Lead Scoring Case Study

Business Problem

01

Overview

- X Education sells online courses to industry professionals, attracting leads through website interactions and referrals.
- Current lead conversion rate is approximately 30%, highlighting a need for a more efficient process.

02

Mission

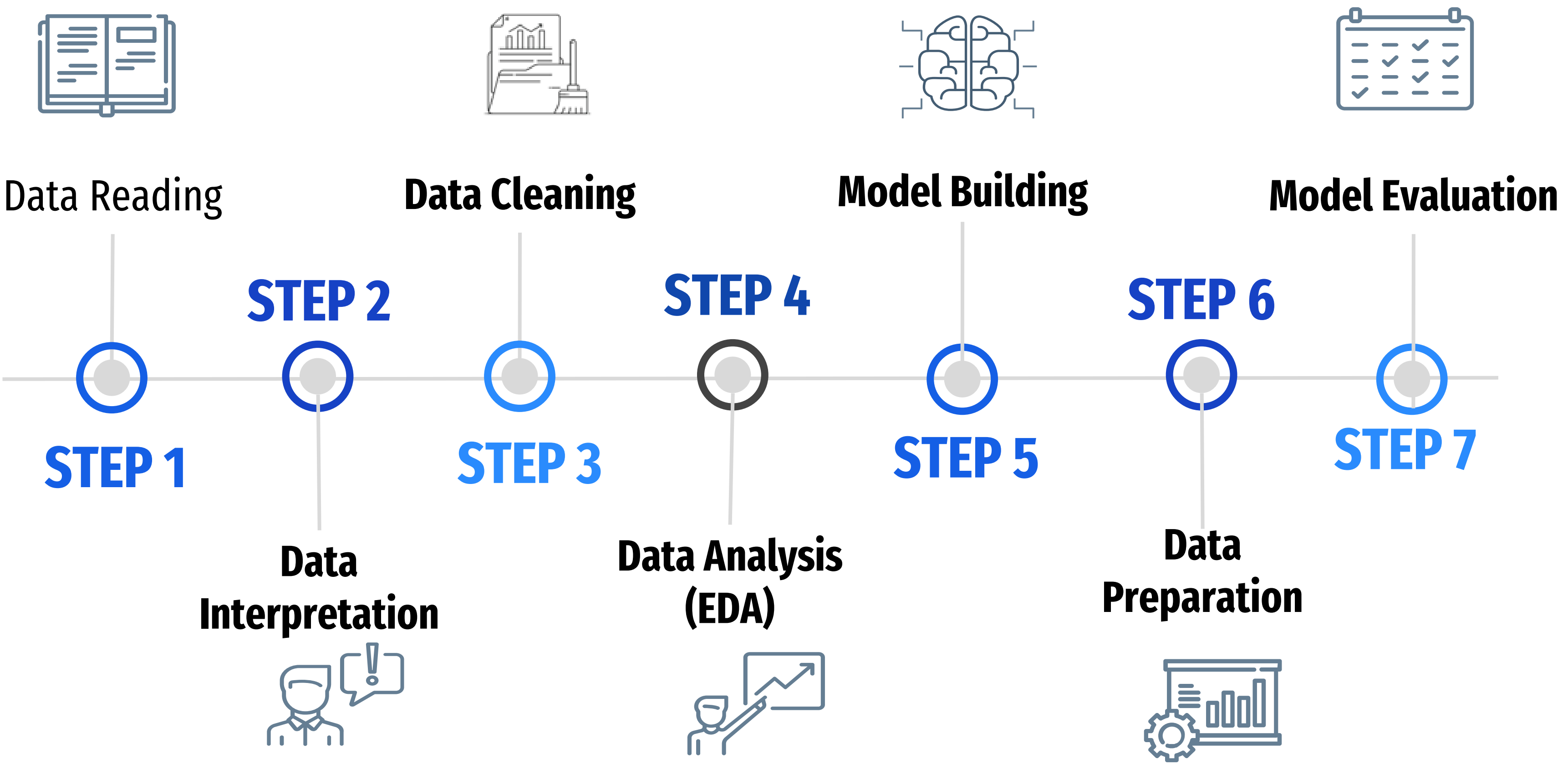
- Enhance X Education's lead conversion process by identifying and prioritizing 'Hot Leads,' aiming to achieve a target lead conversion rate of around 80%.

03

Core values

- Prioritize potential leads by assigning lead scores, ensuring the sales team focuses on high-conversion probability leads, thus optimizing the lead conversion process.

STEPS PERFORMED



Data CLeaning

Observation

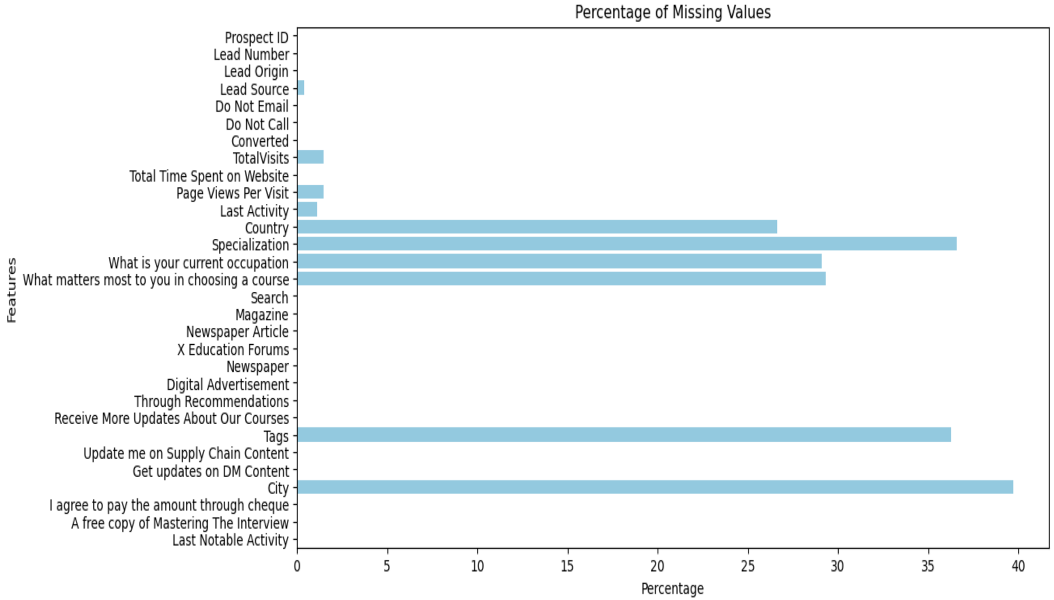
The graph reveals a substantial number of missing or null values in the dataset.

Objective

Ensure a cleaner dataset for robust analysis and modeling.

Approach

- Drop columns: Remove features with excessive missing values.
- Imputation: Fill missing values using appropriate techniques.

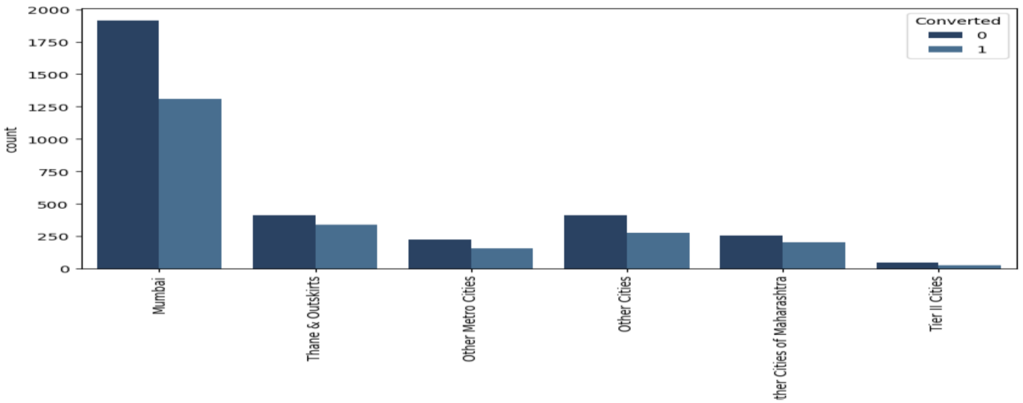


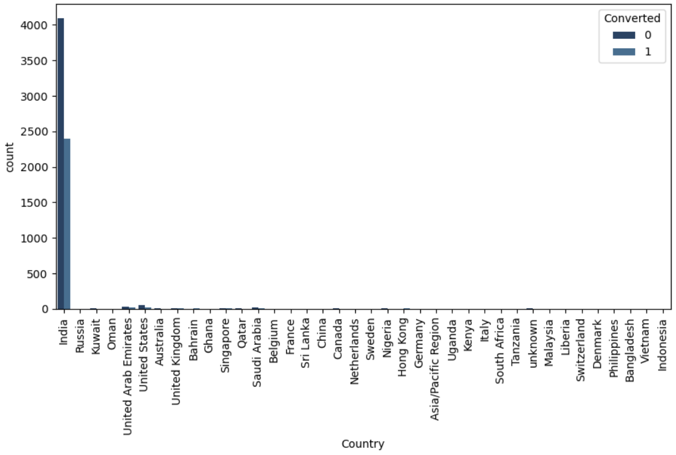
How did you hear about X Education	Lead Profile	Lead Quality
78.46%	74.19%	51.59%

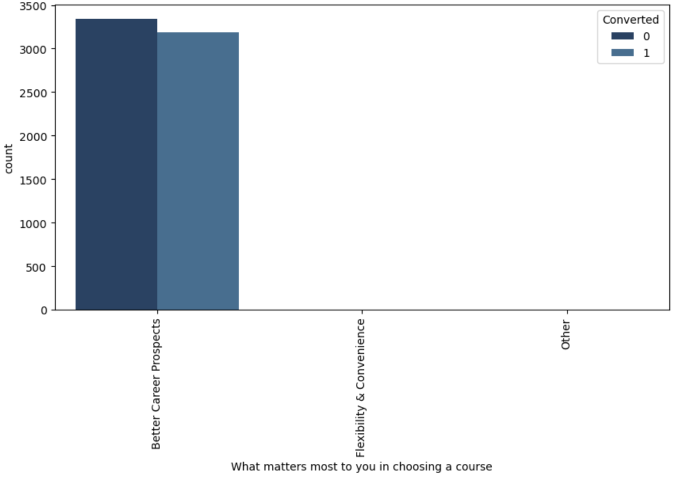
Data Cleaning

Dropping Columns

- The graphs displayed are for the features City, Country, and "What matters most to you in choosing a course.“
- These variables exhibit significant skewness and are unlikely to provide meaningful insights.
- Therefore, they have been dropped from the analysis



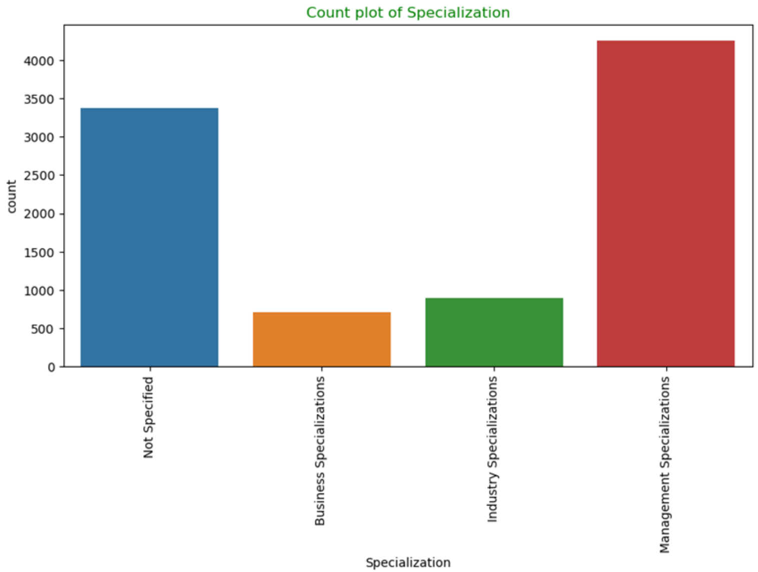




Data Analysis (EDA) Univariate Analysis

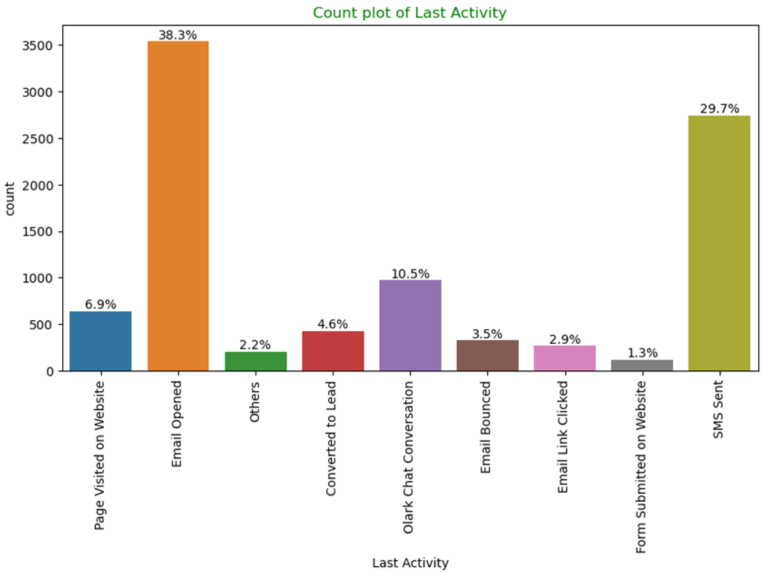
Specialization

Majority of customer chooses management specialization.



Lead Activity

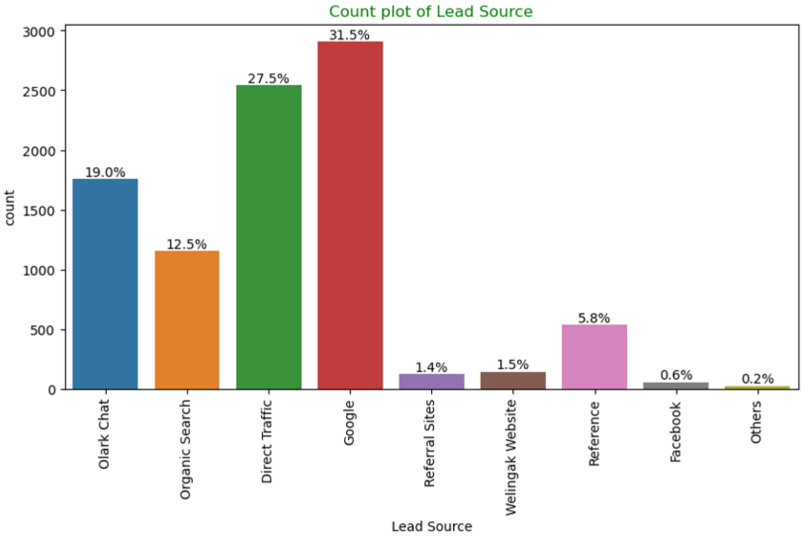
68% of customers contribution in SMS Sent & Email Opened activities



Data Analysis (EDA) Univariate Analysis

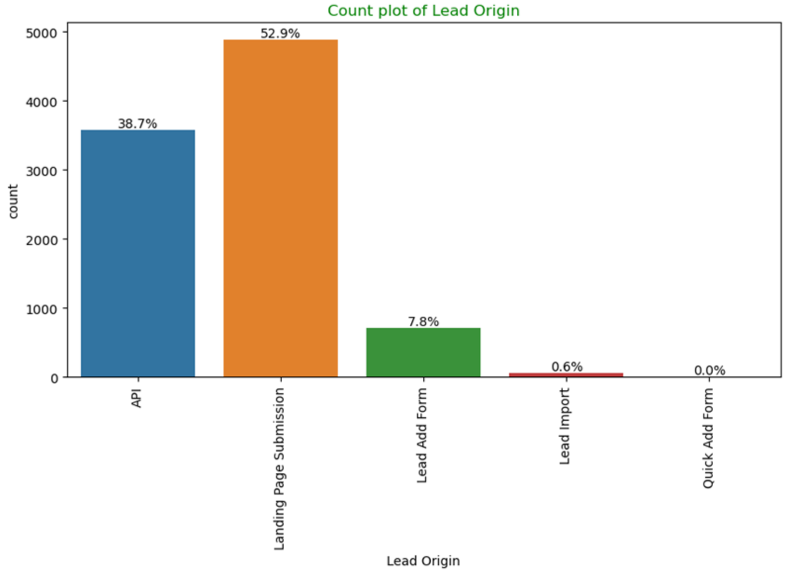
Lead Score

58% Lead source is from Google & Direct Traffic combined



Lead Origin

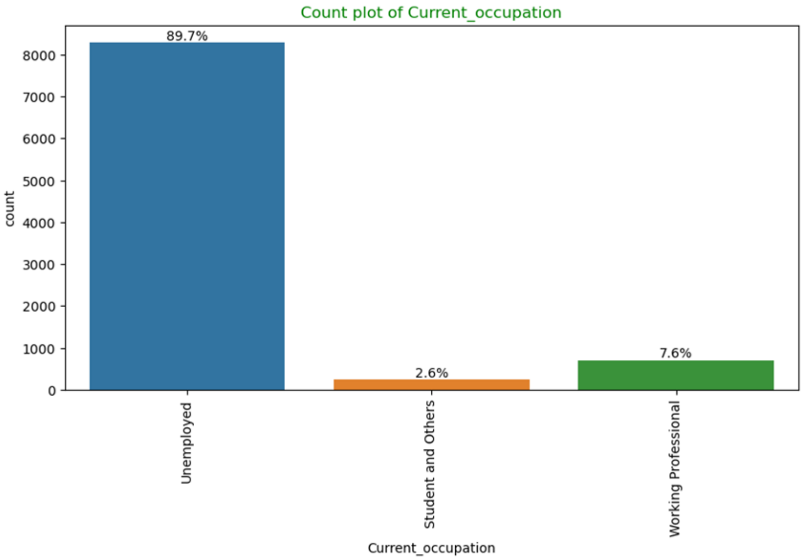
Landing Page Submission identified 53% customers, API identified 39%.



Data Analysis (EDA) Univariate Analysis

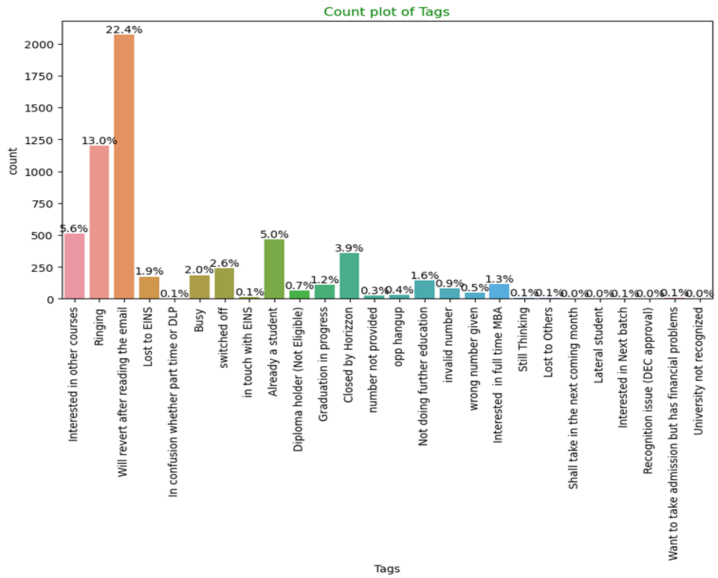
Current_Occupation

It has 90% of the customers as Unemployed



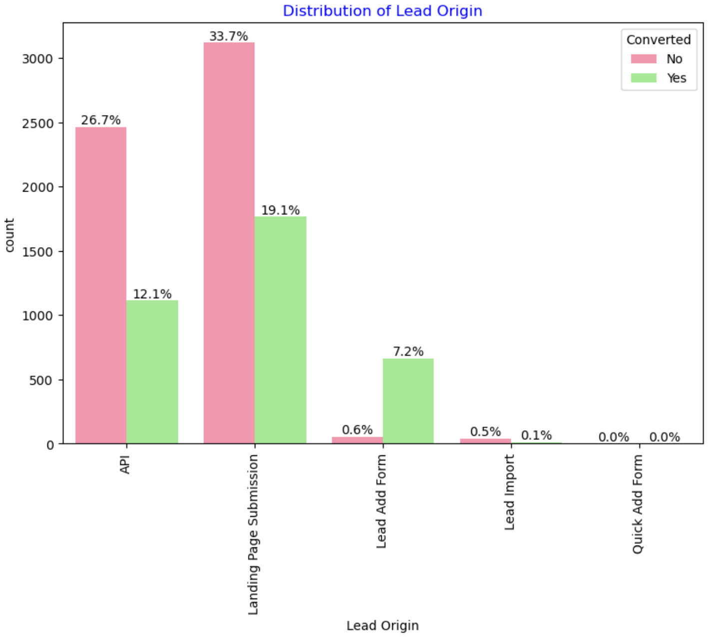
Tags

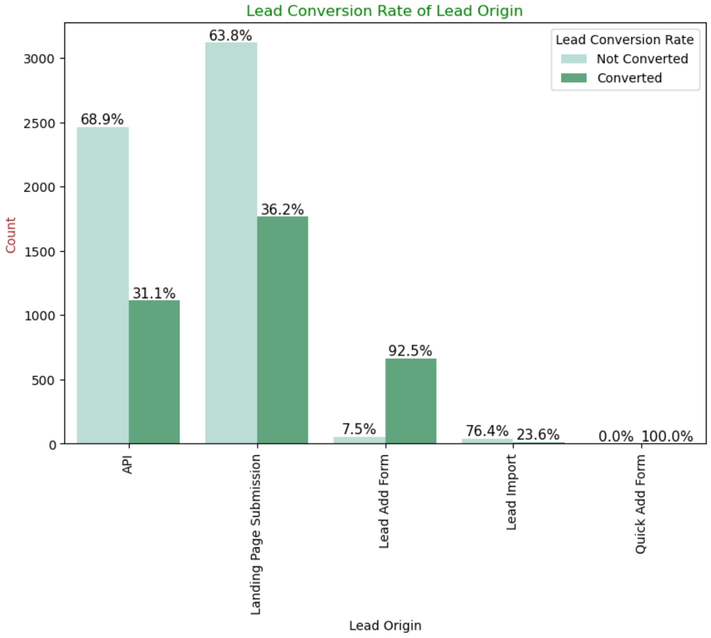
Will revert back contributes to 22.4% of lead conversion



Data Analyst (EDA) Bivariate Analysis

Lead Origin vs Conversation Rate

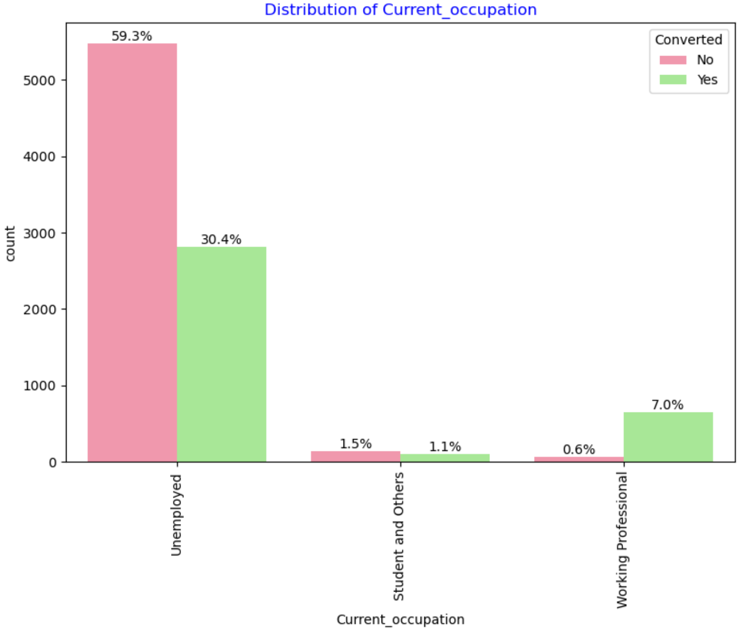


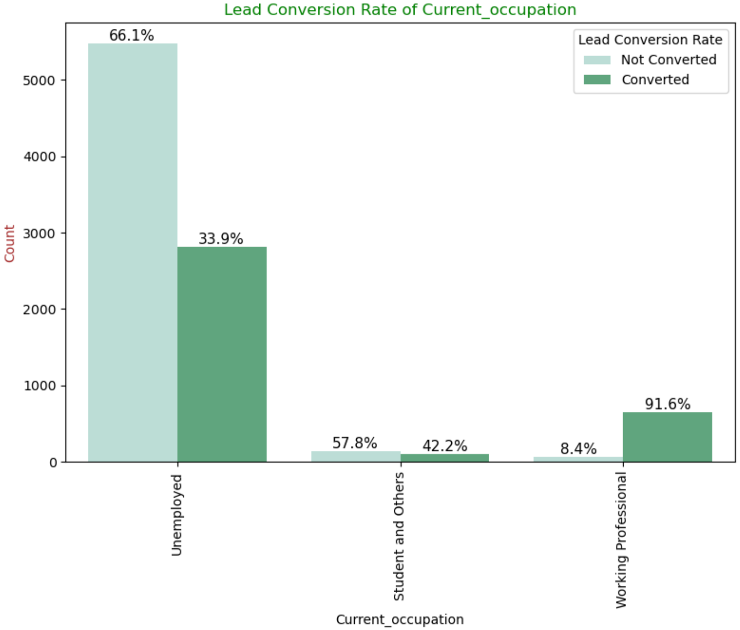


Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%.The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

Data Analyst (EDA) Bivariate Analysis

Current_Occupation vs Conversation Rate

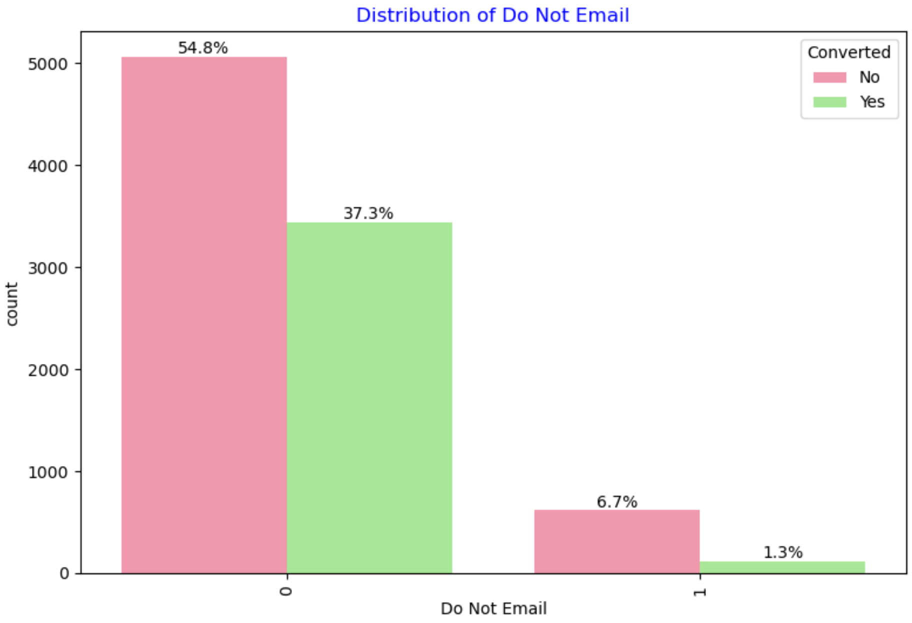


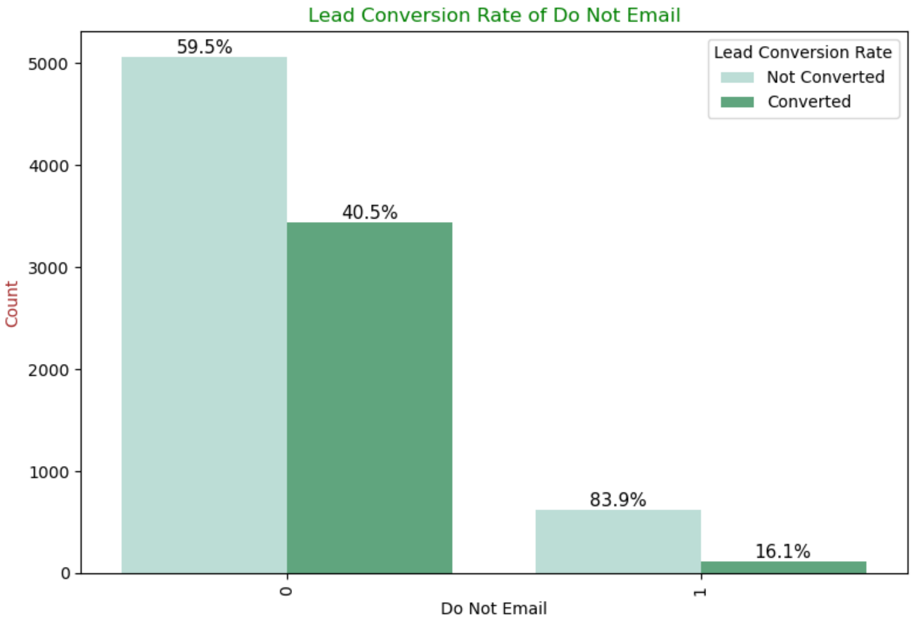


Around 90% of the customers are Unemployed with lead conversion rate (LCR) of 34%. While Working Professional contribute only 7.6% of total customers with almost 92% lead conversion rate (LCR).

Data Analyst (EDA) Bivariate Analysis

Do Not Email vs Conversation Rate

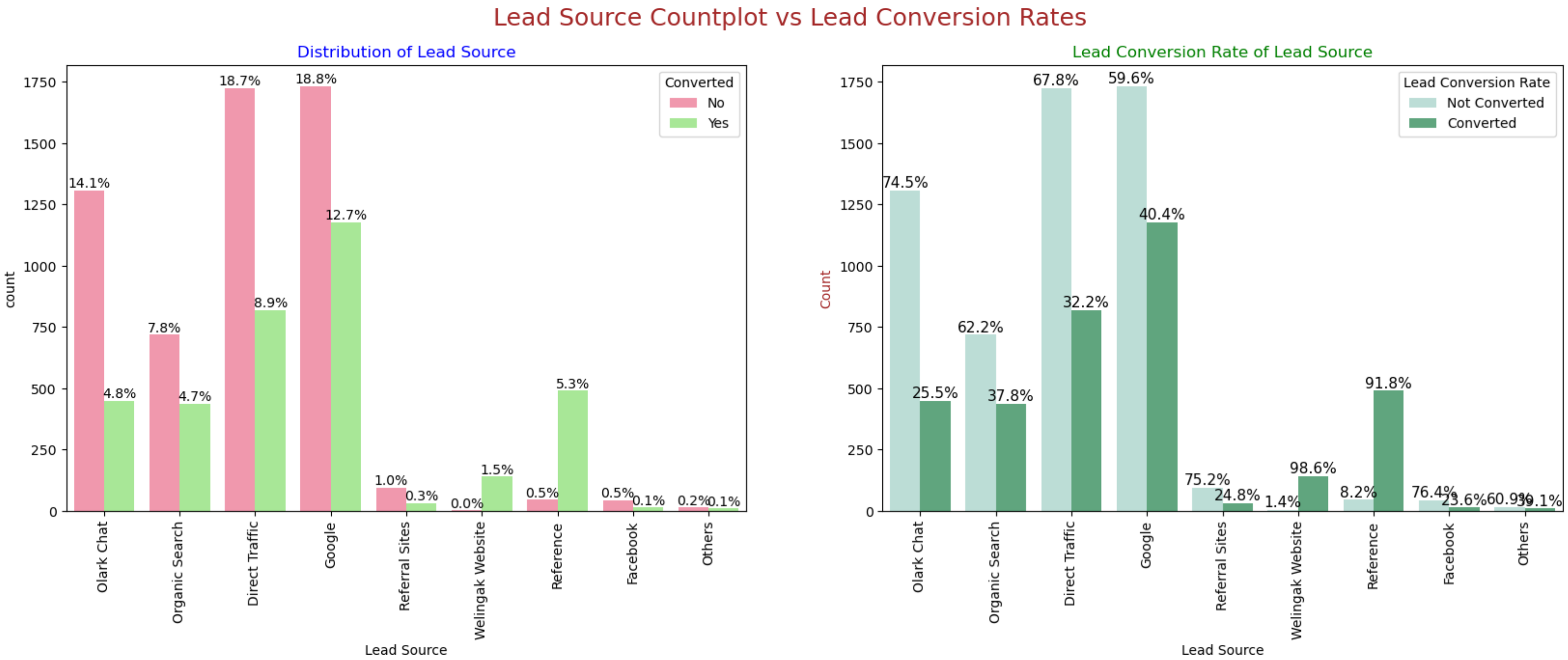




92% of the people has opted that they don't want to be emailed about the course.

Data Analyst (EDA) Bivariate Analysis

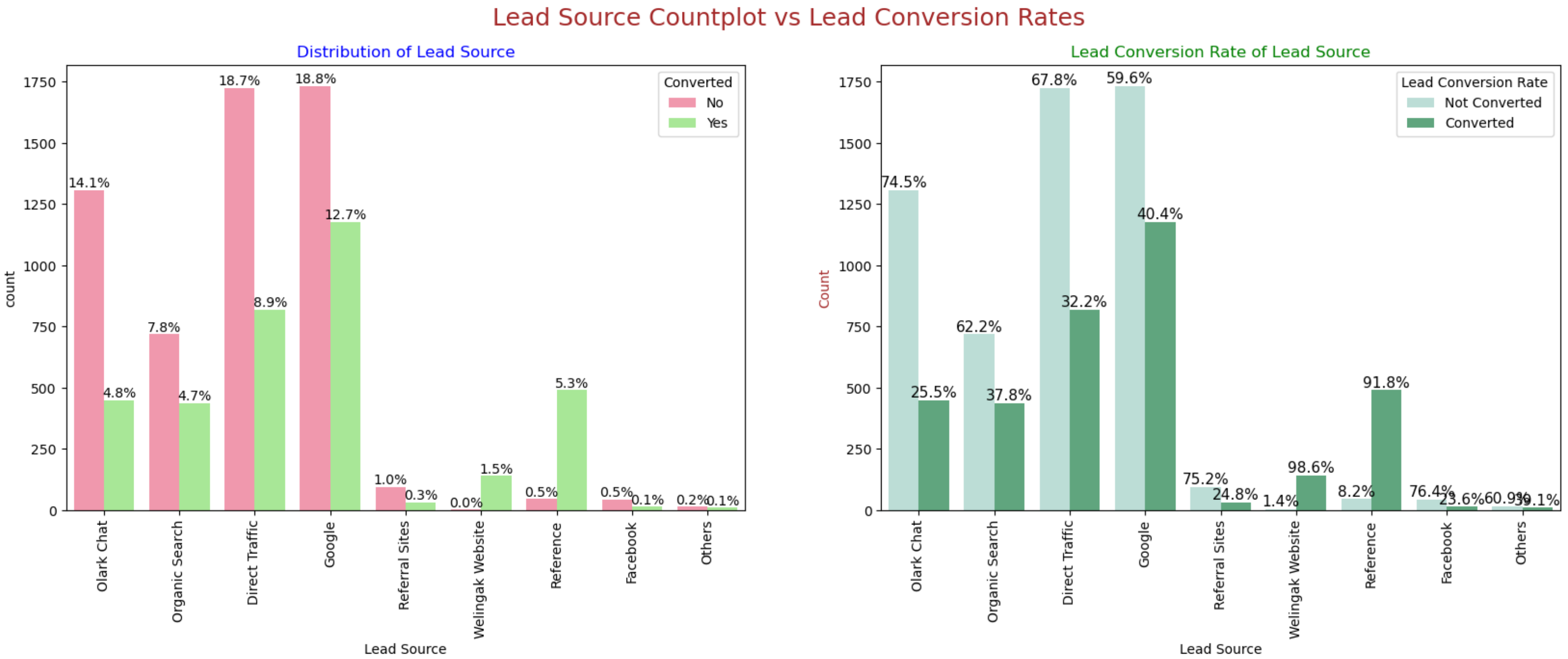
Do Not Email vs Conversation Rate



Google has LCR of 40% out of 31% customers , Direct Traffic contributes 32% LCR with 27% customers which is lower than Google, Organic Search also gives 37.8% of LCR but the contribution is by only 12% of customers, Reference has LCR of 91% but there are only around 5% of customers through this Lead Source.

Data Analyst (EDA) Bivariate Analysis

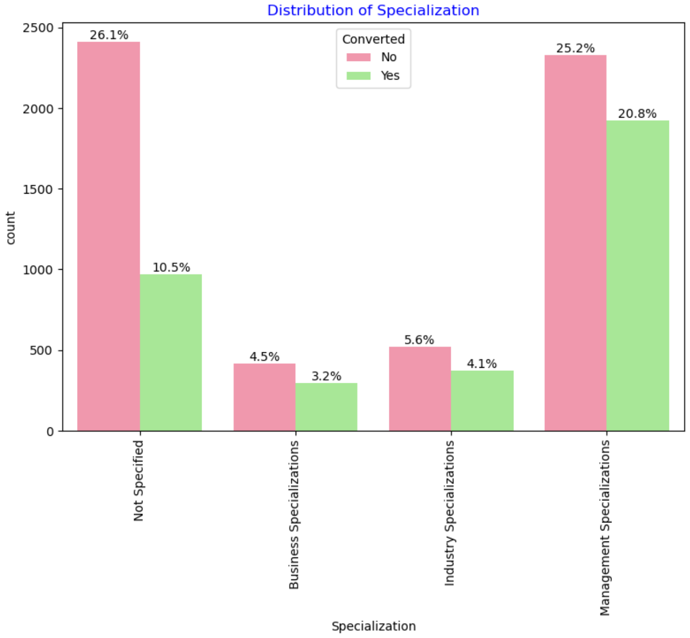
Lead Source vs Conversation Rate

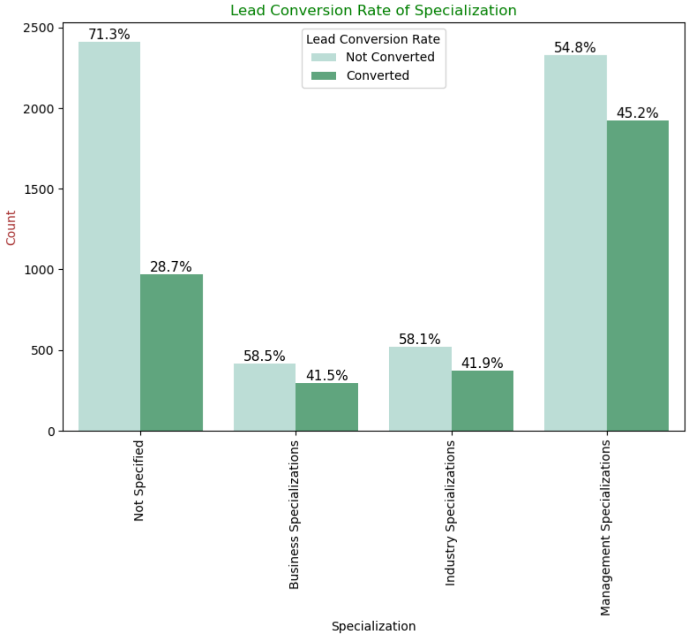


Google has LCR of 40% out of 31% customers , Direct Traffic contributes 32% LCR with 27% customers which is lower than Google, Organic Search also gives 37.8% of LCR but the contribution is by only 12% of customers, Reference has LCR of 91% but there are only around 5% of customers through this Lead Source.

Data Analyst (EDA) Bivariate Analysis

Specialization vs Conversation Rate

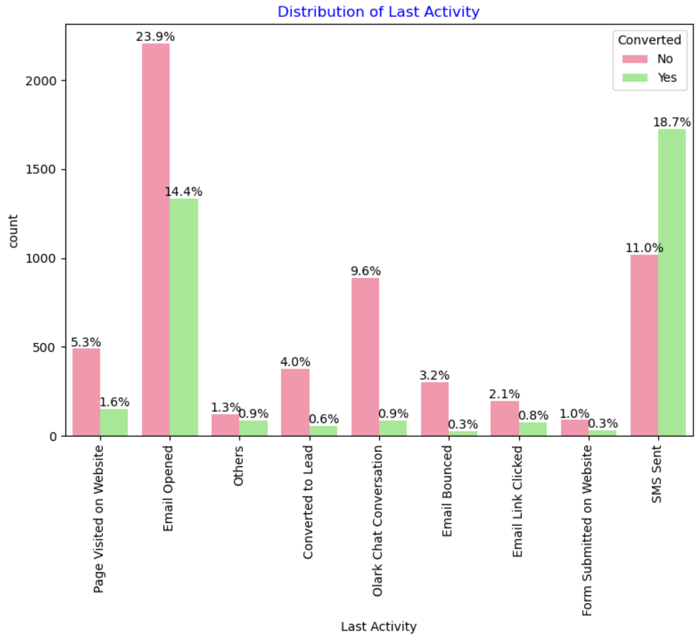


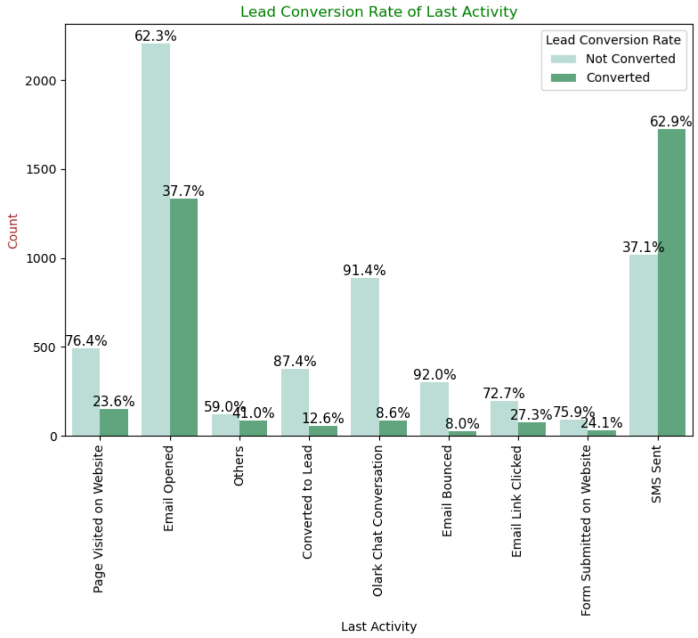


Management Specialization shows good contribution.

Data Analyst (EDA) Bivariate Analysis

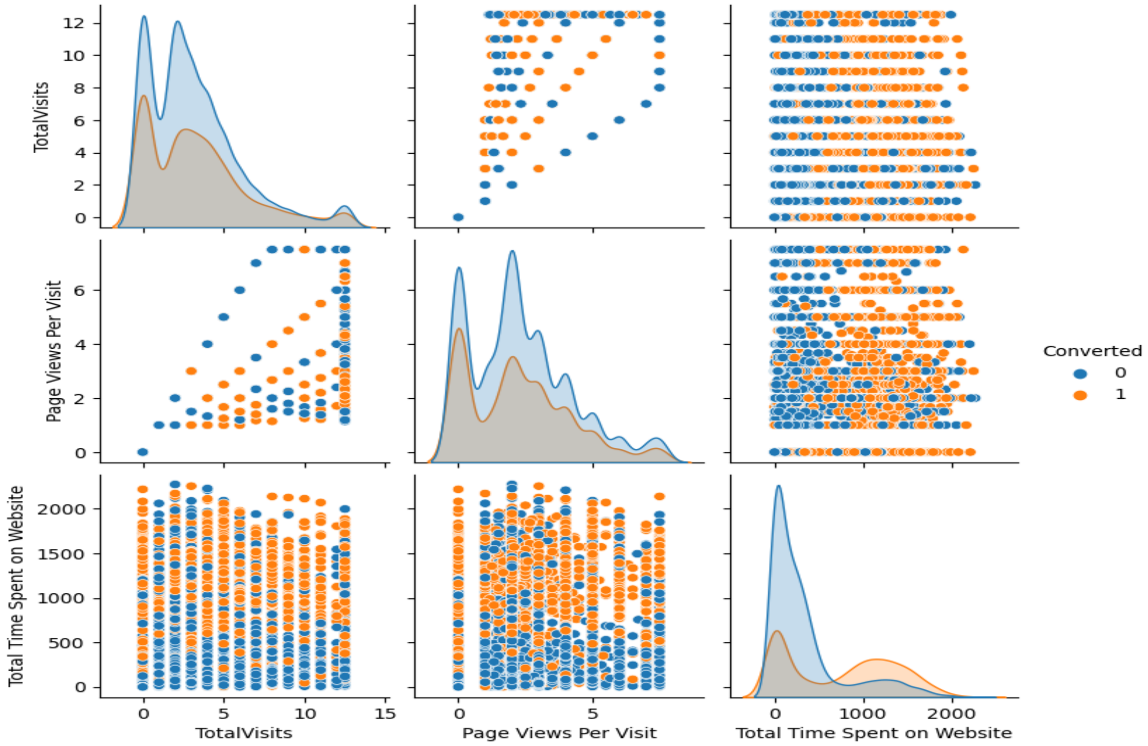
Last Activity vs Conversation Rate





'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities, 'Email Opened' activity contributed 38% of last activities performed by the customers with 37% lead conversion rate.

Data Analyst (EDA) Bivariate Analysis



The graph illustrates the behavior of various numerical variables concerning conversions.

Model Building

01	Train – Test Split	Split data into training and testing sets
02	Feature Scaling	Ensures all features contribute proportionally by scaling them, preventing domination by features with larger magnitudes
03	Feature Selection Using RFE	RFE systematically removes less important features, ranking them based on their impact on model performance
04	Coarse Tuning	Train Logistic Regression model on selected features
05	Manual Fine Tuning	Check model summary. Remove features with high p-values (>0.05) and VIF (>5).
06	Performance Metrics	Assess how well the model predicts outcomes compared to actual results. Use accuracy, precision, recall.

Model Evaluation

01	Confusion Matrix	Understand true positives, true negatives, false positives, and false negatives
02	**Feature Scaling**	Ensures all features contribute proportionally by scaling them, preventing domination by features with larger magnitudes
03	**ROC Curve**	Plot the ROC curve to visualize the trade-off between true positive rate and false positive rate.
04	**Finding Optimal Cutoff Point/Probability**	Determine the threshold maximizing model performance
05	**Performance Metrics**	Assess how well the model predicts outcomes compared to actual results. Use accuracy, precision, recall.

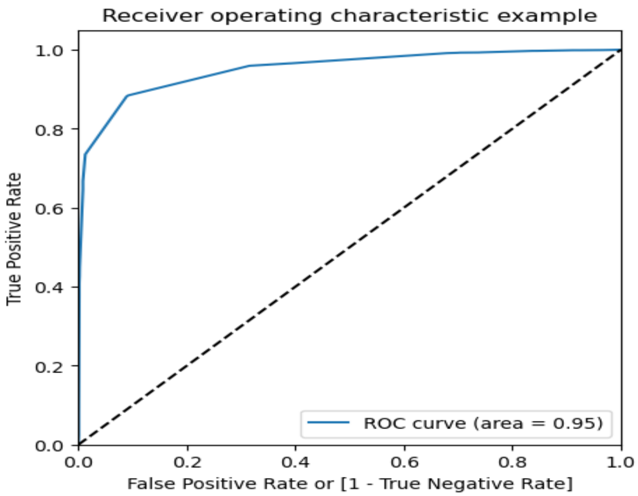
Model Evaluation

ROC Curve

Visualize the ROC curve to observe the dynamic relationship between true positive rate and false positive rate across different classification thresholds.

A value of 0.95 reflects a strong ability of the model to distinguish between positive and negative instances

This high ROC-AUC score suggests a robust performance, reinforcing the model's reliability in making accurate predictions.



The Area Under the ROC curve is 0.95, indicating a highly predictive model.

Model Evaluation

Performance		Train Set	Test Set
1	Accuracy	90%	90.12%
2	Sensitivity	88%	89.41%
3	Specificity	92.13%	90.58

Top Features

01 Tags_Closed by Horizzon

02 Tags_Lost to EINS

03 Tags_Will revert after reading the email

Inference

This highlights the relevance of tag-based information in predicting successful lead conversions.

The Logistic Regression model places higher importance on specific tags, such as "Closed by Horizzon," "Lost to EINS," and "Will revert after reading the email."

Insights from Model Feature Analysis

Inference

Logistic Regression assigns significance to tags such as "Closed by Horizzon," "Lost to EINS," and "Will revert after reading the email."

These tags serve as indicators of the current status of a lead, portraying the actions or decisions taken by the lead.

Relying solely on the importance of tags may not align with an optimal strategy for proactive lead engagement.

Solely relying on the importance of tags implies that the sales team is expected to initiate outreach based on these tag indications.

To further explore relevant features and maintain a proactive approach to lead engagement, a Decision Tree model is employed.

Decision Trees offer insights into features beyond tag importance, providing a more comprehensive understanding of factors influencing lead conversions.

Decision Tree

Performance		
1	Accuracy	91.7%
2	Sensitivity	86.21%
3	Specificity	95.29%

Top Features

01 Lead Source_Organic Search

02 Lead Source_Facebook

03 Lead Origin_Lead Add Form

Inference

Decision Tree emphasizes the importance of online channels (Organic Search, Facebook) and Lead Add Forms in predicting lead conversions.

This suggests that online search, social media platforms, and interactions through specific lead forms are influential factors in predicting conversions.

Conclusion

The Decision Tree's emphasis on lead sources and forms suggests that online channels and specific forms play a crucial role in conversion prediction.

On the other hand, the Logistic Regression model underscores the importance of tag-related information.

Combining insights from both models, a comprehensive strategy might involve prioritizing lead sources and forms, especially those identified by the Decision Tree, while also paying attention to tags as indicated by Logistic Regression.

This holistic approach could enhance the overall effectiveness of lead conversion strategies.