Lead Scoring Case Study

By

Yash Sahindrakar and Sagar Patil

Problem Statement

- X Education sells online courses to industry professionals. ⍰ X Education gets a lot of leads, but 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.

Objective

- X education is interested in learning about the most promising leads.
- They intend to create a model that detects hot leads for that purpose.
- Model deployment for future usage.

Methodology

➢ Data Cleaning and manipulation

1. Verify and deal with duplicate data.
2. Verify and handle missing and NA data.
3. Remove columns from the analysis if they have a significant number of missing data.
4. If required, value impugnation.

➢ EDA

1. Analysis of univariate data: value count, variable distribution, etc.
2. Bivariate data analysis: patterns between the variables and correlation coefficients, etc.Data encoding, feature scaling, and dummy variables.

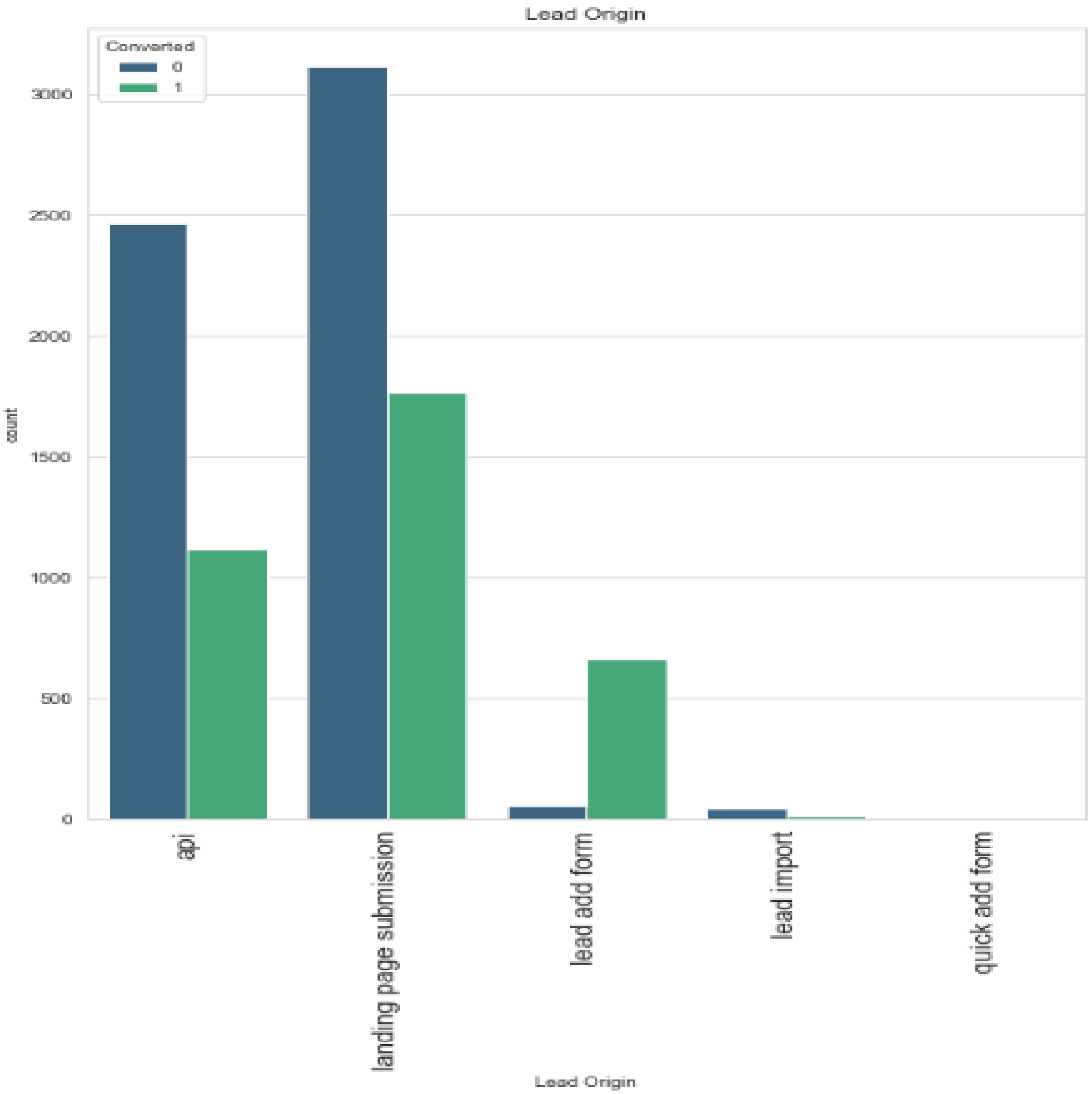
➢ Model building

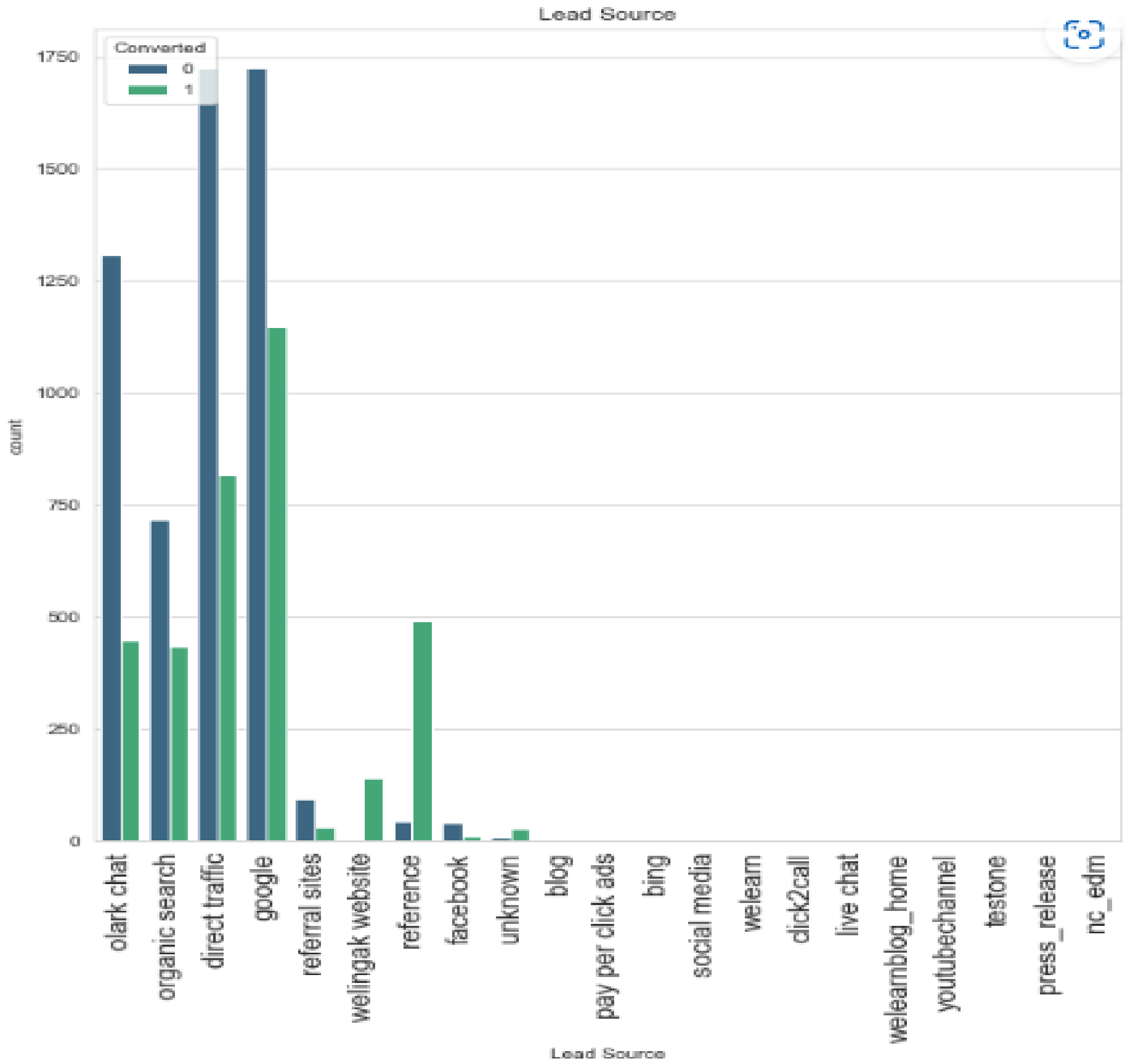
1. Model is built using the logistic regression approach.
2. Model Validation by checking various parameters.
3. Conclusions and recommendations based on the model.

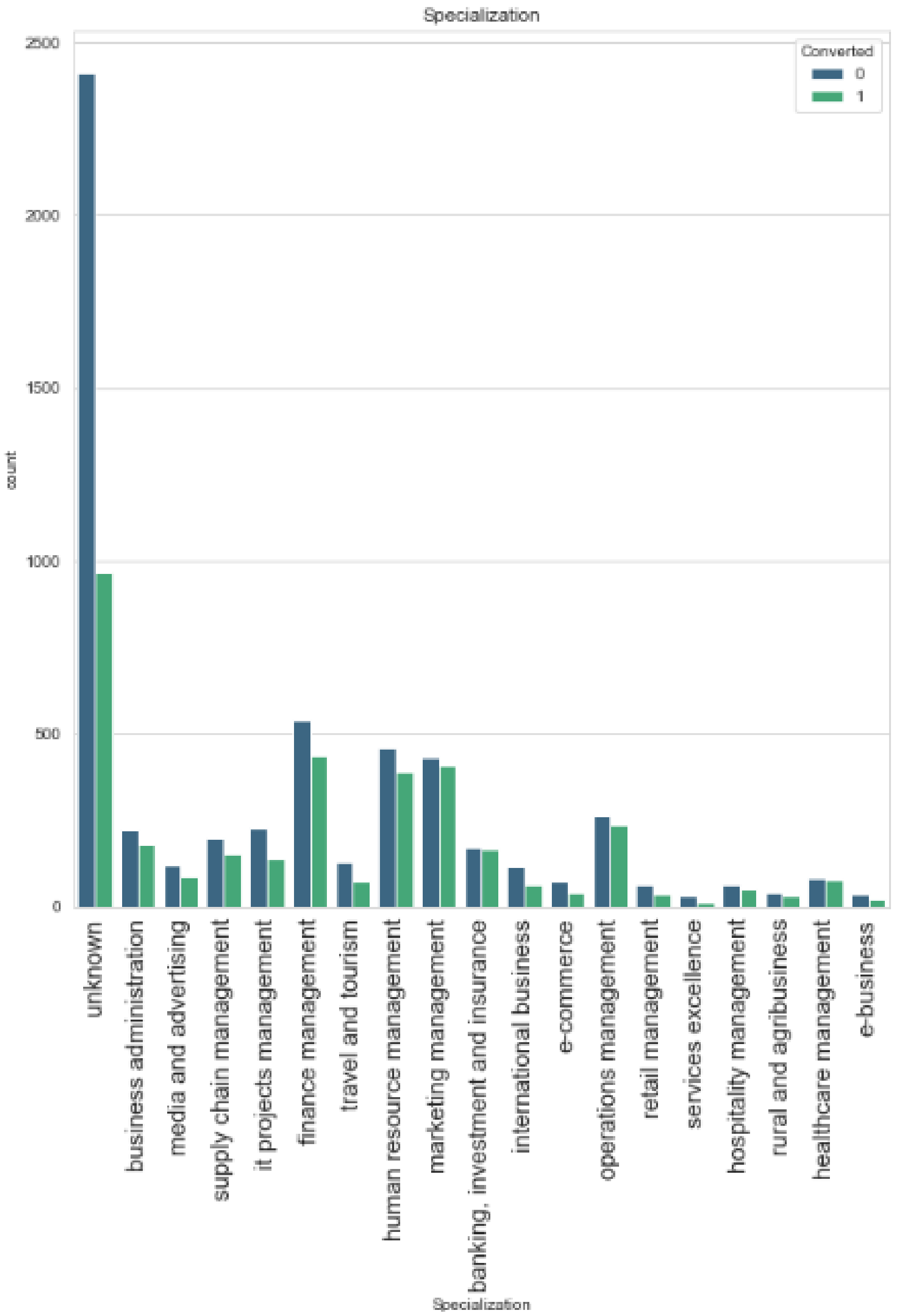
Data Cleaning

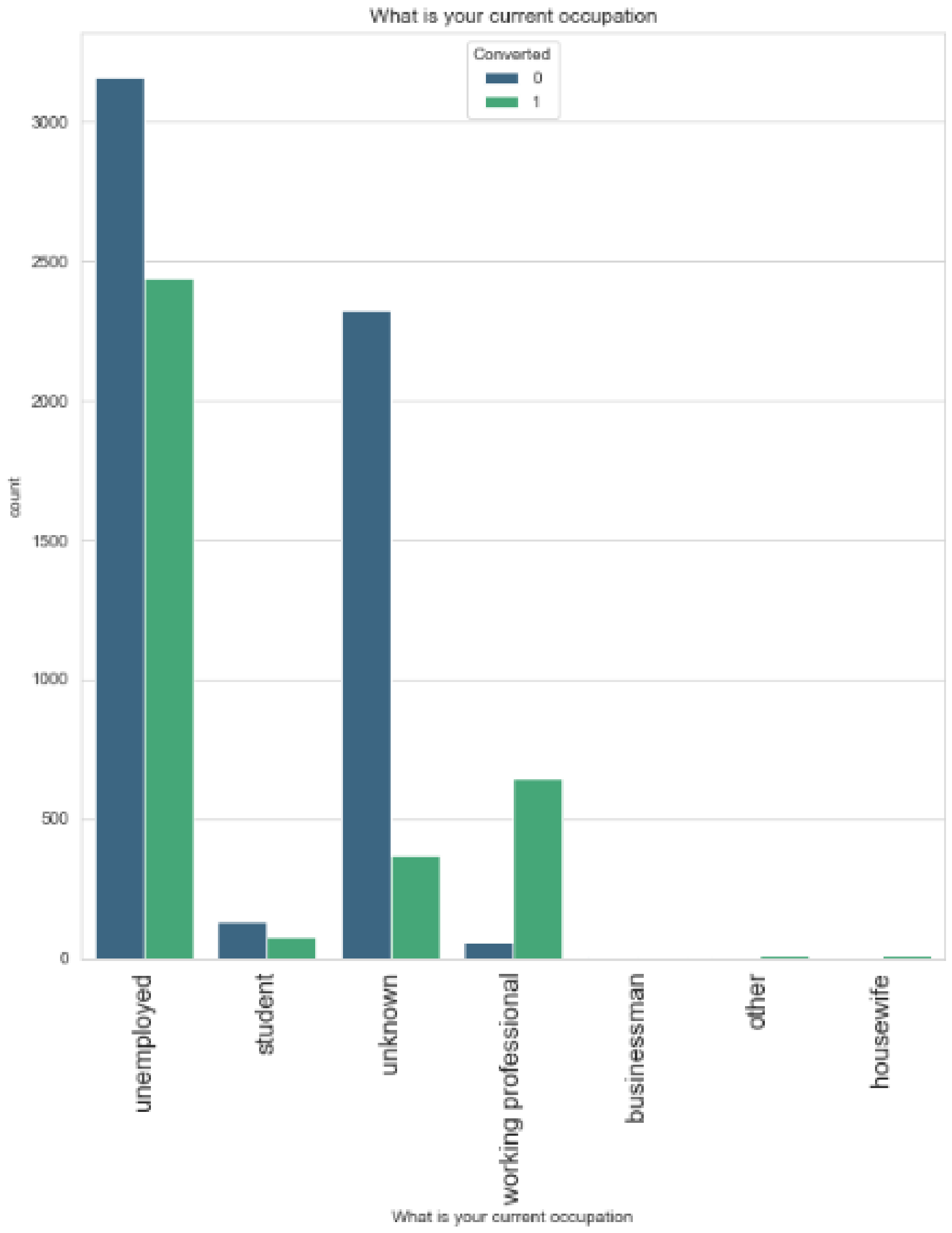
- Some of the columns have missing values.
- Columns with missing values of more than 35% have been dropped.
- Columns with missing values between 20% to 35% have been imputed with various techniques.
- Columns named ‘Specialization’, ‘What matters most to you in choosing a course’, ‘Country’, 'What is your current occupation’ have been imputed with ‘unknown’ as to completely fill up all null values.
- Some of the categorical columns have been imputed with mode.
- Numerical columns have been imputed with median and mean depending upon the case.

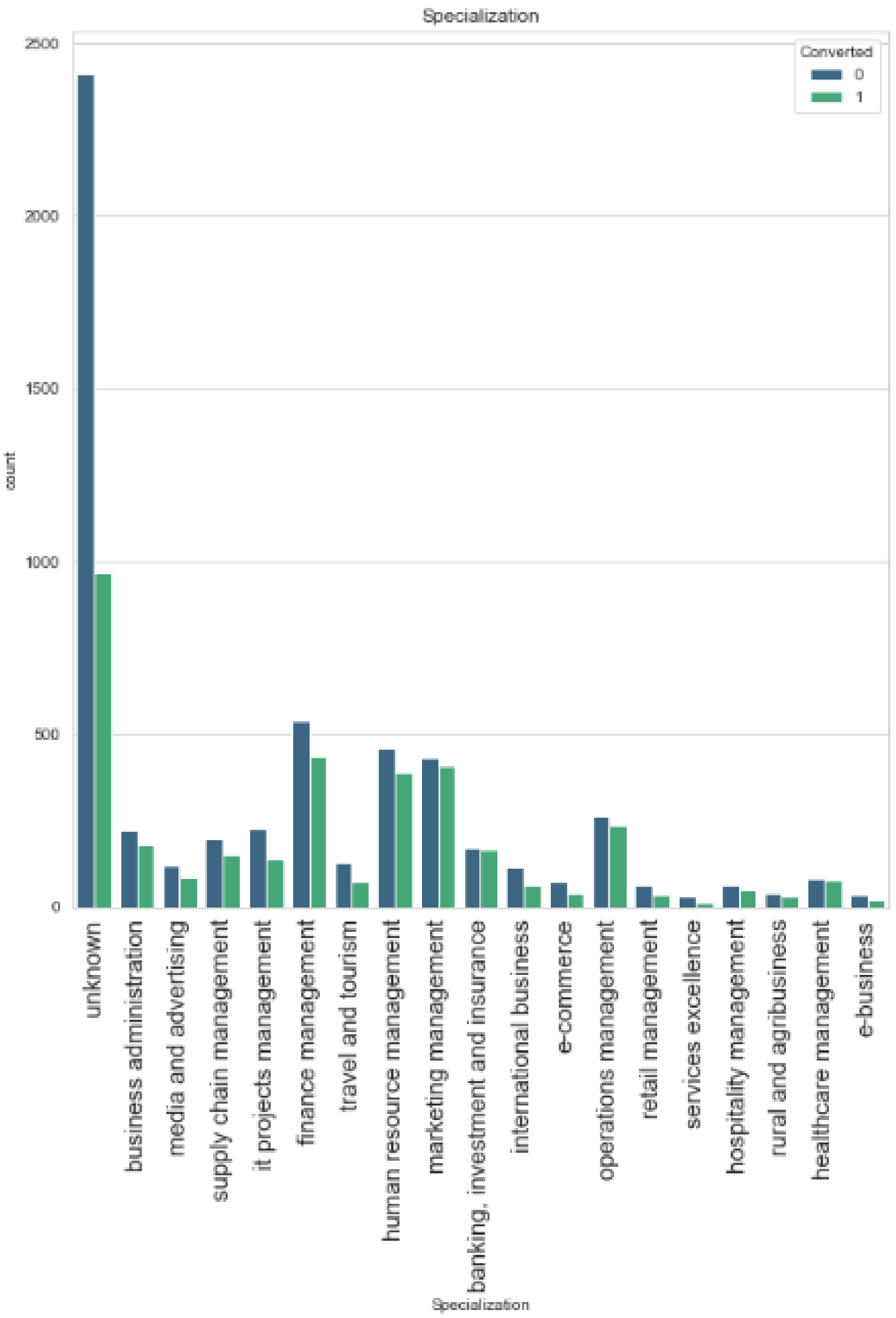
EDA

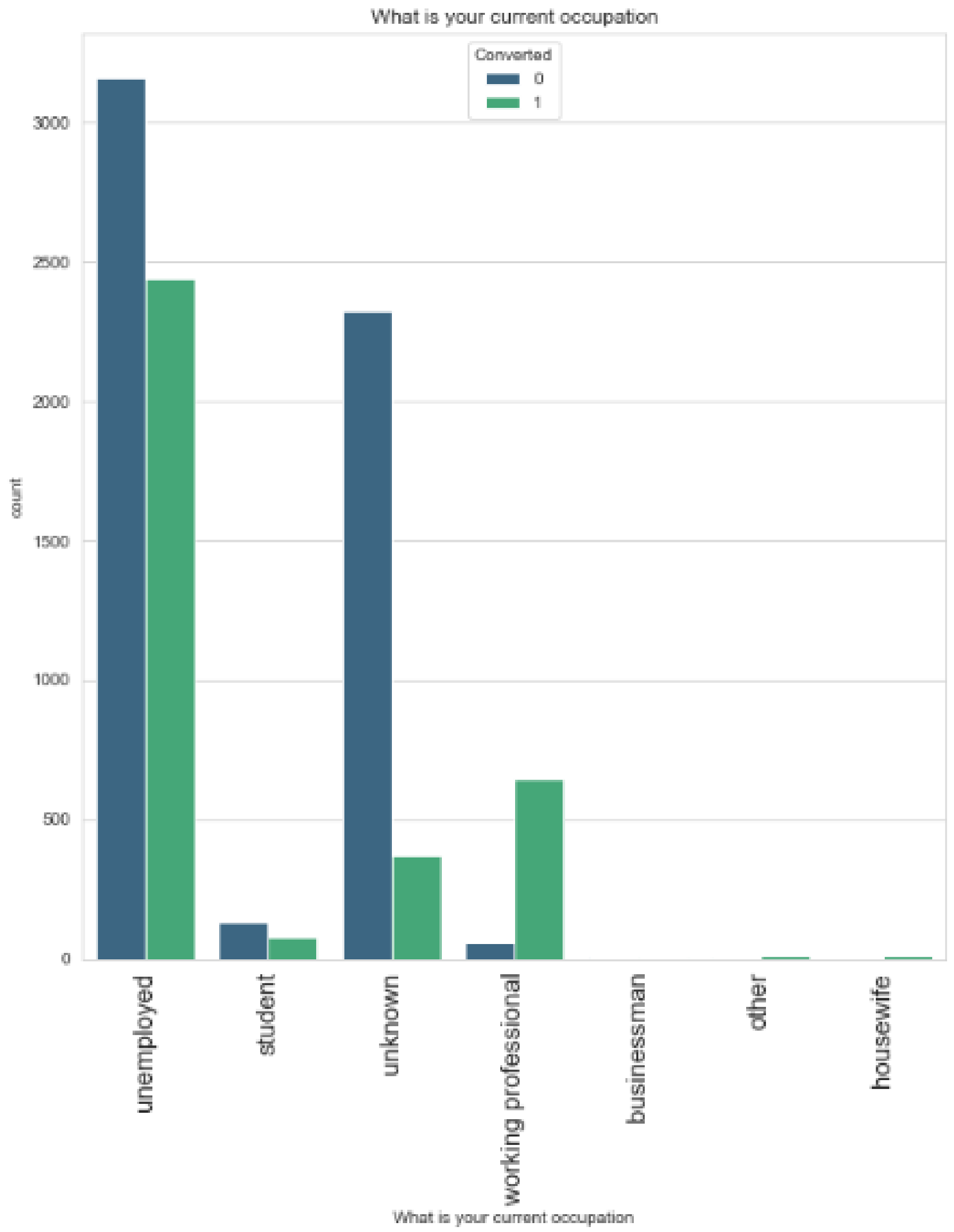

Lead Origin
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
api
landing page submission
lead add form
lead import
quick add form
Lead Origin


Lead Source
Converted
0
1
count
0
250
500
750
1000
1250
1500
1750
olark chat
organic search
direct traffic
google
referral sites
welingak website
reference
facebook
unknown
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source


Specialization
Converted
0
1
2500
2000
1500
1000
500
0
count
unknown
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization


What is your current occupation
Converted
0
1
3000
2500
2000
1500
1000
500
0
count
unemployed
student
unknown
working professional
businessman
other
housewife
What is your current occupation

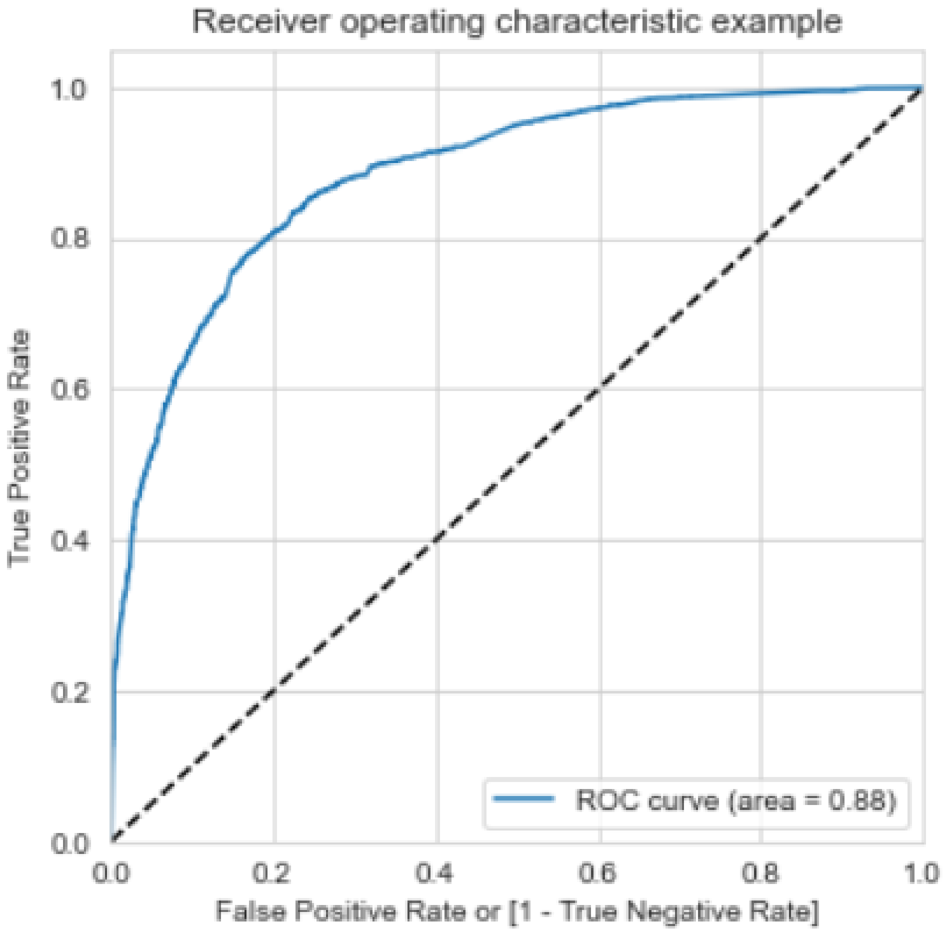

Specialization
Converted
0
1
count
2500
2000
1500
1000
500
0
unknown
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization

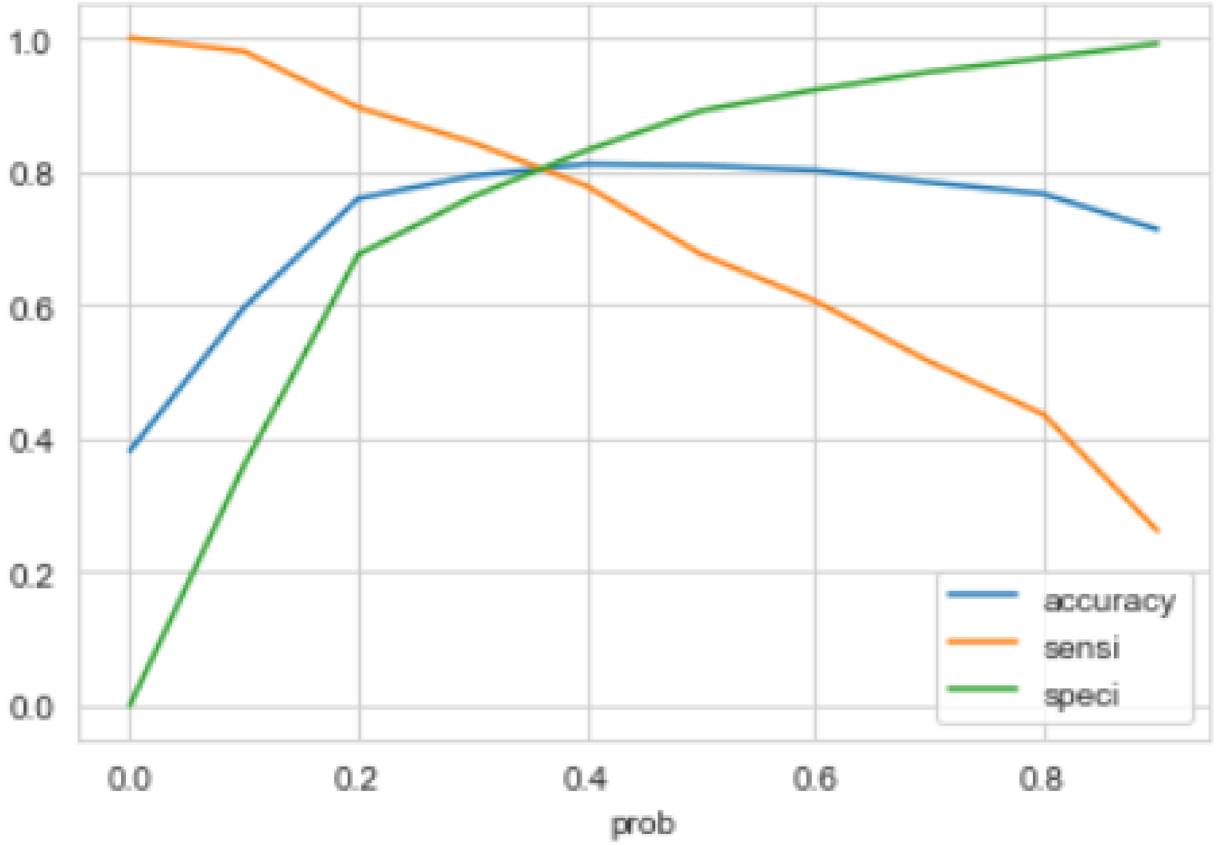

What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
unemployed
student
unknown
working professional
businessman
other
housewife
What is your current occupation

Model Building

- Creating Training and Test Sets from the Data.
- A train-test split is the first fundamental stage in the regression process; we have selected a 70:30 split.
- Utilize RFE while choosing features.
- Building a model by deleting the variable whose p-value is larger than 0.05 and whose VIF value is more than 5 after running RFE with 15 variables as result forecasts for the test data set 81% overall accuracy.

ROC Curve and Cut-off Optimization





- ➢ Determining the Best Cutoff Point

- The chance where sensitivity and specificity are balanced is the optimal cutoff probability.
- The second graph makes clear that 0.36 is the ideal cutoff.

Conclusion and Recommendations

- The factors that affected potential purchasers the most were discovered to be: Total time spent on the Website.
- The total quantity of visitors. When Google, direct traffic, organic search, or the Welingak website was the top source when the last activity occurred, it was either an SMS or an Olark chat.
- If the lead origin is Lead Add format with these in mind, X Education may succeed since they have a very high possibility of persuading nearly all prospective customers to alter their minds and purchase their courses.