LEAD SCORING CASE STUDY

BY-

Sanyogita Gupta

Rakshit Moundekar

Vamsikrishna

Topics

- Introduction
- Problem Statement
- Business Requirement
- Problem Approach
- EDA
- Model Building
- Model Evaluation
- Observation
- Conclusion
- Summary




Introduction

- THIS CASE STUDY AIMS TO GIVE YOU AN IDEA OF APPLYING MODEL BUILDING IN A REAL BUSINESS SCENARIO. IN THIS CASE STUDY, APART FROM APPLYING THE TECHNIQUES THAT WE HAVE LEARNT IN THE EARLIER MODULES, WE WILL ALSO DEVELOP THE UNDERSTANDING OF HANDLING THE REAL-LIFE DATA IN MULTIPLE SERVICES AND UNDERSTAND HOW DATA IS USED TO MINIMIZE THE RISK OF LOSING LEADS.

Problem Statement

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these 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. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%. 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 Requirement-

- 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

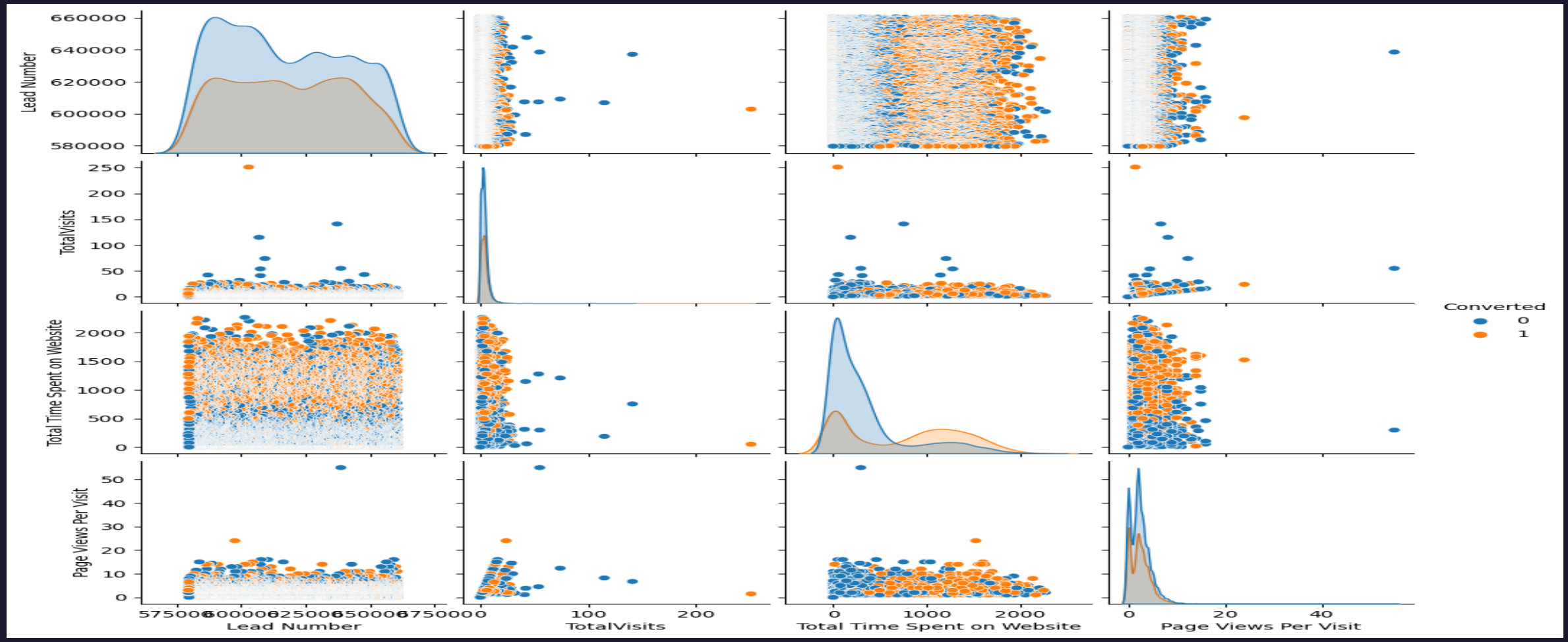
- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE R-squared VIF and p-values)
- Model Evaluation
- Making predictions on test set

DATA CLEANING and EDA

First we have imported the dataset and cleaned the impurities such as null values and to make data handling easier we dropped the unnecessary columns. When the dataset is clean and ready for analysis, we will start by EDA (Exploratory data analysis).

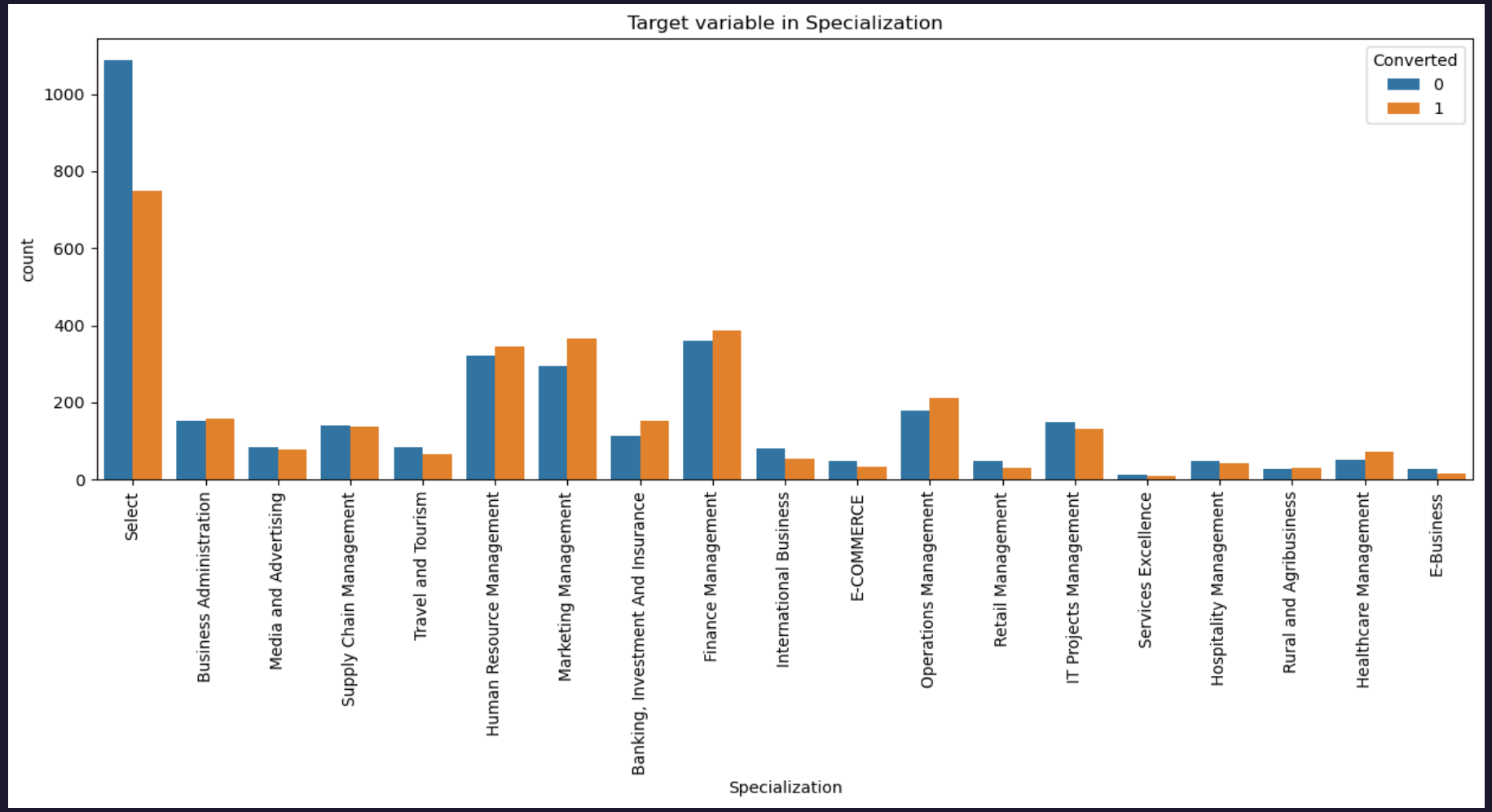
Analysis of the cleaned data-

Here we can see the data is bit scattered but with minimum number of outliers.



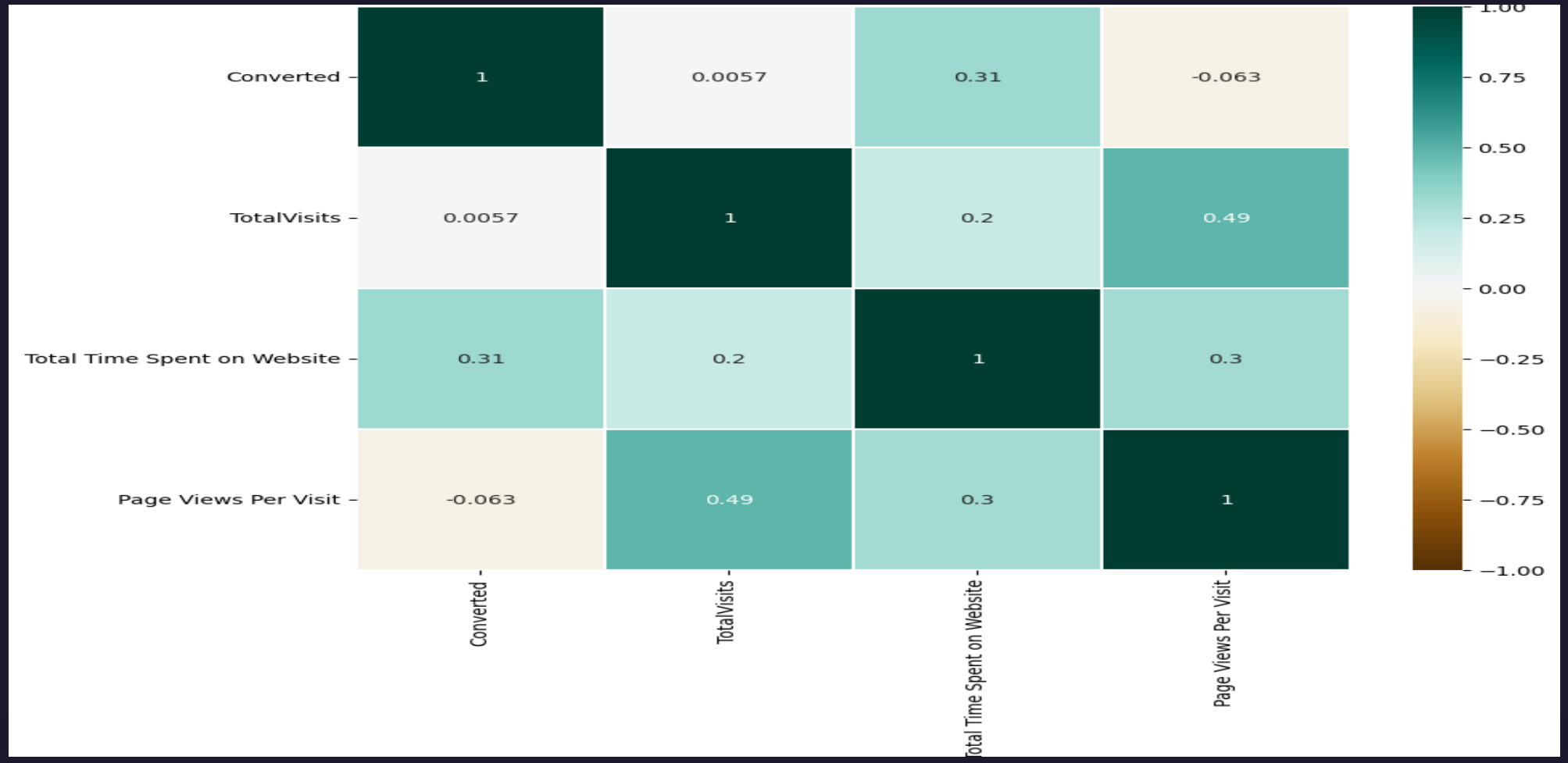
Specialization-

It is evident that leads from HR, Finance and Marketing management field have high chances to convert.



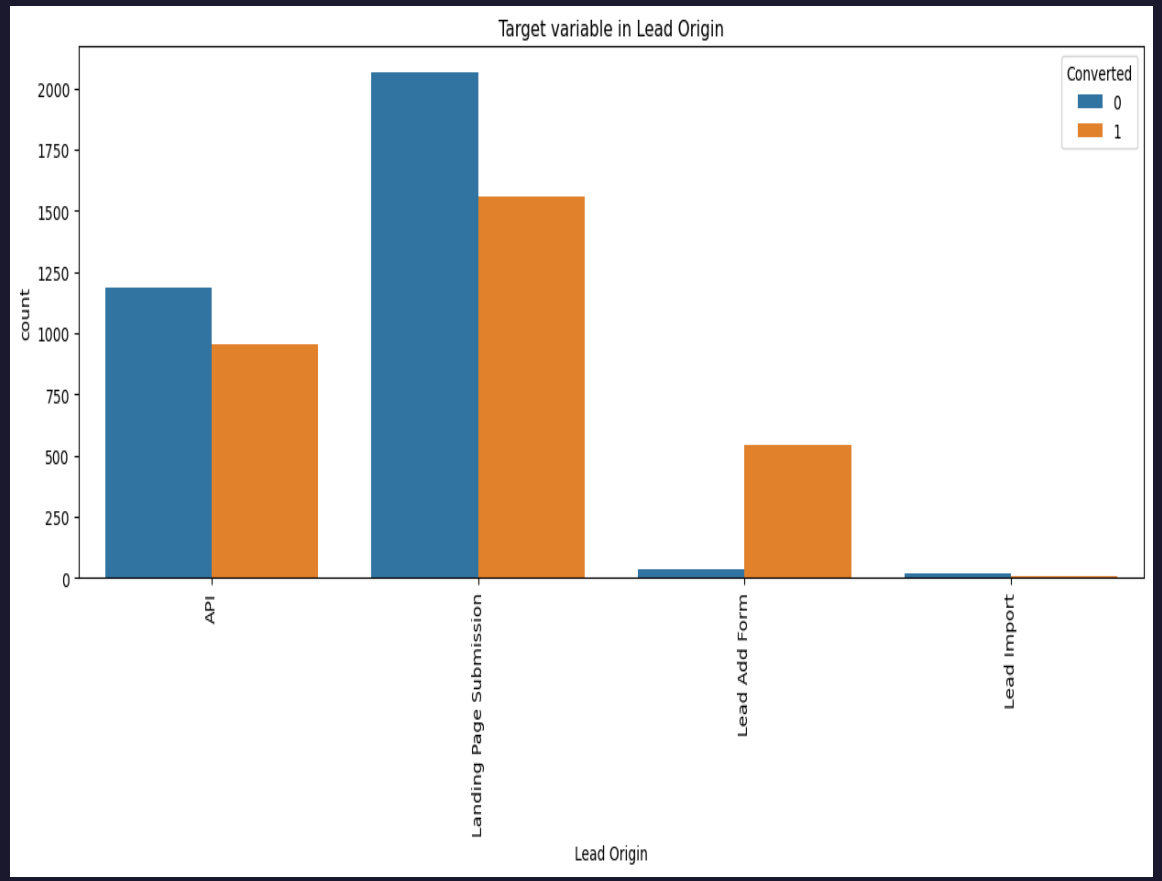
Select	1942
Finance Management	976
Human Resource Management	848
Marketing Management	838
Operations Management	503
Business Administration	403
IT Projects Management	366
Supply Chain Management	349
Banking, Investment And Insurance	338
Media and Advertising	203
Travel and Tourism	203
International Business	178
Healthcare Management	159
Hospitality Management	114
E-COMMERCE	112
Retail Management	100
Rural and Agribusiness	73
E-Business	57
Services Excellence	40

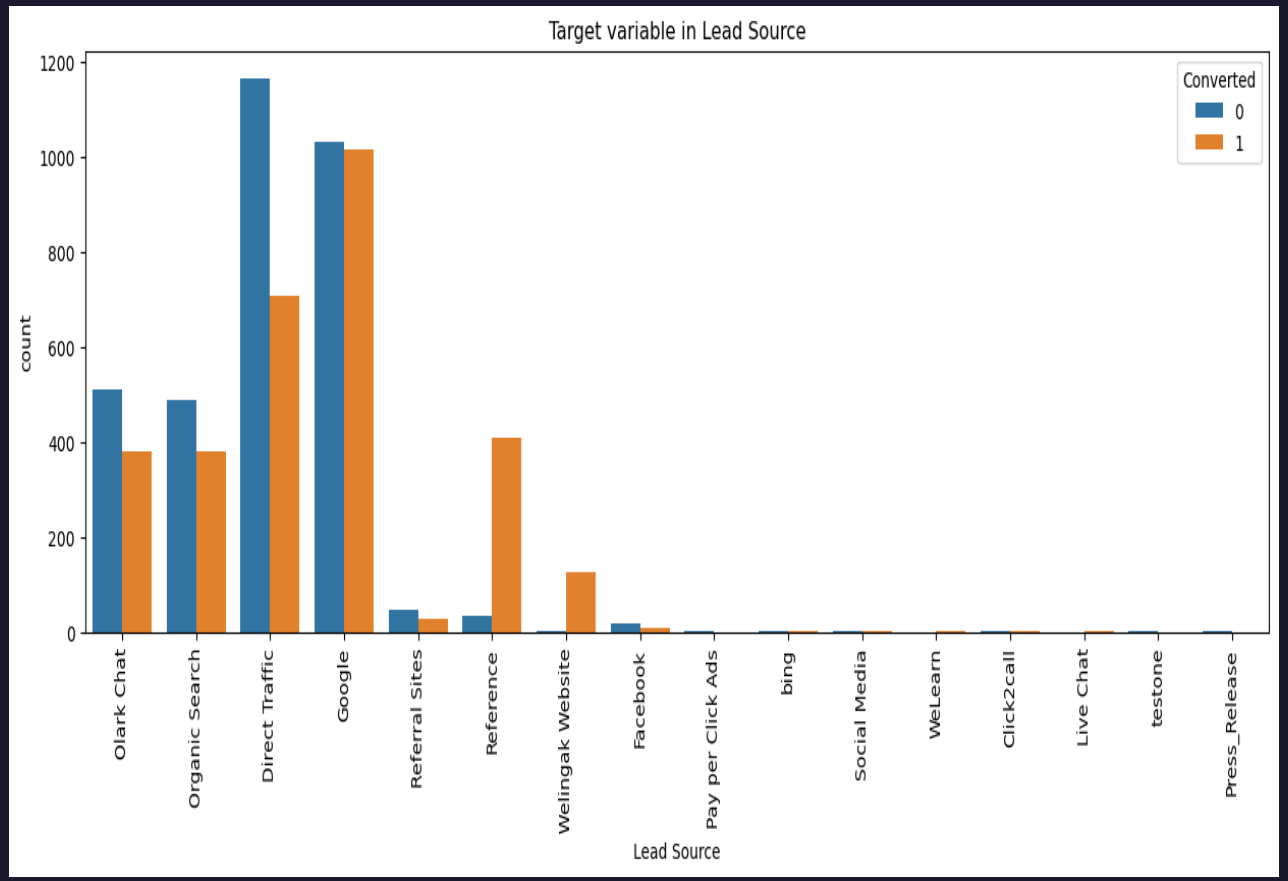
Below is the Heatmap of correlation between categorical variables and we can clearly notice a decent relationship between variables



Lead Source & Lead origin

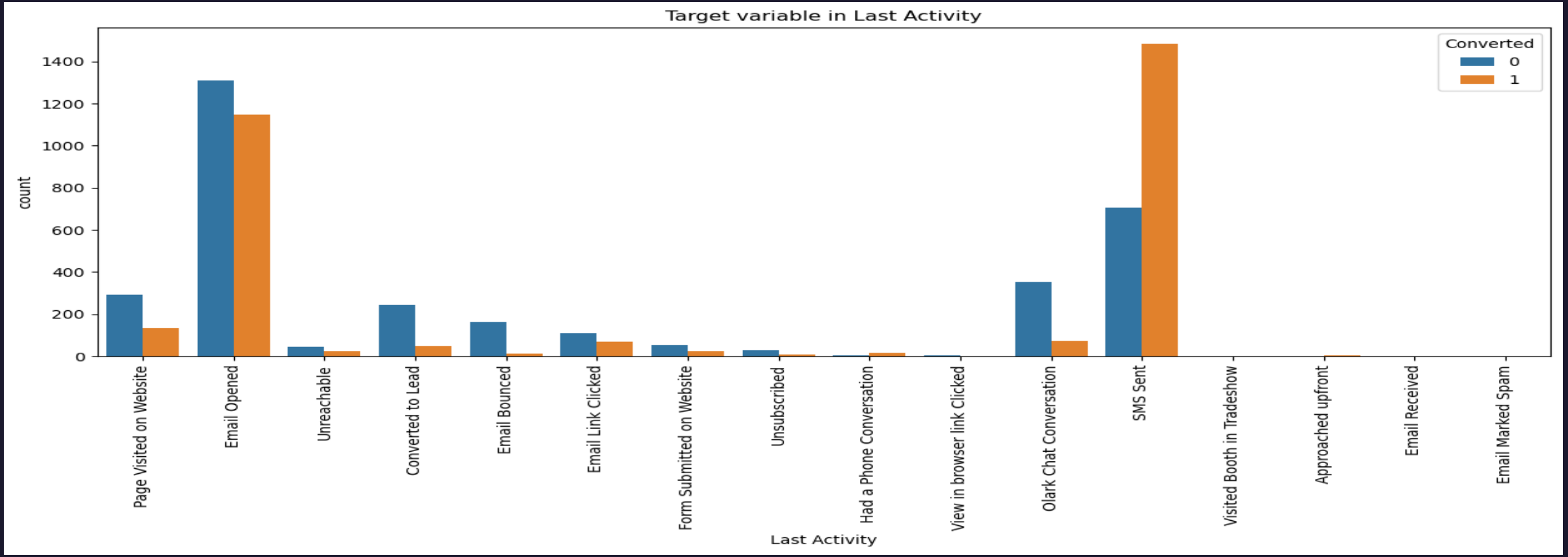
In lead source, the leads through google & direct traffic have high probability to convert. On the other hand, in Lead origin most number of leads are coming from landing on submission page.





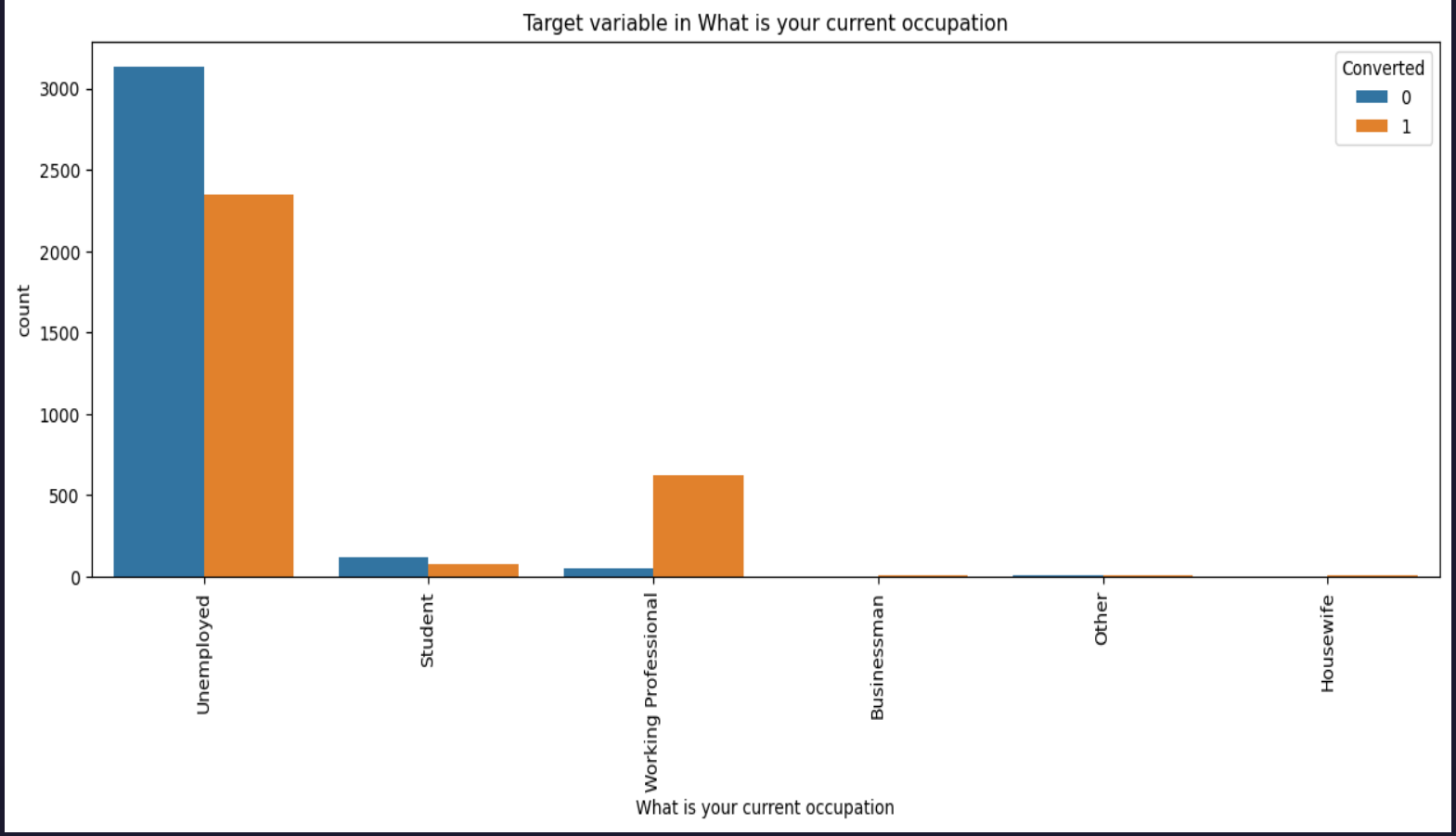
Last Activity

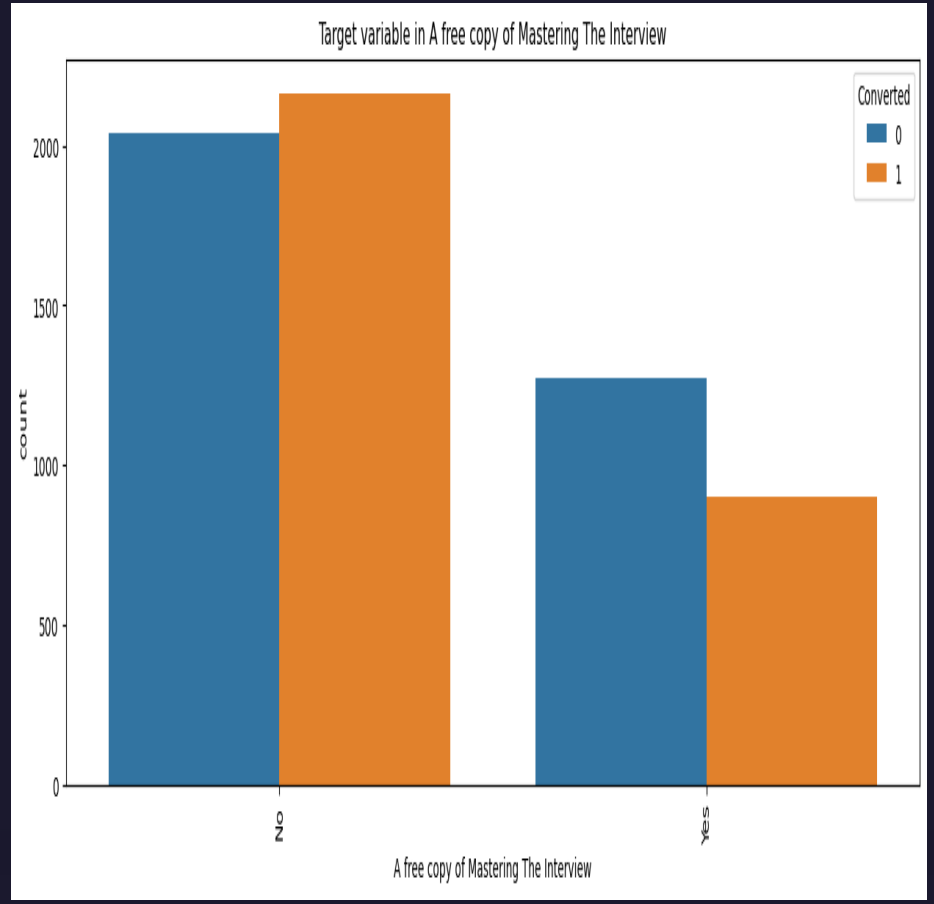
Here, leads who are opening their emails have high probability to convert, same trend is seen with Sending SMS.



Lead Occupation.

Leads which are Unemployed are more interested to join the course than others. Also people who have given a free copy of mastering the interview have high chances to convert.

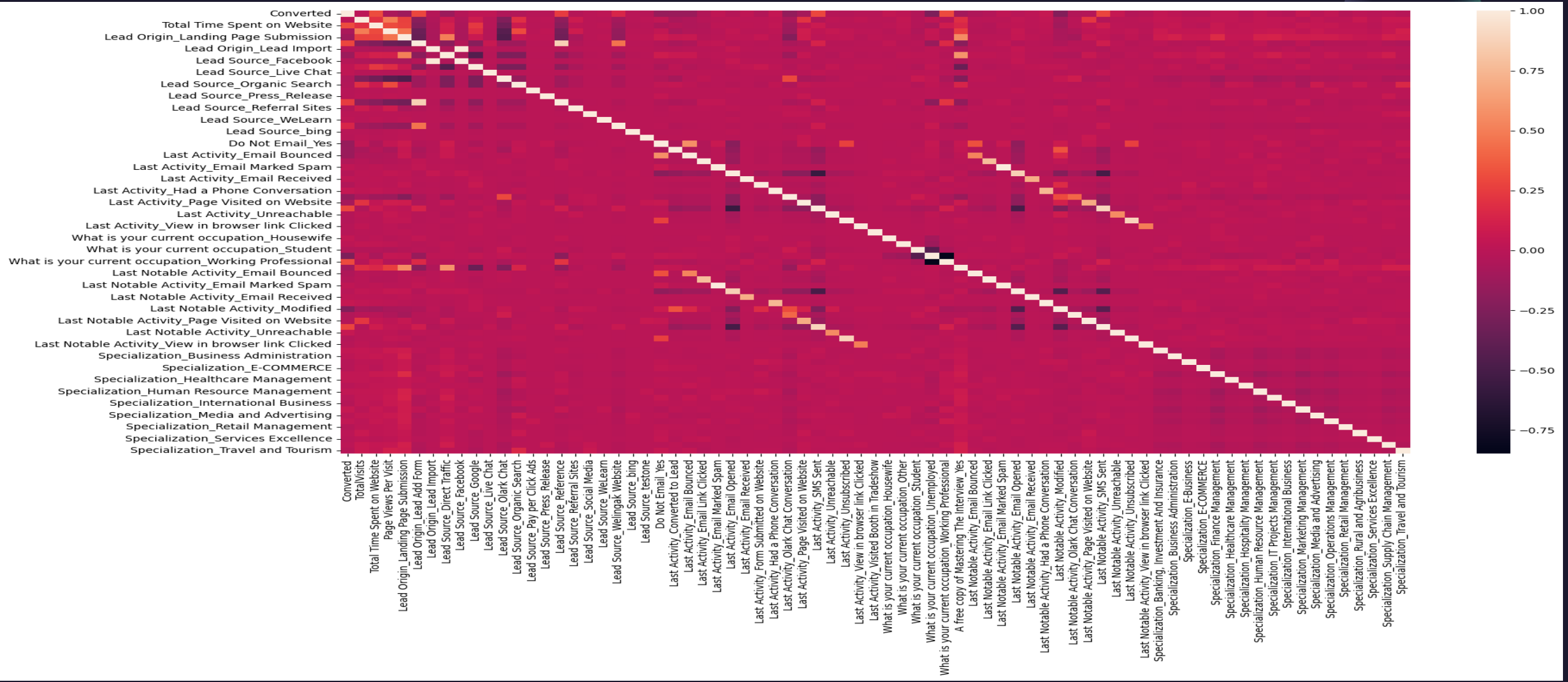




Dummy variable creation and Model Building.

- After cleaning the data we have 69% of data and that is good enough for analysis
- The next step is to dealing with the categorical variables present in the dataset. So first take a look at which variables are actually categorical variables and create the dummy variables of them.
- After that we will split the data into Train and test data and after doing the split, we will be performing the Scaling of train data
- We got 48% of conversion rate after the split and featuring the data.

Checking the correlation of the dummy variables.



Model Building.

- After creating dummy variables and splitting them, we will move to building the model. For this we will be using Logistic regression as we need to estimate the probability of leads conversion.
- For this we are using RFE (Recursive feature elimination) and create models one by one. Then we can check the variables who have a p-value higher than 0.05 and VIF (Variation inflation factor) more than 5 (assumed), we need to drop those variables unless we get a model which has all p-values under 0.05 and VIFs under 5 (assumed).

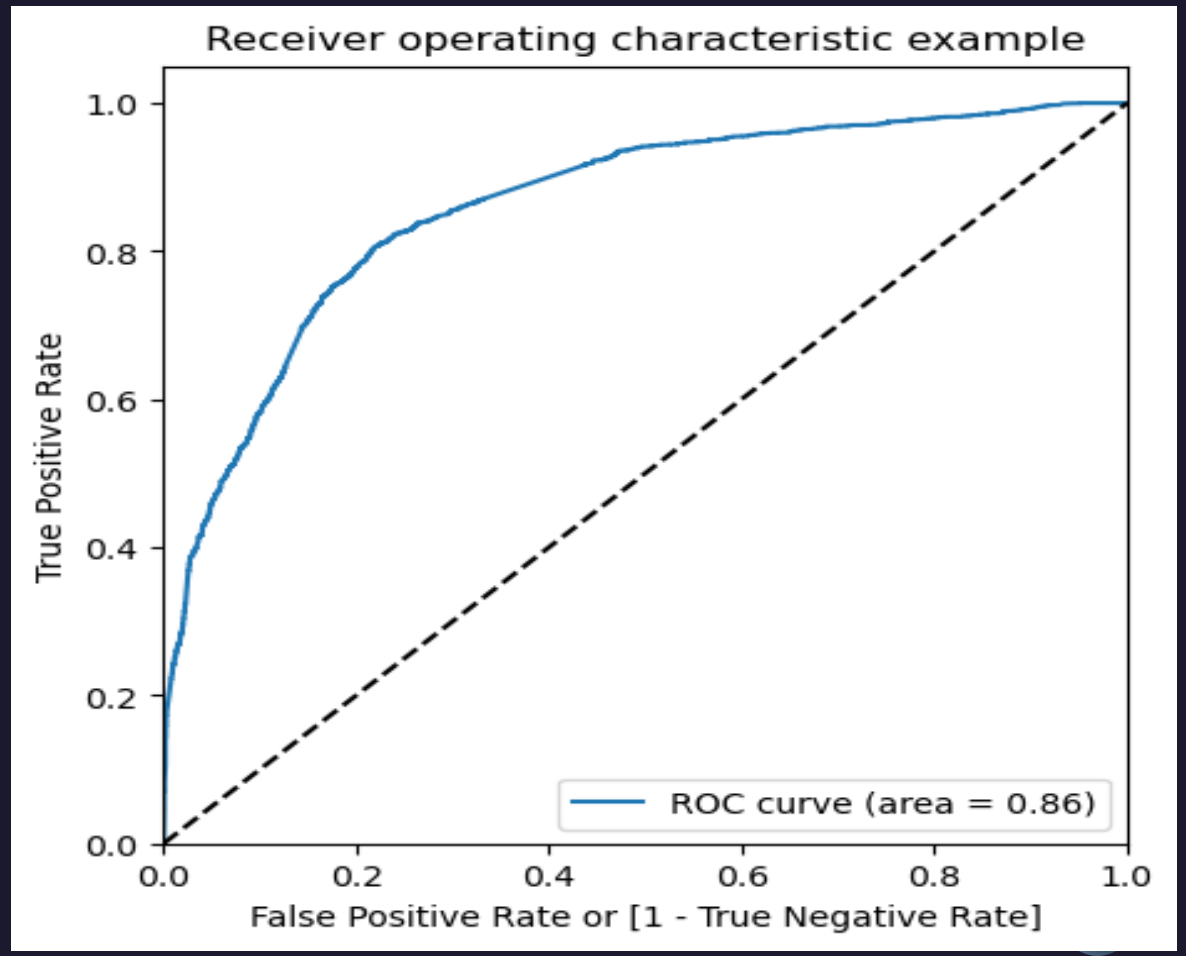
	coef	std err	z	P>|z|
const	0.2040	0.196	1.043	0.297
TotalVisits	11.1489	2.665	4.184	0.000
Total Time Spent on Website	4.4223	0.185	23.899	0.000
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Model Evaluation

PLOTTING THE ROC CURVE

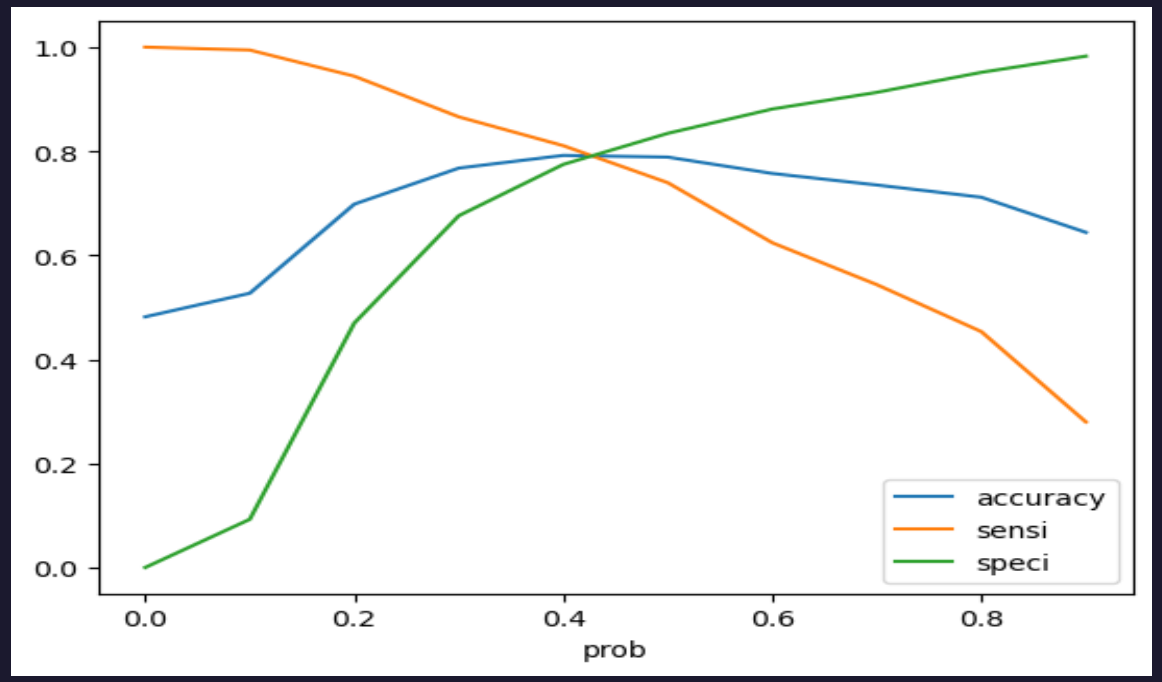
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The area under the curve of the ROC is 0.86 which implies quite good measure. It depicts to have a good model. Let's also check the sensitivity and specificity trade-off to find the optimal cutoff point.

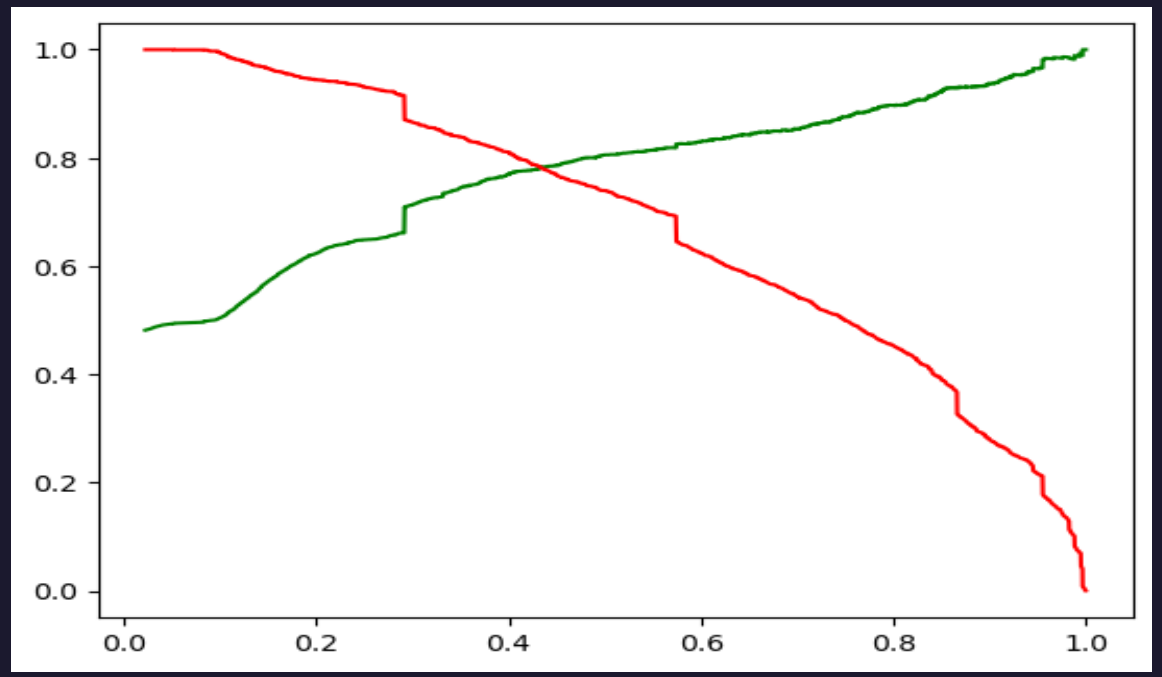


Model Evaluation

FINDING OPTIMAL CUTOFF POINT

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- After calculating the tradeoff between Precision and recall, we got the value 0.42
- This means, we can safely choose to consider any Prospect Lead with Conversion Probability higher than 42 % to be a hot Lead





Observations-

TRAIN DATA

After doing the model evaluation, below are the key values we have observed on training data.

- Accuracy : 80%
- Sensitivity : 77%
- Specificity : 80%

TEST DATA

After doing the model evaluation, below are the key values we have observed on testing data.

- Accuracy : 80%
- Sensitivity : 77%
- Specificity : 80%

FINAL KEY FEATURES

- 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 based on observations

As we can see there are a lot of leads generated in the initial stage but only a few of them come out as paying customers.

To increase the percentage of converting leads, we can nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

- ✓ For this we can follow-
- ❑ Sort out the best prospects from the leads which we have generated.
- ❑ 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' are the parameters which contribute most in the process to convert a lead.
- ❑ Then, we need to keep a track of leads handy so that we can inform them about new courses, services, job offers and future higher studies.
- ❑ Need to build proper plan to understand the needs of each lead, this will ultimately help us to go a long way to capture the leads as prospects.
- ❑ Focus on converted leads more and give more importance to them.
- ❑ Hold question-answer sessions with leads to extract the right information we require from them.
- ❑ Make more follow-up sessions with the leads to understand their intention and thinking behind joining.

Summary

We observed 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 interpret 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. Maximum leads are unemployed and maximum conversion is shown in working professional.

Thank You