Lead Score Case Study

Collaborators

Soundari Govindan

Kunal Garg

Jeena Elizabeth Jose

Problem Statement

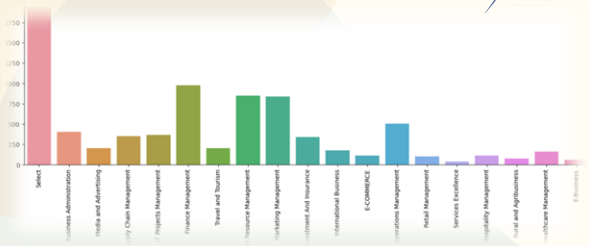
An education company- X Education is looking for potential customers who would use its Education platform for upskilling. The company notes the leads – leads are the ones who shows an interest to join the different courses.

X Education requires to build a model where in ,we need to assign a lead score to each of the leads.

Customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

In order to get a better lead conversion rate, the company needs to identify the ‘hot leads’.

Hot Leads are the most potential leads who will eventually be the customers.



Analysis - Flow

Step 1.Importing the Data
Step 2.Inspecting the Data
Step 3.Data Cleaning and Preparation
  3.1 Exploratory Data Analysis
  3.2 Checking Outliers
Step 4. Test – Train Split
Step 5.Feature Selection
Step 6.Model Building
Step 7.Plotting the ROC Curve
Step 8.Finding Optimal cut off
Step 9. Prediction making

Test – Train Split

Target Variable (y) – Converted
Independent Variables (X) – Total 66 variables
Splitting Train Data – 60 %
Split of Test Data – 40 %
Assume Random State – 100
Feature Scaling – Minmax Scaler

Model Building

- Feature Selection – RFE a logistic regression model was built in Python using the function GLM() under statsmodel library.

- Some variables were removed first based on an automated approach, i.e. RFE (Running RFE with 20 variables)

- Manual approach based on the VIFs and p-values are used for further process.

- After dropping some columns which has p-values above 0.05 and very high VIF we have obtained our final model with 14 variables.

Features of final model

	Features	VIF
3	Lead Source_Olark Chat	2.06
12	Last Notable Activity_Modified	1.79
10	Specialization_Others	1.73
2	Lead Origin_Landing Page Submission	1.68
6	Last Activity_Olark Chat Conversation	1.62
11	Specialization_Select	1.57
8	Last Activity_SMS Sent	1.55
1	Total Time Spent on Website	1.27
0	Do Not Email	1.22
4	Lead Source_Reference	1.14
5	Lead Source_Welingak Website	1.12
9	Last Activity_Unsubscribed	1.09
7	Last Activity_Other_Activity	1.01
13	Last Notable Activity_Unreachable	1.01

Metrics for Probability Cut Off

	prob	accuracy	sensi	speci
0.0	0.0	0.382439	1.000000	0.000000
0.1	0.1	0.605988	0.973103	0.378644
0.2	0.2	0.751470	0.914025	0.650803
0.3	0.3	0.791514	0.871278	0.742118
0.4	0.4	0.804188	0.769933	0.825402
0.5	0.5	0.805658	0.684438	0.880726
0.6	0.6	0.797024	0.617675	0.908090
0.7	0.7	0.768736	0.490394	0.941106
0.8	0.8	0.748898	0.395293	0.967876
0.9	0.9	0.698935	0.233910	0.986913

Finding optimal cut off

Optimal cut off was arrived using Sensitivity, Specificity, Accuracy and precision and recall Trade-off.

- The optimum point to take it as a cut off probability is 0.38

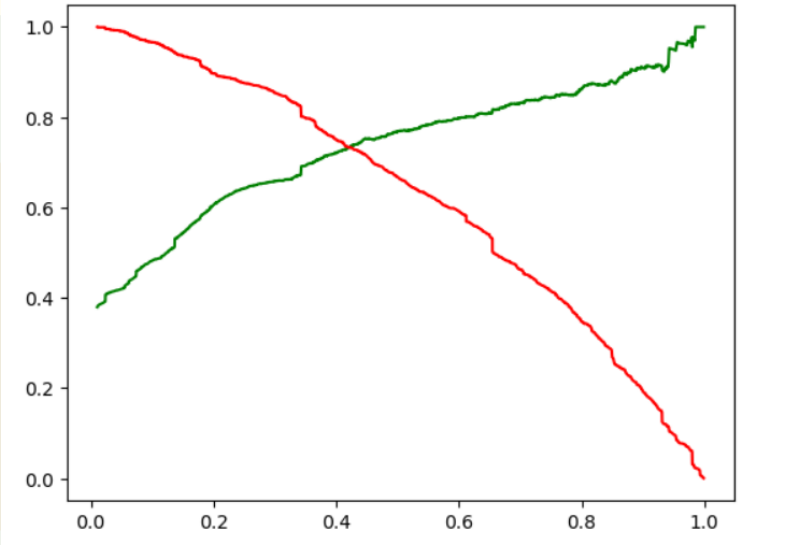


Fig 1. Precision Recall Trade off curve

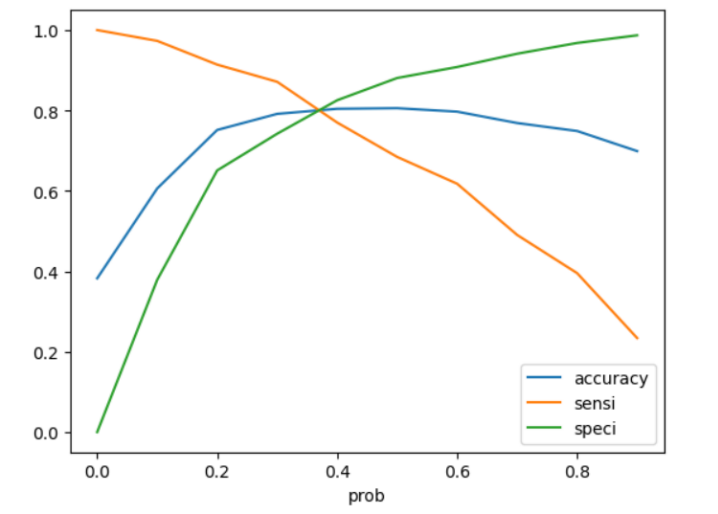
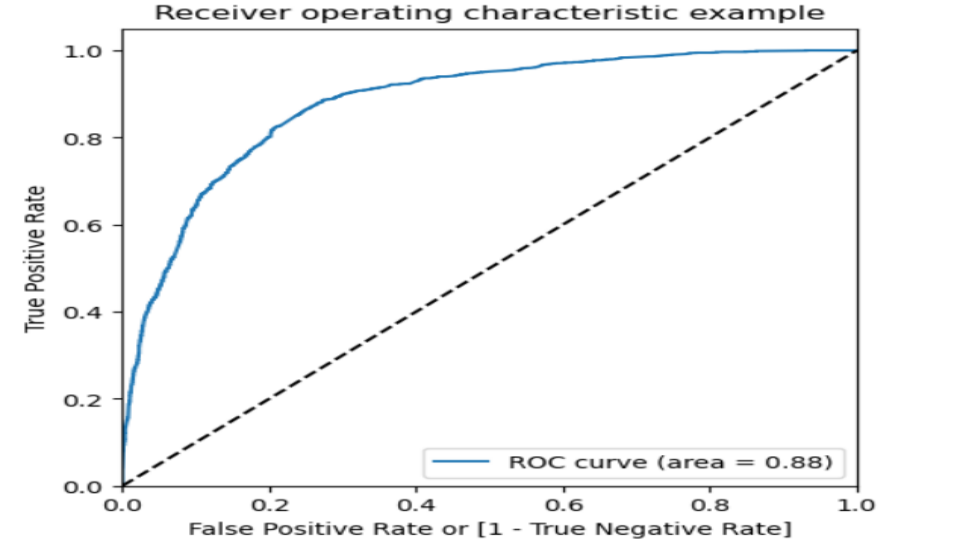


Fig 1. Cut off taking accuracy, sensitivity and specificity

ROC Curve

ROC Curve indicates the trade off between sensitivity and specificity.
The closer the curve follows the left-hand border and top border of the ROC space, the more accurate the test.

- ROC Curve area is 0.88, which is good.



Making predictions on the Test Data

A cut-off point of .38 is chosen based on the precision and recall tradeoff curve.

- Based on the predictions on the test dataset we have obtained the values for Sensitivity (77.67%),Specificity (80.72%)

Results

Trained Data	Test Data
Accuracy – 79.57 %	Accuracy – 79.58 %
Sensitivity – 85.06 %	Sensitivity – 77.67 %
Specificity – 76.17 %	Specificity – 80.72 %
Precision – 78.03 %	Precision – 70.53 %
Recall – 68.44 %	Recall – 77.67 %

Final Model

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2100	0.131	1.599	0.110	-0.047	0.467
Do Not Email	-1.8113	0.207	-8.760	0.000	-2.217	-1.406
Total Time Spent on Website	1.0818	0.042	25.456	0.000	0.999	1.165
Lead Origin_Landing Page Submission	-1.3078	0.134	-9.766	0.000	-1.570	-1.045
Lead Source_Olark Chat	1.2008	0.134	8.979	0.000	0.939	1.463
Lead Source_Reference	3.3460	0.258	12.954	0.000	2.840	3.852
Lead Source_Welingak Website	5.4736	0.734	7.460	0.000	4.036	6.912
Last Activity_Olark Chat Conversation	-1.0007	0.192	-5.211	0.000	-1.377	-0.624
Last Activity_Other_Activity	2.3496	0.507	4.635	0.000	1.356	3.343
Last Activity_SMS Sent	1.2918	0.080	16.157	0.000	1.135	1.448
Last Activity_Unsubscribed	1.5783	0.456	3.461	0.001	0.685	2.472
Specialization_Others	-2.2970	0.173	-13.276	0.000	-2.636	-1.958
Specialization_Select	-1.1077	0.139	-7.950	0.000	-1.381	-0.835
Last Notable Activity_Modified	-0.8740	0.087	-10.015	0.000	-1.045	-0.703
Last Notable Activity_Unreachable	1.5022	0.494	3.043	0.002	0.535	2.470

Conclusion

Company should focus more on the below features to get a lead converted.

- Last Activity
- Lead Source
- Total Time spent on the website

- As we see, Lead Source_Welingak Website and reference has the highest positive impact on the lead conversion, company should focus more on the referral leads during the aggressive calling phase.