Lead Scoring Case Study

Presented By:

- Samrudhi Yeginwar
- Debabrata Panda
- Keerthi Gururaj GS

Problem Statement

- X education sells online courses to industry professionals.
- X education gets lots of lead and its result at target i.e getting converted into hot leads is at rate poor level.
- For an instance, if out of 100 leads only 30 gets converted.
- To make this process more efficient, company wants to identify most eligible(potential) leads which is also known as “Hot Leads”.
- If this process of getting converted into hot leads is more by identifying successfully then this is directly proportional to the sales number and communicating with those leads makes easier for the team to give a call.

Business Objective

- X education wants to identify hot leads instead of giving a call to everyone.
- To achieve this company wishes to make someone to build models and get accurate results.
- And also for the purpose of future in knowing the most potential leads and converting them.

Methodology

- Firstly, data understanding and applying steps accordingly to achieve the result.

➢ Data handling, cleaning and manipulation :

1. Importing all libraries and checking for the missing values and null values.
2. After that, handling them by dropping some “not required” values for analysis.
3. Imputation if necessary and also outliers handled.

➢ Now EDA(Exploratory data analysis).

1. Univariate data analysis is done.
2. Bivariate data analysis also done with correlation coefficients and pattern between variables.

➢ After that creating Dummy variables and scaling the data.

➢ Skill (technique) used : logistic regression used for predicting and making models.

➢ Validation of the models is done .

➢ Model prediction and presenting them with accurate graphs using visualization

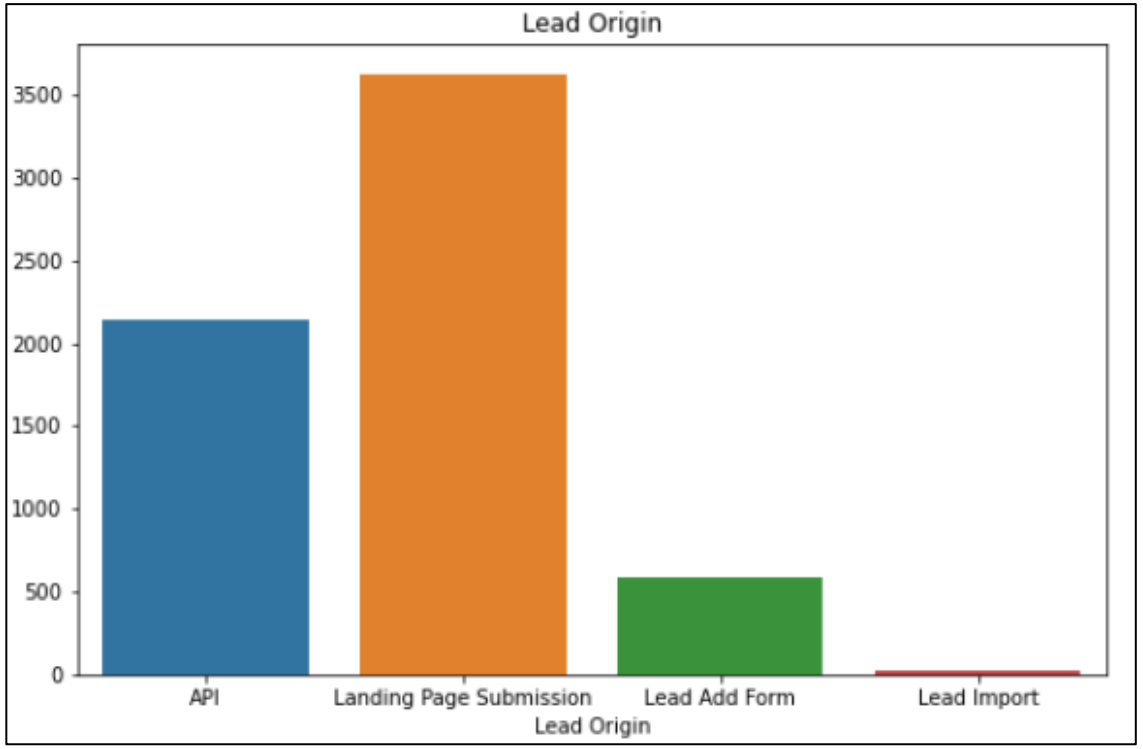
➢ Conclusion

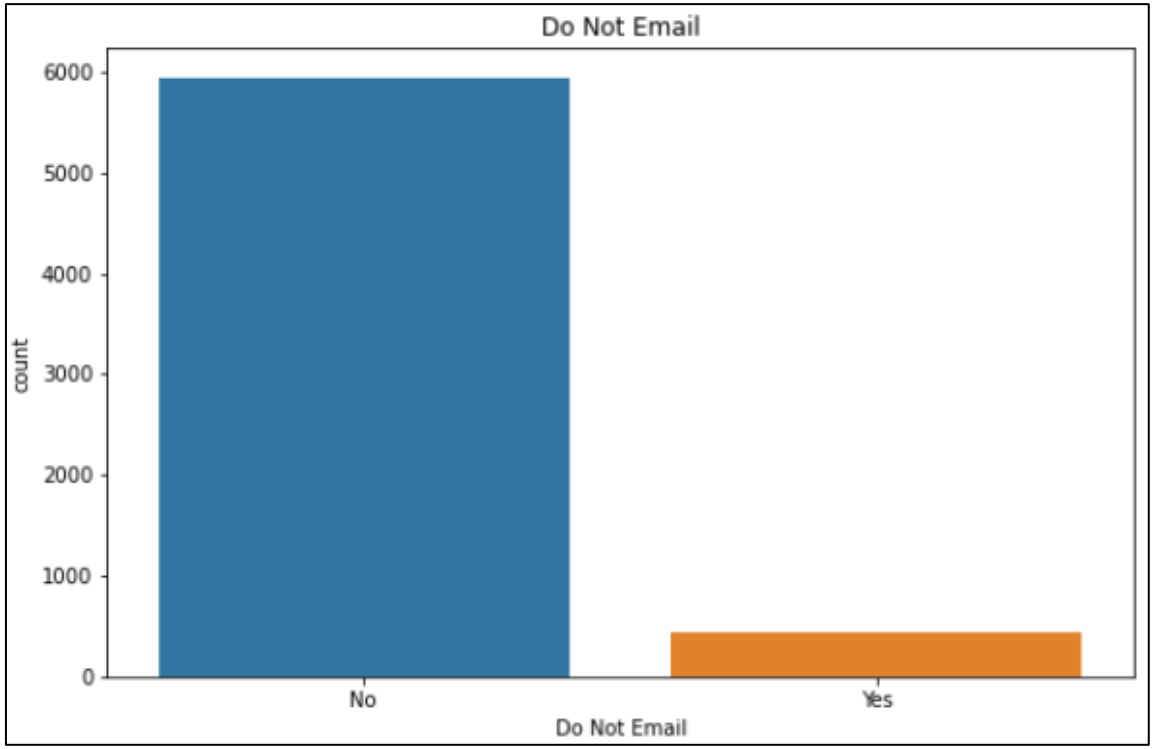
Data Analysis

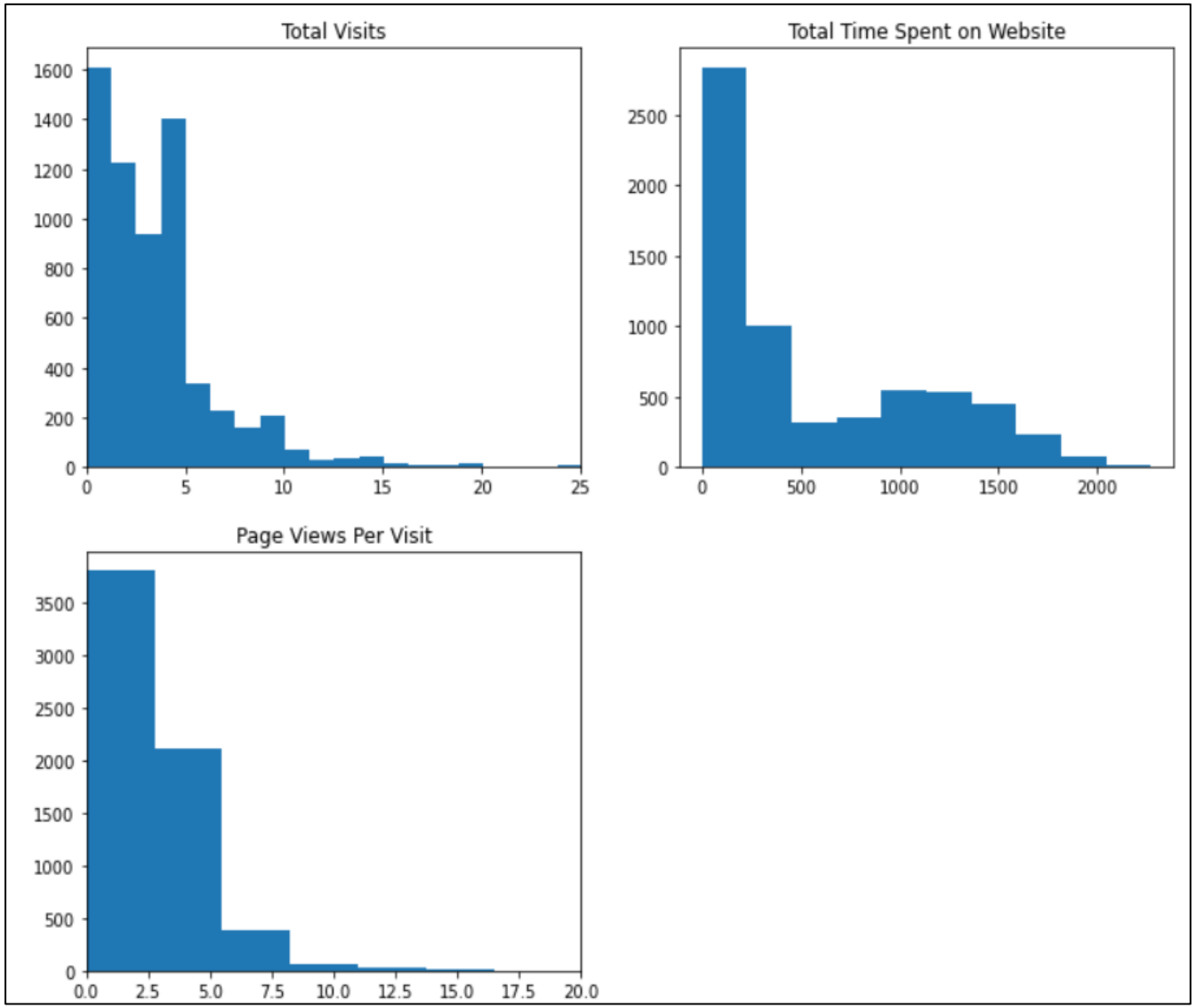
- Total number of rows and columns are like (9240,37).
- As you can see there are a lot of column which have high number of missing values. Clearly, these columns are not useful. Since, there are 9000 datapoints in our data frame, let's eliminate the columns having greater than 3000 missing values as they are of no use to us.
- The variable City and Country aren't really of much use in our analysis. So we can drop it.
- Here we can see three columns have select out of them Lead Profile and How did you hear about X Education have a lot of rows which have the value Select which is of no use to the analysis so we drop them.
- Also notice that when you got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article etc. Since most of the values are No, we drop these columns.
- The column what matters most to you in choosing a course has the value Better Career Prospects 6528 times. So we drop this column.
- Also the columns Prospect ID and Lead Number are of no use in the analysis, so we drop them.

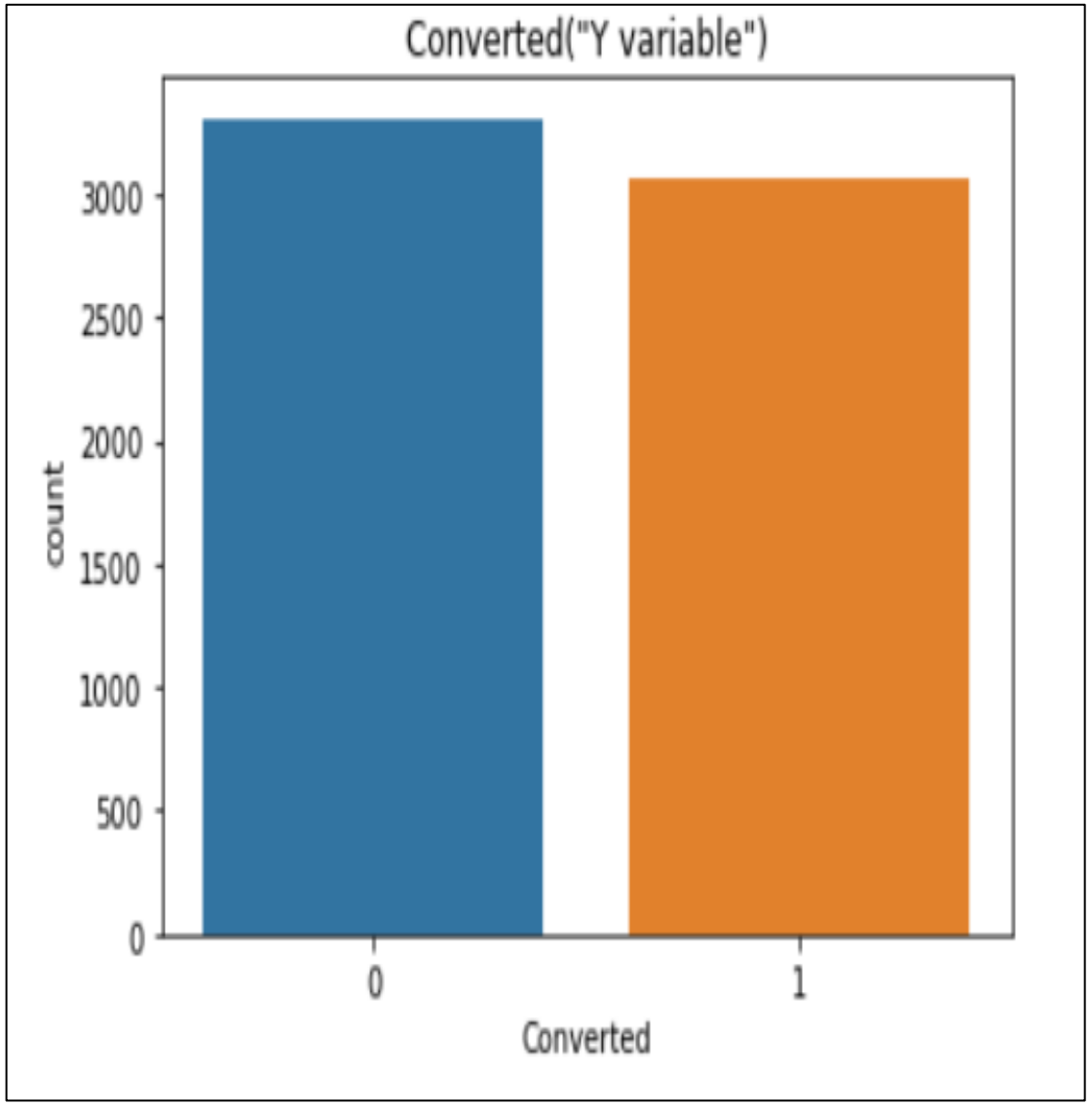
EDA (Exploratory Data Analysis)

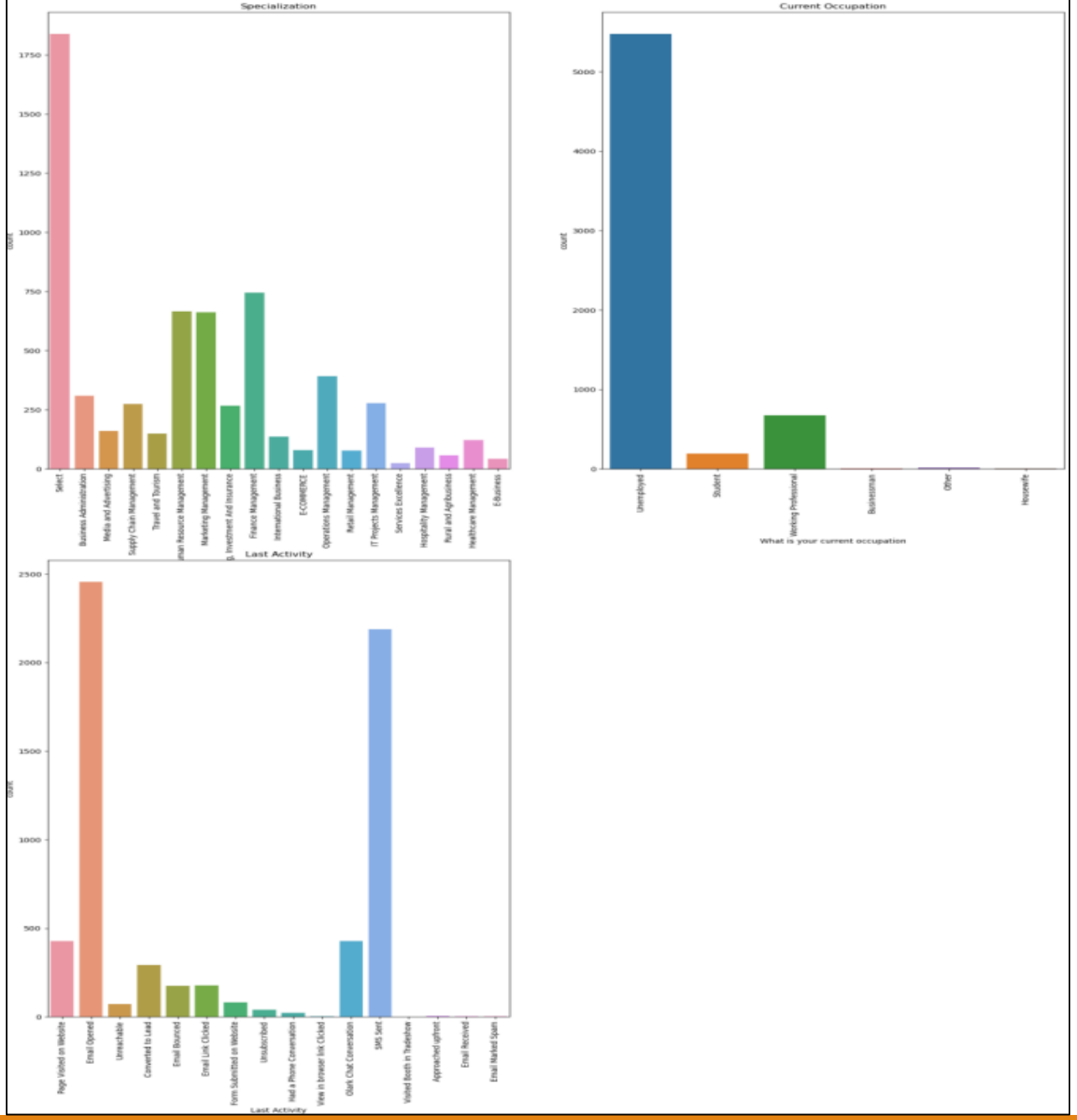
- Here are some univariate and bivariate data analysis done :

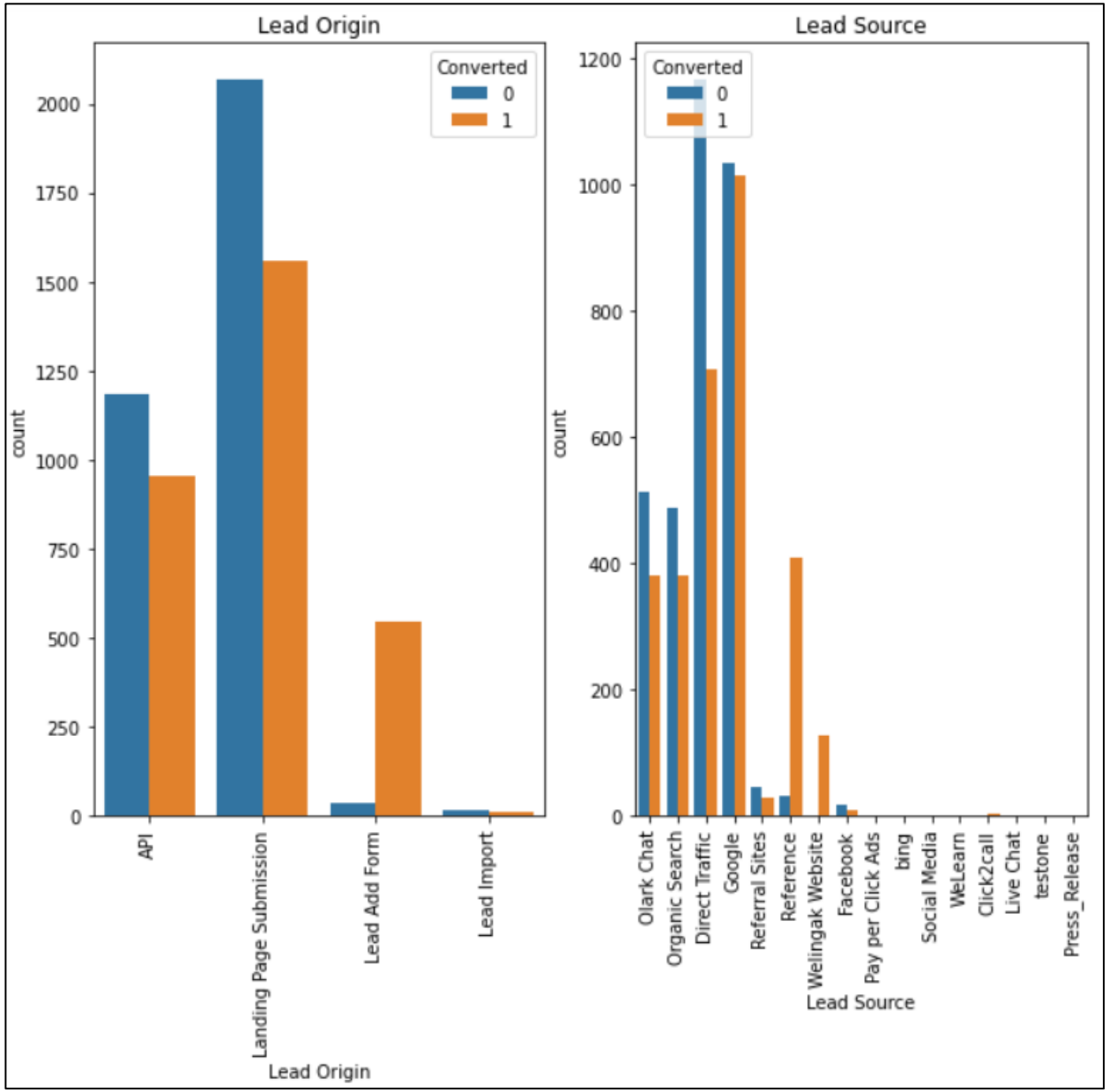





Total Visits
Total Time Spent on Website
Page Views Per Visit


Converted("Y variable")
count
Converted
0
1


Specialization
count
Select
Business Administration
Media and Advertising
Supply Chain Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
Finance Management
International Business
E-COMMERCE
Operations Management
Retail Management
IT Projects Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Last Activity
Current Occupation
count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Olark Chat Conversation
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Email Received
Email Marked Spam
Last Activity


Lead Origin
Converted
0
1
count
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin
Lead Source
Converted
0
1
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
testone
Press_Release
Lead Source

Correlation Among Variables

Data Conversion

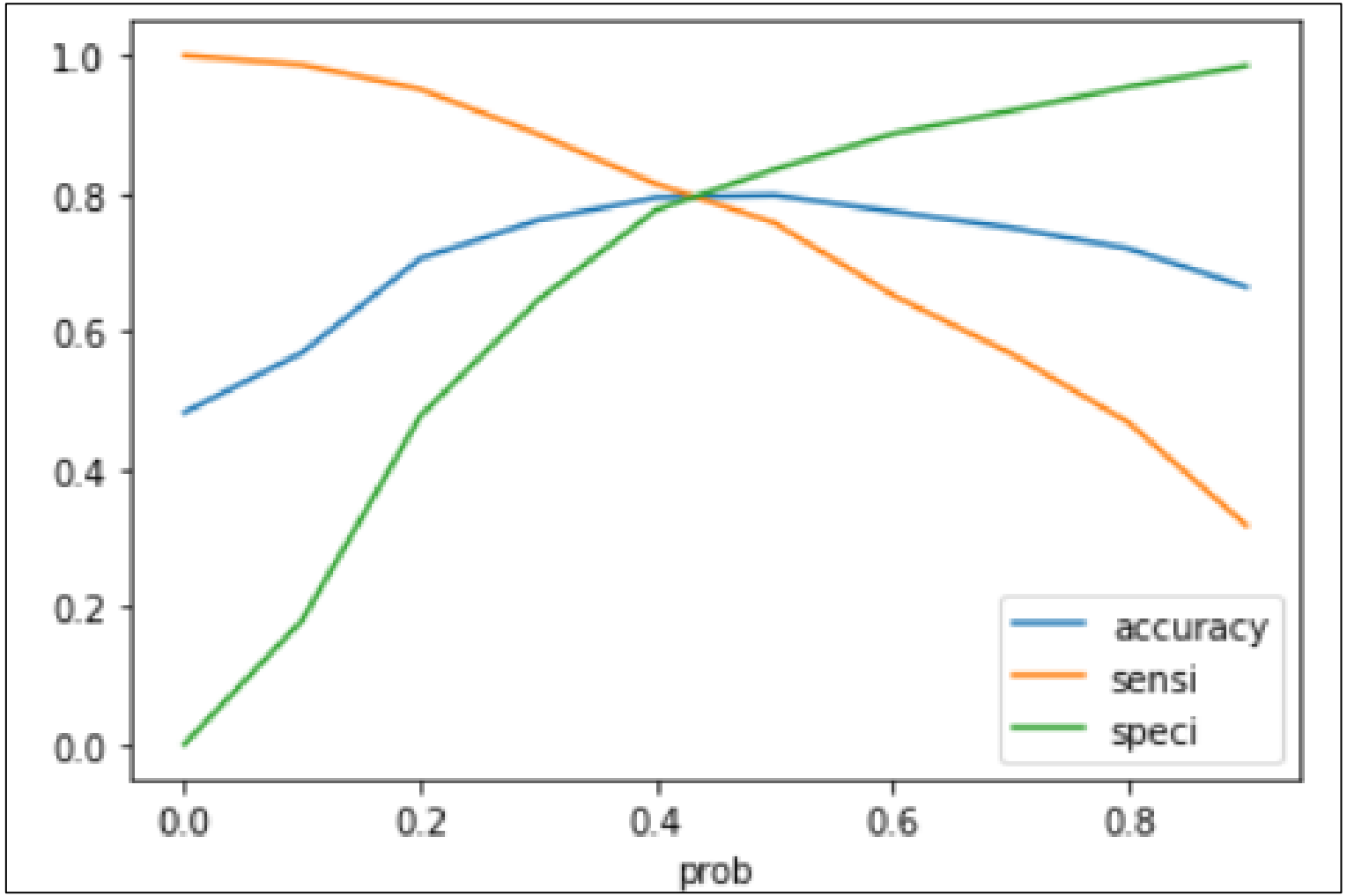
- Dummy variables are created for object type variables.
- After creating dummy variables number of rows and columns are like (4461,73).
- After scaling and model building with the help of split_test_train method we try to validate and predict based on the accuracy rate.

Model Building :

1. Splitting the data into test and train sets.
2. For that we try to split into (70-30) ratio.
3. Use RFE method for feature selection and running this with step 15 variables.
4. Building models with dropping P-values having greater than 0.05 and VIF value greater than 5.
5. And then prediction if data is done on test data set.
6. On the whole, we achieved accuracy level with 79.735 percent.
7. Sensitivity(75.70%), specificity(83.47%).

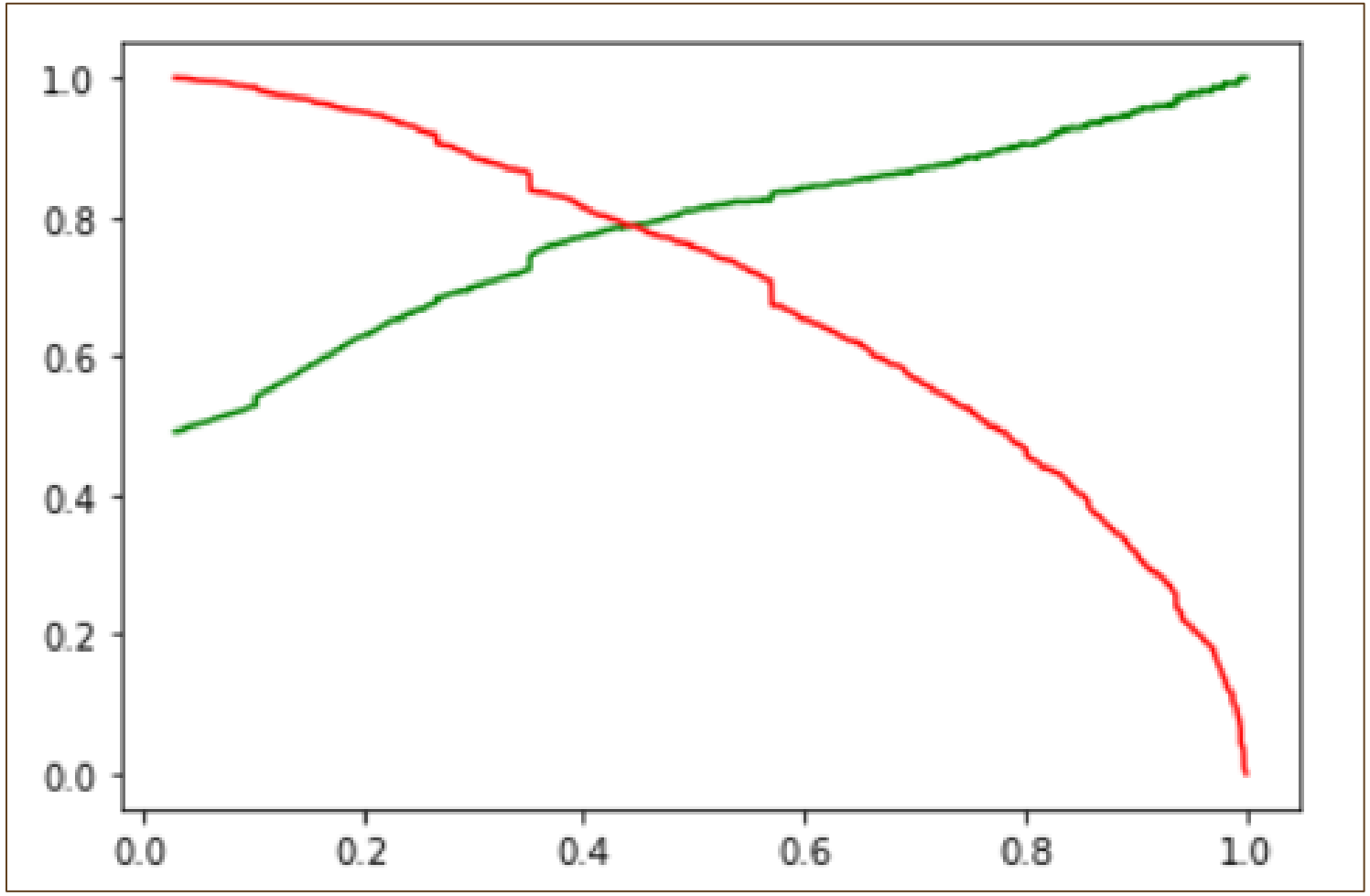
Model Evaluation on Train Data Set

- Accuracy = 79.62%
- Sensitivity = 80.13%
- Specificity = 79.15%



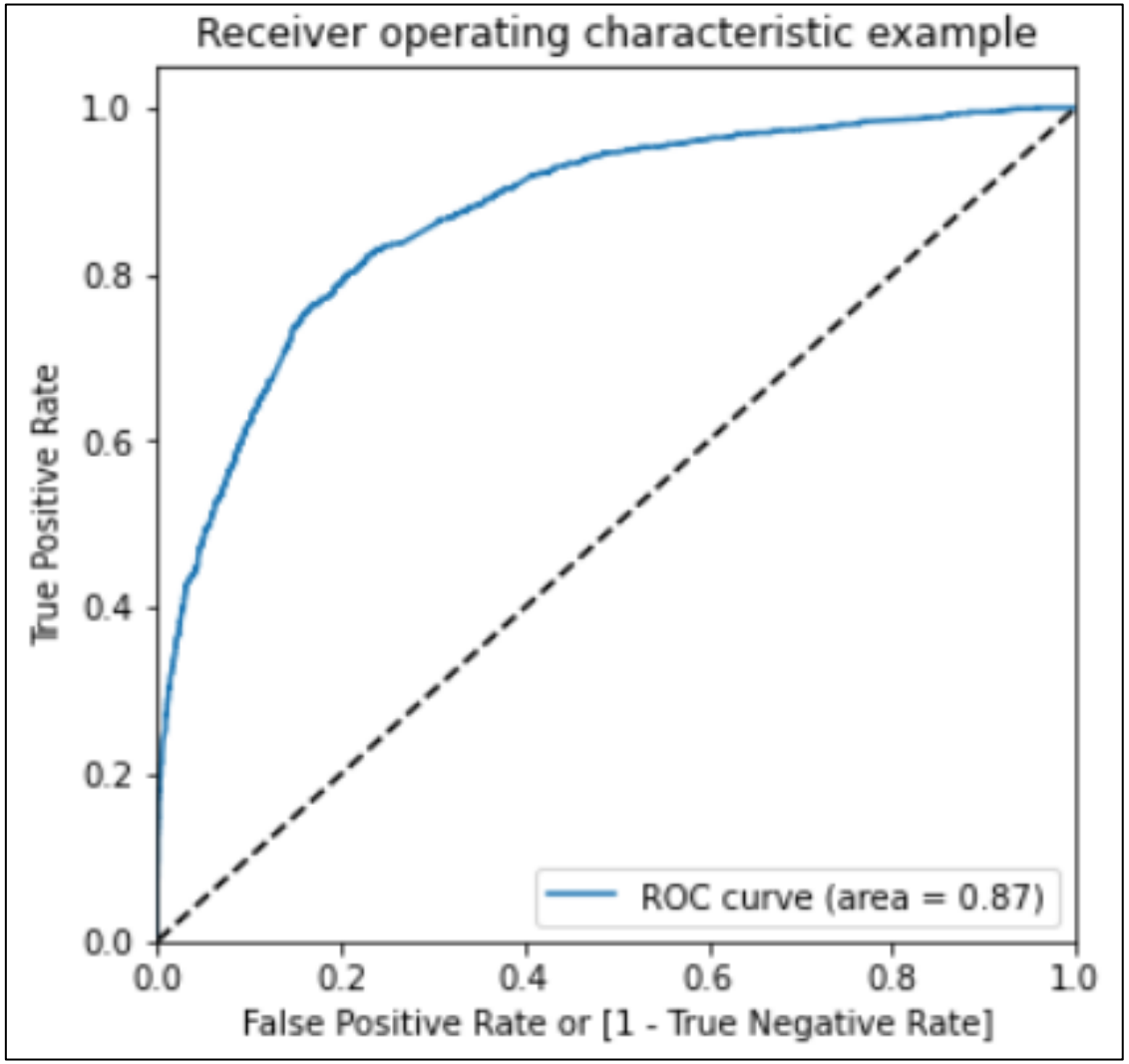
Model Evaluation

- The graph depicts optimal cutoff based on Precision and Recall:
  - Precision = 78%
  - Recall = 80%



ROC Curve

- Finding the optimal cut-off :
- And ROC curve from the graph we can say that :
  - The ROC Curve should be a value close to 1.
  - We are getting a good value of 0.87 indicating a good predictive model.



Model Evaluation on Test Data Set

- Accuracy = 75.83%
- Sensitivity = 88.86%
- Specificity = 63.85%

Conclusion

Logistic Regression Model-

- The model has an accuracy of close to 80%.
- Accuracy, sensitivity, specificity, precision, and recall curves were used to select the threshold.
- The model has a sensitivity of 81% and a specificity of 80%.
- The model differentiates between legitimate, promising leads and leads with lower conversion rates.
- Overall, this model is accurate.

Thank You!!!