

Lead Scoring Case Study Presentation

Submitted By
Amrita
Upadhyay

Afzal
Ahmed

Agam
Bhatnagar

Business Objective

In this project, our main objective is to develop a logistic regression model for X Education that assigns a lead score ranging from 0 to 100 to each potential lead. This lead score is a crucial tool for prioritizing leads based on their probability of conversion. By setting a target conversion rate of 80%, we aim to assist X Education in efficiently identifying and focusing on high-potential leads. Additionally, we will provide solutions to the company's existing challenges and offer recommendations on leveraging the lead scoring model effectively. The flexibility of the model allows it to adapt to any future changes in the company's requirements, ensuring its continued relevance and utility.

Business Understanding

- ▶ X Education operates as an online education provider, catering to professionals seeking industry-specific courses. The company employs diverse online channels, including popular search engines like Google, to market its courses effectively. Interested potential customers explore the X Education website, where they have the opportunity to peruse the available courses.
- ▶ X Education's dedicated sales team initiates contact with these leads through either phone calls or emails, aiming to successfully convert them into paying customers.
- ▶ X Education typically achieves a lead conversion rate of around 30%, reflecting the company's ongoing efforts to engage and convert potential customers into active participants in their educational programs.

Steps performed

- ▶ Reading & understanding the dataset
- ▶ Cleaning the data
- ▶ Performed EDA
- ▶ Feature scaling
- ▶ Splitting the data into test & train data
- ▶ Prepare the data for modelling
- ▶ Model building
- ▶ Model evaluation-specificity & Sensitivity or precision recall
- ▶ Making predictions on the test
- ▶ Assigning lead score
- ▶ Hot leads Determination
- ▶ Feature Importance Determination

Data Cleaning

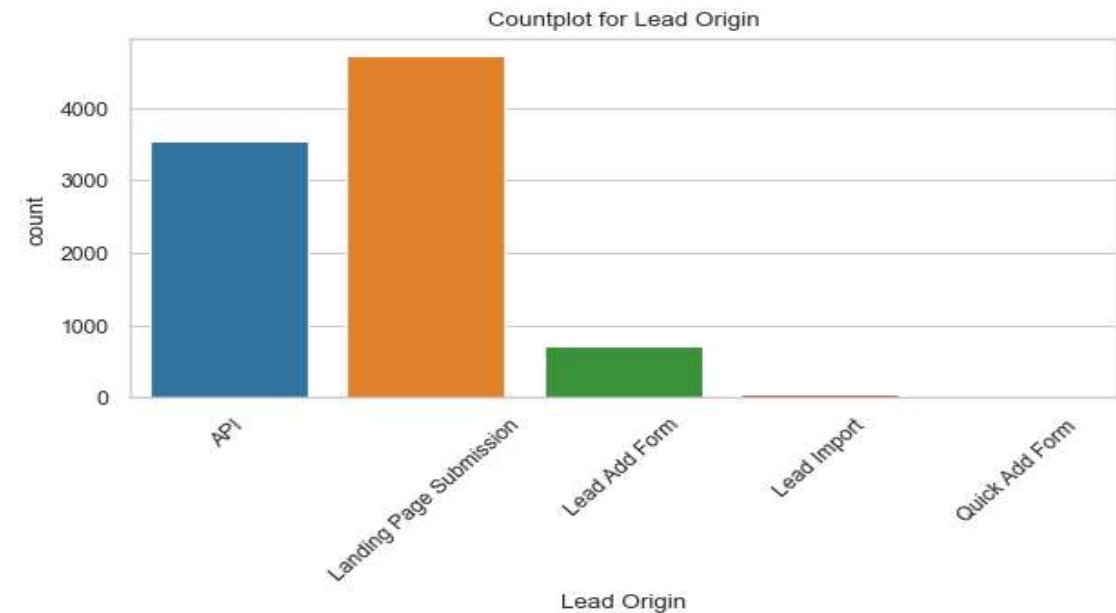
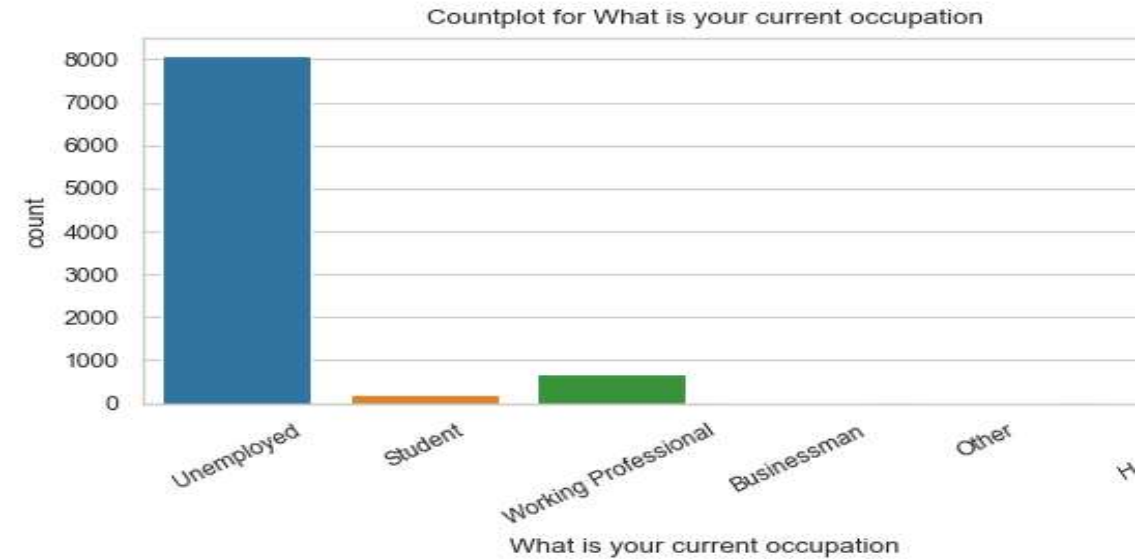
- Checked the datatypes and null values.
- There were few columns which were having a high number of missing values which are dropped right away.
- Subsequently, a strategic approach was taken to address columns that, are important for the model building , had a chances of containing null values. Followed Imputation techniques, utilizing mean and median values based on the variable types (continuous or categorical).
- Also columns with no use were dropped before further analysis
- Moreover, outliers detected during the analysis underwent removal to enhance the data's overall quality. Following these processes, a substantial 98% of the data was retained. The subsequent analysis was conducted on this refined dataset, ensuring a comprehensive exploration of insights and patterns.

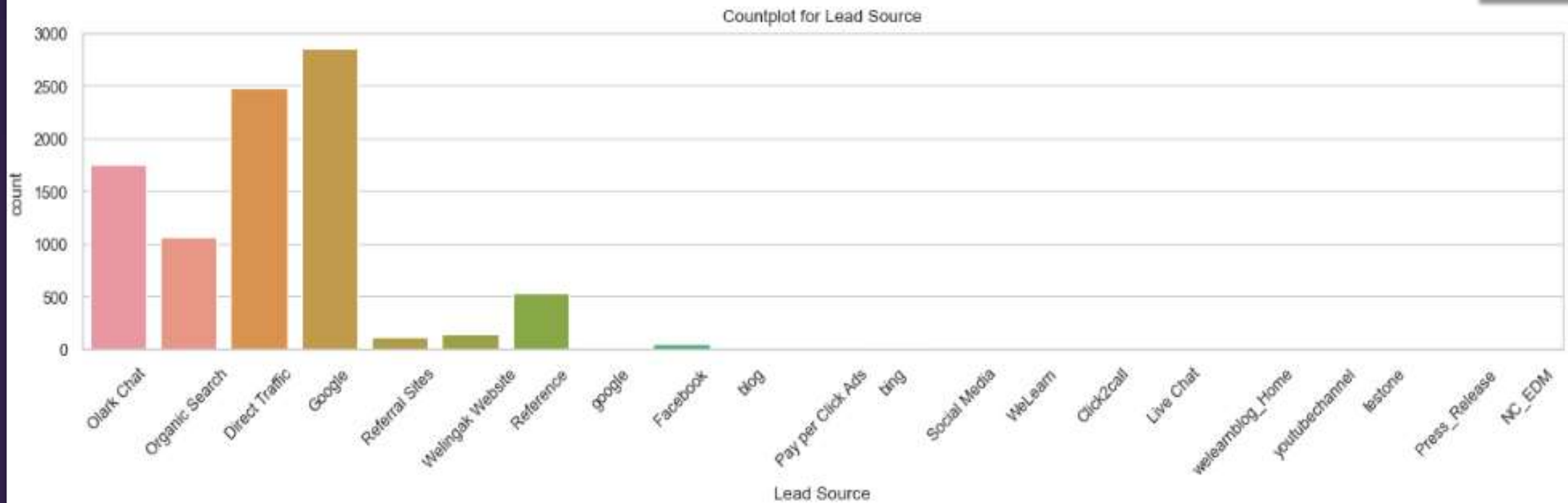
EXPLORATORY DATA ANALYSIS

- EDA was performed after cleaning the data by plotting different types of graphs and analysing both continuous and categorical columns .
- Univariate, bivariate analysis was done against the target variable for better understanding. It analysed and we can see which factor affected the lead and helped in conversion
- The lead origin attributed to Landing Page Submission exhibits the highest conversion rate compared to other origins. Also majority of students were unemployed which is also a significant factor.
- Google stands out with the highest conversion rate among various channels.
- Leads with the Last Activity recorded as "SMS sent" demonstrate the most favorable conversion rates.
- Leads associated with Specialization, particularly those categorized as unknown or labeled as "Select," exhibit the highest conversion rates.
- Working professionals display a notably higher conversion rate in comparison to other occupation categories.

UNIVARIATE ANALYSIS

- ▶ What's your Current_Occupation : A significant proportion of customers, are unemployed based on the current occupation information
- ▶ Lead Origin :The majority of customers, identified through 'Landing Page Submission' as the lead origin, followed by API

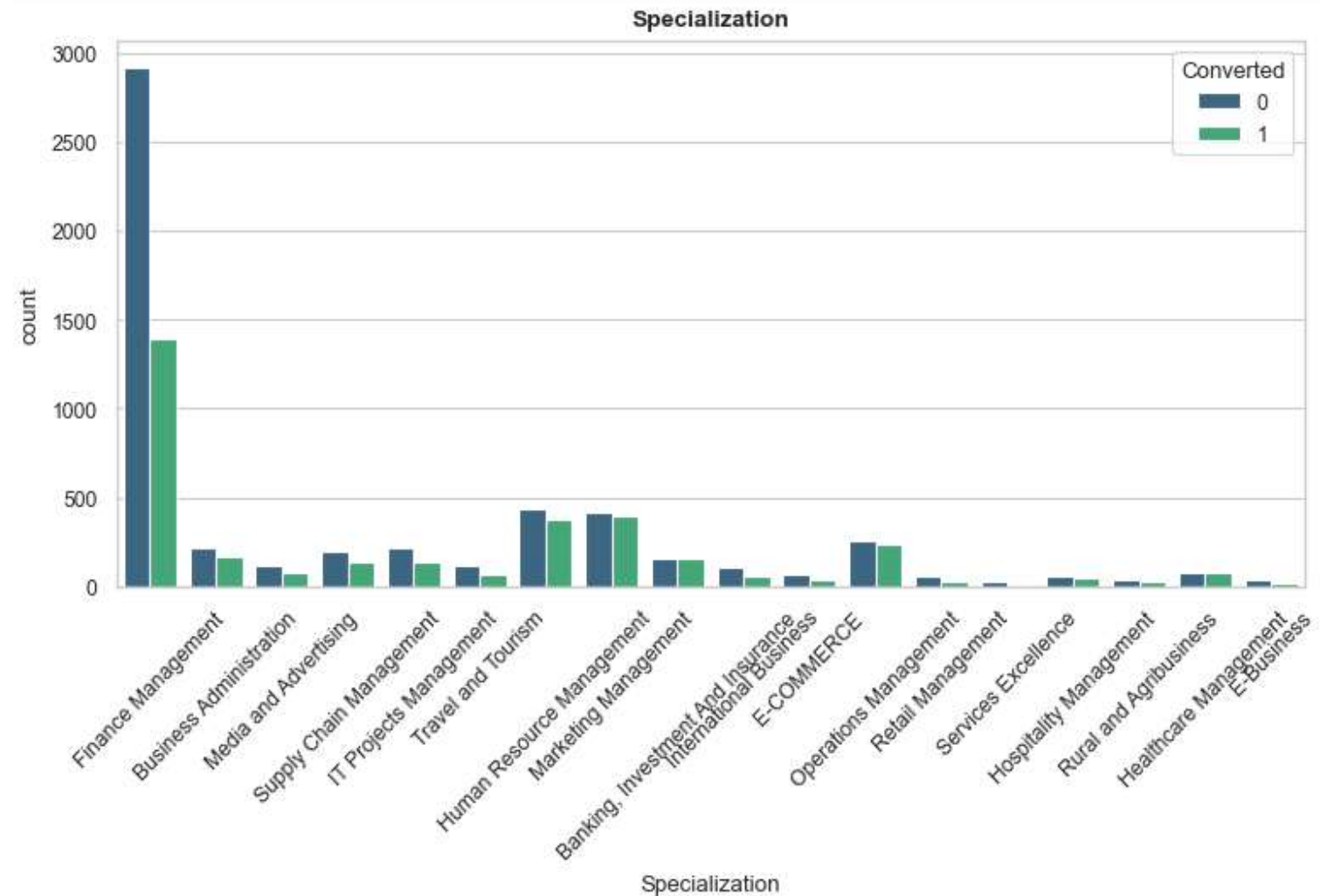




Lead Source: The primary lead source is Google , followed by Direct Traffic .

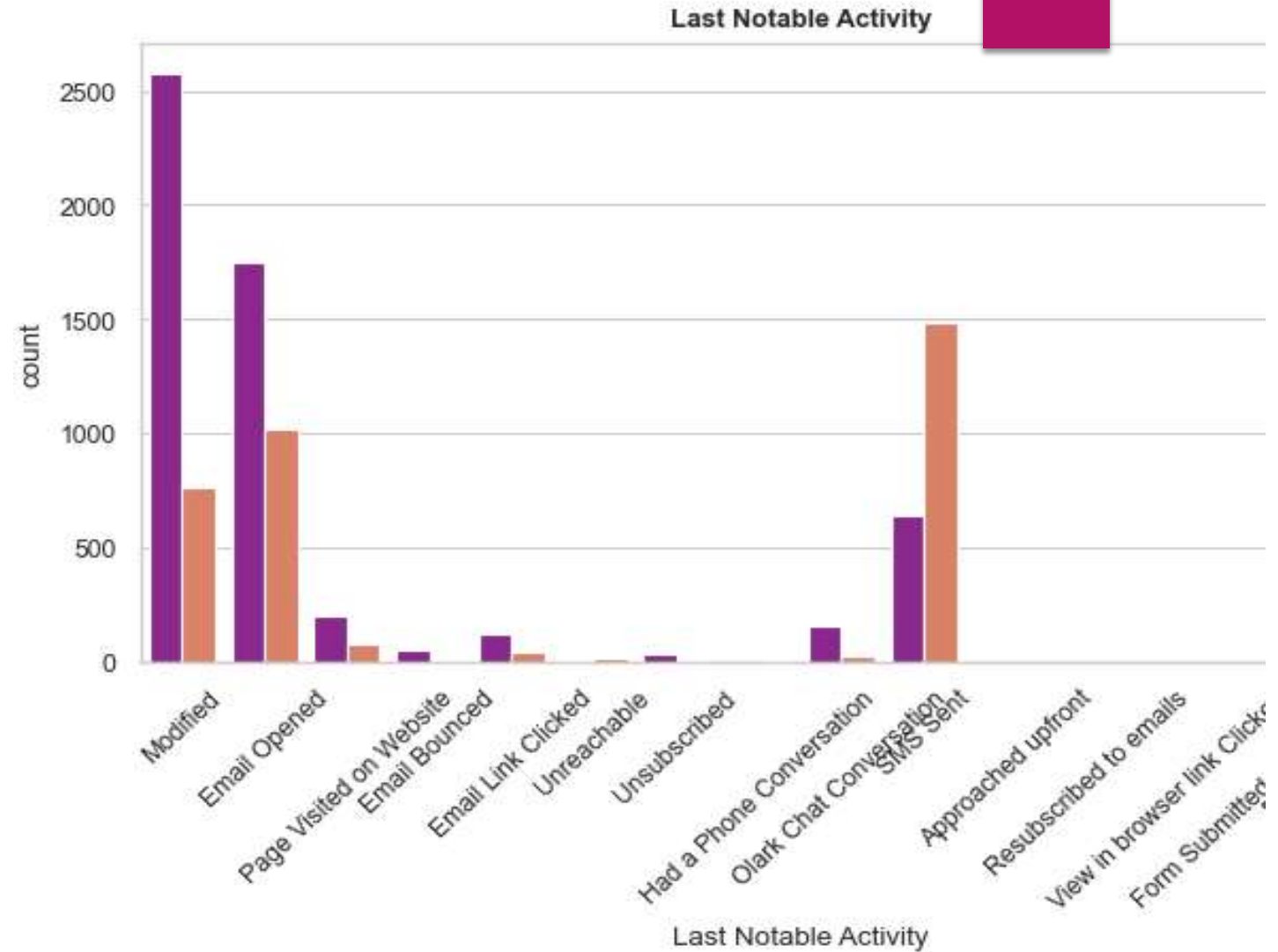
BIVARIATE ANALYSIS

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



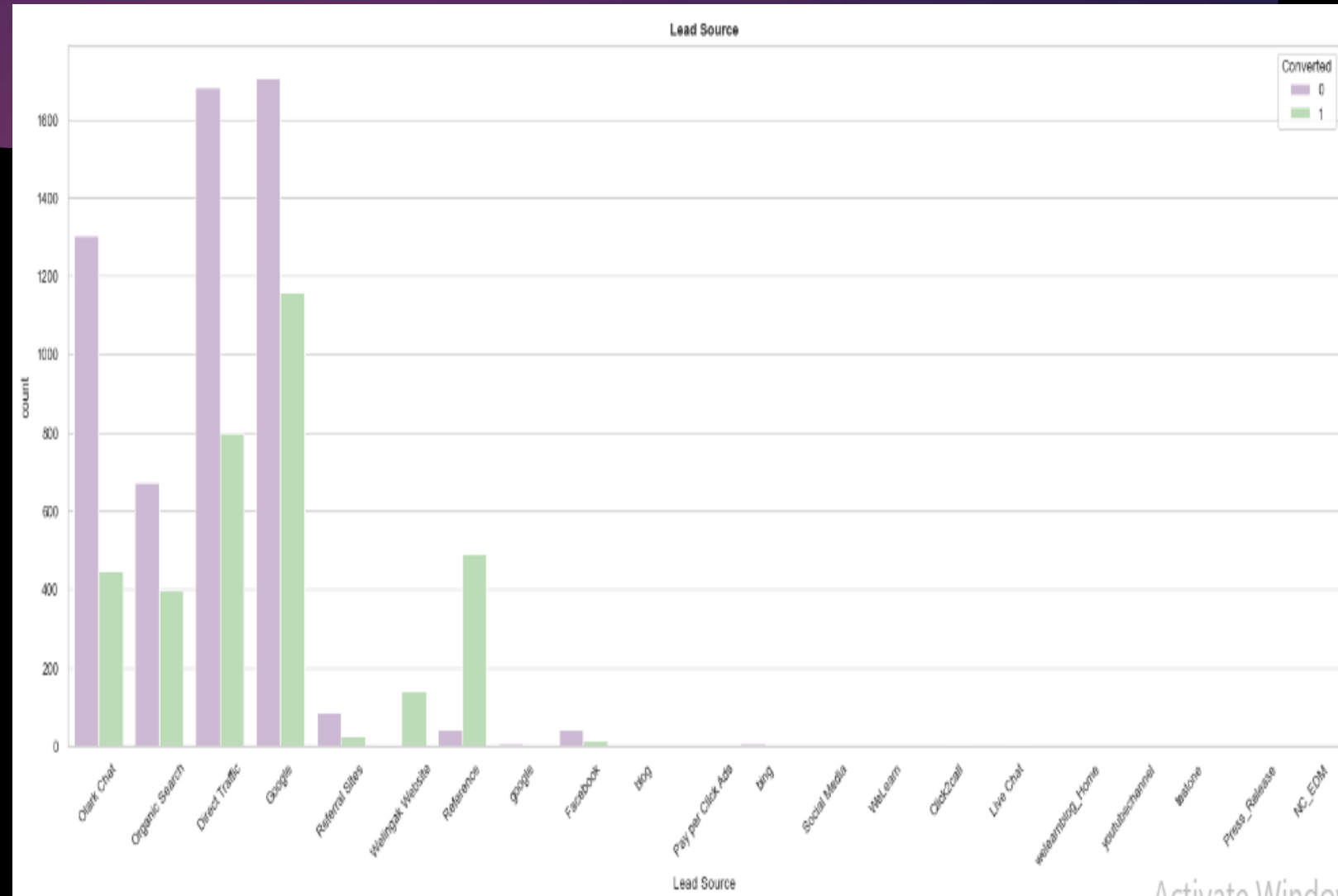
Bivariate Analysis

► Last Notable Activity: SMS Sent and Email Opened are the most effective Last Activity types with high conversion rate.



Lead Source

- Lead Source: Google is the most effective Lead Source with followed by Direct Traffic and Organic Search, Reference also have high conversion rate.





CORRELATION ANALYSIS

Correlation Analysis

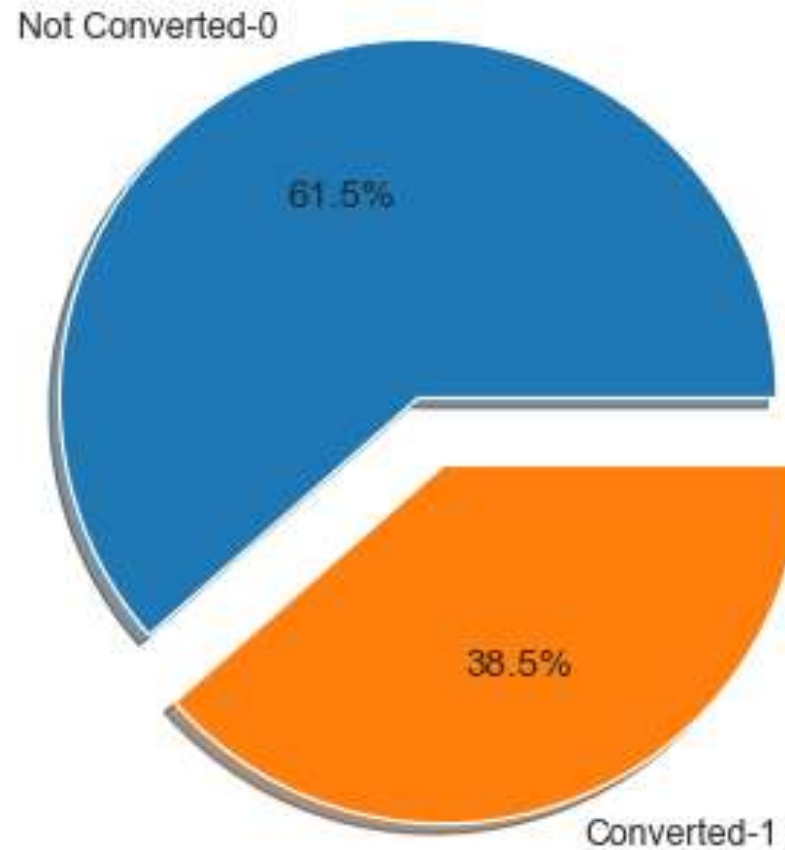
- There is a strong positive correlation between 'Total Visits' and 'Page Views per Visit', indicating
- that customers who visit the website more frequently tend to view more pages per visit.




Data Imbalance Analysis

- ▶ As per observation, the data is not overly unbalanced. In the data, 61.5% of leads are not converted, whereas remaining 38.5% of leads are converted.

Data Imbalance analysis





DATA PREPARATION & Model Building

Data Preparation & Model building

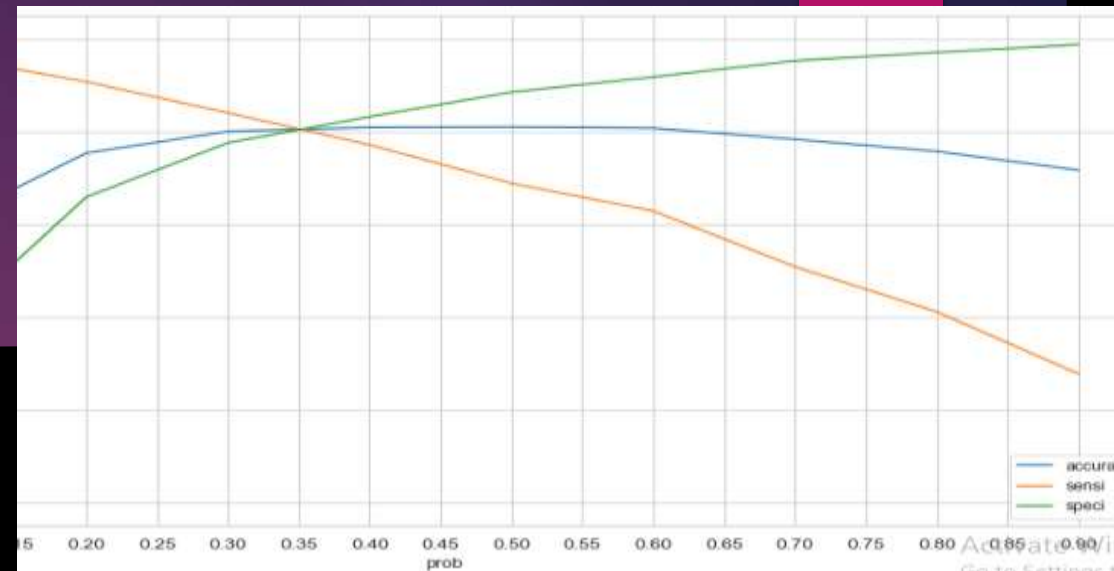
- ▶ Categorical columns at the binary level were changed into a 1/0 format in the next steps, aligning them with the logistic regression model requirements. Additionally, categorical variables like Lead Origin, Lead Source, Last Activity, Specialization, and Current Occupation underwent one-hot encoding, creating dummy features.
- ▶ To facilitate model training and assessment, the dataset was split into training and testing sets in a 70:30 ratio. This allowed the model to be trained on a portion of the data and evaluated on unseen data to gauge its generalization performance.
- ▶ To ensure uniformity in feature scales and prevent any single feature from dominating others, feature scaling was applied using the standardization method.
- ▶ Feature selection was applied using RFE technique and then the elimination was done according the steps followed for fetching the column having high P value & VIF, a final model was obtained after occurrence of 3 times until both VIF and p values reached under acceptable range.



MODEL EVALUATION

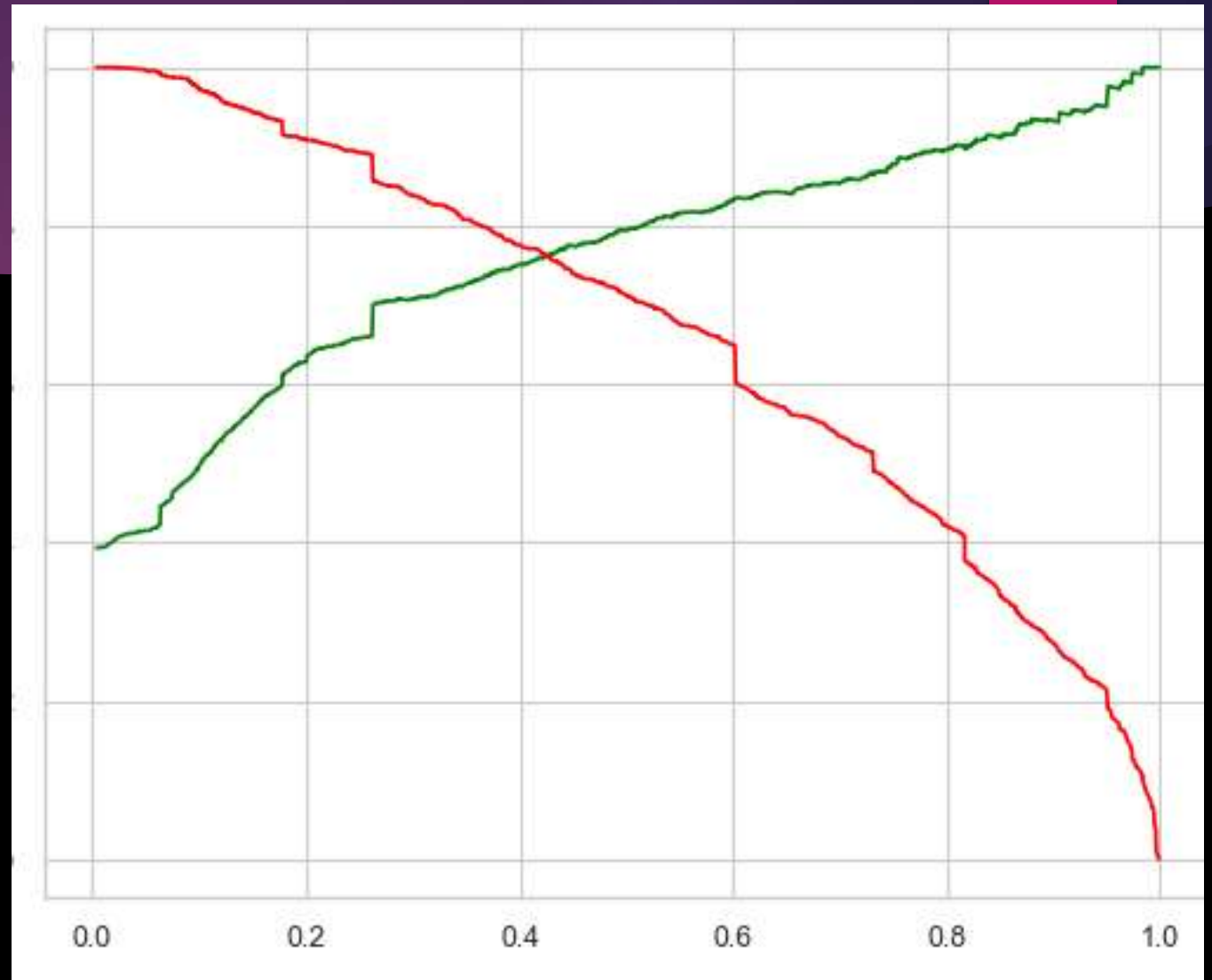
Model Evaluation(Train)

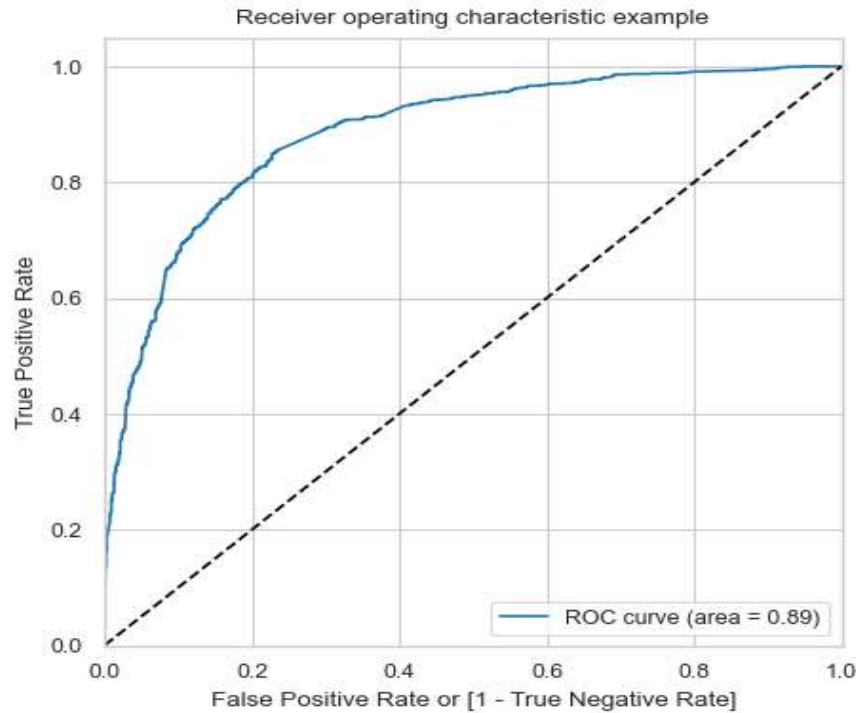
- ▶ Accuracy : 80.36%
- ▶ Sensitivity/Recall : 81.66%
- ▶ Specificity : 79.55%
- ▶ Precision/Positive predictive value: 71.13%
- ▶ Negative predictive value: 87.55%



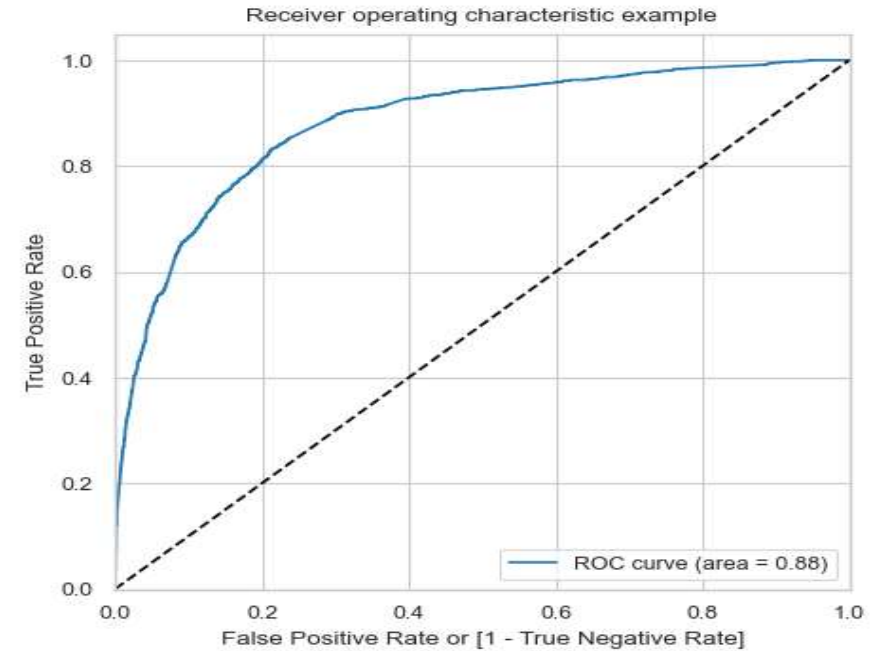
Model Evaluation(Test)

- Accuracy : 80.354%
- Sensitivity/Recall : 81.89%
- Specificity : 79.67%
- Precision/Positive predictive value: 72.32%





- ▶ ROC curve for train set
- ▶ The Area under ROC curve was found to be 0.89 out of 1, indicating that the model is a good predictor.



- Roc curve for test set
- The Area under ROC curve was found to be 0.88 out of 1, indicating that the model is a good predictor.

Lead score for pred train set

Prospect Id	Converted	Converted_Prob	final_Predicted	final_predicted	Lead_Score
0	0	0.18	NaN	0.00	18
1	0	0.31	0.00	NaN	31
2	1	0.67	1.00	NaN	67
3	0	0.09	NaN	0.00	9
4	1	0.50	1.00	NaN	50

- ▶ Every lead in the Original Dataframe has their Lead Score calculated
- ▶ The customers with a high lead score have a higher chance of conversion and low lead score have
- ▶ a lower chance of conversion.

Relative importance of feature

- ▶ Each feature's relative importance is calculated on a scale of 100, with 100 being the highest score.
- ▶ Features with High Positive Values are those that have a major impact on a lead's likelihood of conversion.
- ▶ In a similar way, traits with high negative values make the less contribution.

	index	0
2	Total Time Spent on Website	100.00
3	Lead Origin_Lead Add Form	87.88
4	What is your current occupation_Working Profes...	57.28
6	Lead Source_Welingak Website	48.88
5	Lead Source_Olark Chat	32.28
1	TotalVisits	25.44
7	Last Activity_Olark Chat Conversation	-25.50
9	Last Notable Activity_Email Opened	-31.80
0	Do Not Email	-34.90
11	Last Notable Activity_Olark Chat Conversation	-38.87
8	Last Notable Activity_Email Link Clicked	-39.41

Conclusion

- ▶ 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional' and 'Total Time Spent' are effective factors that contribute to a good conversion rate.
- ▶ Working professionals and Unemployed customers tend to have higher conversion rates.
- ▶ Referral leads generated by old customers have a significantly higher conversion rate
- ▶ Google and Direct Traffic are channels that are showing good conversion rates.
- ▶ Leads whose 'Last Activity' is 'SMS Sent' or 'Email Opened' tend to have a higher conversion rate.
- ▶ The 'Others' specialization category is the most common among customers followed by Finance Management, HR Management and Marketing Management.

RECOMMENDATION

- ▶ As per analysis 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 tend 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.
- ▶ 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. By focusing on popular specializations like Marketing Management and HR Management, the ability to offer targeted content and resources is enhanced.



Thank You