Lead Scoring Case Study using logistic regression

SUBMITTED BY :

1. Aniket Mishra
2. MadhuSree Kodali
3. Saksham Jain

Contents

- Problem statement
- Problem approach
- EDA
- Correlations
- Model Evaluation
- Observations
- 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

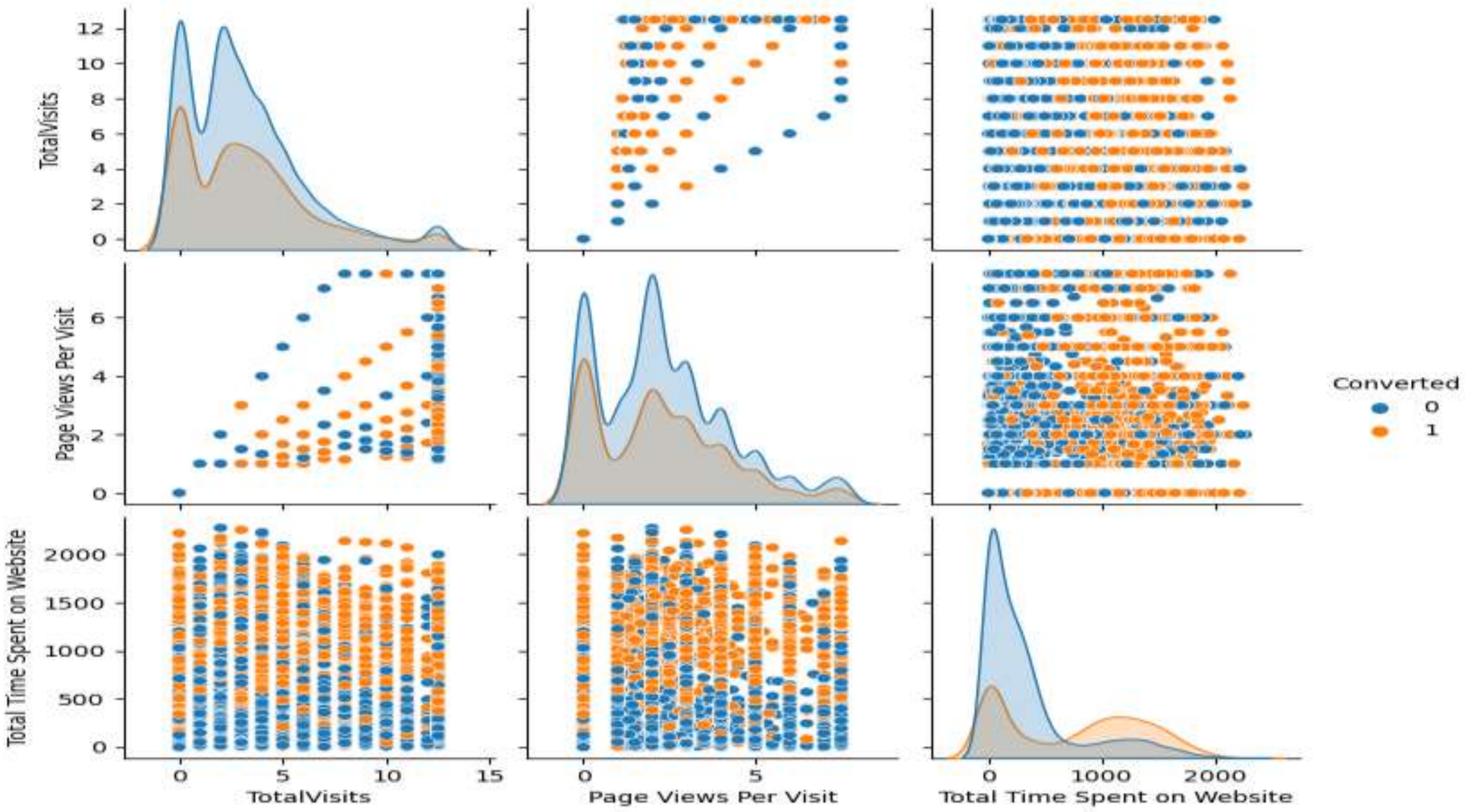
Business Objective

- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- 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 Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

Problem Approach

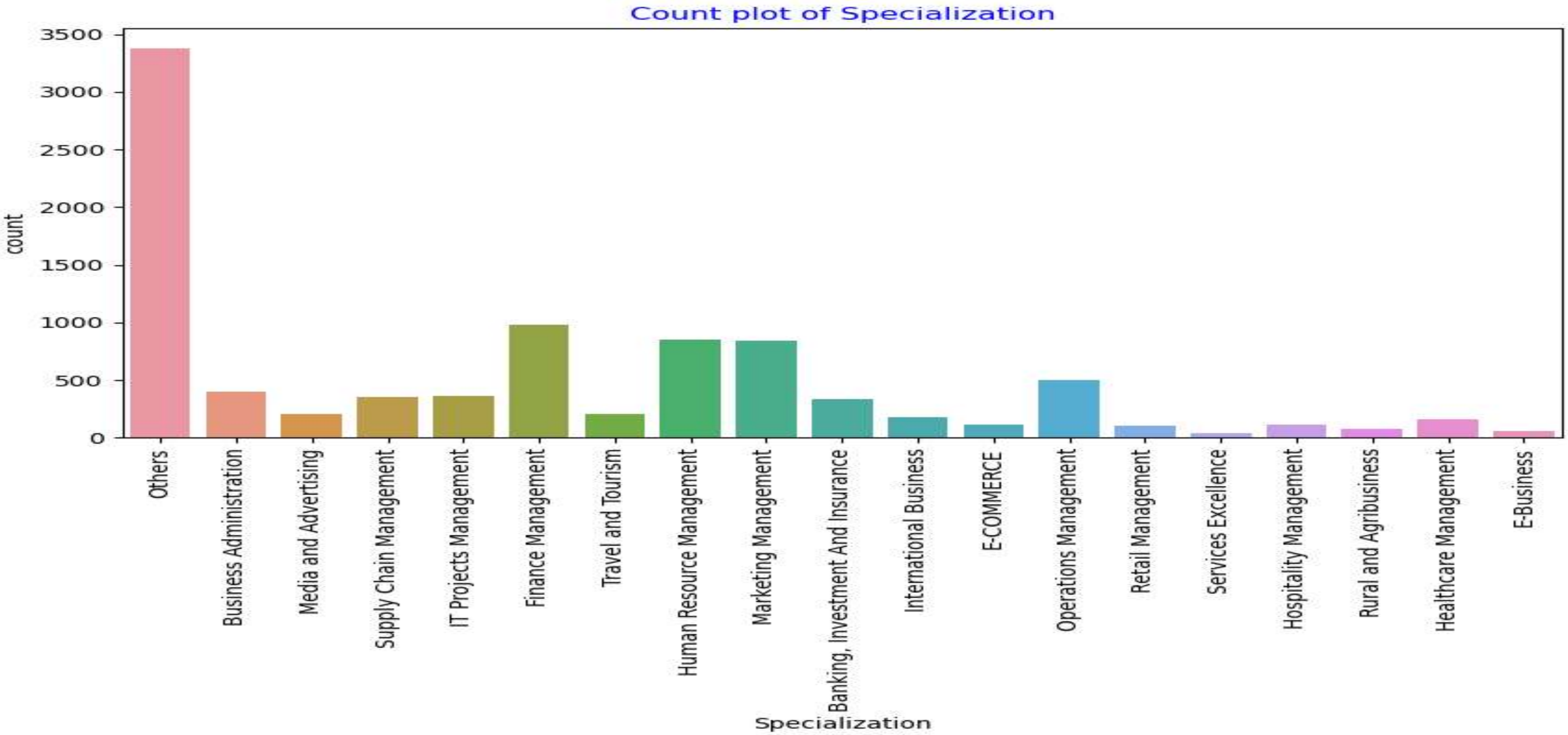
- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE Rsquared VIF and p-values)
- Model Evaluation
- Making predictions on test set

EXPLORATORY DATA ANALYSIS

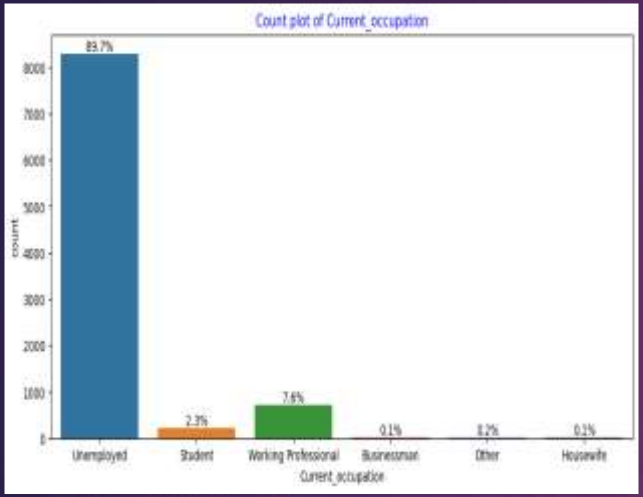


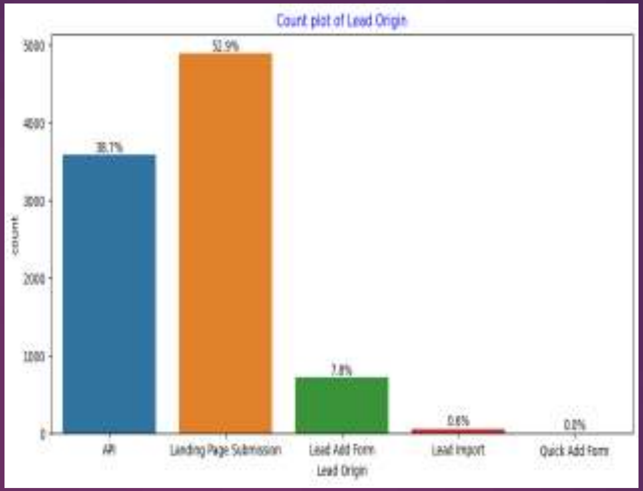
Specialization

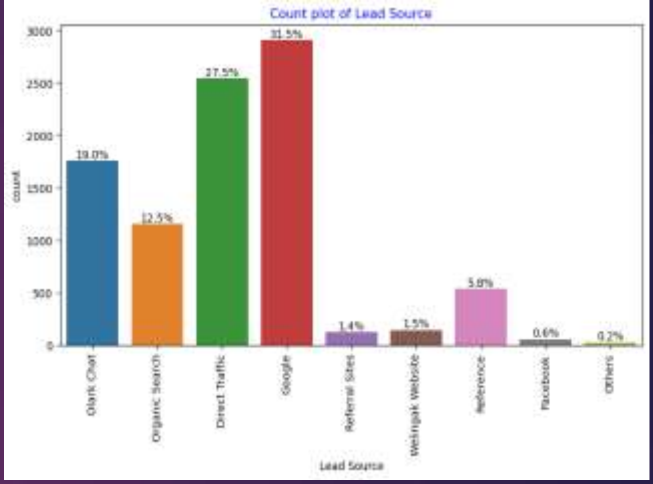
Leads from HR, Finance & Marketing management specializations are high probability to convert



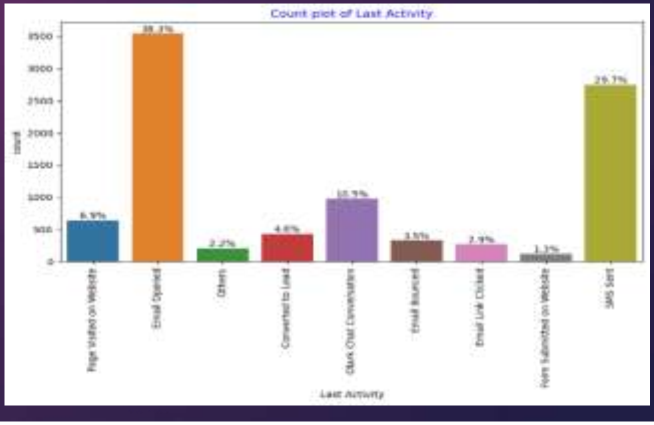
UNIVARIATE ANALYSIS OF CATEGORIC VARIABLE



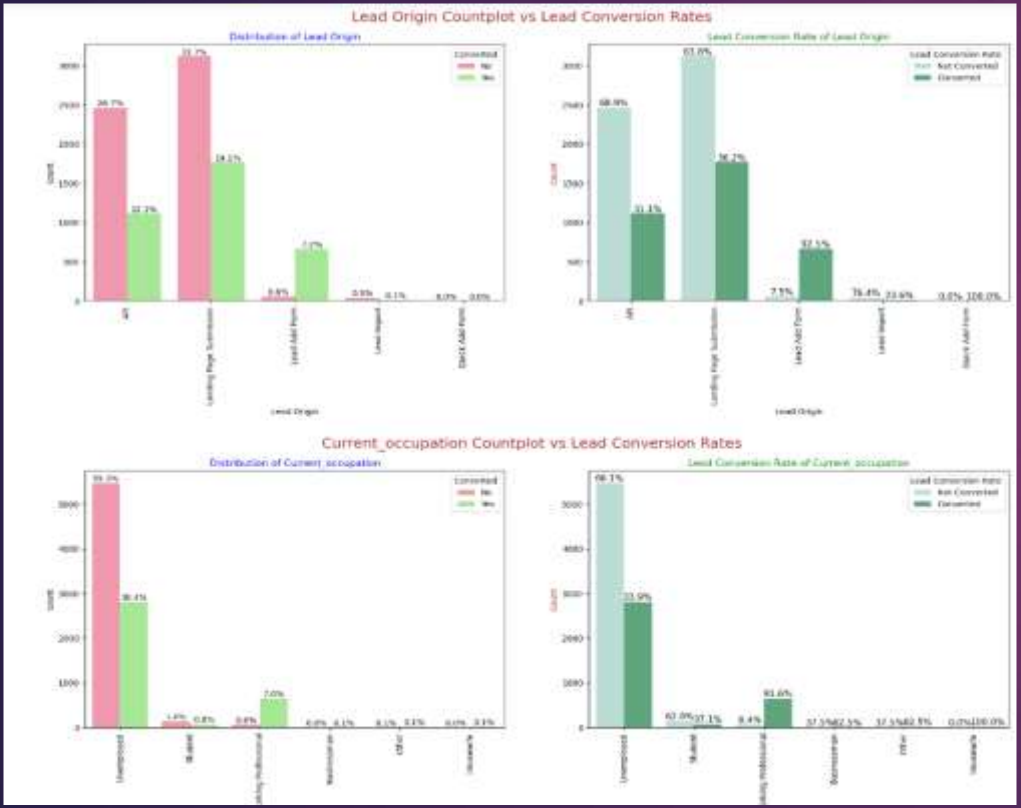


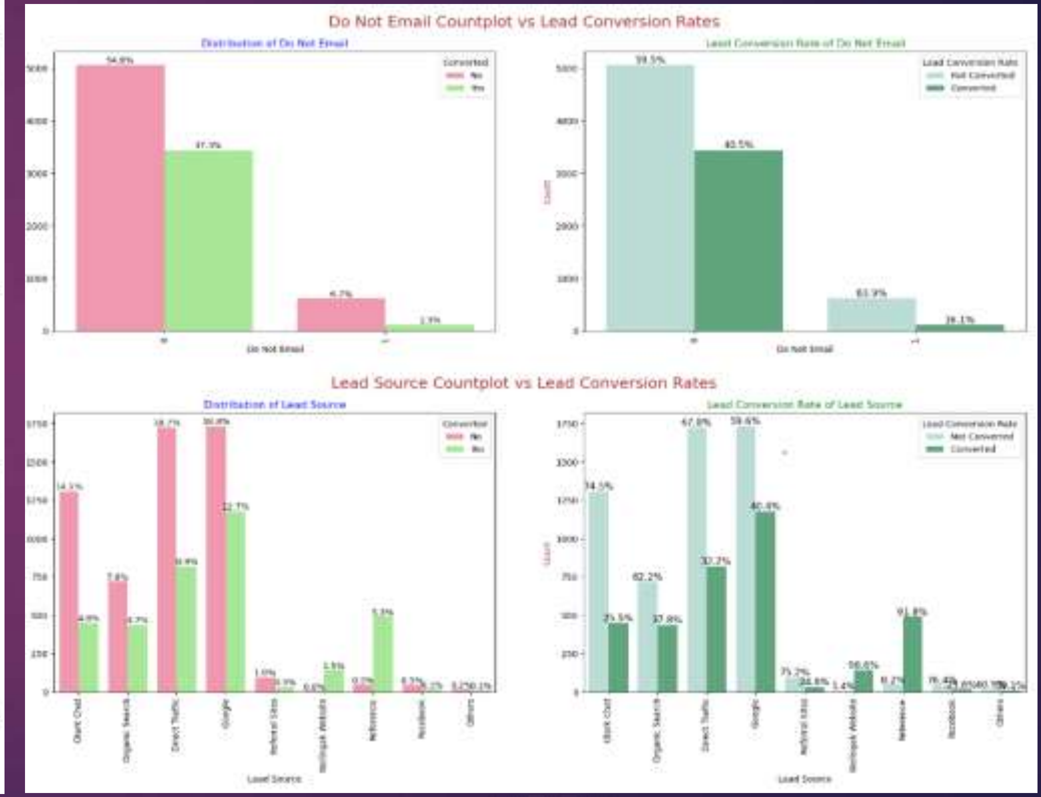


In Categorical Univariate Analysis we get to know the value counts percentage in each column with this we can distinguish which variables can be used in Bivariate analysis.
Univariate Analysis Insights:
Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.
Lead Source: 58% Lead source is from Google & Direct Traffic combined
Current_occupation: It has 90% of the customers as Unemployed
Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.
Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

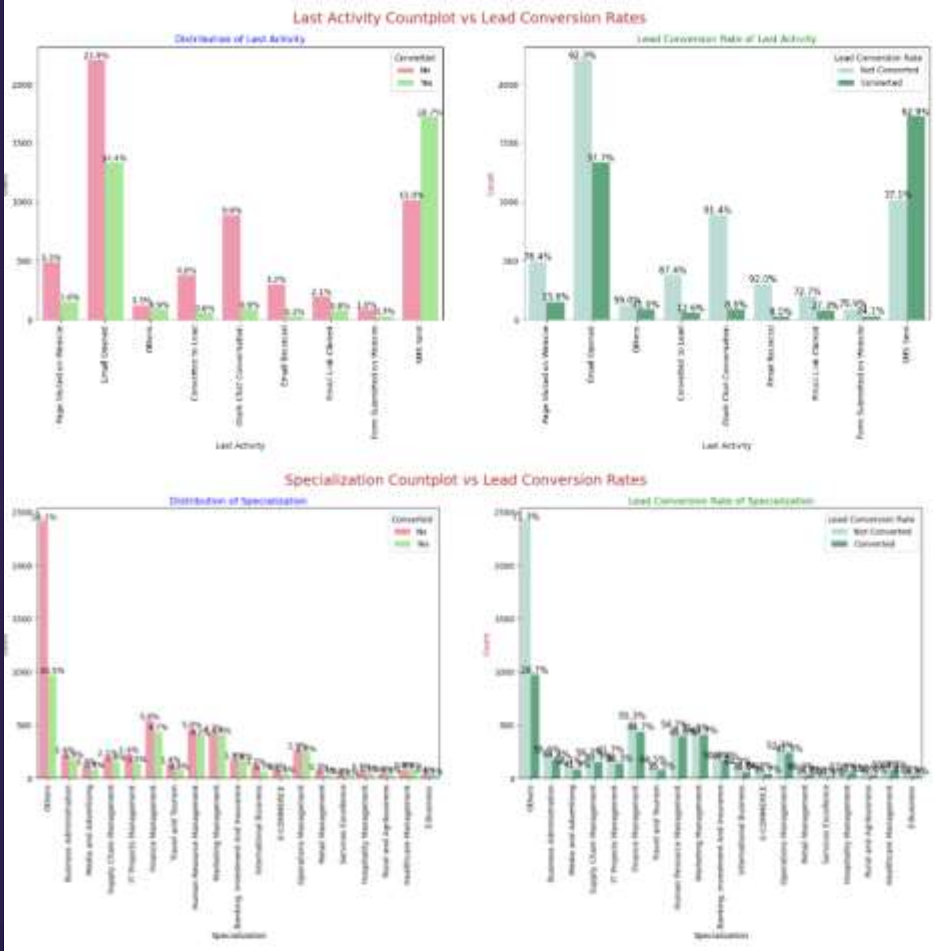


Bivariate Analysis for Categorical Variable





Bivariate Analysis for Categorical Variable



Lead Source: Google has Lead Conversion Rate of 40% out of 31% customers , Direct Traffic contributes 32% Lead Conversion Rate with 27% customers which is lower than Google,Organic Search also gives 37.8% of LCR but the contribution is by only 12.5% of customers ,LCR of 91% but there are only around 6% of customers through this Lead Source.

Lead Origin: Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%.The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

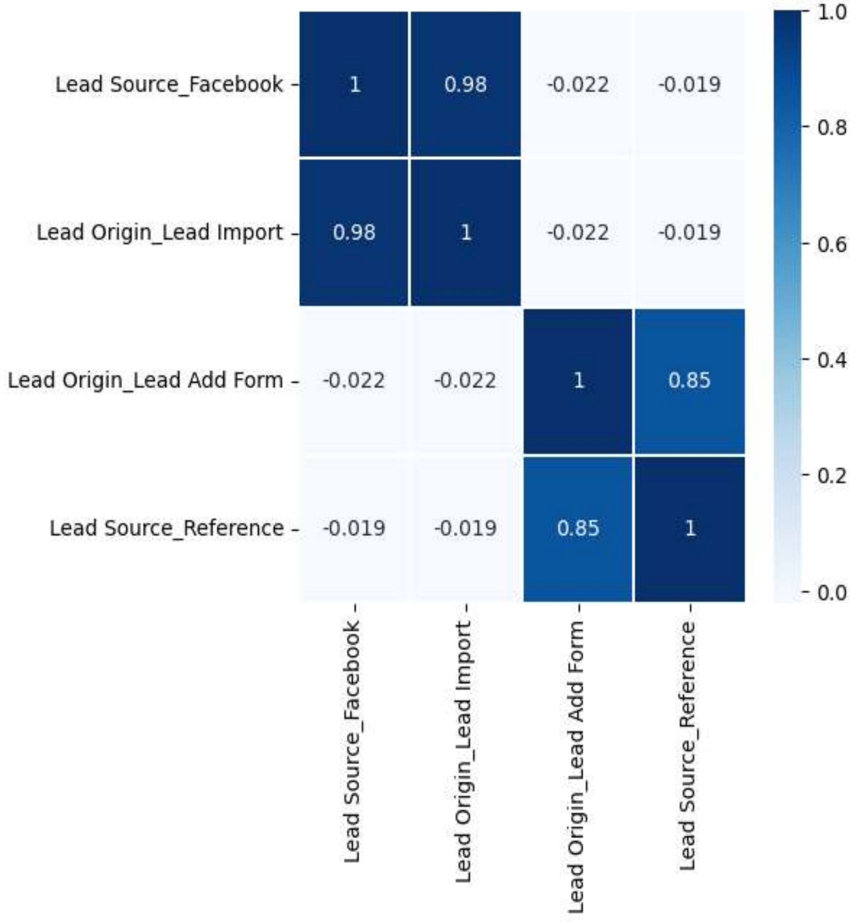
Current_occupation: Around 90% of the customers are Unemployed with lead conversion rate (LCR) of 34%. While Working Professional contribute only 7.6% of total customers with almost 92% lead conversion rate (LCR).

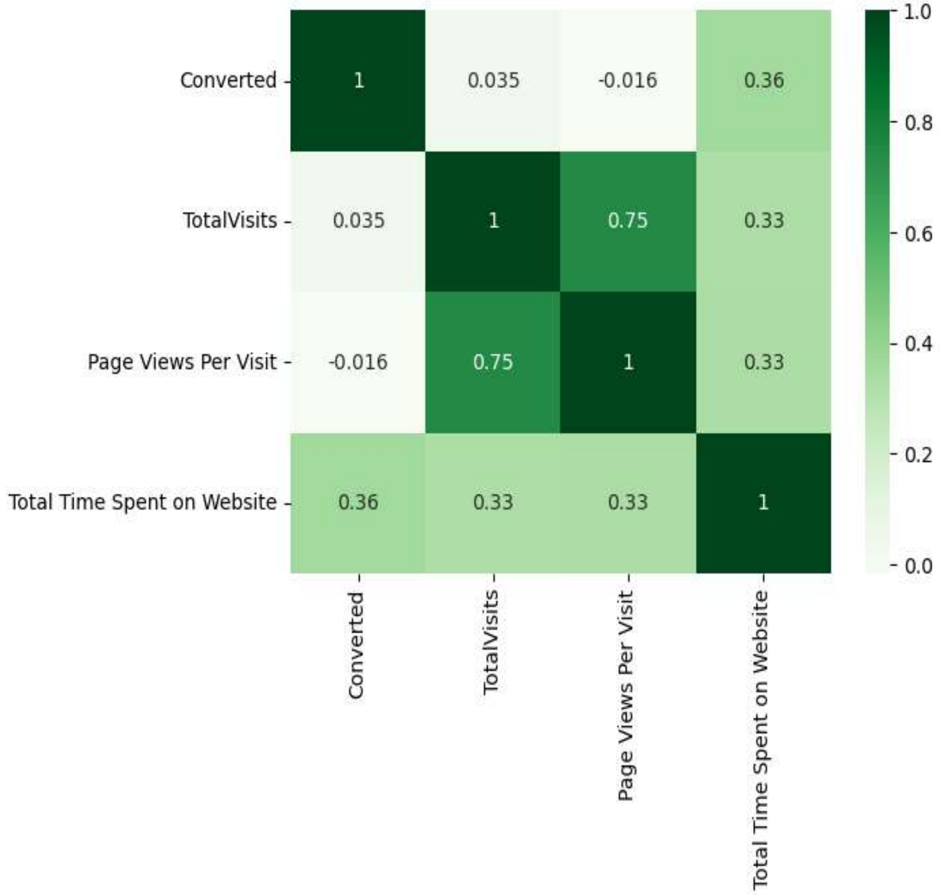
Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.

Last Activity: 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities, 'Email Opened' activity contributed 38% of last activities performed by the customers with 37% lead conversion rate.

Specialization: Marketing Managemt,HR Management,Finance Management shows good contribution.

Correlation



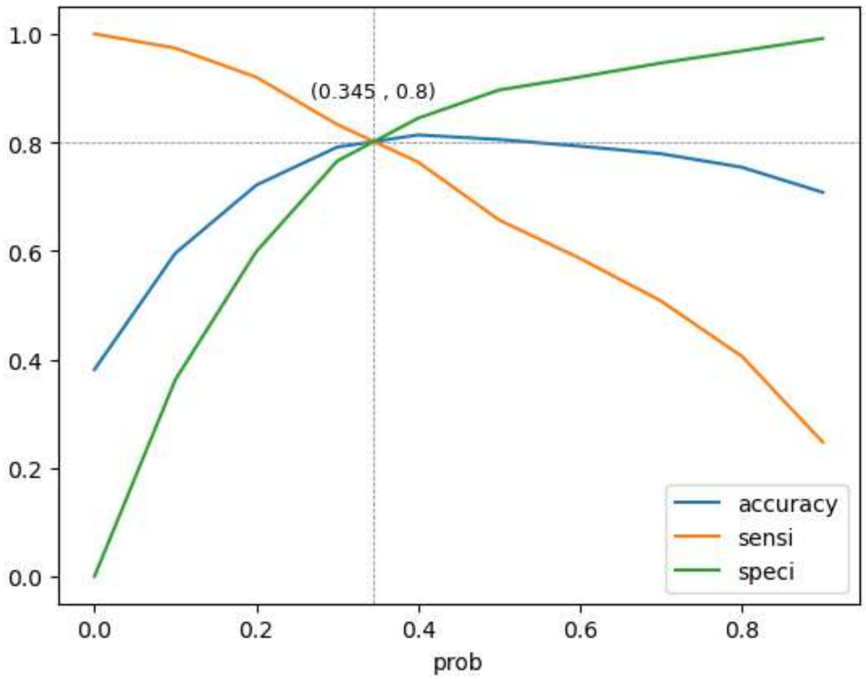


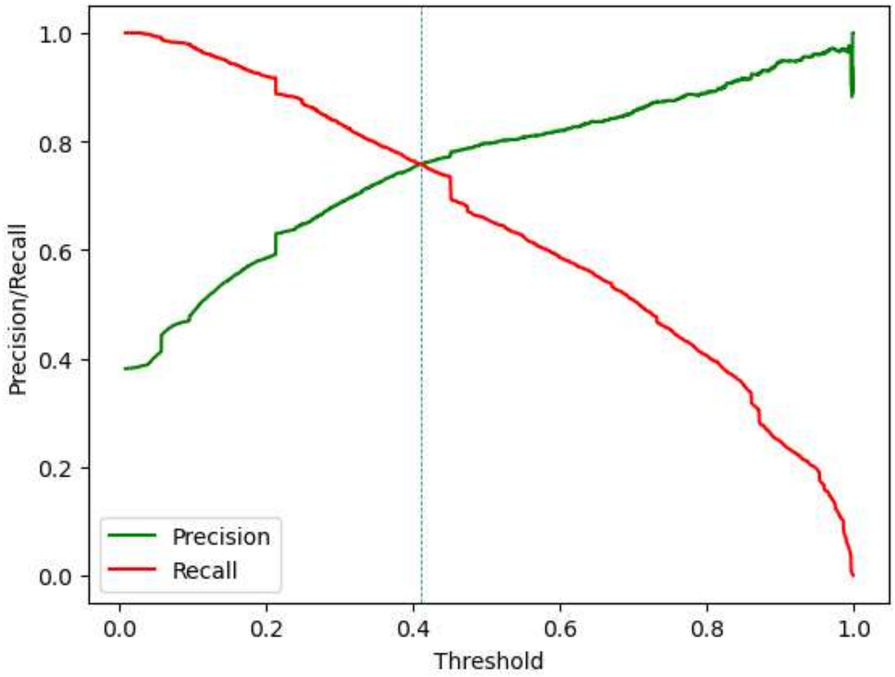
Model Evaluation

ROC curve

0.42 is the tradeoff between Precision and Recall -

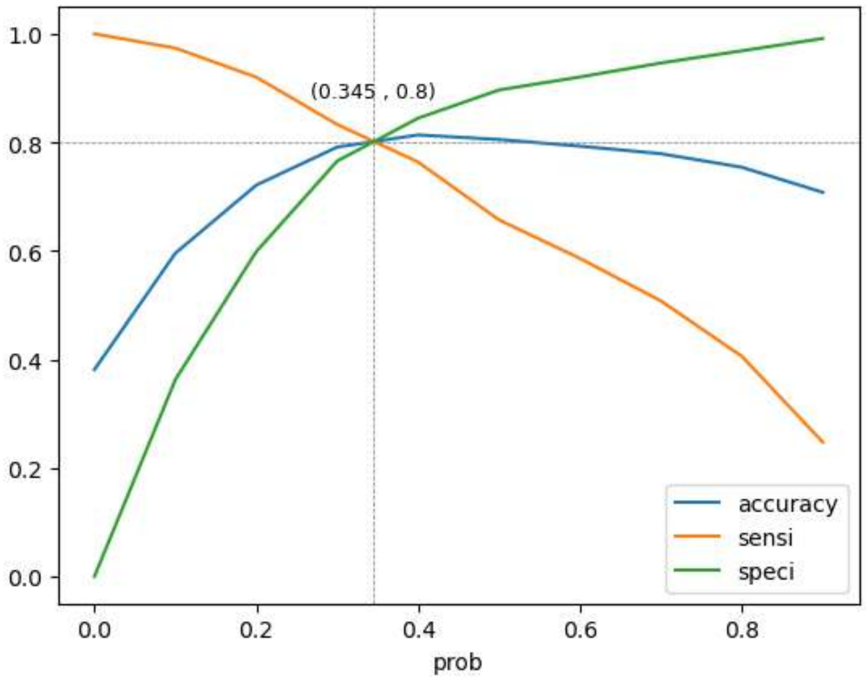
Thus we can safely choose to consider any Prospect Lead with Conversion **Probability higher than 42 % to be a hot Lead**

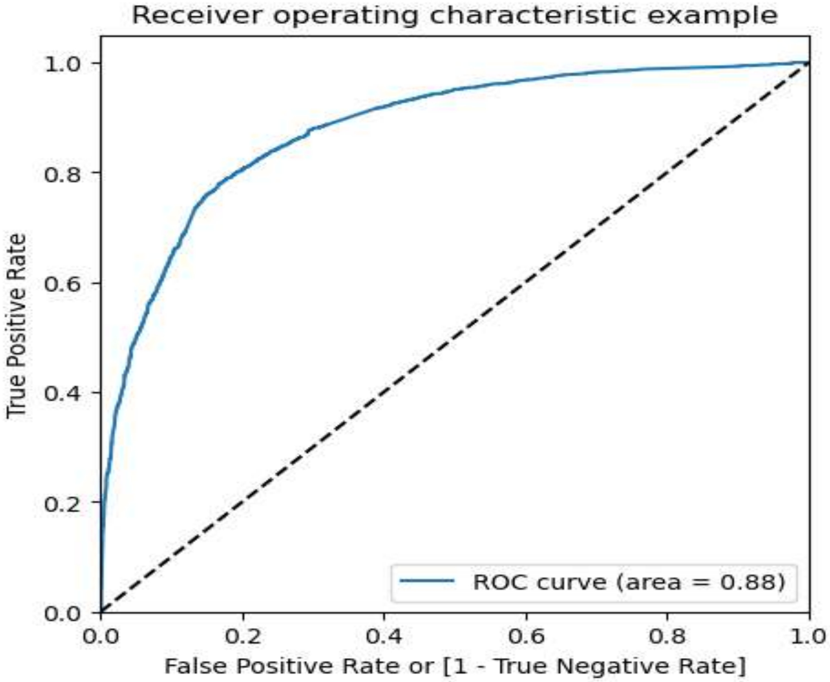




ROC CURVE

From the first graph it is visible that the optimal cut off point is at 0.35.





Observations

Train Data:

Accuracy : 80.46%
Sensitivity : 80.05%
Specificity : 80.71%

Test Data:

Accuracy : 80.34%
Sensitivity : 79.82%
Specificity : 80.68%

Optimal Cutoff Point – 0.345

Final Features list:

- Lead Source_Olark Chat
- Specialization_Others
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- What is your current occupation_Working Professionals
- Do Not Email

Conclusion

- We see that the conversion rate is 30-35% (close to average) for API and Landing page submission. But very low for Lead Add form and Lead import. Therefore we can intervene that we need to focus more on the leads originated from API and Landing page submission.
- We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and welingak website.
- Leads who spent more time on website, more likely to convert.
- Most common last activity is email opened. highest rate = SMS Sent. Max are unemployed. Max conversion with working professional.