upGrad

upGrad & IIITB | Data Science Program

Lead Scoring Case Study

IPSITA PAL

ADITHYA BS

VIKAS KAUSHIK

BATCH ID 3459
DS51

Table of contents

Problem Statement

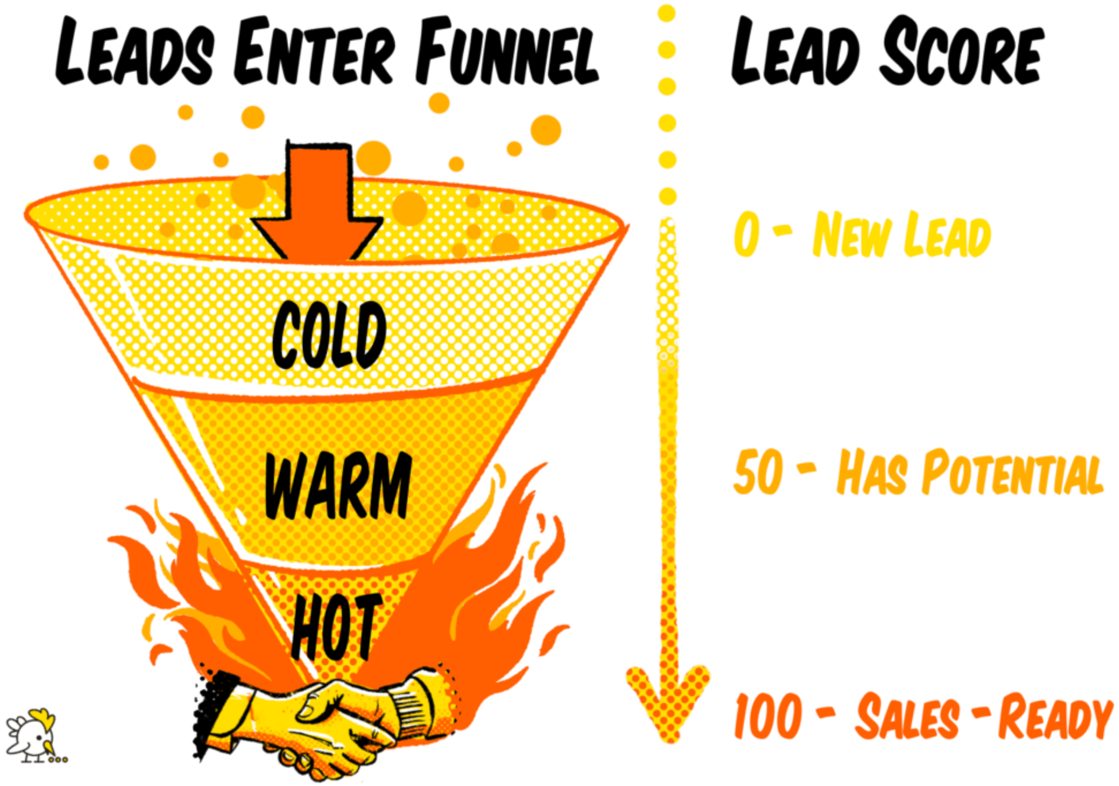
X Education is an online education company targeting industry professionals.

They attract potential customers through various marketing channels and collect leads via website forms and referrals.

However, their lead conversion rate is low, with only 30% of leads being converted. To improve efficiency, the company wants to identify "Hot Leads" with the highest conversion potential.

By focusing their efforts on these leads, the conversion rate is expected to increase. The task at hand is to build a lead scoring model that assigns scores to leads, enabling the sales team to prioritize high-scoring leads.

The CEO aims for a target lead conversion rate of approximately 80%.



Business Goal

The provided dataset contains approximately 9000+ data points, including attributes like Lead Source, Total Time Spent on Website, Total Visits, and Last Activity.

The target variable is 'Converted,' indicating whether a lead was converted (1) or not (0). Categorical variables may have a level called 'Select,' which is equivalent to a null value.

The case study aims to construct a logistic regression model that assigns a lead score between 0 and 100. This score will help the company target potential leads effectively, with higher scores indicating a higher likelihood of conversion.

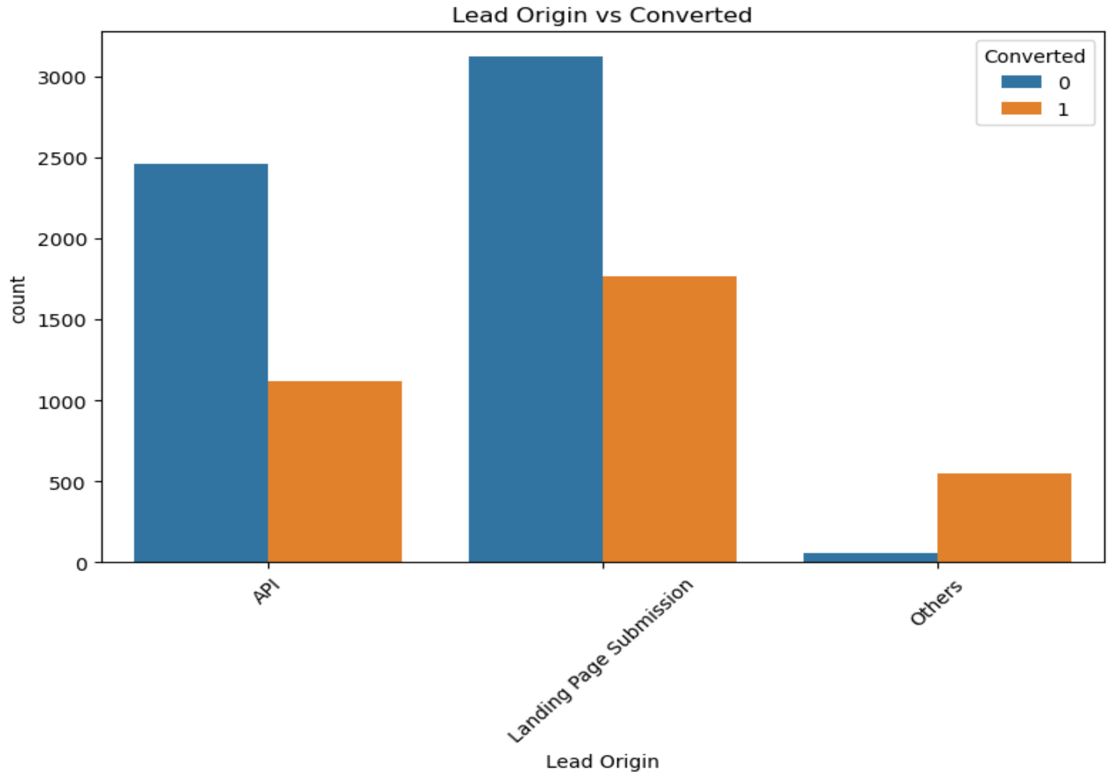
The model should also accommodate future changes in the company's requirements and handle additional problems as needed.

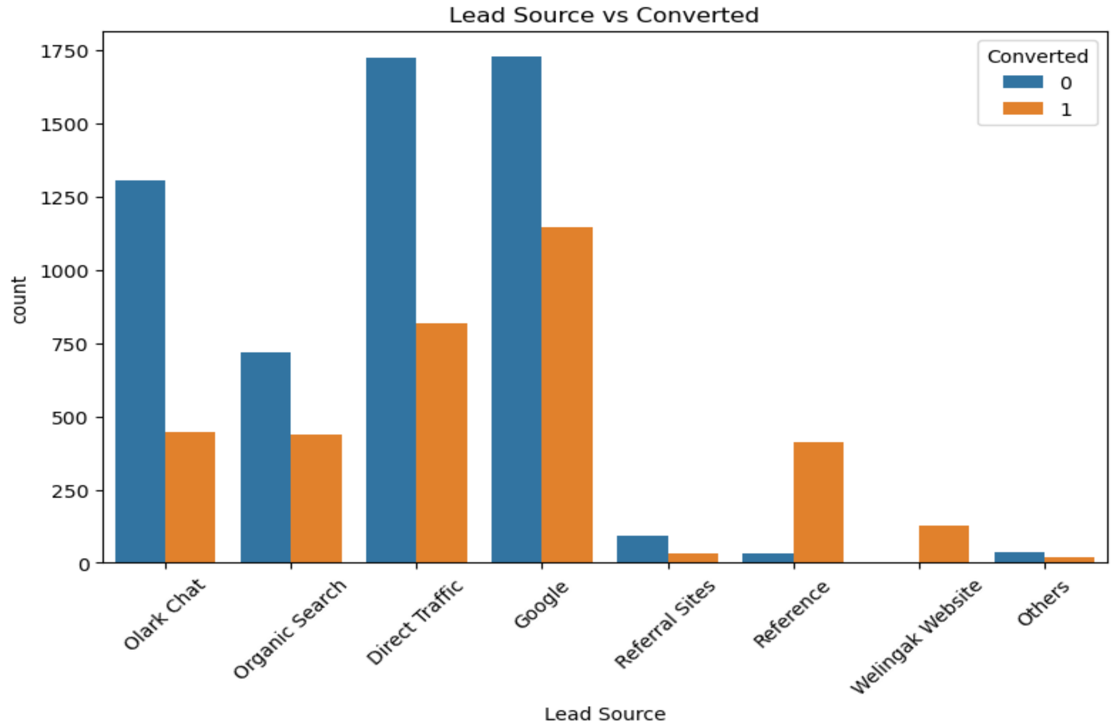
Strategy

- Load necessary libraries.
- Reading and understanding the dataset.
- Clean and drop unnecessary columns.
- Imputing, Caping variables wherever necessary
- Visualization
- Preparing dataset, creating dummies, converting yes/no
- Splitting data into train-test.
- Model building
- Evaluating the model
- Lead Scoring
- Summary

Exploratory Data Analysis

Categorical variables





Insight:

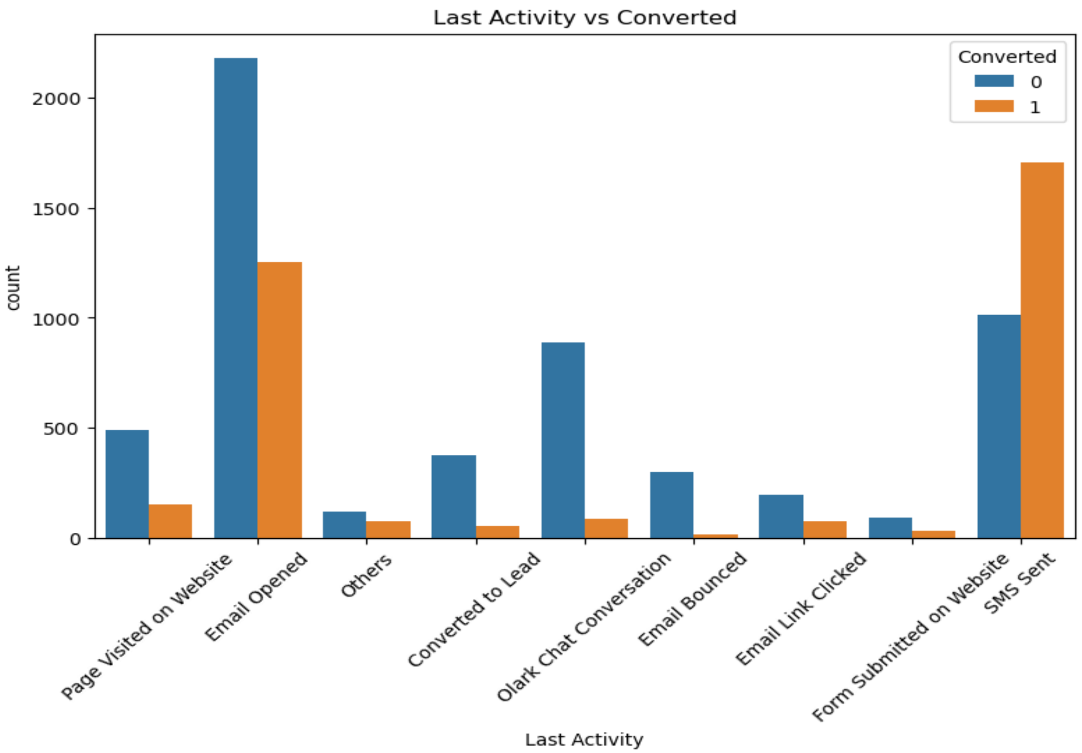
‘Lead Origin’: API and Landing Page Submission have relatively higher conversions compared to other Lead Origins

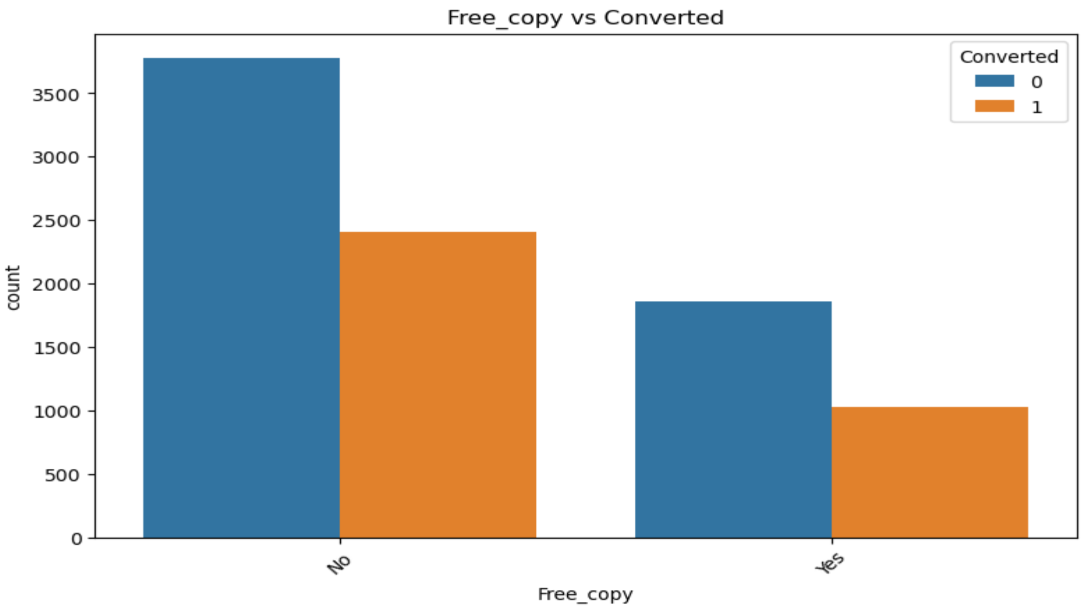
Insight:

‘Lead Source’: Google and Direct traffic have higher conversions compared to other Lead Sources. Referral Sites have the lowest of all.

Exploratory Data Analysis

Categorical variables





Insight:

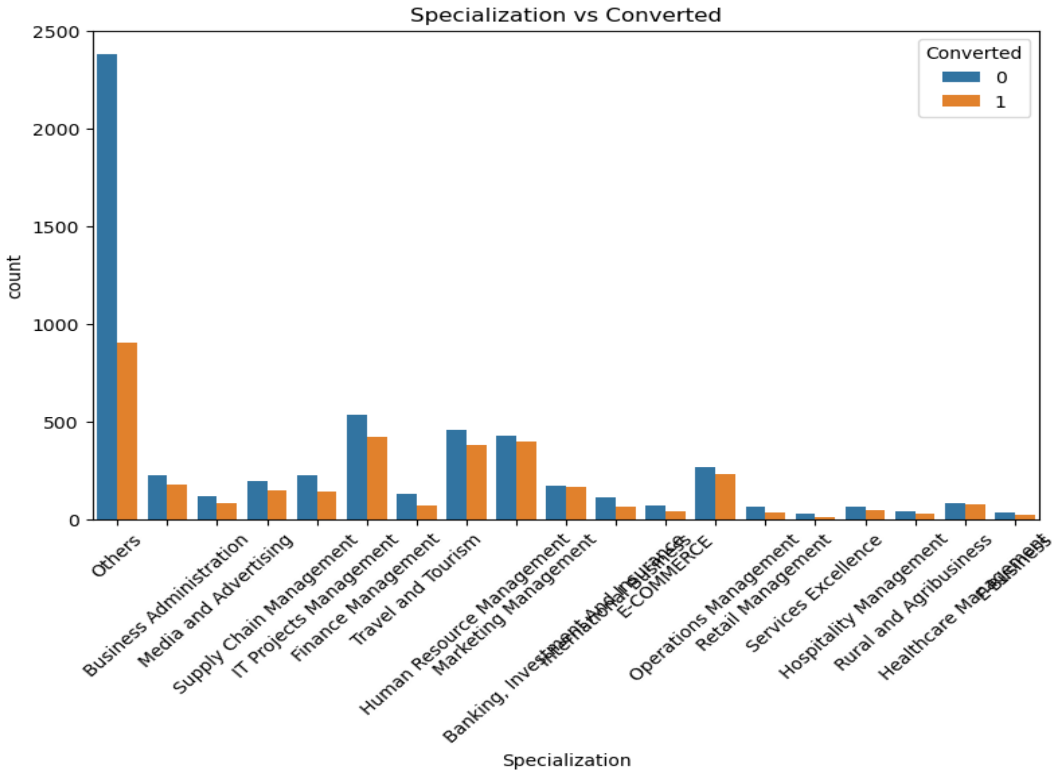
‘Last Activity’: Email Opened and SMS sent contribute the higher leads and Email Bounced/ Email Link clicked has the lower conversion rate.

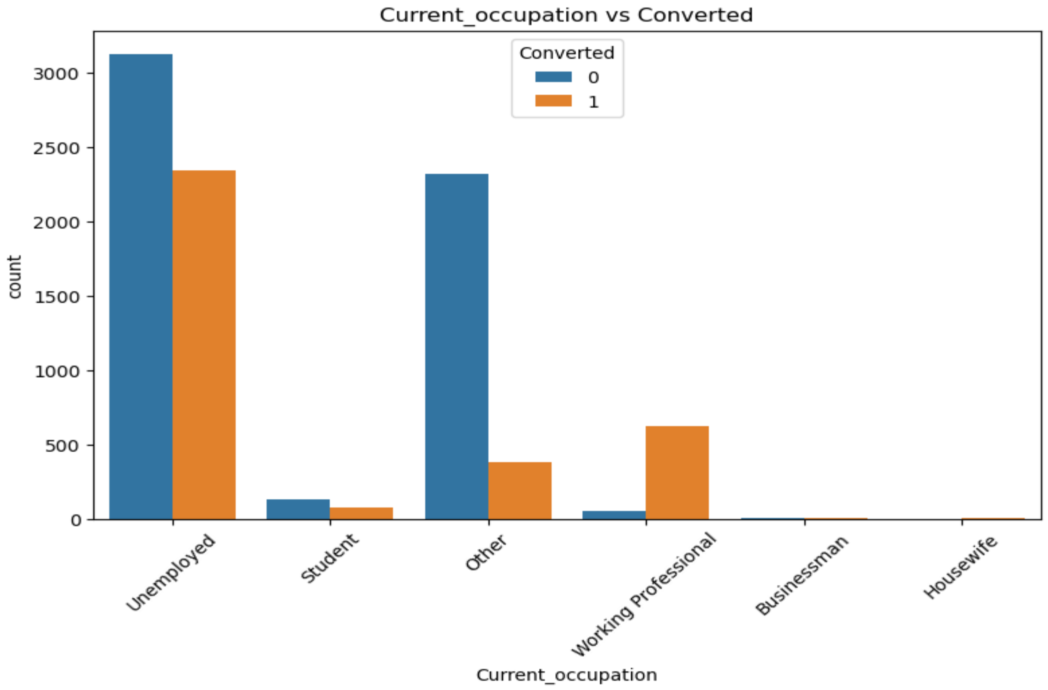
Insight:

‘Free copy’: Giving a free copy for the interview doesn't seems to have much impact with lead conversion.

Exploratory Data Analysis

Categorical variables





Insight:

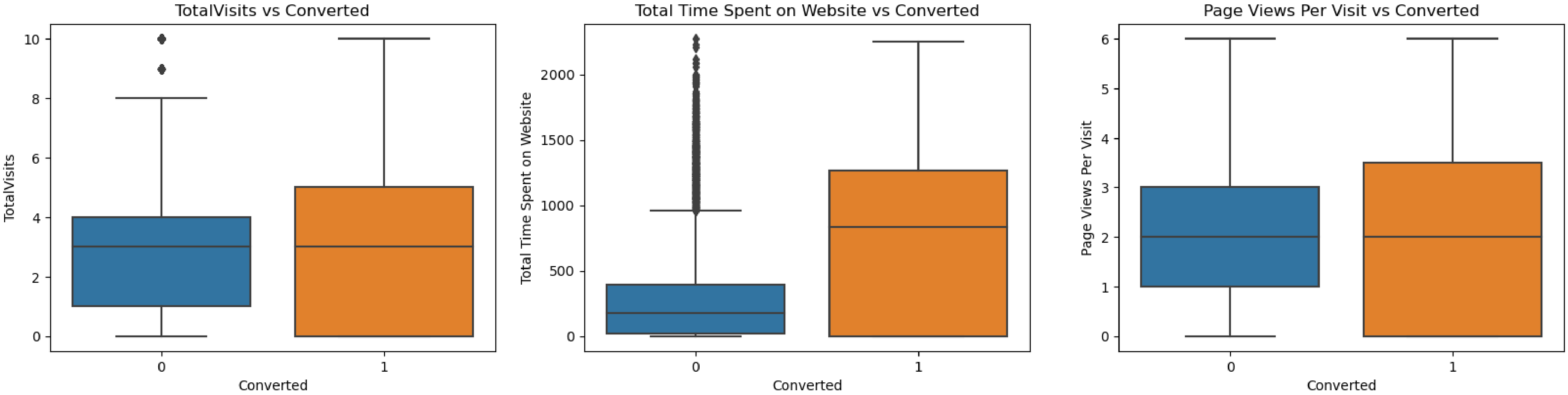
‘Specialization’: Certain Specialization like Finance Management and Human Resource Management can be targeted as they have more lead conversion than rest of Specialization.

Insight:

‘Current occupation’: Unemployed and working professionals have higher conversion rates while Businessman and Housewife have the least conversion rates.

Exploratory Data Analysis

Numerical variables



Insight:

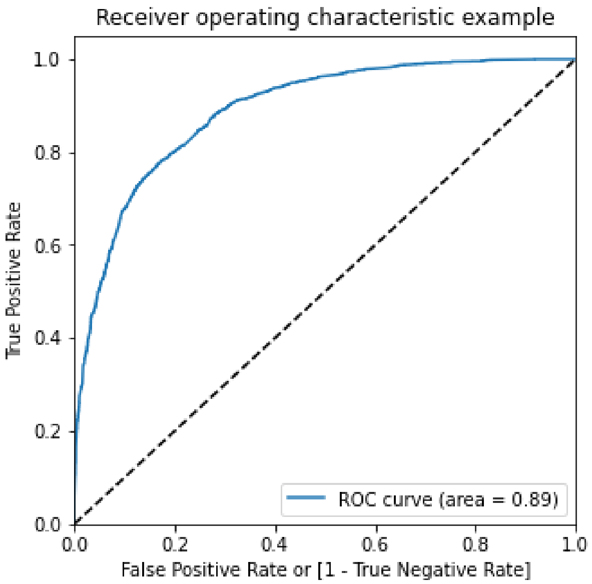
‘Total Visits’: There is no major insight with this visualization.

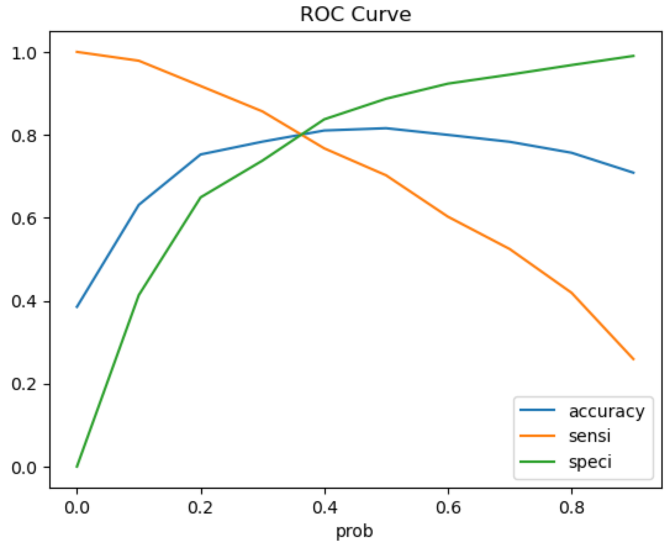
‘Total Time Spent on Website’: Customers spending more time at website have higher conversion rate.

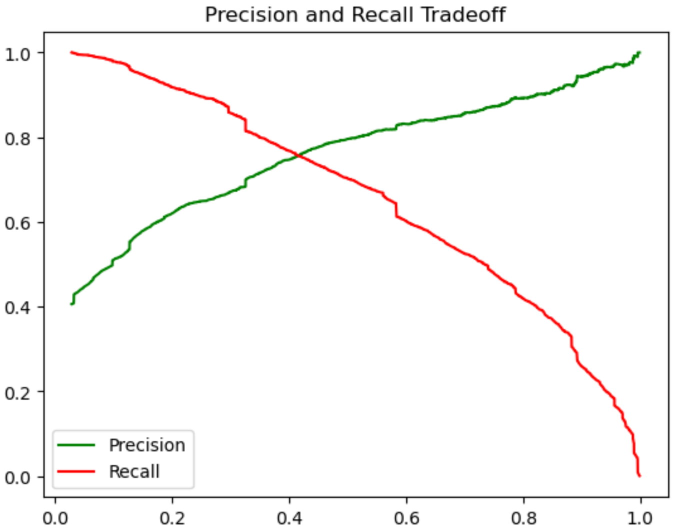
‘Page Views Per Visit’: There is no major insight with this visualization, but the overall conversion rate was higher for Total Visits and Total Time Spent on Website.

Model Evaluation

The logistic regression models were constructed using the Manual Feature Reduction process, wherein variables with a p-value greater than 0.05 and a VIF higher than 5 were dropped. A total of 3 models were built, leading to the final stable model (Logm3) consisting of 13 variables.







The area under the curve of the ROC is 0.89 which is a good model.

By evaluating the tradeoff between sensitivity and specificity, the optimal cutoff point for probability was determined to be 0.38. Additionally, a precision-recall tradeoff curve was plotted, resulting in a cutoff of 0.44. Therefore, a cutoff probability of 0.44 was utilized.

Summary

Train Data:

Accuracy: 81.62%
Sensitivity: 74.08%

Test Data:

Accuracy: 82.07%
Sensitivity: 74.11%

- The company should give more importance to leads originating from the "Landing Page" as they have a higher conversion rate.
- Focus on targeting individuals who are unemployed, working professionals, and students.
- Active leads who have opened emails should be a priority.
- Total Time Spent on Website is a significant factor that should be considered.
- The conversion rate is 95% for lead scores greater than or equal to 95 and 87.02% for lead scores greater than or equal to 80.
- There are 447 leads from the given dataset that can be contacted and have a high chance of conversion if we target lead scores of 80% or higher.
- The top 3 features contributing to lead scores are:
- Lead Source: Welingak Website
- Lead Source: Reference
- Current Occupation: Working Professional

Overall, the model achieved an accuracy of 82.07% on the test set, with sensitivity and specificity close to 74%. By focusing on the recommended features and targeting leads with high scores, the company can effectively increase the chances of conversion.

Recommendations

Optimize Lead Conversion:

- Prioritize high-quality leads based on a selected threshold.
- Develop engaging emails and SMS messages to enhance lead interaction.
- Improve website interactivity for increased user engagement.
- Optimize parameters strongly correlated with lead conversion.
- Offer targeted discounts to specific leads to incentivize conversion.
- Gather customer feedback to improve lead quality and boost conversion rates.

Focus on High-Potential Leads:

- Target leads from "Welingak Website" and "Reference" sources.
- Give priority to leads with the occupation of "Working Professional."
- Contact leads who have received SMS and opened emails.

Additional Considerations:

- Pay attention to leads with substantial website browsing time.
- Reach out to leads who previously displayed conversion potential.
- Utilize the provided table to identify positive parameters and address negative factors hindering lead conversion, leading to overall improvement.

Thank you!