Lead Scoring Case Study

Prediction of potential leads

1. Business Context and Terminologies

- **Service provided:** X Education company markets and sells online courses.
- **Leads:** Individuals who shows interest in online courses by visiting related website as well as who are undergone or undergoing the courses.
- **Potential Leads or Hot Leads:** Leads who has the highest chance for conversion to actual lead.
- **Converted leads:** Leads who had converted to actual leads.
- **Conversion Rate:** (Successful converted lead count) / (contacted Lead count)

2. Problem Statement and Business Goals

- **Problem Statement:** On day to day life, lakhs of people shows interest in online courses. However, it doesn't mean they are ready to buy it. Sales team contact every lead without considering the outcome which is not an efficient process. Hence, business want to make this process efficient and improve it's conversion rate by calling the potential candidates first.

- **Business goal:** Currently, the typical conversion rate is about 30% whereas business set the goal of around 80%.

3. Analysis Approach

- Initially, inspect the data and make it ready for analysis by handling null records, transforming data points as required and drop the irrelevant columns which are not necessary for analysis by performing EDA. Further, transform categorical variables into numeric by creating dummy variables. On top of that, split the dataset into train and test plus perform min max scaling on numeric variables. Begin building the logistic regression model(with target variable as Converted) within 15 to 20 variables and reduce it to 15 by using RFE and further reduce variables manually until it satisfies the criteria of p-value<0.05 and VIF<5. Plotting the ROC curve to find the optimal cut-off point. Finally, evaluate the model based on Predicted value on both Train and Test data set, creating confusion matrix and calculating Specificity, Sensitivity, Precision and Recall.

4. Business Results

- Below all are the 10 variable which effects the convertible factor of lead

6 Positively effected variable

Higher TotalVisits, Total Time Spent on Website, LeadOrigin_API, LeadSource_ReferenceLeadSource_Welingak Website and LastActivity_SMS Sent can result as potential or hot lead.

4 Negatively effected variable

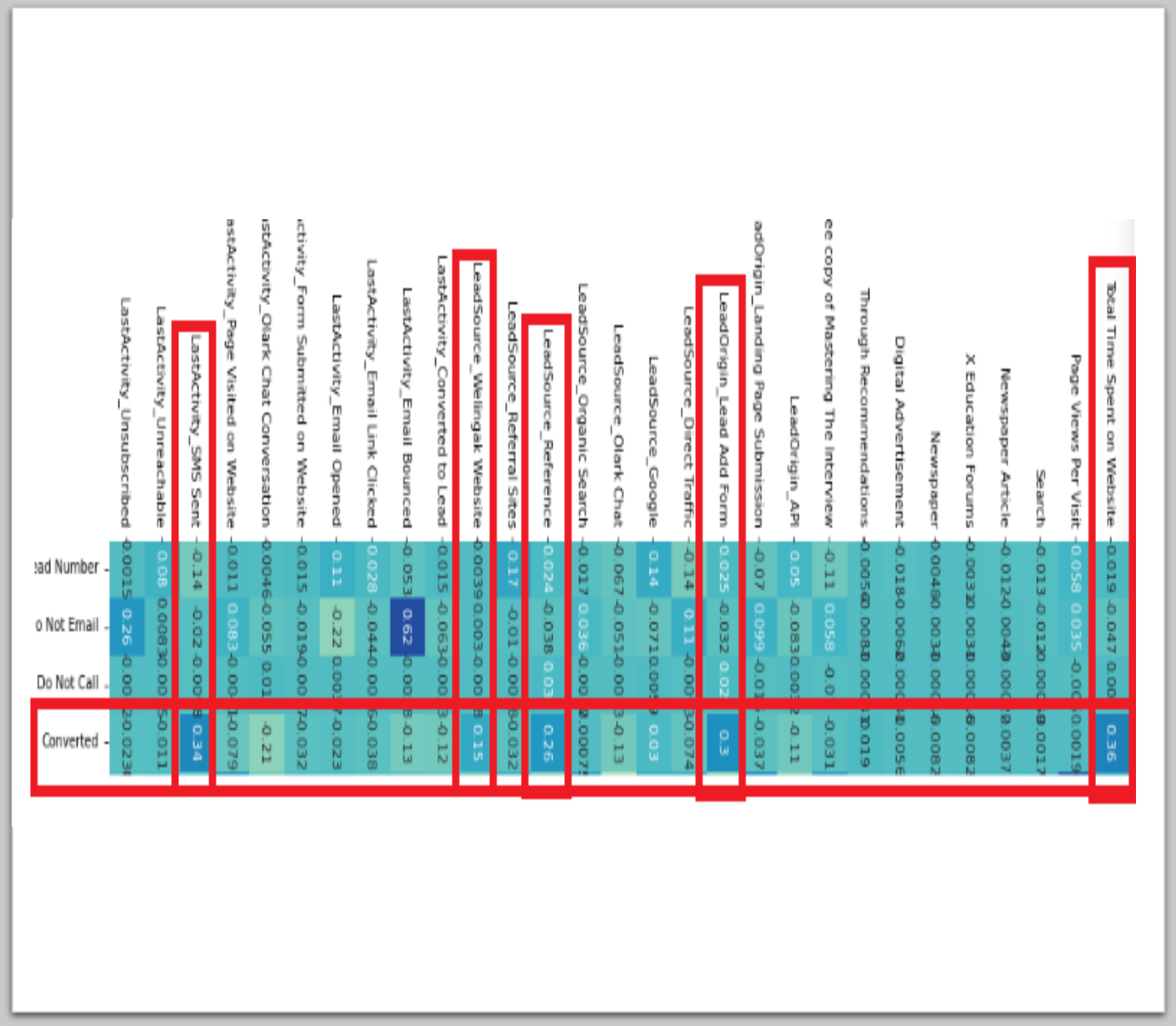
Lower Do Not Email, LeadSource_Referral Sites, LastActivity_Converted to Lead and LastActivity_Page Visited on Website can result as potential or hot lead.

Based on above 10 variable inputs of a lead, Model can easily predict whether any particular lead has potential to be converted or not (1 or 0 in predicted column)

With sensitivity of 80.4% , specificity of 80.7%

5. Correlated with target variable

- Total Time Spent on Website, LeadOrigin_Lead Add Form, LeadSource_reference, LeadSource_Welingak Website and LastActivity_SMS Sent variables are highly corelated with Converted variable which means these are potentially variables which helps in building the model.



6. Ignore highly correlated variables

- LeadOrigin_Lead Add Form and LeadSource_reference are highly correlated on top of that LeadOrigin_Lead Add Form is also highly corelated with LeadSource_Welingak Website. Therefore, LeadOrigin_Lead Add Form variable is removed from the model due to higher correlationships.

7. Final Model

- After rigorous building of model, on 5th try, I mean 5th model satisfied all the conditions like p value<0.05 and VIF<5
- Please see the model on right hand side with it's statistics and variable influence factor coefficient to know how much each variable impact the final outcome.

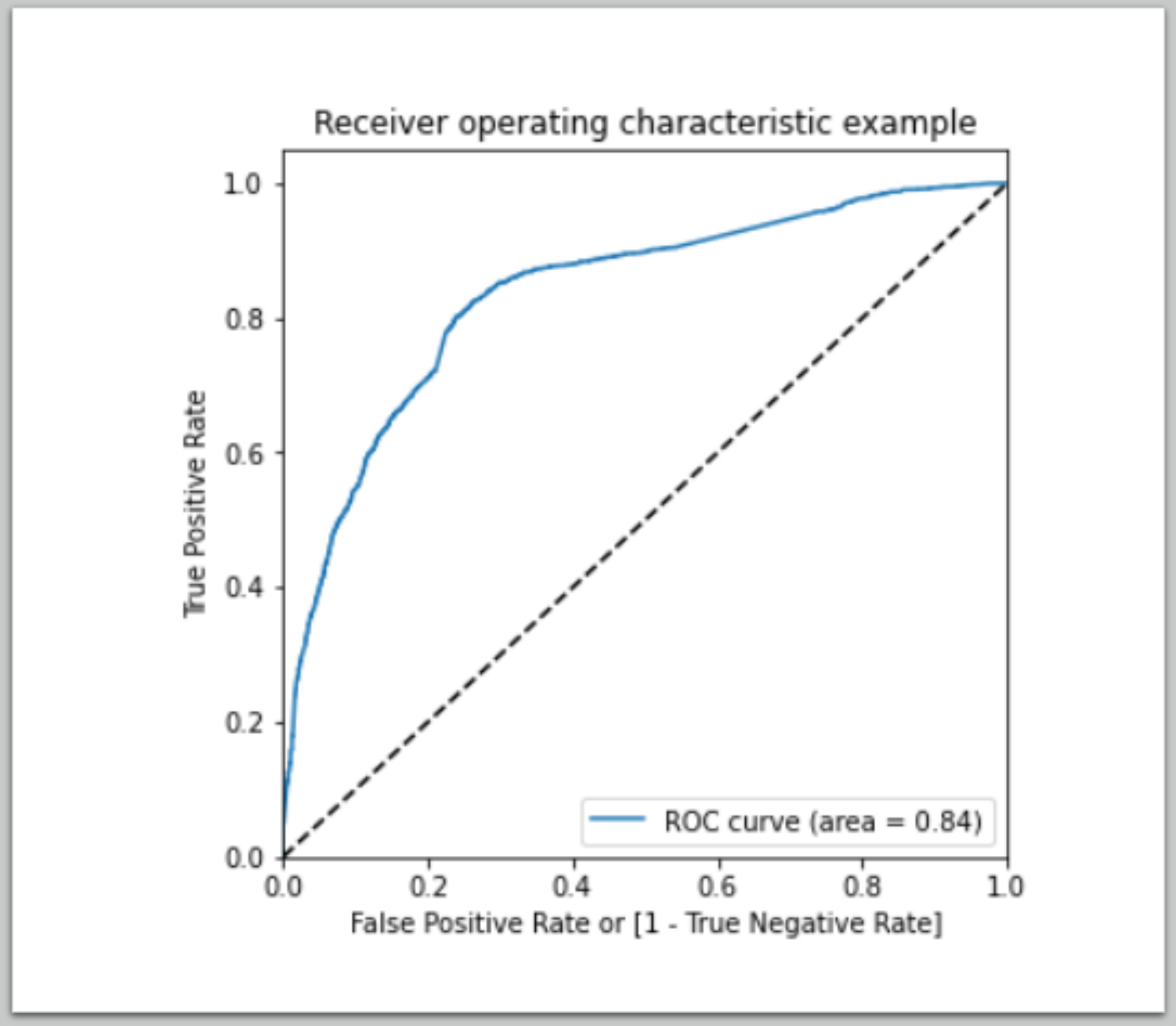
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6320
Model:	GLM	**Df Residuals:**	6309
Model Family:	Binomial	**Df Model:**	10
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3023.4
Date:	Mon, 27 Feb 2023	**Deviance:**	6046.8
Time:	01:06:31	**Pearson chi2:**	6.46e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3121
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1497	0.076	-28.410	0.000	-2.298	-2.001
Do Not Email	-1.3751	0.150	-9.194	0.000	-1.668	-1.082
TotalVisits	5.3061	1.884	2.817	0.005	1.614	8.998
Total Time Spent on Website	4.0394	0.141	28.591	0.000	3.762	4.316
LeadOrigin_API	0.3853	0.071	5.415	0.000	0.246	0.525
LeadSource_Reference	4.0108	0.215	18.627	0.000	3.589	4.433
LeadSource_Referral Sites	-0.8017	0.311	-2.581	0.010	-1.411	-0.193
LeadSource_Welingak Website	5.5528	0.722	7.688	0.000	4.137	6.968
LastActivity_Converted to Lead	-0.9866	0.191	-5.163	0.000	-1.361	-0.612
LastActivity_Page Visited on Website	-0.3968	0.142	-2.800	0.005	-0.675	-0.119
LastActivity_SMS Sent	1.2571	0.070	18.034	0.000	1.120	1.394

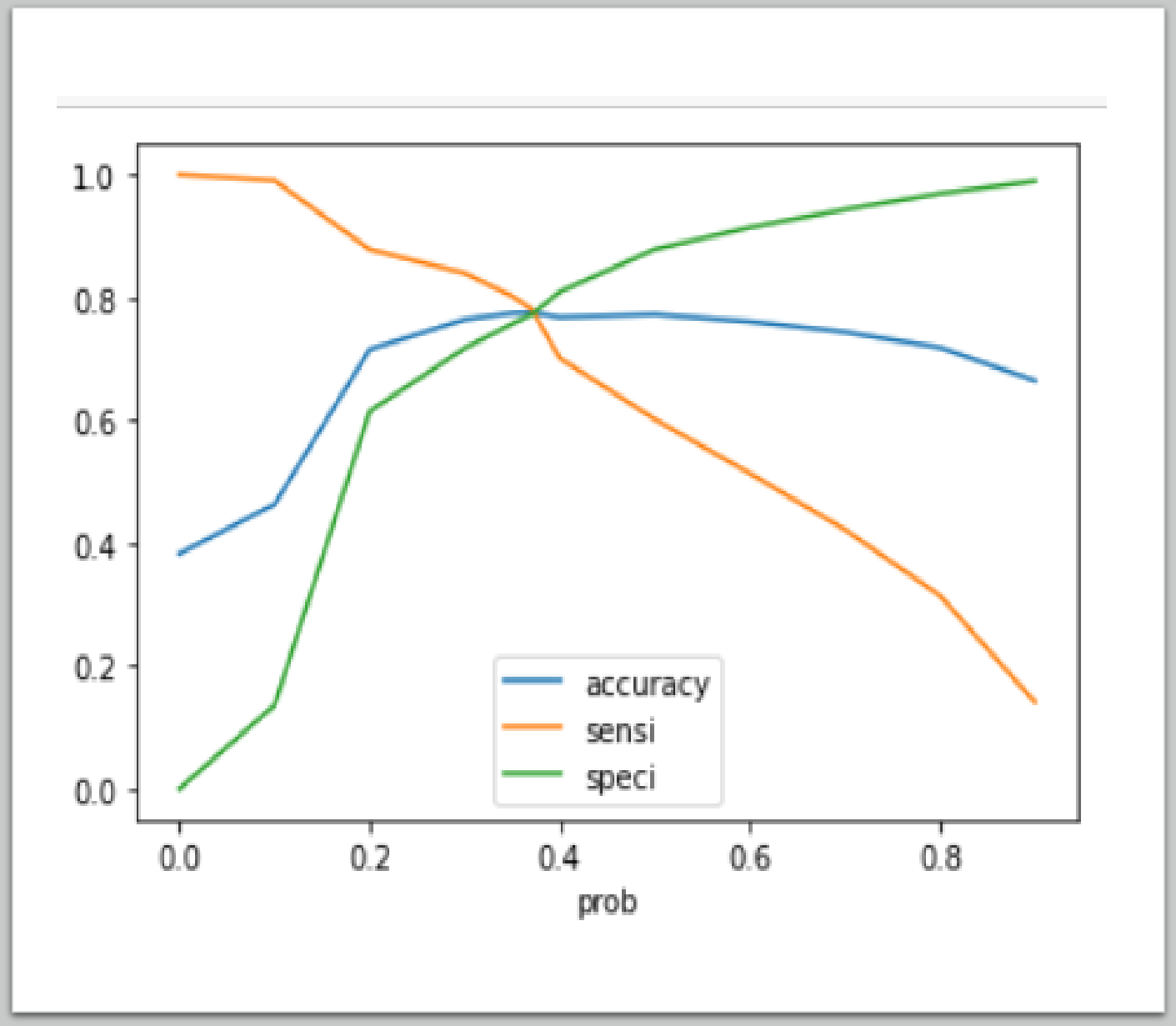
8. Model Evaluation – ROC curve

- ROC curve plotted as shown in the figure
- Since area under ROC curve is about 84% , model performance should be excellent.



9. Optimal Cut-off Point

- Plot among accuracy, sensitivity and specificity intersect at the optimal cut-off point which minimizes the cost of false positives and false negatives. it means model will predict with less errors at this point.
- Just below 0.4 is the optimum point to take it as a cutoff probability.



Thank You