LEAD SCORING CASE STUDY

CONTENTS

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.

--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%.

--its lead conversion rate is very poor.

-- the company wishes to identify the most potential leads, also known as ‘Hot Leads’

BUSINESS OBJECTIVE

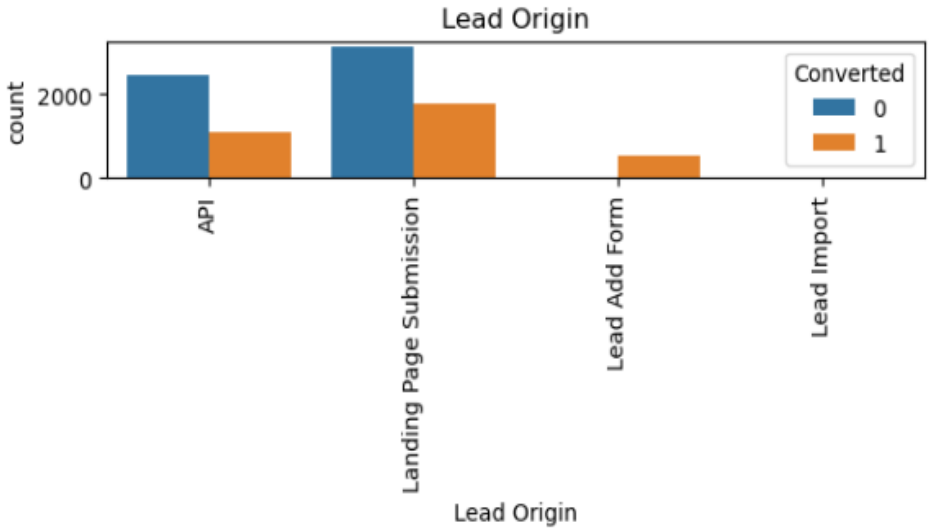
-- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

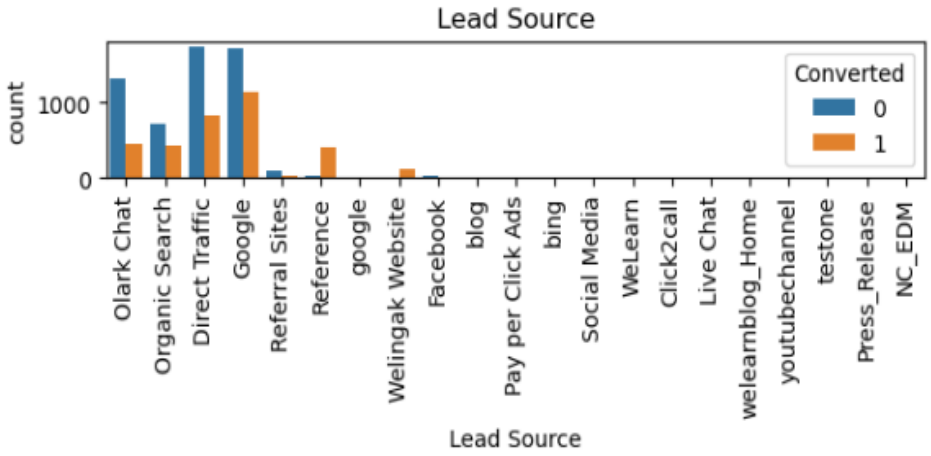
-- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

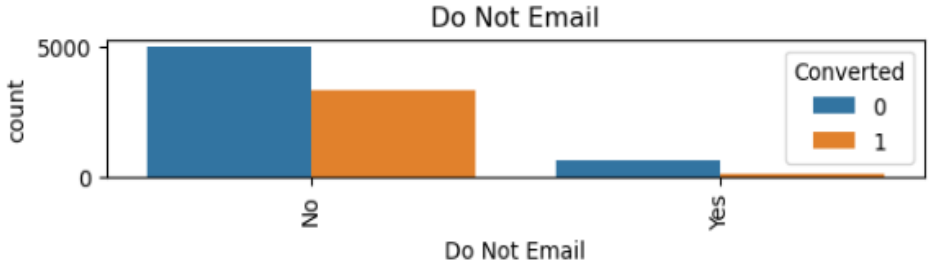
Problem Approach

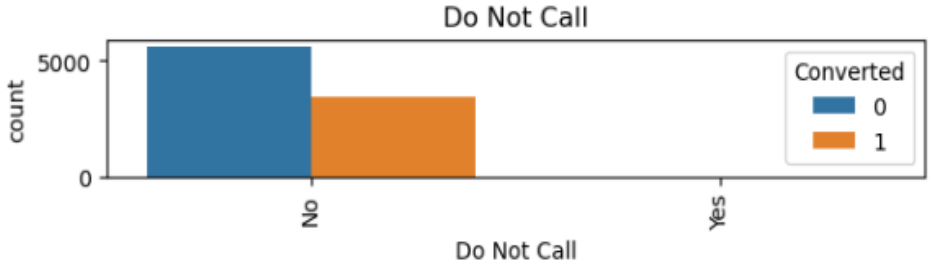
Step-1 : Reading and Understanding data
Step-2 : Visualization of the data
Step-3 : Data Preparation
Step-4 : Splitting data into Training and Test sets
Step-5 : Building Logistic Model
Step-6 : Prediction
Step-7 : Model Evaluation
Step-8 : Optimize Cut off (ROC Curve)
Step-9 : Prediction on Test
step-10: set Precision-Recall
step-11: Precision and recall tradeoff
step-12: Prediction on Test set
step-13: checking over all accuracy
step-14: Conclusion

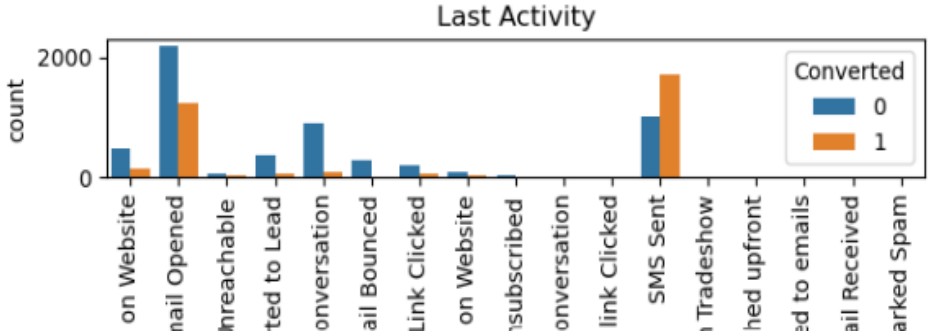
EDA- Exploratory data analysis

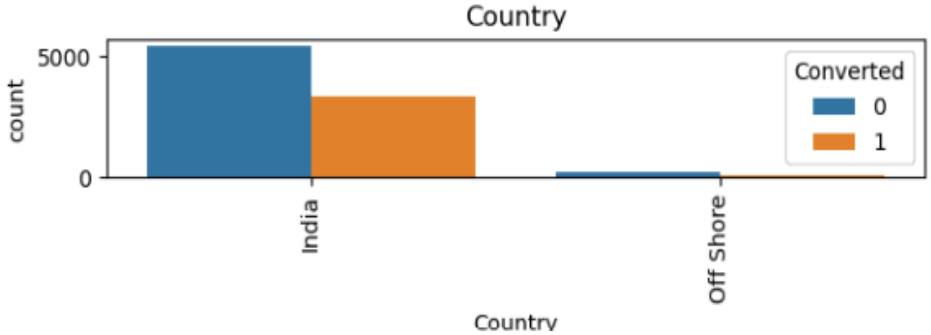




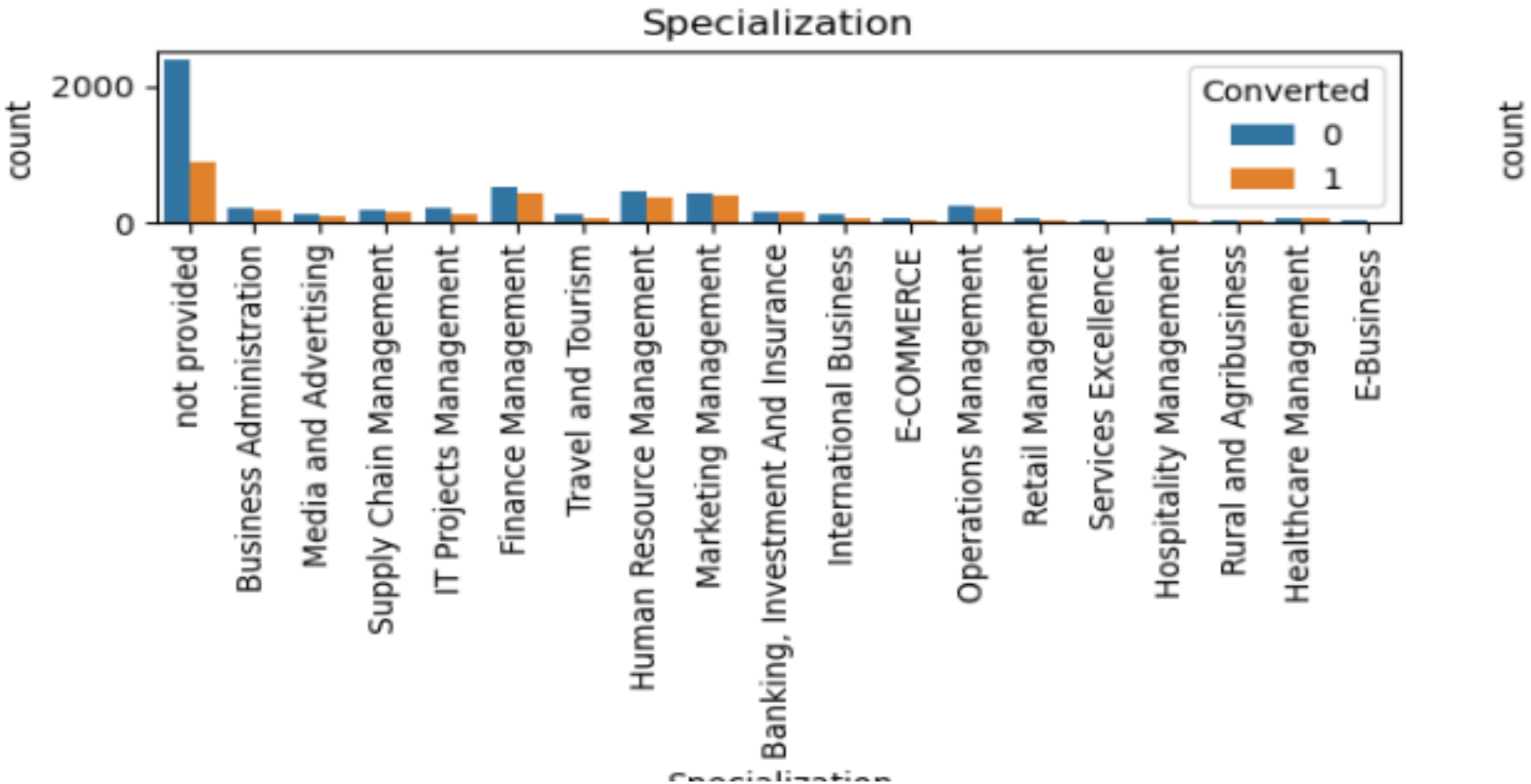






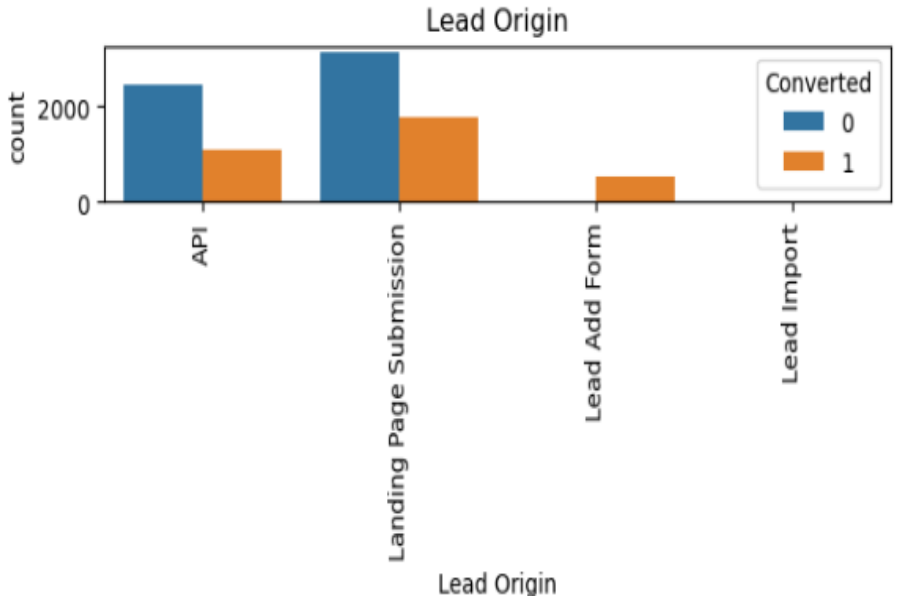


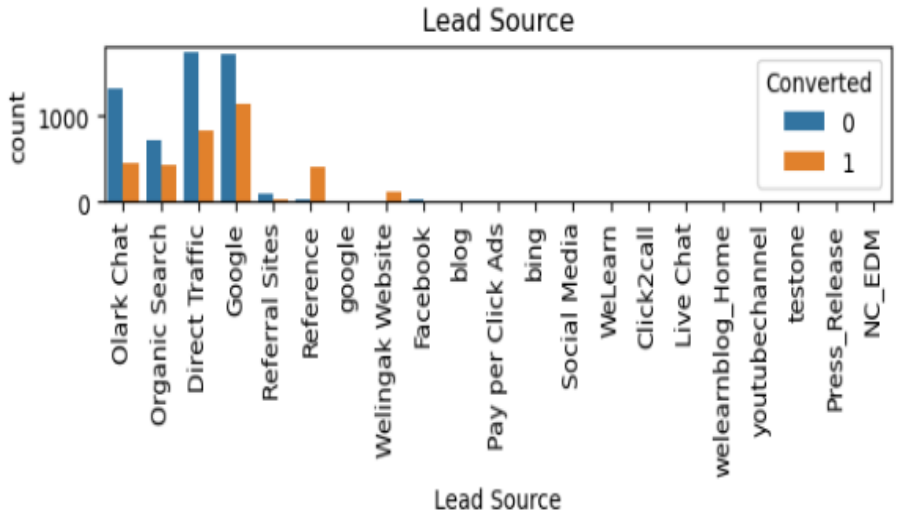
With Specialisation information provided by leads we can infere HR, Finance & Marketing management specializations are high probability to convert



LEAD SOURCE & LEAD ORIGIN

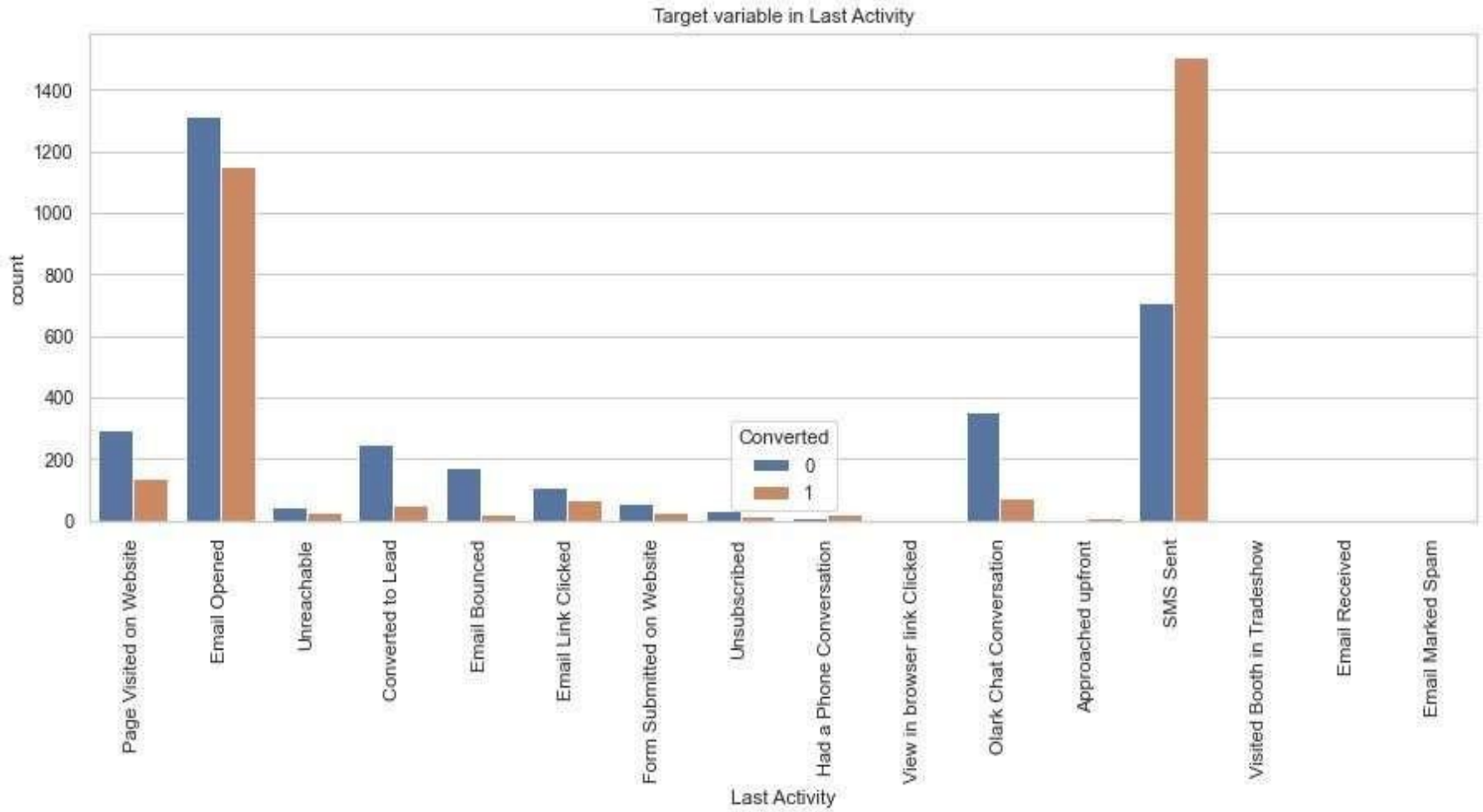
With Lead origin and Lead score information provided by leads we can infer Landing Page submission , Lead Add from , Google source , direct traffic and reference is something to be concentrated





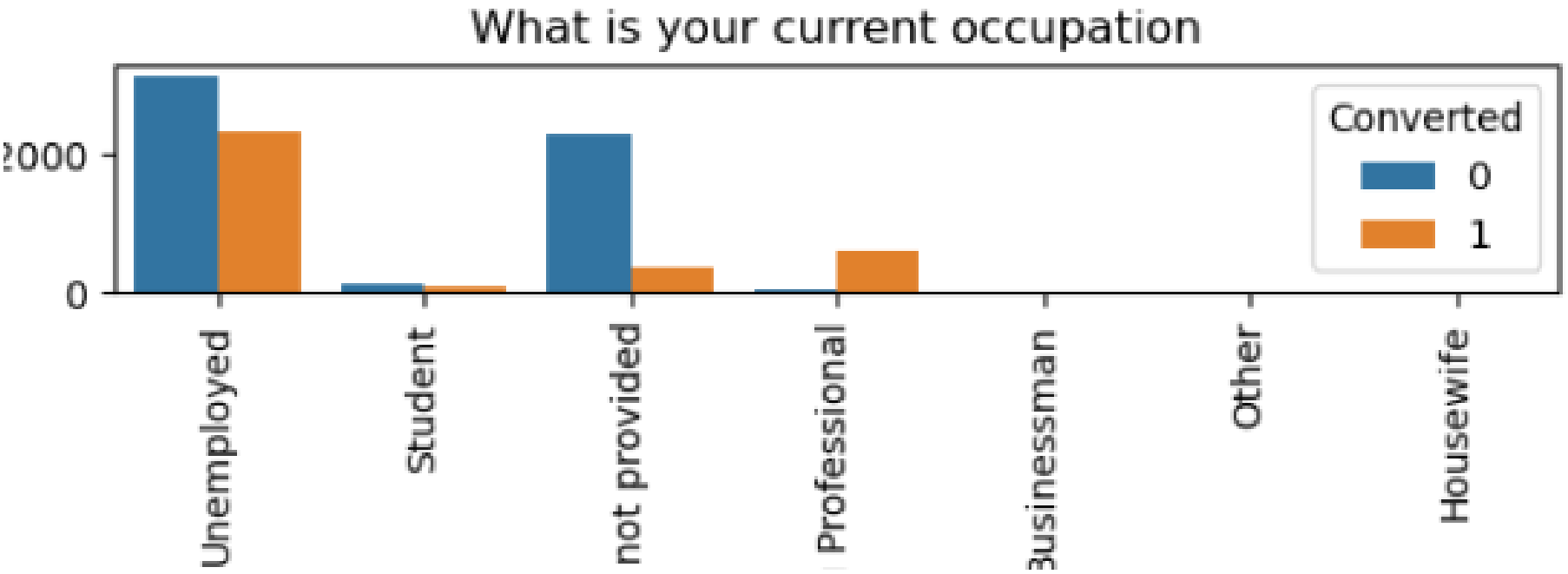
Last activity information

SMS sent leads . Email opened leads have high probability to turning to hot leads

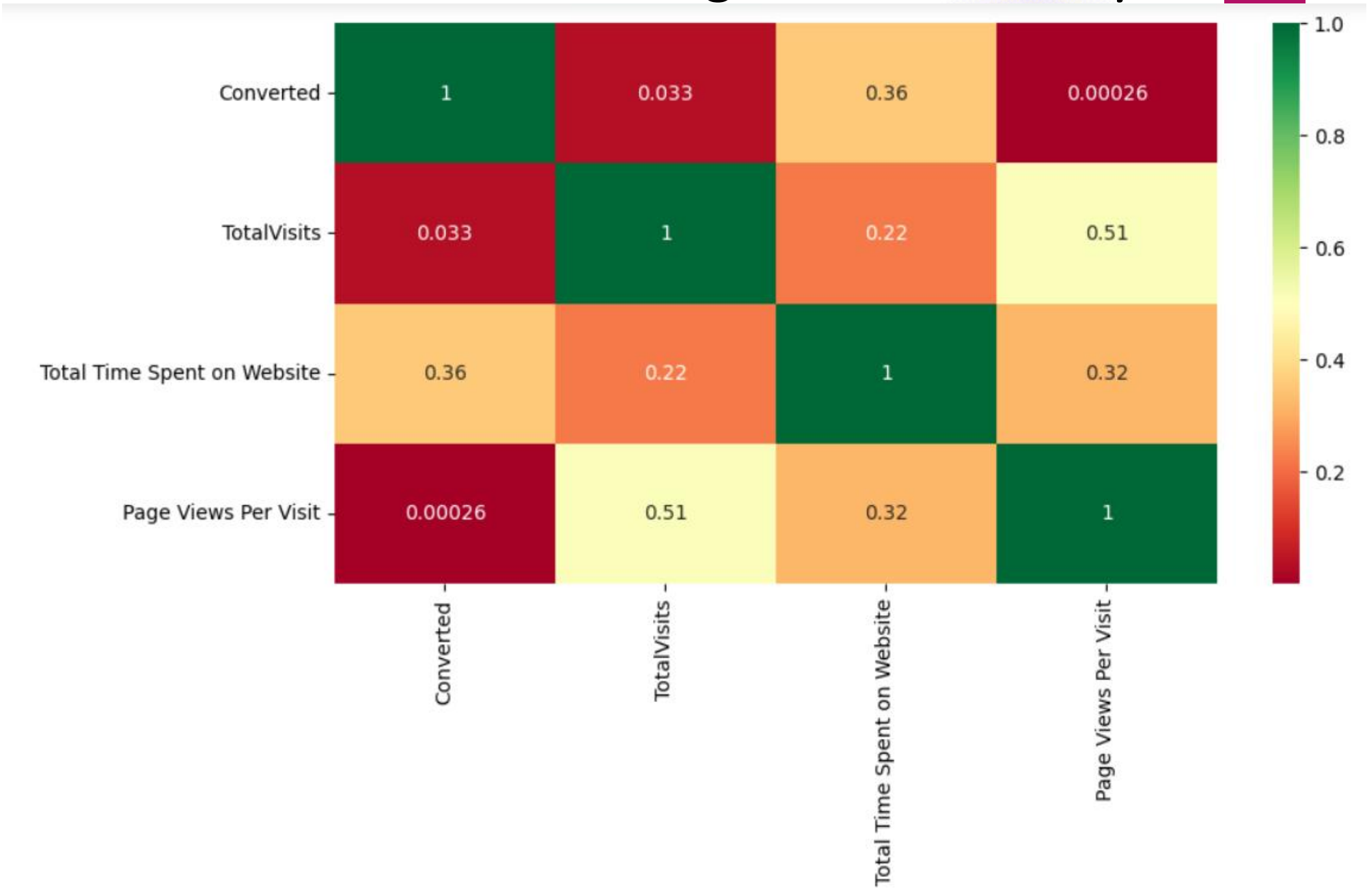


Current occupation status of leads

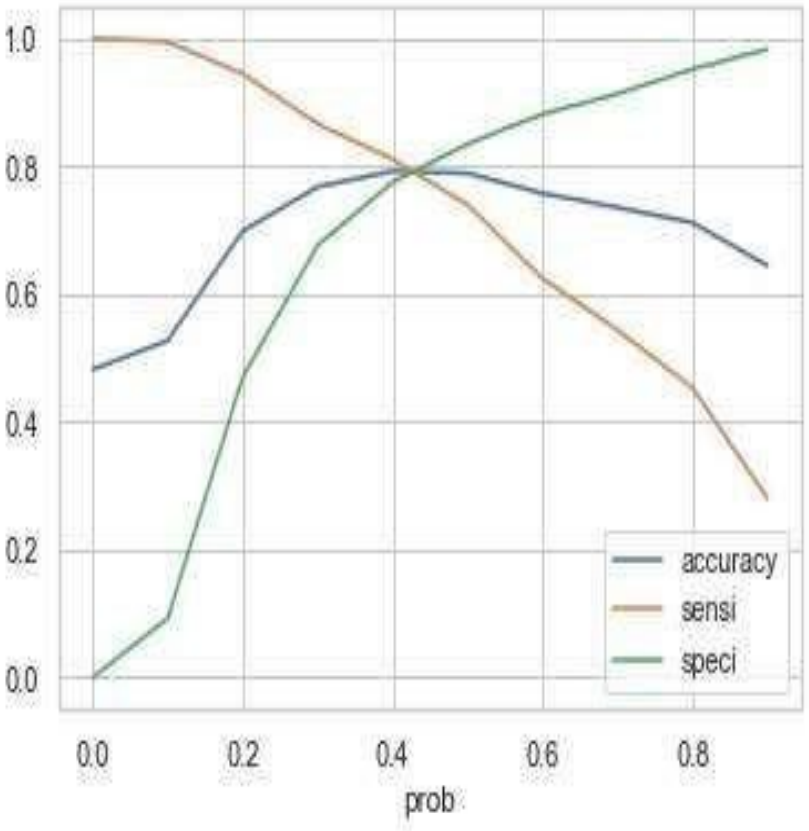
Leads with unemployed status are high probable to turn as hot leads

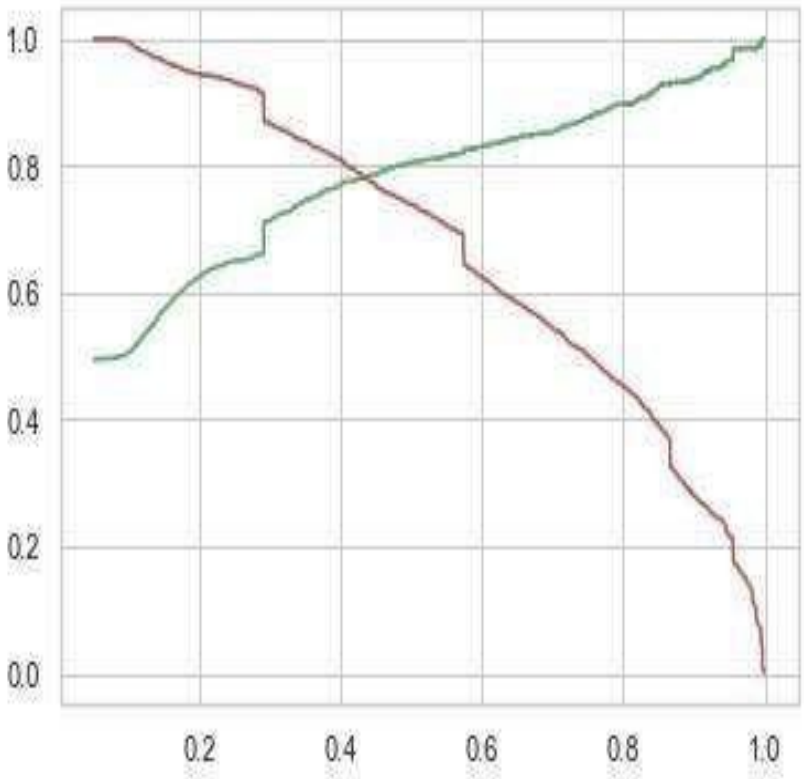


Correlation and checking Multicollinearity



MODEL EALUATION





OBSERVATIONS

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

Train Data:

Accuracy : 82%
Sensitivity : 76.3%
Specificity : 80%

Test Data:

Accuracy : 80.5%
Sensitivity : 75%
Specificity : 73%

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 inference 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.