Lead Score Case Study

JITESH KURIAN
NAVITHA MUNIRAJU
SANJEEV KUMAR

Problem Statement

X Education is an education company which sells online courses to industry professionals .Whoever are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

The company has 30% conversion rate though the process of turning leads in to customers are found having interest in taking the course. 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 Goal

- The company will need a model which will select the most promising leads.
- 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The model built should have the lead conversation rate around 80% or more.
- The higher the lead score the more promising the lead to get converted , the lower it is the lesser the chances of conversion

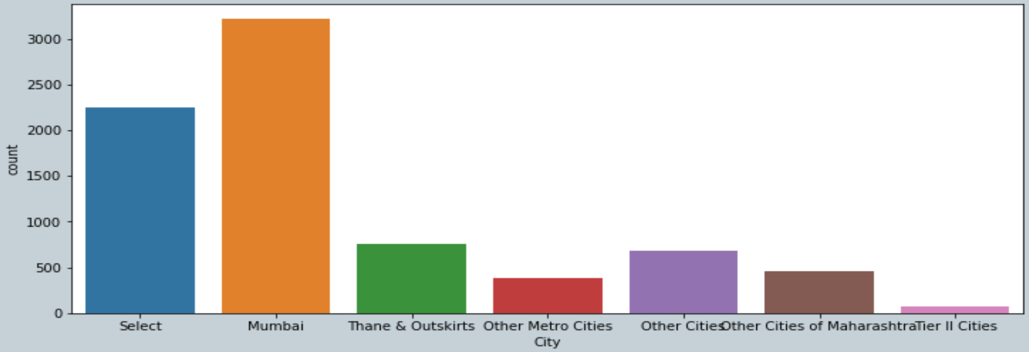
Analysis Approach

- Import the data.
- Data Cleaning and Preparation.
- Prepare the data for modelling.
- Dummy variable creation.
- Test-train Split.
- Scaling.
- Looking at the correlations.
- Model Building
- Model Evaluation.
- Finding the optimal cutoff.
- Making Predictions on the Test Set.
- Precision-Recall View and Tradeoff.
- Summary.

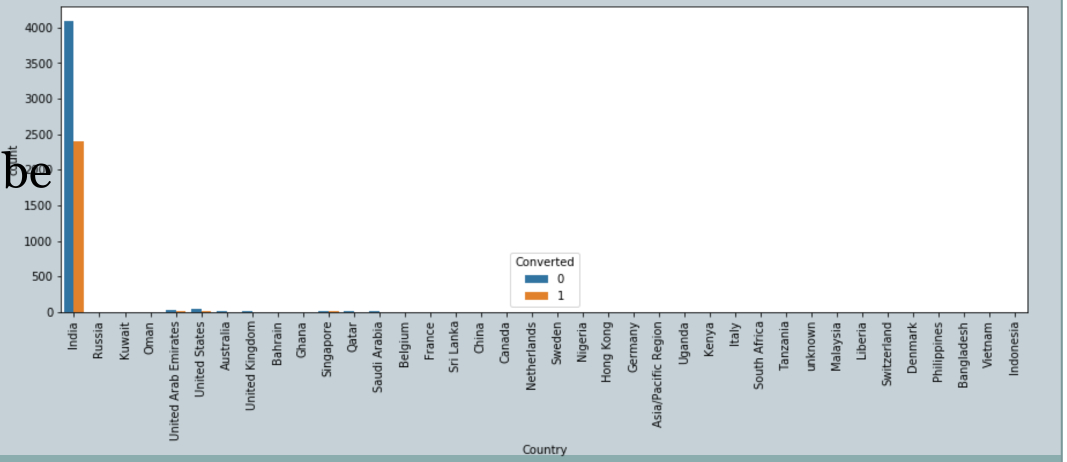
Analysis Approach

Exploratory Data Analysis

1. From our Analysis we could make that variable City would be of no use and hence its better to drop.



2. From our Analysis we could make that variable Country would be of no use and hence its better to drop.

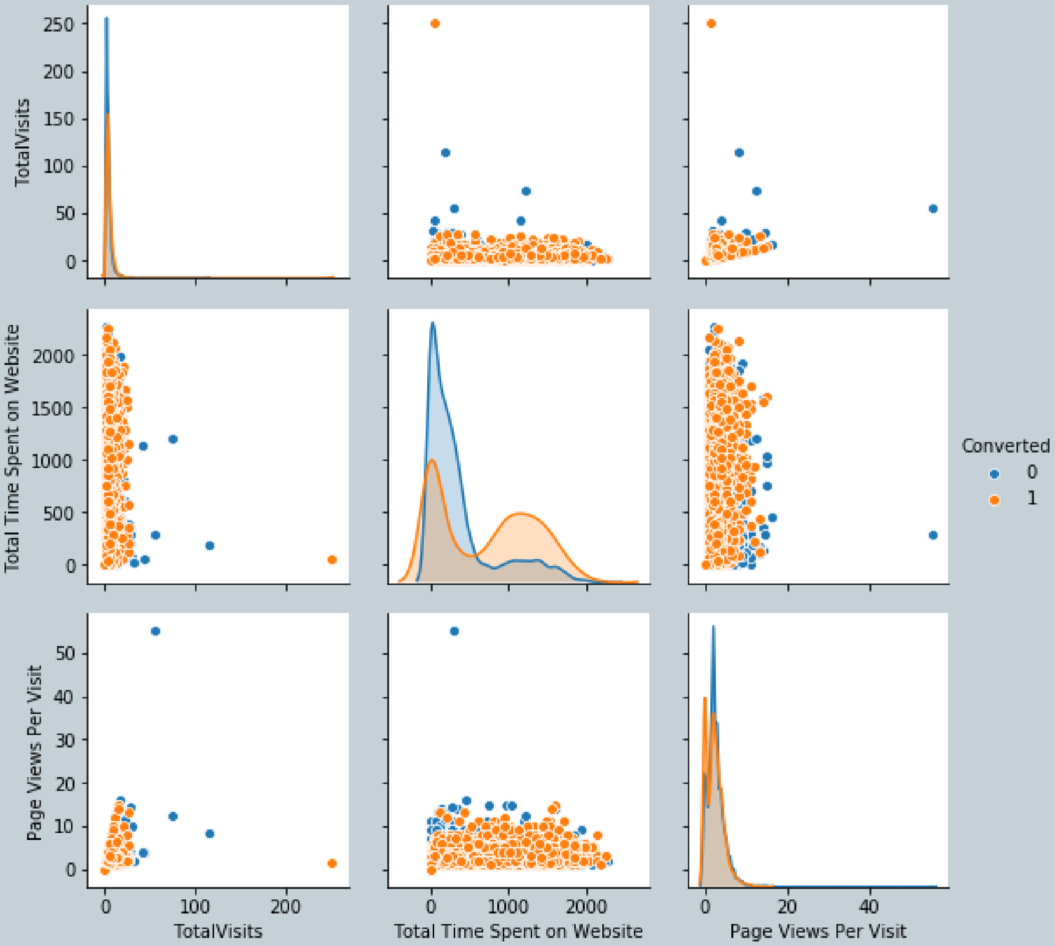


Similarly there are many other columns in which only one value was majorly present for all the data points. Since all of these values are No, its best that we drop these columns as they wont help our analysis. Below are the columns.

1. Do Not Call
2. Search
3. Magazine
4. Newspaper Article
5. X Education Forums
6. Newspaper
7. Digital Advertisement
8. Through Recommendations
9. Receive More Updates About Our Courses
10. Update me on Supply Chain Content
11. Get updates on DM Content
12. I agree to pay the amount through cheque

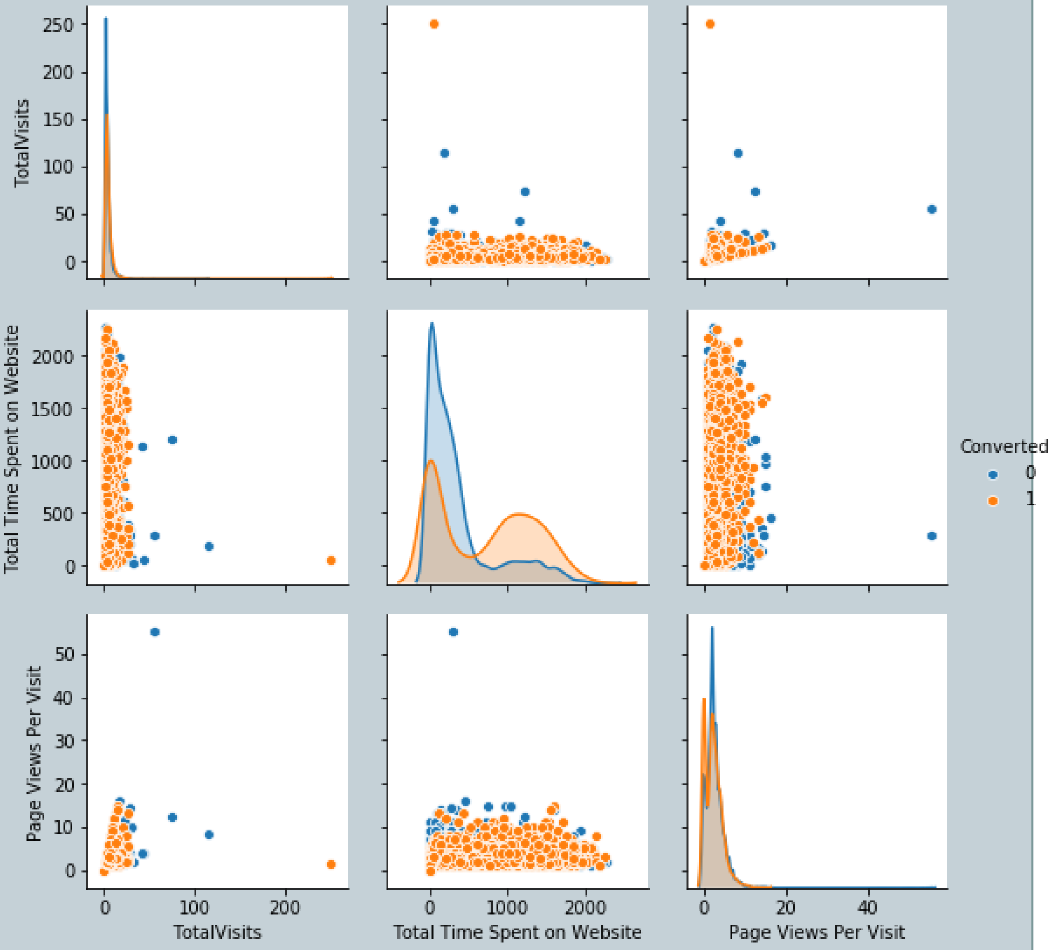
Data Modelling

This is a Pairplot which indicates three most important factors or features



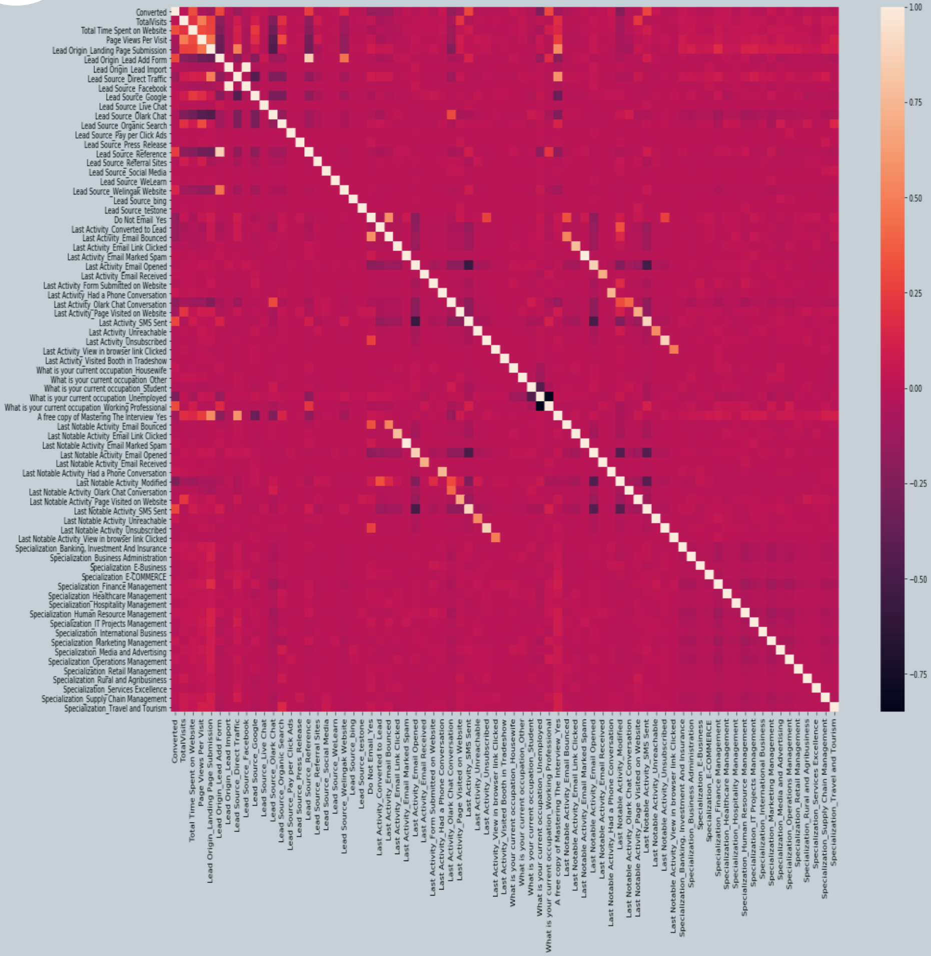
Data Modelling

This Pair plot indicated the detailed information of the columns TotalVisits, Total Time Spent on Website, Page Views Per Visit



Looking at the correlations

This HeatMap represents the corelation between all the columns/features.



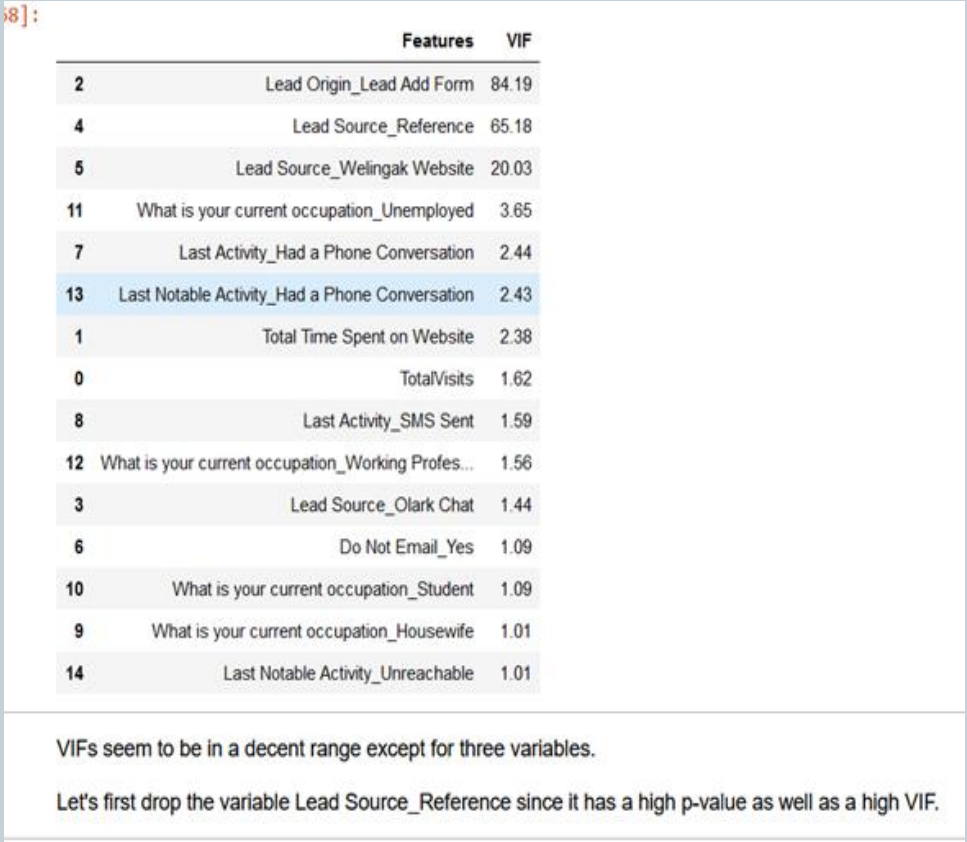
Model Building

- There are a lot of variables present in the dataset which we cannot deal with. So the best way to approach this is to select a small set of features from this pool of variables using RFE.
- Once we have all the variables selected by RFE and since we care about the statistics part, i.e. the p-values and the VIFs, let's use these variables to create a logistic regression model using statsmodels.

Model Building

- Now let us see the VIF dataframe .

- This seems to be good except for 3 variables and hence we can drop which has high p-value .



8]:

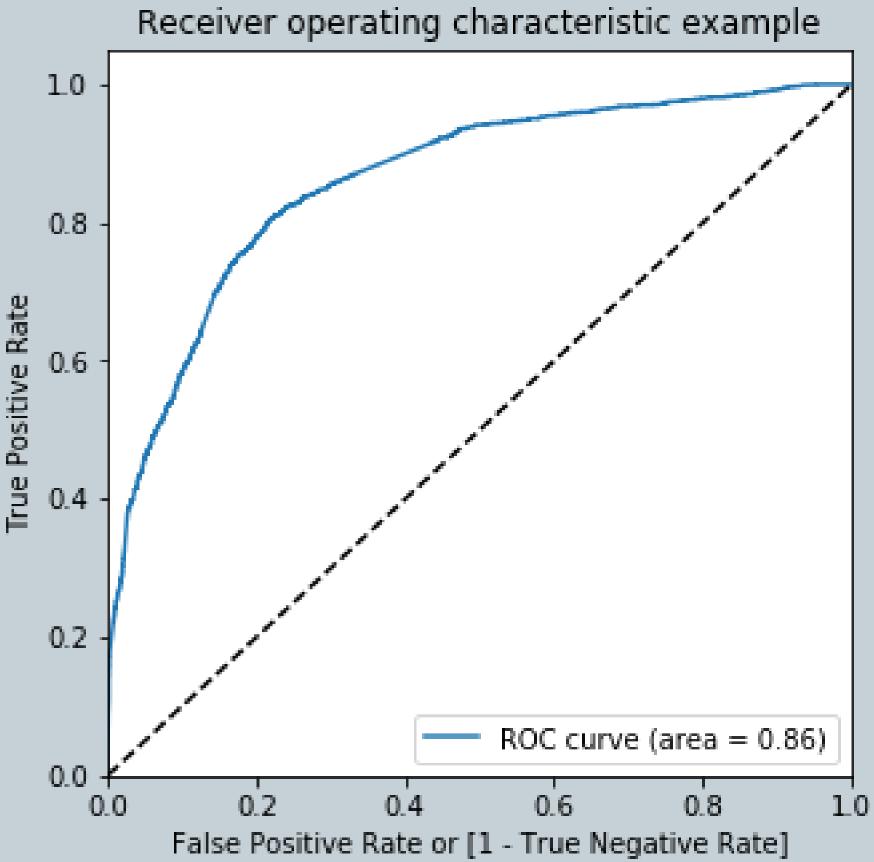
	Features	VIF
2	Lead Origin_Lead Add Form	84.19
4	Lead Source_Reference	65.18
5	Lead Source_Welingak Website	20.03
11	What is your current occupation_Unemployed	3.65
7	Last Activity_Had a Phone Conversation	2.44
13	Last Notable Activity_Had a Phone Conversation	2.43
1	Total Time Spent on Website	2.38
0	TotalVisits	1.62
8	Last Activity_SMS Sent	1.59
12	What is your current occupation_Working Profes...	1.56
3	Lead Source_Olark Chat	1.44
6	Do Not Email_Yes	1.09
10	What is your current occupation_Student	1.09
9	What is your current occupation_Housewife	1.01
14	Last Notable Activity_Unreachable	1.01

VIFs seem to be in a decent range except for three variables.

Let's first drop the variable Lead Source_Reference since it has a high p-value as well as a high VIF.

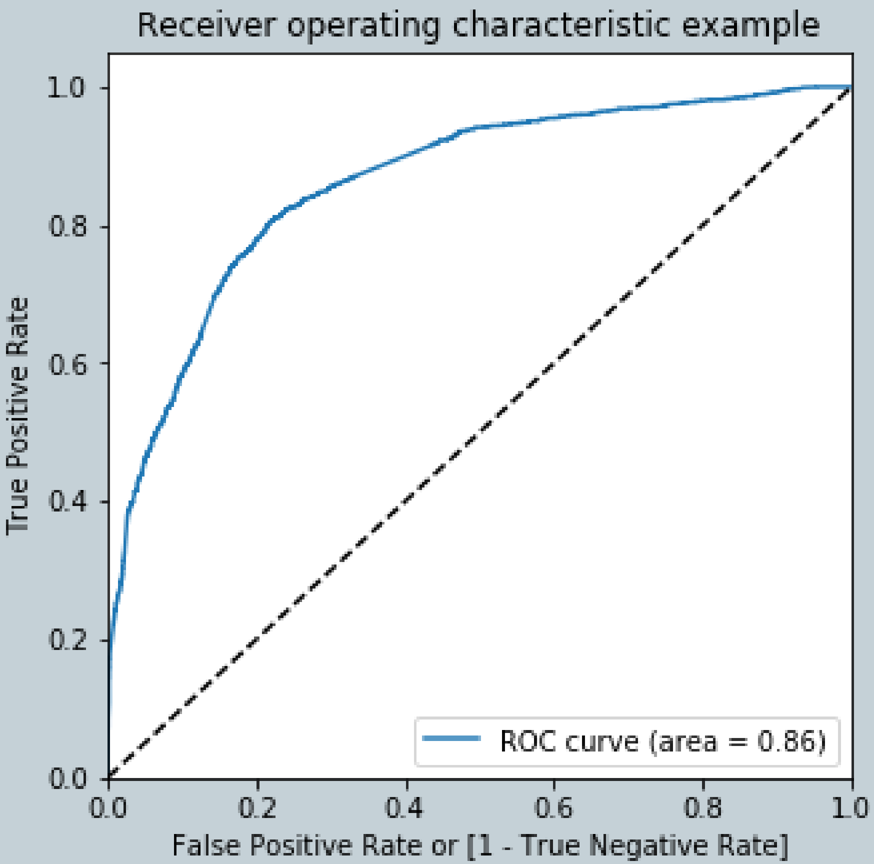
ROC Curve

The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. Let's also check the sensitivity and specificity tradeoff to find the optimal cutoff point.



Precision and Recall Tradeoff

The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. Let's also check the sensitivity and specificity tradeoff to find the optimal cutoff point.

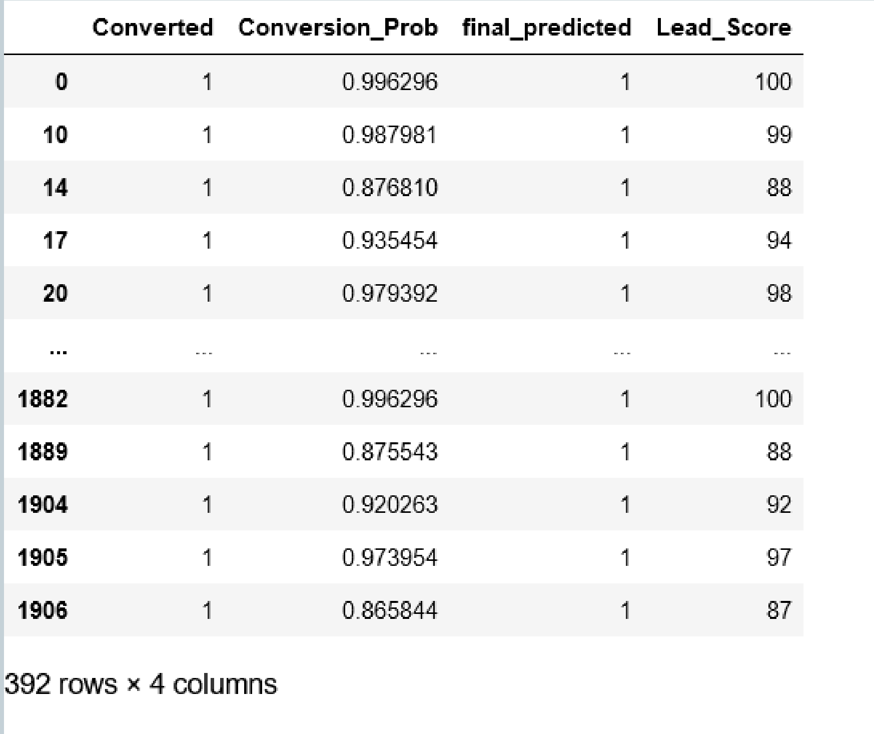


Final Results

- Accuracy: 0.7866108786610879
- Sensitivity: 0.767467248908297
- Specificity: 0.8042168674698795
- Precision: 0.7828507795100222

- By this we can see that the model has given the Lead conversion of 80%.

By this we can see that the leads that can be contacted is 392.



	Converted	Conversion_Prob	final_predicted	Lead_Score
0	1	0.996296	1	100
10	1	0.987981	1	99
14	1	0.876810	1	88
17	1	0.935454	1	94
20	1	0.979392	1	98
...	...	...	...	...
1882	1	0.996296	1	100
1889	1	0.875543	1	88
1904	1	0.920263	1	92
1905	1	0.973954	1	97
1906	1	0.865844	1	87

392 rows × 4 columns