Lead Score Case Study

Business Problem

- An education company named X Education sells online courses to industry professionals.
- Although 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.
- The objective is to build a model to identify the hot leads and achieve lead conversion rate to 80%

Business Objective

"The Business Objective Is To Build A Logistic Regression Model To Identify The Hot/Potential Leads And Achieve The Lead Conversion Rate To 80%."

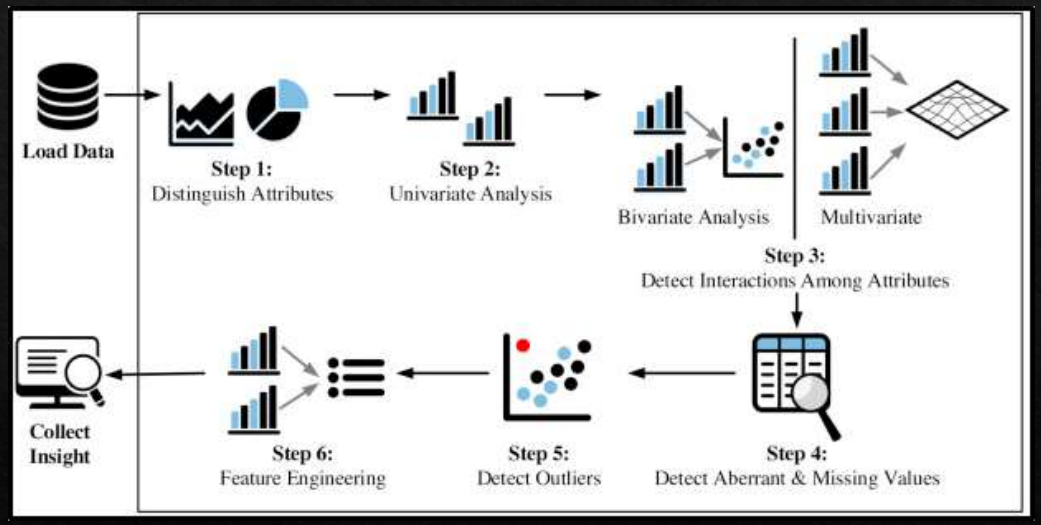
Understanding Dataset

- We got a file named “Leads.csv” provided with a leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- To learn more about the dataset we got the data dictionary.
- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 meansit was converted and 0 meansit wasn’t converted.
- Another thing that to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

Data Knowledge

- Dataset used : “Leads.csv “
- Total number of customers present : 9240
- Total number of features : 37
- Model used : Logistic Regression
- After initial analysis, we see that there are multiple factors that influence conversion rate
- The target column in our dataset : “Converted” .
- We need to reduce the features to maximize the conversion rate.

Technical Approach for Business Problem



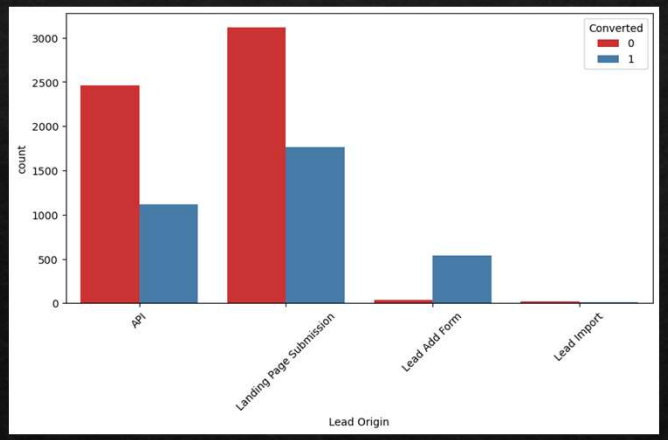
Data Cleaning

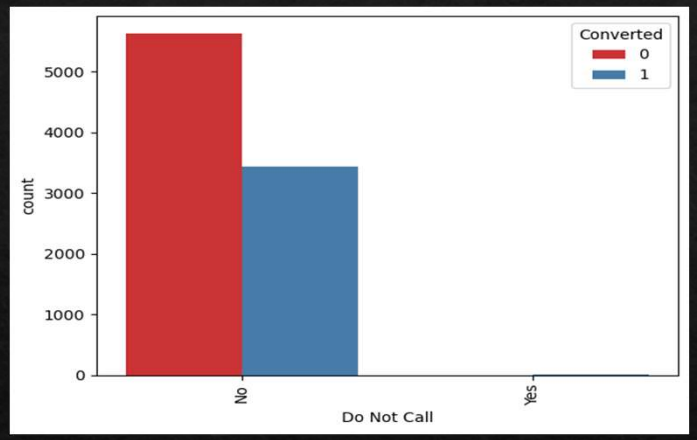
- Handling ' Select' variable:

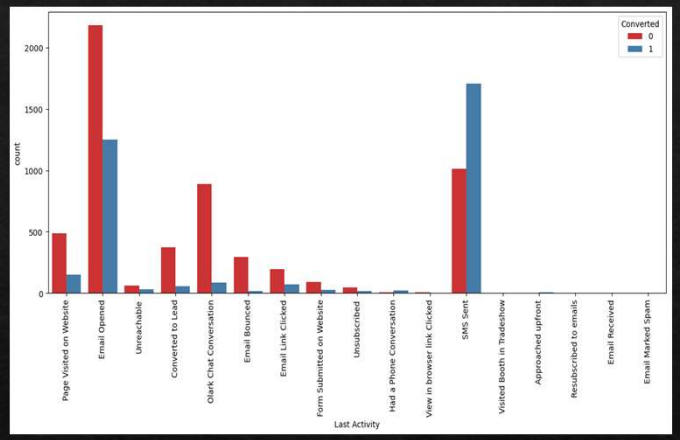
  "Select" variable indicates that the user has not selected any option

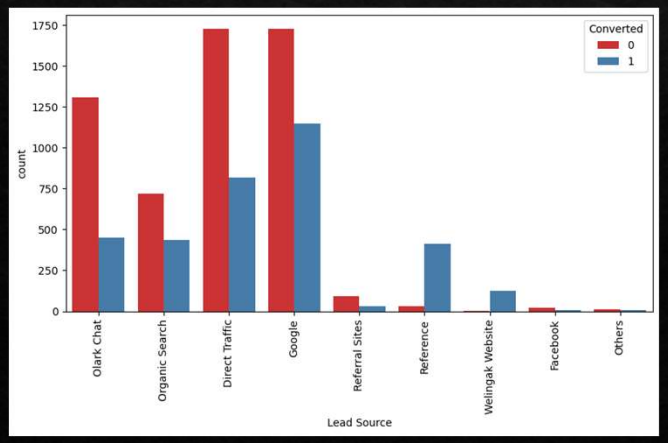
  We impute the same with null values.
- Dropping Score and Activity variables
- Dropping column with high null values.

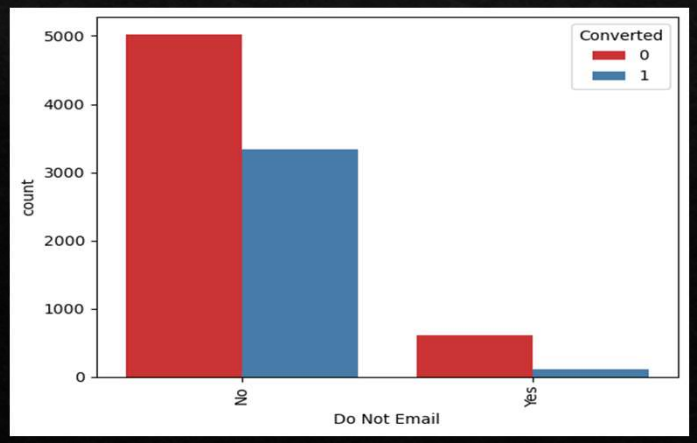
EDA

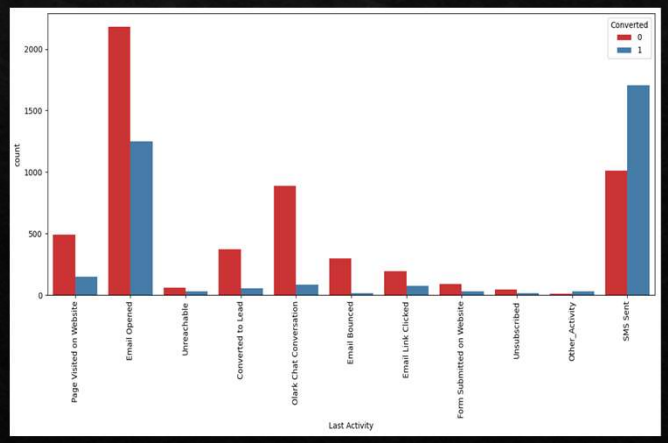

Converted
0
1
count
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


Converted
0
1
count
No
Yes
Do Not Call

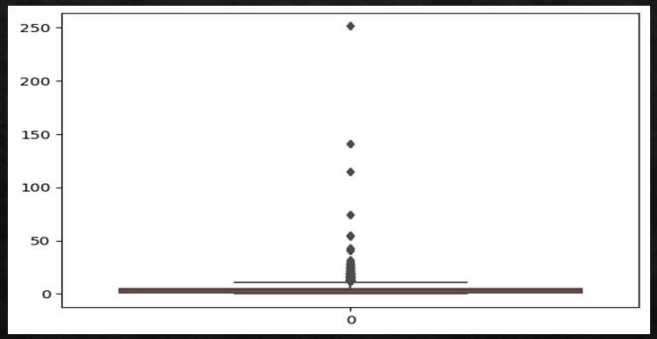

Converted
0
1
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

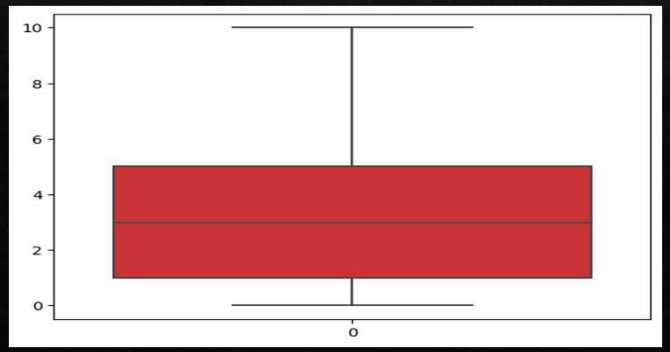

Converted
0
1
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source


Converted
0
1
count
No
Yes
Do Not Email

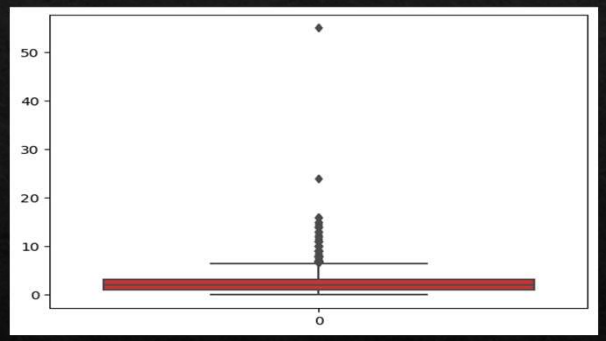

Converted
0
1
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Other_Activity
SMS Sent
Last Activity

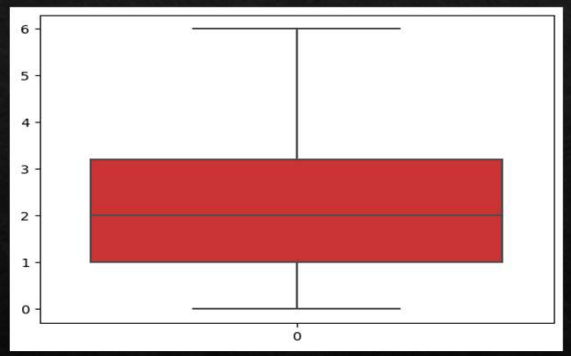
Outlier Treatment





- As we can see there are a number of outliers in the data. We will cap the outliers to 95% value for analysis for total visits





Data Preprocessing

- We split the data set into training and test data set
- Division of data into test data and train data to check the stability of the model.
- We have randomly sampled 70% of the data as the test data and 30% of the data as test data.
- Division of Train Data into X and Y where X has all the features and Y has the target variable – Converted.
- We perform scaling to normalize the data within a particular range

Model Bulding

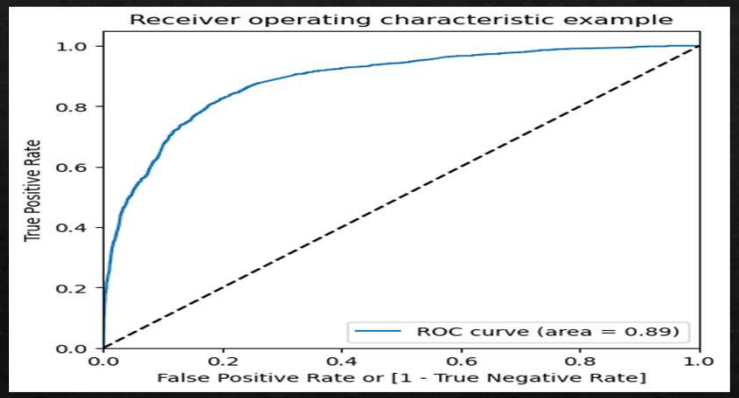
- We have used Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain. RFE uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.
- In this step we made the model stable by using stats library, where we checked the p-values to be less than 0.05 and vif values to be under 5. Variance inflation factor(vif) is used to treat the multicollinearity.
- Once the stable model was created, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0.

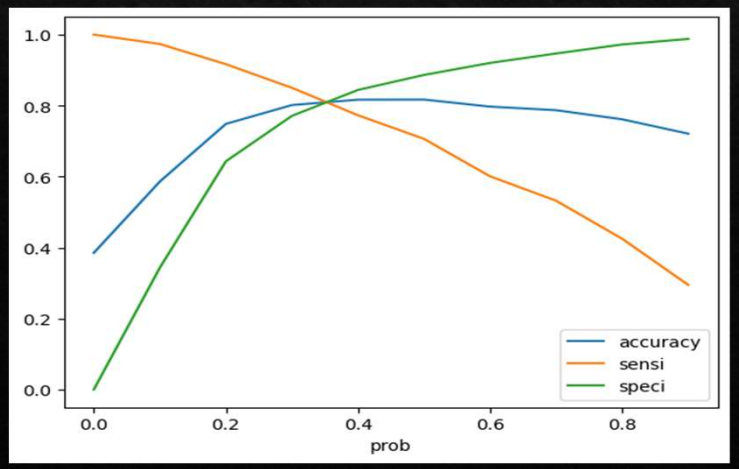
Model Building

- Once the stable model was created, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0.
- We calculated the confusion matrix on this predicted column to the actual converted column. We also calculated the metrics sensitivity, specificity, precision, recall and accuracy. We also plotted roc curve to find the area under the curve.
- After the final model the confusion matrix is

```
[[3461 444]
[ 719 1727]]
```

ROC Curve And Optical Cut-Off Probability





- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- From the curve above, 0.34 is the optimum point to take it as a cutoff probability.

Model Test Performance

- Training Data Set

  Accuracy: 81.0%

  Sensitivity: 81.7%

  Specificity: 80.6%

- Test Data Set

  Accuracy: 80.4%

  Sensitivity: 80.4%

  Specificity: 80.5%

Recommendations

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" 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 who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls 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 last activity was "Olark Chat Conversation" as they are not likely to get converted

Recommendations

- 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.

Thank You