Lead Score Case Study

Abhigyan

Batch DS C45

Problem statement

X Education is an education company which sells online courses to industry professionals. X Education gets a lot of leads, but its lead conversion rate is very poor i.e. around 30%

Expectations

X Education wants to make process more efficient by identify the leads that are most likely to convert into paying customers. The company needs to build a model where in a lead score can be generated 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. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Inputs provided

A leads dataset from the past with 9240 data points is provided . This dataset consists of 37 attributes. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Many of the categorical variables have a level called 'Select' which is as good as a null value.

Analysis Approach

1. **Data understanding, preparation and EDA** - All data quality checks are performed, and all data quality issues are addressed in the right way. Dummy variables are also created wherever applicable. The data is converted to a clean format suitable for analysis in Python.
2. **Model building** - Logistic regression models are made using appropriate variables and the best one is chosen based on key performance metrics post model parameters tunning.
3. **Model evaluation** – The selected model is evaluated using both evaluation metrics viz sensitivity-specificity view, and precision-recall view

Data understanding, preparation and EDA

- 6 columns having missing values more than 30% are dropped.
- “Country” and “City” column are dropped as not of much importance for online education analysis.
- “Lead Profile” and “How did you hear about X Education” column have 63% and 72% of rows as value `Select` respectively which is of no use to the analysis hence dropped.
- `Do Not Call`, `Search`, `Magazine`, `Newspaper Article`, `X Education Forums`, `Newspaper`, `Digital Advertisement`, `Through Recommendations`, `Receive More Updates About Our Courses`, `Update me on Supply Chain Content`, `Get updates on DM Content`, `I agree to pay the amount through cheque` and `What matters most to you in choosing a course` columns have very high imbalance i.e one value was majorly present for all the data points. Hence dropped (13 columns) as they won't help with our analysis.

Data understanding, preparation and EDA

- “What is your current occupation” is still having a lot of null values. Since we have already lost so many feature variables, so we are keeping the variable but dropping the null rows for the column
- Remining 5 columns have a smaller number of missing values, hence row with missing values are dropped one by one for each columns.
- Post all the cleaning, 69% rows of original data set is retained.
- “Prospect ID” and “Lead Number” are not of any use in analysis hence dropped.
- Dummy variables are created for 11 categorical variables
- Dummy variables are combined with cleaned data set and original columns are dropped from which dummy variables are created.

Data understanding, preparation and EDA

- Target variable(“Converted”) is separated as Y data frame and remining variable are under X data frame
- Prepared data set is divided into train set & test set with train_size=0.7, test_size=0.3, random_state=100
- Min-Max scaling is done for 3 continuous variable ("TotalVisits", "Total Time Spent on Website", "Page Views Per Visit“)
- Correlation matrix is created with data set

Model building

- There are 74 variables which is tough to deal manually hence RFE is used to select to most suitable 15 variables for model building.
- Logistic regression models are made with the finalised variables in such a way that P-values of the model is less than 0.05 and VIF is less than 5.
- 5th Model is the best fit and its equation is :

P = 1/(1+e^(-A))

Where,

A = 0.204 + 11.1489* TotalVisits + 4.4223*(Total Time Spent on Website) + 4.2051*(Lead Origin_Lead Add Form) + 2.7846*(Last Notable Activity_Unreachable) + 2.7552*(Last Activity_Had a Phone Conversation) + 2.1526*(Lead Source_Welingak Website) + 1.4526*(Lead Source_Olark Chat) + 1.1856*(Last Activity_SMS Sent) - 1.50307*(Do Not Email_Yes) – 2.5445*(What is your current occupation_Unemployed) – 2.3578*(What is your current occupation_Student)

Model building

- P value & VIF for the finalised model is :

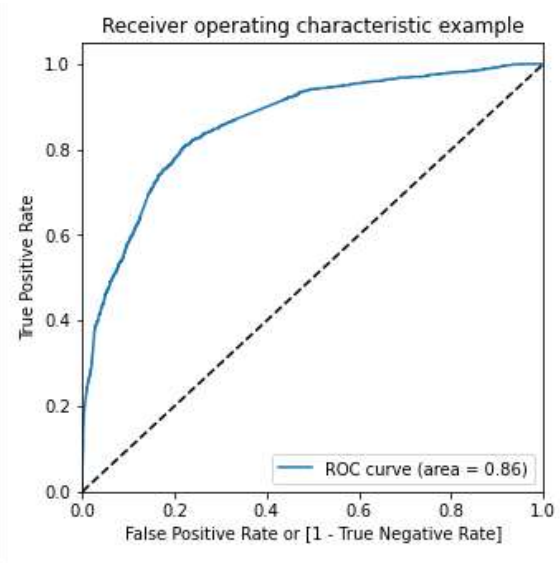
	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Fri, 13 Jan 2023	**Deviance:**	4158.1
Time:	17:37:36	**Pearson chi2:**	4.80e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3642
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

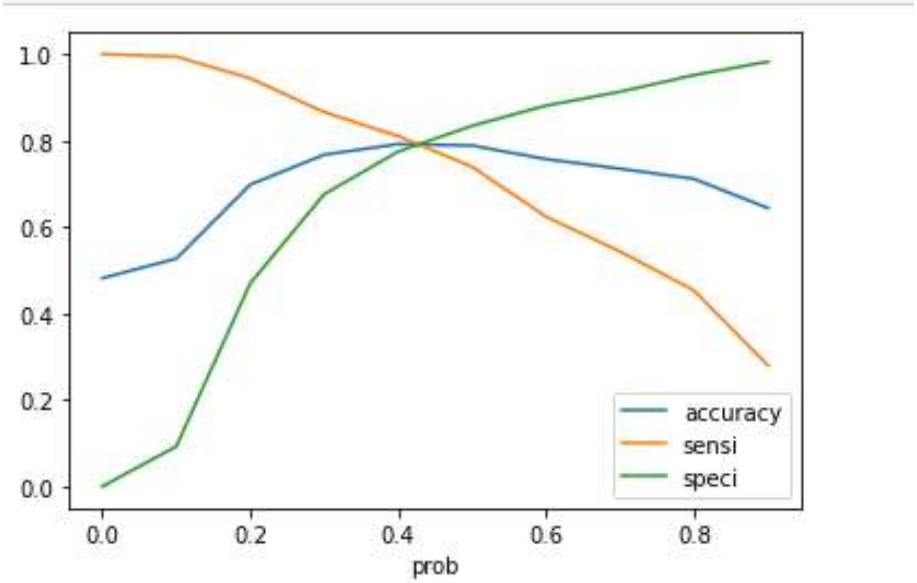
Model evaluation: Sensitivity-Specificity view

- The accuracy on the train set is 0.7886, sensitivity = 0.7394 and specificity = 0.8343 with probability cut off = 0.5
- Area under ROC curve is 0.86



Model evaluation: Sensitivity-Specificity view

- From accuracy, sensitivity & specificity graph, optimal value of probability cut off is coming as 0.42



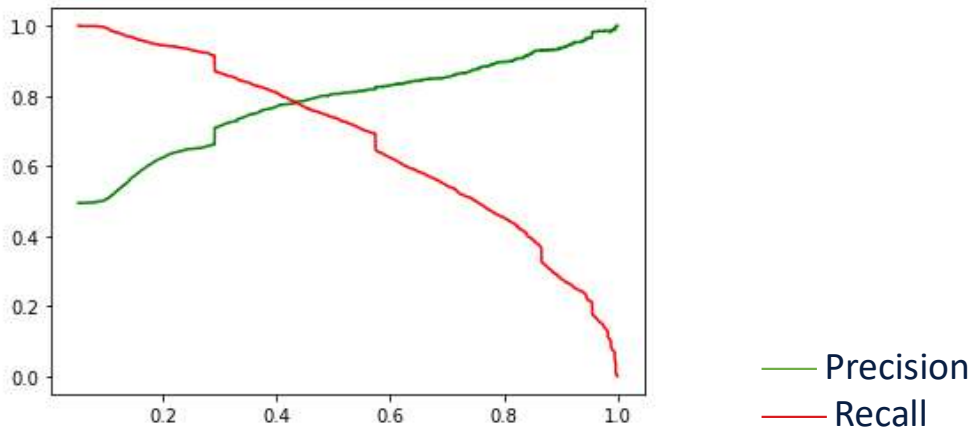
Model evaluation: Sensitivity-Specificity view

- The accuracy on the train set is 0.7908, sensitivity = 0.7934 and specificity = 0.7885 with probability cut off = 0.42

- The accuracy on the test set is 0.7845, sensitivity = 0.7795 and specificity = 0.7892 with probability cut off = 0.42

Model evaluation: Precision-Recall view

- The accuracy on the train set is 0.7886, precision = 0.8058 and specificity = 0.7394 with probability cut off = 0.5
- From Precision & Recall graph, optimal value of probability cut off is coming as 0.44.



Model evaluation : Precision-Recall view

- The accuracy on the train set is 0.7895, sensitivity = 0.7840 and specificity = 0.7771 with probability cut off = 0.44

- The accuracy on the test set is 0.7866, sensitivity = 0.7828 and specificity = 0.7646 with probability cut off = 0.44

Observation

- The model is evaluated with both Sensitivity-Specificity view & Precision-Recall view and its accuracy, sensitivity, specificity, precision and recall is very descent and almost same for train set and test set.
- The area under ROC curve is 0.86, hence the model is very good.
- Probability cut off for Sensitivity-Specificity view is 0.42
- Probability cut off for Precision-Recall view is 0.44
- The equation for log odds is:

  0.204 + 11.1489* TotalVisits + 4.4223*(Total Time Spent on Website) + 4.2051*(Lead Origin_Lead Add Form) + 2.7846*(Last Notable Activity_Unreachable) + 2.7552*(Last Activity_Had a Phone Conversation) + 2.1526*(Lead Source_Welingak Website) + 1.4526*(Lead Source_Olark Chat) + 1.1856*(Last Activity_SMS Sent) - 1.50307*(Do Not Email_Yes) – 2.5445*(What is your current occupation_Unemployed) – 2.3578*(What is your current occupation_Student)

Conclusion

- Below are the variables which contribute most towards the probability of a lead getting converted in decreasing order:
    a. **TotalVisits** –Higher the count, higher is the probability of lead converting successfully. It has highest coefficient viz 11.1489.
    b. **Total Time Spent on Website** - Higher the time spent on website, higher is the probability of lead converting successfully. Its coefficient is 4.4223.
    c. **Lead Origin** – If "Lead origin" = "Lead Add Form", probability of lead converting successfully is high. Its coefficient is 4.2051.
    d. **Last Notable Activity** - If "Last Notable Activity" = "Unreachable", probability of lead converting successfully is high. Its coefficient is 2.7846
    e. **Last Activity** - If "Last Activity" = "Had a Phone Conversation", probability of lead converting successfully is high. Its coefficient is 2.7552
    f. **Lead Source** - If "Lead Source" = "Welingak Website", probability of lead converting successfully is high. Its coefficient is 2.1526
    g. **Variable: Lead Source** - If "Lead Source" = "Olark Chat", probability of lead converting successfully is high. Its coefficient is 1.4526
    h. **Last Activity** - If "Last Activity" = "SMS Sent", probability of lead converting successfully is high. Its coefficient is 1.1856

- Below are the variables which degrades the probability of a lead getting converted in decreasing order:
    a. **What is your current occupation**– if "What is your current occupation" = "Unemployed", probability of successful conversion is negative. It has highest negative coefficient viz is -2.3578
    b. **What is your current occupation**– if "What is your current occupation" = "Student", probability of successful conversion is negative. Its coefficient is -2.5445.
    c. **Do Not Email** – if "Do Not Email "= "Yes", probability of successful conversion is negative. Its coefficient is -1.50307