Lead score case Study

Submitted by-

1. Sukanya Bakshi
2. Atharva Pathak
3. Apeksha Khare

Lead Score Case Study for X Education

- **Problem Statement :**

X Education sells online courses to industry professionals. 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%.

- **Business Goal:**

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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%.

Strategy :

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

UpGrad

Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into
- Clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

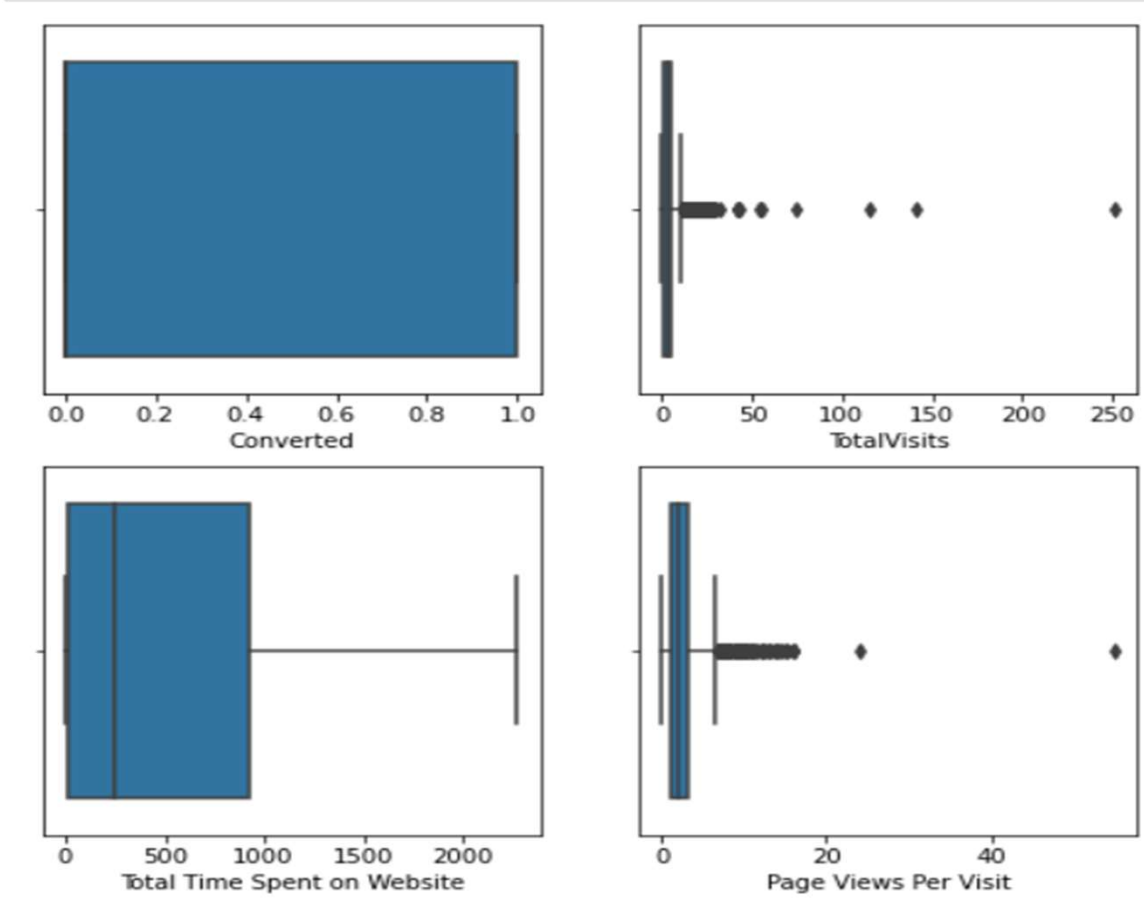
Model Building

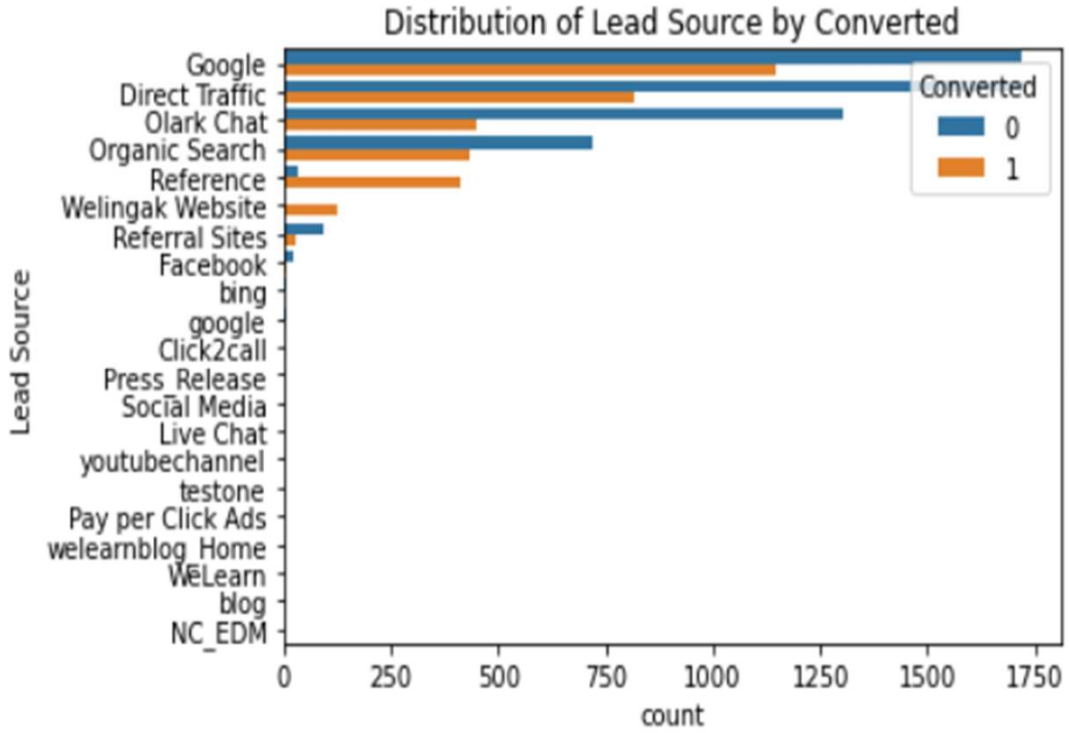
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

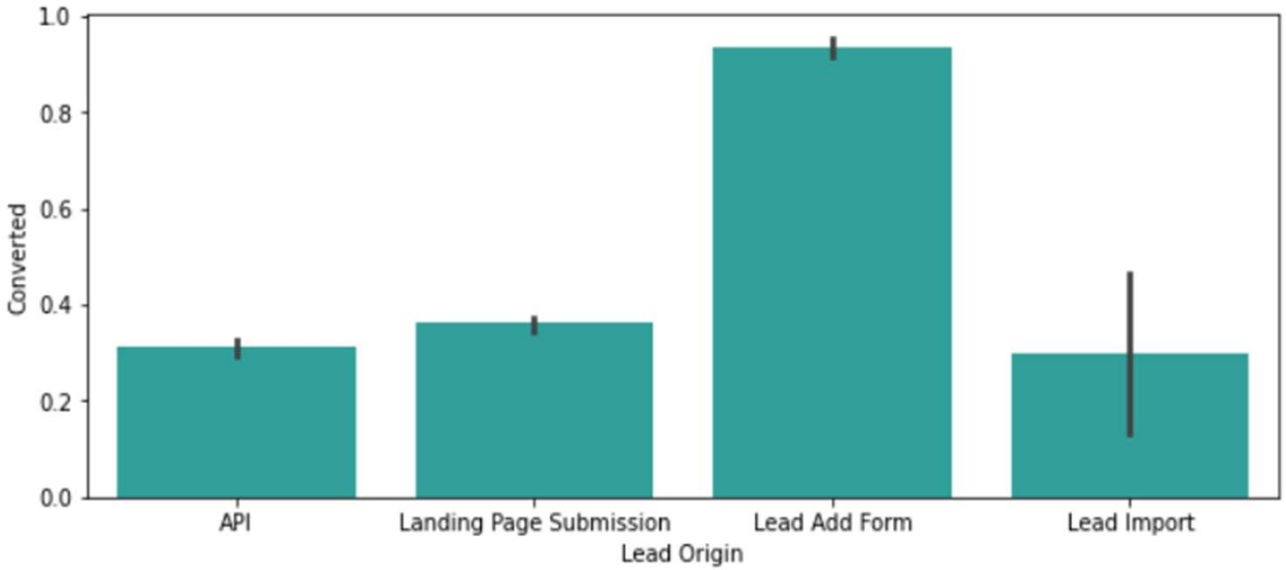
- Determine the lead score and check if target final predictions amounts to 91% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Exploratory Data Analysis :

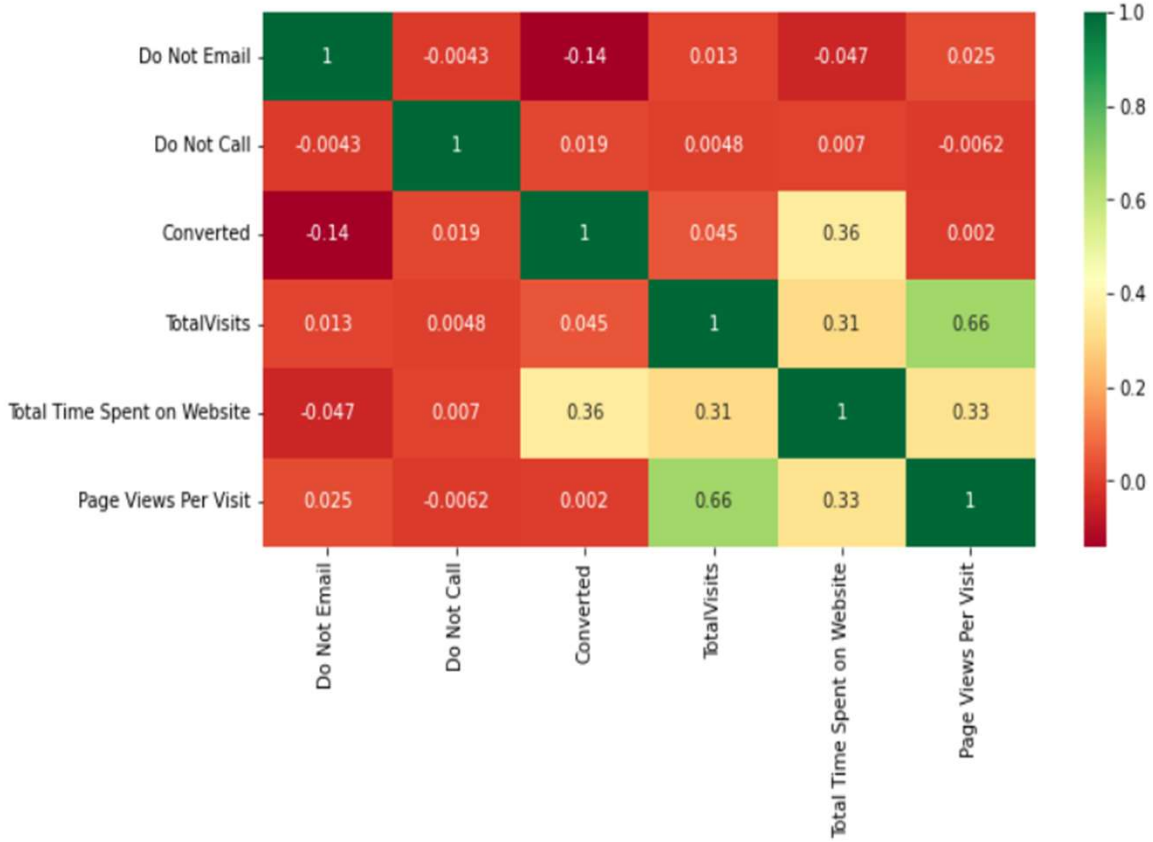



Distribution of Lead Source by Converted
Google
Direct Traffic
Olark Chat
Organic Search
Reference
Welingak Website
Referral Sites
Facebook
bing
google
Click2call
Press_Release
Social Media
Live Chat
youtubechannel
testone
Pay per Click Ads
welearnblog_Home
WeLearn
blog
NC_EDM
Lead Source
Converted
0
1
0
250
500
750
1000
1250
1500
1750
count

The conversion rate is high for Lead add Form.



Correlation Matrix:



Variables Impacting the Conversion Rate

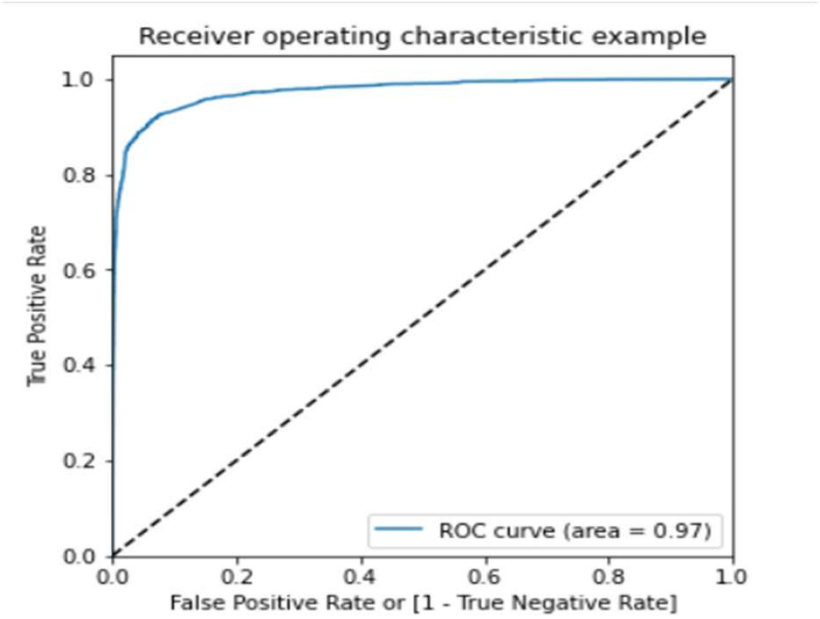
- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  - Google
  - Direct traffic
  - Organic search
  - Welingak website
- When the last activity was:
  - SMS
  - Olark chat conversation
- When the lead origin is Lead add format

An ROC curve demonstrates several things:

>> It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

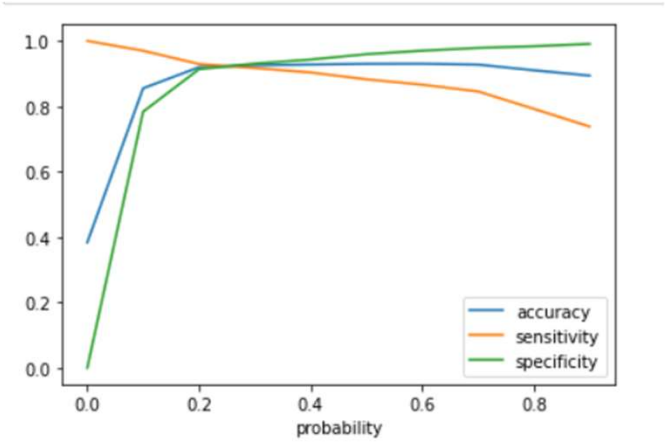
>>The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

>> The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.3 based on Accuracy, Sensitivity and Specificity



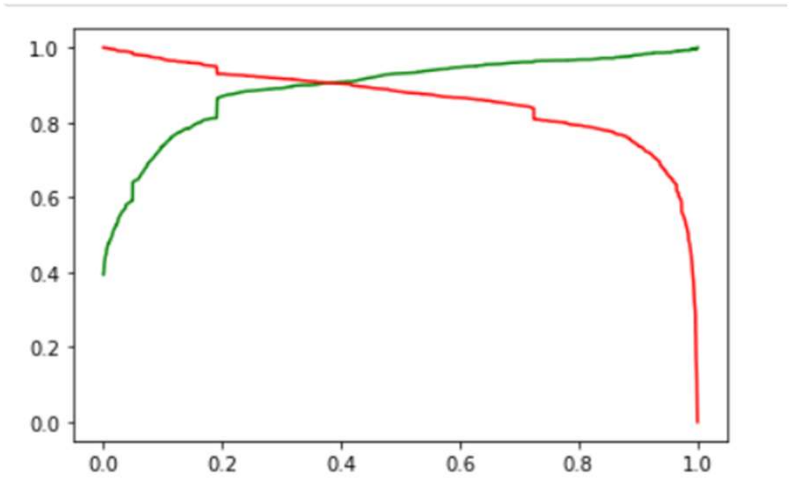
Confusion Matrix

3753	159
286	2148

- Accuracy - 92%
- Sensitivity - 88 %
- Specificity -95%

Model Evaluation- Precision and Recall on Test Dataset :

The graph depicts an optimal cut off of 0.42 based on Precision and Confusion Matrix Recall



Confusion Matrix

1606	116
102	897

- Precision - 88 %
- Recall - 89%

Conclusion:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 92%, 88% and 95% which are approximately closer to the respective values calculated using trained set.
- The top 3 variables that contribute for lead getting converted in the model are –
- Total time on website
- TotalVisits
- Lead Source
- Hence overall this model seems to be good.