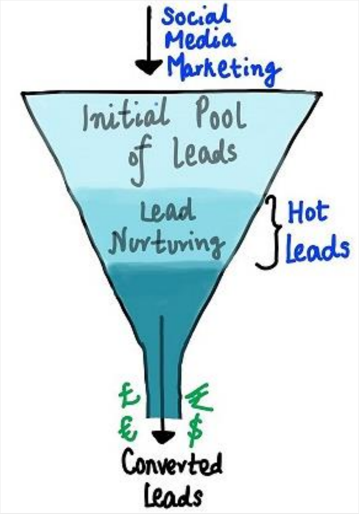
Lead Scoring – Case Study

X Education Company

Abhishek Singhal
Jitendra Mulinti
Prem Kumar R
Hemant Badgujar

Problem Statement

- An education company named X Education wants to improve on the lead conversion rate
- Current lead conversion rate at X Education is 30%(current) and are aiming to move to 80%
- Identify ‘Hot Leads’



Goal As A Data Scientist

- Build a model wherein a lead score will be assigned to each of the leads such that the customers with higher lead score will have a high probability of conversion.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Approach

Data Reading

Data Cleansing

- Removed columns with high null values (>40%)
- Imputed columns with most common response
- Removed rows with null values

Data Preparation

- Converted categorical data
- Created Dummy variables
- Removed outliers from the continuous variables

Created model on train data (70%) and predicted results for test data (30%)

Feature Scaling

Looked at the feature correlations

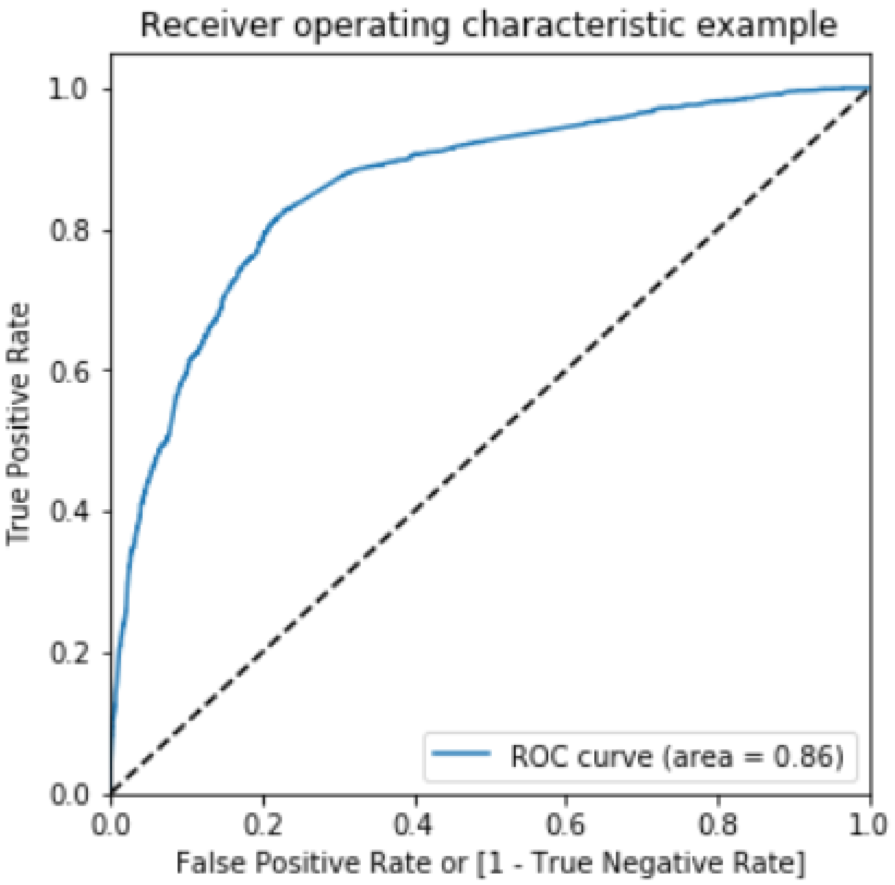
Model Building with GLM using RFE

ROC Curve to check the overall classification accuracy

Predictions on Test set of Data

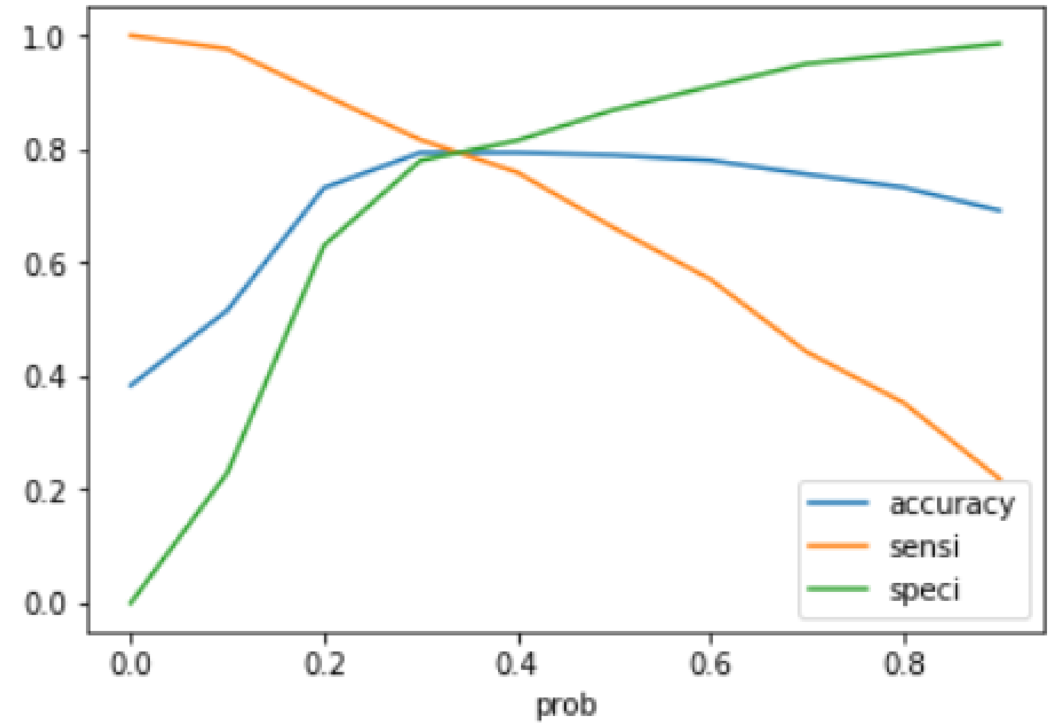
ROC Curve

- The ROC Curve plotted between True Positive Rate and False Positive Rate helps in understanding the overall accuracy of classification model.
- Our aim here is to maximize the true positive rate and minimize the false positive rate thereby signifying the high area under the ROC curve
- Our model has area under ROC curve as 0.86 which is pretty good



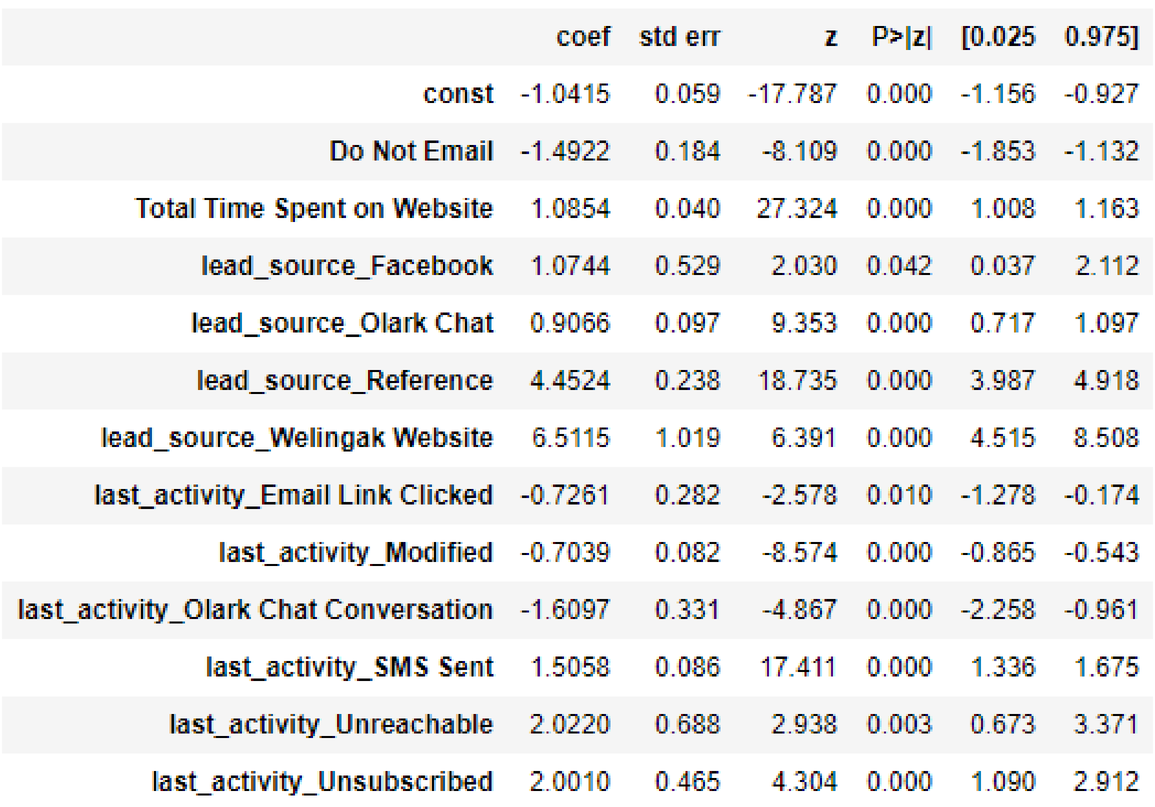
Finding Optimal Cut

- Plotted ‘Accuracy’, ‘Sensitivity’ and ‘Specificity’ against probabilities
- Optimal probability is the point at which all the three curves meet.
- For our model, optimal probability cut was found out to be at 0.3
- So, we assumed that the predicted probability more than 0.3 can be considered safely as lead getting converted.



Predicted Results

- Features selected in the model which contribute towards the conversion rate are;
    - Lead Source
    - Last Activity of Customer
    - Leads selecting “Do Not Email” option (negatively related)
    - Total time spent by leads on website
- Top three dummy variables contributing highly at conversion rate;
    - Lead Source as “Welingak Website”
    - Lead Source as “Reference”
    - Last Activity of customer as “Unreachable”



	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0415	0.059	-17.787	0.000	-1.156	-0.927
Do Not Email	-1.4922	0.184	-8.109	0.000	-1.853	-1.132
Total Time Spent on Website	1.0854	0.040	27.324	0.000	1.008	1.163
lead_source_Facebook	1.0744	0.529	2.030	0.042	0.037	2.112
lead_source_Olark Chat	0.9066	0.097	9.353	0.000	0.717	1.097
lead_source_Reference	4.4524	0.238	18.735	0.000	3.987	4.918
lead_source_Welingak Website	6.5115	1.019	6.391	0.000	4.515	8.508
last_activity_Email Link Clicked	-0.7261	0.282	-2.578	0.010	-1.278	-0.174
last_activity_Modified	-0.7039	0.082	-8.574	0.000	-0.865	-0.543
last_activity_Olark Chat Conversation	-1.6097	0.331	-4.867	0.000	-2.258	-0.961
last_activity_SMS Sent	1.5058	0.086	17.411	0.000	1.336	1.675
last_activity_Unreachable	2.0220	0.688	2.938	0.003	0.673	3.371
last_activity_Unsubscribed	2.0010	0.465	4.304	0.000	1.090	2.912

Predicted Results

Metrics	Train Data	Test Data
Accuracy	78%	79%
Sensitivity	82%	82%
Specificity	76%	78%

- Sensitivity of the model is around 82% which is maintained on the test data as well.
- It measures the proportion of actual positives being correctly predicted as such
- We're believing that the 'X Education company's' target of **80% conversion rate** is met with our model with an accuracy power of 78%