

# LEAD SCORING CASE STUDY

---



Submitted by - Aanand Purbey,  
Priyanshu Pandey, and Priyanka noni



# 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.

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%.





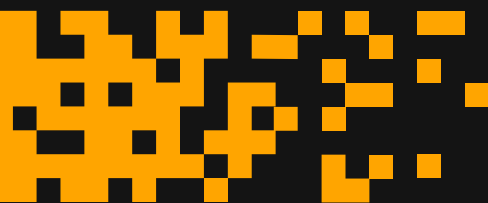
# GOAL OF CASE STUDY

There are quite a few goals for this case study:

1. 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.
2. There are some more problems presented by the company which your 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. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.



machine learning



# PROBLEM SOLVING STRETEGY

D



## DATA SOURCING

1. Read the data from source and convert data and clean for suitable analysis and remove any duplicate entries.
2. Outlier treatment
3. Perform EDA
4. Standardization

T



## TRAIN AND TEST

1. Feature scaling and numeric data
2. Splitting data into train and test set.

M



## MODEL BUILDING

1. Feature selection using REF
2. Determine the optimal model using logistic regression
3. Calculate the metrics like accuracy, sensitivity, specificity, precision and recall, and evaluate the model.

R




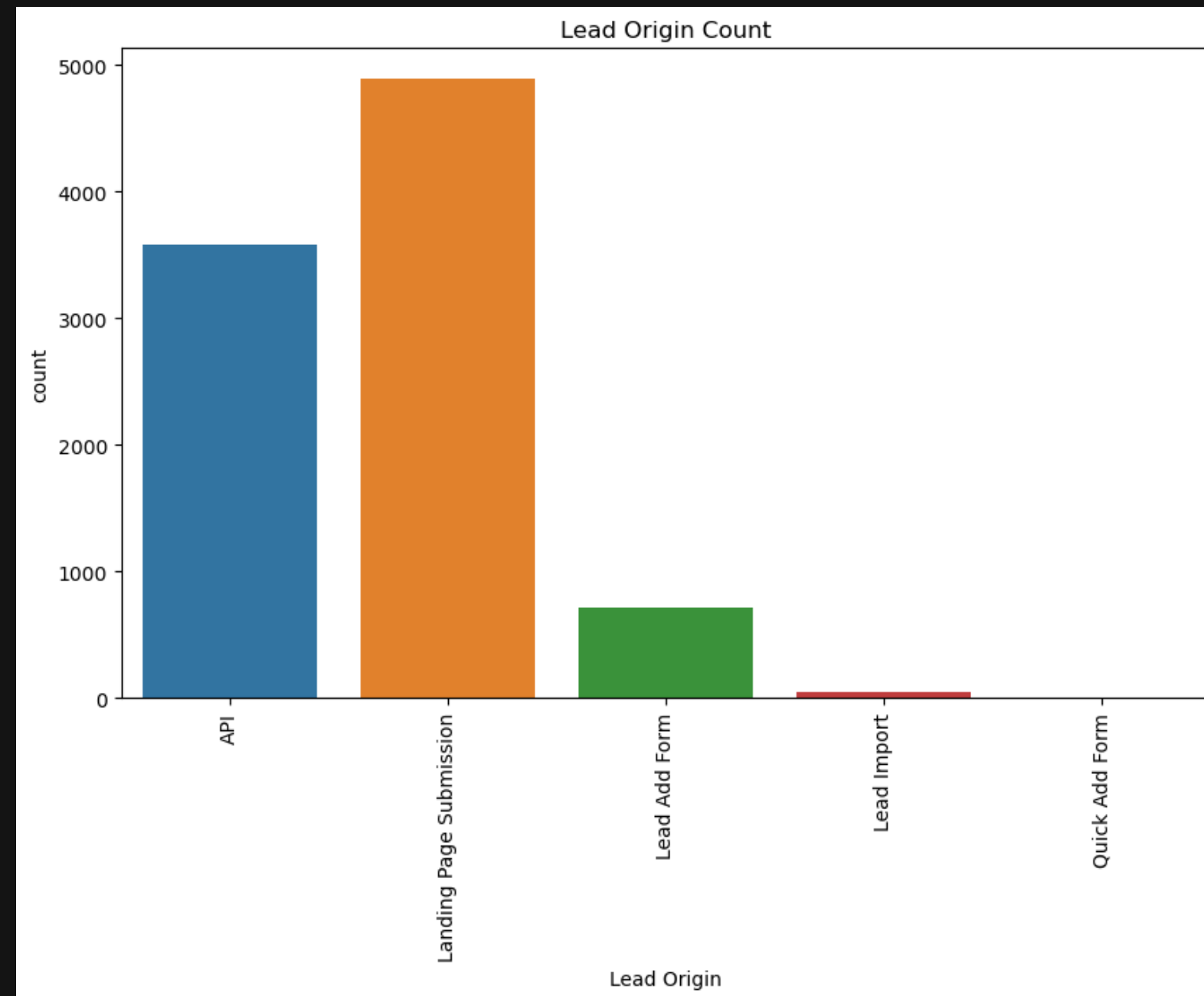
## RESULT

1. Determine the Lead score and check if target predictions amounts to 80% conversion rate.
2. Evaluate the final prediction on the test set using cut off threshold.



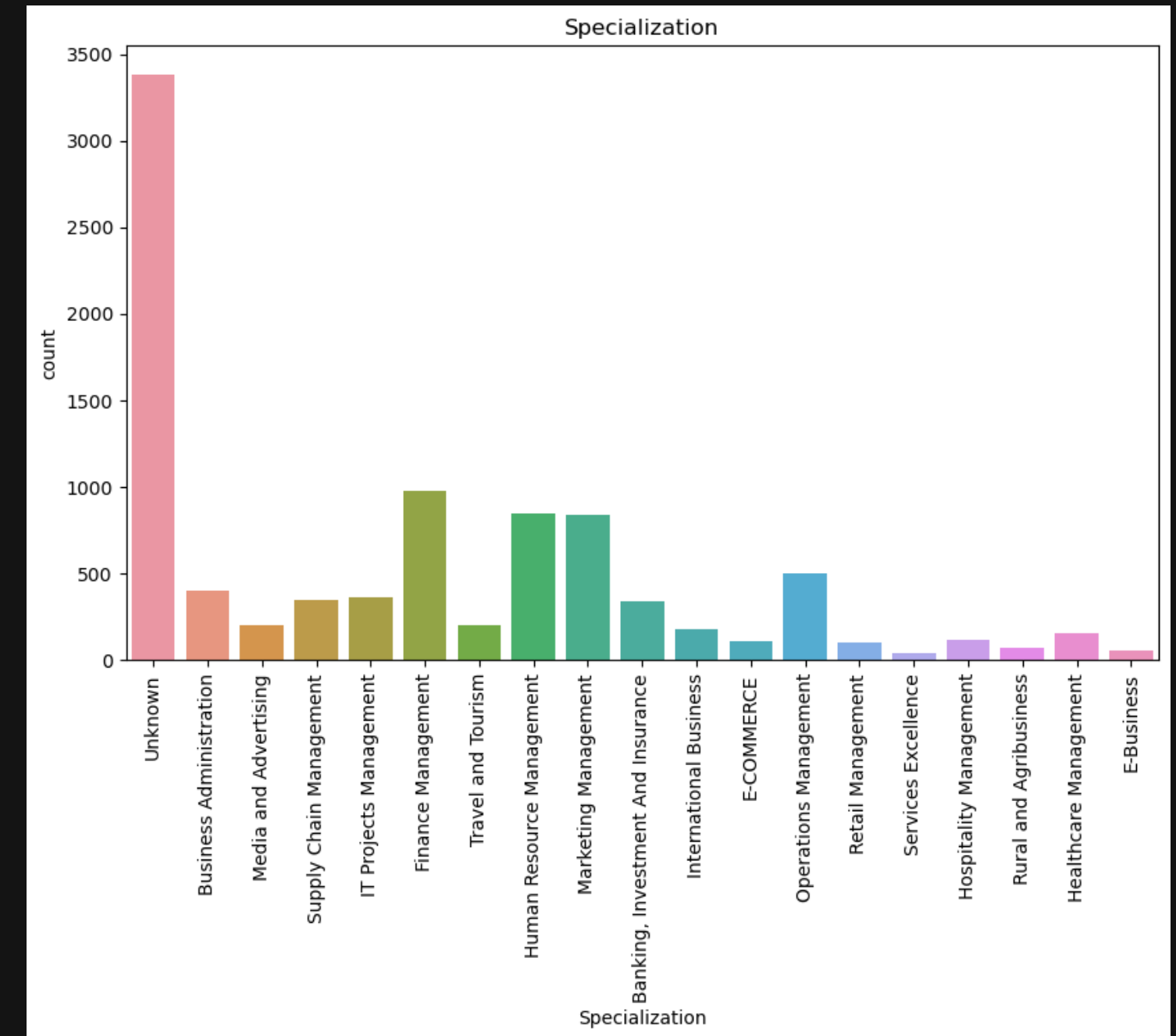
# PERFORMING EDA

 In this we have plotted the Lead origin count

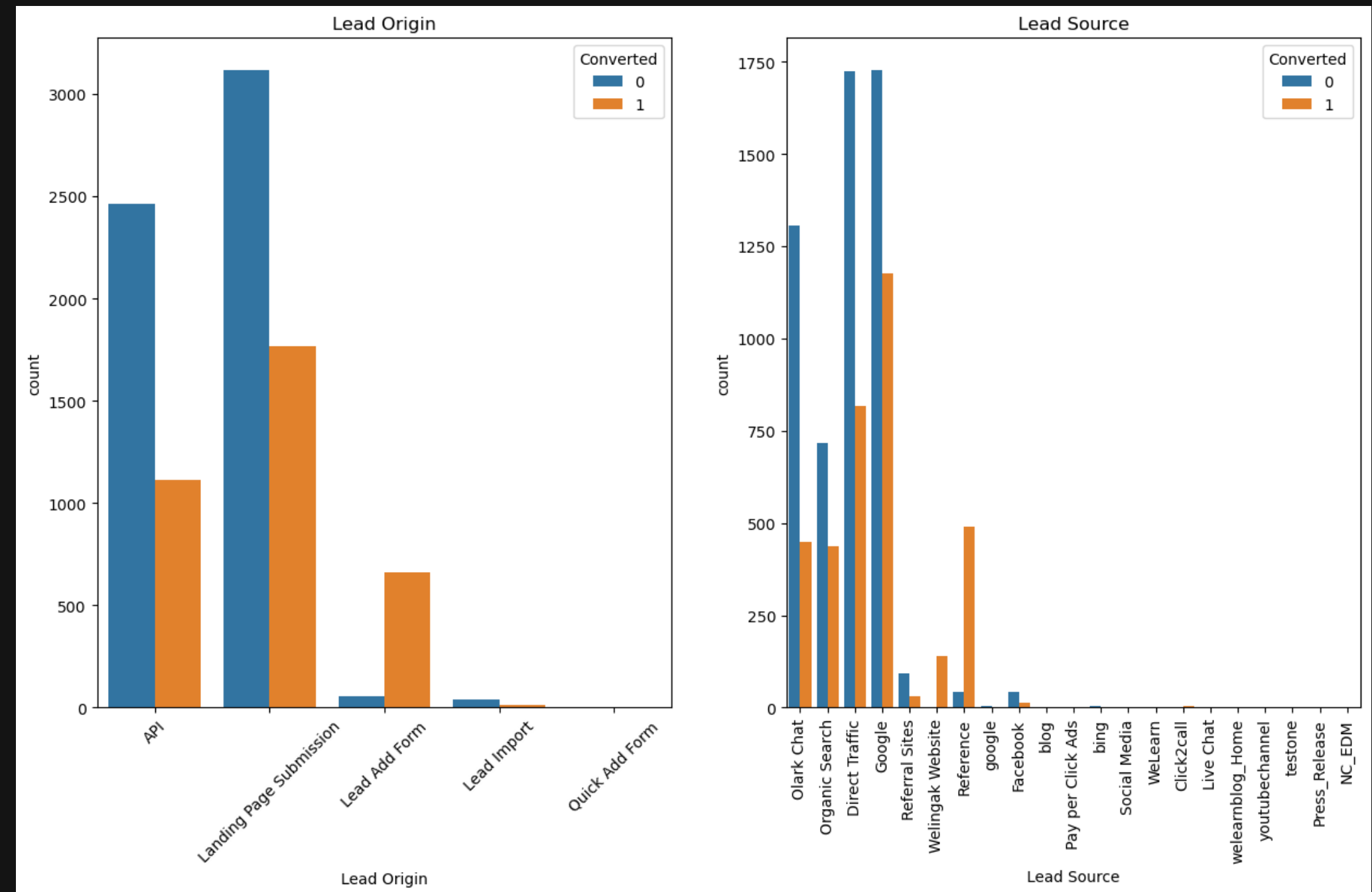


# SPECIALIZATION

1. From the specialization plotting we can see that a lot of people have not filled their specialization
2. People from the management profession in any genre are most likely to be lead.

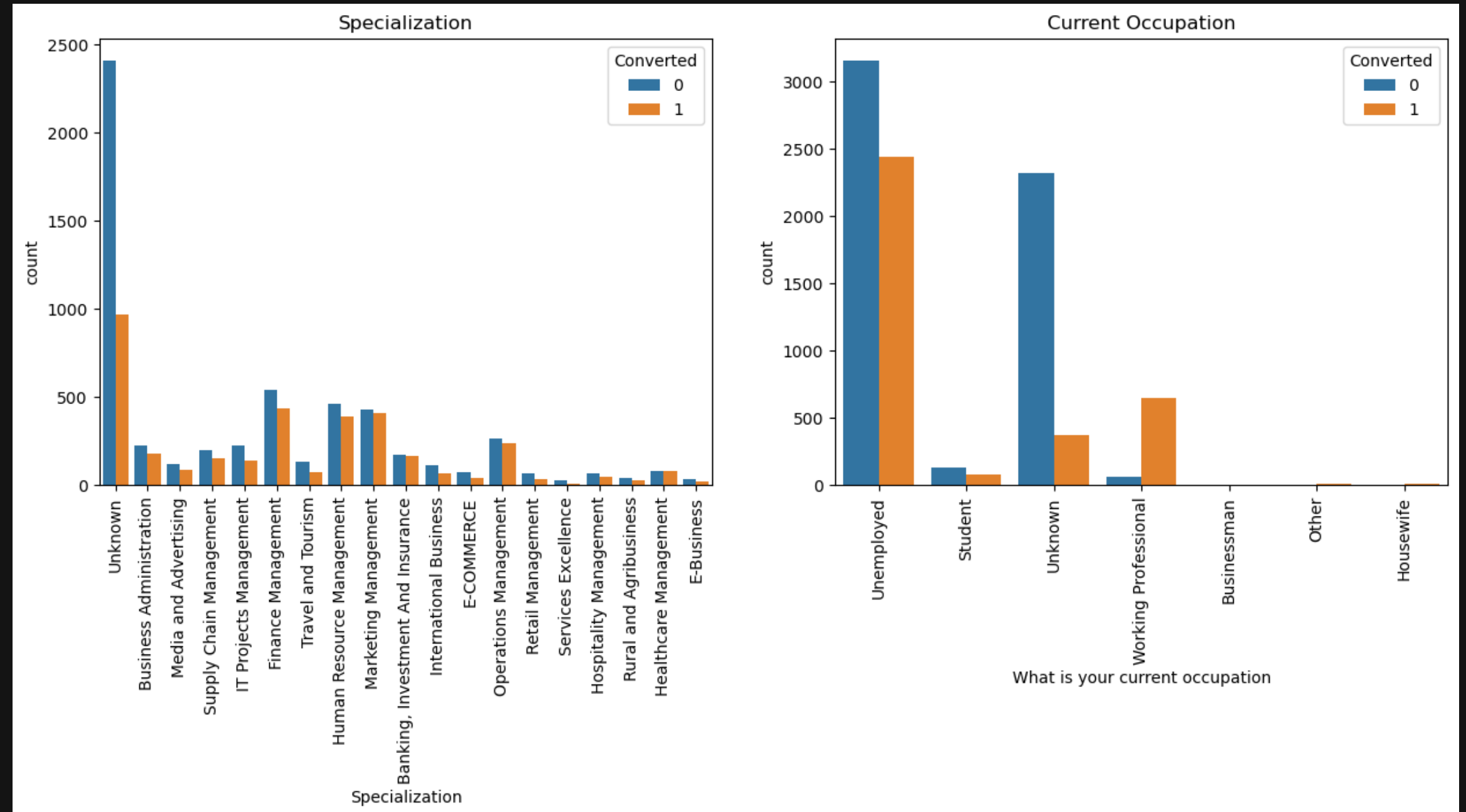


- Added form is more effective way to convert people but it is significantly less in count.
- Landing Page Submission has highest count of people who did not convert. Still it is second best effective way to convert people.
- Reference helps most in converting people followed by Google.
- Olark chat and referral sites perform lowest in conversion of people.



- Management profession like finance ,HR, Marketing and Operation have very good count of conversion compared to other specializations.

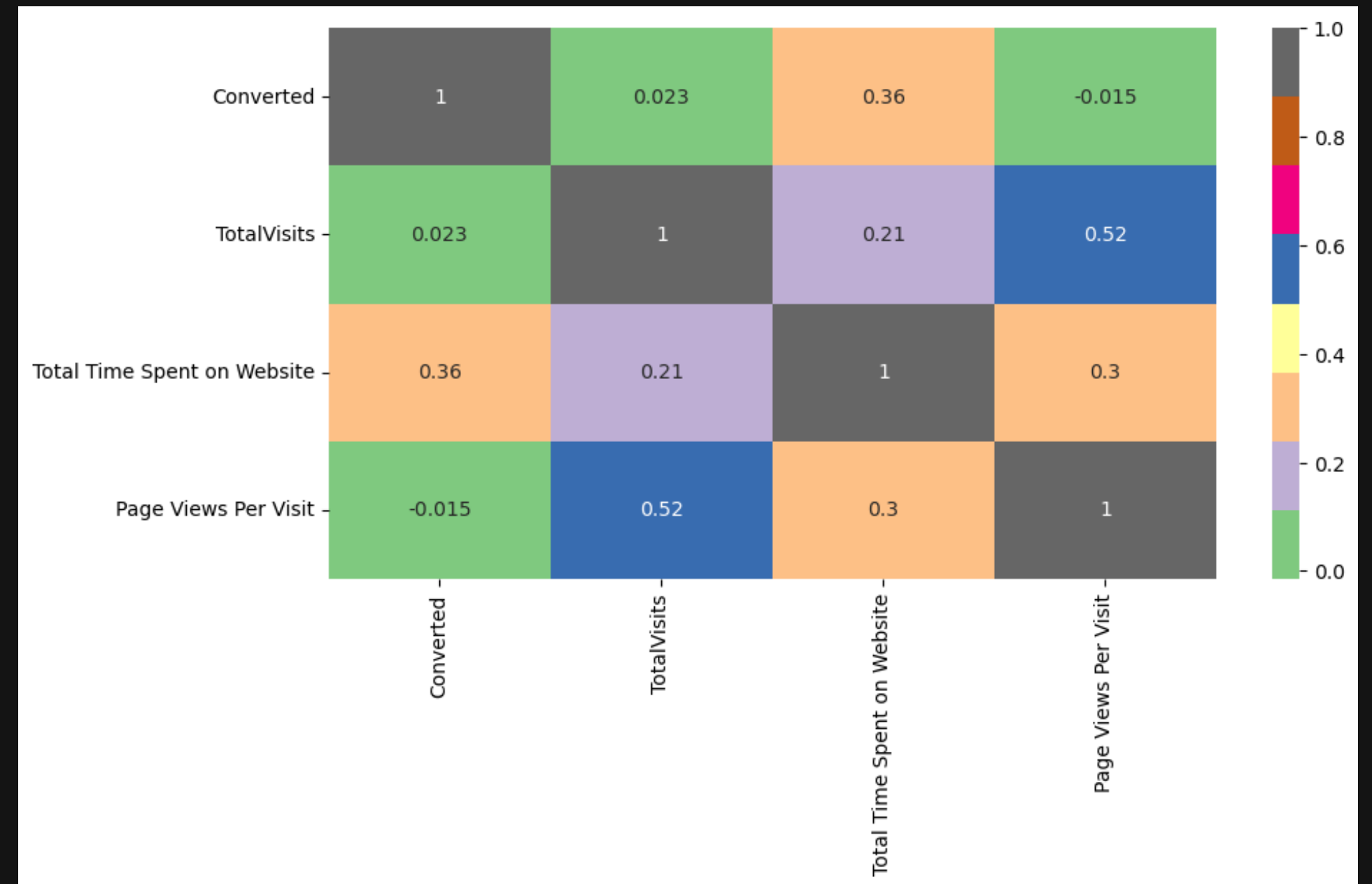
- Working profession shows excellent count of conversion whereas unemployed people have hogher count for being converted.





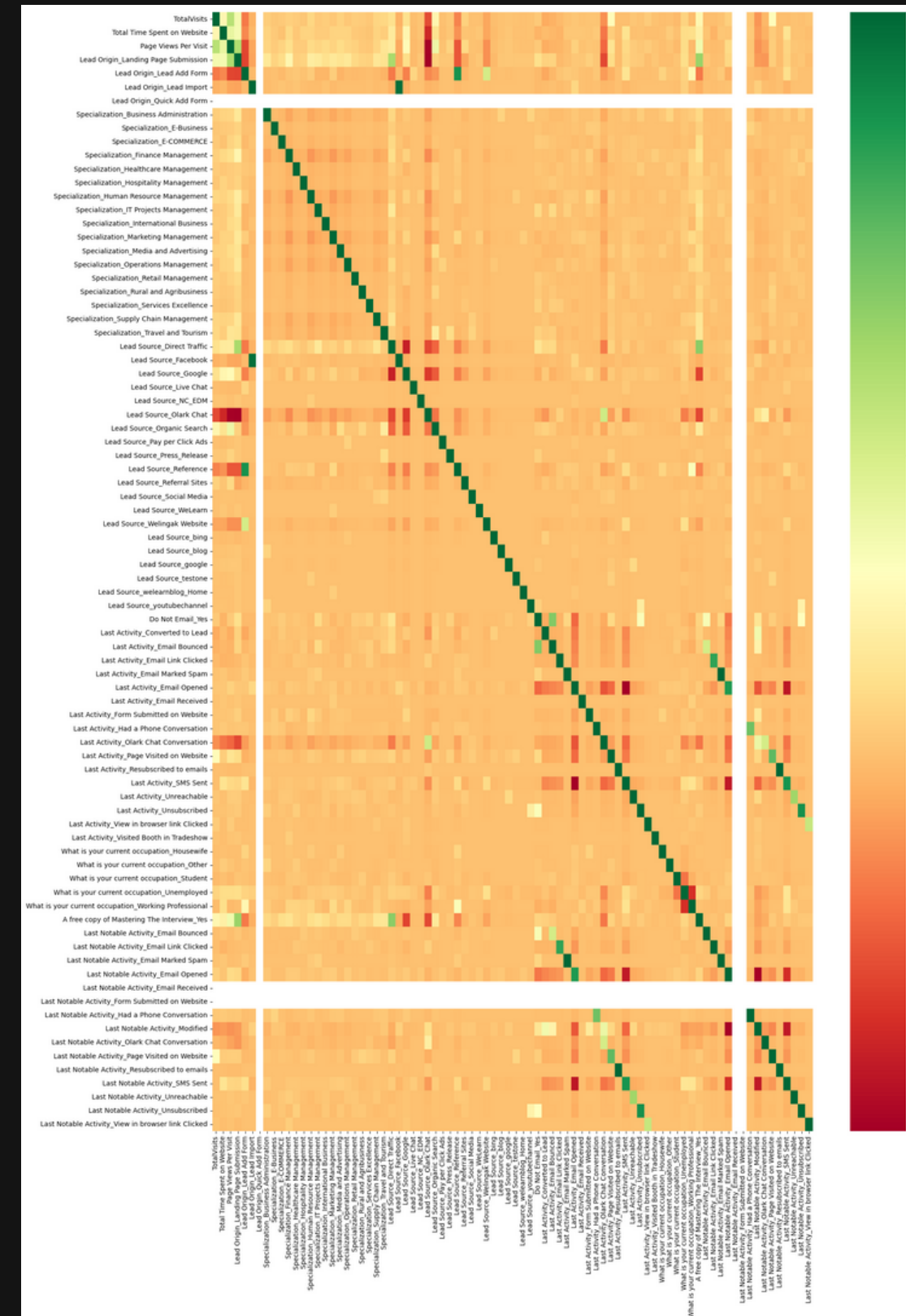
# MULTI VARIATE ANALYSIS

- There are 0.36 correlation of 'Total Time Spent on Website' with target variable 'Converted'.
- 'Page Views Per Visit' have -0.015 correlation with target variable.



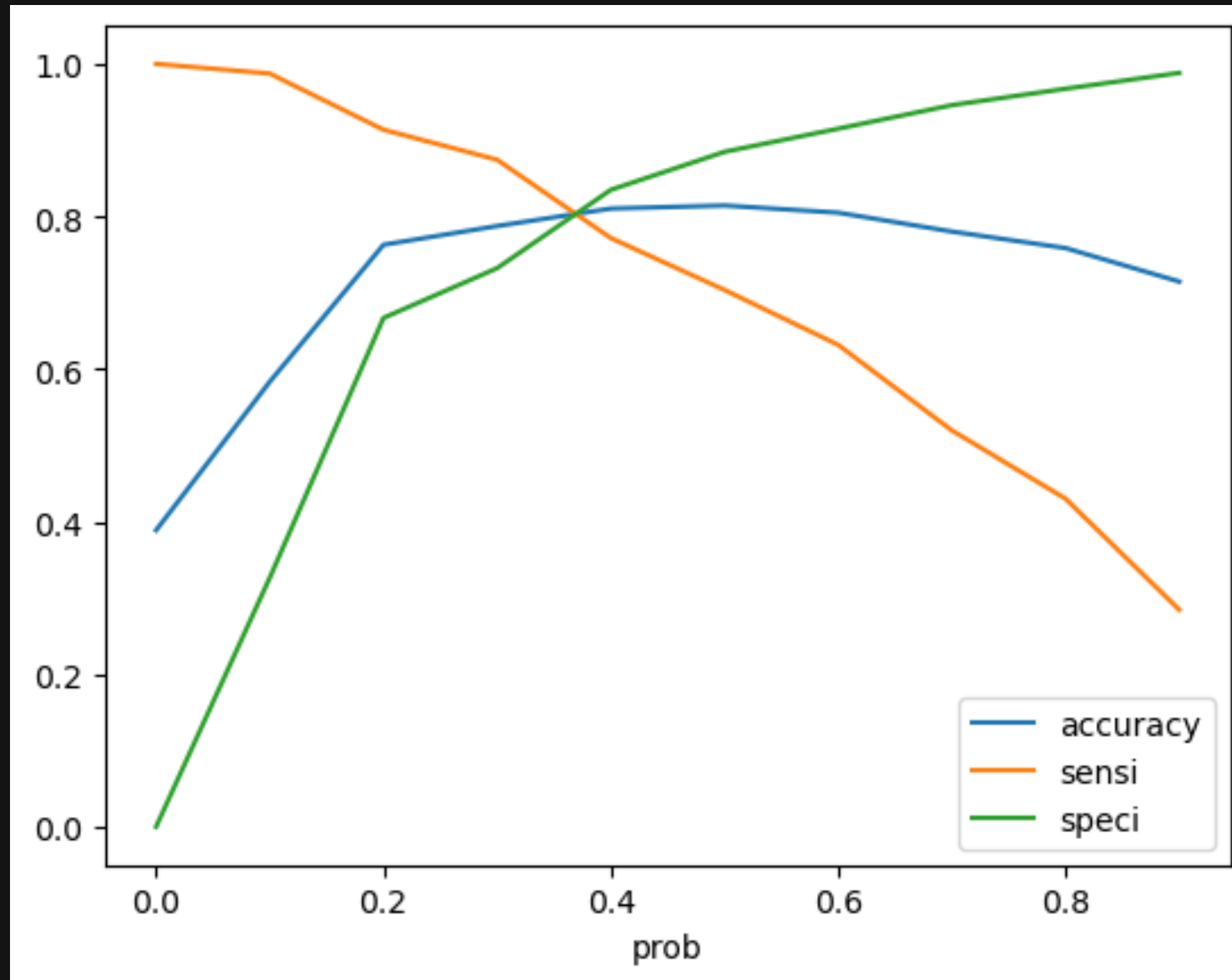
# HEAT MAP

- Since there are a lot of variables it is difficult to drop variable. We will do it after RFE.!



# MODEL EVALUATION

The graph depicts an optimal cut off of 0.37 based on accuracy, sensitivity and specificity.





# CONCLUSION

1. While we have checked both sensitivity-specificity as well as precision and Recall metric, we have considered the optimal cut off based on sensitivity and specificity for calculating the final prediction.
2. Accuracy, sensitivity and specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.
3. Also the lead score calculated shows the conversion rate on the final predicted model is around 80% and 79% in the test set
4. Hence overall model seems to be good.

