Lead Scoring Case Study

Submitted By

Bhagyashree Sharma

Problem Statement

Introduction:

- 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. Moreover, the company also gets leads through past referrals. 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%.

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Business Goal

- Company wishes to identify the most potential leads, also known as “Hot Leads”
- The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and
- customer with lower lead score have a lower conversion chance
- The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. 80%

Overall Approach

1. Data cleaning and imputing missing values
2. Exploratory data analysis: univariate, bivariate, and multivariate analysis
3. Feature scaling and dummy variable creation
4. Logistic regression model building
5. Model evaluation: specificity, sensitivity, precision, and recall
6. Conclusion and recommendation

Problem Solving Methodology

DATA CLEANING AND PREPARATION

->. Read data from source

->Convert data into clean format suitable for analysis

-> Remove duplicate data Outlier treatment

->Exploratory data analysis

SPLITTING THE DATA AND FEATURE SCALING

->Splitting the data into train and test dataset

-> Feature scaling of numerical variables

MODEL BUILDING

-> Feature selection using RFE, VIF and p-value

-> Determine optimal model using Logistic Regression

-> Calculate various evaluation metrics

RESULT

-> Determine Lead score and check if target final prediction is greater than 80% conversion rate

-> Evaluate final prediction on test set

Data Conversion

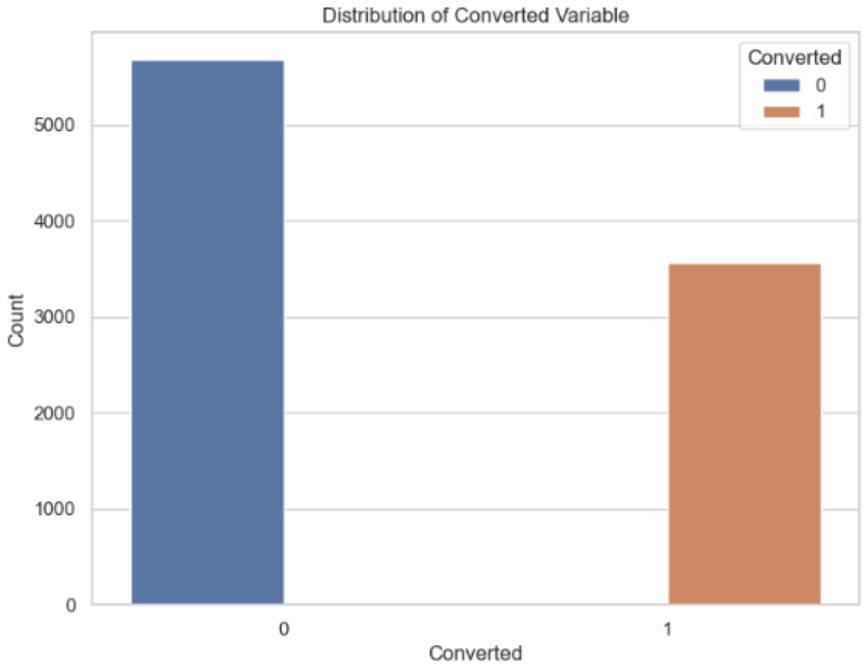
1. Converting the variable with values "Yes/No" to 1/0s.
2. Converting the 'Select' values with NaNs.
3. Dropping the columns having more than 70% of null values.
4. Dropping unnecessary columns.
5. Dropping the rows as the null values were less than 2%.

Data Manipulation

- Total number of rows: 37 and total number of columns: 9240.
- Single-value features like "Magazine", "Receive More Updates About Our Courses", "Update my supply", "Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque", etc., have been dropped.
- Removed the "ProspectID" and "Lead Number" which are not necessary for the analysis.
- After checking the value counts for some object type variables, it was found that some features have enough variance and have been dropped. These features include "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper, Article", "XEducation Forums", "Newspaper", "Digital Advertisement", etc.
- Dropped columns having more than 35% missing values, such as "How did you hear about X Education" and "Lead Profile".

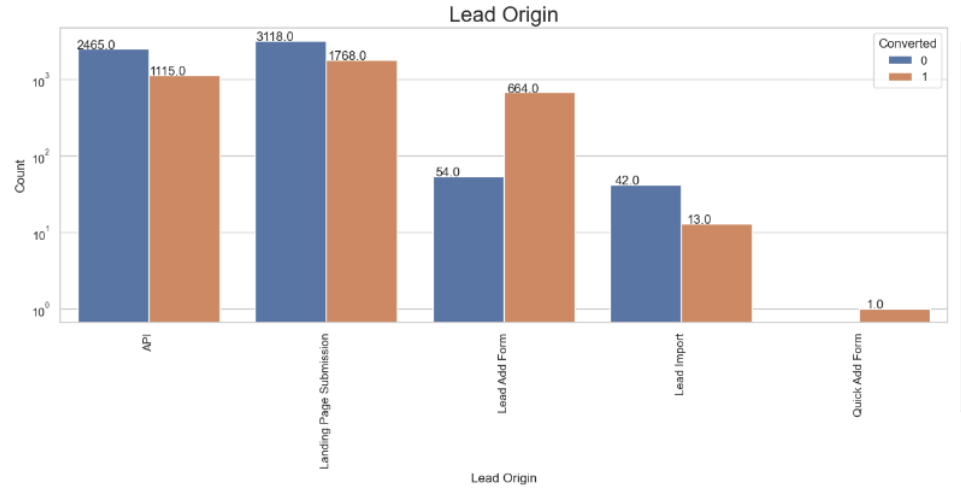
Exploratry Data Analysis

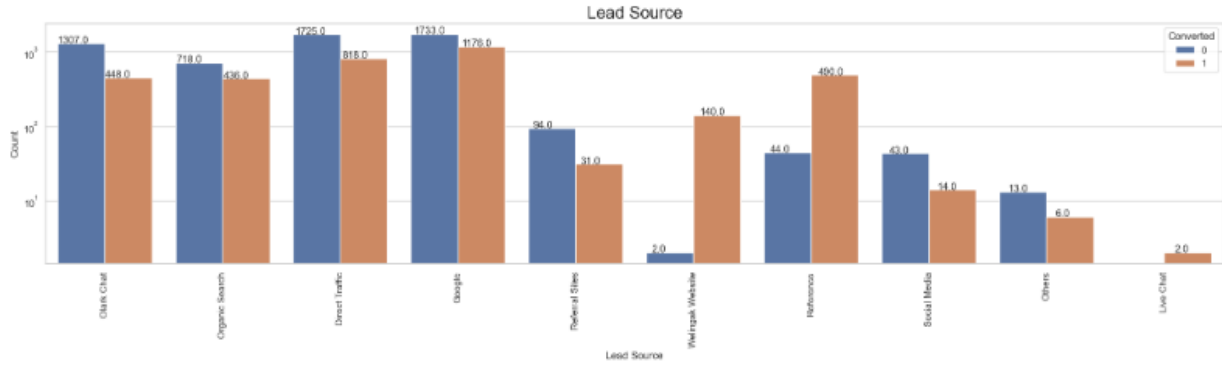
Univariate Analysis :



From above graph observation: We have around 30% of Conversion Rate

Categorical Variables Analysis

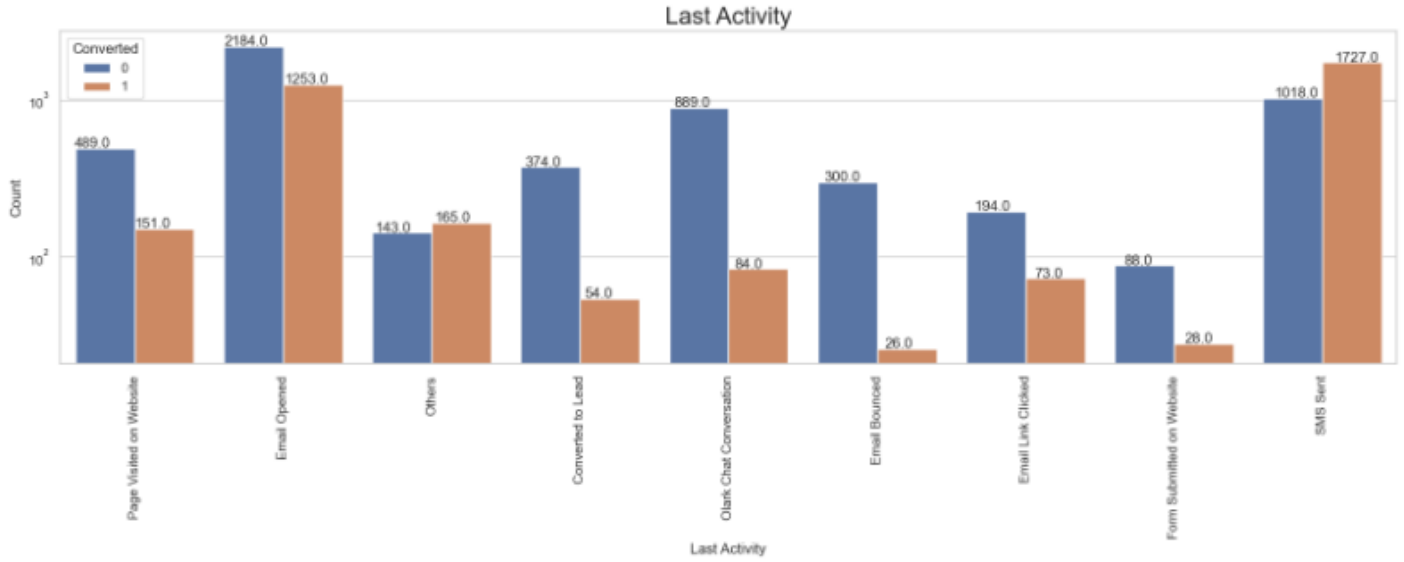




- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and Welingak Website is maximum
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are
- considerable
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

- Do Not Email: A higher number of leads are generated from people who opted for the email option

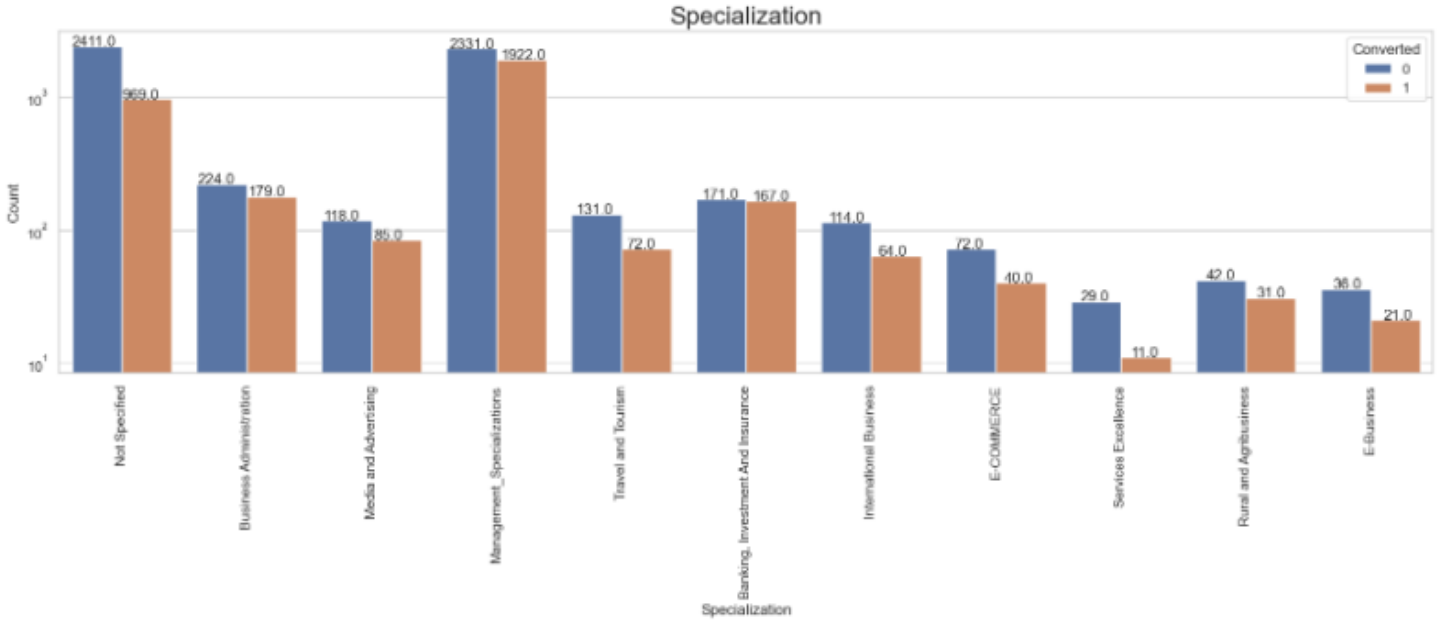
- Lead Activity:
- 1. The conversion rate for leads with the last activity of 'SMS Sent' is approximately 63%.
- 2. The most common last activity for leads is 'Email Opened'.



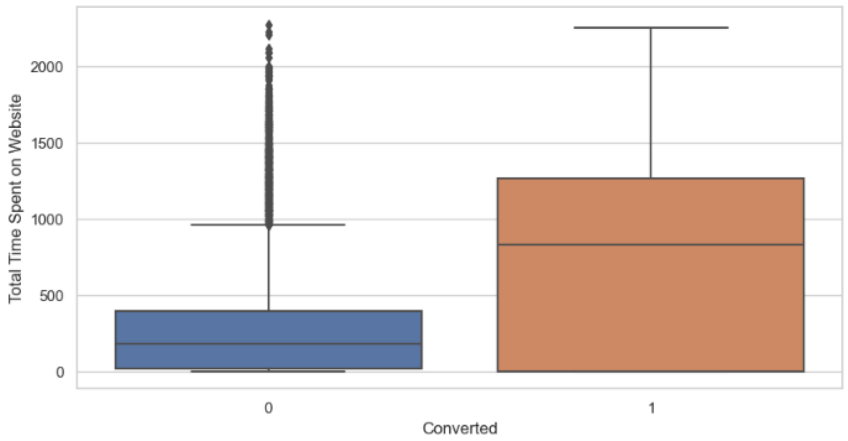
Current Occupation: 'Unemployed' leads are generating a higher number of leads and have approximately a 45% conversion rate.
The conversion rate is higher for 'Working Professionals'.

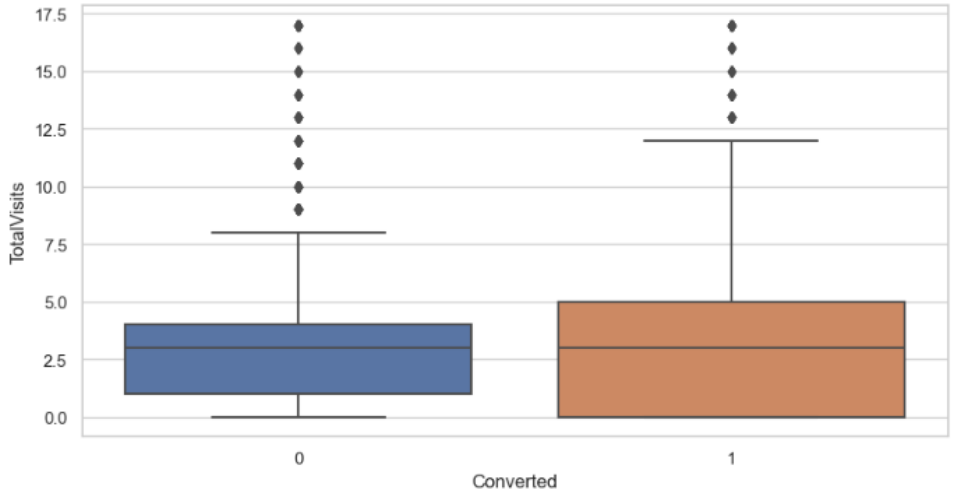
Specilization:
From the count plot of 'Specialization', we observe that 'Management' specialization has the highest number of leads generated.

The 'Other' category also contributes significantly to the number of leads generated.



Numerical Variable Analysis





Observation:

Total Time Spent on Website:

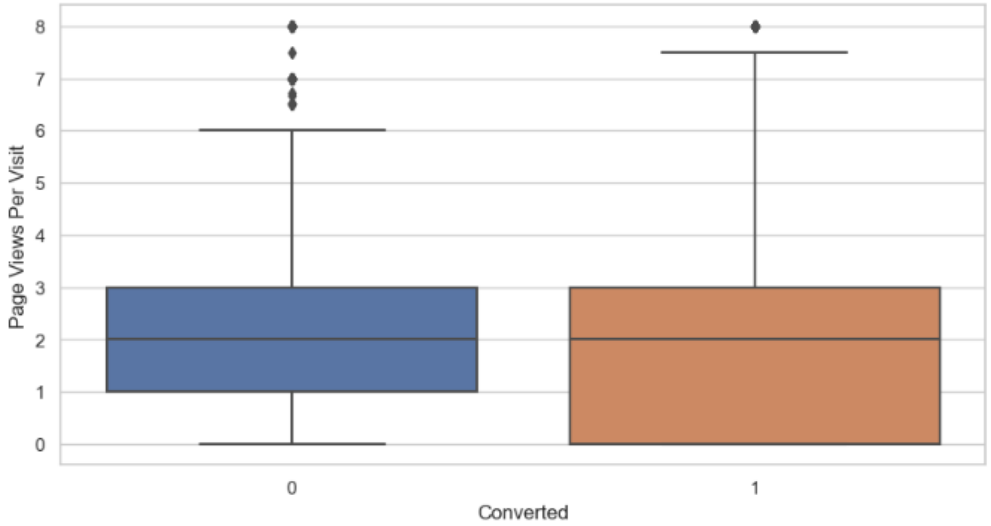
The above box plot shows Leads spending more time on website are more likely to opt for curses or converted.

TotalVisits:

we can see that median for converted and non-converted is approx same.

Page Views Per Visit:

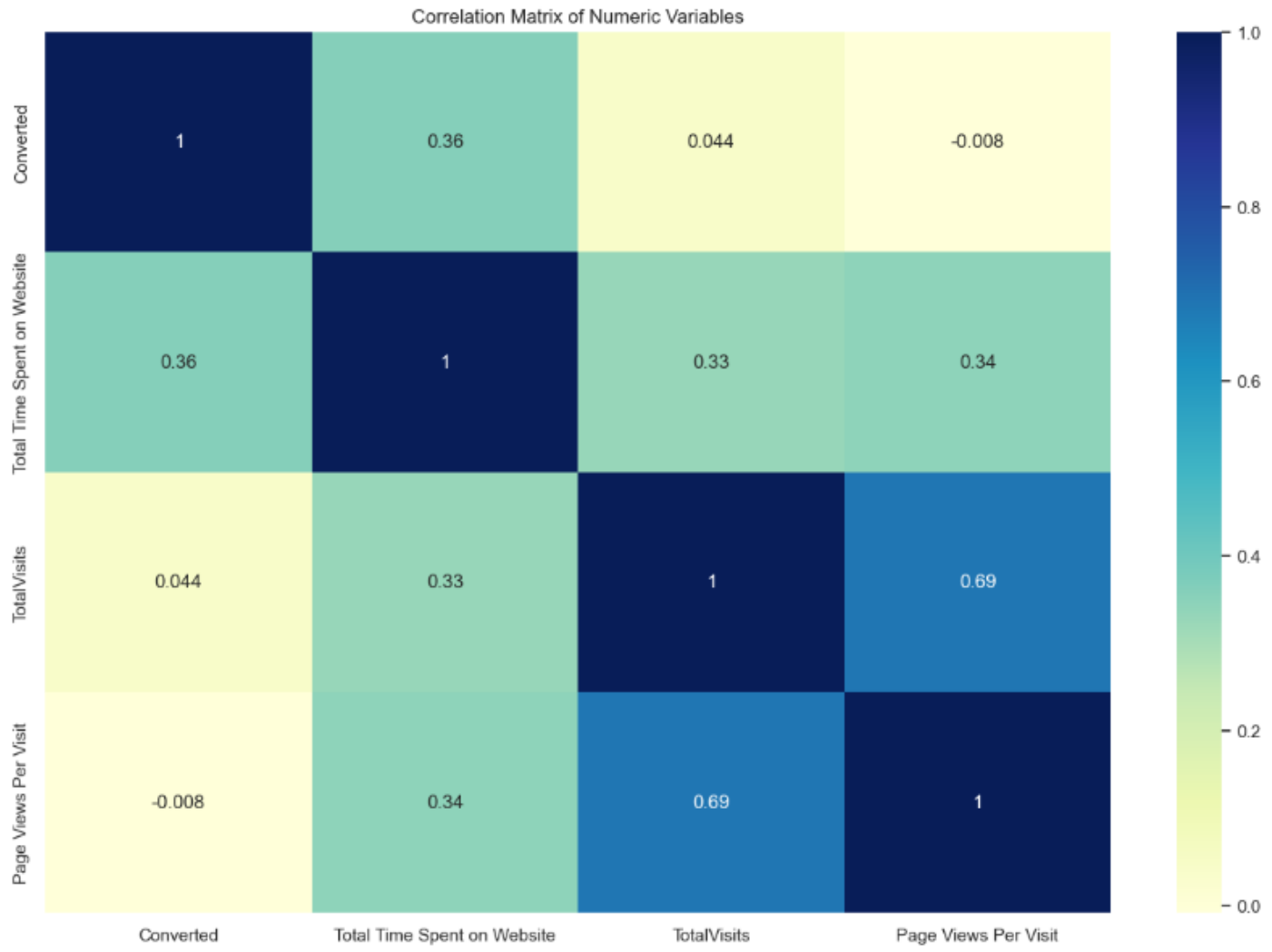
we can see that median for converted and non-converted is approx same.



Bivariate Analysis

Observation
--> 'TotalVisits' and 'Page Views per Visit' are highly correlated with correlation of .69

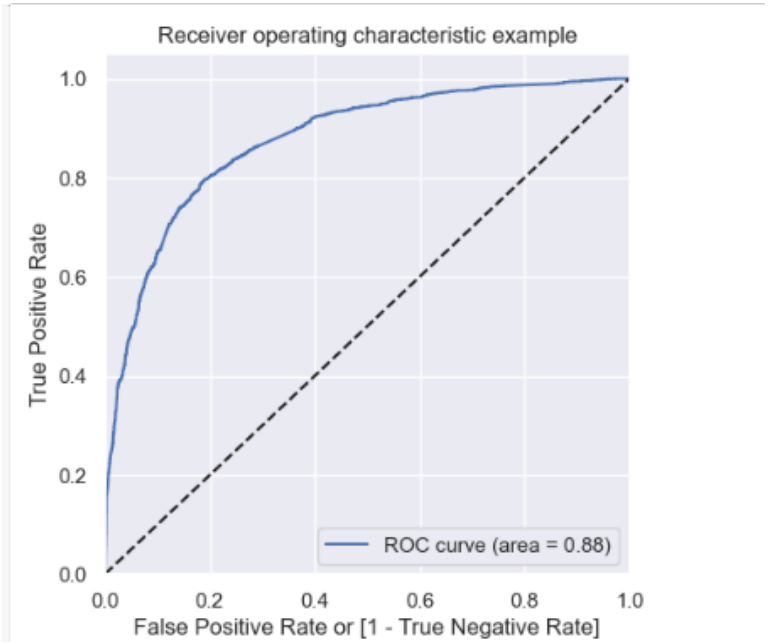
-->'Total Time Spent on Website' has correlation of 0.36 with target variable 'Converted'.

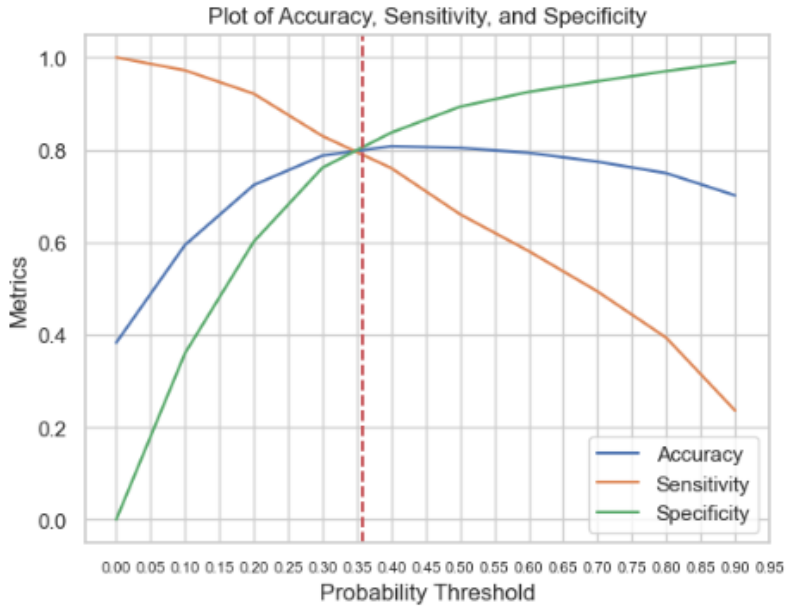


Model Building

- Split the data into test and training sets with a ratio of 70:30.
- Employed Recursive Feature Elimination (RFE) to select the top 15 variables.
- Built the model by removing variables with a p-value greater than 0.05 and VIF greater than 5.
- Made predictions on the test dataset.
- Achieved an overall accuracy of 80.0%.

ROC Curve & Optimal Cut Off Point





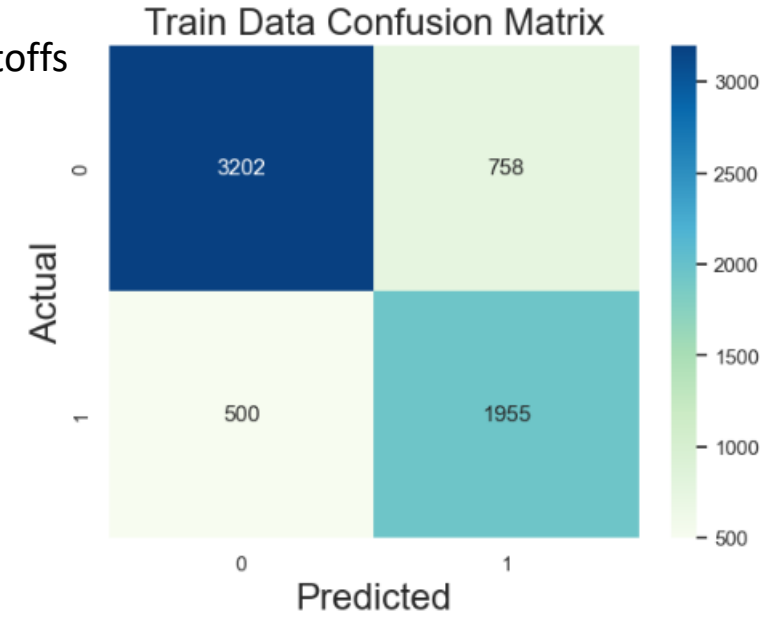
Observations:

- Finding Optimal Cut off Point
- 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 0.358.

Model Evaluation

Calculation of accuracy sensitivity and specificity for various probability cutoffs

	Probability	Accuracy	Sensitivity	Specificity
0.0	0.0	0.382697	1.000000	0.000000
0.1	0.1	0.594076	0.971894	0.359848
0.2	0.2	0.723772	0.921385	0.601263
0.3	0.3	0.787529	0.829735	0.761364
0.4	0.4	0.807794	0.760489	0.837121
0.5	0.5	0.804209	0.660692	0.893182
0.6	0.6	0.793453	0.580855	0.925253
0.7	0.7	0.774279	0.493686	0.948232
0.8	0.8	0.749493	0.393483	0.970202
0.9	0.9	0.701481	0.236253	0.989899



Overall Accuracy	80.30%
Sensitivity	79.63%
Specificity	80.85%
Precision	72.06%
Recall	79.63%

Lead Score for the Data Frame:

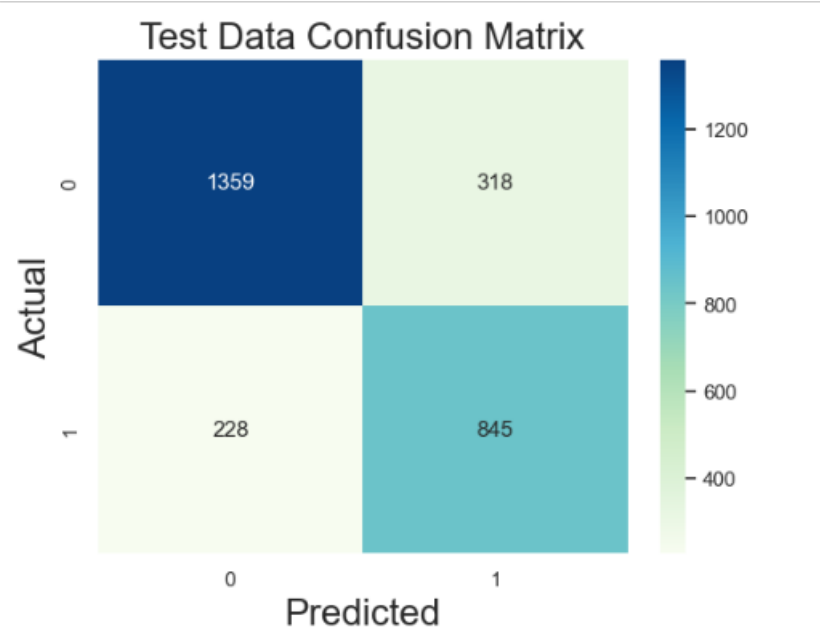
	Prospect ID	Converted	Converted_Prob	Lead_Score	final_Predicted
0	1124	0	0.303593	30	0
1	4778	0	0.985691	99	1
2	1012	0	0.169345	17	0
3	3103	1	0.298306	30	0
4	5094	0	0.238621	24	0

Model Prediction

Top Features of Model

```
----------------------Top Features ----------------------

Lead Source_Olark Chat                                  23.09
Lead Source_Reference                                   60.86
Lead Source_Welingak Website                           100.00
Last Activity_Email Opened                              11.21
Last Activity_Olark Chat Conversation                  -15.72
Last Activity_Others                                    20.70
Last Activity_SMS Sent                                  31.01
What is your current occupation_Student                  6.85
What is your current occupation_Working Professional    44.37
Specialization_Not Specified                           -11.04
Do Not Email                                           -23.83
Total Time Spent on Website                             18.23
A free copy of Mastering The Interview                  -8.51
dtype: float64
```



Overall Accuracy	80.14%
Sensitivity	78.75%
Specificity	81.03%
Precision	72.65%
Recall	78.75%

Recommendations & Results

- To improve the potential lead conversion rate X-Education will have to mainly focus important features responsible for good conversion rate are :-

- Lead Source_Welingak Website : As conversion rate is higher for those leads who got to know about course from 'Welingak Website',so company can focus on this website to get more number of potential leads.

- Lead Origin_Lead Add Form: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it to get more number of leads cause have a higher chances of getting converted.

- What is your current occupation_Working Professional : The lead whose occupation is 'Working Professional' having higher lead conversion rate ,company should focus on working professionals nad try to get more number of leads.

- Last Activity_SMS Sent: Lead whose last activity is sms sent can be potential lead for company.

- Total Time Spent on website: Leads spending more time on website can be our potential lead.

Conclusion

- The logistic regression model is used to predict the probability of conversion of a customer.
- While we have calculated both sensitivity-specificity as well as Precision-Recall metrics, we have considered optimal cut off on the basis of sensitivity-specificity for final prediction.
- Lead Score calculated shows the conversion rate of final predicted model is around 80% in test data as compared to 80% in train data
- In Business terms, this model has capability to adjust with the company's requirements in coming future
- TOP variables that contributes for lead getting converted in the model are:
- Lead Source_Welingak Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Hence Overall this model seems to be good