X Education - Lead Scoring Case Study

Team Members: Mukesh chede, Sankalp Shah & Gaurav Singh Bisht

Table of Contents

.

Background of X Education Company

- 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.
- 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%.

Problem Statement & Objective of the Lead Case Study

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

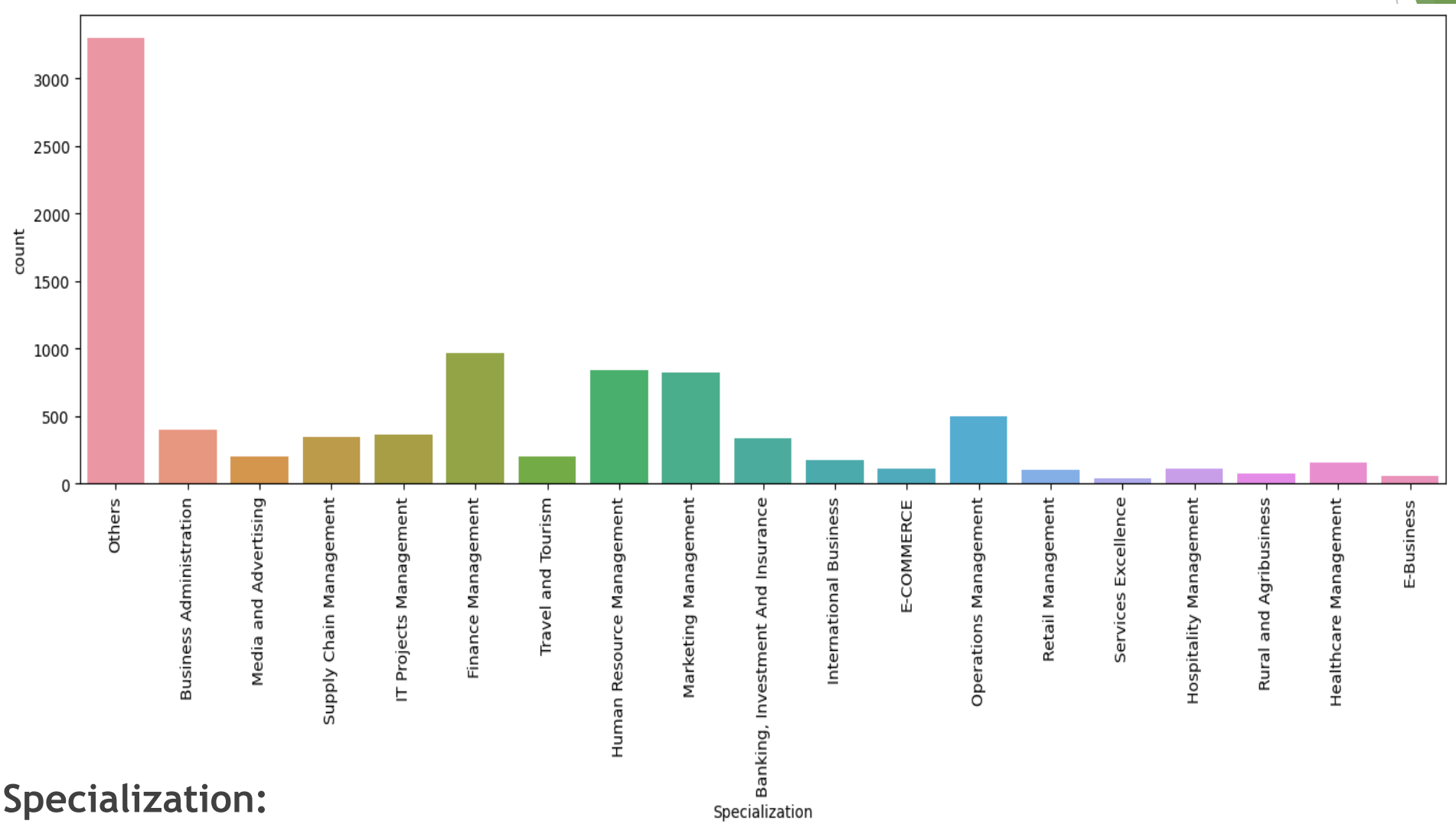
Objective of the Study:

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

Step Involved: -

- Understanding the domain/ variables.
- Import/ Load the data.
- Check the structure/ metadata.
- Missing value check.
- Visualising the Data
- Scaling the feature variable
- Model Building
- Evaluating Model and confusion matrix
- Plotting ROC Curve
- Making Predictions on test set
- Test set Model Evaluation

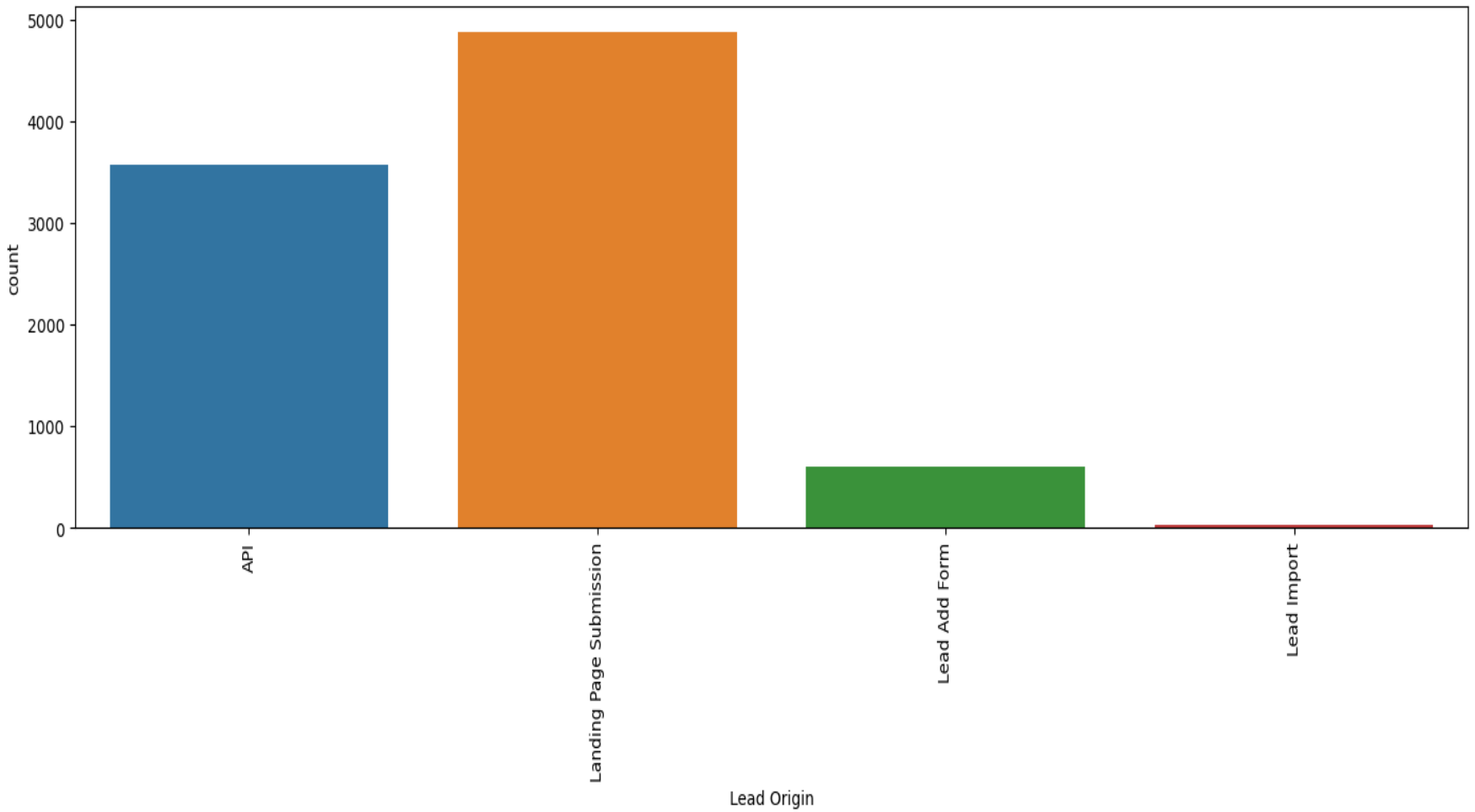
EDA



Specialization:

- Marketing Management, Human Resource Management, Operation Management, Finance Management shows good contribution in Leads conversion than other specialization.

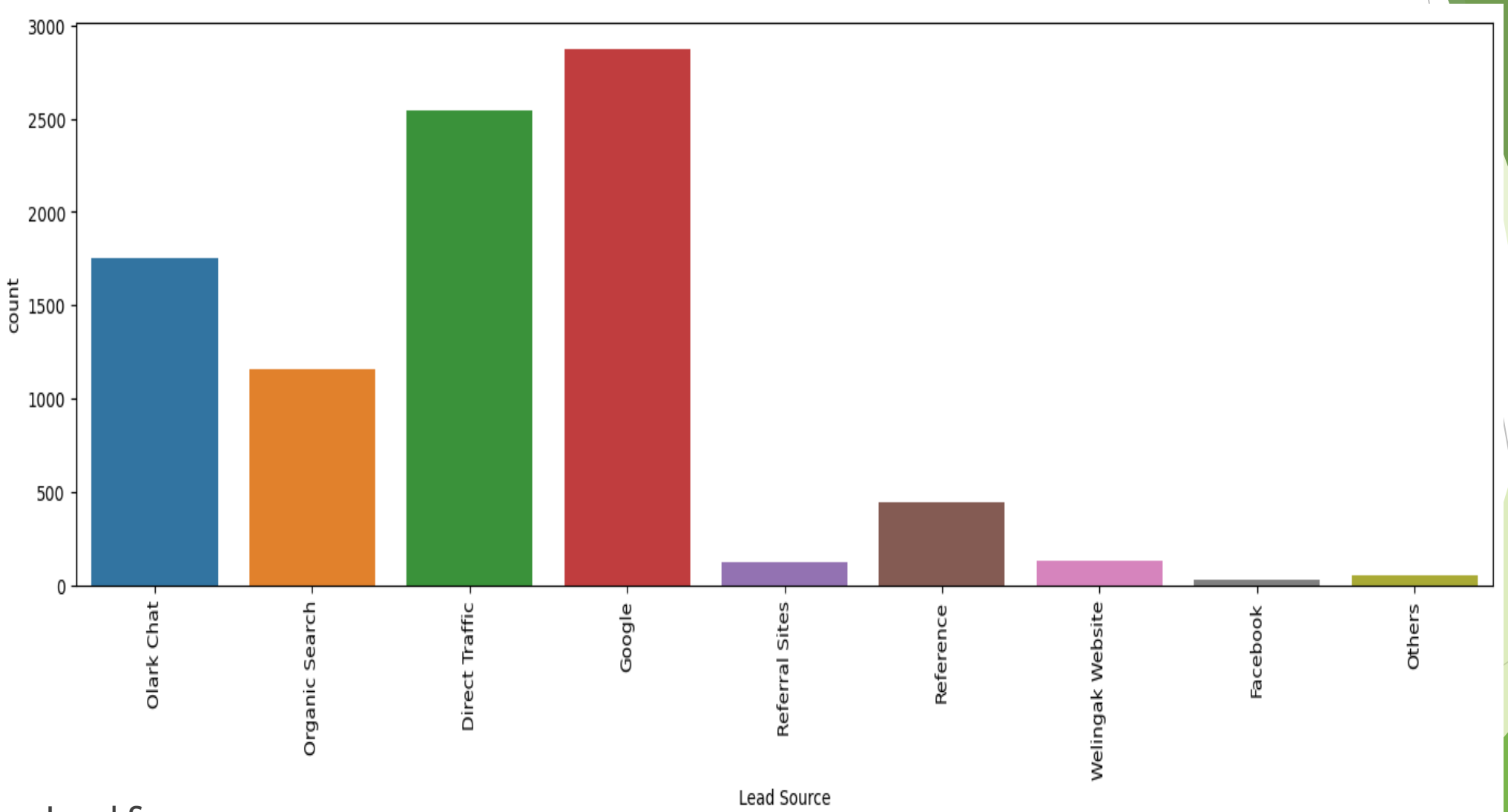
EDA



Lead Origin:

- Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%.
- The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.

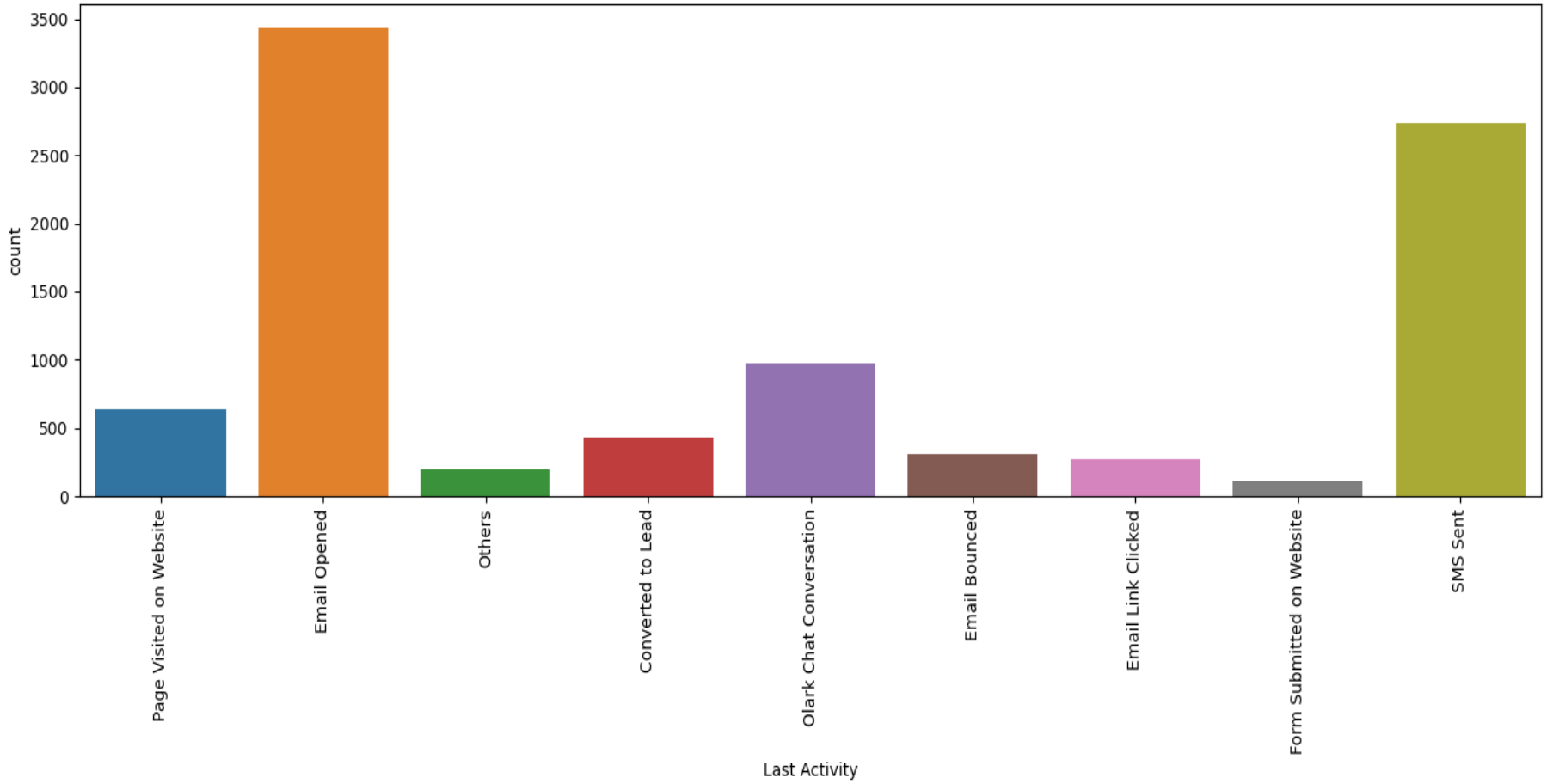
EDA



Lead Source:

Google, Direct Traffic and Olark Chat shows good contribution in Lead Source

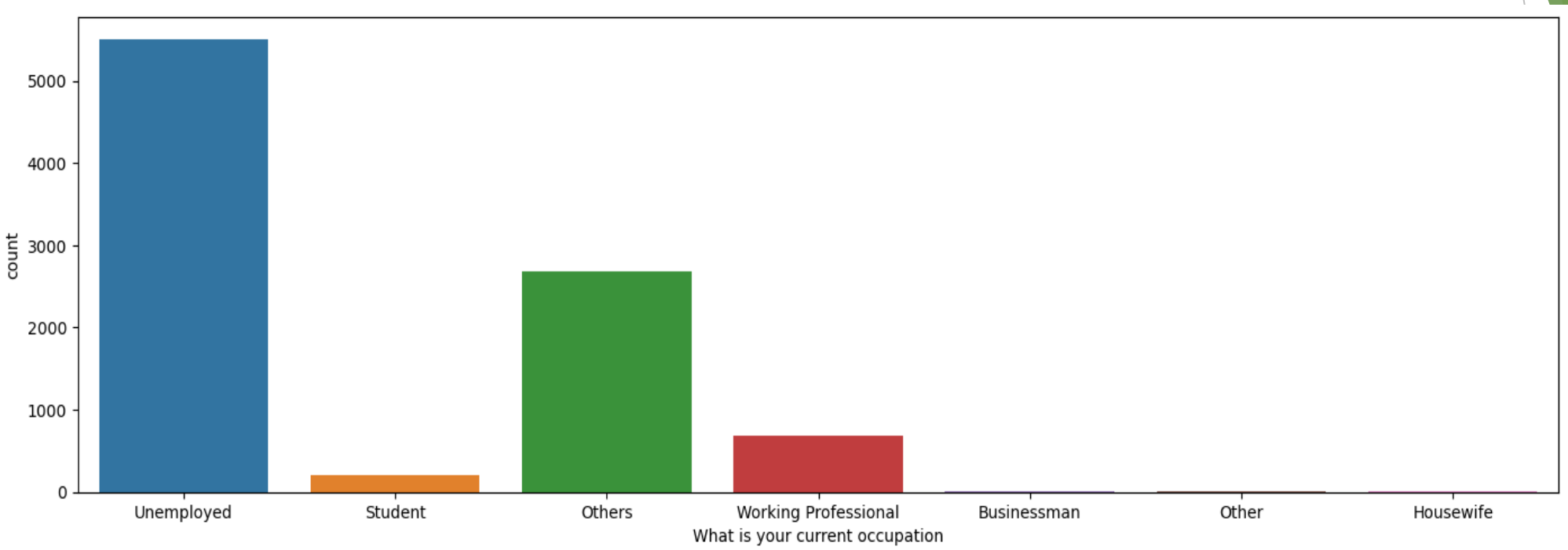
EDA



Last Activity:

Email Opened, SMS Sent and Olark Chart Conversation shows good contribution in Leads conversion than other specialization.

EDA



What is your current occupation:

- Mostly customers are unemployed.
- Only a few customers are belonging to Businessman, Housewife and Students.

Data Preparation before Model building

- Checking Null Values and outliers in the data.
- Binary level categorical columns(Do Not Email & Do Not Call) were mapped to 1 / 0.
- Created dummy features for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, What is your current occupation.
- Splitting Train & Test Sets in 70:30 % ratio.
- Feature scaling
- MinMax Scaler method was used to scale the features
- Checking the correlations.

Model Building

Feature Selection:

- Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.
- It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

```
rfe = RFE(logreg, n_features_to_select=15)
```

- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns(in this case 15 columns were selected).

Model Building

- Model 3 was stable with the threshold
  - p-values < 0.05 and
  - VIFs less than 5
- Hence, logm3 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

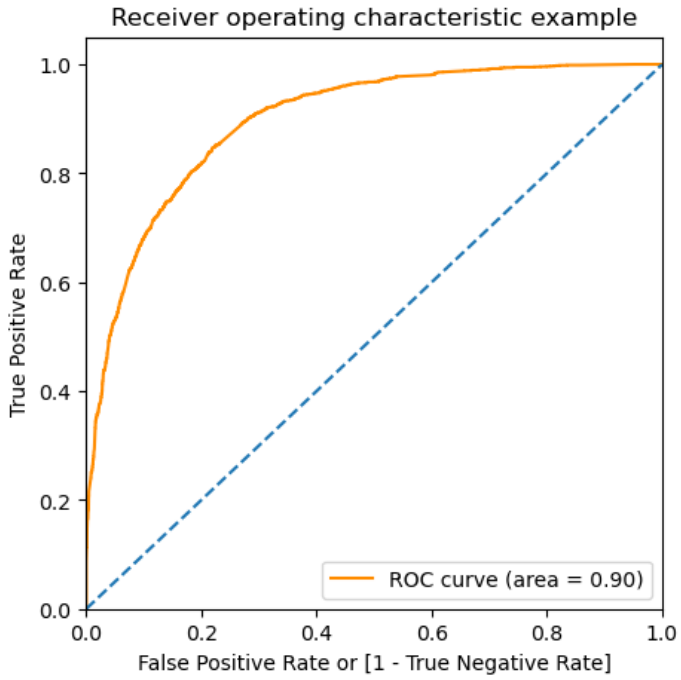
Model Evaluation

ROC Curve –

Train Data Set

● Area under ROC curve is 0.90 out of 1 which indicates a good predictive model.

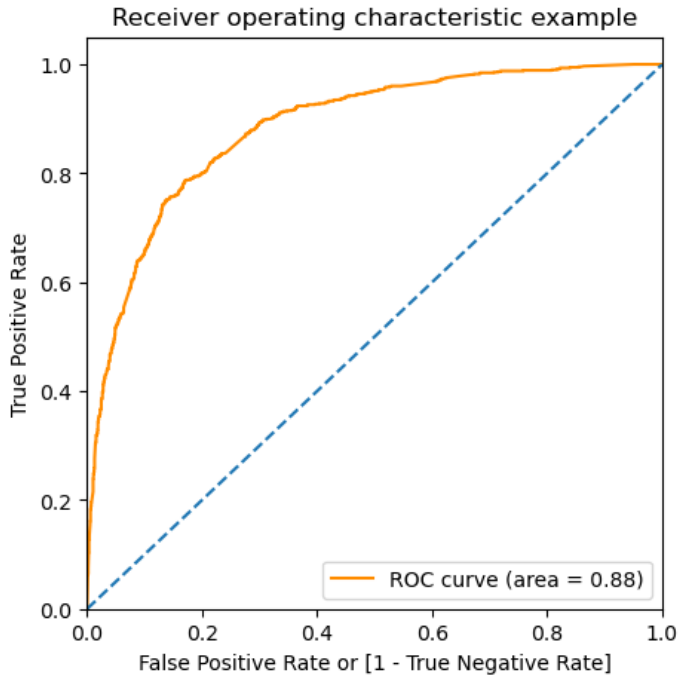
● The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve – Test Data Set

● Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.

● The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



- Using a cut-off value of 0.345, the model has achieved following data for Test set: -
  - **Accuracy : 80.66%**
  - **Sensitivity : 79.44%**
  - **Specificity : 80.37%**

Recommendations:

•The company **should make contact** to the leads coming from the lead sources "Welingak Websites" ,"Companies Websites" and "Add Form" as these are more likely to get converted.
•The company **should make calls** to the leads who are the "working professionals" as they are more likely to get converted.
•The company **should make calls** to the leads coming from the lead sources "Olark Chat" and spending less time in websites as these are more likely to get converted.
•The company **should make intereaction** to the leads whose last activity was SMS Sent as they are more likely to get converted.
•The company **should not make calls** to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
•The company **should not make calls** to the leads whose Specialization was "Others" as they are not likely to get converted.
•The company **should not make calls** to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.