Lead Scoring Case Study

SUBMITTED BY:

JITENDRA AHUJA

MADHAVA SHYAM NIRAGHATAM

SURBHIT SRIVASTAVA

Problem Statement

An education company named X Education sells online courses to industry professionals

The company markets its courses on several websites and search engines like Google.

The company also gets leads through past referrals, but is struggling with poor conversion rate and the typical lead conversion rate at X education is around 30%.

To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads' and that will help sales team to focus on communicating with the potential leads rather than calling every lead.

Expected Outcomes:

1. Identify the most potential leads

2. Target lead conversion rate of around 80%, based on potential leads or hot leads

Methodology

Reading and understanding data

Data cleaning and preparation

EDA

Dummy variables and encoding the data

Splitting data in train and test sets

Feature scaling

Model building – Logistic regression

Model Evaluation

Conclusion

Reading and understanding data

The dataset contains 9240 rows and 37 columns

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	What is your current occupation	What matters most to you in choosing a course	Searc
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unemployed	Better Career Prospects	N
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployed	Better Career Prospects	N
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select	Student	Better Career Prospects	N
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	India	Media and Advertising	Word Of Mouth	Unemployed	Better Career Prospects	N

Data Cleaning and Preparation

Checked for columns with null values

Dropped columns with more than 3000 null values (considering the columns had no critical business implications on the lead conversion)

Dropped columns not relevant for regression analysis like City, Country, Prospect ID, and Lead number

Dropped columns ‘How did you hear about X Education‘ and 'Lead Profile’ as they had more than 50% values as “Select”

Dropped columns with most values as “No” like ‘Newspaper’, ‘Digital Advertisement’ etc.

Dropped rows with null values

After all the cleanup, we were able to retain 69% of original data and the same was used for model building

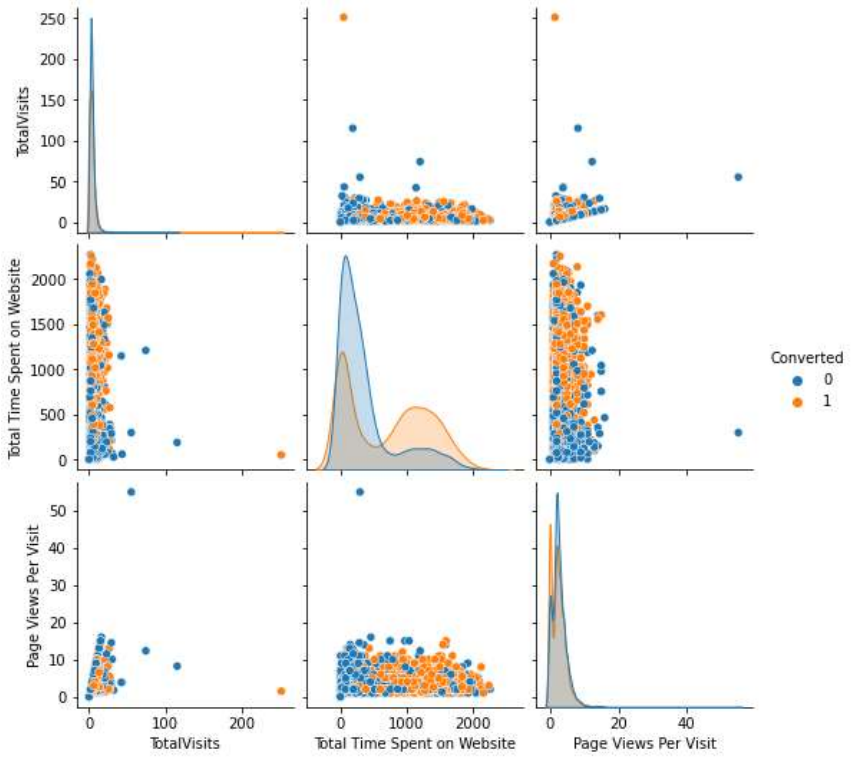
EDA

We did basic info check on data set

Checked the statistical features

Data cleaning

Pair plots to check relation between numerical variables



Dummy variables and encoding the data

Created dummy variables for categorical variables like ‘Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity‘ and then dropped the original column from dataset

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Live Chat	Lead Source_Olark Chat	Lead Source_Organic Search
0	0	0.0	0	0.0	0	0	0	0	0	0	0	1	0
1	0	5.0	674	2.5	0	0	0	0	0	0	0	0	1
2	1	2.0	1532	2.0	1	0	0	1	0	0	0	0	0
3	0	1.0	305	1.0	1	0	0	1	0	0	0	0	0

Splitting data and Feature Scaling

Split data in train and test set with 30% in test and 70% in train

Did feature scaling using MinMaxScaler for features 'TotalVisits', 'Page Views Per Visit', and 'Total Time Spent on Website'

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Live Chat	Lead Source_Olark Chat	Lead Source_Organic Search	Le Source_P per Cli A
8003	0.015936	0.029489	0.125	1	0	0	1	0	0	0	0	0	
218	0.015936	0.082306	0.250	1	0	0	1	0	0	0	0	0	
4171	0.023904	0.034331	0.375	1	0	0	1	0	0	0	0	0	
4037	0.000000	0.000000	0.000	0	0	0	0	0	0	0	1	0	
3660	0.000000	0.000000	0.000	0	1	0	0	0	0	0	0	0	

Model building – Logistic regression

We used hybrid approach to build model by starting with RFE and then manually eliminating the features not adding value to model

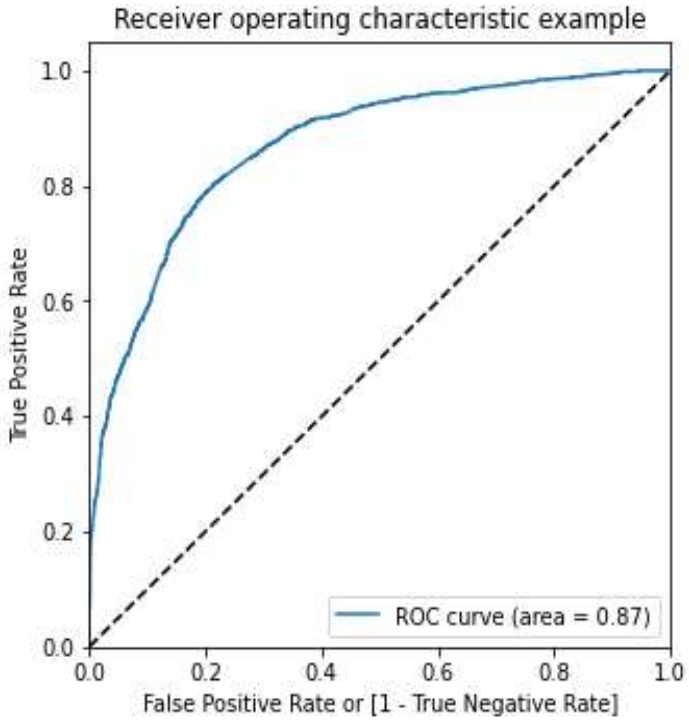
Identified features using RFE with 20 features as starting point

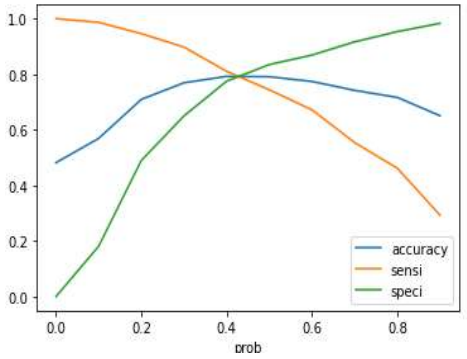
We used statsmodels to fit the model on identified features and subsequently eliminating the redundant feature using p value and VIF

Running prediction on test data set

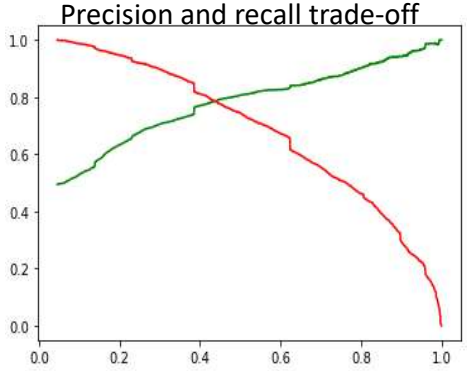
Achieved model accuracy of 78% on test data set

Model Evaluation





Precision and recall trade-off



Area under ROC is 87%

Optimal cutoff point if 0.42

Sensitivity: 79.9%

Specificity: 78.7%

Precision: 80.7%

Recall: 74.4%

From Precision and recall trade-off we get the optimal cutoff of 0.44, which is used to make prediction on test set

Conclusion

The model used total of 14 features and gave an accuracy of 79%.

In order of importance, the top few features were

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Do Not Email_Yes

The measure of Sensitivity (with around 79.9%) and specificity (with around 78.7%) are quite close to the target of 80% conversion rate as expected by the company.

Thank You