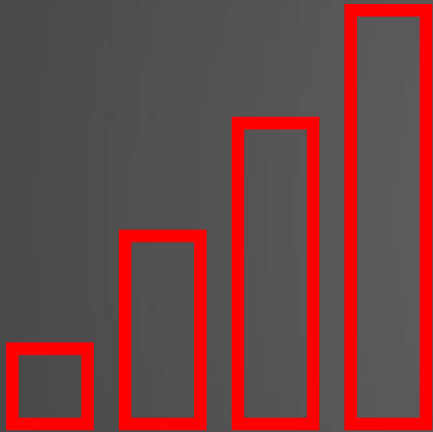


X-Education: Lead Scoring Case Study



TEAM MEMBERS:
JATIN SETHI
JANAKI T R
JANAGIRAMAN T

TABLE OF CONTENTS

Problem Statement

Problem Approach

EDA

Correlations

Model Evaluation

Observations

Conclusion

Accessing and Reviewing the historical data supplied by the Company.

Univariate, Bivariate, and Heatmap analysis for both numerical and categorical variables.

Executing prerequisites for RFE and Logistic Regression.

Data Gathering

Data Cleaning


Performing EDA

Data Preparation

Model Building

Removing duplicates, handling null values, and eliminating unnecessary columns, among other tasks.

Handling Outliers, Standardizing Features



Identifying the top 15
features through Recursive
Feature Elimination (RFE)

Reduction of Columns and
Reconstruction of Models

Assessing the Accuracy of
Our Final Model Using the
PCA-Built Model



Feature
Selection

Model
Building

Model
Improvement

Final Model

Verifying
with PCA

Constructing models through
Recursive Feature Elimination
(RFE) for chosen variables.

Analysis of the Final
Model and its
Performance on Test Data



PROBLEM STATEMENT

- X Education is an educational firm that offers online courses aimed at industry professionals. Each day, numerous professionals visit their website to explore available courses. After browsing, visitors fill out a form on the site, which allows the company to categorize them as leads.
- Following the acquisition of these leads, the sales team begins their outreach efforts through calls, emails, and other forms of communication. While some leads successfully convert, the majority do not.
- Typically, X Education sees a lead conversion rate of approximately **30%**. This indicates that if the company generates 100 leads in a single day, around 30 of those leads result in conversions. To enhance this process, the company aims to pinpoint the most promising leads, referred to as Hot Leads.
- By effectively identifying this group of leads, the conversion rate is expected to improve, as the sales team can concentrate on engaging with the most promising prospects rather than reaching out to all leads indiscriminately.



BUSINESS OBJECTIVES

- Lead X has requested the development of a model that assigns a lead score ranging from 0 to 100. This will help them identify high-potential leads and enhance their overall conversion rate.
- The CEO aims to reach a lead conversion rate of 80%.
- Additionally, the model should be capable of addressing future challenges, such as actions needed during peak times, optimal resource utilization, and strategies to implement after reaching the target.



PROBLEM APPROACH

- Loading the data and examining the data frame
- Data preprocessing
- Exploratory Data Analysis (EDA)
- Creation of dummy variables
- Splitting into training and testing sets
- Scaling features
- Analyzing correlations
- Constructing the model (including RFE, R squared, VIF, and p- values)
- Assessing the model
- Generating predictions on the test set

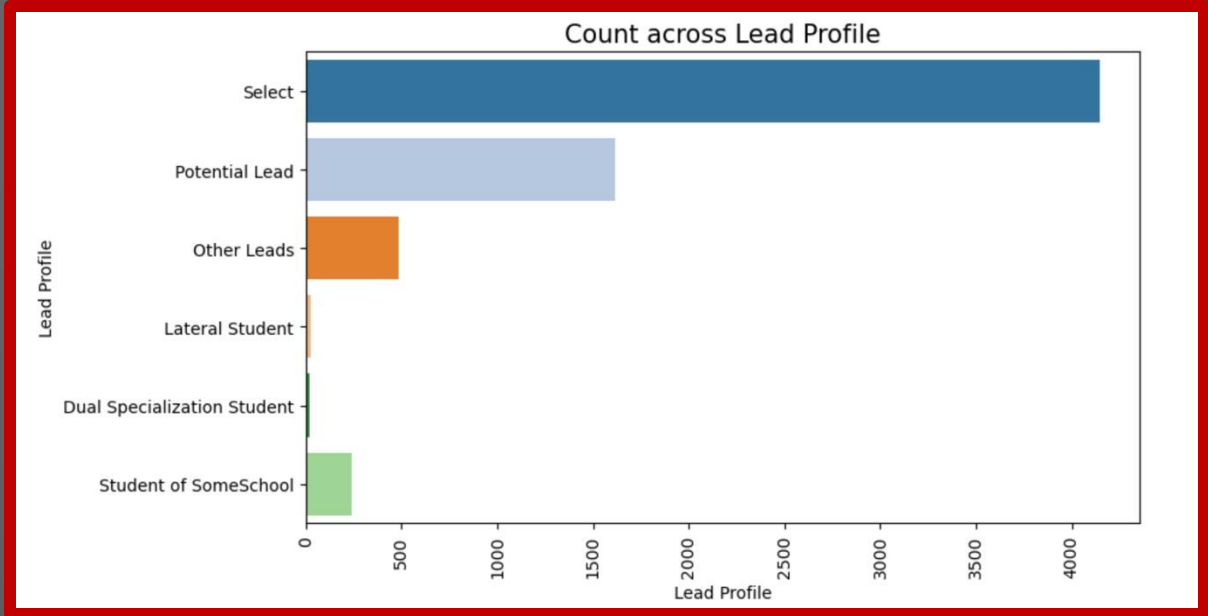
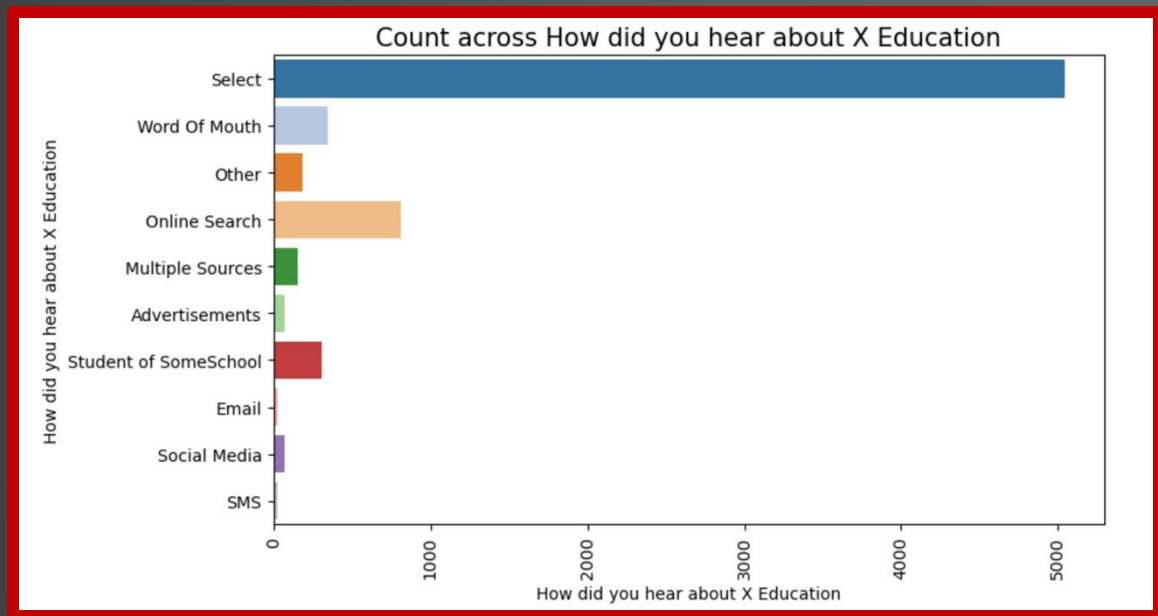
OUTLIER ANALYSIS

- Total Visits, Page Views Per Visit : Both these variables contain outliers, rest seems fine



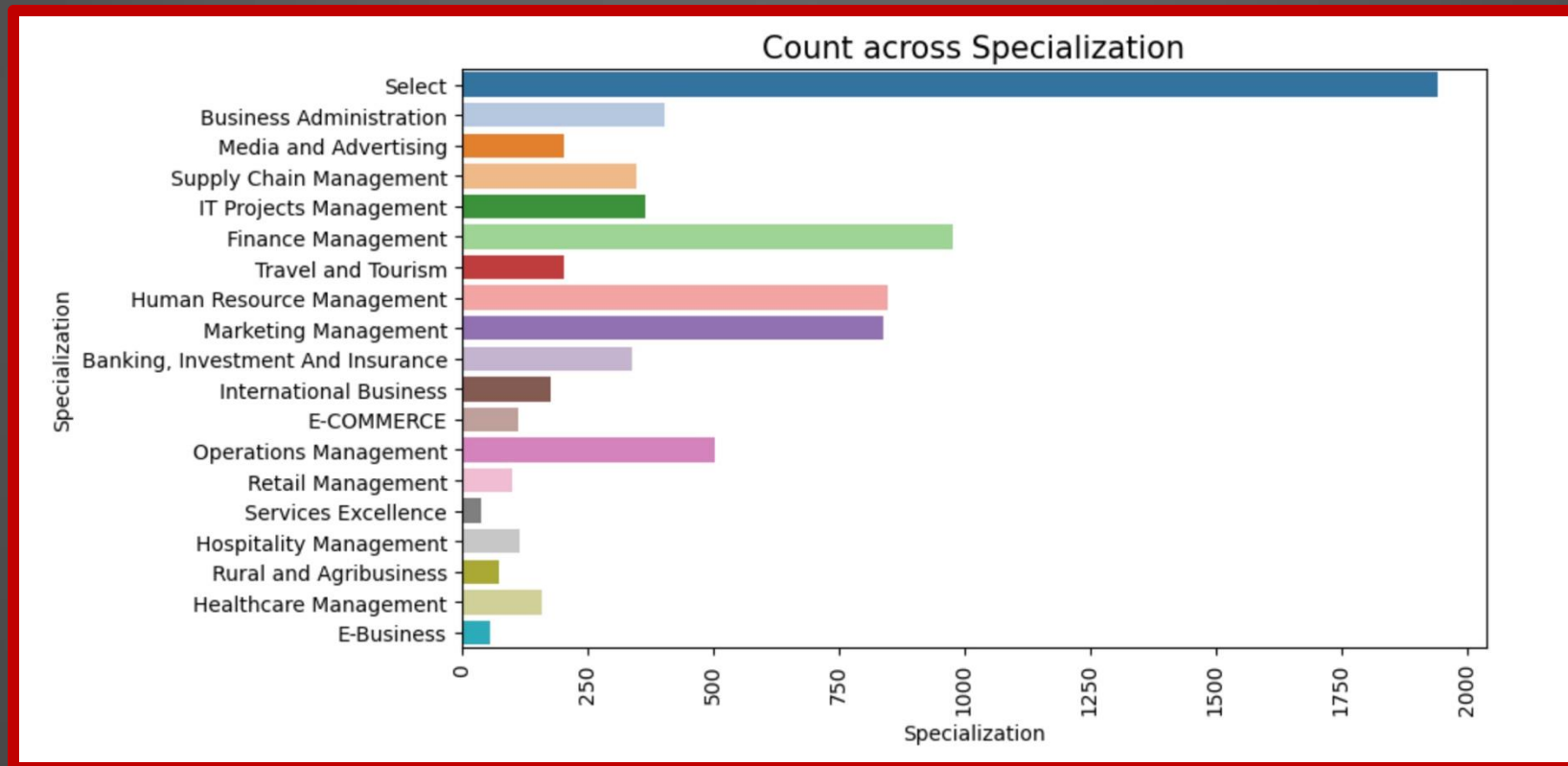
EDA - DATA CLEANING

- There are a few columns in which there is a level called 'Select' which is taking care



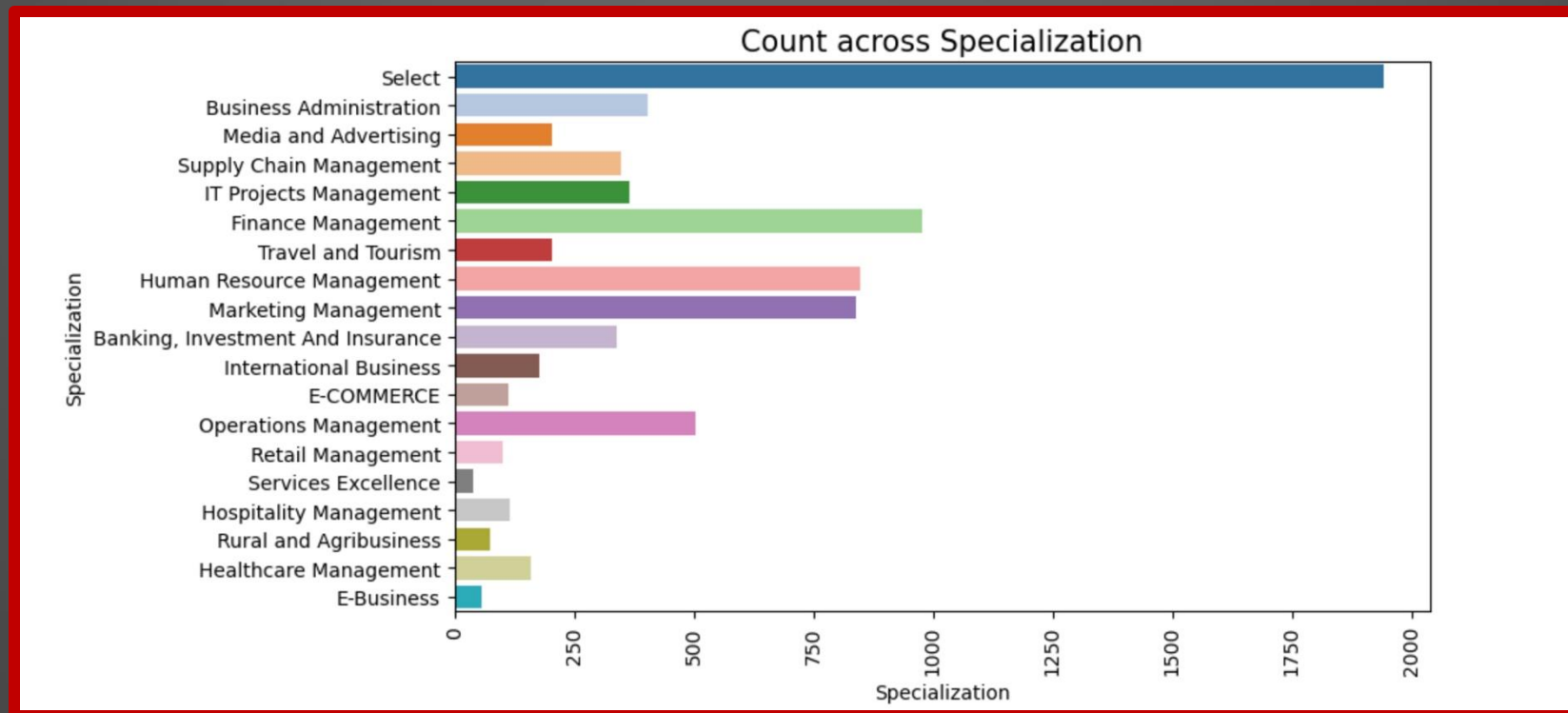
EDA - DATA CLEANING

- There are a few columns in which there is a level called 'Select' which is taking care



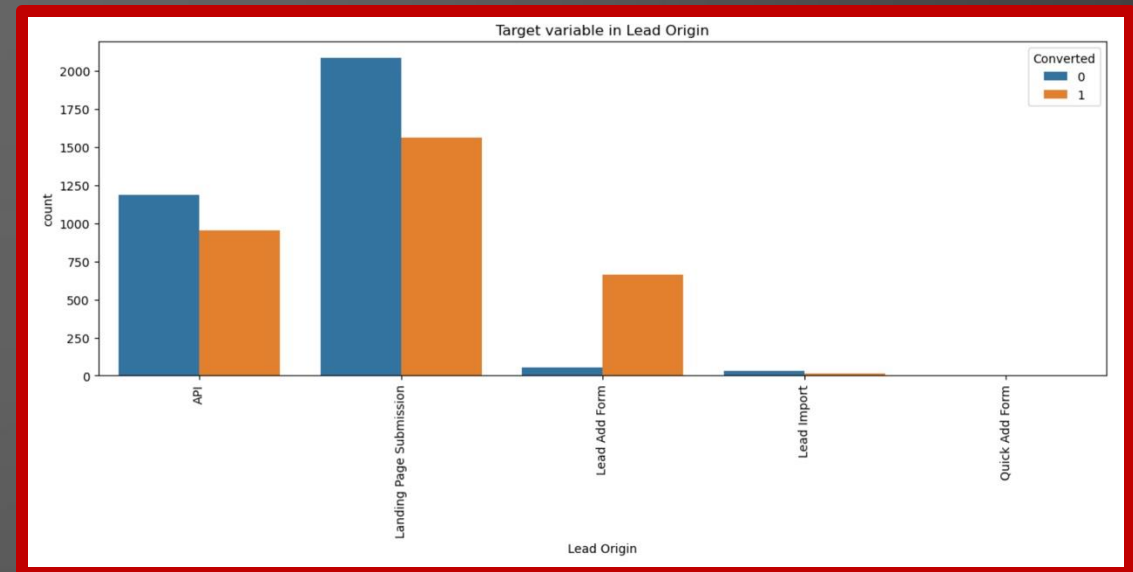
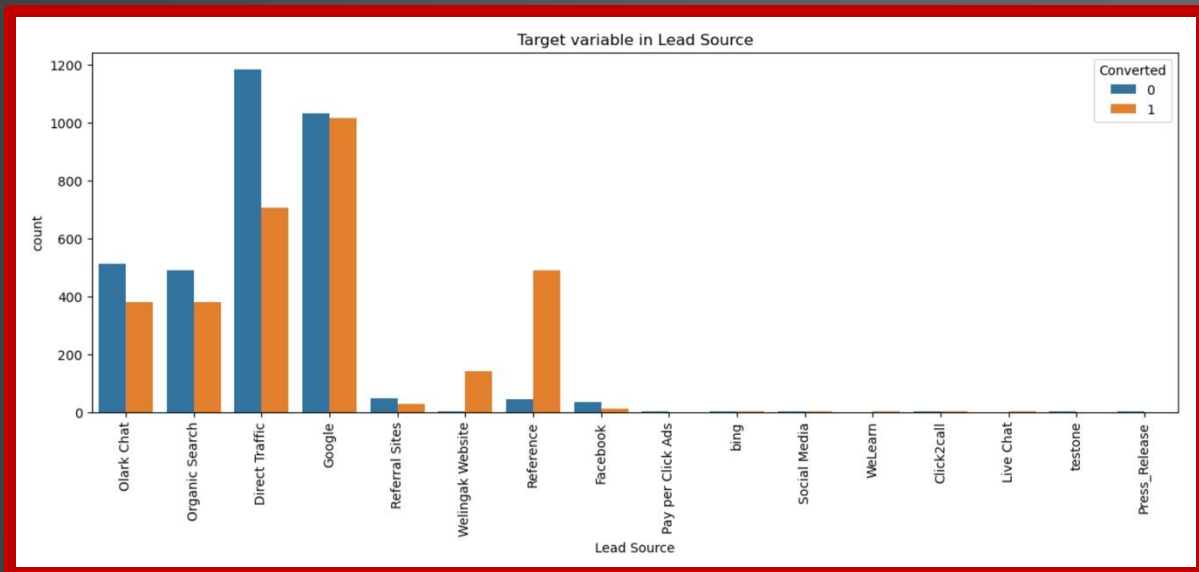
SPECIALIZATION

- Leads from HR, Finance & Marketing management specializations are high probability to convert



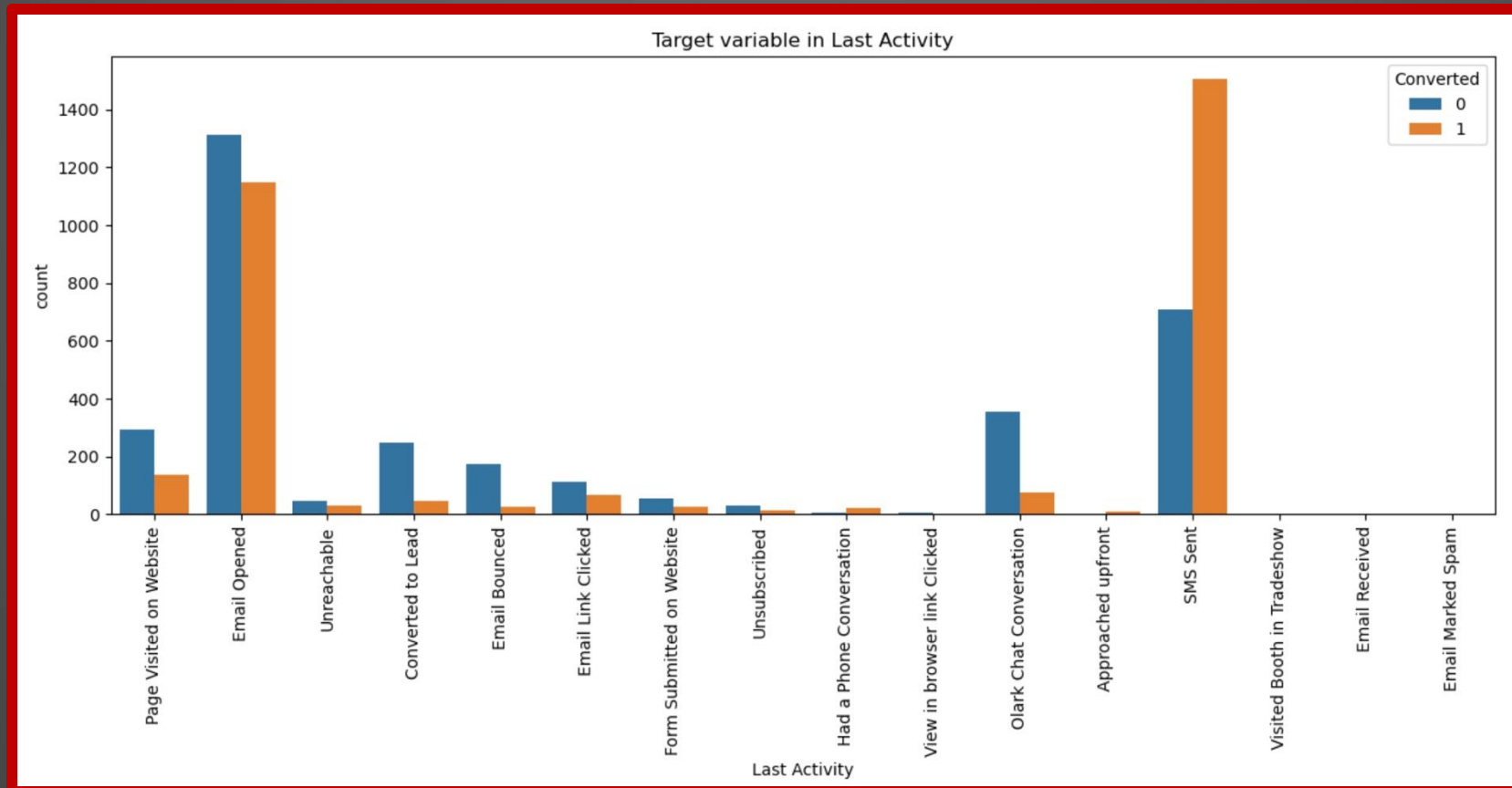
LEAD SOURCE AND LEAD ORIGIN

- In lead source the leads through google & direct traffic high probability to convert
- Whereas in Lead origin most number of leads are landing on submission



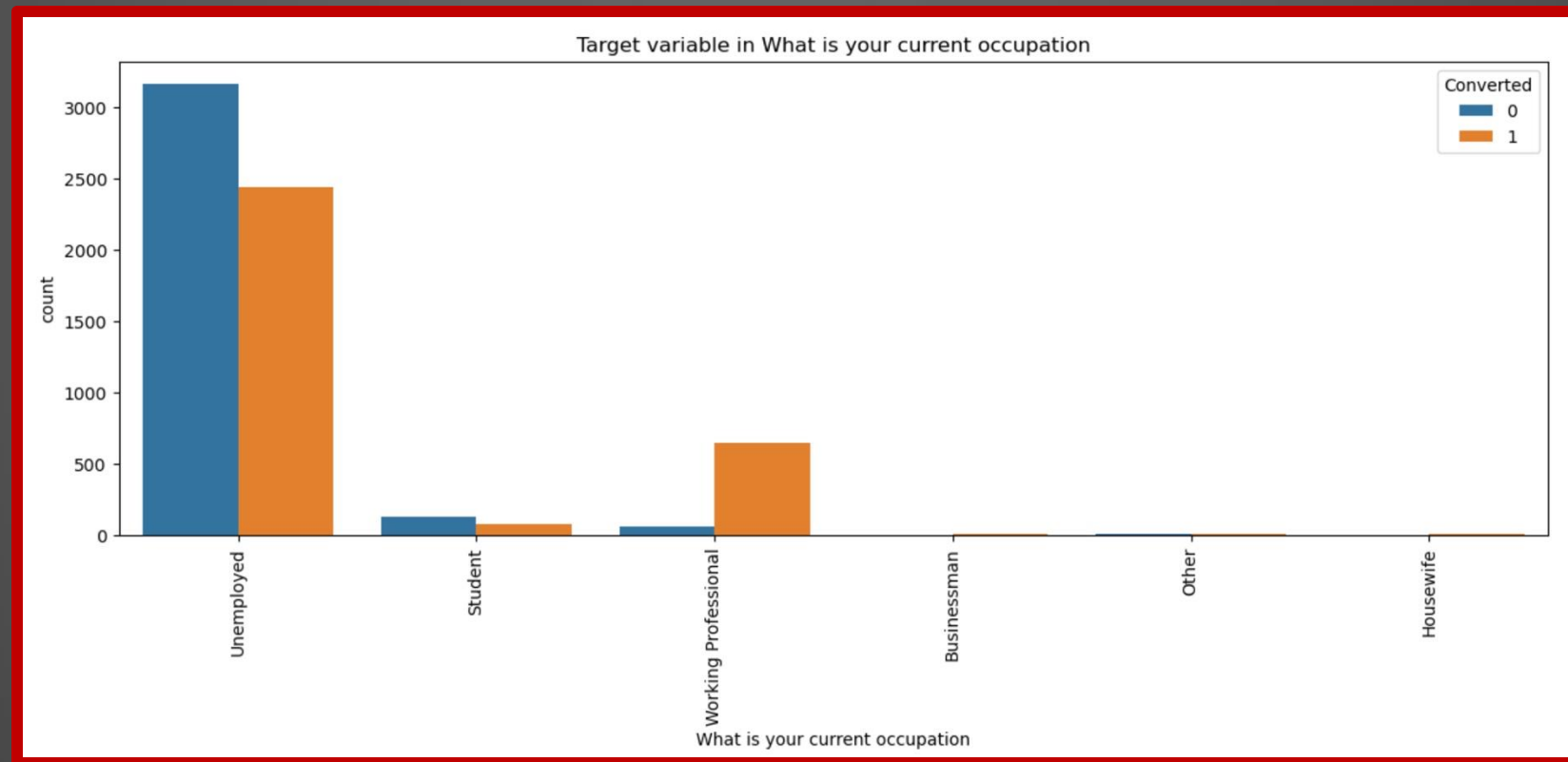
LAST LEAD ACTIVITY

- Leads which are opening email have high probability to convert, Same as Sending SMS will also benefit.



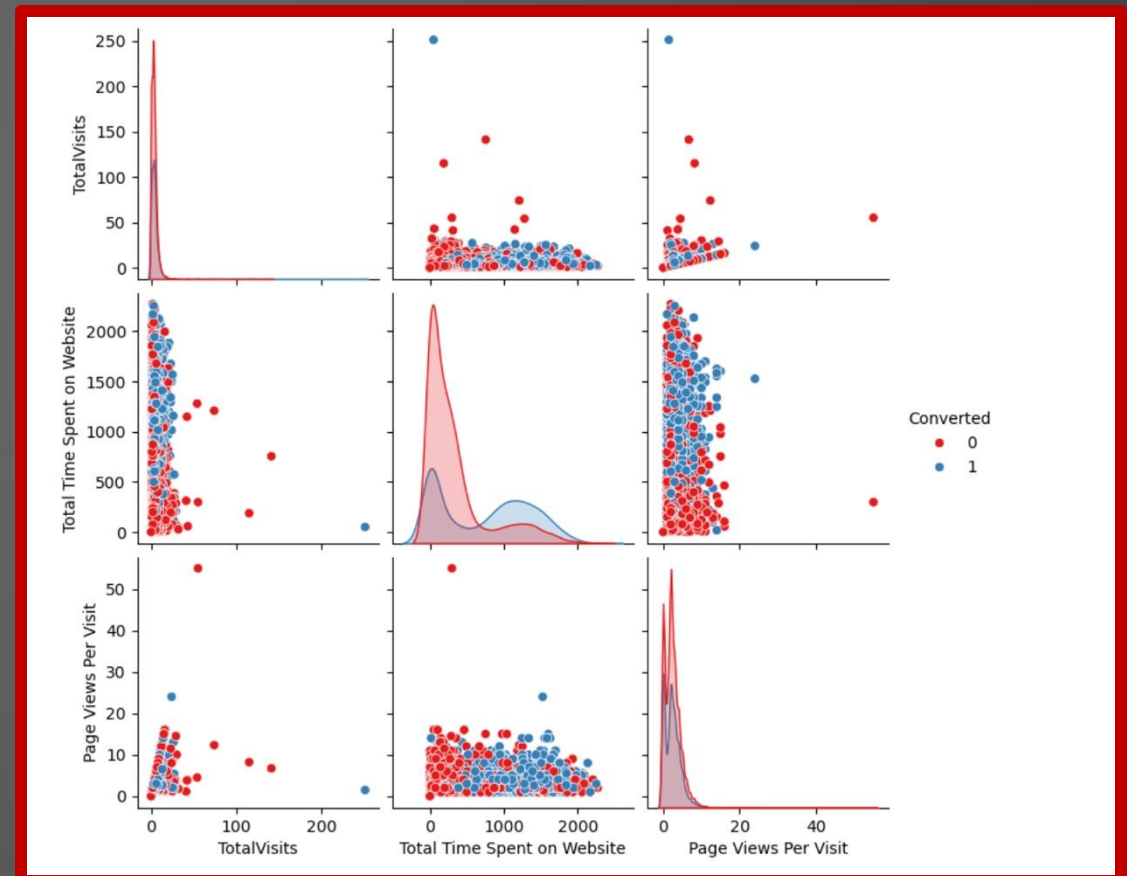
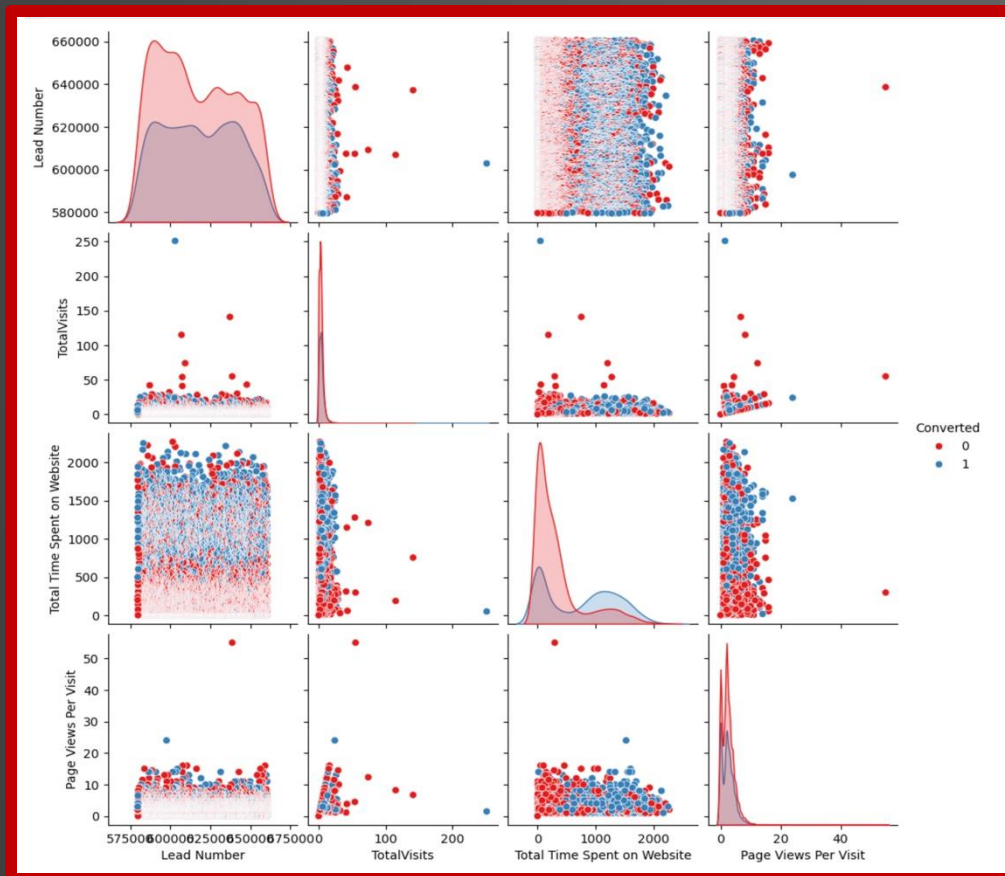
LAST WHAT IS YOUR OCCUPATION

- Leads which are Unemployed are more interested to join the course than others.



BIVARIATE ANALYSIS

- Bivariate analysis of Numerical Variables



CORRELATION

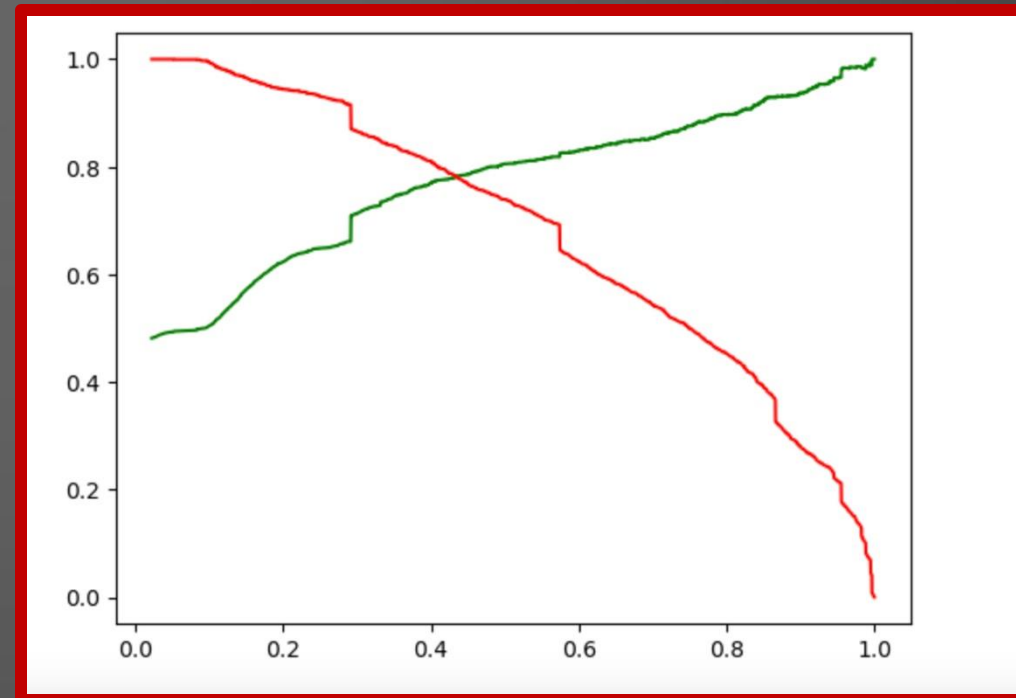
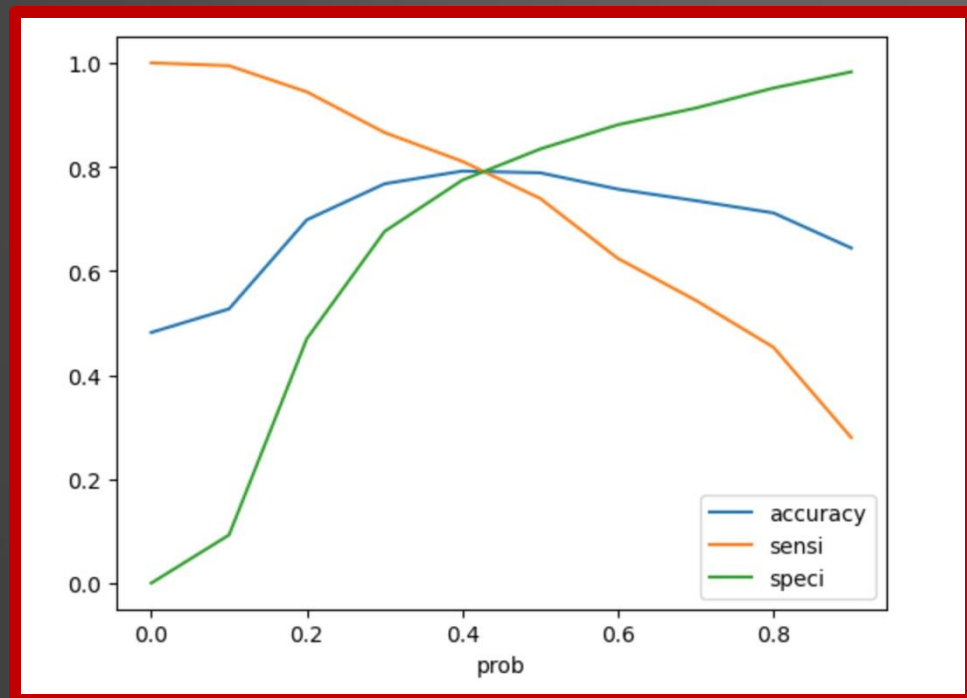
- There is no correlations between variables



MODEL EVALUATION

ROC CURVE

- **0.42 is the trade off between Precision and Recall -**
Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than 42 % to be a hot Lead



OBSERVATIONS

Train Data:

Accuracy : 80%

Sensitivity : 77%

Specificity : 80%

Test Data:

Accuracy : 80%

Sensitivity : 77%

Specificity : 80%

Final Features list:

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

CONCLUSION

- The conversion rate for API and Landing page submissions stands at 30-35%, which is approximately average. Conversely, the rates for Lead Add forms and Lead imports are notably low. This indicates a need to concentrate our efforts on leads generated through API and Landing page submissions.
- The majority of leads are sourced from Google and direct traffic, with the highest conversion rates coming from referrals and the Welingak website.
- Leads who spend more time on the website are significantly more likely to convert.
- The most frequent last activity recorded is the opening of emails, while the highest conversion rate is associated with SMS messages sent. Additionally, the largest group of leads consists of unemployed individuals, but the highest conversion rate is seen among working professionals.



Thank You !!