LEAD SCORING CASE STUDY

Problem Statement and Business Goal

- An education company named X Education sells online courses to industry professionals.
- 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%.

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, that is, it is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps followed:

1. Reading and understanding the data
2. Cleaning the data
3. Preparing the data for Model Building
4. Model Building
5. Model Evaluation
6. Making Predictions on the Test Set

Dataset Summary

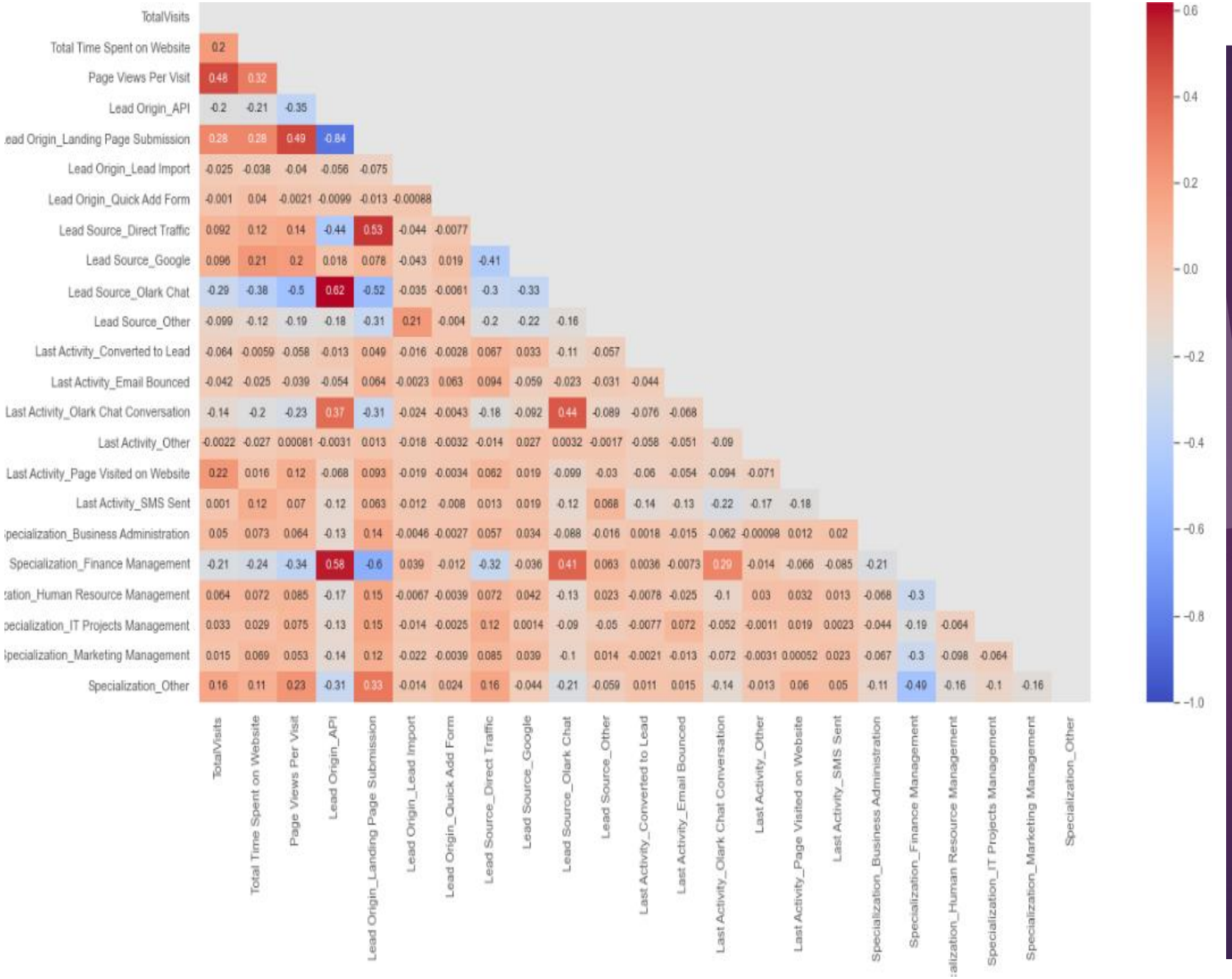
	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Dataset status after cleaning of data

	Lead Origin	Lead Source	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	A free copy of Mastering The Interview	Last Notable Activity
0	API	Olark Chat	No	0	0.0	0	0.0	Page Visited on Website	Select	Unemployed	No	Modified
1	API	Organic Search	No	0	5.0	674	2.5	Email Opened	Select	Unemployed	No	Email Opened
2	Landing Page Submission	Direct Traffic	No	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Yes	Email Opened
3	Landing Page Submission	Direct Traffic	No	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	No	Modified
4	Landing Page Submission	Google	No	1	2.0	1428	1.0	Converted to Lead	Select	Unemployed	No	Modified

Performing the scaling

Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Lead Origin_API	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Origin_Quick Add Form	...	Last Activity_Other	La Activity_Pag Visited c Websi
660737	0	0.0	0	0.0	0	1	0	0	0	0	...	0	
660728	0	5.0	674	2.5	0	1	0	0	0	0	...	0	
660727	1	2.0	1532	2.0	1	0	1	0	0	0	...	0	
660719	0	1.0	305	1.0	0	0	1	0	0	0	...	1	
660681	1	2.0	1428	1.0	0	0	1	0	0	0	...	0	



Correlation Table between variables

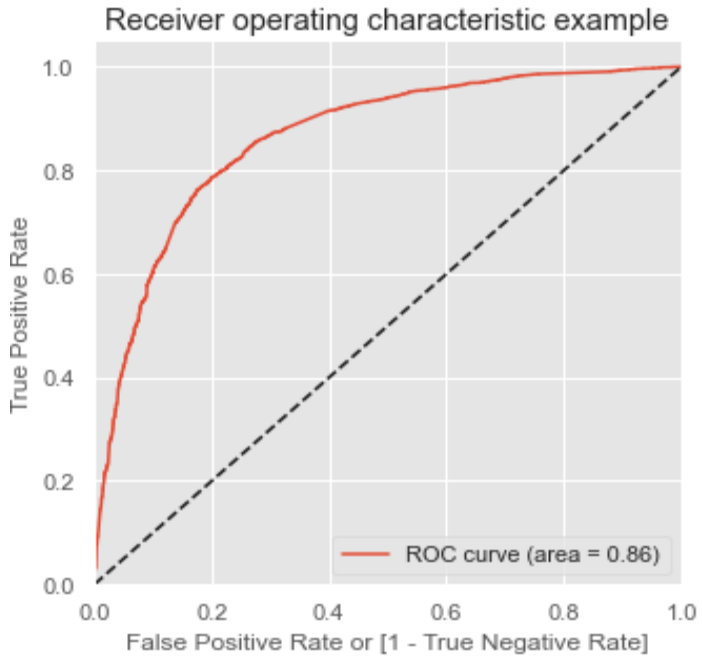
Generalized Linear Model Regression Results

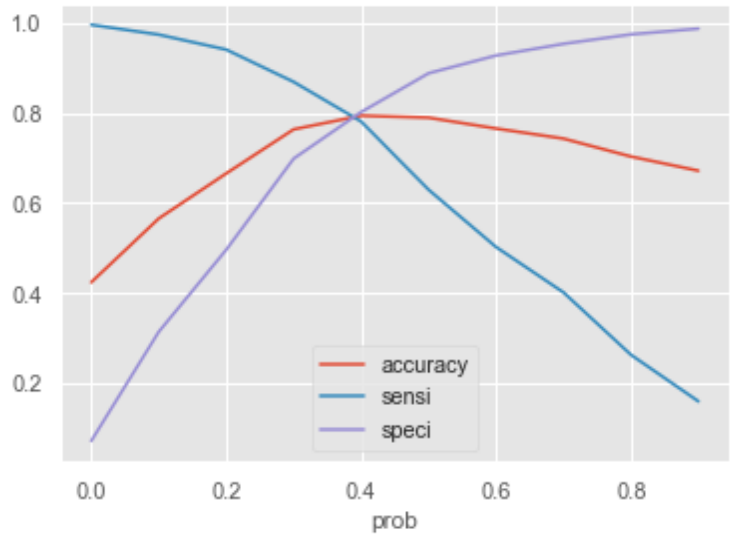
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6447
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2906.3
Date:	Sat, 12 Nov 2022	**Deviance:**	5812.7
Time:	11:30:53	**Pearson chi2:**	6.69e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.1558	0.319	9.901	0.000	2.531	3.781
TotalVisits	0.1858	0.056	3.316	0.001	0.076	0.296
Total Time Spent on Website	1.1059	0.038	28.764	0.000	1.031	1.181
Page Views Per Visit	-0.1511	0.051	-2.938	0.003	-0.252	-0.050
Lead Origin_API	-3.7302	0.301	-12.395	0.000	-4.320	-3.140
Lead Origin_Landing Page Submission	-4.0368	0.310	-13.019	0.000	-4.645	-3.429
Lead Origin_Lead Import	-3.7940	0.500	-7.594	0.000	-4.773	-2.815
Lead Source_Direct Traffic	-0.2438	0.119	-2.046	0.041	-0.477	-0.010
Lead Source_Google	0.1157	0.110	1.056	0.291	-0.099	0.330
Lead Source_Olark Chat	1.0167	0.158	6.434	0.000	0.707	1.326
Lead Source_Other	0.0735	0.259	0.283	0.777	-0.435	0.582
Last Activity_Converted to Lead	-1.0793	0.200	-5.396	0.000	-1.471	-0.687
Last Activity_Email Bounced	-1.8659	0.280	-6.658	0.000	-2.415	-1.317
Last Activity_Olark Chat Conversation	-1.7821	0.168	-10.610	0.000	-2.111	-1.453
Last Activity_Other	-0.2684	0.137	-1.963	0.050	-0.536	-0.000
Last Activity_Page Visited on Website	-0.7414	0.148	-5.023	0.000	-1.031	-0.452
Last Activity_SMS Sent	1.0878	0.075	14.536	0.000	0.941	1.234
Specialization_Business Administration	-0.2308	0.165	-1.397	0.163	-0.555	0.093
Specialization_Finance Management	-0.5756	0.102	-5.618	0.000	-0.776	-0.375
Specialization_Human Resource Management	-0.1508	0.129	-1.171	0.241	-0.403	0.102
Specialization_Other	-0.1518	0.103	-1.481	0.139	-0.353	0.049

Generalized Linear Model Regression Results



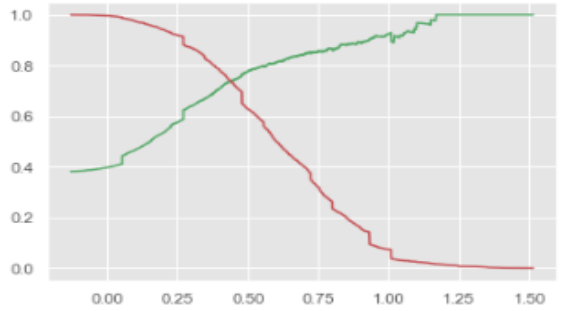


	Converted	Converted_Prob	Lead Number	Converted_pred	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted
0	0	0.176616	609431	0	1	1	0	0	0	0	0	0	0	0	0
1	0	0.053991	631817	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0.016089	596164	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0.181098	646570	0	1	1	0	0	0	0	0	0	0	0	0
4	0	0.028289	643974	0	1	0	0	0	0	0	0	0	0	0	0

Making Predictions on Test Set

ACCURACY, PRECISION, RECALL

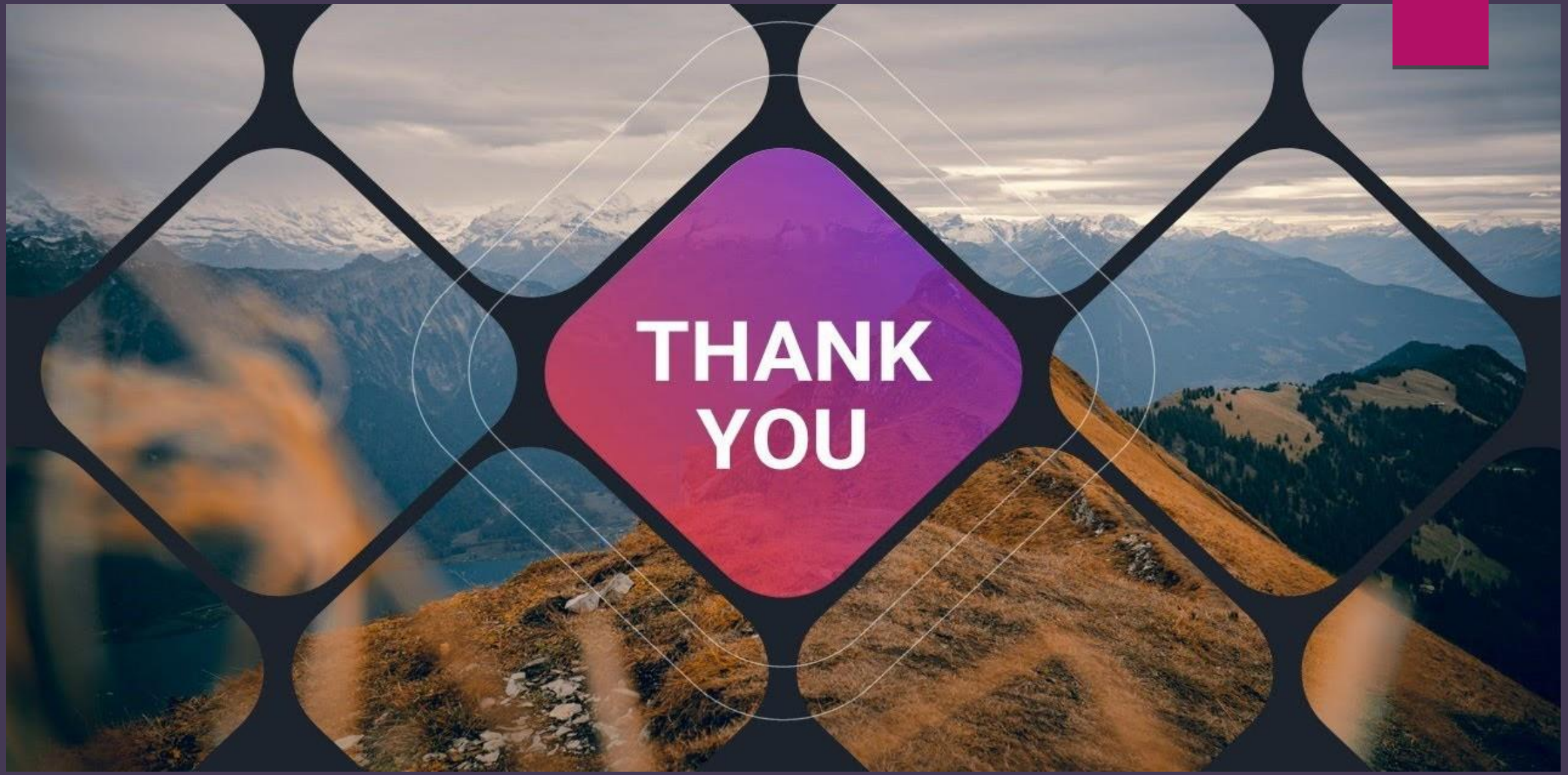
```
In [112]: # Printing the Metrics Accuracy, Sensitivity, Specicity

          acc,sensi,speci=metrices_(y_train_pred_final.Converted,y_train_pred_final.final_predicted)
          print('Accuracy: {}, Sensitivity {}, specifitiy {}  '.format(acc,sensi,speci))

          Accuracy: 0.7857142857142857, Sensitivity 0.8102189781021898, specifitiy 0.7706146926536732

In [113]: confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.Converted_pred )
          confusion

Out[113]: array([[3556,  446],
                 [ 913, 1553]], dtype=int64)

In [114]: # Finding the Precision Score

          precision_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)

Out[114]: 0.6851851851851852

In [115]: # Finding the Recall Score

          recall_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)

Out[115]: 0.8102189781021898

In [116]: p, r, thresholds = precision_recall_curve(y_train_pred_final.Converted, y_train_pred_final.Converted_Prob)

In [117]: plt.plot(thresholds, p[:-1], "g-",label='Precision')
          plt.plot(thresholds, r[:-1], "r-",label='Recall')
          # plt.savefig('precision-recall_curve',dpi=300,transparent=True)
          plt.show()
```



- Here we got 0.37 as the Cut-off as Precesion-Recall Threshold.

Summary

- During the initial stage (top), a lot of leads are generated, but from the bottom, only a few become paying customers.
- The middle stage involves nurturing the leads well (i.e. educating the leads about the product, communicating constantly, etc.) in order to get a higher conversion rate.
- These factors contribute most towards the probability of a lead getting converted: the 'Total Visits', 'Total Time Spent on Website', and the 'Page View Per Visit'.
- Providing job offers, information, or courses that are relevant to the interests of the leads and Charting the needs of each lead will go a long way toward converting leads into prospects.
- Need to focus on converted leads. Hold question-and-answer sessions with leads to extract accurate details about them. Discern the leads' intentions and mentality towards enrolling in online courses through continued inquiries and appointments.


THANK
YOU