Lead Score Case Study

Group Members
1. Anurag Ghosh
2. Sweta Seal

Problem Statement

- X Education sells online courses to industry professionals.
- 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.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, 'Get updates on DM Content’ have been dropped.
- Removing the “Prospect ID” , “Lead Number”, 'Country','I agree to pay the amount through cheque','A free copy of Mastering The Interview','City' which are not necessary for the analysis.

  Imputed the missing the values with MODE
- Dropping the columns having more than 30% as missing value such as ‘Specialization’, How did you hear about X Education, Tags, ‘Lead Quality’, ’Lead Profile’, ‘Asymmetrique Activity Index’, ‘Asymmetrique Profile Index’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Profile Score’.

Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 73

Model Building

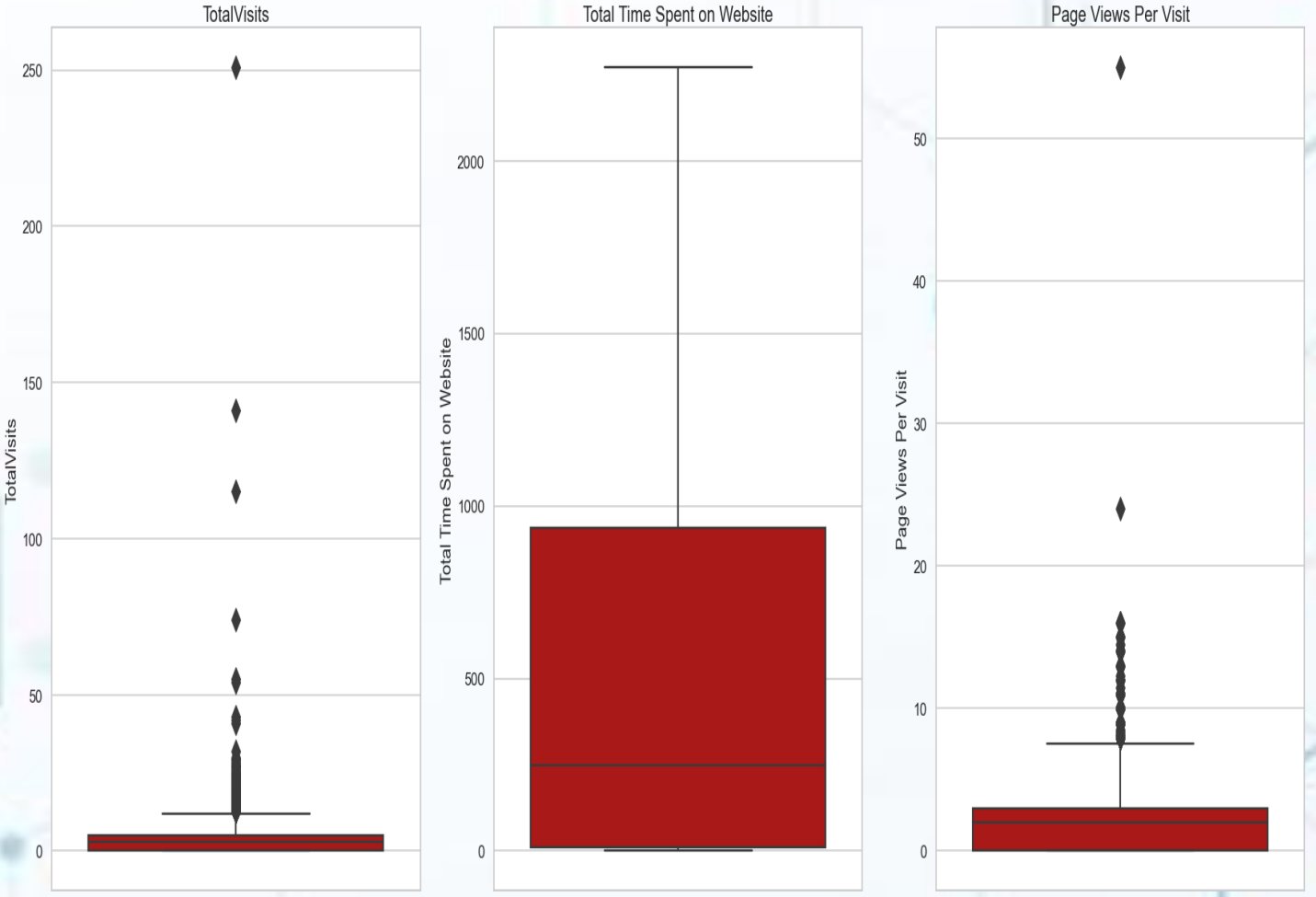
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 80.51%

Approach of the analysis

I. We started our analysis with our cleaned dataset by converting all the binary variables to ‘0’ and ‘1’ and multiple categories into dummy variables.

II. Next, we checked the outliers of the dataset by boxplot.

III. Outliers in logistic model is very sensitive hence we need to deal with care so that we donot it without lose any valuable information., which was achieved by creating bin



Correlation

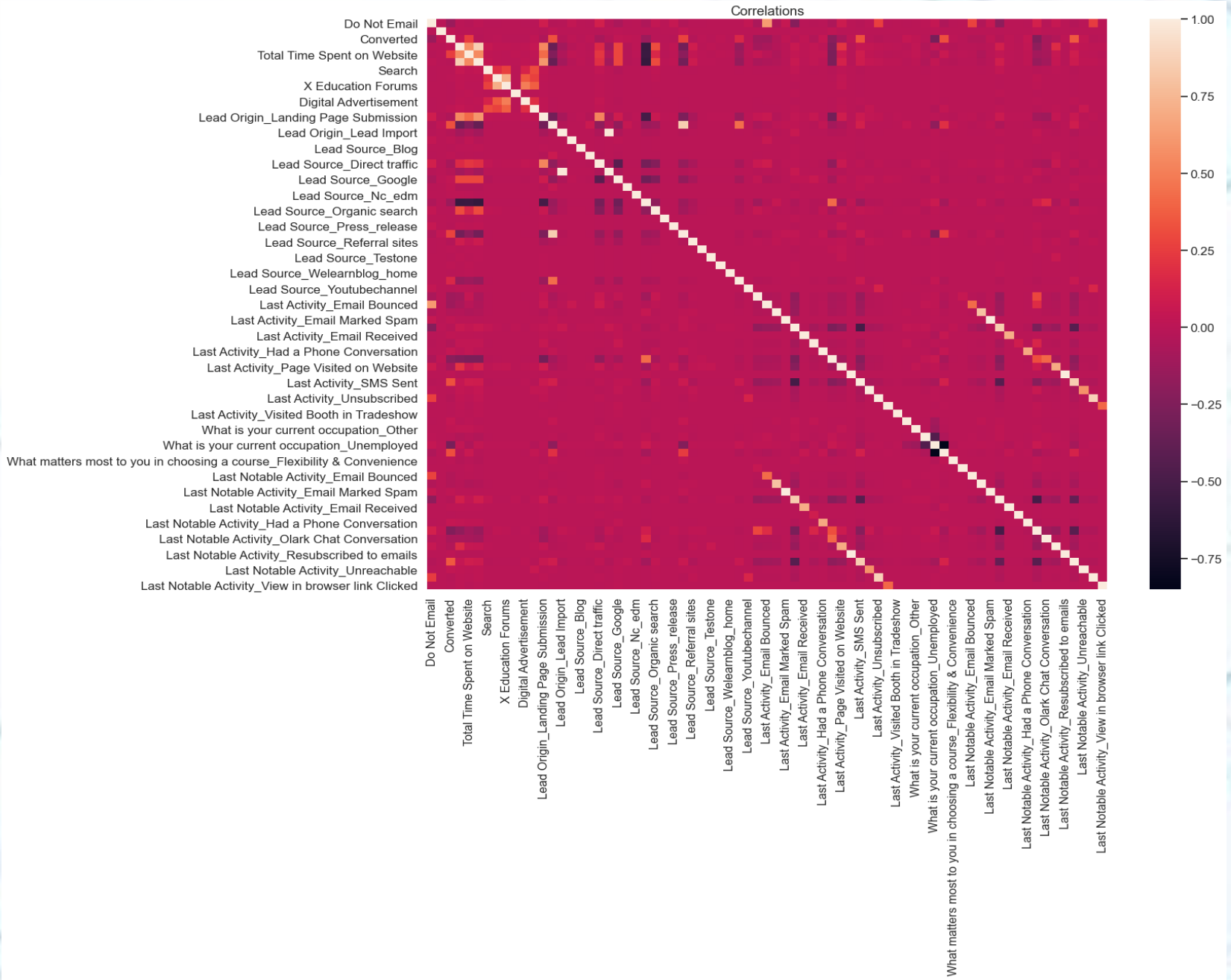
After fixing the outliers and dummy creation we proceed with our next step of analysis which is data preparation.

a) We split the dataset into train and test set and applied standardization on the features.

b) Standardization is required in order to keep all the variables in same scale which will help us in computation in more efficient way.

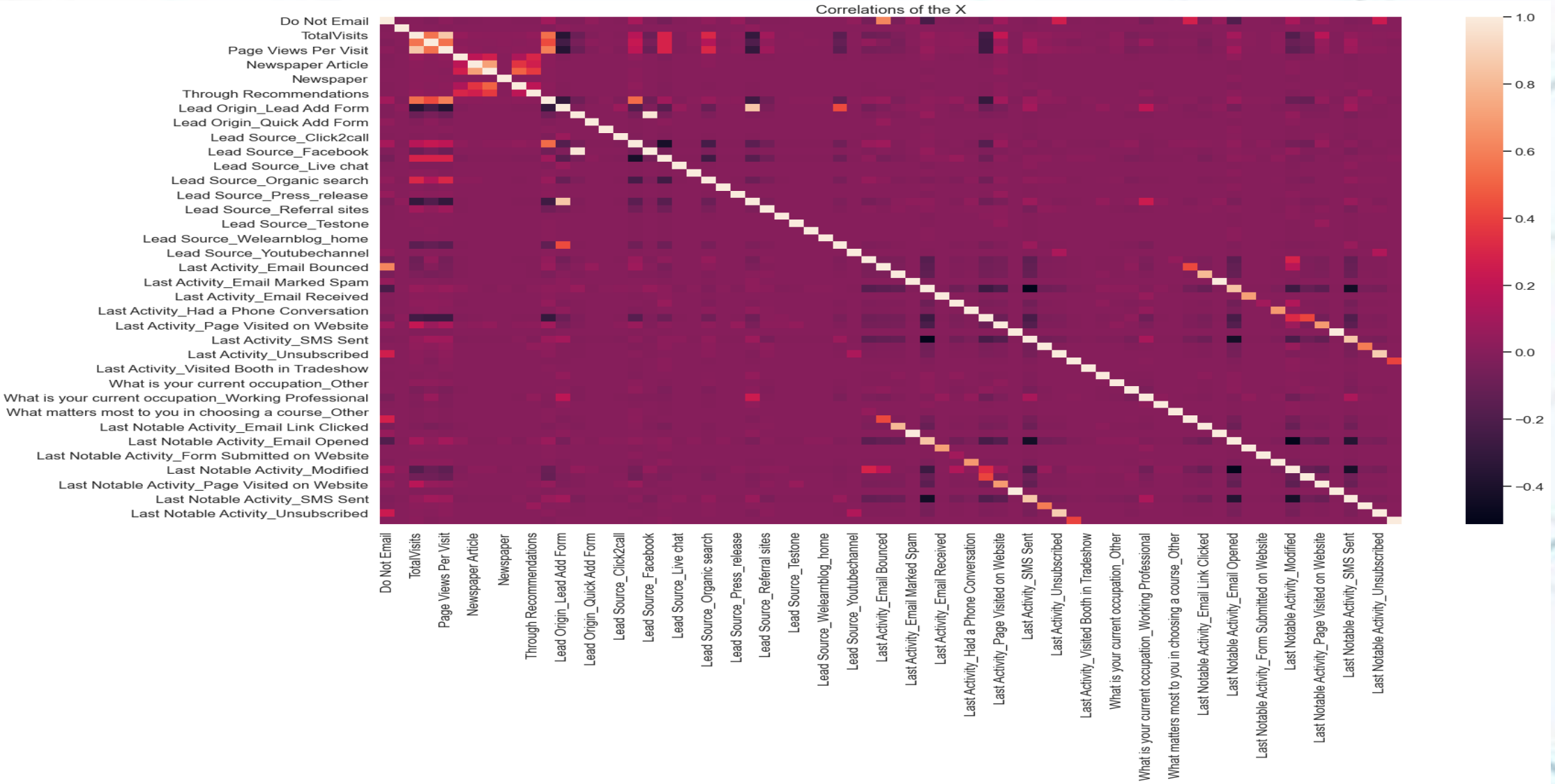
c) Checked the correlation of the dataset by heatmap. Attached heatmap is showing the correlation of all features present in the dataset.

d) There are some high correlations in the heatmap which we dropped.



Correlation

- After dropping those high correlations features, we plotted heatmap again.



Final model visualization with VIF

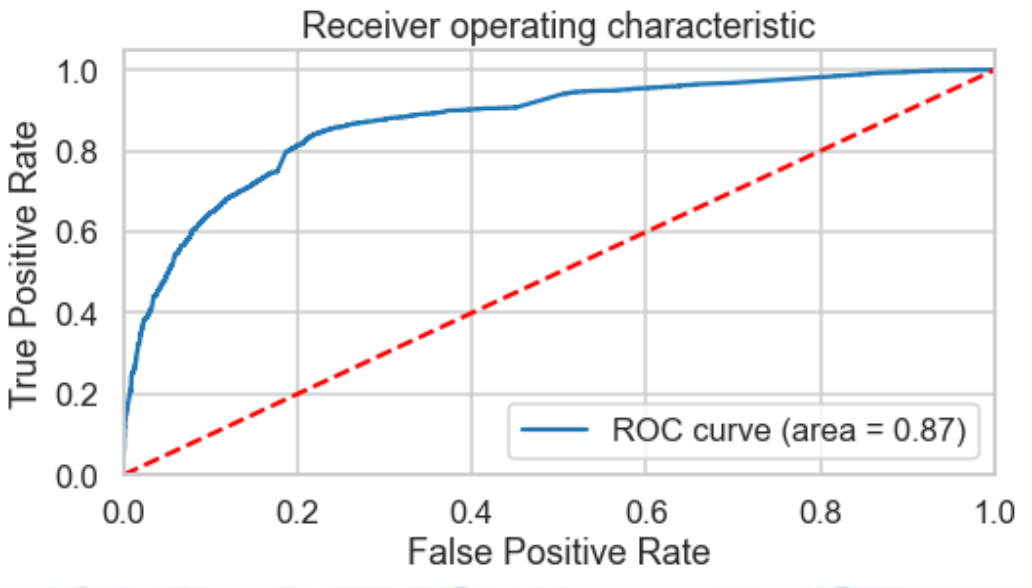
	Features	VIF
3	Page Views Per Visit	2.76
1	TotalVisits	1.98
2	Total Time Spent on Website	1.82
11	Last Notable Activity_Modified	1.56
9	Last Notable Activity_Email Opened	1.41
4	Lead Origin_Lead Add Form	1.40
5	Lead Source_Welingak website	1.24
6	Last Activity_Converted to Lead	1.16
7	What is your current occupation_Working Profes...	1.16
13	Last Notable Activity_Page Visited on Website	1.14
0	Do Not Email	1.13
8	Last Notable Activity_Email Link Clicked	1.02
12	Last Notable Activity_Olark Chat Conversation	1.01
10	Last Notable Activity_Had a Phone Conversation	1.00

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6453
Model Family:	Gaussian	**Df Model:**	14
Link Function:	identity	**Scale:**	0.14069
Method:	IRLS	**Log-Likelihood:**	-2827.8
Date:	Thu, 05 Aug 2021	**Deviance:**	907.90
Time:	21:31:36	**Pearson chi2:**	908.
No. Iterations:	3		
Covariance Type:	nonrobust		

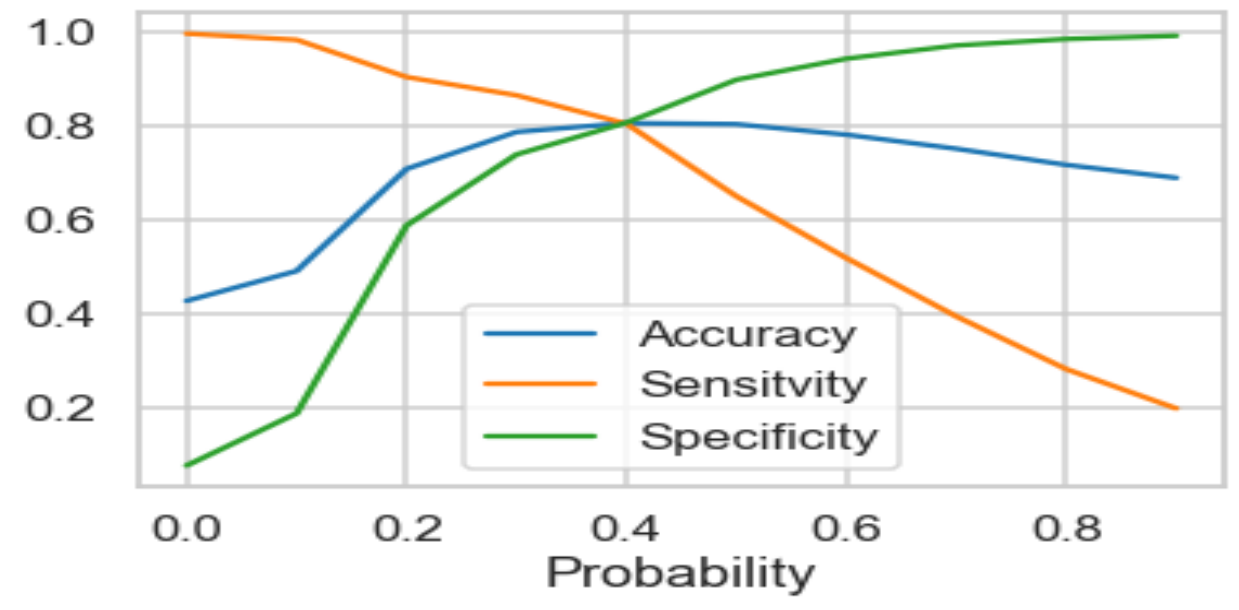
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4083	0.012	32.752	0.000	0.384	0.433
Do Not Email	-0.1877	0.018	-10.491	0.000	-0.223	-0.153
TotalVisits	1.0230	0.261	3.915	0.000	0.511	1.535
Total Time Spent on Website	0.7259	0.021	35.097	0.000	0.685	0.766
Page Views Per Visit	-0.9736	0.143	-6.790	0.000	-1.255	-0.693
Lead Origin_Lead Add Form	0.4950	0.020	24.308	0.000	0.455	0.535
Lead Source_Welingak website	0.1923	0.044	4.397	0.000	0.107	0.278
Last Activity_Converted to Lead	-0.1247	0.023	-5.317	0.000	-0.171	-0.079
What is your current occupation_Working Professional	0.3459	0.018	18.997	0.000	0.310	0.382
Last Notable Activity_Email Link Clicked	-0.2916	0.036	-8.145	0.000	-0.362	-0.221
Last Notable Activity_Email Opened	-0.2237	0.013	-17.428	0.000	-0.249	-0.199
Last Notable Activity_Had a Phone Conversation	0.2268	0.114	1.997	0.046	0.004	0.449
Last Notable Activity_Modified	-0.3054	0.013	-24.110	0.000	-0.330	-0.281
Last Notable Activity_Olark Chat Conversation	-0.3494	0.036	-9.739	0.000	-0.420	-0.279
Last Notable Activity_Page Visited on Website	-0.3034	0.027	-11.122	0.000	-0.357	-0.250

ROC Curve

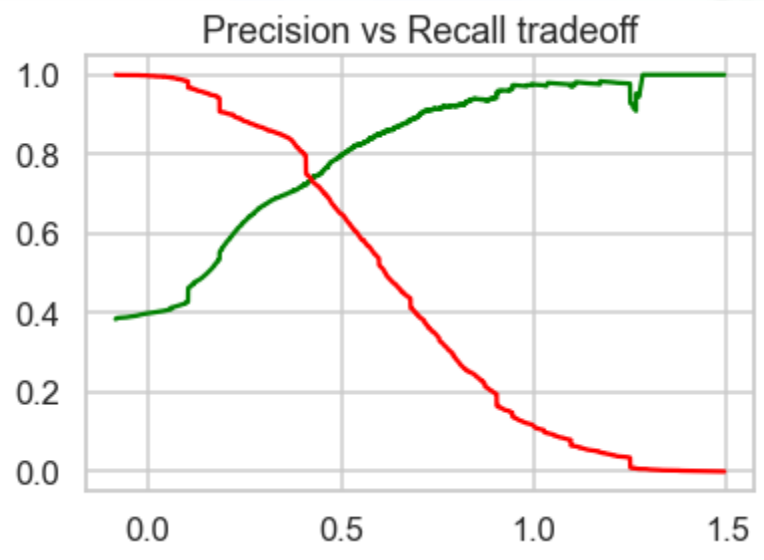


- After building the final model making prediction on it(on train set), we created ROC curve to find the model stability with auc score(area under the curve) As we can see from the graph plotted on the right side, the area score is 0.88 which is a great score.
- And our graph is leaned towards the left side of the border which means we have good accuracy

Finding the optimal cutoff point



- Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff.
- We found that on 0.4 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected.
- To verify our answer we plotted this in a graph – line plot which is on the right side and we stand corrected that the meeting point is close to 0.4 and hence we choose 0.4 as our optimal probability cutoff



- We used this cutoff point to create a new column in our final dataset for
- predicting the outcomes.
- 1. After this we did another type of evaluation which is by checking Precision and
- Recall
- 2.As we all know, Precision and Recall plays very important role in build our model
- more business oriented and it also tells how our model behaves.
- 3. Hence, we evaluated the precision and recall for this model and found the score
- as 0.73 for precision and 0.79 for recall.
- 4. Now, recall our business objective - the recall percentage I will consider more
- valuable because it is okay if our precision is little low which means less hot lead
- customers but we don't want to left out any hot leads which are willing to get
- converted hence our focus on this will be more on Recall than Precision.
- 5. i.e We get more relevant results - as many as hot lead customers from our model

Prediction on test set

- Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set and the new predictions values were saved in new dataframe.
- After this we did model evaluation i.e. finding the accuracy, precision and recall.
- The accuracy score we found was 0.80, precision 0.77 and recall 0.71 approximately.
- This shows that our test prediction is having accuracy , precision and recall score in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

Conclusion

Valuable Insights -

- The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.
- 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.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

a) **Total Visits**
b) **Page Views Per Visit**
c) **Total Time Spent on Website**
d) **Lead Origin_Lead Add Form**