Summary of Lead Score Case Study

Cleaning data:

In this data set lot of ‘Select’ variables in categorical data set. We had to remove them from the dataset.

We drop columns 'Lead Quality','Asymmetrique Activity Index','Asymmetrique Profile Score','Asymmetrique Activity Score','Asymmetrique Profile Index','Tags’ where null values are more than 40%.

In this data set, we drop column Country where more than 95% is India..

We drop City column where more than 40% is Mumbai and 28% is ‘Select’.

EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

Dummy Variables:

The dummy variables were created and later on the dummies with ‘not provided’ elements were removed. For numeric values we used the MinMaxScaler

Train-Test split:

The split was done at 70% and 30% for train and test data respectively

Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 78%, 77% and 76% respectively.

Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 78%.

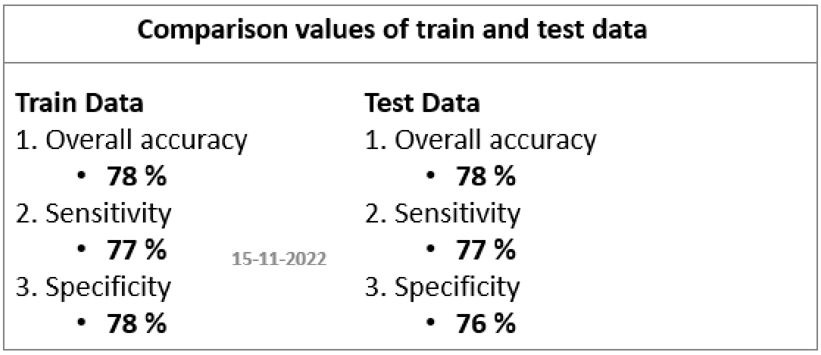
Precision – Recall:

This method was also used to recheck and a cut off of 0.44 was found with Precision around 77% and recall around 78% on the test data frame.

Learnings

- Existing conversion ratio is 38%
- According to lead origin 'landing page submission' worked well
- According to lead source 'Google ' management can spend more money on google ads compare to all

- According to last activity ' Email Opened and sms sent' has worked well
- According to the observation 'Newspaper and Newspaper article and Magazine' investing money in this is not a great idea.
- **After model building, we achieved lead conversion of 77%**



Comparison values of train and test data

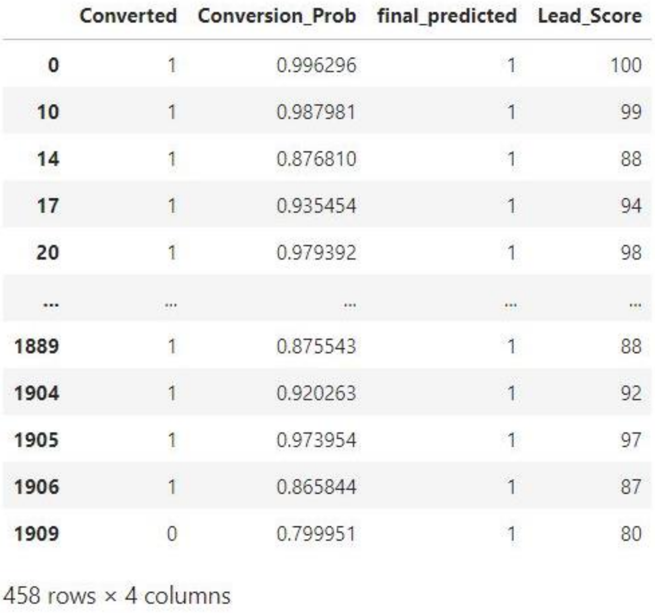
Train Data	**Test Data**
1. Overall accuracy • **78 %**	1. Overall accuracy • **78 %**
2. Sensitivity • **77 %**	2. Sensitivity • **77 %**
3. Specificity • **78 %**	3. Specificity • **76 %**

15-11-2022

Important Variables and VIF values in the model

TotalVisits	**11.14891**
Total Time Spent on Website	4.422291
Lead Origin_Lead Add Form	4.205123
Last Notable Activity_Unreachable	2.784594
Last Activity_Had a Phone Conversation	2.75522
Lead Source_Welingak Website	2.152559
Lead Source_Olark Chat	1.452589
Last Activity_SMS Sent	1.185594
const	0.204037
Do Not Email_Yes	-1.50368
What is your current occupation_Student	-2.35778
What is your current occupation_Unemployed	-2.54446

Hot Leads



	Converted	Conversion_Prob	final_predicted	Lead_Score
0	1	0.996296	1	100
10	1	0.987981	1	99
14	1	0.876810	1	88
17	1	0.935454	1	94
20	1	0.979392	1	98
...	...	...	...	...
1889	1	0.875543	1	88
1904	1	0.920263	1	92
1905	1	0.973954	1	97
1906	1	0.865844	1	87
1909	0	0.799951	1	80

458 rows × 4 columns