Lead Score Case Study

- Sagar Sonone

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

- Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

- EDA

1. Univariate data analysis
2. Bivariate data analysis

- Data preparation : logistic regression used for the model making and prediction.
- Validation of the model
- Model presentation.
- Conclusions and recommendations.

Data cleaning and data manipulation.

- Checking for the number of unique values in the categorical columns
- Dropping the columns with highly skewed data
  - Dropping ,(I agree to pay the amount through cheque, Get updates on DM Content, Update me on Supply Chain Content,Receive More Updates About Our Courses,Magazine) all the same data
  - Dropping(Do Not Call,What matters most to you in choosing a course,Search,Newspaper Article,X Education Forums,Newspaper,Digital Advertisement,Through Recommendations) these columns with highly skewed data
  - Dropping the tags because it is unnecessary
- Finding the percentage of null values present in the columns
- Dropping the columns with null values which is higher than 40 percent of the total values
- Standardizing the values
- Mapping Binary categorical variables

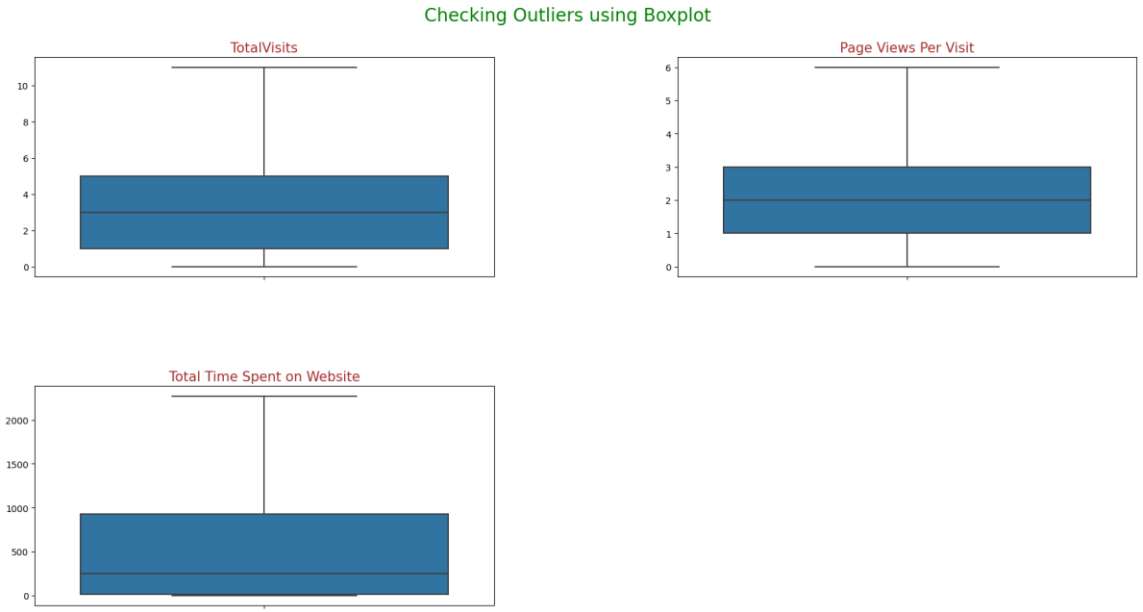
Data cleaning and data manipulation.

Spotting for the outliers

There are outliers present in the Total visits and Page Views Per Visit columns so we have to treat them we can cap the outliers

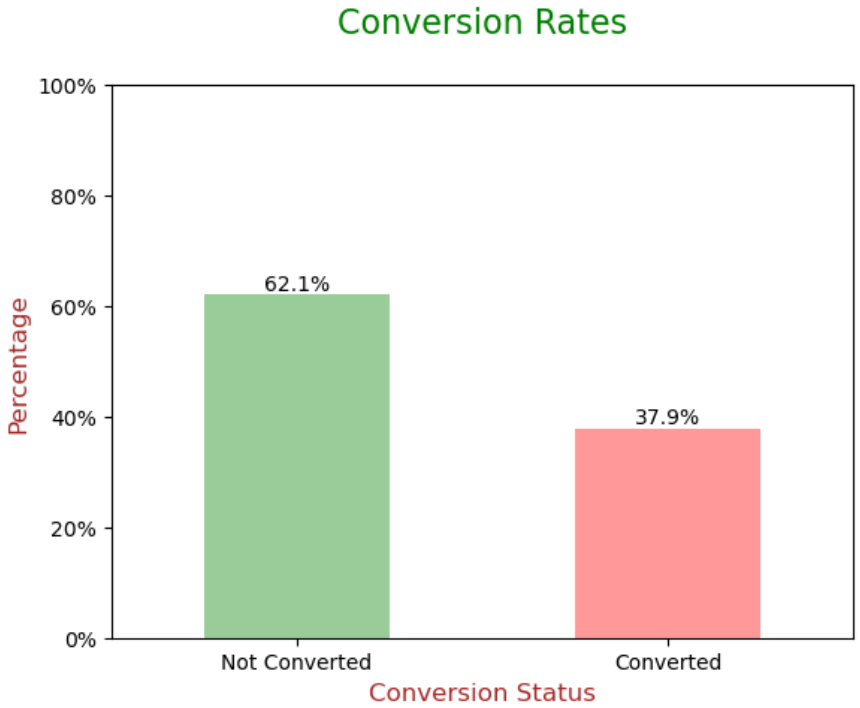
Data cleaning and data manipulation.

Capping and flooring the outliers



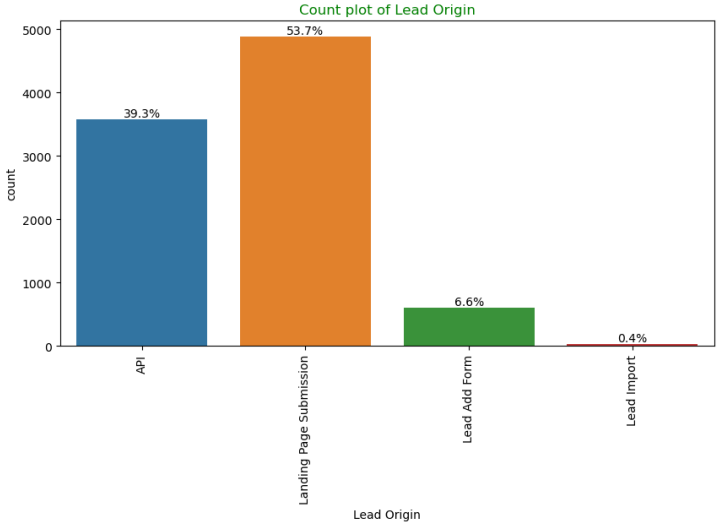
All the data has been cleaned and no further cleaning is necessary

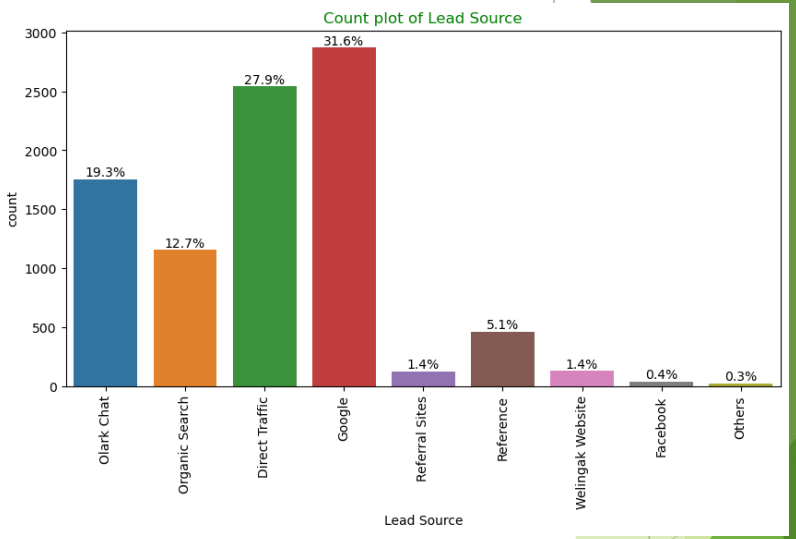
Exploratory Data Analysis (EDA)



Insights there are only 37.9% of people that are successfully converted into customers and there are 62.1 % of people that are not converted into the customers

EDA : Univariate data analysis

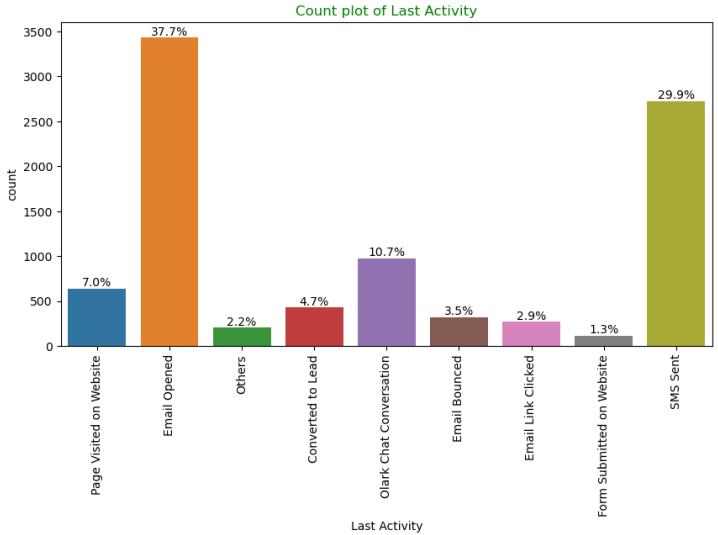


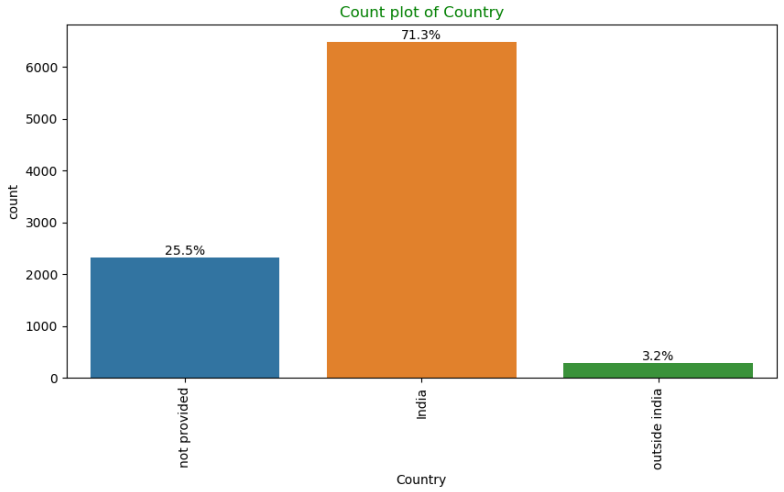


Insights

- Most of our leads are from olark Chat, organic Search and google
- The lead origin plot shows the lead import have lowest no. of individual and landing page submissions are the highest

EDA : Univariate data analysis

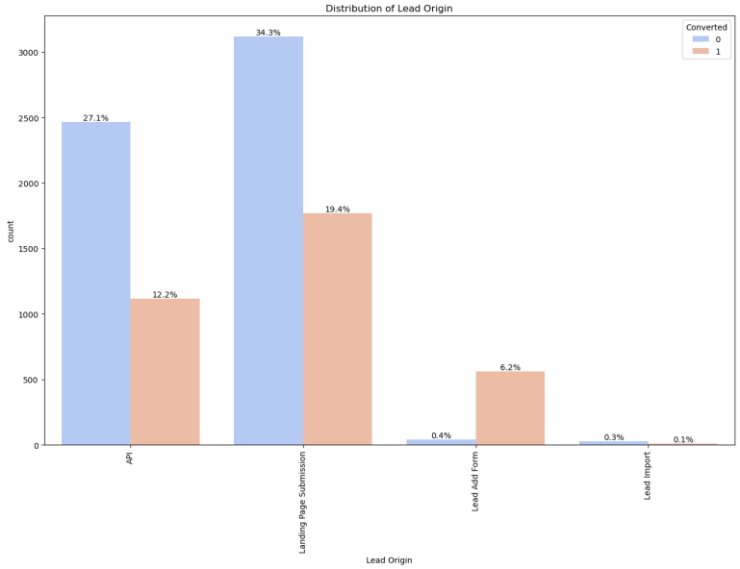


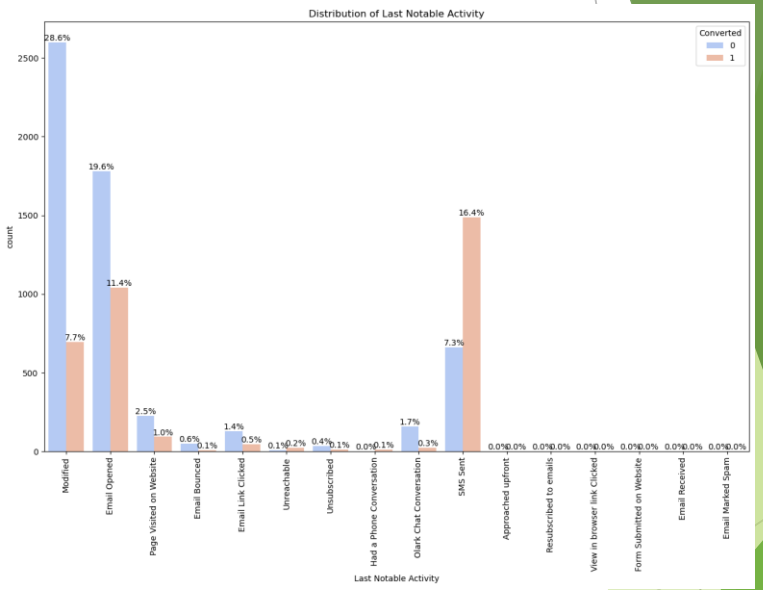


Insights

- Most of our visitors are from india
- Most of the activity is email opened and sms sent and converted to leads are very less

EDA : Bivariate data analysis





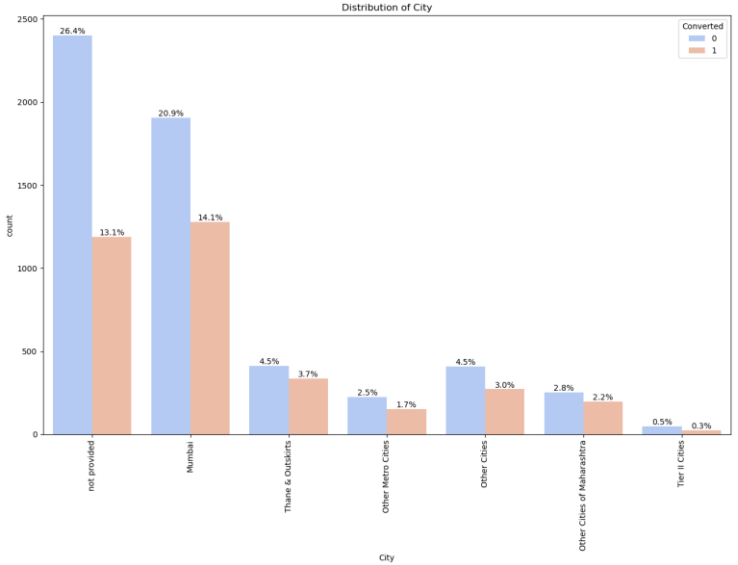
Lead Origin Distribution:

API Leads: 34.3% converted, 27.1% not converted.
Landing Page Submission: 19.4% converted, 12.2% not converted.
Lead Add Form: 62.7% converted, 0.4% not converted.
Lead Import: 0.1% converted, 0.3% not converted.

Last Notable Activity:

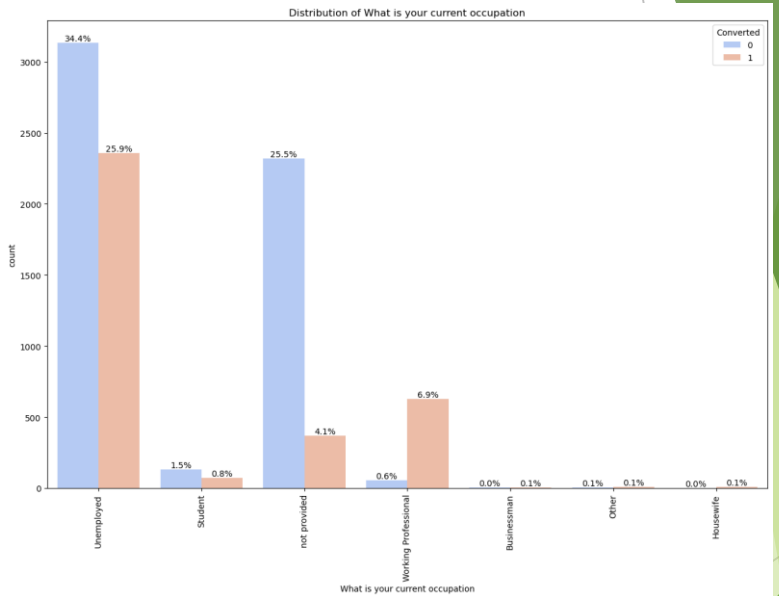
Modified: Highest count, both converted and not converted leads.
Email Opened: Significant count, relatively lower conversion rate.
SMS Sent: Fewer leads, notably higher conversion rate.
Other Activities: Fewer leads, generally low conversion rates.

EDA : Bivariate data analysis



City Distribution:

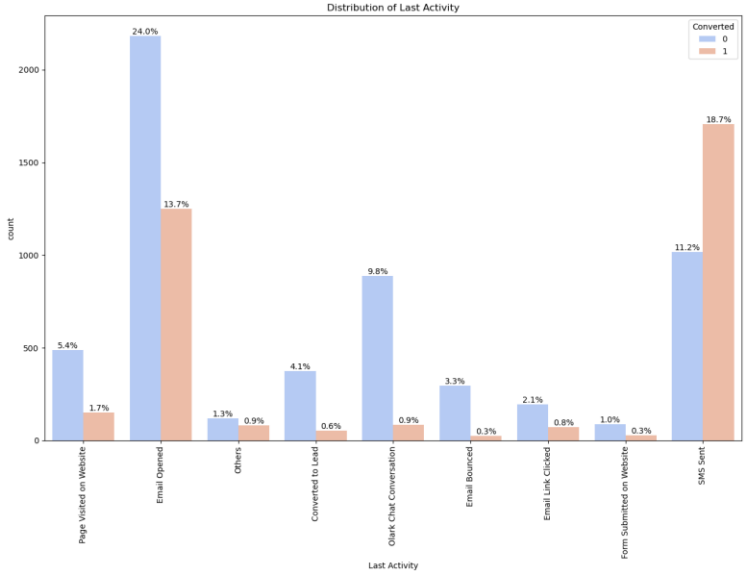
Majority from "not specified," followed by Mumbai. Conversion rates consistent across cities, highest in "Other Cities" category.

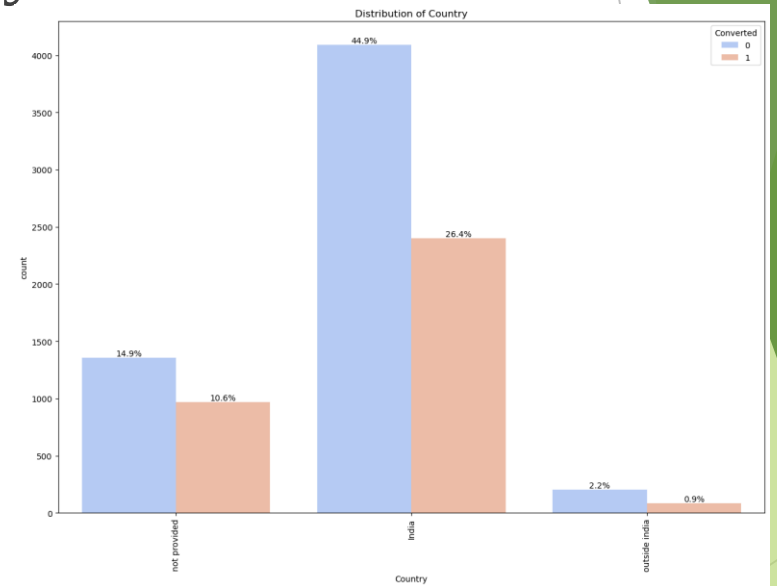


Occupation Impact:

Most leads are unemployed, followed by working professionals. Working professionals have high conversion rates, students low.

EDA : Bivariate data analysis





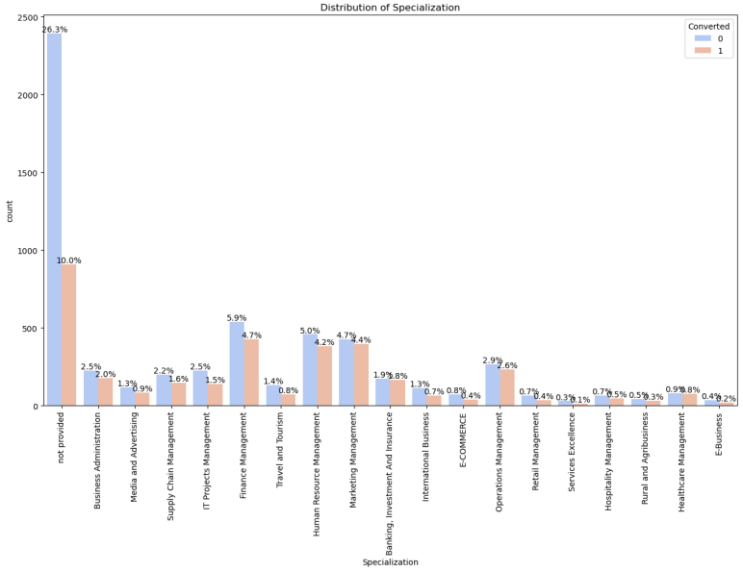
Last Activity Insights:

"Page Visited on Website" most common, "SMS Sent" has high conversion.
Understanding last activity context crucial for lead management.

Country Distribution:

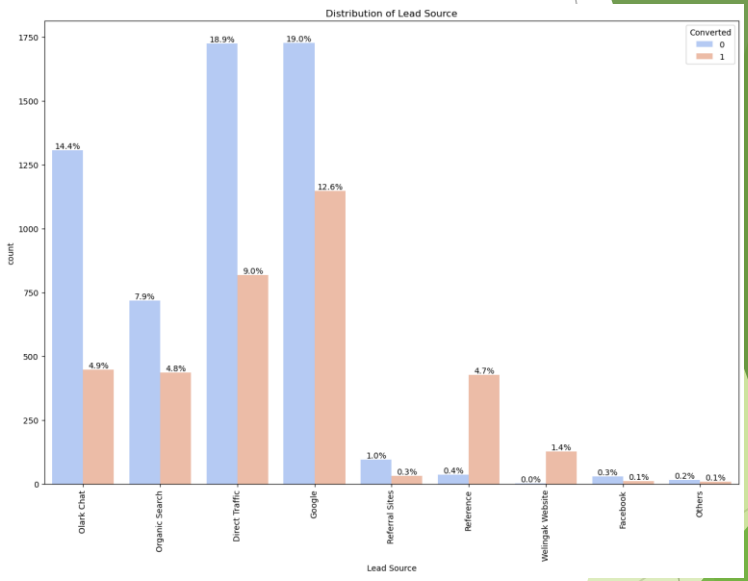
Majority from "non-specified" country.
Conversion rates consistent across countries, highest in "non-specified" category.

EDA : Bivariate data analysis



Specialization Impact:

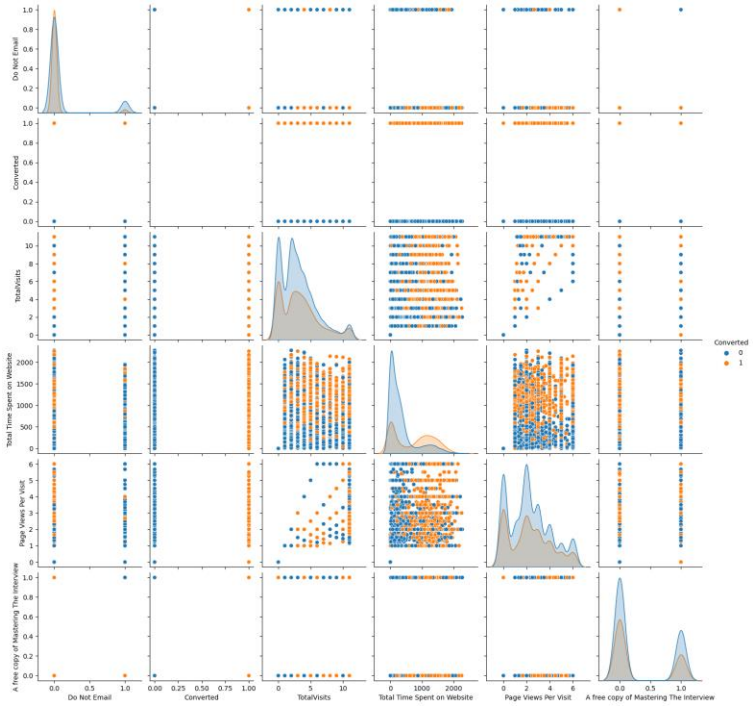
"Not Specified" specialization dominant but low conversion. Some specializations show higher conversion despite fewer leads.

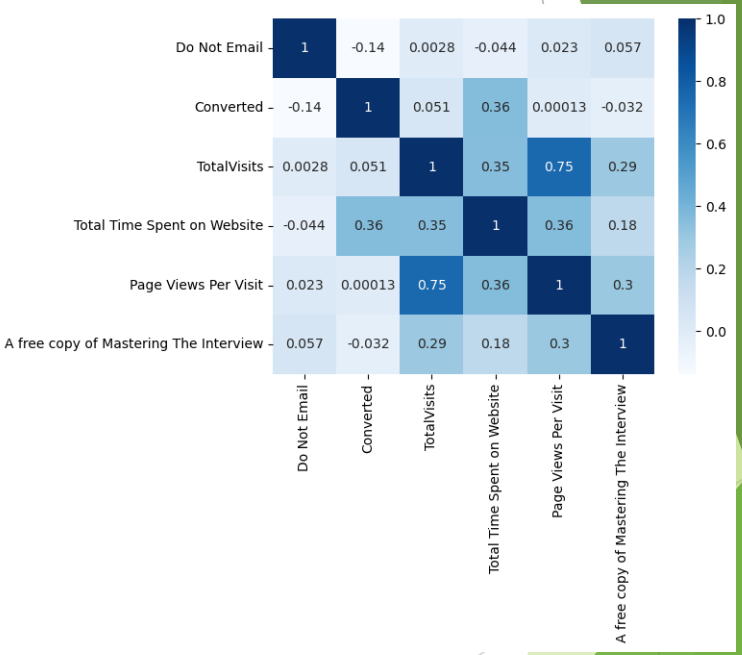


Distribution of Lead Source :

3 topmost lead sources with highest conversion is from google and direct traffic

EDA : Bivariate Analysis for Numerical Variables





Converted is highly corelated to the total no. of visits suggesting the direct connection

Data preparation : logistic regression used for the model making and prediction.

Data preparation:

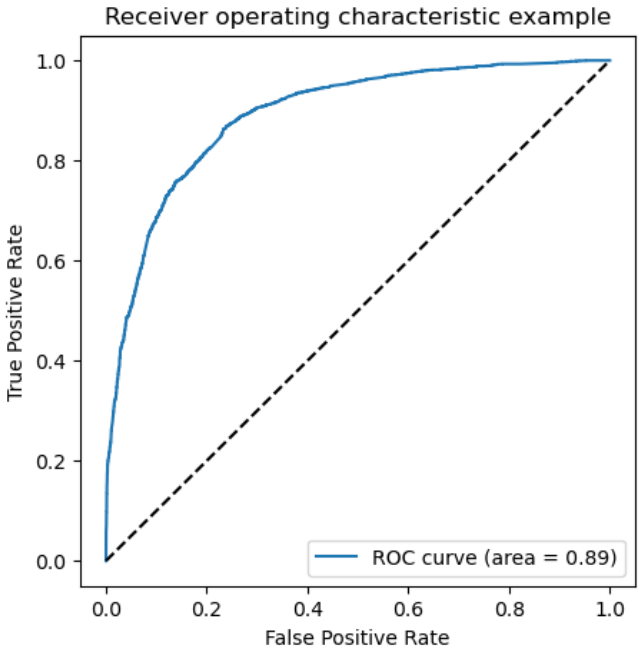
1. Creating the dummy variable
2. Splitting the data : Train-Test Split: The dataset was divided into 70% training and 30% testing subsets.
3. Scaling of the data

Model Building:

RFE identified the top 15 relevant variables, and the remaining variables were manually removed based on VIF and p-value criteria (VIF < 5 and p-value < 0.05 were retained).

1. We will Build Logistic Regression Model for predicting categorical variable : 6 models are created
2. Feature Selection Using RFE (Coarse tuning)
3. Manual fine-tuning using p-values and VIFs

Validation of the model

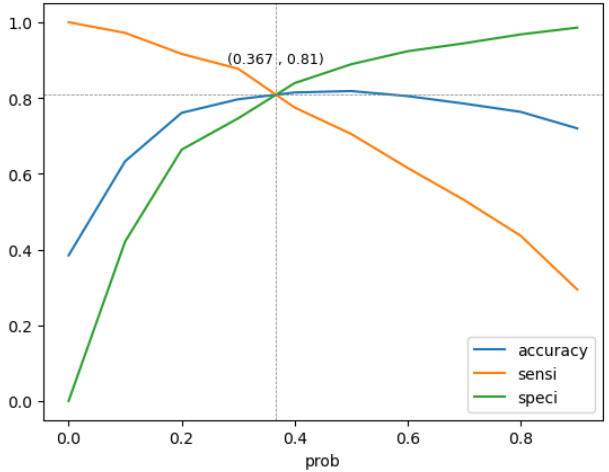


Area under ROC curve is 0.89 out of 1 which indicates a good predictive model

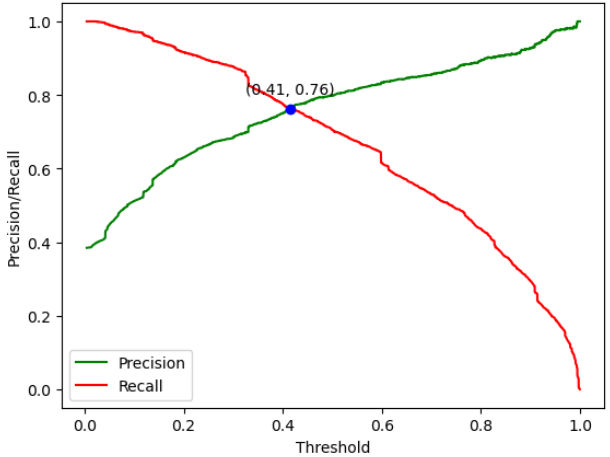
A confusion matrix was constructed, and the optimal cutoff value (determined via ROC curve) resulted in approximately 80% accuracy, sensitivity, and specificity.

- Confusion Matrix, Accuracy, Sensitivity and Specificity, Threshold determination using ROC & Finding Optimal cutoff point, Precision and Recall

Model presentation



From the graph it is visible that the optimal cut off is at 0.367



The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimize the performance of the model based on business requirement, Here our probability threshold is 0.42 from above curve.

- Prediction: 80% for the metrics we are getting with the sensitivity-specificity cut-off threshold of 0.359. So, we will go with sensitivity-specificity view for our Optimal cut-off for final predictions.
- Precision-Recall: As we can see in above metrics when we used precision-recall threshold cut-off of 0.42 the values in True Positive Rate, Sensitivity, Recall have dropped to around 76%, but we need it close to 80% as the Business Objective.

Conclusions and recommendations

Key Findings: Analysis revealed the most influential factors for potential buyers:
Top 3 features that contributing positively to predicting hot leads in the model are:

- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Lead Origin_Lead Add Form

Top 3 features that contributing negatively to predicting hot leads in the model are:

- Lead Origin_Landing Page Submission
- Specialization_not provided
- What is your current occupation_not provided

Strategies for Increasing Lead Conversion Rates:

- Feature Focus , Quality Leads
- Tailored Messaging, Optimize Channels
- Budget Allocation, Incentives for Referrals
- Target Working Professionals, Enhance Website Appeal