LEAD SCORE CASE STUDY

Presented by: Rudri Dave
Saba Afreen

Problem Description

- X Education is an organization which provides online courses for industry professionals. The company makes its courses on several popular websites like Google.
- X Education wants to select most promising leads that can be converted to paying customers.
- The company requires a model to be built for selecting most promising leads.
- Lead score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the less are the chances of conversion.
- The model to be built in lead conversion rate around 80% or more.

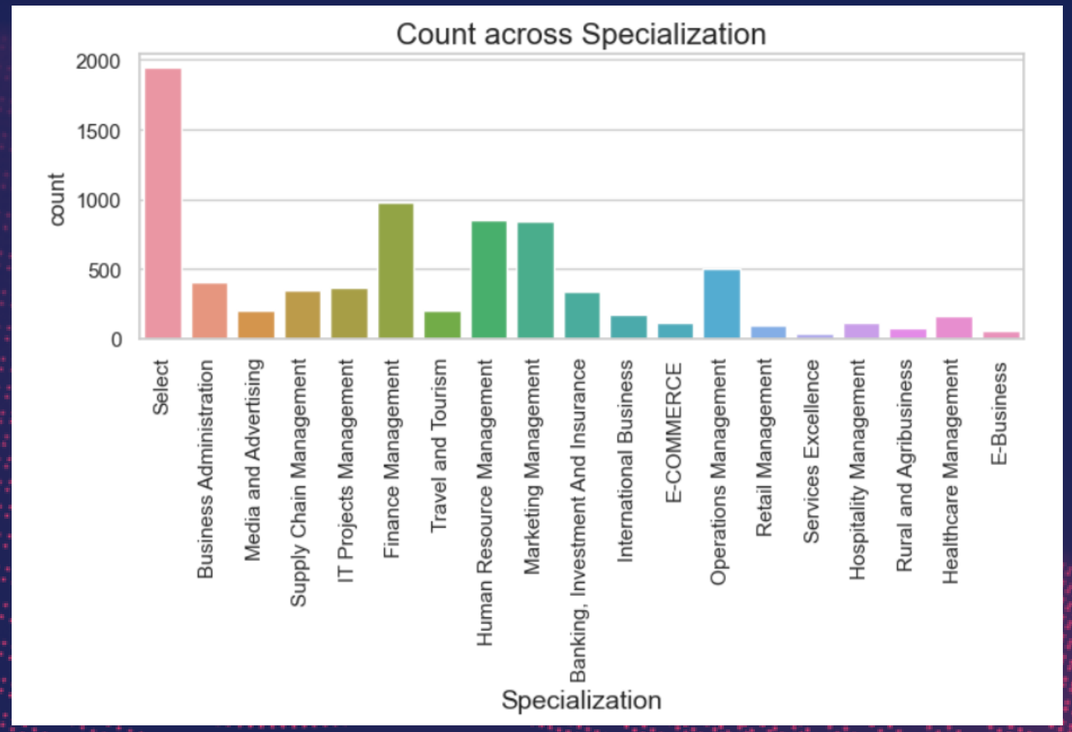
DATA ANALYSIS APPROACH

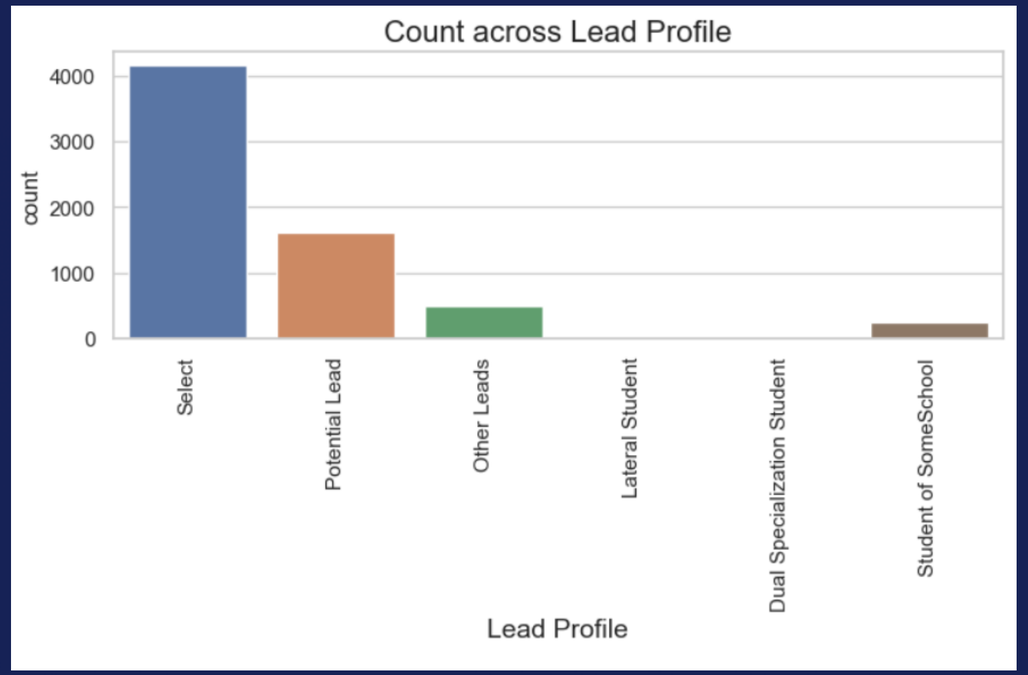
- IMPORT DATA
- CLEANING AND PREPARING DATA FOR FURTHER ANALYSIS
- EDA FOR FIGURING OUT MOST HELFUL ATTRIBUTES FOR CONVERSION
- SCALING FEATURES
- PREAPRE THE DATA FOR MODEL BUILDING
- BUILDING A LOGISTIC REGRESSION MODEL
- ASSIGN A LEAD SCORE FOR EACH LEADS
- TEST THE MODEL ON TRAIN SET
- EVALUATING MODEL BY DIFFERENT MEASURES AND METRICS
- TESTING THE MODEL ON TEST SET
- MEASURING THE ACCURACY OF MODEL AND OTHER METRICS USED FOR EVALUATION

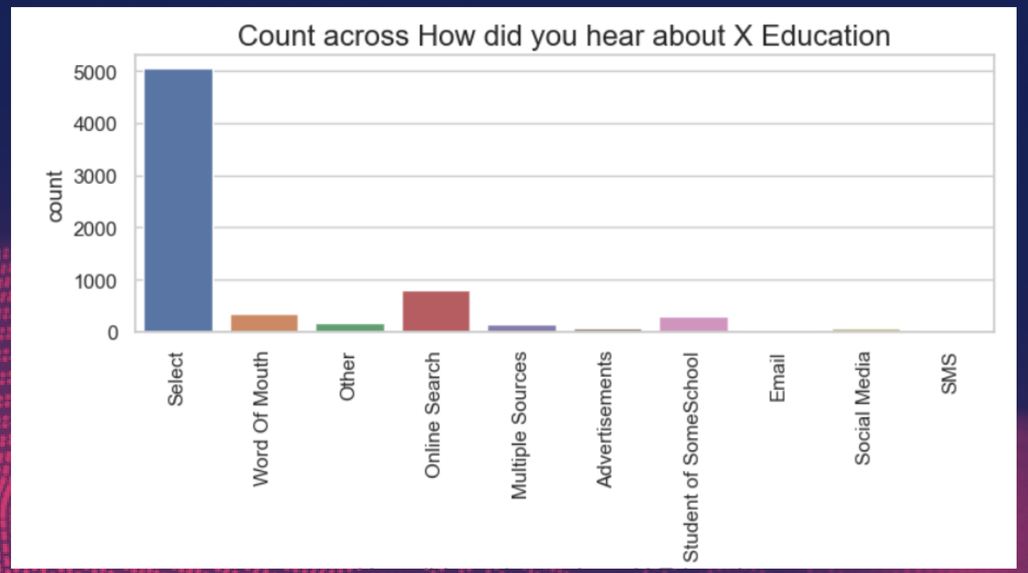
Data Manipulation

- Total number of rows = 37, Total number of columns = 9240
- Eliminating the columns having greater than 3000 missing values
- Single value features like "Magazine", "Do Not Call","Receive More Updates About our Courses", "Supply Chain Content" etc. have been dropped
- Removing the "Prospect ID" and "Lead Number" which is not necessary for analysis

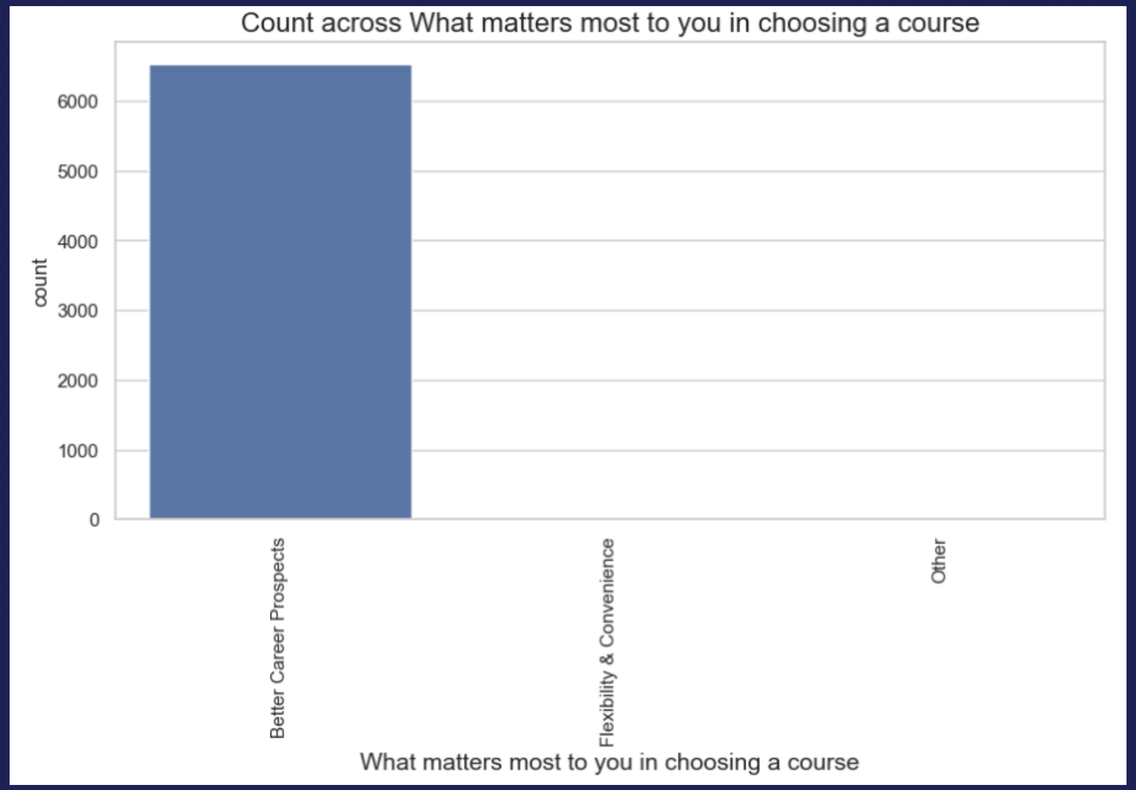
EXPLORATORY DATA ANALYSIS

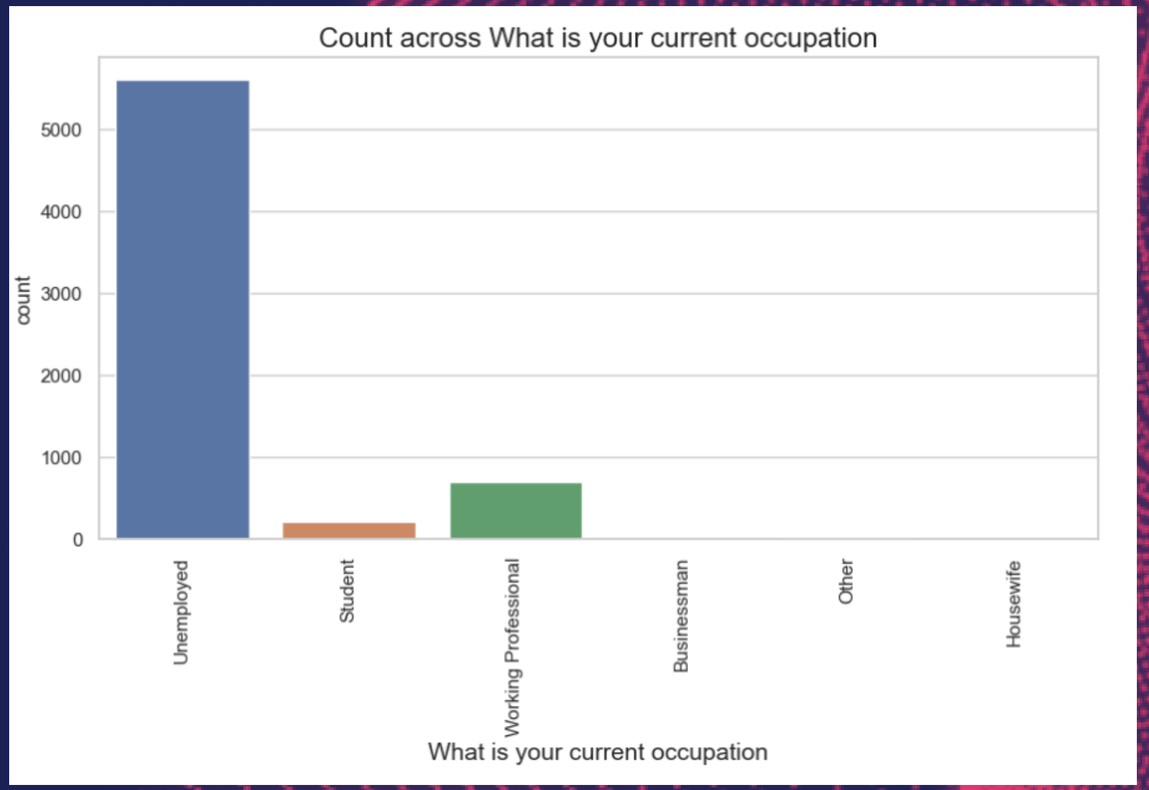




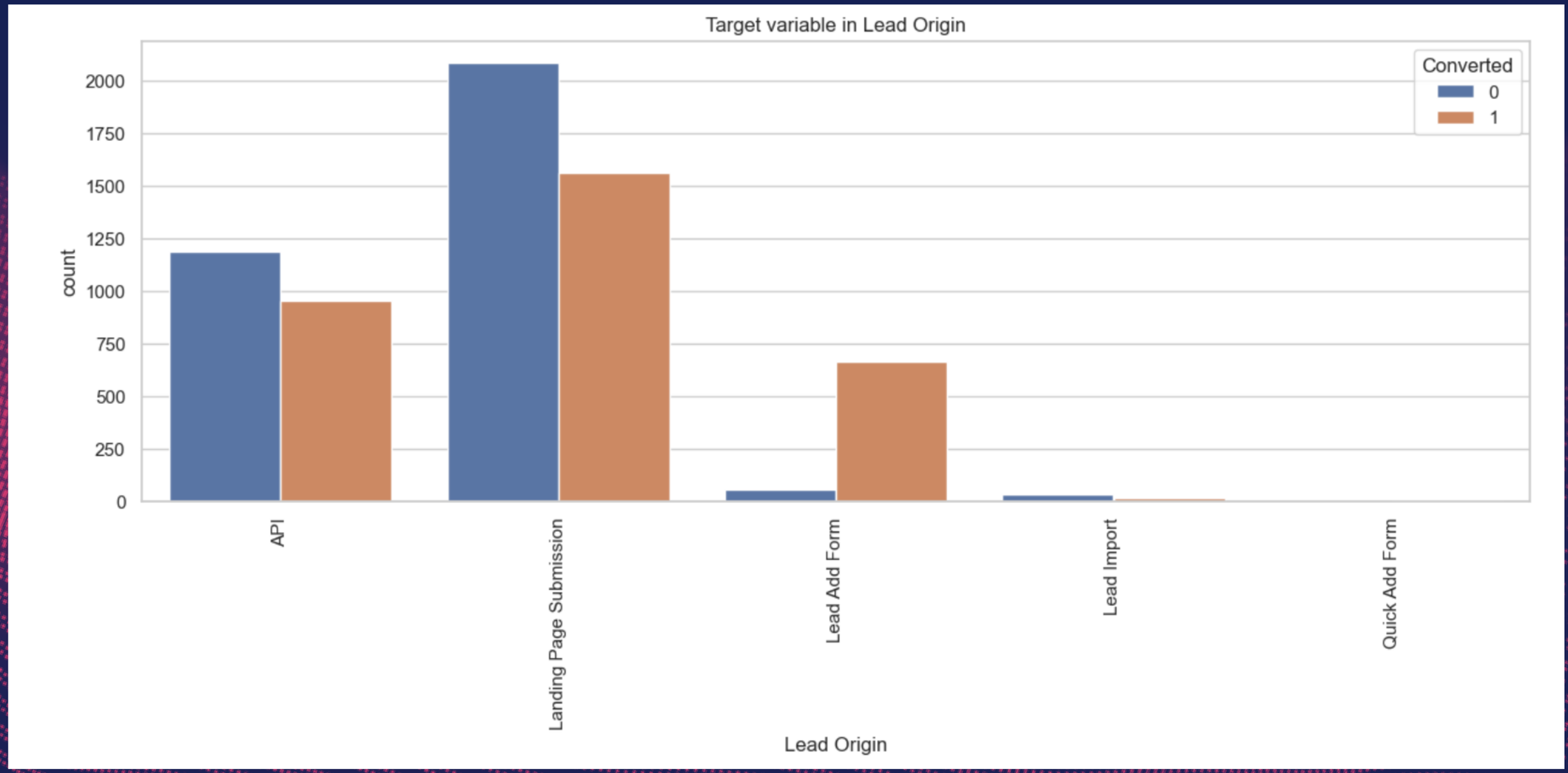


Exploratory Data Analysis

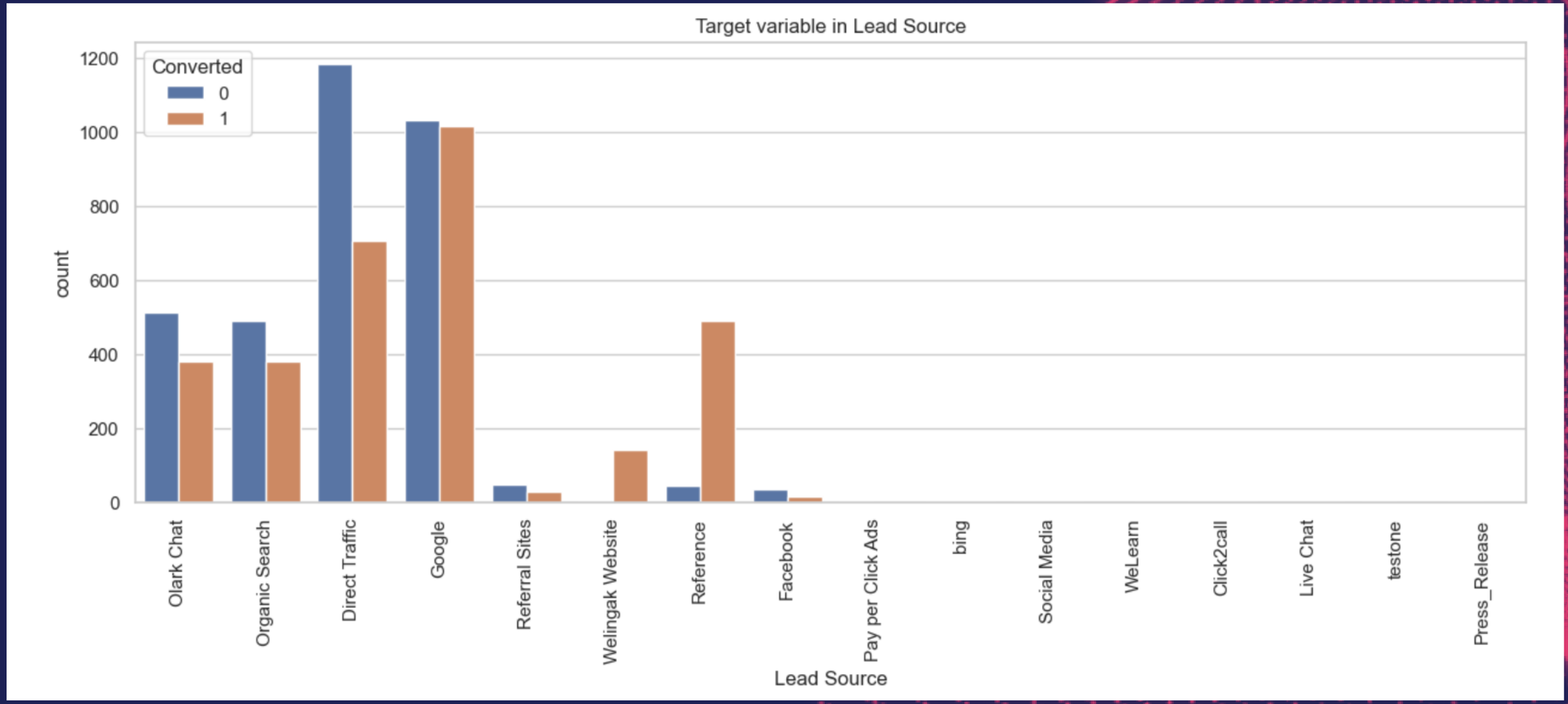




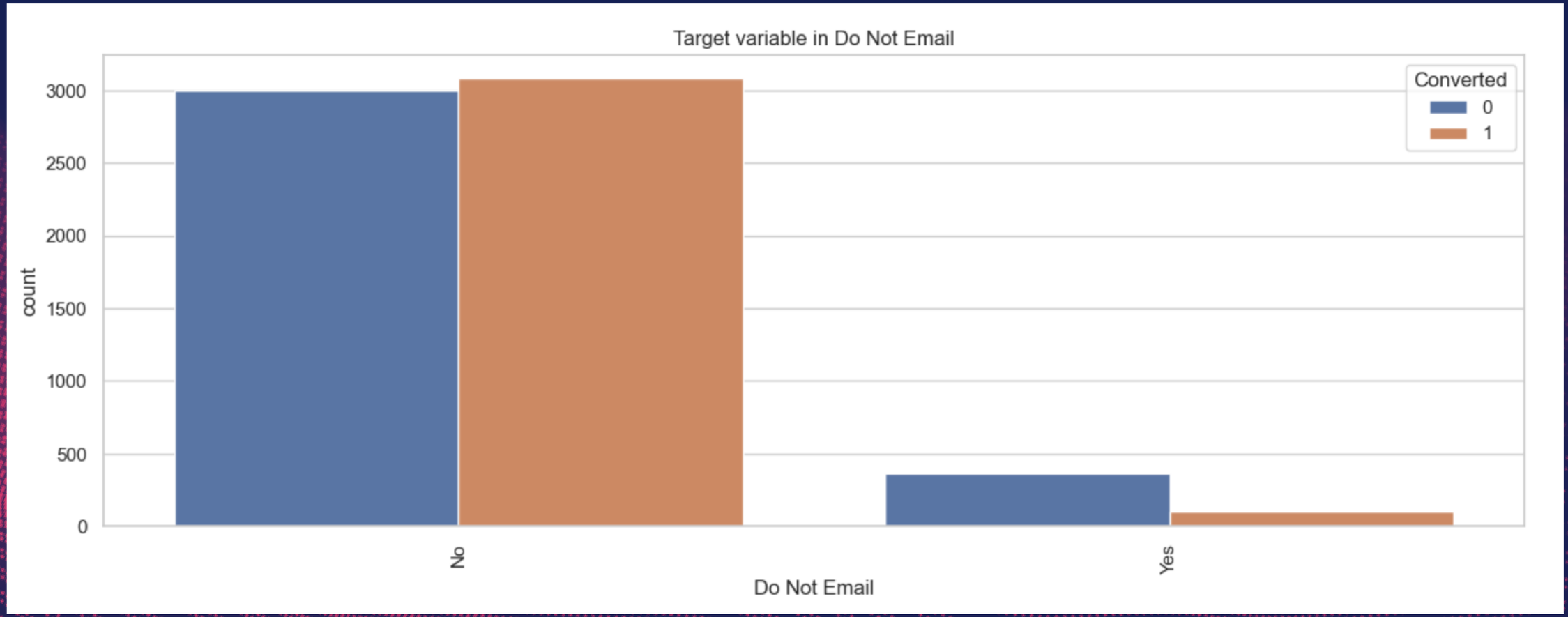
Categorical Variable Relation



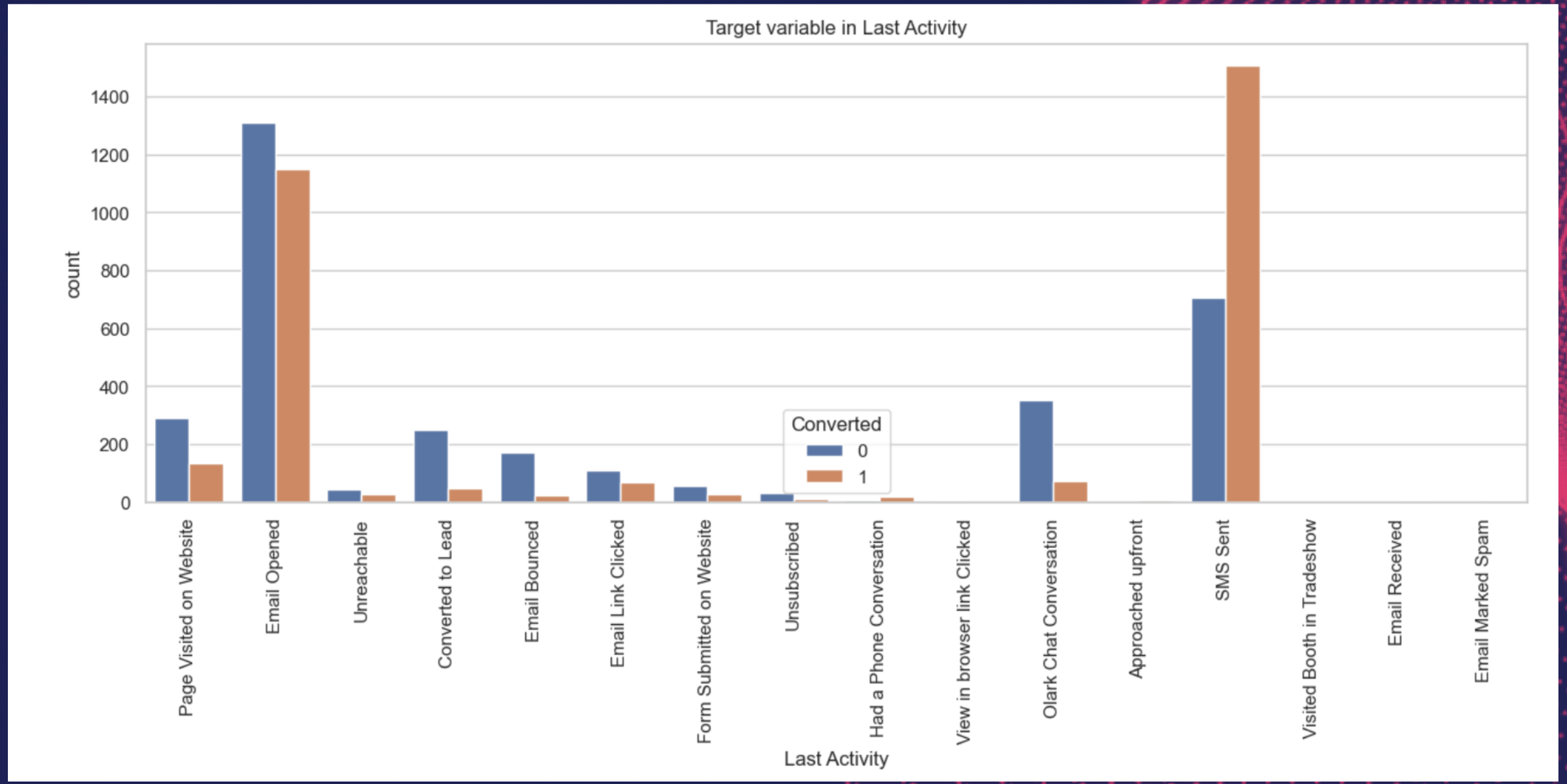
Categorical Variable Relation



Categorical Variable Relation



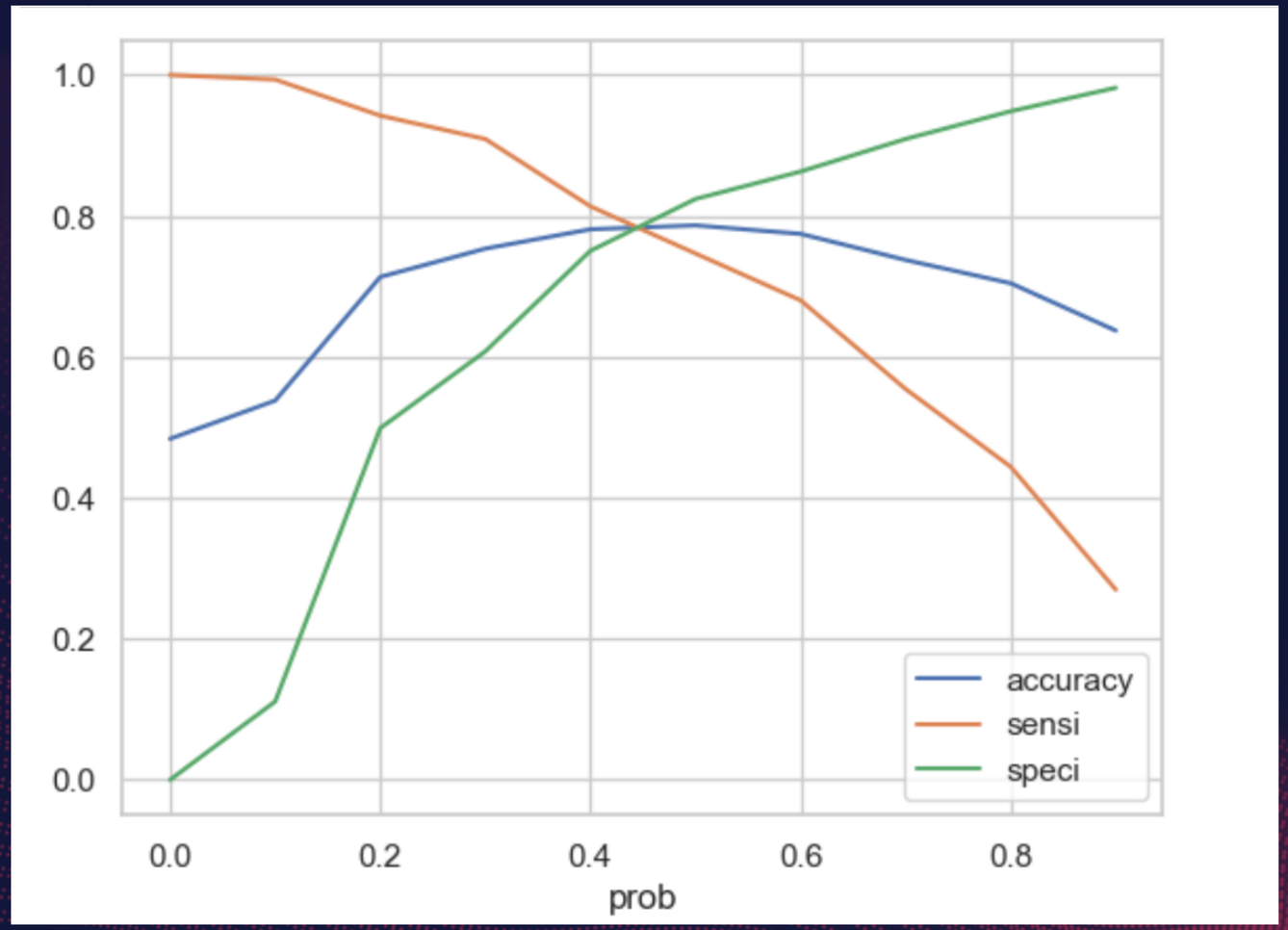
Categorical Variable Relation



MODEL BUILDING

- Splitting the data into Training and Testing Sets
- The basic step for regression is performing a train-test spilt and we have selected 70:30 ratio
- Used RFE for feature selection
- Running RFE with 15 variables as the output
- We have built a model by removing variables whose p-value > 0.05 and VIF > 5
- Predictions on Test Data Set

ROC CURVE (TRAIN MODEL)



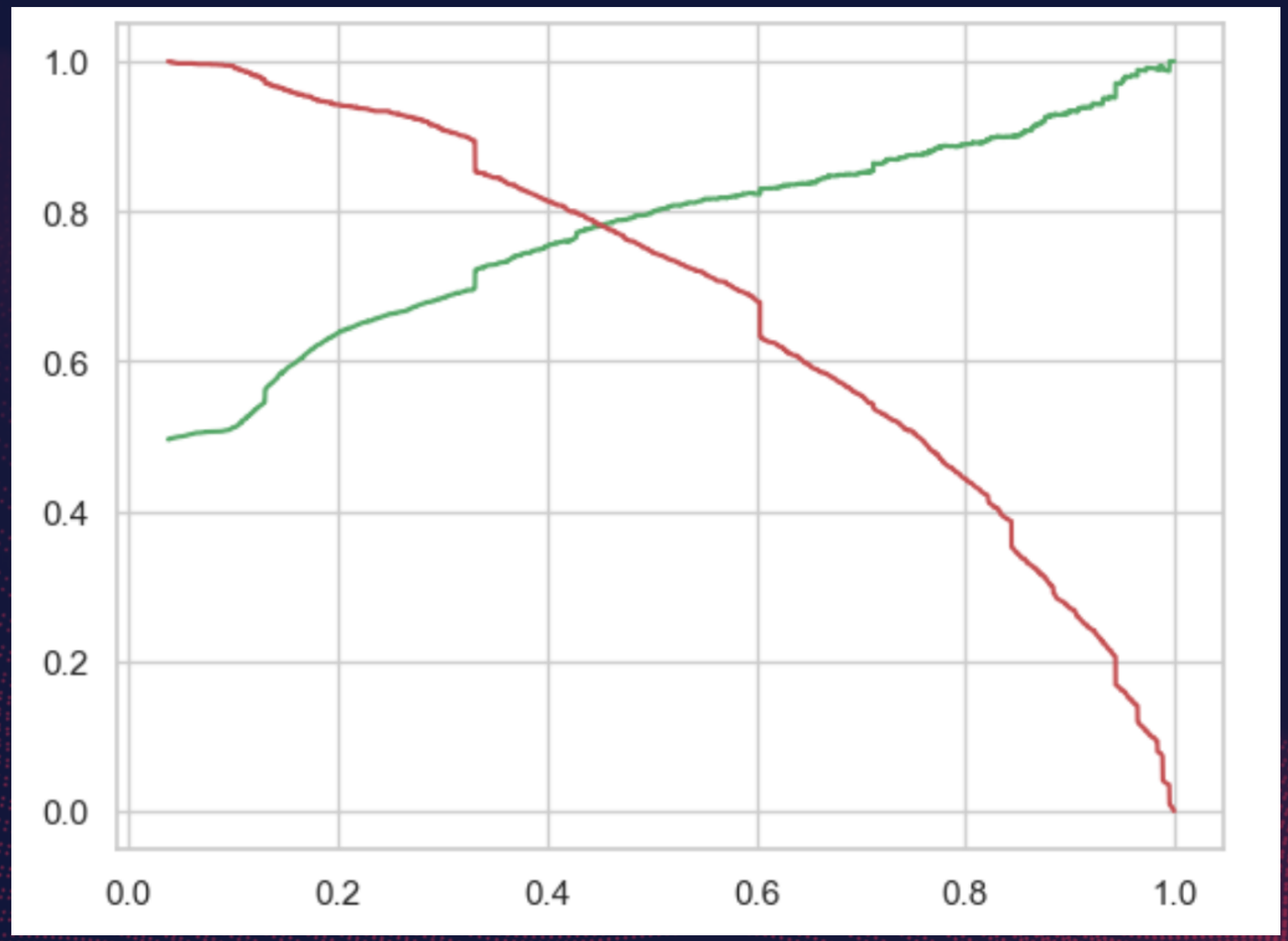
CUTOFF VALUE : 0.42

ACCURACY: 78%

SENSITIVITY: 80%

SPECIFICITY: 76%

MODEL EVALUATION (TEST MODEL)



CUTOFF VALUE : 0.44

OVERALL ACCURACY: 78%

PRECISION: 77%

RECALL: 78%

Conclusion

- It was found that the variables that mattered the most in potential buyers are:
- Total Visits
- Total Time Spent on the Website
- Lead Score is also a notable feature that should be focused on

- The below 3 categorical variables should be focused on the most to increase the probability of lead conversions:
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Last Activity_Had a Phone Conversation

- Most of the leads joined course for their better career path and majority of them are from Finance Management background.

- One thing that can convert the leads to HOT LEADS is through repetitive calls and email enagagement.

- Almost all of the leads current occupation is unemployes, hence more focus should be given to unemployed leads.

THANK YOU!