LEAD SCORE CASE STUDY

By:

Shraddha Sail

PROBLEM

- X-education, an Ed-tech company, sells online courses to working professionals.
- But the lead conversion rate is very low, about 30%
- The company wants to increase the lead conversion rate to around 80%, and wants to identify the potential leads.

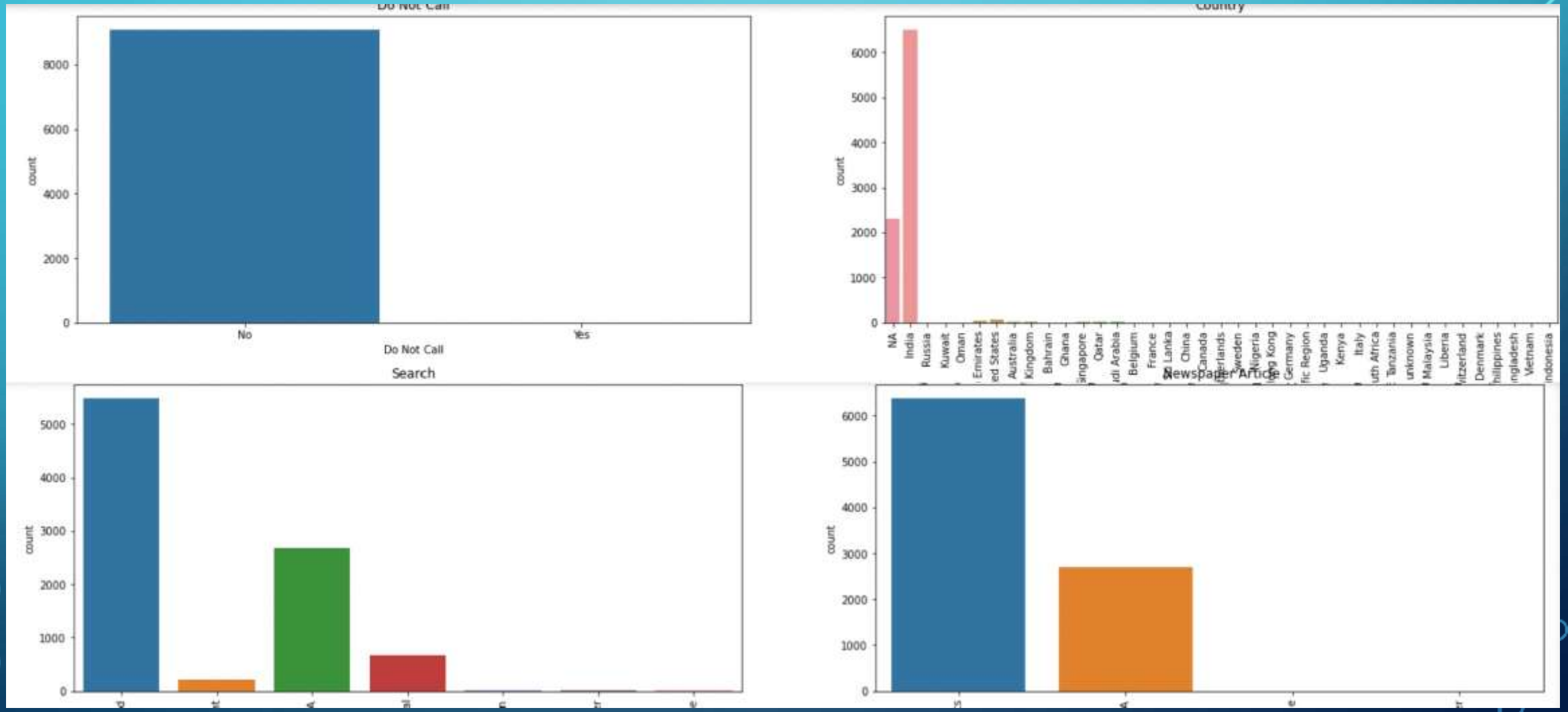
OBJECTIVE

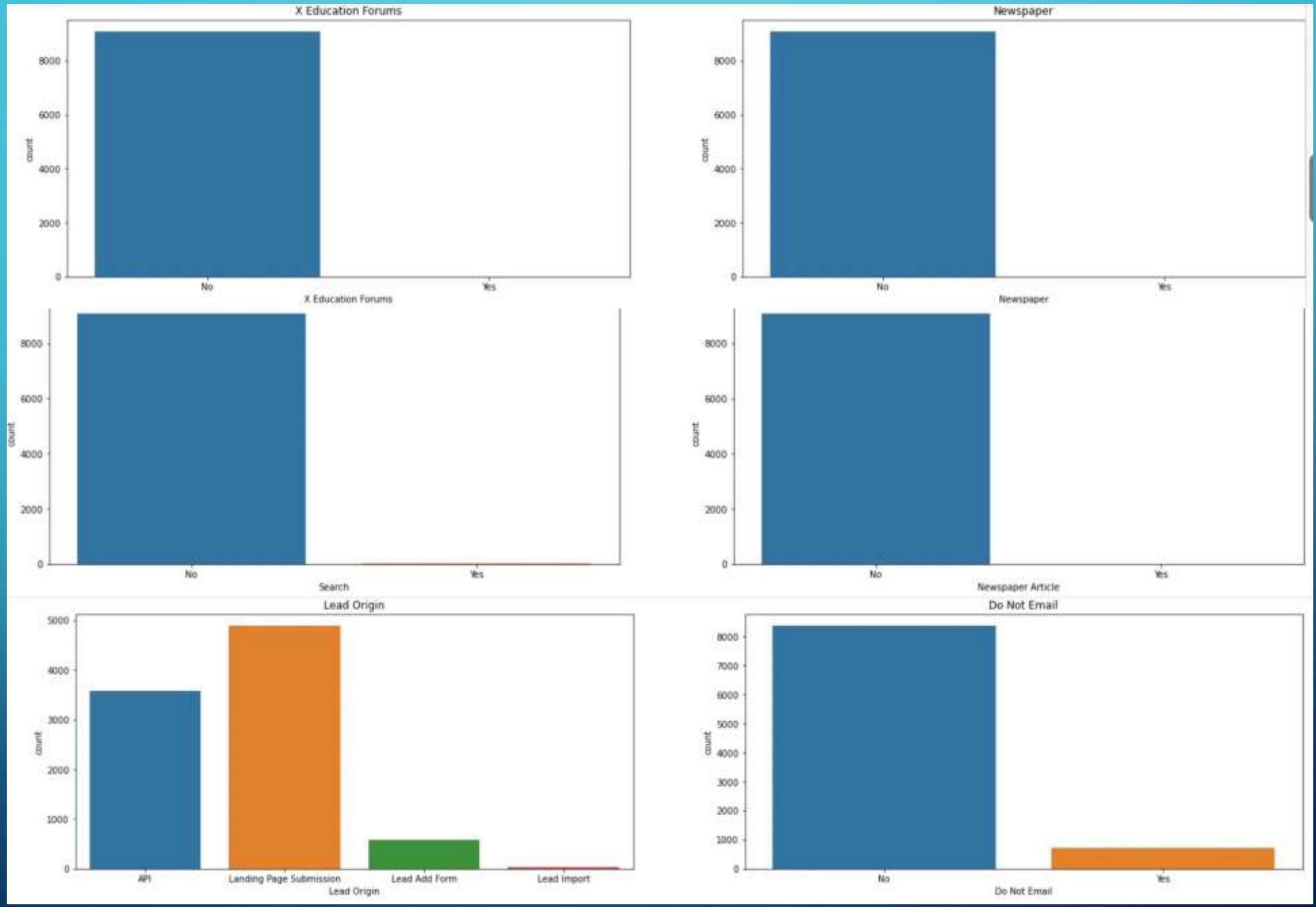
- X-education wants to identify the hot leads.
- Build and deploy a model that point out the necessary variables.

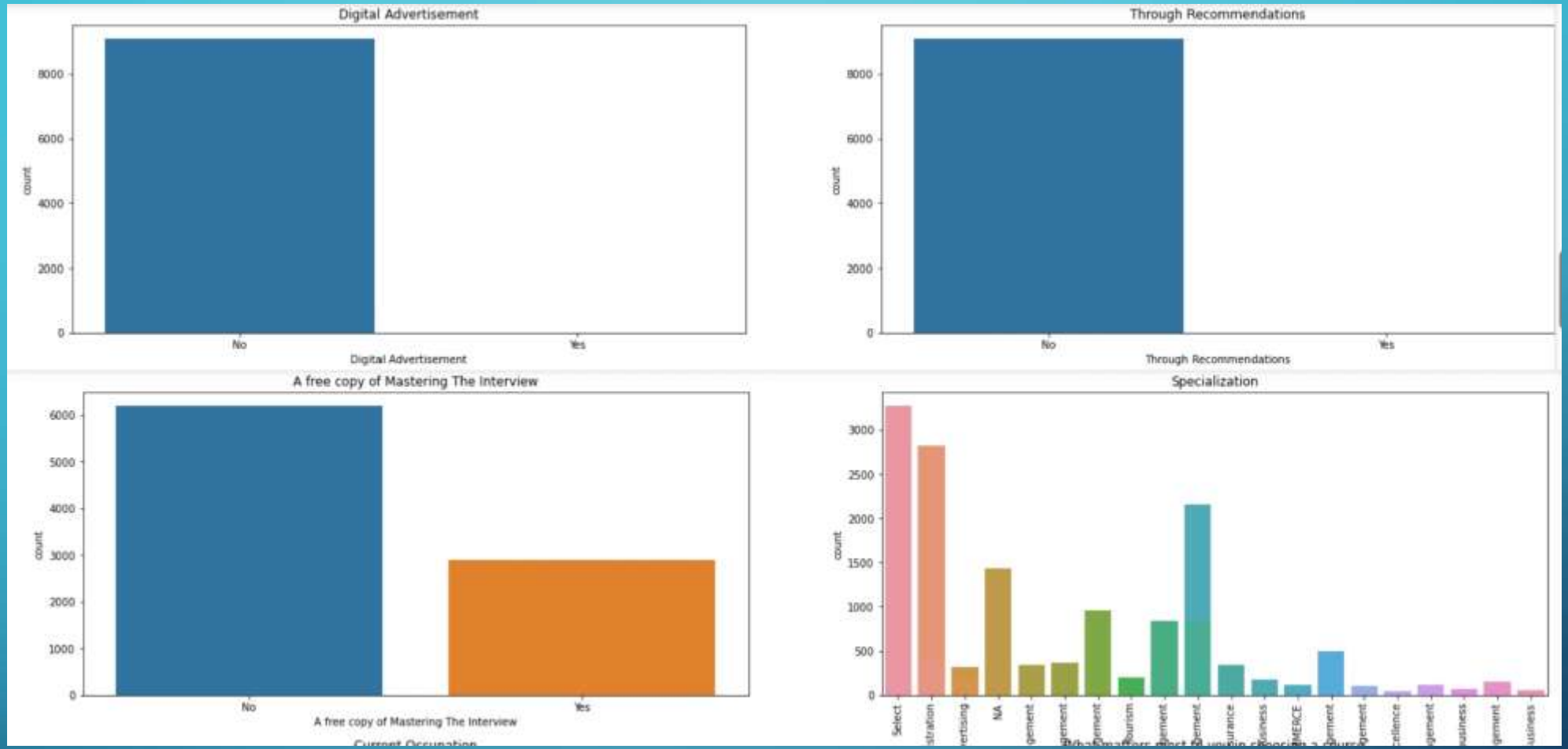
APPROACH

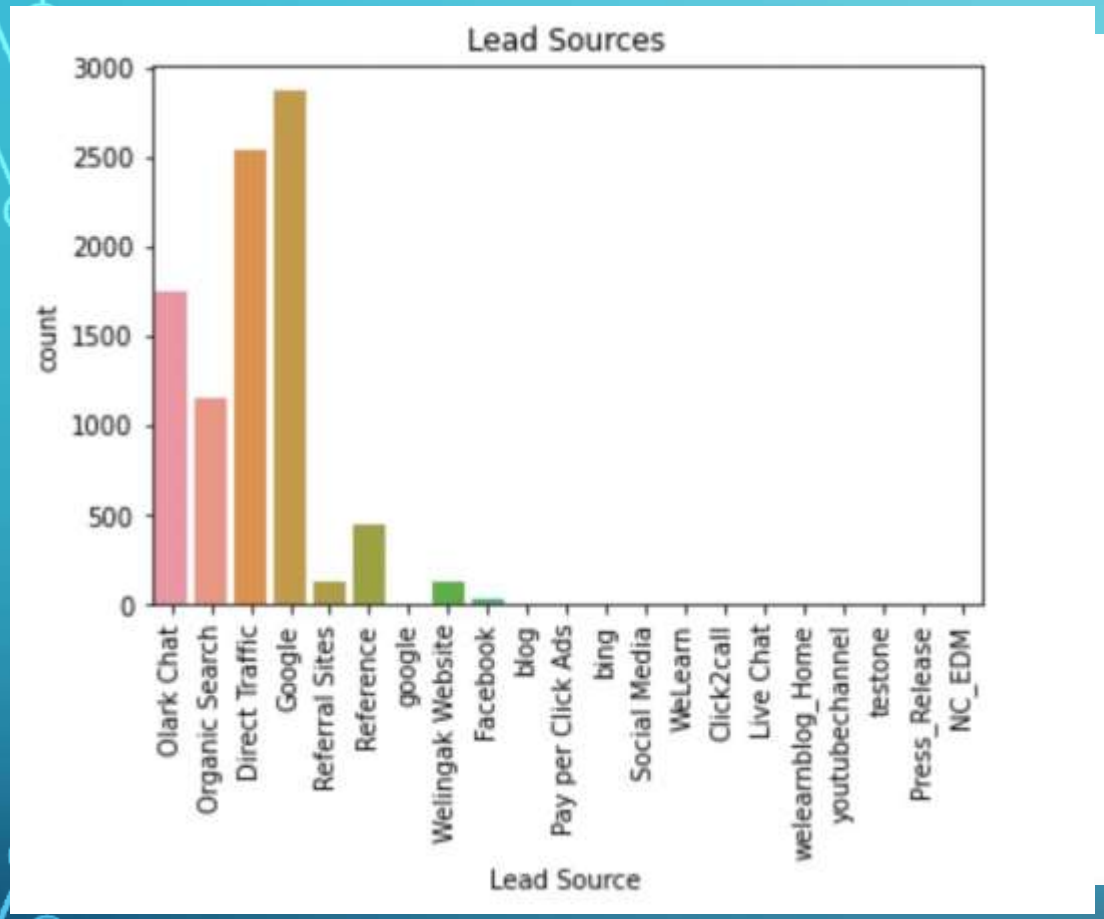
- The first step is to clean the data, to set the context, to identify missing values and make the data fit enough for EDA.
- Next comes the Exploratory Data Analysis- doing analysis to find patterns in the data and some statistical values, correlation etc.
- Perform Feature Scaling and Dummy Variables for building variables for the model.
- Splitting the dataset, fitting a linear regression model and check accuracy.
- Use the testing set to see the accuracy of the model.

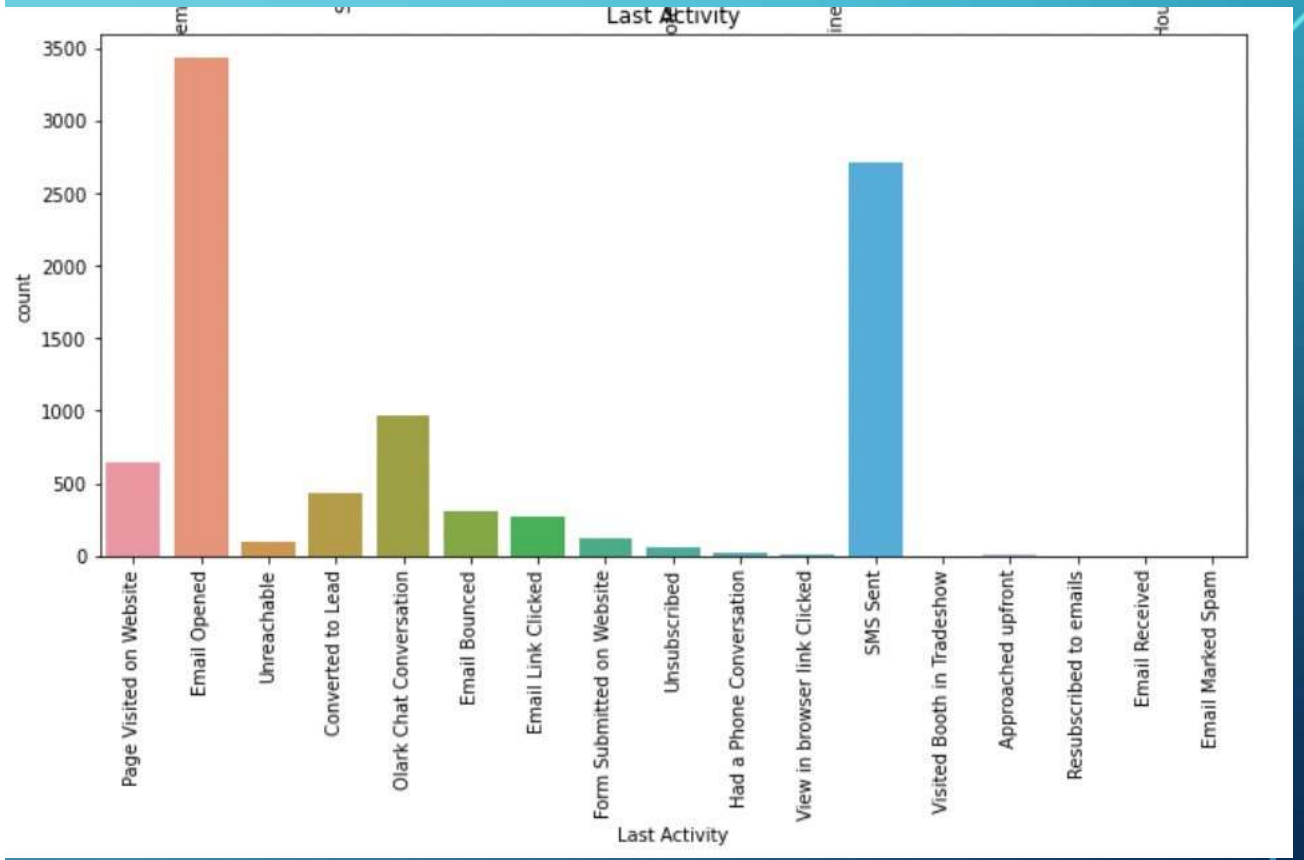
EDA
Categorical Variables



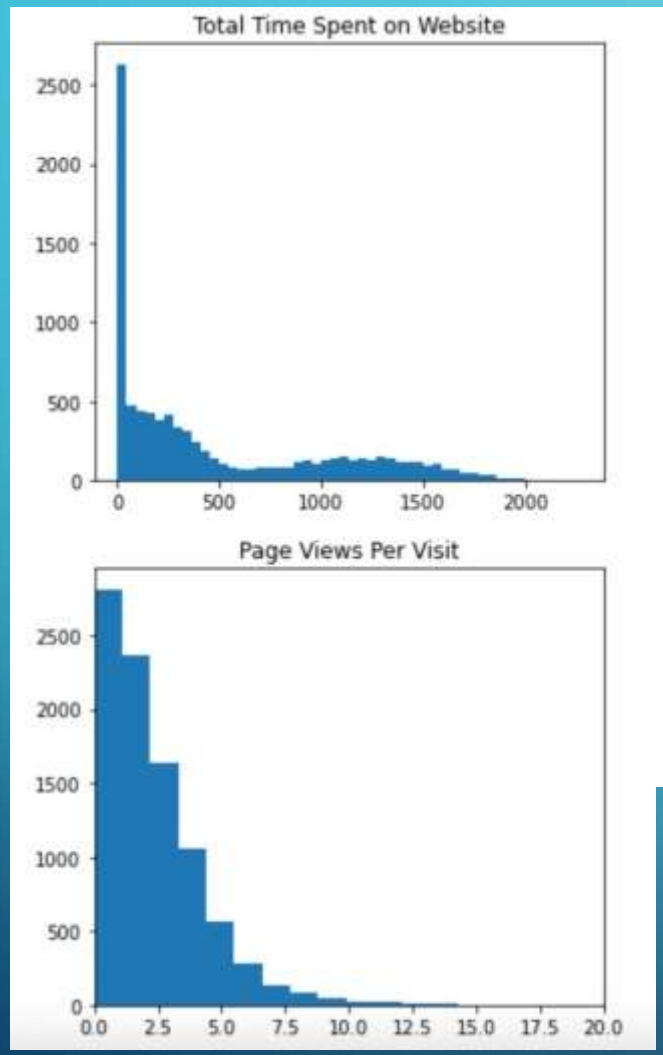

X Education Forums
count
No
Yes
Newspaper
count
No
Yes
X Education Forums
count
No
Yes
Search
Newspaper
count
No
Yes
Newspaper Article
Lead Origin
count
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin
Do Not Email
count
No
Yes
Do Not Email

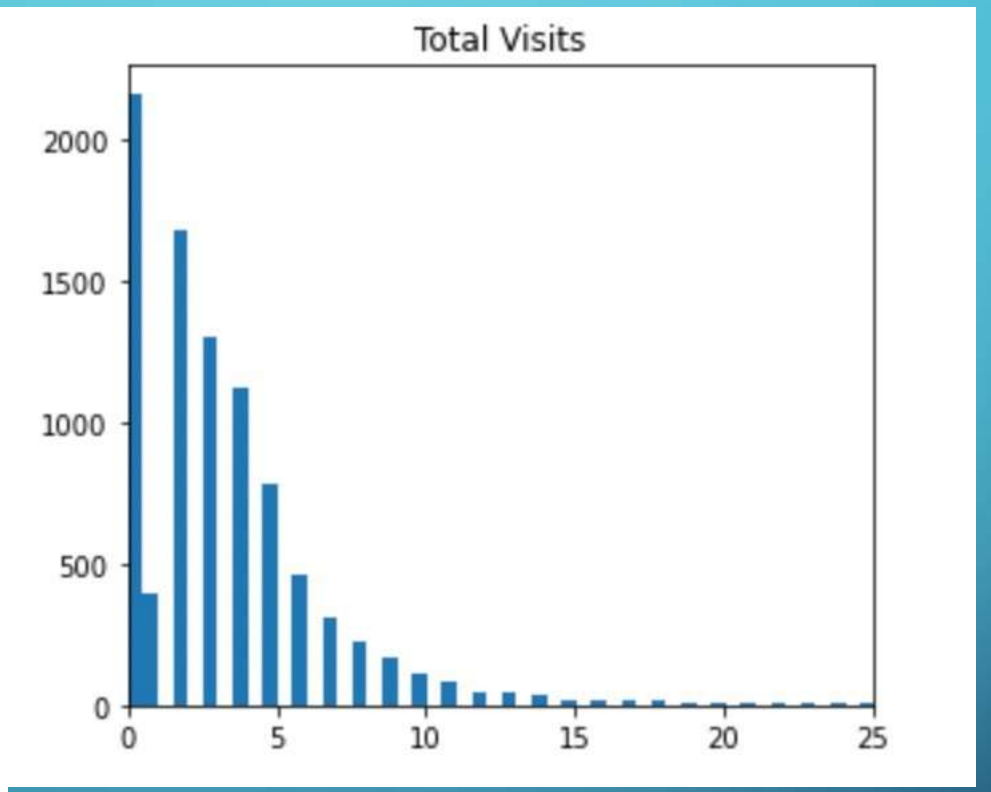

Digital Advertisement
count
8000
6000
4000
2000
0
No
Yes
Digital Advertisement
Through Recommendations
count
8000
6000
4000
2000
0
No
Yes
Through Recommendations
A free copy of Mastering The Interview
count
6000
5000
4000
3000
2000
1000
0
No
Yes
A free copy of Mastering The Interview
Specialization
count
3000
2500
2000
1500
1000
500
0
Select
NA


Lead Sources
count
Lead Source
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM


count
Last Activity
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

Numerical Variables





DATA CONVERSION

- The numerical variables are normalized
- Only the variables with object as datatype were converted into dummy variables.
- Dimension of dummy variable table: (9074x137)

BUILDING THE MODEL

- Split the data 75% (training set) and 25% (testing set)
- Feature selection by Recursive Feature Elimination (RFE)
- Used 15 variables as initial RFE output.
- Build the model keeping in mind the p-value< 0.05 and variance inflation factor< 5
- Repeat the process of model building till we get all the proper p-values and vifs. Here, we had 2-iterations, the later one being the final model.
- Predict using the testing set.
- Prediction accuracy achieved ~80%

CONCLUSION

The variables that directly impact the lead conversion (Ordered in descending order) :

>The total time spent on website.

>Total number of visits.

- When the lead source was: Direct traffic, Welingak website, Organic search
- When the last activity was: Olark chat conversation
- When the lead origin is Lead add format.

Prioritizing these factors would lead to a high lead conversion rate.