Lead Scoring Case Study

Analyzed by:

Swathi Somayaji

Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score 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.

Goals:

1. Create a model in such a way that the customers with high lead score have higher conversion chance and low lead score have lower conversion chance. The ballpark of the target lead conversion rate is around 80%.

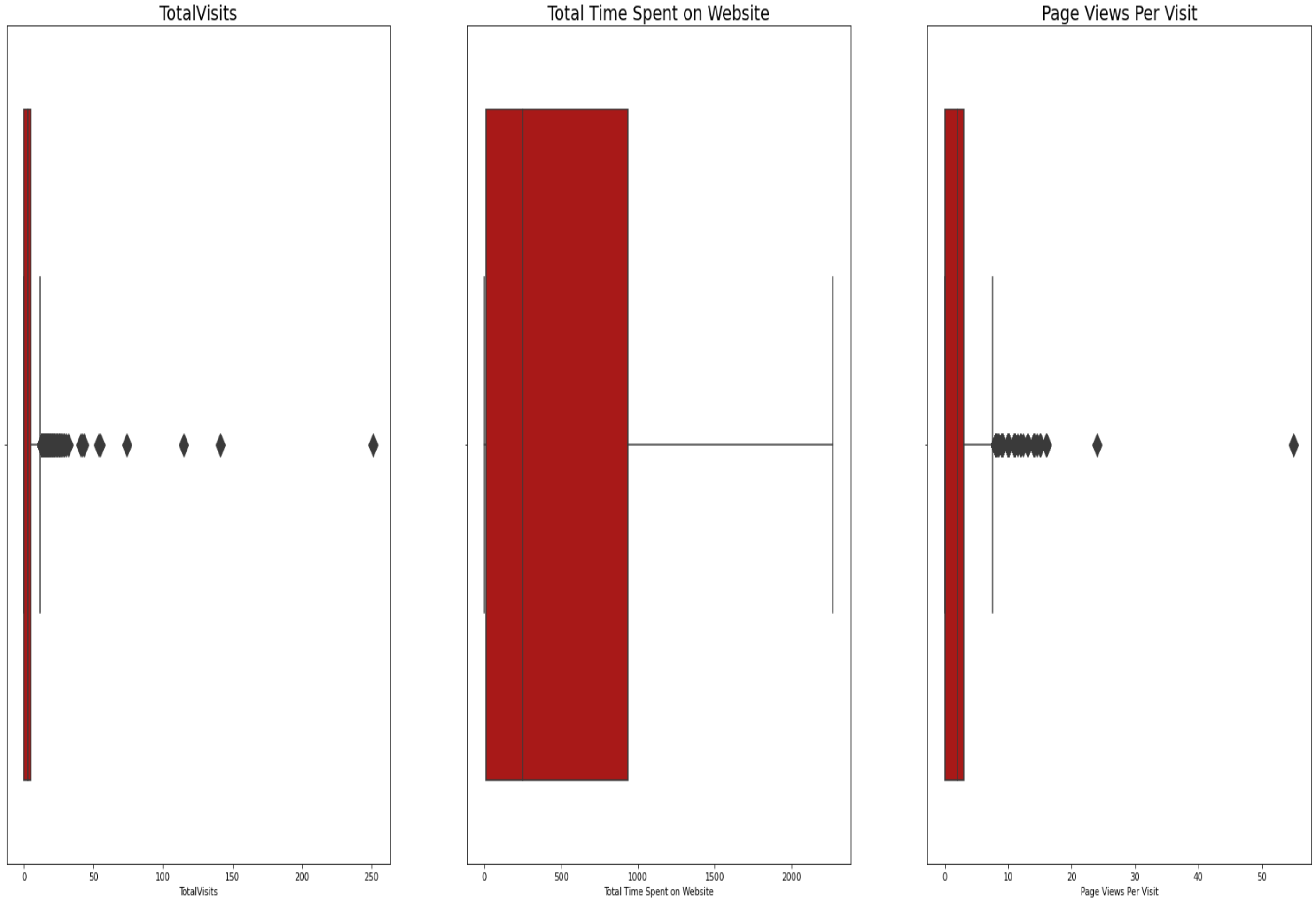
2. The model should be able to adjust if the company’s requirement changes in near future.

Approach of the analysis

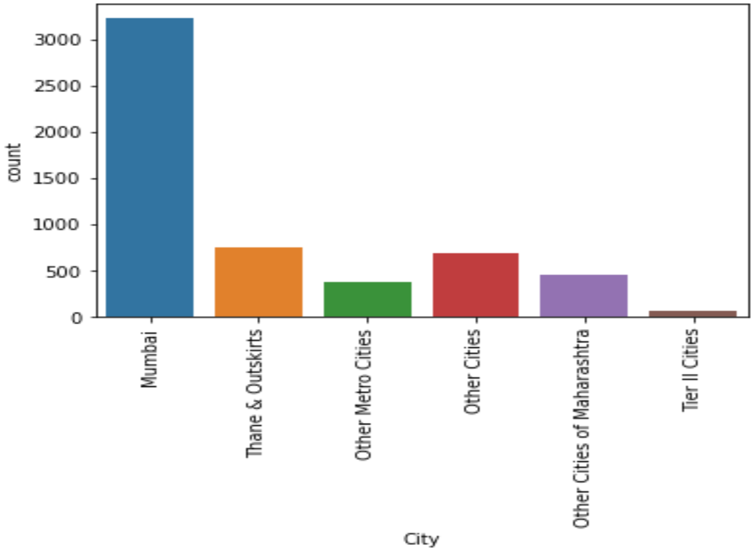
The dataset had: 37- Columns

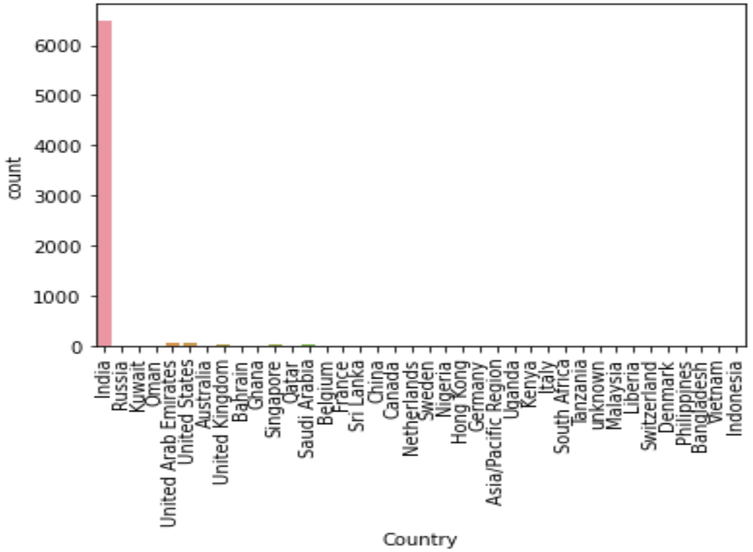
9240- Entries

- I started this analysis with taking care of missing values and if the missing value percentage was above 45% , it was not considered for further analysis
- Mistaken inputs were corrected(Select and google variables)
- outliers of the dataset: The visualization of those outliers we can see on the graph attached on the right side.
- Outliers in logistic model is very sensitive hence we need to deal with it without losing our valuable information. This can be achieved by creating bins. Hence, we did it.
- cleaned dataset by converting all the binary variables to '0' and '1' and multiple categories into dummy variables.

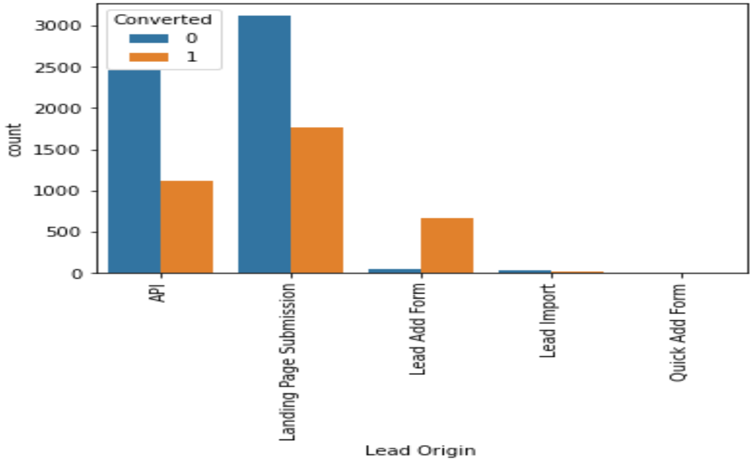


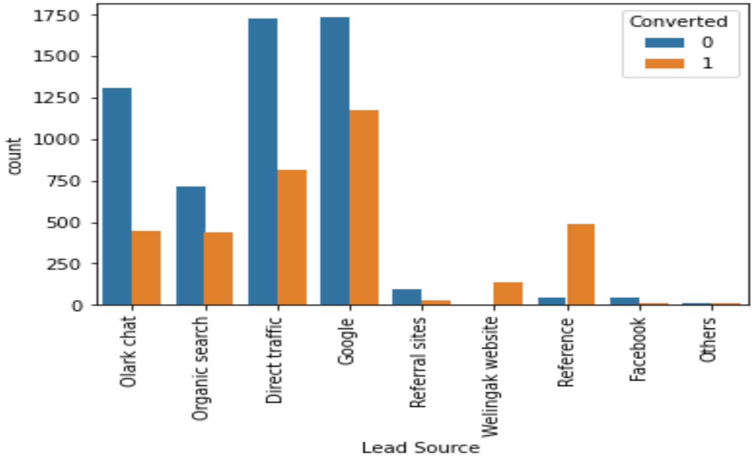
- Most of the data were from India.
- Mumbai was the major city.



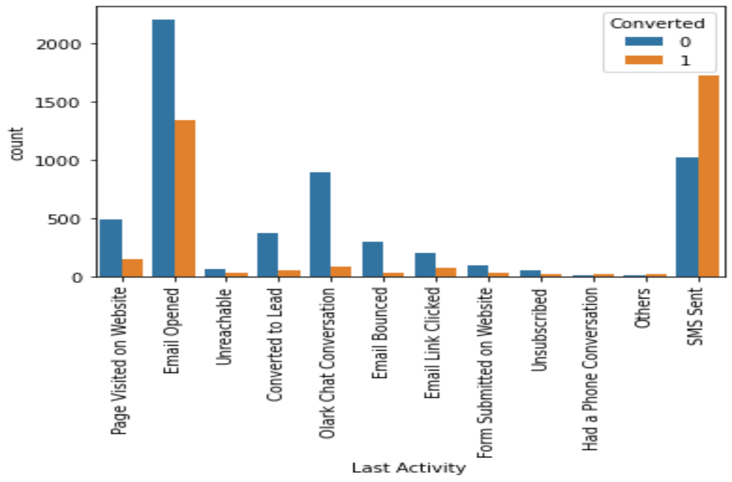


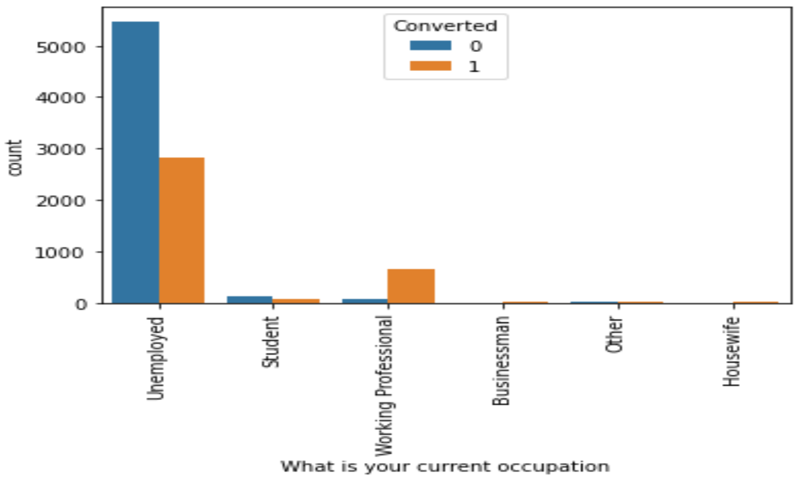
- lead conversion was High in API and Landing Page Submission origin in Lead Origin.
- Google and Olark Chat had high Conversion rate in Lead Source.

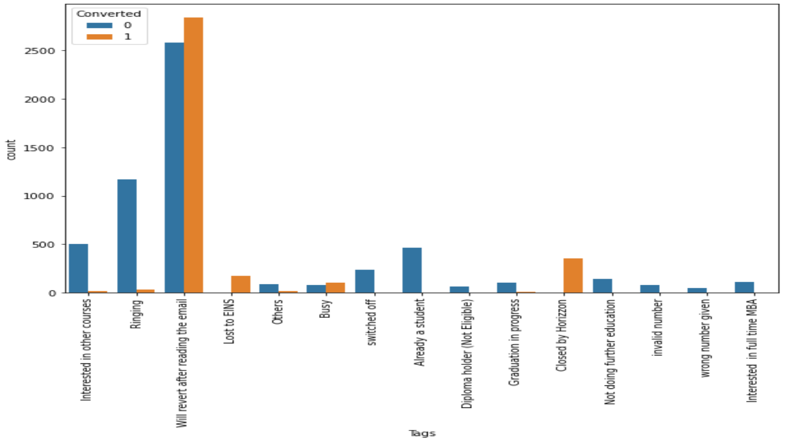


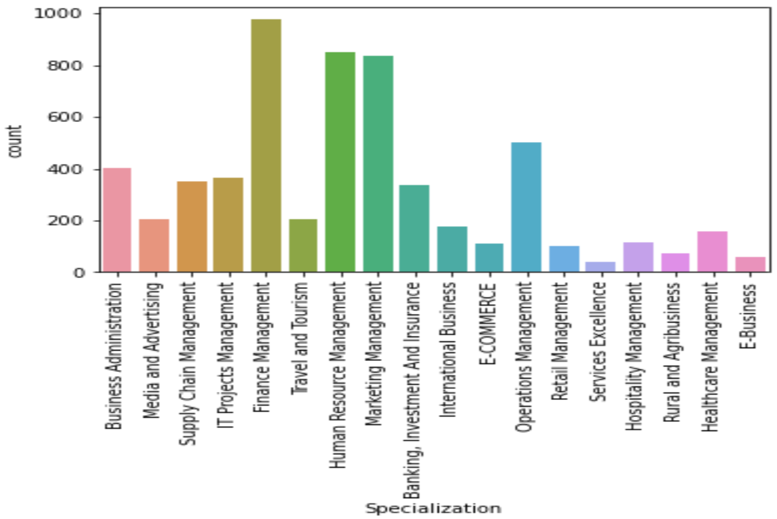


- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is High.
- Majority of the people were unemployed, and they have high Conversion rate.
- 'Will revert after reading the mail' Tag has high conversion rate
- Specialization had high frequency in management.
- Conversion Ratio:38.53



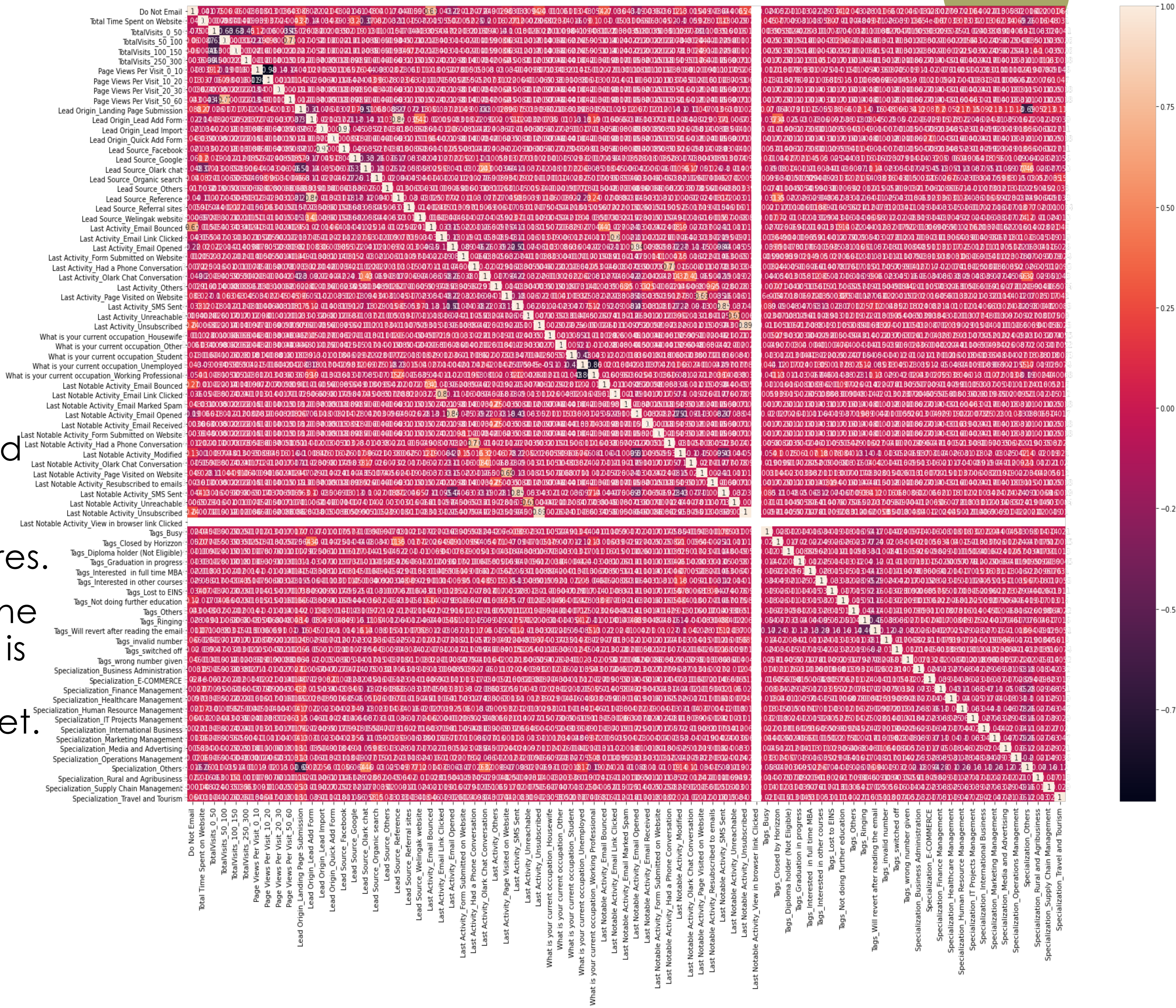






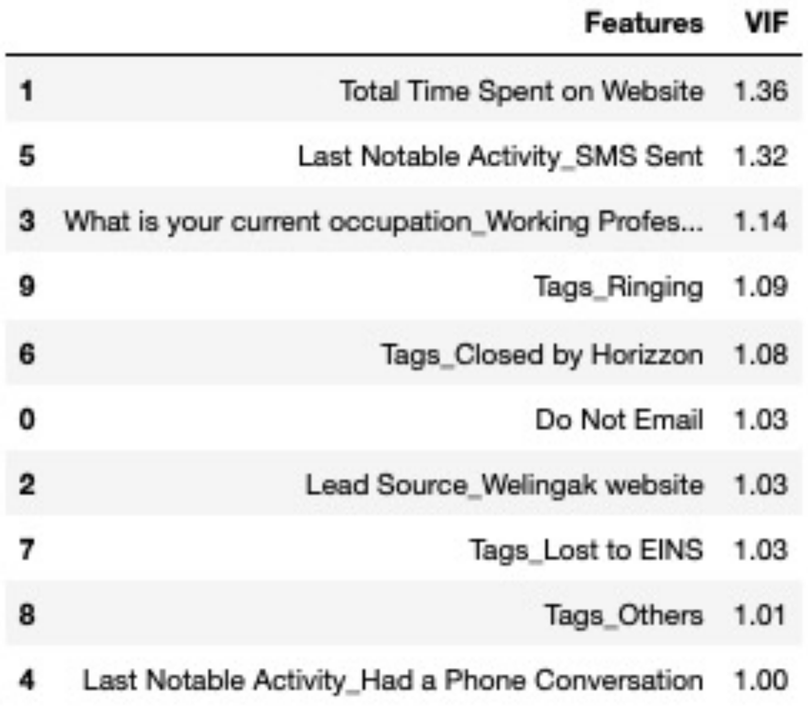
Data preparation:

- After fixing the outliers and dummy creation we proceed with next step of analysis which is data preparation.
- Split the dataset into train and test set to 70-30 and do MinMaxScalar() standardization on the features.
- Checked the correlation of the dataset. Attached heatmap is showing the correlation of all features present in the dataset.
- There are some high correlations in the heatmap which we dropped.

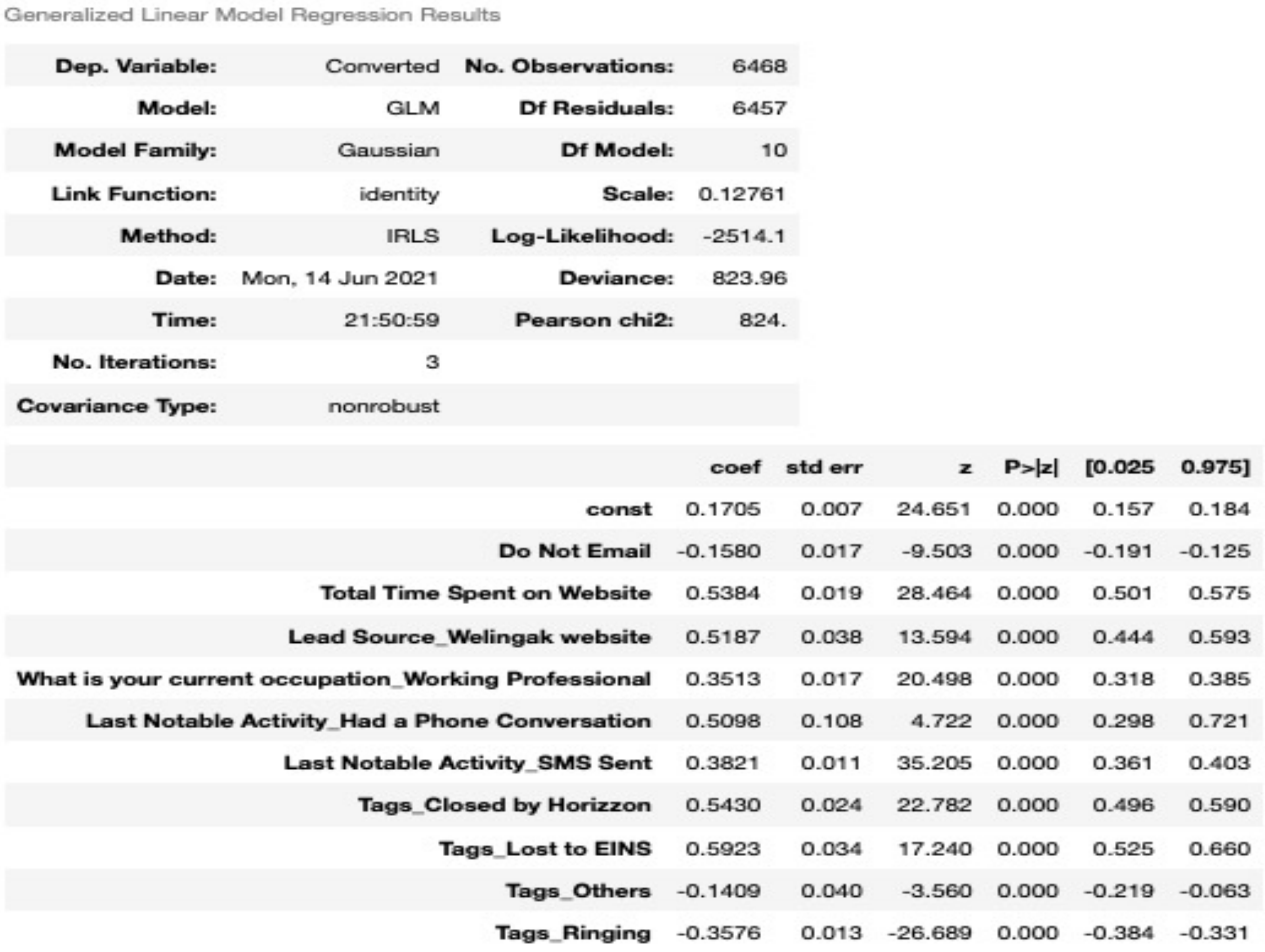


Model Building and Testing

- Created the model with rfe count 15 and manually compared the model evaluation
- with respect to P values and VIF values and choose the final model with 10 variables as has more stability and accuracy than the other.
- For the final model I checked the optimal probability cut-offs by finding points
- conversion rate: 38.54%
- Accuracy:0.83
- Sensitivity:0.70
- Specificity:0.91



	Features	VIF
1	Total Time Spent on Website	1.36
5	Last Notable Activity_SMS Sent	1.32
3	What is your current occupation_Working Profes...	1.14
9	Tags_Ringing	1.09
6	Tags_Closed by Horizzon	1.08
0	Do Not Email	1.03
2	Lead Source_Welingak website	1.03
7	Tags_Lost to EINS	1.03
8	Tags_Others	1.01
4	Last Notable Activity_Had a Phone Conversation	1.00

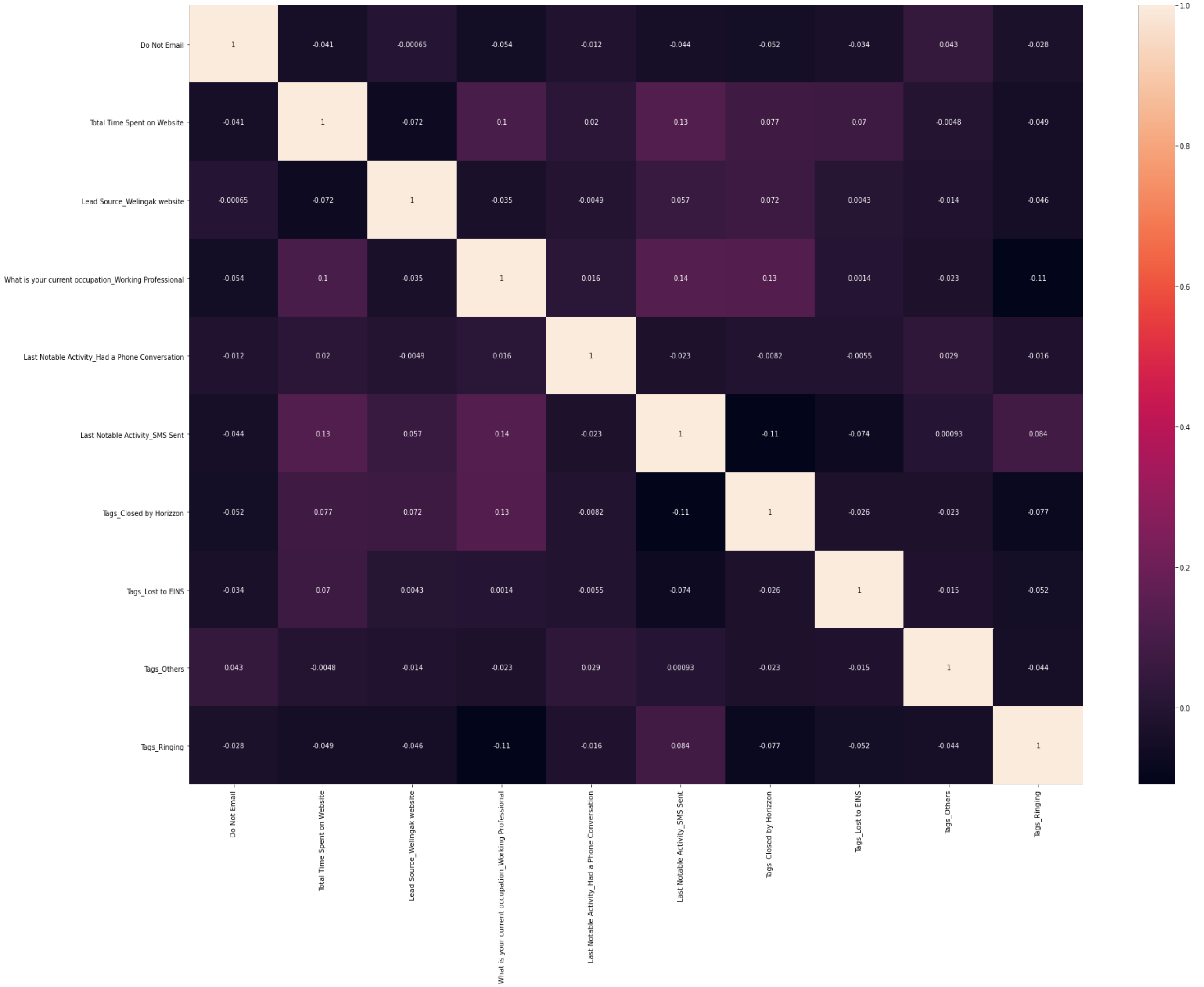


Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6457
Model Family:	Gaussian	**Df Model:**	10
Link Function:	identity	**Scale:**	0.12761
Method:	IRLS	**Log-Likelihood:**	-2514.1
Date:	Mon, 14 Jun 2021	**Deviance:**	823.96
Time:	21:50:59	**Pearson chi2:**	824.
No. Iterations:	3		
Covariance Type:	nonrobust		

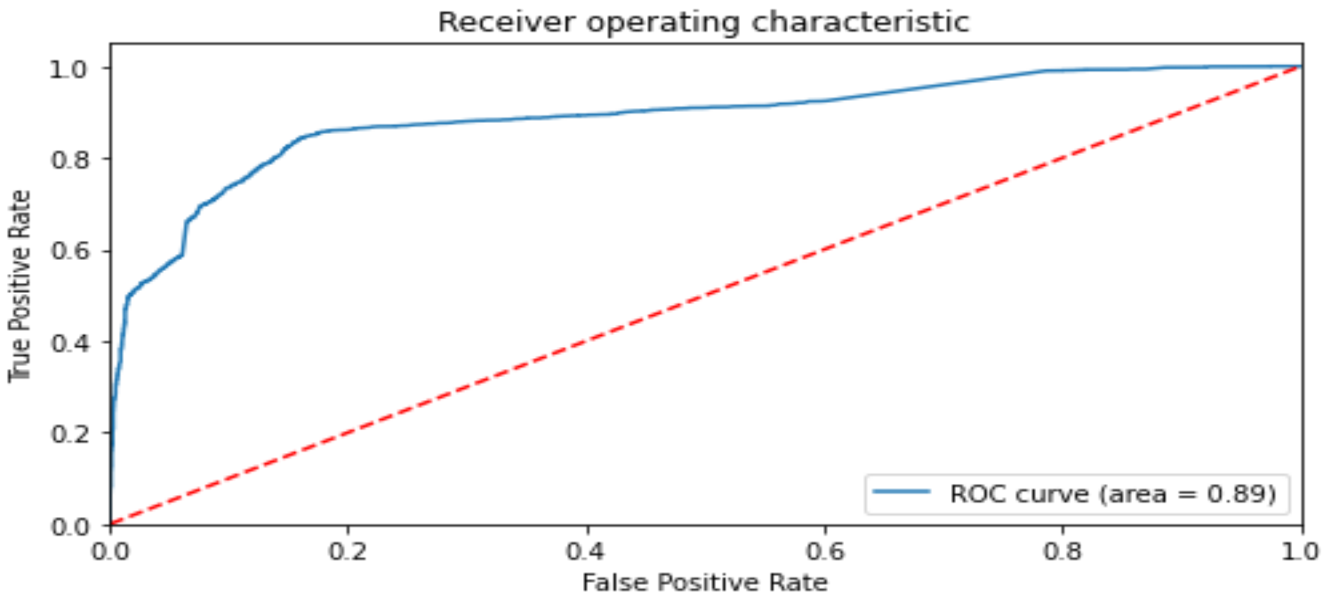
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.1705	0.007	24.651	0.000	0.157	0.184
Do Not Email	-0.1580	0.017	-9.503	0.000	-0.191	-0.125
Total Time Spent on Website	0.5384	0.019	28.464	0.000	0.501	0.575
Lead Source_Welingak website	0.5187	0.038	13.594	0.000	0.444	0.593
What is your current occupation_Working Professional	0.3513	0.017	20.498	0.000	0.318	0.385
Last Notable Activity_Had a Phone Conversation	0.5098	0.108	4.722	0.000	0.298	0.721
Last Notable Activity_SMS Sent	0.3821	0.011	35.205	0.000	0.361	0.403
Tags_Closed by Horizzon	0.5430	0.024	22.782	0.000	0.496	0.590
Tags_Lost to EINS	0.5923	0.034	17.240	0.000	0.525	0.660
Tags_Others	-0.1409	0.040	-3.560	0.000	-0.219	-0.063
Tags_Ringing	-0.3576	0.013	-26.689	0.000	-0.384	-0.331

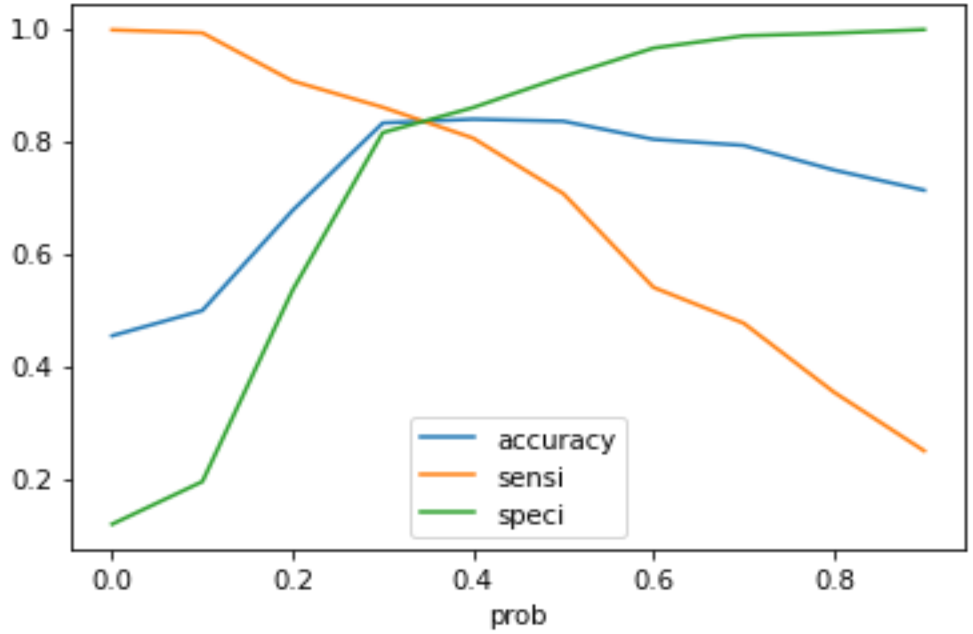
- No sign of multicollinearity shown from Heatmap in Final Model
- Accuracy:0.83
- Sensitivity: 0.70
- Specificity:0.91



Evaluating the Model

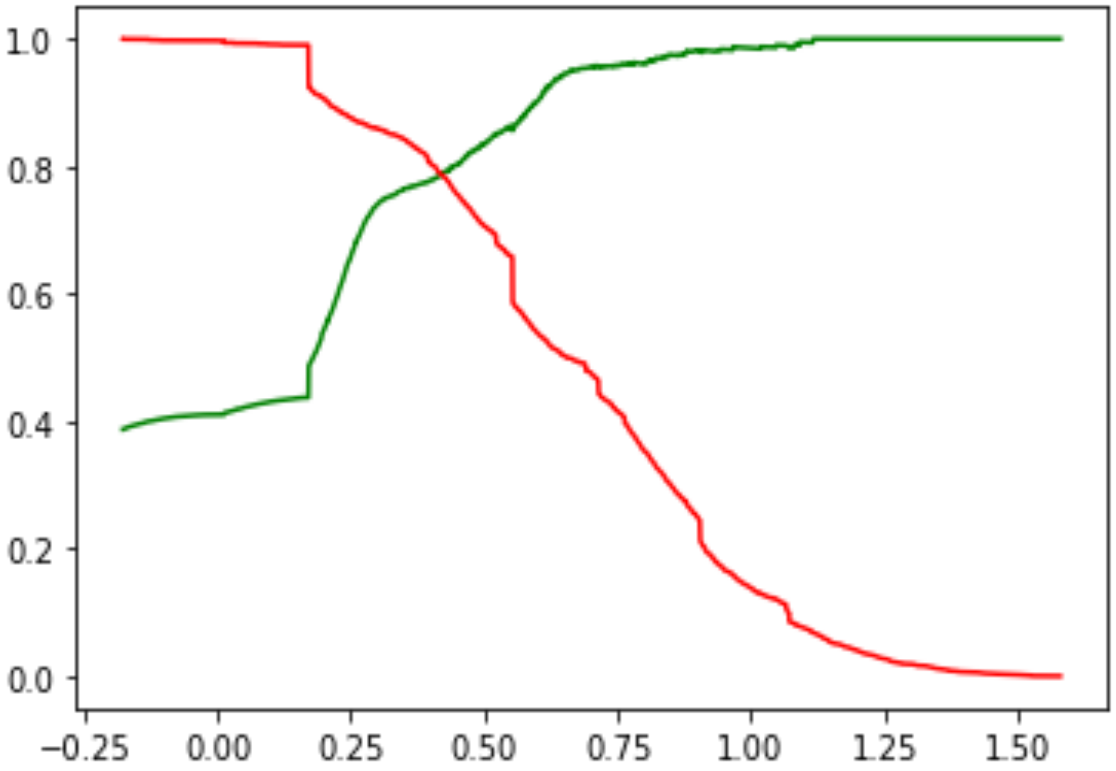
- The area under the curve of the ROC is 0.89 which is quite good. So it seems to have a good model.
- And our graph is leaned towards the left side of the border which means we have good accuracy.
- After looking at sensitivity and specificity tradeoff 0.35 was found out to be the the optimal cutoff point
- Sensitivity: 0.85
- Specificity:0.84





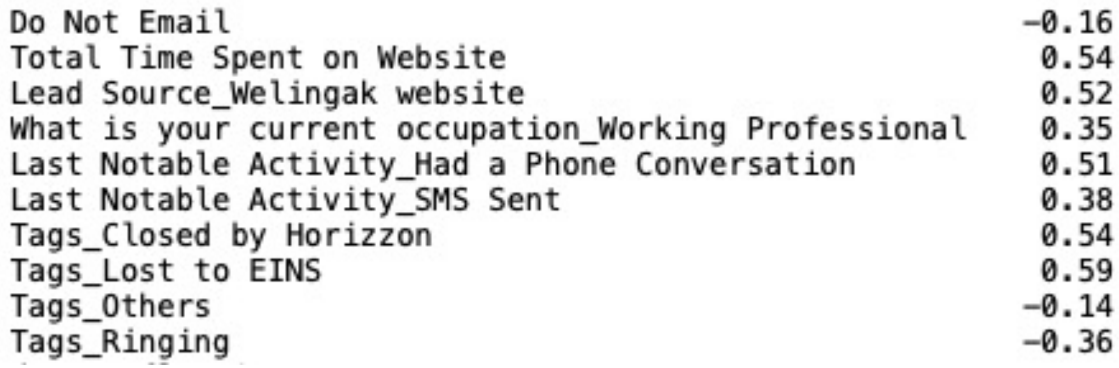
Precision and recall tradeoff

- Graph which will show us the tradeoff between Precision and recall.
- We found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.35.
- Accuracy:0.84
- Precision:0.76
- Recall:0.84

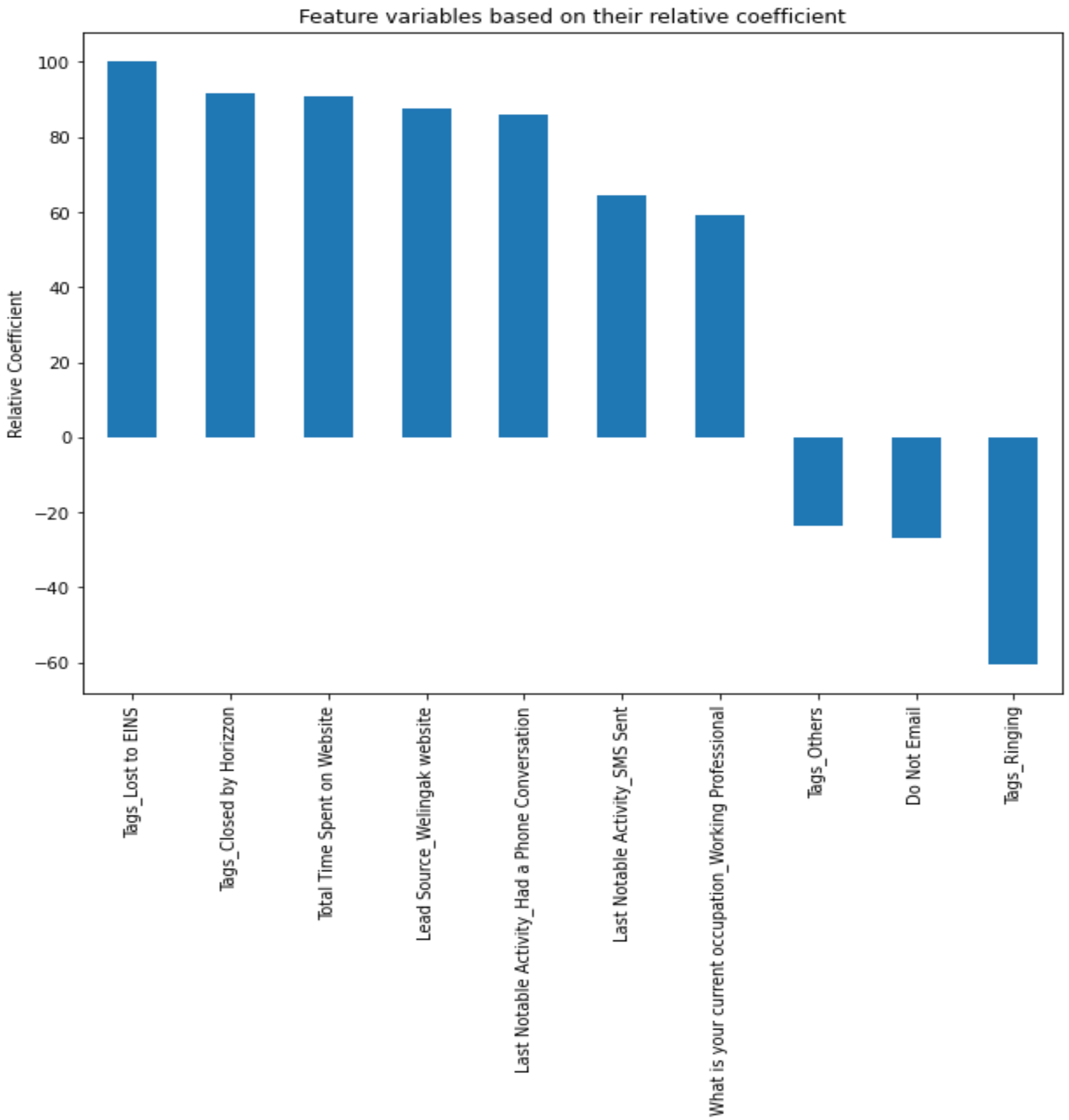


Predictions on the Test Set

- Accuracy:0.84
- Sensitivity:0.86
- Specificity:0.84
- The Accuracy, Precision and Recall score we got from test set in acceptable range
- Sensitivity of the prediction over test data set is 86%
- Conversion Ratio:43.5%



```
Do Not Email                                              -0.16
Total Time Spent on Website                                0.54
Lead Source_Welingak website                               0.52
What is your current occupation_Working Professional       0.35
Last Notable Activity_Had a Phone Conversation             0.51
Last Notable Activity_SMS Sent                             0.38
Tags_Closed by Horizzon                                    0.54
Tags_Lost to EINS                                          0.59
Tags_Others                                               -0.14
Tags_Ringing                                              -0.36
```



Conclusion:

- Test set is having accuracy, recall/sensitivity in an acceptable range.
- In business terms, the model is having stability an accuracy with adaptive environment skills. Means it will adjust with the company’s requirement changes made in coming future.
- Top features for good conversion rate:

  1.Tags_Lost to EINS

  2.Tags_Closed by Horizzon

  3.Total Time Spent on Website
- Top 3 variables that need improvement to convert a lead are:

  1.Lead Source_Welingak website

  2.Last Notable Activity_Had a Phone Conversation

  3. Tags_Ringing