Lead Scoring Case Study using logistic regression

SUBMITTED BY :

1. **Ishan Srivastava**
2. **Kanak Gupta**

Contents

Problem Statement

1. **X Education** is an online learning platform that offers courses to industry professionals.
2. Every day, many professionals visit the website, browse through courses, and fill out a form to express interest. Once the form is submitted, the individual is recorded as a **lead**.
3. After acquiring leads, the **sales team** reaches out through calls and emails to convert them into paying customers. However, only a fraction of these leads actually convert.
4. The current **lead conversion rate** at X Education is around **30%.** This means that out of 100 acquired leads, only about **30 get converted**.
5. To improve efficiency, the company aims to identify **high-potential leads (Hot Leads)**.
6. By successfully identifying **Hot Leads**, the sales team can focus on the most promising leads instead of reaching out to everyone, leading to a higher conversion rate and better resource utilization.

Business Objective

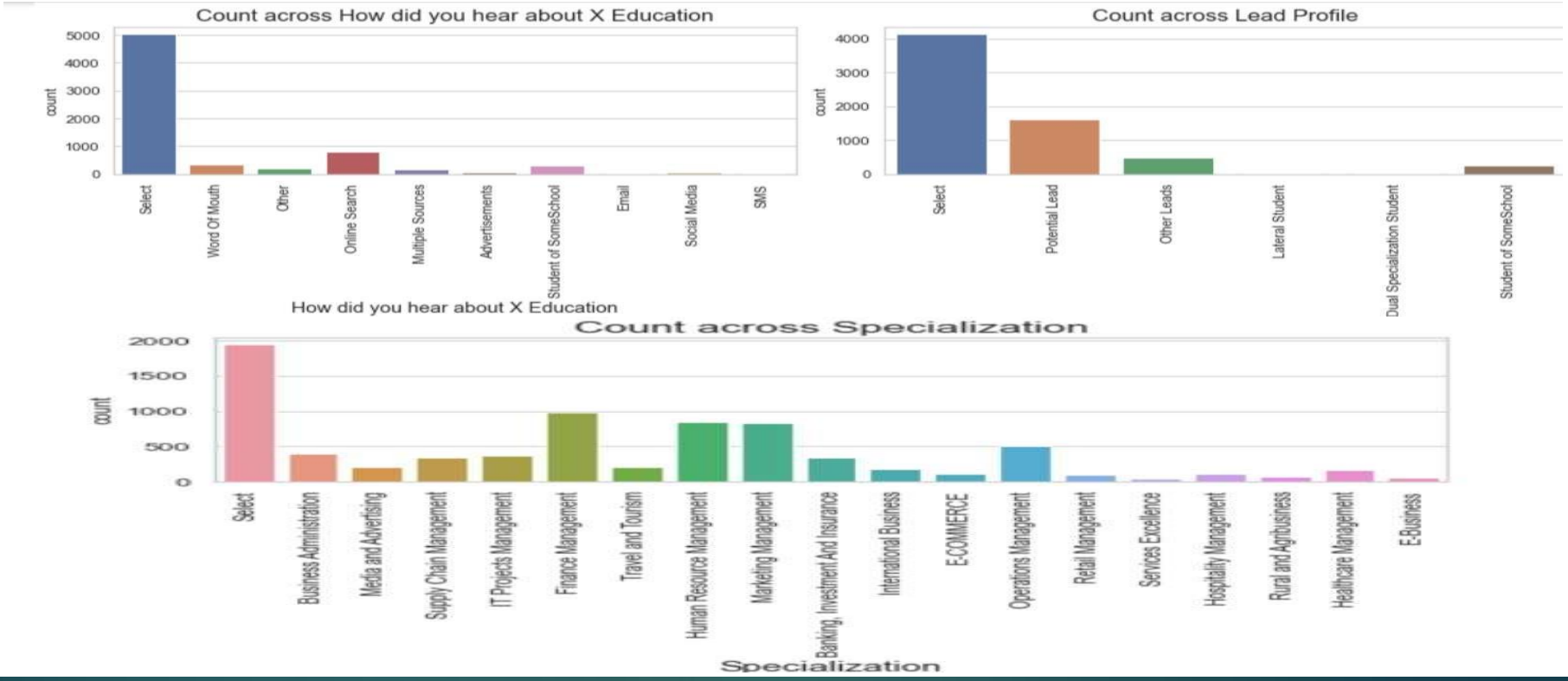
1. **X Education** wants to develop a model that assigns a **lead score (0-100) to each lead**. This will help in identifying **Hot Leads** and improving the **conversion rate**.
2. The **CEO's goal** is to increase the **lead conversion rate to 80%**.
3. The model should be designed to handle **future constraints**, including:
4. **Peak time strategies** to optimize lead conversion.
5. **Efficient utilization of manpower** to maximize productivity.
6. **Post-target strategies** to maintain performance after achieving conversion goals.

Problem Approach

1. 1.Data Import & Inspection – Load and check data quality.
2. 2.Data Preparation – Handle missing values and clean data.
3. 3.EDA – Analyze trends, distributions, and key variables.
4. 4.Dummy Variable Creation – Convert categorical variables.
1. 5.Test-Train Split – Split data into training and test sets.
2. 6.Feature Scaling – Normalize numerical variables.
3. 7.Correlation Analysis – Identify relationships between variables.
4. 8.Model Building – Use RFE, check R-squared, VIF, p-values, and fit Logistic Regression.
5. 9.Model Evaluation – Assess accuracy, precision, recall, F1-score, and ROC-AUC.
6. 10.Predictions on Test Set – Predict lead conversions and compare results.

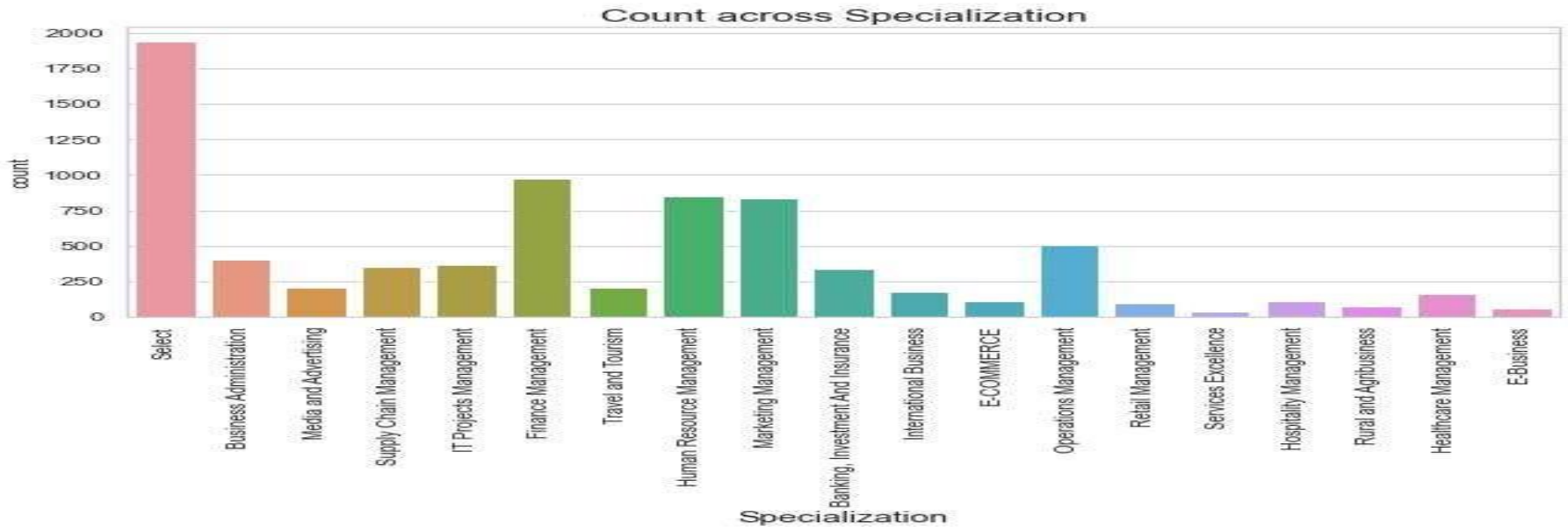
EDA – Data Cleaning

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



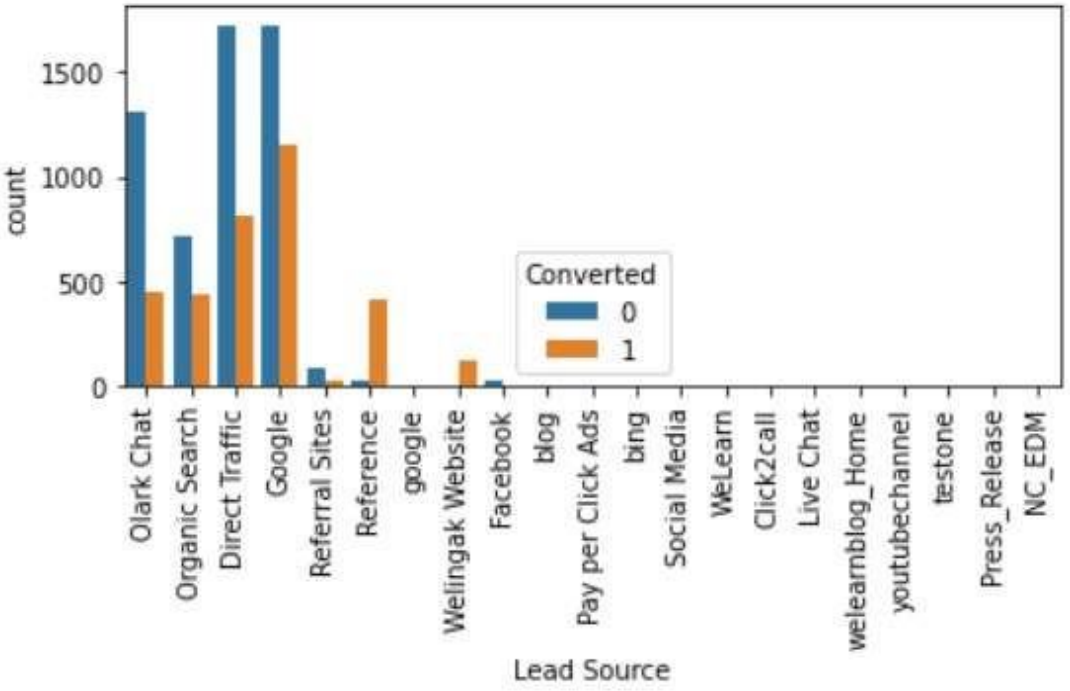
Specialization

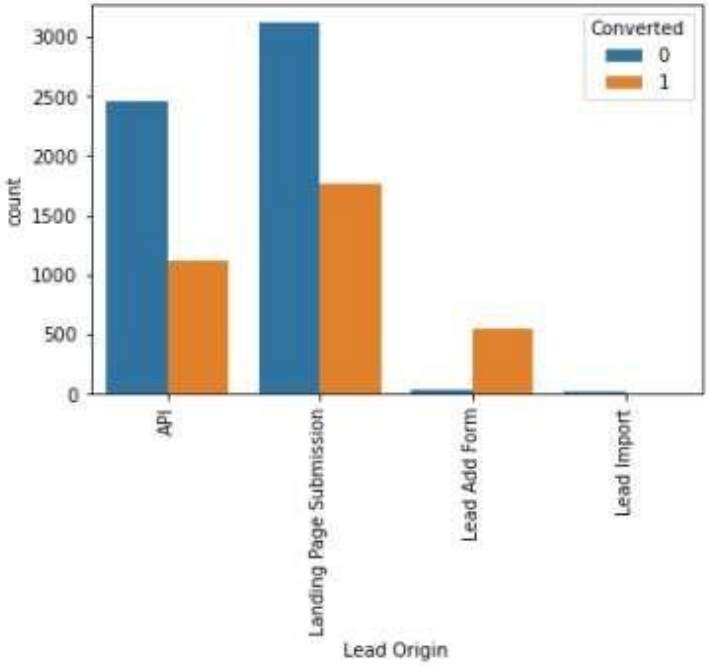
Leads from HR, Finance & Marketing management specializations are high probability to convert



Lead Source & Lead origin

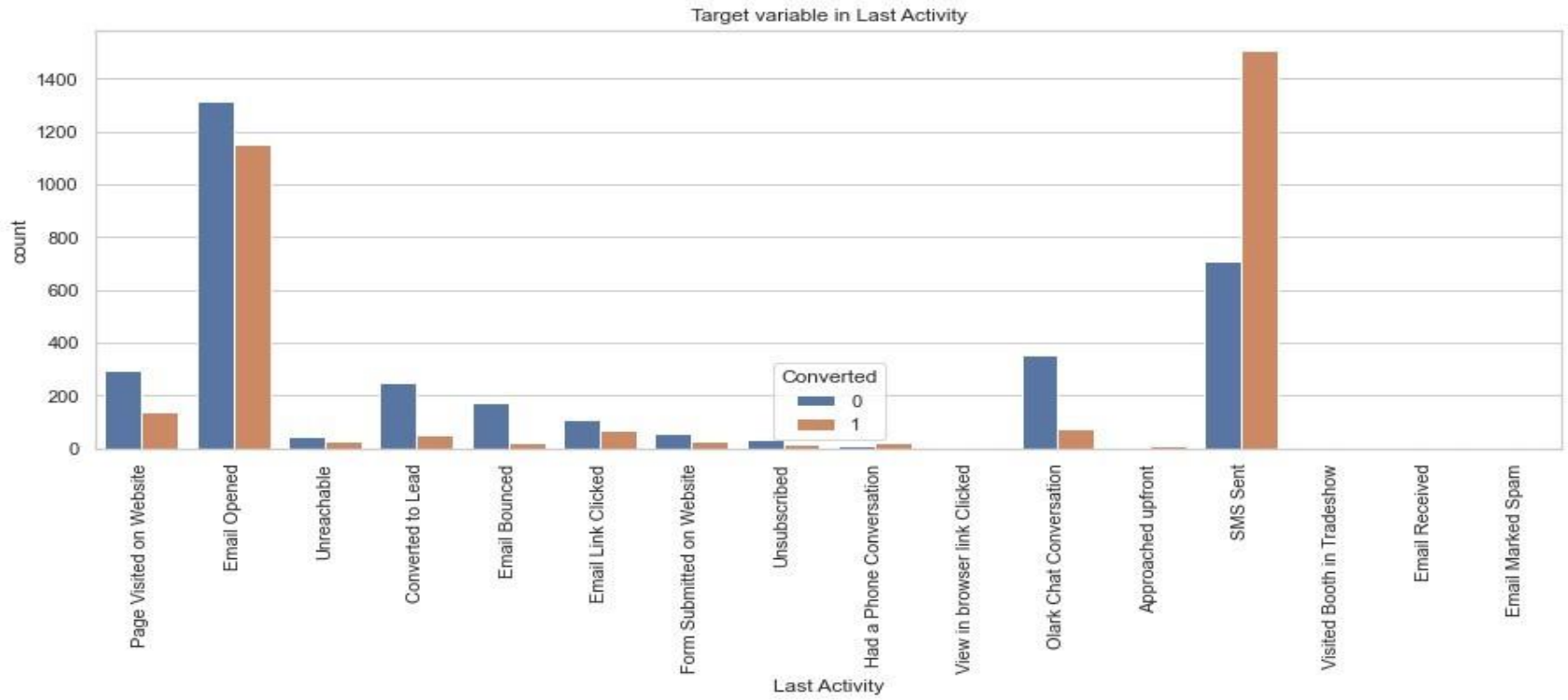
In lead source the leads through google & direct traffic high probability to convert Whereas in Lead origin most number of leads are landing on submission





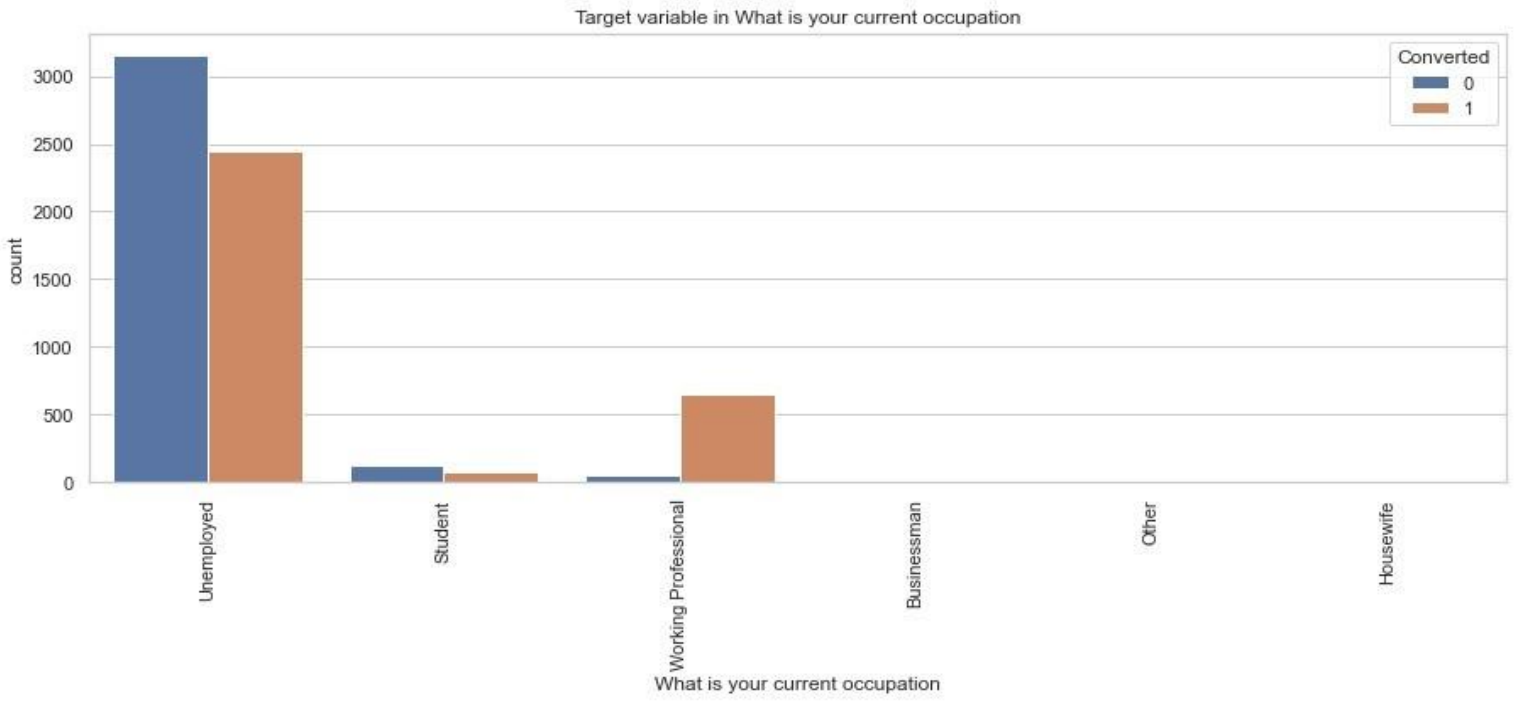
Last lead Activity

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



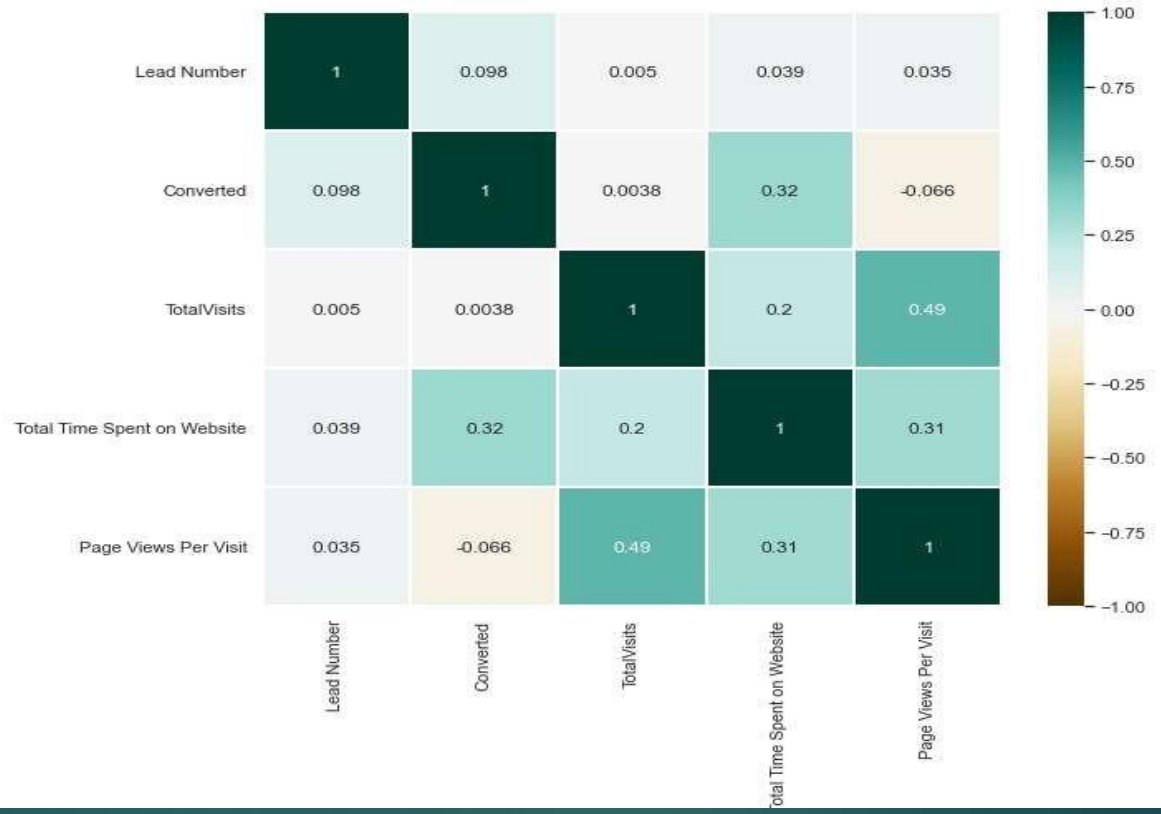
Last What is Your Occupation

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



Correlation

There is no correlation between the variables



Conclusion

•The conversion rate is 30-35% for leads coming from API and Landing Page Submissions, which is around the average. However, it is significantly lower for Lead Add Form and Lead Import. Hence, the focus should be on API and Landing Page Submission leads for better conversions.
•The highest number of leads are generated from Google and direct traffic, but the best conversion rates come from referrals and the Welingak website.
•Leads who spend more time on the website have a higher likelihood of converting.
•The most common last activity is email opened, but the highest conversion rate is observed for SMS sent.
•The majority of leads are unemployed, while working professionals have the highest conversion rate.