Lead Score Case Study

Harish JK

Anshuman Padhy

Problem Statement

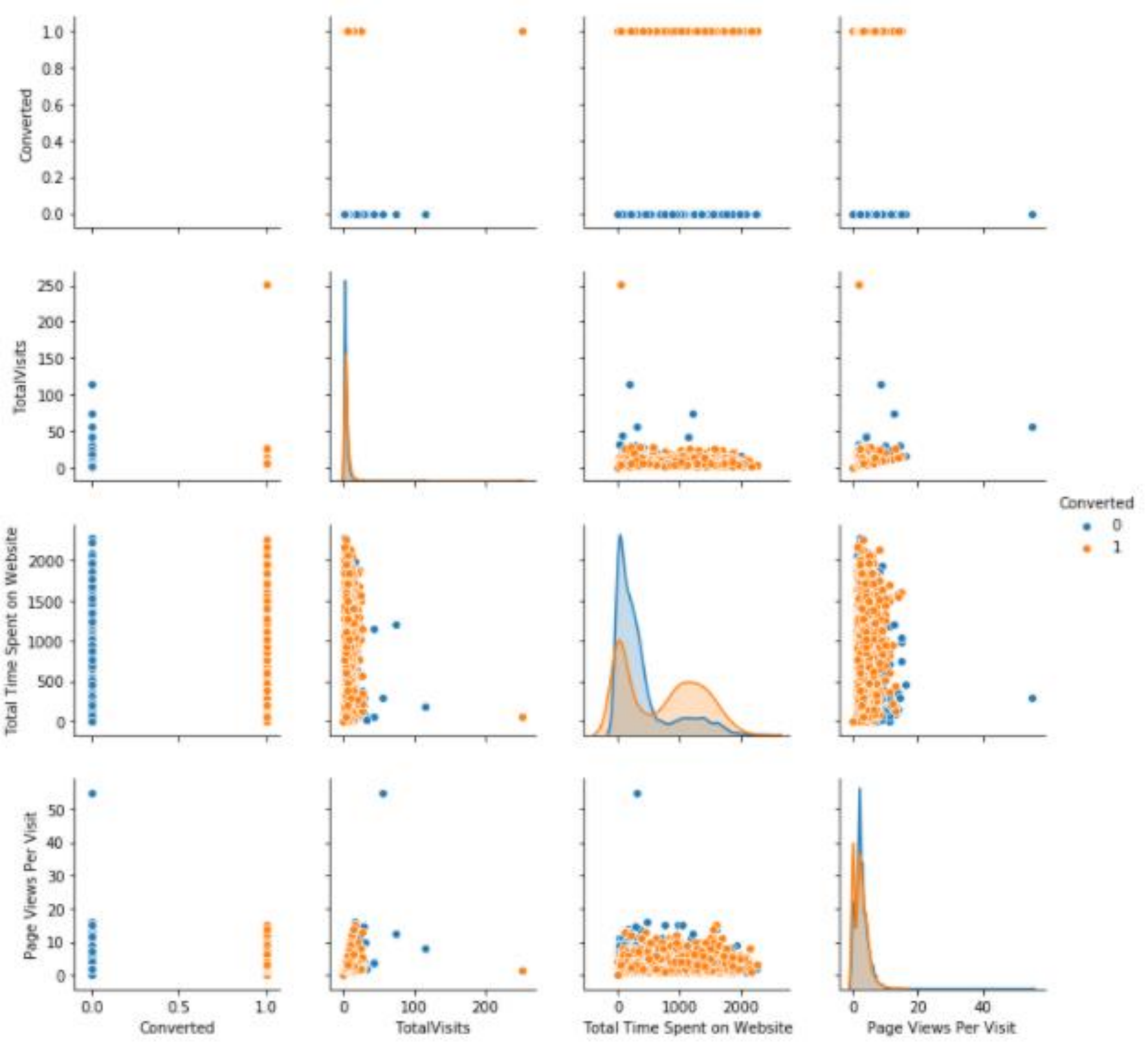
- Creating a Logistic Regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the Education company to target potential leads for their business.
- To achieve a target lead conversion rate around 80%.

Data Visualization

- Plotted a pair plot to visualize the relationships between each variables in a data set.

- After plotting we can get some insights of the relation between each variables.

- Here we have four variables in our pair plot.

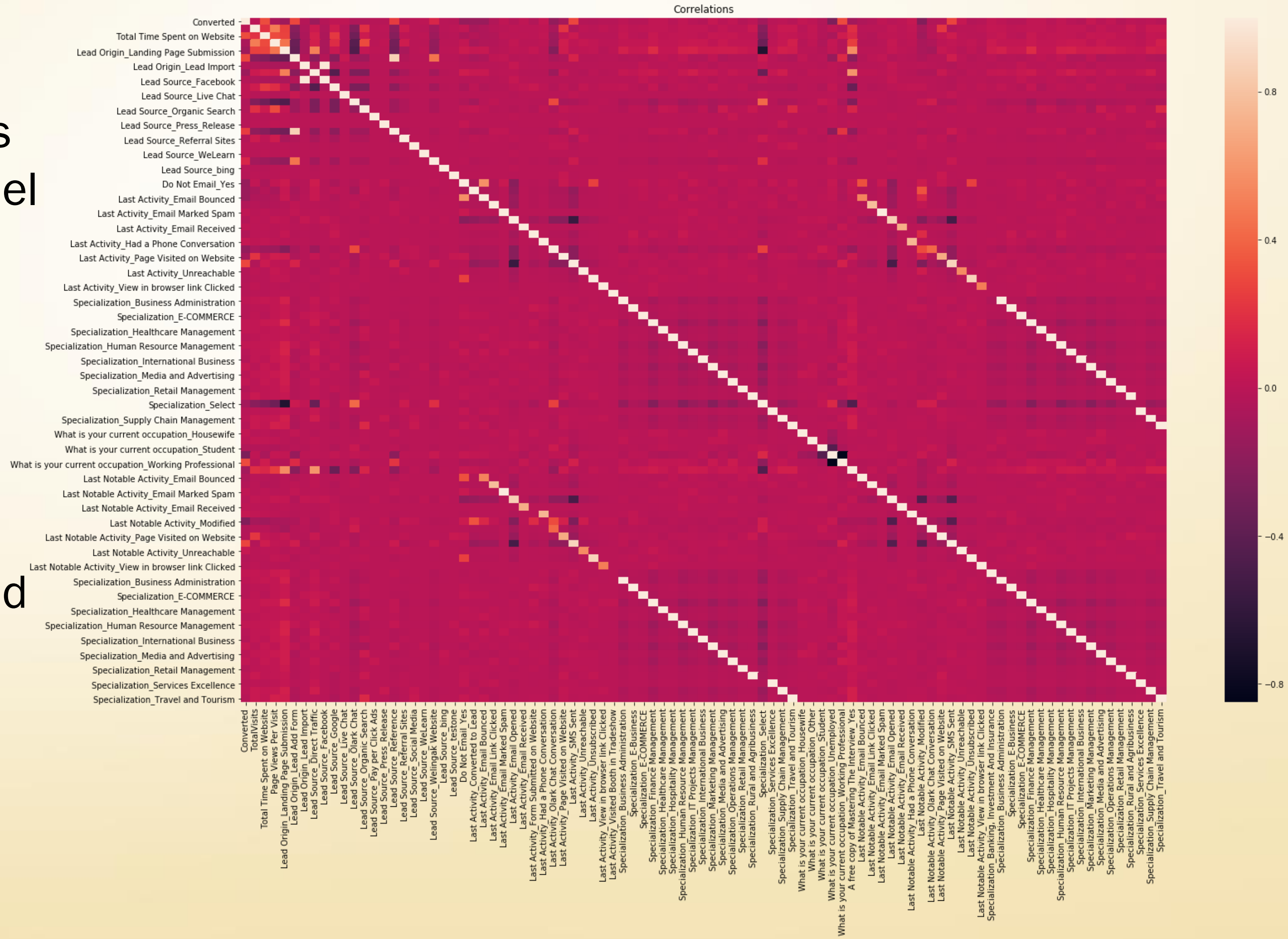


Correlation

- After creating the dummy variables in our dataset to increase the model accuracy, it generates a set of new features

- Plotted a heat map on these features to get some insights.

- Looks like its difficult to understand the correlation between due to more no of features.

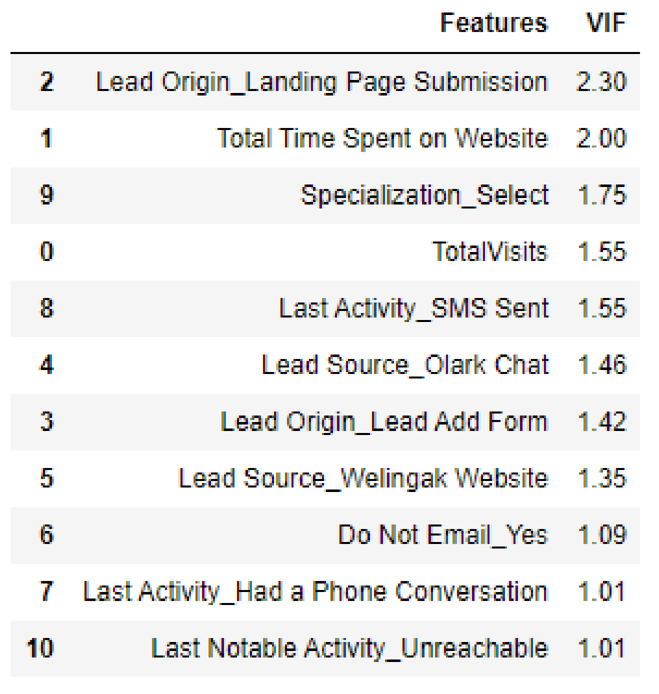


Building the model

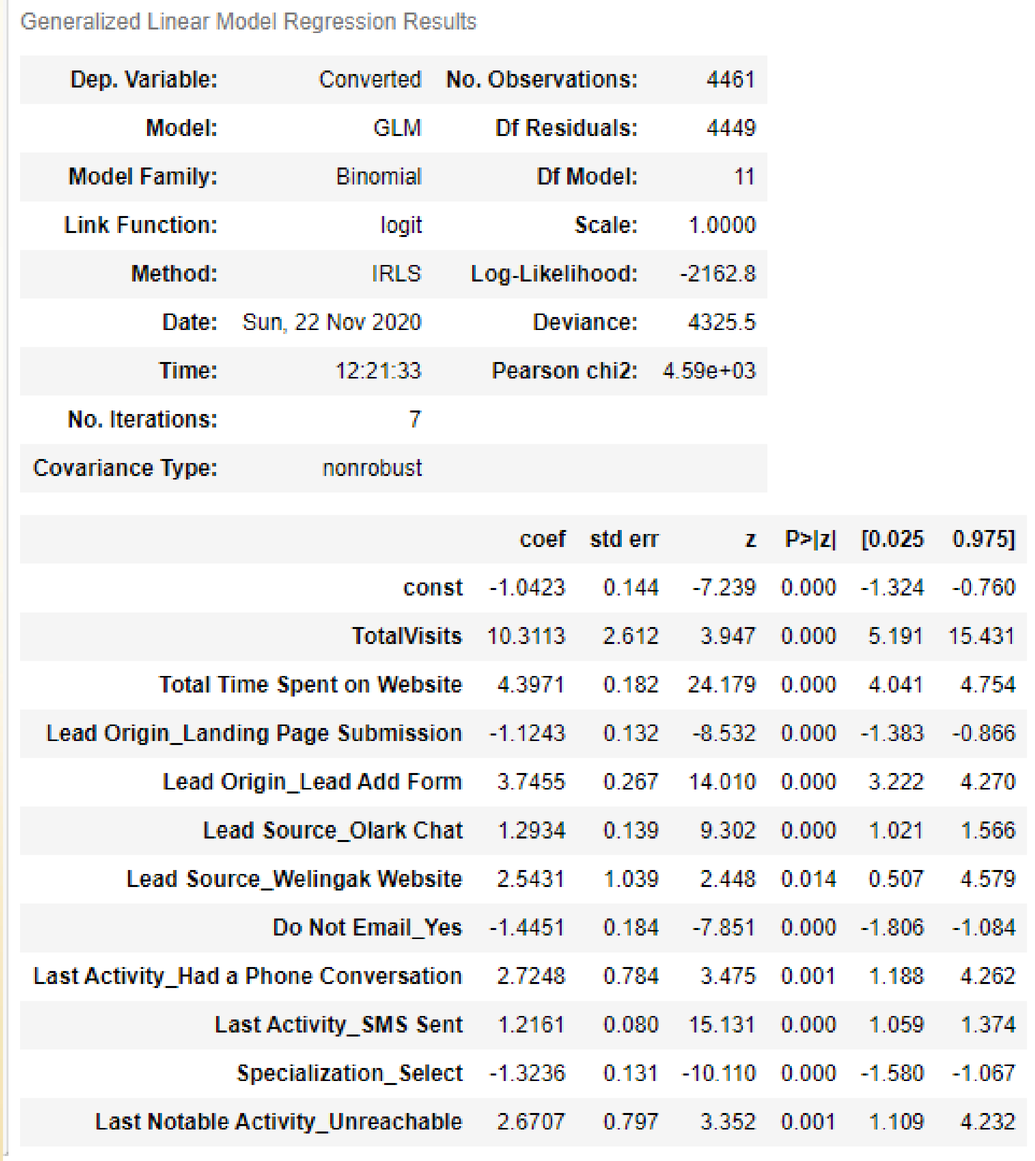
- We have started to build our logistic regression model by using RFE method.
- We have chosen RFE count 16 for the model and went dropping the unwanted features one by one until we reach the model stable.
- Also we need to consider that our model should consists of features with low p-value and low VIF values less that 5.0
- Adding a constant using statsmodel and predicting the scores.

Our Model

- Finally we have build our model using RFE method and achieved a good accuracy and significant variables with less than 0.05 p-value and low VIF value of 5.0



	Features	VIF
2	Lead Origin_Landing Page Submission	2.30
1	Total Time Spent on Website	2.00
9	Specialization_Select	1.75
0	TotalVisits	1.55
8	Last Activity_SMS Sent	1.55
4	Lead Source_Olark Chat	1.46
3	Lead Origin_Lead Add Form	1.42
5	Lead Source_Welingak Website	1.35
6	Do Not Email_Yes	1.09
7	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2162.8
Date:	Sun, 22 Nov 2020	**Deviance:**	4325.5
Time:	12:21:33	**Pearson chi2:**	4.59e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

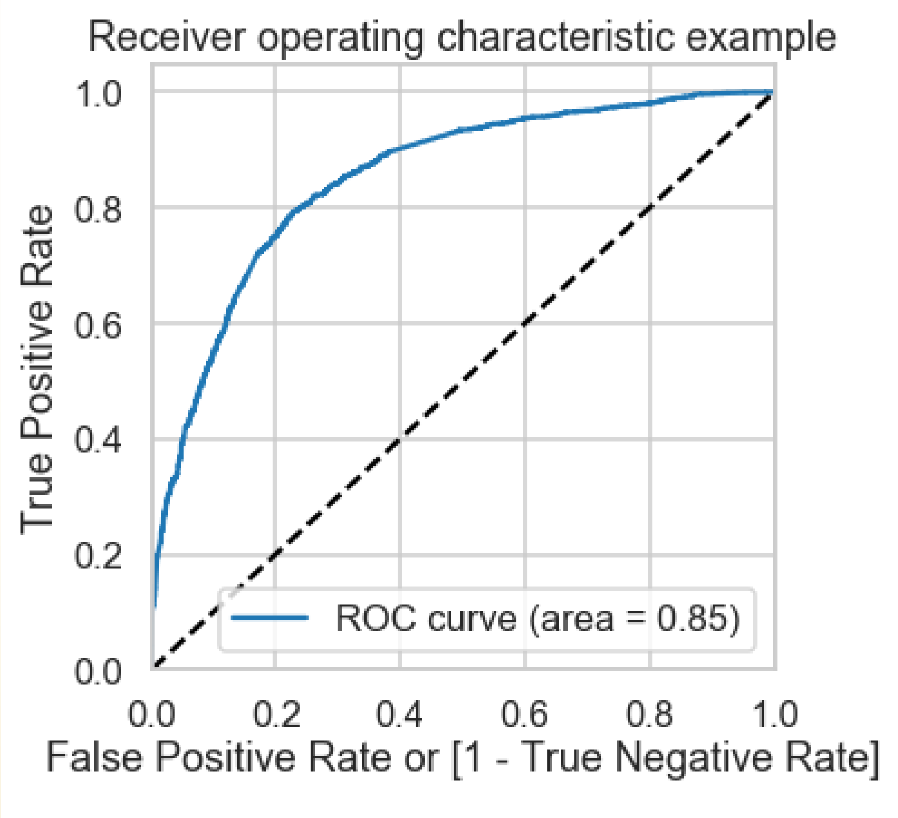
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0423	0.144	-7.239	0.000	-1.324	-0.760
TotalVisits	10.3113	2.612	3.947	0.000	5.191	15.431
Total Time Spent on Website	4.3971	0.182	24.179	0.000	4.041	4.754
Lead Origin_Landing Page Submission	-1.1243	0.132	-8.532	0.000	-1.383	-0.866
Lead Origin_Lead Add Form	3.7455	0.267	14.010	0.000	3.222	4.270
Lead Source_Olark Chat	1.2934	0.139	9.302	0.000	1.021	1.566
Lead Source_Welingak Website	2.5431	1.039	2.448	0.014	0.507	4.579
Do Not Email_Yes	-1.4451	0.184	-7.851	0.000	-1.806	-1.084
Last Activity_Had a Phone Conversation	2.7248	0.784	3.475	0.001	1.188	4.262
Last Activity_SMS Sent	1.2161	0.080	15.131	0.000	1.059	1.374
Specialization_Select	-1.3236	0.131	-10.110	0.000	-1.580	-1.067
Last Notable Activity_Unreachable	2.6707	0.797	3.352	0.001	1.109	4.232

ROC Curve

- After the model building, we are plotting a ROC curve to find the stability of the model with auc score.

- Area under score (auc) of our model is 0.85. As seen in the graph we have achieved a good score.

- Also if you look into our graph, our graph is leaned towards the left side of the Y axis which means we have a very good accuracy in our model.

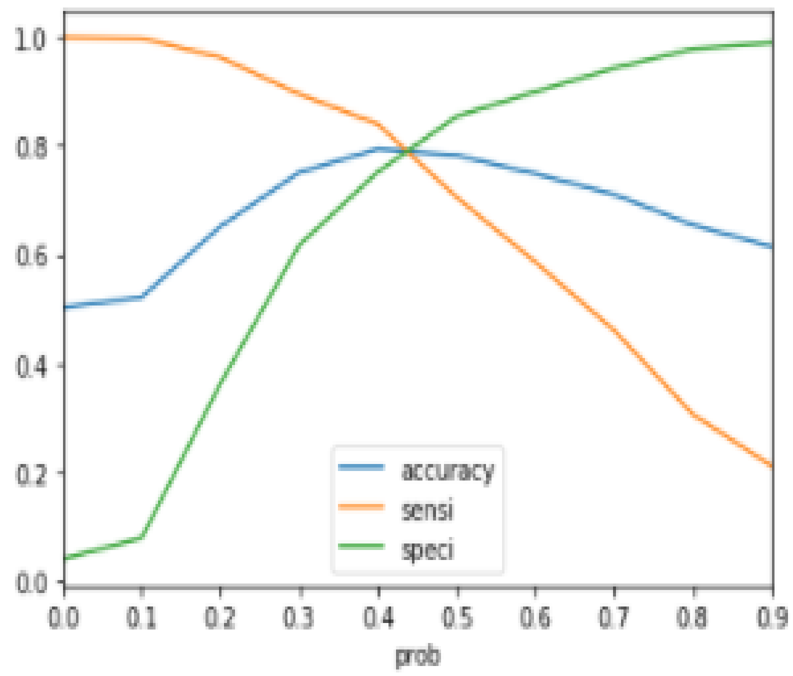


Finding Optimal Cut-off Point

- Now we have created range of points for which we have found accuracy, sensitivity and specificity for each points.

- Looking at the points we can set a valuable cut off point as we can consider 0.4 as cut-off point.

- In order to verify we have plotted all these in a graph, we have got 0.4 as a optimum cut-off point which meets all three lines.



From the curve above, 0.4 is the optimum point to take it as a cutoff probability.

Precision and Recall

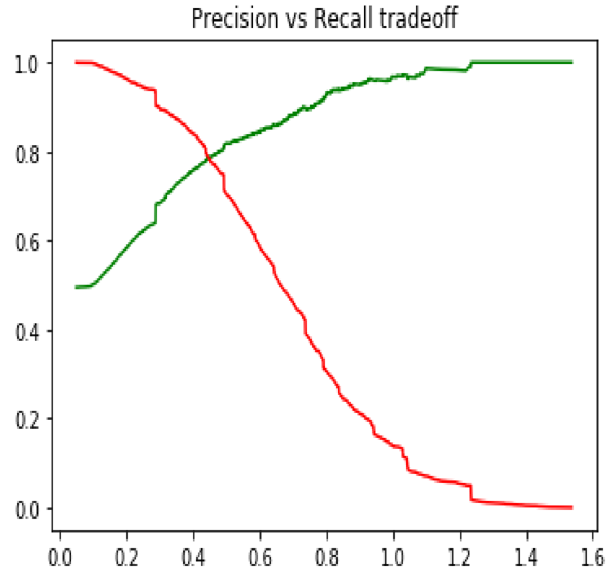
- After achieving the cut-off point we can create a new column for the predicted values.
- Precision and Recall are very important in any model and it contributes more towards business point of view. Also it delivers how our model behaves.
- We have evaluated the precision and recall scores for our model and it is found to be 0.78 and 0.74 respectively.
- According to our business requirement I will focus more on Recall percentage since we don’t want to lose any hot leads which are willing to get converted.
- Since focusing more Recall we get more results of hot leads which gets converted.

Precision and Recall trade-off

- Precision-Recall tradeoff occur due to increasing one of the parameter (precision or recall) while keeping the model same.

- This graph clearly helps us to find out the meeting point of Precision and Recall.

- We can conclude that the meeting point is approximately around 0.5



As we can see that there is a trade off between Precision and Recall and the meeting point is nearly at 0.5

Predictions on Test set

- We have build a model on train set and predicted scores, let us run the model with same features on test set to check the accuracy of the model.
- The new prediction on test set is stored in a new data frame.
- Again accuracy, precision and recall are calculated. We get 0.76 accuracy, 0.73 precision and 0.80 as recall score.
- This shows that model is having an acceptable range of scores on the test set.
- Finally lead score is created on the test set to identify the hot leads. High the lead score higher the chance of lead conversion and low the lead score lower the chance of lead conversion.

Conclusion

- As expected we have got promising scores of Accuracy, Precision and Recall in our test set and also we have compared the same with the train set also.
- As per the business needs we have got a higher Recall score when compared to Precision.
- In Business terms, the model can be adjusted according to the company requirement in the future.
- Thus our 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 :
  - ✓ TotalVisits
  - ✓ Total Time Spent on Website
  - ✓ Lead Origin_Lead Add Form