Lead Score

Case Study

Problem statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business objective

- X Education aims to identify its most promising leads.
- To achieve this, they plan to develop a model that detects high-potential leads.
- The model will be deployed for ongoing use in the future.

Methodology

- Perform data cleaning and manipulation:
  - Address duplicate data.
  - Handle missing or “SELECT” values.
  - Drop columns with excessive missing values and limited relevance for analysis.
  - Impute missing values where necessary.
  - Identify and manage outliers.

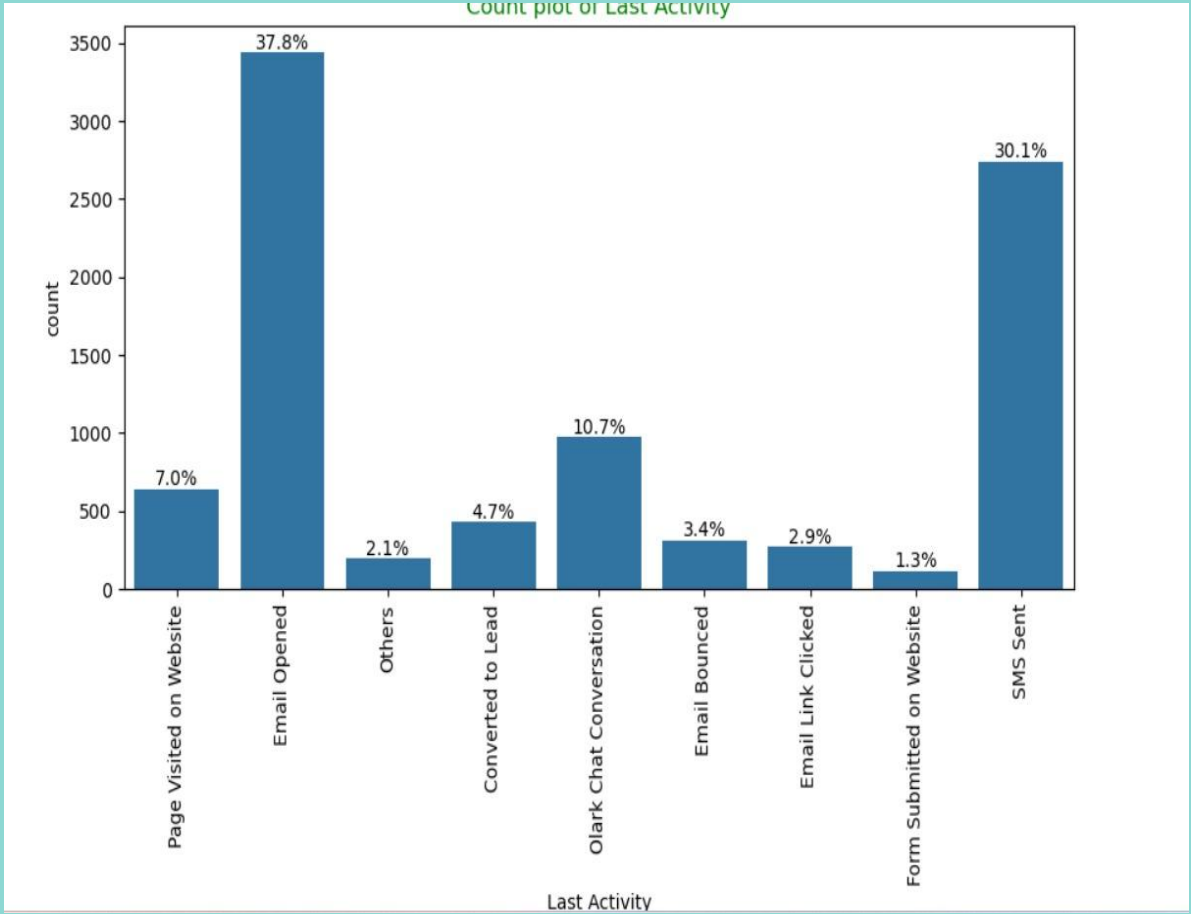
- Conduct exploratory data analysis (EDA):
  - Perform univariate analysis, including value counts and variable distribution.
  - Carry out bivariate analysis, examining correlations and patterns between variables.

- Apply feature scaling, dummy variables, and data encoding.
-
- Use logistic regression for model building and prediction.
-
- Validate the model, present results, and provide conclusions and recommendations.

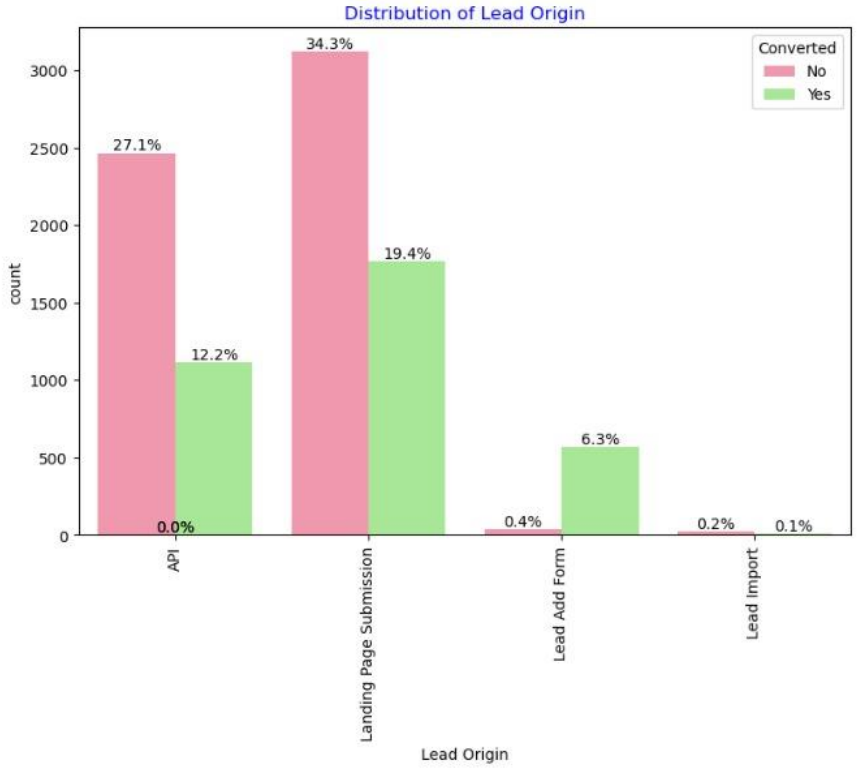
Data Manipulation

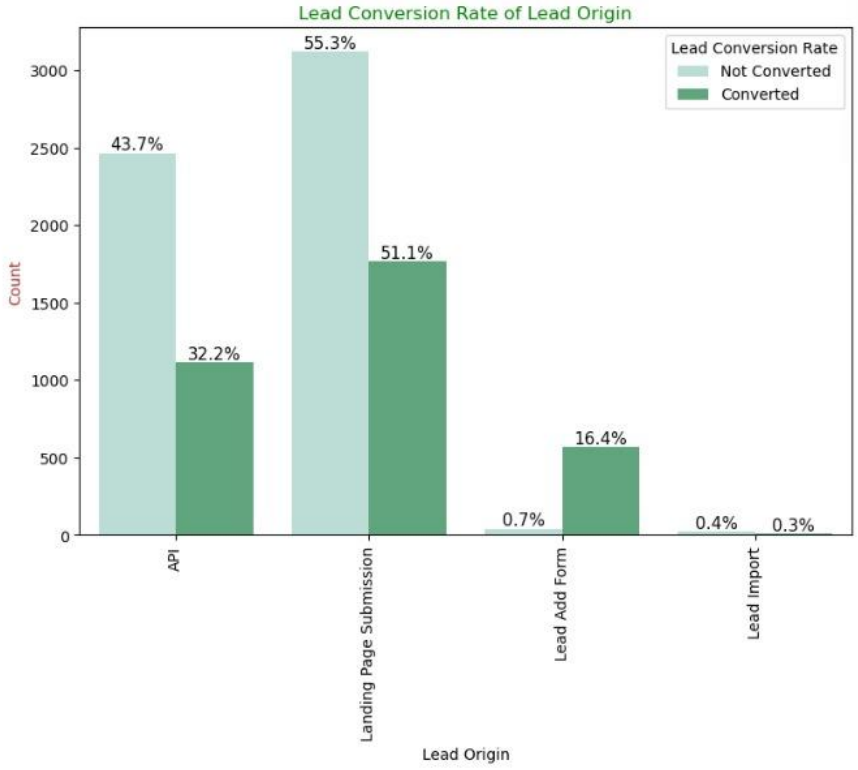
- The dataset has 37 rows and 9240 columns. Features with only a single value, like "Magazine," "Receive More Updates About Our Courses," "Update me on Supply," "Chain Content," and "I agree to pay the amount through cheque," were removed.
-
- Unnecessary columns like "Prospect ID" and "Lead Number" were also dropped as they are not relevant for analysis.
-
- After reviewing the value counts for certain object-type variables, we removed features with little variation, such as "Do Not Call," "What matters most to you in choosing a course," "Search," "Newspaper Article," "X Education Forums," and "Digital Advertisement."
-
- Columns with more than 40% missing values, such as "How did you hear about X Education" and "Lead Profile," were also dropped.

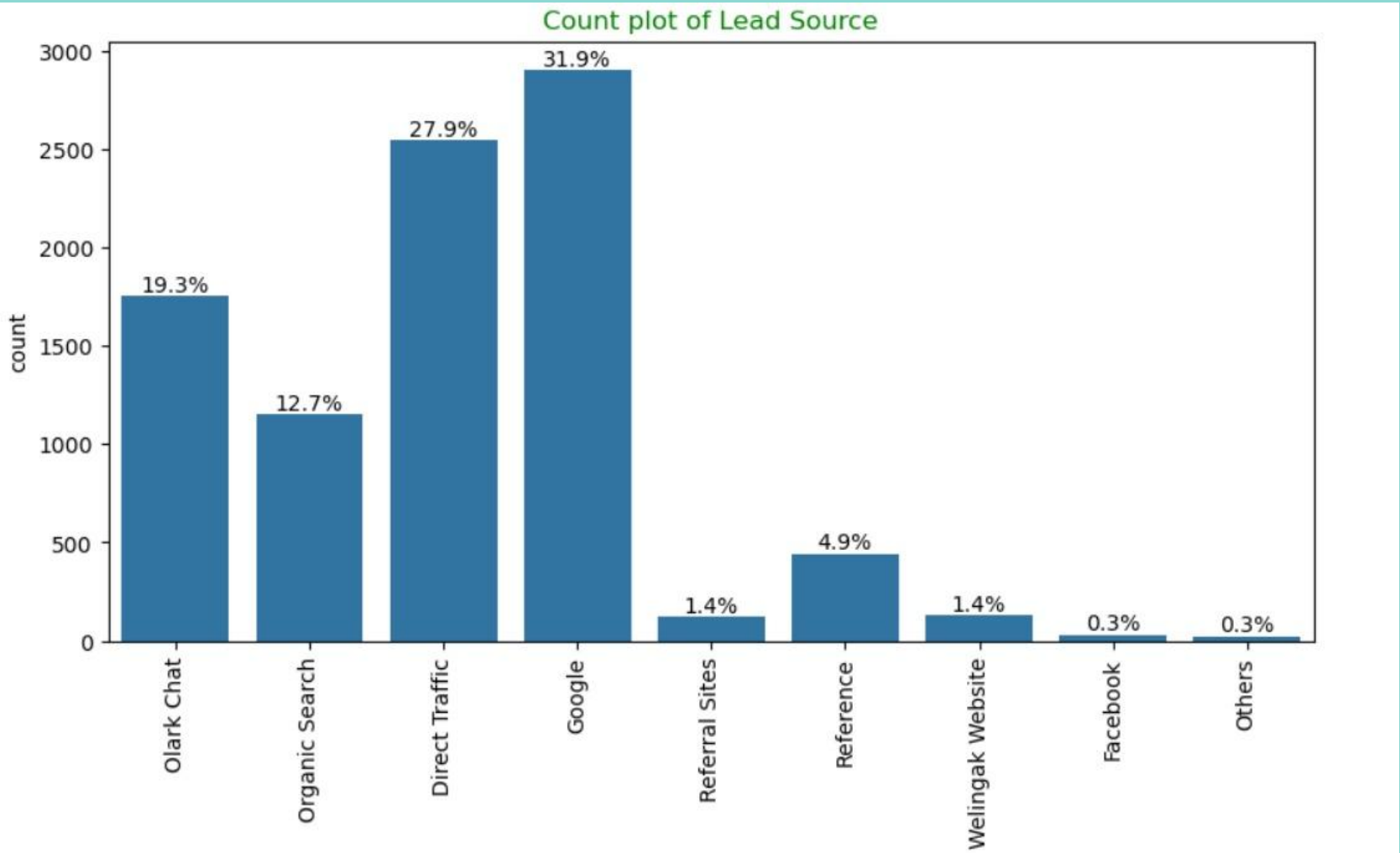
EDA


Count plot of Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
37.8%
30.1%
10.7%
7.0%
4.7%
3.4%
2.9%
2.1%
1.3%
Page Visited on Website
Email Opened
Others
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
SMS Sent
Last Activity

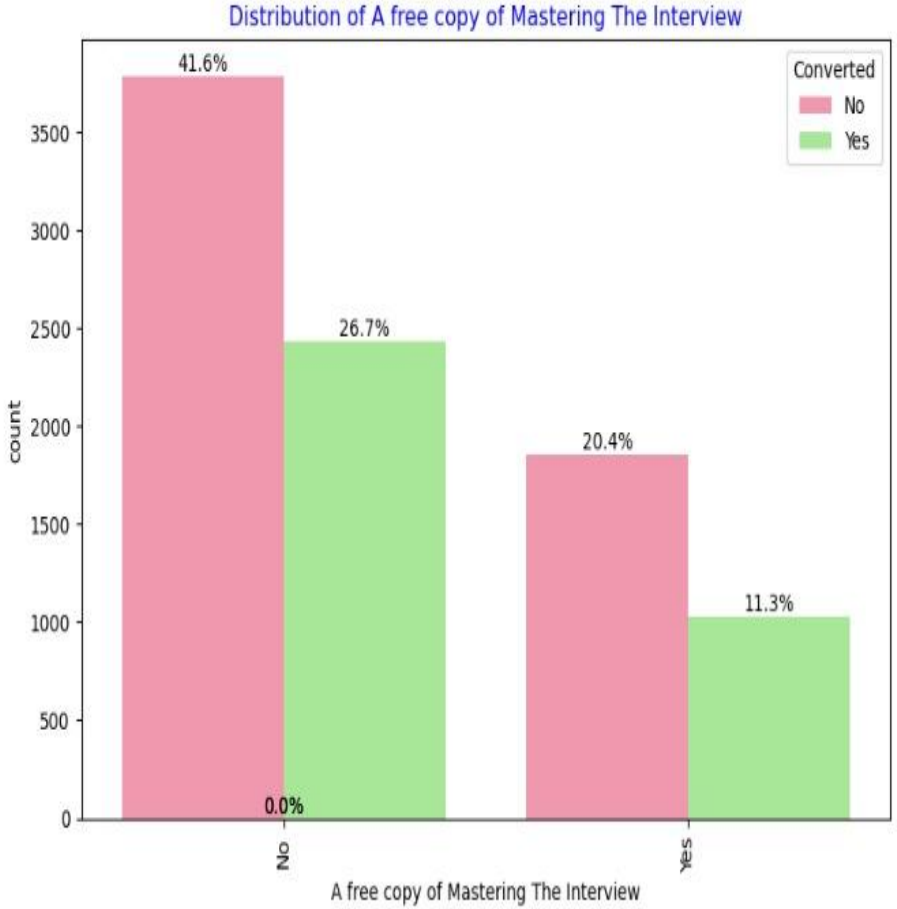
Lead Origin Countplot vs Lead Conversion Rates

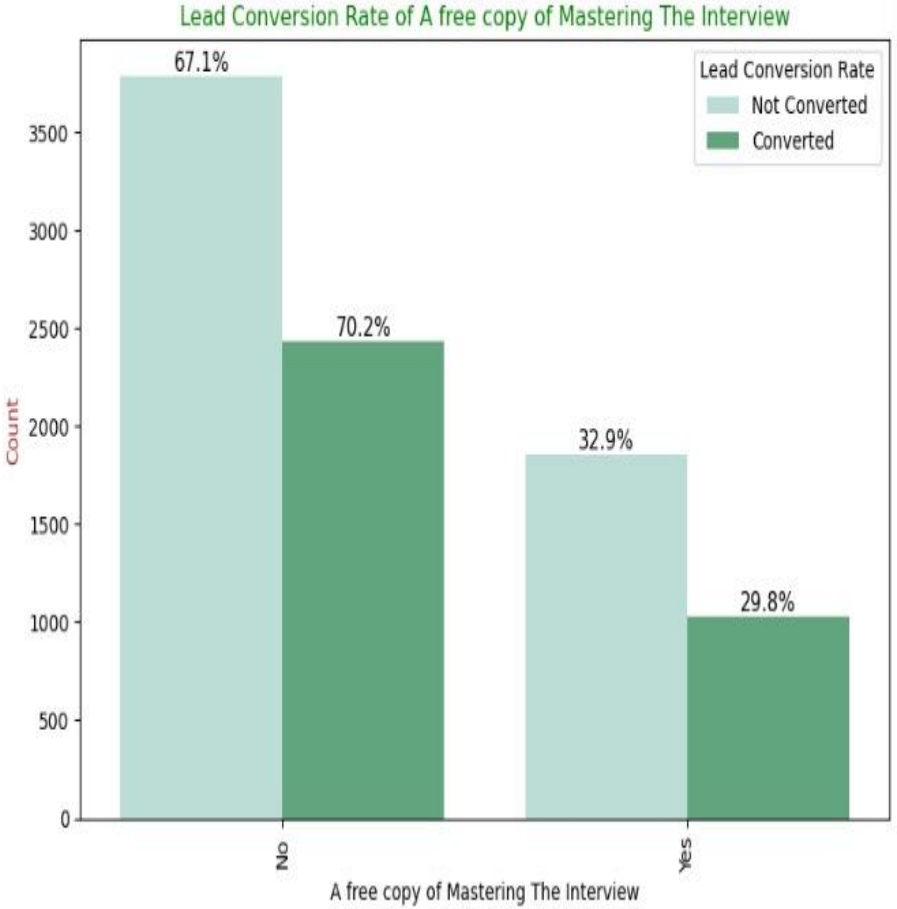




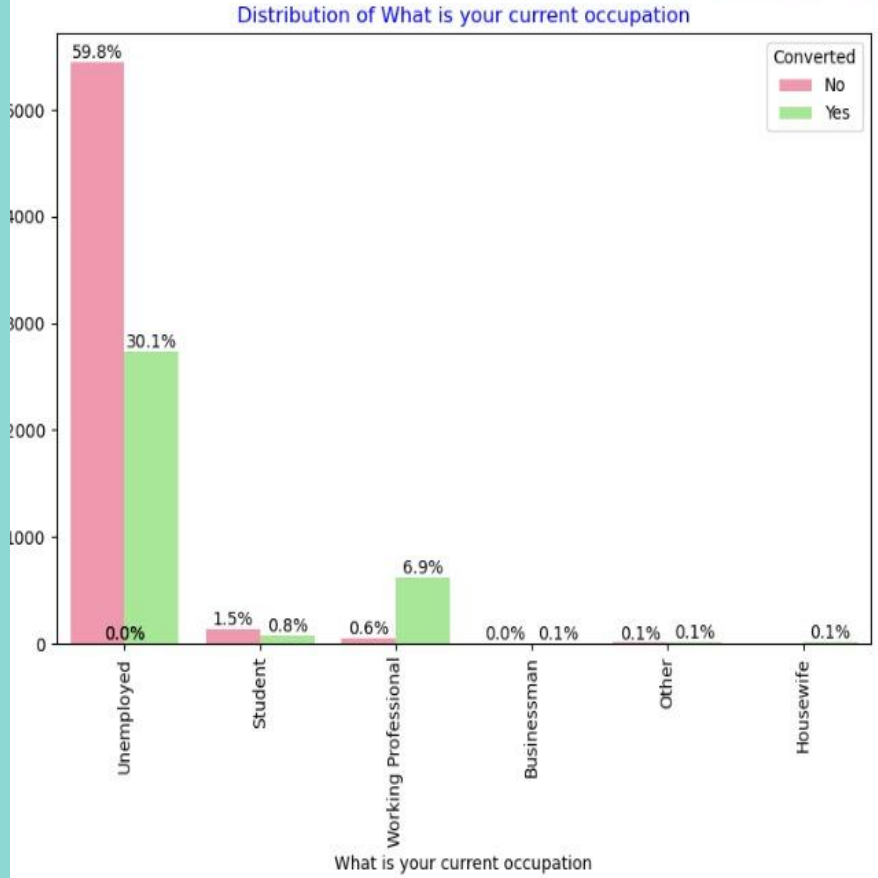

Count plot of Lead Source
count
3000
2500
2000
1500
1000
500
0
19.3%
12.7%
27.9%
31.9%
1.4%
4.9%
1.4%
0.3%
0.3%
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others

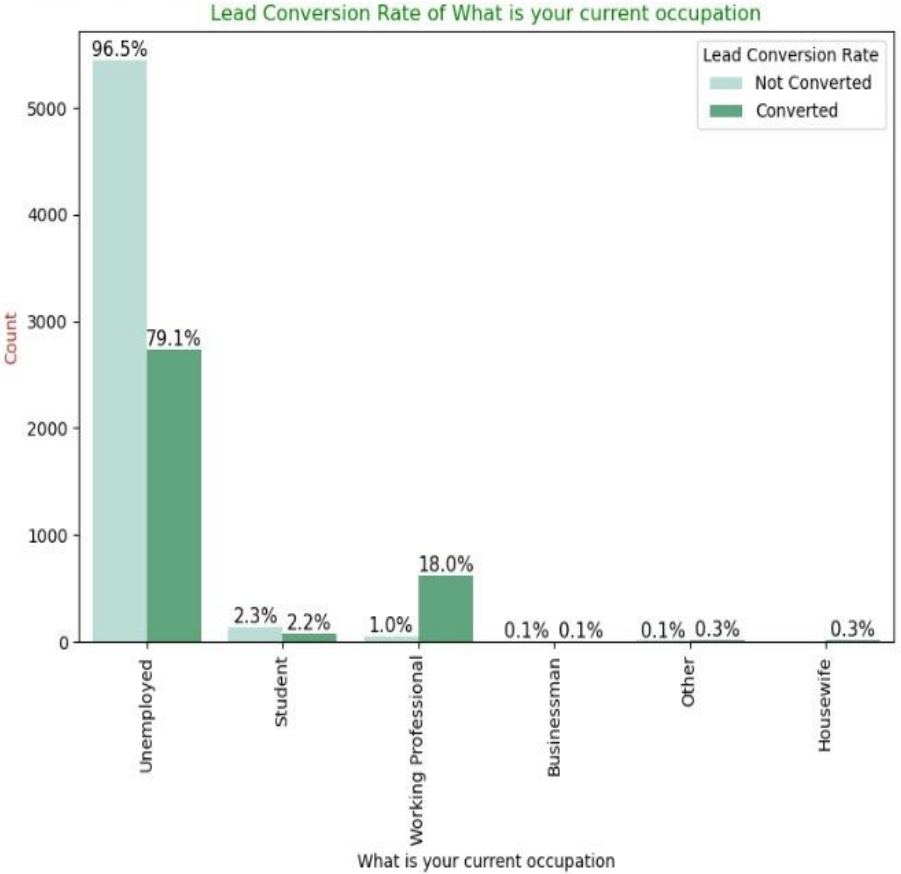
A free copy of Mastering The Interview Countplot vs Lead Conversion Rates





What is your current occupation Countplot vs Lead Conversion Rates

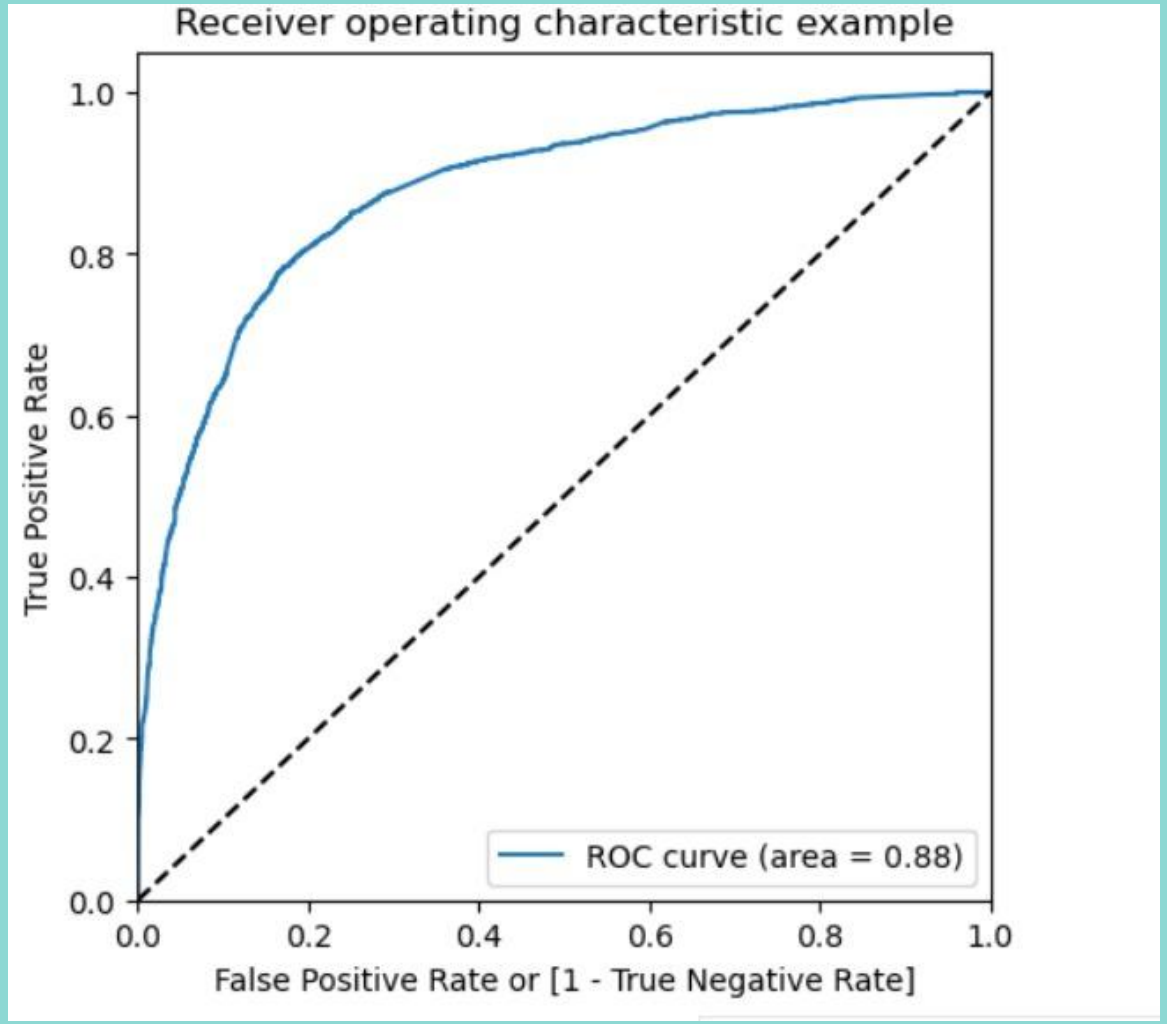


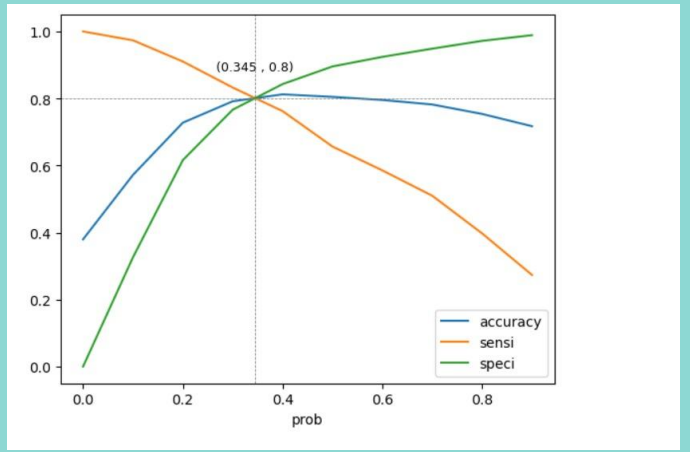


Model Building

- The data was split into training and testing sets, using a 70:30 ratio for regression analysis.
- Feature selection was done using RFE (Recursive Feature Elimination).
- RFE was run, and 15 key variables were selected.
- The model was built by removing variables with a p-value higher than 0.05 and a VIF (Variance Inflation Factor) greater than 5.
- The model was tested on the test dataset, achieving an overall accuracy of 80%.

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.88)
0.0
0.2
0.4
0.6
0.8
1.0


(0.345 , 0.8)
accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

The key factors that influence potential buyers, in order of importance, are:

- The total time spent on the website.
- The total number of visits.
- The lead source, especially when it's from: a. Google b. Direct traffic c. Organic search d. Welingak website
- The last activity, especially when it was: a. SMS b. Olark chat conversation
- When the lead origin is from a lead ad format.
- When the person’s current occupation is a working professional.

By focusing on these factors, X Education has a strong chance of convincing potential buyers to purchase their courses.