LEADS SCORING CASE STUDY SUBMISSION

Submitted By:

Barkha Garg

Suryateja Rallapalli

Objective

Business Context:

The objective of this case study to identify the potential leads that are likely to convert into paying customers for the company named X Education. This company sells online courses to industry professionals.

The company markets its courses on several websites and search engines like Google. Once the people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When they fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Objective:

This conversion rate is not so impressive. The aim of this case study is to identify the most potential leads, also known as ‘Hot Leads’. If the leads are identified successfully, the conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

We will build a model to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance. The target is to achieve a conversion rate of 80%.

Approach

Approach:

The analysis was carried out using the leads scoring data provided. The language used to analyse and visualize is Python and the scripting tool used is Jupyter Notebook.

The entire process was divided into sequence of steps (individual tasks) with output of tasks helping us infer some useful insights about the problem and also look at them visually for a better understanding. These steps included

a) Reading & Cleaning the Data – Dropping high nulls count columns, imputing missing values, Removing Outliers

b) Dummy Values Creation – Dummy values were created for categorical data so that the dataset has all the features/variables in numeric format

c) Feature Scaling (Standardisation) – This was done to get all variables on same scale and have an unbiased data

d) Applying the technique of RFE to reduce the number of features recursively and pick only the features that are not multi collinear and give a model with a good accuracy score

e) Building the Logistics Model and predict the output for test data

f) Assessing the ROC curve and the accuracy, sensitivity and specificity to get the optimal cut off and using that to predict the final output

At the end of these tasks/steps, some useful inferences were drawn and finally arrive at identifying the key factors to maximize Leads conversion.

Data Cleaning

Data Loading & Data Cleaning

Step 1: Loading the Leads data (Leads.csv)

a) The request data was loaded in Data Frame**.**

b) The total count of records in loaded data frame was 9240

Step 2: Cleaning the Data

a) There were a lot of columns for which the % of null values was > 30. These columns were dropped as this is quite a high % for null values. (Tags, Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score)

b) There were few columns which had nulls between 15% and 30%, but they had a lot of missing values (value 'Select') which means this information is not available which is as good as null. And the total % of nulls + this value 'Select' was over 30. Such columns were dropped too.

City, Lead Profile, Specialization, How did you hear about X Education

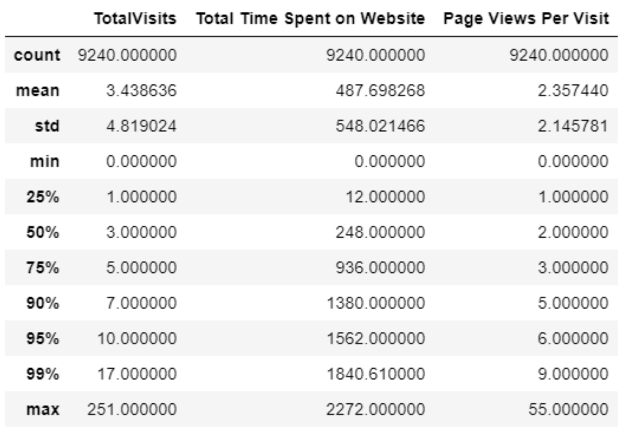
c) Lead Source, TotalVisits, Page Views Per Visit and Last Activity have a very small number of nulls (<2%). So these were imputed wit the mean value of that column (for numeric variable) of the most frequent value for that column (for categorical variable)

d) Variables Update me on Supply Chain Content, Receive More Updates About Our Courses, Get updates on DM Content, I agree to pay the amount through cheque, Magazine had only one value. These would not add any variation to the data, so these columns were also dropped

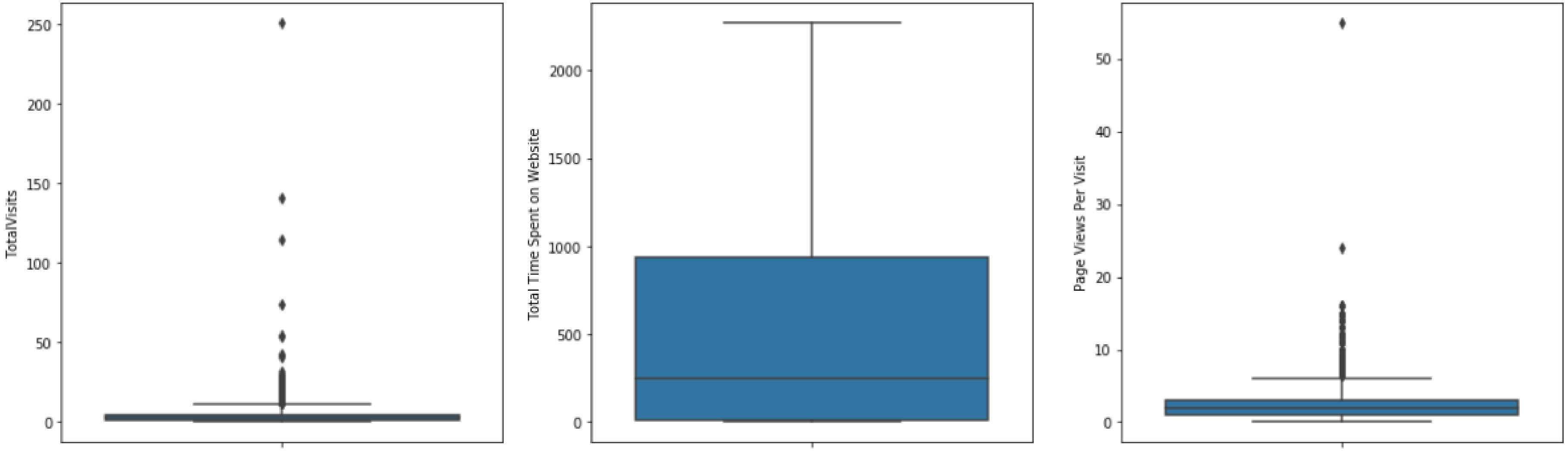
e) The continuous variable TotalVisits, Total Time Spent on Website and Page Views Per Visit had presence of some outliers. These outliers were removed and we still had 80% of the data to pursue the analysis.

Data Cleaning (Cntd..)

Data with Outliers



	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9240.000000	9240.000000	9240.000000
mean	3.438636	487.698268	2.357440
std	4.819024	548.021466	2.145781
min	0.000000	0.000000	0.000000
25%	1.000000	12.000000	1.000000
50%	3.000000	248.000000	2.000000
75%	5.000000	936.000000	3.000000
90%	7.000000	1380.000000	5.000000
95%	10.000000	1562.000000	6.000000
99%	17.000000	1840.610000	9.000000
max	251.000000	2272.000000	55.000000

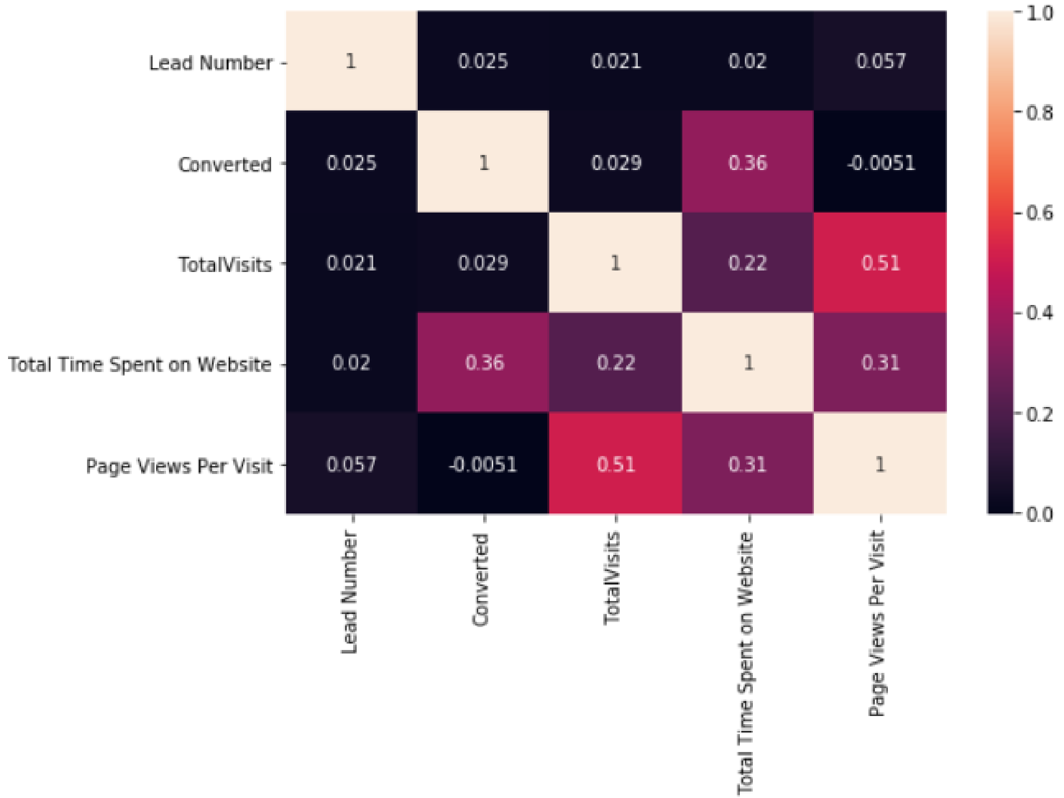


Data Cleaning (Cntd..)

Since the people may visit and leave the website any number if times, we will NOT remove the outliers and go ahead with original data.

Corelation Heatmap

Not much correlation was seen among variables in original dataset, so the continuous variables were not dropped



Feature Scaling

Step 1: The variables that had only 2 values Yes or No, were mapped to Numeric binary values 0 & 1 (Do Not Email, Newspaper Article, Do Not Call, Search, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, A free copy of Mastering The Interview)

Dummy Variable Creation

Step 2: For Categorical Variables (viz., Lead Origin, Lead Source, Last Activity, Last Notable Activity) that had more than 2 unique values , were processed to generate dummy columns from the values

After above steps, all variables we have in data came to numeric datatype.

Feature Scaling:

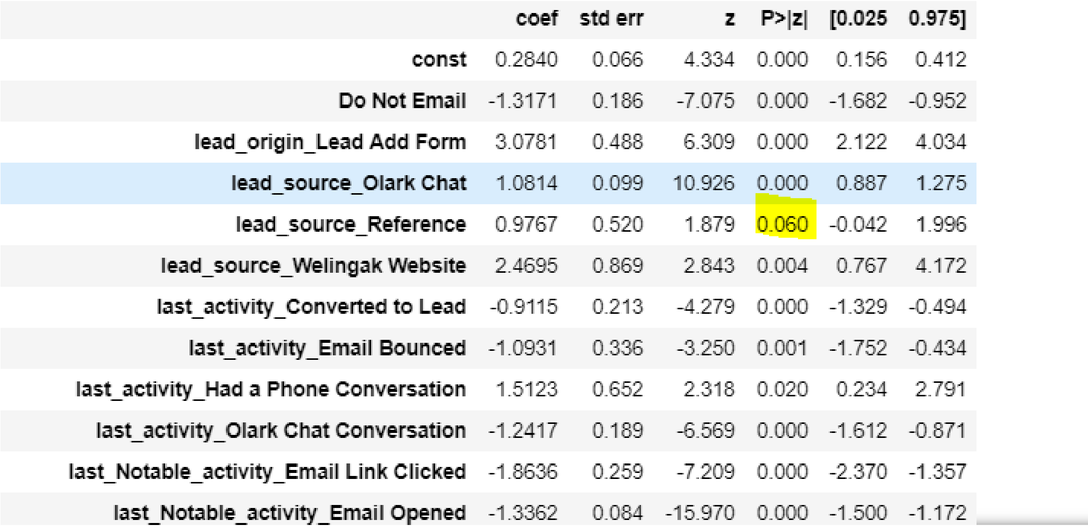
Step 3: The continuous variables TotalVisits, Total Time Spent on Website, Page Views Per Visit had values which are higher in magnitude than other numeric influence. To suppress the influence of biasness from these values these variables were standardized (using standard deviation).

RFE (Recursive Feature Elimination)

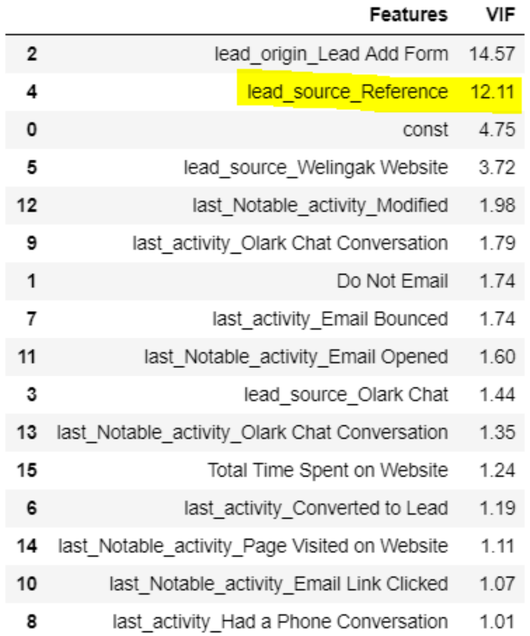
- There were 60+ attributes in the dataset after dummy variable creation and feature scaling.

- There were only around 18 variables in the cleaned data before Dummy variable creation. So, the RFE we began was with 15 variables in the first Model

Model 1:



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2840	0.066	4.334	0.000	0.156	0.412
Do Not Email	-1.3171	0.186	-7.075	0.000	-1.682	-0.952
lead_origin_Lead Add Form	3.0781	0.488	6.309	0.000	2.122	4.034
lead_source_Olark Chat	1.0814	0.099	10.926	0.000	0.887	1.275
lead_source_Reference	0.9767	0.520	1.879	0.060	-0.042	1.996
lead_source_Welingak Website	2.4695	0.869	2.843	0.004	0.767	4.172
last_activity_Converted to Lead	-0.9115	0.213	-4.279	0.000	-1.329	-0.494
last_activity_Email Bounced	-1.0931	0.336	-3.250	0.001	-1.752	-0.434
last_activity_Had a Phone Conversation	1.5123	0.652	2.318	0.020	0.234	2.791
last_activity_Olark Chat Conversation	-1.2417	0.189	-6.569	0.000	-1.612	-0.871
last_Notable_activity_Email Link Clicked	-1.8636	0.259	-7.209	0.000	-2.370	-1.357
last_Notable_activity_Email Opened	-1.3362	0.084	-15.970	0.000	-1.500	-1.172



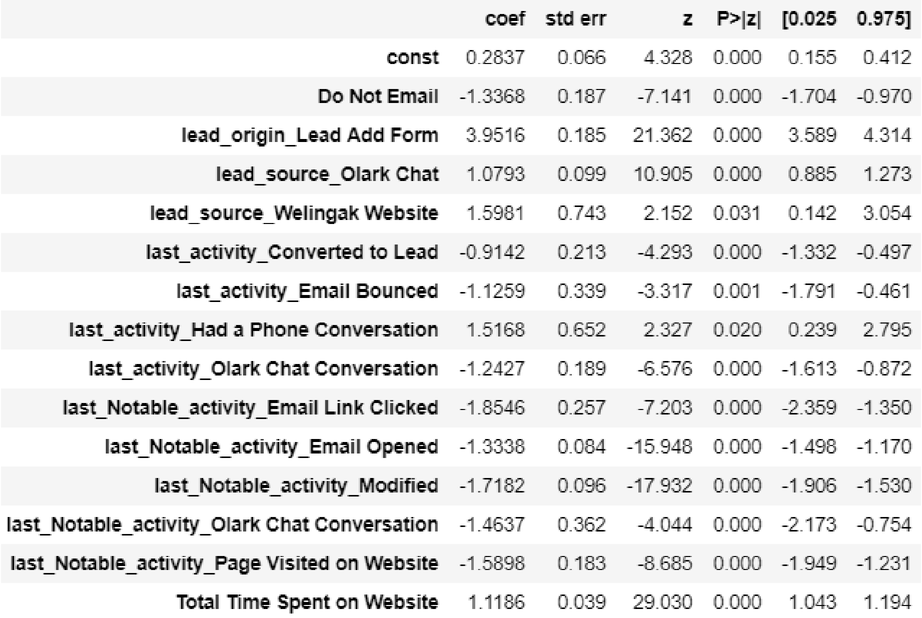
	Features	VIF
2	lead_origin_Lead Add Form	14.57
4	lead_source_Reference	12.11
0	const	4.75
5	lead_source_Welingak Website	3.72
12	last_Notable_activity_Modified	1.98
9	last_activity_Olark Chat Conversation	1.79
1	Do Not Email	1.74
7	last_activity_Email Bounced	1.74
11	last_Notable_activity_Email Opened	1.60
3	lead_source_Olark Chat	1.44
13	last_Notable_activity_Olark Chat Conversation	1.35
15	Total Time Spent on Website	1.24
6	last_activity_Converted to Lead	1.19
14	last_Notable_activity_Page Visited on Website	1.11
10	last_Notable_activity_Email Link Clicked	1.07
8	last_activity_Had a Phone Conversation	1.01

This model gave a good accuracy score of 80%

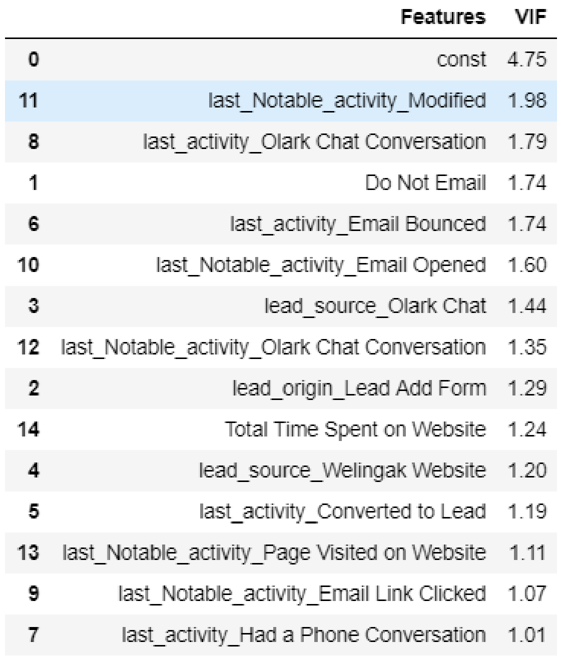
But as we see, there are variables with VIF > 5 in this model. So this is not the optimal model. Also, the variable lead_source_Reference has a high p-value and a high VIF score. So this variable was dropped and a new model was built.

RFE (Cntd..)

Model 2:



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2837	0.066	4.328	0.000	0.155	0.412
Do Not Email	-1.3368	0.187	-7.141	0.000	-1.704	-0.970
lead_origin_Lead Add Form	3.9516	0.185	21.362	0.000	3.589	4.314
lead_source_Olark Chat	1.0793	0.099	10.905	0.000	0.885	1.273
lead_source_Welingak Website	1.5981	0.743	2.152	0.031	0.142	3.054
last_activity_Converted to Lead	-0.9142	0.213	-4.293	0.000	-1.332	-0.497
last_activity_Email Bounced	-1.1259	0.339	-3.317	0.001	-1.791	-0.461
last_activity_Had a Phone Conversation	1.5168	0.652	2.327	0.020	0.239	2.795
last_activity_Olark Chat Conversation	-1.2427	0.189	-6.576	0.000	-1.613	-0.872
last_Notable_activity_Email Link Clicked	-1.8546	0.257	-7.203	0.000	-2.359	-1.350
last_Notable_activity_Email Opened	-1.3338	0.084	-15.948	0.000	-1.498	-1.170
last_Notable_activity_Modified	-1.7182	0.096	-17.932	0.000	-1.906	-1.530
last_Notable_activity_Olark Chat Conversation	-1.4637	0.362	-4.044	0.000	-2.173	-0.754
last_Notable_activity_Page Visited on Website	-1.5898	0.183	-8.685	0.000	-1.949	-1.231
Total Time Spent on Website	1.1186	0.039	29.030	0.000	1.043	1.194



	Features	VIF
0	const	4.75
11	last_Notable_activity_Modified	1.98
8	last_activity_Olark Chat Conversation	1.79
1	Do Not Email	1.74
6	last_activity_Email Bounced	1.74
10	last_Notable_activity_Email Opened	1.60
3	lead_source_Olark Chat	1.44
12	last_Notable_activity_Olark Chat Conversation	1.35
2	lead_origin_Lead Add Form	1.29
14	Total Time Spent on Website	1.24
4	lead_source_Welingak Website	1.20
5	last_activity_Converted to Lead	1.19
13	last_Notable_activity_Page Visited on Website	1.11
9	last_Notable_activity_Email Link Clicked	1.07
7	last_activity_Had a Phone Conversation	1.01

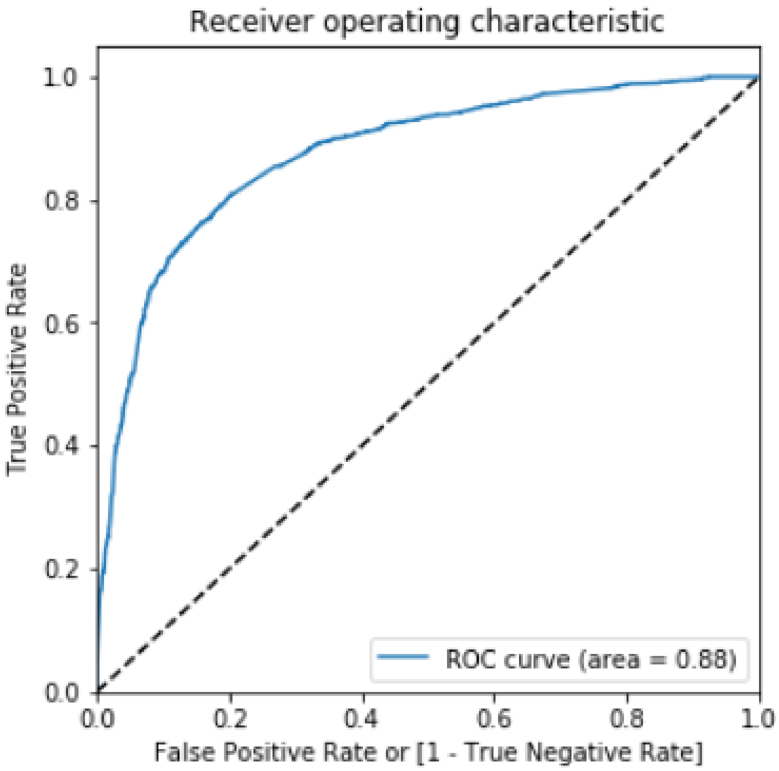
The accuracy with this model was also 80% which is the same as Model 1

And all variables have a VIF under 5 and p-value less than 0.05. So this model was accepted as the optimal model with 14 variables

ROC Curve

ROC Curve:

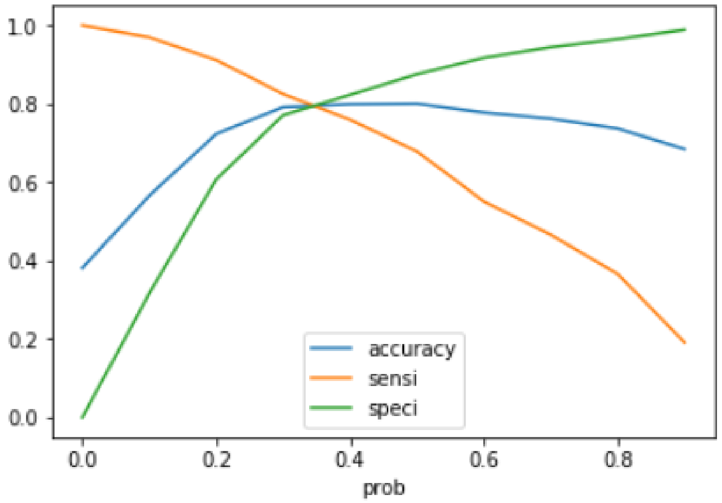
- For the model built, we plotted the ROC curve to assess the tradeoff between True Positives vs False Positives. The area under the curve was good with the curve inclining towards the top left corner.
- This appears quite promising.



Optimal Cut-Off (Accuracy, Sensitivity & Specificity Trade - Off)

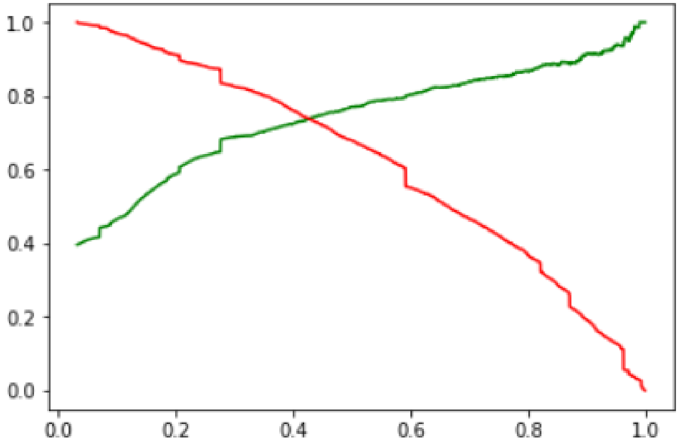
Accuracy, Sensitivity & Specificity Trade Off:

- For the model built, we also plotted the curve to assess the tradeoff between Accuracy, Sensitivity and Specificity.
- The intersection of the 3 values showed at around 0.35
- So 0.3 was chosen as the optimal cut-off for prediction of output on our test data



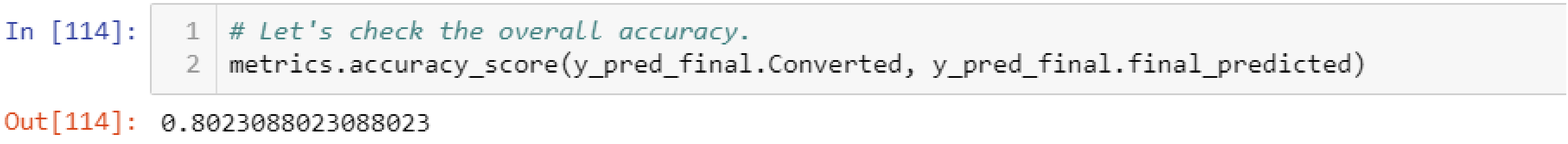
Precision-Recall Tradeoff

- This looked somewhat close to what we got in the previous graph



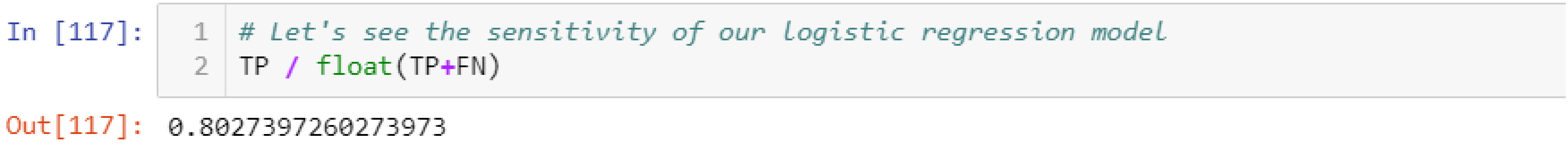
Test Output

➢ The accuracy-score, sensitivity & specificity of our model was close to 80% which is what was desirable. Hence the model build is good and can be used for predictions



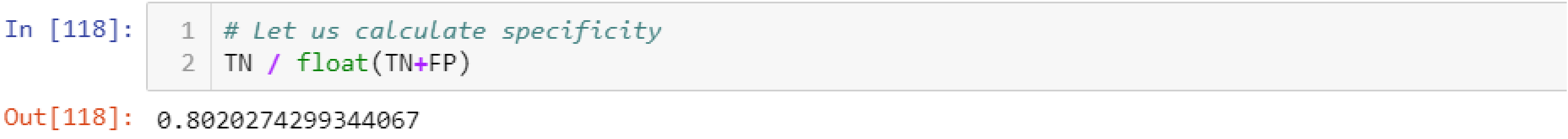
```
In [114]:  # Let's check the overall accuracy.
           metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)

Out[114]:  0.8023088023088023
```



```
In [117]:  # Let's see the sensitivity of our logistic regression model
           TP / float(TP+FN)

Out[117]:  0.8027397260273973
```



```
In [118]:  # Let us calculate specificity
           TN / float(TN+FP)

Out[118]:  0.8020274299344067
```

Conclusion

Desired Lead Score

As we saw, the cut-off probability to classify a Lead as Hot Lead came out to be 0.35.

So X Education can consider a lead with a lead score or 35 or more as a Hot Lead and may attempt to convert those to paying customers.

Top Parameters

On sorting the parameters built from the model, we get the output as shown:

- Top 3 variables are:

    i) Lead origin

    ii) Lead Source

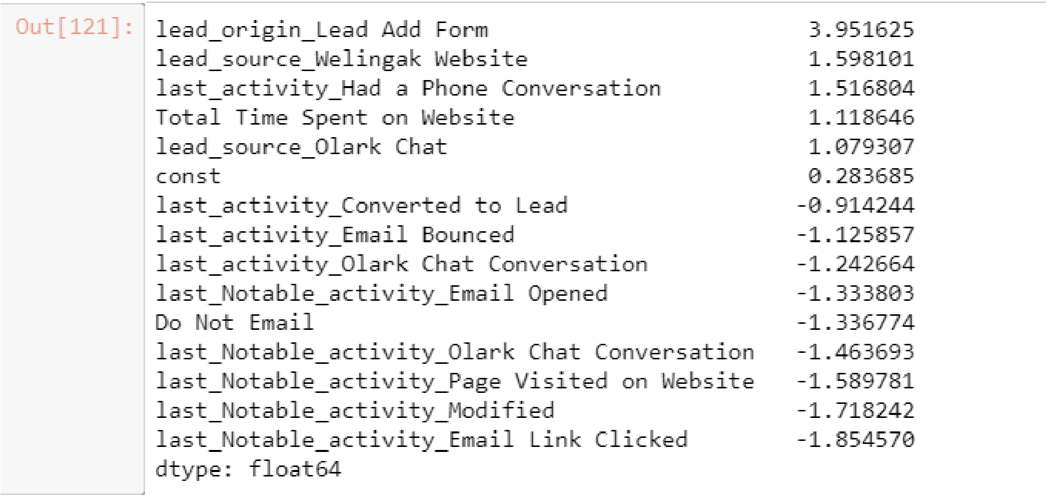
    iii) Last Activity

- Top 3 categorical/dummy variables

    i) lead_origin_Lead Add Form

    ii) lead_source_Welingak Website

    iii) last_activity_Had a Phone Conversation



```
Out[121]: lead_origin_Lead Add Form                     3.951625
          lead_source_Welingak Website                  1.598101
          last_activity_Had a Phone Conversation        1.516804
          Total Time Spent on Website                   1.118646
          lead_source_Olark Chat                        1.079307
          const                                         0.283685
          last_activity_Converted to Lead              -0.914244
          last_activity_Email Bounced                  -1.125857
          last_activity_Olark Chat Conversation        -1.242664
          last_Notable_activity_Email Opened           -1.333803
          Do Not Email                                 -1.336774
          last_Notable_activity_Olark Chat Conversation  -1.463693
          last_Notable_activity_Page Visited on Website  -1.589781
          last_Notable_activity_Modified               -1.718242
          last_Notable_activity_Email Link Clicked     -1.854570
          dtype: float64
```