Lead Scoring Case Study

Group Member: 1)Simran Deshmukh

2) Prashant Rai Bhumihar

3) Tejesh Suresh

4) Aeman Zahoor Masoodi

Problem Statement

- XEducation specializes in selling online courses tailored for industry professionals. However, the company faces a significant challenge with its low lead conversion rate. For example, out of 100 leads acquired daily, only about 30 are converted into paying customers.

- To address this, X Education aims to identify its most promising leads, referred to as Hot Leads. By focusing its sales efforts on these high-potential prospects, the company hopes to boost its lead conversion rates and overall efficiency.

Objective

- The primary goal is to build a model that identifies potential hot leads and deploy this model for future use

Solution Approach

The solution is implemented through the following steps:

1. Data Cleaning and Preparation:

- Duplicate Data: Identified and handled duplicates.
- MissingValues:Addressed missing data by dropping unnecessary columns or imputing values as needed.
- Irrelevant Features: Removed columns that either had low variance or were not useful for analysis, such as "Do Not Call," "DigitalAdvertisement," etc.
- Outliers: Reviewed and managed outliers to maintain data quality

2. Exploratory DataAnalysis (EDA)

- UnivariateAnalysis: Examined individual variables for their value distributions.
- Bivariate Analysis: Assessed relationships between variables, using correlation coefficients and patterns

3. DataTransformation

- Feature Scaling: Normalized numerical variables.
- Encoding: Created dummy variables for categorical data

4. Model Development

- DataSplitting: Divided data into training (70%) and testing (30%) sets.
- Feature Selection: Used Recursive Feature Elimination (RFE) to identify the top 19 variables.
- ModelRefinement: Iteratively removed variables with high p-values (>0.05) or high Variance Inflation Factors (VIF > 5).
- ModelAccuracy:Achieved an overall accuracy of 80%

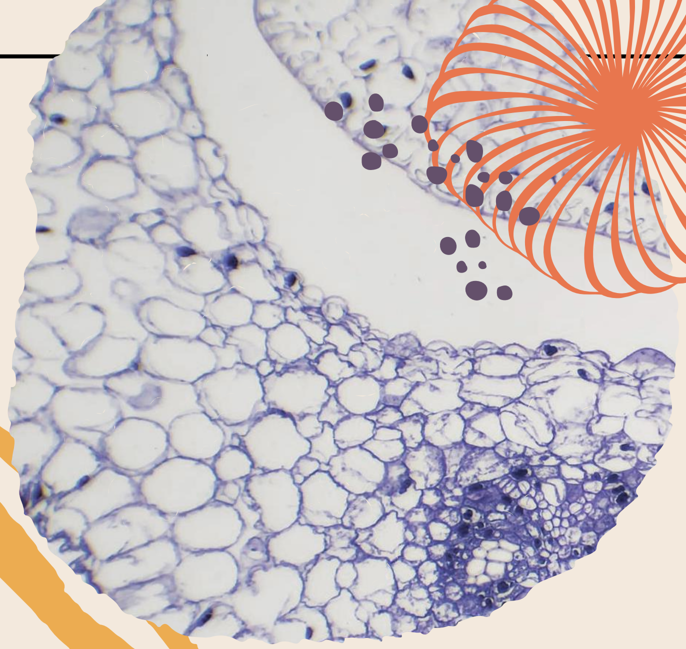
5. ModelValidation

- ROC Curve:Used the Receiver Operating Characteristic (ROC) curve to identify the optimal cutoff point for balanced sensitivity and specificity
- Optimal Cutoff Probability: Determined to be 0.86

Data Manipulation

1. Initial Dataset Overview there are Rows:37 ,Columns:9,240.
2. Eliminating Features with Single Values

- Certain columns that contained only one unique value and provided no meaningful variation were removed. Examples include:
- "Magazine"
- "Receive More Updates About Our Courses"
- "Update Meon Supply"
- "Chain Content"
- "Get Updates on DM Content"
- "I Agree to Pay the Amount Through Cheque"



3. Removing Irrelevant Identifiers

- Columnssuchas "Prospect ID" and "Lead Number," which did not contribute to the analysis, were discarded.

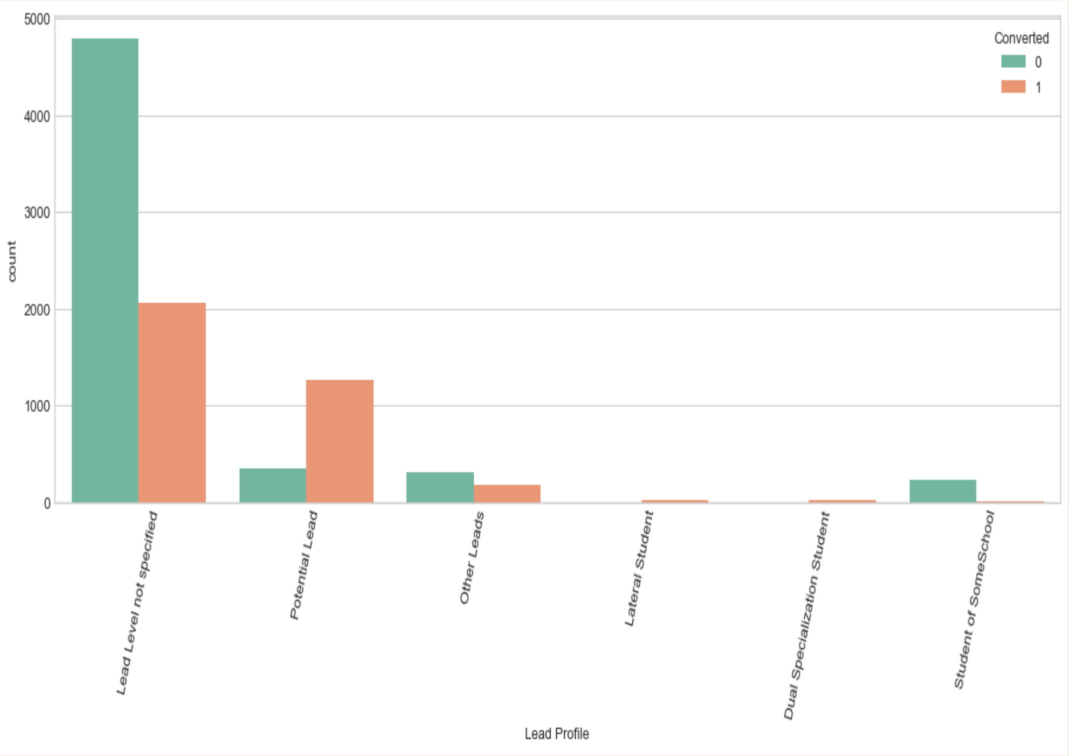
4. Dropping Low-Variance Features

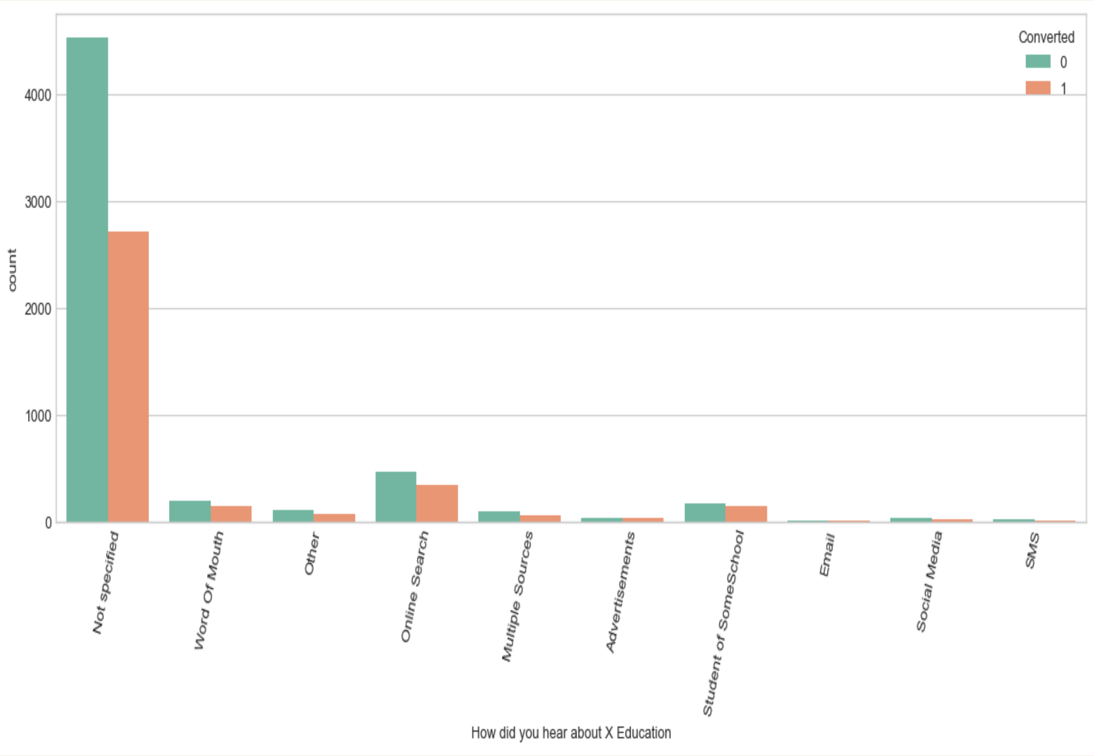
- Features with minimal variability were identified and excluded. These included:
- "DoNotCall"
- "What Matters Most toYouin Choosing a Course"
- "Search"
- "NewspaperArticle"
- "XEducationForums"
- "Newspaper"
- "DigitalAdvertisement"

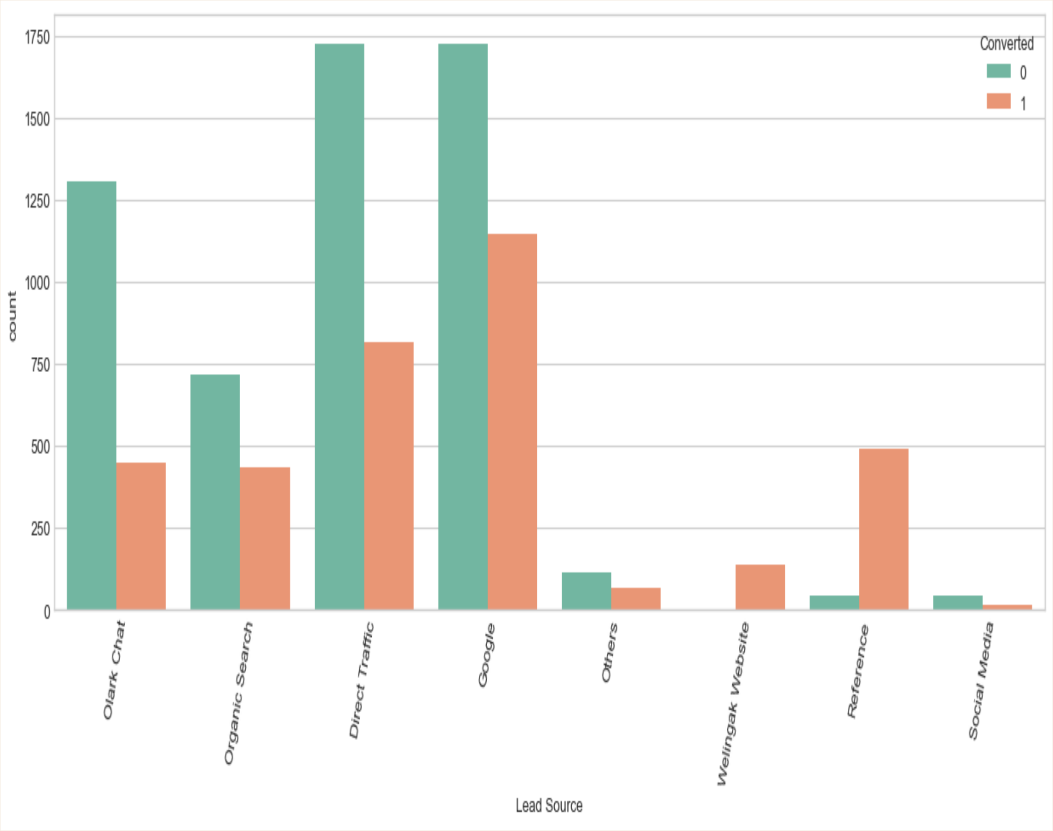
5. Handling Columns with Excessive Missing Data

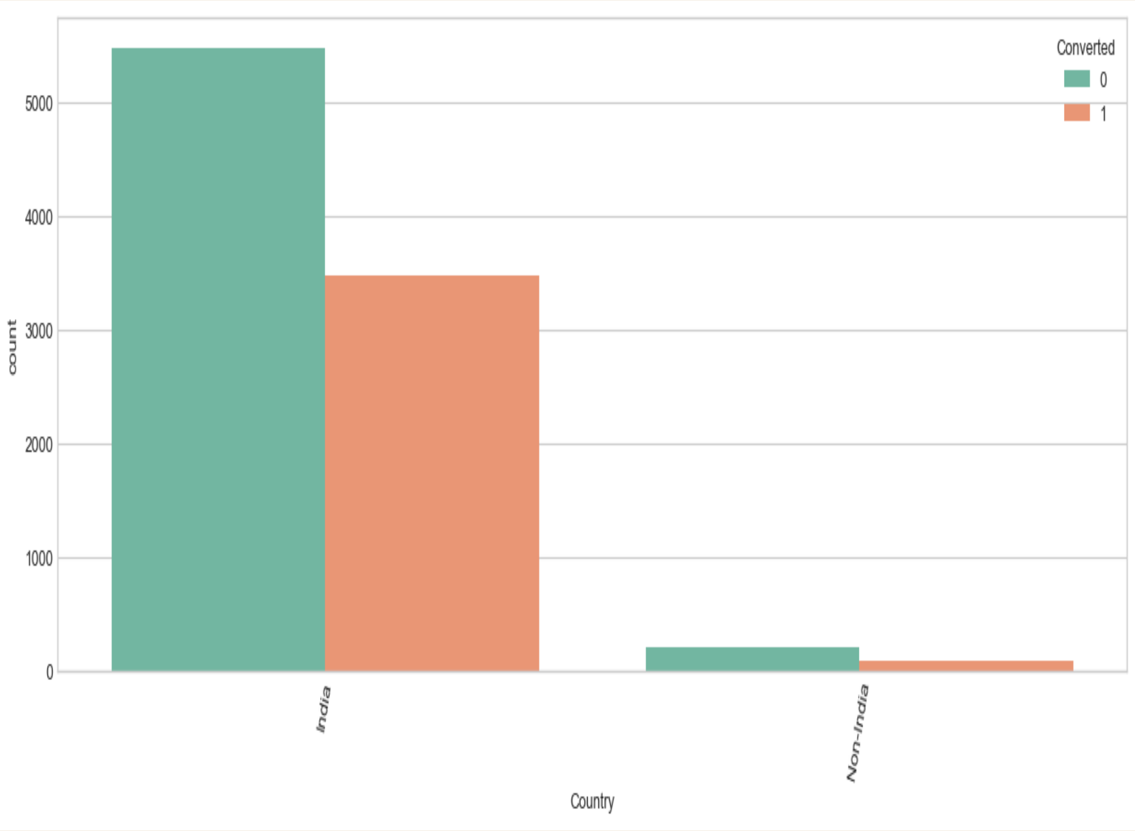
- Columns with over 35%missing values were dropped, including:
- "How Did You Hear About X Education"
- "Lead Profile“
- By applying these steps, the dataset was significantly streamlined, ensuring only relevant and high-quality data was retained for further analysis.

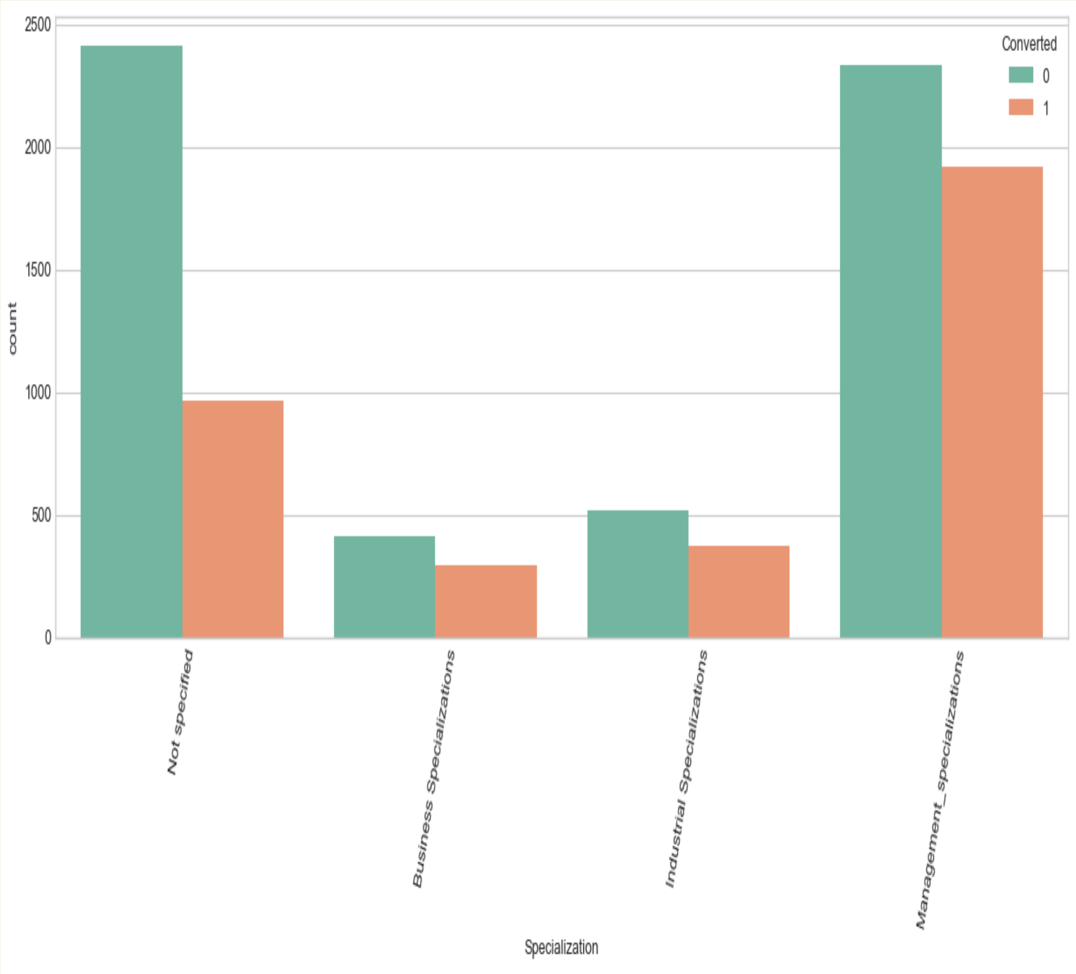
EDA

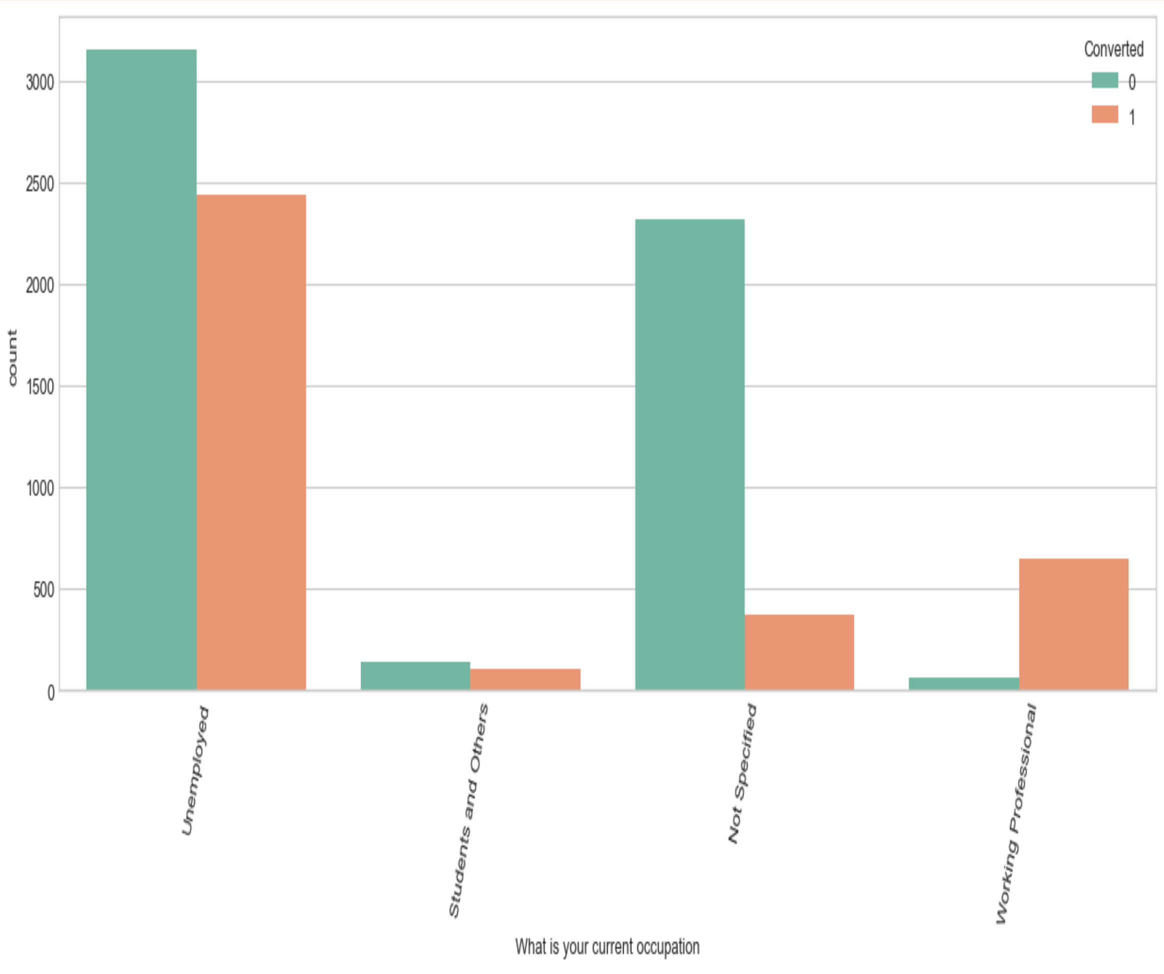

Converted
🟩 0
🟧 1
count
5000
4000
3000
2000
1000
0
Lead Level not specified
Potential Lead
Other Leads
Lateral Student
Dual Specialization Student
Student of SomeSchool
Lead Profile

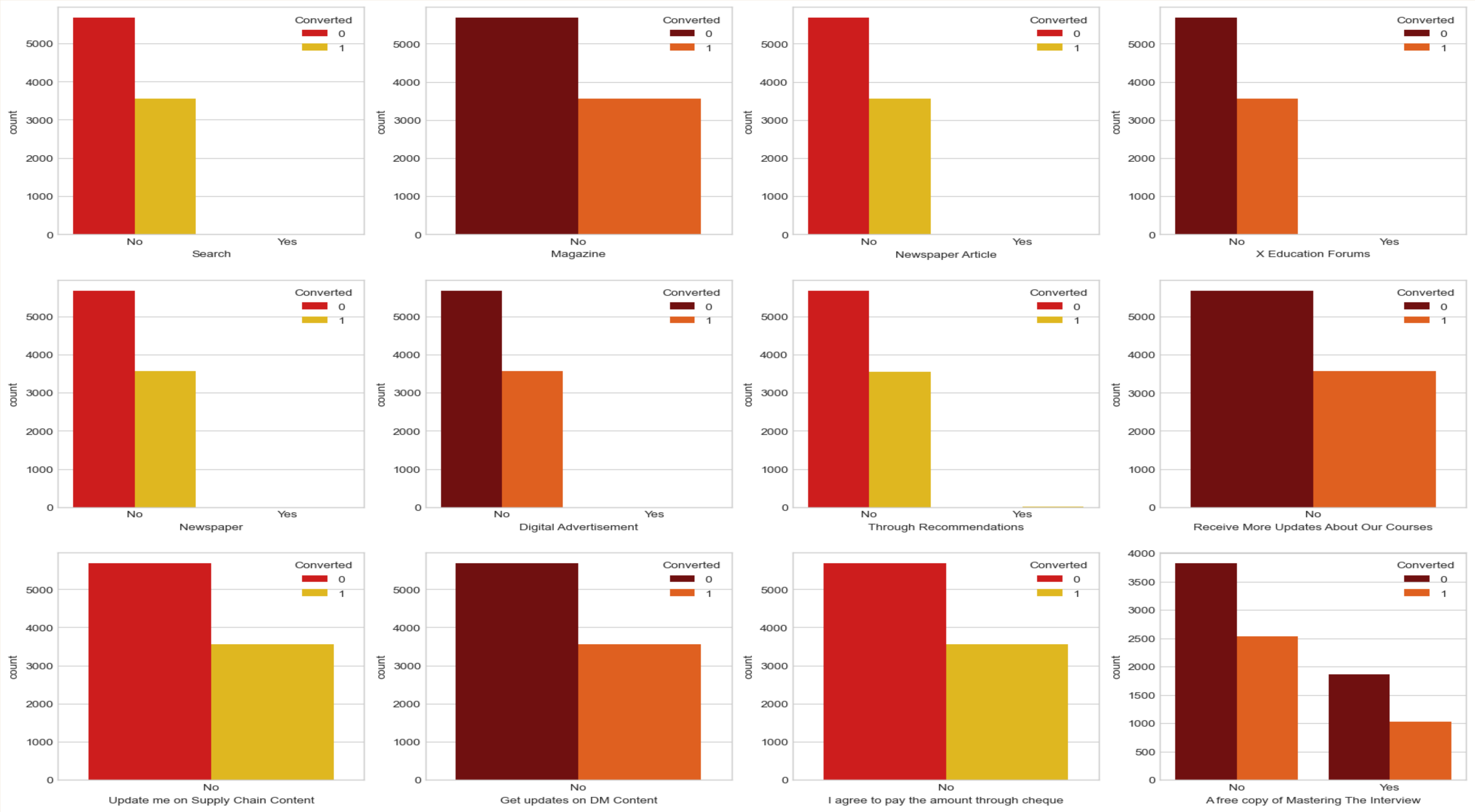

Converted
🟩 0
🟧 1
count
4000
3000
2000
1000
0
Not specified
Word Of Mouth
Other
Online Search
Multiple Sources
Advertisements
Student of SomeSchool
Email
Social Media
SMS
How did you hear about X Education


Converted
🟩 0
🟧 1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Others
Welingak Website
Reference
Social Media
Lead Source


Converted
🟩 0
🟧 1
count
5000
4000
3000
2000
1000
0
India
Non-India
Country


Converted
0
1
count
2500
2000
1500
1000
500
0
Not specified
Business Specializations
Industrial Specializations
Management_specializations
Specialization


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Students and Others
Not Specified
Working Professional
What is your current occupation


count
0
1000
2000
3000
4000
5000
Converted
0
1
No
Yes
Search
Magazine
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendations
Receive More Updates About Our Courses
Update me on Supply Chain Content
Get updates on DM Content
I agree to pay the amount through cheque
A free copy of Mastering The Interview
500
1500
2500
3500

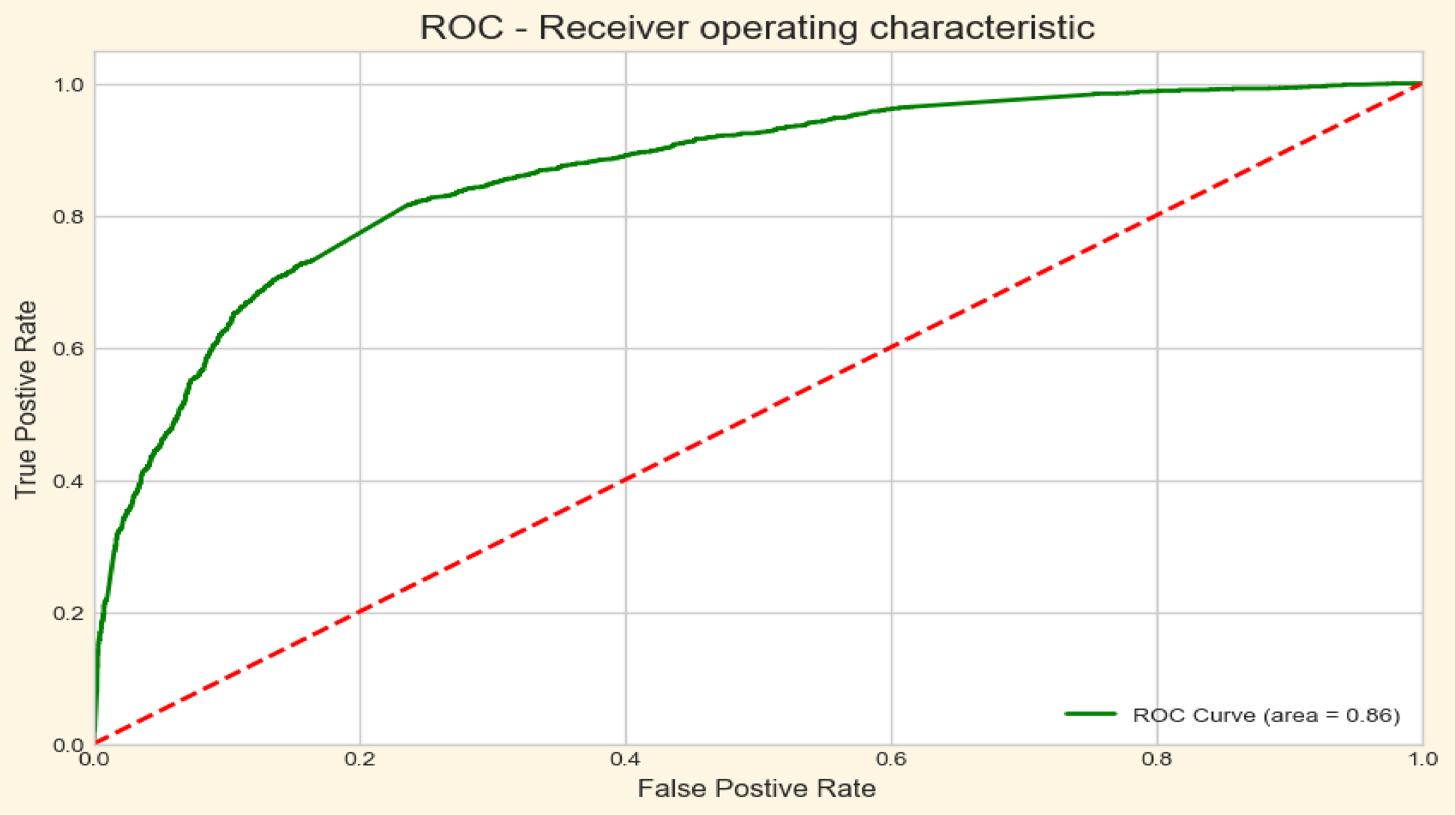
Data Conversion

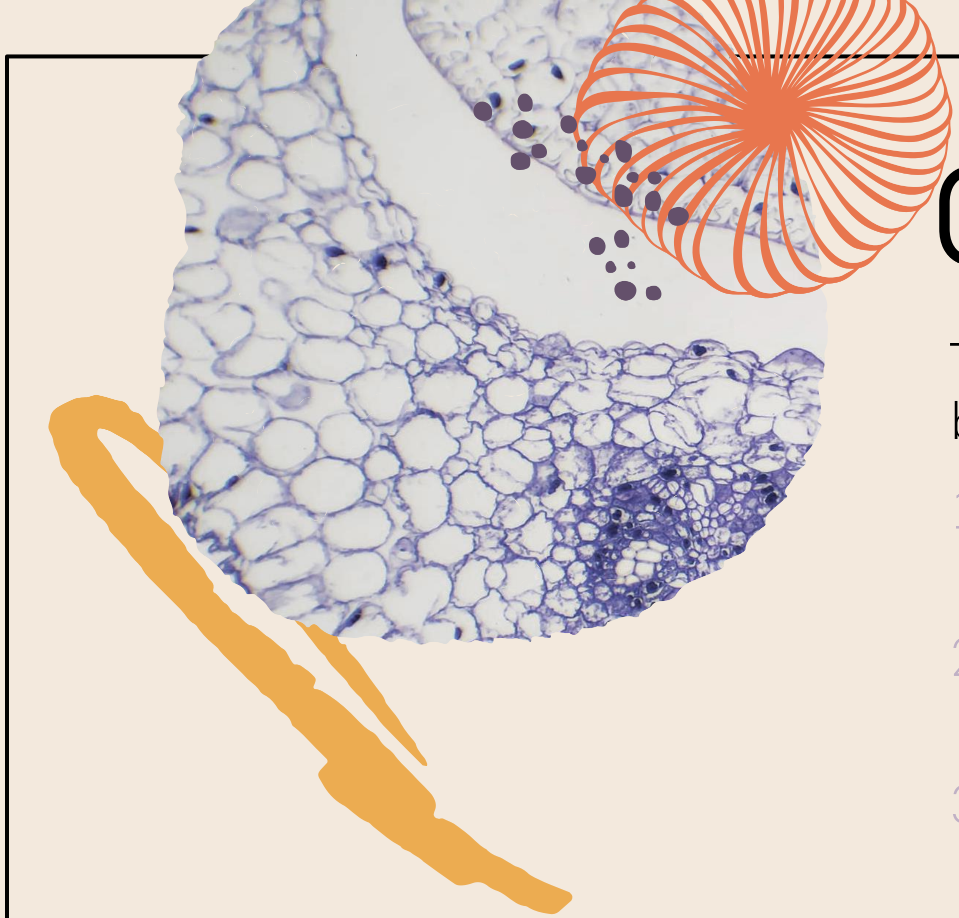
- Numerical Data:All numeric variables have been normalized to ensure consistent scaling and comparability.
- Categorical Data: Dummy variables were created to represent non-numeric (object type) categories, converting them into a usable format for analysis.
- Dataset Overview:The dataset contains 8,792 rows (individual entries) and 43 columns (features) ready for analysis

Building Model

- Data Splitting: The dataset was divided into training (70%) and testing (30%) sets for model evaluation.
- Feature Selection: Recursive Feature Elimination (RFE) was used to narrow down the features, selecting the 15 most relevant variables.
- Model Optimization:The model was refined by systematically removing variables with a p-value > 0.05 (indicating low statistical significance) or a VIF > 5 (indicating multicollinearity).
- Predictions: The refined model was tested on the test dataset to make predictions.
- Performance:The model achieved an overall accuracy of 80% on the test data.

ROC Curve


ROC - Receiver operating characteristic
1.0
0.8
0.6
0.4
0.2
0.0
True Postive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Postive Rate
ROC Curve (area = 0.86)



Conclusion

The analysis revealed key factors influencing potential buyers, listed in order of importance:

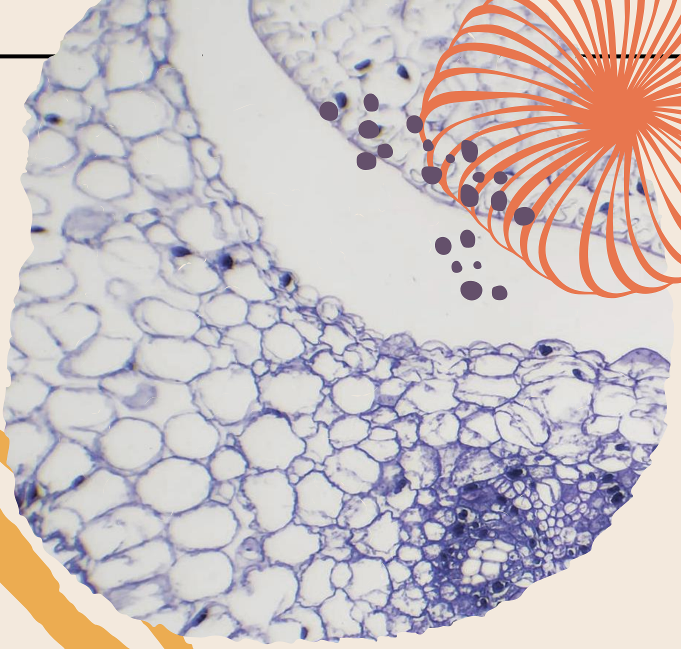
1. Time Spent on the Website:The more time users spend on the website, the higher the likelihood of conversion.
2. Number ofVisits: Frequent visits indicate stronger interest.
3. Lead Sources: Buyers are more likely to convert when leads come from:

   Google

   DirectTraffic

   OrganicSearch

   TheWelingkarWebsite



4. LastActivity: Certain actions like:

   Receiving an SMS

   Engaging in an Olark Chat Conversation

   show strong potential for conversion.

5. Lead Origin: Leads originating from LeadAd Forms are highly valuable.

6.Occupation: Working Professionals are significantly more likely to purchase courses.

By focusing on these key variables, X Education can enhance its strategies and effectively convert a majority of potential buyers into actual customers, driving growth and success

Thank you