Lead Scoring

Ashish Sunil

Ashish Srivastava

Arun Anil Kumar

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- After survey, X Education company's typical lead (people fill up a form providing their email address or phone number) conversion rate is around 30% and this rate is very poor. Hence the CEO has given a ballpark of the target lead conversion rate to be around 80%.

Basic Information on Data

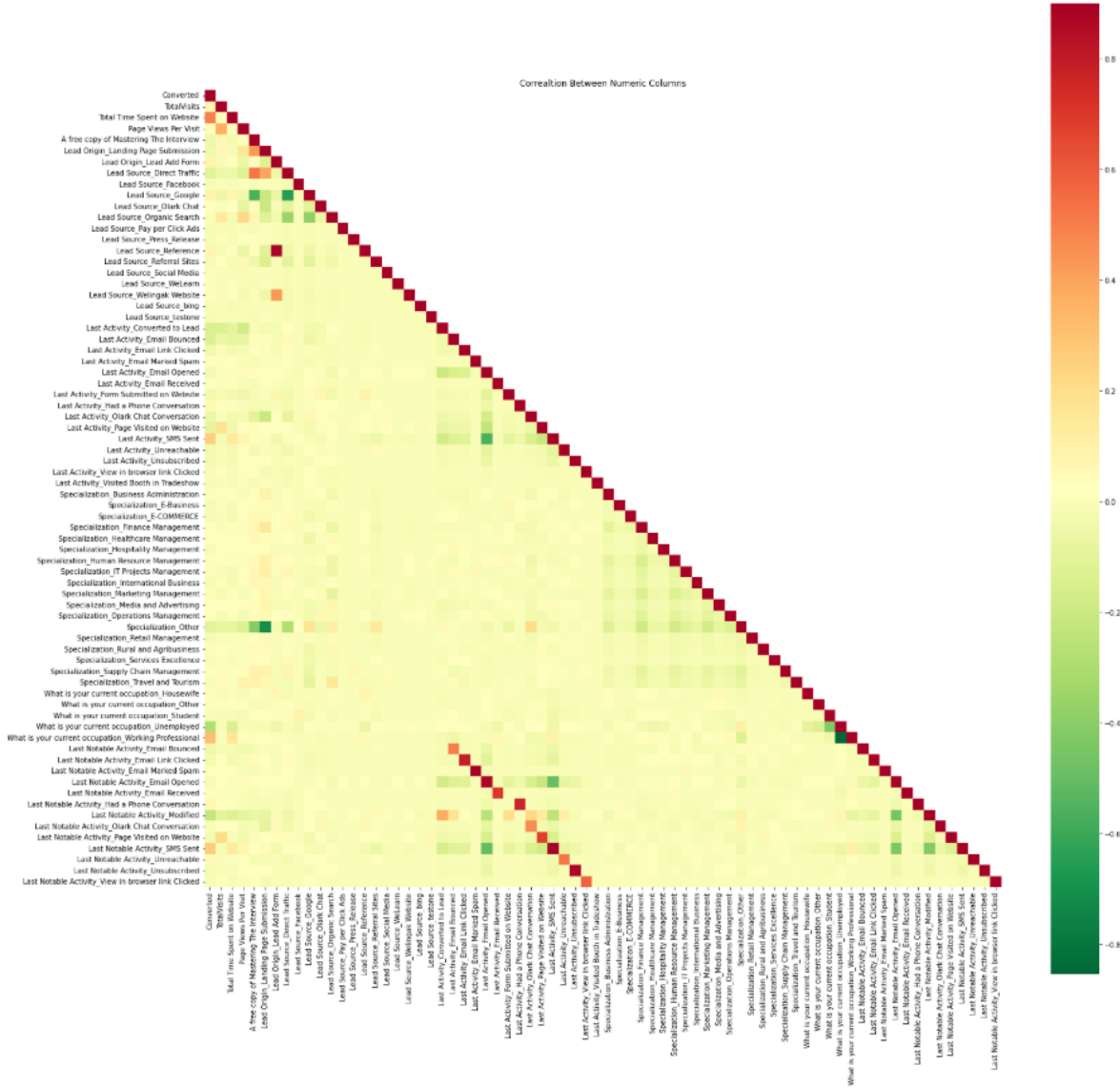
- There are 9240 records with 37 columns in given dataset.
- There are few columns present in the dataset which has null values.
- Some columns have default value as ‘Select’ which is not adding any information in dataset.

Data Cleaning

- Checked null % of each column and then decided that to drop such columns which has more than 25 % of null values. Dropped 8 columns in the above process.
- Dropped all rows having null values, having highly imbalanced dataset and which do not add any values to the dataset. Along with that, considered that the variable city, the country & lead number which do not give any useful information so that they were dropped

Correlation Matrix

- The darker shades in the correlation matrix shows higher relationship between the variables



Finalized Model

- For modeling purpose, we split the data in 70% for train & 30% for test data.
- We have scaled the numeric variables by StandardScaler.
- By using RFE method, we try the find the 25 most significant variables then 20 and after that 15 variables.
- After building the first model we have dropped the columns which have high p-values and VIF values.
- For dropping them we consider following conditions as per the standards:
  - p-value should be less than 0.05.
  - VIF also should be less than 0.05.

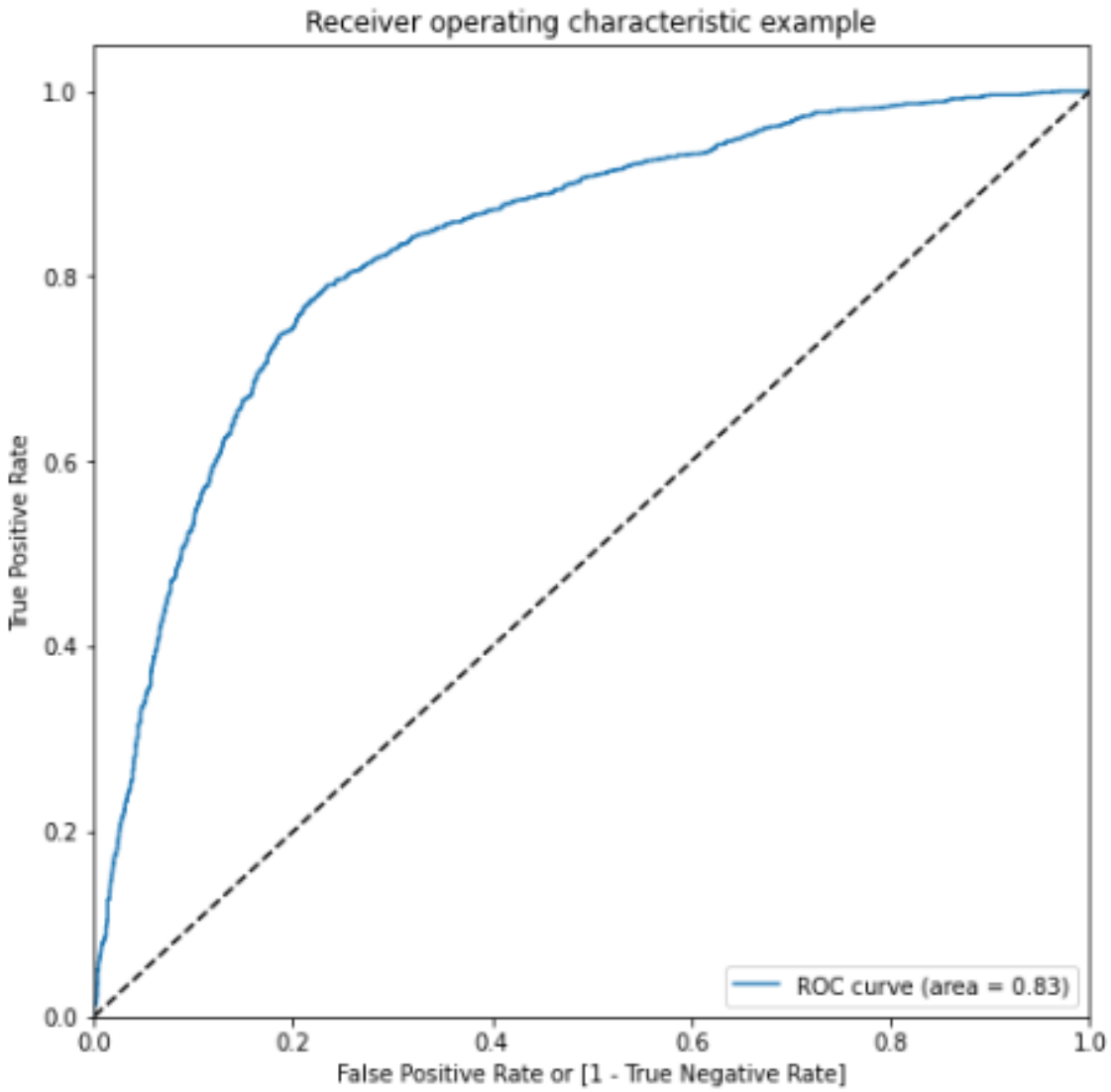
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	3447
Model:	GLM	Df Residuals:	3429
Model Family:	Gaussian	Df Model:	17
Link Function:	Identity	Scale:	0.16644
Method:	IRLS	Log-Likelihood:	-1791.6
Date:	Sun, 16 Feb 2025	Deviance:	570.73
Time:	09:38:32	Pearson chi2:	571.
No. Iterations:	3	Pseudo R-squ. (CS):	0.3819
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.6582	0.022	30.108	0.000	0.615	0.701
TotalVisits	0.0330	0.008	3.900	0.000	0.016	0.050
Total Time Spent on Website	0.2244	0.007	31.578	0.000	0.210	0.238
Page Views Per Visit	-0.0225	0.008	-2.651	0.008	-0.039	-0.006
Lead Origin_Landing Page Submission	-0.1314	0.024	-5.489	0.000	-0.178	-0.085
Lead Source_Direct Traffic	-0.0841	0.016	-5.200	0.000	-0.116	-0.052
Last Activity_Converted to Lead	-0.2561	0.031	-8.385	0.000	-0.316	-0.196
Last Activity_Email Bounced	-0.2872	0.039	-7.323	0.000	-0.364	-0.210
Last Activity_Email Link Clicked	-0.1618	0.045	-3.630	0.000	-0.249	-0.074
Last Activity_Form Submitted on Website	-0.1698	0.056	-3.056	0.002	-0.279	-0.061
Last Activity_Olark Chat Conversation	-0.1891	0.036	-5.209	0.000	-0.260	-0.118
Last Activity_Page Visited on Website	-0.1978	0.026	-7.553	0.000	-0.249	-0.146
Last Activity_Unreachable	-0.2440	0.069	-3.550	0.000	-0.379	-0.109
Last Activity_Unsubscribed	-0.1652	0.082	-2.013	0.044	-0.326	-0.004
Specialization_Other	-0.2155	0.026	-8.337	0.000	-0.266	-0.165
What is your current occupation_Housewife	0.4908	0.236	2.075	0.038	0.027	0.954
Last Notable Activity_Had a Phone Conversation	0.3321	0.136	2.433	0.015	0.065	0.600
Last Notable Activity_Unreachable	0.5510	0.136	4.045	0.000	0.284	0.818

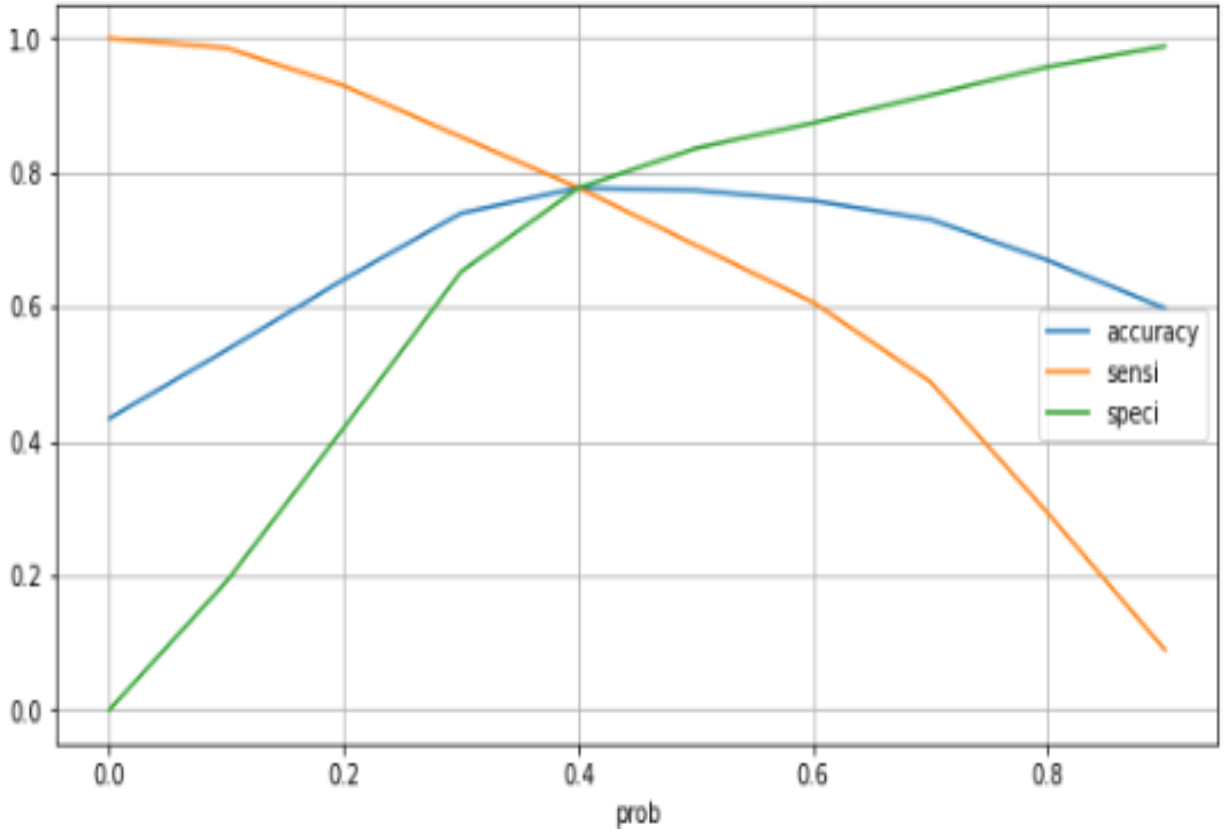
ROC Curve

- As per the observation area under ROC curve is 0.83 which is good.
- But if we have look on specificity & sensitivity of model at threshold value of 0.5, there is significant difference between them. Hence 0.5 is not the proper threshold value for the model.



Sensitivity and Specificity Curve

- As per the sensi-speci curve, optimal threshold value is nearly 0.4
- Using this value, the model is giving 77.72% accuracy. Also, sensitivity and specificity at that point is 77.74% & 77.7% respectively



Final Accuracy Metrics

```
Train Data Accuracy     :77.72 %
Train Data Recall       :77.74 %
Train Data Specificity  :77.7 %
Test Data Accuracy      :75.85 %
Test Data Recall        :78.15 %
Test Data Specificity   :73.87 %
```

- The final prediction on the test data set to bring the accuracy of that indicates that the model predicts 75.85% right conversion.
- The sensitivity is and specificity is very near to each other.

We have developed logistic regression model which has accuracy 75.85%.

THANKYOU !