

X Education - Lead Scoring Case Study

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

Table of Contents

Background of X Education Company

Problem Statement & Objective of the Study

Suggested Ideas for Lead Conversion

Analysis Approach

Data Cleaning

EDA

Data Preparation

Model Building (RFE & Manual fine tuning)

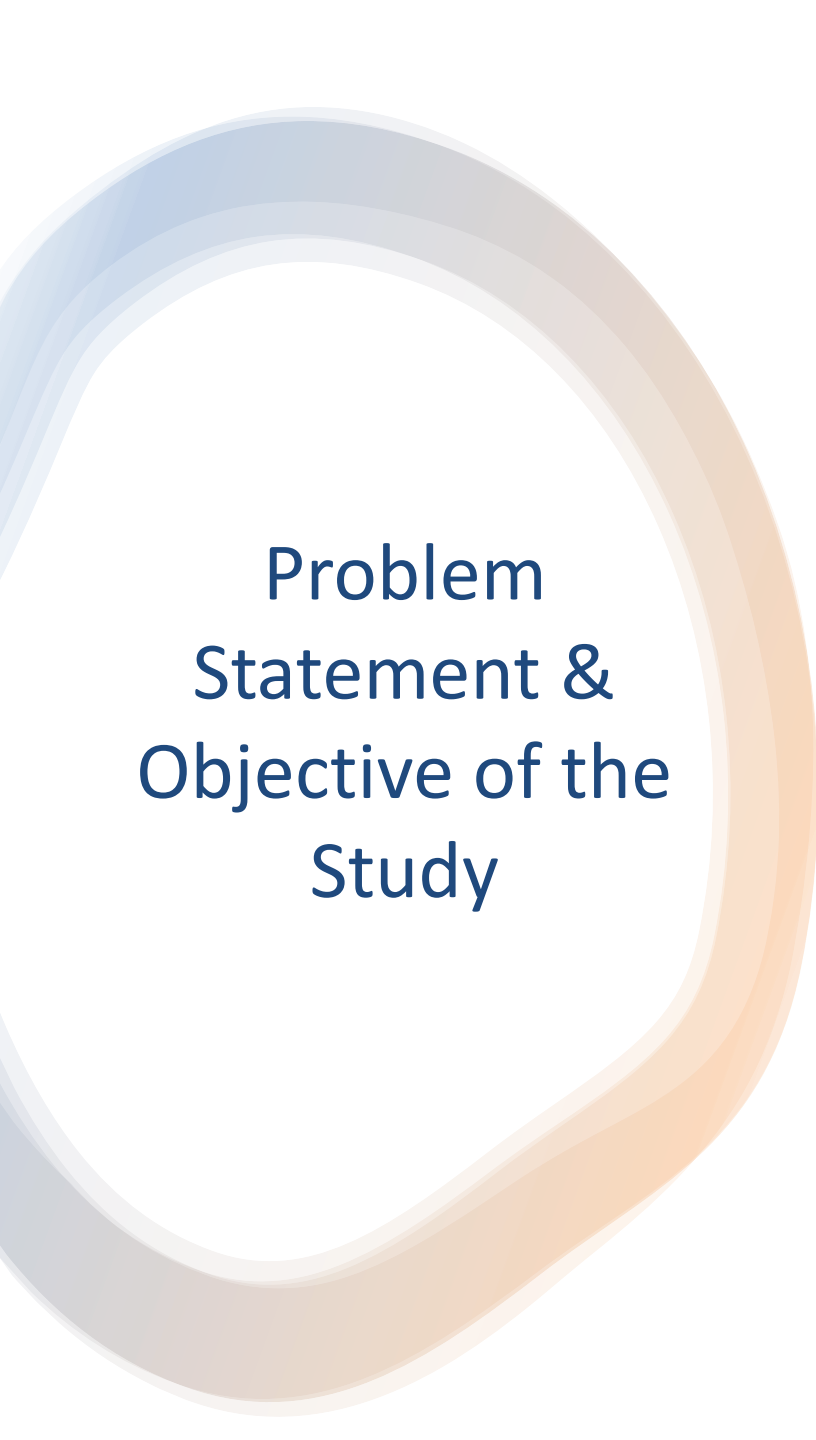
Model Evaluation

Summary



Background of X Education Company

- 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.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- 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.
- The typical lead conversion rate at X education is around 30%.



Problem Statement & Objective of the Study

- **Problem Statement:**
 - X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
 - X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
 - 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:**
 - To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
 - 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.
 - The CEO has given a ballpark of the target lead conversion rate to be around 80%.



Problem 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:**
 - 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

- The dataset contains no duplicate entries in the `Prospect ID` and `Lead Number` columns.
- Both `Prospect ID` and `Lead Number` serve as unique identifiers and do not contribute to the analysis; therefore, they can be excluded from the model.
- The dataset includes multiple instances where the value 'Select' is used, indicating missing data. These values should be converted to `NA` for better handling.
- Columns with more than 45% missing data should be dropped to improve the quality and reliability of the analysis.
- The dataset shows that approximately 97% of the entries are from India, making this column redundant. Therefore, it was removed from the analysis.

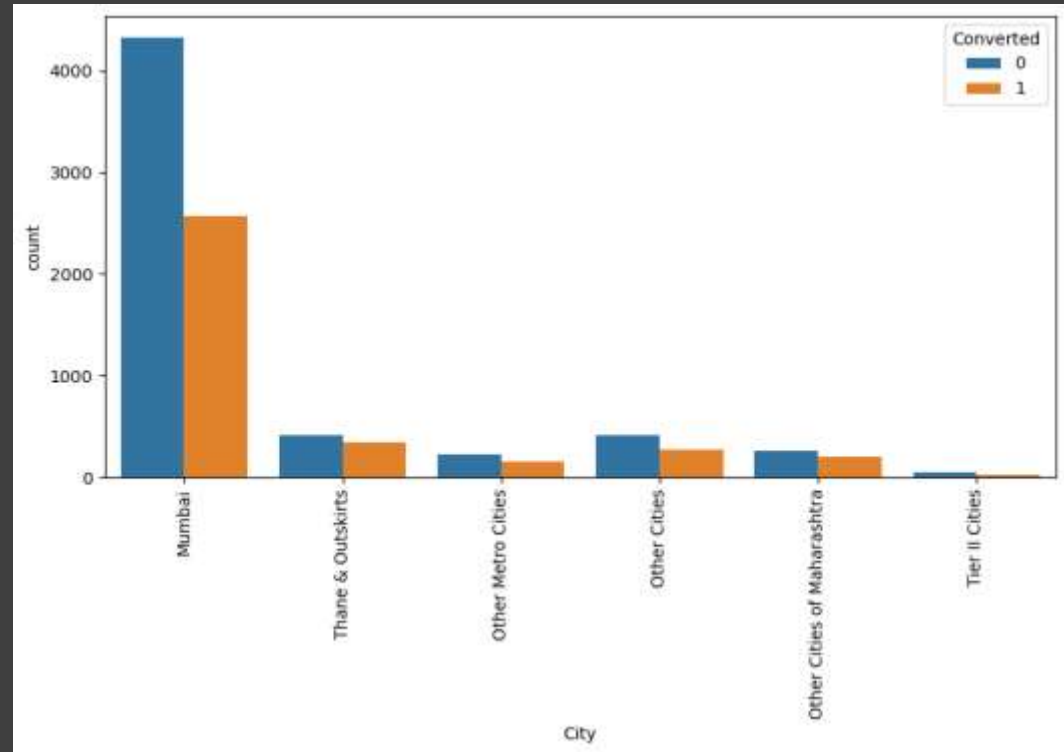


Data Cleaning

- Mumbai is the most frequently occurring city in the data, with a count close to the number of `NaN` entries.
- Missing (`NaN`) values in the city column were replaced with 'Mumbai' to ensure consistency.
- Some of the Specializations are about Management with different Specializations in them. We will combine those under one Specialization as Management
- We will replace NaN values with mode (Unemployed) in the What is your current occupation column.
- We will drop What matters most to you in choosing a course column as Most of the values are Better Career Prospects.

Exploratory Data Analysis (EDA)

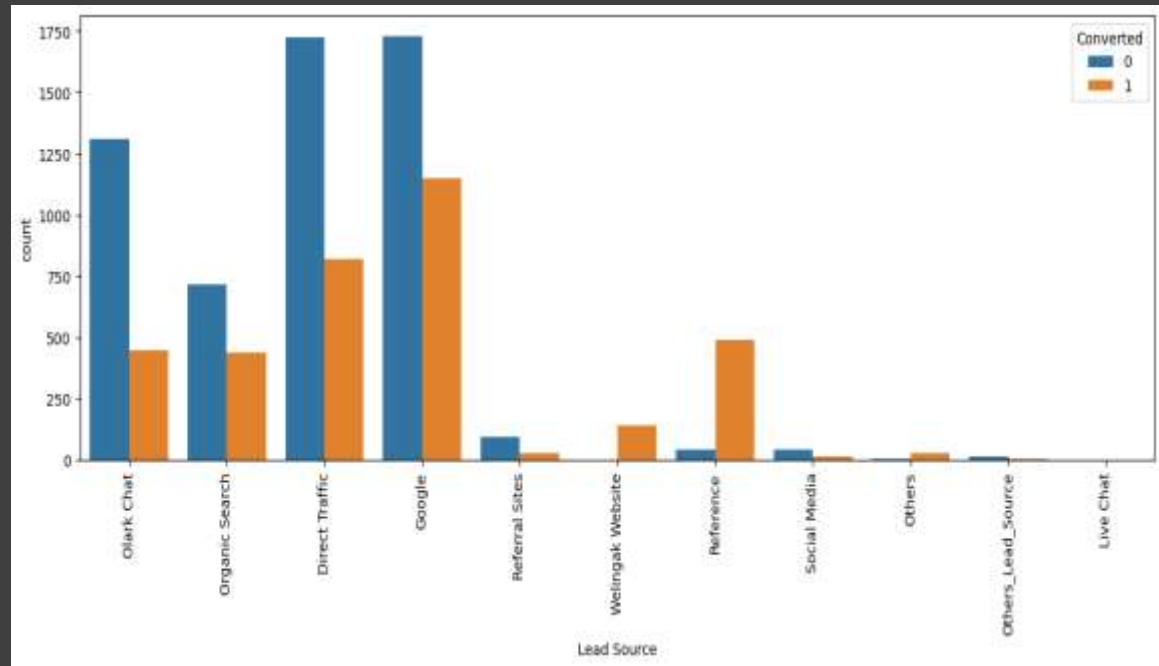
Categorical Variables



Mumbai is the most frequently occurring city in the data.

Exploratory Data Analysis (EDA)

Categorical Variables

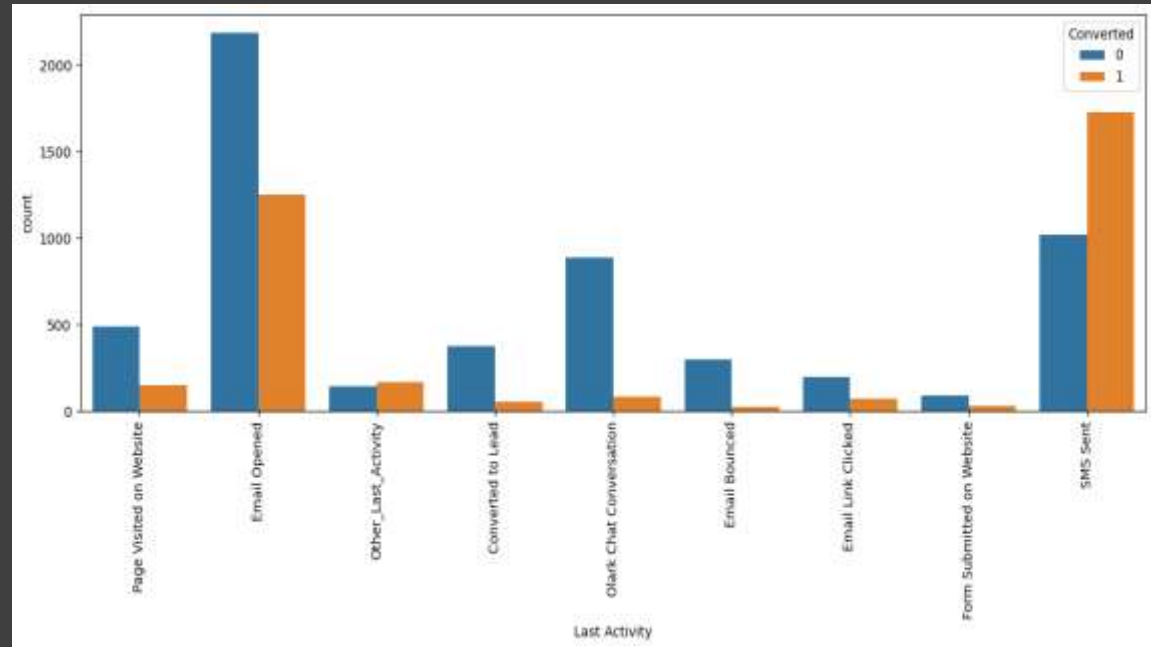


Lead Source :

- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of leads through reference and leads through welingak website is high.
- To improve the overall lead conversion rate, the focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and Google leads and generating more leads from reference and welingak website.

Exploratory Data Analysis (EDA)

Categorical Variables

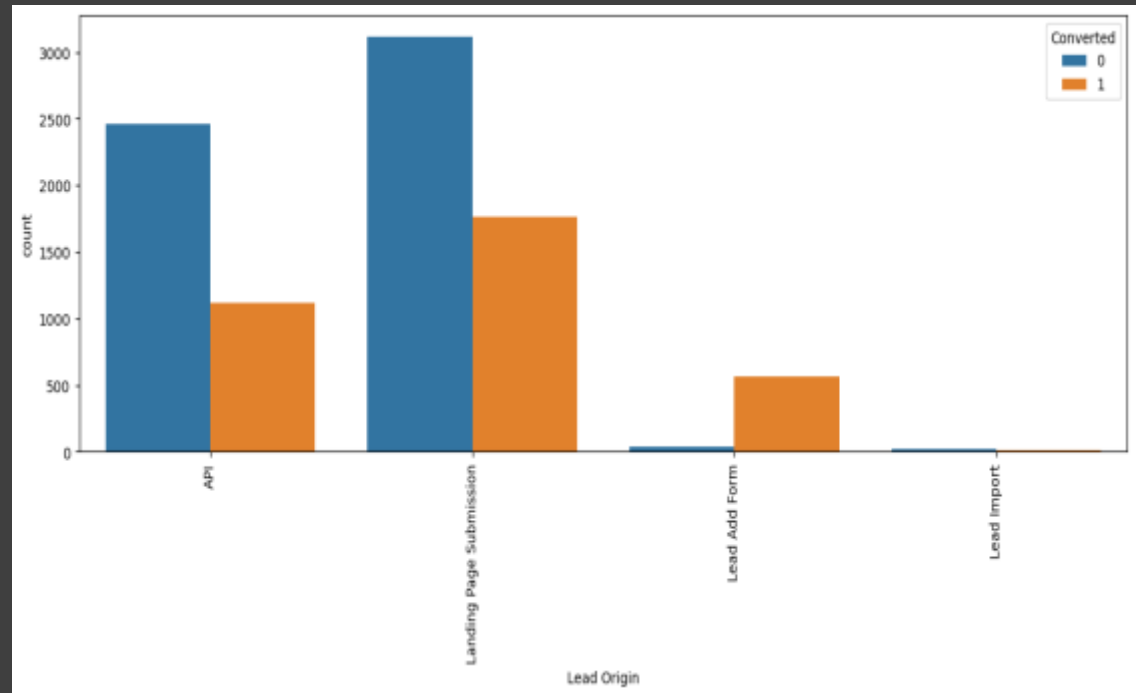


Last Activity:

- 68% of customers contribution in SMS Sent & Email Opened activities.
- '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.

Exploratory Data Analysis (EDA)

Categorical Variables

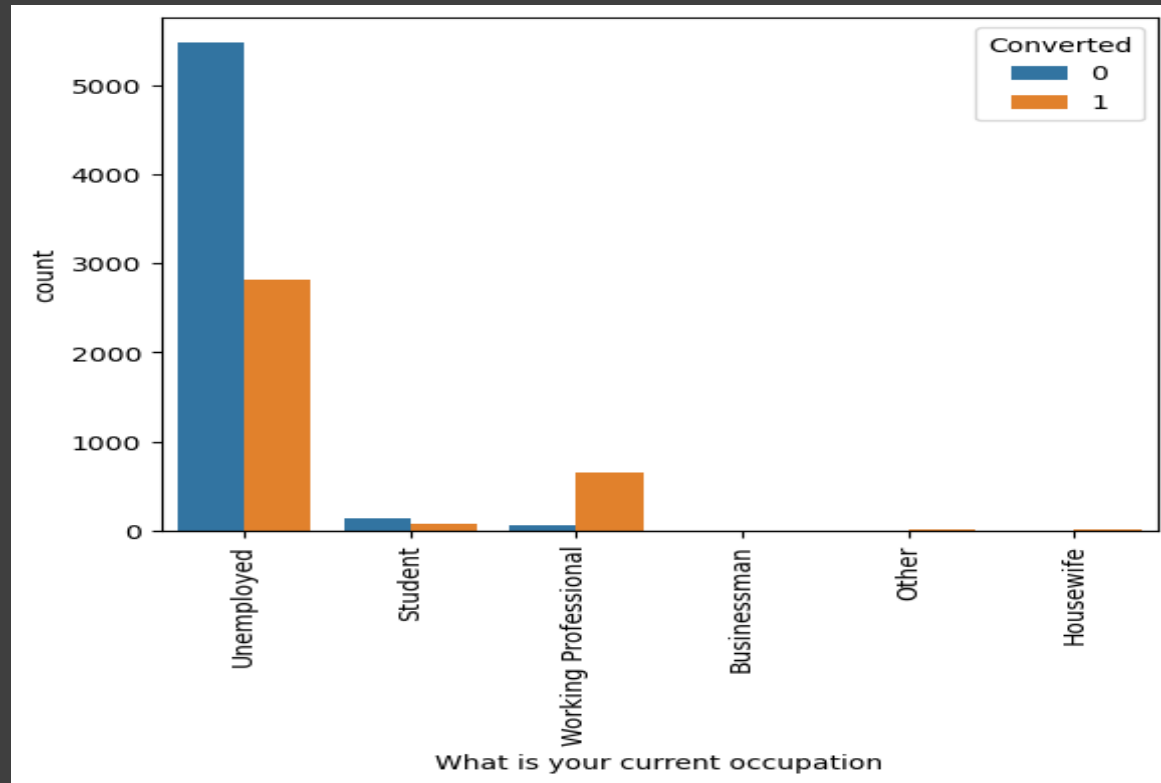


Lead Origin:

- "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- 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%.

Exploratory Data Analysis (EDA)

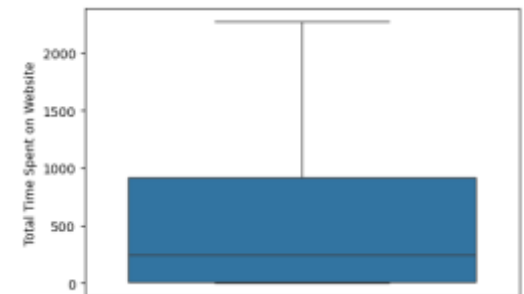
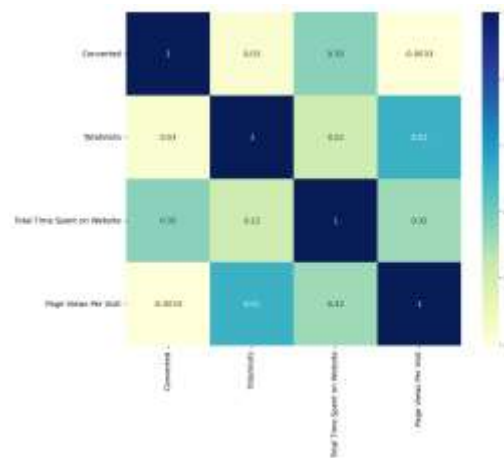
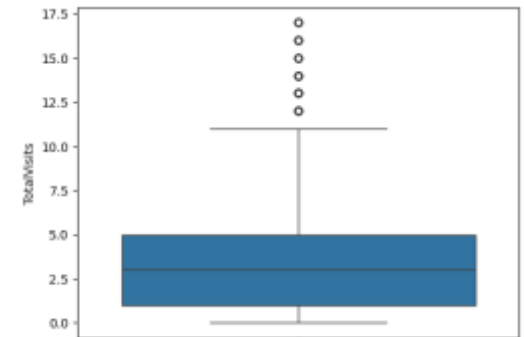
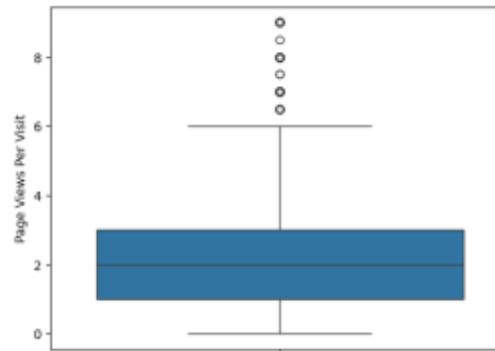
Categorical Variables



Current_occupation: It has 90% of the customers as Unemployed.

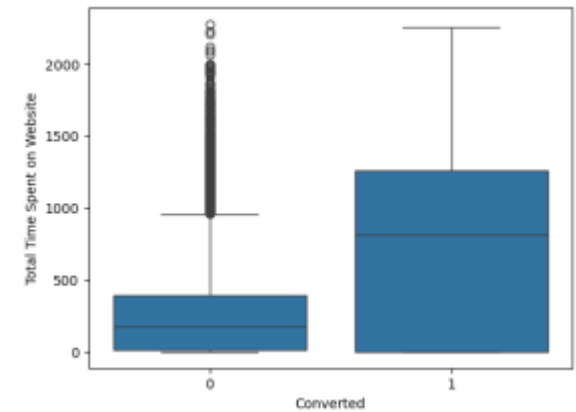
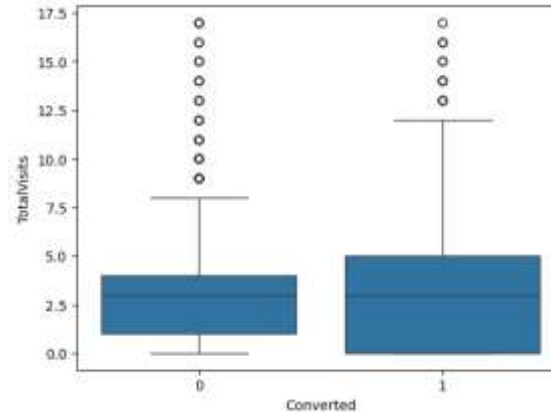
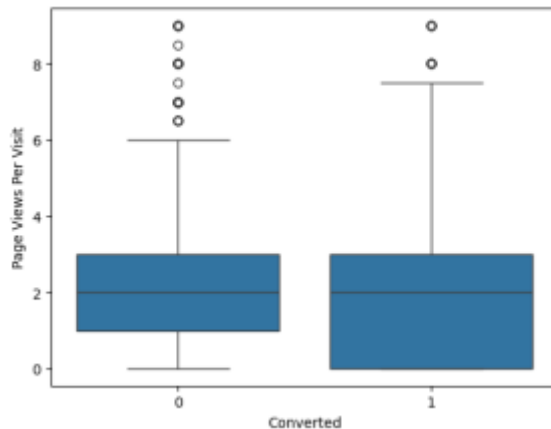
EDA – Bivariate Analysis for Numerical Variables

- **Total Visits:** There is a presence of outliers, Treatment done by remove top & bottom 1% of the Column Outlier values.
- **Total Time Spent on Website** : Since there are no major Outliers for the above variable we don't do any Outlier Treatment for this above Column.
- **Page Views Per Visit** : There is a presence of outliers. Treatment done by Remove top & bottom 1%.



EDA – Bivariate Analysis for Numerical Variables

- **Total Visits VS converted** : Median for converted and not converted leads are the same. Nothing conclusive can be said based on Total Visits.
- **Total Time Spent on Website VS Converted** : Leads spending more time on the website are more likely to be converted. Website should be made more engaging to make leads spend more time.
- **Page Views Per Visit VS Converted** : Median for converted and unconverted leads is the same. Nothing can be said specifically for lead conversion from Page Views Per Visit





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, What is your current occupation, Lead Source, Last Activity, Specialization, Last Notable Activity and Tags.
- Converted the list of boolean columns to Integers.
- **Splitting Train & Test Sets:**
 - 70:30 % ratio was chosen for the split
- **Feature scaling**
 - Standardization method was used to scale the features



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 – 52 columns & Post RFE – 15 columns



Model Building

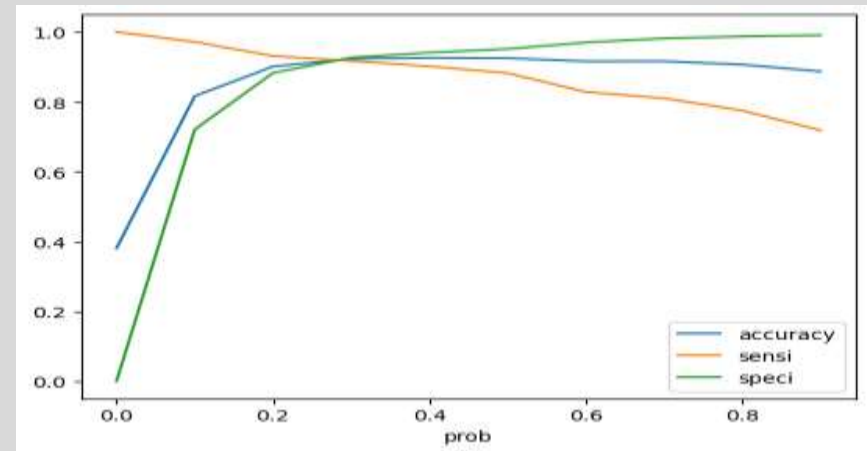
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 3 looks stable after four iteration with: 0 significant p -values within the threshold (p -values < 0.05)
- No sign of multicollinearity with VIFs less than 5
- Hence, logm3 will 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

Optimal Cutoff Point

From the curve Below,
0.3 is the optimum
point to take it as a
cutoff probability

- Accuracy sensitivity and specificity :

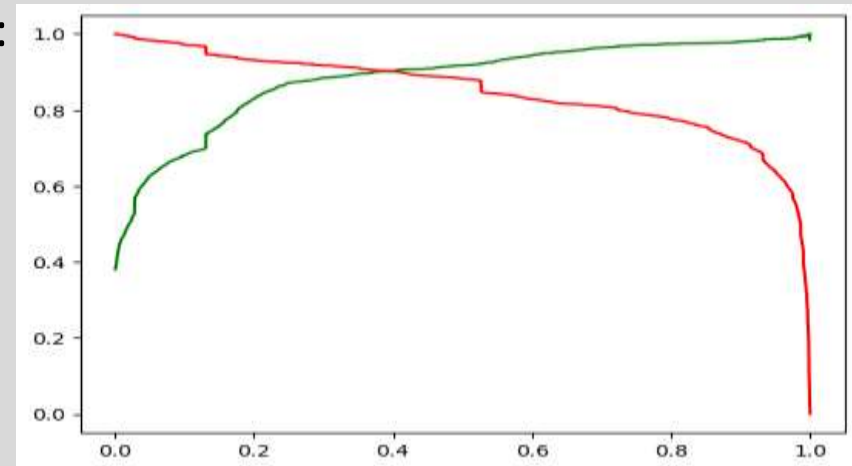


We have the following values for the Train Data:

- Accuracy : 90.81%
- Sensitivity : 92.05%
- Specificity : 90.10%

Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value, Negative Predictive Values, Precision & Recall.

- Precision and recall :

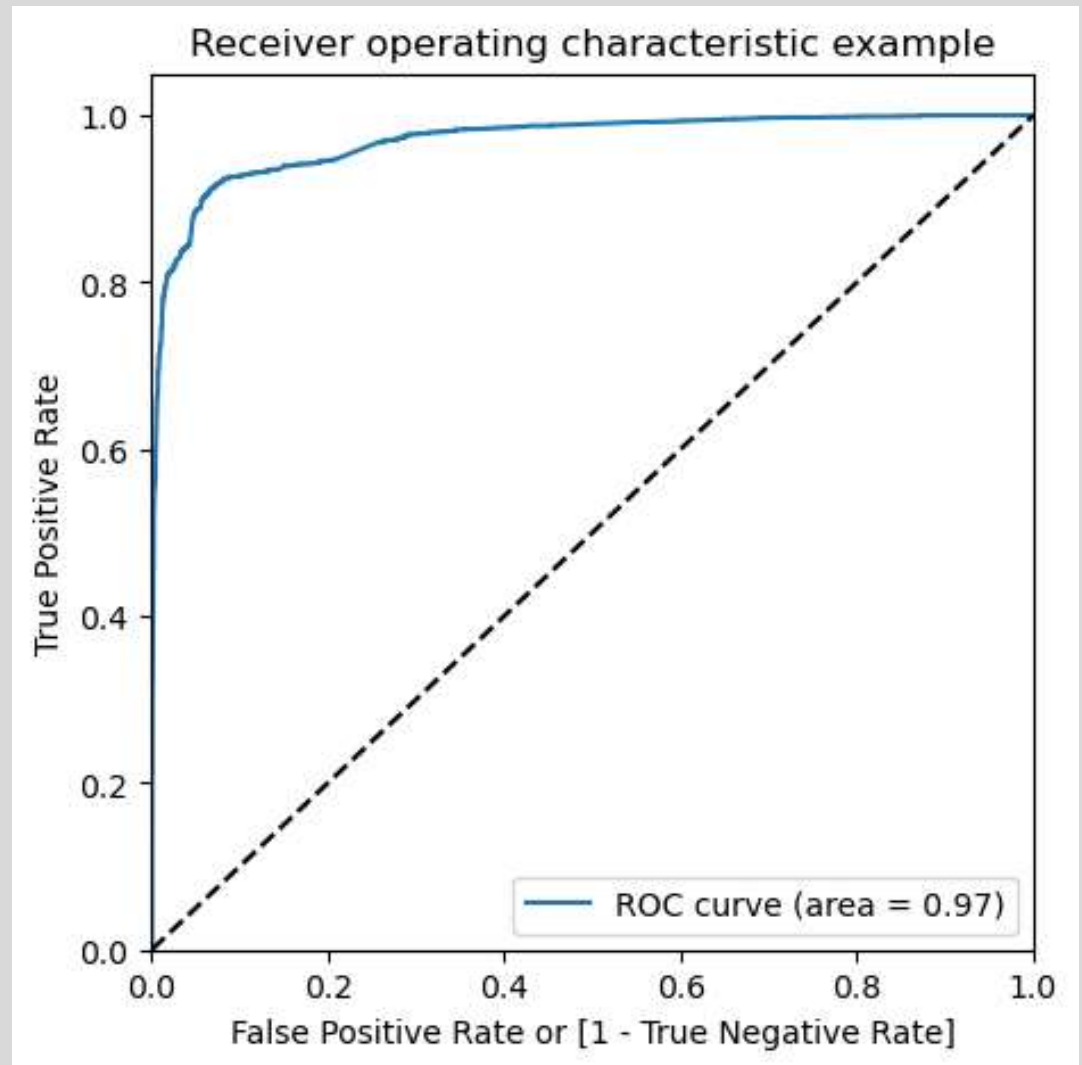


Model Evaluation –

Train Data Set

ROC CURVE :

The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.



Metric	Train Data	Test Data
Accuracy	89.15%	90.81%
Sensitivity	91.98%	92.05%
Specificity	93.25%	90.10%

Comparison

The model performs well on both the training and test datasets, demonstrating high accuracy, sensitivity, and specificity.

This indicates that the model predicts the conversion rate effectively, giving the CEO confidence to make informed decisions based on these results.



Summary

1. Lead scoring case study for X Education has been done using a logistic regression model to meet the constraints as business requirements.
2. The regression model aims to identify key factors influencing lead conversion and predict the probability of lead conversion.
3. The dependent variable is whether a lead converts (1) or not (0).
4. In the initial stage of data preparation there are a lot of leads. But few of them are converted to paying customers.
5. Most number of leads are from country India and Mumbai city has the most number of leads.
6. Certain columns include a value labelled "Select," which indicates that the student has not made a selection for that specific option.
7. The high number of total visits & Total time spent on the platform may increase the chances of lead being converted.
8. The leads have joined courses for Better Career Prospects, most of them having Specialization in Finance Management. Leads from HR, Finance & marketing management specializations have a high probability of conversion.
9. Talking to the last notable Activity, making improvements in customer engagement through email & calls will help to convert leads. As the leads that are opening emails have a high probability of converting, Same as Sending SMS will also benefit.
10. Most of the leads' current occupation is Unemployed, which means gave more focus on unemployed leads.