
B
A
Lead Scoring Case
Study

Problem Statement

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. Moreover, the company also gets leads through past referrals. 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%.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Objective

- ✓ Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- ✓ There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution Methodology

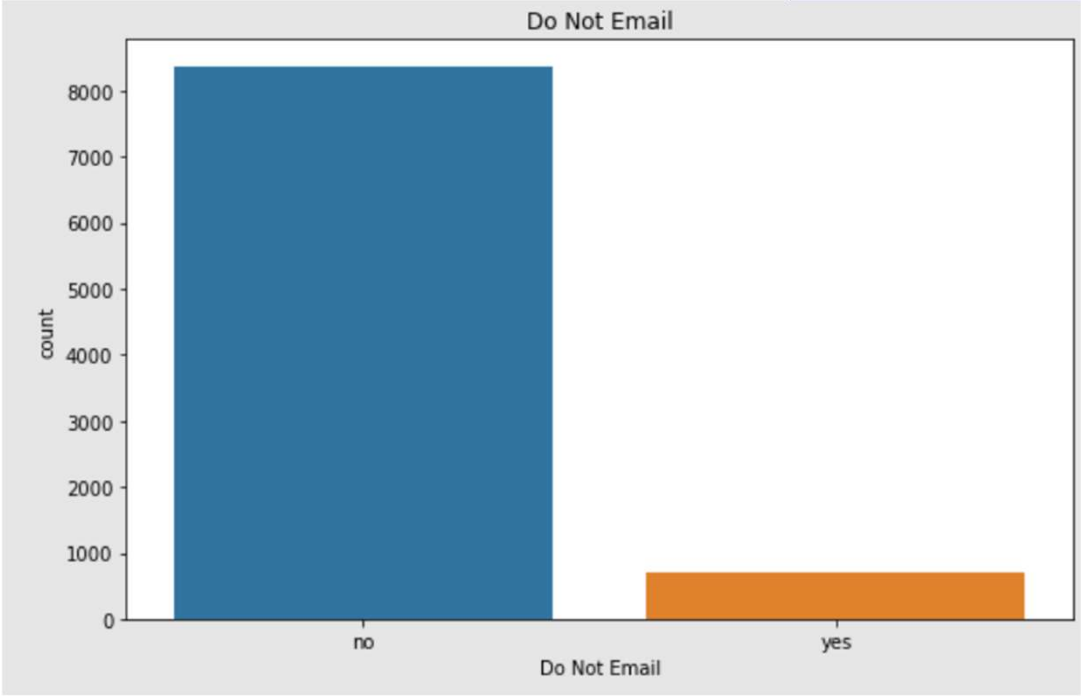
- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.

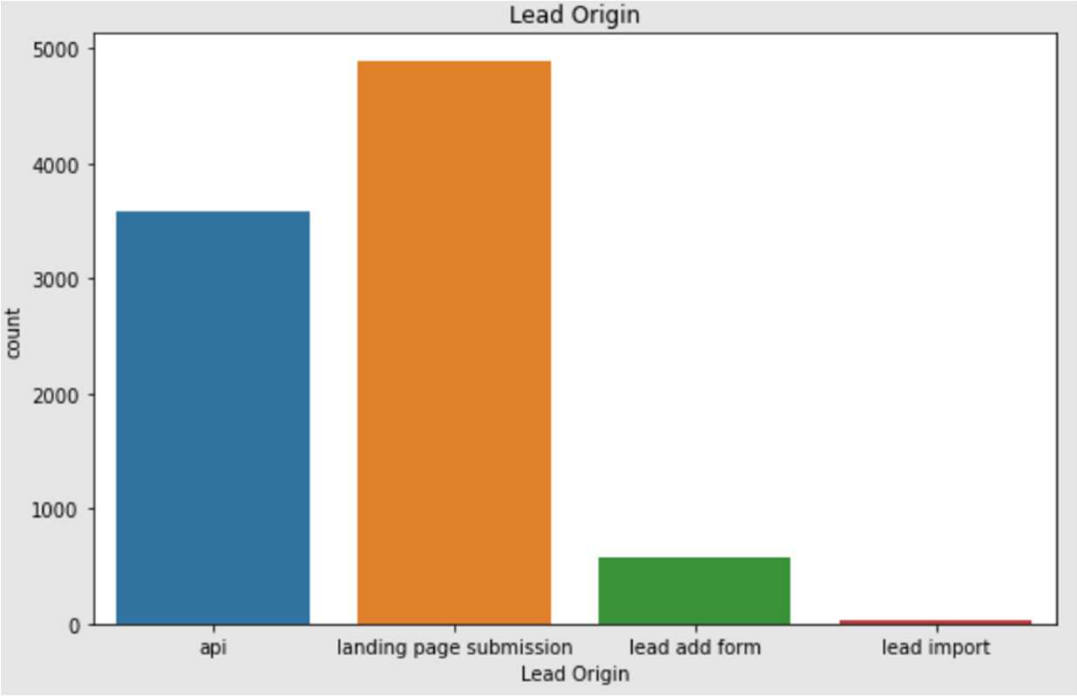
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

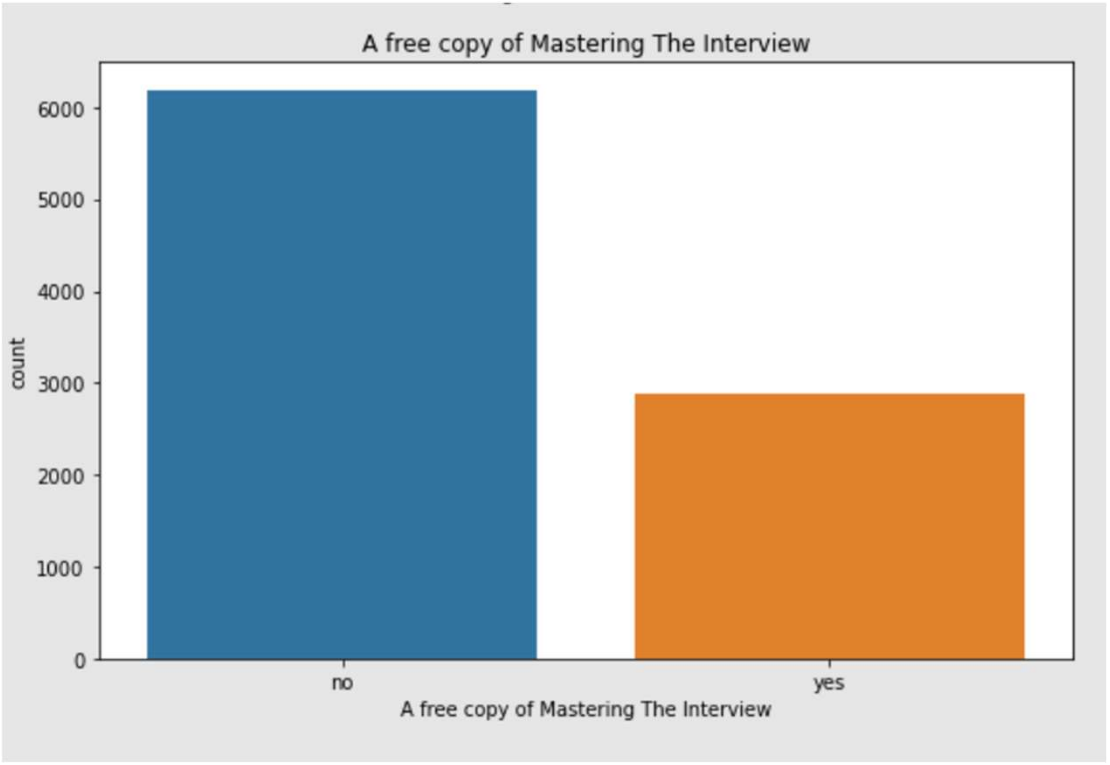
Data Manipulation

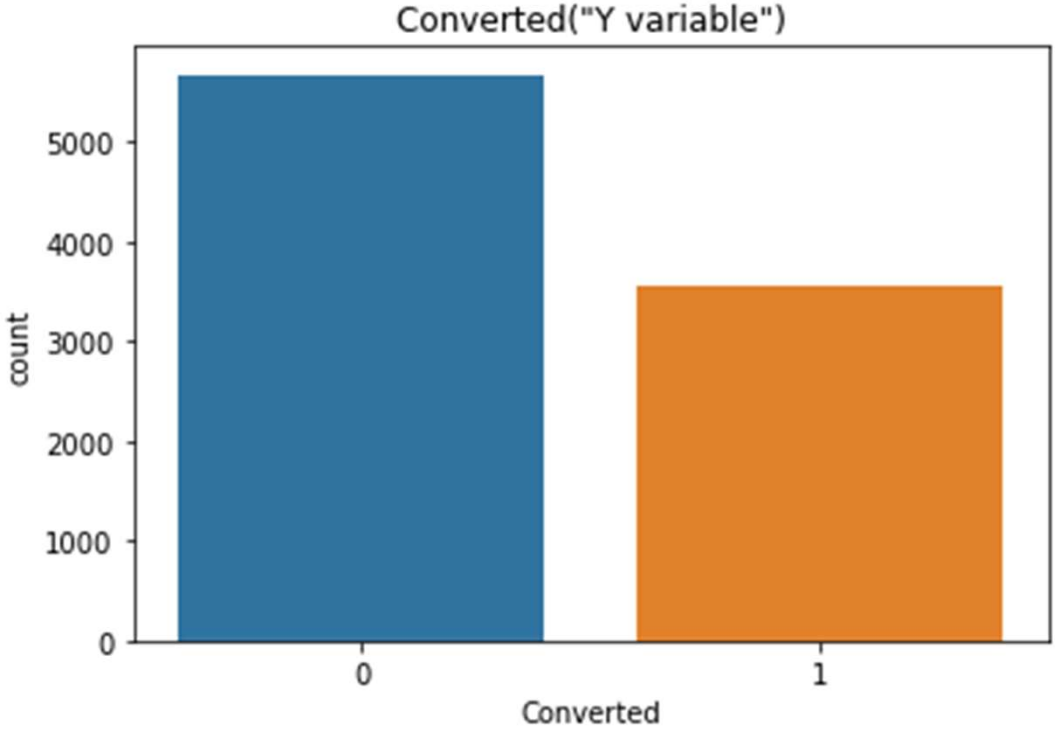
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”,
- “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through
- cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of
- the features which has no enough variance, which we have dropped, the features are:
- “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper
- Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear
- about X Education’ and ‘Lead Profile’.

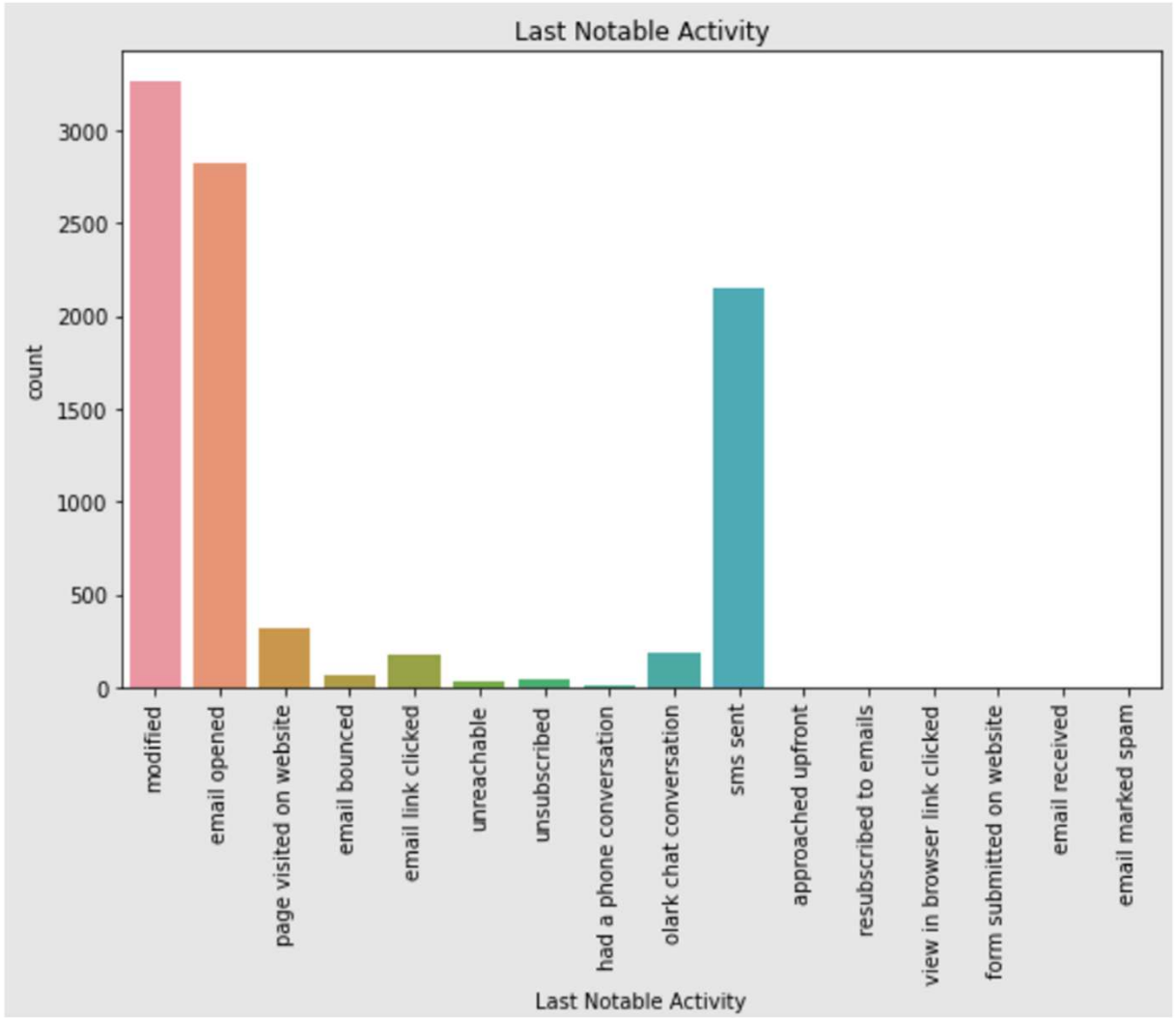
EDA


Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

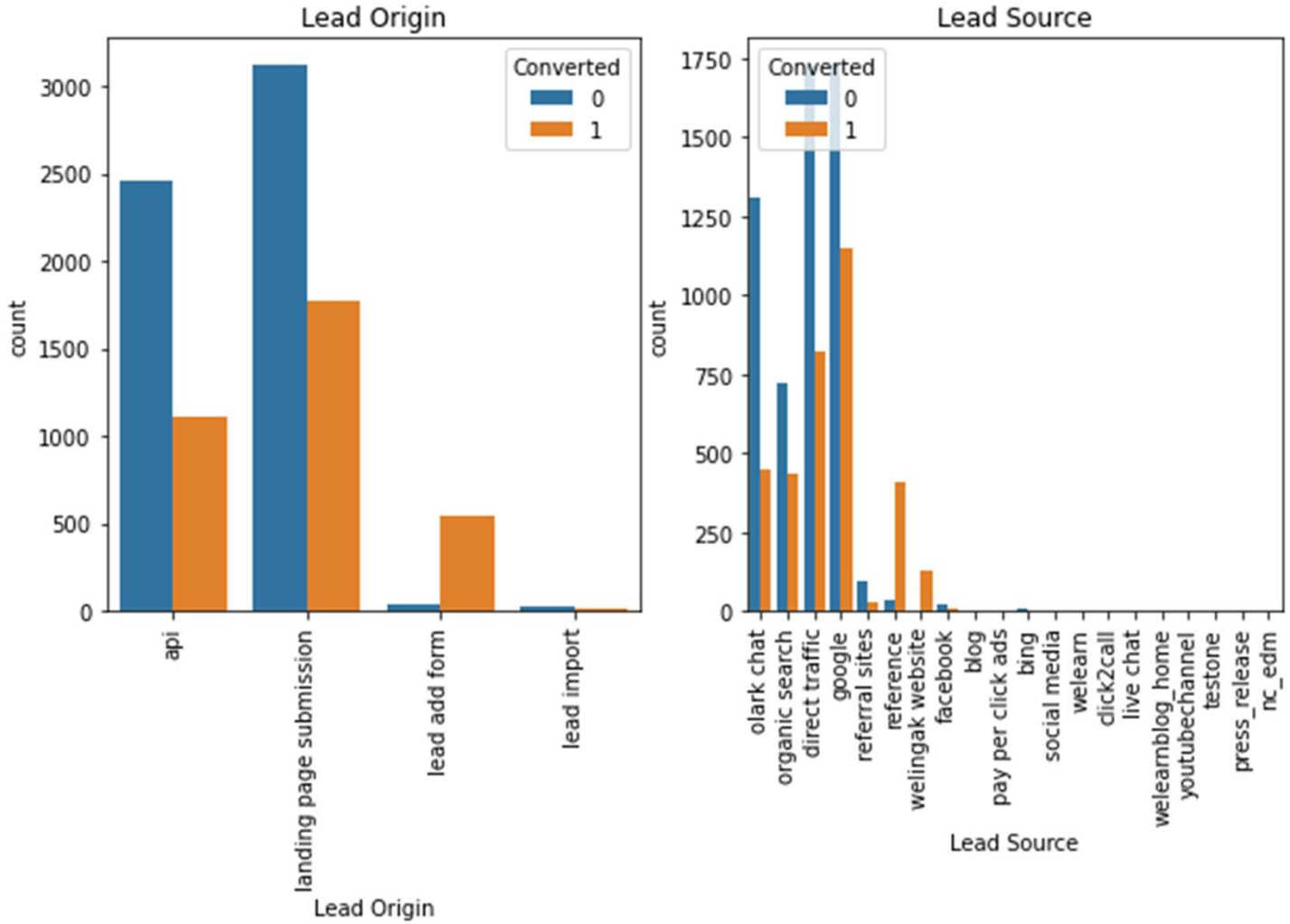

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

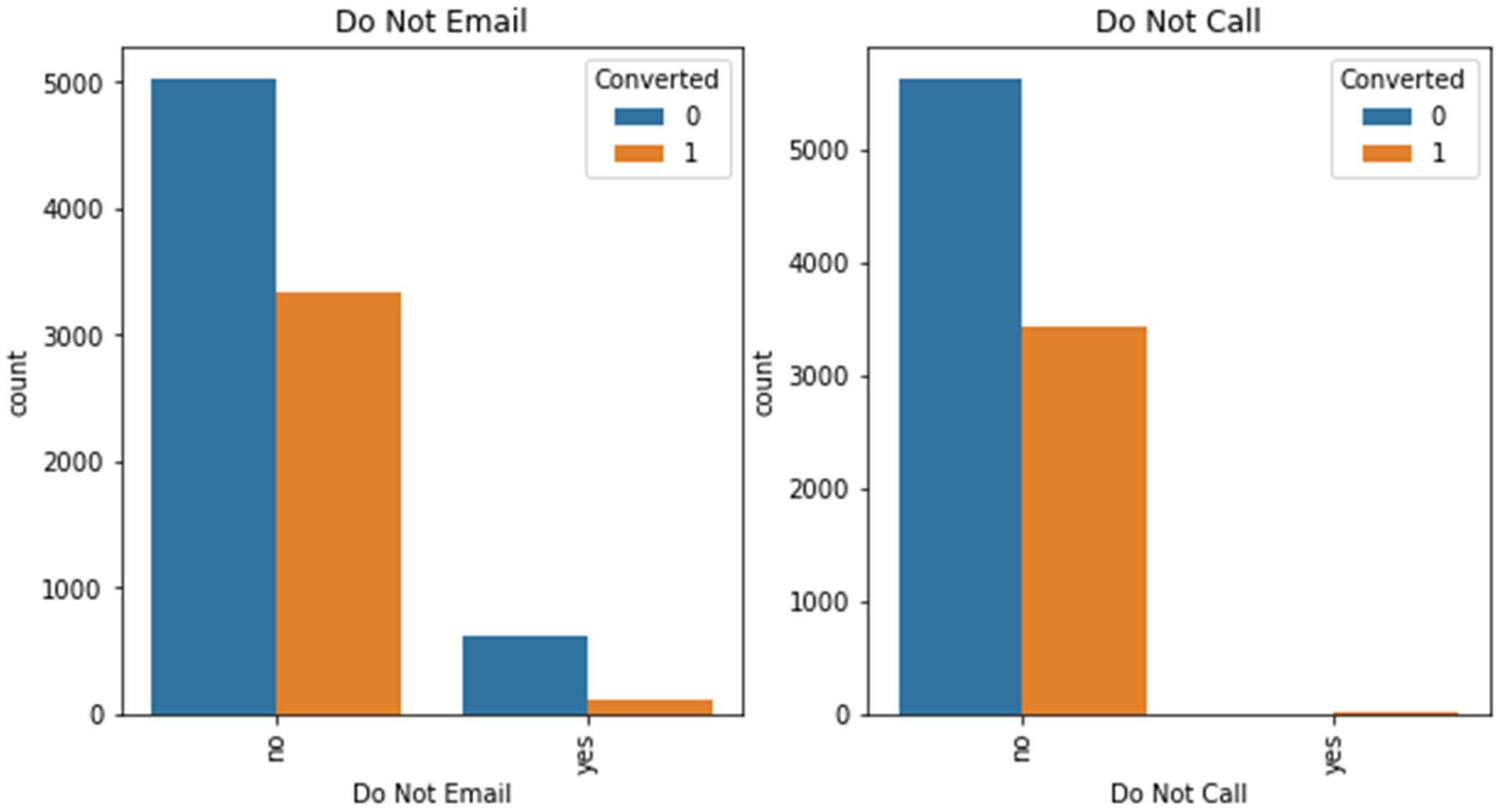

A free copy of Mastering The Interview
6000
5000
4000
3000
2000
1000
0
count
no
yes
A free copy of Mastering The Interview

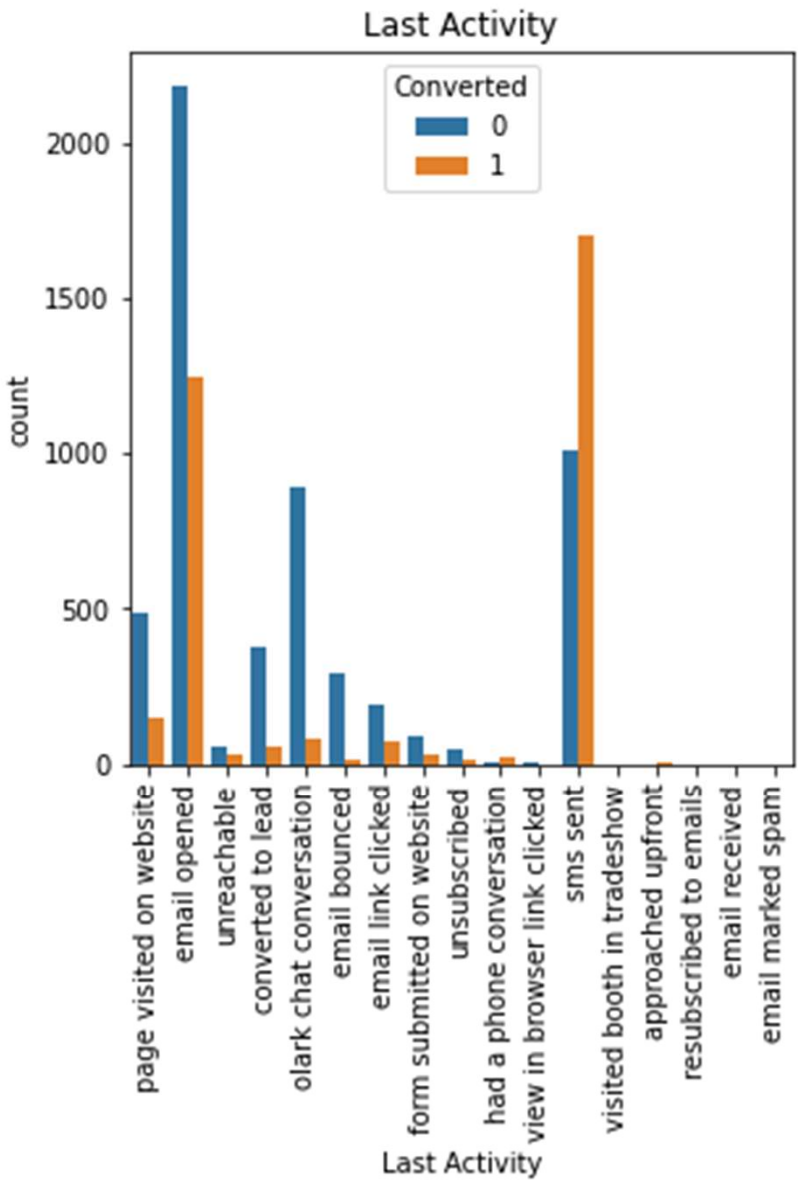

Converted("Y variable")
5000
4000
3000
2000
1000
0
count
0
1
Converted

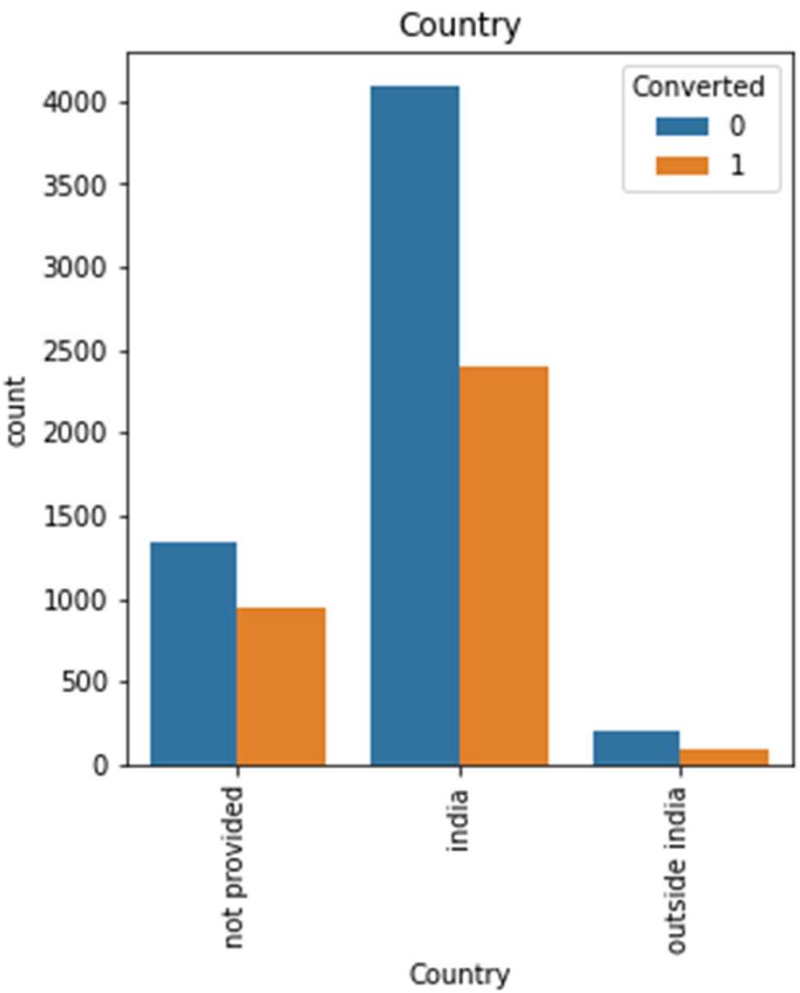

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity

Categorical Variable Relation




Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email
Do Not Call
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Call


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity


Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

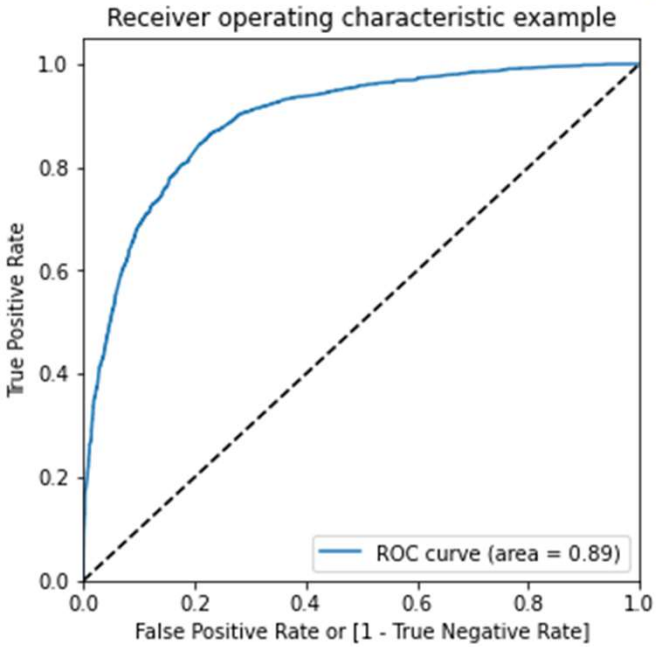
Data Conversion

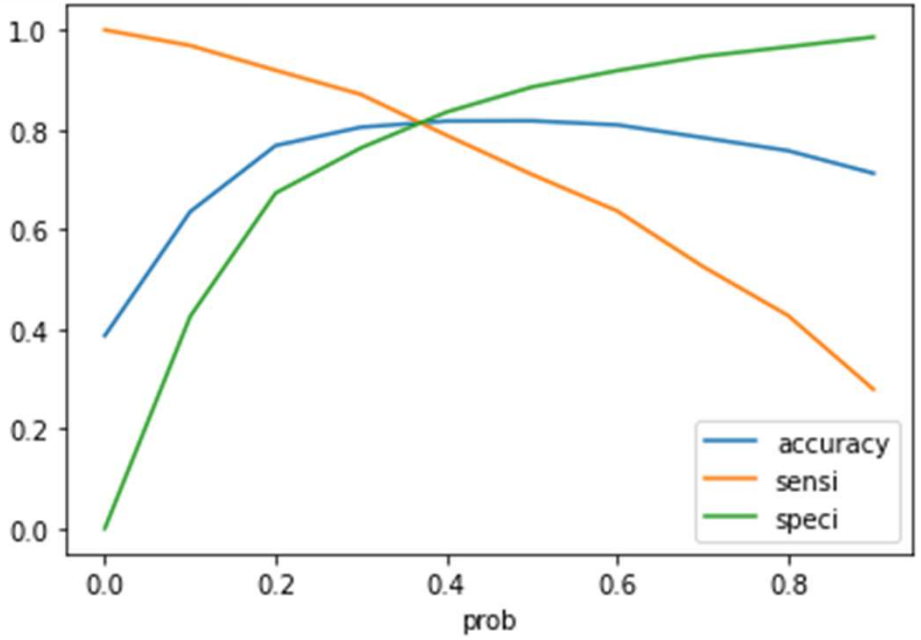
- ✓Numerical Variables are Normalised
- ✓Dummy Variables are created for object type variables
- ✓Total Rows for Analysis: 8792
- ✓Total Columns for Analysis: 43

Model Building

- ✓ Splitting the Data into Training and Testing Sets
- ✓ The first basic step for regression is performing a train-test split, we have chosen 70:30
- ✓ ratio.
- ✓ Use RFE for Feature Selection
- ✓ Running RFE with 15 variables as output
- ✓ Building Model by removing the variable whose pvalue is greater than 0.05 and vif
- ✓ value is greater than 5
- ✓ Predictions on test data set
- ✓ Overall accuracy 81%

ROC Curve





1. Finding Optimal Cut off Point
2. Optimal cut off probability is that
3. probability where we get balanced sensitivity and specificity.
4. From the second graph it is visible that the optimal cut off is at 0.35.

Summary and Insights

- The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost accurate.
- We have high recall score than precision score which is a sign of good model.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
  A. Lead Origin_Lead Add Form
  B. Total Time Spent on Website

Conclusion

The total time spend on the Website.

1. Total number of visits.
2. When the lead source was:
    A. a. Google
    B. b. Direct traffic
    C. c. Organic search
    D. d. Welingak website
3. When the last activity was:
    a. SMS
    b. Olark chat conversation
4. When the lead origin is Lead add format.
5. When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.


THANK YOU