Leads Scoring Case Study

Abhisek De
Vipin Panthri

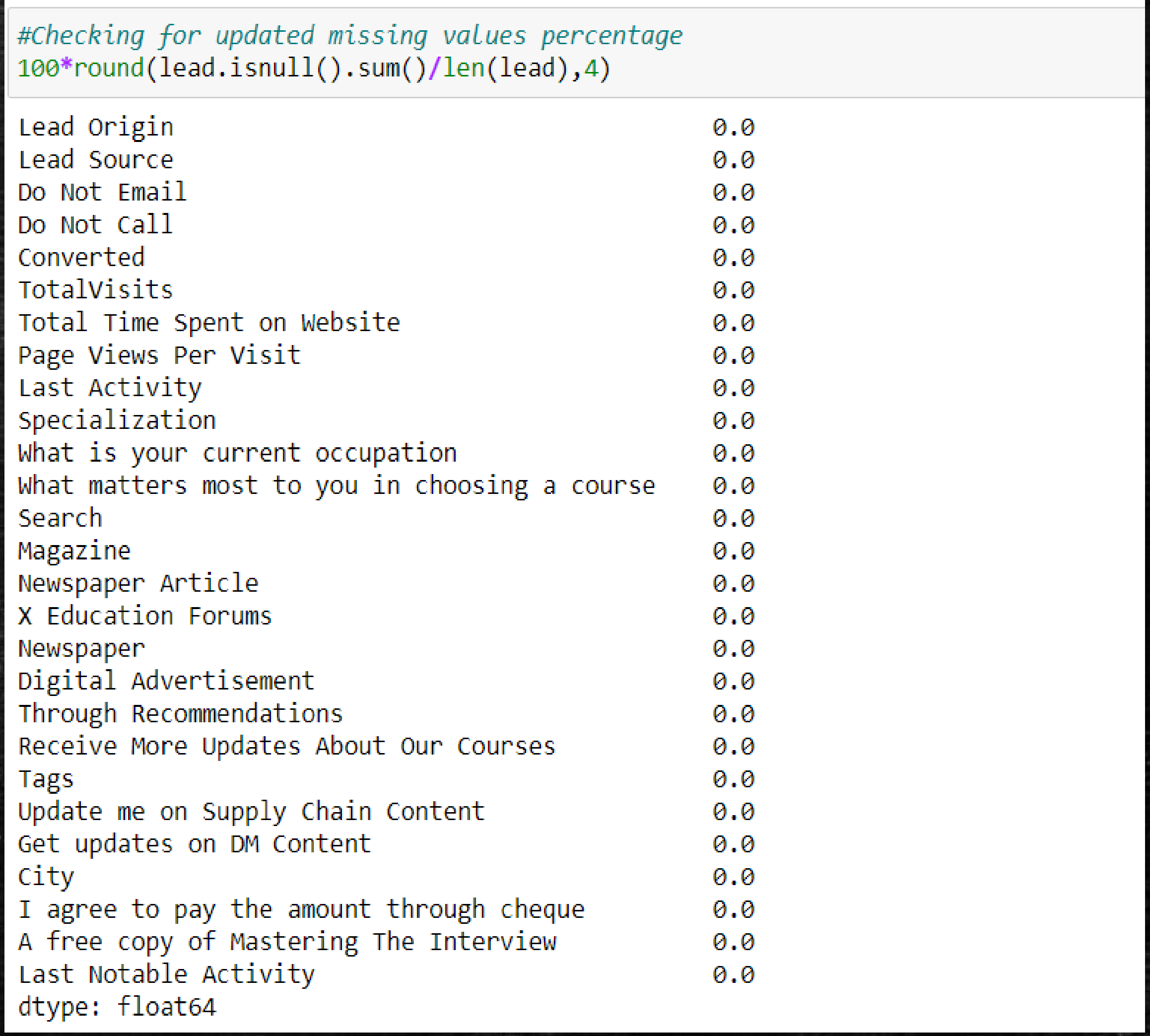
Problem statement

- We need to build a model wherein we 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.
- A ballpark of the target lead conversion rate to be around 80%.
- The model should adjust to company's requirement changes in the future.

Analysis of the Approach

Data Cleaning:

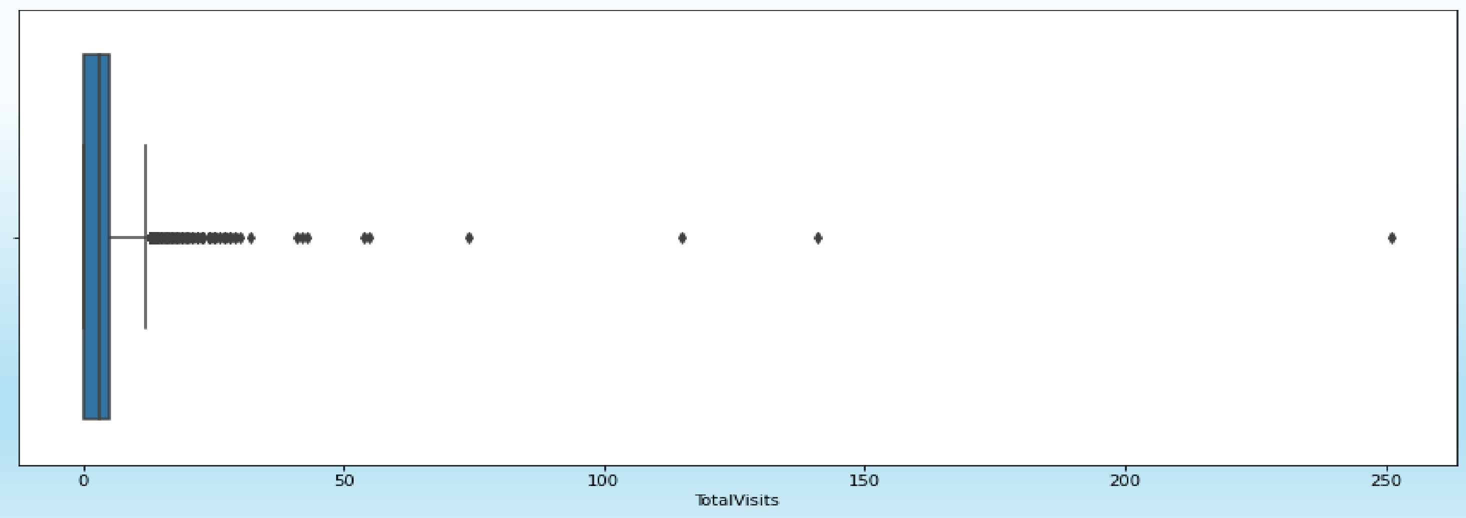
- We first cleaned the data and dealt with the missing values. So this is what the final dataframe looked like.
- To achieve this we dropped the columns having more than 40% missing values.
- Imputed the remaining missing values with the mode of that column.
- Dropped other redundant columns and columns having data imbalance.



```python
#Checking for updated missing values percentage
100*round(lead.isnull().sum()/len(lead),4)
```

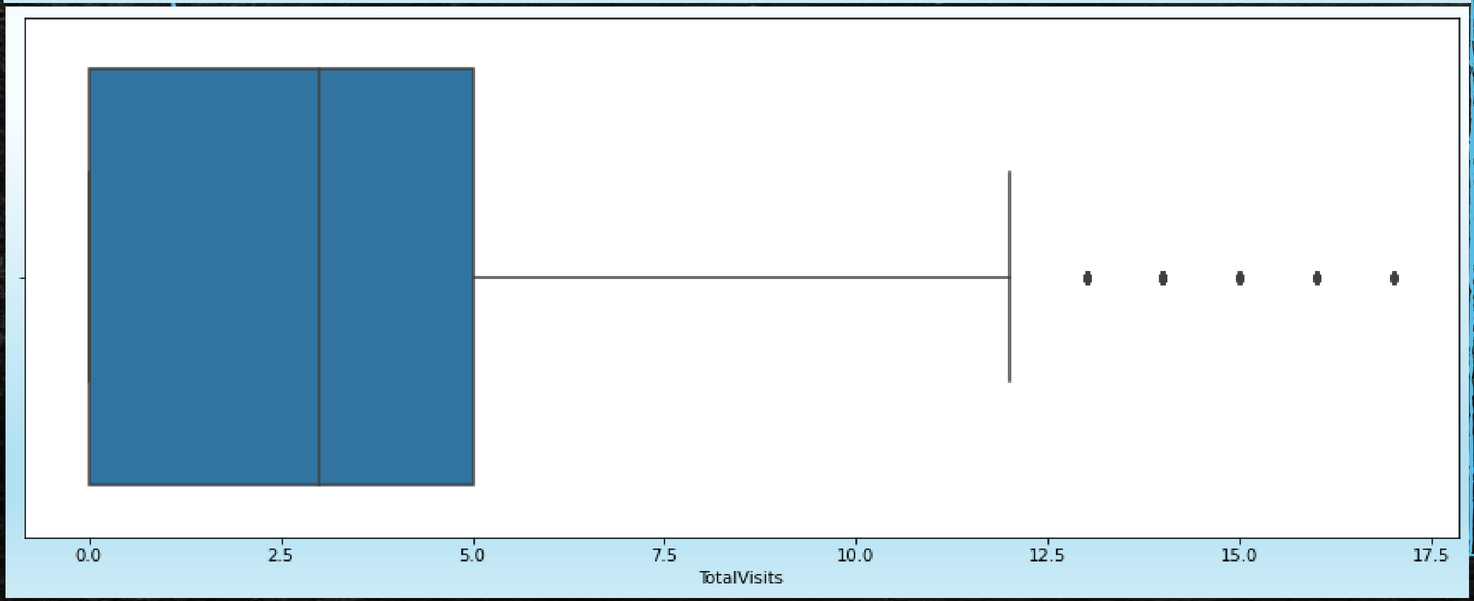
```
Lead Origin                                      0.0
Lead Source                                      0.0
Do Not Email                                     0.0
Do Not Call                                      0.0
Converted                                        0.0
TotalVisits                                      0.0
Total Time Spent on Website                      0.0
Page Views Per Visit                             0.0
Last Activity                                    0.0
Specialization                                   0.0
What is your current occupation                  0.0
What matters most to you in choosing a course    0.0
Search                                           0.0
Magazine                                         0.0
Newspaper Article                                0.0
X Education Forums                               0.0
Newspaper                                        0.0
Digital Advertisement                            0.0
Through Recommendations                          0.0
Receive More Updates About Our Courses           0.0
Tags                                             0.0
Update me on Supply Chain Content                0.0
Get updates on DM Content                        0.0
City                                             0.0
I agree to pay the amount through cheque         0.0
A free copy of Mastering The Interview           0.0
Last Notable Activity                            0.0
dtype: float64
```

Outlier Analysis



Before Outlier Treatment:

As we can see there are many points which are outliers.

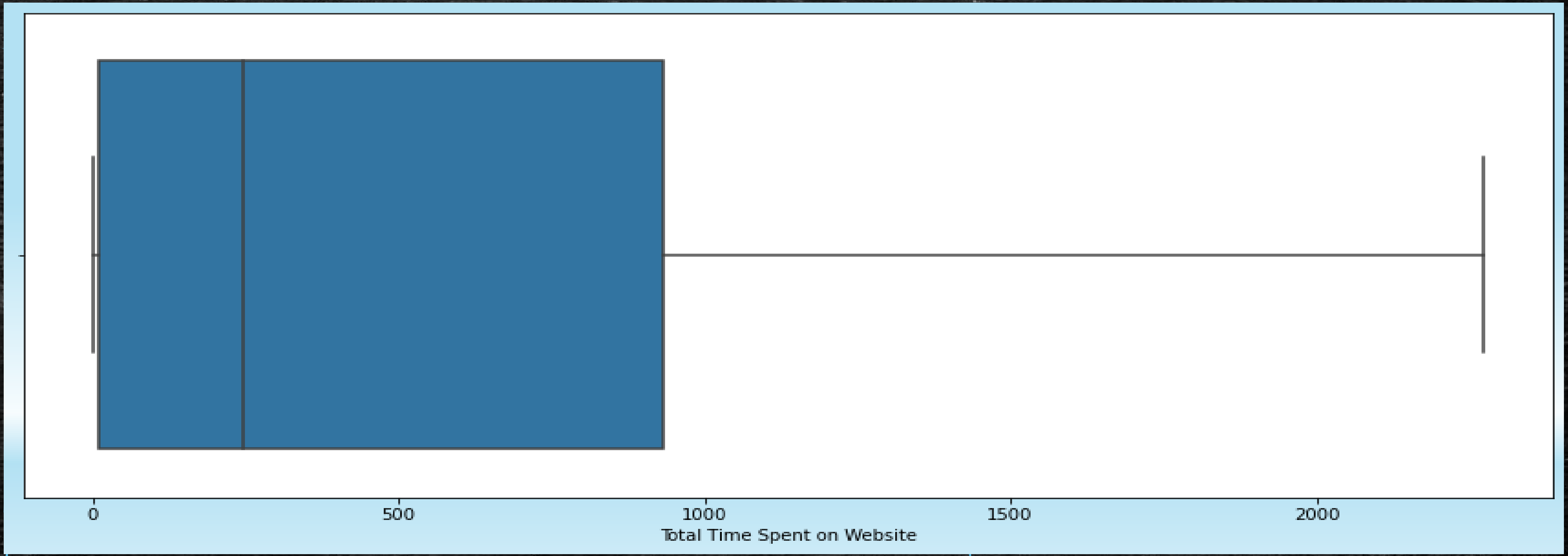


After Outlier Treatment:

As we can see here that the points which were too high in magnitude were removed.

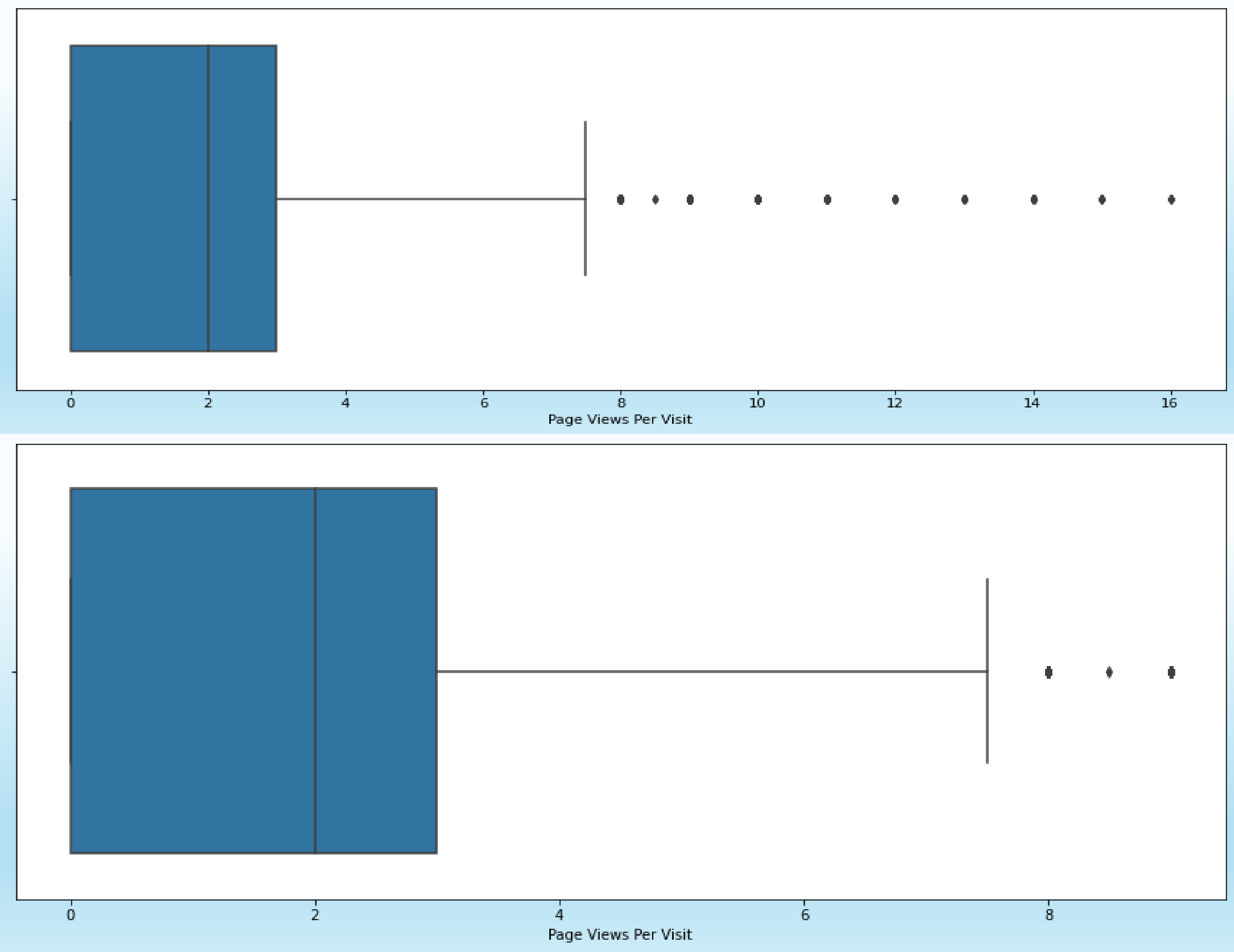
We did this with the help of IQR (Inter-Quartile Range Method where we chose to remove 5% of the lower range values and 10% of the upper range values.

Outlier Analysis



As we can see here that this variable doesn’t have outliers. Hence outlier treatment was not done here.

Outlier Analysis



Before Outlier Treatment:

As we can see there are many points which are outliers

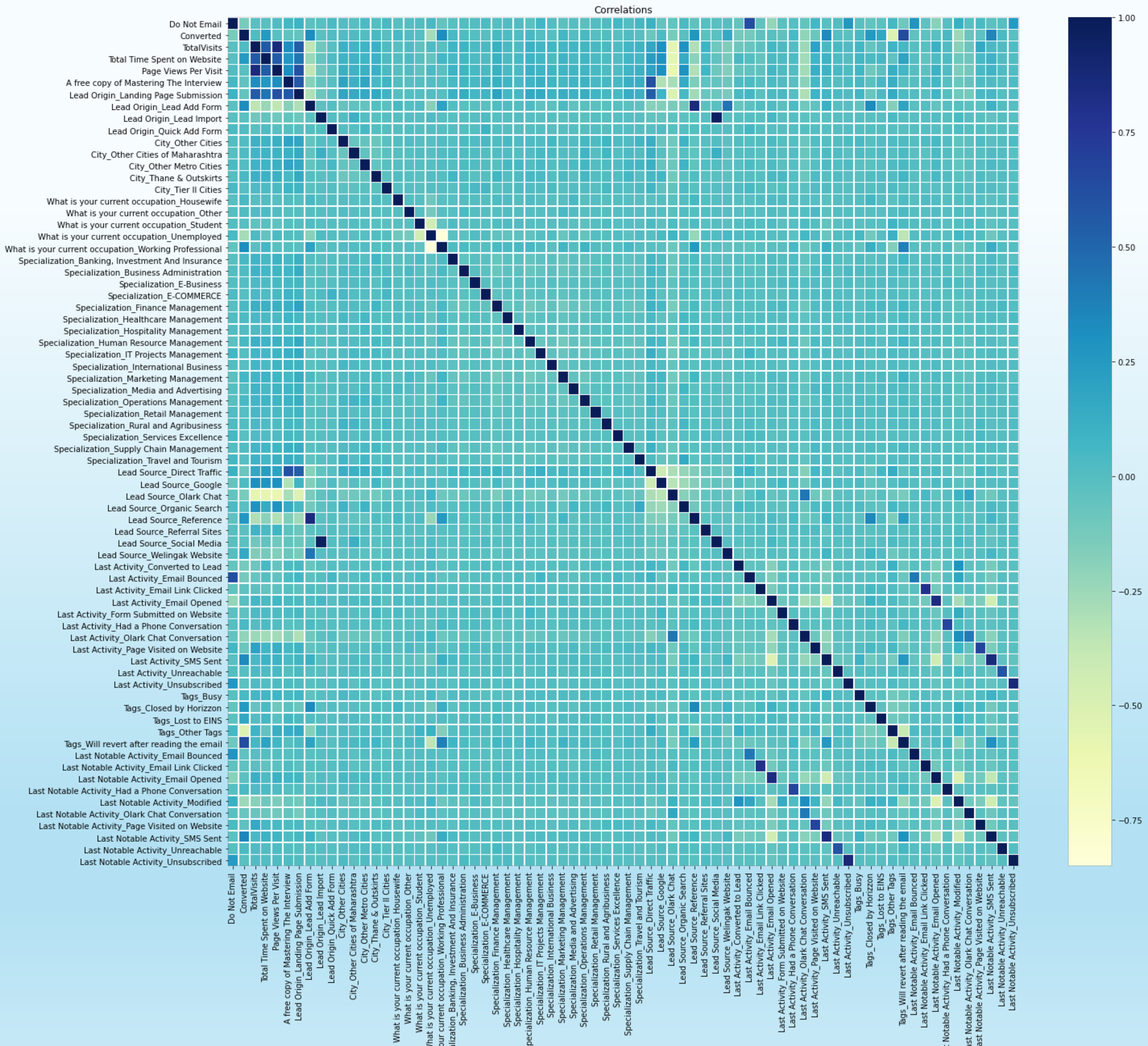
After Outlier Treatment:

As we can see here that the points which were too high in magnitude were removed.

Here we simply chose to remove 5% of the lower range values and 1% of the upper range values

Correlation(Heatmaps)

- After Dummy variable creation and removing redundant columns from it, we plot a heatmap to check for most correlated variables in the dataset.
- We will check those variables after creating our model to see their impact. RFE automatically deals with such correlations hence as of now we just visualize it.



Model Building

- We used RFE to build our model since there are many columns in the dataset and doing a backward or a forward variable selection would be cumbersome. Hence RFE deals with the insignificant variables.

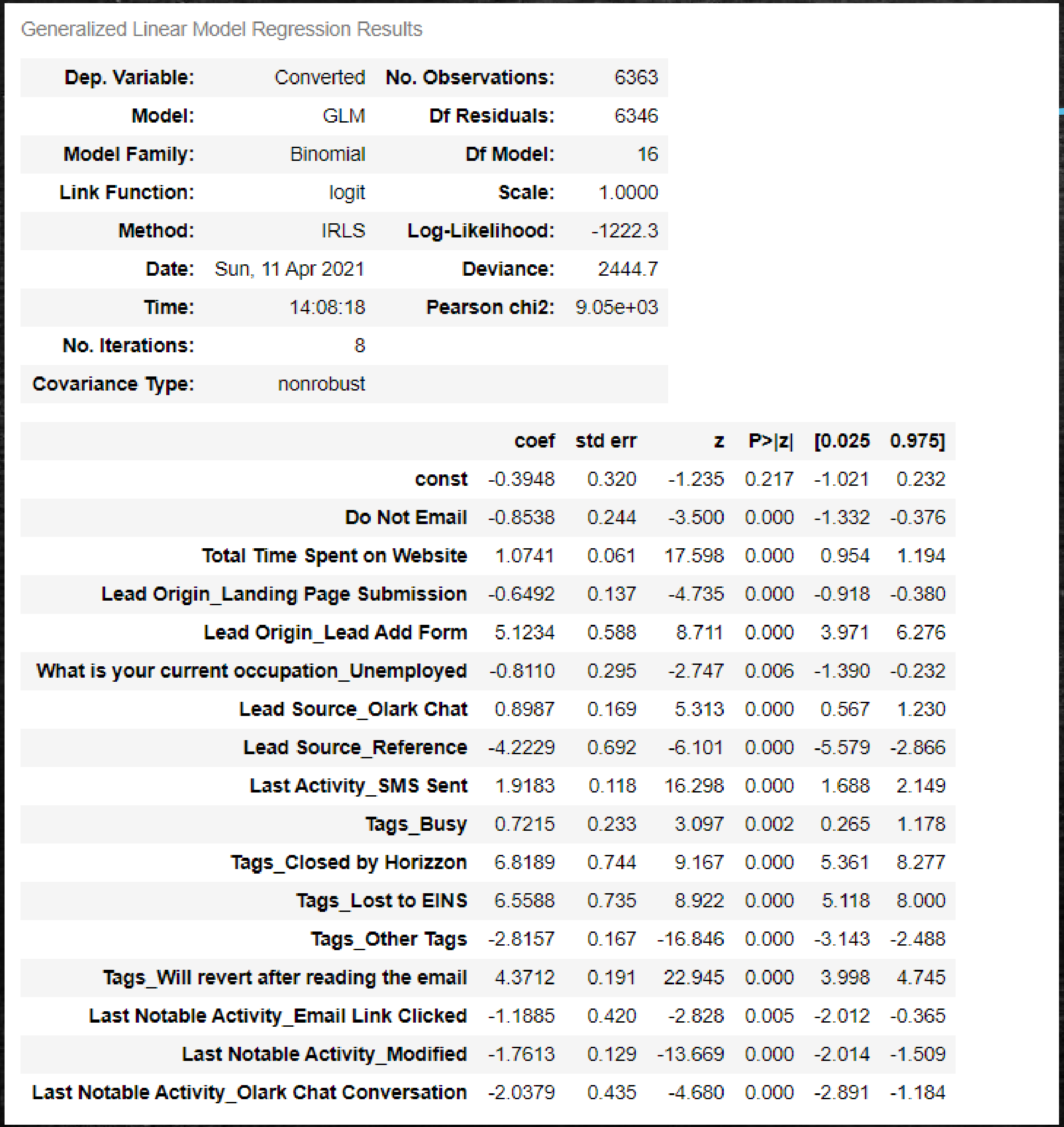
- We start our model building with 20 variables. Hence RFE selects 20 most significant variables.

- After that we drop those variables which have P-Values greater than 5%.

- Lastly, we drop those variables which have VIF values greater than 5.

- Our 5th model is the final model which has all parameters under check.

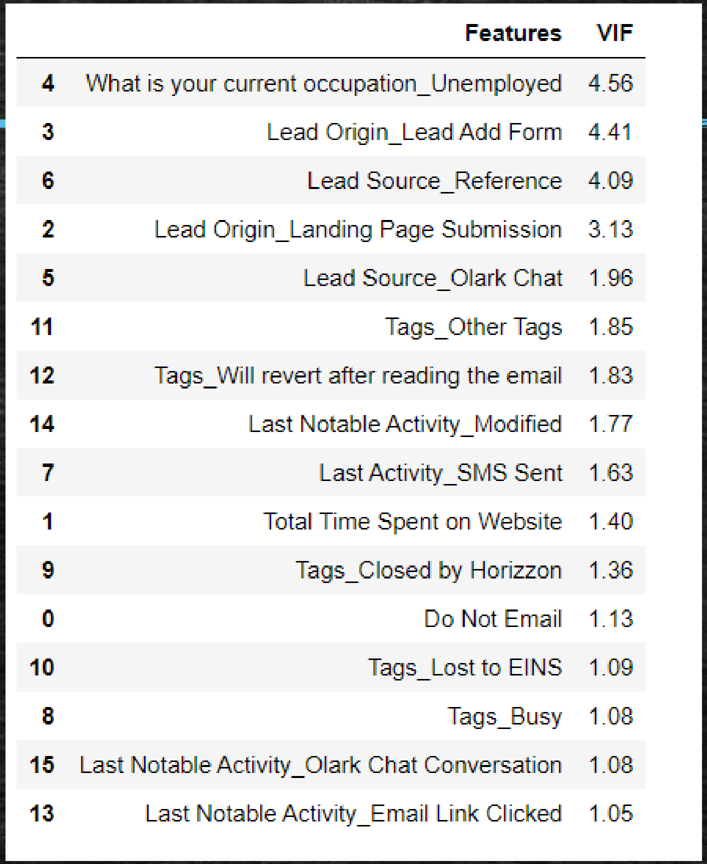
Final Model Visualizations



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6363
Model:	GLM	**Df Residuals:**	6346
Model Family:	Binomial	**Df Model:**	16
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1222.3
Date:	Sun, 11 Apr 2021	**Deviance:**	2444.7
Time:	14:08:18	**Pearson chi2:**	9.05e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

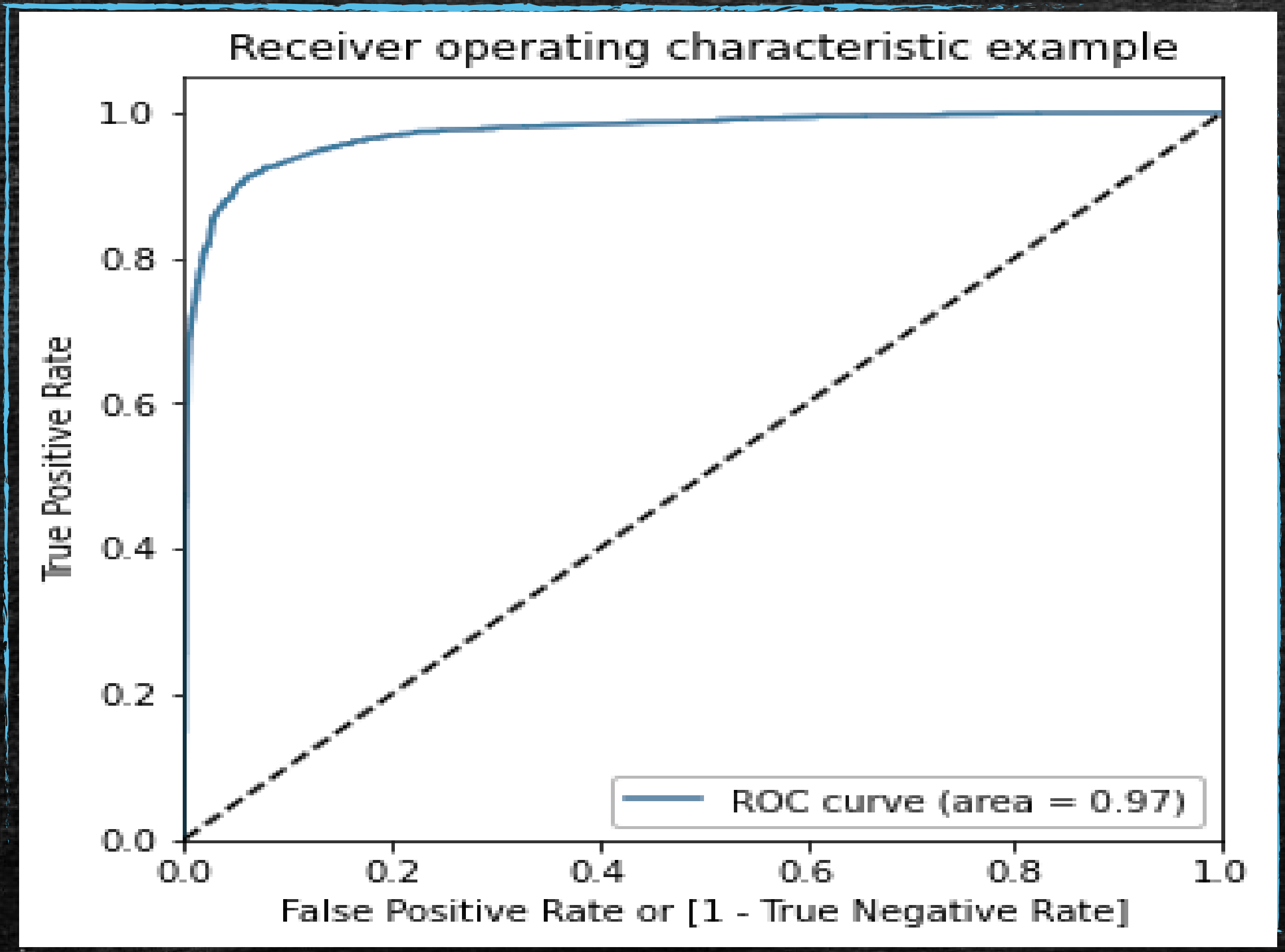
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.3948	0.320	-1.235	0.217	-1.021	0.232
Do Not Email	-0.8538	0.244	-3.500	0.000	-1.332	-0.376
Total Time Spent on Website	1.0741	0.061	17.598	0.000	0.954	1.194
Lead Origin_Landing Page Submission	-0.6492	0.137	-4.735	0.000	-0.918	-0.380
Lead Origin_Lead Add Form	5.1234	0.588	8.711	0.000	3.971	6.276
What is your current occupation_Unemployed	-0.8110	0.295	-2.747	0.006	-1.390	-0.232
Lead Source_Olark Chat	0.8987	0.169	5.313	0.000	0.567	1.230
Lead Source_Reference	-4.2229	0.692	-6.101	0.000	-5.579	-2.866
Last Activity_SMS Sent	1.9183	0.118	16.298	0.000	1.688	2.149
Tags_Busy	0.7215	0.233	3.097	0.002	0.265	1.178
Tags_Closed by Horizzon	6.8189	0.744	9.167	0.000	5.361	8.277
Tags_Lost to EINS	6.5588	0.735	8.922	0.000	5.118	8.000
Tags_Other Tags	-2.8157	0.167	-16.846	0.000	-3.143	-2.488
Tags_Will revert after reading the email	4.3712	0.191	22.945	0.000	3.998	4.745
Last Notable Activity_Email Link Clicked	-1.1885	0.420	-2.828	0.005	-2.012	-0.365
Last Notable Activity_Modified	-1.7613	0.129	-13.669	0.000	-2.014	-1.509
Last Notable Activity_Olark Chat Conversation	-2.0379	0.435	-4.680	0.000	-2.891	-1.184



	Features	**VIF**
4	What is your current occupation_Unemployed	4.56
3	Lead Origin_Lead Add Form	4.41
6	Lead Source_Reference	4.09
2	Lead Origin_Landing Page Submission	3.13
5	Lead Source_Olark Chat	1.96
11	Tags_Other Tags	1.85
12	Tags_Will revert after reading the email	1.83
14	Last Notable Activity_Modified	1.77
7	Last Activity_SMS Sent	1.63
1	Total Time Spent on Website	1.40
9	Tags_Closed by Horizzon	1.36
0	Do Not Email	1.13
10	Tags_Lost to EINS	1.09
8	Tags_Busy	1.08
15	Last Notable Activity_Olark Chat Conversation	1.08
13	Last Notable Activity_Email Link Clicked	1.05

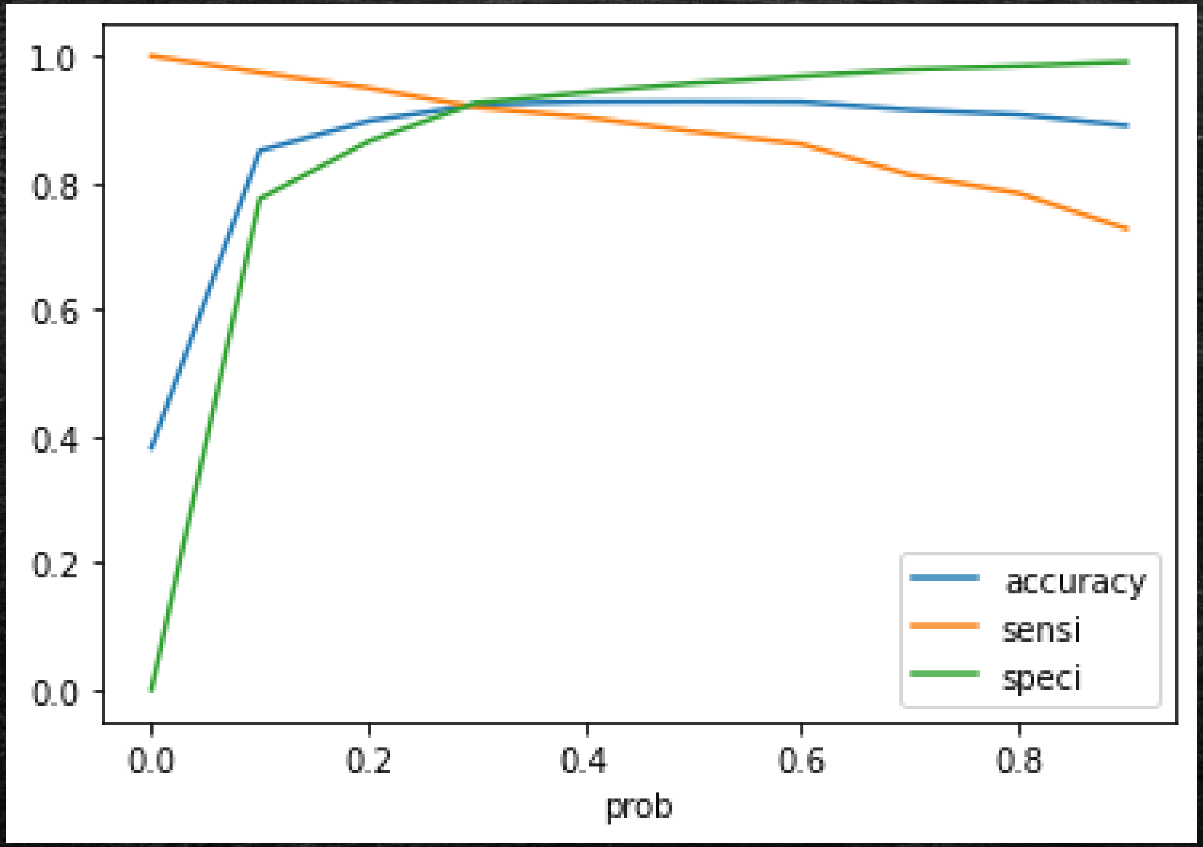
All the P-Values and the VIF Values are in the acceptable range. Hence this model is statistically good.

Model Evaluation



- Now that the model has been built, we need to check the stability of the model. To do this we plot the ROC Curve and the AUC score (Area under the curve).
- As we can see that the area under the curve is 0.97 which means that the model has good predictive power.
- Also, the graph is curved towards the upper left of the border approaching 1 which only goes to say that the model has a very good accuracy.

Finding Optimal Cut off Point



- We have created range of probabilities for ranging from 0.0 to 0.9 where we can find the best balance of metrics like accuracy, sensitivity and specificity.

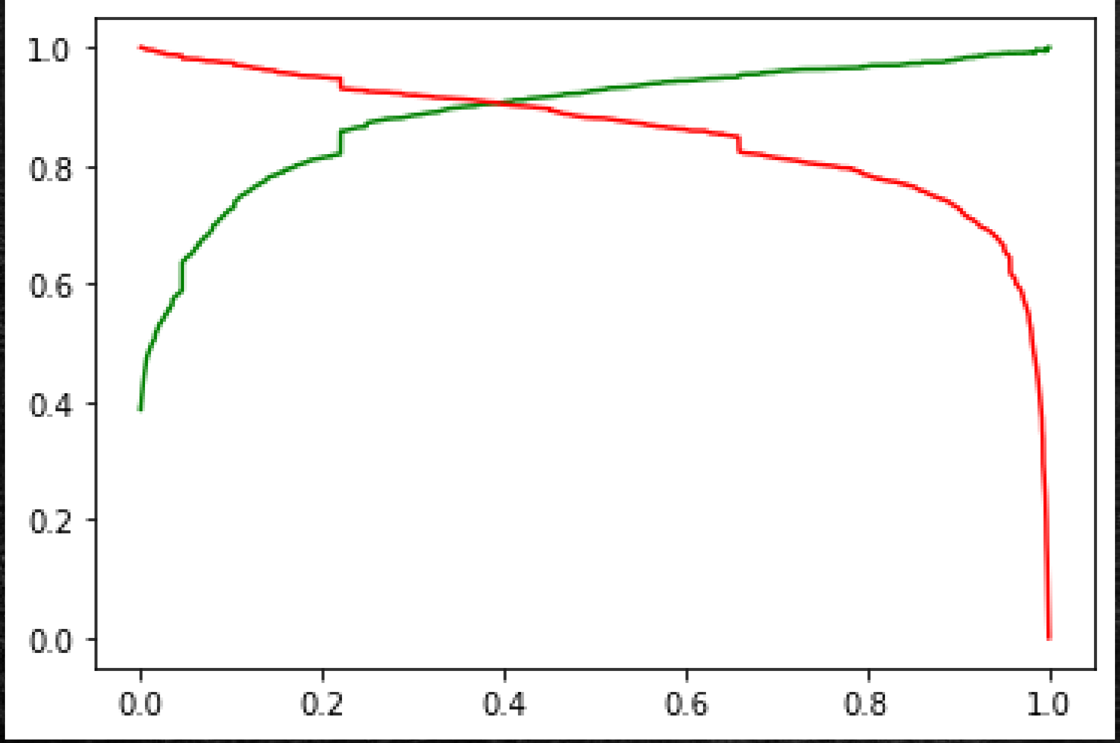
- For this we plot this graph, and we can see that the point of intersection is somewhere around 0.3. Hence that is the ideal optimum cutoff probability.

- When the probability is 0.3, the values of accuracy, sensitivity and specificity are in very close range.

Precision and Recall

- Now after finding the optimal point, we append a column in the dataframe which predicts the outcomes.
- We now check other essential parameters like Precision and Recall since it is important from business point of view as they depict our model behavior.
- We see that our Recall (Sensitivity) and Precision values comes out to be 0.92 and 0.88 approx. respectively.
- If we look at the business objective of this assignment, we need to correctly identify the actual hot leads by our model. Hence, we will prioritize recall over precision because recall is the measure of identifying the actual positive correctly.
- Hence with our final model which is having a recall value of 0.92, the model seems perfect.

Precision-Recall Curve



- There is a trade off between precision and recall because while one goes up the other goes down.
- Mathematically both are inversely proportional.
- Hence, we need to find an optimal value for this trade off.
- For that we plot Precision-Recall Curve.
- From the graph we see that the intersection point is approximately at 0.4

Predictions on Test Set

- Before we do any kind of predictions we need to scale the test dataset.
- We should only transform the test dataset and not fit it.
- After transforming, we predict the outcomes with the help of optimal cut-off probability.
- When we create the confusion matrix and find out metrics like accuracy, sensitivity and specificity of the dataset, we see that the test dataset also exhibits good predictive power.
- The accuracy is 0.92, sensitivity is 0.90, specificity is 0.93.
- The precision is 0.90 as well.
- We then add the column lead score to this dataframe. Higher the lead score, greater is the chance of getting converted while a lower lead score means that the conversion chances are feeble.

Conclusions:

- We see that the value of sensitivity and precision are high which means that the model is rightly able to detect the actual converted and non converted cases really well. In this manner our business objective is getting fulfilled.
- Hence this model is able to adjust to the company's requirement in the future, if any.
- This only goes to say that our final model is stable.
- The final model parameters are as shown

Parameters in the final model	
const	-0.394782
Do Not Email	-0.853794
Total Time Spent on Website	1.074097
Lead Origin_Landing Page Submission	-0.649220
Lead Origin_Lead Add Form	5.123421
What is your current occupation_Unemployed	-0.811015
Lead Source_Olark Chat	0.898705
Lead Source_Reference	-4.222895
Last Activity_SMS Sent	1.918348
Tags_Busy	0.721461
Tags_Closed by Horizzon	6.818901
Tags_Lost to EINS	6.558759
Tags_Other Tags	-2.815719
Tags_Will revert after reading the email	4.371225
Last Notable Activity_Email Link Clicked	-1.188471
Last Notable Activity_Modified	-1.761277
Last Notable Activity_Olark Chat Conversation	-2.037923

Recommendations:

- To increase the conversion rate, the sales team can particularly focus on the following features:

1.Tags_Closed by Horizzon
2.Tags_Lost to EINS
3.Lead Origin_Lead Add Form
4.Tags_Will revert after reading the email
5.Last Activity_SMS Sent
6.Total Time Spent on Website

- Also, if they want to call the number of potential leads (i.e., the customers who have been predicted as 1 by the model) then the sales team need to see the final dataframe and call the respective prospect ID customer who have a lead score above 30 since there are customers who have been predicted as 1. Below it there are no customers who are predicted as 1 by the final model.

- To reduce the number of useless phone calls, the sales team need to not focus on these features since they lower the conversion rate chances:
- 1.Lead Source_Reference
- 2.Tags_Other Tags
- 3.Last Notable Activity_Olark Chat Conversation
- 4.Last Notable Activity_Modified
- 5.Last Notable Activity_Email Link Clicked
- Also, to avoid wastage of time, sales team should call only those Prospect ID customers having a lead score of greaterthan 90 since they have a higher chance of getting converted.

Thank you