Lead Scoring Case Study

Presentation

Akhila Setiraju

Introduction

- In this case study, we will build a logistic regression model for X Education to assign a lead score between 0 and 100 to each lead. The lead score will help the company identify potential leads and prioritize them based on their likelihood of conversion. Our aim is to help X Education achieve their target conversion rate of 80%. Additionally, we will address the other problems presented by the company and provide recommendations on how to utilize the lead scoring model effectively to achieve their business goals. The model should also be able to adjust to any changes in the company's requirements in the future.

BUSINESS UNDERSTANDING

- X Education is a provider of online courses for industry professionals.
- The company markets its courses through various digital channels, including search engines like Google.
- Interested customers visit the X Education website to explore the available courses.
- Some visitors complete a form on the website, providing their email address or phone number to indicate interest. These individuals are identified as leads.
- The X Education sales team contacts these leads via phone or email to try and convert them into paying customers.
- While some leads do convert to paying customers, the majority do not.
- The average lead conversion rate for X Education is around 30%.

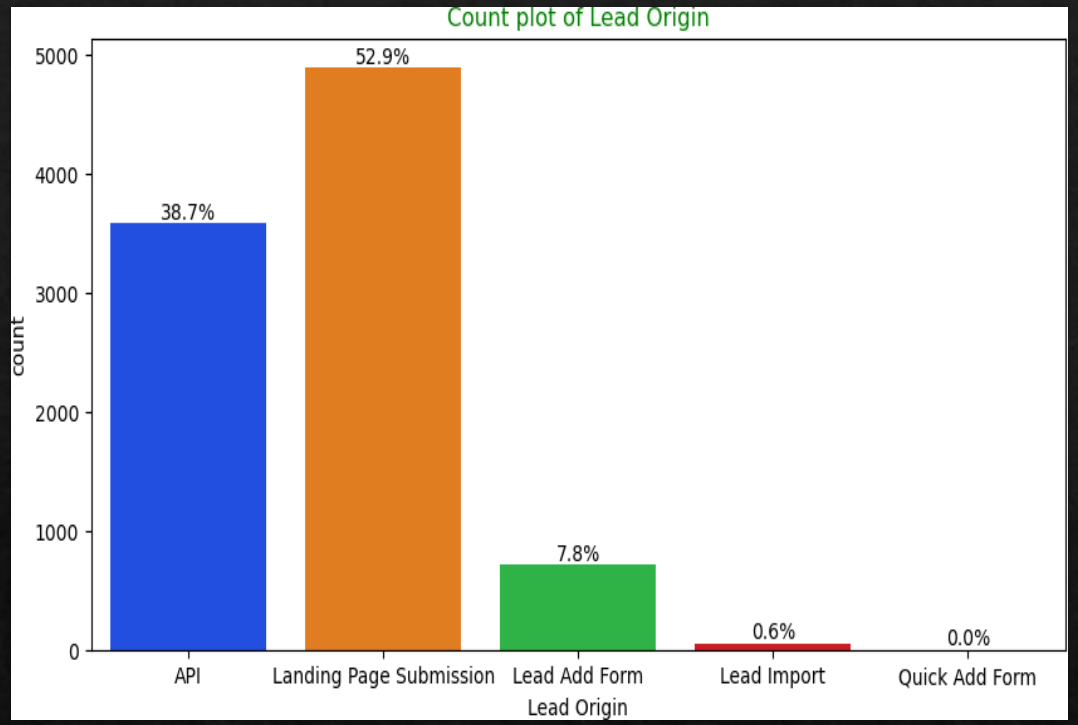
Data Understanding

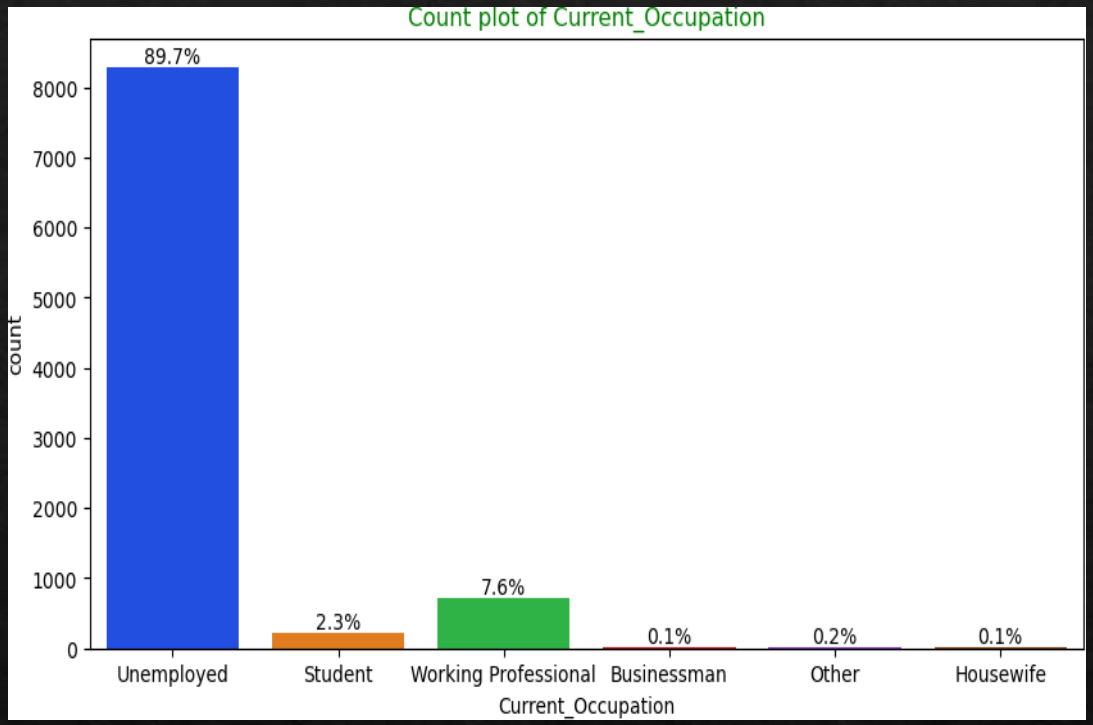
- The dataset consists of two files: 'Leads.csv' and 'Leads Data Dictionary.xlsx'.
- The 'Leads.csv' file contains around 9000 data points. The target variable of the dataset is the column 'Converted', which indicates whether a past lead was converted or not. The values in the 'Converted' column are binary, where 1 means the lead was converted and 0 means it wasn't converted.
- The 'Leads Data Dictionary.xlsx' file provides a data dictionary that explains the meaning of the variables in the 'Leads.csv' file.

Data Cleaning Insights

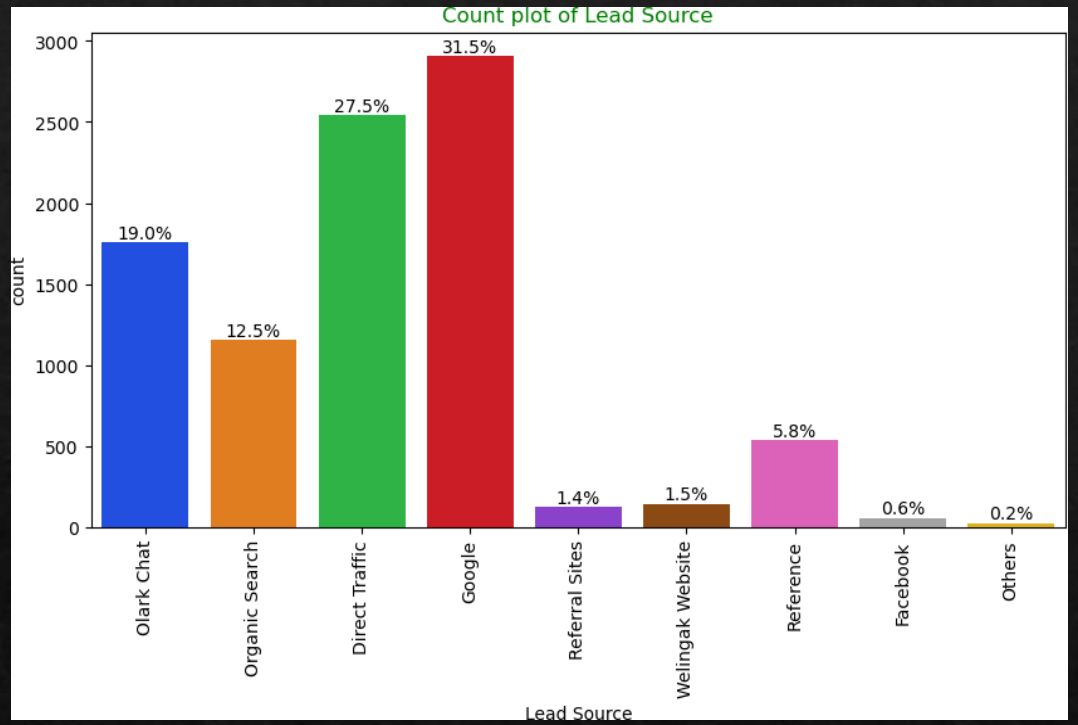
- The "Select" level represents null values for certain categorical variables when customers did not choose any option from the list.
- Columns with more than 40% null values were removed.
- Missing values in categorical columns were addressed based on value counts and specific considerations.
- Columns that did not contribute to the study objective (such as 'City', 'Tags', 'Country', 'What matters most to you in choosing a course') were dropped.
- Imputation was applied to some categorical variables.
- Columns irrelevant to modeling ('Prospect ID', 'Lead Number', and 'Last Notable Activity') were removed.
- Numerical data was imputed using the mode after examining the distribution.
- Skewed categorical columns were reviewed and dropped to prevent bias in logistic regression models.
- Outliers in 'TotalVisits', 'Total Time Spent on Website', and 'Page Views Per Visit' were identified and capped.
- Low frequency values were grouped into an "Others" category.
- Data standardization involved checking and correcting casing styles in columns (e.g., "Lead Source" had both Google and google).

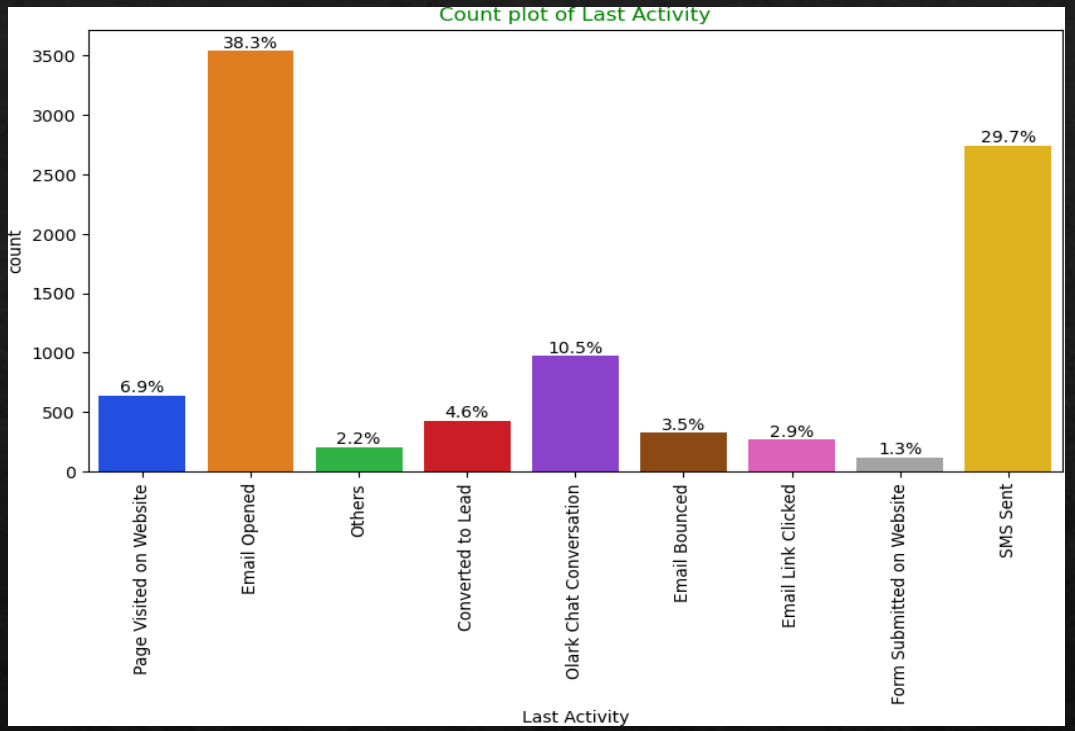
EDA
Univariate Analysis



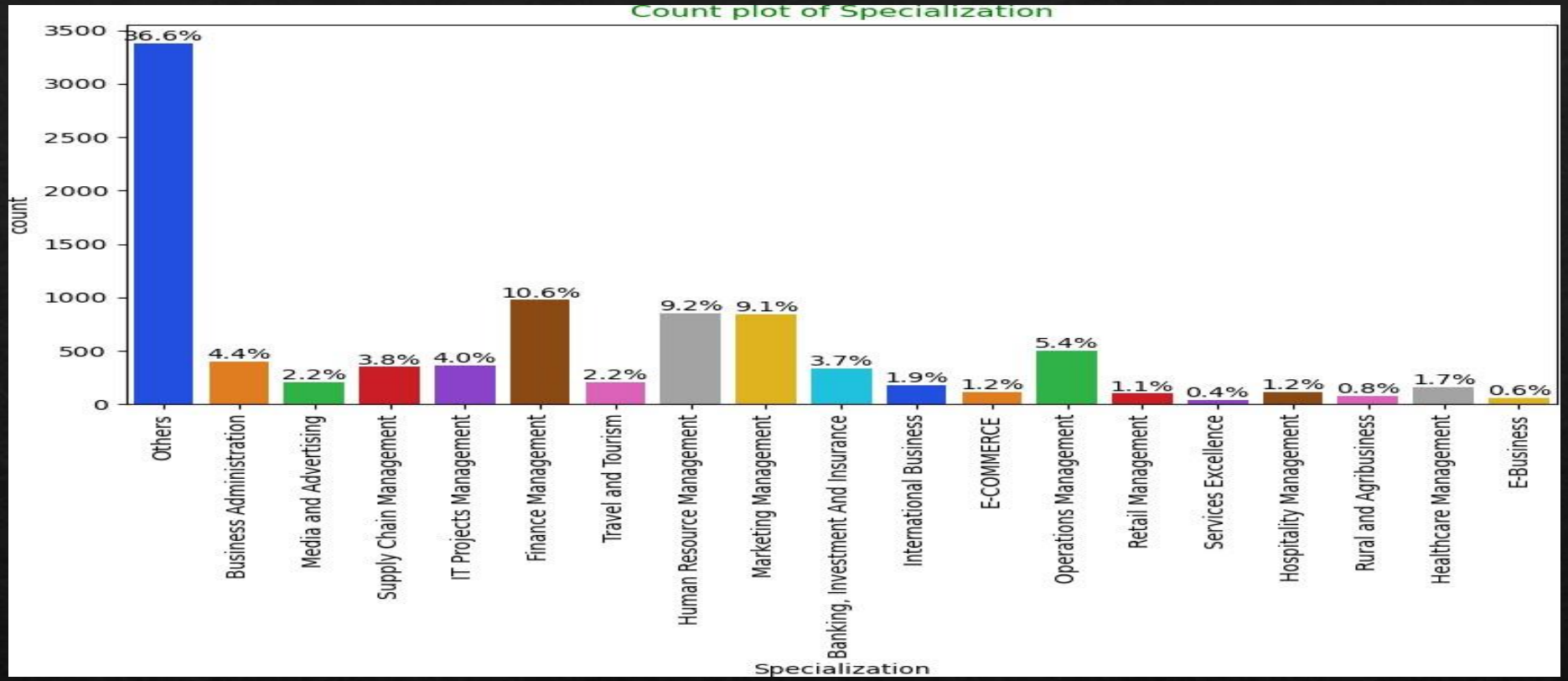


Lead Origin: Most customers, 52.9%, were identified through 'Landing Page Submission', with 'API' being the second most common source at 38.7%.
Current Occupation: A large majority of customers, 89.7%, are unemployed based on the current occupation data.



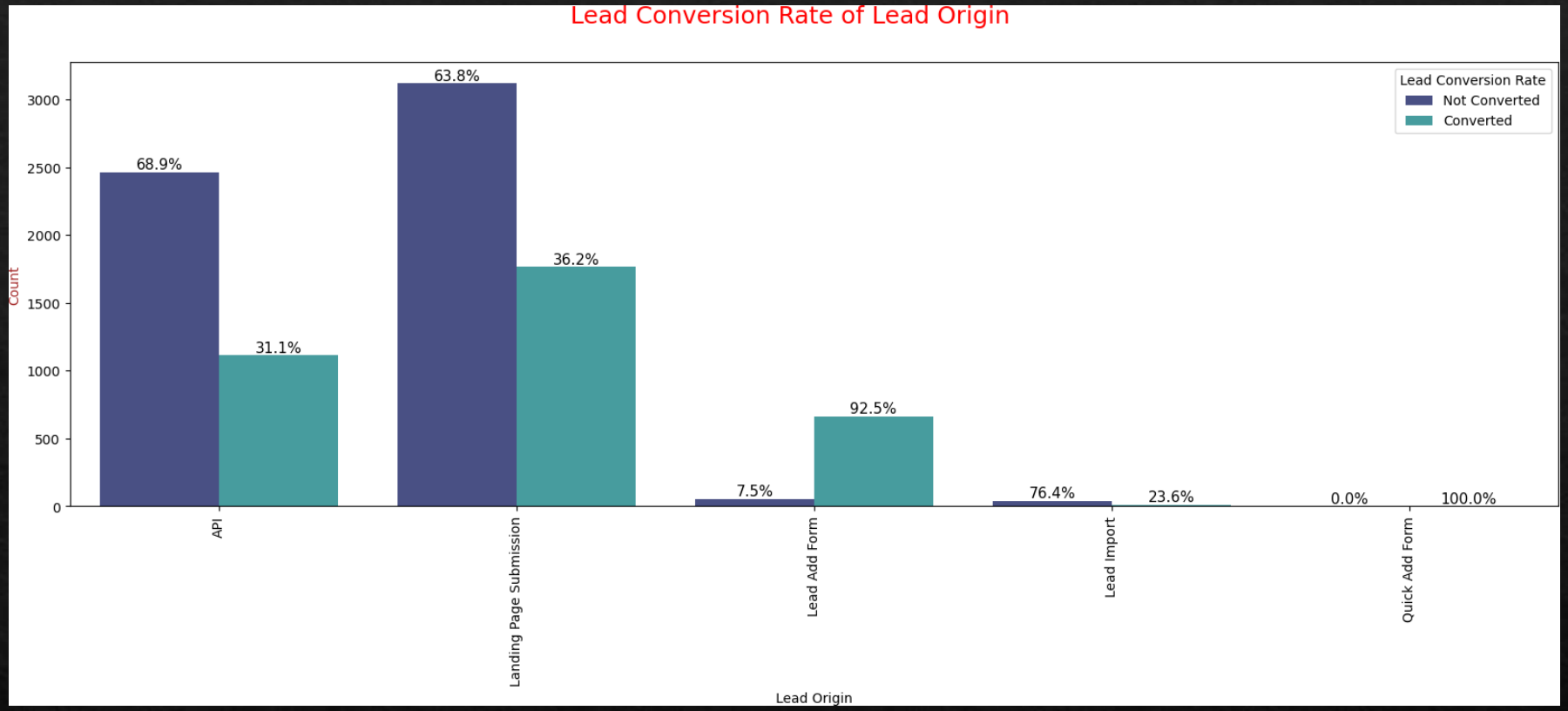


- Last Activity: Email is the most common last activity, with 38.3% of customers having opened an email, and 29.7% having sent an SMS
- Lead Source: The primary lead source is Google at 31.5%, followed by Direct Traffic at 27.5%.

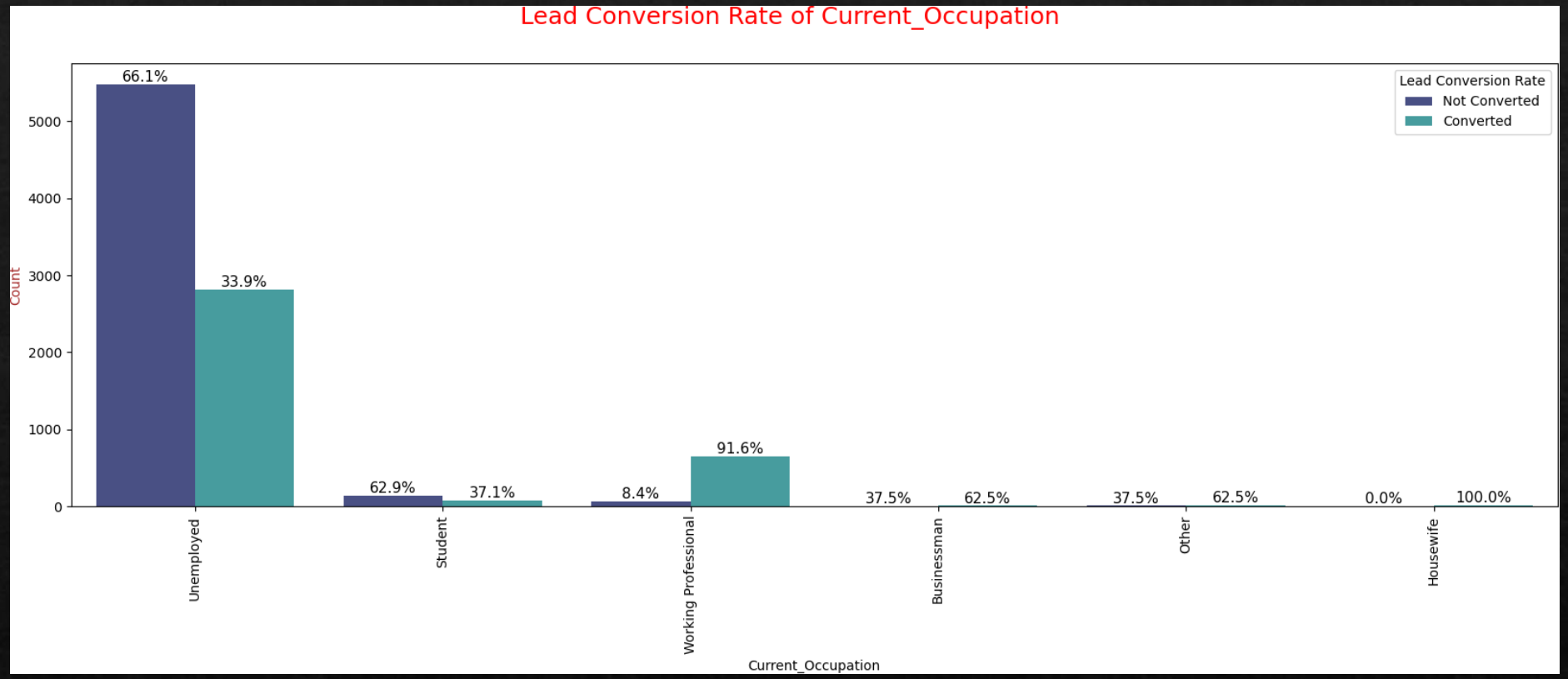


- Specialization: The 'Others' specialization category is the most common among customers at 36.6%, followed by Finance Management at 10.6%, HR Management at 9.2%, Marketing Management at 9.1%, and Operations Management at 5.4%.

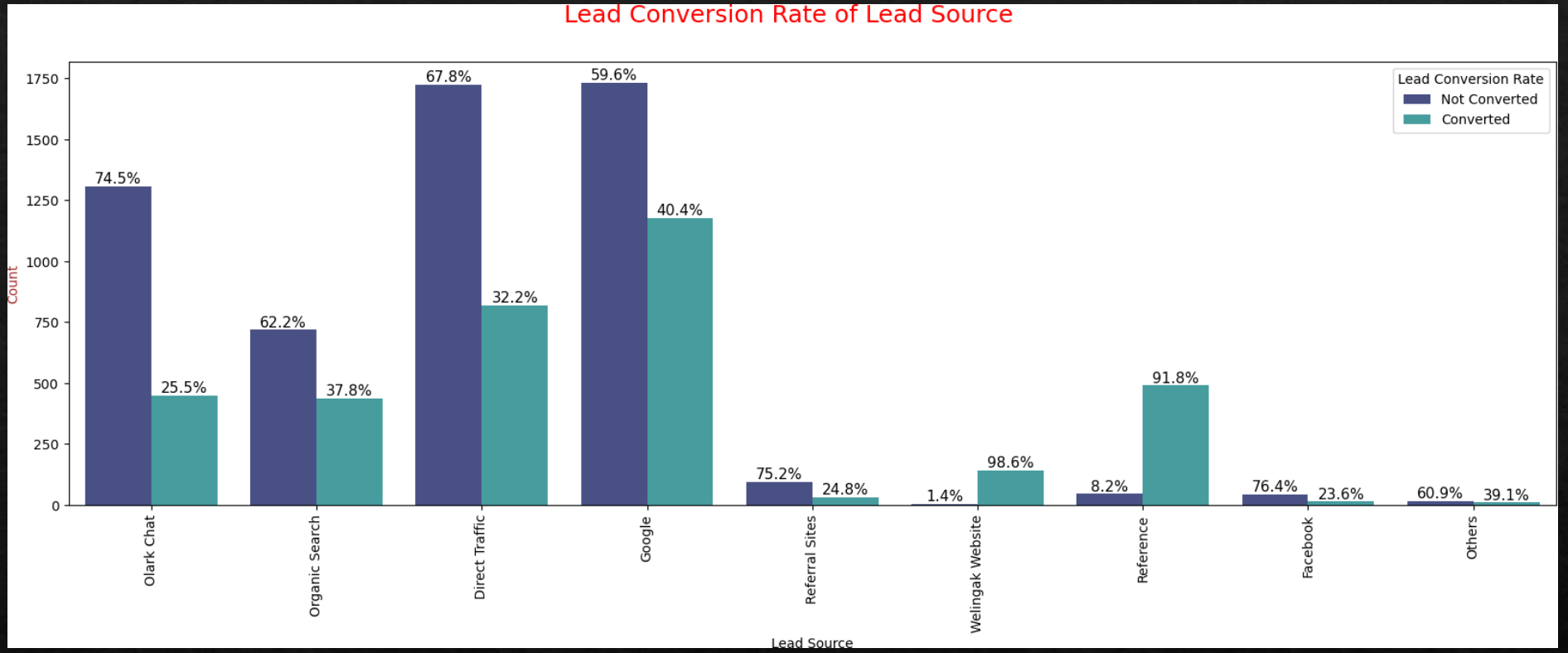
EDA
Bivariate Analysis



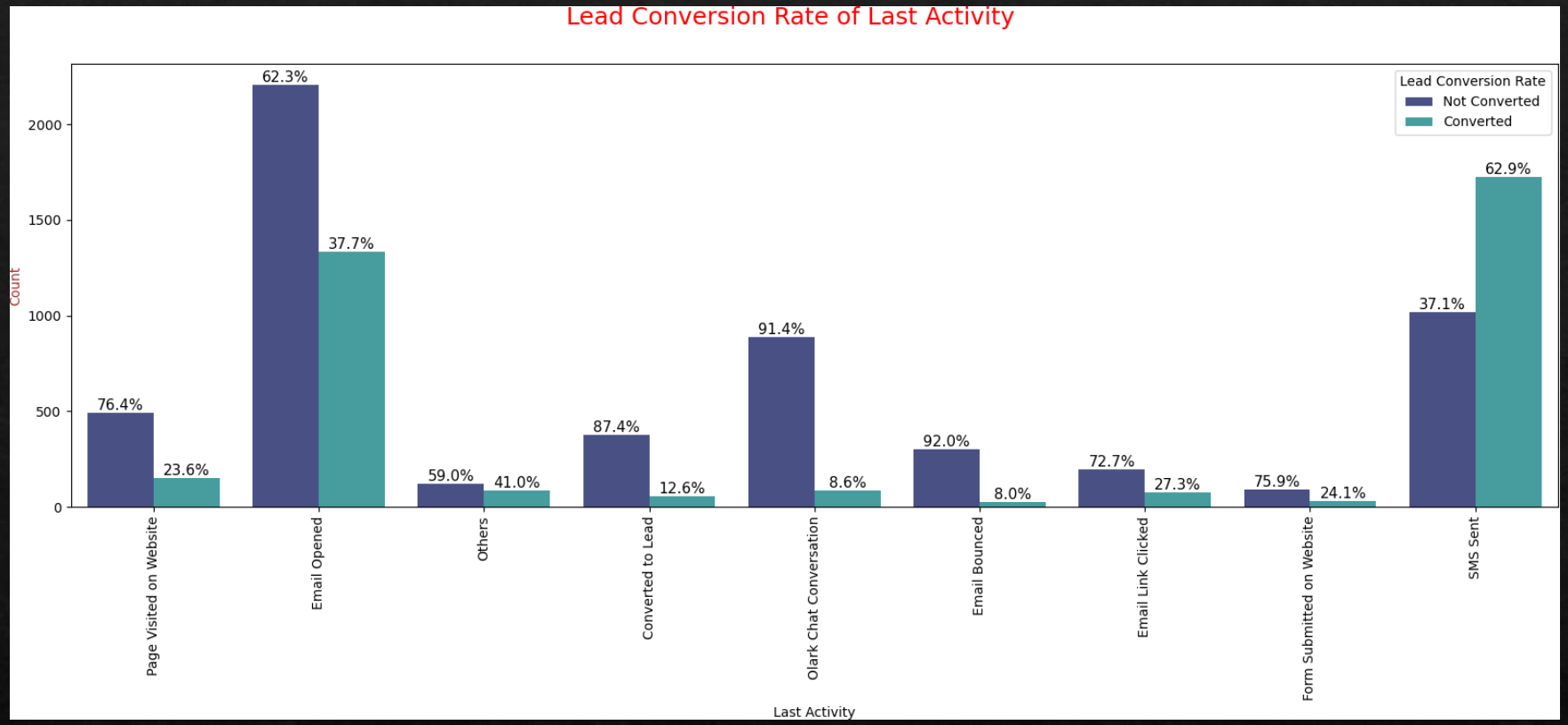
Lead Origin: 'Landing Page Submission' is the most effective Lead Origin with a Lead Conversion Rate (LCR) of 36.2%, followed by 'API' at 31.1%.



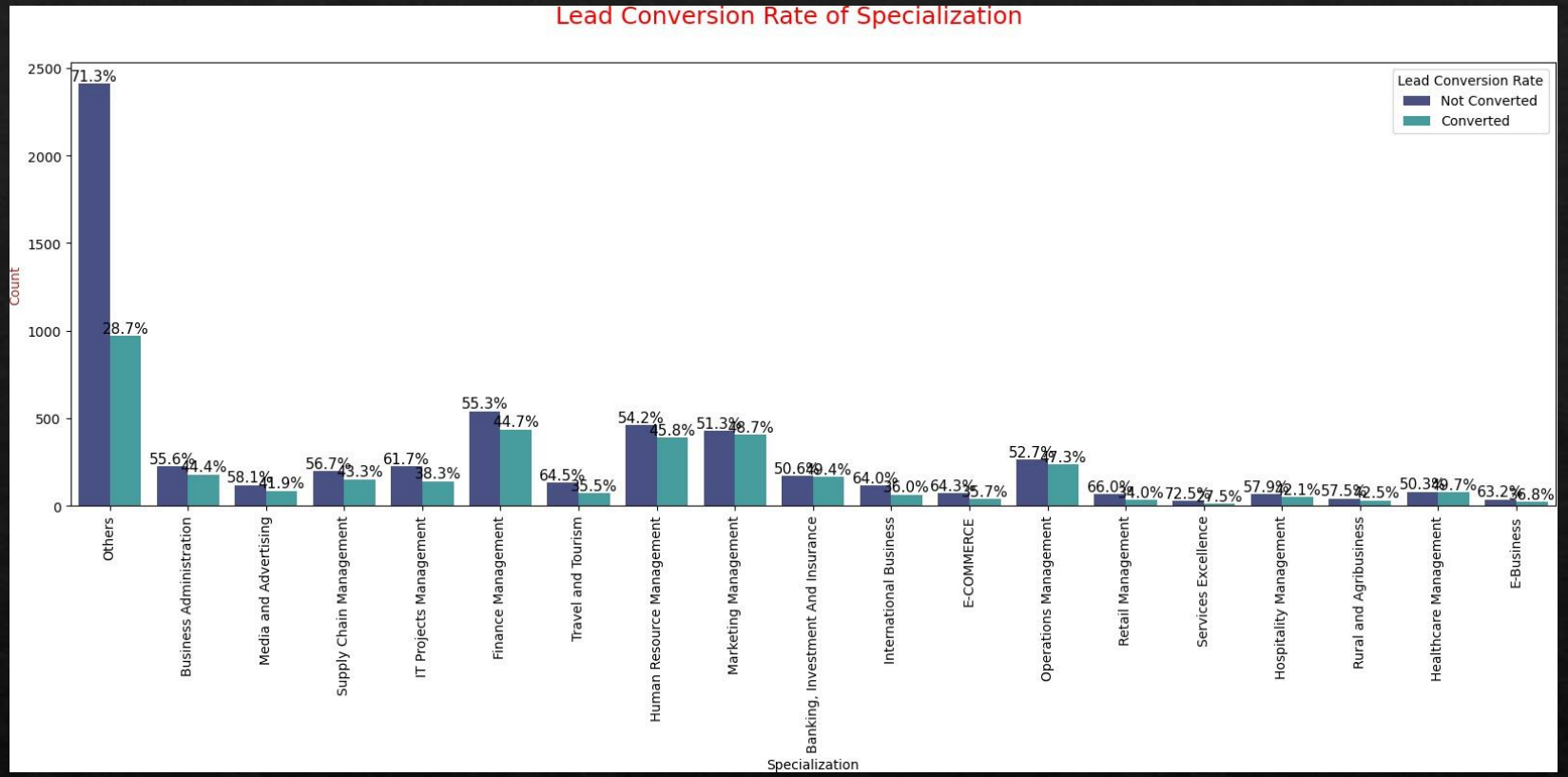
Current_Occupation: Working Professionals have a significantly higher LCR at 91.6% compared to Unemployed people at 33.9%.



Lead Source: Google is the most effective Lead Source with an LCR of 40.4%, followed by Direct Traffic at 32.2% and Organic Search at 37.8% (contributing to only 12.5% of customers). Reference has the highest LCR at 91.8%, but there are only 5.8% of customers through this Lead Source.

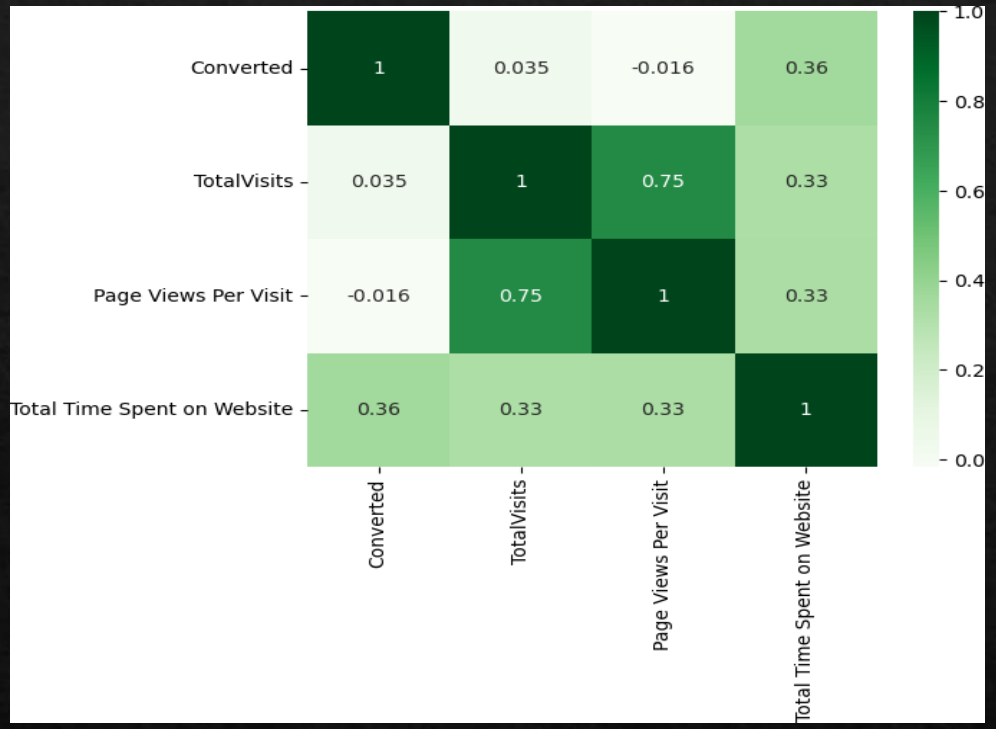


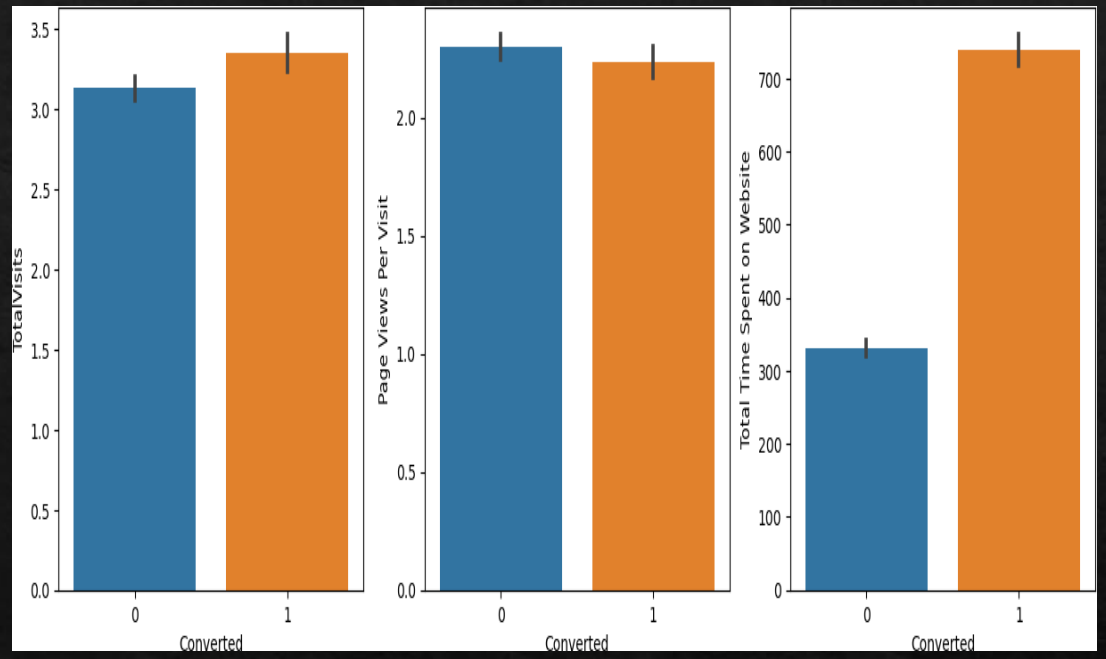
Last Activity: SMS Sent and Email Opened are the most effective Last Activity types with LCRs of 62.9% and 37.7% respectively.



Specialization: Marketing Management, HR Management, Finance Management and Operations Management all show good LCRs, indicating a strong interest among customers in these specializations

Correlation Analysis





- There is a strong positive correlation between 'Total Visits' and 'Page Views per Visit', indicating that customers who visit the website more frequently tend to view more pages per visit.
- Customers who spend more time on the website have a higher LCR, indicating that increasing the time spent on the website can lead to higher conversion rates.

Data Preparation

- Binary level categorical columns were mapped to 1/0 in previous steps to make them compatible with the logistic regression model.
- Dummy features were created for categorical variables such as Lead Origin, Lead Source, Last Activity, Specialization, and Current_Occupation, using one-hot encoding.
- The train and test sets were split in a 70:30 ratio to train the model and evaluate its performance on unseen data.
- Feature scaling was performed using the standardization method to ensure that all features were on the same scale and no feature dominated the others.
- Correlated predictor variables, such as Lead Origin_Lead Import and Lead Origin_Lead Add Form, were dropped to avoid multicollinearity issues.

Model Building

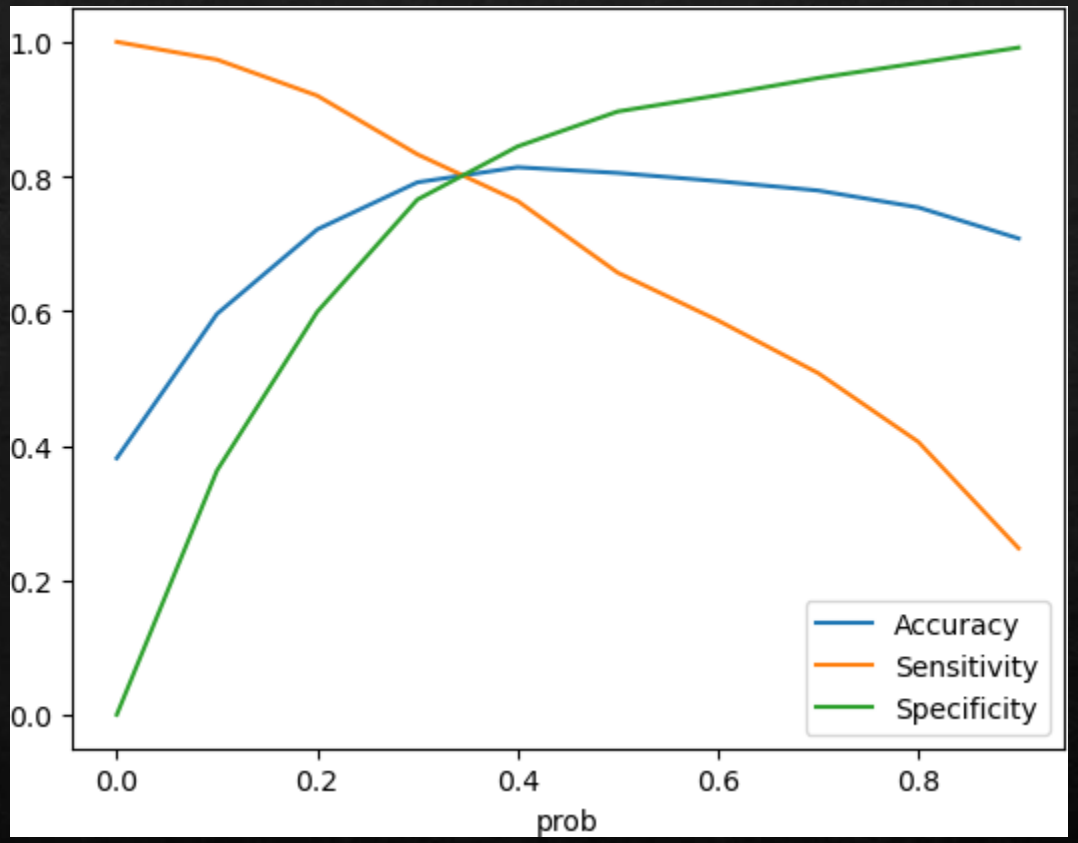
Model Building

- The data set has a large number of features and dimensions which can reduce model performance and increase computation time.
- Recursive Feature Elimination (RFE) is performed to select only the important columns.
- Pre RFE, the data set had 48 columns and post RFE it has 15 columns.
- Logistic Regression Model - 1 is a basic model.
- Manual feature reduction process was used in Logistic Regression Model - 2 and 3 to build models by dropping variables with p-value greater than 0.05.
- Logistic Regression Model - 4 is stable after four iterations with:
- Significant p-values within the threshold (p-values < 0.05)
- No sign of multicollinearity with VIFs less than 5
- Logistic Regression Model - 4 (LRMod4) is the final model used for model evaluation and making predictions.

Model Evaluation

CONFUSION MATRIX - 1		
Actual/Predicted	not_converted	converted
not_converted	**3588**	**414**
converted	**846**	**1620**

Accuracy	0.8052
Sensitivity	0.6569
Specificity	0.8966
False Positive Rate	0.1034
Precision	0.7965
Recall	0.6569
Negative Predictive Value	0.8092

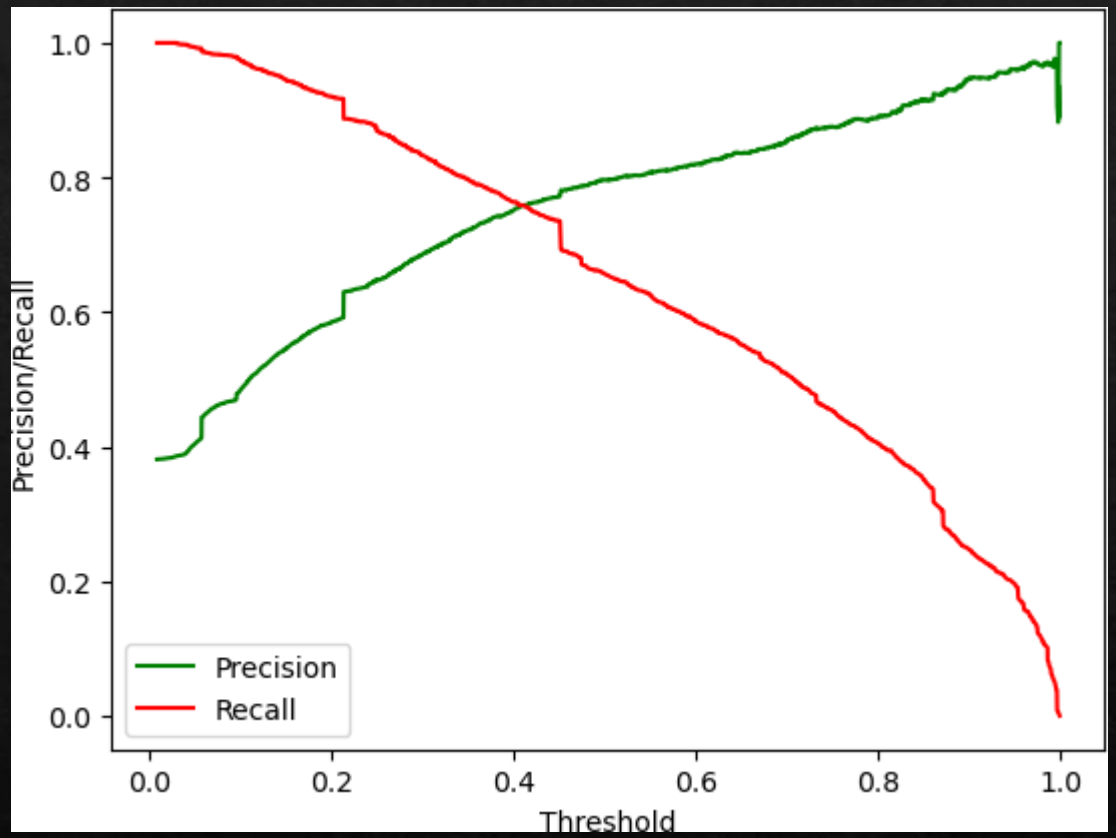


Inference:

- **Based on the curve analysis, a cutoff probability of 0.35(approx.) is suggested as the optimal point for classification**

CONFUSION MATRIX - 2		
Actual/Predicted	not_converted	converted
not_converted	**3064**	**938**
converted	**412**	**2054**

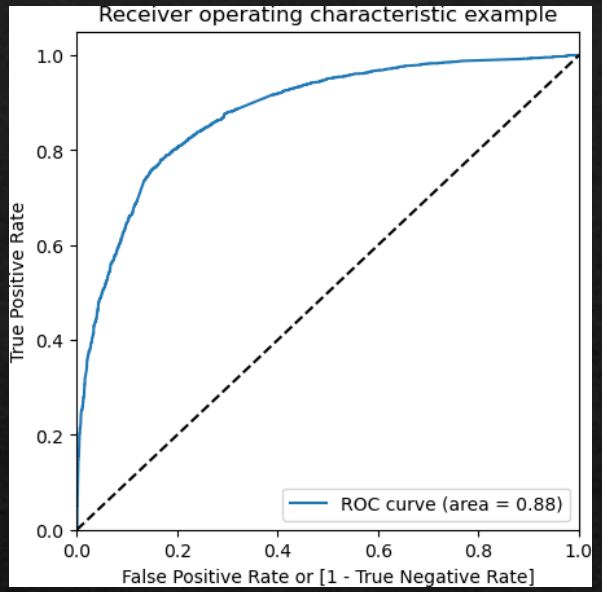
Accuracy	0.8057
Sensitivity	0.7972
Specificity	0.8108
False Positive Rate	0.1892
Precision	0.722
Recall	0.7972
Negative Predictive Value	0.8665



Inference:

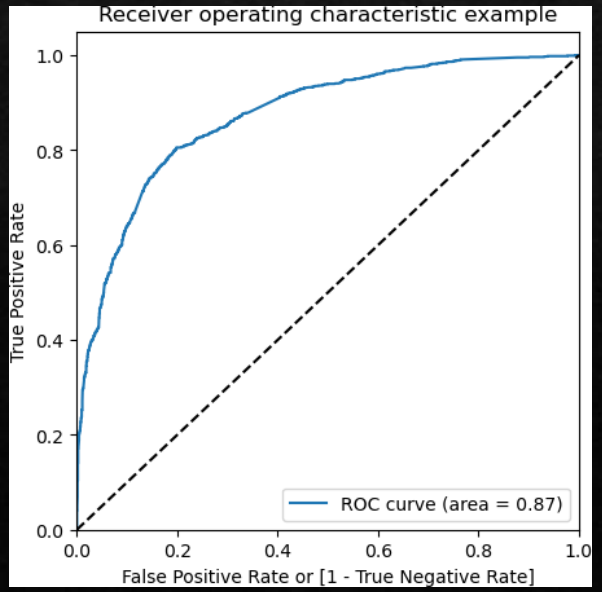
- Based on the precision-recall curve, a threshold of 0.4 provides a good balance between
- precision and recall.

Making Prediction On Test Data



ROC Curve – Train Data Set

The Area under ROC curve was found to be 0.88 out of 1, indicating that the model is a good predictor.
The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve – Test Data Set

The Area under ROC curve was found to be 0.87 out of 1, indicating that the model is a good predictor.
The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.

Conclusion

CONFUSION MATRIX - 3		
Actual/Predicted	not_converted	converted
not_converted	**1359**	**318**
converted	**227**	**868**

Accuracy	0.8034
Sensitivity	0.7927
Specificity	0.8104
False Positive Rate	0.1896
Precision	0.7319
Recall	0.7927
Negative Predictive Value	0.8569

Inference:

- Train Data Set:
  - Accuracy: 80.57%
  - Sensitivity: 79.72%
  - Specificity: 81.08%
- Test Data Set:
  - Accuracy: 80.34%
  - Sensitivity: 79.27%
  - Specificity: 81.04%
- The evaluation metrics of the model are consistently close to each other, indicating that the model is performing consistently across different evaluation metrics in both the test and train datasets. This consistency suggests that the model is reliable and is not overfitting to the training data. It also implies that the model is generalizing well to new data, which is important for real-world applications. The similar performance across evaluation metrics also means that there are no significant biases in the model's predictions. This is a positive sign for the model's performance and provides confidence in its ability to make accurate predictions in the future.

Recommendations

- Features such as 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional', and 'Total Time Spent on Website' have a high conversion rate and should be utilized more in lead generation efforts.
- Working professionals should be aggressively targeted as they have a higher probability of converting and are likely to have better financial situations to pay for services.
- Referral leads generated by old customers have a significantly higher conversion rate and should be incentivized with discounts or other rewards to encourage more referrals.
- Increasing the frequency of media usage such as Google ads or email campaigns can save time and increase the conversion rate.
- Leads whose 'Last Activity' is ‘SMS Sent’ or ‘Email Opened’ tend to have a higher conversion rate and should be targeted more frequently.
- Analyzing the behavior of customers who spend more time on the website can help improve the user experience and increase conversion rates, and company should focus on creating engaging content and user- friendly navigation to encourage customers to spend more time on the website.
- Understanding the most popular specializations can help tailor course offerings and marketing campaigns to
- specific groups of customers. Providing targeted content and resources for popular specializations such as Marketing Management and HR Management can also help attract and retain customers in those fields