LEAD SCORING CASE STUDY

1. SHIVANI RAUT
2. SHIWANI JAMDAGNI
3. SHIVAM BANERJEE

CONTENTS

- Problem Statement
- Problem Approach
- EDA - Univariate & Bivariate analysis
- Model Building
- Model Evaluation
- Observation & Conclusion

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. They have process of form filling on their website after which the company that individual as a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%. Now, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUISNESS OBJECTIVE

- The lead X wants us to build a model to give every lead a lead score from 0 -100 . So that they can identify the hot leads and increase their conversion rate as well. And the CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like – 1) Peak time actions required, 2) How to utilize full man power & 3) After achieving target what should be the approaches.

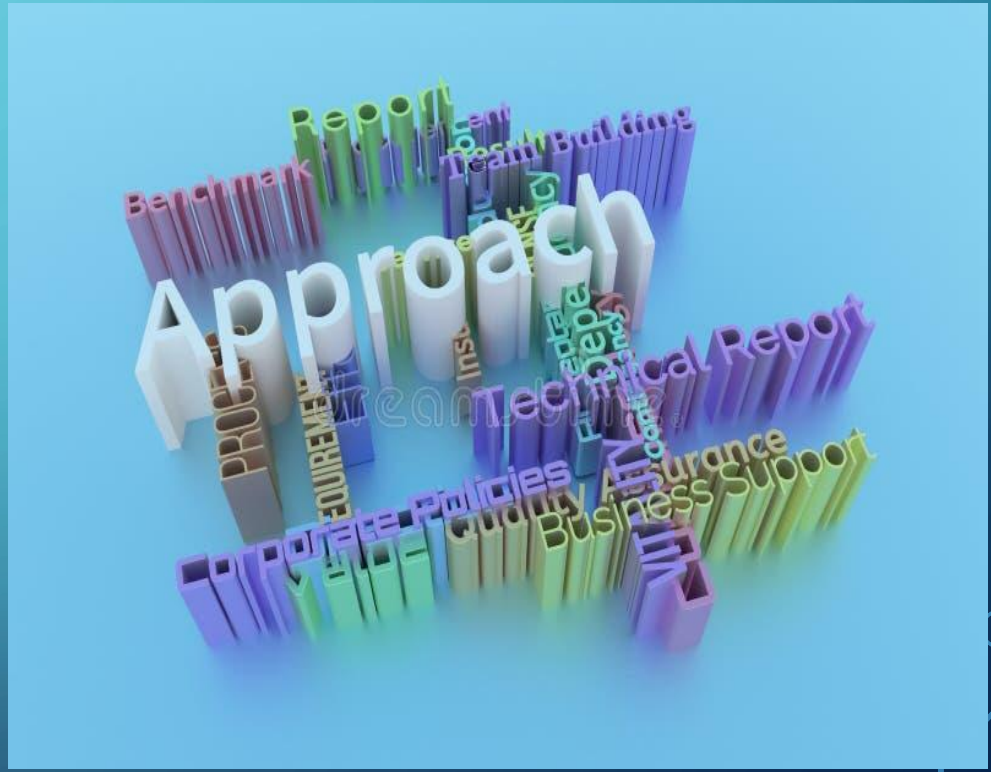
PROBLEM APPROACH

Importing data → Data cleaning and Data preparation. → EDA for figuring out most helpful attributes for conversion.

Splitting data into Train and Test dataset. → Scaling features. → Building a logistic regression model.

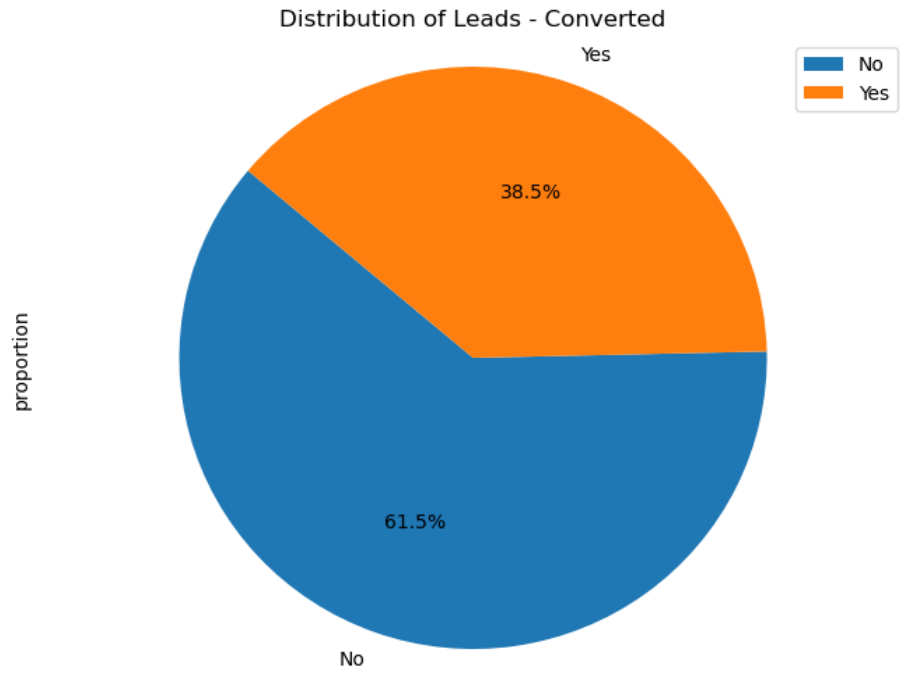
Variable Selection through RFE and manual tuning. → Train model evaluation and performance check. → Building Test model based on Train dataset final model

Test model evaluation and performance check.



EXPLORATORY DATA ANALYSIS (EDA)

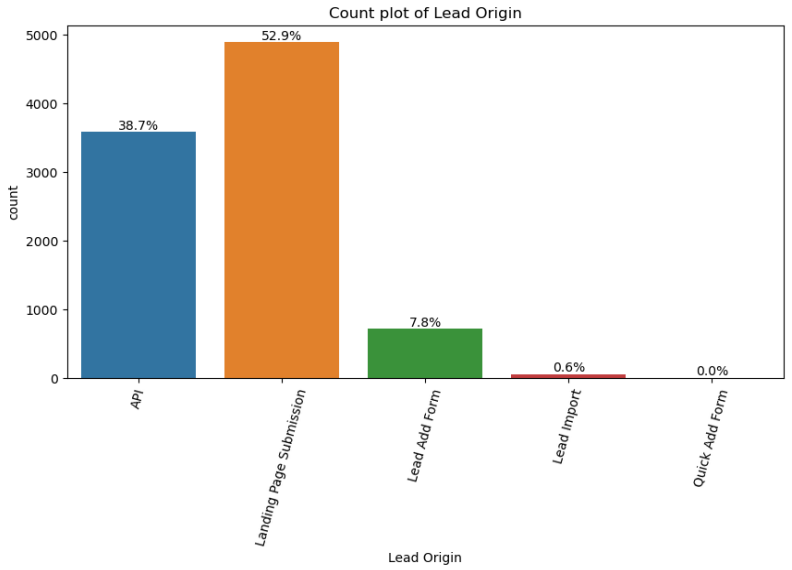
IMBALANCE RATIO



- Majority of the people don't converted which is 61.5% compared to 38.5% who converted.
- The imbalance between the leads that non-converted to converted is 1.59:1.

UNIVARIANT ANALYSIS – CATEGORICAL VARIABLES

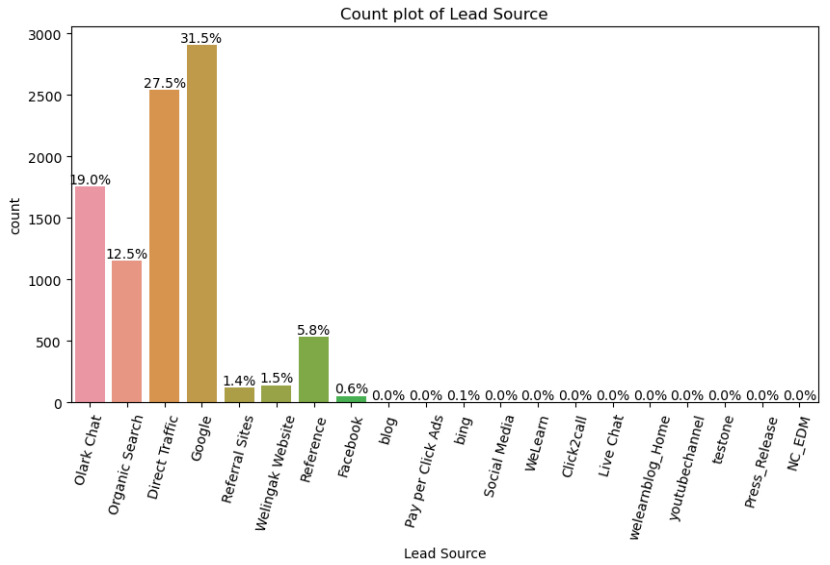
Lead Origin



The most lead origin from Landing Page Submission followed by API.

Quick Add Form generates least leads.

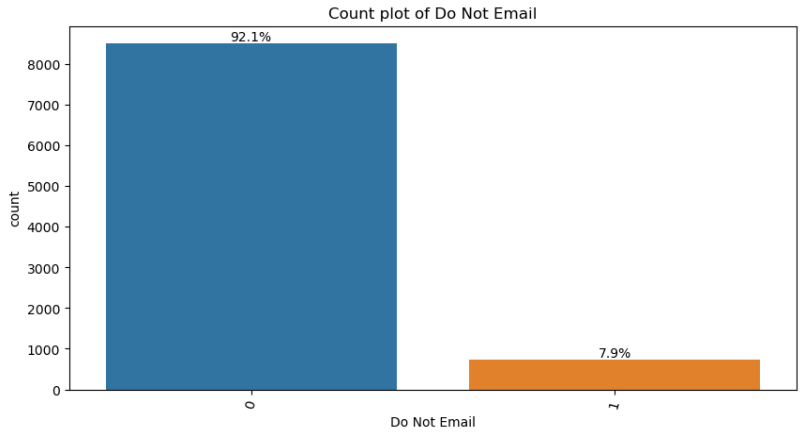
Lead Source



Google has highest rate of lead source count followed by Direct Traffic.

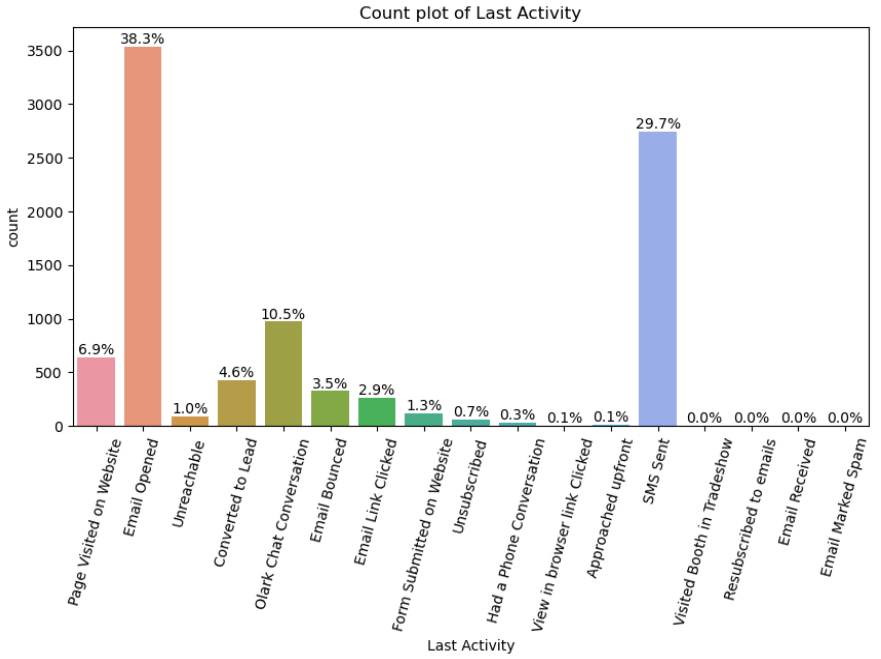
Very few leads are generate from source like blogs, Bing, live chat, etc.

Do Not Email



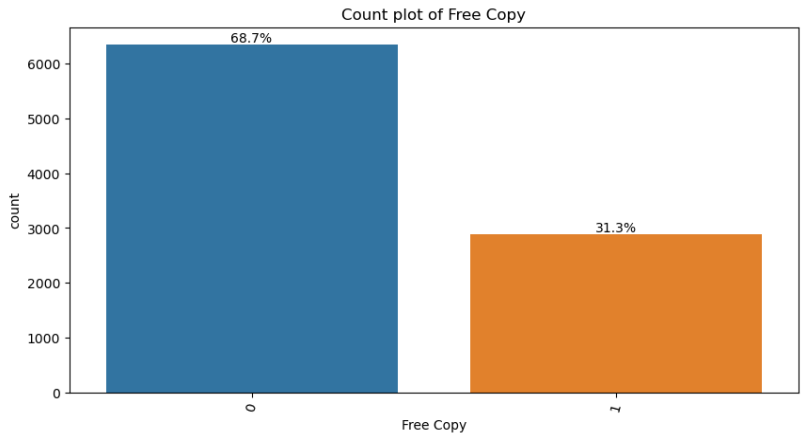
Most of the people don't want any emails regarding courses.

Last Activity



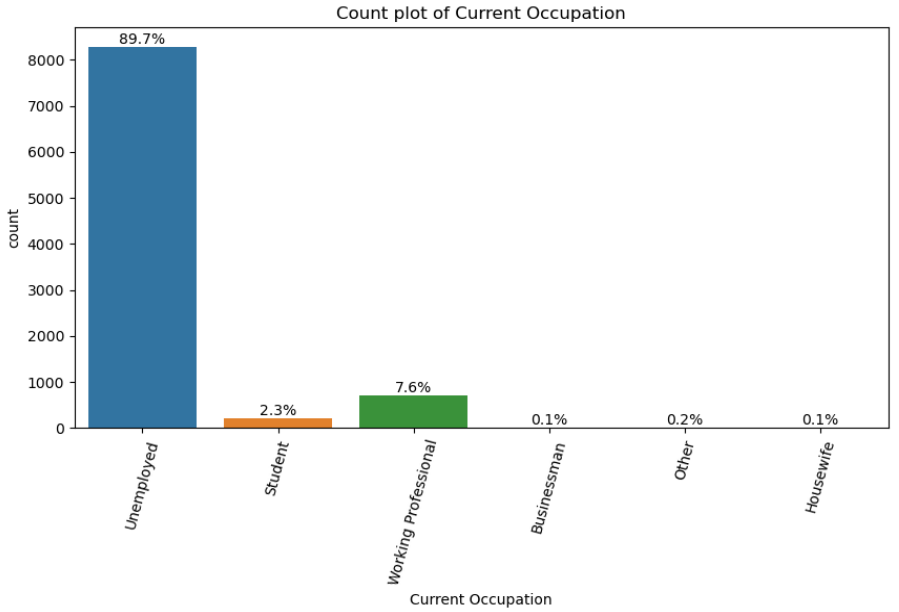
SMS is a promising method to get higher confirmed method leads and Emails also has high conversions.

A free copy of Mastering The Interview



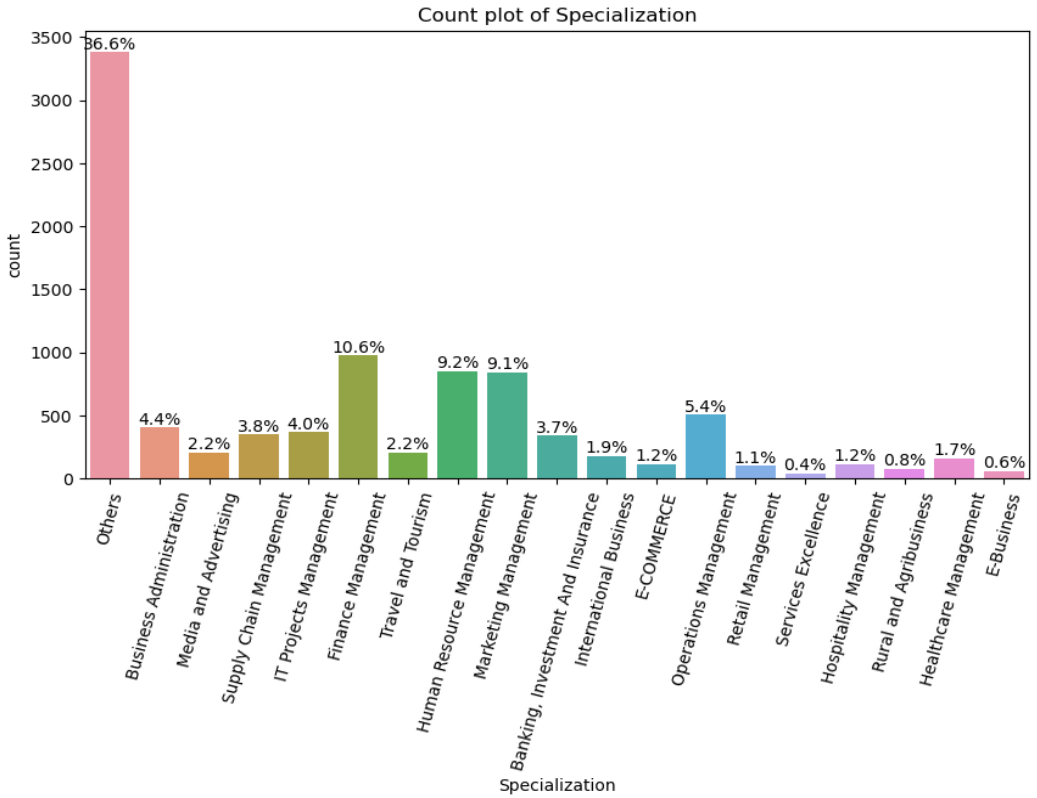
Less than 32% prefer to have a copy of the interview.

Current Occupation



Unemployed people leads are more interested in joining the course that others.

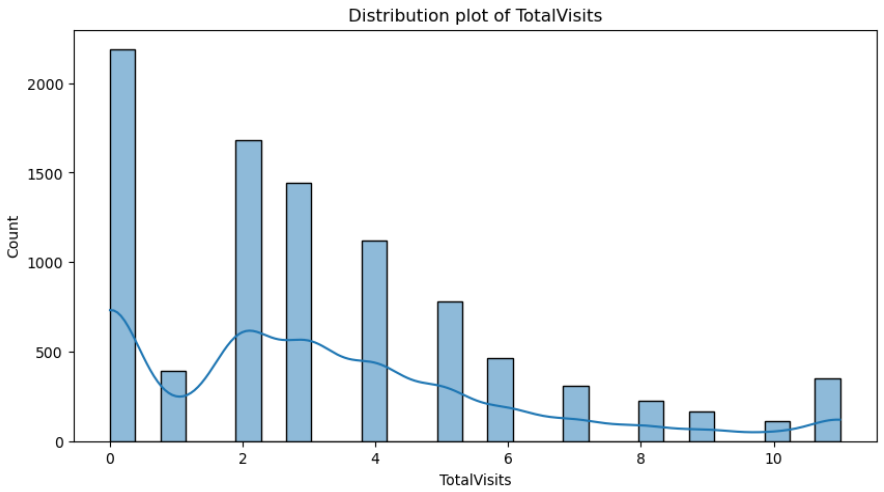
Specialization



On one hand Human Resources Management, Marketing Management, Finance Management, etc. has high count and promising leads are people from these specializations.

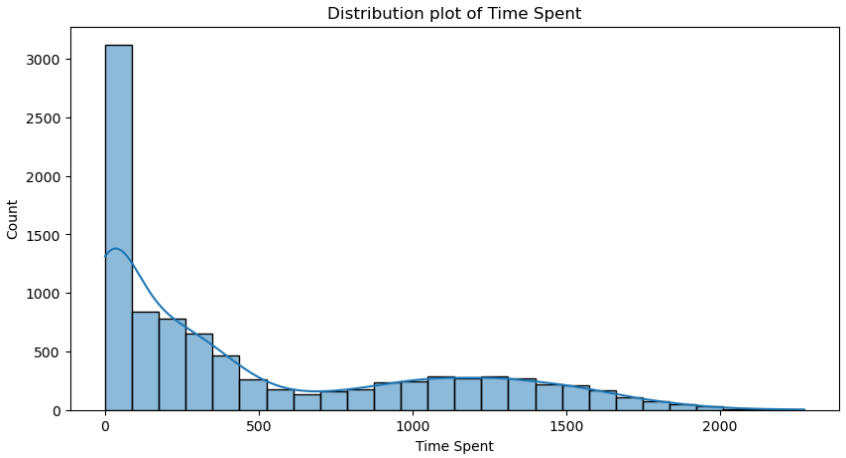
UNIVARIANT ANALYSIS – NUMERICAL VARIABLES

Total Visits



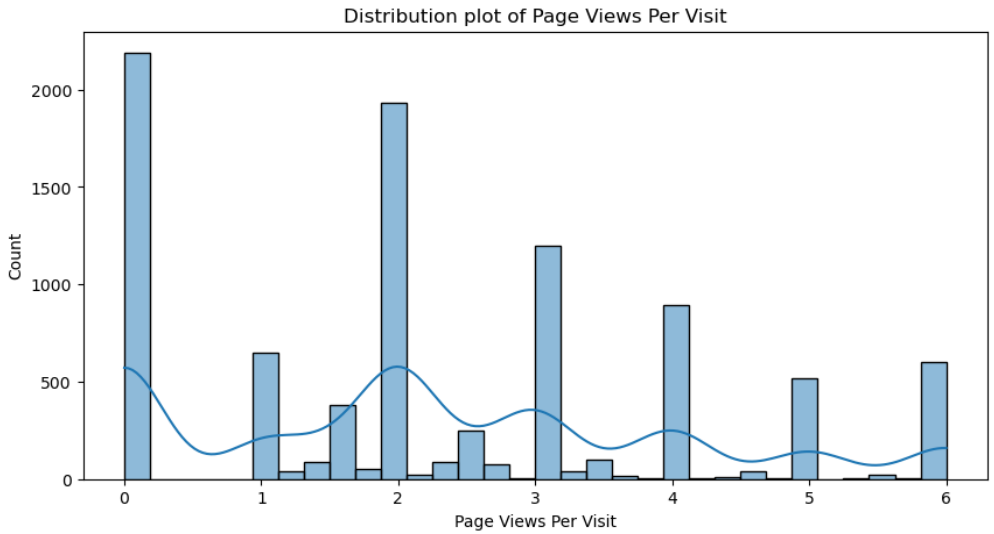
There are many people who don't visit the website or just visit it once.

Time Spent



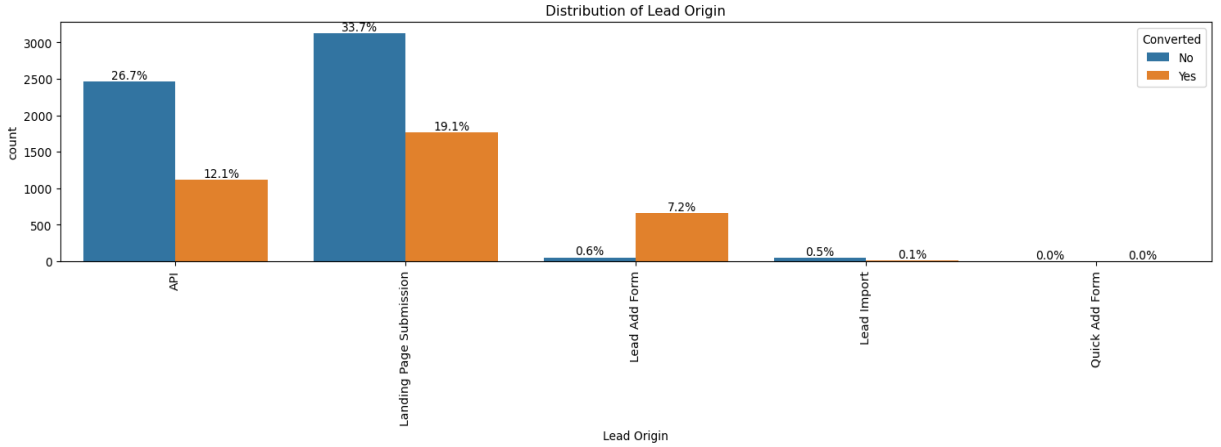
There are very few customer who spend a long time on the website.

Page Views Per Visit



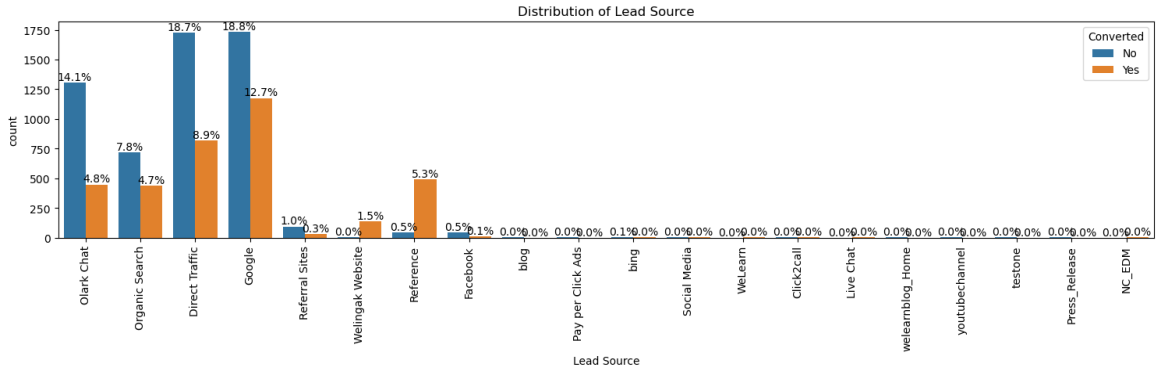
Most of the people visit 2 pages per visit before loosing their interest in the website.

BIVARIANT ANALYSIS – CATEGORICAL VARIABLES



Lead Origin

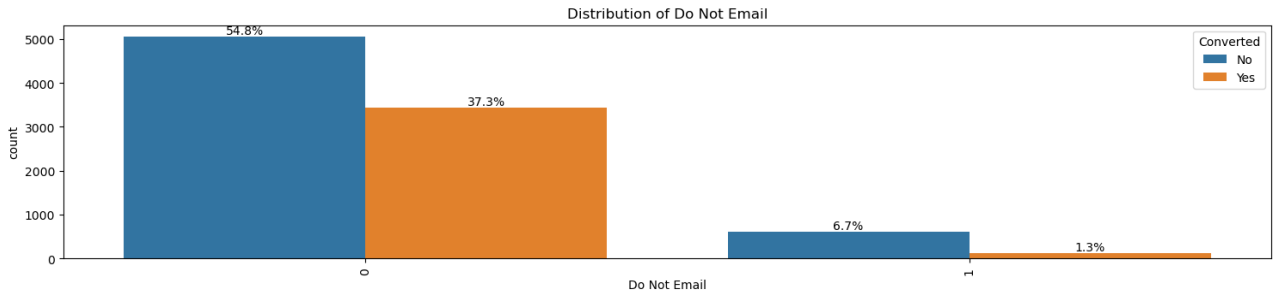
There is higher conversion rate for the leads from Lead Add Form.



Lead Source

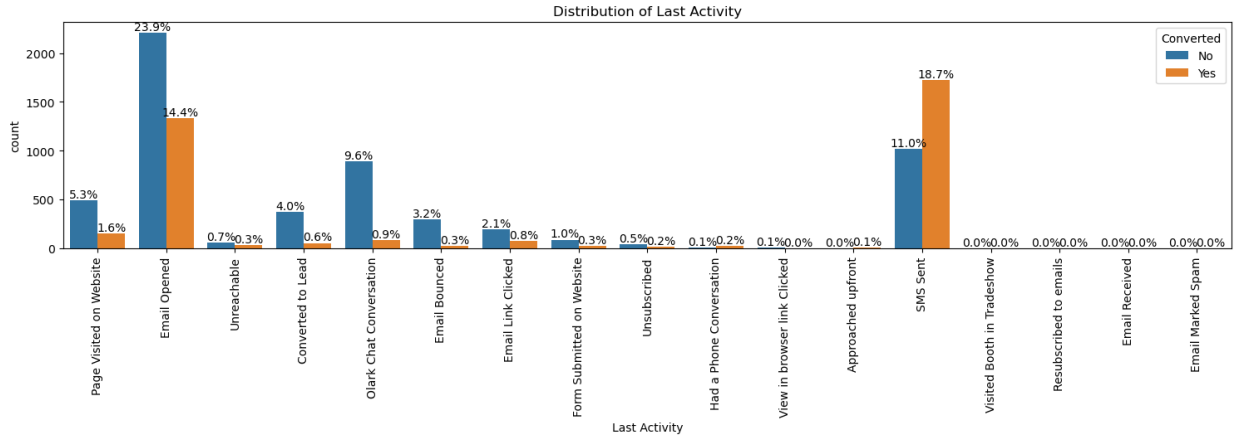
Highest conversion rate is for the reference based lead source.

This followed by Welingak Website.



Do Not Email

Selecting Do Not Email option doesn't have much impact on conversion rate.



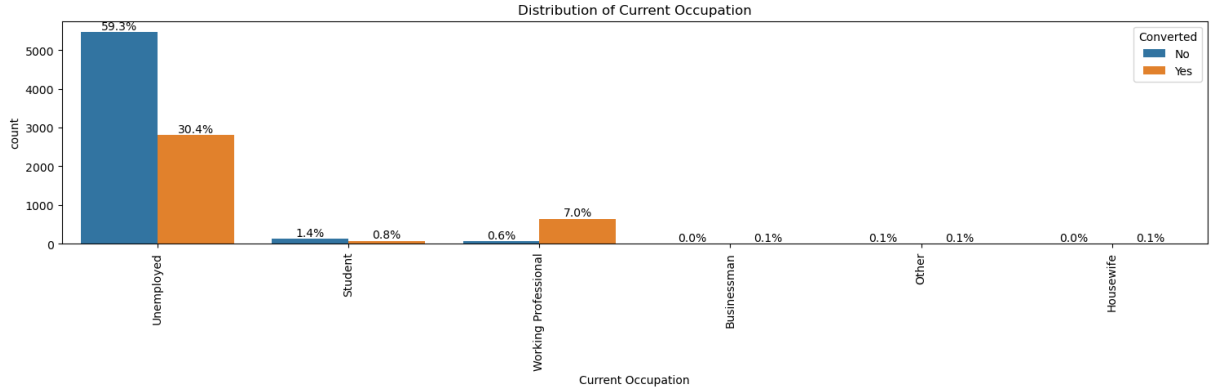
Last Activity

Leads with last activity as SMS sent have high conversion rate.

Whereas, leads with last activity as Email Opened have less conversion rate.

Free Copy

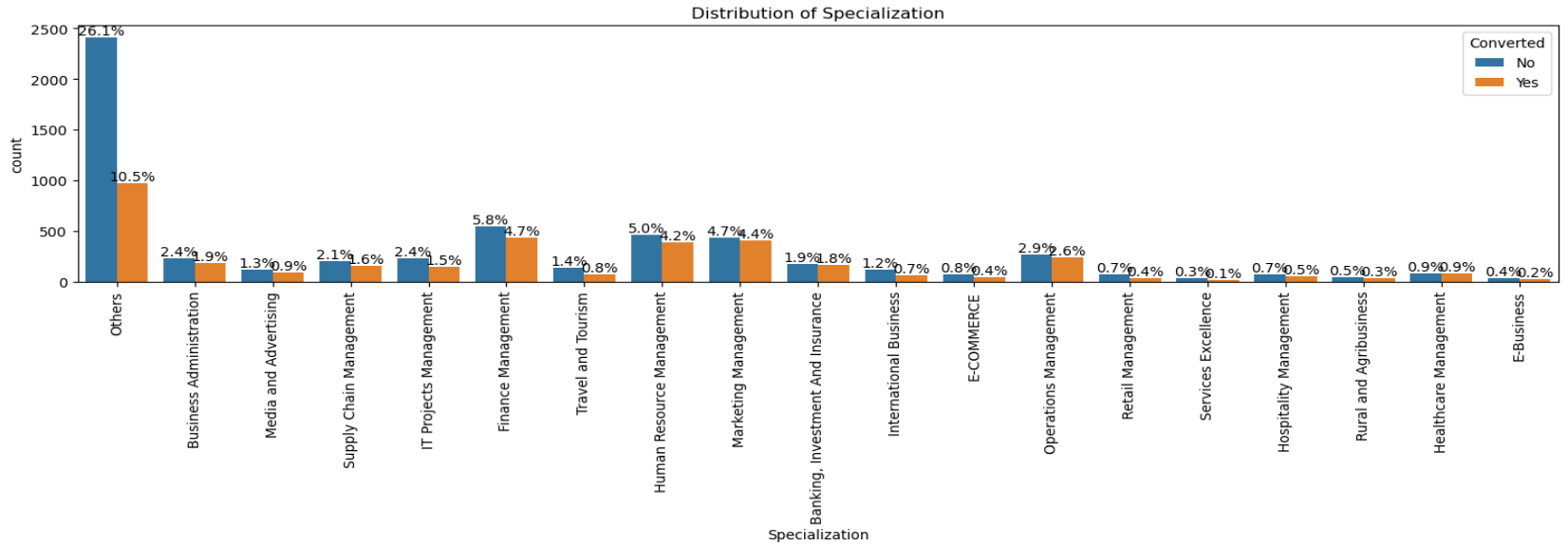
Free interview copy doesn't help increase conversion rate.



Current Occupation

Working Professionals have a high conversion rate.

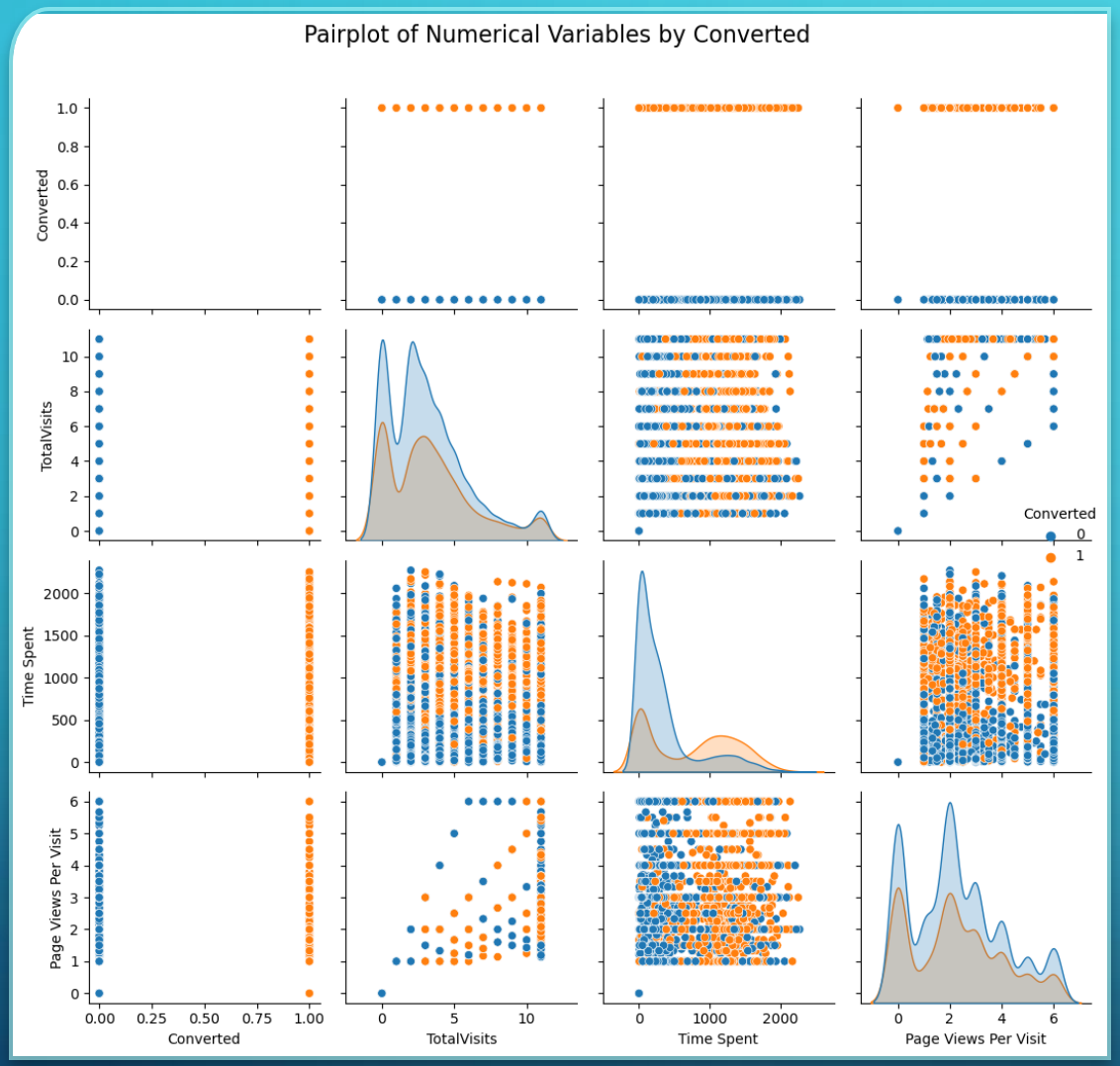
Students have low conversion rate.

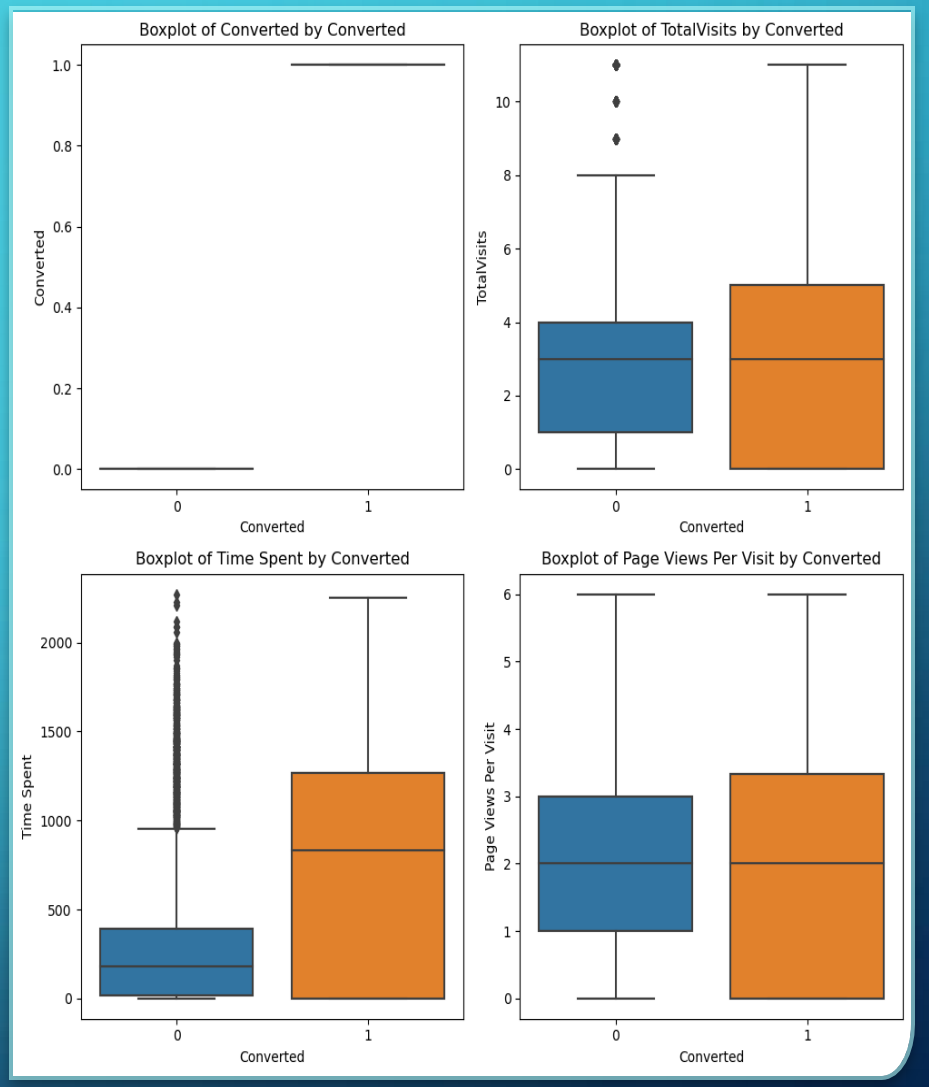


Specialization

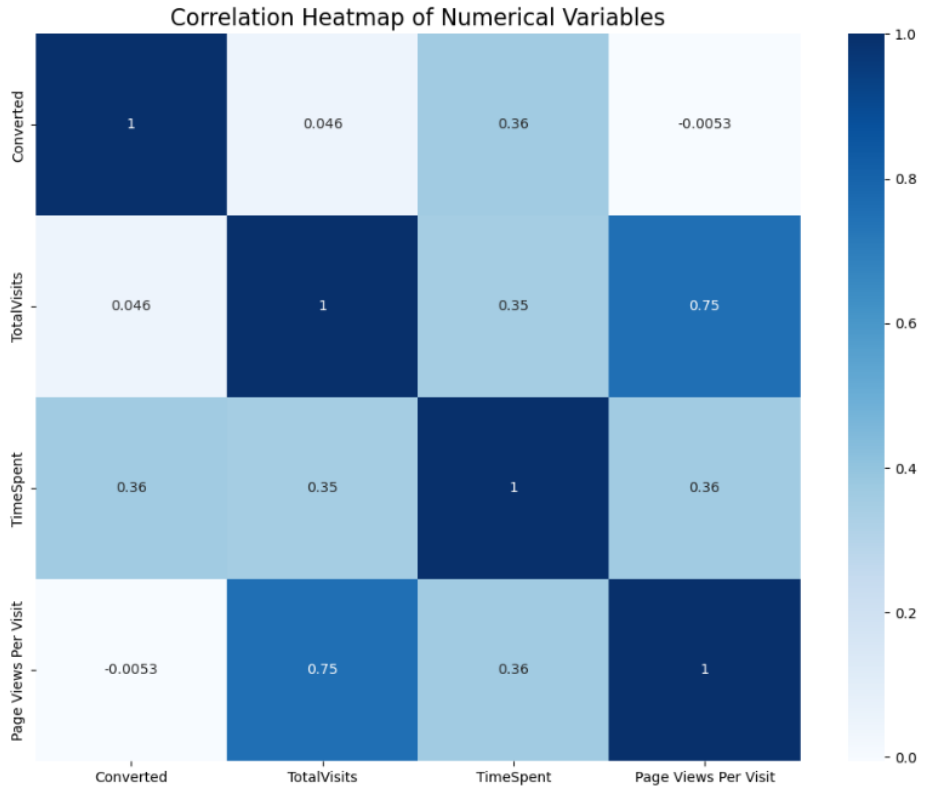
There is no specific trend regarding specializations. But human resources management, marketing management etc. has high conversion rates and promising leads are people from these specializations.

BIVARIANT ANALYSIS – NUMERICAL VARIABLES





- Leads which convert usually spend longer time on the website and view higher number of pages.



CORRELATIONS

There are no variables with two high variables.

Total Visits and Pages Views Per Visit – 0.75

OBSERVATION

EXPLORATORY DATA ANALYSIS:

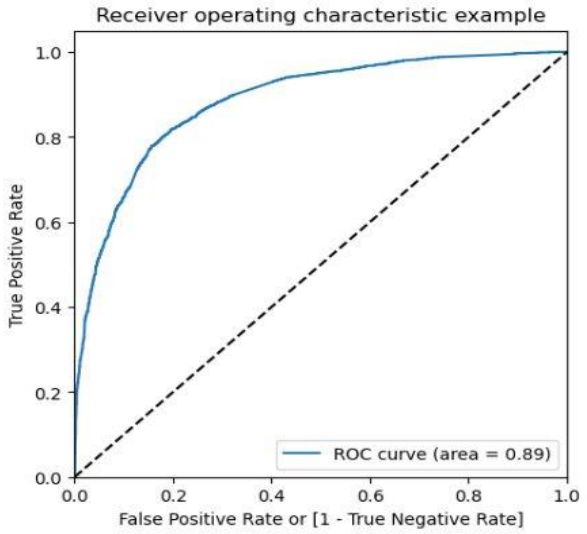
- It can be helpful in conversion to target and approach the people spending higher than average time on websites.
- SMS messages can be highly impactful on lead conversion.
- Landing page submissions can be helpful in finding out more leads.
- The human resources management, marketing management etc. has high conversion rates and promising leads are people from these specializations.
- Two good sources for higher conversions are references and offers.
- Working professionals based leads have higher conversion rate.

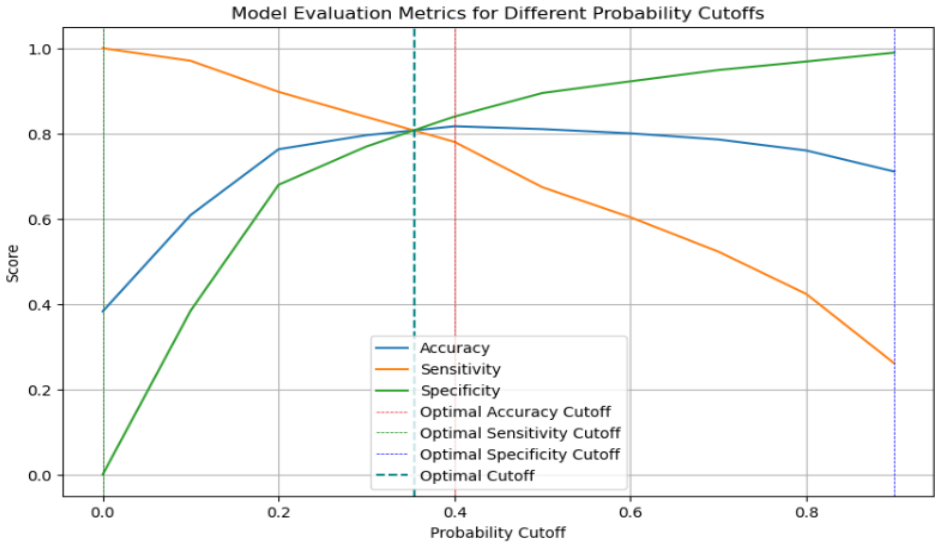
MODEL BUILDING AND EVALUATION

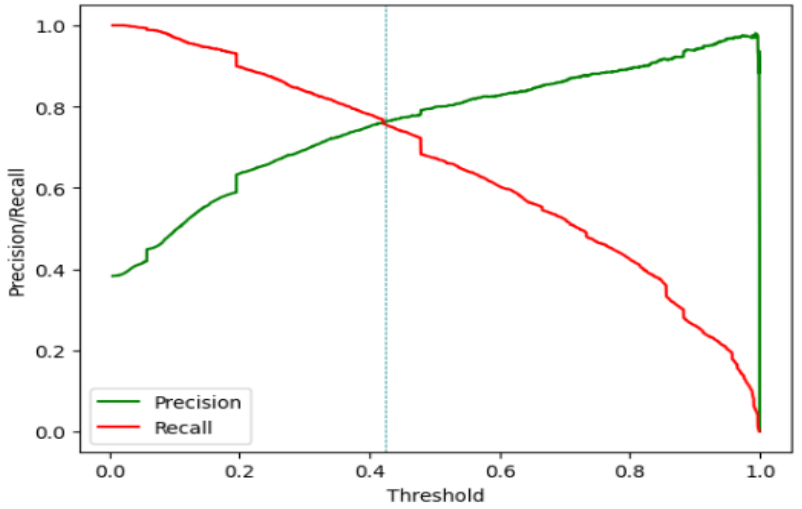
MODEL BUILDING STEPS

- Split into train set and test set. In ratio of 80:20
- Scaling variables in a train set using MinMaxScaler.
- Building the first model.
- Eliminating less relevant variables by using RFE with 15 variables.
- Improving model by removing the variables whose p-value is greater than 0.05 and VIF value is greater than 5.
- Using train set for prediction.
- Accuracy and other metric evaluation.
- Using test set for prediction.
- Precision and recall analysis on test predictions.

MODEL EVALUATION (TRAIN DATASET)





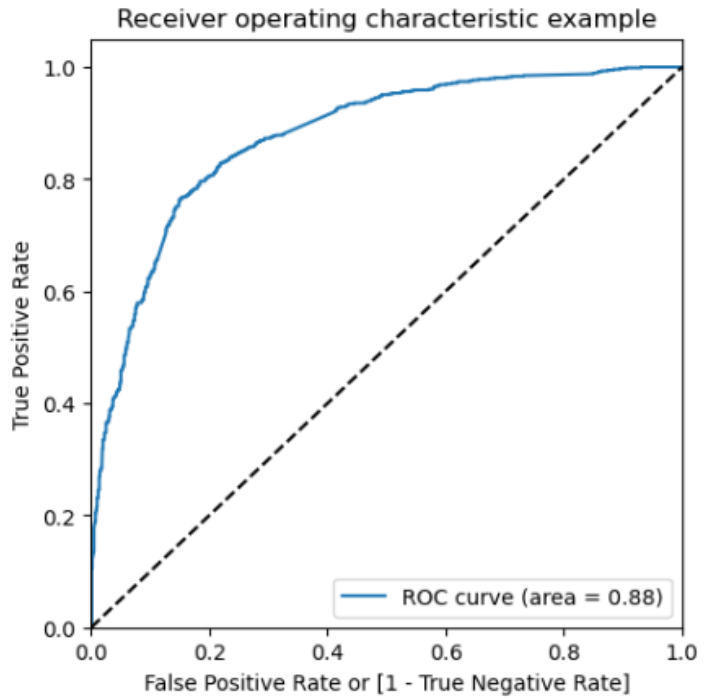


EVALUATIONS

- Accuracy: 80.96% ≈ 81%
- Sensitivity: 80.80%
- Specificity: 85.12%

3700	863
544	2285

MODEL EVALUATION (TEST DATASET)



Area under ROC curve is 0.88 out of 1 which indicates a good predictive model

EVALUATIONS

- Accuracy: 80.25%
- Sensitivity: 80.28%
- Specificity: 80.19%

896	220
145	587

OBSERVATION

- The probability of conversion of a customer is predicted by using the logistic regression model.
- The train and test data sets show above 80% accuracy.
- Accuracy, sensitivity & specificity measures and precision & recall curves are the parameters from which the threshold has been selected.
- Promising leads and leads that have less changes of getting converted are being correctly found by the model.
- 80.73% specificity and 80.46% sensitivity are also shown by the model.
- Since, the train and test result are similar and they fulfill the business requirement this model is valid.

OBSERVATIONS

FINAL FEATURE LIST:

1. Time Spent
2. Lead Origin - Lead Add Form
3. Current Occupation - Working Professional
4. Lead Source - Welingak Website
   - Olark Chat
5. Last Activity - Had a Phone Conversation
   - SMS Sent
   - Converted to Lead
   - Olark Chat Conversation
6. TotalVisits
7. Page Views Per Visit
8. Lead Origin - Landing Page Submission
9. Specialization - Other
10. Do Not Email

Keeping this features in mind X Education can increase their lead conversion rate.

CONCLUSION

- Lead Origin-Lead Add Form, Total Visits, and Time Spent are the three most important variables affecting lead conversion.
- Since working professionals are more likely to take advantage of upskilling possibilities, targeting them seems promising.
- Using text messages or phone calls could increase the conversion rate of customers.
- A significant percentage of website visits do not become leads.
- Moreover, those connected to the most recent Activity-Olark Chat Conversation has a decreased chance of converting.


THE END