

LEAD SCORING CASE STUDY

Submitted By:

Shubham Dhar
Shriya Gandreti
Shruti kumar

PROBLEM STATEMENT

- An education company named X sells online courses to industry professionals. They have a process of online form filling on their website.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Once acquired, employees from the sales team start making calls, writing emails, etc. and some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30% which is poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

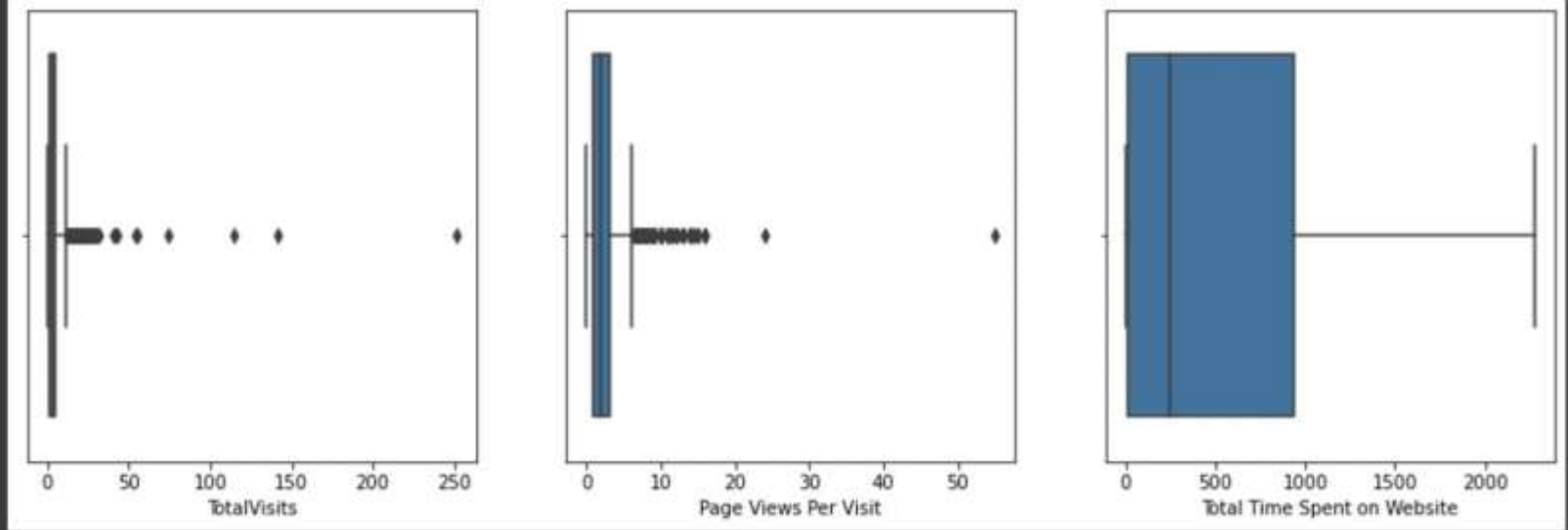
BUSINESS OBJECTIVE

- X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers
- The company requires to build a model wherein a lead score is assigned 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 wants to achieve the target lead conversion rate to be around 80%.

ANALYSIS APPROACH

- Data Collection
 - Importing the data and inspecting the data frame created initially.
 - Gathering the statistics related to dataframe namely shape, description of rows and columns etc.
- Data Cleaning
 - Checking for null value columns in the dataset.
 - Dropping columns with large count of NULL values.
 - Dropping unnecessary columns.
 - Imputation of NaN values in columns like 'Select', 'Last Activity' etc.

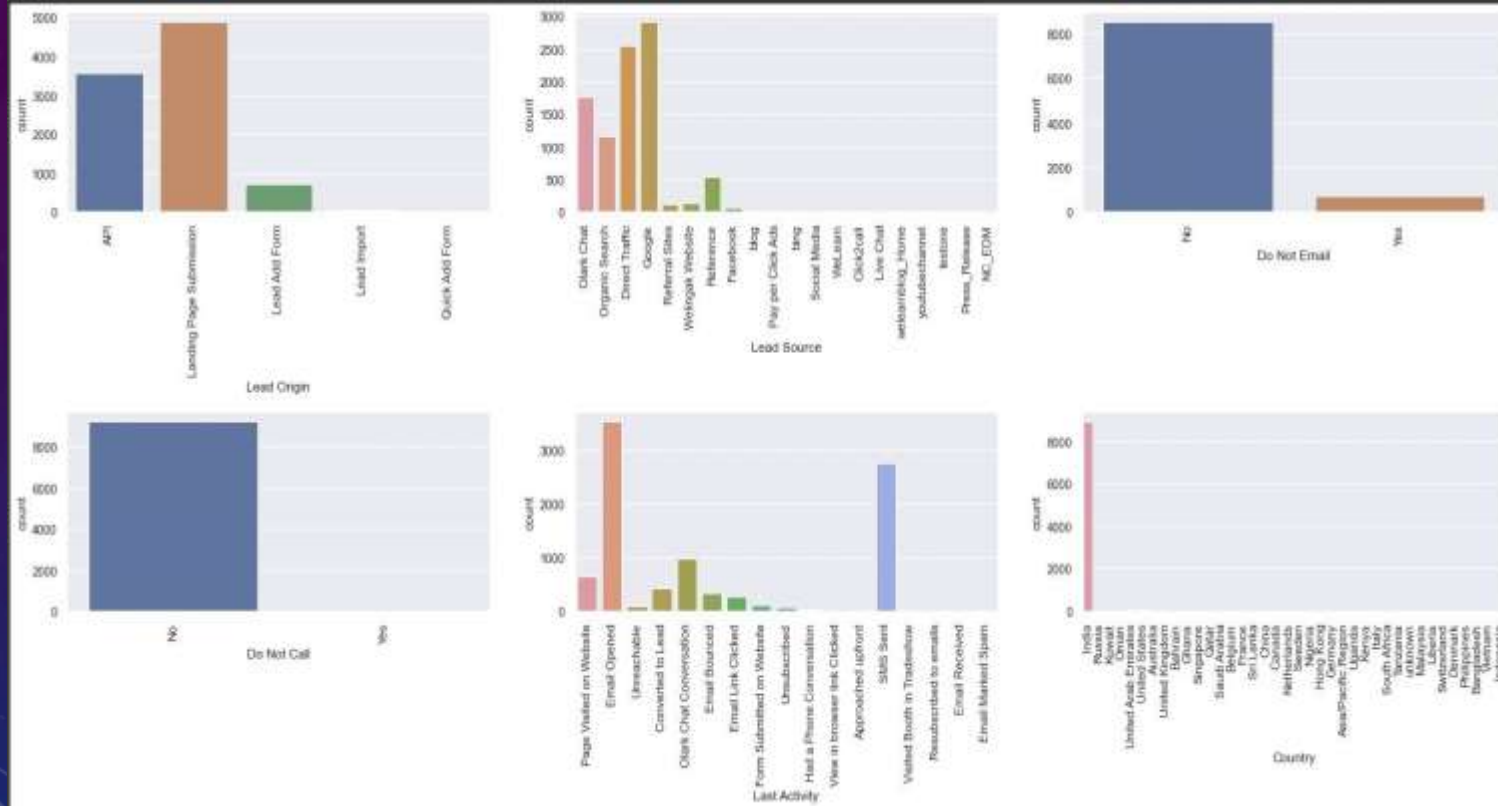
- Exploratory Data Analysis (EDA)
 - Checking for outliers.



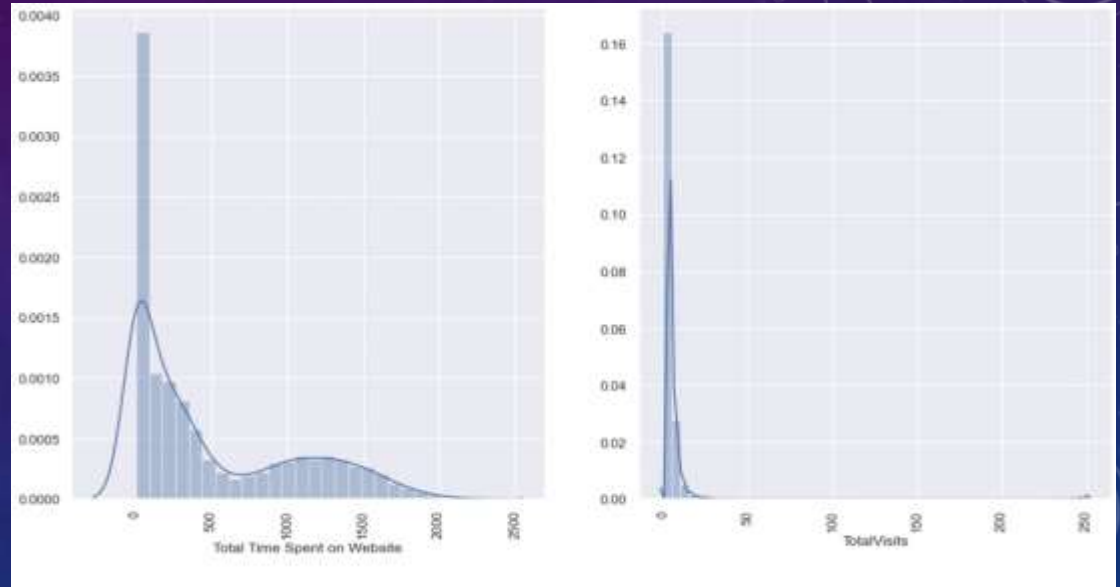
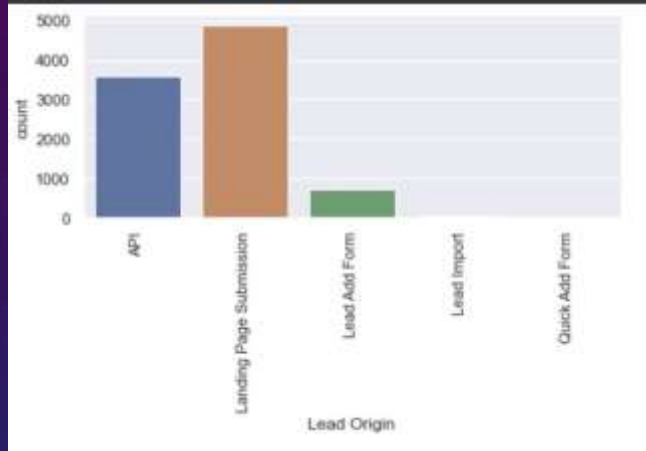
- Correlation analysis.

	TotalVisits																							
Time Spent on Website	0.2																							
Page Views Per Visit	0.48	0.32																						
Lead Origin_API	-0.2	-0.21	-0.35																					
Page Submission	0.28	0.28	0.49	-0.84																				
Origin_Lead Import	-0.025	-0.038	-0.04	-0.056	-0.075																			
Origin_Quick Add Form	-0.001	0.04	-0.0021	-0.0099	-0.013	-0.00088																		
Source_Direct Traffic	0.092	0.12	0.14	-0.44	0.53	-0.044	-0.0077																	
Lead Source_Google	0.096	0.21	0.2	0.018	0.078	-0.043	0.019	-0.41																
Source_Olark Chat	-0.29	-0.38	-0.5	0.62	-0.52	-0.035	-0.0061	-0.3	-0.33															
Lead Source_Other	-0.099	-0.12	-0.19	-0.18	-0.31	0.21	-0.004	-0.2	-0.22	-0.16														
_Converted to Lead	-0.064	-0.0059	-0.058	-0.013	0.049	-0.016	-0.0028	0.067	0.033	-0.11	-0.057													
Activity_Email Bounced	-0.042	-0.025	-0.039	-0.054	0.064	-0.0023	0.063	0.094	-0.059	-0.023	-0.031	-0.044												
Click Chat Conversation	-0.14	-0.2	-0.23	0.37	-0.31	-0.024	-0.0043	-0.18	-0.092	0.44	-0.089	-0.076	-0.068											
Last Activity_Other	-0.0022	-0.027	0.0008	0.0031	0.013	-0.018	-0.0032	-0.014	0.027	0.0032	-0.0017	-0.058	-0.051	-0.09										
Time Visited on Website	0.22	0.016	0.12	-0.068	0.093	-0.019	-0.0034	0.062	0.019	-0.099	-0.03	-0.06	-0.054	-0.094	-0.071									
Last Activity_SMS Sent	0.001	0.12	0.07	-0.12	0.063	-0.012	-0.008	0.013	0.019	-0.12	0.068	-0.14	-0.13	-0.22	-0.17	-0.18								
Business Administration	0.05	0.073	0.064	-0.13	0.14	-0.0046	-0.0027	0.057	0.034	-0.088	-0.016	0.0018	-0.015	-0.062	-0.00098	0.012	0.02							
Finance Management	-0.21	-0.24	-0.34	0.58	-0.6	0.039	-0.012	-0.32	-0.036	0.41	0.063	0.0036	-0.0073	0.29	-0.014	-0.066	-0.085	-0.21						
Resource Management	0.064	0.072	0.085	-0.17	0.15	-0.0067	-0.0039	0.072	0.042	-0.13	0.023	-0.0078	-0.025	-0.1	0.03	0.032	0.013	-0.068	-0.3					
Projects Management	0.033	0.029	0.075	-0.13	0.15	-0.014	-0.0025	0.12	0.0014	-0.09	-0.05	-0.0077	0.072	-0.052	-0.0011	0.019	0.0023	-0.044	-0.19	-0.064				
Marketing Management	0.015	0.069	0.053	-0.14	0.12	-0.022	-0.0039	0.085	0.039	-0.1	0.014	-0.0021	-0.013	-0.072	-0.0031	0.00052	0.023	-0.067	-0.3	-0.098	-0.064			
Specialization_Other	0.16	0.11	0.23	-0.31	0.33	-0.014	0.024	0.16	-0.044	-0.21	-0.059	0.011	0.015	-0.14	-0.013	0.06	0.05	-0.11	-0.49	-0.16	-0.1	-0.16		
	TotalVisits	Time Spent on Website	Page Views Per Visit	Lead Origin_API	Page Submission	Origin_Lead Import	Origin_Quick Add Form	Source_Direct Traffic	Lead Source_Google	Source_Olark Chat	Lead Source_Other	_Converted to Lead	Activity_Email Bounced	Click Chat Conversation	Last Activity_Other	Time Visited on Website	Last Activity_SMS Sent	Business Administration	Finance Management	Resource Management	Projects Management	Marketing Management	Specialization_Other	

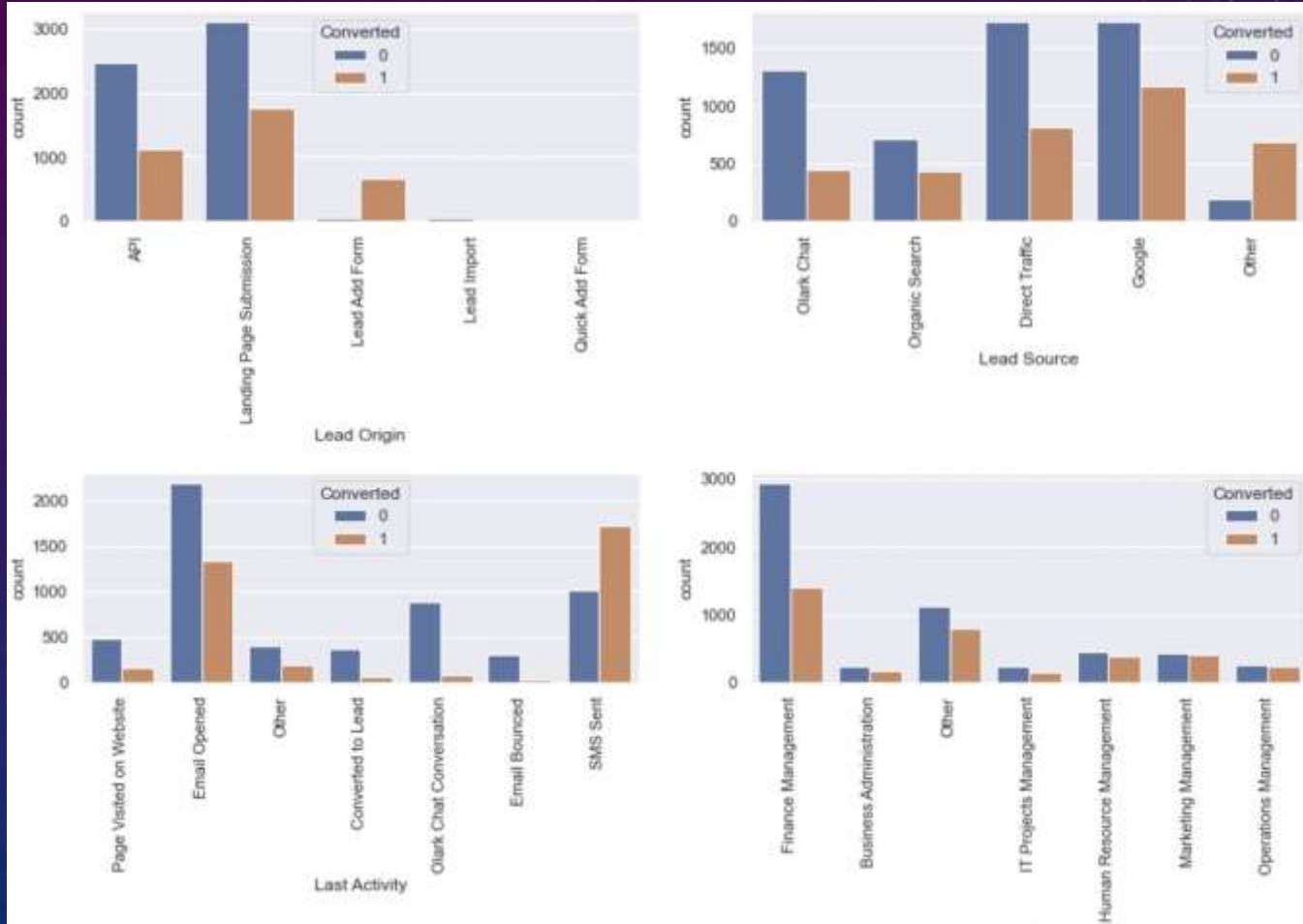
- Countplot to check data distribution.



- Univariate Analysis - Categorical and Continuous.

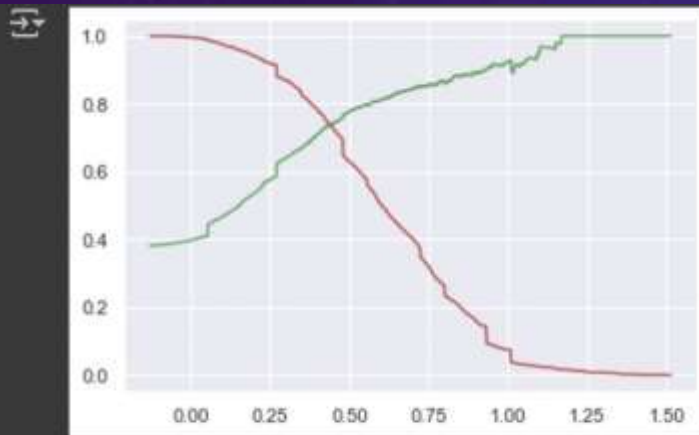


- Bivariate Analysis

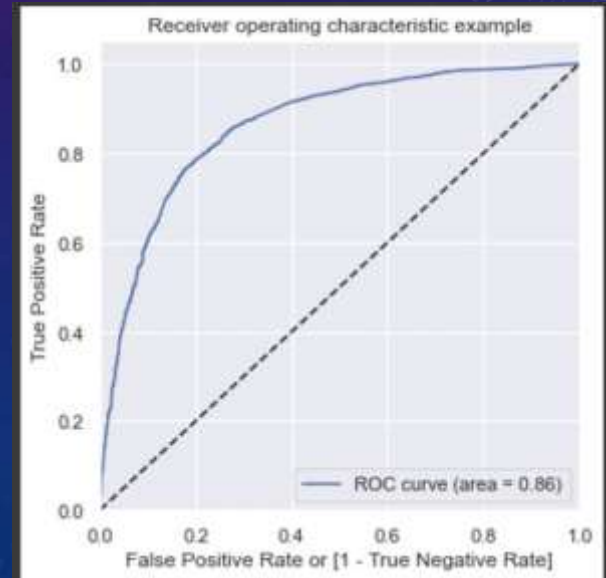


- Dummy Variable Creation
 - Creating dummies for categorical variables.
- Data Preparation and Feature Scaling
 - Splitting dataset into test and training datasets for model evaluation.
 - Scaling and transforming datasets for standardisation.
- Model Building and training
 - Creating a Logistic Regression model.
 - Training the model using training dataset.
 - Model building using different VIF values and p-values by dropping non required columns.

- **Model Evaluation**
 - Making predictions on training dataset.
 - Checking for accuracy, confusion matrix , precision and other parameters.
 - Graphs plotted between various output parameters.
- **Test Dataset Prediction**
 - Making predictions on test dataset.
 - Checking for accuracy, confusion matrix , precision and other parameters.



Here we got 0.37 as the Cut-off as Precesion-Recall Thresholdm



RESULTS

- We can conclude the following that the variables that important the most in the potential buyers are:
 - The total time spent on the Website.
 - Total number of visits.
 - When the lead source was: a. Google b. Direct traffic c. Organic search d.
 - Olark Chat
 - When the last activity was: a. SMS b. Olark chat conversation When the lead origin is Leadadd format.

Training Dataset

Accuracy: 78.57%

Sensitivity: 81.02%

Specificity: 77.06%

Test Dataset

Accuracy: 69.84%

Sensitivity: 27.85%

Specificity: 97.25%