Lead Scoring Case Study

Ramya D & Mounica N, DSC 27

June 2021

Contents

- Key Objectives
- Inspecting Applications Dataset
- Data Cleaning and Transformation
- Exploratory Data Analysis
- Splitting dataset
- Model Building
- Model Evaluation – Train Dataset
- Model Evaluation – Test Dataset
- Conclusion

Key Objectives

Identify potential hot leads among different a pool of online visitors

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. For this purpose, we will build a predictive model to classify the visitors into potential hot leads and non leads.

Inspecting Dataset

Given dataset contains information of visitors of the X Education website along with their responsive actions.

- The dataset ‘Leads.csv' contains all the information of the visitors. This data talks about whether a visitor is a potential lead or not.

- A total of 37 attributes for 9240 visitors with their history of being converted to a lead or not is given in this csv file

- All the applicants have a unique Lead ID and Prospect ID which is the id given by the X Education to that visitor. The column “Converted” is an indicator variable (has values 0/1) which gives us the information whether that visitor was converted as a customer or not.

- A quick run of .info() method on this data frame shows that there are 4 float64 columns, 3 int columns and 30 object type columns.

Data Cleaning and Transformation

- Since the columns Lead Number and Prospect ID are not relevant for modelling, we dropped these columns in this phase. Also, there are few columns which have only a constant value. We dropped these columns as well as they do not have any predictive information in them.

- In the next step we observed that there are quite a few columns with missing percentages in the range of 20% to 70% and decided to drop all those columns who have a missing percentage greater than 35% . Prior to this step we replaced “Select”, “select” values in the dataset with nan so that those observations are properly addressed in the data cleaning stage.

- For a few of the remaining columns that have a missing percentage values less than 35%, we decided to create a level called “not provided” so that we do not lose information either due to dropping or imputing. For the rest we have imputed the missing values with the mode of the respective categorical columns.

- After this step, we reduced the levels in categorical columns with an exceedingly high number of levels to a reasonable number of levels with appropriate data transformations.

Data Cleaning and Transformation

- Next, we converted categorical columns with binary responses to numeric columns (0-1 mapping) and data types of the rest of object columns to category.

- For the categorical variables that have more than 3 levels, we use one hot encoding to create dummy variables for each of those variables.

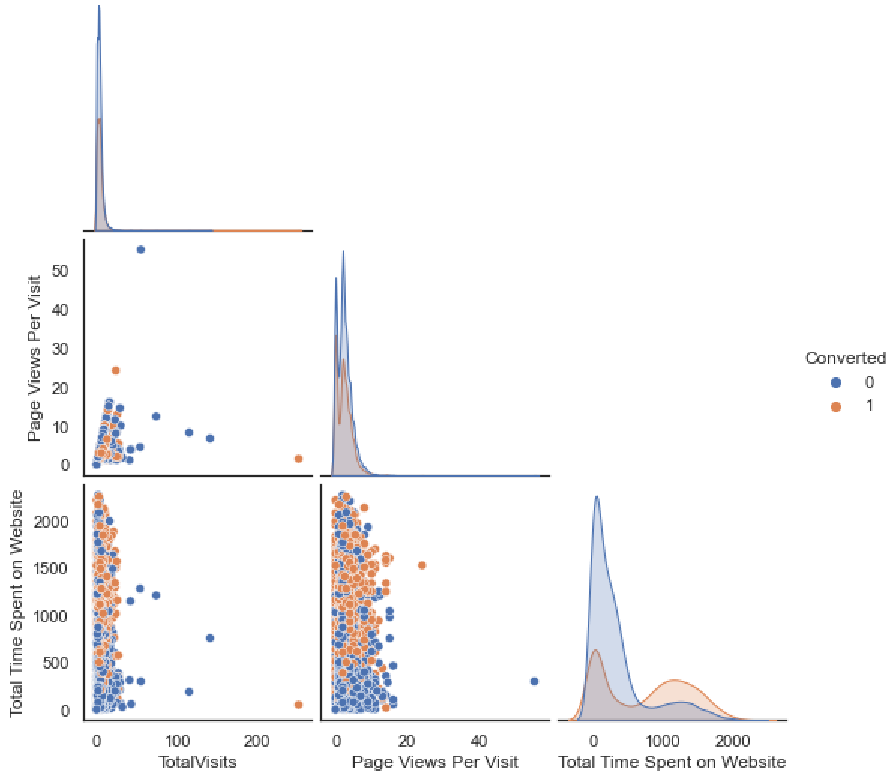
- After all the above data processing steps, we did a outlier analysis on the three numerical variables present in the dataset. We found that out of these 3 columns, there are not outliers present in one column, and the rest of the two columns have outliers in the order of 200. Since this is a significant number of observations, instead of dropping these rows we decided to transform the numerical variable to categorical variable by binning.

- For this purpose, we used the quantiles of the column to come with the relevant boundaries for the level. After creating these bins, we again transformed these two categorical variables using one hot encoding

Exploratory Data Analysis – Numerical Columns

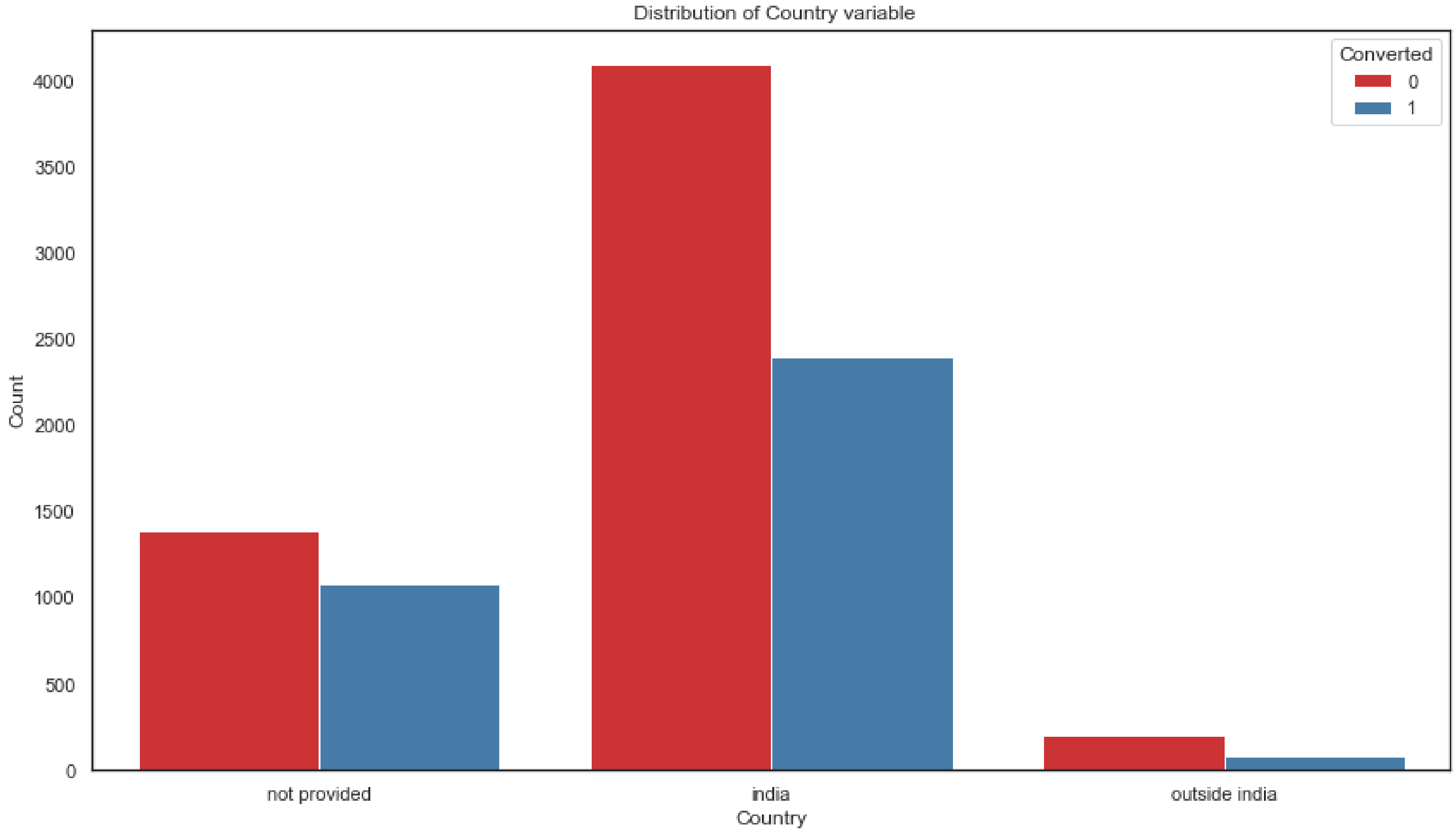
- It is clear from the pair plot below that there is no distinct trend among both the populations of visitors who got converted and who didn't with respect to the variables Total Visits and Page Views per visits.

- However, from the density plot of Total Time spent on the website, visitors who got converted tend to spend a high time on the website and the visitors who don't get converted spend almost a very little time (mode just after near to 0) on the platform.



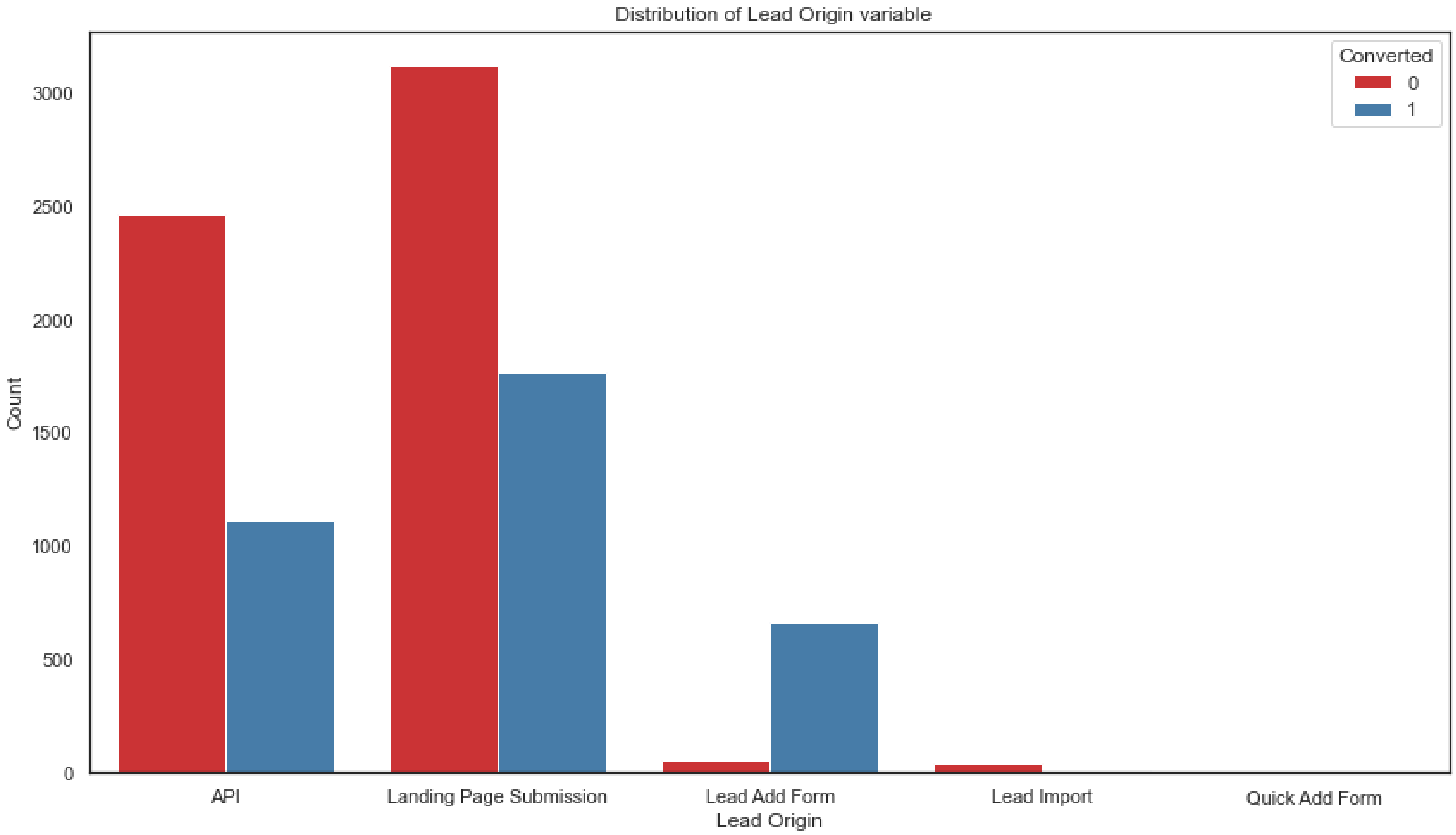
Exploratory Data Analysis – Categorical Columns

- The distribution of the categorical column country is almost the same in the cases of bot converted and non converted vistors.



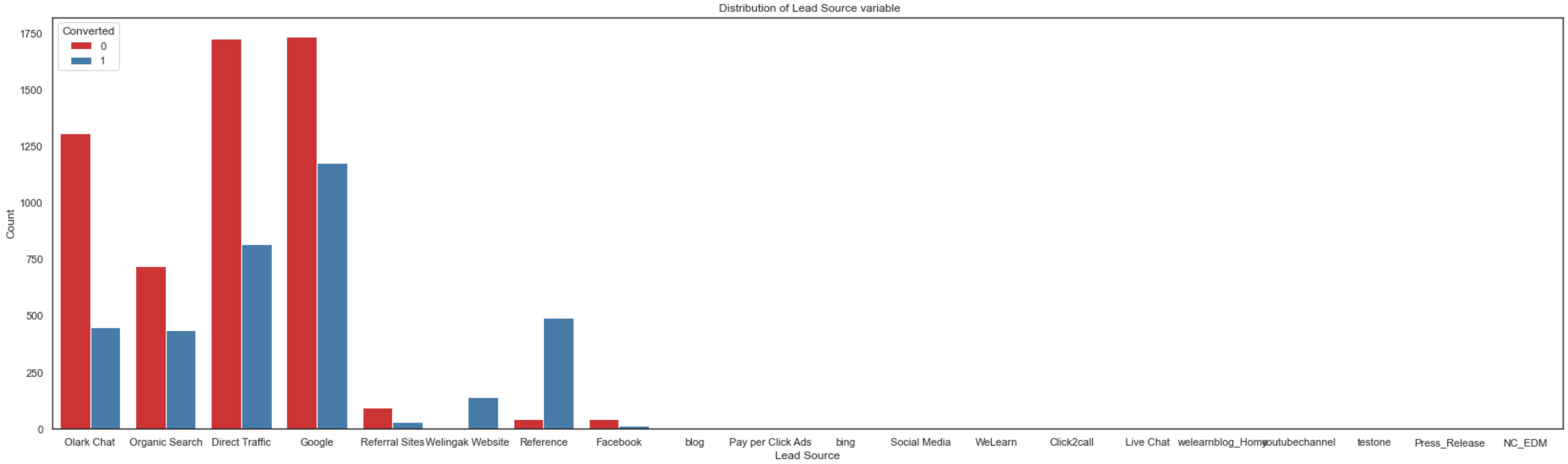
Exploratory Data Analysis – Categorical Columns

- Majority of the visitors have Lead Origin as API and Landing Page Submission. An interesting point to note here is that visitors who has Lead Origin as Lead add form tend to have a higher rate of conversion.



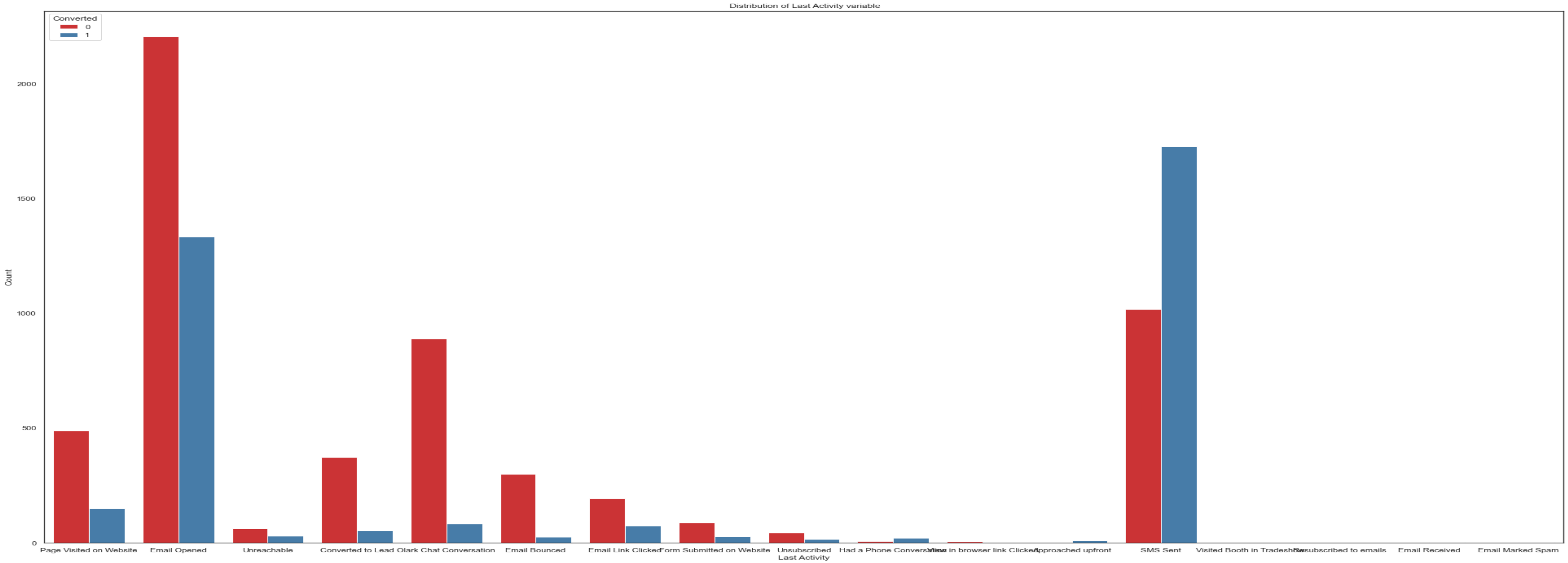
Exploratory Data Analysis – Categorical Columns

- Majority of the visitors have Lead Source as olark chat,organic search, direct traffic and google. An interesting point to note here is that visitors who has Lead Source as referred tend to have a higher rate of conversion.



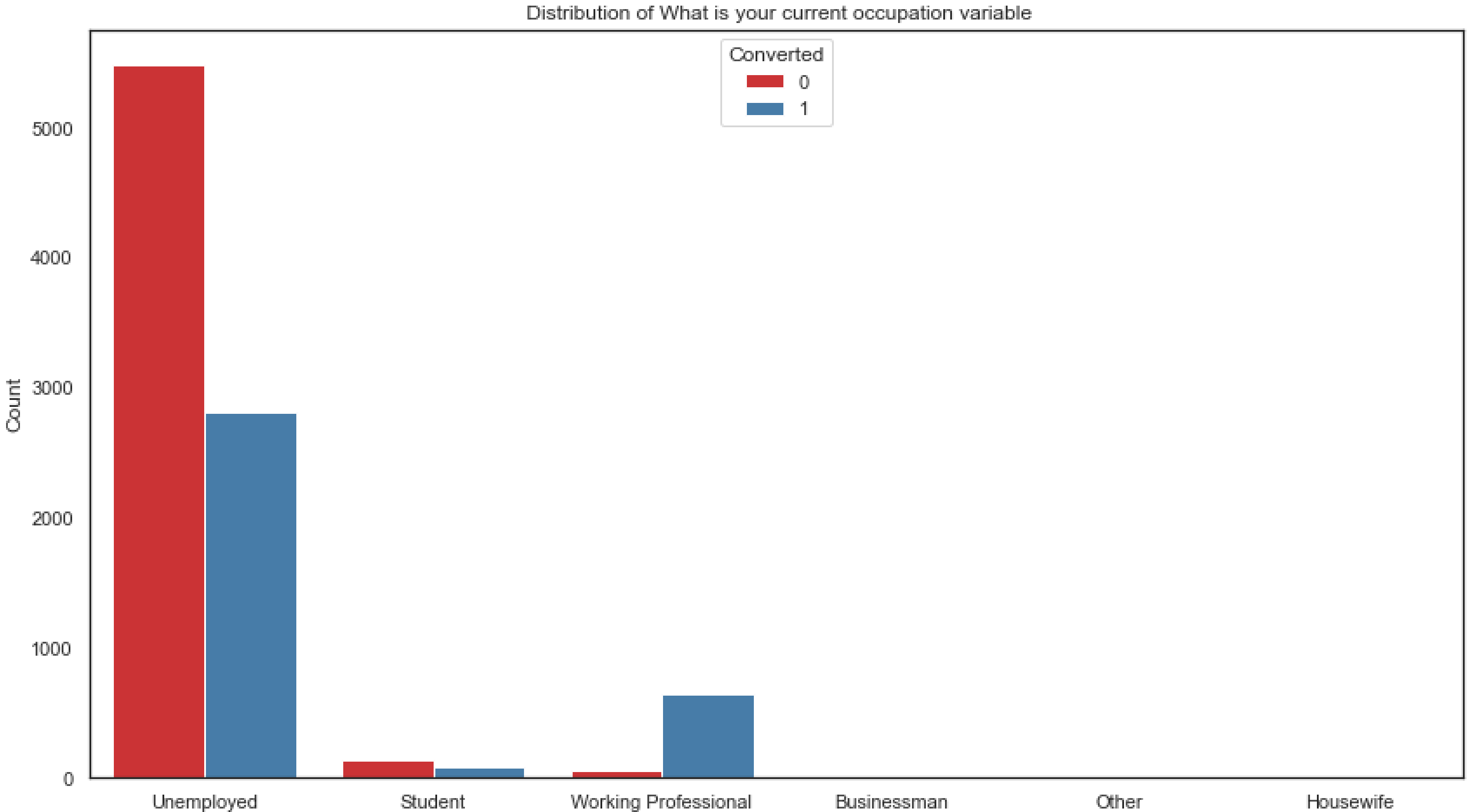
Exploratory Data Analysis – Categorical Columns

- From the bar graph below we can see that visitors who has Last Activity as sms sent generally get converted more.



Exploratory Data Analysis – Categorical Columns

- From the bar graph below we can see that visitors who are working professionals in What is your current occupation generally tend to get converted more.



Scaling And Splitting The Dataset

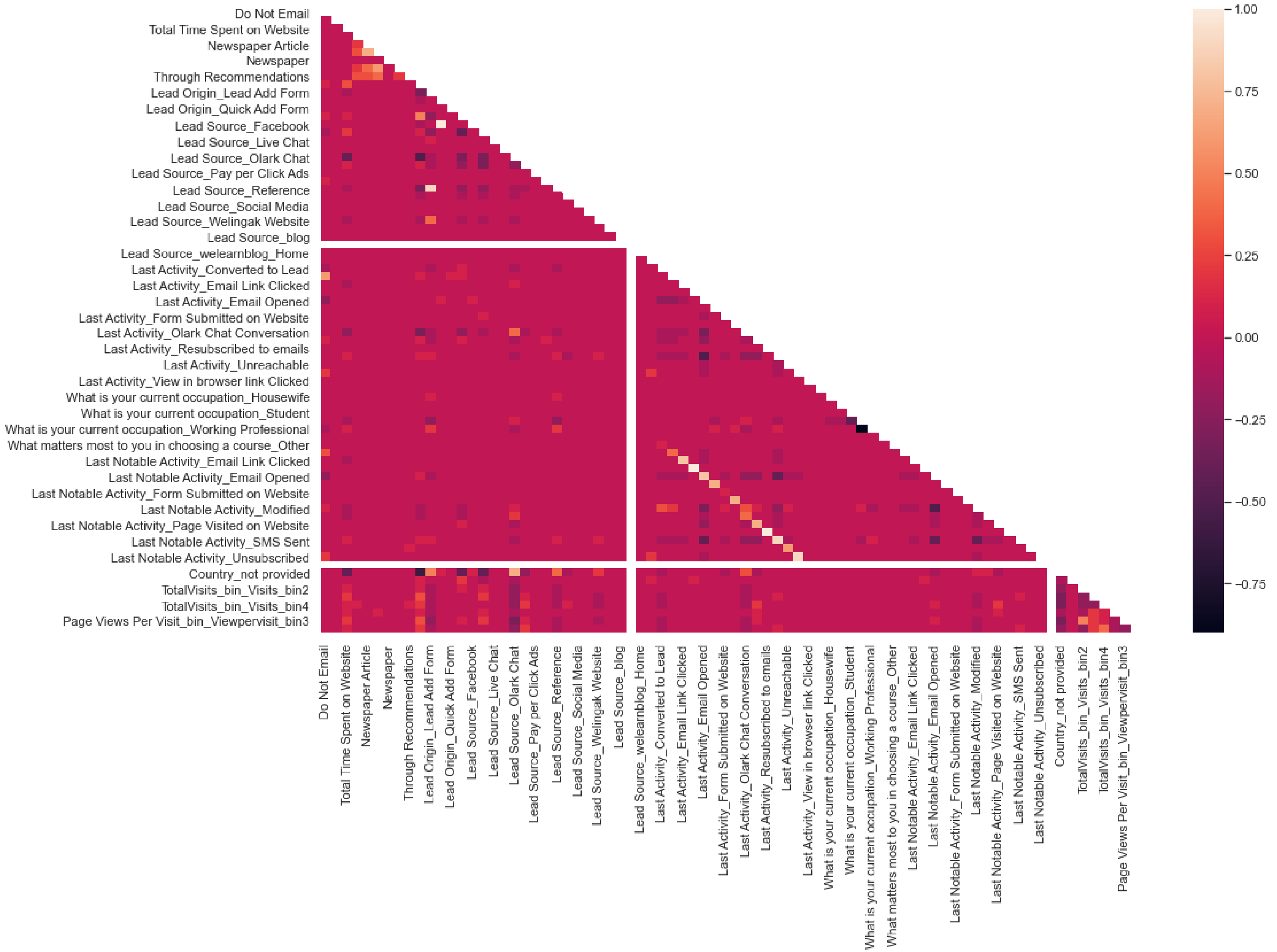
- Split the dataset into train and test dataset with a a split of 70% : 30% and scaled the train dataset using the standard transformer .

- Did a detailed analysis on correlations and plotted heatmaps to fine the dependencies between the variables.

Train (70%)

Test (30%)

Correlation Analysis



Model Building

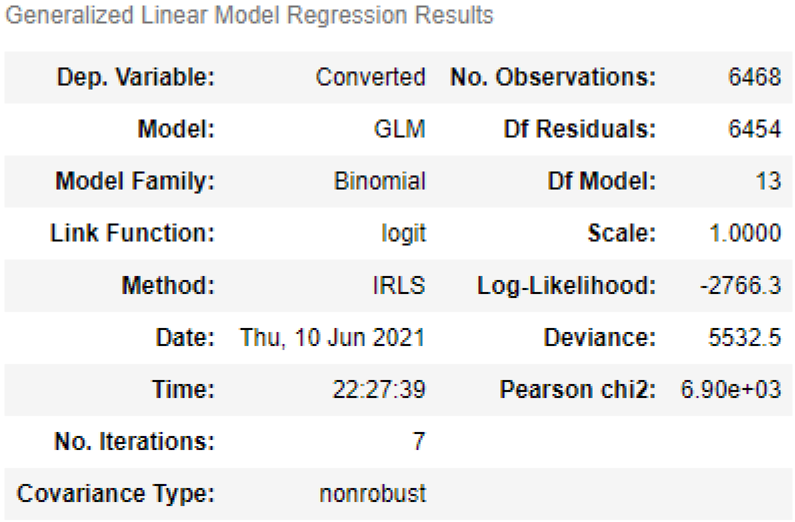
- Used the standard logistic regression model with recursive feature elimination (RFE) to select the top 20 variables.

- After feature selection step, we started iteratively updating the model by removing a variable at time so that there is no multi collinearity present in the model and all the VIF values are < 2 in the model.

- Next, we removed all the variables which have an insignificant coefficient. For achieving this we had a threshold of 5% of p-values.

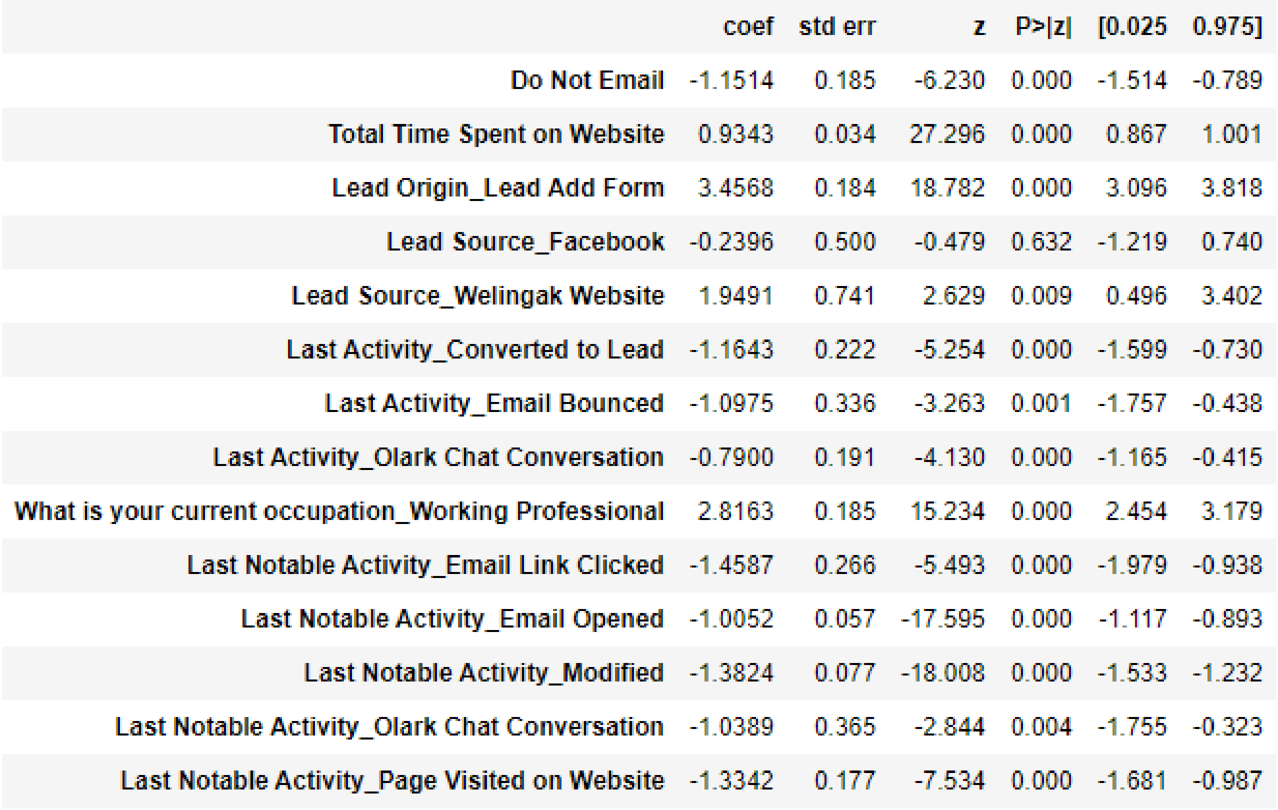
- At the end of eight such model building iterations our final model has 14 variables.

Model Building : Final Model Summary



Generalized Linear Model Regression Results

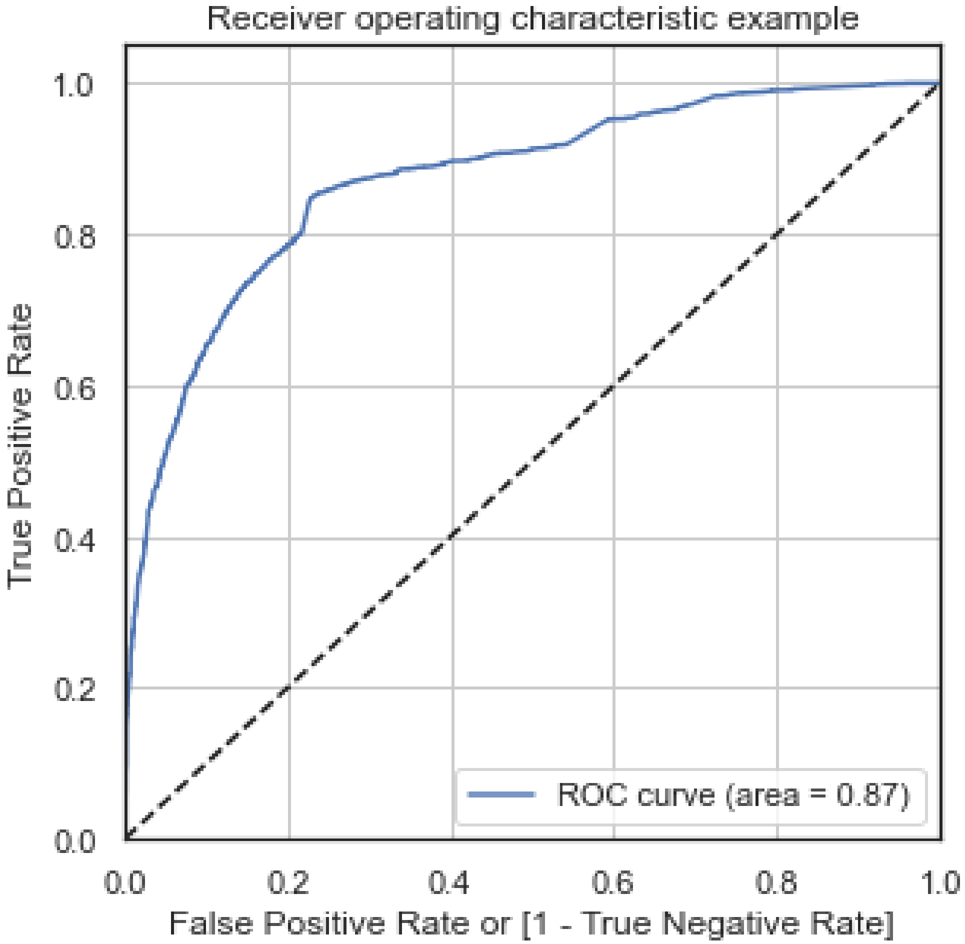
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6454
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2766.3
Date:	Thu, 10 Jun 2021	**Deviance:**	5532.5
Time:	22:27:39	**Pearson chi2:**	6.90e+03
No. Iterations:	7		
Covariance Type:	nonrobust		



	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Do Not Email	-1.1514	0.185	-6.230	0.000	-1.514	-0.789
Total Time Spent on Website	0.9343	0.034	27.296	0.000	0.867	1.001
Lead Origin_Lead Add Form	3.4568	0.184	18.782	0.000	3.096	3.818
Lead Source_Facebook	-0.2396	0.500	-0.479	0.632	-1.219	0.740
Lead Source_Welingak Website	1.9491	0.741	2.629	0.009	0.496	3.402
Last Activity_Converted to Lead	-1.1643	0.222	-5.254	0.000	-1.599	-0.730
Last Activity_Email Bounced	-1.0975	0.336	-3.263	0.001	-1.757	-0.438
Last Activity_Olark Chat Conversation	-0.7900	0.191	-4.130	0.000	-1.165	-0.415
What is your current occupation_Working Professional	2.8163	0.185	15.234	0.000	2.454	3.179
Last Notable Activity_Email Link Clicked	-1.4587	0.266	-5.493	0.000	-1.979	-0.938
Last Notable Activity_Email Opened	-1.0052	0.057	-17.595	0.000	-1.117	-0.893
Last Notable Activity_Modified	-1.3824	0.077	-18.008	0.000	-1.533	-1.232
Last Notable Activity_Olark Chat Conversation	-1.0389	0.365	-2.844	0.004	-1.755	-0.323
Last Notable Activity_Page Visited on Website	-1.3342	0.177	-7.534	0.000	-1.681	-0.987

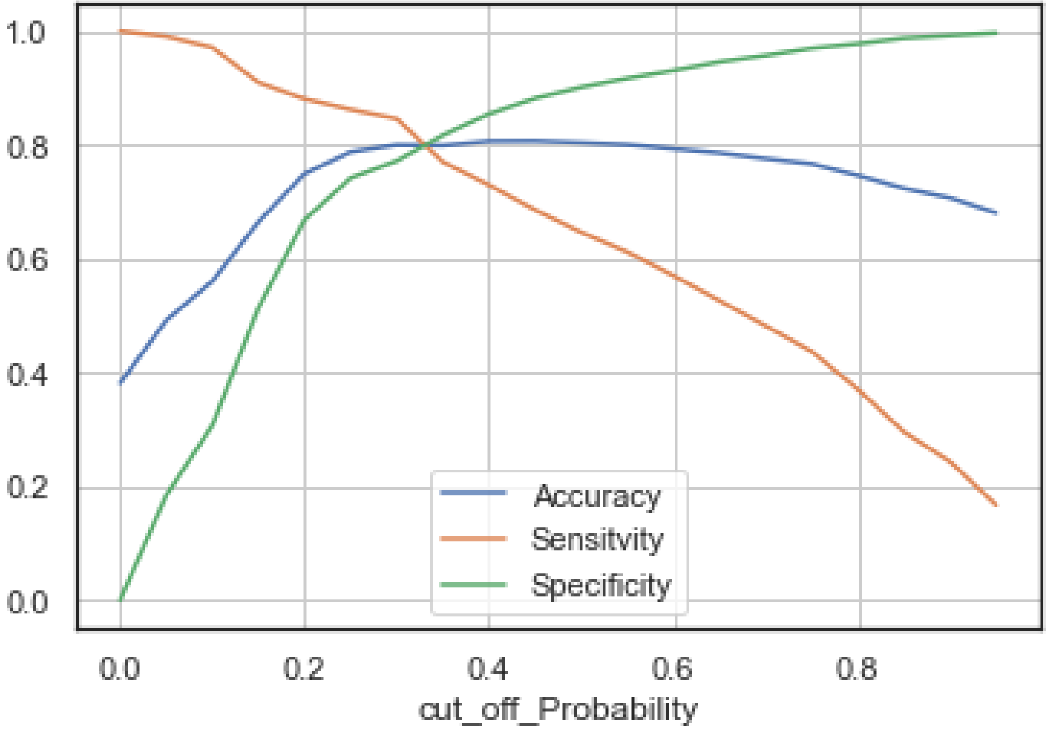
Model Evaluation On Train Data

- Our final model has AUCROC of 87%



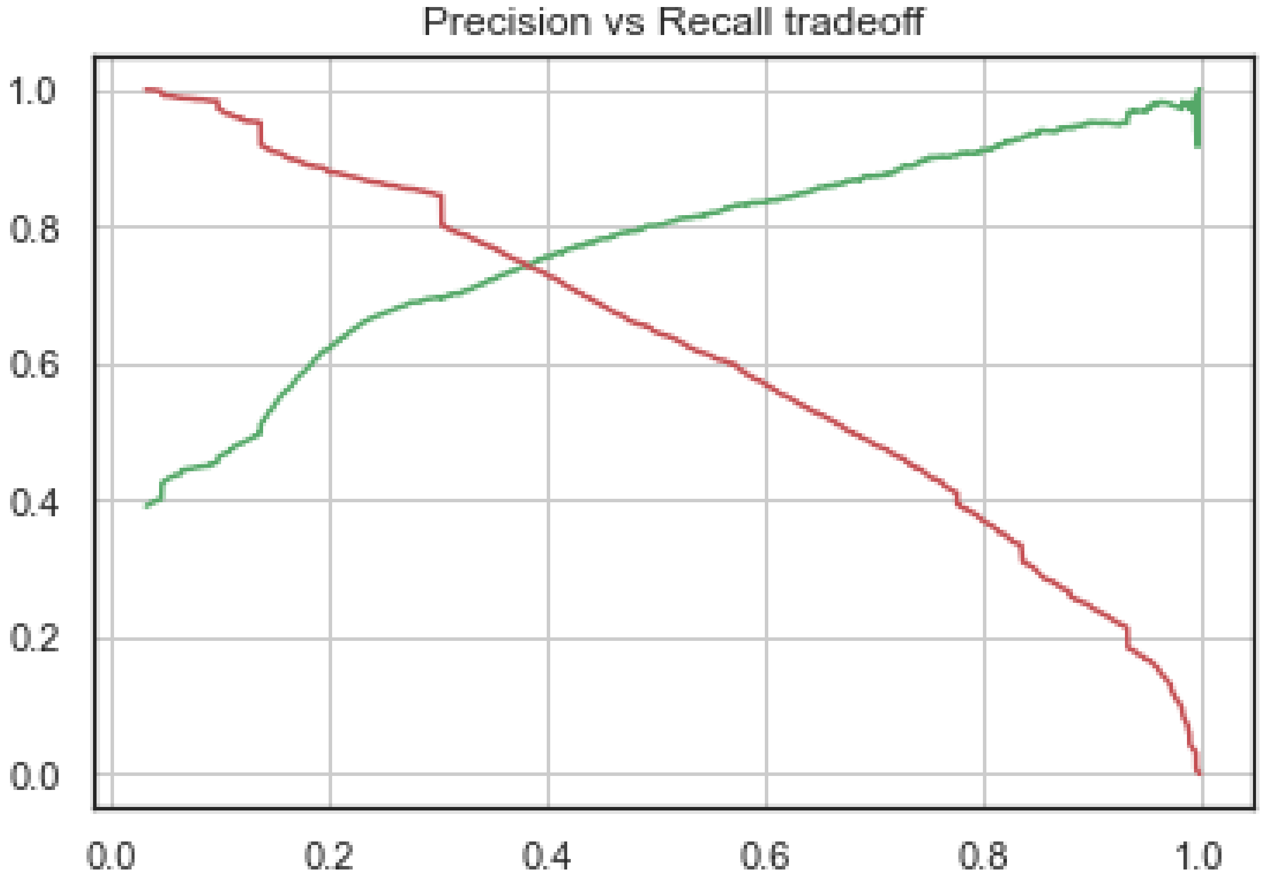
Model Evaluation On Train Data

- To get the final predictions we plotted sensitivity, specificity, and accuracy of the model with thresholds starting from 0 to 0.95 with a step size of 0.05. From the graph we found that 0.35 as the optimal threshold for our model.
- For this optimal threshold of 0.35 the model showed a precision of 0.72 and recall of 0.77. Overall accuracy for the train dataset stands at 80%



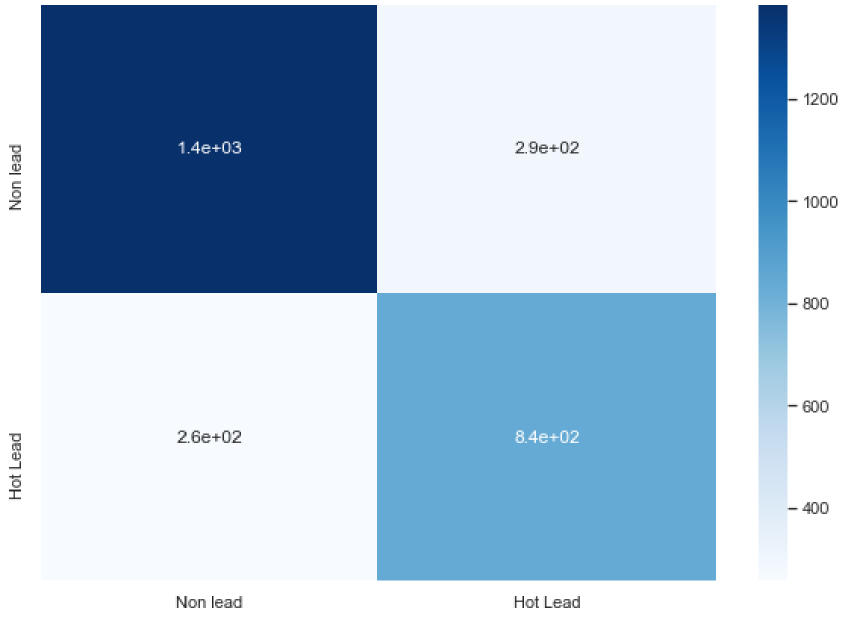
Model Evaluation On Train Data :

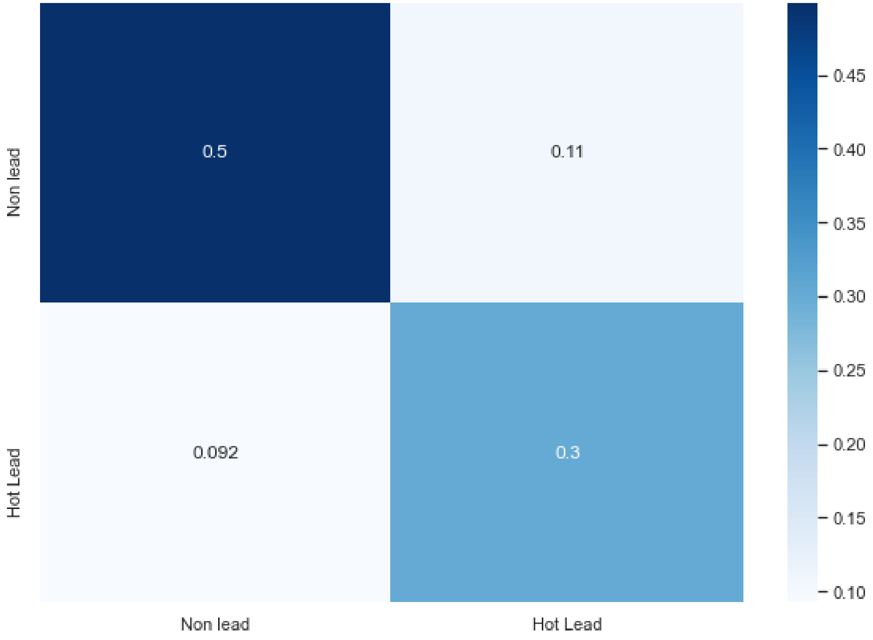
- From the precision recall analysis we did, we found that both these values coincide almost at a threshold of 0.4.



Model Evaluation On Test Data

- To evaluate the model on the test dataset, we used the fitted scalar from the train dataset to transform the numeric variable in the test dataset.
- On the test dataset also, the accuracy stands at 80% showing that the final model was neither under fit nor overfit.
- For the test dataset, we have sensitivity, specificity, precision and recall as 0.77, 0.82, 0.74 and 0.77 respectively.





Conclusion

- The Accuracy, Precision and Recall score we got from test set in acceptable range.
- We have high recall score than precision score which we were exactly looking for.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.
- The top three variables in our model which contribute most towards the probability of a lead getting converted are Total Time Spent on Website, Lead Origin_Lead Add Form, Last Notable Activity_Modified with absolute z values of 27.296, 18.782 and 18.008, respectively. Their coefficients are 0.9343, 3.4568 and -1.3824.