Lead Score Case Study

Team Members:

1. Anjali
2. Diptej
3. Amit

Problem Statement

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.

The company requires to build a model wherein you need to assign a lead score 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.

This will help them convert the most promising leads as their paid customers.

Analysis Approach

Steps in overall building analysis & model building:

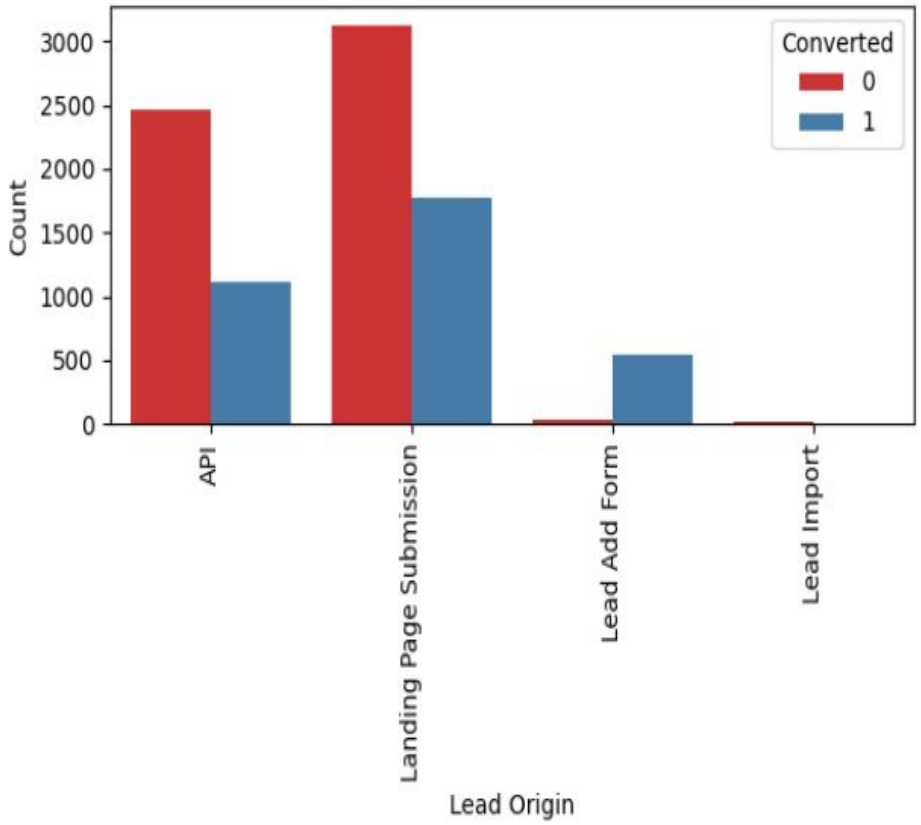
1. Importing & understanding the data statistics.
2. Data cleaning
3. Exploratory Data Analysis:
    a. To understand the relationship between variables
    b. To find outliers.
4. Data preparation:
    a. Binary conversation.
    b. Dummy variable creation
    c. Feature Scaling
    d. Train-Test Split
5. Feature selection by RFE
    a. Model Building
    b. Fine tuning the features
6. Model Evaluation on different metrics.
7. Testing on the test set data
8. Assigning the Lead Score to each lead.
9. Comparing the various metrics for train & test set data.

Important Highlights of EDA

Lead Origin w.r.t Converted

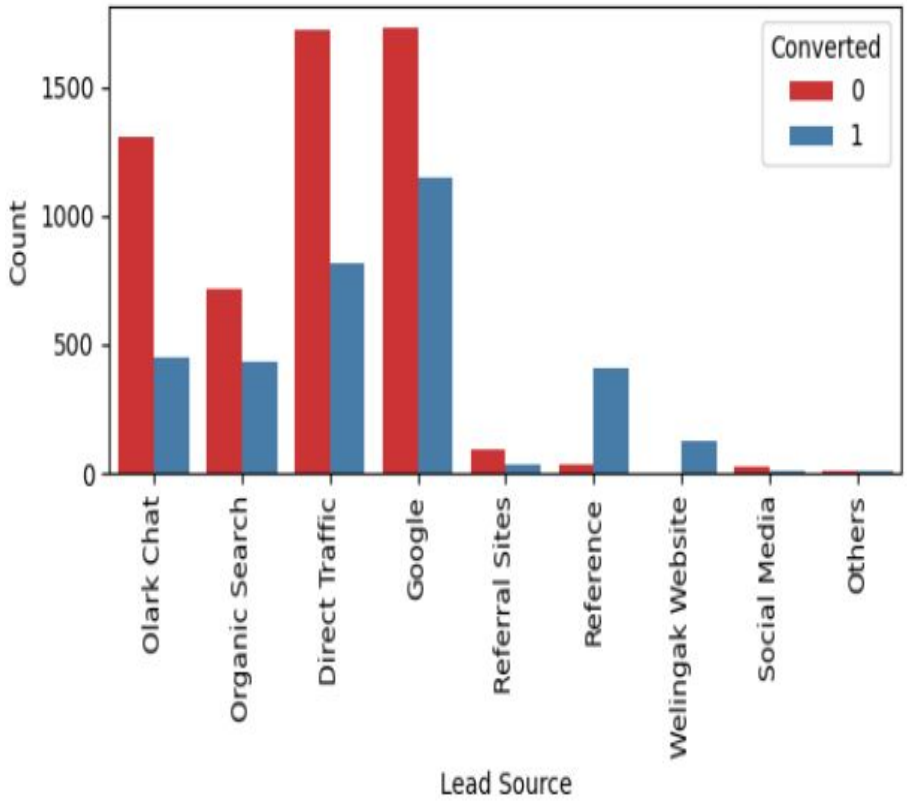
- Lead originating from ‘Landing Page Submission’ has approximately 50% chances of conversion.

- Leads originating from ‘Lead Add Form’ has a very good conversion rate.



Lead Source w.r.t Converted

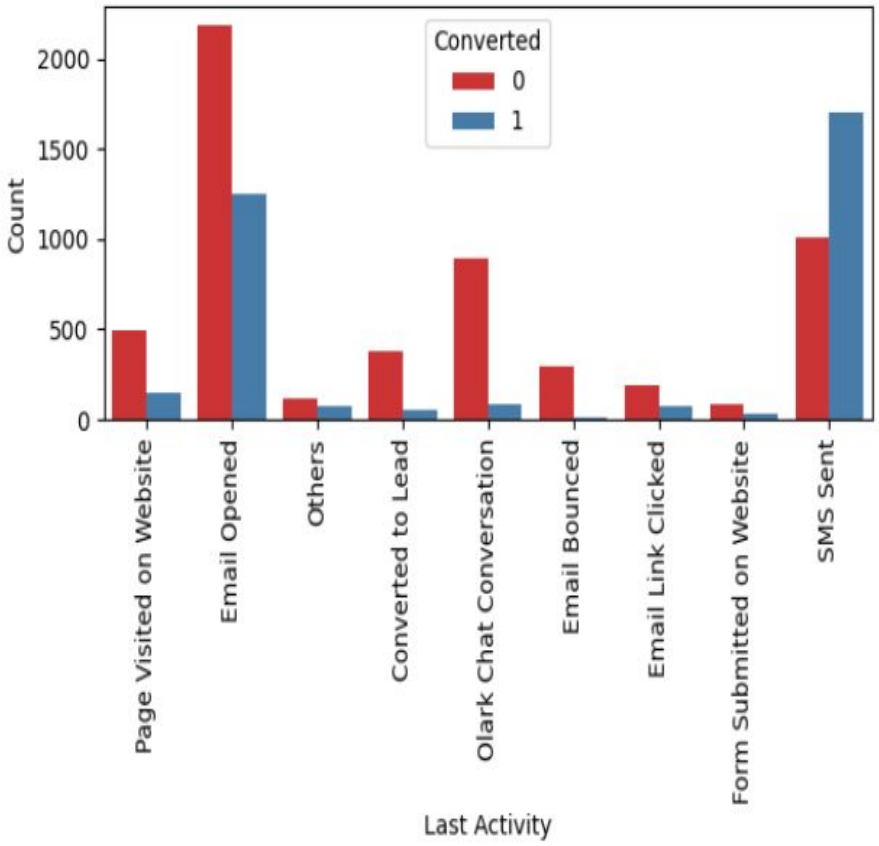
- Google as lead source has more than 50% chances of conversion.
- People coming to X Education through references most certainly gets converted into a customer.



Last Activity w.r.t Converted

- People with ‘Olark Chat Conversation’, ‘Email Bounced’ has a very low chances of conversion.

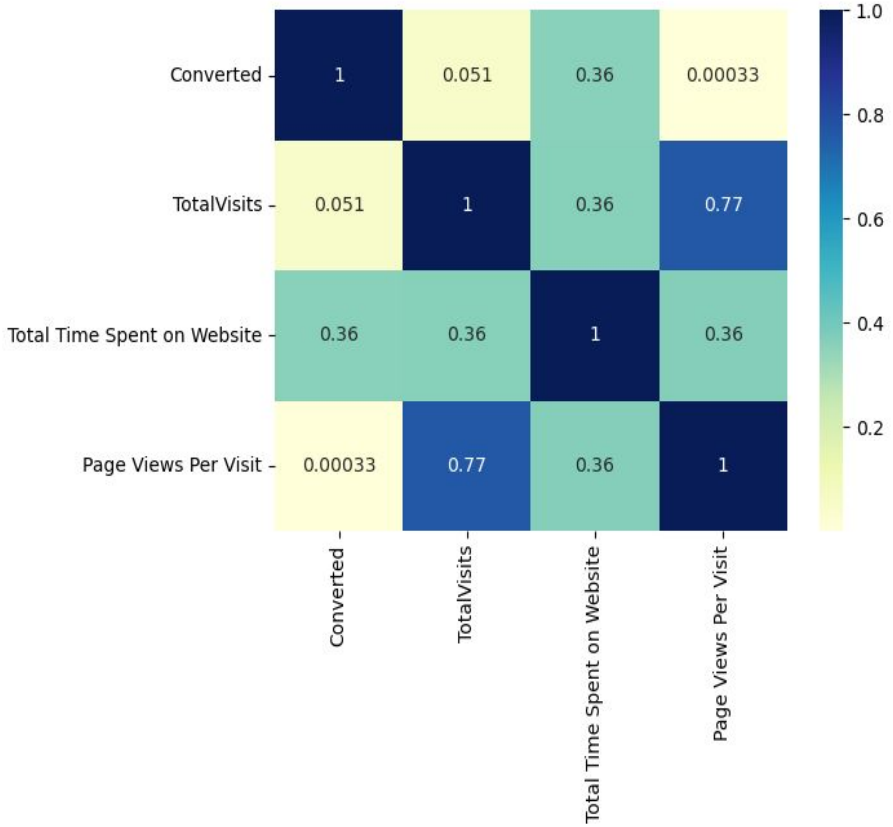
- People with last activity as ‘SMS Sent’ and ‘Email Opened’ has a good rate of conversion overall.



Correlation Matrix

- Conversion has a moderate correlation with ‘Total Time Spent on Website’ by a customer with a positive value of 0.36.

- Conversion rate has a very minimal correlation with number of visits of a customer on website or with number of page views on every visit.



Model Building

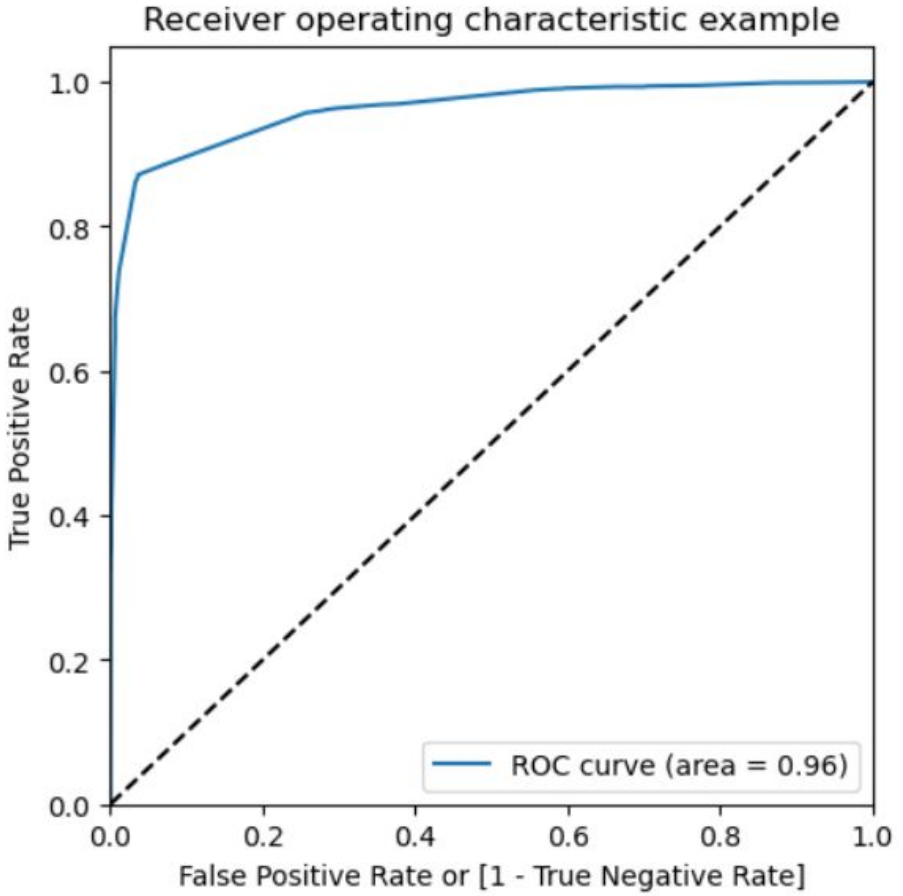
Steps followed in the Logistic Regression Model:

1. Splitting of data in Train & Test sets.
2. Scaling of numerical variable with StandardScaler()
3. Feature selection using RFE
4. Building first model with StatsModel
5. Based on p-values:
    a. Dropping less relevant columns.
    b. Checking VIF values to check strength of correlation between independent variables.
6. Finding overall model accuracy and other metrics.
7. Finding an optimal cut-off.
8. Model Evaluation
9. Predictions on the test set data.

Important Model Highlights

ROC Curve

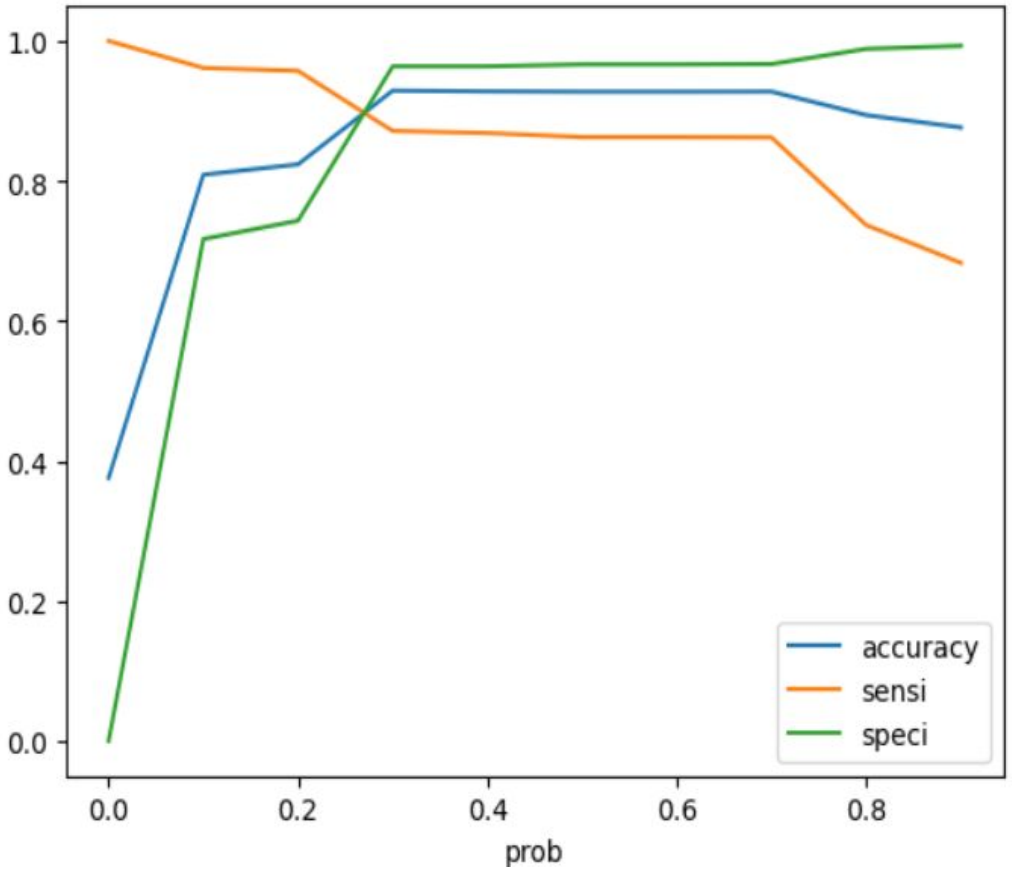
The ROC Curve should be a value close to 1. We are getting a good value of 0.96 indicating a good predictive model.



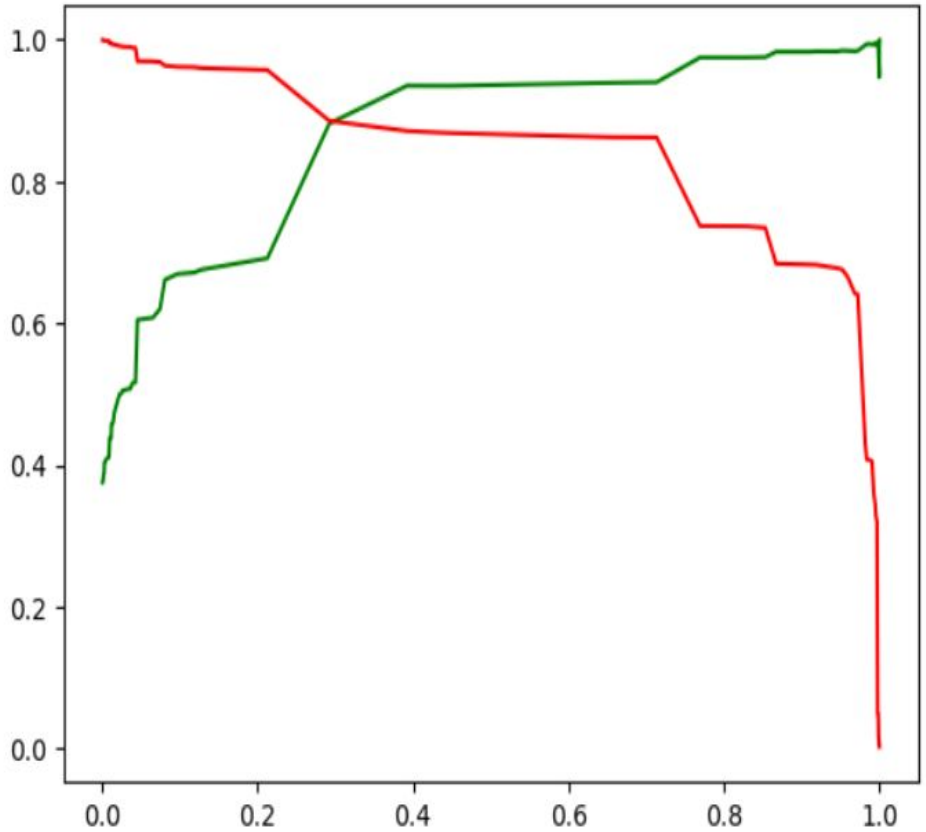
Optimal Cut-off

We plot accuracy, sensitivity and specificity at various probabilities to find the optimal cut-off.

In this case, our cut-off value is 0.3.



Precision-Recall Trade-off Curve



Final Metrics Values

Train Data Set

- Accuracy : 92.8 %
- Sensitivity : 87.1 %
- Specificity : 96.3 %
- Precision: 93.9 %
- Recall: 86.2 %

Test Data Set

- Accuracy : 91.4 %
- Sensitivity : 84.9 %
- Specificity : 95.5 %
- Precision: 92.3 %
- Recall: 84.9 %