Lead Scoring Case Study

Sreekanth Padmanabhankutty
Kudapu Sree Durga
Srinivasacharyulu Daroori

Problem Statement

Question of Interest

- X Education offers online courses to professionals in various industries.
- The company generates a significant number of leads, but its conversion rate is low; for instance, out of 100 leads acquired in a day, only about 30 are converted.
- To enhance this process, X Education aims to identify the most promising leads, referred to as "Hot Leads."
- By successfully pinpointing these leads, the conversion rate is expected to improve, allowing the sales team to concentrate their efforts on

Business Goals

- Develop logistic regression model capable of pinpointing promising leads
- Identify the deficiencies in the current SOP

General Workflow

Too build a logistic regression model for assigning lead scores between 0 and 100, follow these steps:

1.Data Preparation:

1. Collect data on leads, including features that may influence conversion (e.g., demographic information, engagement metrics (time spent on website etc).
2. Preprocess the data by handling missing values, encoding categorical variables, and normalizing numerical features.

2.Model Development:

1. Split the dataset into training and testing sets.
2. Use a logistic regression algorithm to train the model on the training set.
3. Adjust the model to output scores between 0 and 100, possibly by multiplying the predicted probabilities (which range from 0 to 1) by 100.

3.Evaluation:

1. Evaluate the model using appropriate metrics (e.g., accuracy, precision, recall, ROC-AUC) on the testing set.
2. Adjust hyperparameters and refine the model based on evaluation results.

4.Handling Future Changes:

1. Ensure the model is flexible enough to incorporate new features or adapt to changes in data distributions.
2. Document potential adjustments needed for future requirements.

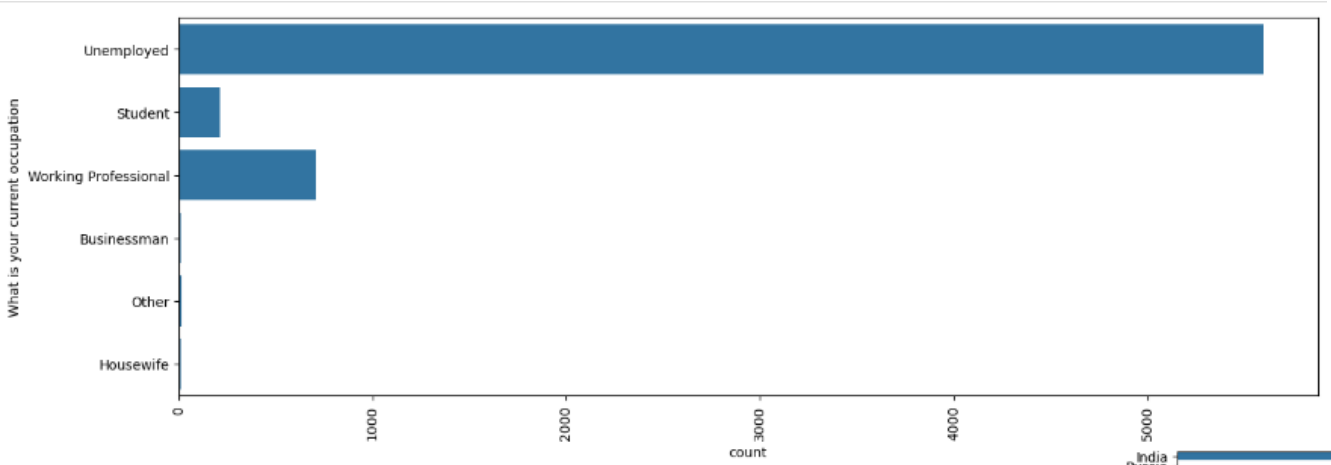
5.Documentation and Recommendations:

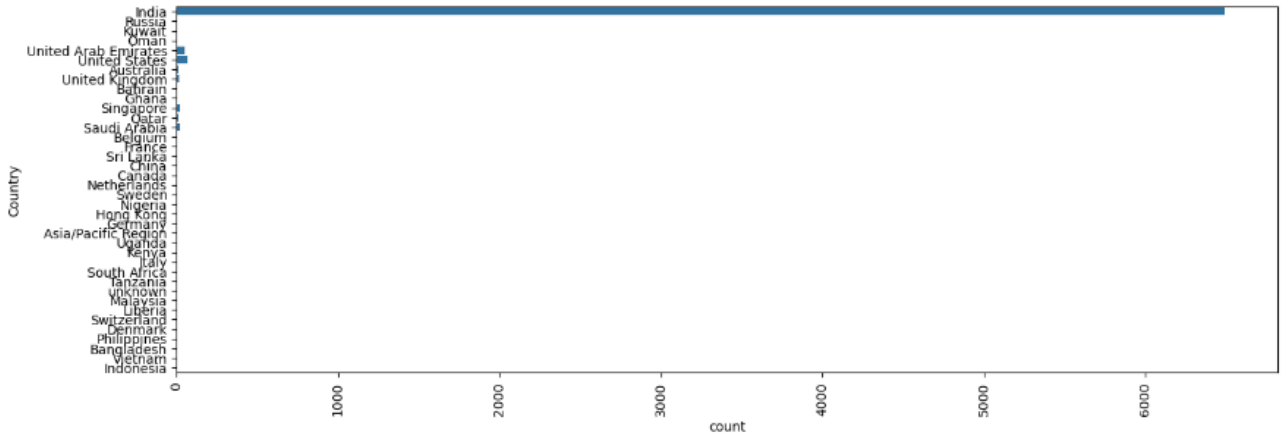
1. Create a presentation (PPT) summarizing the model development process, evaluation results, and recommendations for implementation.

Handling Data

- Total Rows: 37, Total Columns: 9,240.
- Features with single value, such as “Magazine,” “Receive More Updates About Our Courses,” “Update Me on Supply,” “Chain Content,” “Get Updates on DM Content,” and “I Agree to Pay the Amount Through Cheque,” have been removed.
- The “Prospect ID” and “Lead Number” have also been eliminated, as they are unnecessary for the analysis.
- The “Country” and “City” have also been eliminated, as they are highly skewed an has non relevance for the analysis.
- Upon reviewing the value counts for several object-type variables, we identified certain features with insufficient variance that we decided to drop, including: “Do Not Call,” “What Matters Most to You in Choosing a Course,” “Search,” “Newspaper Article,” “X Education Forums,” “Newspaper,” and “Digital Advertisement.”
- There are variables with missing data , we have tried to address the missing data by imputing as “other”, again the skewness of the data was observed and based on that decision was taken to delete or keep the data
- We also removed columns with over 35% missing values, such as “How Did You Hear About X Education” and “Lead Profile.”

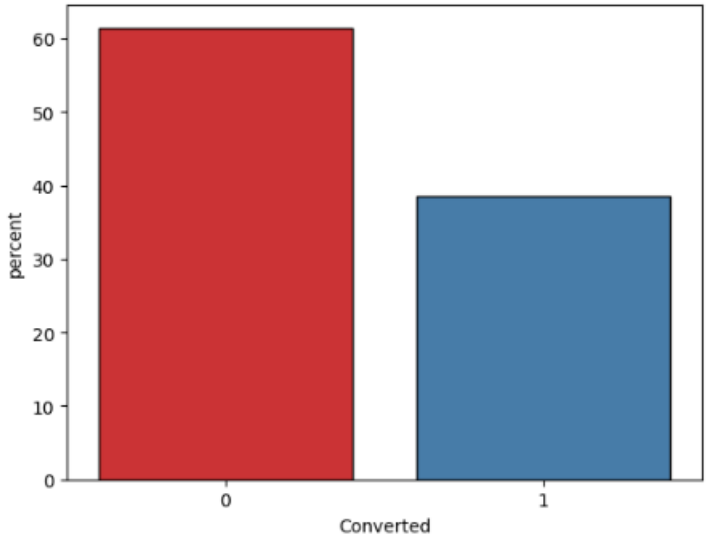
EDA





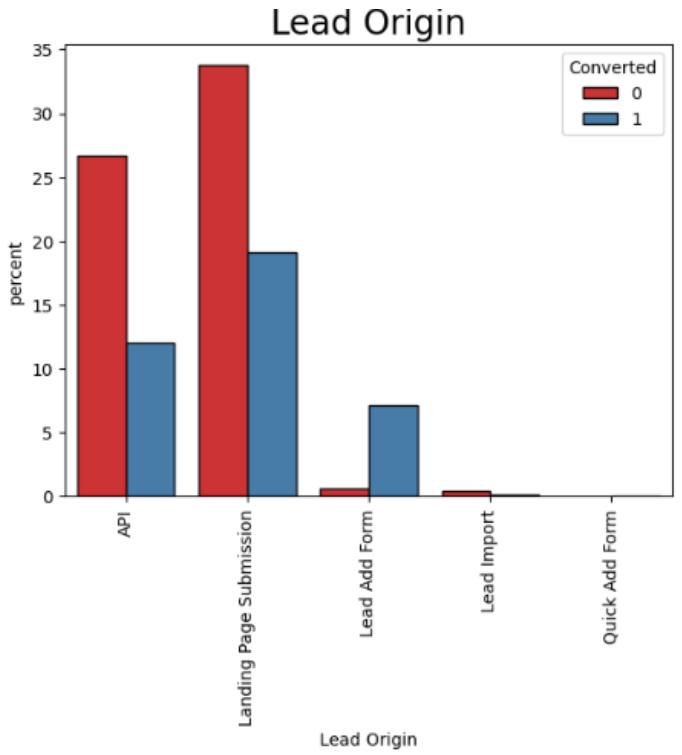
Again a highly skewed data with more one data set being very dominant , we will replace the missing values with frequent one ("India) in this case

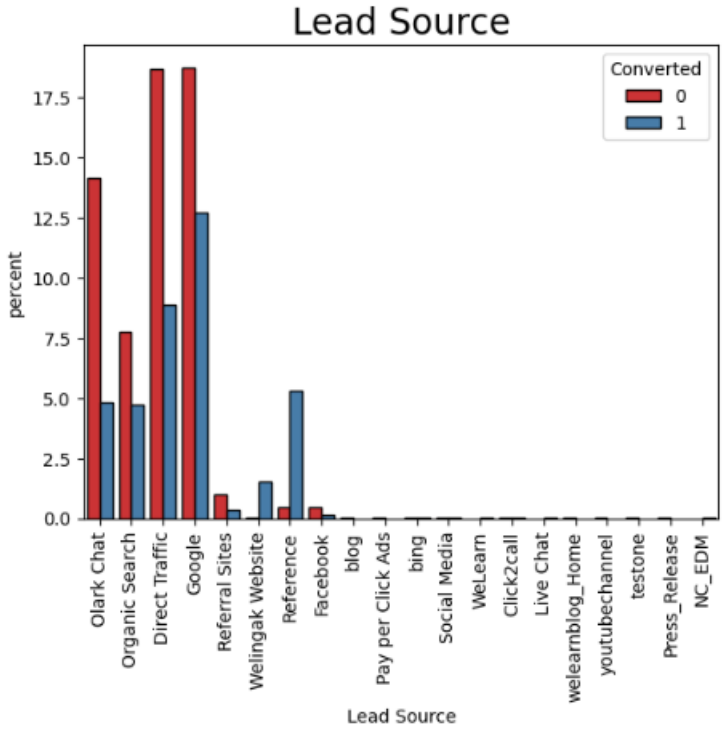
EDA



Note

60% of leads are not converted





Observerations

API and landing page contntibutes to maximum lead generation

Intrestingly Lead Add form has a better conversion rate compared to others lead mechanisms

Data Preparation

Dividing the Data into Training and Testing Sets

The initial step for regression involves splitting the data, which we have done using a 70:30 ratio.

Applying RFE for Feature Selection

We executed Recursive Feature Elimination (RFE) to select 15 output variables.

Model Development

We built the model by excluding variables with a p-value greater than 0.05 and a Variance Inflation Factor (VIF) exceeding 5.

Making Predictions on the Test Dataset

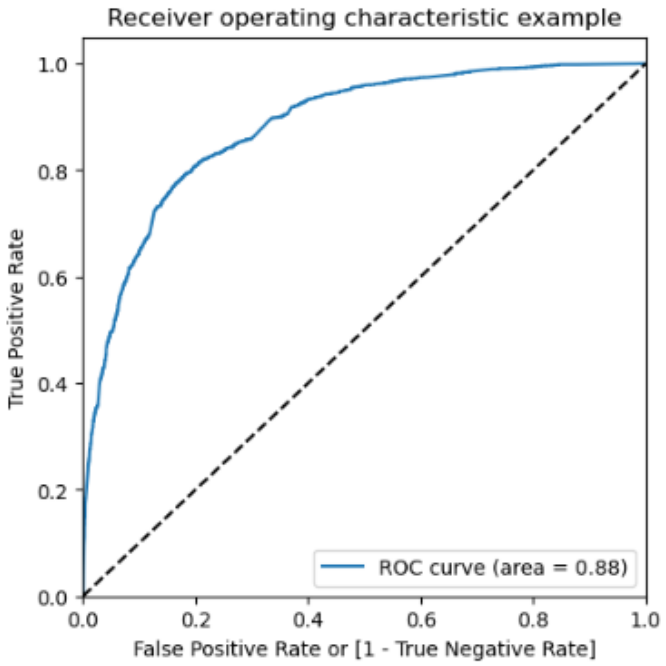
The overall accuracy achieved is 81%.

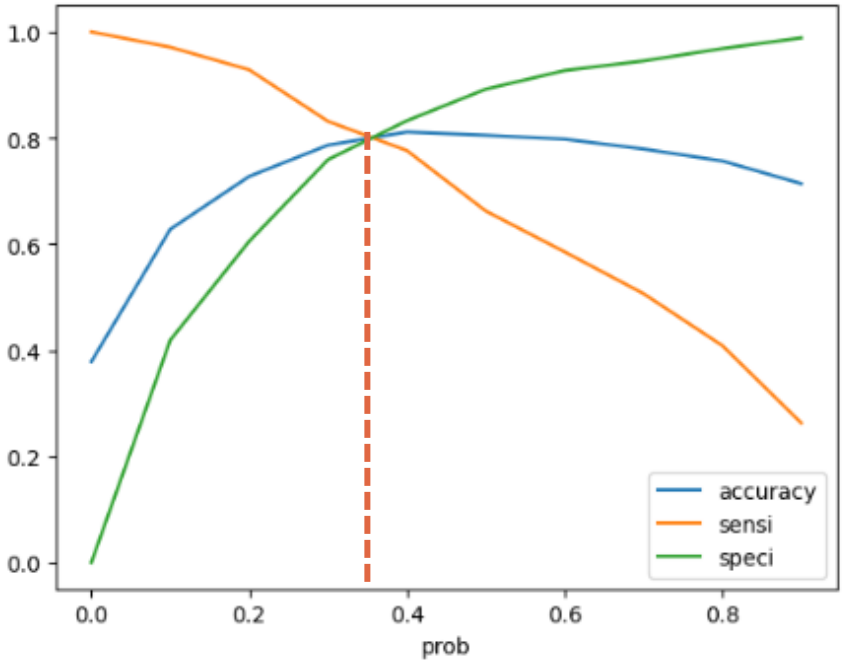
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8545	0.107	-7.995	0.000	-1.064	-0.645
Lead Origin_Lead Add Form	2.6297	0.198	13.274	0.000	2.241	3.018
Lead Source_Direct Traffic	-0.6327	0.083	-7.642	0.000	-0.795	-0.470
Lead Source_Organic Search	-0.2844	0.115	-2.467	0.014	-0.510	-0.058
Lead Source_Welingak Website	1.9058	0.748	2.548	0.011	0.440	3.372
Last Activity_Converted to Lead	-1.1370	0.235	-4.845	0.000	-1.597	-0.677
Last Activity_Email Bounced	-1.6156	0.316	-5.114	0.000	-2.235	-0.996
Last Activity_Email Opened	0.3204	0.113	2.833	0.005	0.099	0.542
Last Activity_Olark Chat Conversation	-1.0595	0.198	-5.362	0.000	-1.447	-0.672
Last Activity_SMS Sent	1.3820	0.116	11.951	0.000	1.155	1.609
What is your current occupation_Other	-1.1768	0.088	-13.345	0.000	-1.350	-1.004
What is your current occupation_Working Professional	2.2971	0.188	12.225	0.000	1.929	2.665
TotalVisits	-0.1509	0.042	-3.611	0.000	-0.233	-0.069
Total Time Spent on Website	0.9475	0.038	24.974	0.000	0.873	1.022

	Features	VIF
0	Lead Origin_Lead Add Form	1.61
1	Lead Source_Direct Traffic	1.52
6	Last Activity_Email Opened	1.51
8	Last Activity_SMS Sent	1.50
9	What is your current occupation_Other	1.50
11	TotalVisits	1.49
2	Lead Source_Organic Search	1.33
7	Last Activity_Olark Chat Conversation	1.27
3	Lead Source_Welingak Website	1.26
12	Total Time Spent on Website	1.23
10	What is your current occupation_Working Profes...	1.18
4	Last Activity_Converted to Lead	1.13
5	Last Activity_Email Bounced	1.12

Finalized Features

Operating Characteristic Curve





Area under the ROC curve looks with 0.88

Optimal cut off is around 0.35

Summary

Key Features which are critical for conversion

- Lead Origin- Lead Add Form
- What is your current occupation -Working Professional
- Lead Source -Welingak Website
- Last Activity- SMS Sent
- Total Time- Spent on Website
- Last Activity -Email Opened

Thank you