Lead Score Case Study

Batch - DS 70

By Aparna Sarkar,

R M Aparna

& Anushka Gandhi

Highlights

- ✓ Problem Statement
- ✓ Business Objective
- ✓ Solution Approach
- ✓ Data Preparation
- ✓ EDA
- ✓ Data Preparation for model building
- ✓ Model Building and Evaluation
- ✓ ROC curve
- ✓ Precision Recall Trade off
- ✓ Final Results
- ✓ Conclusion
- ✓ Business Insights

Problem Statement

X Education aims to boost its lead conversion rate from 30% to an ambitious 80%. The objective is to construct a model capable of assigning lead scores between 0 and 100, empowering the sales team to prioritize interactions with high-potential leads.

Business Objective

The primary business objective for X Education is to improve their lead conversion rate by identifying and prioritizing the most promising leads, referred to as "Hot Leads." This involves:

1.Assigning a Lead Score:
Developing a predictive model to assign a score to each lead based on their likelihood of converting into paying customers.

2.Optimizing Sales Efforts:
Using the lead scores to enable the sales team to focus their time and resources on high-potential leads rather than contacting all leads indiscriminately.

3.Achieving the Target Conversion Rate:
The CEO has set a target lead conversion rate of approximately **80%**, which the model and associated strategies are designed to achieve.

Solution Approach

1.Data Cleaning and Manipulation:

- •Identify and handle duplicate records.
- •Address ‘NA’ and missing values appropriately.
- •Remove columns with significant amount of missing data that are not relevant for analysis.
- •Perform null value imputation when necessary.
- •Detect and handle outliers in the dataset.

2.Exploratory Data Analysis (EDA):

- **•Univariate Analysis:** Analyse individual variables (e.g., value counts, distributions).
- **•Bivariate Analysis:** Examine relationships between variables, including correlation coefficients and patterns.

3.Feature Engineering:

- •Scale features and create dummy variables or encode categorical data for compatibility with the model.

4.Classification Technique:

- •Use logistic regression to build the model and make predictions.

5.Model Evaluation:

- •Evaluate the model's performance using appropriate metrics and techniques.

6.Model Presentation:

- •Present the model results and key findings effectively.

7.Conclusions and Recommendations:

- •Summarize insights and provide actionable recommendations based on the analysis.

Data Manipulation

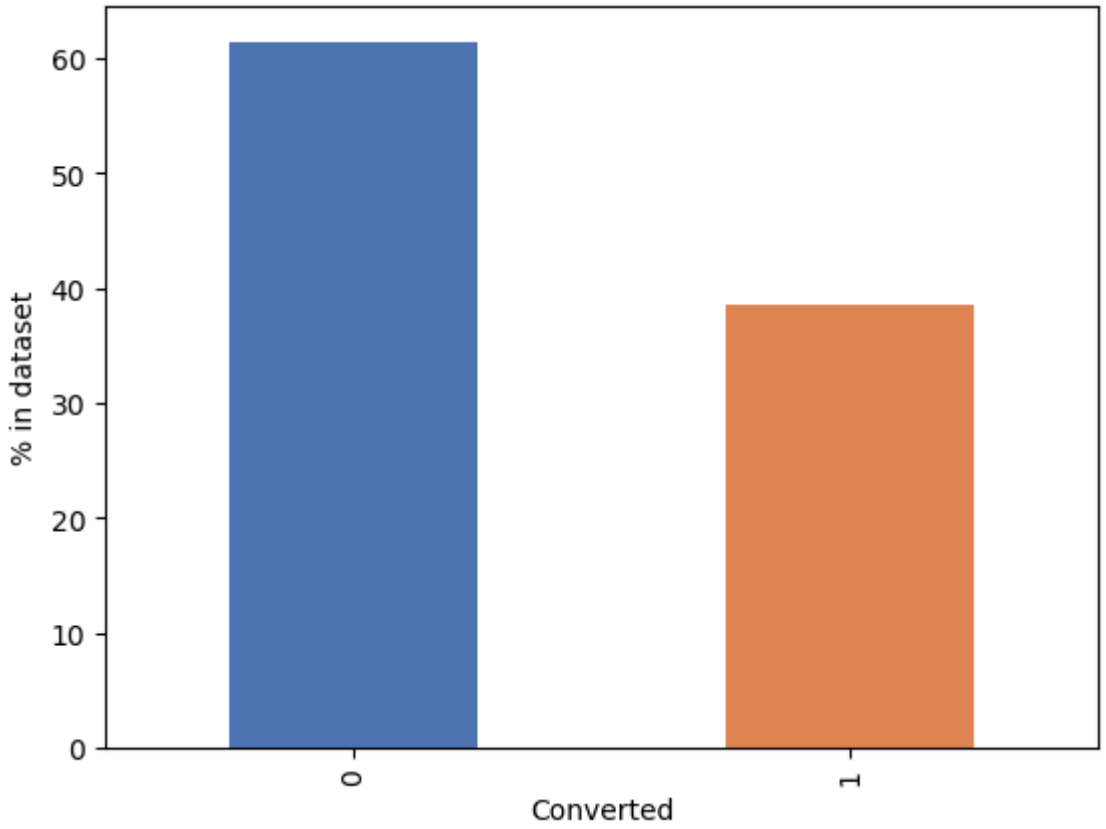
- The **Lead Number** and **Prospect Id** columns, deemed irrelevant for analysis, were removed.

- Columns with more than 35% missing values were dropped.

- Columns with only a single unique value were excluded.

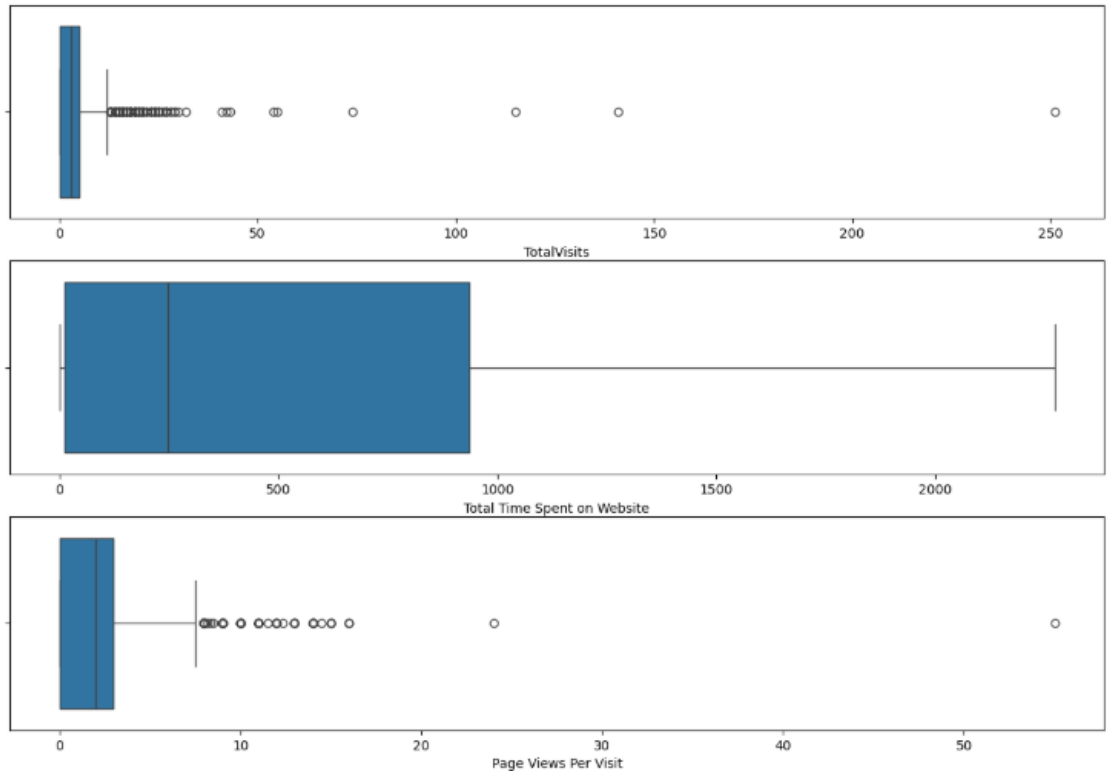
- Missing values in columns like **Lead Source**, **Total Visits**, **Page Views Per Visit**, **Last Activity**, **Country**, **What is your current occupation**, and **What matters most to you in choosing a course** were filled with appropriate imputed values.

EDA



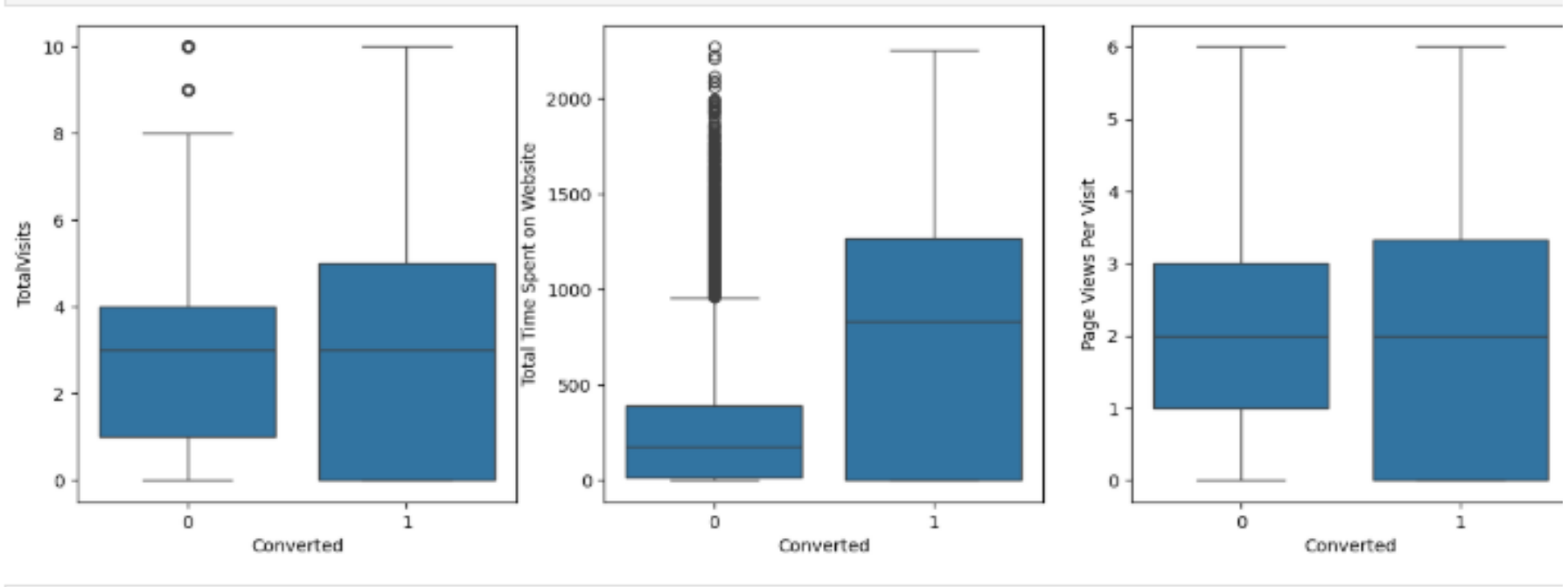
The percentage of non-converted leads is higher than the percentage of converted leads .
This indicates an imbalance in the dataset, with a greater number of leads not converting.

Univariate Analysis Of Numerical variables



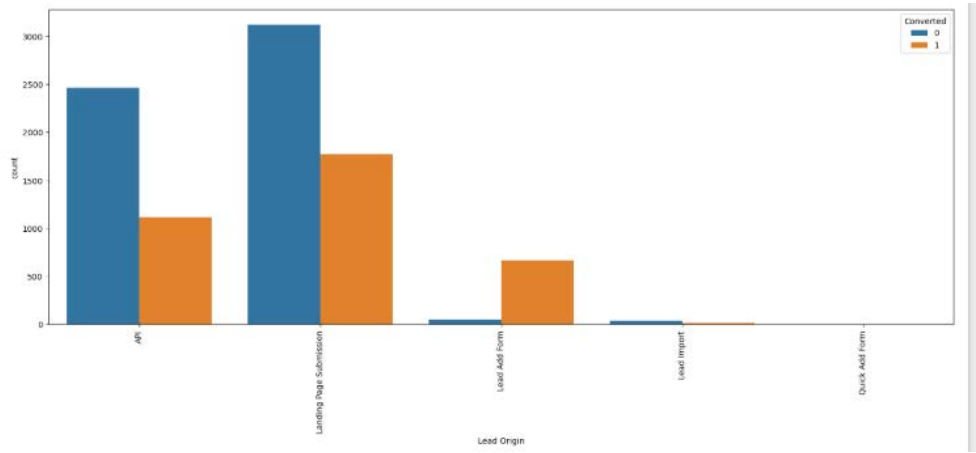
Total Visits and **Page Views Per Visit** both show extreme outliers, indicating a few users had significantly higher interaction levels compared to the majority of the dataset. These users may represent unique cases that require outlier handling.

Bivariate Analysis Of Numerical variables

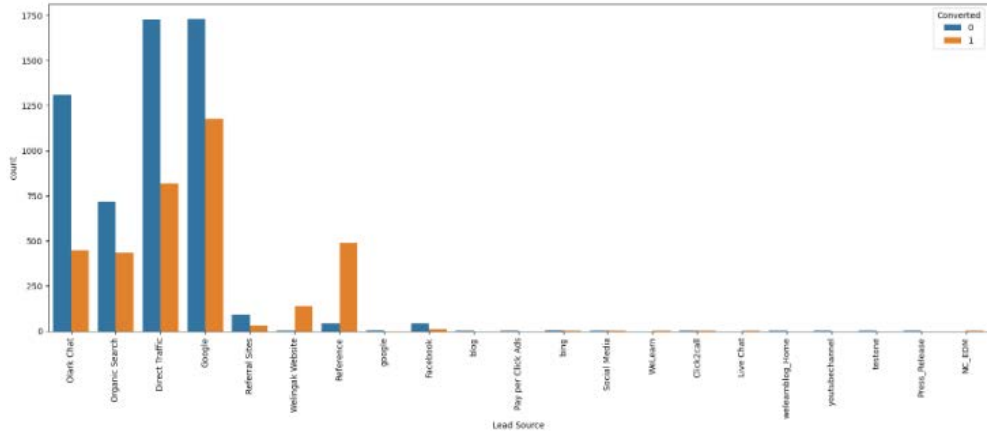


People spending more time on websites are most likely to be converted.

Bivariate Analysis Of Categorical variables



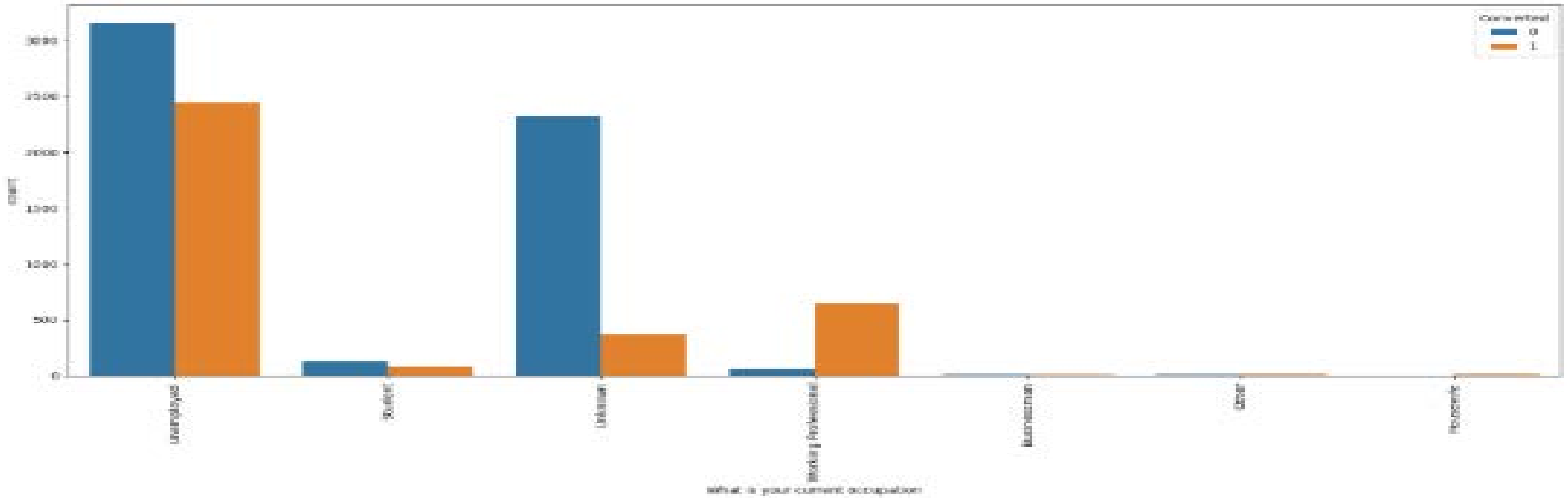
The graph shows that **Landing Page Submission** has the highest number of converted leads, indicating it is the most effective lead origin for driving conversions.



Very high conversion rates for lead sources ‘Reference’ & ‘Welingak website’

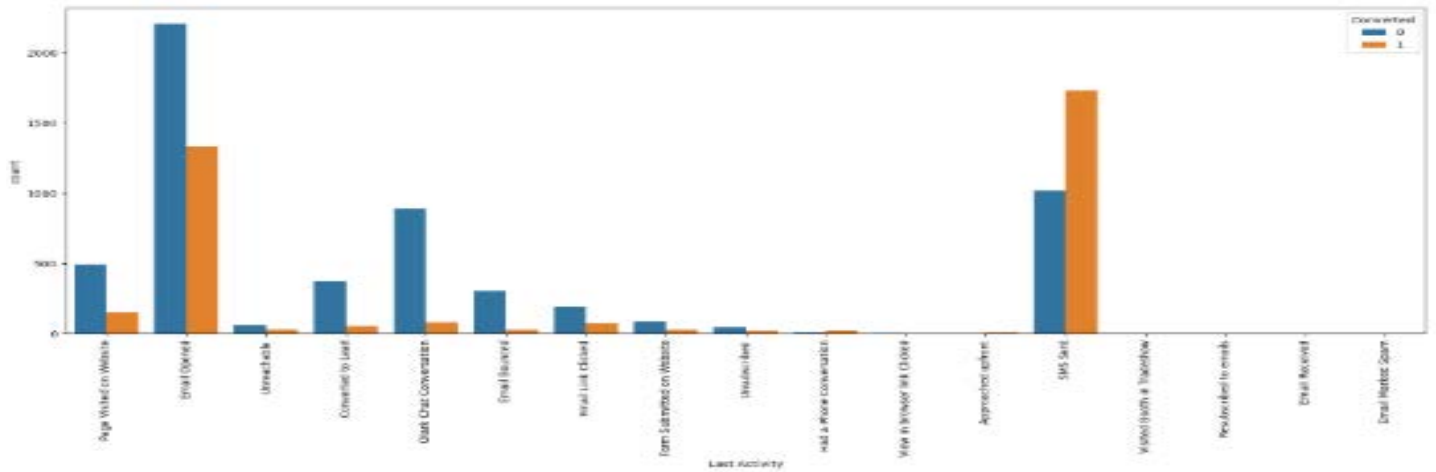
Most leads are generated through ‘Direct Traffic’ & ‘Google’

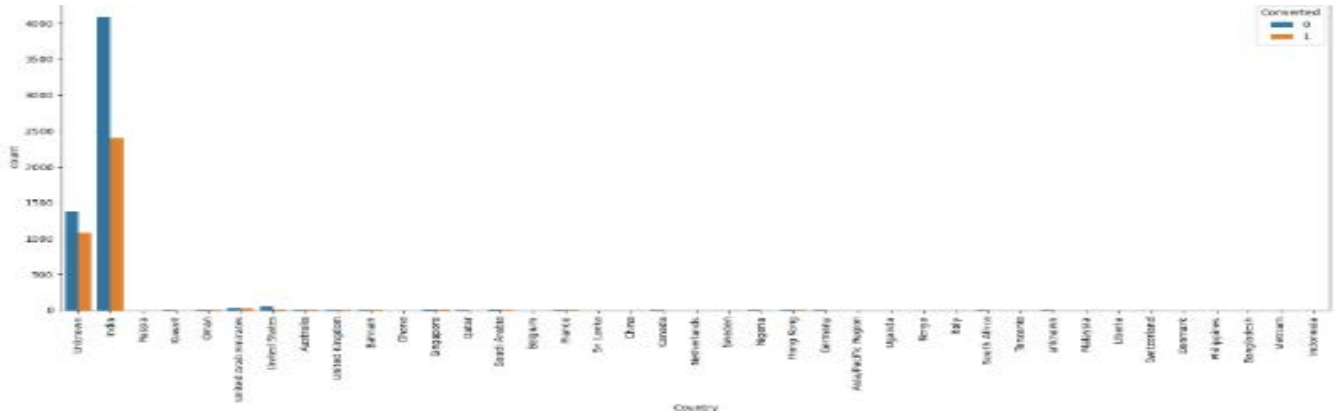
Bivariate Analysis Of Categorical variables



Working professionals are most likely to get converted

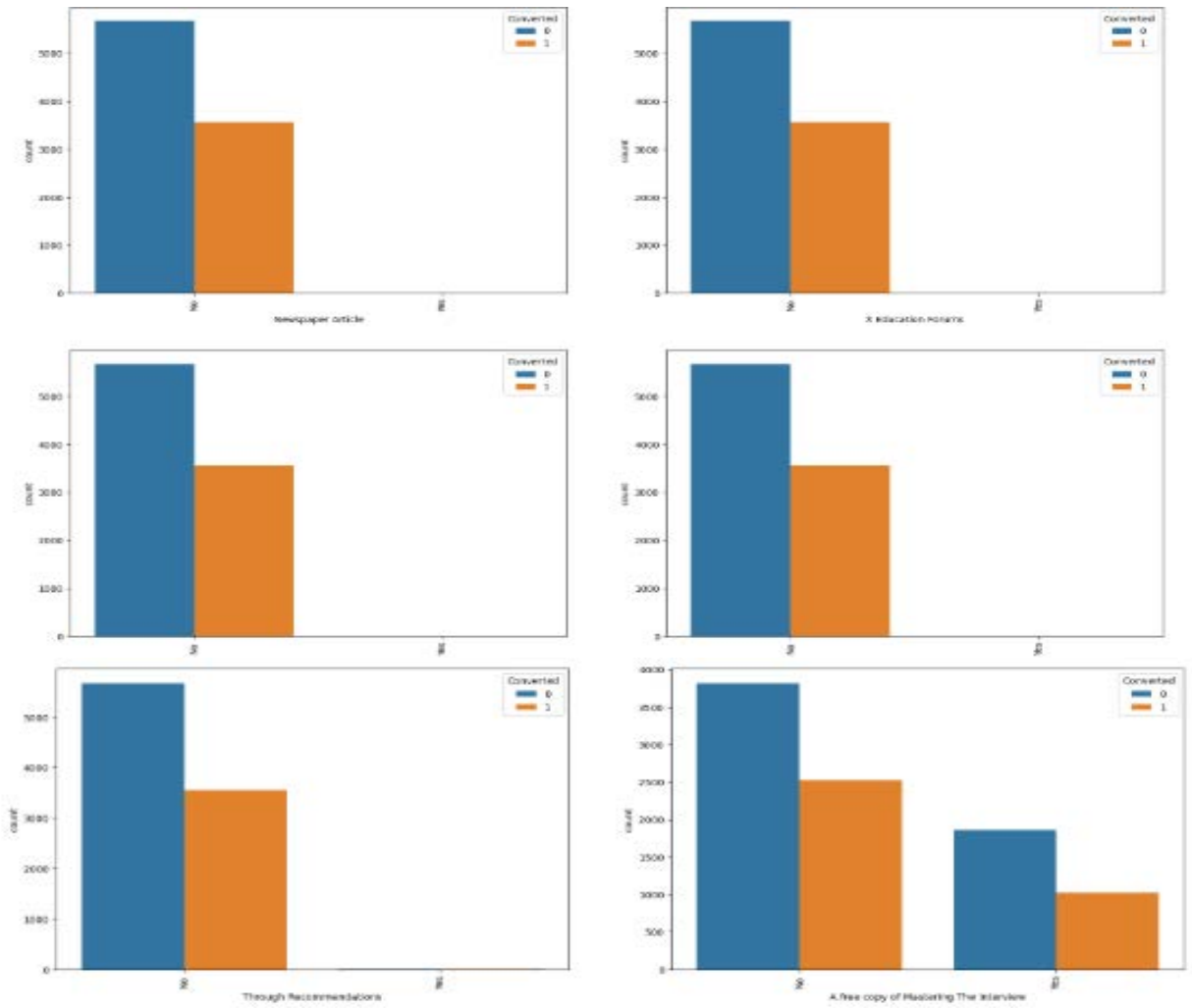
Analysis Of Categorical variables





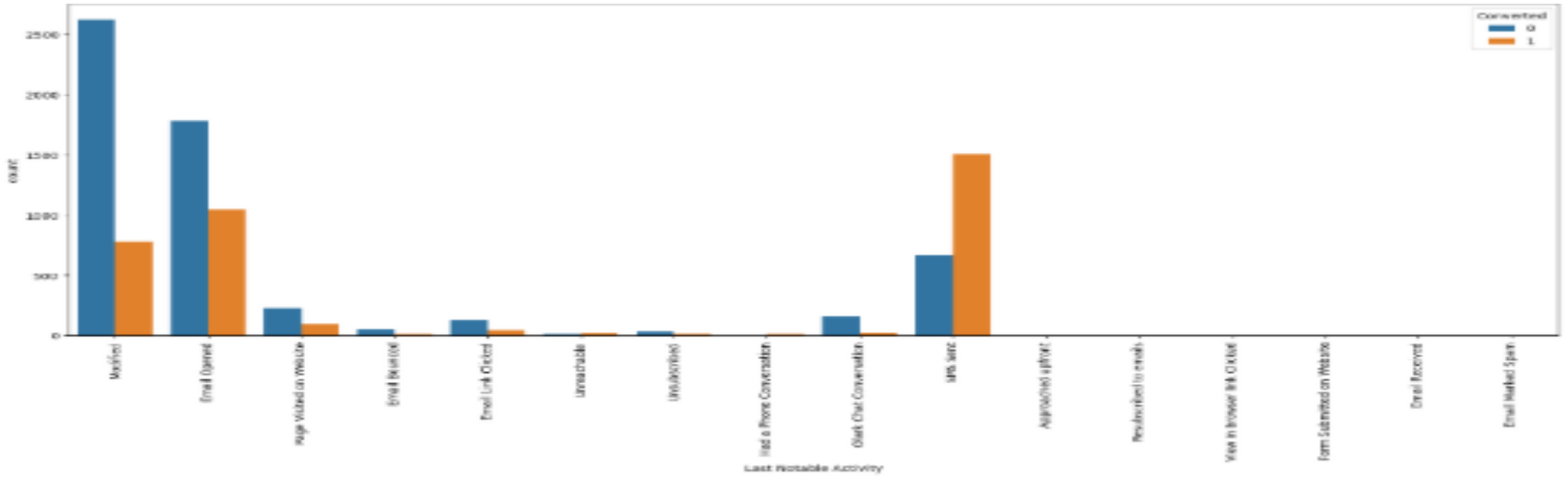
Highest conversion rate is for Last Activity ‘SMS sent’ and country ‘India’

Analysis Of Categorical variables



As all the above variables have most of the values as no, nothing significant can be inferred from these plots.

Analysis Of Categorical variables



In 'Last Notable Activity', we can combine categories after 'SMS Sent' similar to the variable 'Last Activity'.
It has most generated leads for the category 'Modified' while most conversion rate for 'SMS Sent' activity.

Data Preparation for model building

- Creation of Dummy Variables:
  - Dummy variables established for independent variables for ease of interpretation and odds ratio calculation.
  - Binary variables encoded with 1 for 'Yes' and 0for 'No.'

- Train-Test Split:
  - 'Leads' dataset split into Train (70%) and Test (30%) sets.
  - Train set utilized for model training, Test set for model evaluation.

- Feature Scaling:
  - Ensured uniform scale for all variables to prevent dominance by high-magnitude features.
  - Implemented MinMaxScaler for feature scaling

Model Building and Evaluation

1.Model Selection:

● Utilized the Generalized Linear Model (GLM) from the StatsModels library to build the Logistic Regression Model.

2.Initial Model Features:

● Initially built the model from the X_train dataset.

3.Identifying Insignificant Features:

● Found a majority of features to be insignificant, prompting the need for feature selection.

4.Feature Selection Technique: Recursive Feature Elimination (RFE)

● Applied RFE, an optimization technique, to identify the best subset of features.

● RFE repeatedly constructs models, selecting the best-performing features based on coefficients.

● Top 15 features were identified for further model building.

5.Feature Elimination Criteria:

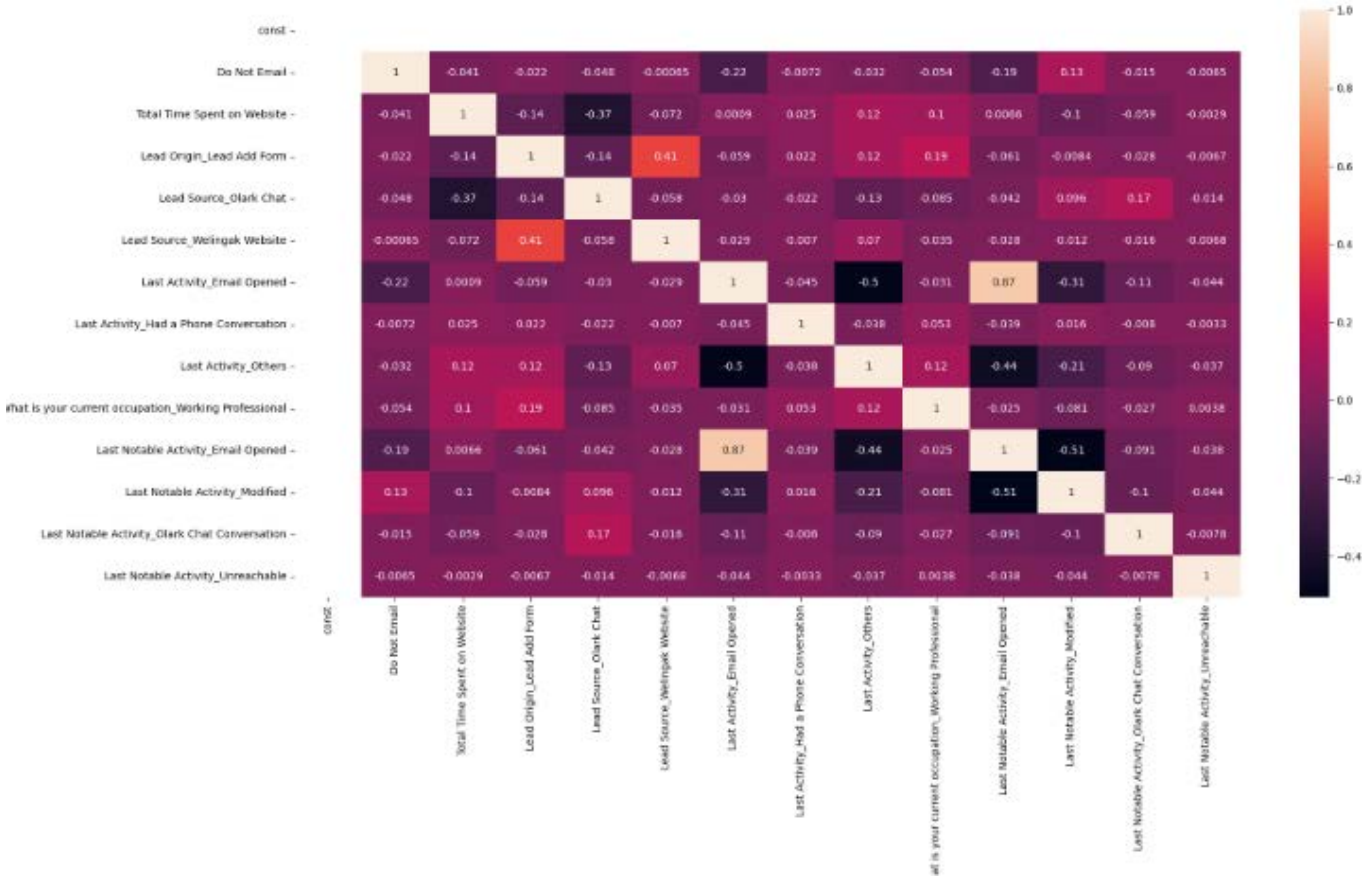
● Insignificant features were systematically dropped based on P-value and Variance Inflation Factor (VIF).

● Accepted P-value set below 0.05, and VIF kept less than 5 for feature retention.

6.Iterative Elimination Process:

● Features were eliminated one by one, considering their statistical significance and multicollinearity.

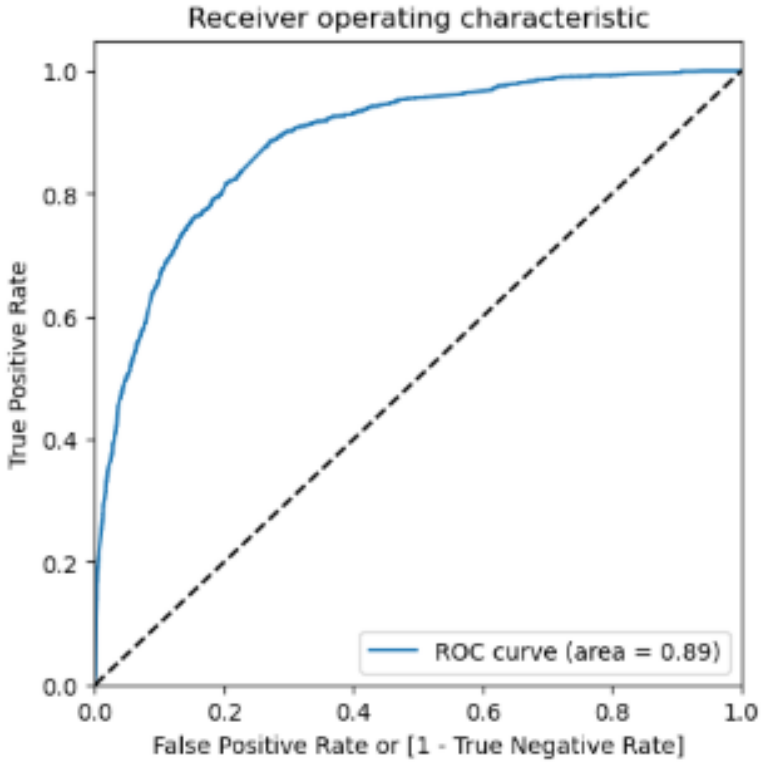
Correlation among Features



	Features	VIF
9	What is your current occupation_Unemployed	2.51
1	TotalVisits	2.40
2	Total Time Spent on Website	2.06
3	Lead Origin_Lead Add Form	1.56
4	Lead Source_Olark Chat	1.54
7	Last Activity_Others	1.51
6	Last Activity_Olark Chat Conversation	1.39
10	What is your current occupation_Working Profes...	1.36
5	Lead Source_Welingak Website	1.24
0	Do Not Email	1.06
8	What is your current occupation_Student	1.05
12	Last Notable Activity_Unreachable	1.01
11	Last Notable Activity_Had a Phone Conversation	1.00

From the above heatmap and VIF score it is evident that there is no significant multicollinearity among the selected features

ROC Curve & Optimal Cutoff

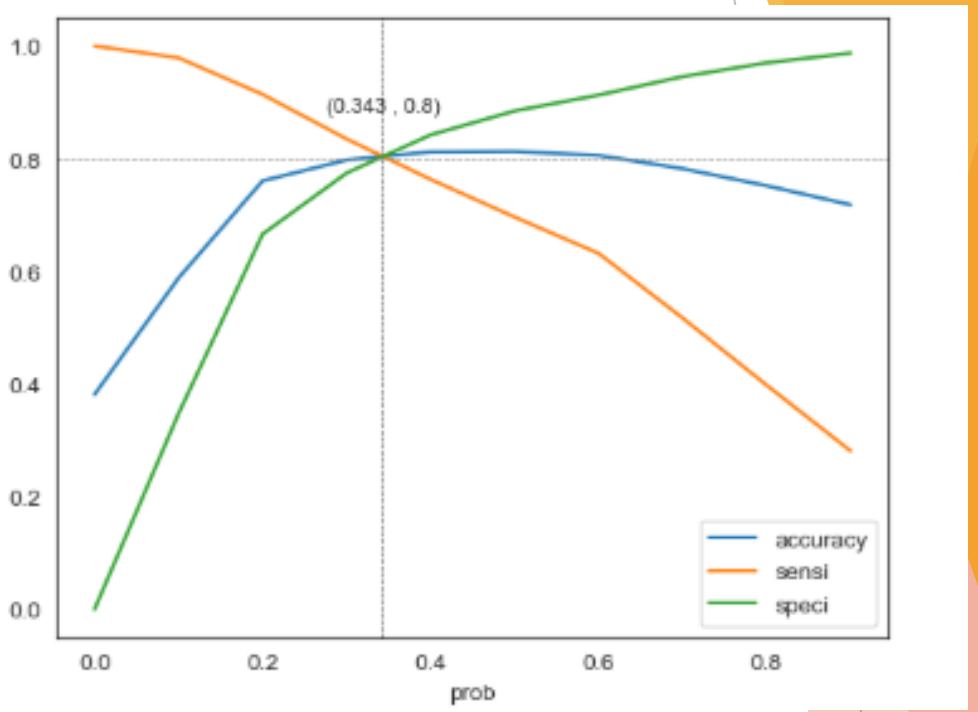


Receiver Operating Characteristics (ROC) Curve:

- **AUC Assessment:** The Area Under the Curve (AUC) of the ROC curve is a pivotal metric for evaluating the model's performance.
- **Goodness of the Model:** The ROC curve's proximity to the upper-left section of the graph signifies the effectiveness of the model.
- **Model Evaluation:** With an AUC value of 0.89, our model demonstrates a high level of discriminative ability and accuracy in distinguishing between positive and negative instances.

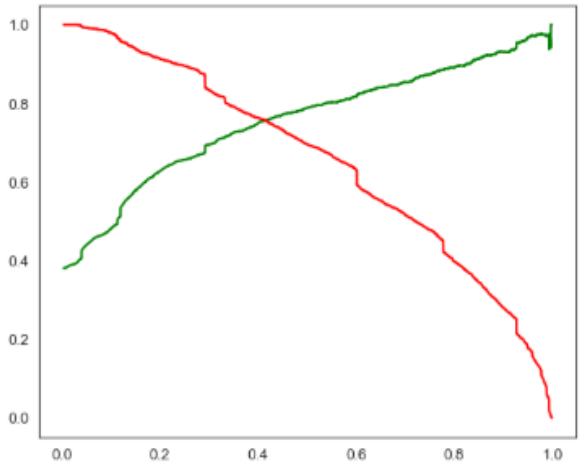
Evaluation metrics

Feature Selection:
● The final model includes 13 key features that satisfy selection criteria, enhancing model efficiency and interpretability.
Prediction Threshold:
● Lead scores with a conversion probability greater than 0.34 are classified as "Converted.“
Test Dataset Prediction:
● Utilized the 0.34 probability threshold to predict conversions for leads in the test dataset.
Confusion Matrix (cut-off 0.43):
● [[3229 , 773]
● [499 , 1967]]
Evaluation Metrics:
● Accuracy: 0.803
● Sensitivity (Recall): 0.7976
● Specificity: 0.8068
Lead Score:
- Lead scores have been computed using conversion probability.



Precision Recall Trade-off

Feature Selection:
● The final model includes 13 key features that satisfy selection criteria, enhancing model efficiency and interpretability.
Prediction Threshold:
● Lead scores with a conversion probability greater than 0.40 are classified as "Converted." using
Test Dataset Prediction:
● Utilized the 0.40 probability threshold to predict conversions for leads in the test dataset.
Confusion Matrix (cut-off 0.40):
● [[3370 , 632]
● [584 , 1882]]
Evaluation Metrics:
● Accuracy : 0.8119
● Sensitivity (Recall): 0.7631, Specificity- 0.8420



Final Results

Accuracy and other metrics yield better values for cutoff 0.34.
For predictions on the test set same cutoff is used

	Train Dataset	Test Dataset
Accuracy	0.8033	0.8109
Sensitivity	0.7976	0.7945
Specificity	0.8068	0.8217
False Positive rate	0.1931	0.1782
Positive Predictive Value	0.7178	0.7442
Negative Predictive Value	0.8661	0.8596
AUC	0.8866	0.8931

	Converted	Lead ID	Converted_prob	final_predicted	lead_score
0	1	4269	0.582367	1	58
1	1	2376	0.927611	1	93
2	1	7766	0.926117	1	93
3	0	9199	0.113907	0	11
4	1	4359	0.777847	1	78

Lead Score computed on test data.

Conclusions

1.Key Predictors of Conversion

1. **Total Time Spent on Website (4.4833):** Strongest positive predictor of lead conversion. Customers spending more time on the website are highly likely to convert.
2. **Lead Origin: Lead Add Form (3.7892):** Leads originating from the "Lead Add Form" show high conversion potential.
3. **Occupation: Working Professional (3.6360):** Working professionals are highly likely to convert compared to other occupations.
4. **Last Notable Activity: Phone Conversation (3.5757):** Leads who had a phone conversation with the team show significantly higher chances of conversion.

2.Negative Influencers of Conversion

1. **Do Not Email (-1.3659):** Leads opting out of emails have a significantly lower likelihood of conversion.
2. **Last Activity: Olark Chat Conversation (-1.1676):** Leads engaging via Olark Chat show a lower probability of conversion.

3.Secondary Predictors

1. **Total Visits (0.8349):** Higher website visits are moderately associated with lead conversion.
2. **Lead Source: Olark Chat (1.6522):** Leads from Olark Chat perform well, though less impactful than other factors.
3. **Unreachable Leads (2.0023):** Leads who were unreachable show a higher conversion rate once re-engaged.

4.Model Accuracy

1. The model's coefficients align with the business objective of identifying high-potential leads.
2. Insights derived from the model can help achieve the target lead conversion rate of 80%.

Business Insights

1.Focus Areas for the Sales Team

1. Prioritize leads who have spent significant time on the website and those from the "Lead Add Form" origin.
2. Target working professionals and students with tailored communication strategies.
3. Emphasize follow-ups with leads who had phone conversations or were initially unreachable.

2.Optimize Marketing Efforts

1. Leverage "Olark Chat" and "Welingak Website" as effective lead sources.
2. Address potential negative interactions with "Do Not Email" customers through alternate channels like phone or text.

3.Improve Engagement Strategies

1. Reduce reliance on Olark Chat for critical interactions, as its effectiveness appears limited in certain cases.
2. Enhance website content to increase "Total Time Spent on Website" and encourage deeper engagement.

4.Data-Driven Lead Prioritization

1. Use the lead scores generated by the model to segment and rank leads by priority.
2. Allocate resources effectively by focusing on "Hot Leads" with the highest likelihood of conversion.