Lead Score Case Study

By

Kamalika Dutta

Rajashree Kuttisankaran

DS C29

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Analysis Approach

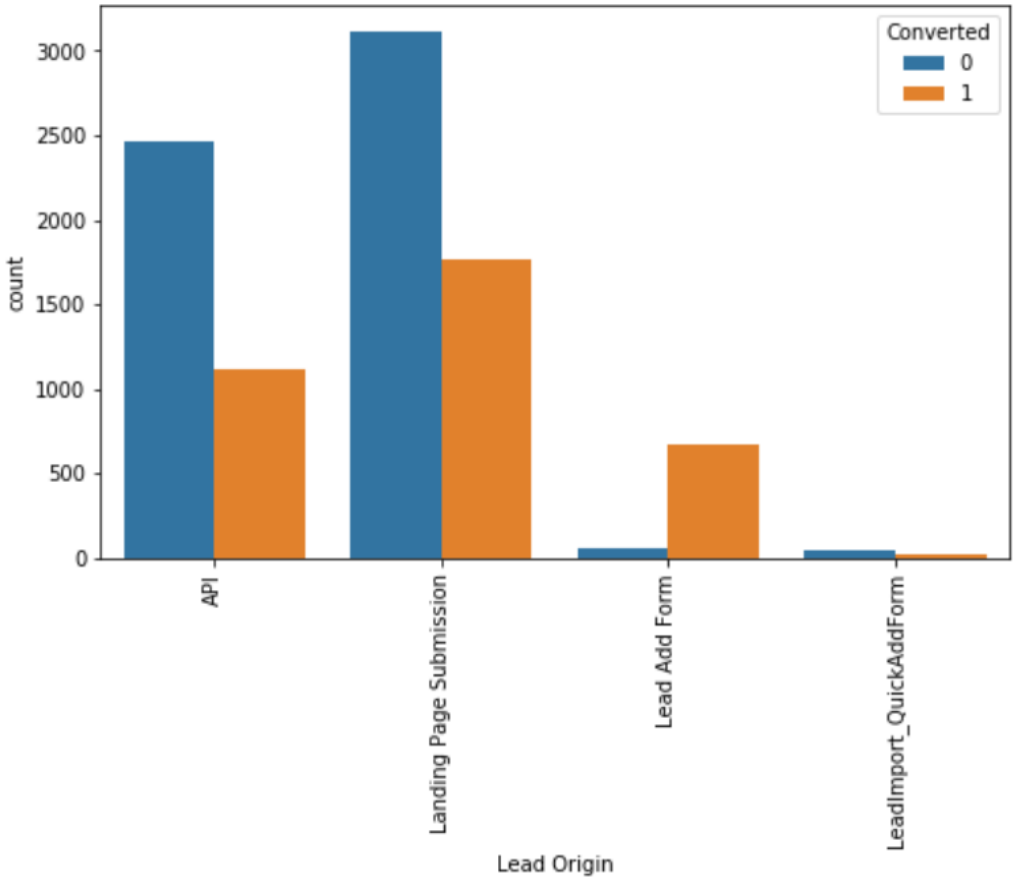
- Data cleaning steps like missing value treatments, deleting columns which are not relevant or having highly skewed data, Renamed few columns with lengthy names

- Converted all the binary variables to ‘0’ and ‘1’ and also multi class variables into dummy variables. There were many categorical variables where few of the categories have a very smaller number of records. Those were merged into one with name as ‘Other’.

- Outlier Treatment - Checked outliers of the dataset wherein few of the continuous variables were found to have it. Since outliers in logistic model is very sensitive so dealt with it by capping the dataset with 99^{th} percentile.

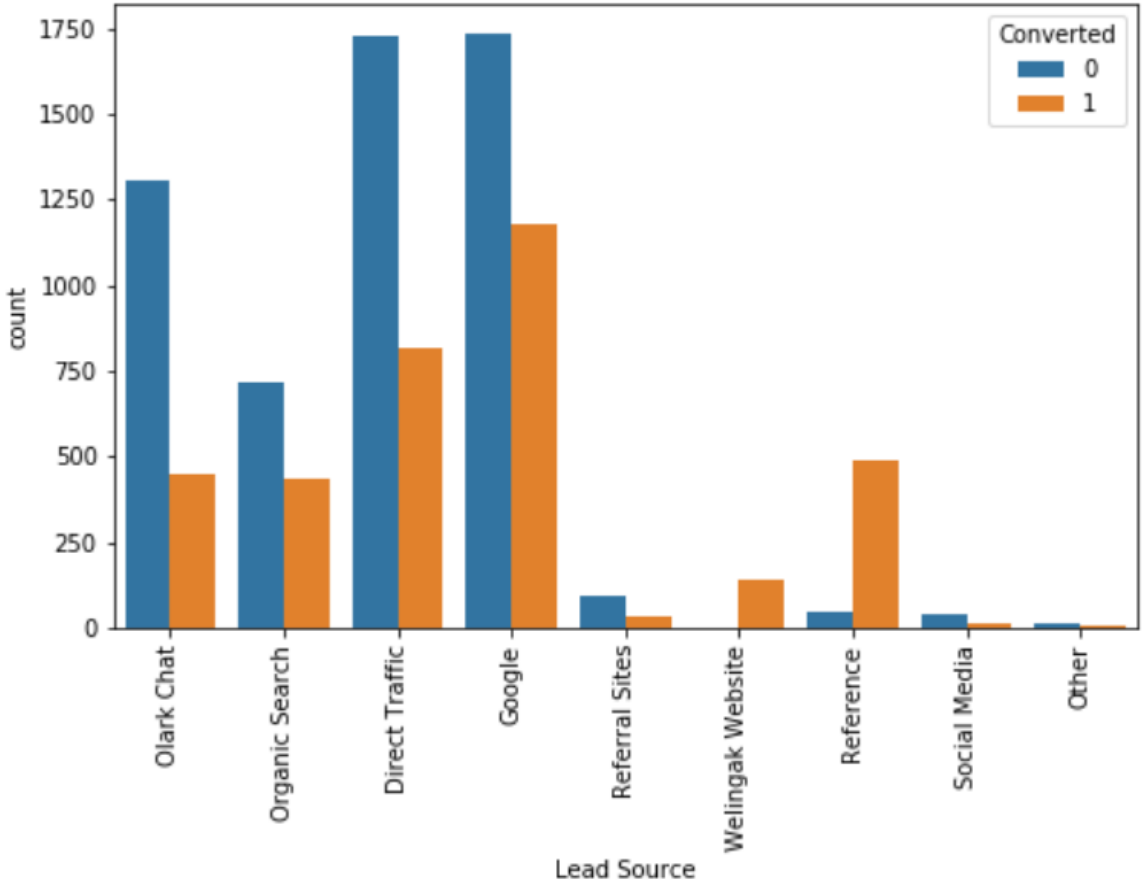
Visualization

1. API and Landing Page Submission has higher number of leads as well as conversion.
2. Lead Add Form has a very high conversion rate but count of leads are not very high.
3. Lead Import and Quick Add Form has got very few leads.



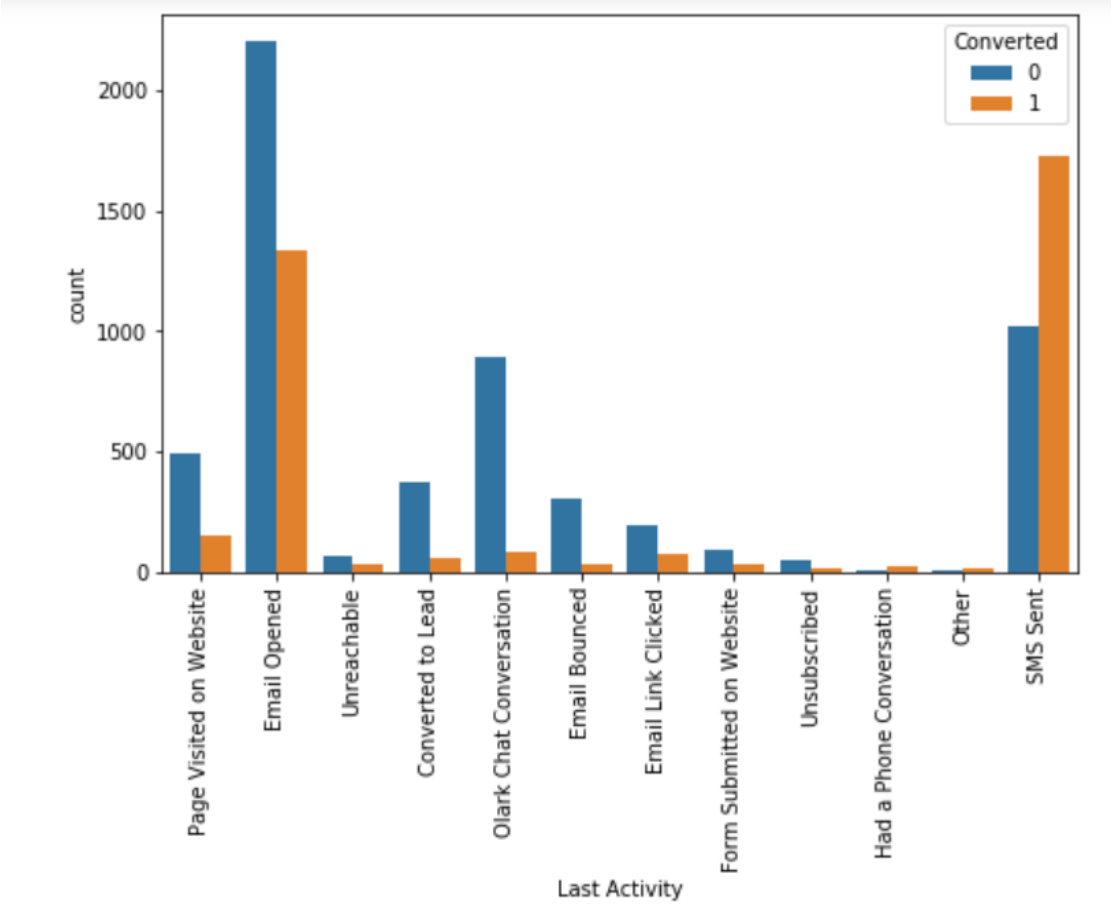
Visualization

1. Google and Direct Traffic has higher number of leads as well as good conversion.
2. Reference has a very high conversion rate but count of leads are not very high.
3. Social Media, Referral Sites and other sites has got very few leads.



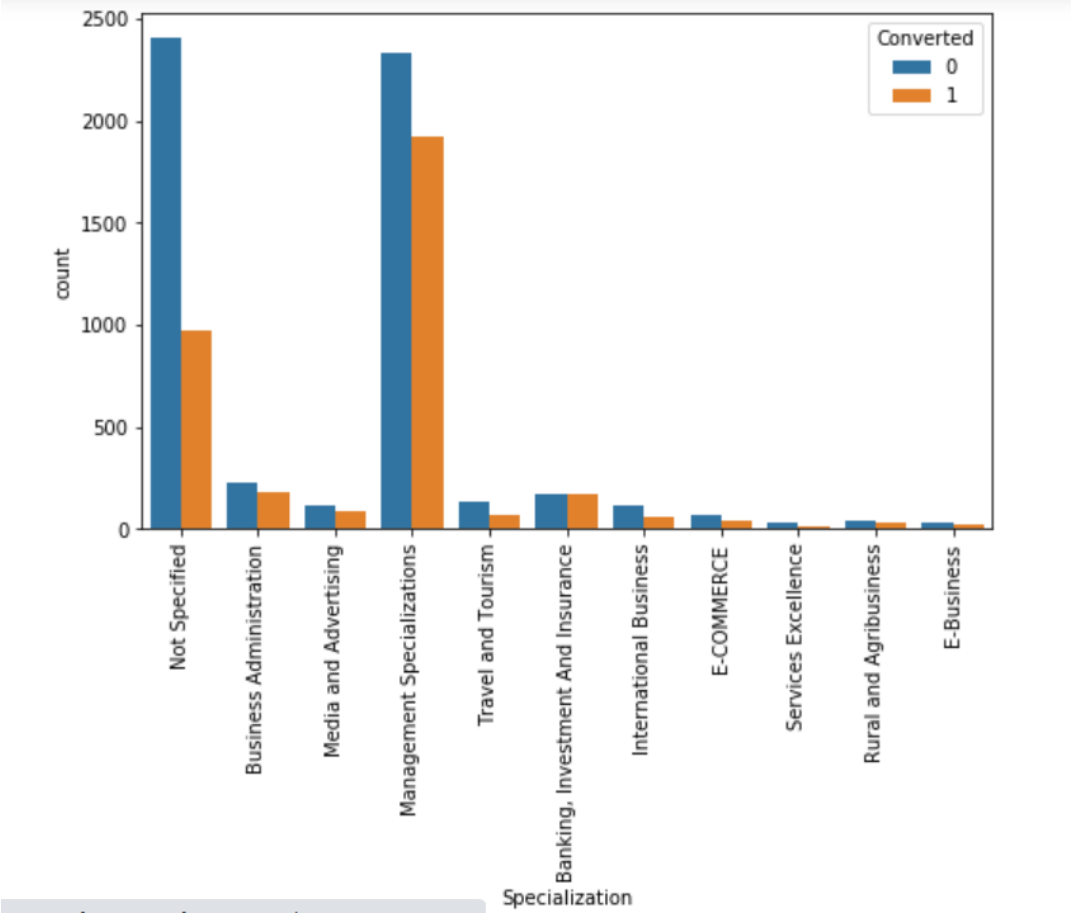
Visualization

Email Opened and SMS sent has higher number of leads as well as conversion. For SMS Sent conversion is very high.



Visualization

Management specialization has very good number of leads as well as conversion.



Analysis Approach

- After creating dummy variables , checked the correlation of the dataset. There were some highly correlated fields which was dropped. After dropping those highly correlated features, we plotted again a heatmap to check and it was confirmed that the highly correlated feature was dropped.

There were few left but did not drop to check them after creating model to know their impact.

- Split train and test data
- Feature scaling

Analysis Approach

- First we build a model with all the features included and found there were many insignificant features present in our model. To do so, we used RFE with count of variables 15.
- It was the first model after choosing features with RFE we got all of them with VIF < 5 and p value < 0.05.
- Evaluated the model by first predicting it.
- We created new datasets with original converted value and the prediction values.

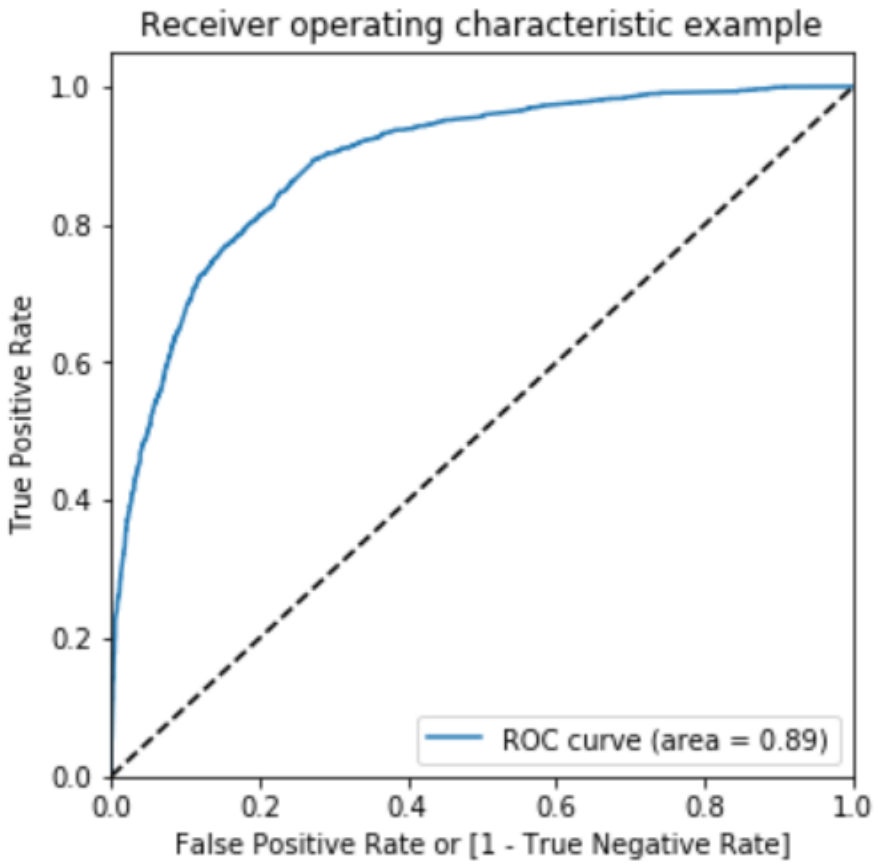
Final Model

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9068	0.157	-12.153	0.000	-2.214	-1.599
Do Not Email	-1.1268	0.181	-6.214	0.000	-1.482	-0.771
Total Time Spent on Website	1.0635	0.040	26.748	0.000	0.986	1.141
Lead Origin_Landing Page Submission	-1.0244	0.128	-7.981	0.000	-1.276	-0.773
Lead Origin_Lead Add Form	2.8748	0.203	14.178	0.000	2.477	3.272
Lead Source_Olark Chat	1.0375	0.121	8.592	0.000	0.801	1.274
Lead Source_Welingak Website	2.4569	0.752	3.269	0.001	0.984	3.930
Last Activity_Email Opened	0.9139	0.097	9.396	0.000	0.723	1.104
Last Activity_Had a Phone Conversation	2.8598	0.682	4.195	0.000	1.524	4.196
Last Activity_Other	1.9069	0.678	2.814	0.005	0.579	3.235
Last Activity_SMS Sent	2.0226	0.100	20.231	0.000	1.827	2.219
Last Activity_Unreachable	1.0177	0.361	2.821	0.005	0.311	1.725
Last Activity_Unsubscribed	1.4640	0.492	2.978	0.003	0.501	2.427
Specialization_Not Specified	-1.0135	0.123	-8.234	0.000	-1.255	-0.772
Occupation_Unemployed	0.9433	0.082	11.473	0.000	0.782	1.104
Occupation_Working Professional	3.3561	0.201	16.706	0.000	2.962	3.750

VIF

	Features	VIF
2	Lead Origin_Landing Page Submission	2.95
13	Occupation_Unemployed	2.75
12	Specialization_Not Specified	2.31
6	Last Activity_Email Opened	2.23
9	Last Activity_SMS Sent	2.12
4	Lead Source_Olark Chat	1.88
3	Lead Origin_Lead Add Form	1.69
14	Occupation_Working Professional	1.36
1	Total Time Spent on Website	1.26
5	Lead Source_Welingak Website	1.26
0	Do Not Email	1.24
11	Last Activity_Unsubscribed	1.07
10	Last Activity_Unreachable	1.03
7	Last Activity_Had a Phone Conversation	1.02
8	Last Activity_Other	1.01

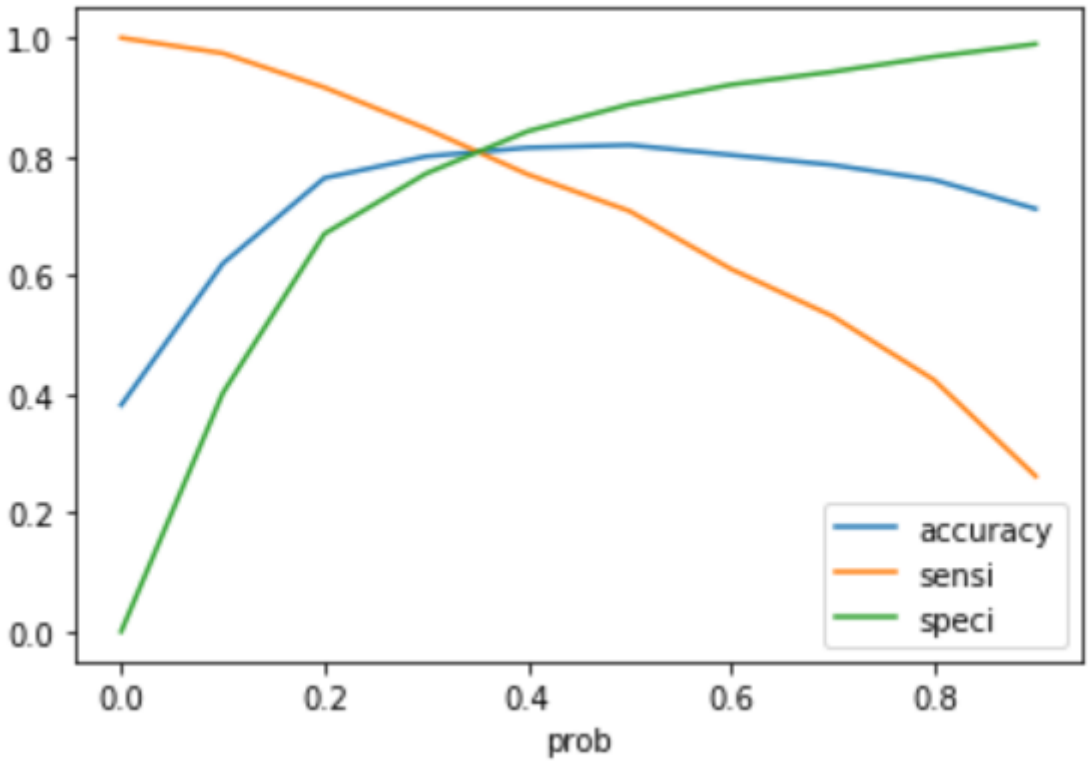
Model Evaluation



Model Evaluation

- As shown in the previous slide, after building the final model we calculated various metrics like sensitivity, specificity etc. using confusion matrix.
- We created ROC curve to find the model stability with AUC score. The area under curve we got as 0.89 which is a very good score.
- Our graph is leaned towards left side of the dotted line which shows that it has great accuracy.

Finding Optimal cut off point



Finding Optimal cut off point

- We created range of points for which we found accuracy, sensitivity and specificity for each one of them to choose for probability cut off.
- We found that on 0.4 all the three parameters are almost in same range and hence we selected the same.
- To verify we plotted line plot.
- After getting the optimum cut off point we again re-calculated sensitivity accuracy etc.

Precision and Recall

- We used this optimum cut off to create a new column for predicting the outcomes.
- We did evaluation through checking precision and recall score.

```
Precision Score 0.7508896797153025
Recall Score 0.7700729927007299
```

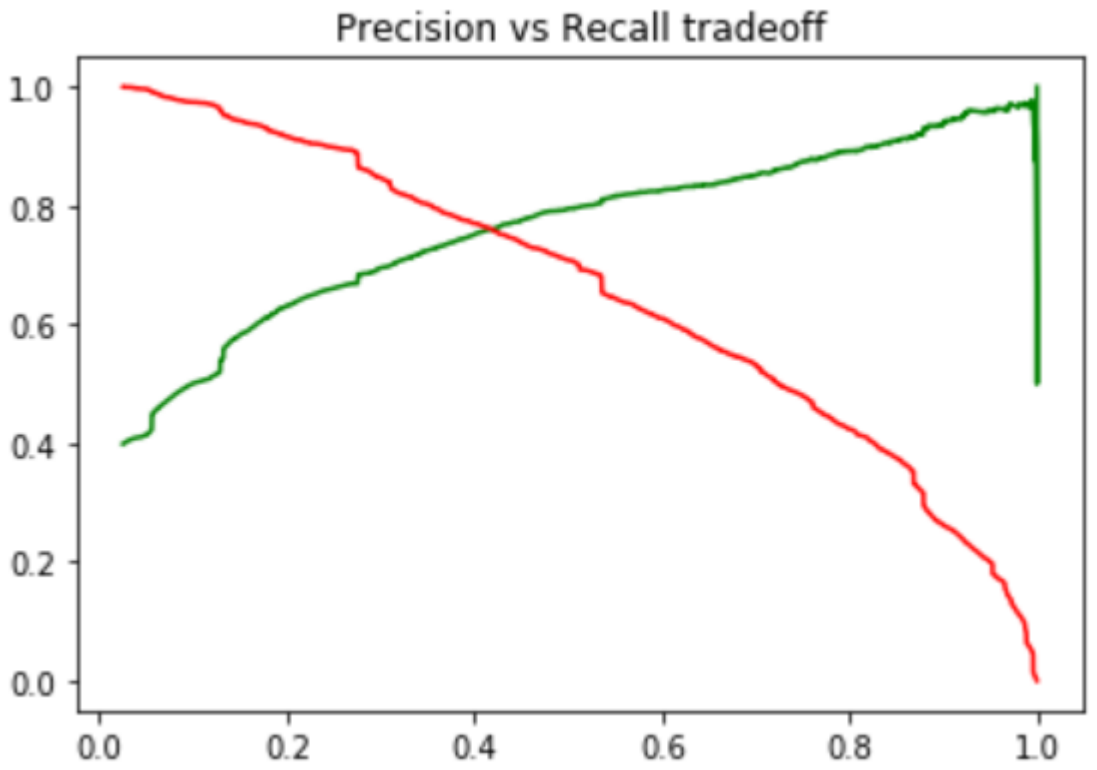
Important point to be noted from the outcomes for precision and recall score -

Our precison percentage is 75% approximately and recall percentage is 77% This means we have very good model which explains relevancy of 75% and true relevant results about 77%.

It is good to consider the recall percentage more valuable than precision because it is okay if precision is a bit low which means less hot lead customers it is not good to leave out any hot leads who are willing to get converted

Precision and Recall trade off

We created a graph to show the trade off between precision and recall. We found that there is a trade off and the meeting point is somewhere after 0.4.



Prediction on Test set

- Before prediction on test set, we scaled the test set using transform method.
- After that we predicted the test data and the new predicted values were saved in a data frame . Again we did model evaluation and found the accuracy score was 81 %.
- Lead score was created on test datasets to identify hot leads. Higher the lead score higher the chance of converted ,lower the lead score lower the chance of converted.
- The below shows the final results on both train and test set. This shows that our test prediction is having accuracy and other parameters in acceptable ranges.

Final Observation: Let us compare the values obtained for Train & Test:

Train Data:

Accuracy : 81.49% Sensitivity : 70.84% Specificity : 88.78%

Test Data:

Accuracy : 81.63% Sensitivity : 76.71% Specificity : 84.85%

Conclusions

- All the metrics accuracy, precision and recall/sensitivity are showing very good results in test set.
- Recall is having higher score than precision which is acceptable from business perspective.
- The model is in stable state.

Important Features

- Occupation_Working Professional
- Last Activity_Had a Phone Conversation
- Lead Origin_Lead Add Form

Thank You