LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

Submitted by:
Minh Nguyen
Nikhil Agarwal
Neha Gaonkar

CONTENTS:

No.	TOPICS
1.	PROBLEM STATEMENT
2.	PROBLEM APPROACH
3.	EDA : DATA CLEANING AND DATA PREPARATION
4.	OBSERVATION
5.	CONCLUSIONS

PROBLEM STATEMENT

1.An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.

2. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

3. The typical lead conversion rate at X education is around 30%. Now, this means 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.

4. 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:

- Lead X wants us to build a model to give every lead a lead score between 0 -100 .So that they can identify the Hot leads and increase their conversion rate as well.

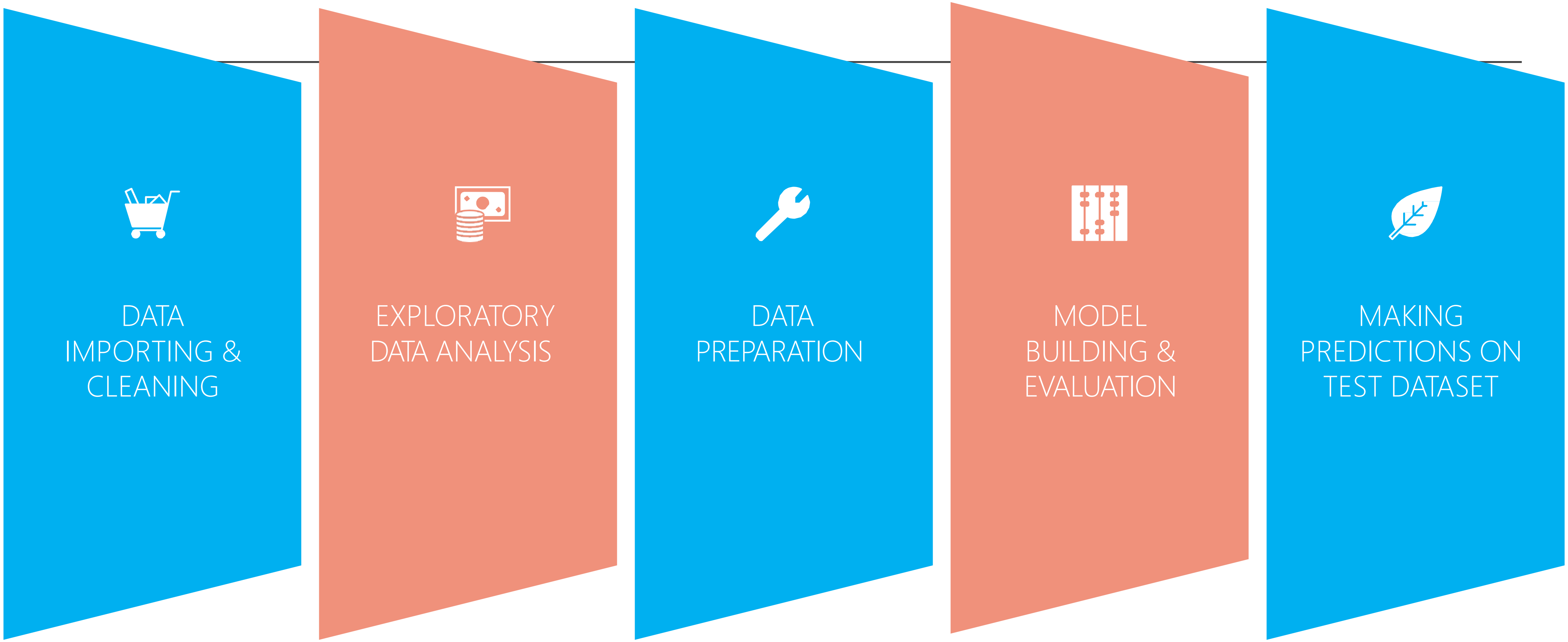
- The CEO want to achieve a lead conversion rate of 80%.

- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full manpower and after achieving target what should be the approaches.

Data understanding

The dataset comprises two files: 'Leads.csv' and 'Leads Data Dictionary.xlsx'. 'Leads.csv' contains approximately 9000 data points, with the target variable being 'Converted', indicating whether a past lead was converted (1) or not (0). The 'Leads Data Dictionary.xlsx' file serves as a data dictionary, offering explanations for the variables present in the 'Leads.csv' file

Steps of Analysis



DATA IMPORTING & CLEANING

EXPLORATORY DATA ANALYSIS

DATA PREPARATION

MODEL BUILDING & EVALUATION

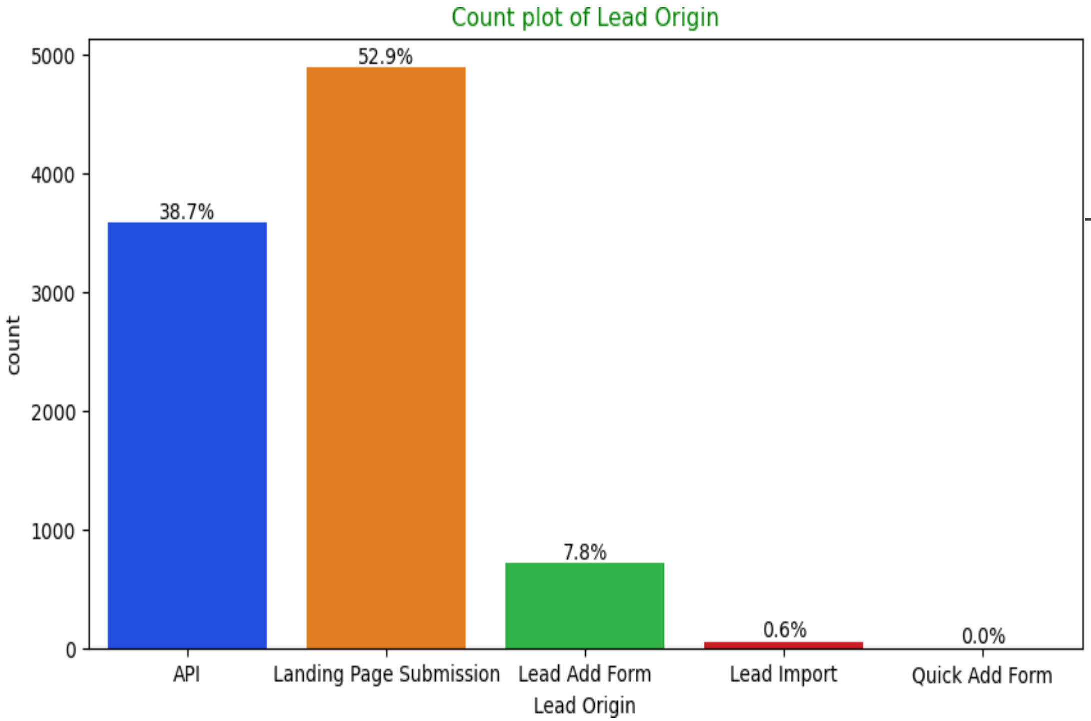
MAKING PREDICTIONS ON TEST DATASET

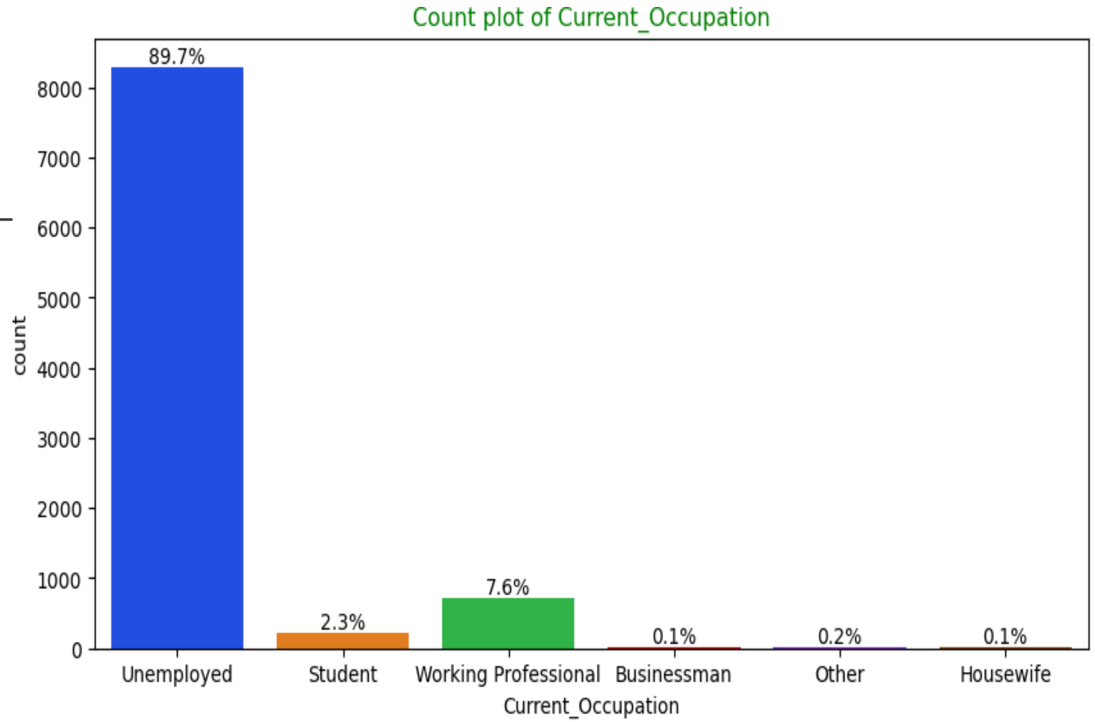
Data Cleaning

- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Imputation was used for some categorical variables.
- Numerical data was imputed with mode after checking distribution.
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in 'TotalVisits’, 'Total Time Spent on Website’ and 'Page Views Per Visit' were treated and capped.
- Low frequency values were grouped together to “Others”.
- Standardizing Data in columns by checking casing styles, etc. (“Lead Source” has Google and google)

EXPLORATORY DATA ANALYSIS

Univariate Analysis

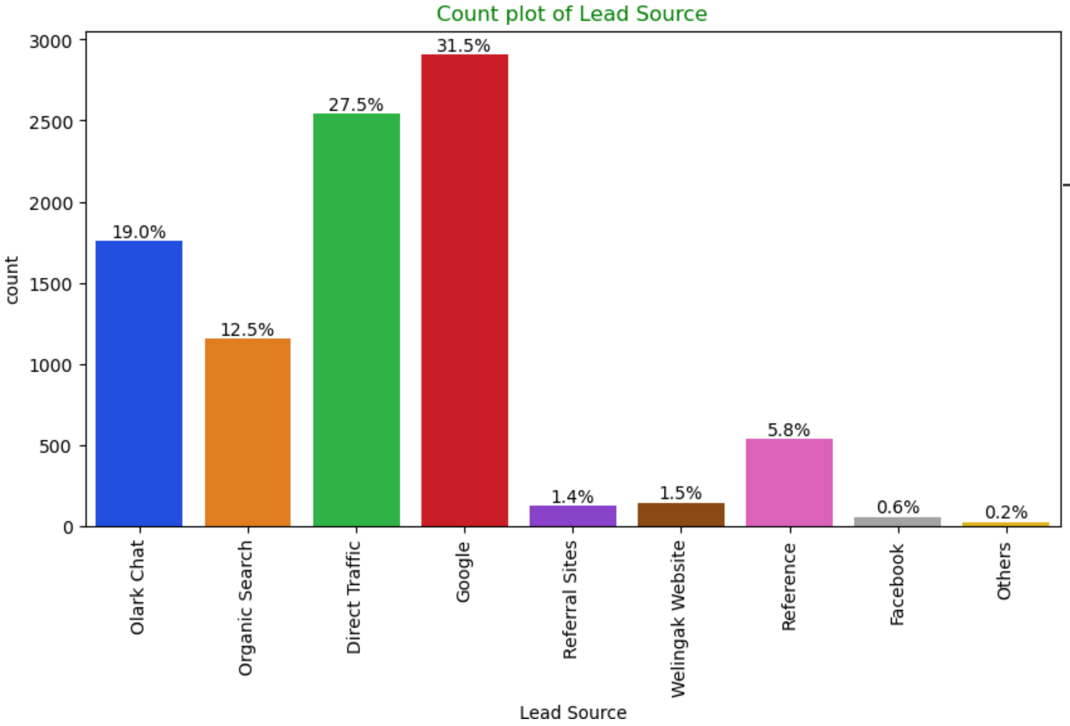


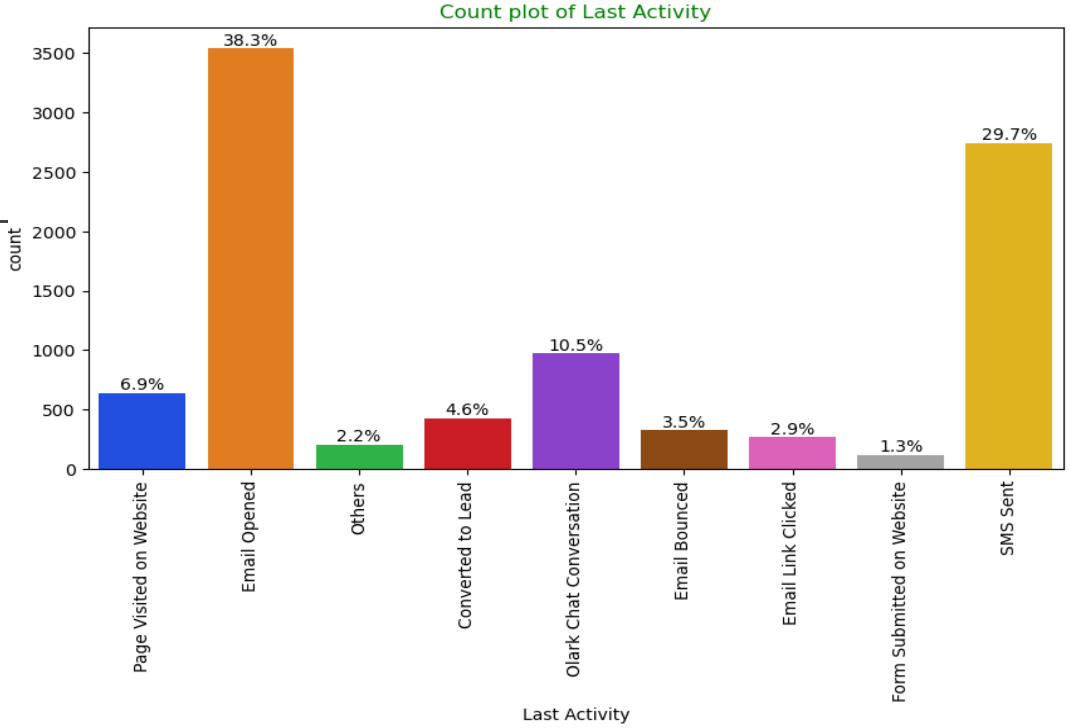


Inference:

- Lead Origin: 'Landing Page Submission' is the predominant lead origin, accounting for 52.9% of customers, followed by 'API' with 38.7%.
- Current_Occupation: The majority of customers, 89.7%, are classified as unemployed based on their current occupation

Univariate Analysis

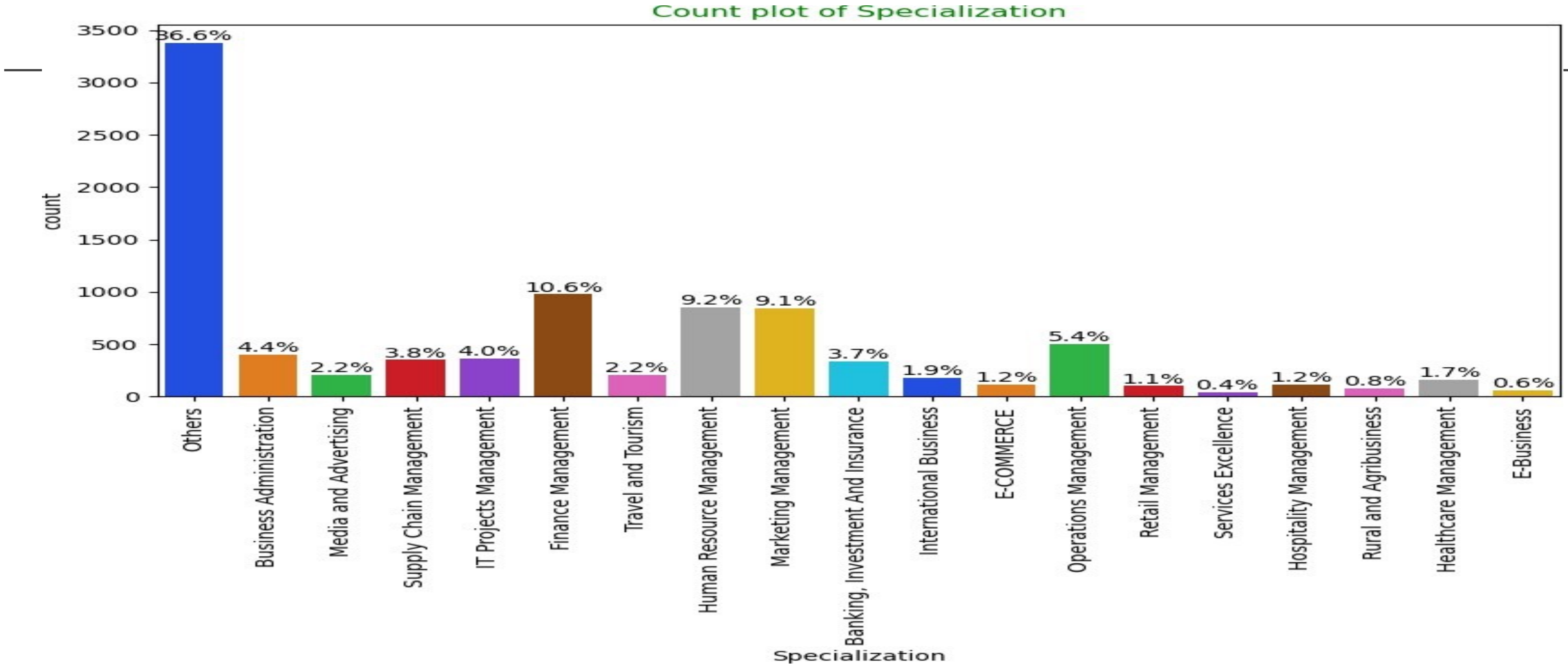




Inference:

- Lead Source: The primary lead source is Google at 31.5%, followed by Direct Traffic at 27.5%.
- 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.

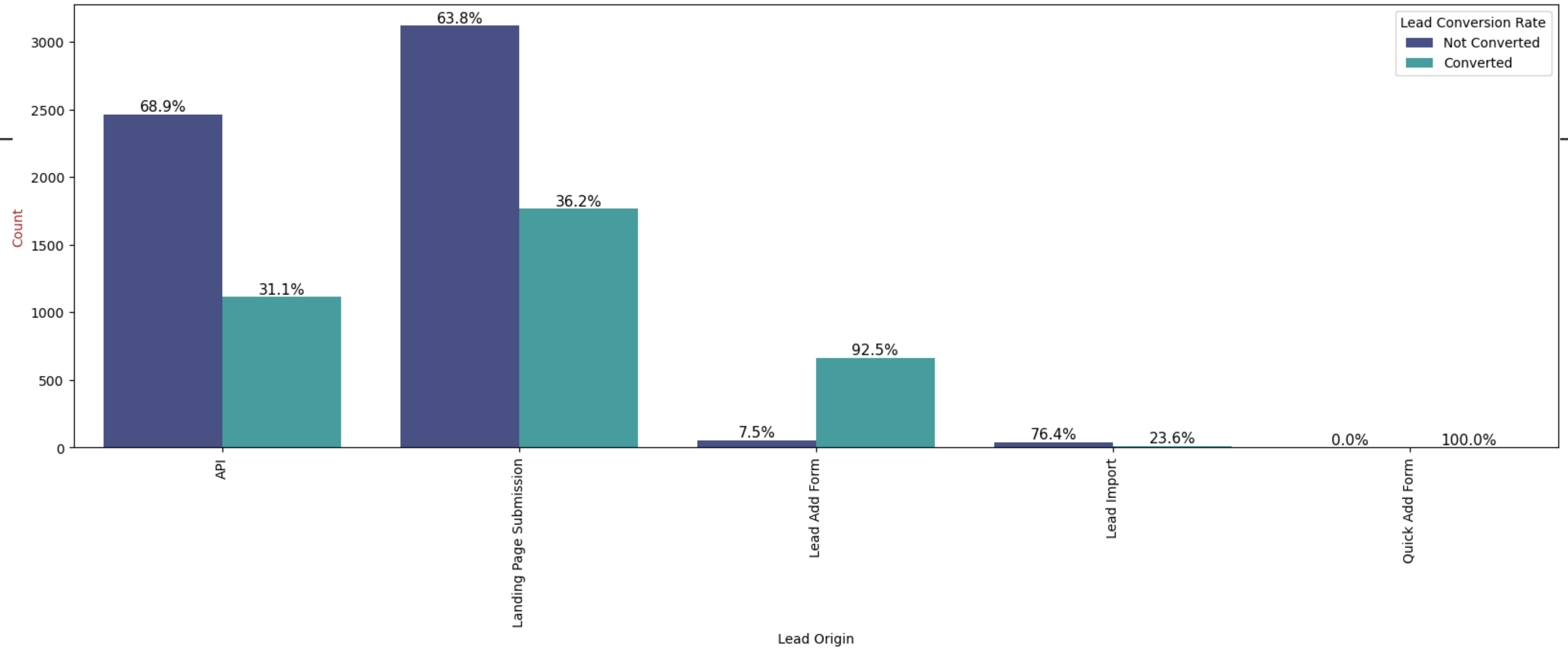
Univariate Analysis



- 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%.

Bivariate Analysis

Lead Conversion Rate of Lead Origin

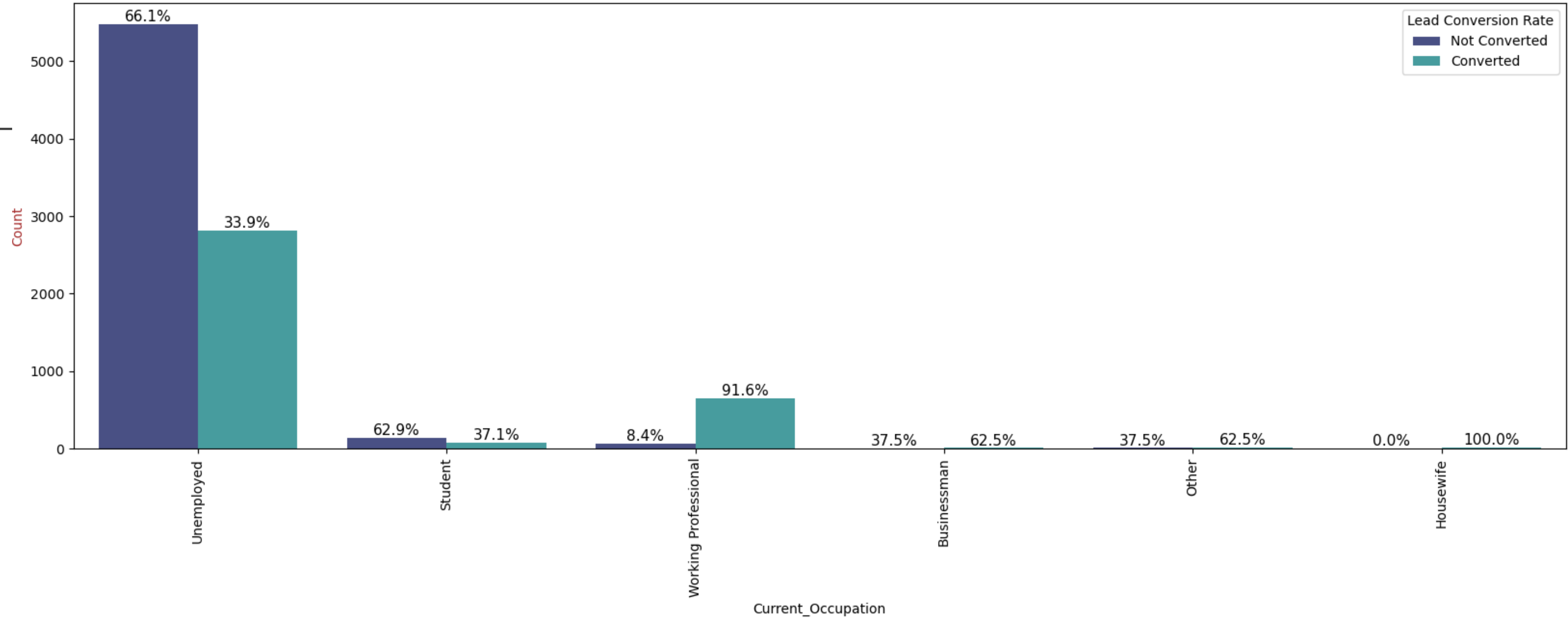


Inference:

- Lead Origin: 'Landing Page Submission' stands out as the most effective lead origin with a Lead Conversion Rate (LCR) of 36.2%, followed closely by 'API' at 31.1%.

Bivariate Analysis

Lead Conversion Rate of Current_Occupation

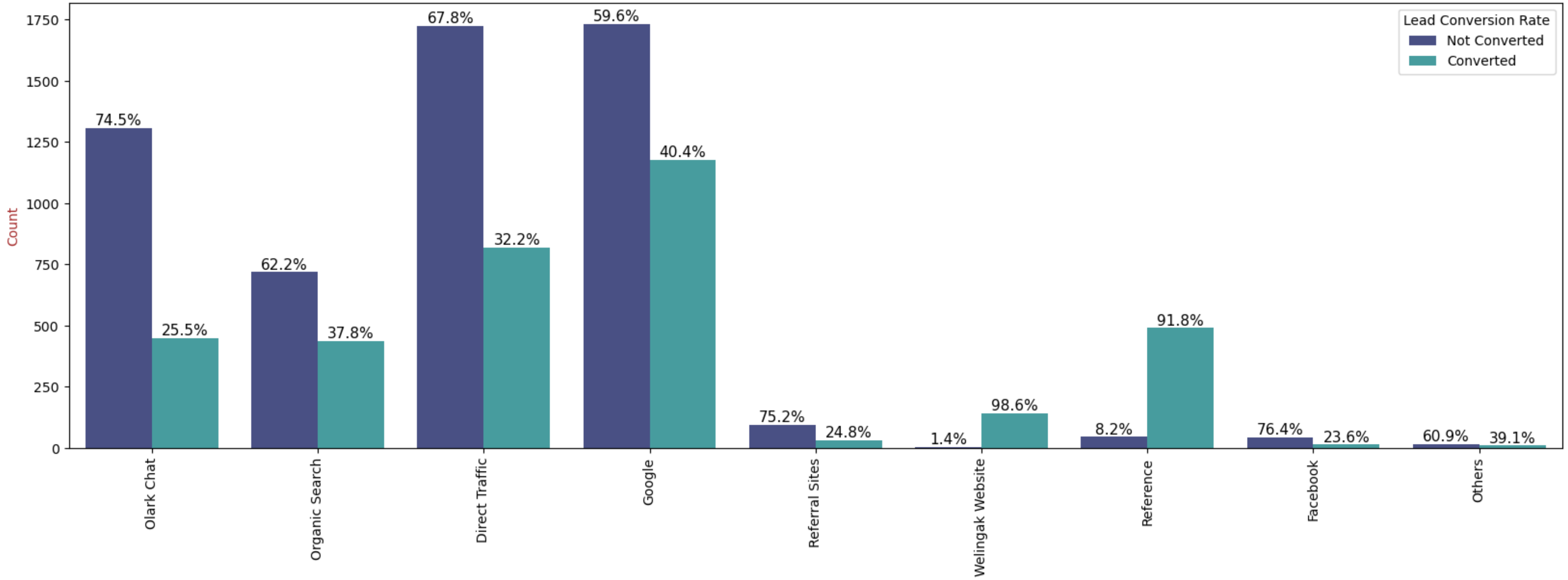


Inference:

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

Bivariate Analysis

Lead Conversion Rate of Lead Source

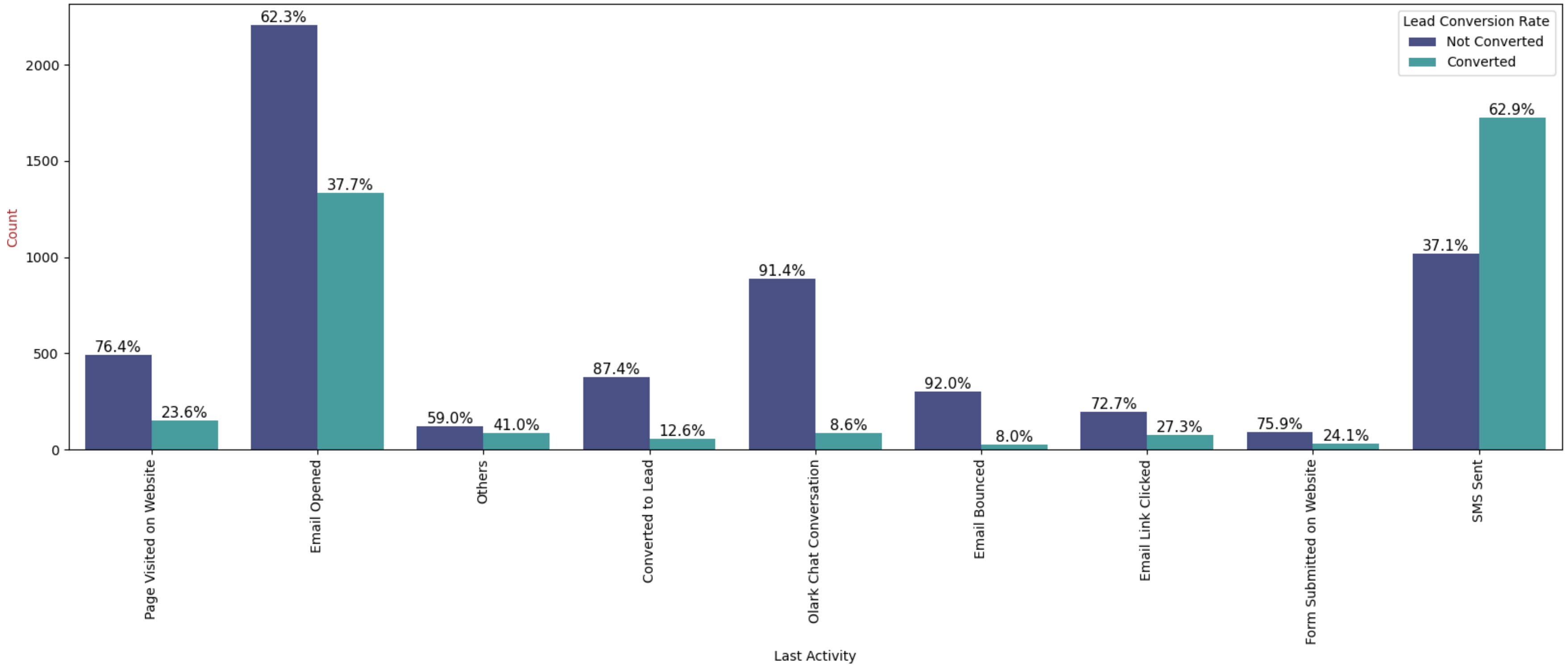


Inference:

- Lead Source: Google emerges as the most effective Lead Source, boasting an LCR of 40.4%. Direct Traffic and Organic Search closely follow with LCRs of 32.2% and 37.8%, respectively, accounting for a combined 12.5% of customers. While Reference holds the highest LCR at 91.8%, it only represents 5.8% of customers through this Lead Source.

Bivariate Analysis

Lead Conversion Rate of Last Activity

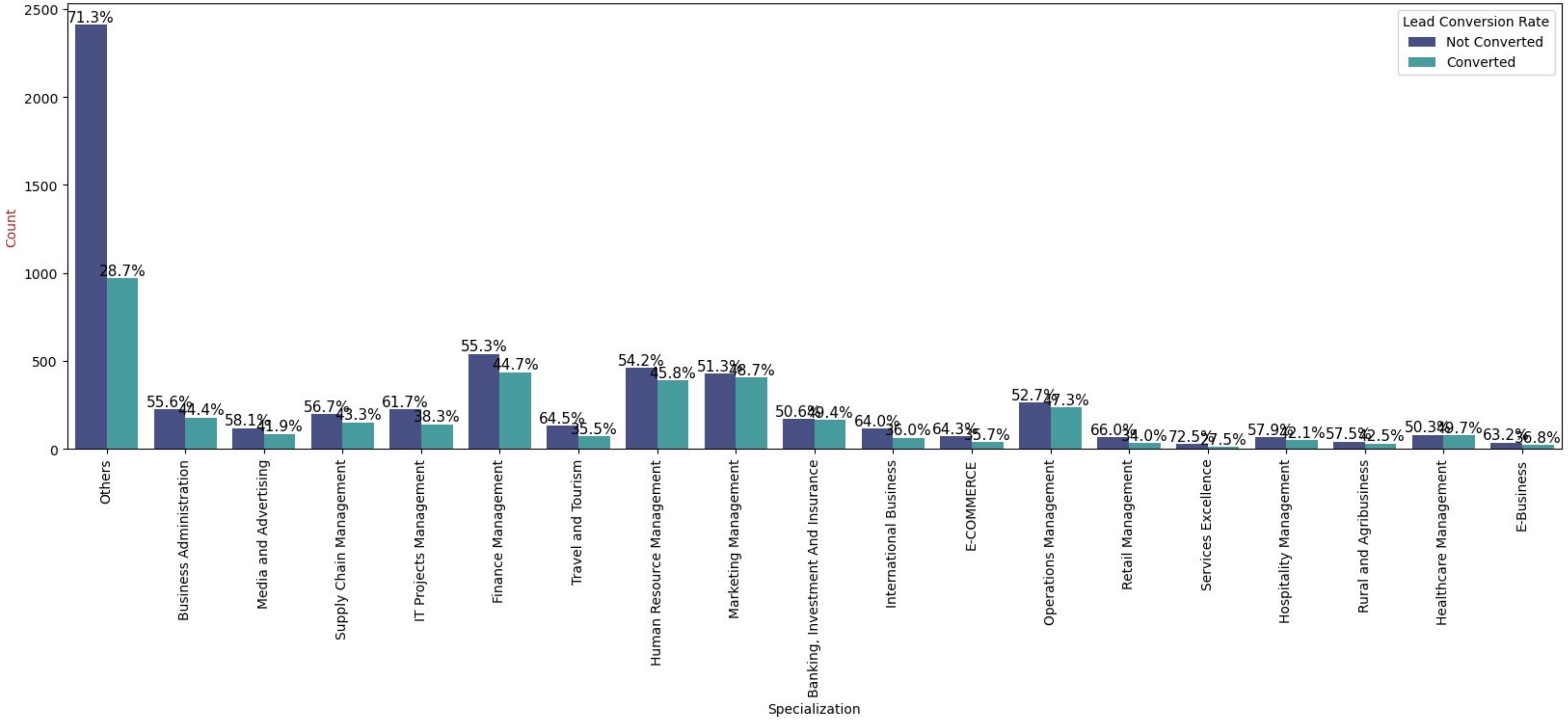


Inference:

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

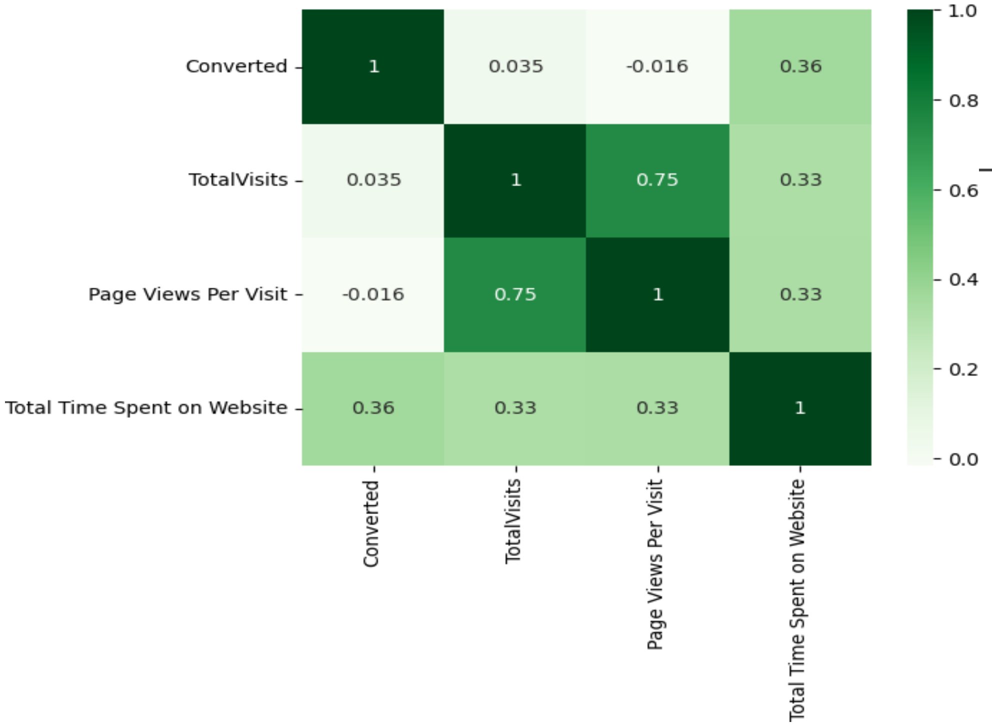
Bivariate Analysis

Lead Conversion Rate of Specialization

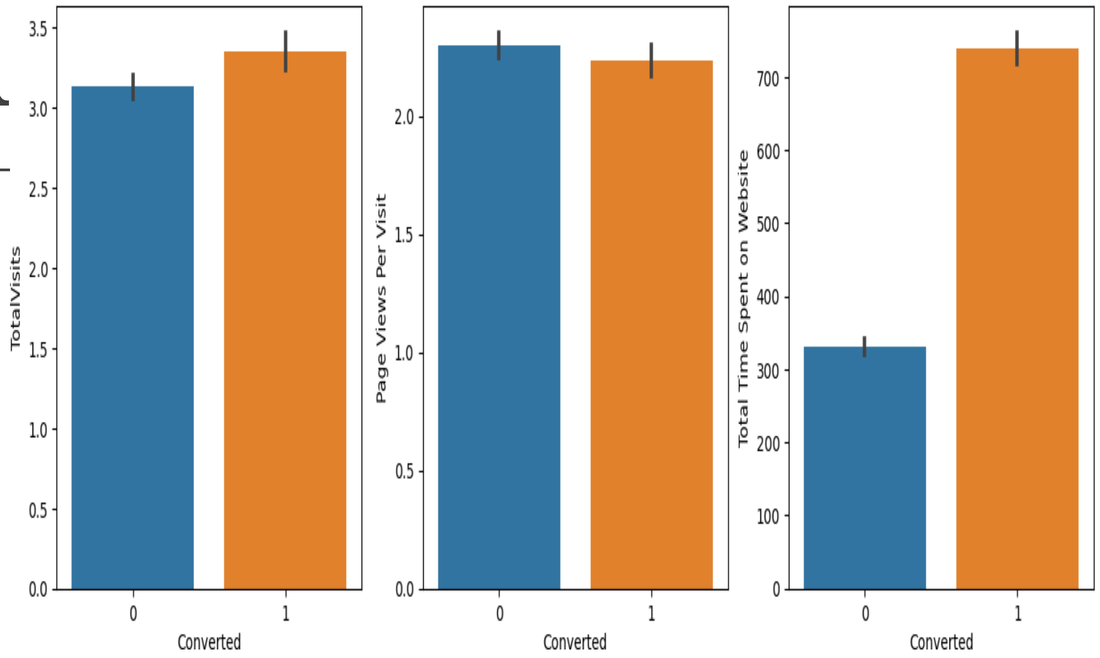


Inference:

- Specialization: Marketing Management, HR Management, Finance Management, and Operations Management exhibit promising Lead Conversion Rates (LCRs), signifying a significant interest among customers in these specific specializations.



A



Inference:

- 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

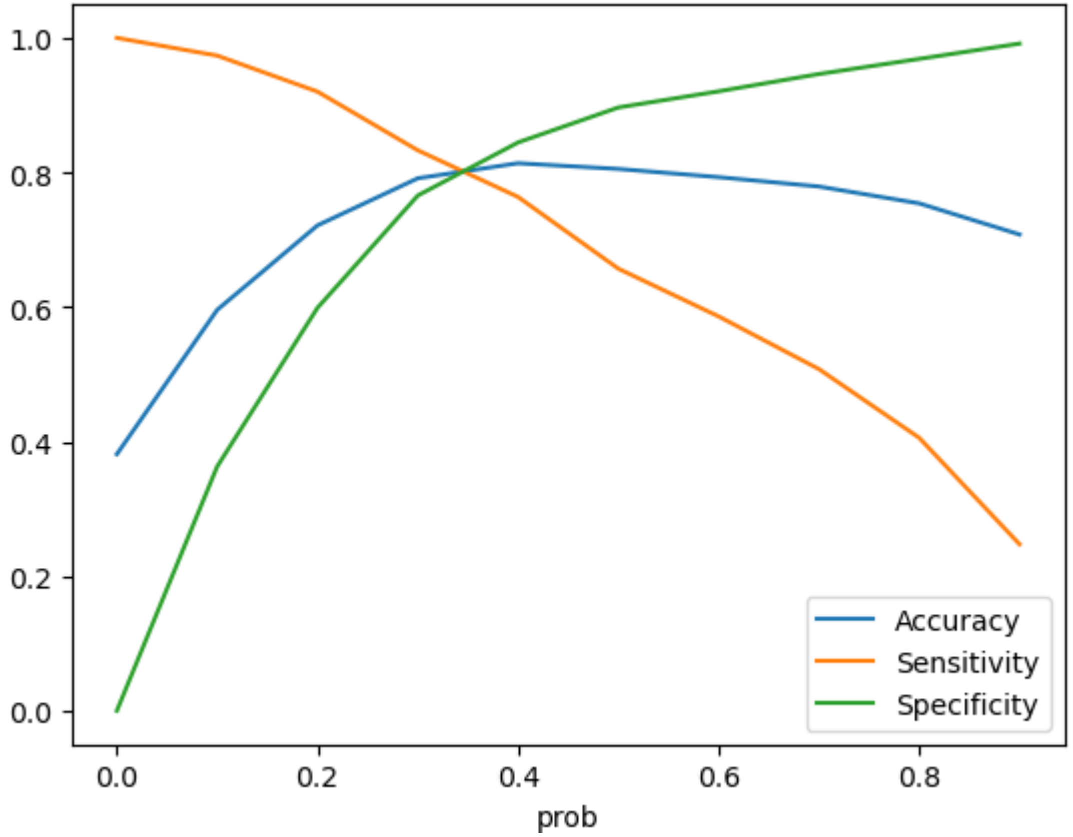
- Binary categorical columns were encoded as 1/0 to align with the logistic regression model requirements. Additionally, dummy features were generated for categorical variables like Lead Origin, Lead Source, Last Activity, Specialization, and Current_Occupation through one-hot encoding.
- The dataset was split into training and testing sets in a 70:30 ratio to train the model and assess its performance on unseen data. Feature scaling was implemented using standardization to ensure uniform scales across all features, preventing any particular feature from dominating the others.
- To address multicollinearity concerns, correlated predictor variables like Lead Origin_Lead Import and Lead Origin_Lead Add Form were removed from the dataset.

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

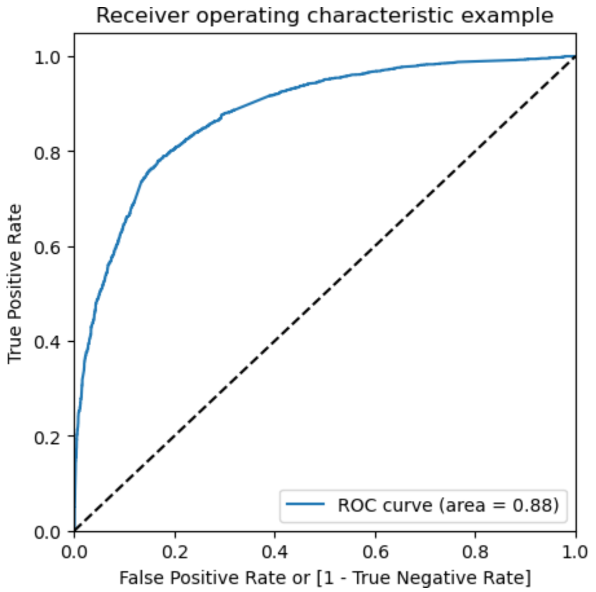
CONFUSION MATRIX - 1		
Actual/Predicted	not_converted	converted
not_converted	**3588**	**414**
converted	**846**	**1620**
Accuracy	0.8052	
Sensitivity	0.6569	
Specificity	0.8966	
False Positive Rate	0.1034	
Precision	0.7965	
Recall	0.6569	
Negative Predictive Value	0.8092	



Inference:

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

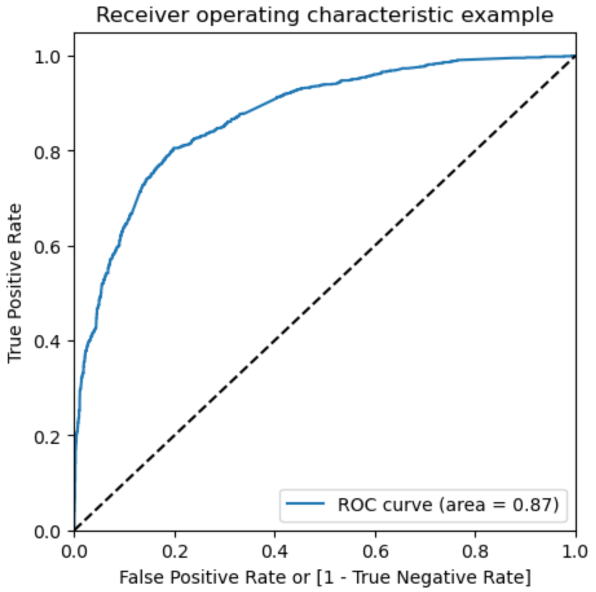
Making Predictions On Test Dataset



ROC Curve – Train Data Set

The Area under the ROC curve (AUC) was determined to be 0.88 out of 1, signifying that the model is a reliable predictor.

The ROC curve is positioned close to the top-left corner of the plot, indicating a high true positive rate and a low false positive rate across various threshold value



ROC Curve – Test Data Set

The Area under the ROC curve (AUC) was determined to be 0.87 out of 1, suggesting that the model is a reliable predictor.

The curve is positioned as close to the top-left corner of the plot as possible, indicating a high true positive rate and a low false positive rate across all threshold values.

CONCLUSION

- Prioritize features like 'Lead Source_Welingak Website', 'Current_Occupation_Working Professional', and 'Lead Source_Reference' in lead generation efforts due to their high conversion rates.
- Target working professionals aggressively, considering their higher conversion probability and potential better financial situations.
- Incentivize referral leads from existing customers with discounts or rewards to encourage more referrals.
- Increase the frequency of media usage, such as Google ads or email campaigns, to save time and boost the conversion rate.

CONCLUSION

- Focus on leads with 'Last Activity' as 'SMS Sent' or 'Email Opened,' as they tend to have higher conversion rates.
- Analyze the behavior of customers spending more time on the website to enhance the user experience and increase conversion rates.
- Tailor course offerings and marketing campaigns based on the popularity of specializations, providing targeted content for fields like Marketing Management and HR Management to attract and retain customers.

Thank You

Submitted by:
Minh Nguyen
Nikhil Agarwal
Neha Gaonkar