LEAD SCORING CASE STUDY

By Nishant Kumar and Niharika Girdhar

PROBLEM STATEMENT

“X Education” sells online courses to industry professionals. They market courses on several websites and search engines like Google.

Once these people arrived at website, they may browse the courses or fill up a form from the course or watch some videos. When these people fill up form providing their email address or phone number, they are classified to be a lead.

Business Goal:

X education needs help in selecting the most promising leads (that are most likely to convert into paying customers.

The company needs a model where we must assign a lead score to each of the leads. Higher score -> ‘Hot Leads’ means have higher chances of conversion and Lower score -> ‘Cold Leads’ means lower conversion rates

The CEO, has given the target lead conversion rate to be around 80%.

STRATEGY

- Data Cleaning and Data Preparation
- Exploratory Data Analysis
- Feature Scaling
- Splitting the data into Train and Test Dataset
- Building a Logistic Regression model and calculate Lead Score
- Model Evaluation by using different metrics:-
  - Specificity
  - Sensitivity
  - Accuracy
- Applying the best model in Test data based on Sensitivity and Specificity Metrics.

Problem Solving And Methodology

Data Cleaning and Preparation

- Check duplicate data if exists then will remove it
- Clean the data by removing null values for analysis
- Outlier Treatment
- EDA

Feature Scaling and Train-Test Split

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

Model Building

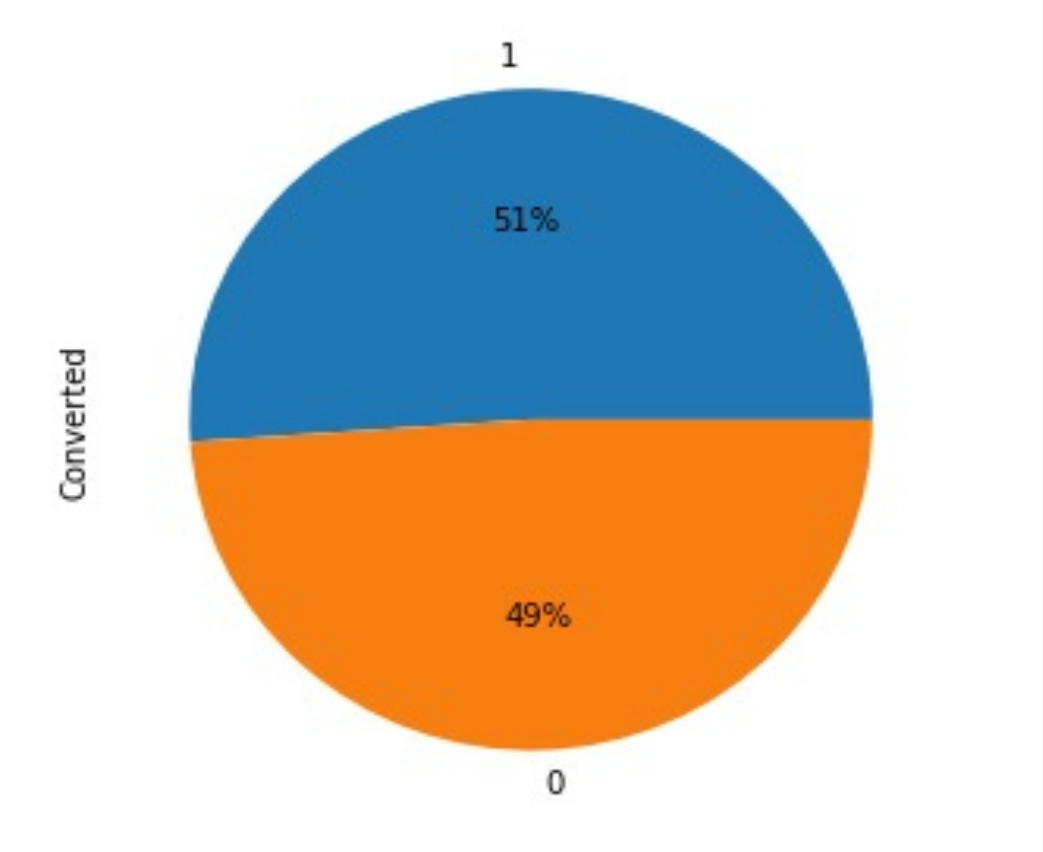
- Feature Selection using RFE
- Develop the optimal model by using Logistic Regression
- Evaluating the model with various metrics:- Sensitivity, Specificity, Accuracy , Precision and Recall

Result

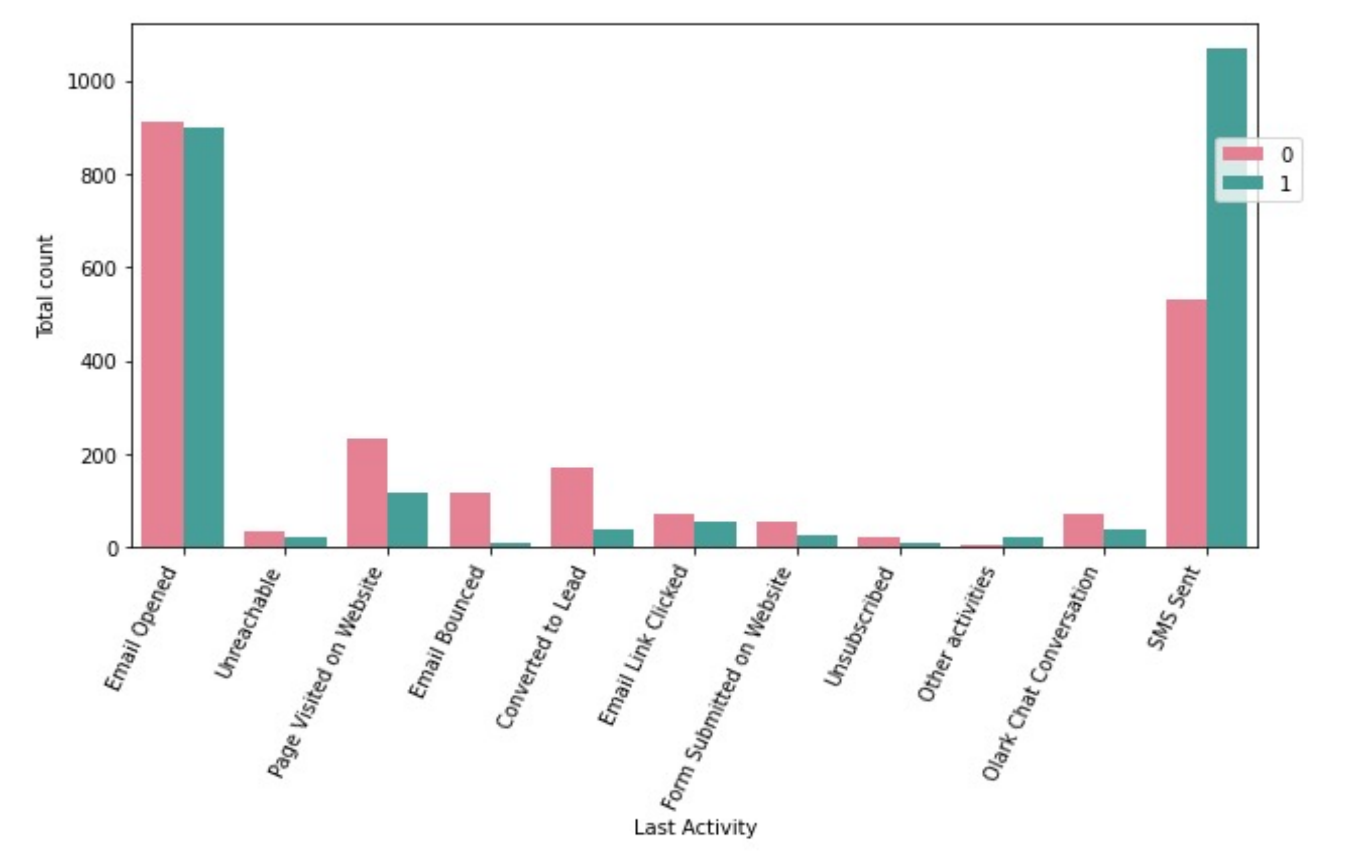
- Assign Lead Score and check if final prediction amount to 80% conversion rate.
- Final Evaluation on Test dataset using cut-off threshold from Sensitivity and Specificity metrics

Exploratory Data Analysis

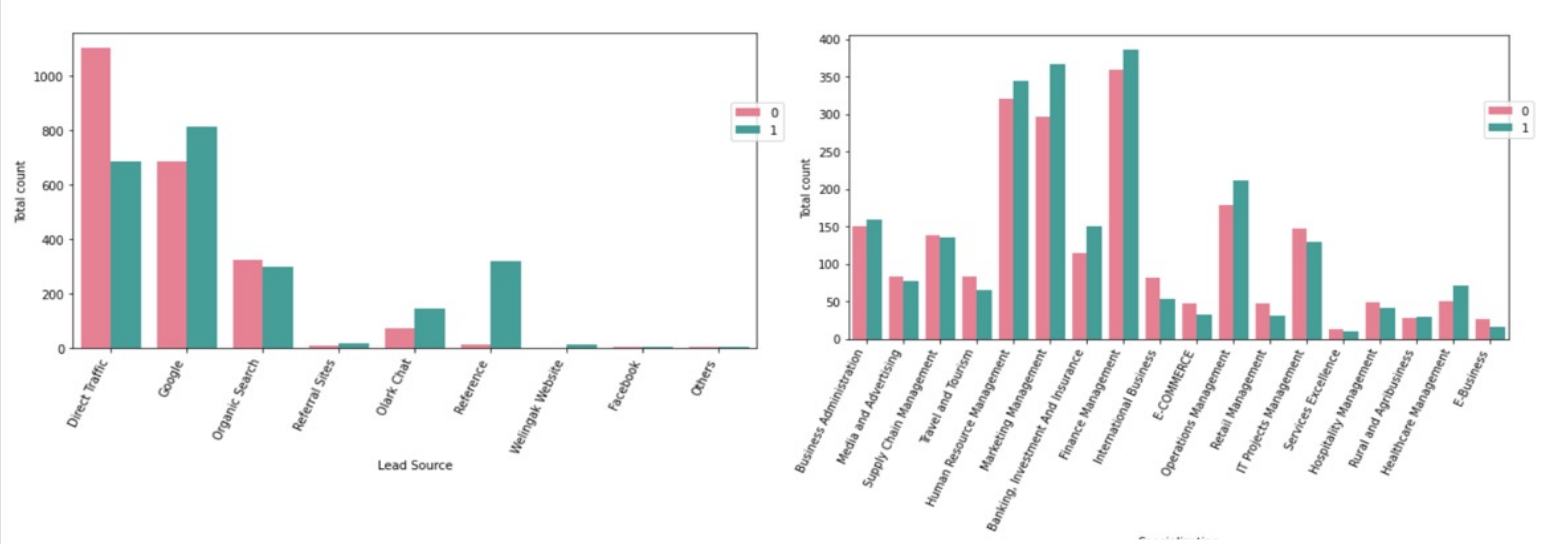
We have around 51% Conversion rate



Conversion rate for leads with last activity as SMS Sent.



Exploratory Data Analysis



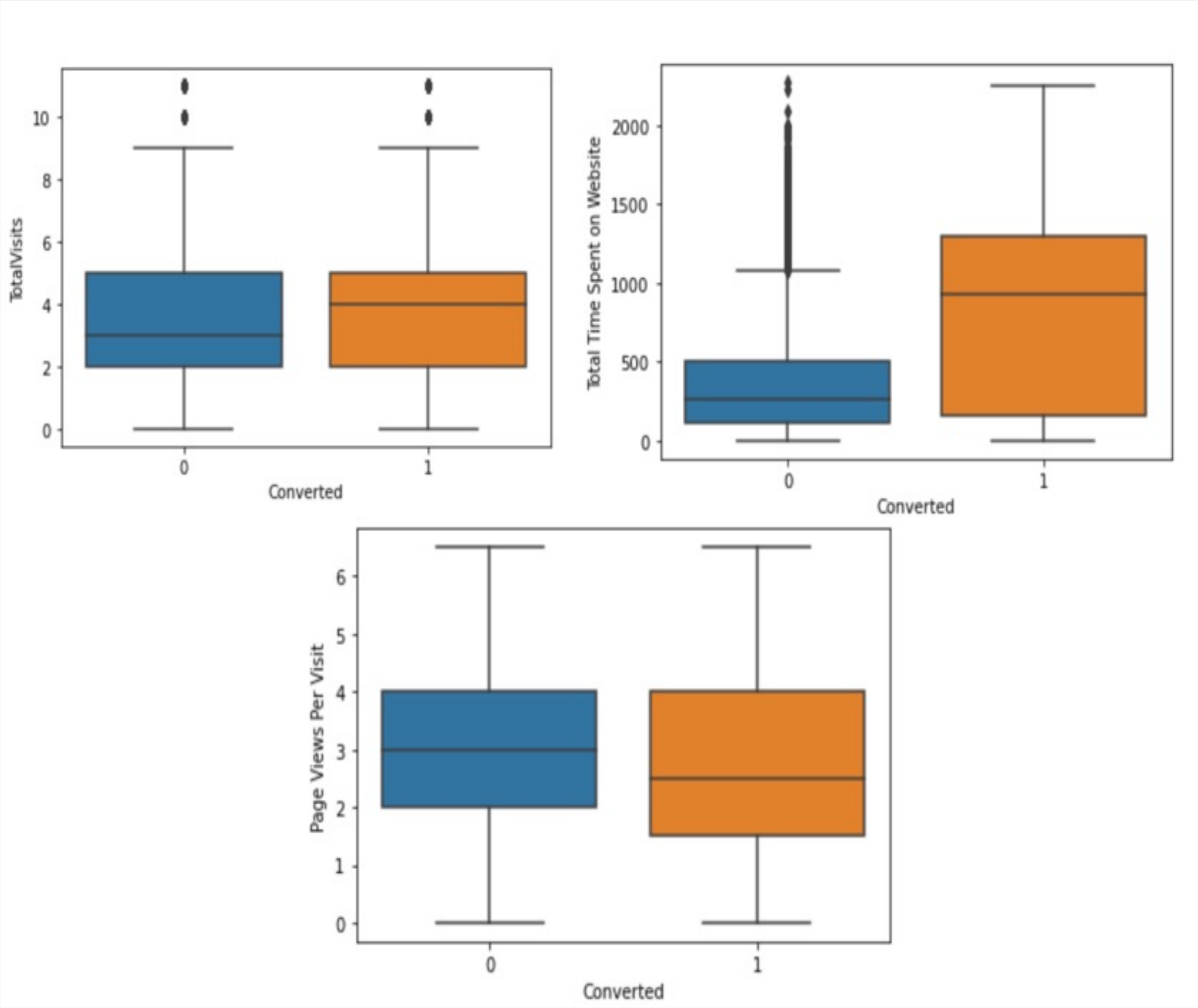
1. Google and Direct Traffic generated maximum number of leads. However, count is less in Direct Traffic as compared to non-converted leads.
2. Company needs to focus on leads who hold specialization in management fields like Finance, HR, Marketing, Operations and the leads in Banking, Investment & Insurance fields.

Approach to the analysis

Our analysis was started with data inspection and data cleaning by removing all the null values. We also checked there were values where leads didn’t select any of the options in survey. So, we treated those values as null.

In addition to this, we checked the outliers in the dataset. In the visualization we found there were many outliers. These were handled by capping these values in 95 percentile.

Outliers in logistic model is very sensitive and hence we dealt it without loosing any information. We also dropped certain features which would not contribute to our analysis as many of them had very less variance.

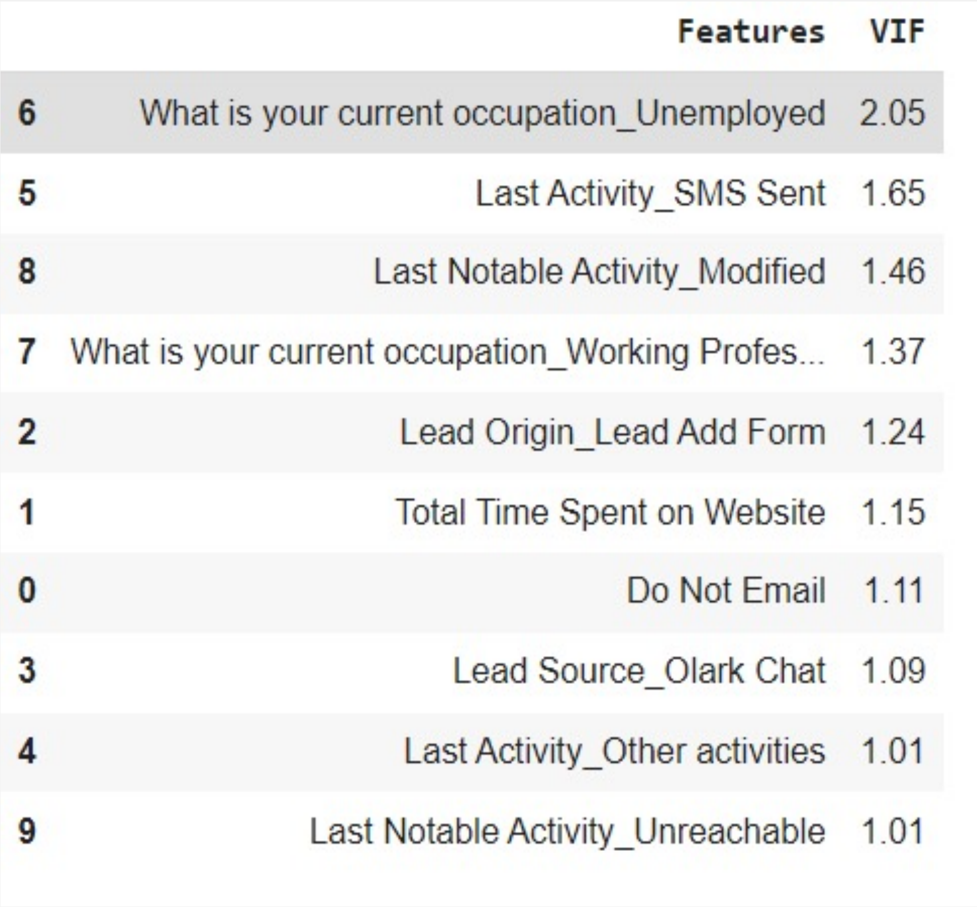


Model Building

We moved to model building phase by using RFE (Recursive Feature Elimination) as we found large number of features in dataset. To deal with these, the best approach is to select small set of features from a pool of features using RFE.

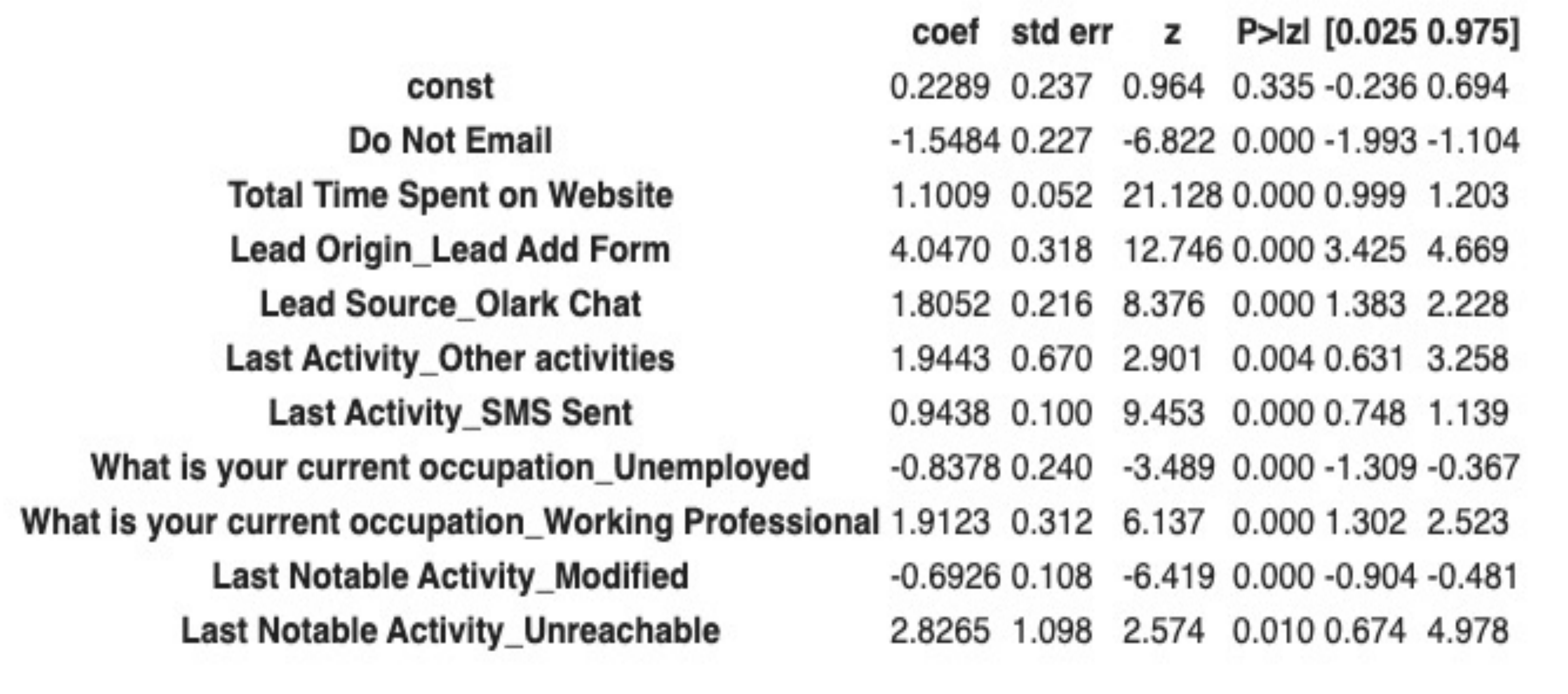
We have used stats model as well to check which of these features are significant to consider. Higher P value means the feature is insignificant, so we dropped those features.

Variance Inflation factor(VIF) was also checked to look how much behaviour of an independent feature is influenced. We set normal threshold of VIF < 5 for significance. Total 10 features were found significant as shown in this slide.



	Features	VIF
6	What is your current occupation_Unemployed	2.05
5	Last Activity_SMS Sent	1.65
8	Last Notable Activity_Modified	1.46
7	What is your current occupation_Working Profes...	1.37
2	Lead Origin_Lead Add Form	1.24
1	Total Time Spent on Website	1.15
0	Do Not Email	1.11
3	Lead Source_Olark Chat	1.09
4	Last Activity_Other activities	1.01
9	Last Notable Activity_Unreachable	1.01

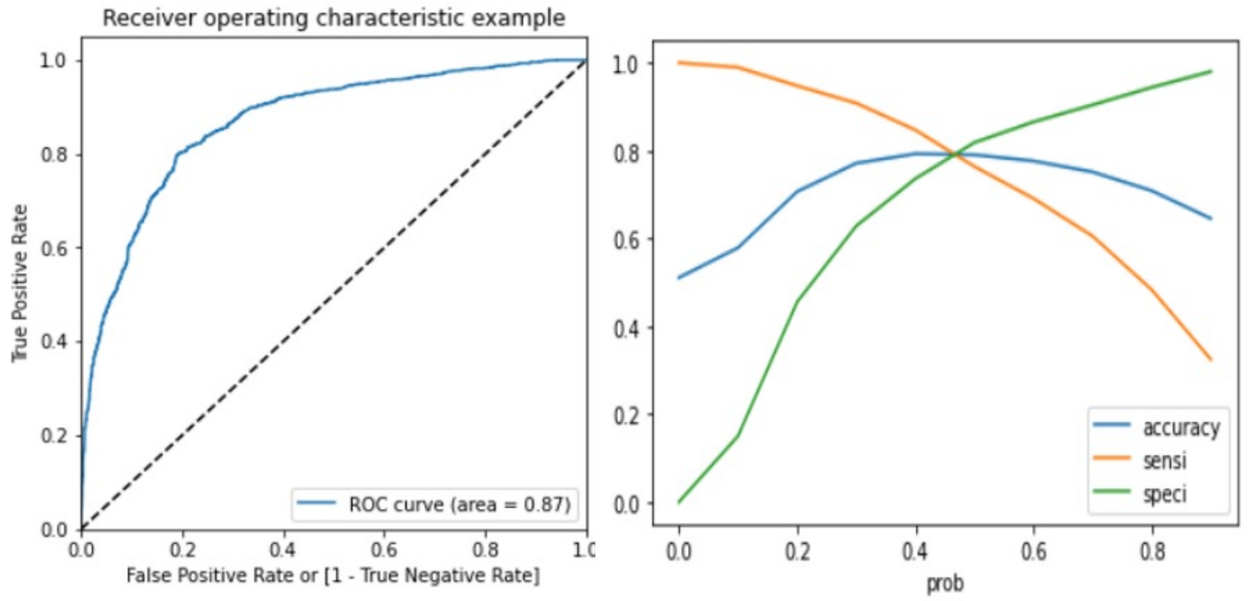
VARIABLES IMPACTING THE CONVERSION RATE



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2289	0.237	0.964	0.335	-0.236	0.694
Do Not Email	-1.5484	0.227	-6.822	0.000	-1.993	-1.104
Total Time Spent on Website	1.1009	0.052	21.128	0.000	0.999	1.203
Lead Origin_Lead Add Form	4.0470	0.318	12.746	0.000	3.425	4.669
Lead Source_Olark Chat	1.8052	0.216	8.376	0.000	1.383	2.228
Last Activity_Other activities	1.9443	0.670	2.901	0.004	0.631	3.258
Last Activity_SMS Sent	0.9438	0.100	9.453	0.000	0.748	1.139
What is your current occupation_Unemployed	-0.8378	0.240	-3.489	0.000	-1.309	-0.367
What is your current occupation_Working Professional	1.9123	0.312	6.137	0.000	1.302	2.523
Last Notable Activity_Modified	-0.6926	0.108	-6.419	0.000	-0.904	-0.481
Last Notable Activity_Unreachable	2.8265	1.098	2.574	0.010	0.674	4.978

MODEL EVALUATION USING ROC

- We then moved on to Model Evaluation step of training dataset using metrices like Accuracy, Sensitivity, Specificity, Precision and Recall. For training dataset:-
  - Accuracy : 80%.
  - Sensitivity : 79.6%,
  - Specificity : 81%.
- Area under ROC curve was 0.87, which reflects a good model
- To find optimal cut off probability to classify the leads to be converted or not, sensitivity and specificity values should be balanced. The graph shows an optimal cutoff is 0.48 based above three metrics.
- A "Lead Score" was also assigned to each leads by multiplying Converted Probability with 100.



PREDICTIONS ON TEST DATASET

- After training the model, we moved to make predictions on test dataset using the same metrics for evaluating training dataset.
- For test dataset:

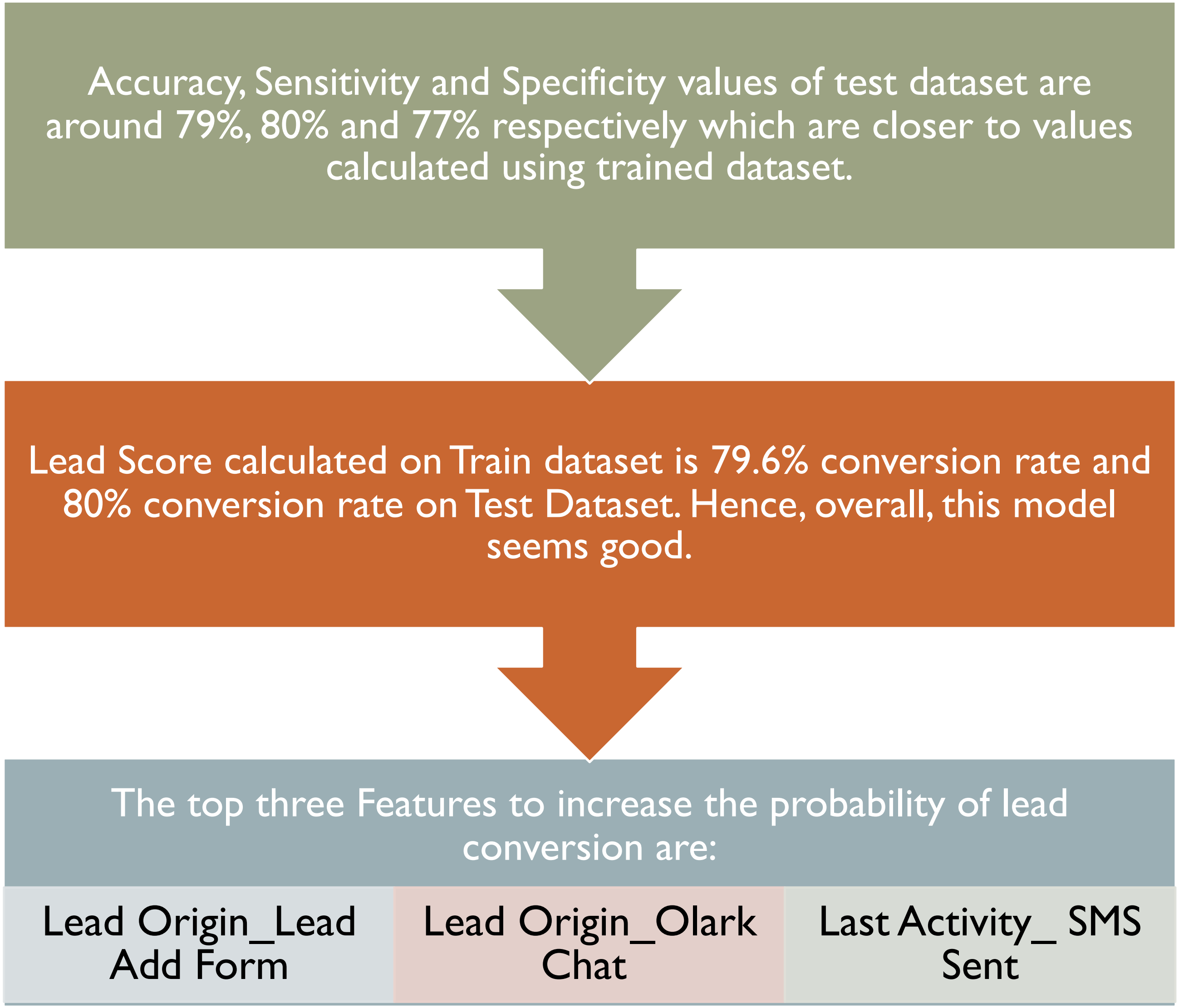
  a. Accuracy : 78.7%.

  b. Sensitivity : 80%

  c. Specificity : 77.1%.

- Our model can predict leads which are truly converted (sensitivity) and nonconverted (specificity) accurately.
- Precision score in test dataset is 79% which means when the model predicts a lead to be converted, its 79% correct.
- Recall score is 80% which means the model correctly identifies 80 % of all converted leads.

CONCLUSION



Accuracy, Sensitivity and Specificity values of test dataset are around 79%, 80% and 77% respectively which are closer to values calculated using trained dataset.

Lead Score calculated on Train dataset is 79.6% conversion rate and 80% conversion rate on Test Dataset. Hence, overall, this model seems good.

The top three Features to increase the probability of lead conversion are:

- Lead Origin_Lead Add Form
- Lead Origin_Olark Chat
- Last Activity_ SMS Sent

RECOMMENDATIONS

Company should focus on the leads who spend their most of the time on website. 'Time Spent On Website' variable will help in that.

Try to nurture more to the leads who holds specialization in management. They have high probability to be as "Hot Leads".

Automated response email system should be developed so that potential leads can receive a quick response.

Phone calls should be minimized for "Hot Leads" to deliver hassle free customer experience.

THANK YOU