Lead Scoring

CASE STUDY

Manasi Mhatre

BUISNESS PROBLEM

X Education, an online course provider for industry professionals, attracts visitors through marketing on various platforms. When potential customers explore courses or submit their contact information, they become classified as leads. The sales team then engages with these leads via calls and emails, achieving a typical conversion rate of around 30%. Additionally, leads are also generated through past referrals.

BUISNESS OBJECTIVE

- The company needs us to create a model that assigns a lead score to each lead. This way, customers with higher lead scores will have a greater chance of conversion, while those with lower scores will have a reduced likelihood of converting.

- The CEO has provided a rough estimate for the target lead conversion rate, aiming for approximately 80%.

GOAL

To analyze and build a robust model which allows sales team to properly identify the hot leads!!

MODEL BUIDLING

Recursive Feature Elimination (RFE) works by recursively removing the least important features and building the model repeatedly until the specified number of features is reached. This process helps in enhancing the model's performance by focusing on the most relevant features, which can lead to better predictions and insights. By eliminating variables that do not contribute significantly to the model, we can simplify the model, reduce overfitting, and improve computational efficiency. This technique is particularly useful in scenarios where we have a large number of features, allowing us to streamline the data and focus on what truly matters for our predictive analysis.

	Features	VIF
0	TotalVisits	2.37
1	Total Time Spent on Website	1.96
10	Last Notable Activity_Modified	1.86
6	Last Activity_Olark Chat Conversation	1.70
9	Last Notable Activity_Email Opened	1.50
2	Lead Origin_Lead Add Form	1.49
3	Lead Source_Direct Traffic	1.44
4	Lead Source_Welingak Website	1.34
11	Last Notable Activity_Olark Chat Conversation	1.34
7	What is your current occupation_Working Profes...	1.17
12	Last Notable Activity_Page Visited on Website	1.14
5	Do Not Email_Yes	1.13
8	Last Notable Activity_Email Link Clicked	1.02

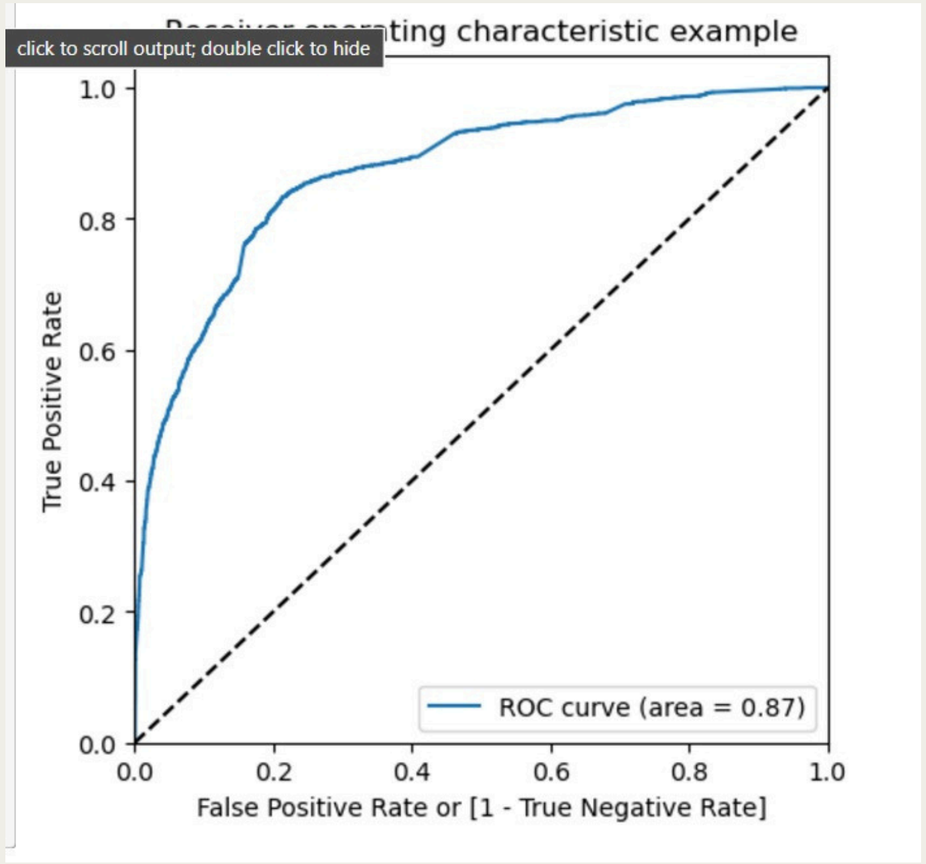
Dep. Variable:	Converted	**No. Observations:**	6293
Model:	GLM	**Df Residuals:**	6279
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2663.0
Date:	Tue, 19 Nov 2024	**Deviance:**	5326.0
Time:	22:40:50	**Pearson chi2:**	6.36e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3810
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2713	0.085	-3.186	0.001	-0.438	-0.104
TotalVisits	-0.2148	0.217	-0.992	0.321	-0.639	0.210
Total Time Spent on Website	4.0568	0.154	26.368	0.000	3.755	4.358
Lead Origin_Lead Add Form	3.7464	0.251	14.906	0.000	3.254	4.239
Lead Source_Direct Traffic	-0.5699	0.078	-7.347	0.000	-0.722	-0.418
Lead Source_Welingak Website	2.4634	1.043	2.362	0.018	0.419	4.507
Do Not Email_Yes	-1.7748	0.175	-10.154	0.000	-2.117	-1.432
Last Activity_Olark Chat Conversation	-0.8634	0.190	-4.548	0.000	-1.236	-0.491
What is your current occupation_Working Professional	2.6939	0.188	14.317	0.000	2.325	3.063
Last Notable Activity_Email Link Clicked	-1.8062	0.265	-6.819	0.000	-2.325	-1.287
Last Notable Activity_Email Opened	-1.3495	0.088	-15.399	0.000	-1.521	-1.178
Last Notable Activity_Modified	-1.8865	0.096	-19.597	0.000	-2.075	-1.698
Last Notable Activity_Olark Chat Conversation	-1.5240	0.365	-4.179	0.000	-2.239	-0.809
Last Notable Activity_Page Visited on Website	-1.7007	0.202	-8.421	0.000	-2.097	-1.305

MODEL EVALUATION

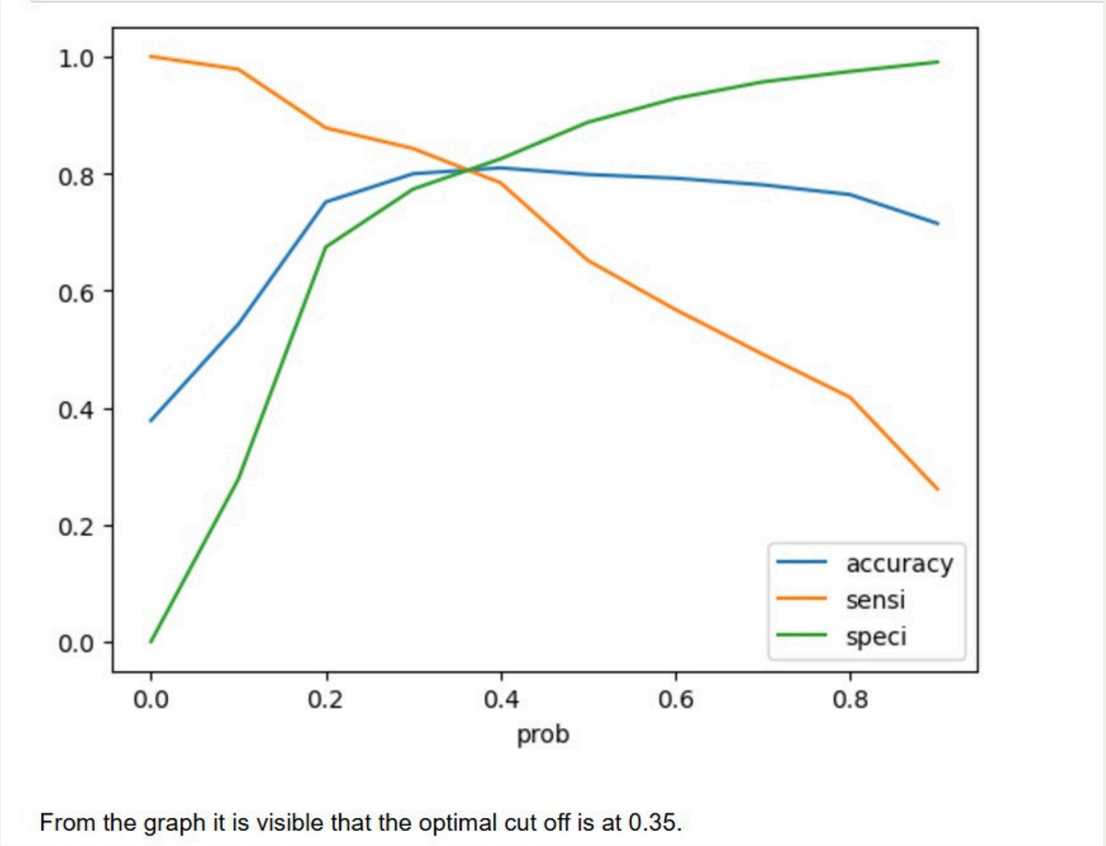
After developing the final model and making predictions on the training set, we generated a ROC curve to assess the model's stability using the AUC score (area under the curve). As illustrated in the graph on the right, the area score is 0.88, which is an impressive result.

Additionally, our graph leans towards the left side of the border, indicating that we have achieved good accuracy.



OPTIMAL THRESHOLD

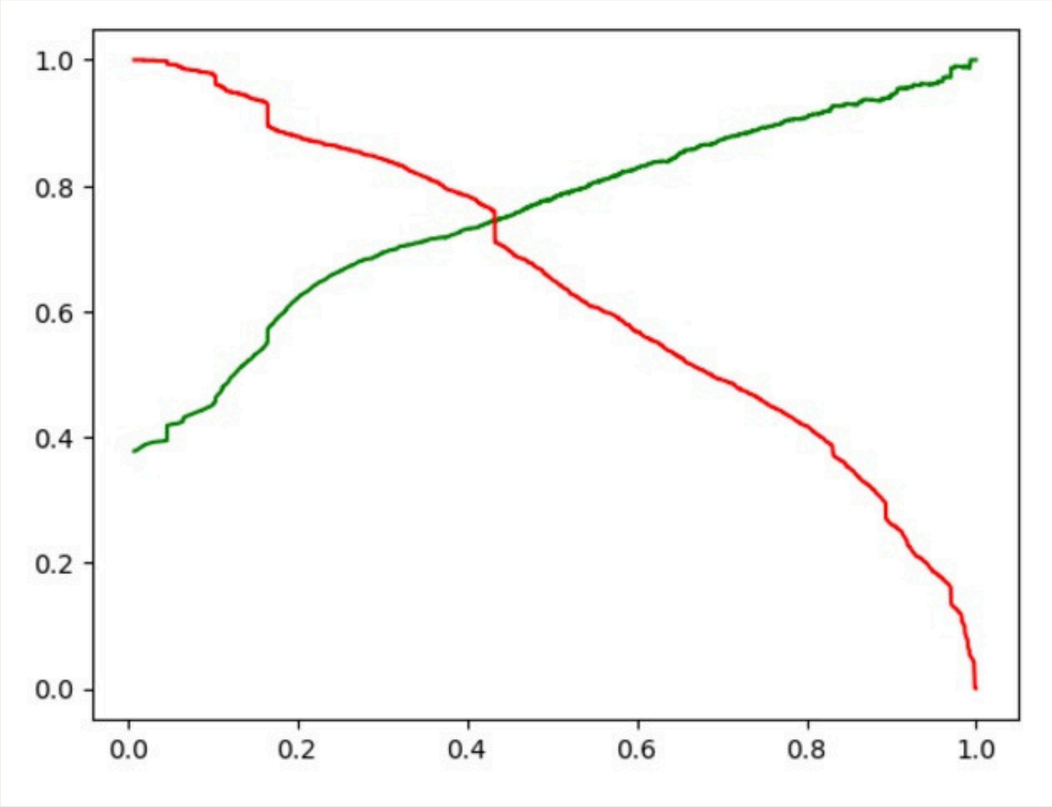
as the optimal threshold for our analysis. By choosing this point, we ensure a balanced performance across all metrics, which is crucial for the robustness of our findings. This decision allows us to confidently interpret the results, knowing that we are minimizing errors and maximizing the reliability of our predictions. Moving forward, this threshold will serve as a benchmark for further analysis



From the graph it is visible that the optimal cut off is at 0.35.

PRECESION AND RECALL TRADE OFF

This balance is crucial because it helps in optimizing the performance of a model, particularly in fields such as information retrieval, medical diagnosis, and fraud detection. By adjusting the threshold at which decisions are made, we can prioritize either precision or recall based on the specific needs of the task. For instance, in a medical diagnosis scenario, a higher recall might be prioritized to ensure that as many potential cases are identified as possible, even if it means a few false positives. Conversely, in a context where precision is more critical, such as spam detection, the focus might be on minimizing false positives to avoid misclassifying legitimate emails. By analyzing the trade-off graph, we can make informed decisions that enhance the effectiveness and reliability of our model.



CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (In descending order) : TotalVisits #The total time spend on the Website. #Lead Origin_Lead Add Form #Lead Source_Direct Traffic #Lead Source_Google #Lead Source_Welingak Website #Lead Source_Organic Search #Lead Source_Referral Sites #Lead Source_Welingak Website #Do Not Email_Yes #Last Activity_Email Bounced #Last Activity_Olark Chat Conversation

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

Thank you!