

Lead Score Case Study By Krishna Katta and Neetima Verma Executive Program in DS C-39

13-06-2022



Problem Statement:

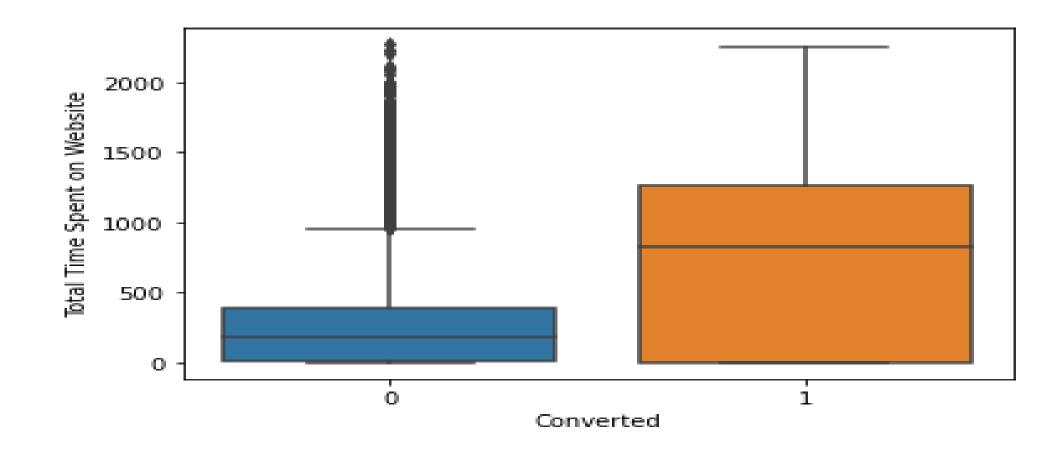
The Data of customer has been provided by an education company with numerus predictors. The task is to construct a model to predict if a particular customer will enroll in the program or not. The customer who enroll in the program is called Hot Leads. This will help the sales team, their attention to potential Hot Leads to increase the conversion rate.



Analysis based on "Total Time Spent on Website":

Observation: The following can be observed.

• Clearly customers who have spend more time on the website have higher probability of enrolling with the Academy.

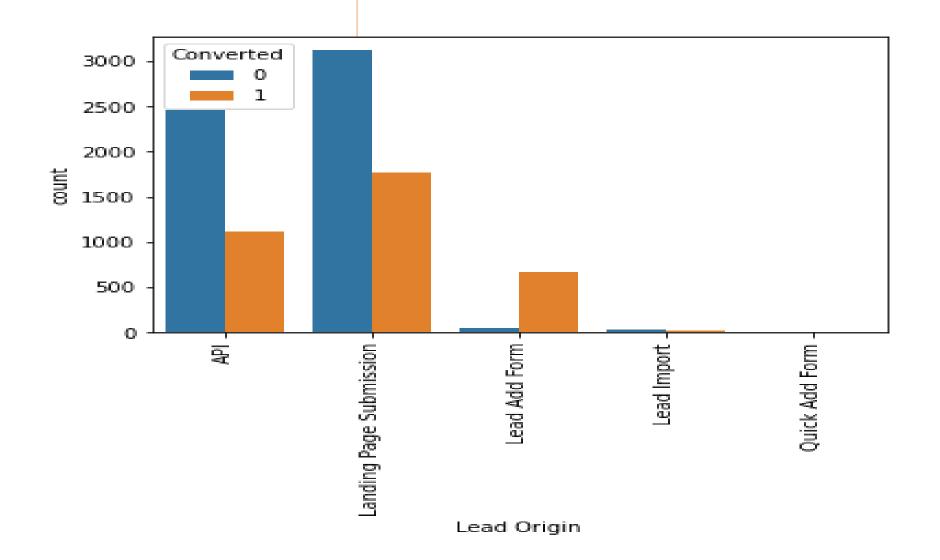




Analysis based on "Lead Origin":

Observation: The following can be observed.

Lead Add Form has high conversion ratio.

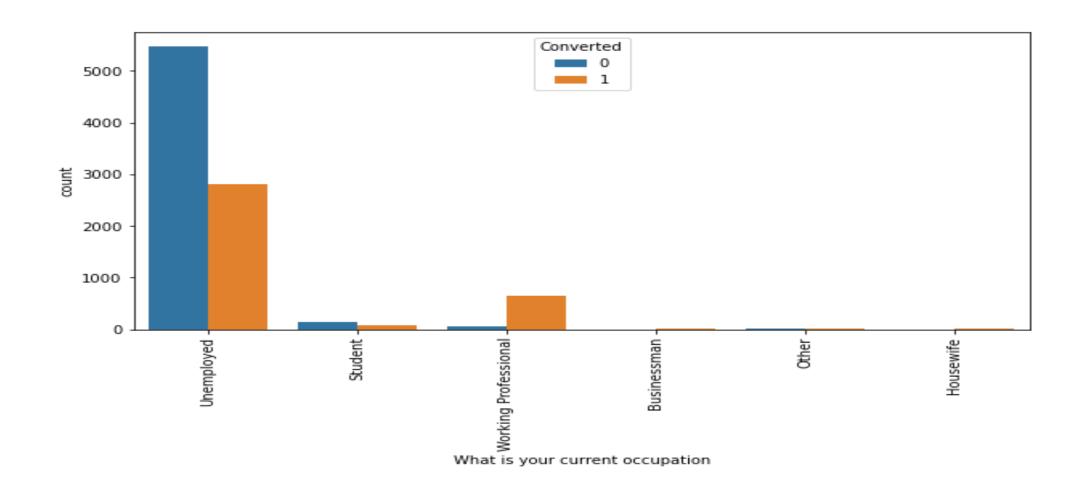


Analysis based on "Occupation":



Observation: The following can be observed.

 Though number of converted customers are high in "Unemployed" category, It is the working professionals who have higher conversion rate

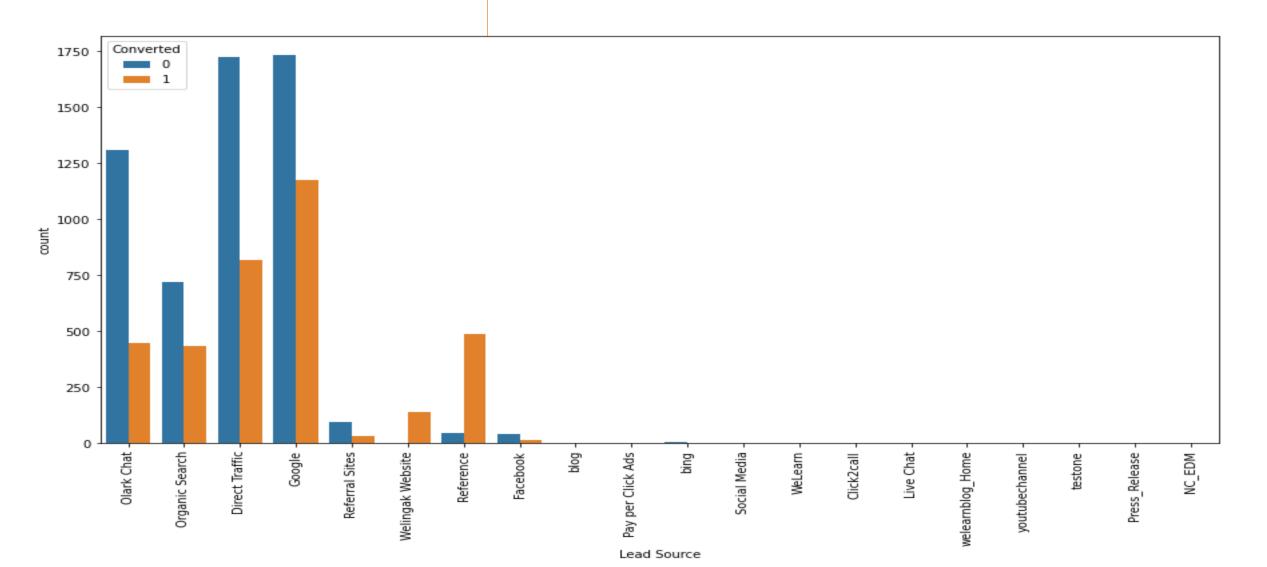


Analysis based on "Lead Source":

Observation: The following can be observed.



 "Reference" leads and "Welingak Website" categories have high conversion rate.



Analysis based on Correlation on RFEs:

Observation: By observing Heatmap, the following can be observed.



From the above Heatmap it is clear that "Total time Spent on Website",
"Lead Origin_Lead Add Form", "Working professionals" and "Lead
Source_Welingak Website" have high correlation with target variable
"Converted"

CO	iive	rteu																	
Do Not Email -	1	-0.14	0.034	-0.046	0.033	-0.021	-0.051	-0.00091	-0.063	0.62	-0.055	-0.0097	-0.05	-0.038	-0.19	-0.011	0.12	-0.013	0.05
Converted -	-0.14	1	0.03	0.36	-0.0033	0.32	-0.13	0.15	-0.12	-0.12	-0.21	0.042	0.31	-0.036	-0.022	0.043	-0.24	-0.073	-0.036
TotalVisits -	0.034	0.03	1	0.22	0.51	-0.16	-0.31	-0.076	-0.071	-0.046	-0.14	-0.0062	-0.0045	-0.0086	0.036	0.022	-0.11	-0.038	0.23
Total Time Spent on Website -	-0.046	0.36	0.22	1	0.32	-0.14	-0.38	-0.079	-0.011	-0.029	-0.19	0.0071	0.099	-0.037	0.0096	0.022	-0.11	-0.051	0.027
Page Views Per Visit -	0.033	-0.0033	0.51	0.32	1	-0.25	-0.5	-0.12	-0.064	-0.036	-0.23	-0.0073	-0.0079	-0.026	0.066	0.014	-0.093	-0.088	0.027
Lead Origin_Lead Add Form -	-0.021	0.32	-0.16	-0.14	-0.25	1	-0.14	0.43	-0.064	-0.029	-0.093	0.04	0.2	-0.0013	-0.048	-0.011	-0.013	-0.03	-0.024
Lead Source_Olark Chat -	-0.051	-0.13	-0.31	-0.38	-0.5	-0.14	1	-0.06	-0.11	-0.022	0.43	-0.016	-0.086	0.045	-0.045	-0.012	0.091	0.17	-0.054
Lead Source_Welingak Website	-0.00091	0.15	-0.076	-0.079	-0.12	0.43	-0.06	1	-0.028	-0.019	-0.04	-0.0041	-0.036	-0.0043	-0.02	-0.0049	-0.019	-0.011	-0.0091
Last Activity_Converted to Lead -	-0.063	-0.12	-0.071	-0.011	-0.064	-0.064	-0.11	-0.028	1	-0.042	-0.076	-0.0073	-0.028	-0.03	-0.15	-0.0086	0.29	-0.031	-0.042
Last Activity_Email Bounced -	0.62	-0.12	-0.046	-0.029	-0.036	-0.029	-0.022	-0.019	-0.042	1	-0.066	-0.0063	-0.042	-0.026	-0.13	-0.0074	0.18	-0.027	-0.036
Last Activity_Olark Chat Conversation -	-0.055	-0.21	-0.14	-0.19	-0.23	-0.093	0.43	-0.04	-0.076	-0.066	1	-0.011	-0.081	-0.047	-0.23	-0.013	0.32	0.41	-0.065
What is your current occupation_Housewife -	-0.0097	0.042	-0.0062	0.0071	-0.0073	0.04	-0.016	-0.0041	-0.0073	-0.0063	-0.011	1	-0.0095	0.02	0.014	-0.0013	0.0021	-0.0047	-0.0062
What is your current occupation_Working Professional	-0.05	0.31	-0.0045	0.099	-0.0079	0.2	-0.086	-0.036	-0.028	-0.042	-0.081	-0.0095	1	-0.016	-0.019	0.0097	-0.08	-0.029	-0.012
Last Notable Activity_Email Link Clicked -	-0.038	-0.036	-0.0086	-0.037	-0.026	-0.0013	0.045	-0.0043	-0.03	-0.026	-0.047	0.02	-0.016	1	-0.092	-0.0054	-0.11	-0.02	-0.026
Last Notable Activity_Email Opened -	-0.19	-0.022	0.036	0.0096	0.066	-0.048	-0.045	-0.02	-0.15	-0.13	-0.23	0.014	-0.019	-0.092	1	-0.026	-0.51	-0.094	-0.13
Last Notable Activity_Had a Phone Conversation	-0.011	0.043	0.022	0.022	0.014	-0.011	-0.012	-0.0049	-0.0086	-0.0074	-0.013	-0.0013	0.0097	-0.0054	-0.026	1	-0.03	-0.0055	-0.0074
Last Notable Activity_Modified -	0.12	-0.24	-0.11	-0.11	-0.093	-0.013	0.091	-0.019	0.29	0.18	0.32	0.0021	-0.08	-0.11	-0.51	-0.03	1	-0.11	-0.14
Last Notable Activity_Olark Chat Conversation -	-0.013	-0.073	-0.038	-0.051	-0.088	-0.03	0.17	-0.011	-0.031	-0.027	0.41	-0.0047	-0.029	-0.02	-0.094	-0.0055	-0.11	1	-0.027
Last Notable Activity_Page Visited on Website -	0.05	-0.036	0.23	0.027	0.027	-0.024	-0.054	-0.0091	-0.042	-0.036	-0.065	-0.0062	-0.012	-0.026	-0.13	-0.0074	-0.14	-0.027	1
	Do Not Email -	Converted -	DtalVisits -	Total Time Spent on Website -	Page Views Per Visit -	Lead Origin_Lead Add Form -	Lead Source_Olark Chat -	Lead Source_Welingak Website -	Last Activity_Converted to Lead -	Last Activity_Email Bounced -	Last Activity_Olark Chat Conversation -	What is your current occupation_Housewife -	is your current occupation_Working Professional –	Last Notable Activity_Email Link Clicked -	Last Notable Activity_Email Opened -	Last Notable Activity_Had a Phone Conversation –	Last Notable Activity_Modified -	Last Notable Activity_Olark Chat Conversation –	Last Notable Activity_Page Visited on Website –

Data preparation



- ➤ Replacing "Select" with nan
- > Dropping columns which has more that 40% null values.
- ➤ Replacing "Yes" and "No" with 1's and 0's
- > Imputing null values in Categorical variables with Mode.
- > Imputing null values in Numerical variables with Mean.
- Dropping variables which had single unique values.
- > Created Dummy variables for all the Categorical variables and Dropping the Categorical Variables.

Test-Train Split



- > Splitting data into X data which are predicted variables and y as target variable.
- > Splitting the data into 70% Training Data and 30% Testing Data.

Feature Scaling



- ➤ Used MinMaxScaler to re-scale the following numerical variables.
 - ✓ TotalVisits
 - ✓ Total Time Spent on Website
 - ✓ Page Views Per Visit



Logistic Regression Model Building

Model Building using GLM logm1 = sm.GLM(y_train,(sm.add_constant(X_train)), family = sm.families.Binomial()) logm1.fit().summary()

Generalized Linear Model Regression Results									
Dep. Variable:	Converted	No. Observations:	6468						
Model:	GLM	Df Residuals:	6384						
Model Family:	Binomial	Df Model:	83						
Link Function:	logit	Scale:	1.0000						
Method:	IRLS	Log-Likelihood:	nan						
Date:	Sun, 20 Jun 2021	Deviance:	nan						
Time:	12:26:33	Pearson chi2:	8.48e+18						
No. Iterations:	100								
Covariance Type:	nonrobust								



Feature Selection Using RFEs

- > Recursive feature elimination carried out on the data set.
- ➤ GLM Models built on using the chosen variables and metrics like Accuracy, Sensitivity and Specificity are compared for different number of variables in the RFE method.
- ➤ 18 variables are chosen based on trial and error method. More than 18 variables does not increase the accuracy of the model by much. Less than 18 Variables reduces the accuracy of the model.





➤ VIF for 18 variables are checked for values greater than 5. None found.

	Features	VIF
3	Page Views Per Visit	3.00
15	Last Notable Activity_Modified	2.56
1	TotalVisits	2.00
9	Last Activity_Olark Chat Conversation	1.99
2	Total Time Spent on Website	1.87
0	Do Not Email	1.86
8	Last Activity_Email Bounced	1.82
5	Lead Source_Olark Chat	1.69
13	Last Notable Activity_Email Opened	1.67
4	Lead Origin_Lead Add Form	1.43
16	Last Notable Activity_Olark Chat Conversation	1.36
7	Last Activity_Converted to Lead	1.25
6	Lead Source_Welingak Website	1.24
11	What is your current occupation_Working Profes	1.17
17	Last Notable Activity_Page Visited on Website	1.16
12	Last Notable Activity_Email Link Clicked	1.05
10	What is your current occupation_Housewife	1.01
14	Last Notable Activity_Had a Phone Conversation	1.01

ROC

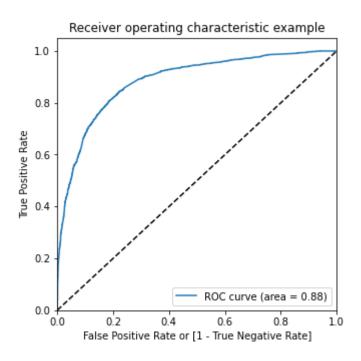


Assuming 0.3 as cutoff probability:

Sensitivity: 0.8381

Specificity: 0.7813

Accuracy: 0.8030

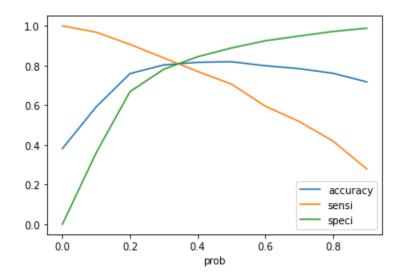


The Shape of the curve confirms that of a good model.





Plotting Sensitivity Vs Specificity Vs Accuracy for cutoff probability values between 0 and 1



Inference: From the curve above 0.35 is the optimum point to take it as cutoff probability.





Lead score is calculated for each customer in the training data set using the probability value of y.

	Converted	Converted_Prob	ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	8.0	0.9	final_predicted	Lead_Score
0	0	0.255123	1871	0	1	1	1	0	0	0	0	0	0	0	0	26
1	0	0.245922	6795	0	1	1	1	0	0	0	0	0	0	0	0	25
2	0	0.299014	3516	0	1	1	1	0	0	0	0	0	0	0	0	30
3	0	0.816244	8105	1	1	1	1	1	1	1	1	1	1	0	1	82
4	0	0.194206	3934	0	1	1	0	0	0	0	0	0	0	0	0	19

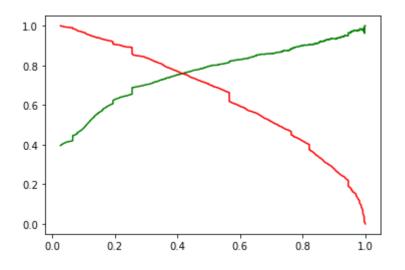
Precision and Recall



Precision Score = 0.7025

Recall Score = 0.8381

Plotting precision score Vs Recall score for various cutoff probabilities between 0 and 1



Inference: From the graph above the intersection point is 0.4 which is close to the value infered from ROC Curve.

Evaluation on the test set



The GLM model built using the training data is used to predict the target variable in the test data.

Accuracy = 0.8170

Sensitivity = 0.8036

Specificity = 0.8258

The model prediction on the test data is satisfactory based on the above Model evaluation

Recommendation



- > To increase conversion rate, the sales team should focus on customers who are predicted as hot leads by the model.
- > To improve the conversion rate the academy should also focus on the variables having high impact on the prediction.