X Education - Lead Scoring Case Study

Identifying high-potential leads to enable marketing to prioritize their conversion efforts.

Team Members : Abhinav Mohan Agarwal, Jeffin Joe Jacob , Ajith Krishnanunni

Background of X Education

- Business Focus: Specializes in offering online courses tailored for industry professionals.
- Lead Generation: Attracts potential customers through website visits, course exploration, form submissions, and video engagement.
- Lead Sources: Obtains leads both through direct interactions on the website and from past customer referrals.
- Conversion Process: Utilizes a sales team to engage and convert leads through calls, emails, and other communication methods.
- Current Lead Conversion Rate: Achieves an average lead conversion rate of approximately 30%.

The problem Statement

Problem Statement - X Education faces a challenge in converting potential leads into paying customers. While numerous leads are generated initially, only a fraction ultimately become paying customers, leading to an unsatisfactory lead conversion rate. To address this issue, the company requires a lead scoring model to identify the most promising leads with a higher likelihood of conversion.

Objective - The primary objective of this case study is to build a logistic regression model capable of assigning a lead score to each potential customer, ranging from 0 to 100. This lead score will enable the company to prioritize its efforts by targeting leads with higher scores, indicating a greater chance of conversion. Conversely, leads with lower scores are less likely to convert. The CEO has set a target lead conversion rate of around 80%, and the model aims to assist in achieving this goal.

Approach

- Acquire the data for analysis.
- Perform data cleansing and preparation.
- Conduct Exploratory Data Analysis.
- Scale the features as needed.
- Partition the data into training and testing datasets.
- Construct a logistic regression model to compute the Lead Score.
- Assess the model using various metrics, including Specificity and Sensitivity or Precision and Recall.
- Apply the most suitable model to the test data based on Sensitivity and Specificity metrics.

Data Cleansing

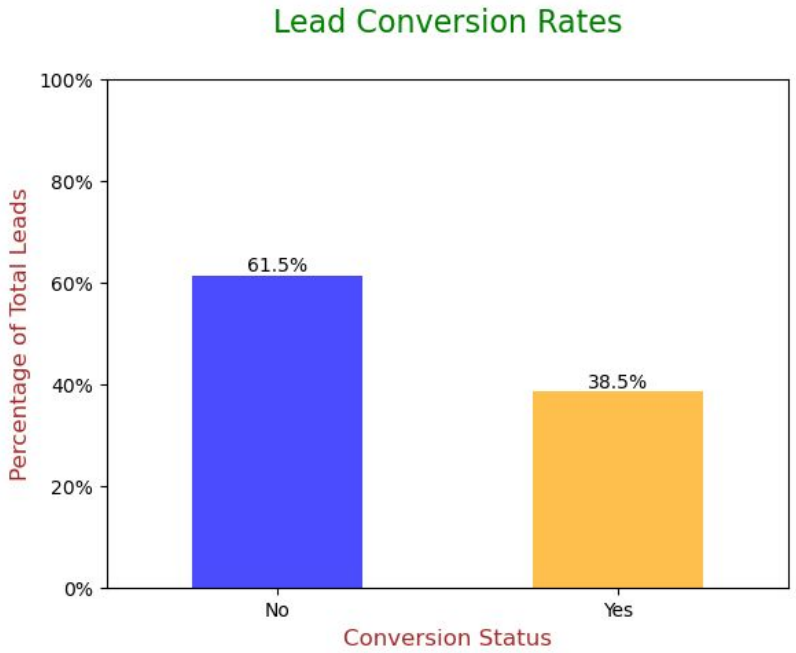
- The presence of the "Select" level in categorical variables indicates that customers didn't make any selections, essentially representing null values.
- Any columns containing more than 40% null values were eliminated from the dataset.
- Missing values within categorical columns were addressed by considering value counts and other relevant factors.
- Unnecessary columns, such as those related to tags and country, were excluded to maintain a focus on the study objectives.
- For some categorical variables, imputation was employed to fill in missing data.
- Certain variables saw the introduction of additional categories to improve data representation.

Data Cleansing

- Columns that held no relevance for modeling, like Prospect ID and Lead Number, or those with only one possible response, were removed.
- Numerical data was imputed using the mode.
- Skewed categorical columns were removed to prevent bias.
- Outliers in TotalVisits and Page Views Per Visit were addressed, and data standardization, correction of invalid values, grouping of low-frequency values under "Others," and encoding of binary categorical variables were carried out as part of a comprehensive cleaning process to improve data quality and precision.

Exploratory Data Analysis

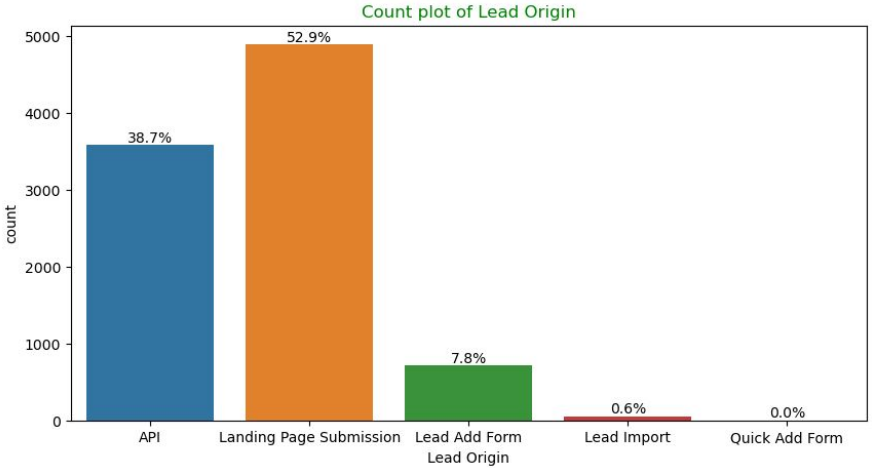
- Lead Conversion Rate in 39%

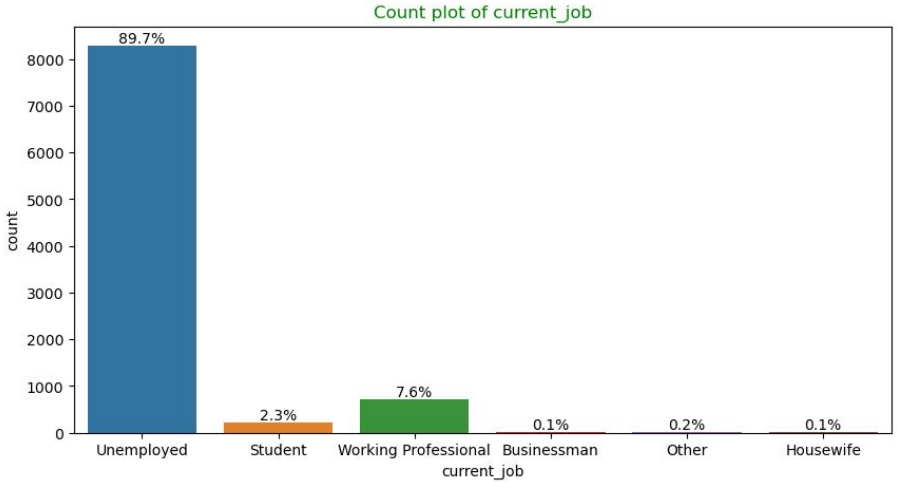


- The conversion rate is 38.5%, which means that only a small portion, 38.5%, has become leads.
- 61.5 % leads were not converted

Exploratory Data Analysis (Univariate)

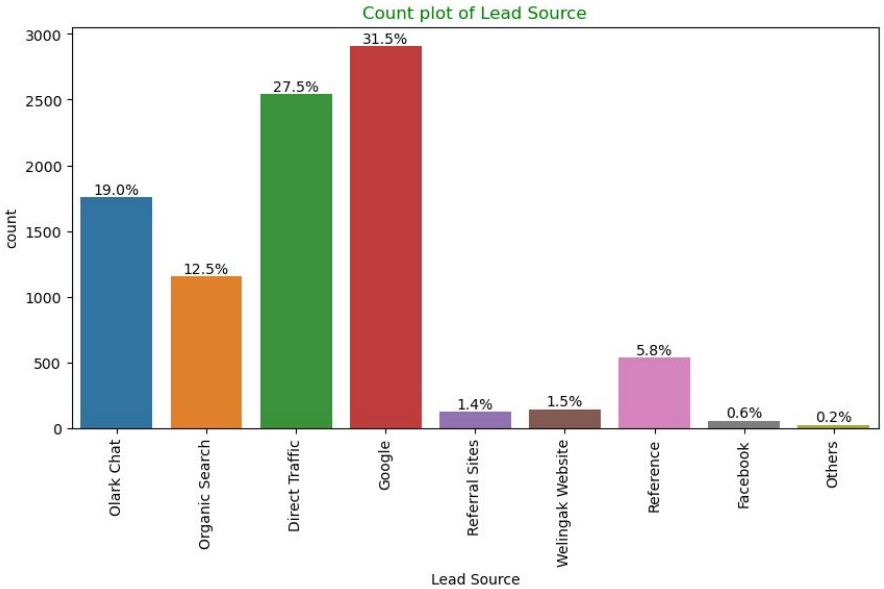
- Majority of the leads are from Landing Page submissions
- Majority of the Customers are unemployed

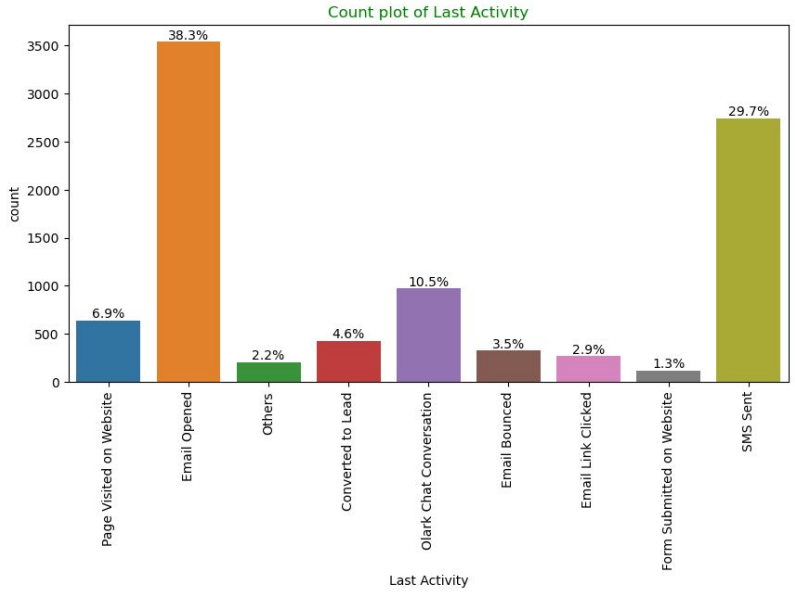




Exploratory Data Analysis (Univariate)

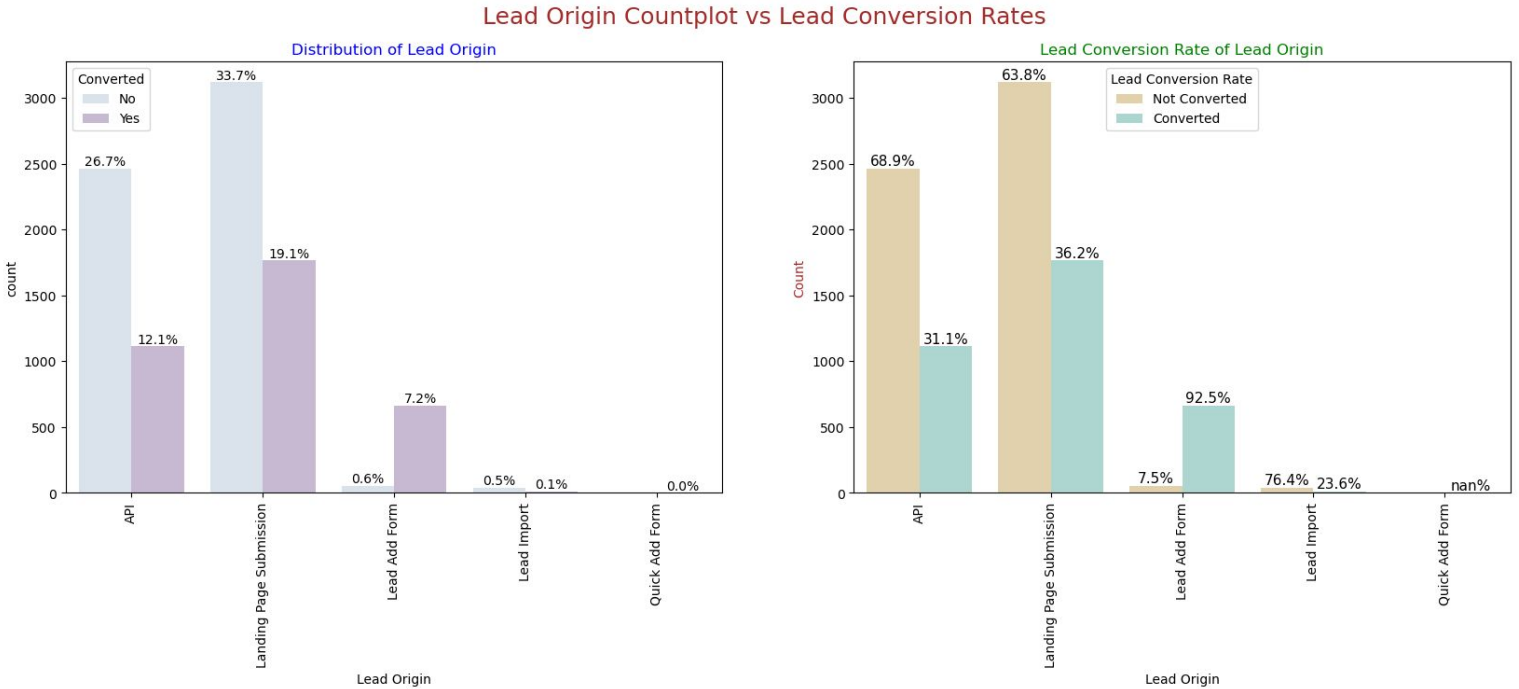
- Google and Direct Traffics are major lead sources
- Major recent activity my customer is “Email Opened”





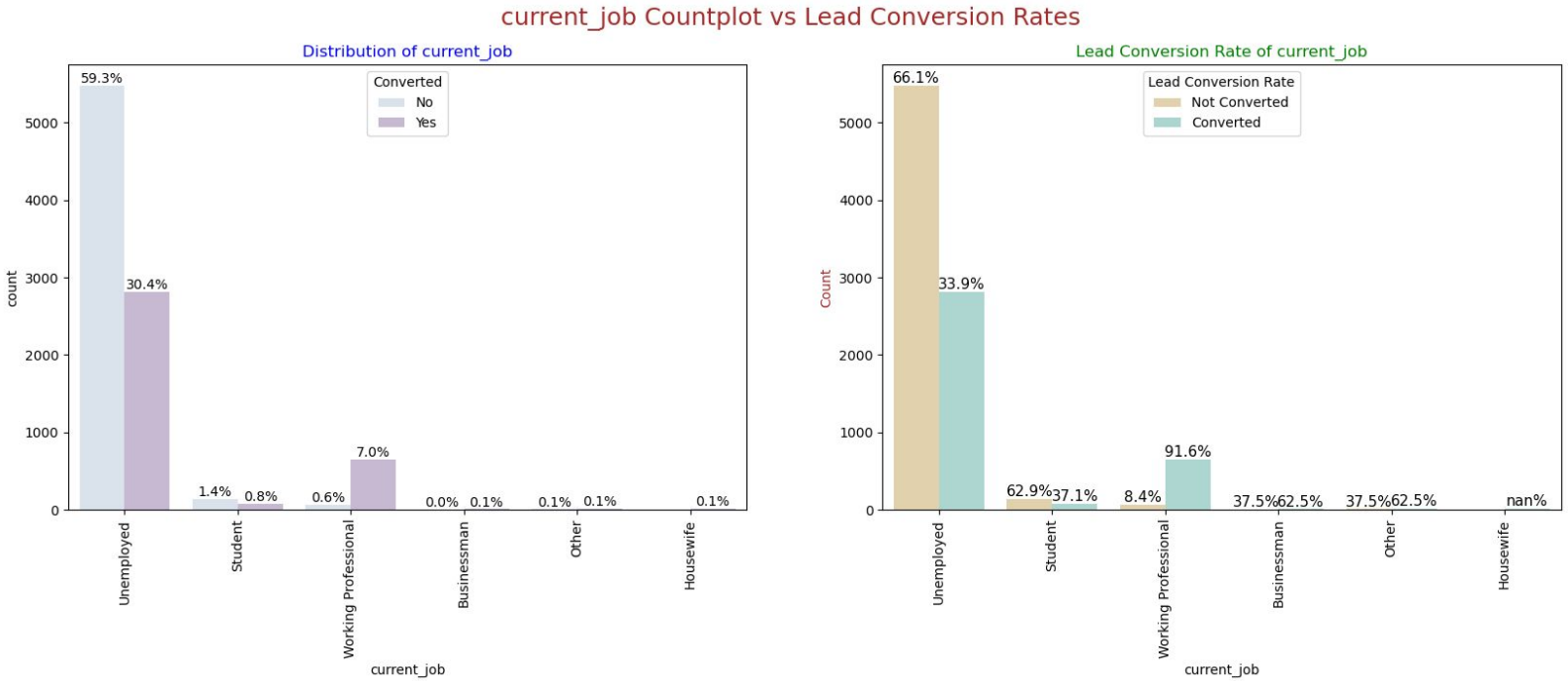
Exploratory Data Analysis (Bivariate categorical)

- 52% leads originated from Landing pages with lead conversion rate
- Leads from Ads has got 92% conversion



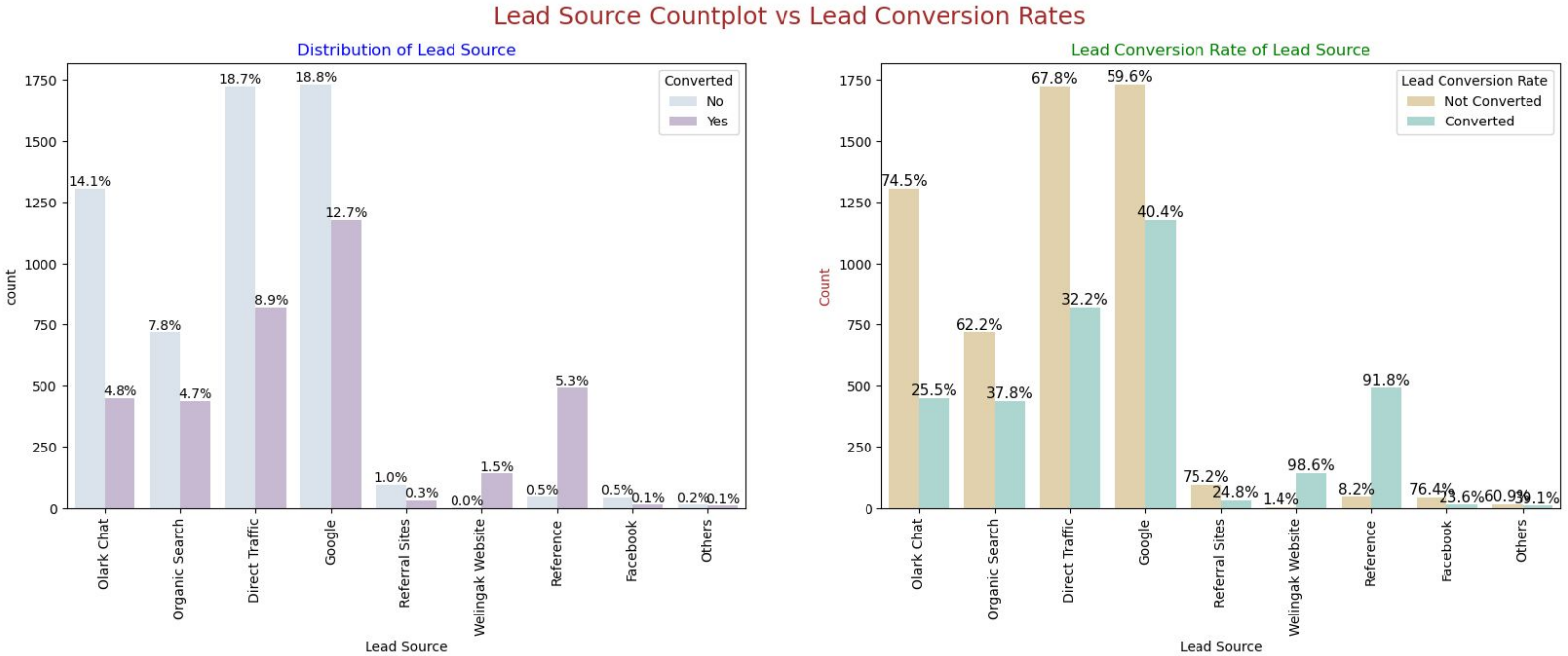
Exploratory Data Analysis (Bivariate categorical)

- 90% customers are unemployed with approx 34% conversion
- Working professionals got highest (92%) conversion



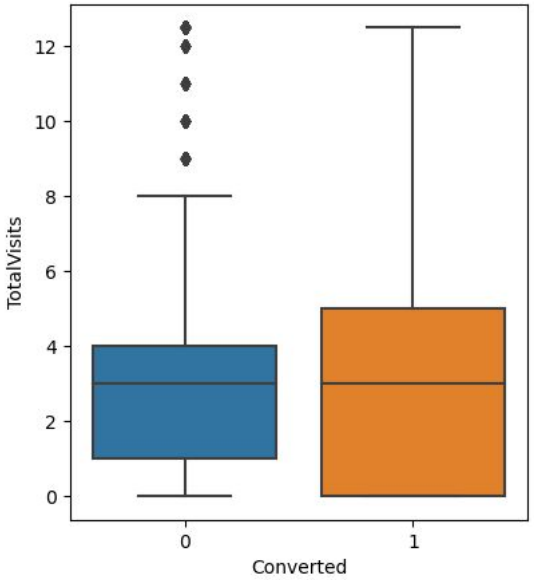
Exploratory Data Analysis (Bivariate categorical)

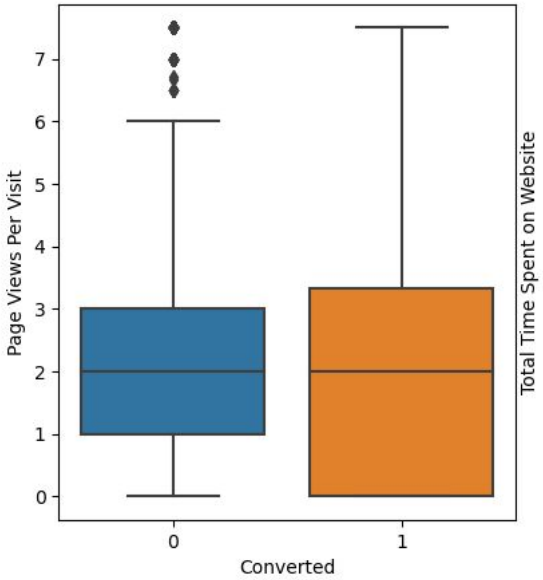
- Google has Conversion rate of 40% which is a decent number
- References has more conversion rate even though overall leads are less

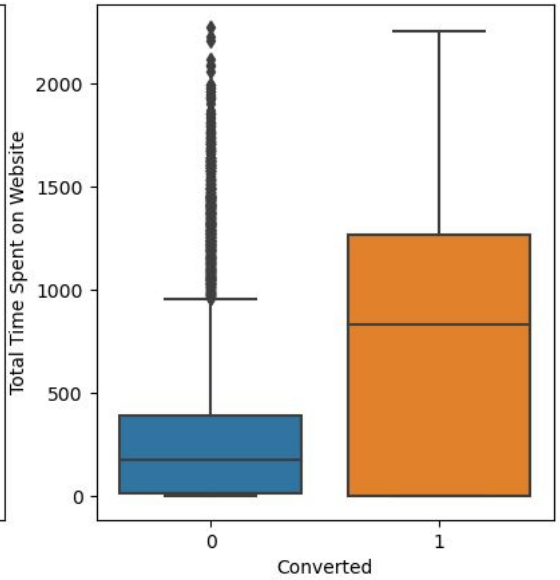


Exploratory Data Analysis (Bivariate numerical)

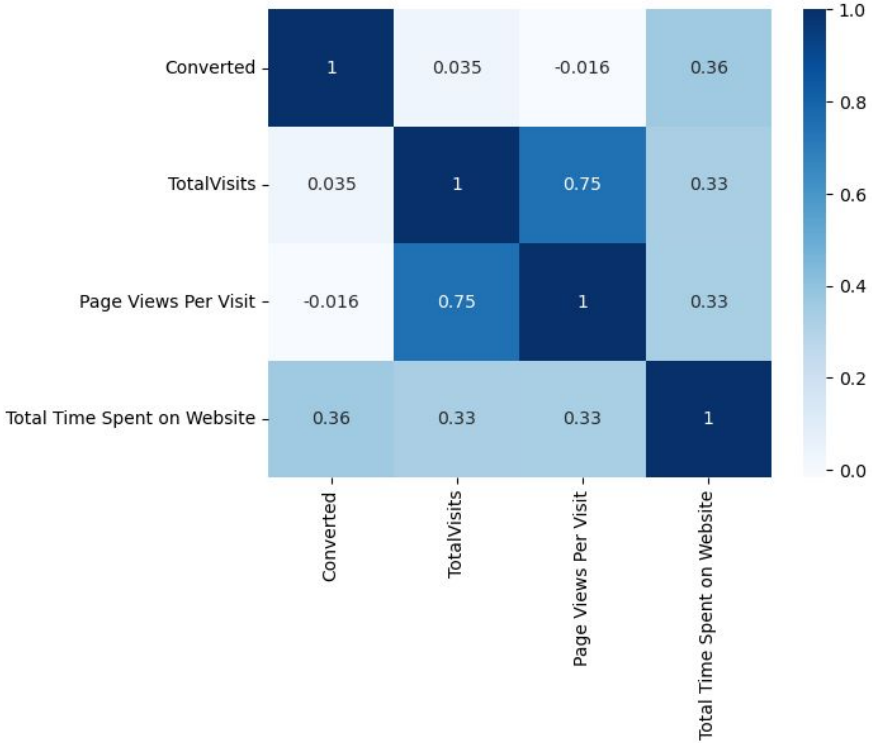
- Old leads where customer spent lot of time on websites has higher conversion rates







Exploratory Data Analysis (Bivariate numerical)



Variables Impacting Model

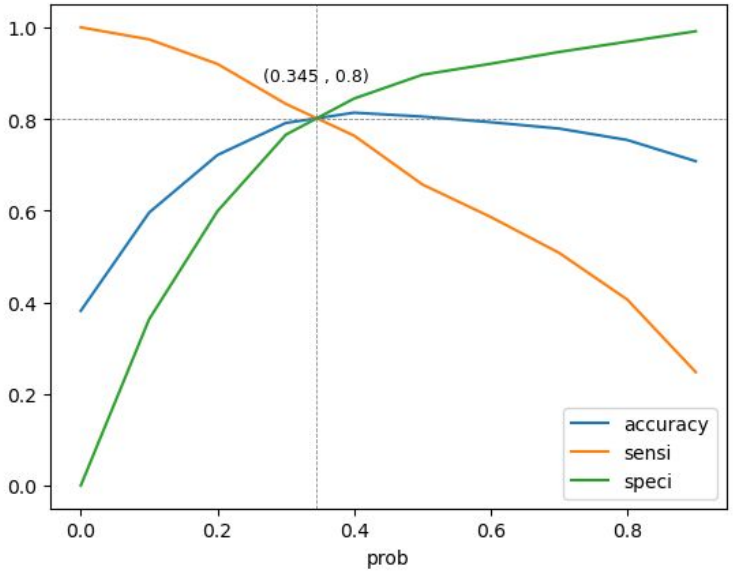
- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin –Lead Page Submission
- Last Activity –Email Bounced
- Current Occupation –No Information
- Current Occupation –Working Professional
- Last Notable Activity –Had a Phone Conversation
- Last Notable Activity -Unreachable

Model Build/Selection

- The dataset contains numerous dimensions and a vast array of features.
- This can lead to decreased model performance and potentially require extensive computational resources.
- Therefore, it's crucial to conduct Recursive Feature Elimination (RFE) to identify and retain only the essential columns.
- Following RFE, we can proceed with manual model fine-tuning
- The models were constructed using a manual feature reduction process, where variables with p-values exceeding 0.05 were eliminated
- After four iterations, Model 4 appears to be consistent or steady

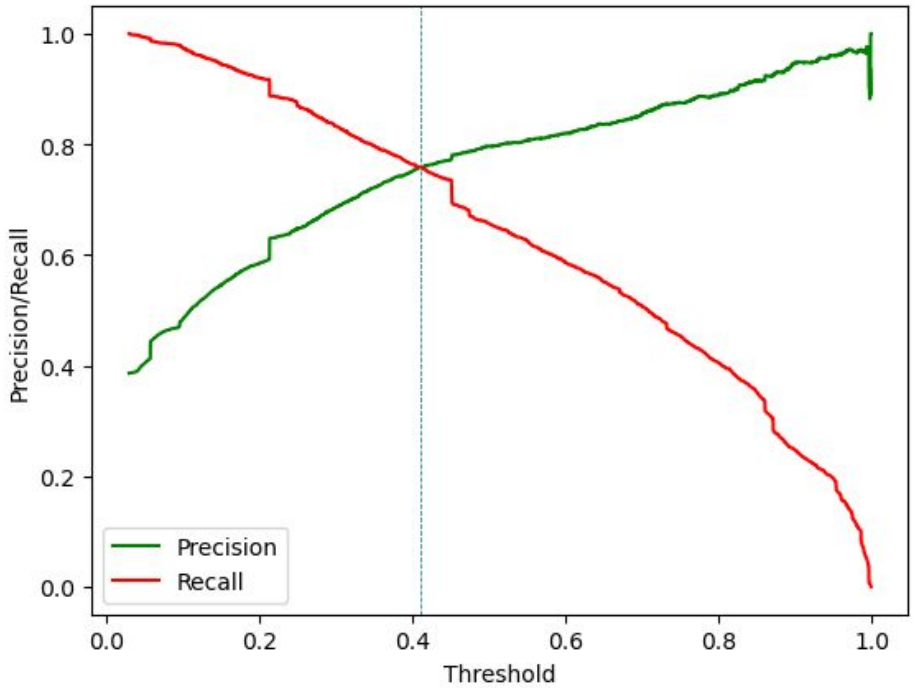
Model Evaluation - Train Dataset

- After examining evaluation metrics from both plots, the decision was made to proceed with a cutoff value of 0.345
- Confusion Matrix & Evaluation Metrics with 0.345 as cutoff



Model Evaluation - Train Dataset contd.

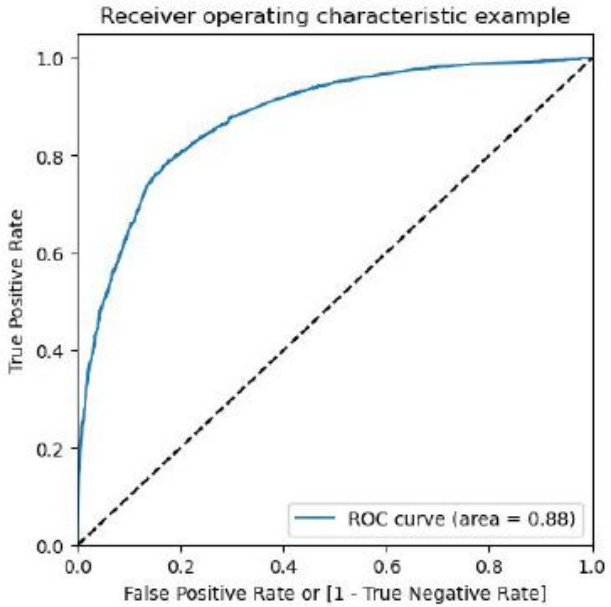
- Confusion Matrix & Evaluation Metrics with 0.41 as cutoff



Model Evaluation

ROC Curve - Train Dataset

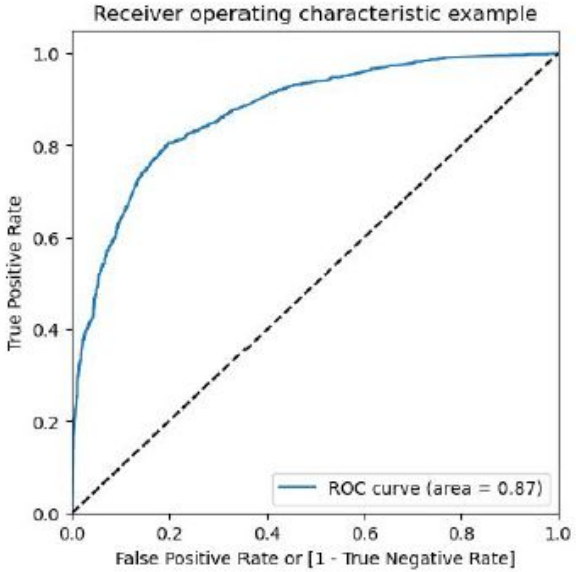
- The curve closely aligns with the upper-left corner of the plot, indicating a model with a consistently high true positive rate and a consistently low false positive rate across all threshold value



Model Evaluation

ROC Curve - Test Dataset

- The curve closely approaches the upper-left corner of the plot, symbolizing a model with a strong true positive rate and a low false positive rate across all threshold values



Model Evaluation

Confusion Matrix

- By employing a cutoff value of 0.345, the model attained a sensitivity of 80.05% in the training set and 79.82% in the testing set.
- In this context, sensitivity reflects the model's ability to accurately identify leads that convert out of all potential leads.
- The CEO of X Education had established a target sensitivity of approximately 80%.
- Furthermore, the model achieved an accuracy of 80.46%, aligning with the study's objectives.

Recommendations

- The success of X Education relies on increasing lead conversion, as per the problem statement. To achieve this, we've developed a regression model to find out which factors impact lead conversion the most.
- Our analysis has shown which features have the most positive impact. These are the ones we should focus on in our marketing and sales efforts to boost lead conversion.
    - Lead Source_Welingak Website: 5.39
    - Lead Source_Reference: 2.93
    - Current_occupation_Working Professional: 2.67
    - Last Activity_SMS Sent: 2.05
    - Last Activity_Others: 1.25
    - Total Time Spent on Website: 1.05
    - Last Activity_Email Opened: 0.94
    - Lead Source_Olark Chat: 0.91
- Negative coefficients :
    - Specialization in Hospitality Management: -1.09
    - Specialization in Others: -1.20
    - Lead Origin of Landing Page Submission: -1.26

Recommendations

- Create tactics to draw in high-quality leads from the best-performing lead sources.
- Concentrate on features with positive coefficients to create precise marketing strategies.
- Tailor messages to engage working professionals effectively.
- Enhance communication channels according to their influence on lead engagement
- Offer incentives or discounts for successful referrals to encourage more references.
- Consider increasing the budget for Welingak Website advertising and other promotional activities.
- Implement aggressive targeting of working professionals, as they exhibit a high conversion rate and typically have the financial capacity to afford higher fees.

Thank You