Lead Score Case Study

By

ANJALI MISHRA

B.BHARATH REDDY

- Problem Statement :

  X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

- Business Goal:

  X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

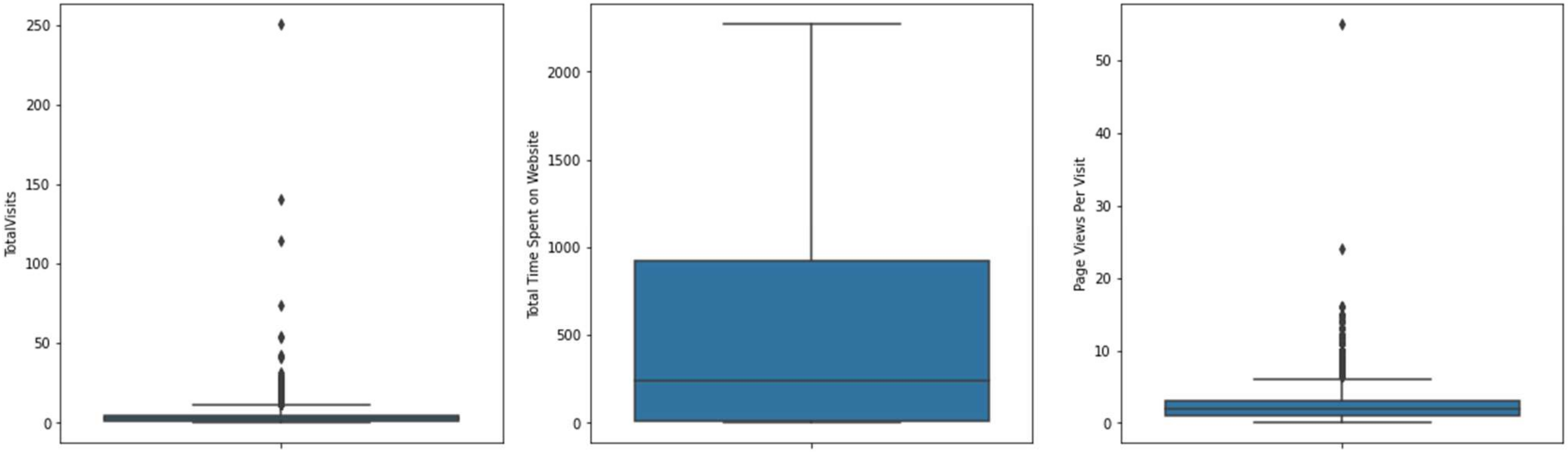
Approach to be followed for the Model Building

1. Read and understand historical lead data
2. Clean the data
3. Prepare data
4. Spilt Train & Test data
5. Feature Scaling
6. Check for correlations between variables
7. Build the model
8. Feature Selection using RFE
9. Assessing the Model with Statsmodel
10. Verify model against metrics
11. Plot ROC Curve
12. Identify optimal cut-off points
13. Make predictions on the Test set
14. Assign Lead Scores
15. Conclude the analysis

Initial observations from Data

- The Leads data provided by Education company has 9240 rows and 37 columns
- There are a few rows with missing values, that needs attention. Such as
    1. Lead Source
    2. Total Visits
    3. Page Views Per Visit
    4. Country
    5. Last Activity
    6. What is your current occupation
    7. What matters most to you in choosing a course
- There are certain columns that have ‘Select’ as label possibly because these are dropdown.
- It looks like there are outliers present in 'Total Time Spent on Website', 'TotalVisits', 'Page Views Per Visit', since the mean is almost twice or more than 50th percentile

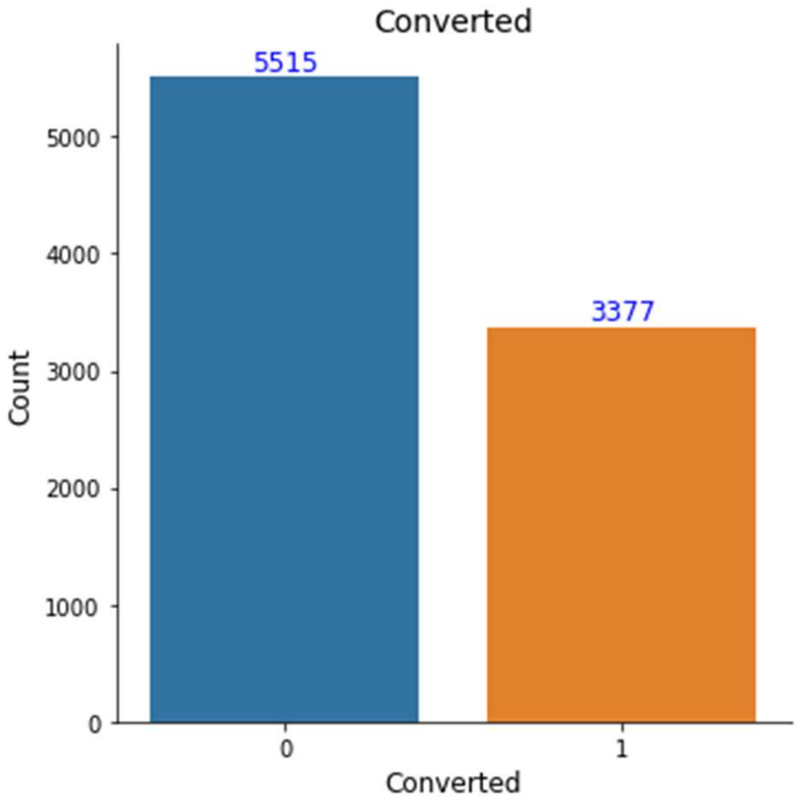
Outlier Treatment



Outliers present in 'Total Time Spent on Website', 'TotalVisits', 'Page Views Per Visit’ so any data above ’99th percentile’ was omitted to limit data from skewing.

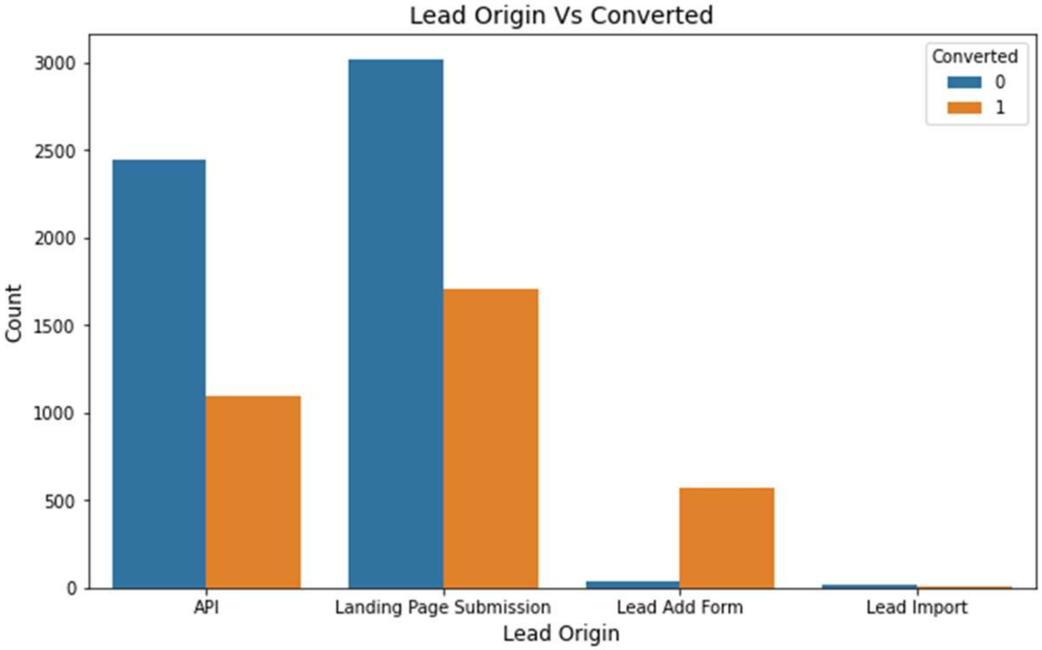
Data Analysis

We have 38% conversion rate so there is no class imbalance.



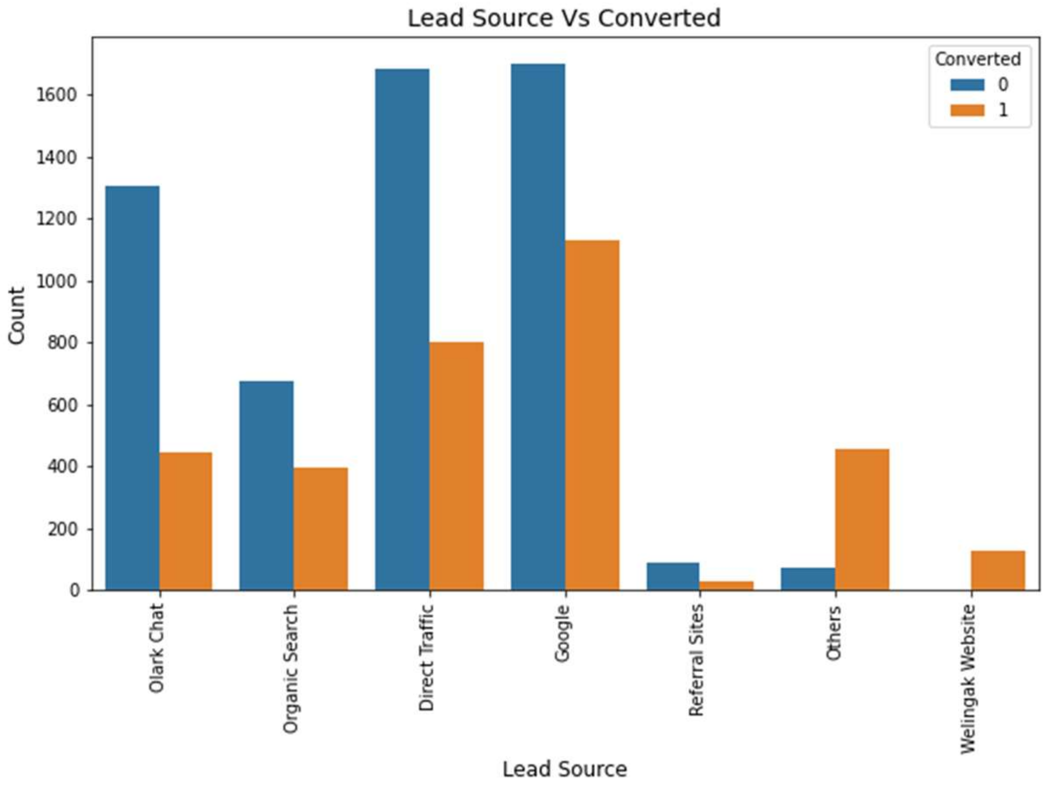
Data Analysis

From the above graph, maximum conversion happened from 'Landing Page Submission' but 'API' has better covertion rate

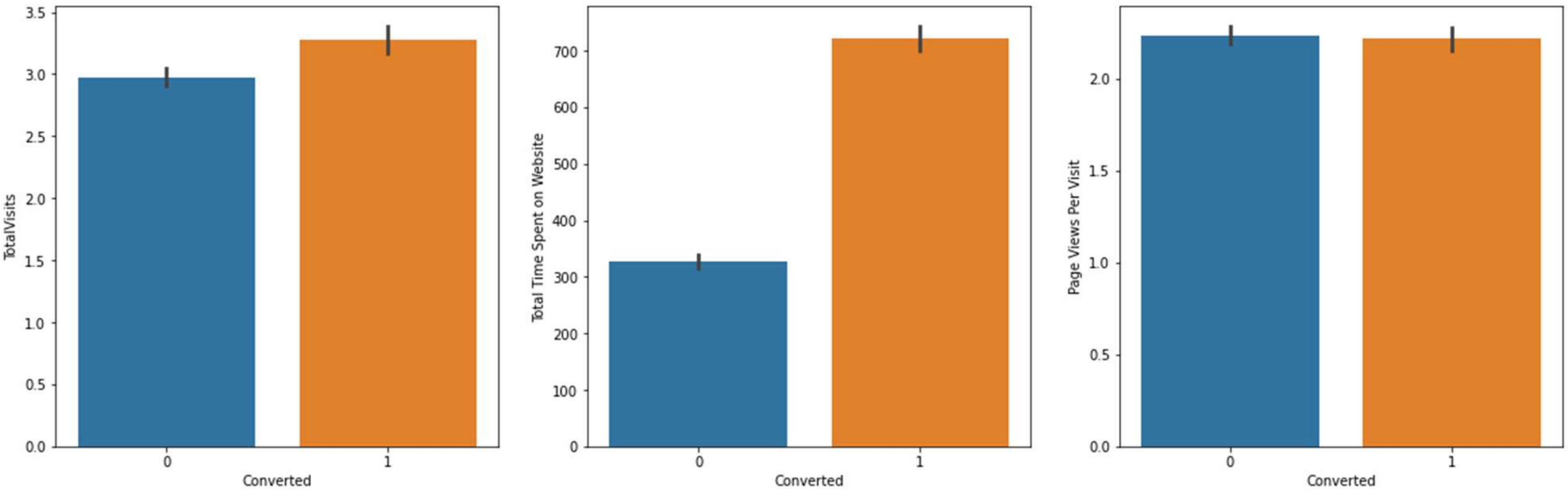


Data Analysis

Major conversion in the lead source is from 'Google', 'Direct Traffic'.



Data Analysis – Numeric Variables

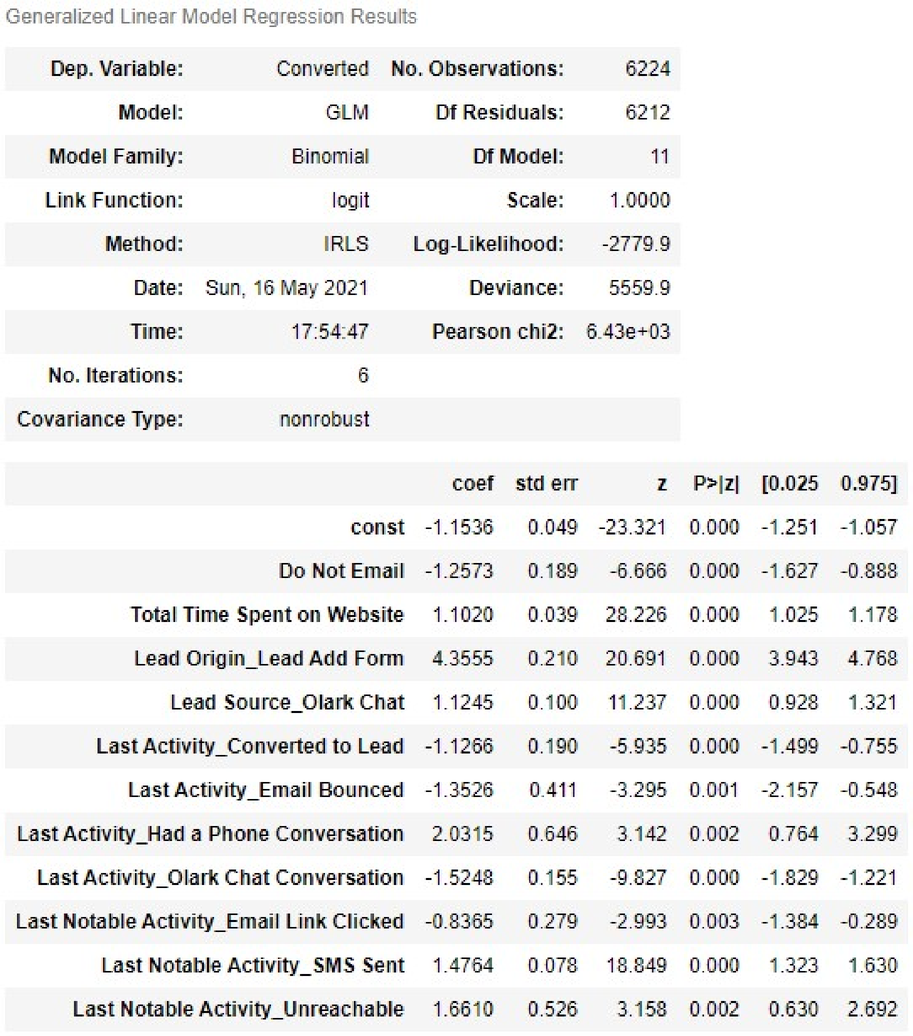


We can observe that the Time spent on the website is good indicator for the conversion rate

Model Selected

These are the top categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion.

- Lead Origin with Lead Add Form (coefficient – 4.356)
- Last activity with Had a Phone Conversation (coefficient – 2.031)
- Last Notable Activity with Unreachable (coefficient – 1.661)

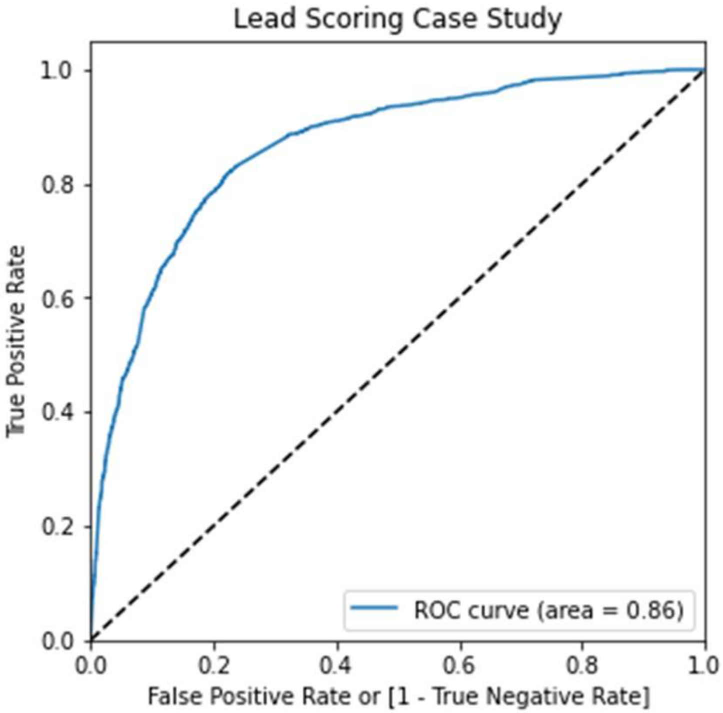


Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6224
Model:	GLM	Df Residuals:	6212
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2779.9
Date:	Sun, 16 May 2021	Deviance:	5559.9
Time:	17:54:47	Pearson chi2:	6.43e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

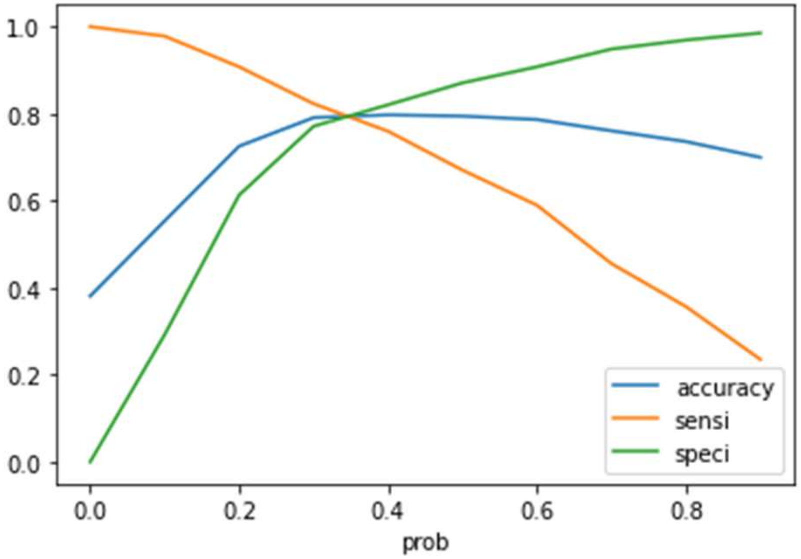
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1536	0.049	-23.321	0.000	-1.251	-1.057
Do Not Email	-1.2573	0.189	-6.666	0.000	-1.627	-0.888
Total Time Spent on Website	1.1020	0.039	28.226	0.000	1.025	1.178
Lead Origin_Lead Add Form	4.3555	0.210	20.691	0.000	3.943	4.768
Lead Source_Olark Chat	1.1245	0.100	11.237	0.000	0.928	1.321
Last Activity_Converted to Lead	-1.1266	0.190	-5.935	0.000	-1.499	-0.755
Last Activity_Email Bounced	-1.3526	0.411	-3.295	0.001	-2.157	-0.548
Last Activity_Had a Phone Conversation	2.0315	0.646	3.142	0.002	0.764	3.299
Last Activity_Olark Chat Conversation	-1.5248	0.155	-9.827	0.000	-1.829	-1.221
Last Notable Activity_Email Link Clicked	-0.8365	0.279	-2.993	0.003	-1.384	-0.289
Last Notable Activity_SMS Sent	1.4764	0.078	18.849	0.000	1.323	1.630
Last Notable Activity_Unreachable	1.6610	0.526	3.158	0.002	0.630	2.692

Model Evaluation - ROC Curve.



Area under ROC curve is 0.86 which is a good indicator for the model

Model Evaluation Optimal Cut-off Point

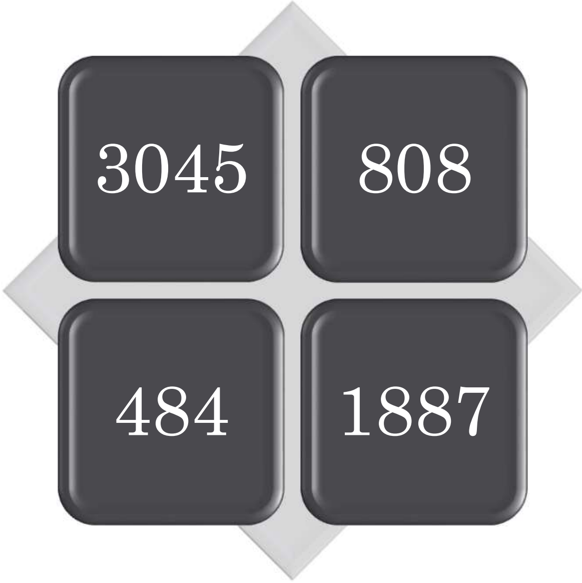


From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Model Evaluation – Confusion Matrix of Train Data

Evaluation Metrics on Train Data

- Accuracy – 79.24%
- sensitivity – 79.58%
- specificity – 79.02%
- false positive rate – 20.97%
- positive predictive value - 70%
- Negative predictive value – 86.28%
- Overall conversion rate predicted – 80%

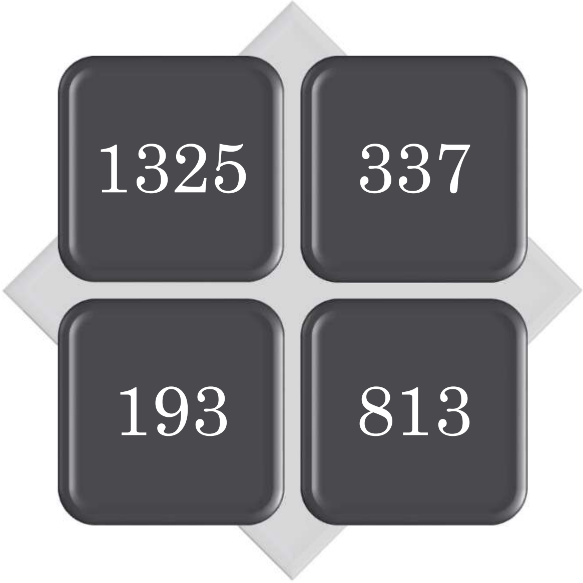


3045	808
484	1887

Model Evaluation – Confusion Matrix of Test Data

Evaluation Metrics on Train Data

- Accuracy – 80.13%
- sensitivity – 80.81%
- specificity – 79.72%
- Precision – 70.69%
- Recall – 80.81%
- Overall conversion rate predicted – 81%



1325	337
193	813

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 80%, 81% and 79.7% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is 80%
- Hence overall this model seems to be good.

THANK YOU