

Lead Scoring Case Study

Prediction of potential leads

1. Business Context and Terminologies

- **Service provided:** X Education company markets and sells online courses.
- **Leads:** Individuals who shows interest in online courses by visiting related website as well as who are undergone or undergoing the courses.
- **Potential Leads or Hot Leads:** Leads who has the highest chance for conversion to actual lead.
- **Converted leads:** Leads who had converted to actual leads.
- **Conversion Rate:** $(\text{Successful converted lead count}) / (\text{contacted Lead count})$

2. Problem Statement and Business Goals

- **Problem Statement:** On day to day life, lakhs of people shows interest in online courses. However, it doesn't mean they are ready to buy it. Sales team contact every lead without considering the outcome which is not an efficient process. Hence, business want to make this process efficient and improve it's conversion rate by calling the potential candidates first.
- **Business goal:** Currently, the typical conversion rate is about 30% whereas business set the goal of around 80%.

3. Analysis Approach

- Initially, inspect the data and make it ready for analysis by handling null records, transforming data points as required and drop the irrelevant columns which are not necessary for analysis by performing EDA. Further, transform categorical variables into numeric by creating dummy variables. On top of that, split the dataset into train and test plus perform min max scaling on numeric variables. Begin building the logistic regression model(with target variable as Converted) within 15 to 20 variables and reduce it to 15 by using RFE and further reduce variables manually until it satisfies the criteria of $p\text{-value} < 0.05$ and $VIF < 5$. Plotting the ROC curve to find the optimal cut-off point. Finally, evaluate the model based on Predicted value on both Train and Test data set, creating confusion matrix and calculating Specificity, Sensitivity, Precision and Recall.

4. Business Results

- Below all are the 10 variable which effects the convertible factor of lead

6 Positively effected variable

Higher TotalVisits, Total Time Spent on Website, LeadOrigin_API, LeadSource_ReferenceLeadSource_Welingak Website and LastActivity_SMS Sent can result as potential or hot lead.

4 Negatively effected variable

Lower Do Not Email, LeadSource_Referral Sites, LastActivity_Converted to Lead and LastActivity_Page Visited on Website can result as potential or hot lead.

Based on above 10 variable inputs of a lead, Model can easily predict whether any particular lead has potential to be converted or not (1 or 0 in predicted column)

With sensitivity of 80.4% , specificity of 80.7%

5. Correlated with target variable

- Total Time Spent on Website, LeadOrigin_Lead Add Form, LeadSource_reference, LeadSource_Welingak Website and LastActivity_SMS Sent variables are highly correlated with Converted variable which means these are potentially variables which helps in building the model.

Total Time Spent on Website	-0.019	-0.047	0.00	0.36
Page Views Per Visit	-0.058	0.035	-0.00	-0.0019
Search	-0.013	-0.0120	0.00	0.0017
Newspaper Article	-0.0120	0.0449	0.00	0.0037
X Education Forums	-0.0030	0.0030	0.00	0.0082
Newspaper	-0.00480	0.0030	0.00	0.0082
Digital Advertisement	-0.0180	0.0050	0.00	0.0056
Through Recommendations	-0.00560	0.0080	0.00	0.019
Free copy of Mastering The Interview	-0.11	0.058	-0.0	-0.031
LeadOrigin_API	-0.05	-0.0830	0.00	-0.11
LeadOrigin_Landing Page Submission	-0.07	0.099	-0.00	-0.037
LeadOrigin_Lead Add Form	-0.025	-0.032	0.02	0.3
LeadSource_Direct Traffic	-0.14	0.11	-0.00	0.074
LeadSource_Google	0.14	-0.0710	0.00	0.03
LeadSource_Olark Chat	-0.067	-0.0510	0.00	0.13
LeadSource_Organic Search	-0.017	0.036	-0.00	0.0074
LeadSource_Reference	-0.024	-0.038	0.03	0.26
LeadSource_Referral Sites	0.17	-0.01	-0.00	0.032
LeadSource_Welingak Website	-0.0039	0.003	-0.00	0.015
LastActivity_Converted to Lead	-0.015	-0.0630	0.00	0.12
LastActivity_Email Bounced	-0.053	0.62	-0.00	0.13
LastActivity_Email Link Clicked	-0.028	-0.0440	0.00	0.038
LastActivity_Email Opened	0.11	-0.22	0.00	0.023
Activity_Form Submitted on Website	-0.015	-0.0190	0.00	0.032
LastActivity_Olark Chat Conversation	-0.0046	-0.055	0.01	-0.21
LastActivity_Page Visited on Website	-0.011	0.083	-0.00	0.079
LastActivity_SMS Sent	-0.14	-0.02	-0.00	0.34
LastActivity_Unreachable	-0.08	0.0080	0.00	0.011
LastActivity_Unsubscribed	-0.0015	0.26	-0.00	0.0234
Lead Number				
Do Not Email				
Do Not Call				
Converted				

6. Ignore highly correlated variables

- LeadOrigin_Lead Add Form and LeadSource_reference are highly correlated on top of that LeadOrigin_Lead Add Form is also highly correlated with LeadSource_Welingak Website. Therefore, LeadOrigin_Lead Add Form variable is removed from the model due to higher correlations.

LeadOrigin_Lead Add Form	-0.023	0.031	-0.16	-0.2	-0.012	-0.0048	0.0034	-0.0058	0.0058	-0.17	-0.22	-0.29	1	-0.17	-0.18	-0.13	-0.1	0.86
LeadSource_Direct Traffic	0.11	-0.011	0.087	0.14	0.022	0.0089	0.02	0.035	0.019	0.6	-0.45	0.52	-0.17	1	-0.43	-0.3	-0.2	-0.14
LeadSource_Google	-0.069	0.0067	0.1	0.22	-0.03	-0.012	-0.0087	-0.015	-0.015	-0.32	0.018	0.079	-0.18	-0.43	1	-0.34	-0.2	-0.16
LeadSource_Olark Chat	-0.053	-0.0086	-0.29	-0.38	-0.0026	-0.0086	0.0061	-0.011	0.0079	-0.3	0.6	-0.52	-0.13	-0.3	-0.34	1	-0.1	-0.11
LeadSource_Organic Search	-0.036	-0.0067	0.18	0.11	0.027	0.02	-0.0048	-0.0083	0.0083	0.14	-0.00056	0.054	-0.1	-0.23	-0.26	-0.18	1	-0.088
LeadSource_Reference	-0.032	0.036	-0.14	-0.17	-0.01	-0.0041	-0.0029	-0.005	-0.005	-0.14	-0.19	-0.25	0.86	-0.14	-0.16	-0.11	-0.08	1
LeadSource_Referral Sites	0.00038	0.0021	0.069	0.0059	-0.0051	-0.0021	-0.0015	-0.0025	-0.0025	-0.044	0.099	-0.08	-0.031	-0.072	-0.08	-0.056	-0.04	-0.027
LeadSource_Welingak Website	-0.012	-0.0023	-0.079	-0.095	-0.0056	-0.0023	-0.0016	-0.0028	0.0028	-0.082	-0.1	-0.14	0.48	-0.079	-0.088	-0.062	-0.04	-0.029
LastActivity_Converted to Lead	-0.063	-0.0039	-0.062	-0.011	-0.0096	-0.0039	-0.0028	-0.0048	0.0048	0.018	-0.0086	0.039	-0.059	0.055	0.035	-0.11	0.04	-0.051
LastActivity_Email Bounced	0.62	-0.0035	-0.045	-0.036	-0.0085	-0.0035	-0.0025	-0.0043	0.0043	0.026	-0.042	0.068	-0.049	0.086	-0.05	-0.022	0.01	-0.045
LastActivity_Email Link Clicked	-0.041	-0.0031	-0.015	-0.035	-0.0075	-0.0031	-0.0022	-0.0038	0.0038	-0.017	0.04	-0.036	-0.0047	-0.016	-0.027	0.056	-0.01	-0.0011
LastActivity_Email Opened	-0.22	0.0046	0.014	0.013	0.026	0.0046	-0.0097	-0.0019	0.0019	0.035	-0.039	0.033	-0.0082	0.019	0.023	-0.033	-0.02	0.0045
ity_Form Submitted on Website	-0.014	-0.002	0.014	0.0032	-0.005	-0.002	-0.0014	-0.0025	0.0025	0.011	-0.047	0.053	-0.014	-0.005	0.061	-0.034	-0.03	-0.0073
ctivity_Olark Chat Conversation	-0.066	0.023	-0.13	-0.19	-0.015	-0.0061	-0.0043	-0.0075	0.0075	-0.19	0.37	-0.31	-0.086	-0.17	-0.087	0.42	-0.07	-0.072
ctivity_Page Visited on Website	-0.07	-0.0047	0.22	0.028	0.018	0.031	0.047	0.023	0.053	0.057	-0.073	0.093	-0.038	0.064	0.029	-0.1	0.01	-0.033
LastActivity_SMS Sent	-0.02	-0.012	-0.0027	0.13	-0.013	-0.012	-0.0084	0.0012	-0.015	0.039	-0.13	0.065	0.14	0.0078	0.0094	-0.13	0.04	0.11
LastActivity_Unreachable	-0.014	-0.0018	0.014	-0.0013	-0.0045	-0.0018	-0.0013	-0.0022	0.0022	0.0065	-0.038	0.039	-0.0026	-0.011	0.032	-0.042	0.01	0.0047
LastActivity_Unsubscribed	0.27	-0.0016	0.0048	0.0027	-0.0038	-0.0016	-0.0011	-0.0019	0.0019	0.042	-0.022	0.03	-0.016	0.019	-0.01	-0.019	0.01	-0.012
	Do Not Email	Do Not Call	TotalVisits	Total Time Spent on Website	Search	Newspaper Article	X Education Forums	Digital Advertisement	Through Recommendations	y of Mastering The Interview	LeadOrigin_API	pin_Landing Page Submission	LeadOrigin_Lead Add Form	LeadSource_Direct Traffic	LeadSource_Google	LeadSource_Olark Chat	LeadSource_Organic Search	LeadSource_Reference

7. Final Model

- After rigorous building of model, on 5th try, I mean 5th model satisfied all the conditions like p value<0.05 and VIF<5
- Please see the model on right hand side with it's statistics and variable influence factor coefficient to know how much each variable impact the final outcome.

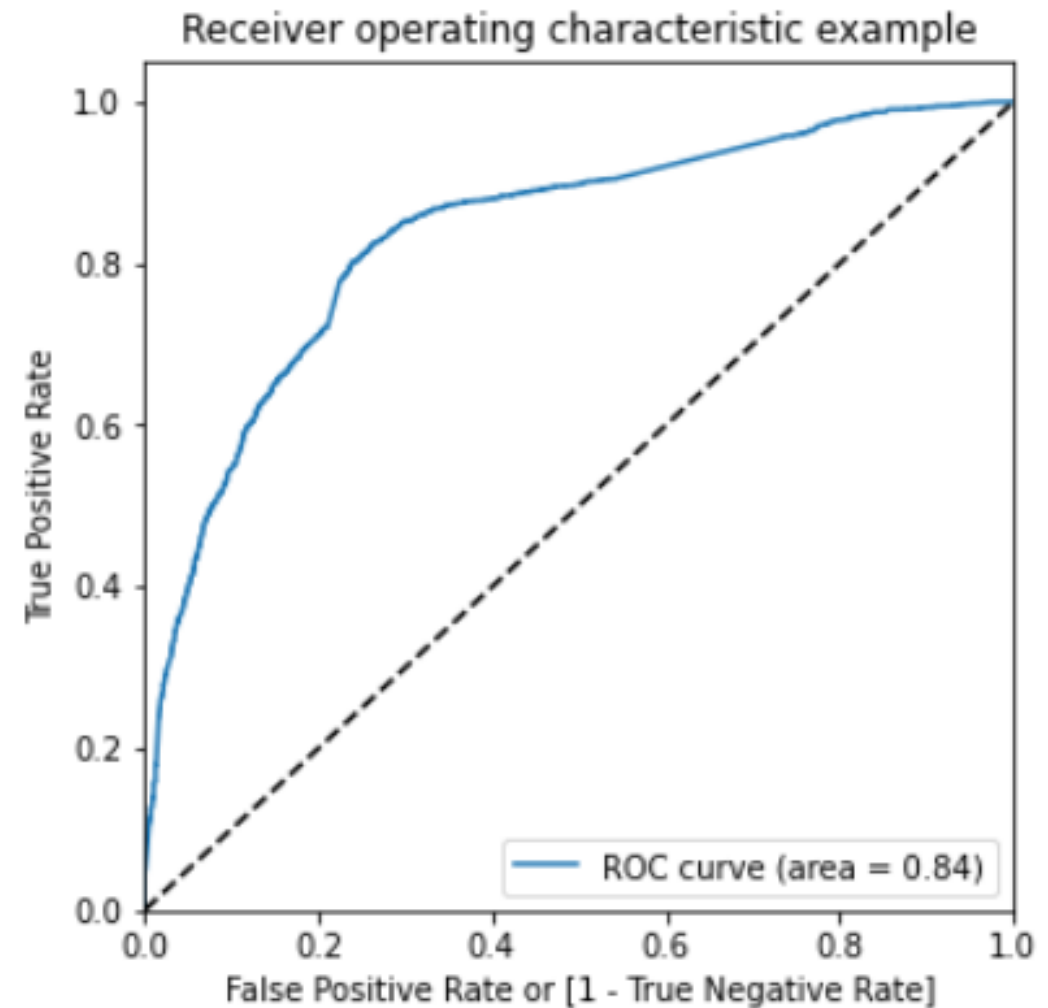
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6320
Model:	GLM	Df Residuals:	6309
Model Family:	Binomial	Df Model:	10
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3023.4
Date:	Mon, 27 Feb 2023	Deviance:	6046.8
Time:	01:06:31	Pearson chi2:	6.46e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3121
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-2.1497	0.076	-28.410	0.000	-2.298	-2.001
Do Not Email	-1.3751	0.150	-9.194	0.000	-1.668	-1.082
TotalVisits	5.3061	1.884	2.817	0.005	1.614	8.998
Total Time Spent on Website	4.0394	0.141	28.591	0.000	3.762	4.316
LeadOrigin_API	0.3853	0.071	5.415	0.000	0.246	0.525
LeadSource_Reference	4.0108	0.215	18.627	0.000	3.589	4.433
LeadSource_Referral Sites	-0.8017	0.311	-2.581	0.010	-1.411	-0.193
LeadSource_Welingak Website	5.5528	0.722	7.688	0.000	4.137	6.968
LastActivity_Converted to Lead	-0.9866	0.191	-5.163	0.000	-1.361	-0.612
LastActivity_Page Visited on Website	-0.3968	0.142	-2.800	0.005	-0.675	-0.119
LastActivity_SMS Sent	1.2571	0.070	18.034	0.000	1.120	1.394

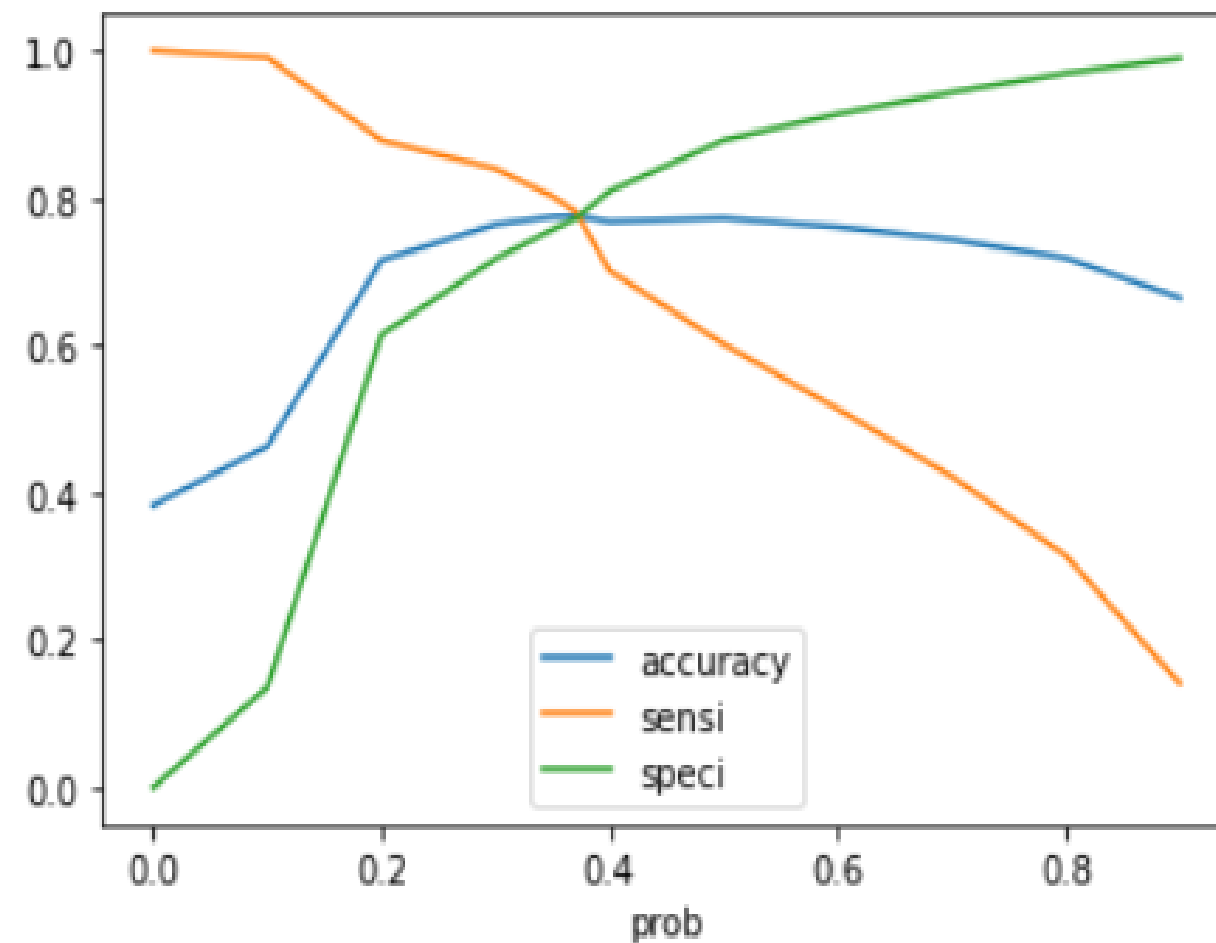
8. Model Evaluation – ROC curve

- ROC curve plotted as shown in the figure
- Since area under ROC curve is about 84%, model performance should be excellent.



9. Optimal Cut-off Point

- Plot among accuracy, sensitivity and specificity intersect at the optimal cut-off point which minimizes the cost of false positives and false negatives. it means model will predict with less errors at this point.
- Just below 0.4 is the optimum point to take it as a cutoff probability.



Thank You