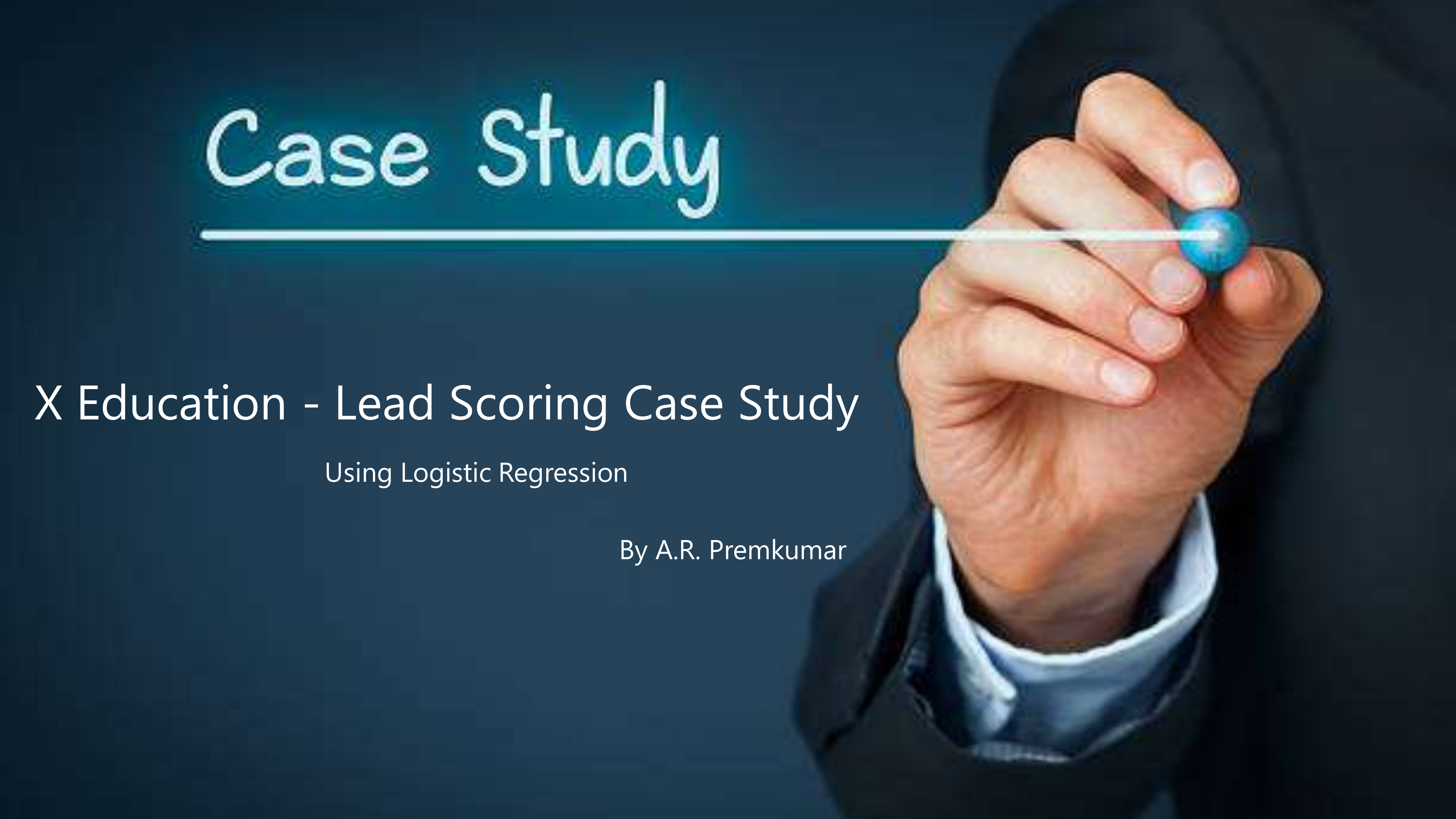


X Education - Lead Scoring Case Study

Using Logistic Regression

By A.R. Premkumar

BUSINESS OBJECTIVE

- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- Our objective is to help select promising leads using logistic regression.

SOLUTION METHODOLOGY

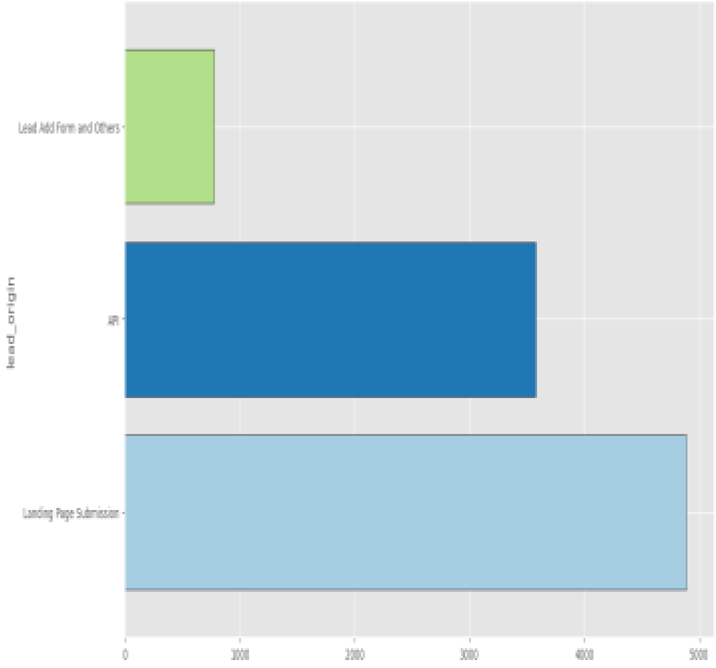
- Data Cleaning and manipulation
- Exploratory Data Analysis
- Model Building
- Model Evaluation
- Model Prediction on Testset
- Inferences
- Recommendation

DATA CLEANING

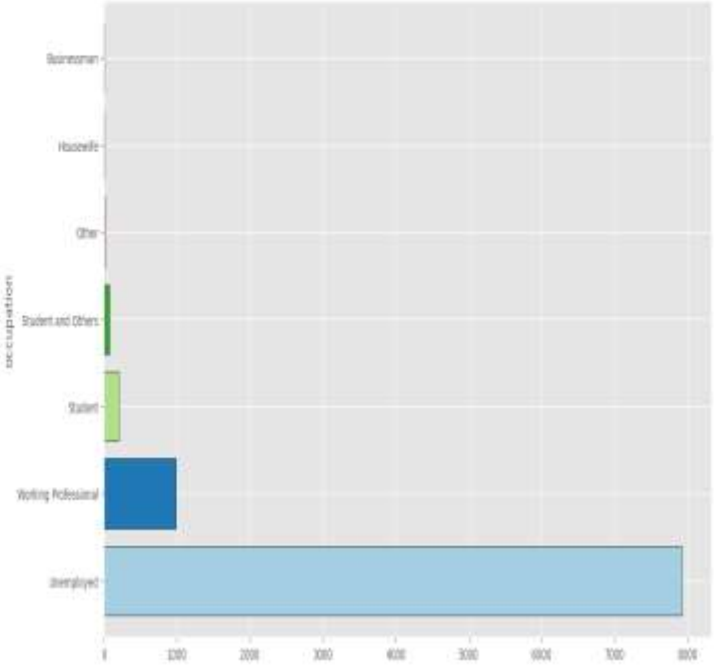
- There a lot of columns with high number of missing values and since we have around 9000+ data points we can eliminate the columns with 30% missing values;
- We dropped City and Country variables since it's of no use to us as the company provides online courses;
- Prospect ID and Lead Number are just records identifier and as hence dropped;
- We dropped all columns which have skewed data points as it wont have any predictability
  - value;
- We have found 48% conversion rate after cleaning the data.

UNIVARIATE ANALYSIS

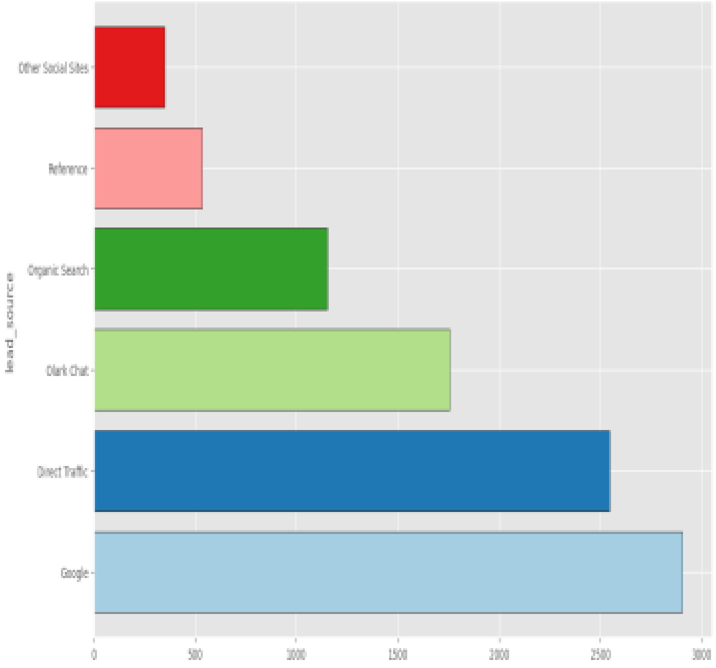
Leads by Origin



Leads by Current Occupation

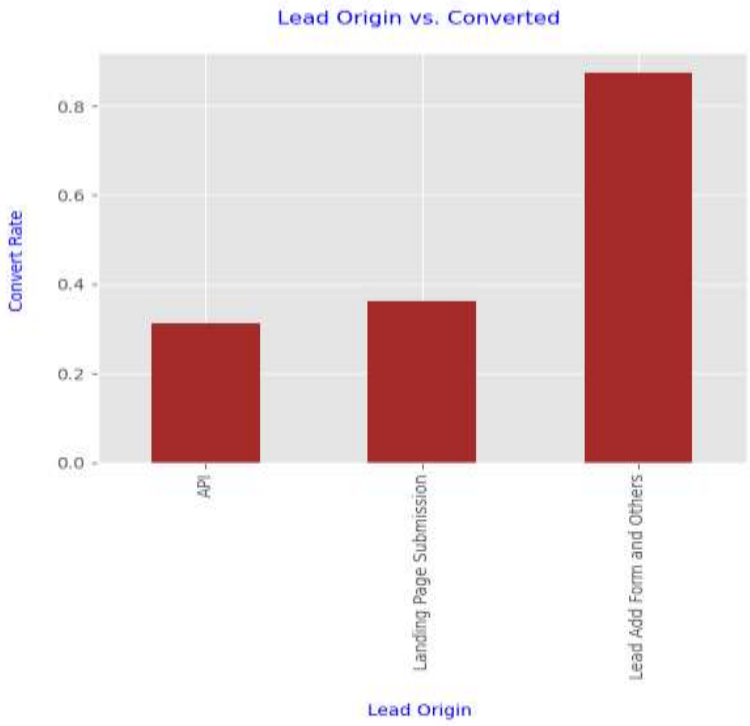


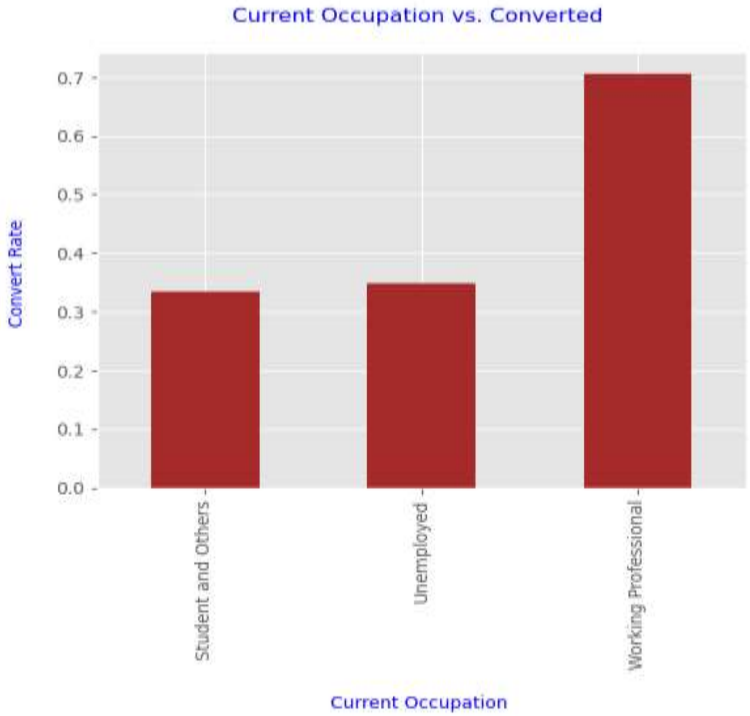
Leads by Source

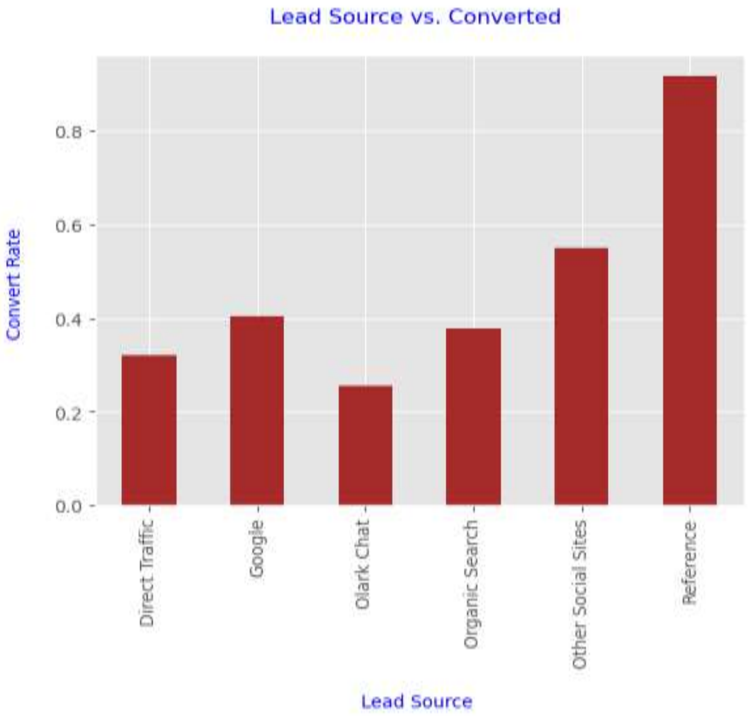


- ➢ Majority of leads are originated from Landing Page Submission followed by API
- ➢ More leads are received from 'Google' and 'Direct Traffic' followed by Olark Chat and Organic Search
- ➢ More leads are received from Unemployed customers

BI-VARIATE ANALYSIS







- Lead originated from Add Form are more likely to be converted
- Working Professional and Housewife are more likely to be converted
- Lead sources from Live Chat, Reference, WeLearn and Welingak Website are more likely to be Converted

MODEL BUILDING

- Slitting the data into train and test split with 70:30 ratio
- Scale numerical feature using MinMax scaler
- Use Recursive feature Elimination (RFE) to identify 15 most important feature
- Use p-value and Variance inflation factor to eliminate statistically insignificant features
- Finally, we ended up with 12 features for the model.
- We created a lead score (i.e. Conversion probability*100) to give a score between 0 and 100. A higher score indicates a hot lead having a higher probability of lead conversion

MODEL EVALUATION

Generalized Linear Model Regression Results

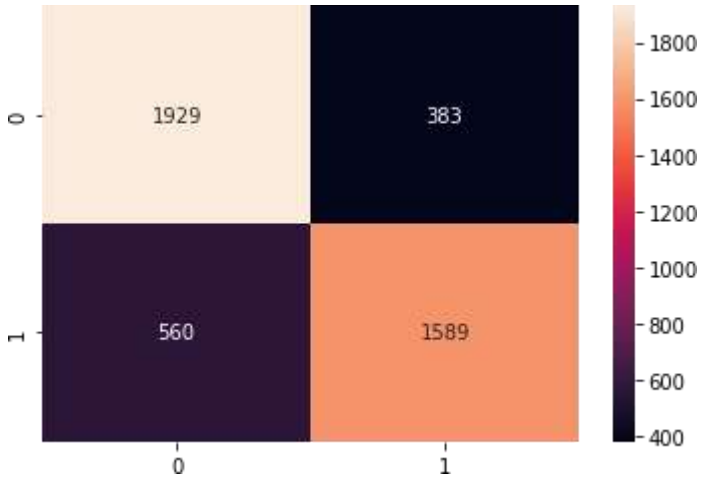
Dep. Variable:	converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6455
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3263.1
Date:	Mon, 11 Mar 2024	**Deviance:**	6526.2
Time:	10:06:27	**Pearson chi2:**	6.71e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.2741
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1774	0.110	1.617	0.106	-0.038	0.393
do_not_email	-1.2295	0.145	-8.490	0.000	-1.513	-0.946
time_on_website	1.0473	0.036	29.205	0.000	0.977	1.118
page_views_per_visit	-0.1052	0.041	-2.596	0.009	-0.185	-0.026
lead_source_Google	0.3538	0.080	4.441	0.000	0.198	0.510
lead_source_Olark Chat	0.6485	0.113	5.751	0.000	0.427	0.869
lead_source_Organic Search	0.2669	0.108	2.463	0.014	0.055	0.479
lead_source_Other Social Sites	1.6310	0.157	10.368	0.000	1.323	1.939
lead_source_Reference	3.8452	0.207	18.564	0.000	3.439	4.251
occupation_Other	-1.6328	0.775	-2.108	0.035	-3.151	-0.114
occupation_Student	-1.1465	0.231	-4.959	0.000	-1.600	-0.693
occupation_Student and Others	-2.7492	0.423	-6.496	0.000	-3.579	-1.920
occupation_Unemployed	-1.3264	0.101	-13.144	0.000	-1.524	-1.129

	Features	VIF
11	occupation_Unemployed	2.96
4	lead_source_Olark Chat	2.15
3	lead_source_Google	1.85
2	page_views_per_visit	1.74
5	lead_source_Organic Search	1.41
1	time_on_website	1.21
7	lead_source_Reference	1.21
6	lead_source_Other Social Sites	1.12
0	do_not_email	1.09
9	occupation_Student	1.04
10	occupation_Student and Others	1.03
8	occupation_Other	1.00

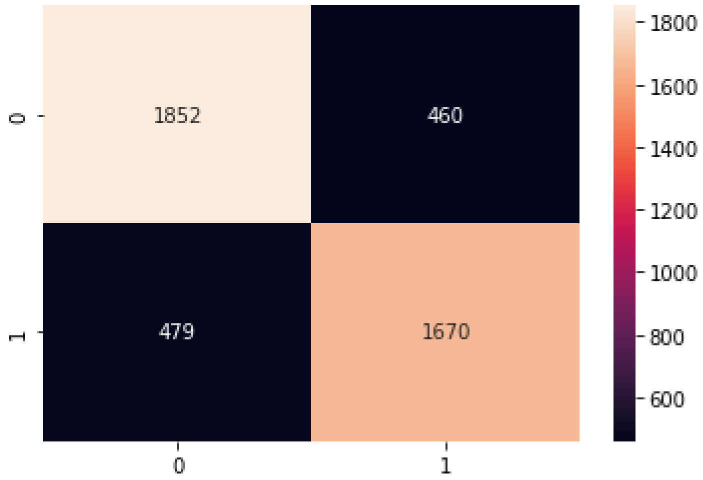
MODEL EVALUATION

TRAINING SET



Accuracy	78.86%
Sensitivity	73.94%
Specificity	83.43%
Precision	80.58%
Recall	73.94%

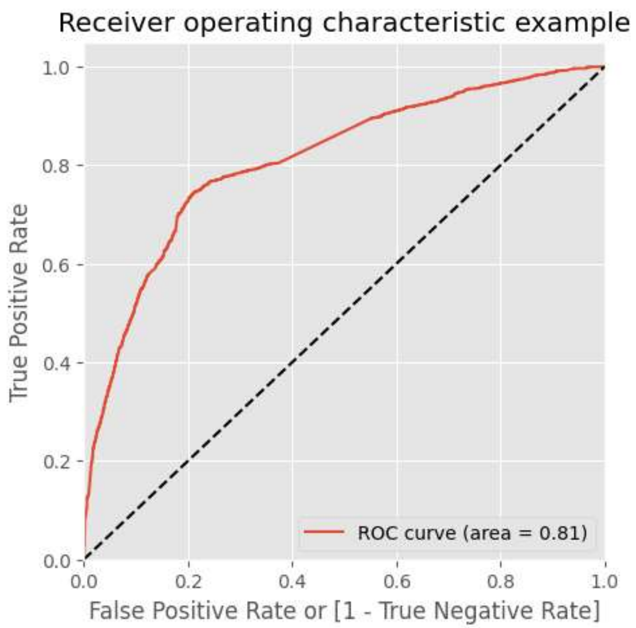
TEST SET

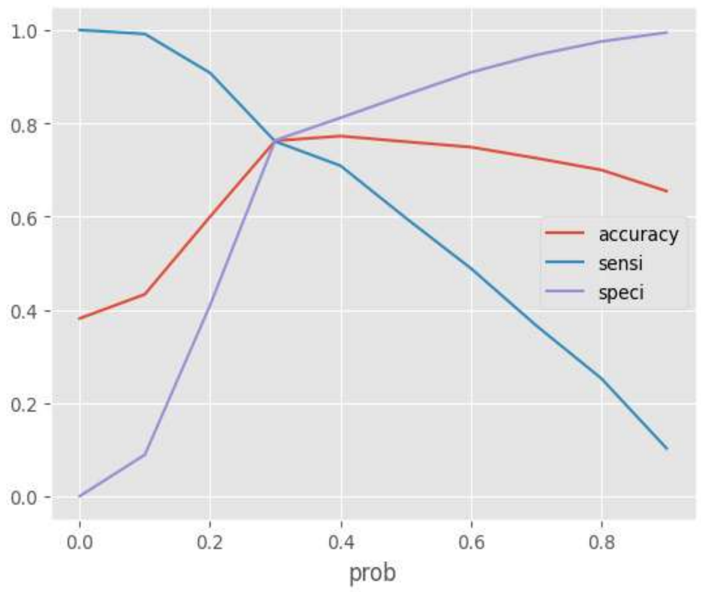


Accuracy	78.95%
Sensitivity	77.71%
Specificity	80.10%
Precision	78.40%
Recall	77.71%

MODEL EVALUATION - ROC/CUTOFF

	prob	accuracy	sensi	speci
0	0	38.00%	100.00%	0.00%
0.1	0.1	43.00%	99.00%	9.00%
0.2	0.2	60.00%	91.00%	41.00%
0.3	0.3	76.00%	76.00%	76.00%
0.4	0.4	77.00%	71.00%	81.00%
0.5	0.5	76.00%	60.00%	86.00%
0.6	0.6	75.00%	49.00%	91.00%
0.7	0.7	73.00%	37.00%	95.00%
0.8	0.8	70.00%	25.00%	98.00%
0.9	0.9	65.00%	10.00%	99.00%





INFERENCES

Top three variables in your model which contribute most towards the probability of a lead getting converted

a. ***TotalVisits,***
b. ***Total Time Spent on Website,***
c. ***Lead Origin_Lead Add Form***

Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion

a. ***Lead Origin_Lead Add Form***
b. ***Last Activity_Had a Phone Conversation***
c. ***Lead Source_Welingak Website***

RECOMMENDATION

Depending on the requirements the model needs to be tweaked such that

Scenario 1:
So when the company has more interns we need have lower cutoff threshold so that our model can predict almost all leads. The flip side to this decrease in threshold will be that we will misclassify some non-conversions as conversions but this is a good tradeoff given we have mode manpower to deal with it.

Scenario 2:
Typically, when the company has less people to call potential customers so its good to have more accurate predictions in which case the model specificity should be much more higher. This would mean form the above graph the we would have to choose a cutoff point which is much higher. The tradeoff of this is that we are going to miss some leads but given that the company has less manpower who can focus

more on correctly predicted leads.

Scenario 3:

The company should focus on sending automated SMS and emails to potential leads during the time they have less manpower which

allows for cost effective lead conversion without manual intervention.

Thank You