

The background is a dark blue gradient with a subtle pattern of white dots. Overlaid on this are several white geometric elements: a large circular scale on the left with degree markings from 140 to 260, and several smaller concentric circles with arrows indicating clockwise or counter-clockwise rotation.

LEAD SCORING CASE STUDY

SUBMITTED BY- PRATIK SINHA

PRIYANKA SANGA

PRATIBHA VERMA

CONTENTS

- Problem Statement
- Objective
- Brief Approach
- Exploratory data Analysis(EDA)
- Model Building
- Model Evaluation
- 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.

MODEL BUILDING APPROACH

- Importing dataset
- Data cleaning and Data preparation
- Handling the Null values. Handling the data rows having 'Select' as the value
- Creating Dummy variables for Categorical Data
- Train-Test dataset split
- Feature Scaling
- Model building with top 15 features selected through RFE
- Model fine tuning
- Model evaluation
- Model prediction on Test dataset

DATA CLEANING

- Dropping the columns which have more than 40% of Null data as these columns would not play any important part in the analysis and might give false results due to major data missing
- Going through dataset having Null data, dropped columns 'How did you hear about X Education', 'What matters most to you in choosing a course', 'Lead Profile', 'City'
- Dropped the rows of the remaining columns which have Null values
- Dropped Lead Number and Prospect ID because of unique data in each record

DATA MANIPULATION

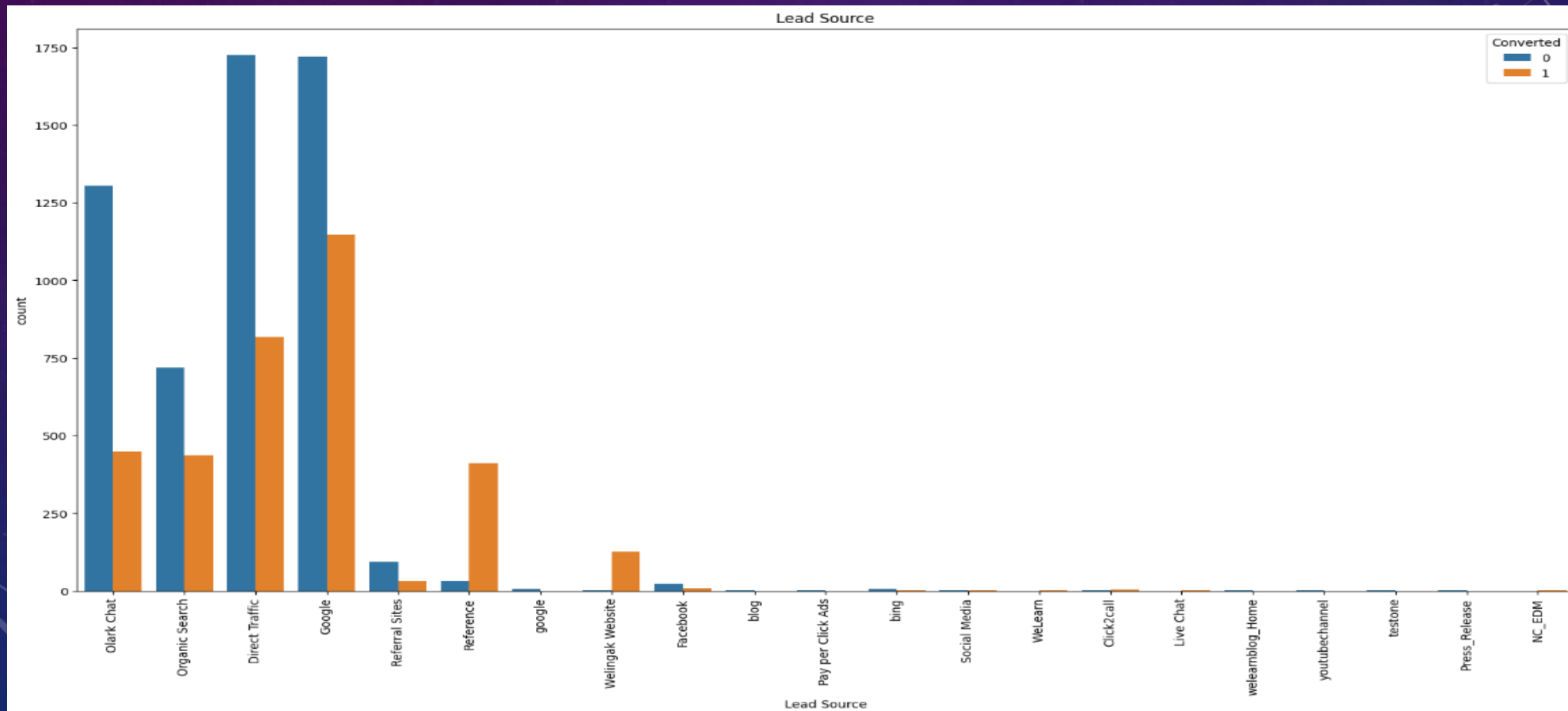
- Some columns like 'Specialization' have around 36% data as either Null or 'Select'. Replaced both with 'Not Specified' as we cannot delete this column due to high relatability with the Conversion rate
- Similar action taken on column 'What is your current occupation ', 'Tags'

EDA- NUMERIC DATA

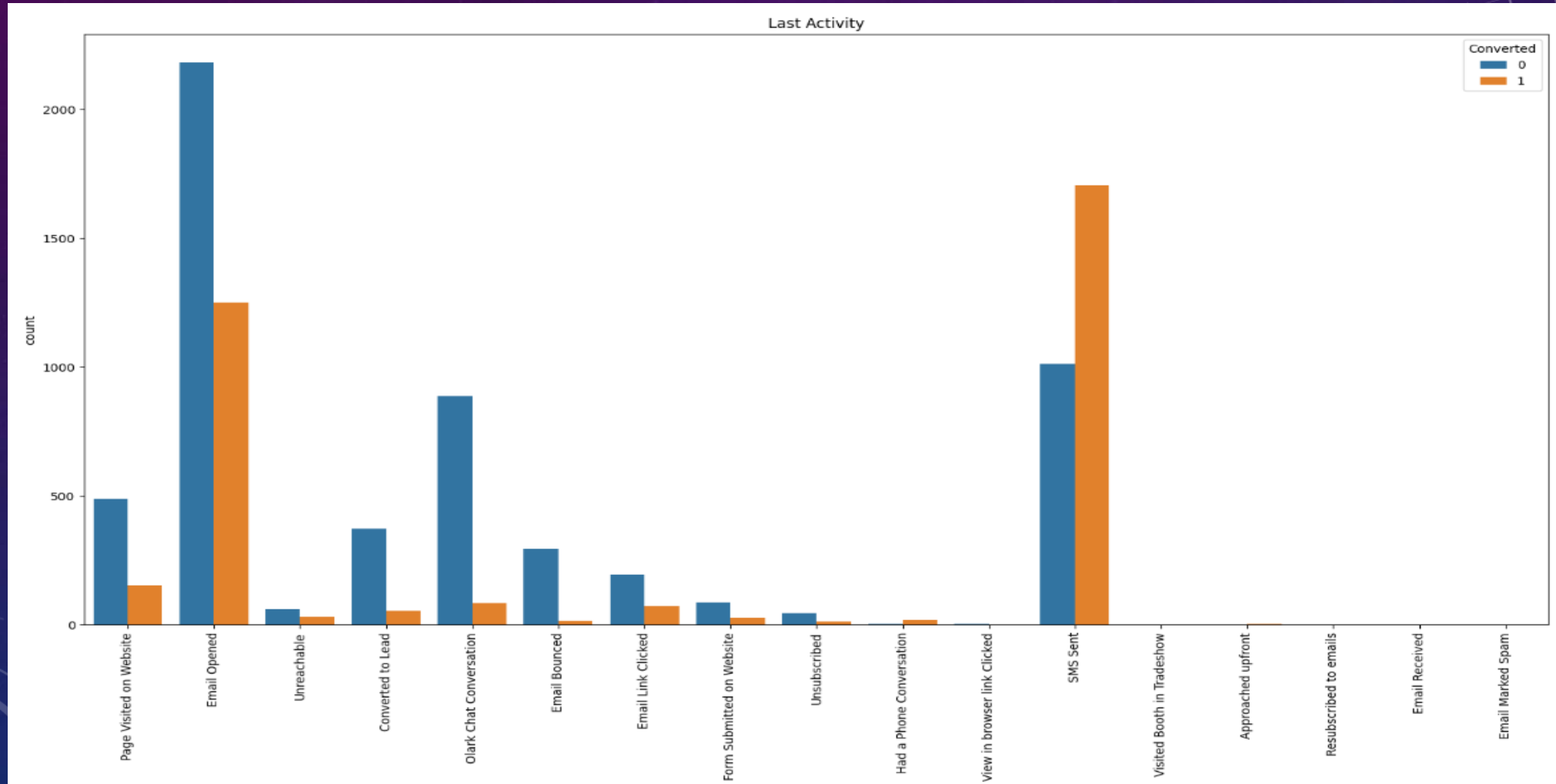
- Conversion rate is high for the people –
 - ✓ who have visited more times on the website
 - ✓ Who have spent more time on the website
 - ✓ Visited more pages in each visit

EDA- CATEGORICAL DATA

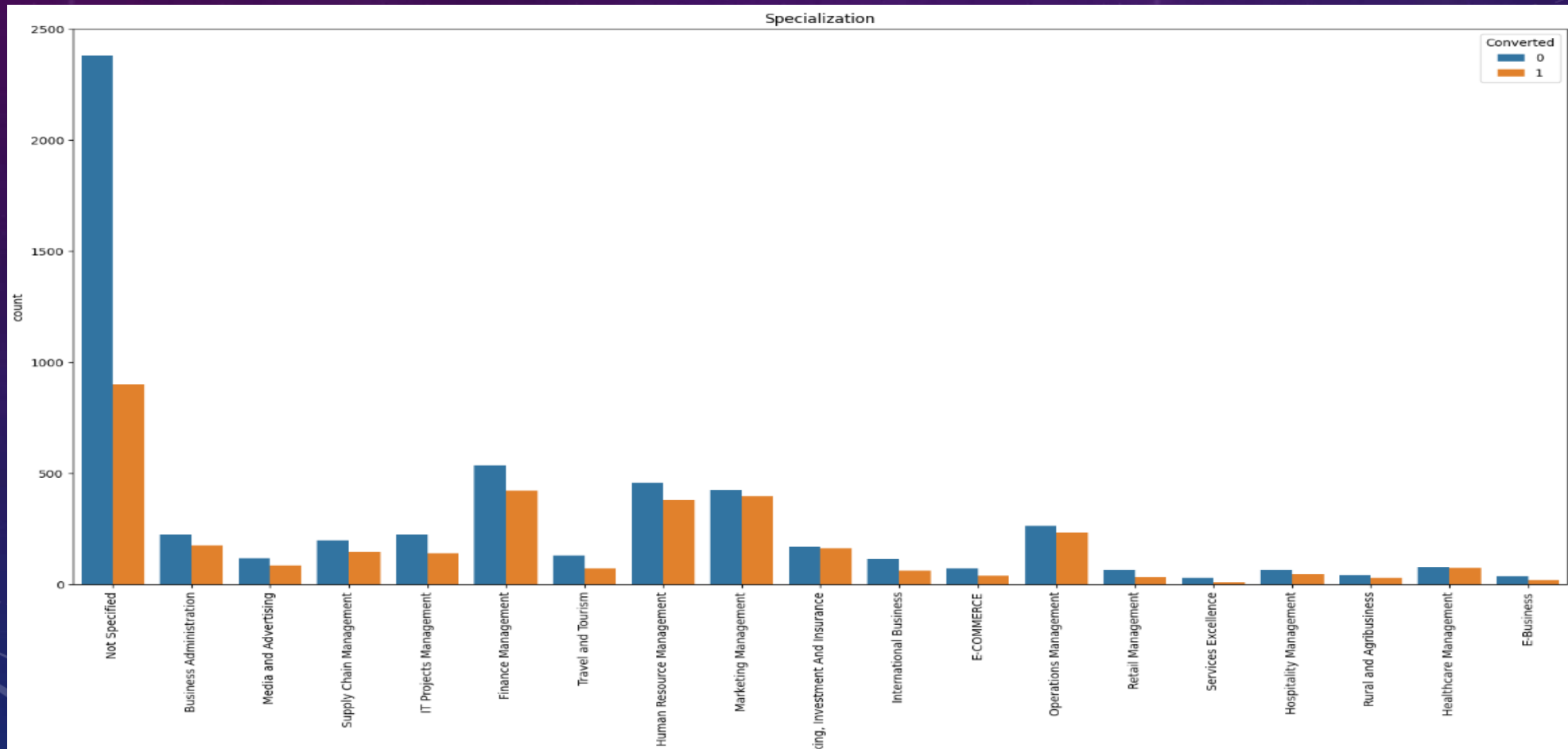
Lead Source having the leads from **Direct Traffic** and **Google** have higher chances of Conversion



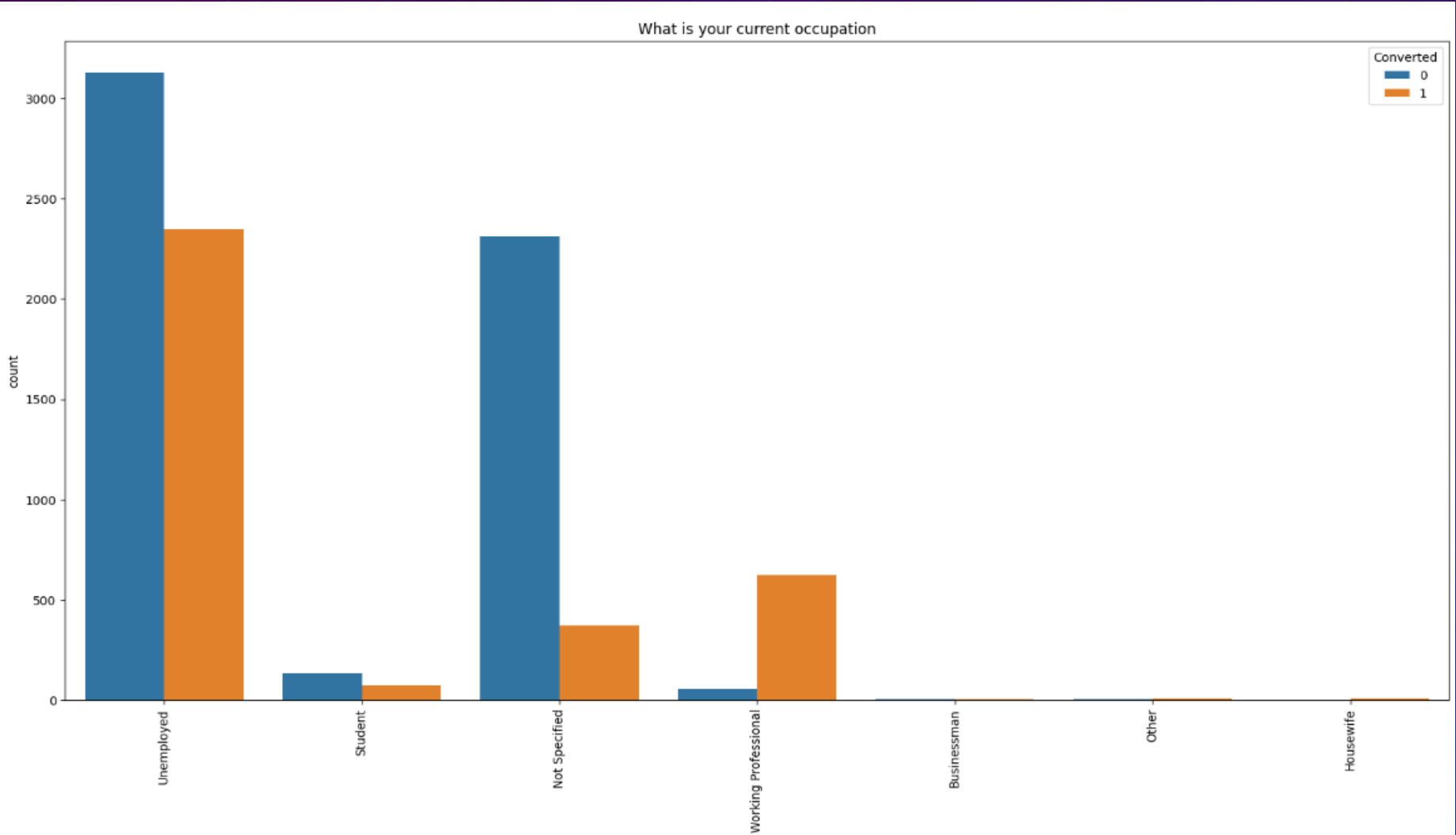
People who have sent SMS have the highest conversion rate followed by those who have opened the email



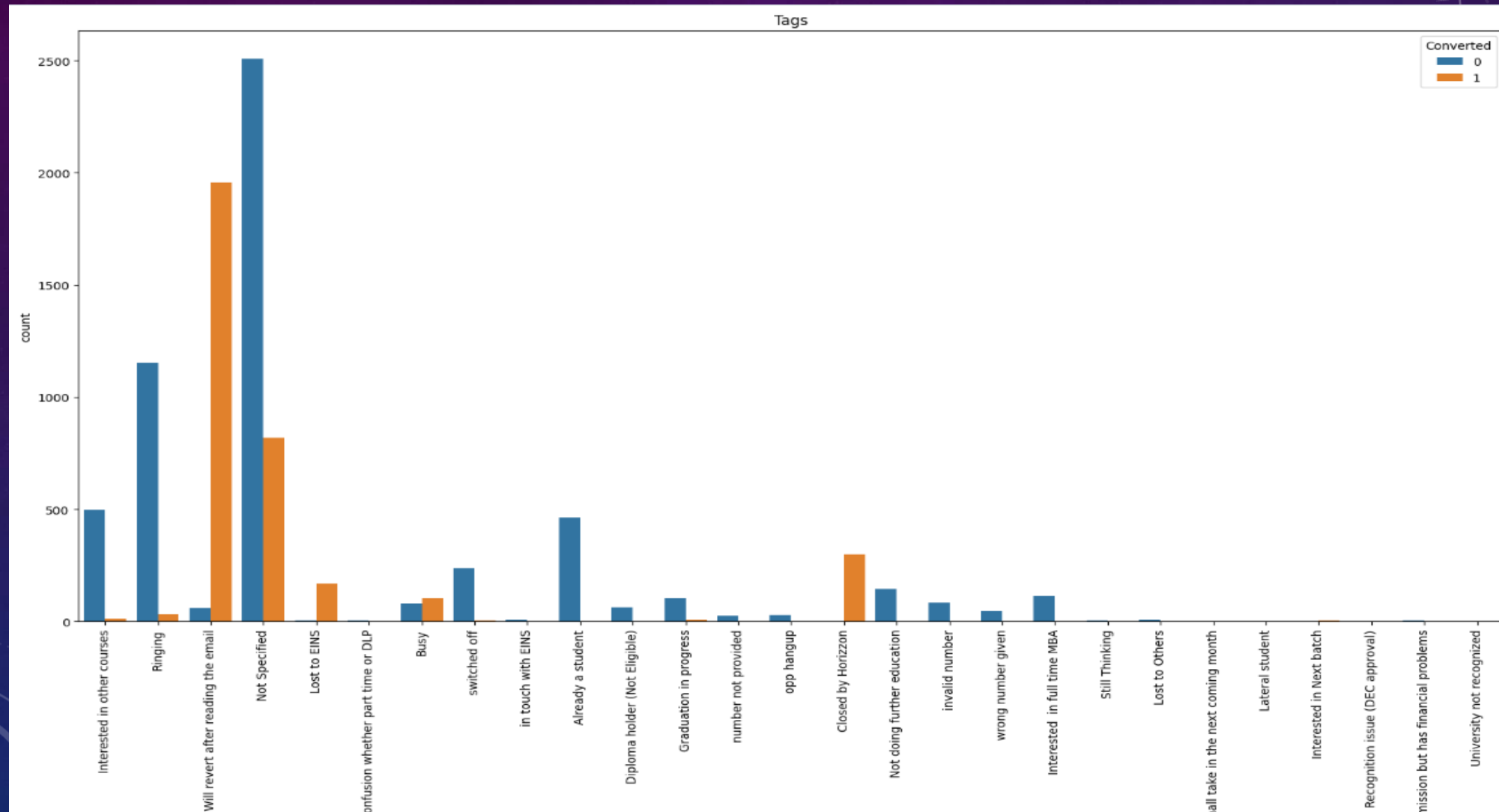
Certain people have not mentioned their Specialization but the conversion rate is the highest among other specializations followed by Finance Management, HR, Marketing and Operations management



People who are **Unemployed** have the high chances to opt for the courses



People who revert to the received email are most likely to opt for the courses



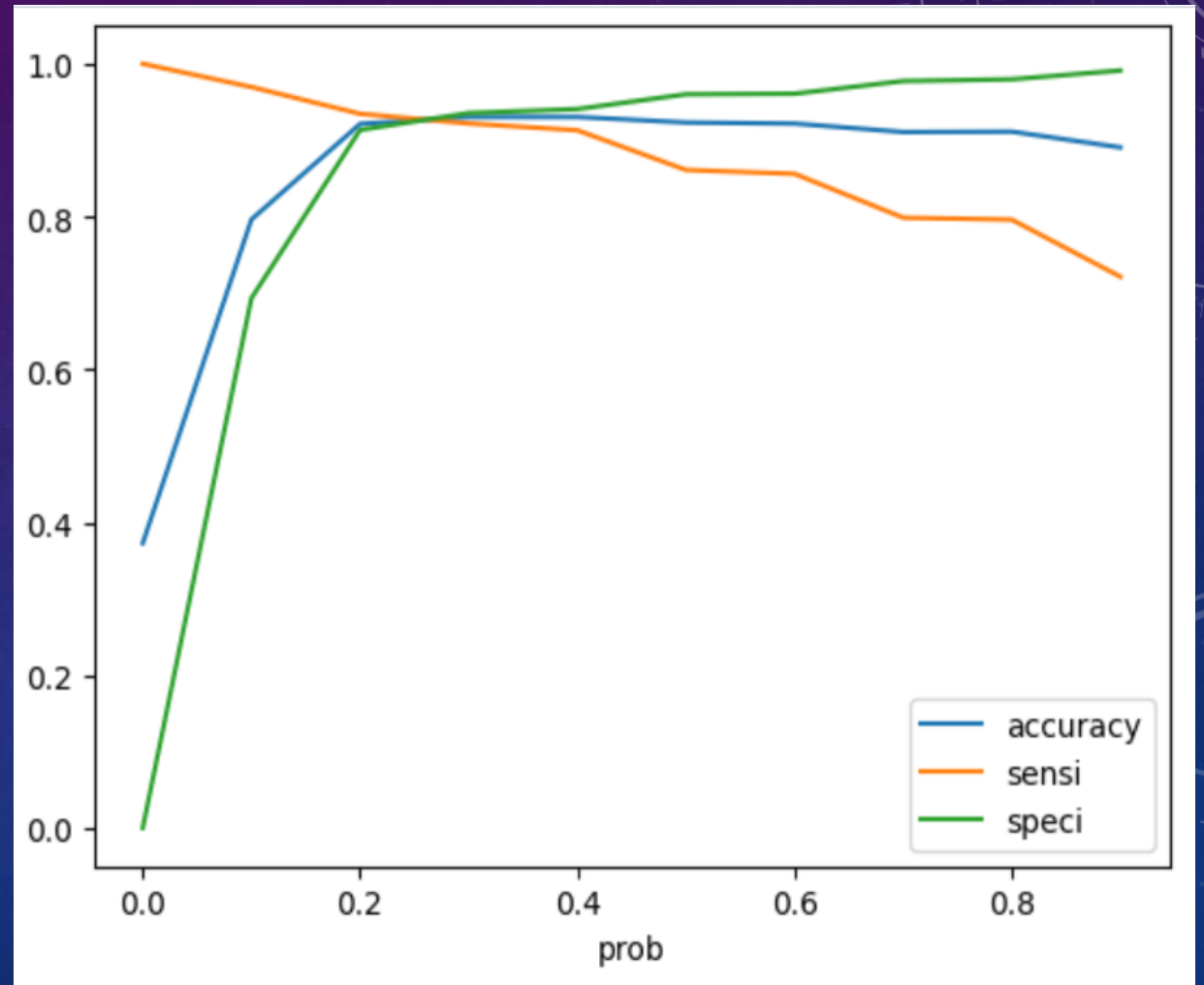
MODEL BUILDING AND EVALUATION

- Feature selection done through RFE Feature selection from which 3 features were deleted by looking at the p-value and VIF score and re-fit the model after deletion of each feature
- Model evaluation on the Train dataset-

Accuracy	92.3%
Sensitivity	86.1%
Specificity	96%

- Plotting the Accuracy, Sensitivity and Specificity we could figure out that the Probability cut-off for getting the optimal result is 0.3
- Results after using cut-off 0.3 increased sensitivity

Accuracy	93.04%
Sensitivity	92.2%
Specificity	93.5%
Precision	92.7%
Recall	86.1%



PREDICTION ON THE TEST DATASET

Accuracy	92.9%
Sensitivity	92.6%
Specificity	93.1%
Precision	89.6%
Recall	92.6%

- Final Features list-
 - ✓ 'Last Activity_SMS Sent'
 - ✓ 'What is your current occupation_Not Specified'
 - ✓ 'Tags_Busy', 'Tags_Closed by Horizzon'
 - ✓ 'Tags_Lost to EINS'
 - ✓ 'Tags_Not Specified'
 - ✓ 'Tags_Ringing'
 - ✓ 'Tags_Will revert after reading the email'
 - ✓ 'Tags_switched off'
 - ✓ 'Last Notable Activity_Modified'

CONCLUSION

- Conversion rate calculated by our model is **93%** which is within the 80% limit set by the customer
- Features responsible for the maximum conversion rate are-
 - ✓ Tags_Closed by Horizzon
 - ✓ Tags_Lost to EINS
 - ✓ Tags_Will revert after reading the email
 - ✓ What is your current occupation_Not Specified
 - ✓ Tags_switched off