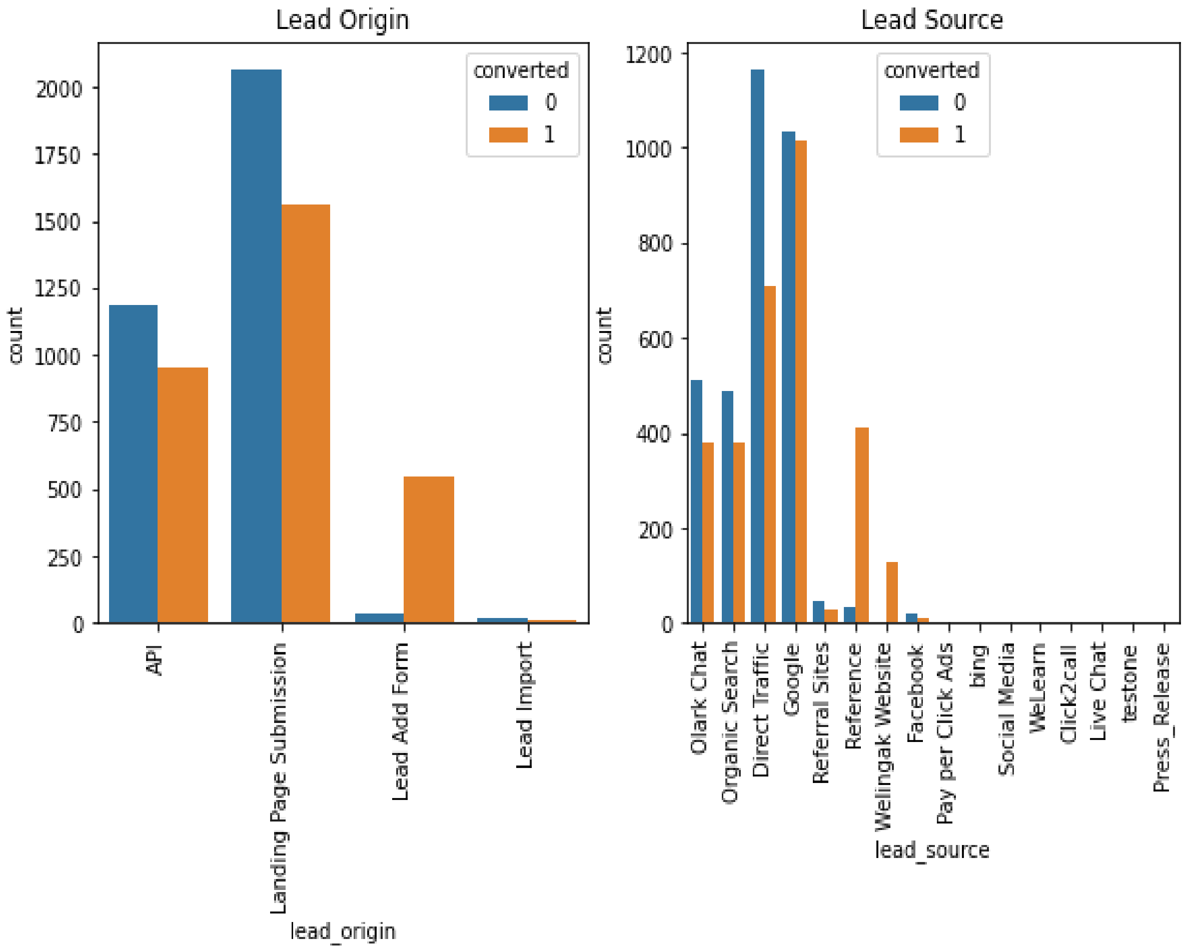
2. Highest lead generated from "Management" Specialization.



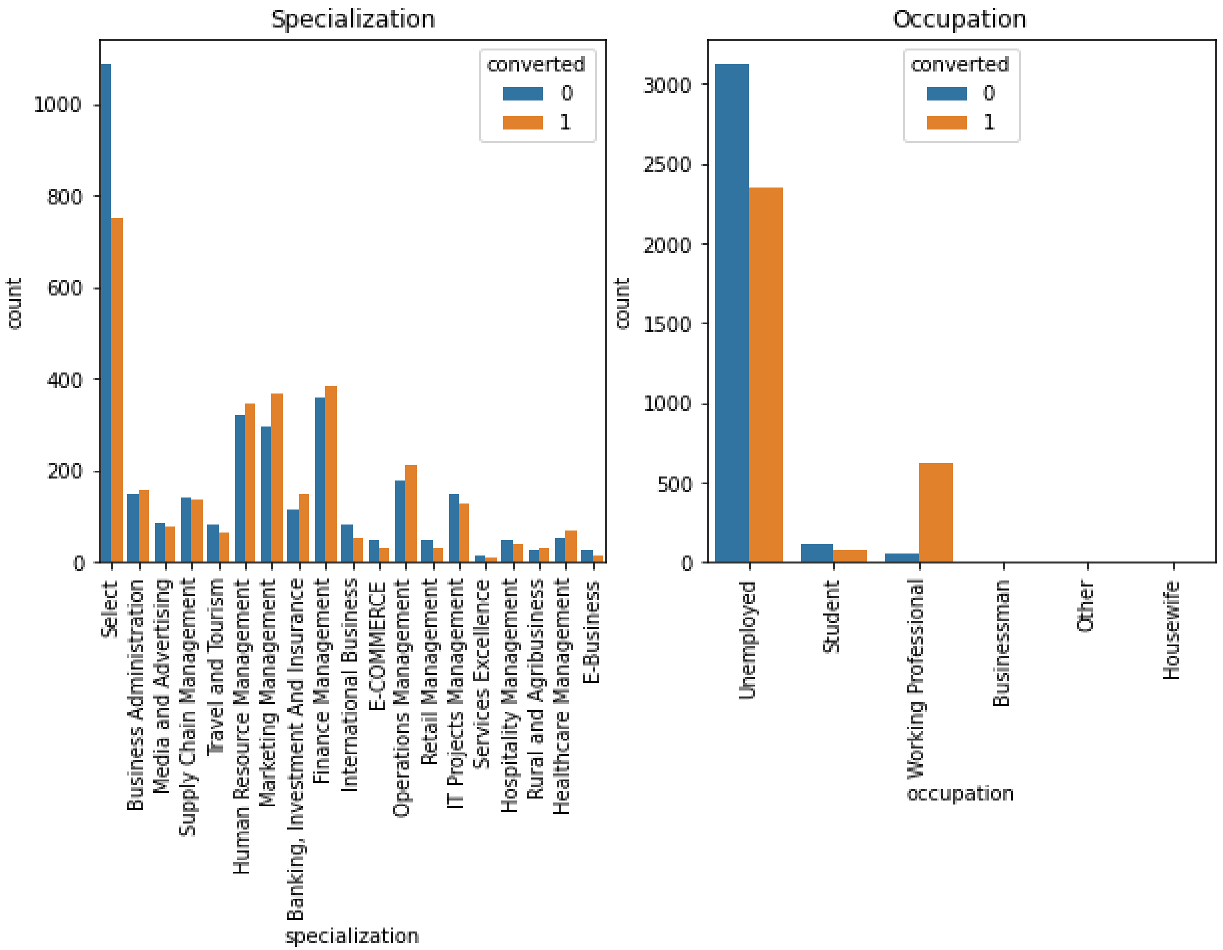
1. Highest lead generated from "Unemployed people" and "Working Professionals".



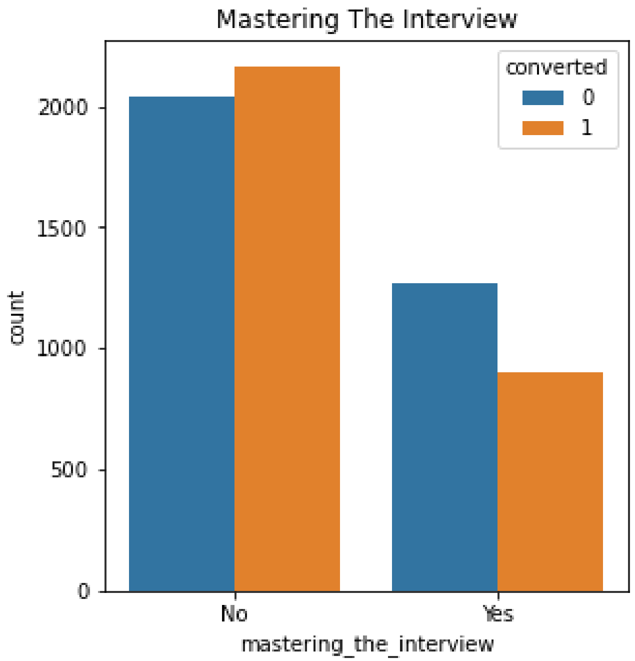
2. Highest lead generated from "Email Opened" and "SMS Sent" activity.

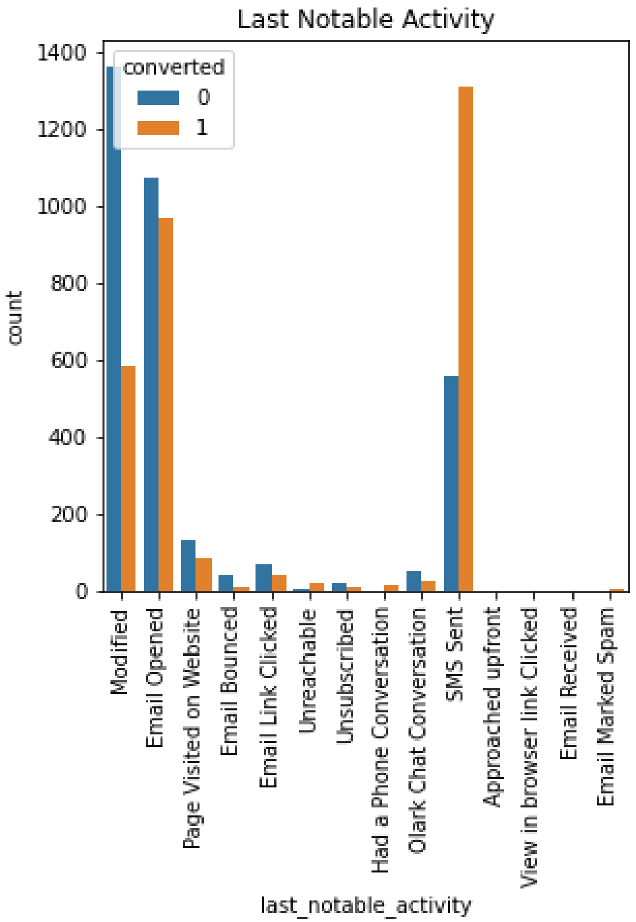


1.The conversion rate is high in "Lead Add From" Origin.
2.The conversion rate is high in "Reference" Source.

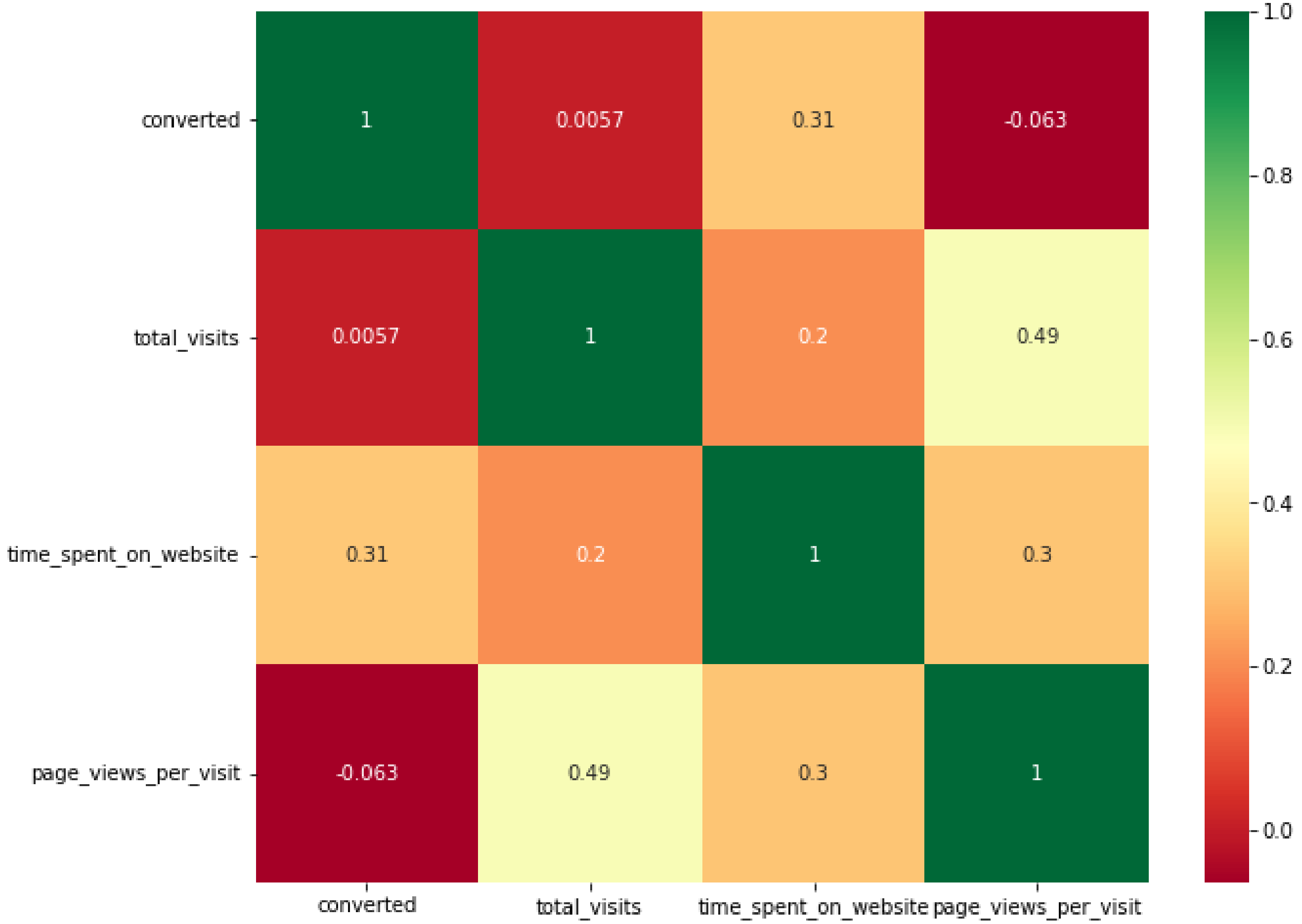


1.The conversion rate is high in all "Management" specializations.
2.The conversion rate is high with "Working Professional occupation.





1. The conversion rate is high in "SMS Sent" activity.

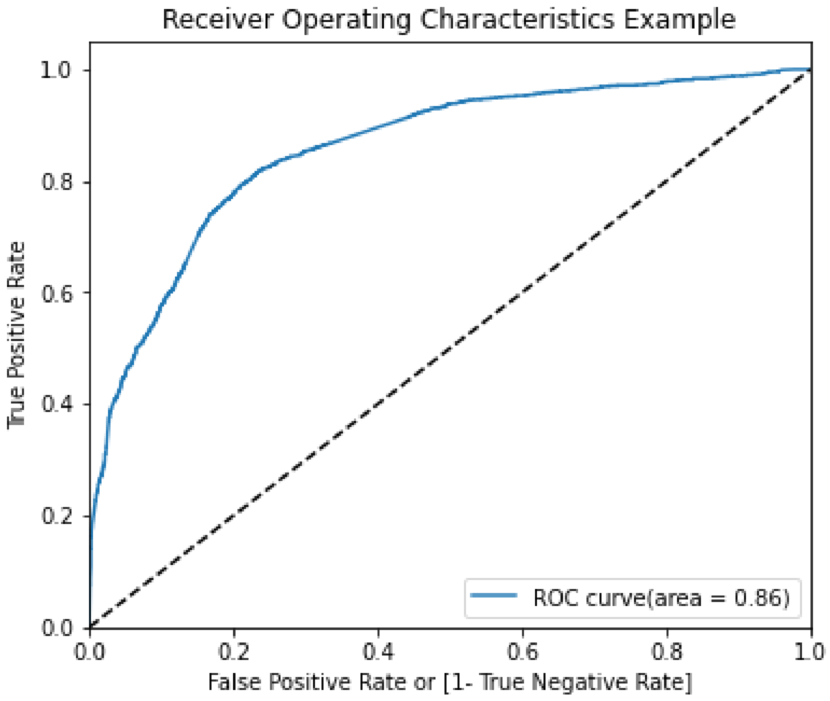


The above heatmap is showing that there are no highly correlated variables present in the dataset.

Final Model Insights

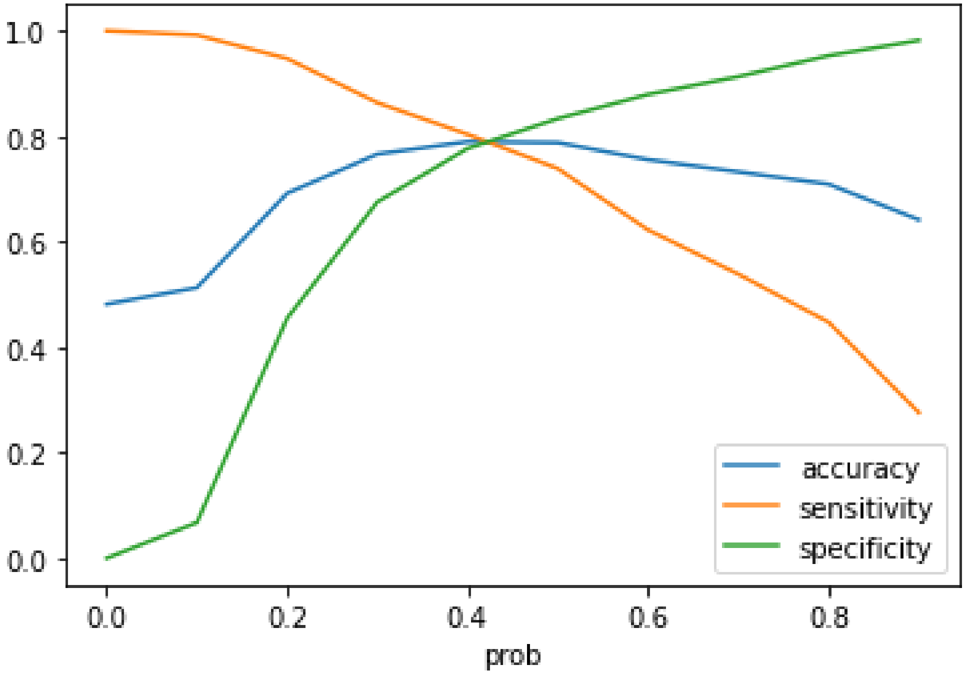
- The Model 5 has significant p-values <0.05
- The multicollinearity between the predictor variables is also low i.e VIF value <5.
- Model 5 has all required values under significant range.
- We took this model 5 for further analysis as final optimum model
- The top three variables from the model are:
    - occupation_Unemployed
    - time_spent_on_website
    - total_visits

ROC For Train data



- The plot showing that the area under the curve of the ROC is 0.86 which is quite good so this model seems to be good model.

Optimal Cut-Off



The plot clearly showing that the optimal values of the three metrics lies around 0.42. So let's choose 0.42 as our cutoff now.

Model Evaluation for train data

- Confusion Matrix: The accuracy of the train data set is 79.01%
- Sensitivity of the train data set is 79.87%
- Specificity of the train data set is 79.15%

Model Evaluation for X-test data set

- Confusion Matrix: The accuracy of the test data set is 76.09%
- Sensitivity of the train data set is 83.18%
- Specificity of the train data set is 69.57%
- Precision value: 78.11%
- Recall value: 78.40%

Model Evaluation for Y-test data set

- Confusion Matrix: The accuracy of the test data set is 76.09%
- Precision value: 71.54%
- Recall value: 83.18%

Conclusion

- **Major Contributors :**

As per the model we build here, It was found that the variables that mattered the most in the potential buyers are:

1. **Total Visits**
2. **Time Spent On Website**
3. **Page Views Per Visit**

By targeting the above variables X Education company can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses. By focussing on these variables company can increase their lead conversion rate up to 80%.

Thank You!