Lead Scoring Case Study

Submitted By:
Dr. Avneet Singh
Ms. Tannu Singh

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google, etc.
- They get leads from various resources like their website, various online forms, past referrals etc.
- Typically they are able to convert only 30% leads.
- So, the lead conversion rate is very poor.

Business Goal

- The company needs a model wherein a lead score is assigned to each of the leads.
- The customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The Company wants to achieve 80% lead conversion ratio.

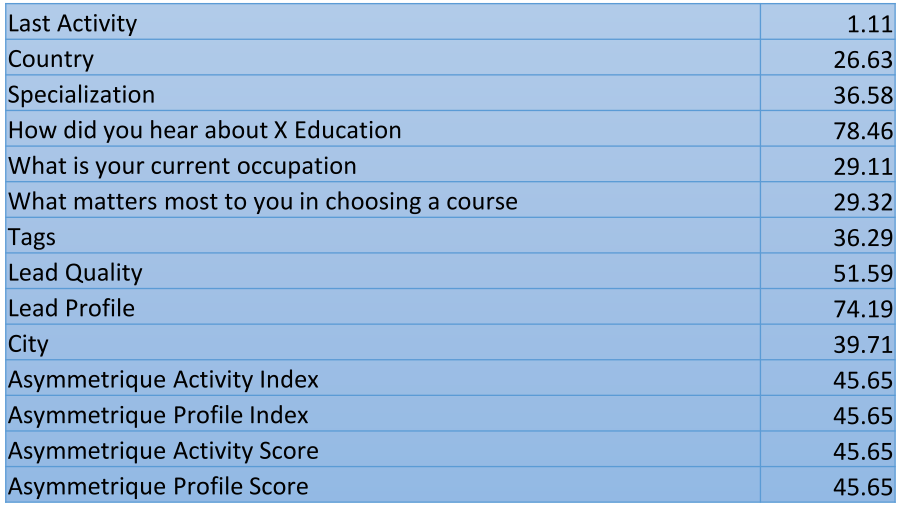
Strategy and Methodology

- Inspect the quality of data for Analysis
- Clean and prepare the data by removing missing values and outliers.
- Perform the exploratory analysis.
- Split the data in test and training set
- Feature Scaling
- Feature selection using RFE
- Model building
- Evaluating the model using various matrices like sensitivity, etc.,
- Applying the best model on testing dataset and assigning a score to each lead.

Missing Values & Outliers

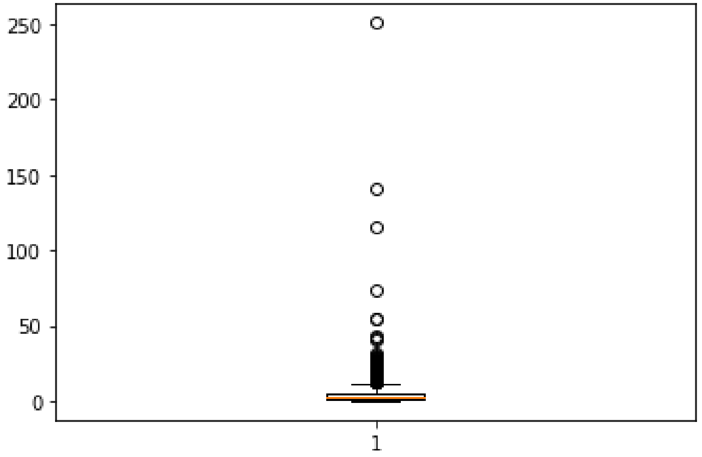
- Certain columns have the value ‘**Select**’ which has been replaced with nan.
- Columns with missing values have been identified and the values have been handled accordingly.
- Columns with more than 40% missing values have been removed.
- Outliers in the “TotalVisits” column have been capped at 95 percentiles.

Columns with Null Values



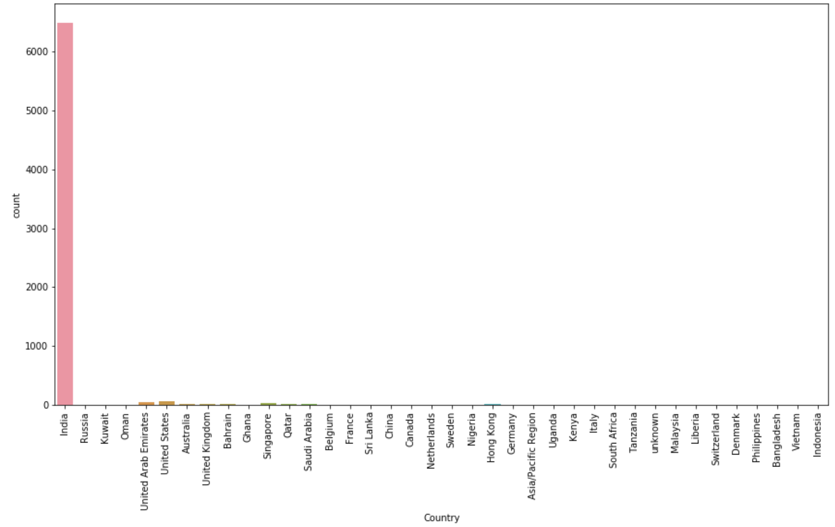
Last Activity	1.11
Country	26.63
Specialization	36.58
How did you hear about X Education	78.46
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Tags	36.29
Lead Quality	51.59
Lead Profile	74.19
City	39.71
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65

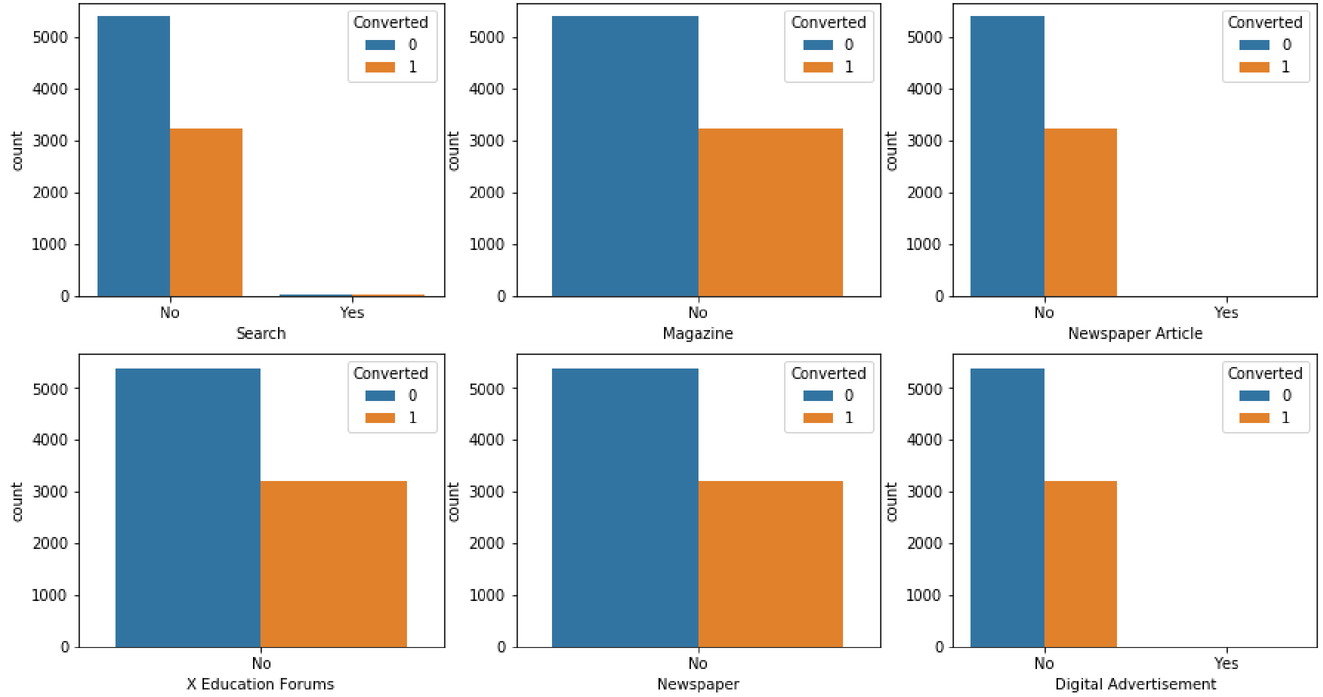
Boxplot of “TotalVisits” column

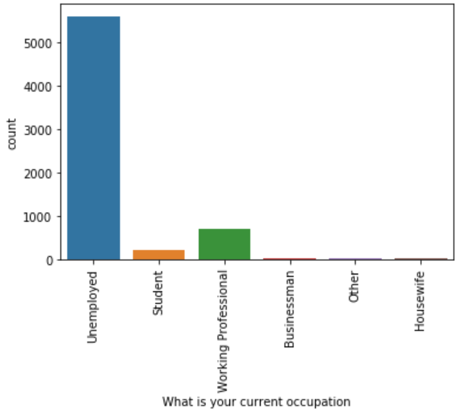


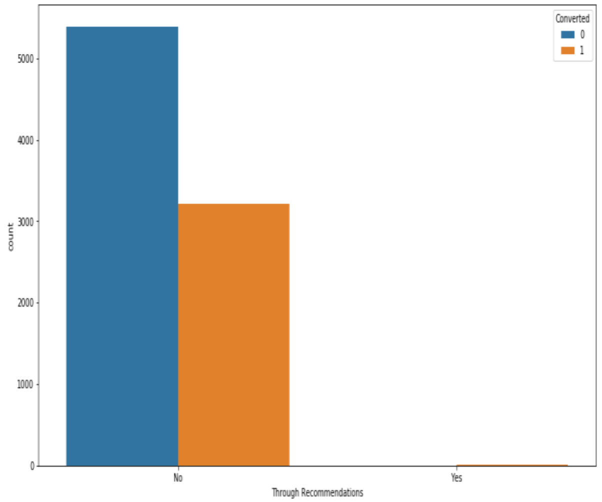
Removing Unwanted Columns

- Highly skewed columns like Country, Current Occupation, Search, Magazine, etc. have been removed.
- Columns like Tags, Lead Quality, Last Activity, Last Notable Activity, etc. which were added by sales team have also been removed.









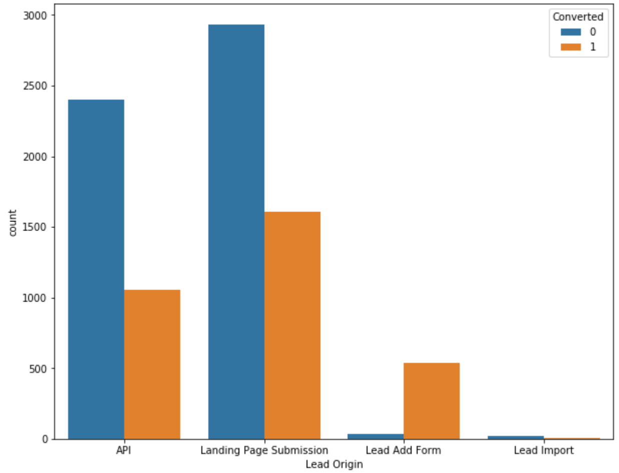
Exploratory Data Analysis Results

- Conversion Ratio of 37% is observed from the cleaned data.

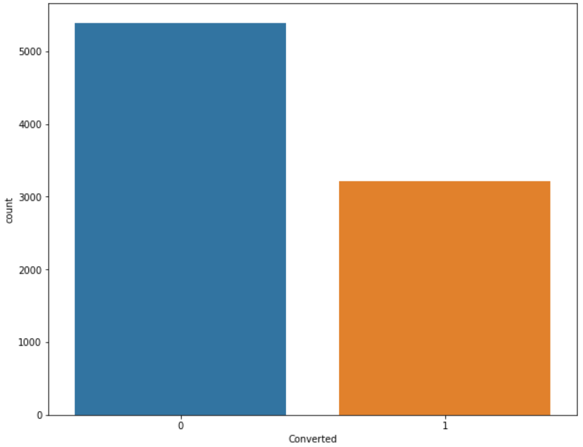
- In the “Lead Score” column, maximum leads have been observed for the “google” category, however “others “category has an appreciable conversion ratio.

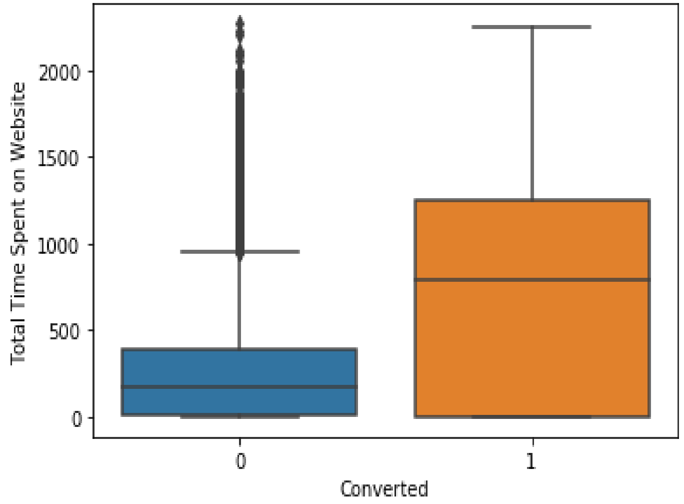
- In the “Lead Origin” column, maximum leads have been seen in “Landing Page Submission” category, but the “Lead Add Form” category has the highest conversion ratio.

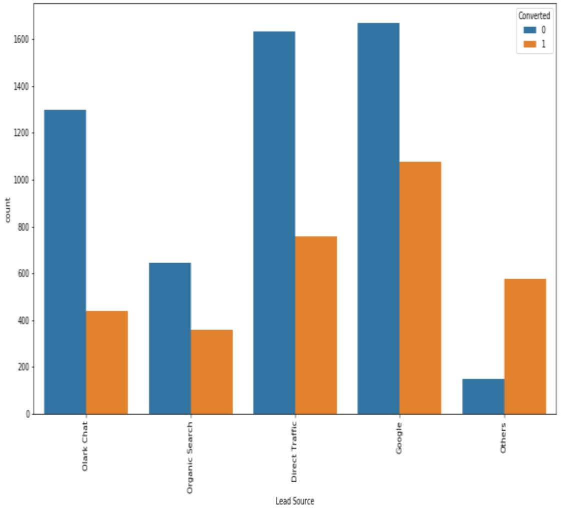
- Leads which gets converted have spent more time on website.



Conversion Ratio



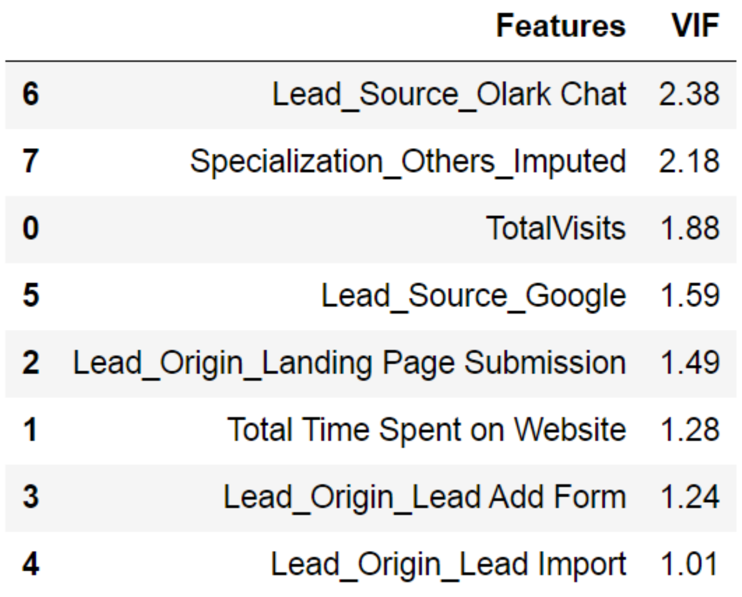




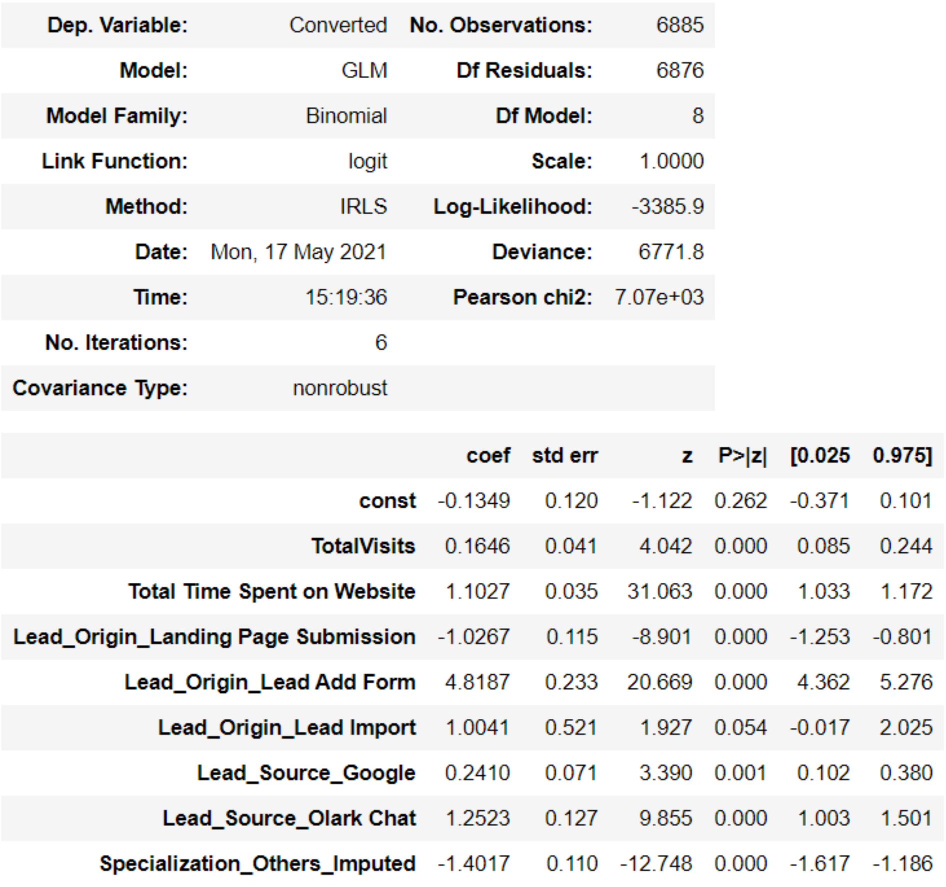
Results of Model

- A model has been developed and 8 variables have been identified that effectively contribute towards the lead conversion.

- P-Value of all the parameters was below 5% and the VIF value is less than 5.



	Features	VIF
6	Lead_Source_Olark Chat	2.38
7	Specialization_Others_Imputed	2.18
0	TotalVisits	1.88
5	Lead_Source_Google	1.59
2	Lead_Origin_Landing Page Submission	1.49
1	Total Time Spent on Website	1.28
3	Lead_Origin_Lead Add Form	1.24
4	Lead_Origin_Lead Import	1.01



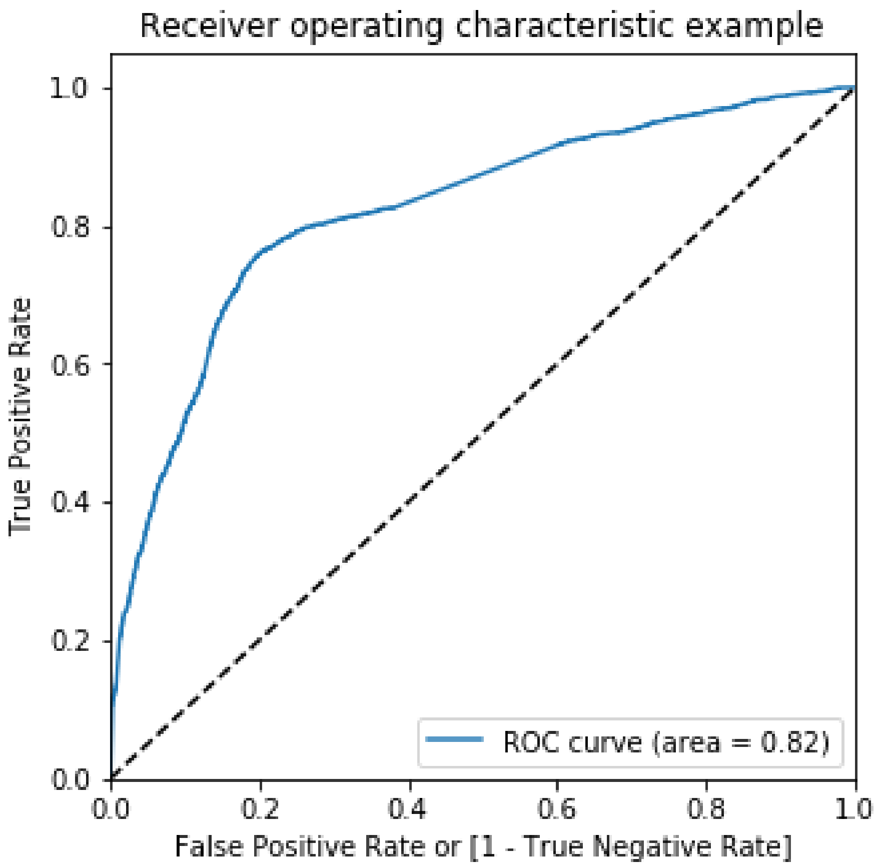
Dep. Variable:	Converted	No. Observations:	6885
Model:	GLM	Df Residuals:	6876
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3385.9
Date:	Mon, 17 May 2021	Deviance:	6771.8
Time:	15:19:36	Pearson chi2:	7.07e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1349	0.120	-1.122	0.262	-0.371	0.101
TotalVisits	0.1646	0.041	4.042	0.000	0.085	0.244
Total Time Spent on Website	1.1027	0.035	31.063	0.000	1.033	1.172
Lead_Origin_Landing Page Submission	-1.0267	0.115	-8.901	0.000	-1.253	-0.801
Lead_Origin_Lead Add Form	4.8187	0.233	20.669	0.000	4.362	5.276
Lead_Origin_Lead Import	1.0041	0.521	1.927	0.054	-0.017	2.025
Lead_Source_Google	0.2410	0.071	3.390	0.001	0.102	0.380
Lead_Source_Olark Chat	1.2523	0.127	9.855	0.000	1.003	1.501
Specialization_Others_Imputed	-1.4017	0.110	-12.748	0.000	-1.617	-1.186

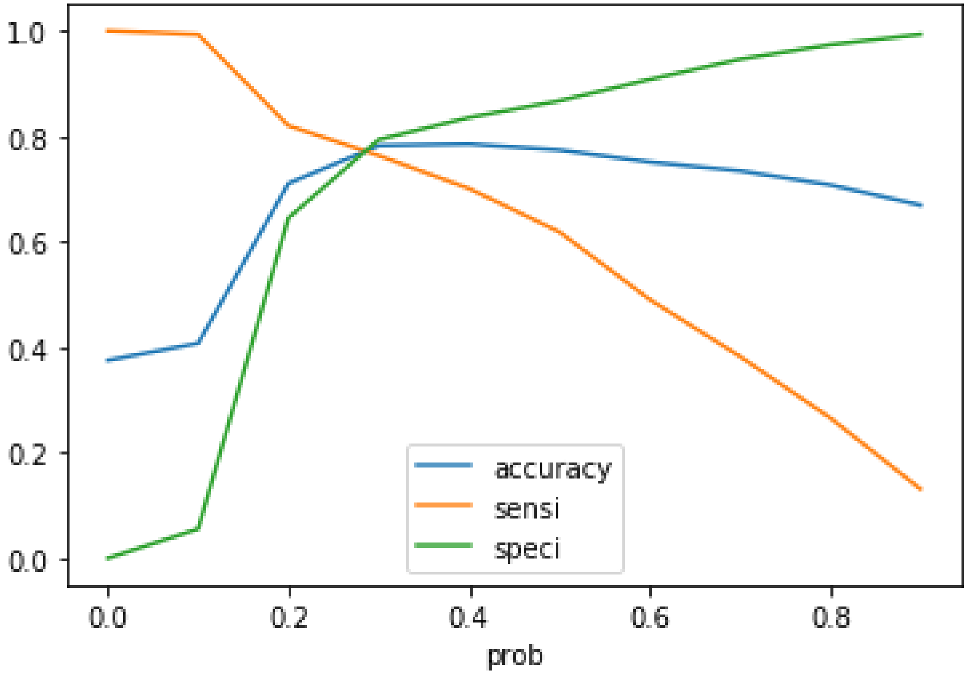
ROC Curve

- It can be seen from the ROC curve that the area under the curve is 0.82 which signifies that the developed model is good.

- The ROC curve is shifted towards the upper left corner, which is desirable for a good model.

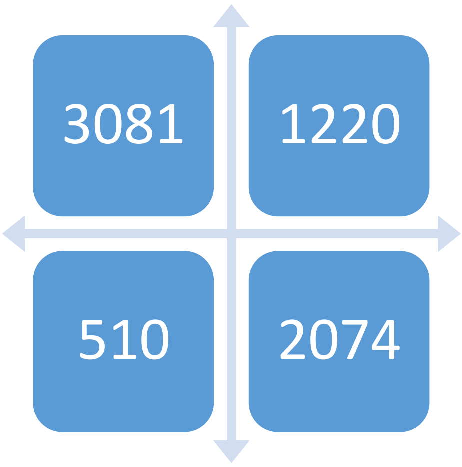


Model Evaluation : Training Set



The graph depicts an optimal cut off of 0.23 based on Accuracy, Sensitivity and Specificity

Confusion Matrix

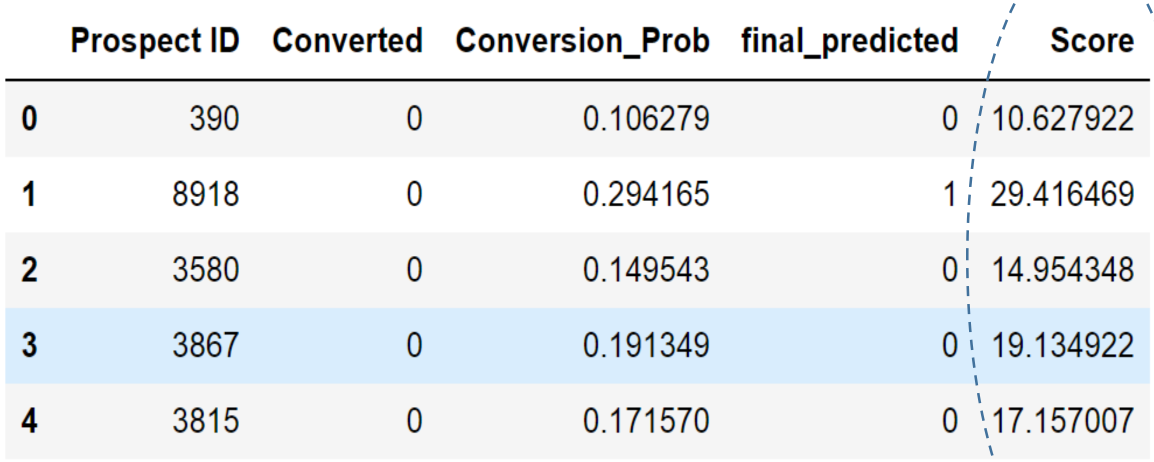


3081	1220
510	2074

Accuracy	0.7487291212781408
Sensitivity	0.8026315789473685
Specificity	0.7163450360381307
False positive rate	0.28365496396186934
Positive predictive value	0.6296296296296297
Negative predictive value	0.8579782790309106

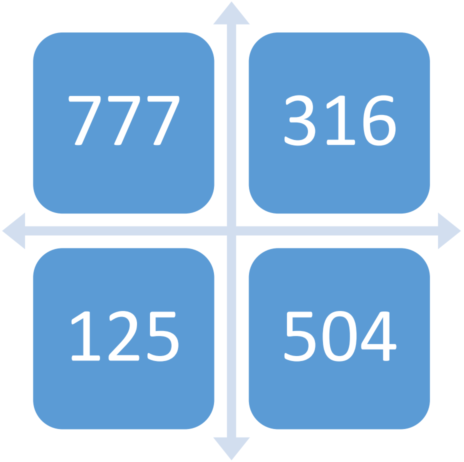
Model Evaluation : Test Set

Final Table with Score column assigned by the model to each lead



	Prospect ID	Converted	Conversion_Prob	final_predicted	Score
0	390	0	0.106279	0	10.627922
1	8918	0	0.294165	1	29.416469
2	3580	0	0.149543	0	14.954348
3	3867	0	0.191349	0	19.134922
4	3815	0	0.171570	0	17.157007

Confusion Matrix



777	316
125	504

Accuracy	0.7487291212781408
Sensitivity	0.8012718600953895
Specificity	0.7108874656907593
False positive rate	0.2891125343092406
Positive predictive value	0.6146341463414634
Negative predictive value	0.8614190687361419

The total time spent on the website, Lead Origin and Lead Source are identified to be the top three variables which contribute most towards the probability of a lead getting converted.