LEAD SCORING CASE STUDY

Group Members

- Satti Srikanth Reddy
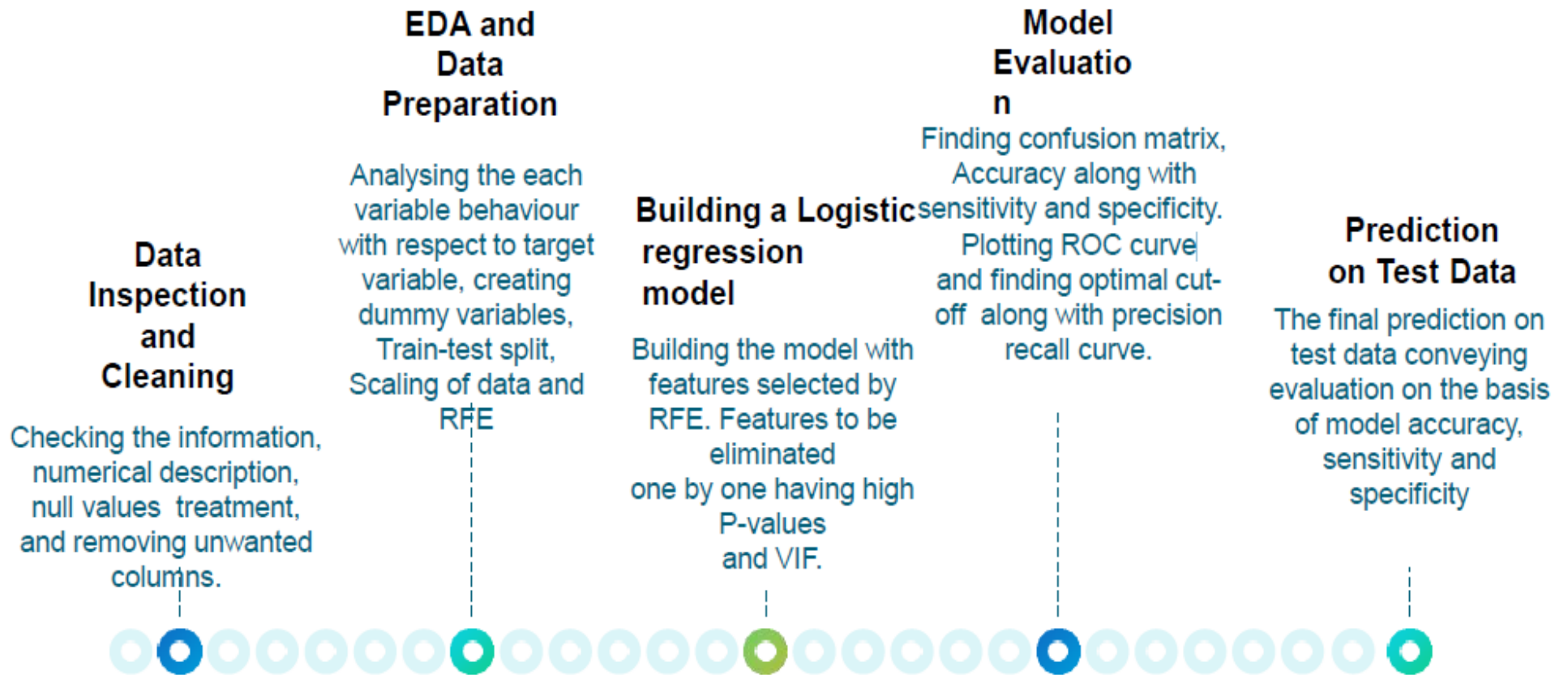- Vivek Sharma
- Mohammed Imran

Problem Statement

➢ X Education sells online courses to industry professionals.

➢ X Education gets a lot of leads, 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.

Business Objective

- ✓ X education wants to know most promising leads.

- ✓ For that they want to build a Model which identifies the hot leads.

- ✓ Deployment of the model for the future use.

Solution Methodology

EDA and Data Preparation

Analysing the each variable behaviour with respect to target variable, creating dummy variables, Train-test split, Scaling of data and RFE

Data Inspection and Cleaning

Checking the information, numerical description, null values treatment, and removing unwanted columns.

Building a Logistic regression model

Building the model with features selected by RFE. Features to be eliminated one by one having high P-values and VIF.

Model Evaluation

Finding confusion matrix, Accuracy along with sensitivity and specificity. Plotting ROC curve and finding optimal cut-off along with precision recall curve.

Prediction on Test Data

The final prediction on test data conveying evaluation on the basis of model accuracy, sensitivity and specificity

Data Manipulation

➢ Total Number of Columns =37, Total Number of Rows =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 32% (null values >= 3000) as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

Categorical Variable Relation – Lead Origin
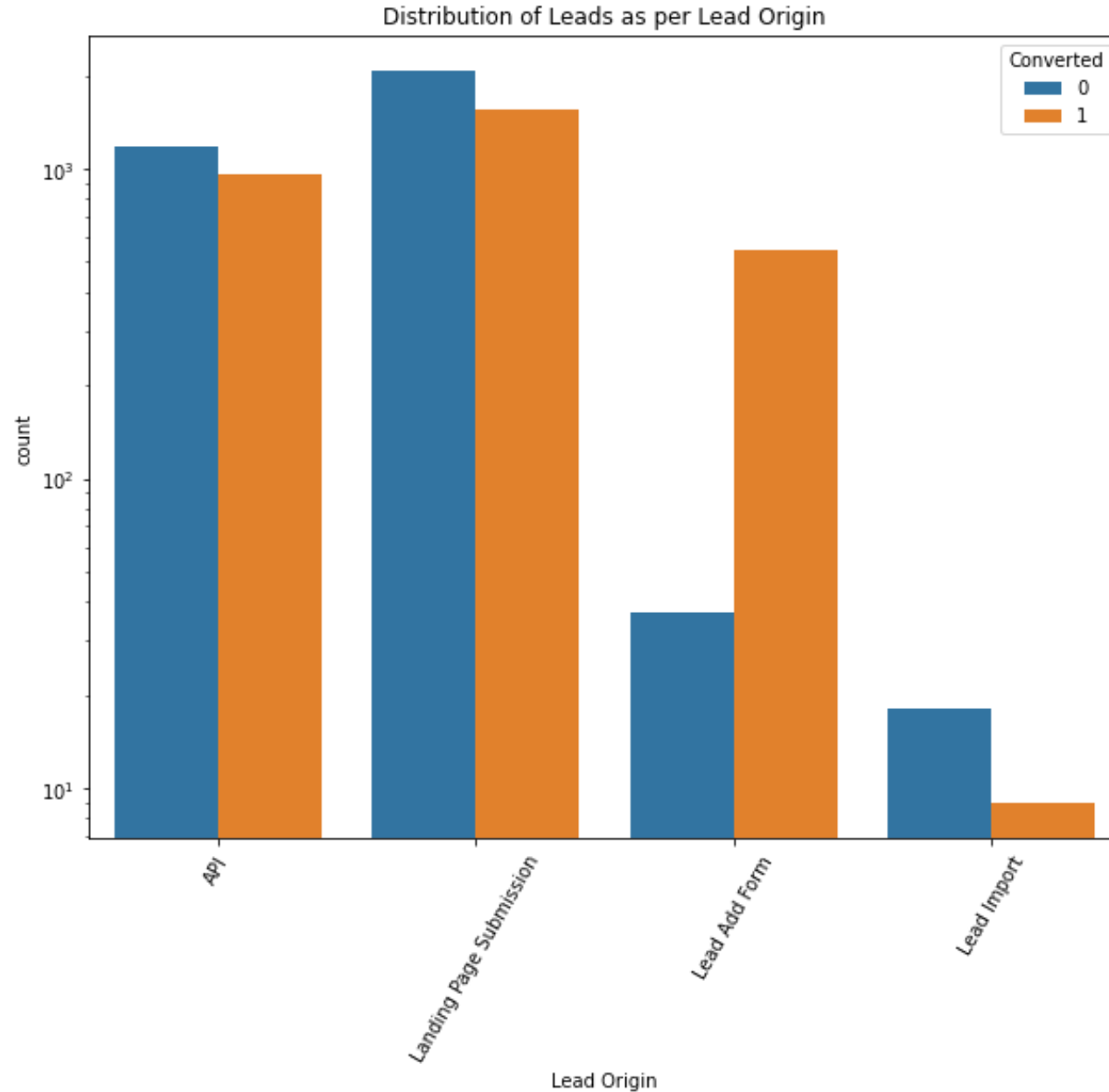
Observation

➢ Most of the Leads originated from submissions on the landing page, followed by API, where around 30% are converted.

➢ Leads from Quick Add Form are 100% Converted, there was just 1 lead from that category.

➢ Leads from the Lead Add Form are the next highest conversions in this category at around 90% of 718 leads.

➢ Lead Import are very less in count and conversion rate is also the lowest

➢ To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page

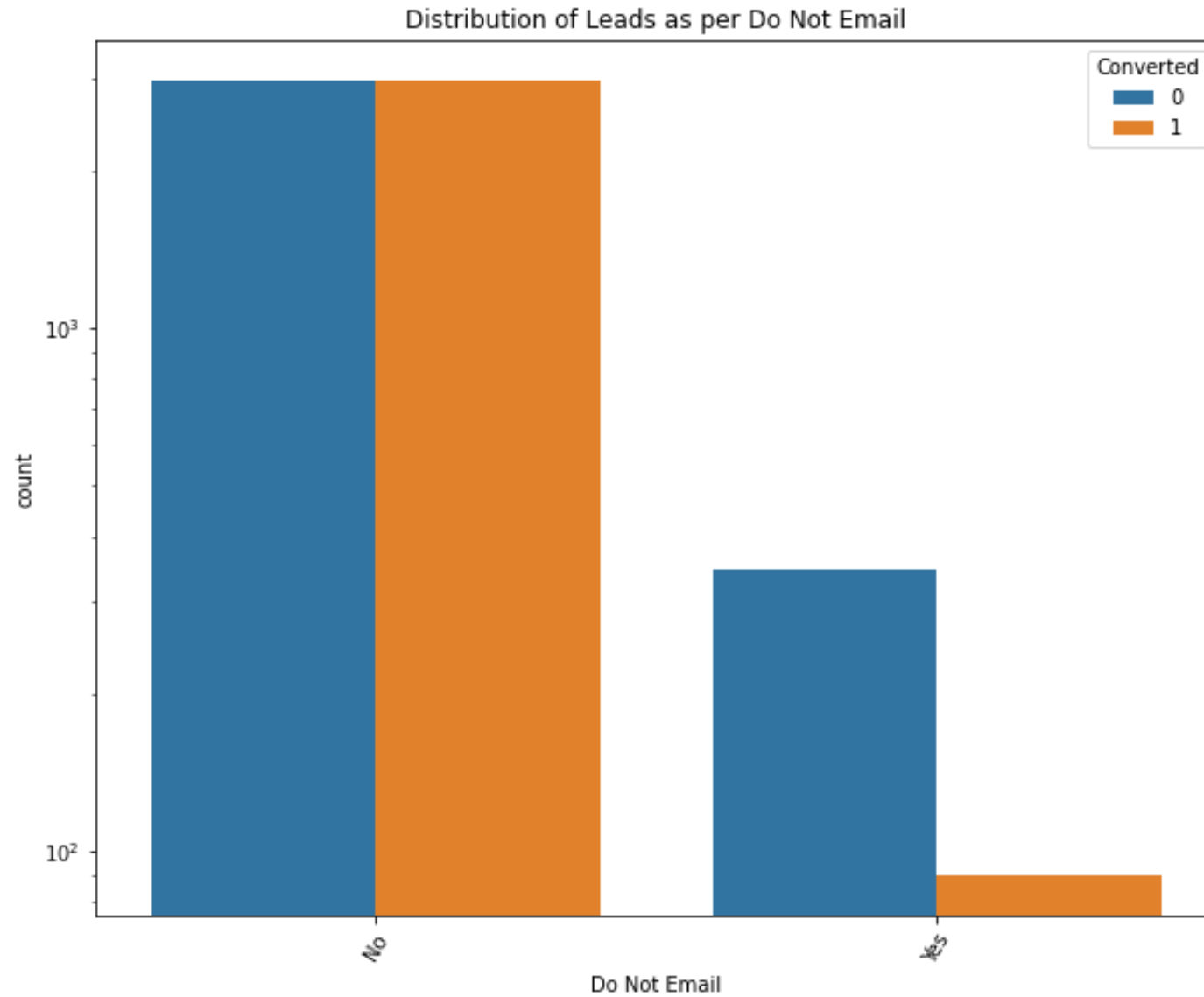➢ Submission origin and generate more leads from Lead Add Form

Categorical Variable Relation – Lead Source

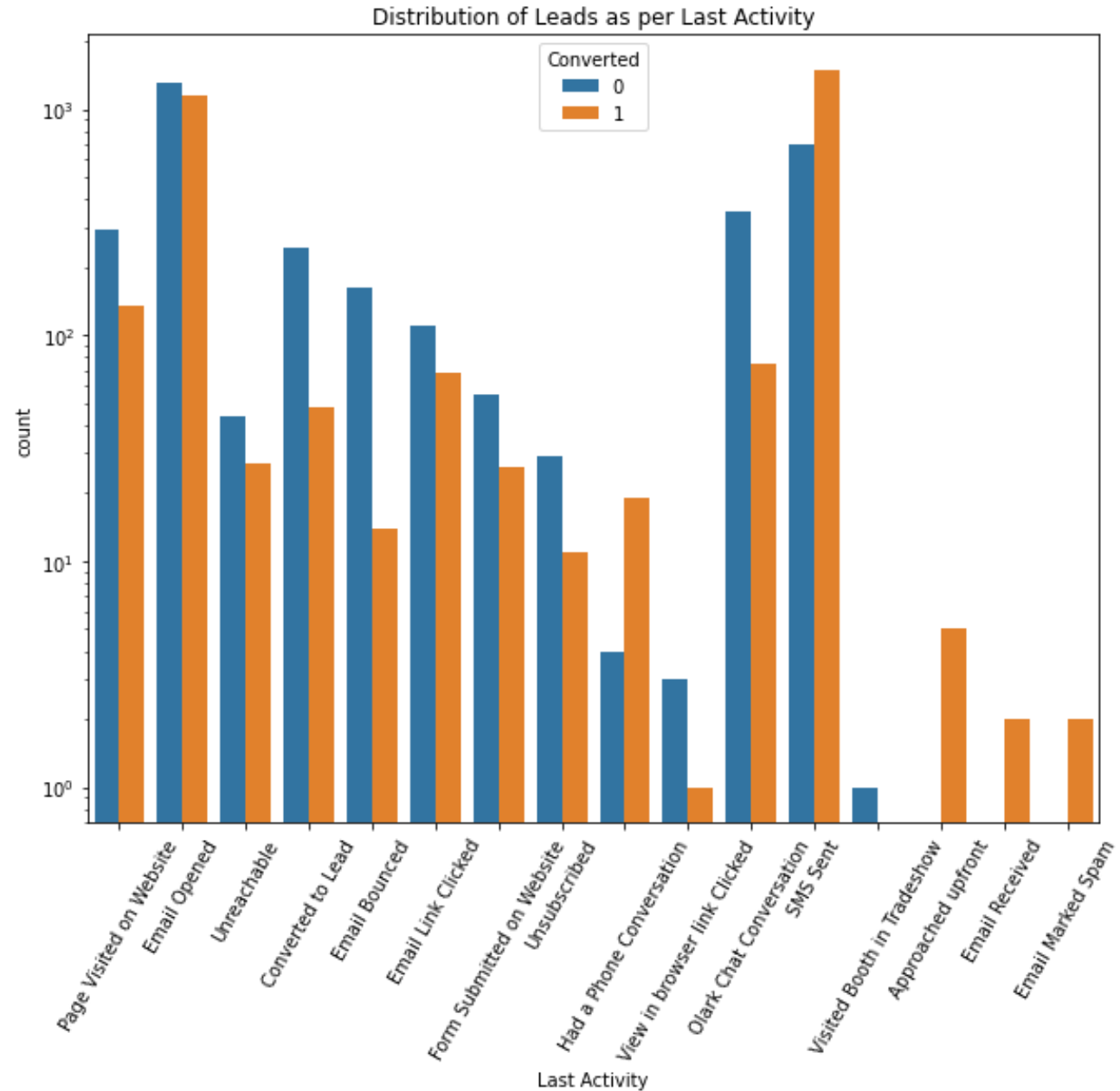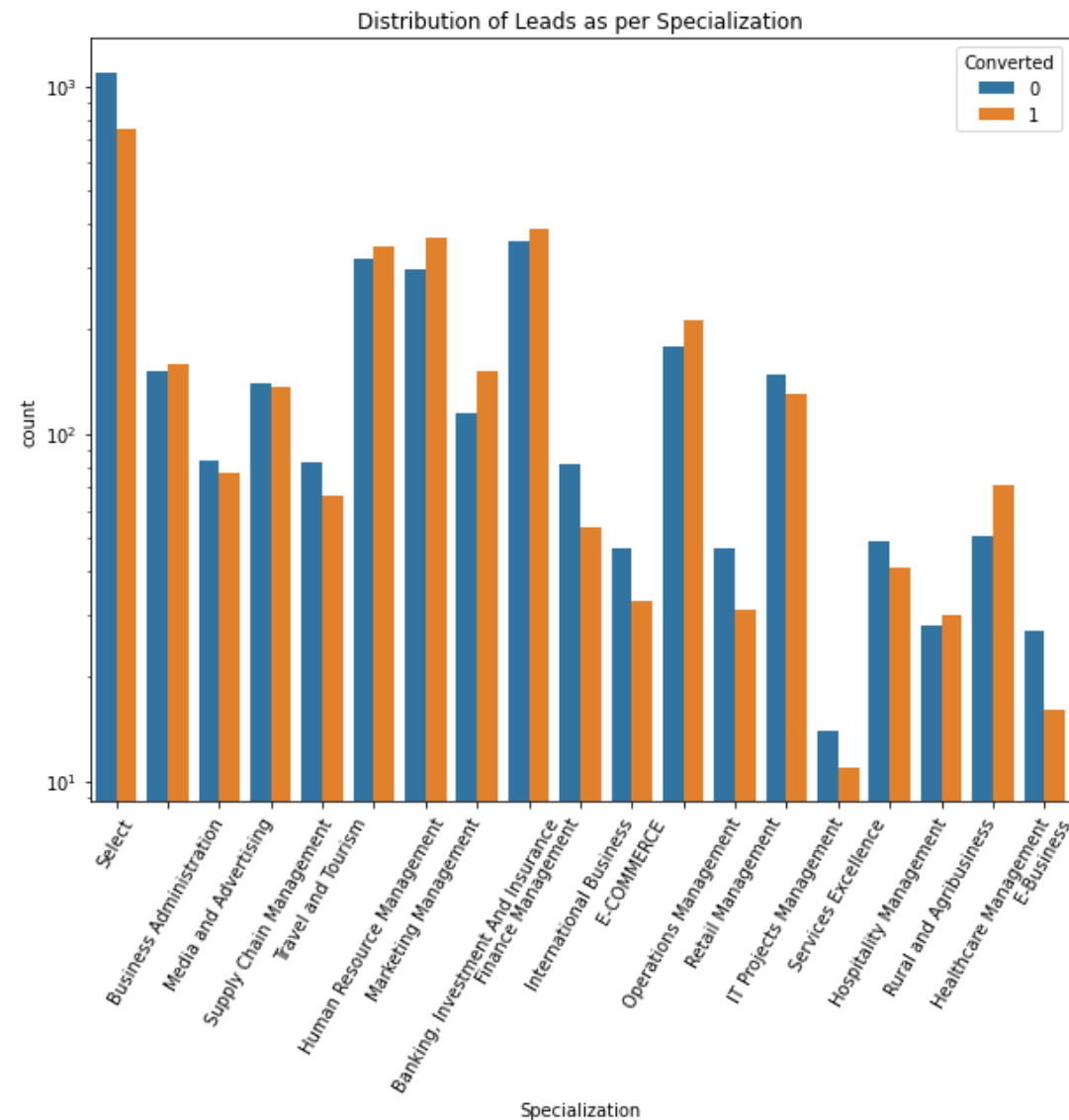Distribution of Leads as per Lead Source
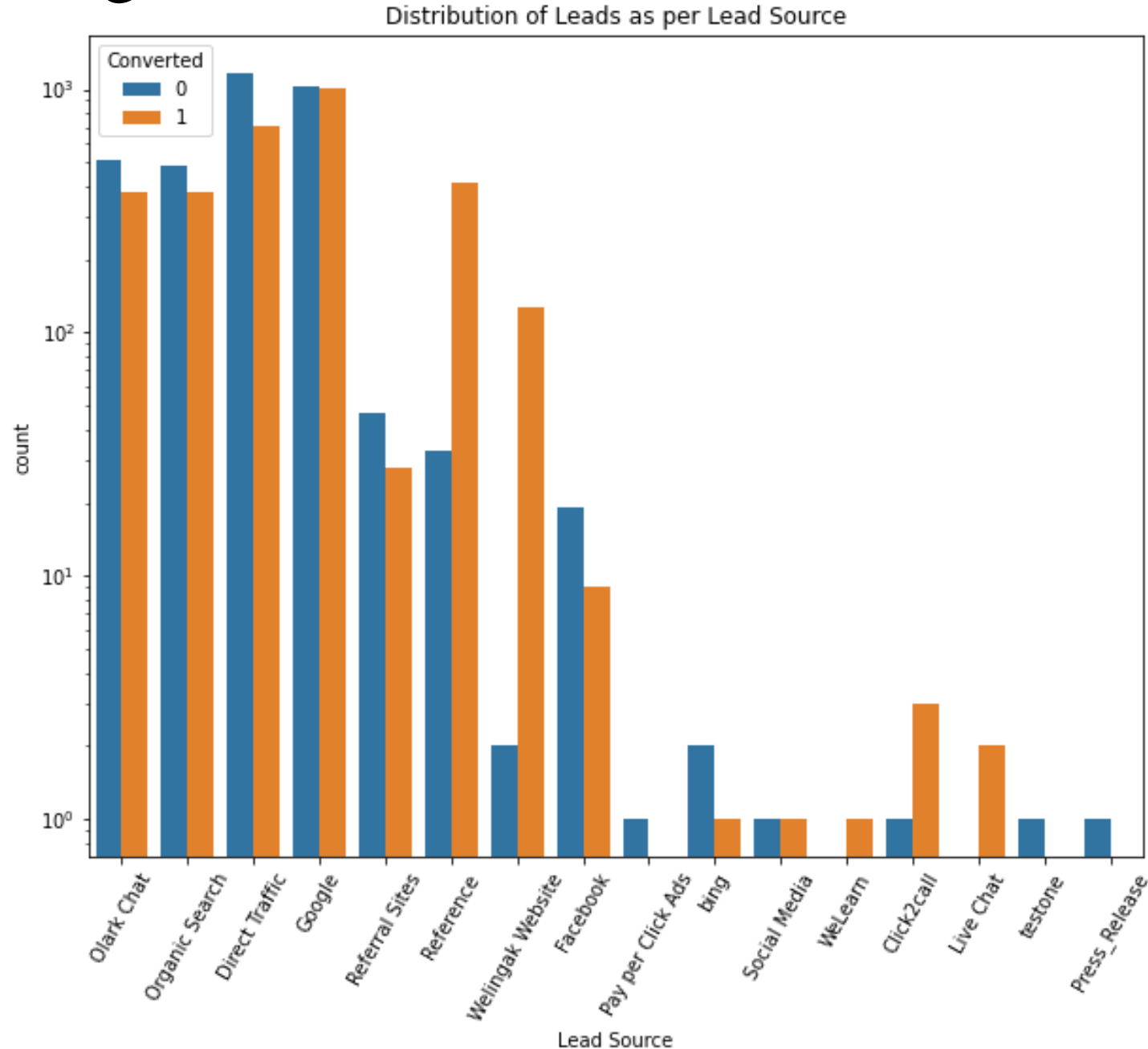
Observation

➢ The source of most leads is Google, and 40% of the leads converted, followed by Direct Traffic, Organic search and Olark chat

➢ A lead that came from a reference has over 90% conversion from the total of 534.

➢ Welingak Website has almost 100% lead conversion rate. This option should be explored more to increase lead

➢ conversion

➢ To increase lead count, initiatives should be taken so already existing members increase their referrals.

DO NOT EMAIL Analysis

Distribution of Leads as per Do Not Email

Last Activity Analysis

Distribution of Leads as per Last Activity

Distribution of Leads as per Specialization

➢**Specialization analysis** - Most of the leads have not mentioned a specialization and around 28% of those converted.
Leads with Banking Investment and insurance and Marketing Management - Over 45% Converted

Distribution of Leads as per Occupation
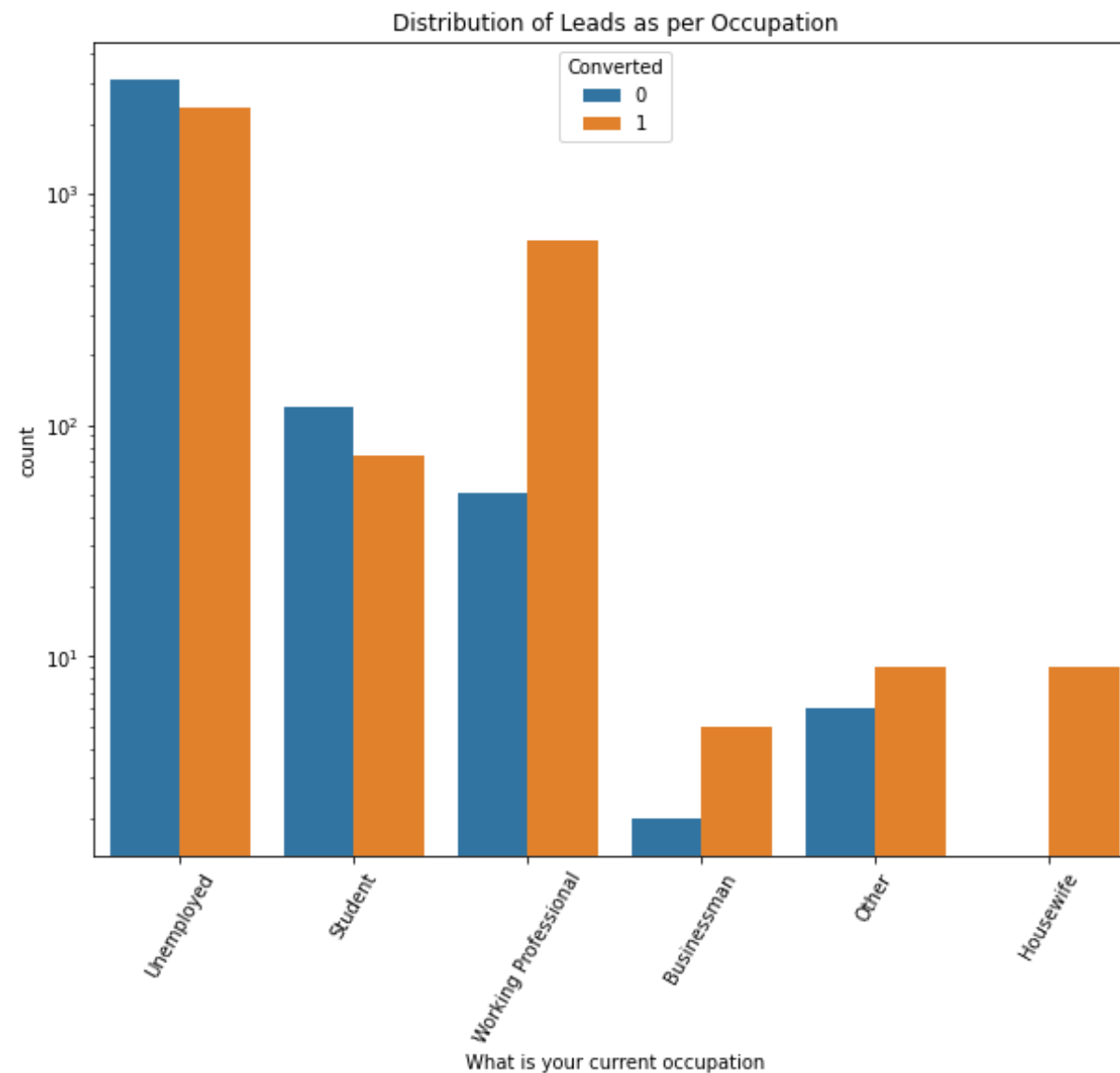
➢ **Current Occupation analysis -**Housewives are less in numbers, but have 100% conversion rate. Working professionals, Businessmen and Other have high conversion rate.

Leads with Unemployed occupation is highest in number, but the conversion rate is low (~40%)

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 p-value is greater than 0.05 and vif value is greater than 5

➢ Predictions on test data set

➢ Overall accuracy 79.09%

Feature Importance

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Top three variables in our model that contribute most towards the probability of a lead getting converted
- Total Visits
- Total Time Spent on Website
- Page Views Per Visit

ROC Curve
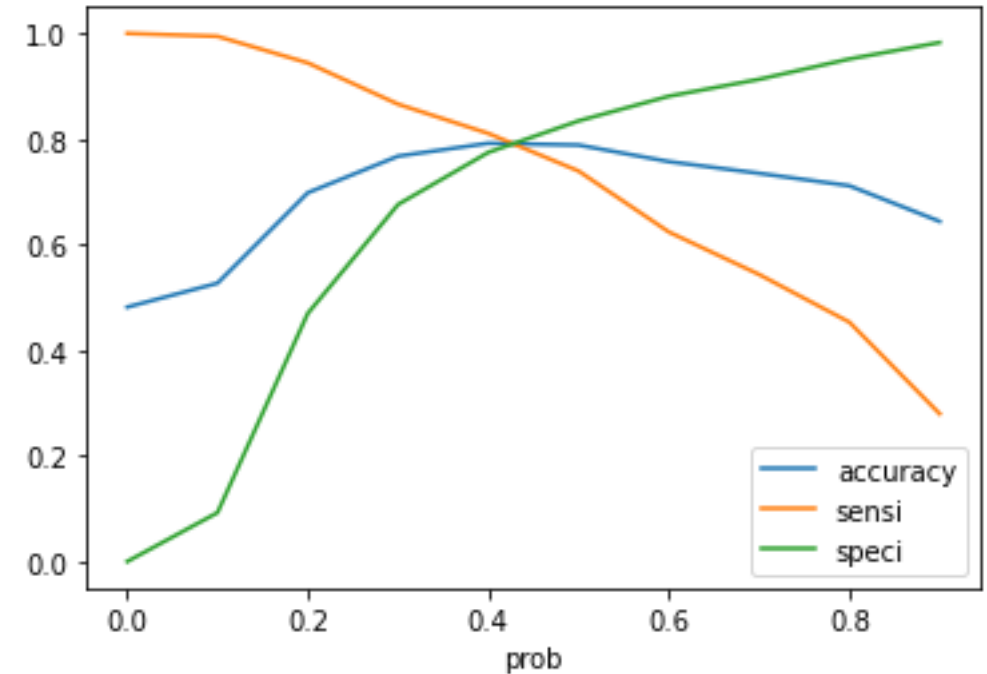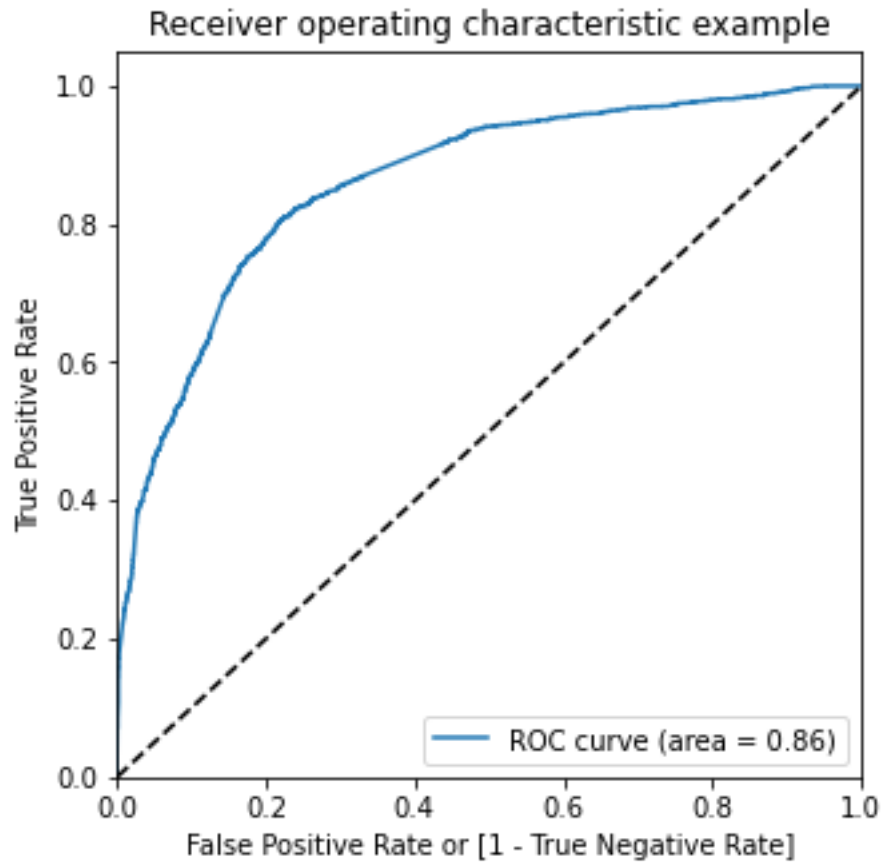
➢ Finding Optimal cut off probability is that probability where we get balanced sensitivity and specificity from the second graph it is visible that the optimal cut off is at 0.42.

Model Evaluation – Train Data

0.42 - Probability Threshold

Accuracy	Sensitivity	Specificity	Recall	Precision
79.09	79.34	78.84	77.71	78.40

Model Evaluation – Test Data

0.44 - Probability Threshold

Accuracy	Sensitivity	Specificity	Recall	Precision
78.45	77.94	78.92	76.75	78.29

Final Prediction

Top 5 Records

	Converted	Conversion_Prob	final_predicted
0	1	0.996296	1
1	0	0.129992	0
2	0	0.703937	1
3	1	0.299564	0
4	1	0.720796	1

✓ **Higher the conversion probability, higher is the chance of a lead getting converted and vice versa,**

✓ **Since, the final probability threshold is 0.42 for deciding if a lead will convert or not, any lead with a conversion probability of 0.42 or above will have a value of '1' in the Final predicted column.**

Recommendation

- 'Total Visits' , 'Total Time Spent on Website' , 'Page Views Per Visit' contribute most towards the probability of a lead getting converted.

- A list of leads should be kept handy so that they can be informed about the new courses, services, job offers and future higher studies.

- Monitoring of each lead should be done utmost care so that lead specific information can be tailored.

- Care should be taken in providing job offerings, information on courses that suits best according to the interest of the leads.

- A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.

- Converted leads should be focused.

- Question-answer sessions with leads should be held to extract the right information required about them.

- Further inquiries and appointments with the leads are to be arranged to determine their intention and mentality to join online courses.