



Lead Scoring - Case Summary

- SANGEETHA
- AISHWARYA
- SURENDER REDDY

Introduction

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

Business and Data Understanding

- ❑ X Education is an online education company that offers courses to industry professionals. The company promotes its courses through various online channels, including search engines like Google.
- ❑ Prospective customers who are interested in the courses visit the X Education website and browse through the available courses.
- ❑ Some of these visitors may fill out a form on the website with their email address or phone number to express interest in the courses. These visitors are classified as leads.
- ❑ X Education's sales team contacts the leads via phone or email to try and convert them into paying customers.
- ❑ While some leads do get converted into paying customers, the majority do not. X Education's typical lead conversion rate is approximately 30%.
- ❑ The 'Leads csv' file contains around 9000 data points. The target variable of the dataset is the column ' which indicates whether a past lead was converted or not. The values in the ' 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.

Phases

DATA IMPORTING &
CLEANING

EXPLORATORY DATA
ANALYSIS

DATA PREPARATION

MODEL BUILDING &
EVALUATION

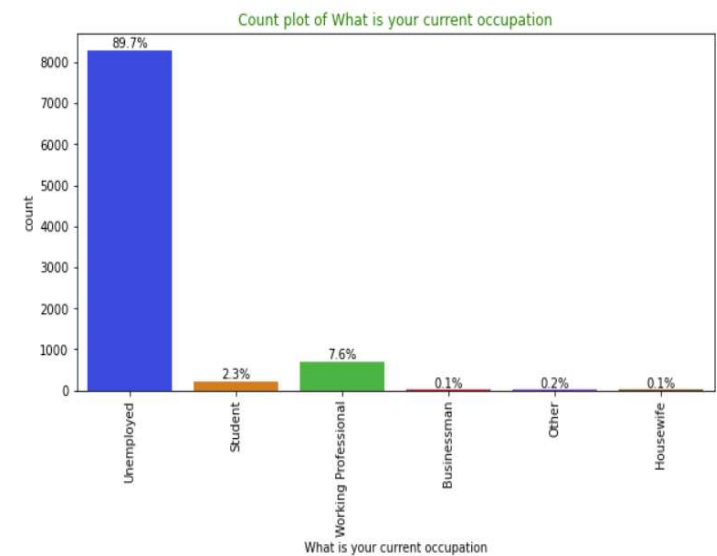
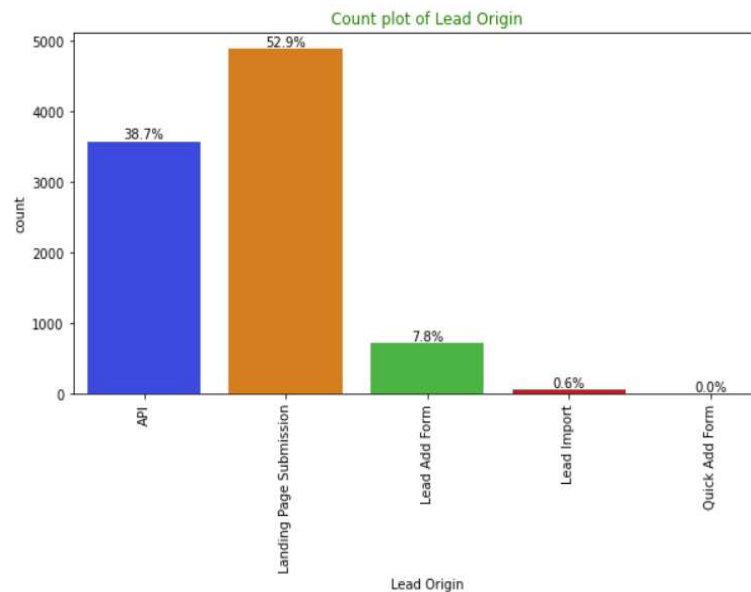
PREDICTIONS ON DATASET

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.
- Drop columns that don't add any insight or value to the study objective (' City',' Tags',' Country', 'What matters most to you in choosing a course').
- Imputation was used for some categorical variables.
- Columns with no use for modeling ('Prospect ID','Lead Number' and 'Last Notable Activity') were dropped.
- 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 ' Total Visits', '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).

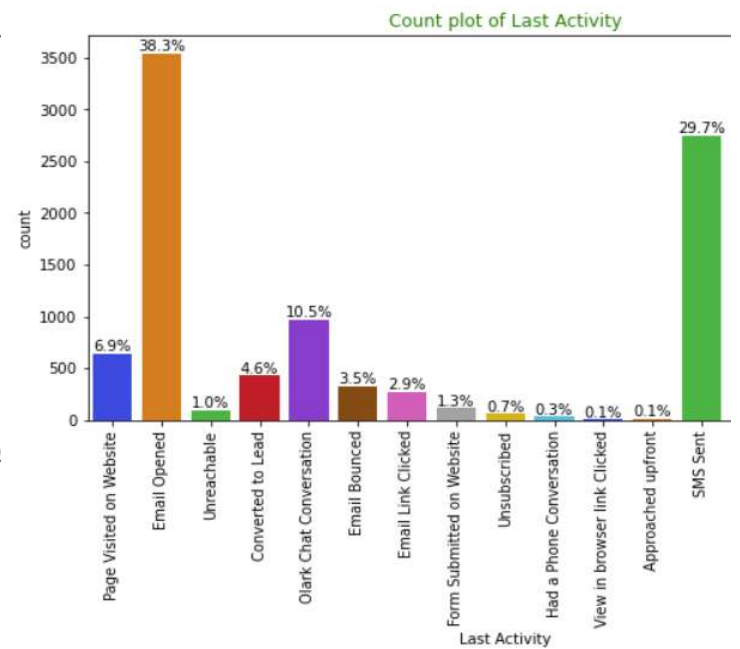
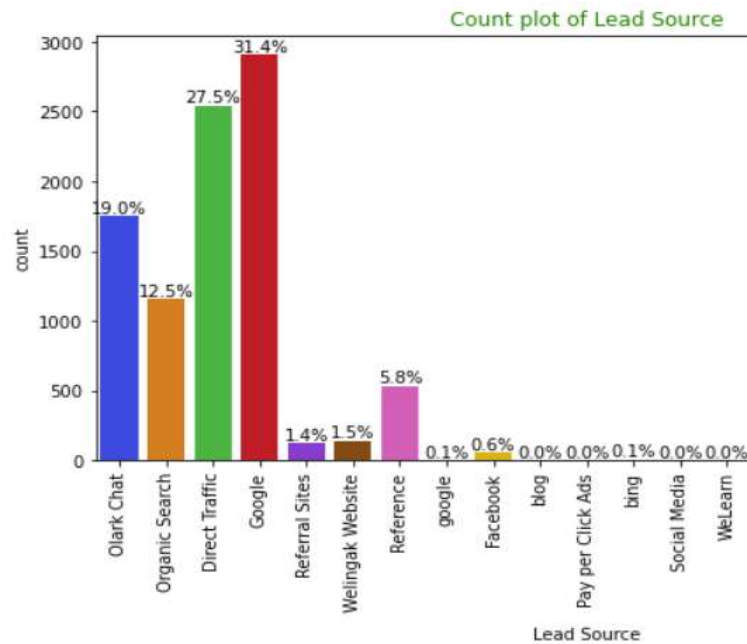
EDA - Univariate Analysis

- ▶ Lead Origin: Majority of customers, 52.9 % were identified through 'Landing Page Submission' as the lead origin, followed by 'API' at 38.7 %.
- ▶ Current_Occupation: A significant proportion of customers, 89.7 % are unemployed based on the current occupation information.



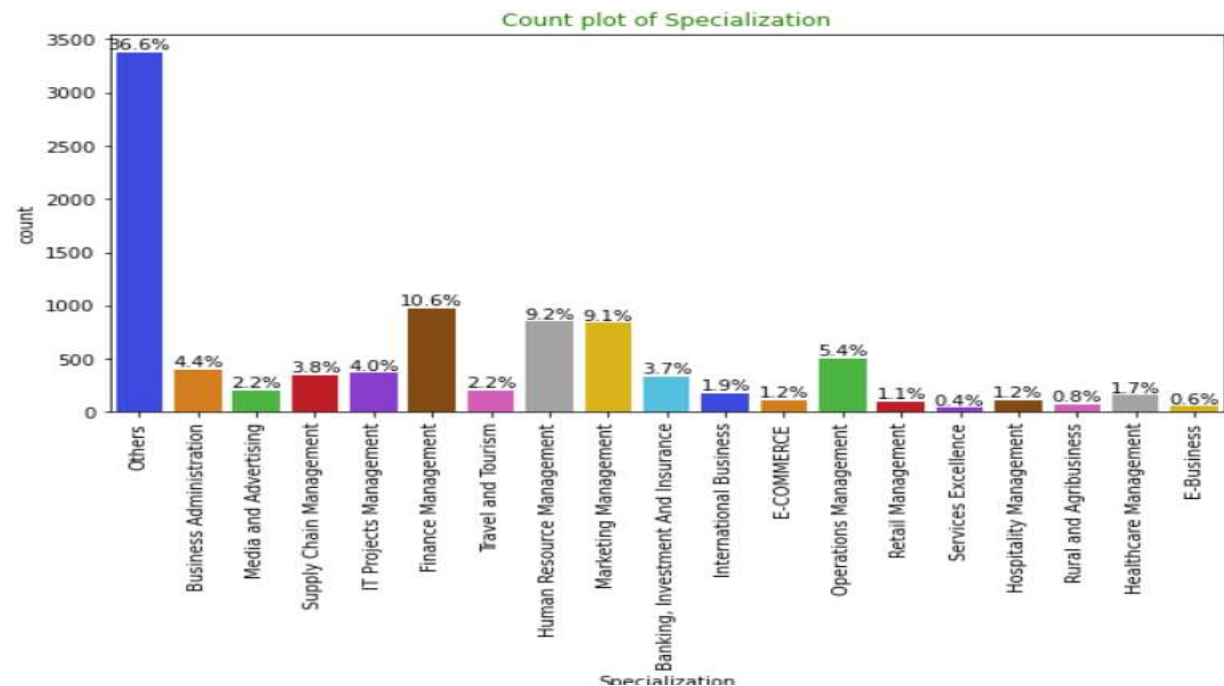
EDA - Univariate Analysis

- ▶ 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.



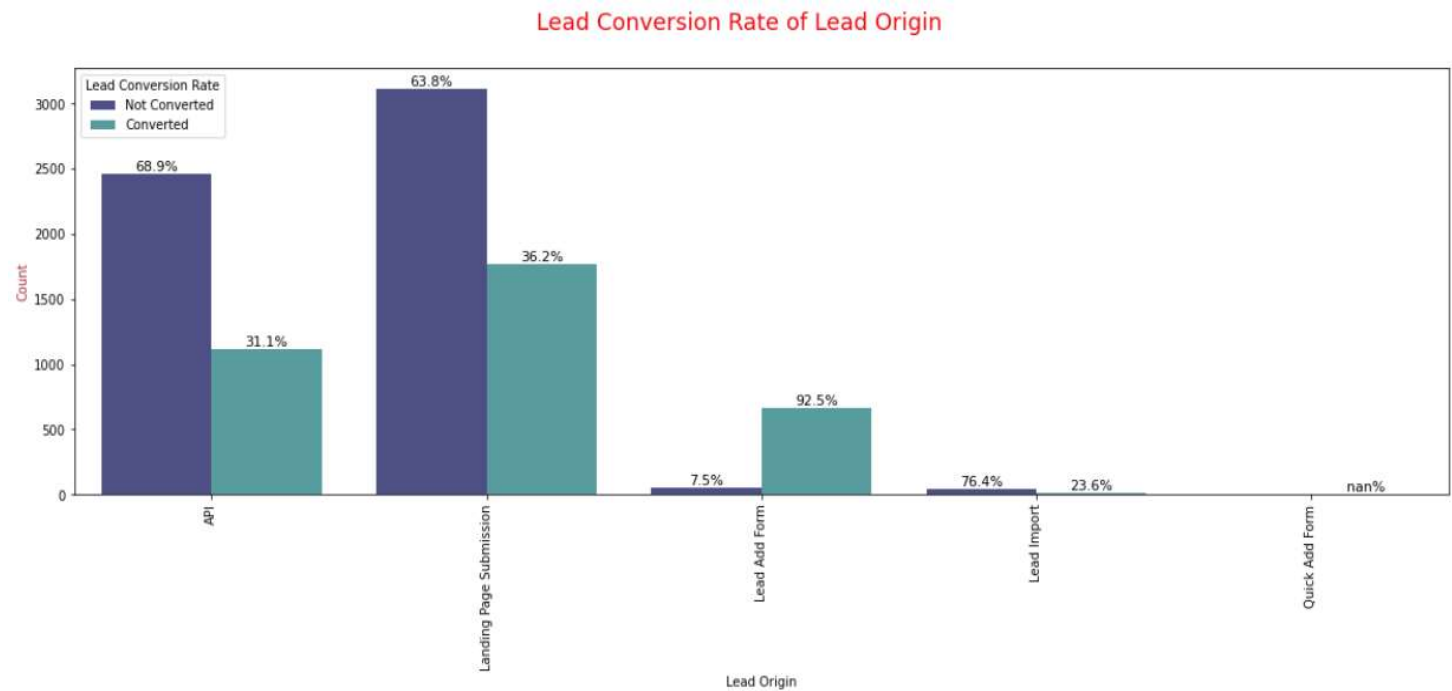
EDA - Univariate Analysis

- 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%.



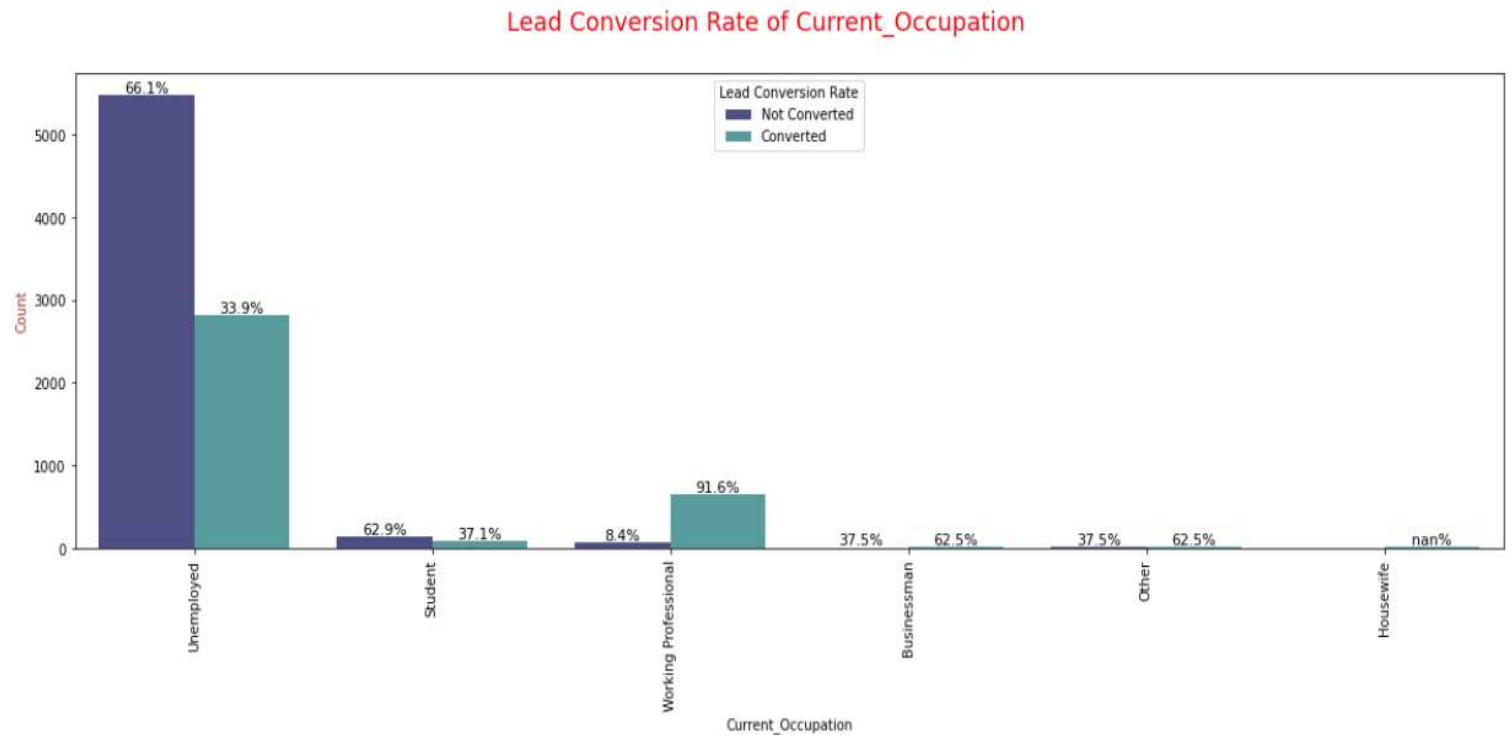
EDA - Bivariate Analysis

- 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 %.



EDA - Univariate Analysis

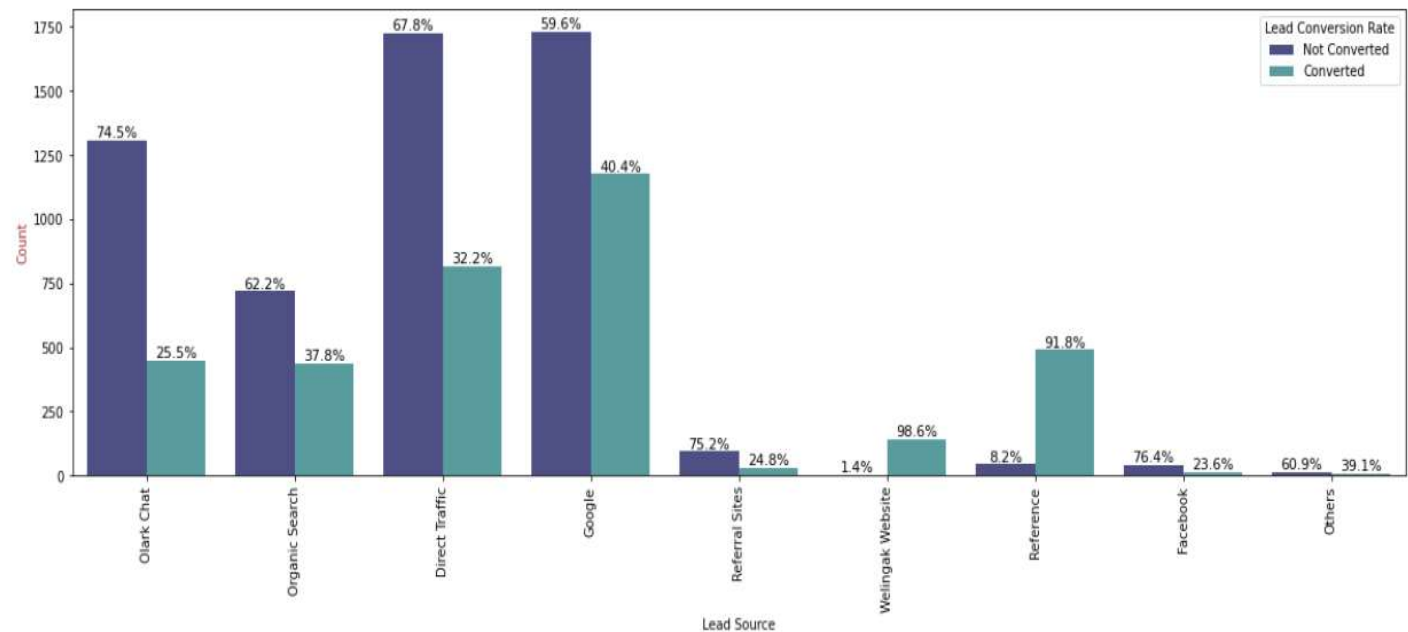
- Current_Occupation :
Working Professionals
have a significantly
higher LCR at 91.6 %
compared to
Unemployed people at
33.9 % .



EDA - Univariate Analysis

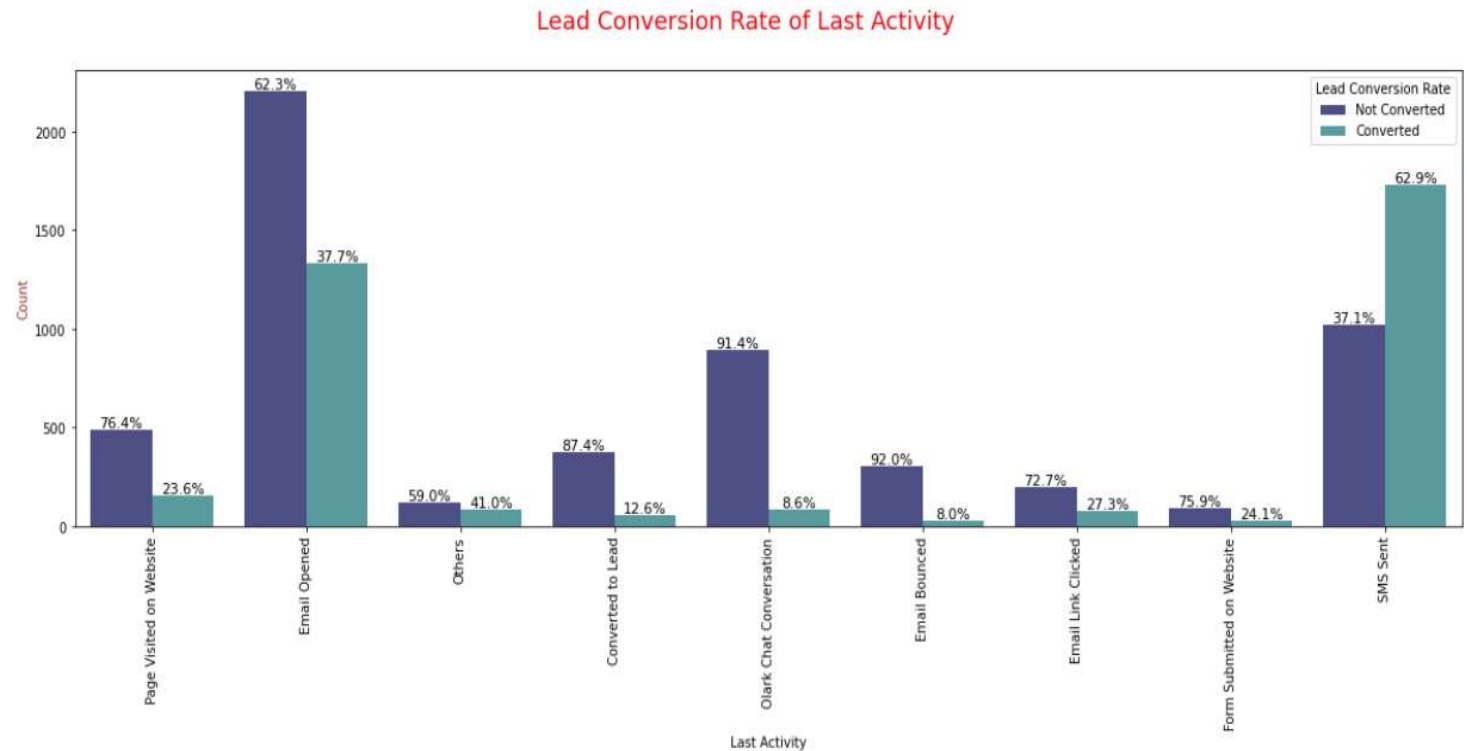
- Lead Source : Google is the most effective Lead Source with an LCR of 40.4 % followed by Direct Traffic at 32.2 % and Organic Search at 37.8 % (contributing to only 12.5 % of customers) Reference has the highest LCR at 91.8 % but there are only 5.8 % of customers through this Lead Source.

Lead Conversion Rate of Lead Source



EDA - Univariate Analysis

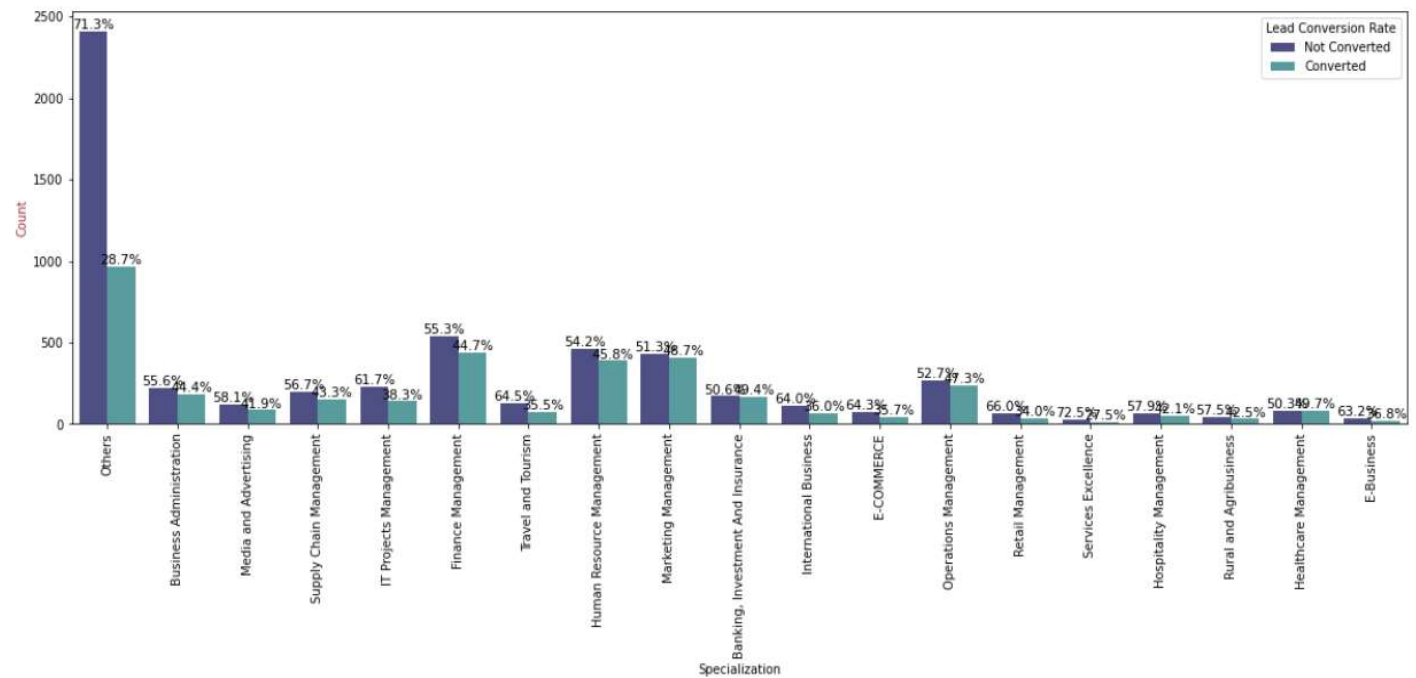
- Last Activity: SMS Sent and Email Opened are the most effective Last Activity types with LCRs of 62.9 % and 37.7 % respectively.



EDA - Univariate Analysis

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

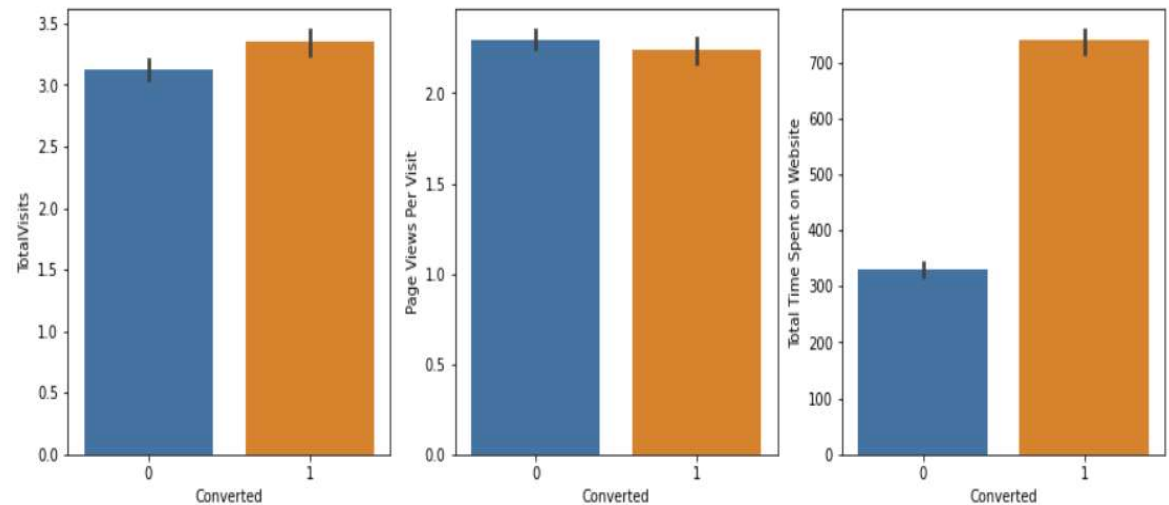
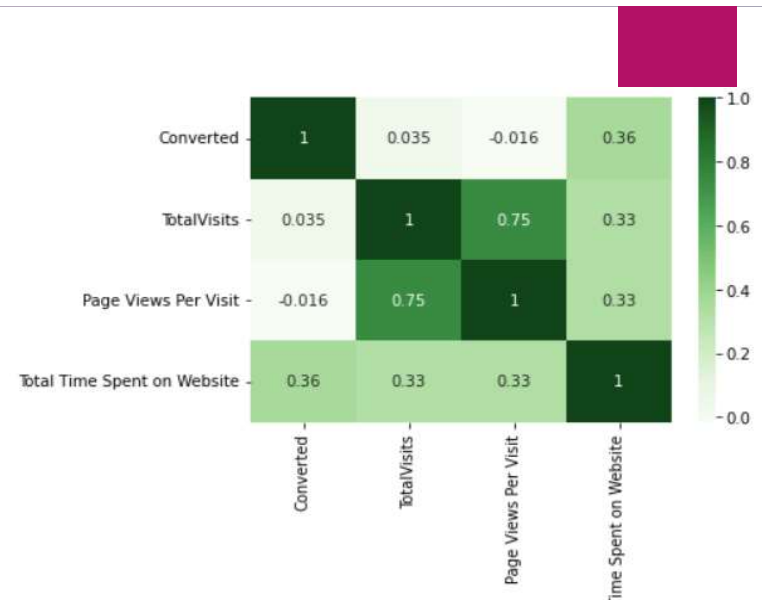
Lead Conversion Rate of Specialization



Correlation Analysis

Below points to be considered.

1. A strong positive correlation between 'Total Visits' & 'Page Views per Visit', indicating that customers who visited the website more frequently tend to view more pages per visit.
2. More time spent by Customers on the website have a higher LCR, indicating that increasing the time spent on the website can lead to higher conversion rates.



- 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

Data Preparation

- ▶ The data set has a large number of features & 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 (LRMod 4) is the final model used for model evaluation & making predictions.

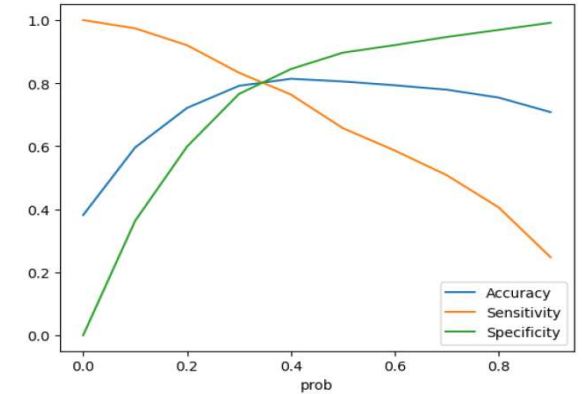
Model Building

Model Evaluation

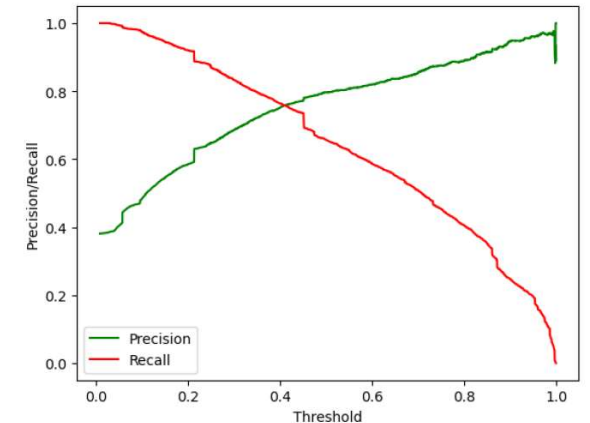
Based on the curve analysis, a cutoff probability of 0.35 (approximation) is suggested as the optimal point for classification.

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

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	



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	



Predictions On Dataset

ROC Curve - Train Data Set :

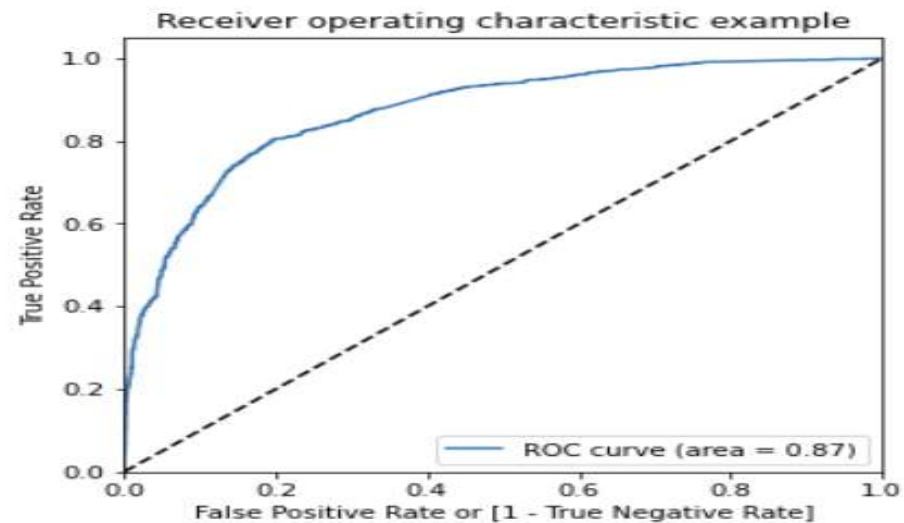
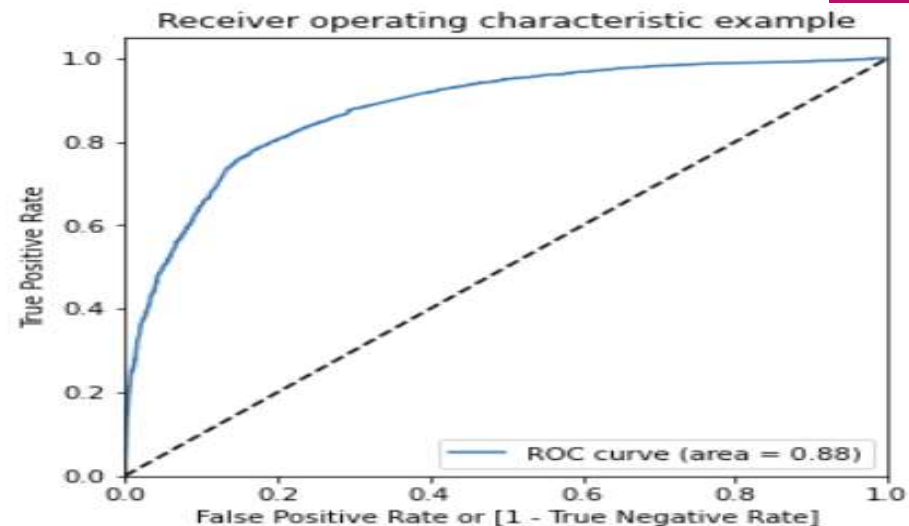
Area under ROC curve is found to be 0.88 out of 1 indicating that model is a good predictor.

The curve is plotted as close to the top left corner of the plot as possible which indicates that model has a high true positive rate & low false positive rate at all threshold values.

ROC Curve - Test Data Set :

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.



Predictions On Dataset

The customers with a high lead score have a higher chance of conversion and low lead score have a lower chance of conversion.

	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
0	4269	1	0.697934	1	70
1	2376	1	0.860665	1	86
2	7766	1	0.889241	1	89
3	9199	0	0.057065	0	6
4	4359	1	0.871510	1	87
5	9186	1	0.503859	1	50
6	1631	1	0.419681	1	42
7	8963	1	0.154531	0	15
8	8007	0	0.072344	0	7
9	5324	1	0.298849	0	30

Conclusion

Evaluation metrics of the model are consistently close to each other indicating that model is performing consistent in both the test & train datasets. This suggests that model is reliable & not overfitting to the training data. Similar performance across evaluation metrics also means that there are no significant biases in the model's predictions. This is positive sign for the model's performance & provides confidence on its ability to make accurate predictions in the future.

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	

Train Data Set:

Accuracy: 80.57%
Sensitivity: 79.72%
Specificity: 81.08%

Test Data Set:

Accuracy: 80.34%
Sensitivity: 79.27%
Specificity: 81.04%

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

Recommendations

- ❑ Features i.e. 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional' & 'Total Time Spent on Website' have a high conversion rate & should be utilized more in lead generation.
- ❑ Working professionals should be aggressively targeted due to high probability of converting & better financial grounds.
- ❑ Referrals by old customers have a significantly higher conversion rate & treating with discounts or other rewards would bring referrals.
- ❑ Utilizing Google ads or email campaigns in more efficient way can save time & also increase the conversion rate.
- ❑ Leads whose 'Last Activity' is 'SMS Sent' or 'Email Opened' tend to have a higher conversion rate.
- ❑ Marketing campaigns to specific groups of customers i.e Targeted / Content based Marketing helps in attracting/ retains customers in respective fields.



Thank you