LEAD SCORING CASE STUDY- LOGISTIC REGRESSION

TEAM MEMBERS

- SUJAY ARYAL
- ANURABHA ROY
- MANISH SHARMA

CONTENTS

- Problem Statement

- Business Goals

- Problem Approach

- Exploratory Data Analysis (EDA)

- Model Building & Evaluation

- Observations & Results

- 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. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

- After which 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 GOALS

- To assign a lead score between 0 -100 to each of the leads inorder to identify the "Hot leads" and increase their conversion rate.

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

- To build a model by assigning lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

PROBLEM APPROACH

- Import dataset and unerstanding the dataframe.
- Data Cleaning & Preparation for EDA
- Exploratory Data Analysis (EDA)
- Dummy Variable creation
- Test-train datasets split & Feature scaling
- Correlations check between features
- Model Building (p-value, VIF, RFE squared) & Evaluation
- Final Predictions on Test Set

EXPLORATORY DATA ANALYSIS (EDA)

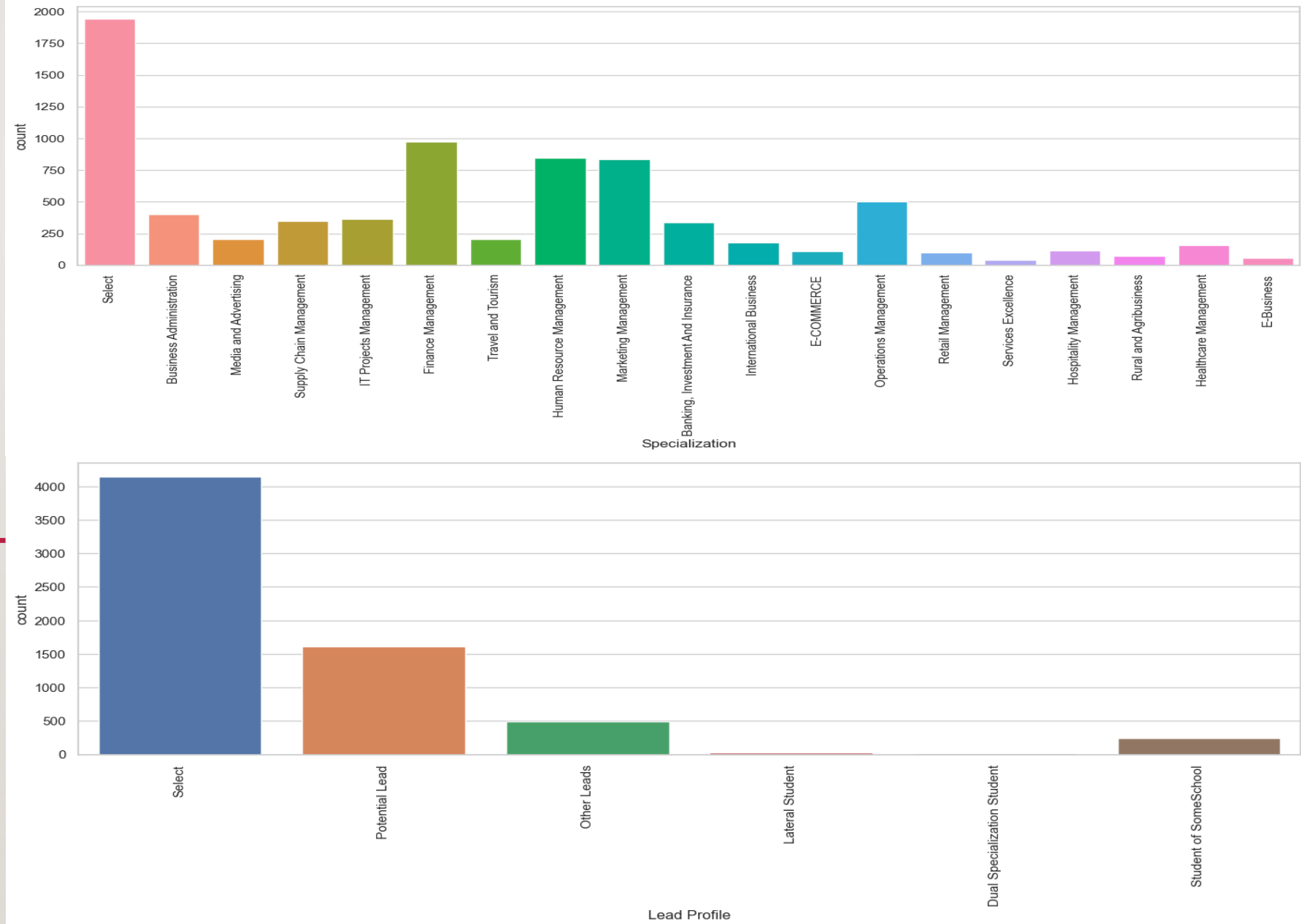There are a lot of 'Select' values present in many columns of categorical variables.

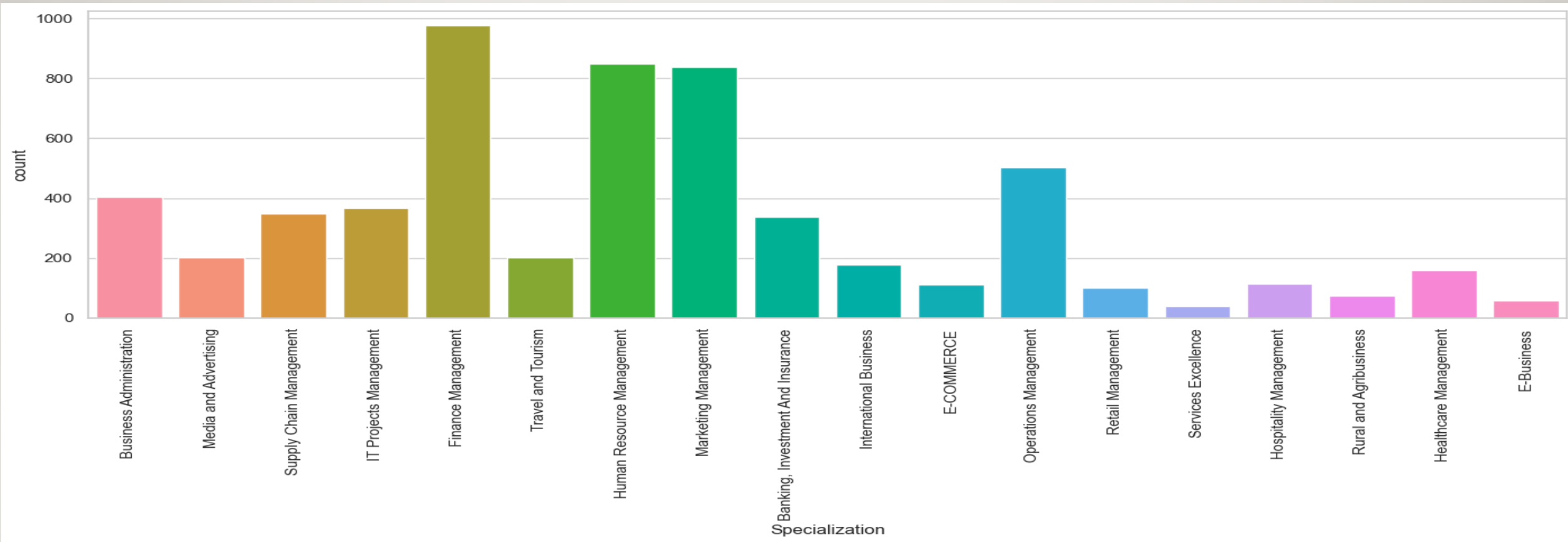It may be because the customer did not select any option from the list, hence it shows 'Select'.

'Select' values are as good as NULL. So, converting them to null values.

ANALYZING FEW COLUMNS WITH 'SELECT' VALUES & REPLACING THEM WITH NULL VALUES LATER
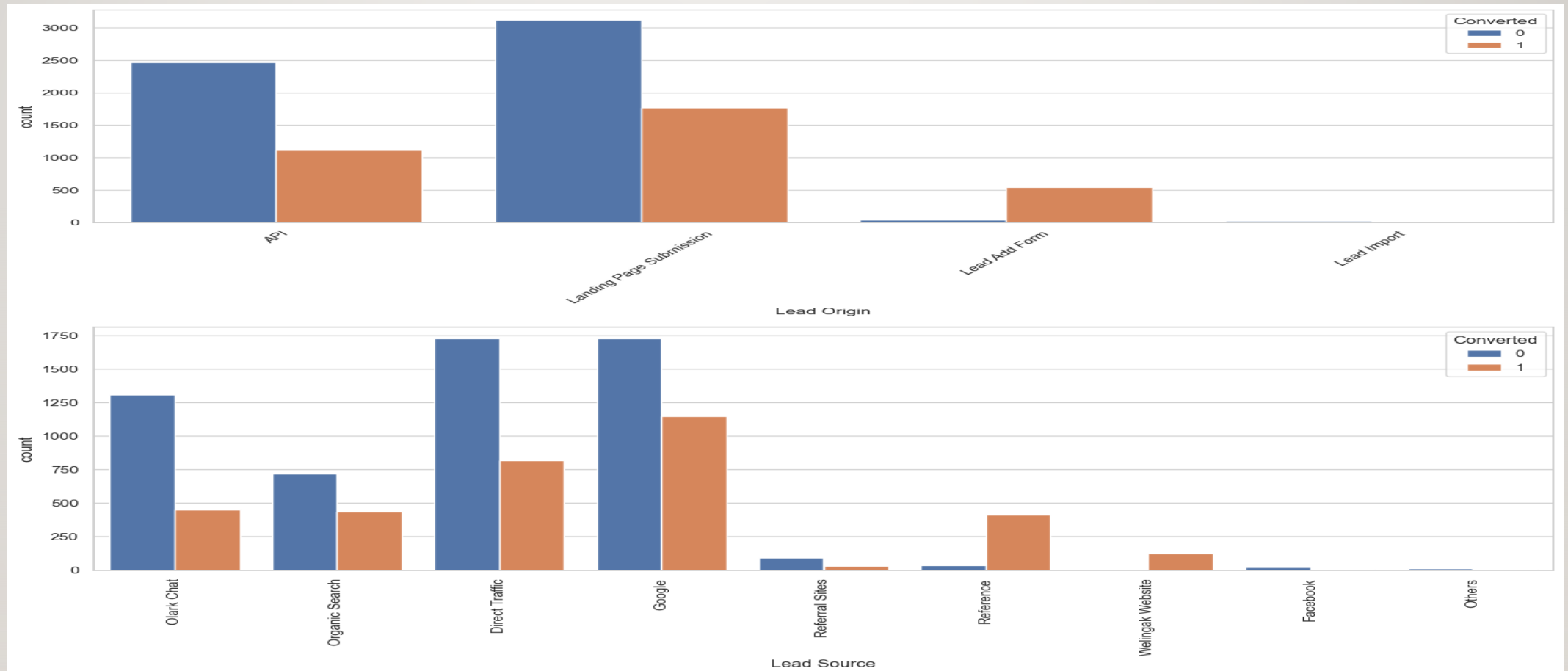
SPECIALIZATION FEATURE

- Leads from Finance Management, Human Resource Management, Marketing Management & Operations Management have high probability chances for successful conversions.

LEAD ORIGIN & LEAD SOURCE

- As seen, in Lead origin most people landing on page submission (Landing Page Submission) have highest number of leads.

- In Lead Source, leads coming via direct traffic and google have high probability of conversion.

LAST ACTIVITY
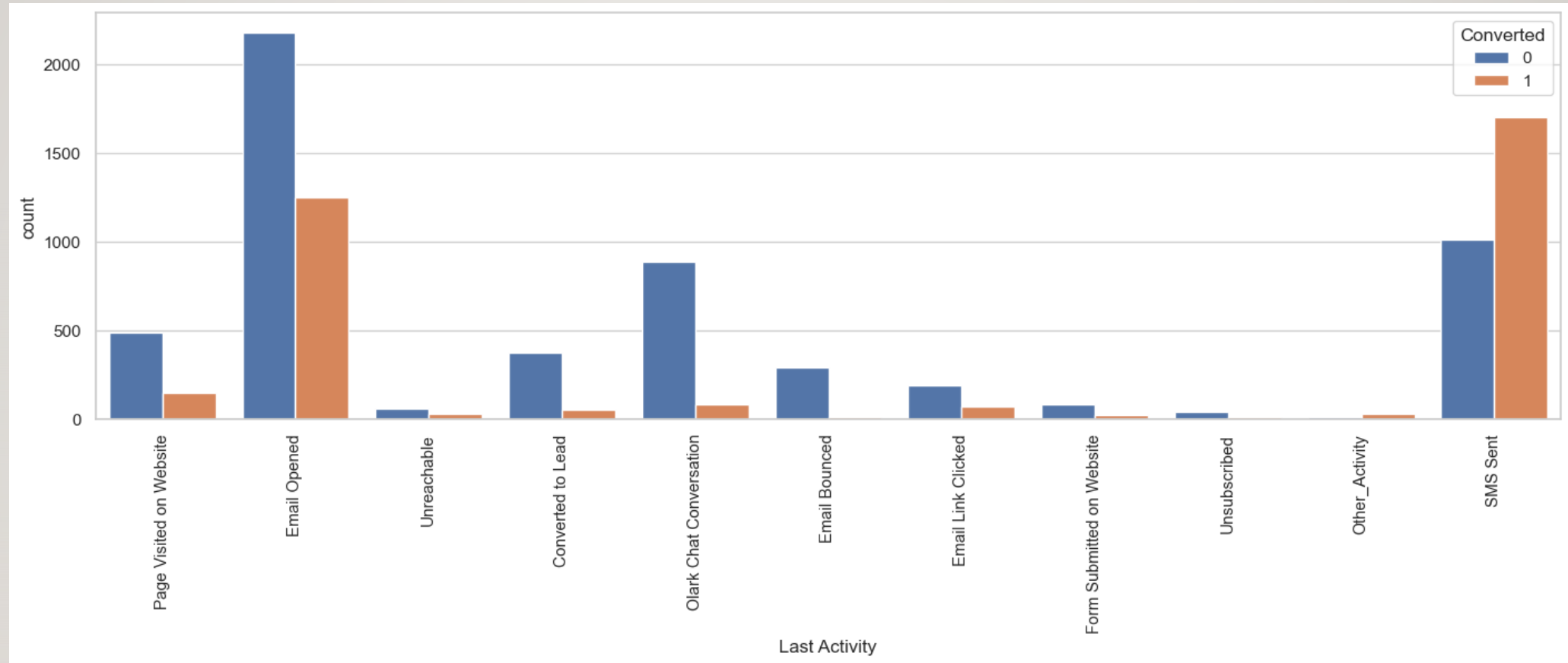
- Leads who have opened and checked Email (Email Opened value) have high chance of conversion.
- Also, if a SMS campaign (SMS Sent value) is organised it will help in attracting more leads and their successful conversion.

WHAT IS YOUR CURRENT OCCUPATION?
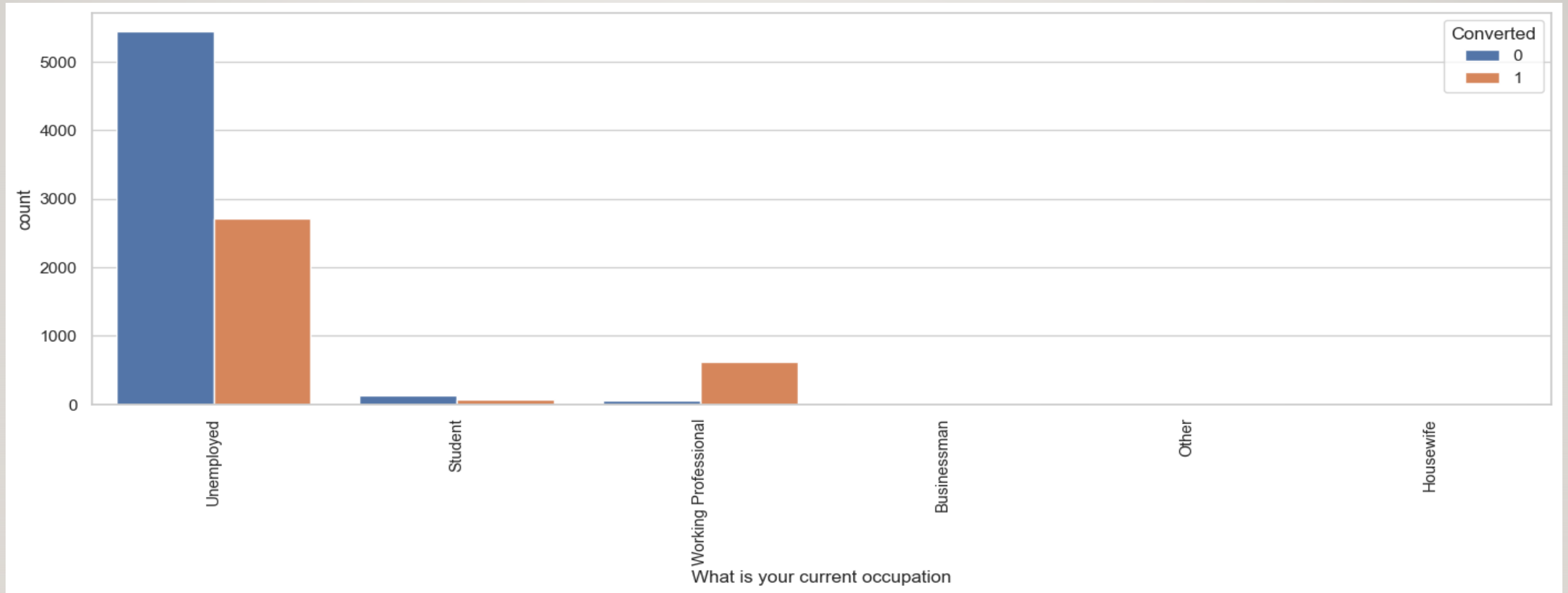
- Unemployed leads are highest but have around 35% conversion rate.
- Working Professional going for the course are more likely to opt for the course.

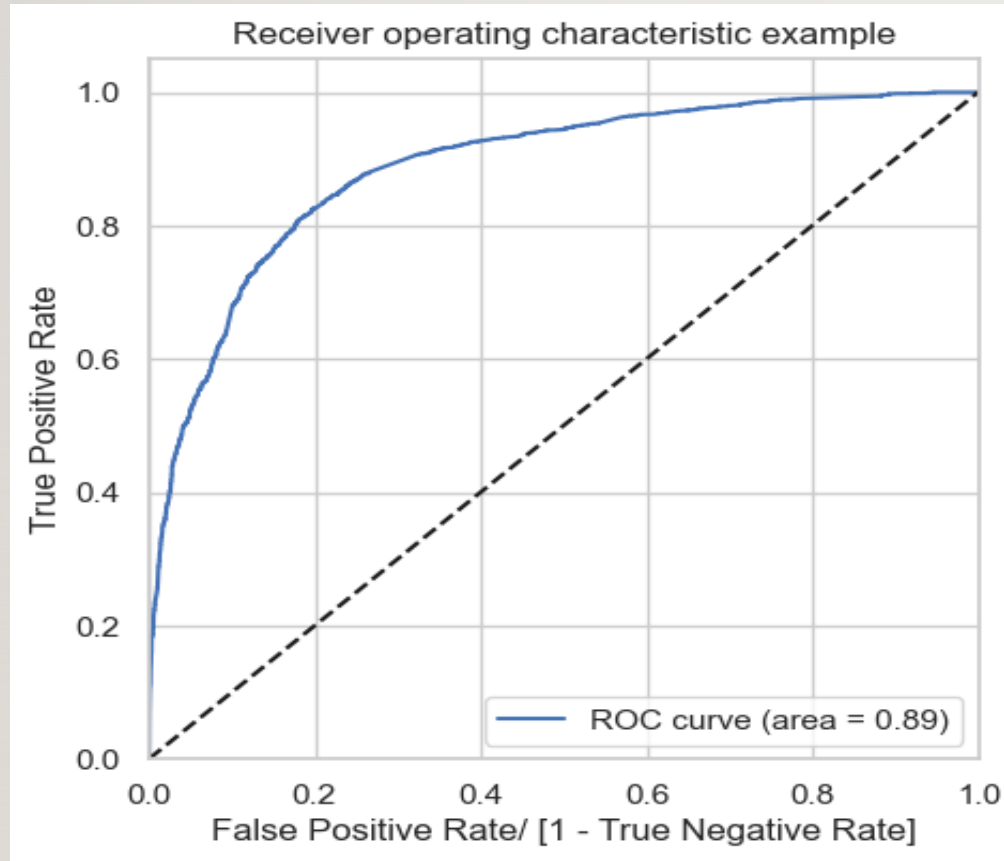So, focus should be on increasing leads for Working professional and help convert unemployed leads using apt strategy.

CORRELATION BETWEEN FEATURES

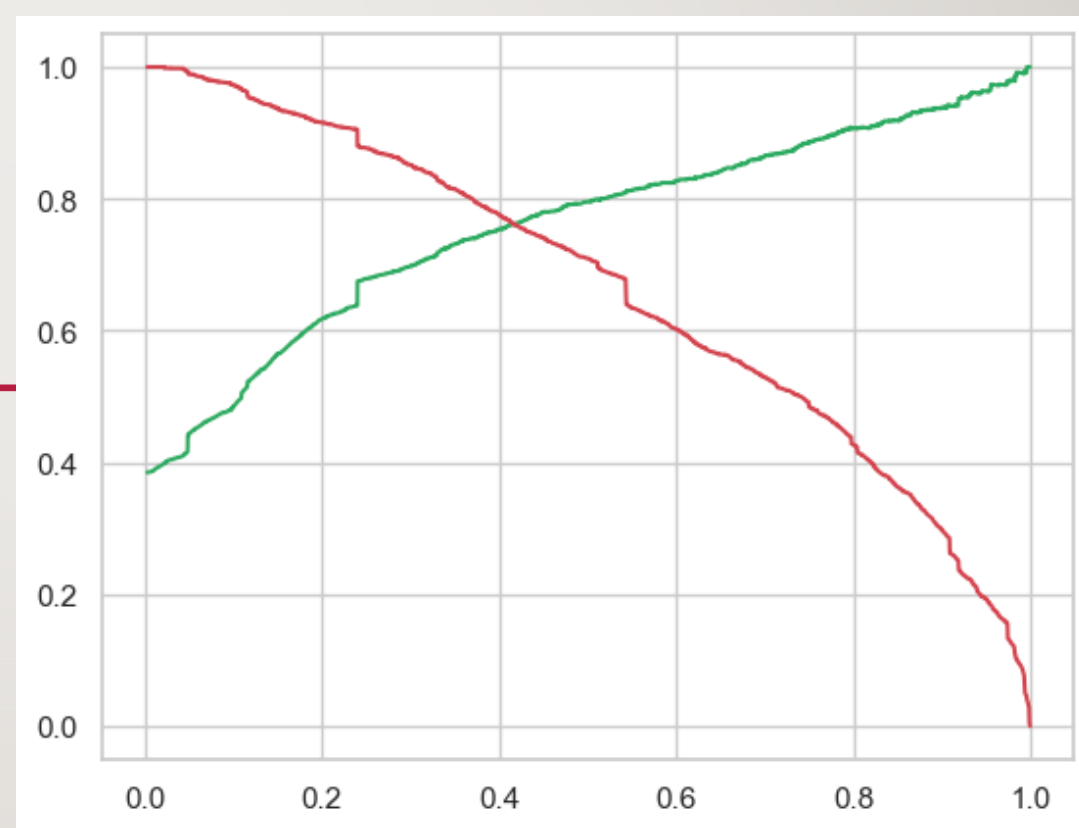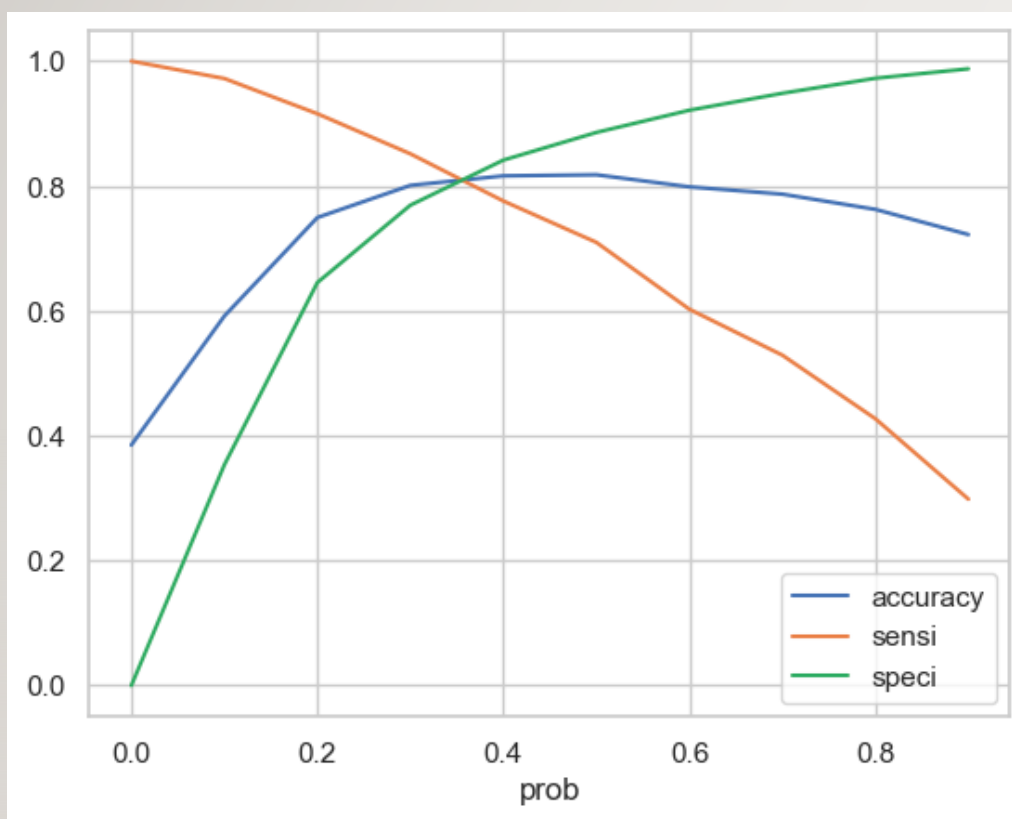- There is no correlation between the features.
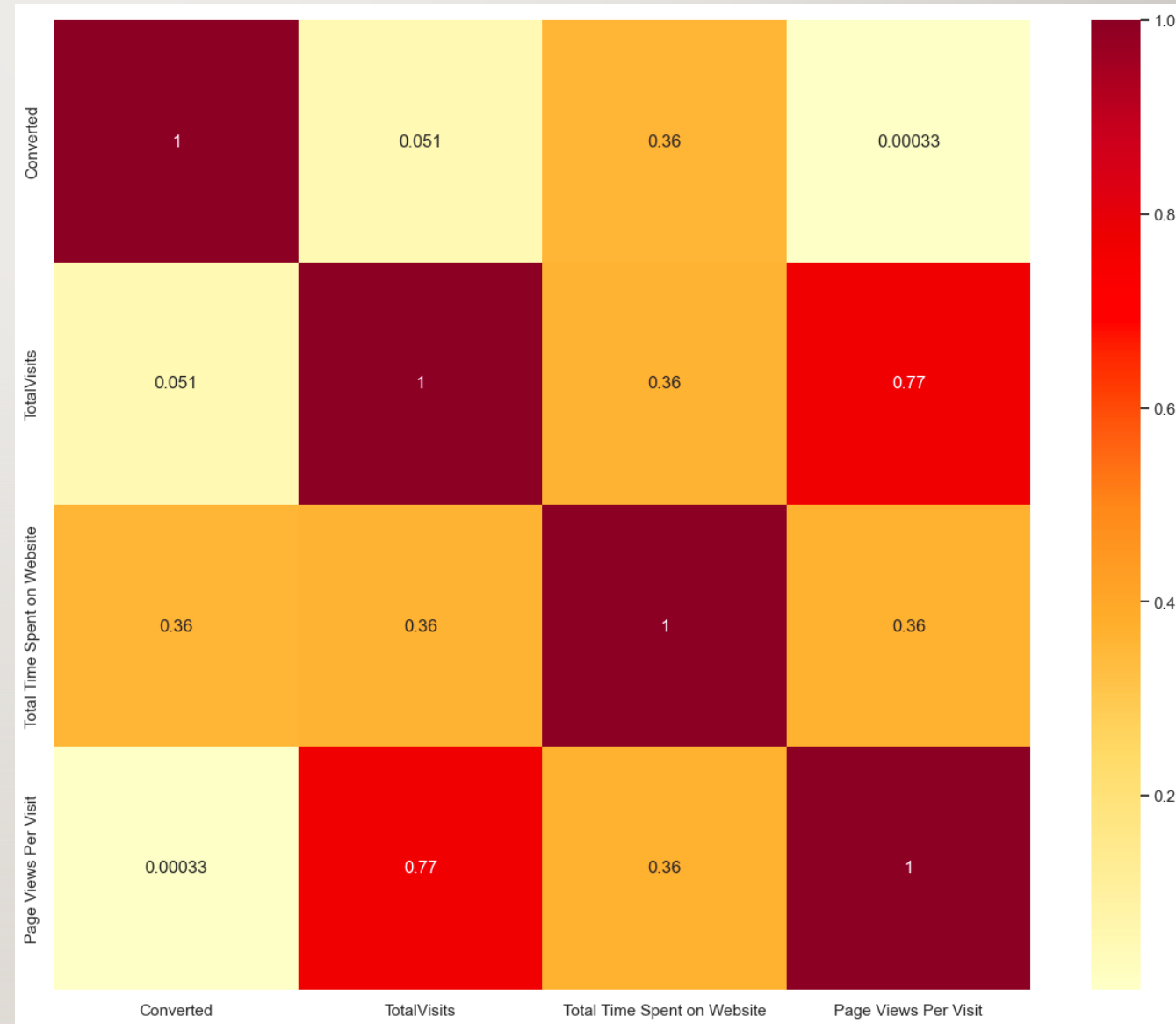
MODEL BUILDING AND EVALUATION
ROC Curve

- The area under the curve of the ROC is 0.89 so we have a good model.

- From the above curve (Left side plot), 0.35 seems to be the optimum point so taking it as cut-off probability/threshold.

- Around 0.41 is the trade-off between precision and recall (Right side plot).

- Focusing on leads with a conversion probability above 42% allows us to prioritize the most promising prospects, significantly increasing our chance of conversion.

OBSERVATIONS & RESULTS

Train Data:

- **Accuracy : 81.3%**

- **Sensitivity : 81.4%**

- **Specificity : 81.2%**

Test Data:

- **Accuracy : 80.7%**

- **Sensitivity : 79.8%**

- **Specificity : 81.1%**

IMPORTANT FEATURES:
1. Lead Source: Welingak Website
2. Lead Source: Reference
3. Current Occupation: Working Professional
4. Last Activity: Other Activity
5. Last Notable Activity: Unreachable
6. Last Activity: Unsubscribed
7. Last Activity: SMS Sent
8. Total Time Spent on Website
9. Lead Source: Olark Chat
10. Last Notable Activity: Modified
11. Last Activity: Olark Chat Conversation
12. Lead Origin: Landing Page Submission
13. Specialization: Others
14. Do Not Email

CONCLUSION

1. **Focus on API and Landing Page Leads:** Conversion rates are higher for these leads compared to Lead Add forms and imports. Prioritize outreach to leads from these sources.

2. **Target High-Engagement Leads:** Leads who spend more time on the website and open emails are more likely to convert. Focus efforts on engaging these leads further.

3. **Leverage High-Converting Channels**: While Google/direct traffic generates the most leads, references and Welingak website have higher conversion rates. Utilize these channels effectively.

4. **Optimize for Working Professionals**: Though unemployed leads are the majority, conversion rates are highest for working professionals. Tailor messaging and outreach to resonate with this segment.

5. **Thus, we have achieved our goal of getting the target lead conversion rate to be around 80%.**