Lead Scoring Case Study

Presented by

ROHIT JOGALE

Contents

About X Education Company

- Industry professionals can purchase online courses from X Education, a company that provides education.
- Many professionals who are interested in the courses visit their website on any given day and search for courses.
- On numerous websites and search engines like Google, the company advertises its courses.
- Upon arriving at the website, these visitors may browse the courses, submit a form for the course, or watch some videos.
- These people are categorized as leads when they fill out a form with their phone number or email address.
- Through this process, some of the leads get converted while most do not.

Problem Statement

- X Education receives a lot of leads, but only about 30% of those leads actually become students.
- By locating the most promising leads, also referred to as Hot Leads, X Education hopes to increase the effectiveness of the lead conversion process.
- Their sales team wants to be aware of this potential group of leads, so rather than calling everyone, they will concentrate more on communicating with them.
- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

Ideas for Increasing Lead Conversion.

- Leads are grouped based on their propensity or likelihood To convert
- This results in a focused group of hot leads.
- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.
- 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.

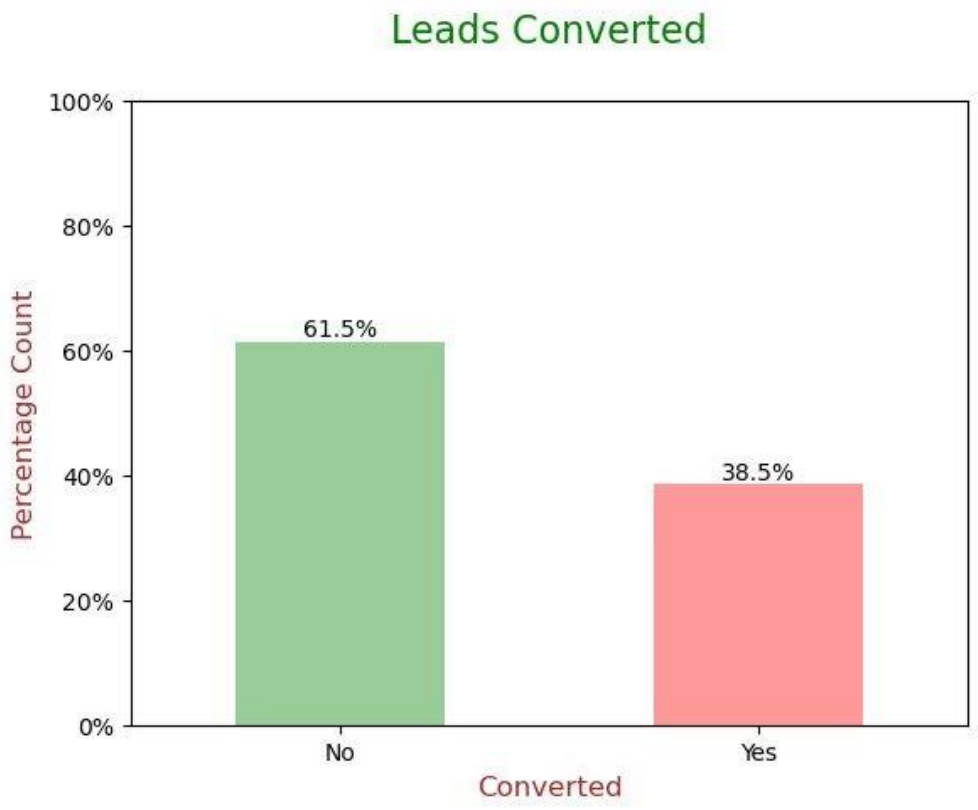
Data cleaning

- **"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 (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

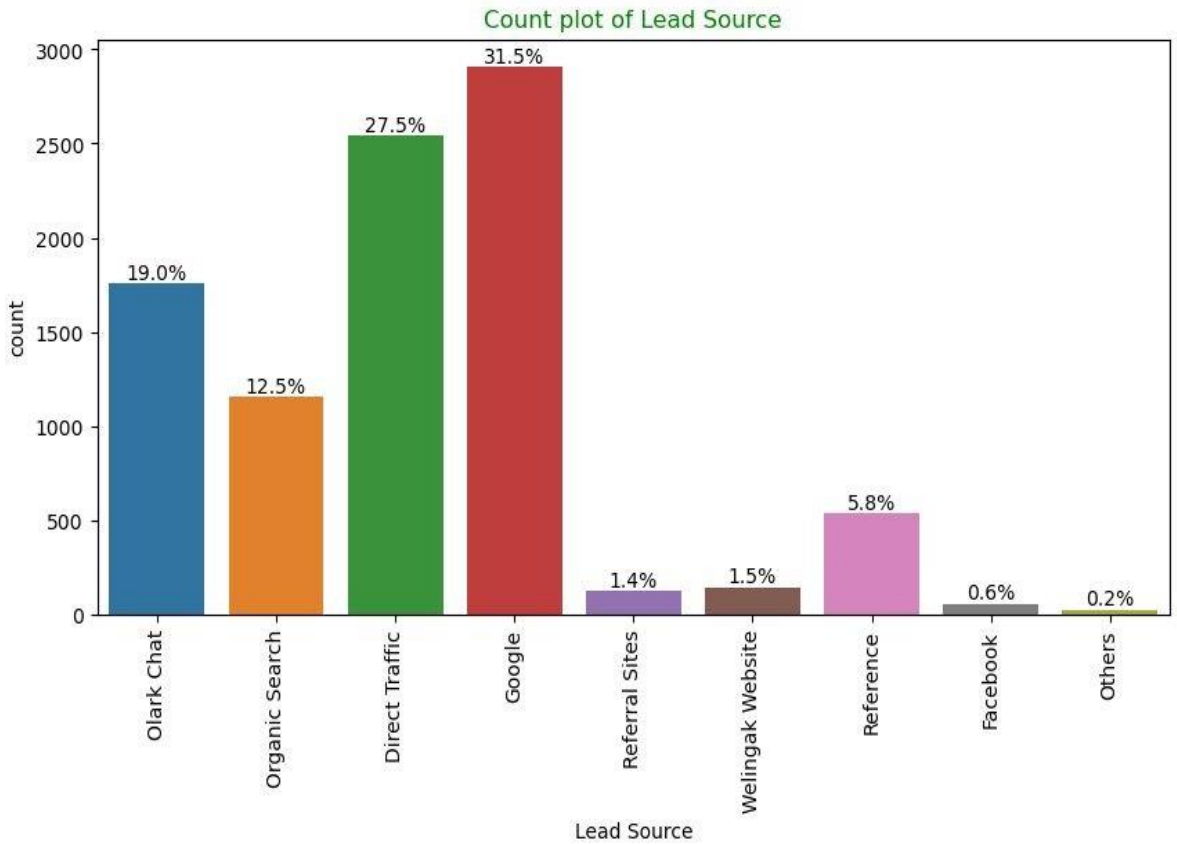
Data Cleaning

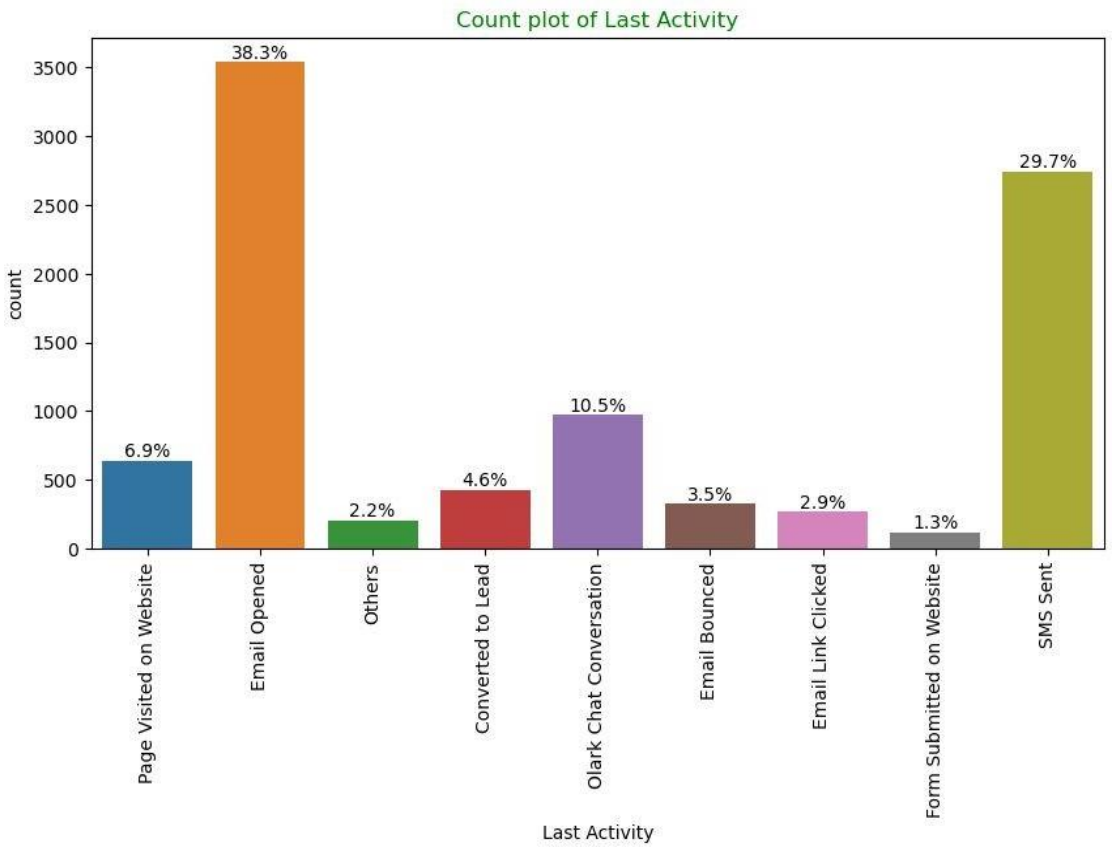
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in **TotalVisits** and **Page Views Per Visit** were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to “Others”.
- Binary categorical variables were mapped.
- Other cleaning activities were performed to ensure data quality and accuracy.
    - Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google)

Exploratory Data Analysis (EDA)



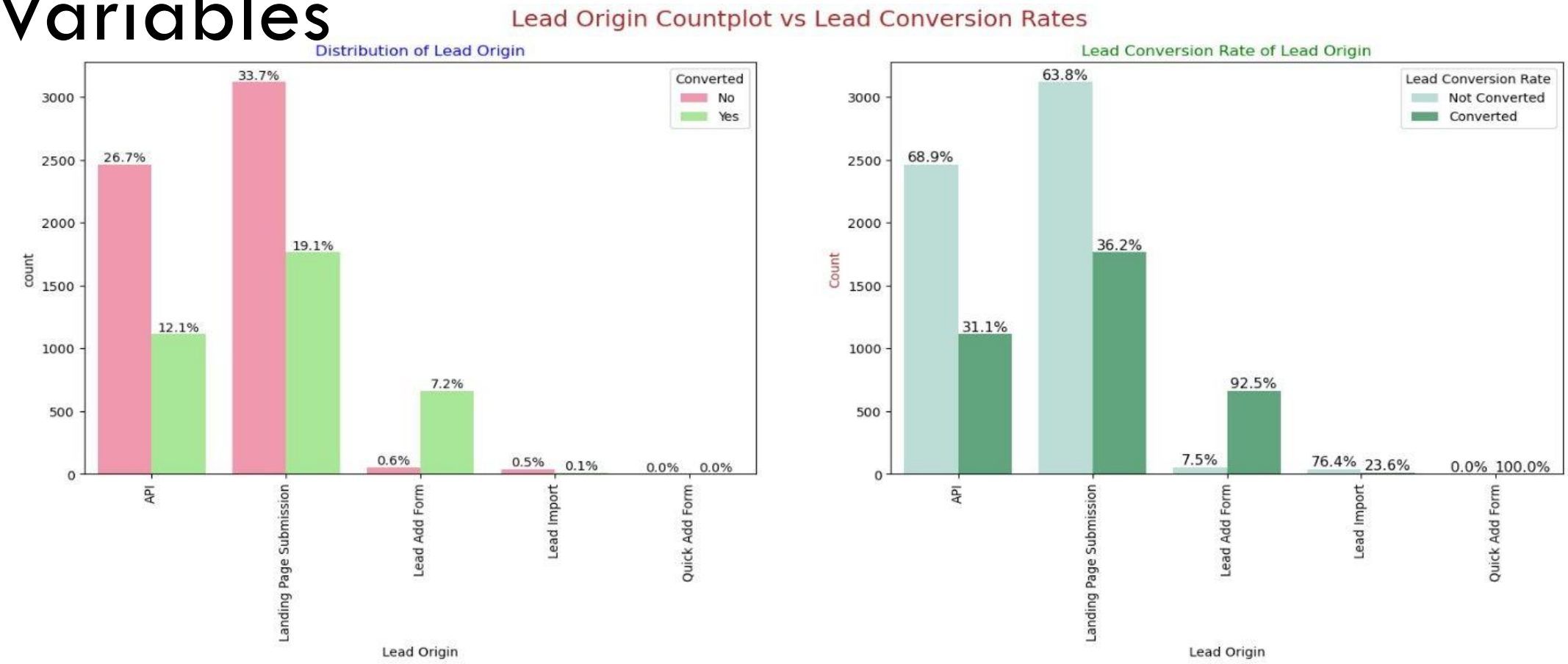
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)





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

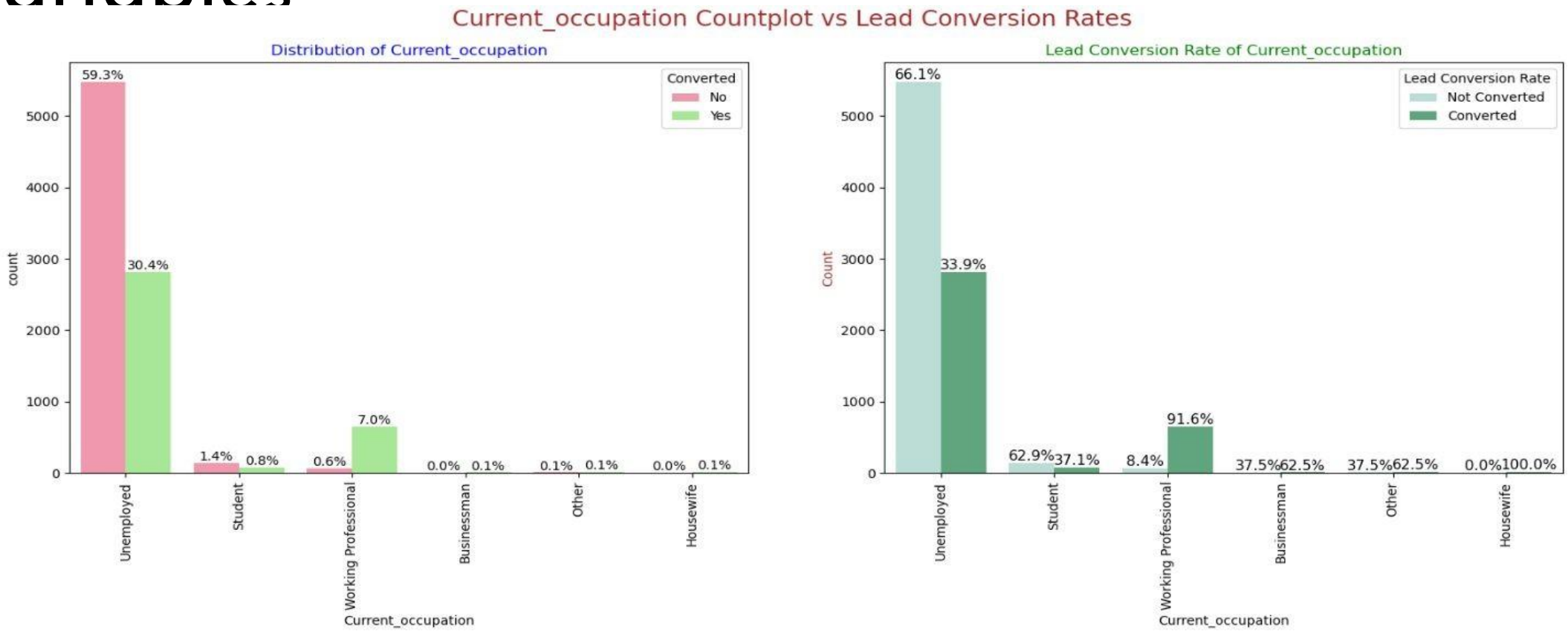
EDA – Bivariate Analysis for Categorical Variables



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%**.

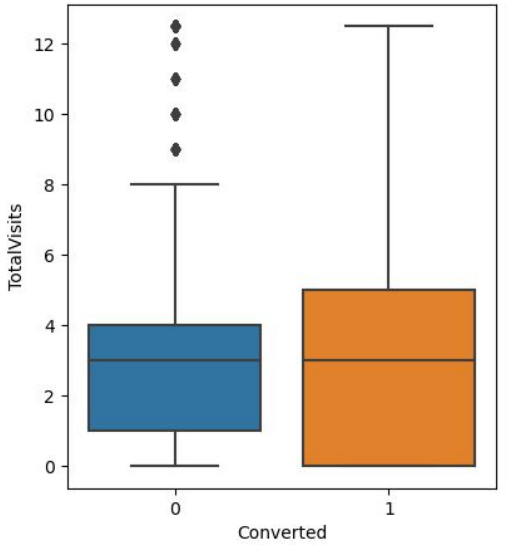
EDA – Bivariate Analysis for Categorical Variables

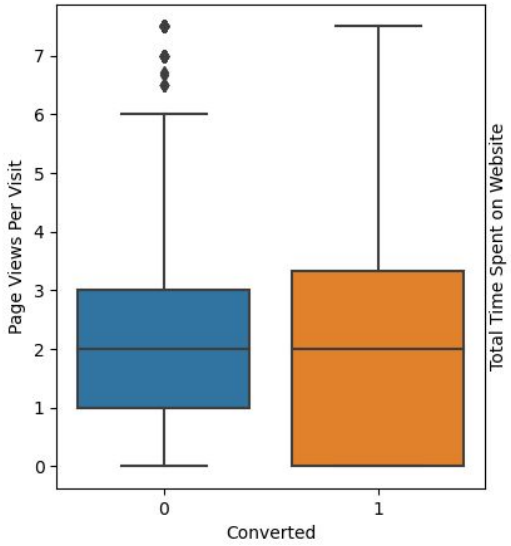


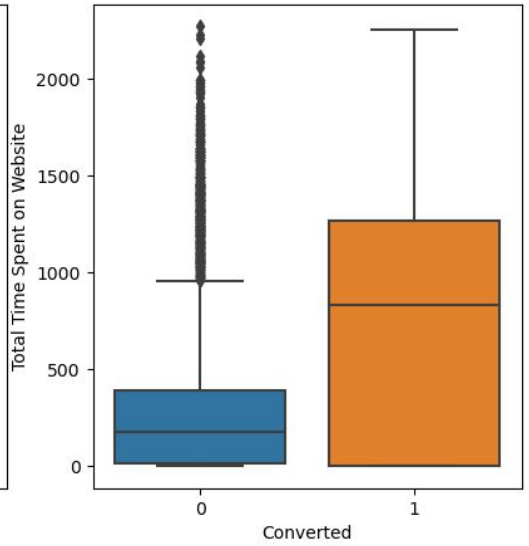
Current_occupation:

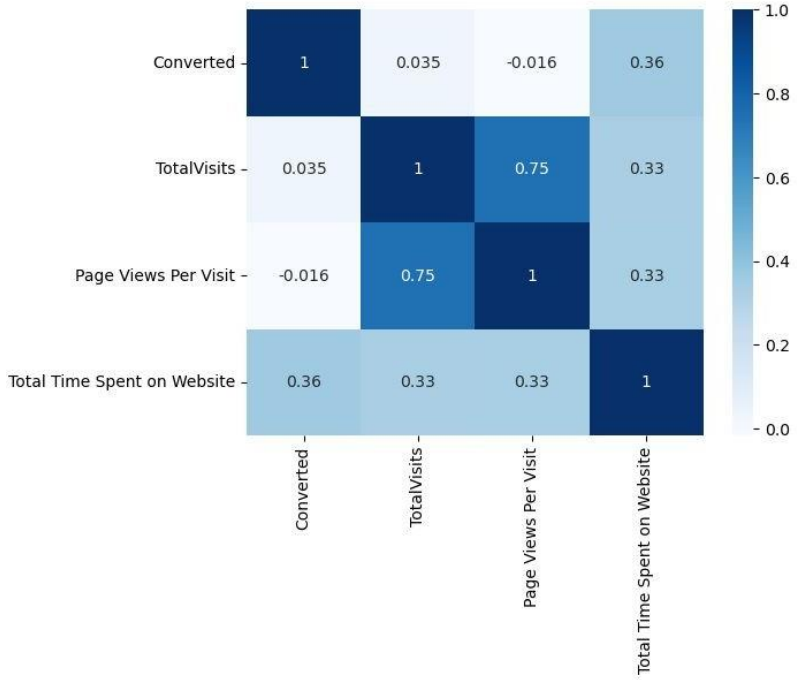
- Around 90% of the customers are *Unemployed*, with **lead conversion rate (LCR) of 34%**.
- While *Working Professional* contribute only 7.6% of total customers with almost **92% Lead conversion rate (LCR)**.

EDA – Bivariate Analysis for Numerical Variables









- 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**

The way to get started is to quit talking and begin doing.

Walt Disney

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 (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

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 4 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, **logm4** 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

It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots

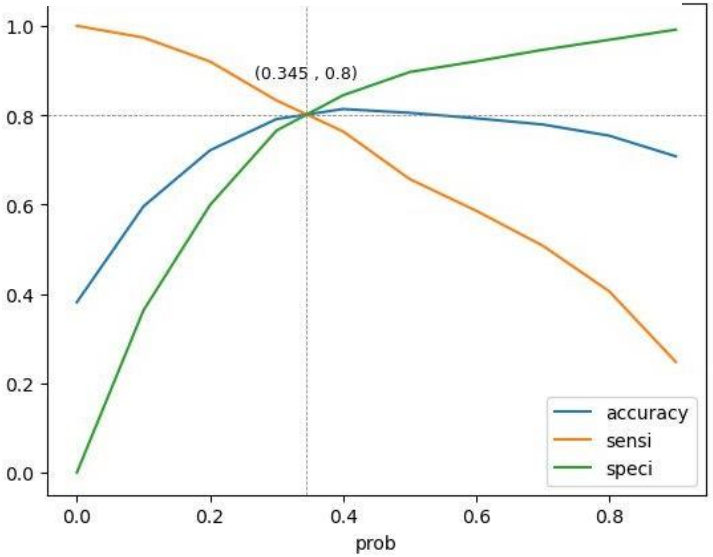
Confusion Matrix & Evaluation Metrics with 0.345 as cutoff

```
*****************************************************

Confusion Matrix
[[3230  772]
 [ 492 1974]]

*****************************************************

True Negative                    :  3230
True Positive                    :  1974
False Negative                   :  492
False Positve                    :  772
Model Accuracy                   :  0.8046
Model Sensitivity                :  0.8005
Model Specificity                :  0.8071
Model Precision                  :  0.7189
Model Recall                     :  0.8005
Model True Positive Rate (TPR)   :  0.8005
Model False Positive Rate (FPR)  :  0.1929
```

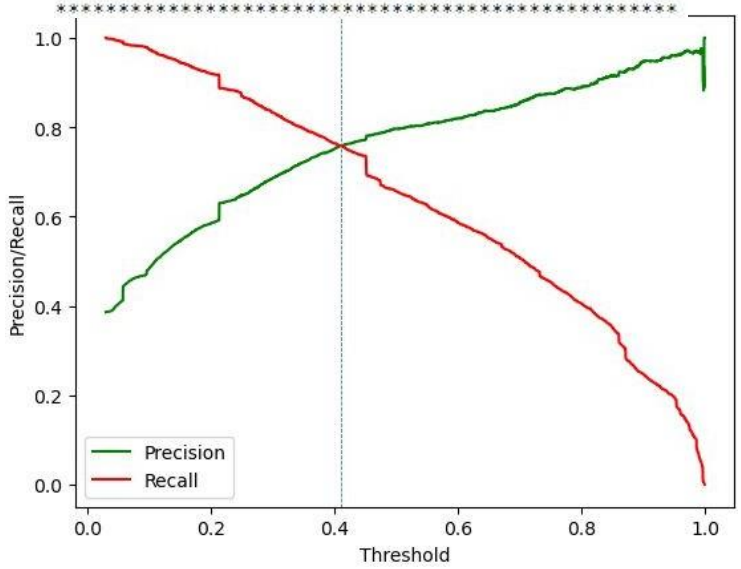
Confusion Matrix & Evaluation Metrics with 0.41 as cutoff

```
*****************************************************

Confusion Matrix
[[3406  596]
 [ 596 1870]]

*****************************************************

True Negative                    :  3406
True Positive                    :  1870
False Negative                   :  596
False Positve                    :  596
Model Accuracy                   :  0.8157
Model Sensitivity                :  0.7583
Model Specificity                :  0.8511
Model Precision                  :  0.7583
Model Recall                     :  0.7583
Model True Positive Rate (TPR)   :  0.7583
Model False Positive Rate (FPR)  :  0.1489
```

****



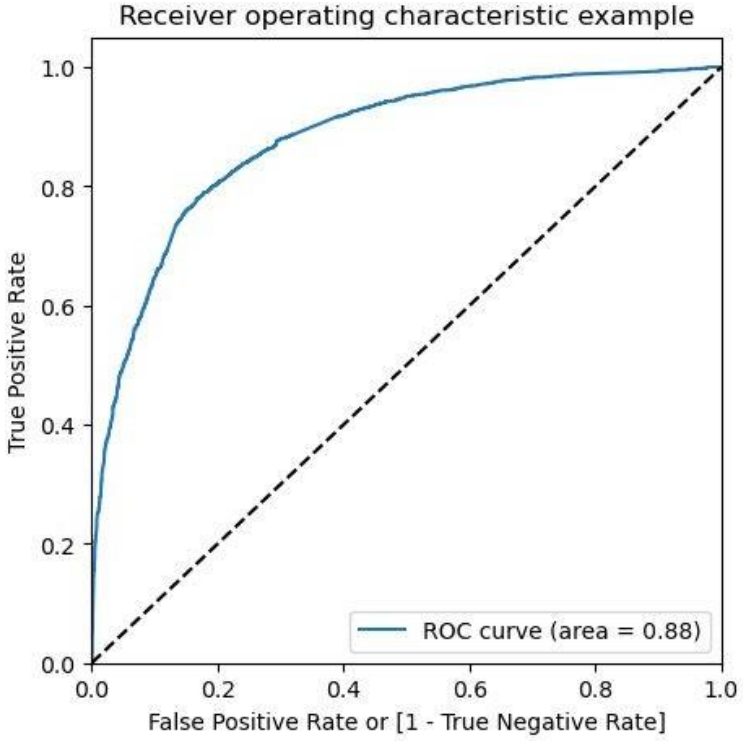
****



Model Evaluation

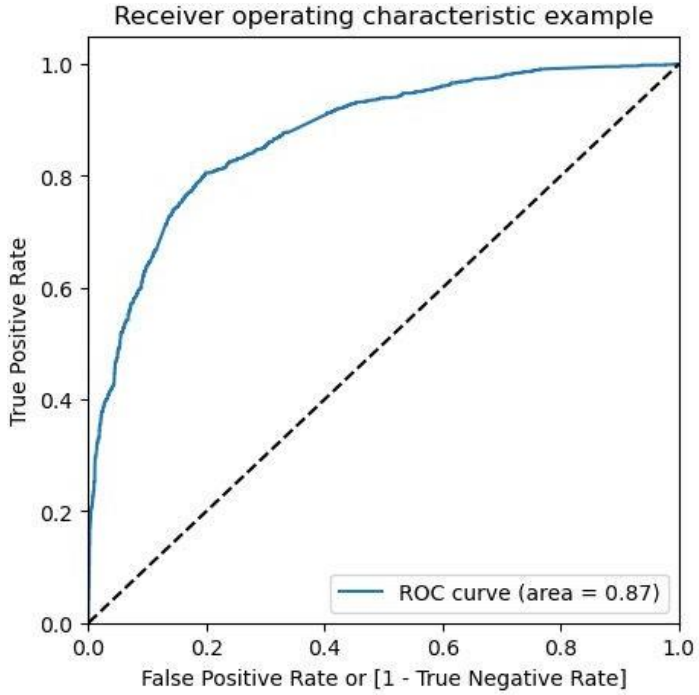
ROC Curve – Train Data Set

- 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.



Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X 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: 5.39
  - Lead Source_Reference: 2.93
  - Current_occupation_Working Professional: 2.67
  - Last Activity_SMS Sent: 2.05
  - Last Activity_Others: 1.25
  - Total Time Spent on Website: 1.05
  - Last Activity_Email Opened: 0.94
  - Lead Source_Olark Chat: 0.91

- We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:
  - Specialization in Hospitality Management: -1.09
  - Specialization in Others: -1.20
  - Lead Origin of Landing Page Submission: -1.26

Thank you