Lead Score Case Study

Submitted by :

Ashish Dhyani

Ranjita Lenka

Lead Score Case Study for X Education Company

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.

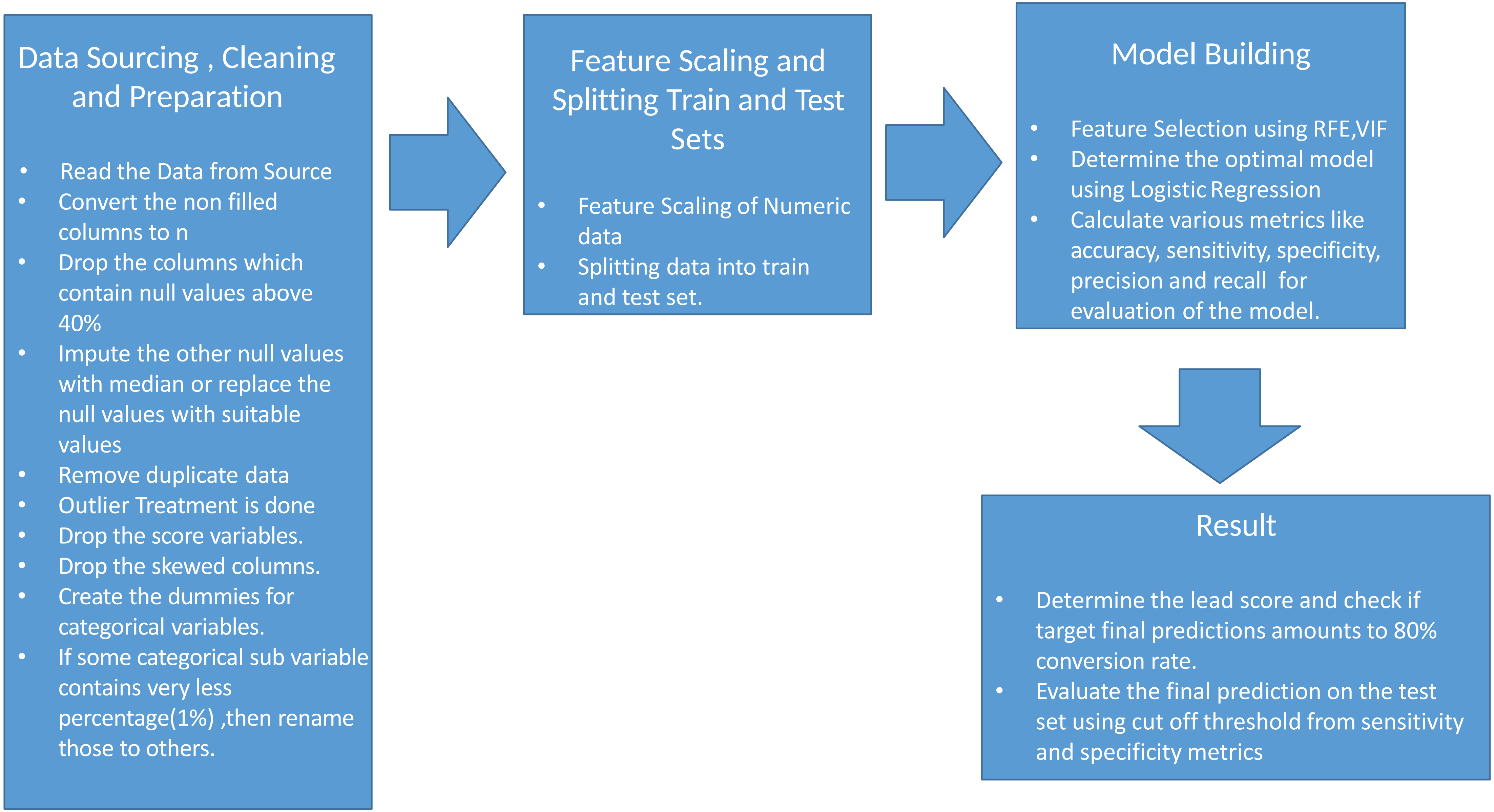
The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Strategy

- Read and Understand the data
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score by using RFE,VIF.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Problem solving methodology



Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert the non filled columns to n
- Drop the columns which contain null values above 40%
- Impute the other null values with median or replace the null values with suitable values
- Remove duplicate data
- Outlier Treatment is done
- Drop the score variables.
- Drop the skewed columns.
- Create the dummies for categorical variables.
- If some categorical sub variable contains very less percentage(1%) ,then rename those to others.

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

Model Building

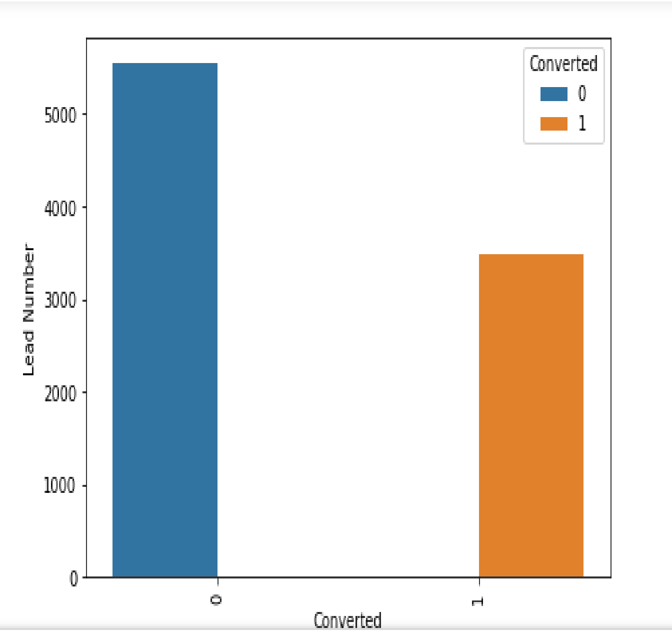
- Feature Selection using RFE,VIF
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall for evaluation of the model.

Result

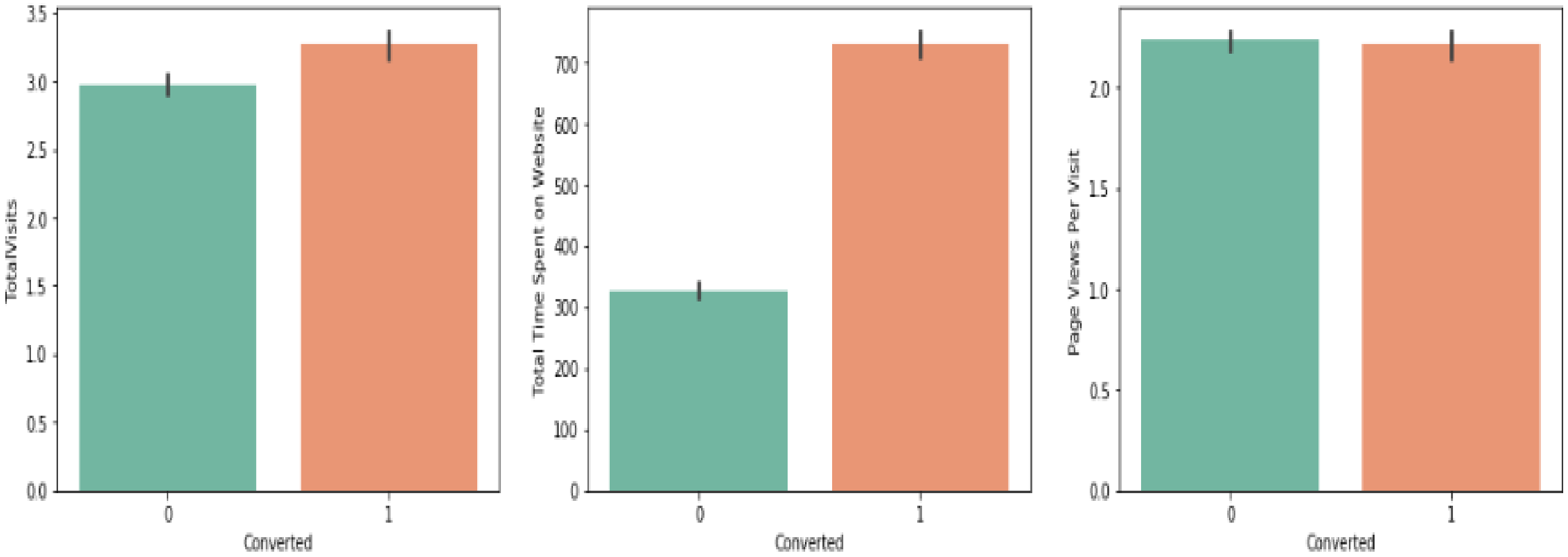
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Exploratory Data Analysis

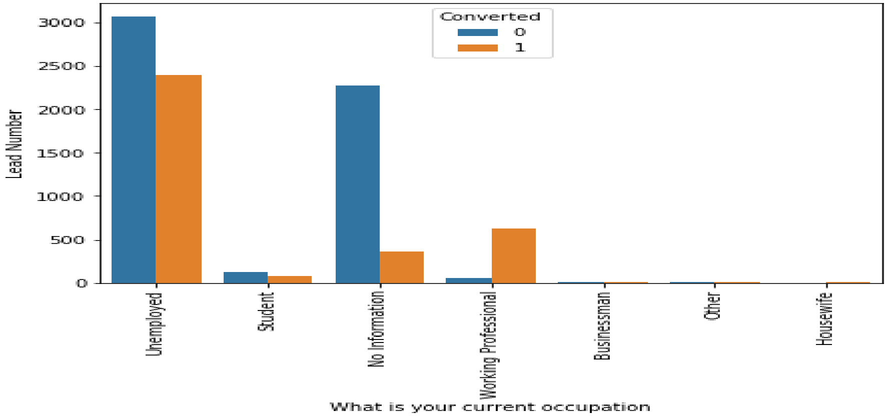
We have around 39% Conversion rate in Total

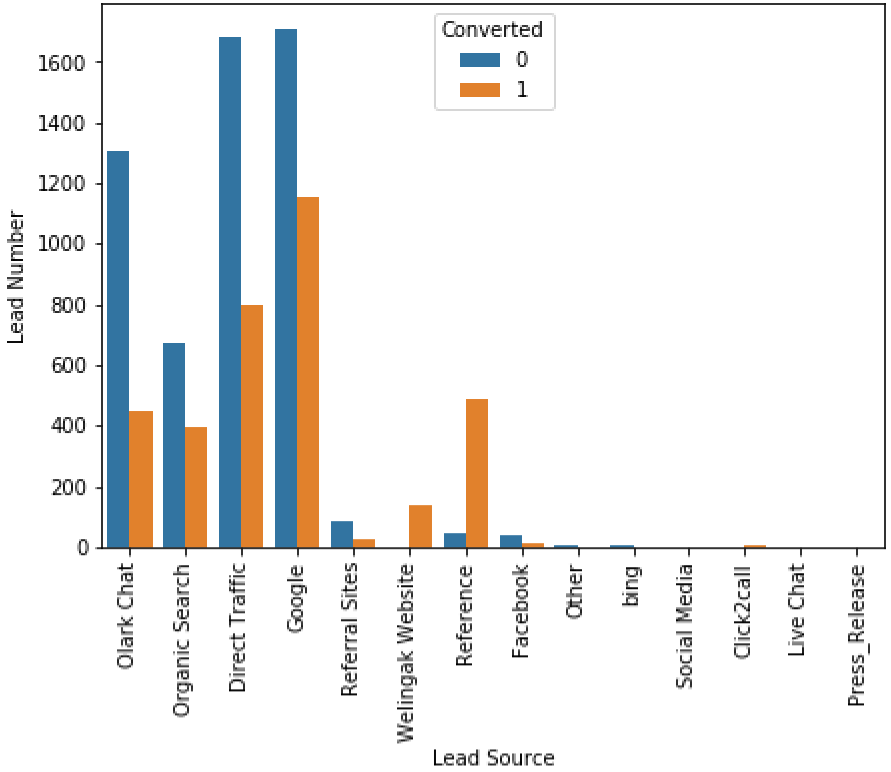


The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

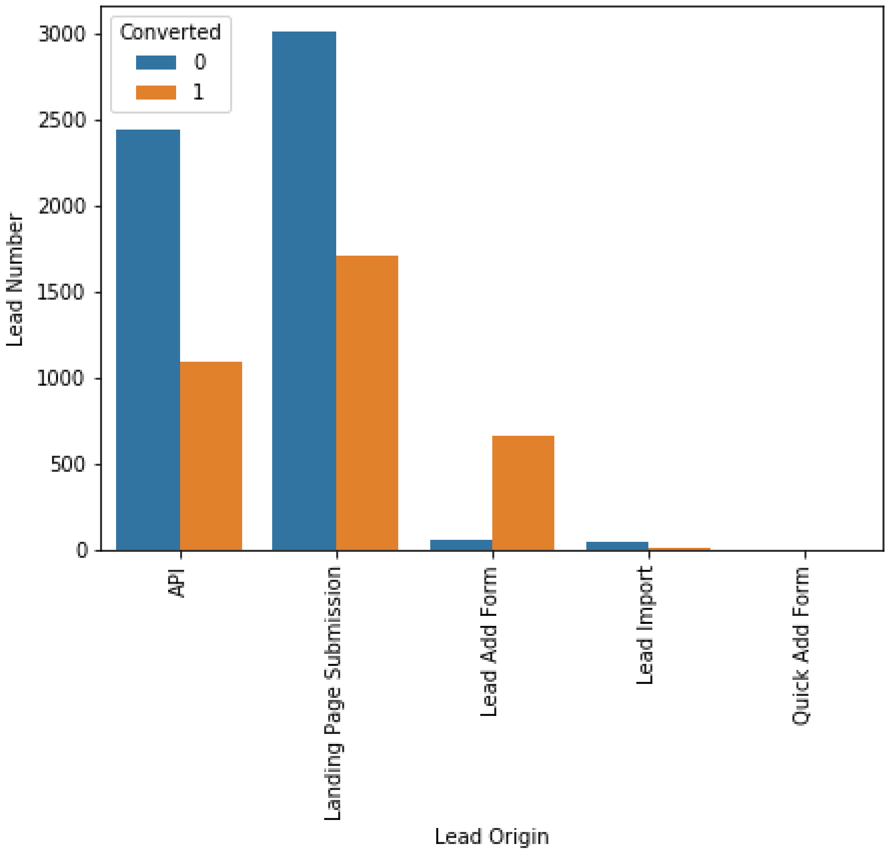


Major conversion in the occupation unemployed and in the lead Source Google

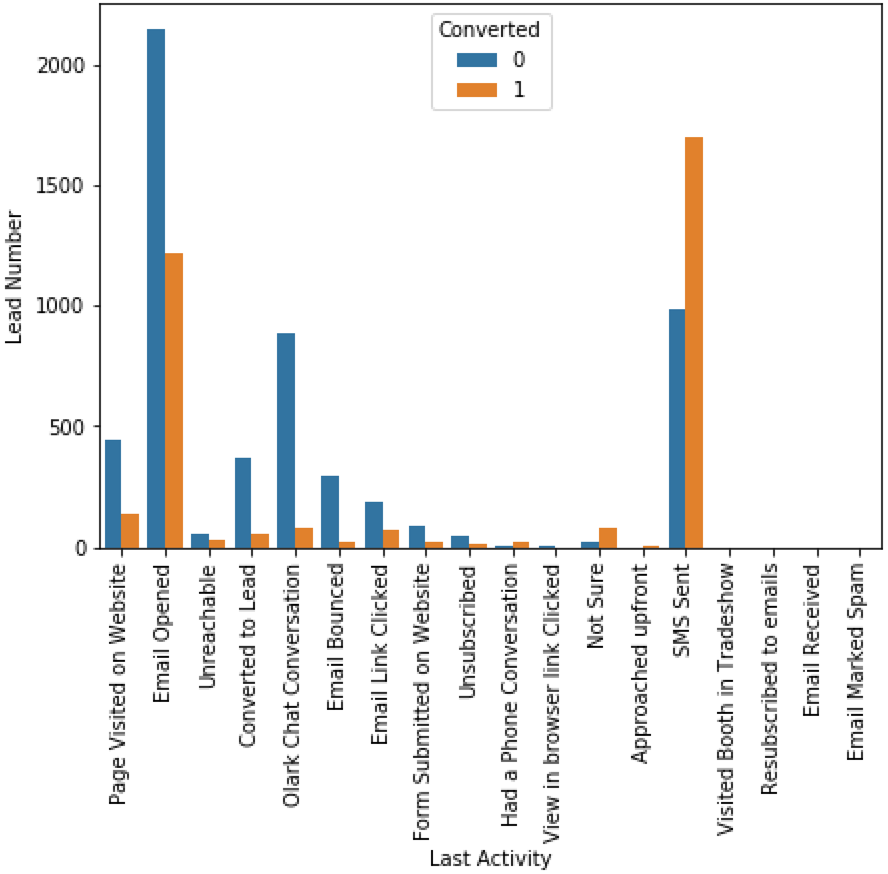




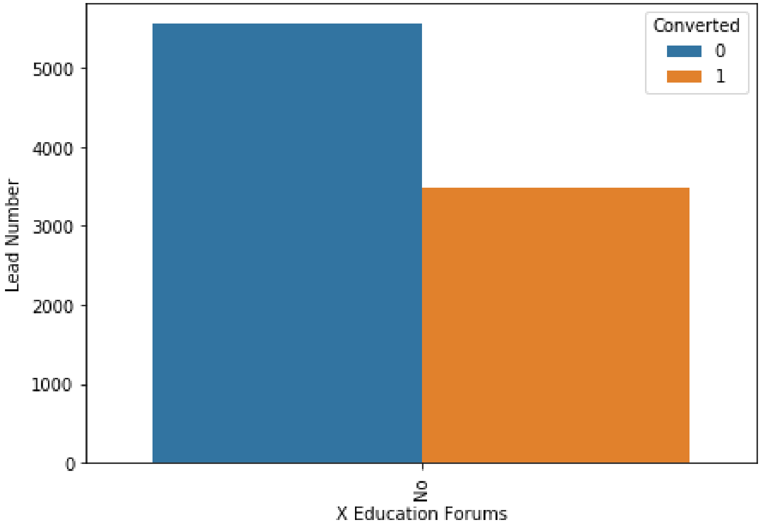
In Lead Origin, maximum conversion happened from Landing Page Submission

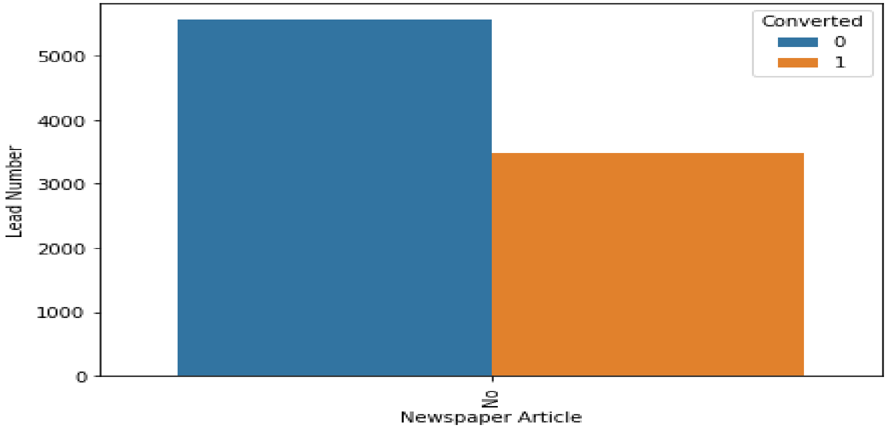


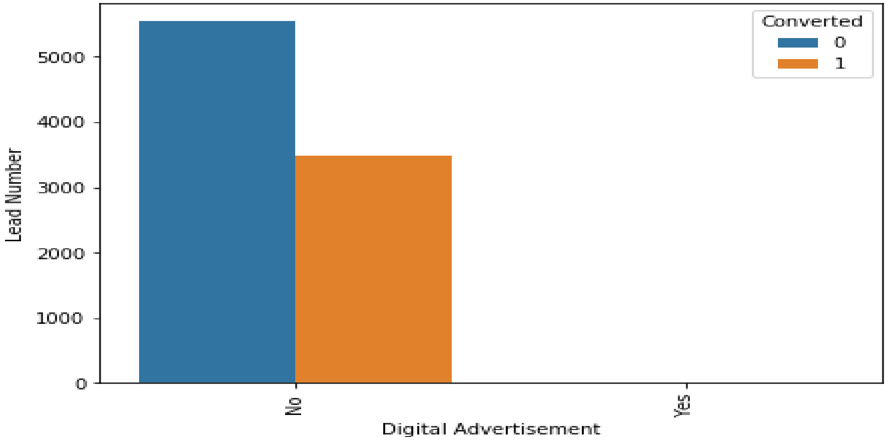
Major conversion has happened from SMS sent



Not much impact on conversion rates through Search, digital advertisements and through recommendations



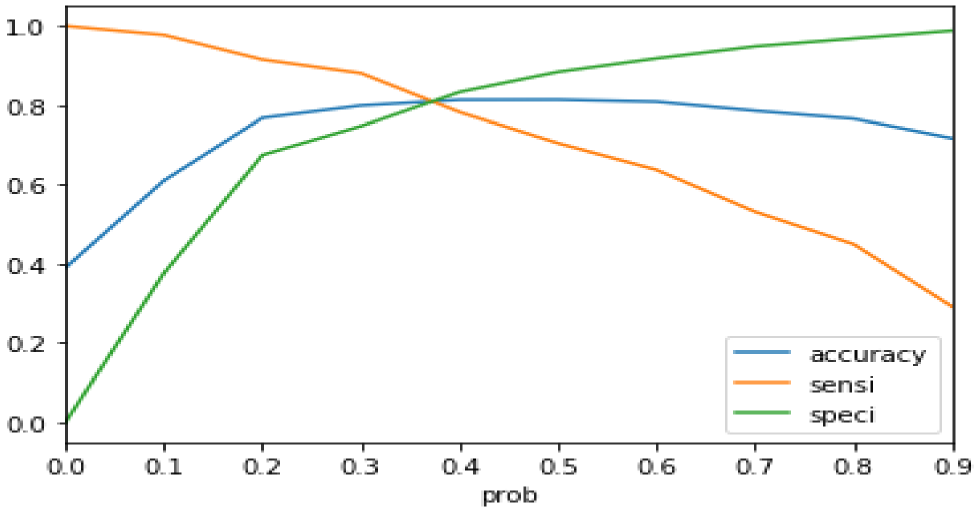




Variables Impacting the Conversion Rate

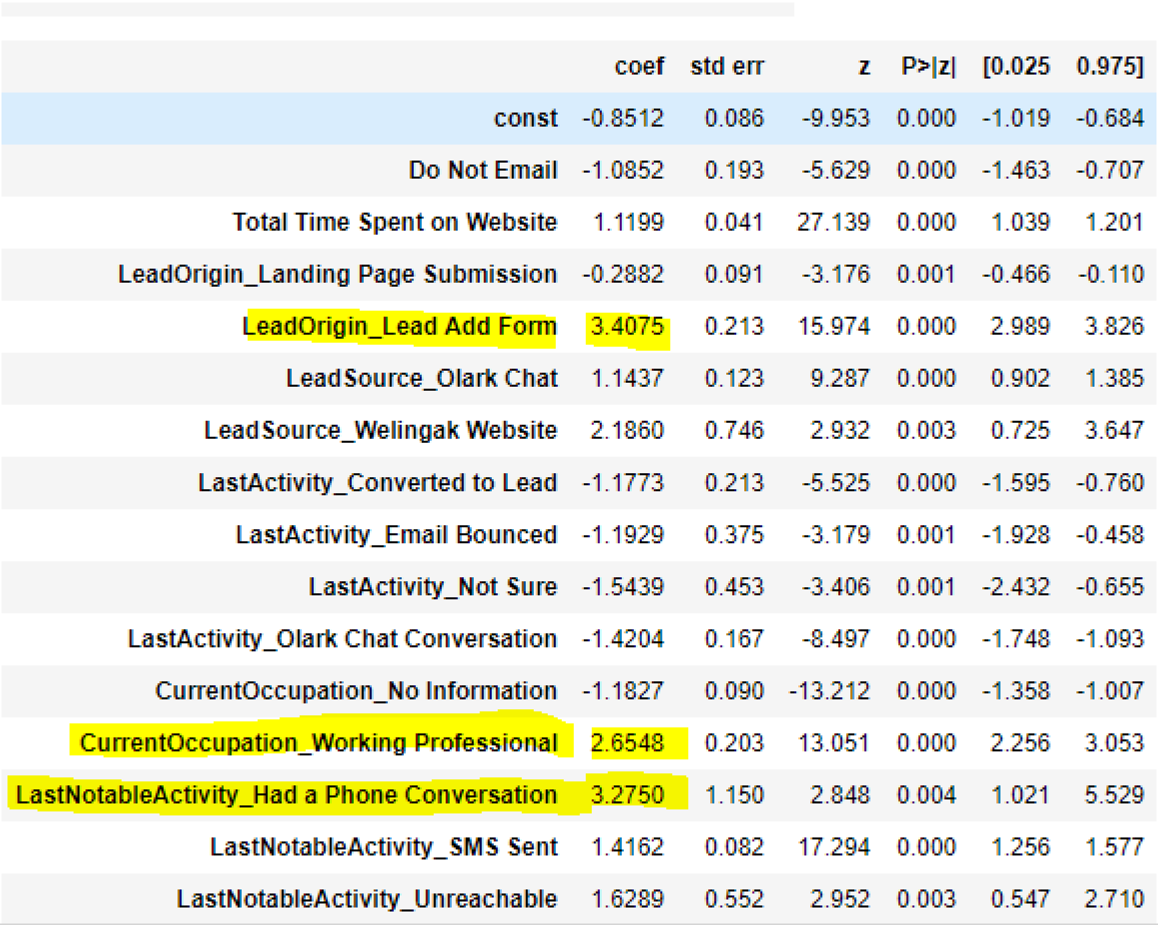
- Do Not Email
- Total Time Spent On Website
- LeadOrigin_API
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- LastActivity_Approached upfront
- LastActivity_Converted to Lead
- Last Activity – Olark Chat Conversation
- LastActivity_Not Sure
- CurrentOccupation_Housewife
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity – Unreachable
- LastNotableActivity_SMS Sent

Finding the optimal cutoff point



From the curve above, 0.37 is the optimum point to take it as a cutoff probability.

Top variables which gets leads converted(highlighted)



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8512	0.086	-9.953	0.000	-1.019	-0.684
Do Not Email	-1.0852	0.193	-5.629	0.000	-1.463	-0.707
Total Time Spent on Website	1.1199	0.041	27.139	0.000	1.039	1.201
LeadOrigin_Landing Page Submission	-0.2882	0.091	-3.176	0.001	-0.466	-0.110
LeadOrigin_Lead Add Form	3.4075	0.213	15.974	0.000	2.989	3.826
LeadSource_Olark Chat	1.1437	0.123	9.287	0.000	0.902	1.385
LeadSource_Welingak Website	2.1860	0.746	2.932	0.003	0.725	3.647
LastActivity_Converted to Lead	-1.1773	0.213	-5.525	0.000	-1.595	-0.760
LastActivity_Email Bounced	-1.1929	0.375	-3.179	0.001	-1.928	-0.458
LastActivity_Not Sure	-1.5439	0.453	-3.406	0.001	-2.432	-0.655
LastActivity_Olark Chat Conversation	-1.4204	0.167	-8.497	0.000	-1.748	-1.093
CurrentOccupation_No Information	-1.1827	0.090	-13.212	0.000	-1.358	-1.007
CurrentOccupation_Working Professional	2.6548	0.203	13.051	0.000	2.256	3.053
LastNotableActivity_Had a Phone Conversation	3.2750	1.150	2.848	0.004	1.021	5.529
LastNotableActivity_SMS Sent	1.4162	0.082	17.294	0.000	1.256	1.577
LastNotableActivity_Unreachable	1.6289	0.552	2.952	0.003	0.547	2.710

Model Evaluation – Sensitivity and Specificity on Train and Test Dataset

- Train data set

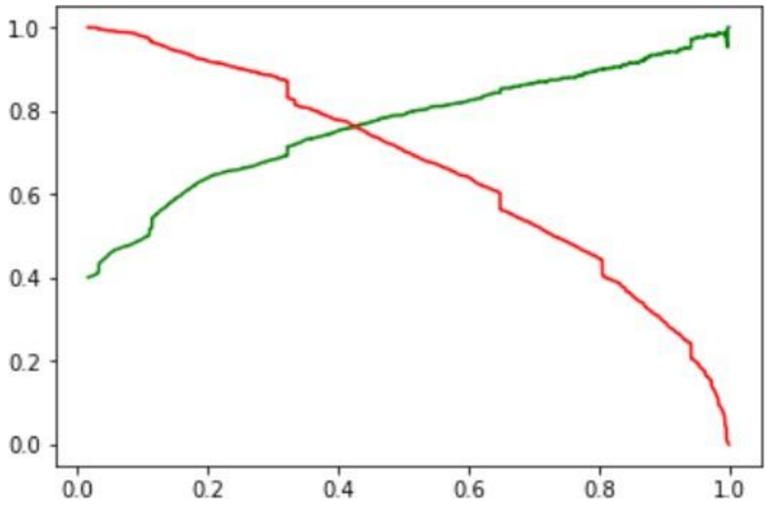
- Accuracy-81.4
- Sensitivity-80.8
- Specificity-81.75

Test Data set:

Accuracy-81.8
Sensitivity-80.79
Specificity-82.41

Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.42 based on Precision and Recall



Confusion Matrix

Train Data set
Precision:79.46
Recall:70.43

- Test Data set
- precision 73.34
- recall 80.78

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 82%, 81% and 82% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 81% (in train set) and 82% in test set
- The top 3 variables that contribute for lead getting converted in the model are
  - Lead Add Form from Lead Origin
  - Had a Phone Conversation from Last Notable Activity
  - Current occupation working professional
- Hence overall this model seems to be good.