LEAD SCORING CASE STUDY

Detection of Hot Leads to concentrate more of marketing efforts on
Them, improving conversion rates for X Education.

– AJAY SINGH

TABLE OF CONTENTS

Background of X Education Company

1. An education company named X Education sells online courses to industry professionals.

2. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

3. The company markets its courses on several websites and search engines like Google.

4. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.

5. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

6. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.

7. Through this process, some of the leads get converted while most do not.

8. The typical lead conversion rate at X education is around 30%.

The problem Statement:

1. X Education gets a lot of leads, its lead conversion rate is very poor at around 30%.

2. X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads.

3. Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

1. To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

2. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

3. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are grouped based on their propensity or likelihood to convert.
- This results in a focused group of hot leads.

Better Communication

- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

Since we have a target of 80% conversion rate, we would want to obtain a high sensitivity in obtaining hot leads

Analysis Approach

Data cleaning

Loading Data Set,
understanding &
cleaning data

EDA

Check imbalance,
Univariate &
Bivariate analysis

Data preparation

Dummy
variables,
test-train split,
feature scaling

Model building

RFE for top 15
feature, Manual
Feature
Reduction
& finalizing
model

Model Evaluation

Confusion
matrix,
Cutoff
Selection,
assigning Lead
Score

Predictions on Test Data set

Compare train vs
test metrics,
Assign
Lead Score and
get
top features

Recommendation

Suggest top 3
features to
focus for
higher
conversion &
areas for
improvement

Data Cleaning

1. "Select" level represents null values for some categorical variables, as customers did not choose anyoption from the list.

2. Columns with over 35% null values were dropped.

3. Missing values in categorical columns were handled based on value counts and certain considerations.

4. Drop columns that don't add any insight or value to the study objective (tags, country)

5. Imputation was used for some categorical variables.

6. Additional categories were created for some variables.

7.Columns with no use for modelling (Prospect ID, Lead Number) or only one category of response were dropped.

8. Numerical data was imputed with mode after checking distribution.

Data Cleaning

1. Skewed category columns were checked and dropped to avoid bias in logistic regression models.

2. Outliers in TotalVisits and Page Views Per Visit were treated and capped.

3. Invalid values were fixed and data was standardized in some columns, such as lead source.

4. Low frequency values were grouped together to "Others".

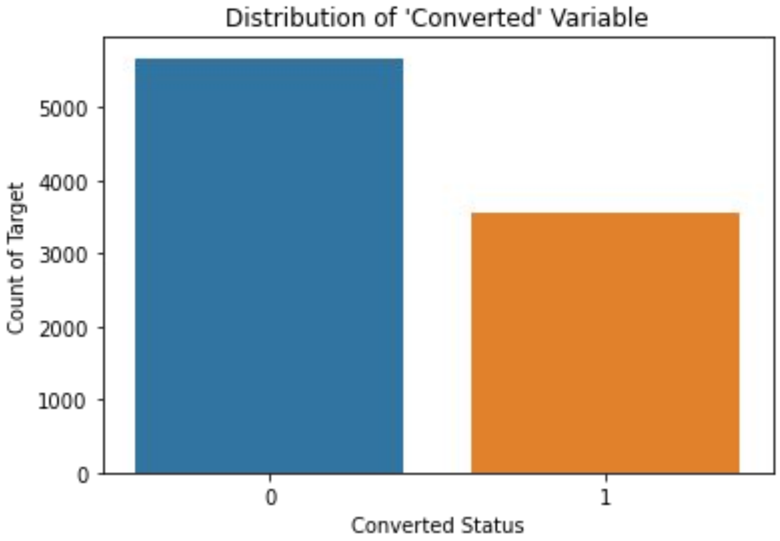
5 .Binary categorical variables were mapped.

6. Other cleaning activities were performed to ensure data quality and accuracy.

7. Fixed Invalid values & Standardizing Data in columns by checking casing styles,etc.
(lead source has Google, google)

EDA

Data is imbalanced while analyzing target variable.

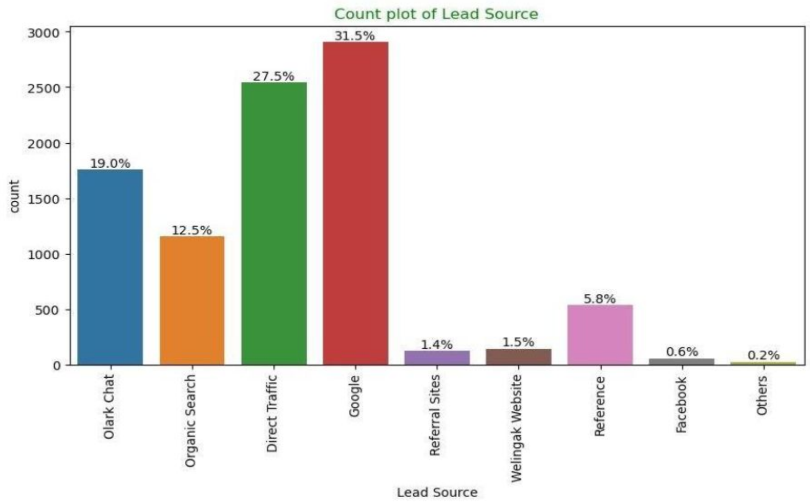


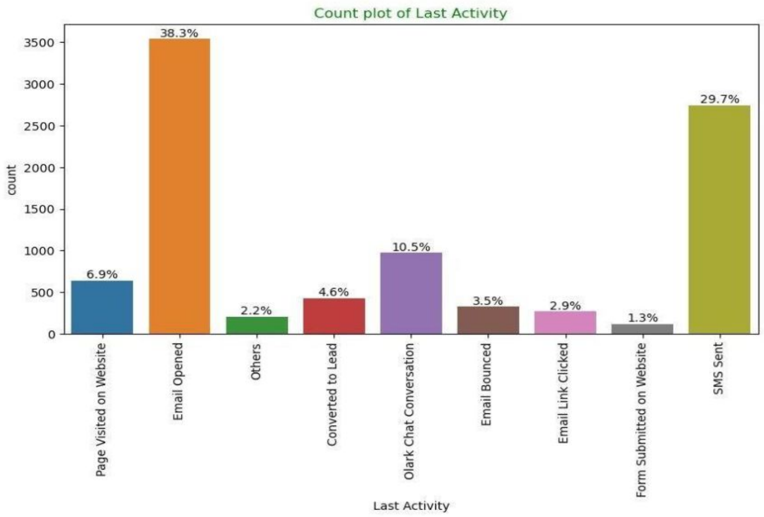
- Conversionrate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)

- While 61 . 5% of the people didn't convert to leads. (Majority)

EDA

- Univariate Analysis - Categorical Variables

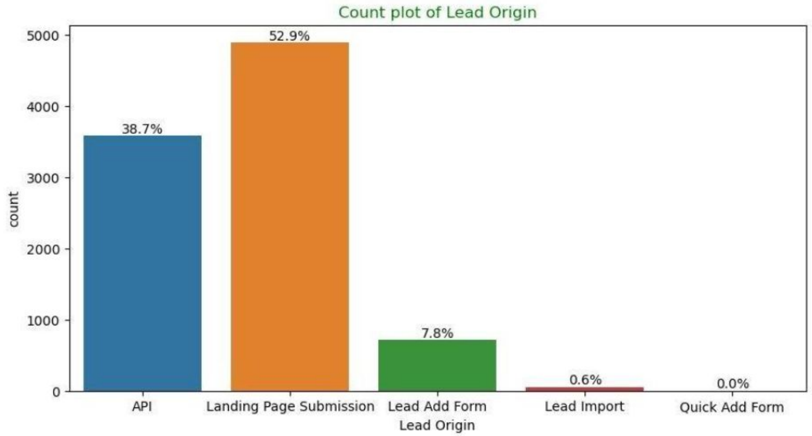


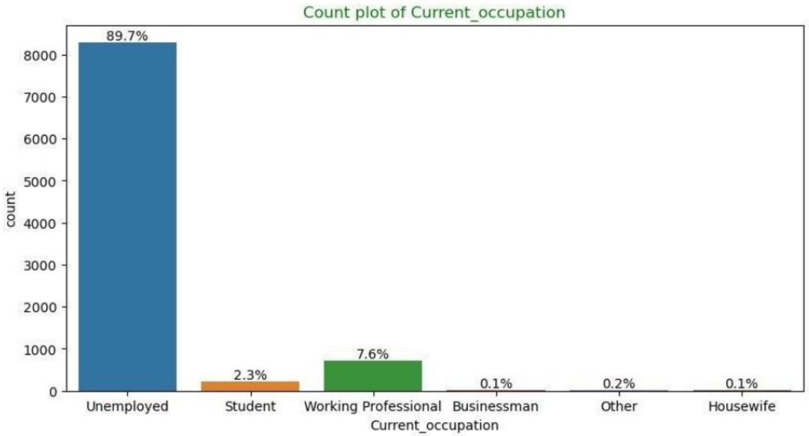


- **Lead Source:** 58% Lead source is from Google & Direct Traffic combined.
- **Last Activity:** 68%of customerscontributionin SMS Sent & Email Opened activities.

EDA

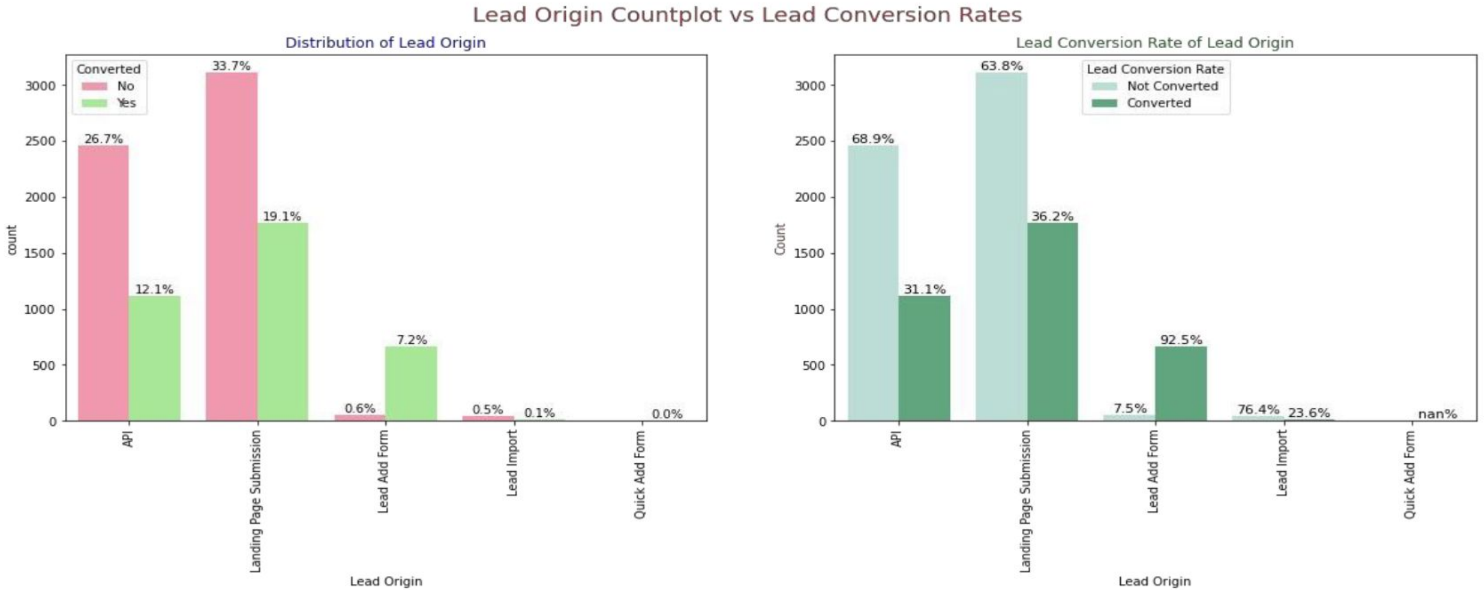
- **Univariate Analysis – Categorical Variables**





- **Lead Origin:** "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- **Current_occupation:** It has 90% of the customers a Unemployed.

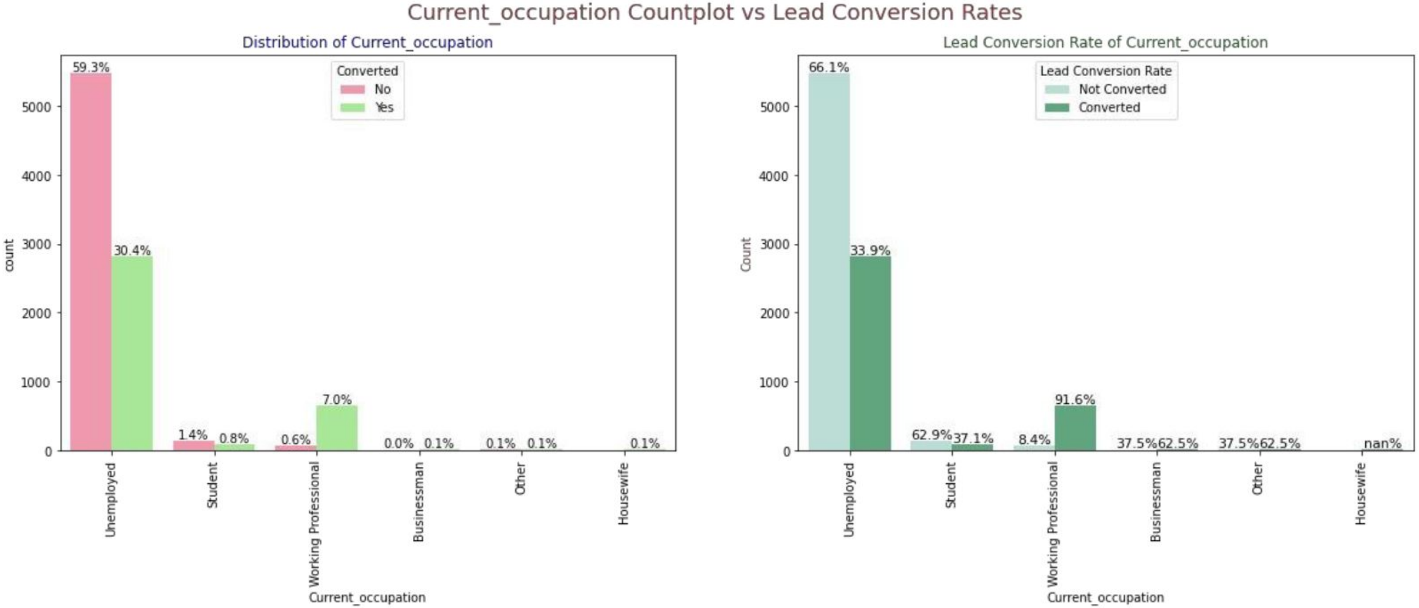
EDA – Bivariate Analysis for Categorical Variables



Lead Origin:

- 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%**.

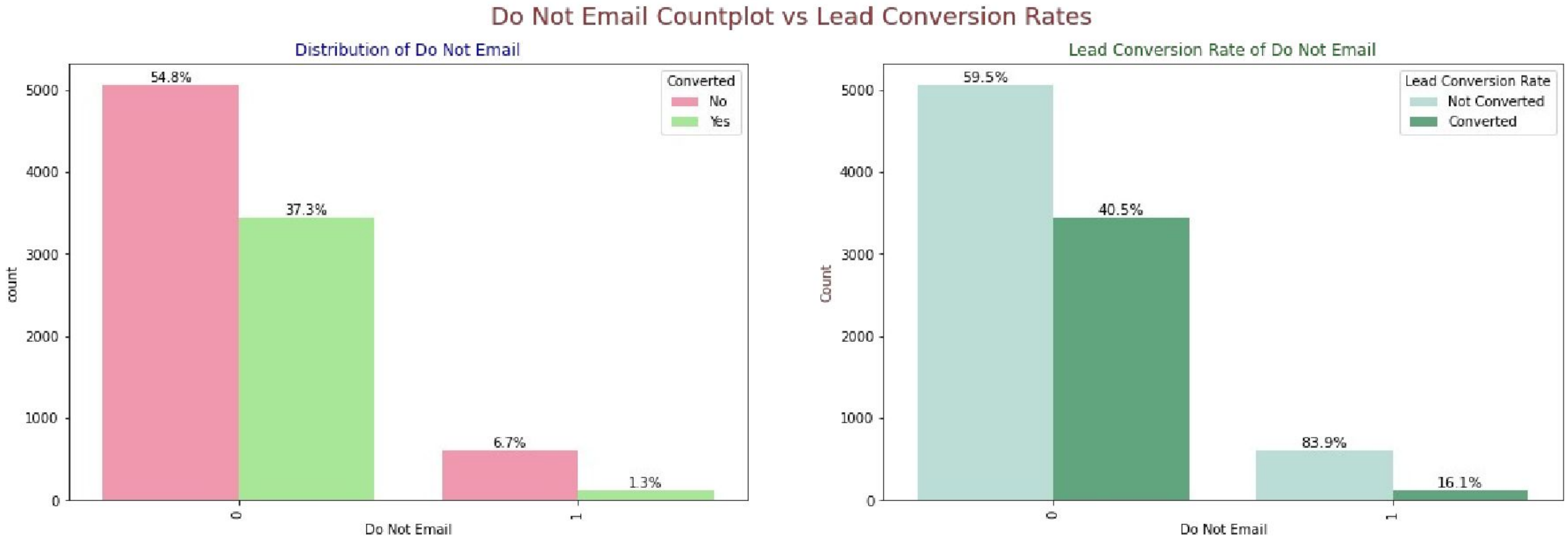
EDA – Bivariate Analysis for Categorical Variables



Current_occupation:

- 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)**.

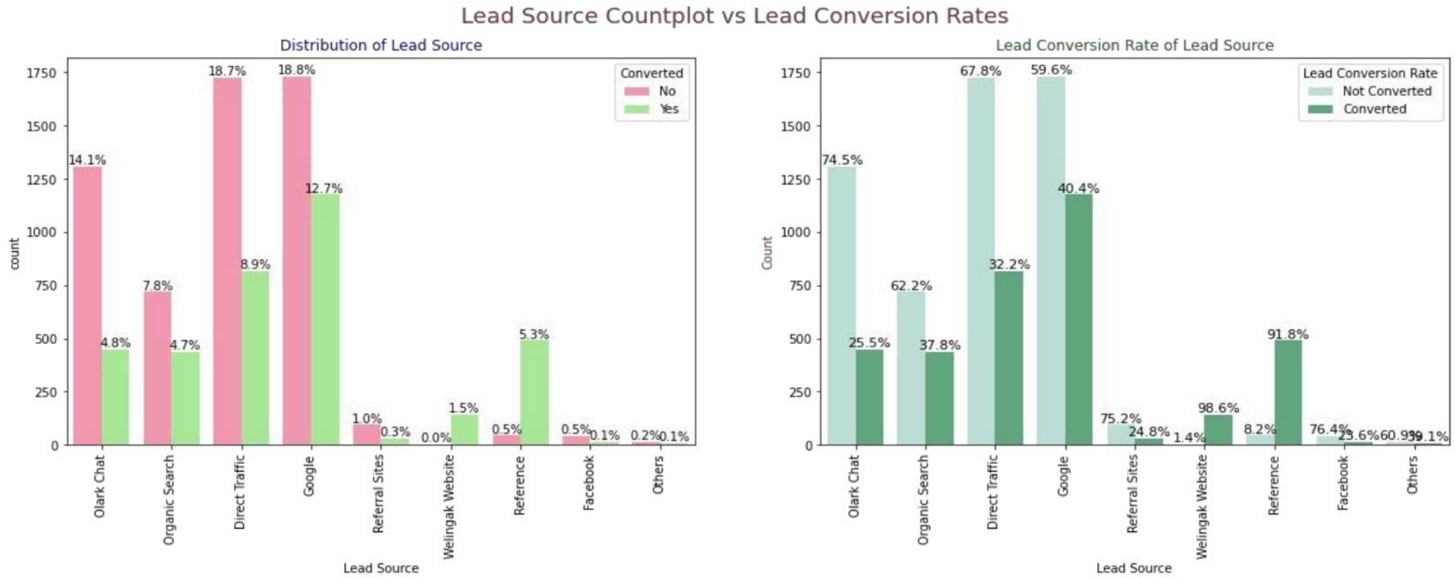
EDA – Bivariate Analysis for Categorical Variables



Do Not Email:

- 92% of the people has opted that they don't want to be emailed about the course & 40% of them are converted to leads.

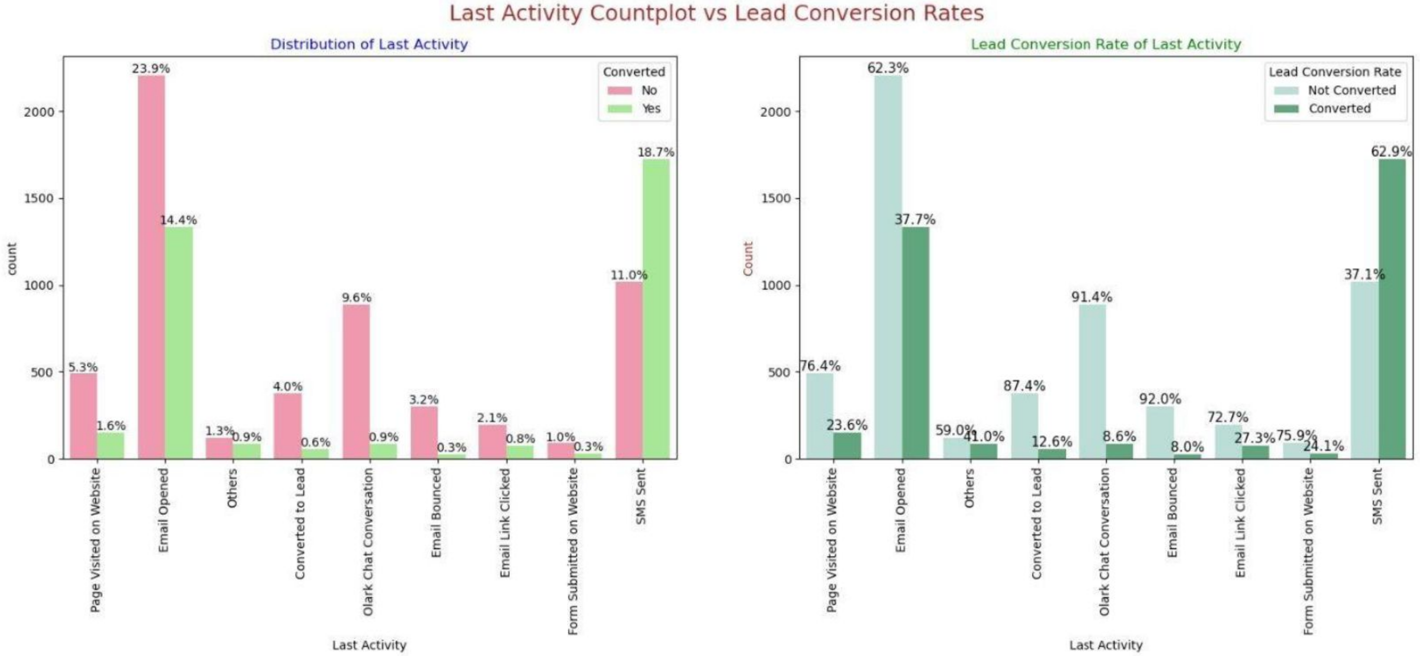
EDA – Bivariate Analysis for Categorical Variables



Lead Source:

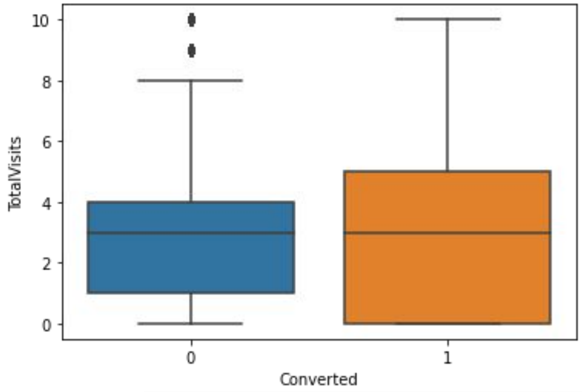
- *Google* has **LCR of 40%** out of 31% customers,
- *Direct Traffic* contributes **32% LCR** with 27% customers, which is lower than Google,
- *Organic Search* also gives **37.8% of LCR**, but the contribution is by only 12.5% of customers,
- *Reference* has **LCR of 91%**, but there are only around 6% of customers through this Lead Source.

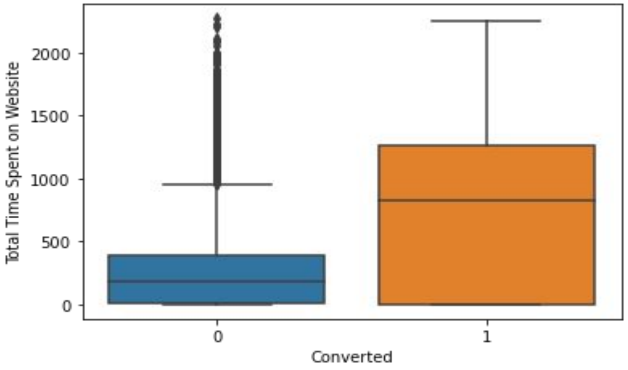
EDA – Bivariate Analysis for Categorical Variables

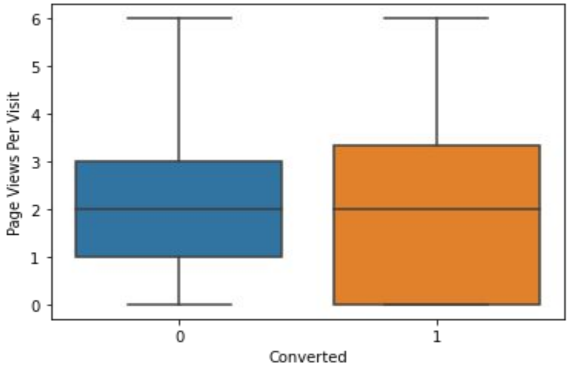


Last Activity:

- *'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.**







• Past Leads who spends more time on the Website have a higher chance of getting successfully converted than those who spends less time as seen in the box-plot

Data Preparation before Model building

- Binary level categorical columns were already mapped to 1 /0 in previous steps

- Created dummy features (one-hot encoded) for categorical variables - Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation

- Splitting Train & Test Sets 70: 30 % ratio was chosen for the split

- Feature scaling Standardization method was used to scale the features

- Checking the correlations Predictor variables which were highly correlated with each other were dropped

Model Building

Feature Selection

• The data set has lots of dimension and large number of features.

• This will reduce model performance and might take high computation time.

• Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.

• Then we can manually fine tune the model.

• RFE outcome Pre RFE - 48 columns & Post RFE - 15 columns

Model Building

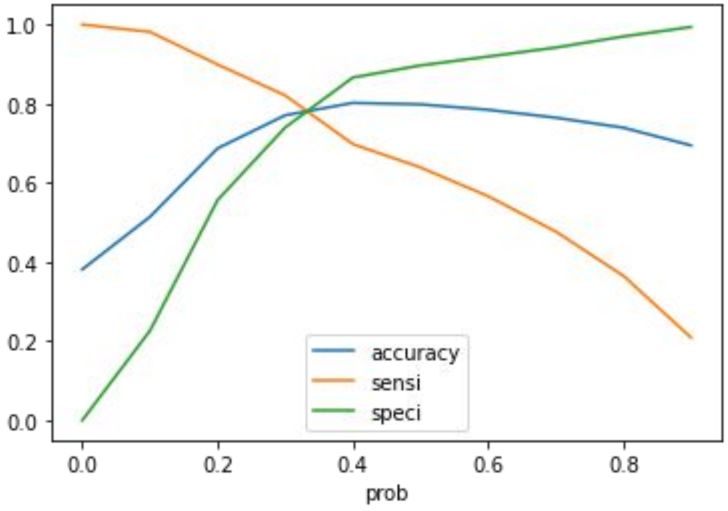
• Manual Feature Reduction process was used to build models by dropping variables with p - value greater than 0.05.

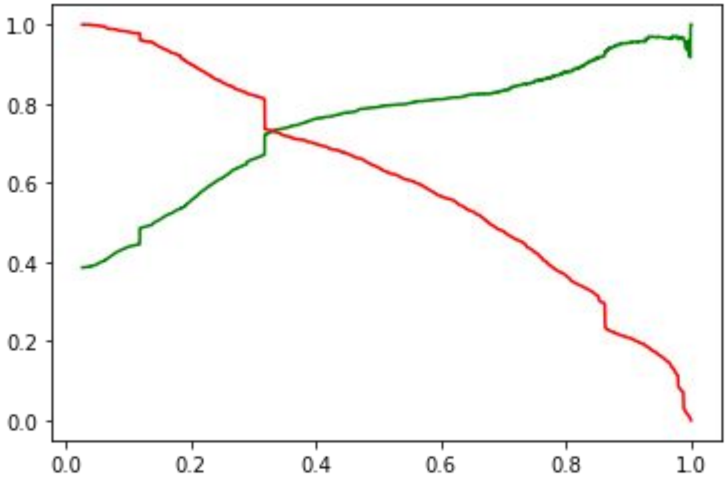
• Model 6 looks stable after four iteration with:
significant p-values within the threshold (p-values < 0.05) and
No sign of multicollinearity with VIFs less than 5

• Hence, logm6 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

Model Evaluation

Train Data Set





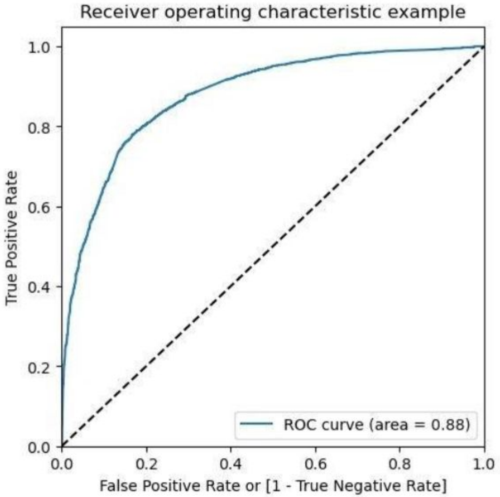
Confusion Matrix & Evaluation Metrics with 0.3 as cutoff

Confusion Matrix & Evaluation Metrics with 0.41 as cutoff

Model Evaluation

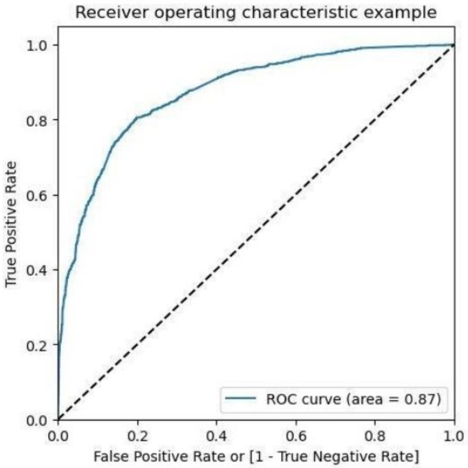
ROC Curve - Train Data Set

- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve - Test Data Set

- Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation

Confusion Matrix & Metrics

Train Data Set

```
array([[2961, 1041],
       [ 442, 2024]], dtype=int64)
```

Test Data Set

```
array([[1236,  441],
       [ 187,  908]], dtype=int64)
```

Using a cut-off value of 0.3, the model achieved a sensitivity of 82.07% in the train set and 82.92% in test set.

Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which converting

The CEO of X Education had set a target sensitivity of around 82%.

The model also achieved an accuracy of 77.07%, which is in line with the study's objectives.

Recommendation based on Final Model

As per the problem statement, increasing lead conversion is crucial for the growth and success Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

• We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

Lead Source_Welingak Website: 5.66

What is your current occupation_Working Professional : 3.75

Lead Source_Reference : 3.66

Lead Source_Others :1.415608

What is your current occupation_Unemployed : 1.247420

What is your current occupation_Student : 1.112550

We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:

Lead Source_Google : 0.345238

Do Not Email : -0.322800