Lead Scoring Case Study

By,

Meera Raghunath

Raghavendra Kothe

Prathmesh Dhule

Problem Statement

- An education company named 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.
- 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Goal

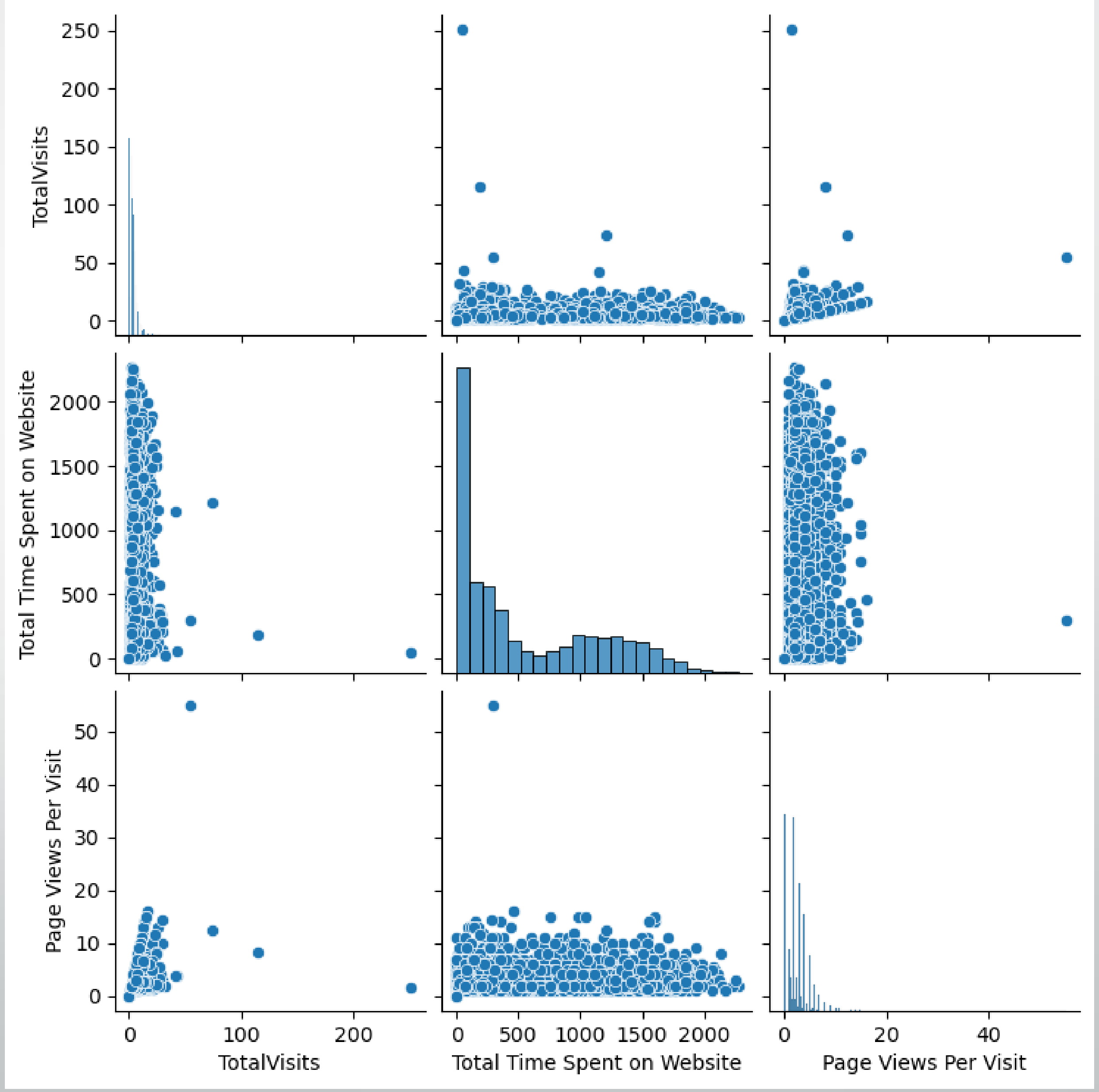
- You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps to be followed

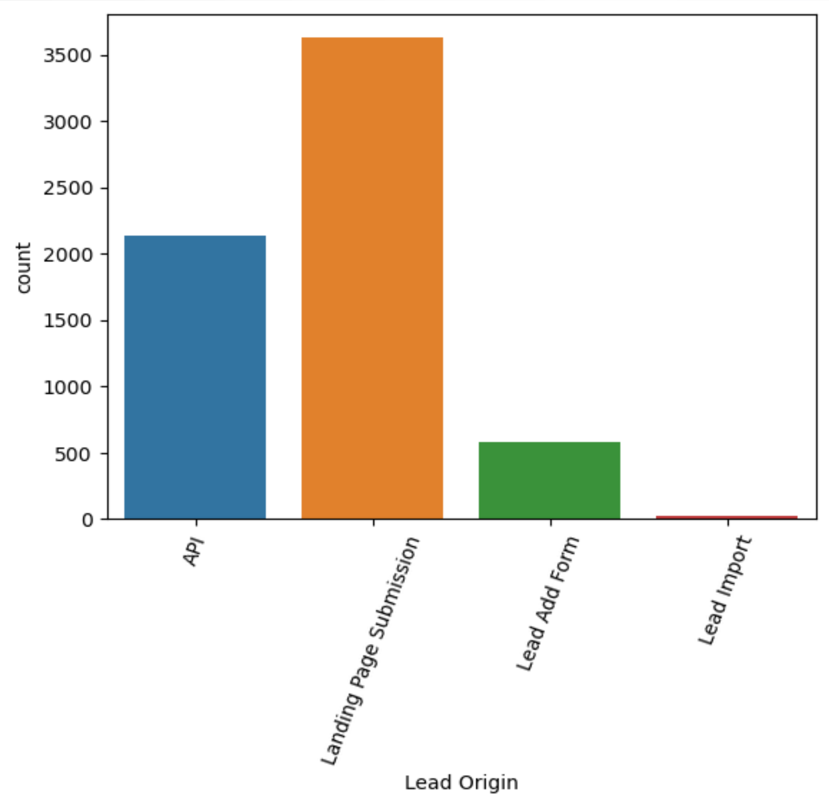
- Step 1: Reading and Understanding Data
- Step 2: Data Cleaning
- Step 3: Exploratory Data Analysis
- Step 4: Creating Dummy Variables
- Step 5: Test Train Split
- Step 6: Feature Rescaling
- Step 7: Feature selection using RFE:
- Step 8: Plotting the ROC Curve
- Step 9: Finding the Optimal Cutoff Point
- Step 10: Computing the Precision and Recall metrics
- Step 11: Making Predictions on Test Set

Exploratory Data Analysis

This is a pairplot of various numerical columns.

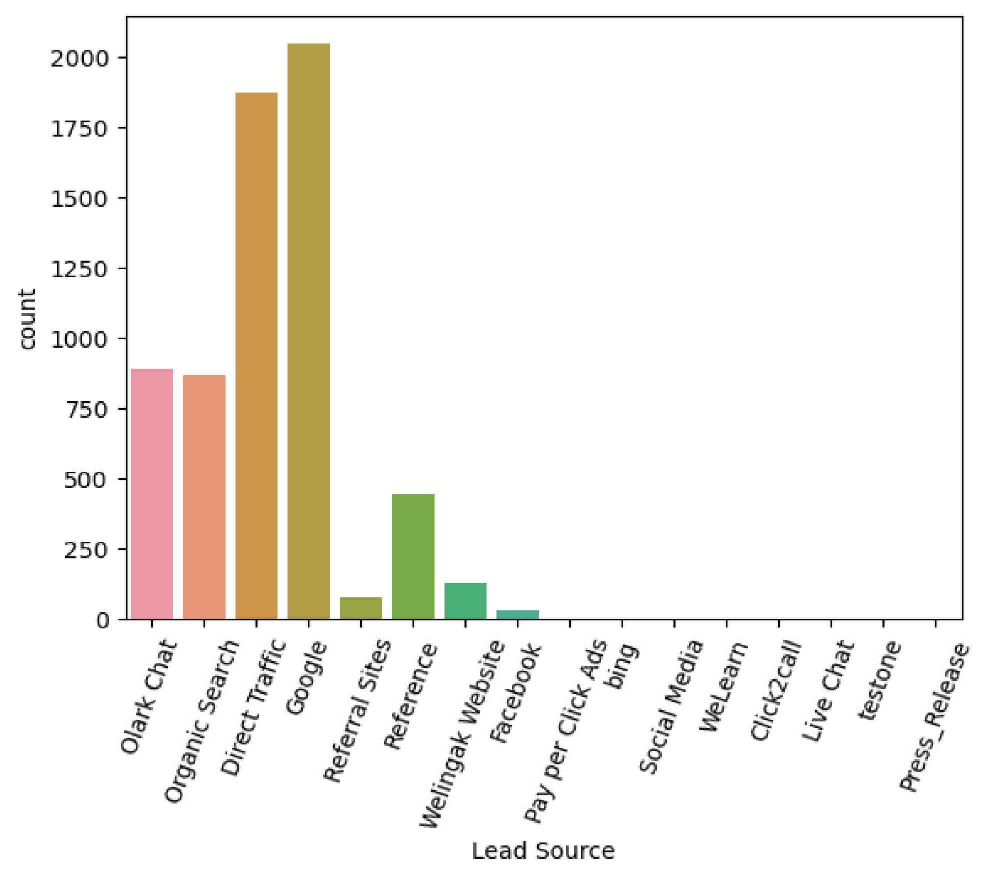


Count vs. Lead Origin



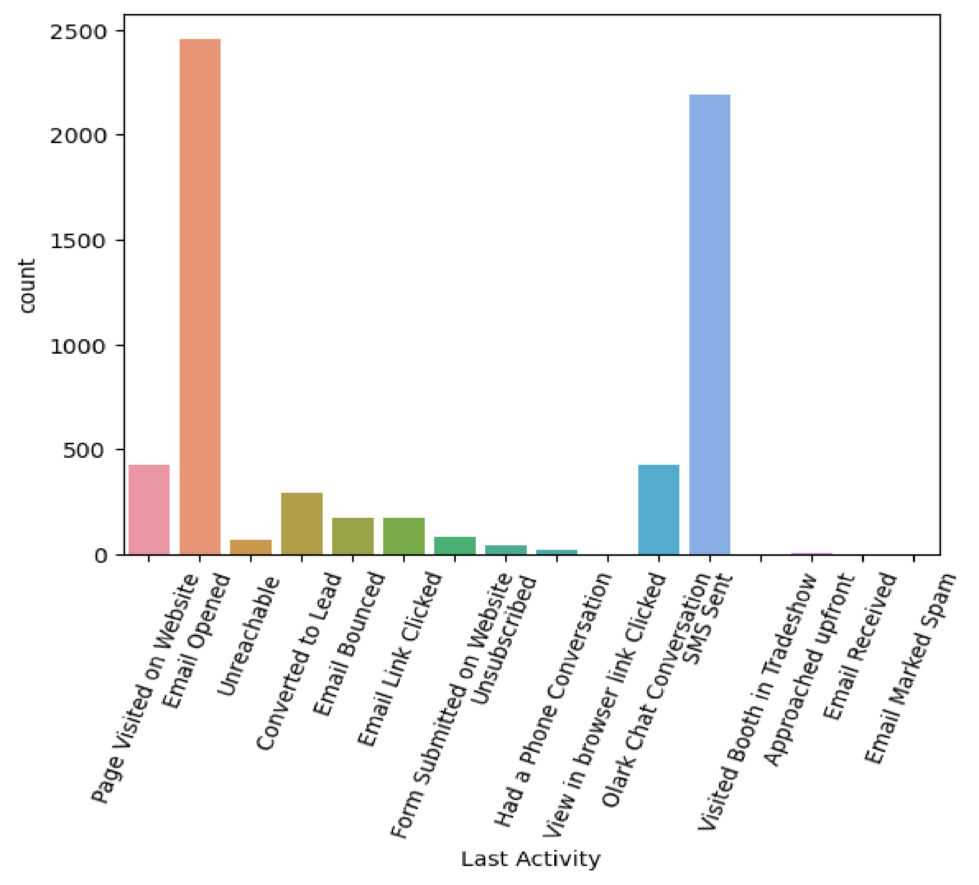
The landing page submission is more common.

Count vs. Lead Source



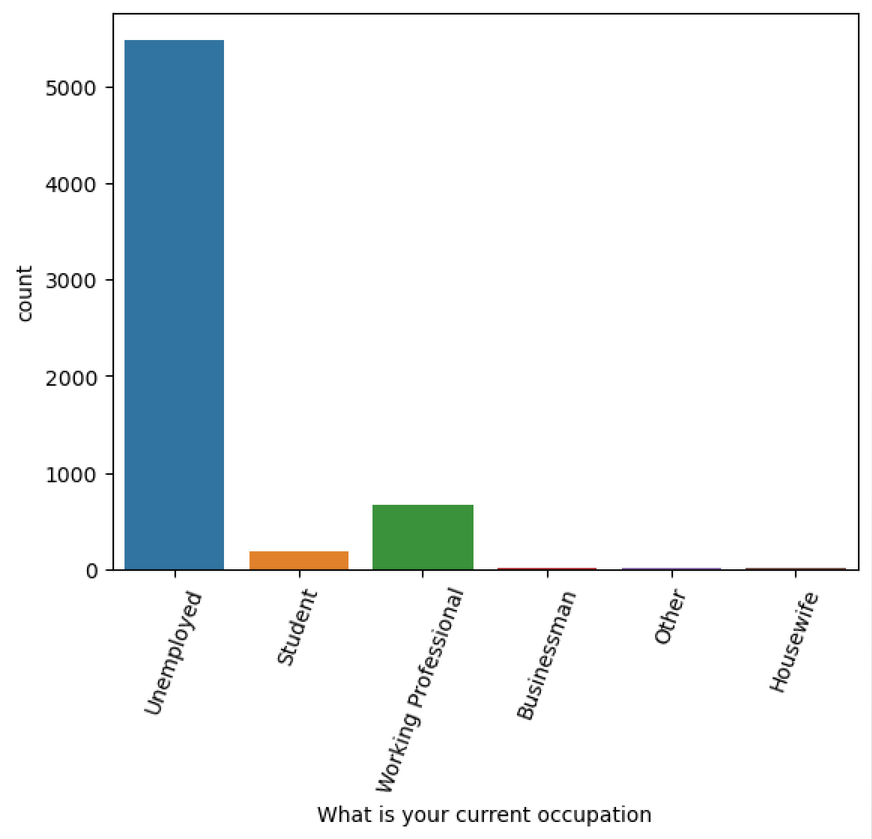
Google and Direct traffic seems to generate more leads.

Count vs. Last Activity



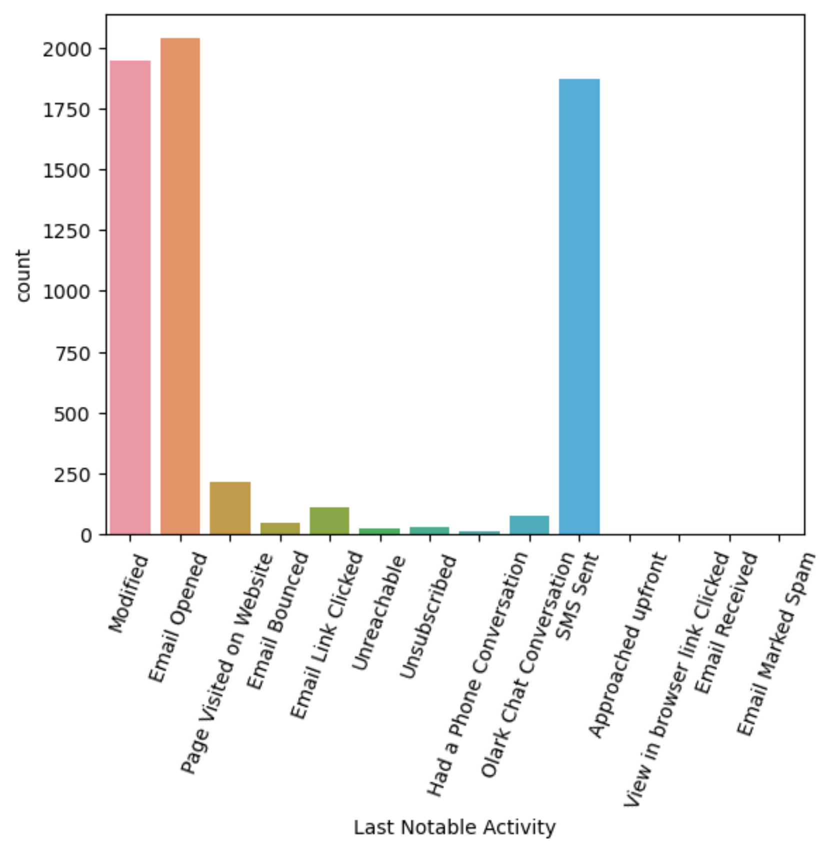
E-mail opened and SMS sent has more potential leads.

Count vs. Occupation



Unemployed people has more probability for becoming leads.

Count vs. Last Notable Activity



Modified, Email opened and SMS sent seems to have more count.

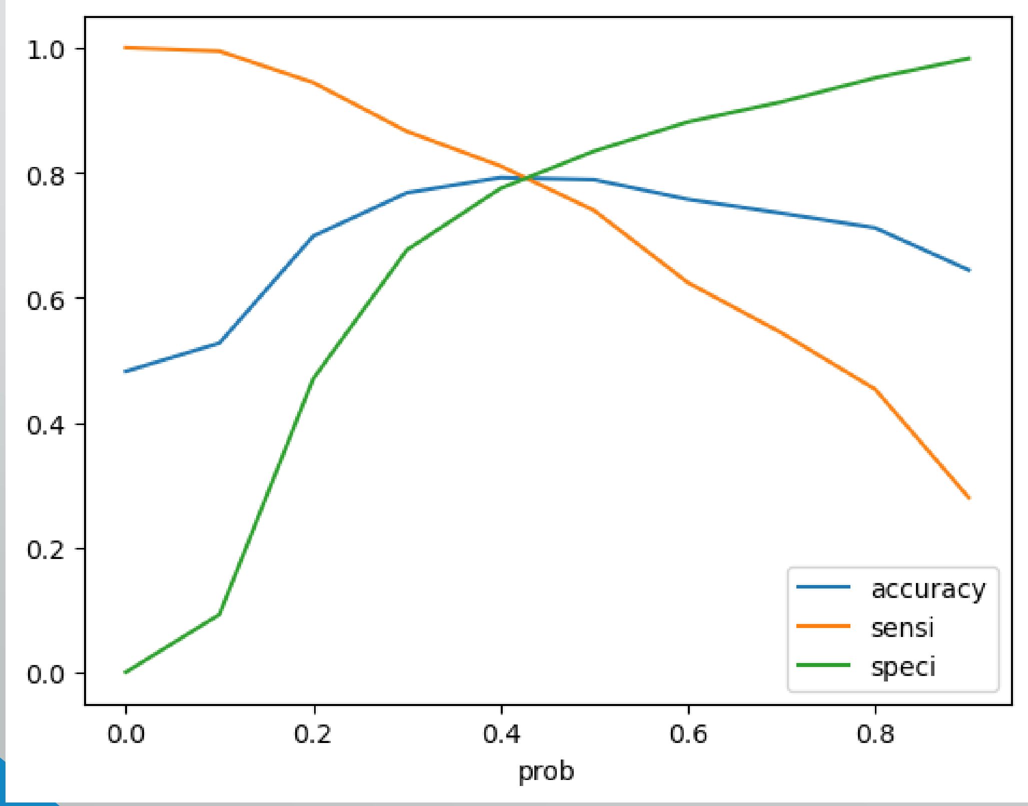
Variables Impacting the Conversion Rate

- Lead Origin_Lead Add Form
- Lead Source_Reference
- Lead Source_Welingak Website
- What is your current occupation_Unemployed
- Last Activity_Had a Phone Conversation
- Last Notable Activity_Had a Phone Conversation
- Total Time Spent on Website
- TotalVisits
- Last Activity_SMS Sent
- What is your current occupation_Working Professional
- Lead Source_Olark Chat
- Do Not Email
- What is your current occupation_Student
- What is your current occupation_Housewife

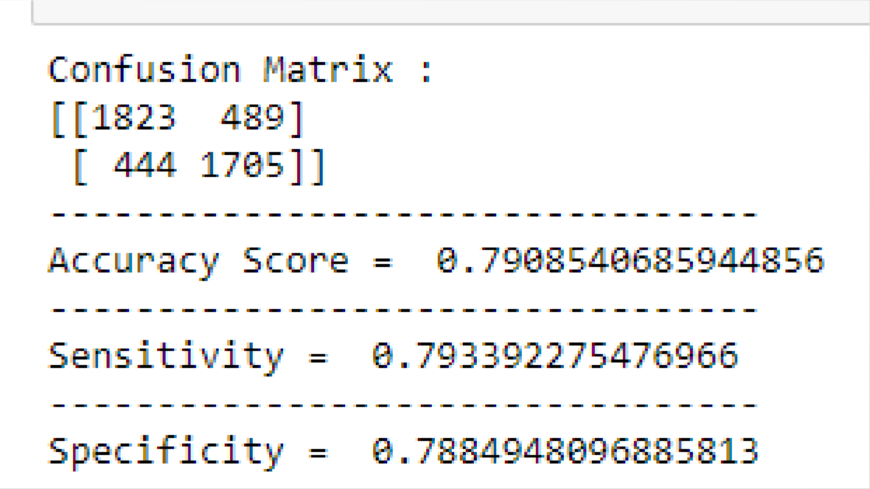
Last Notable Activity_Unreachable

There are 15 variables that has impact on the conversion rate.

Model Evaluation: Sensitivity and Specificity on Train Dataset

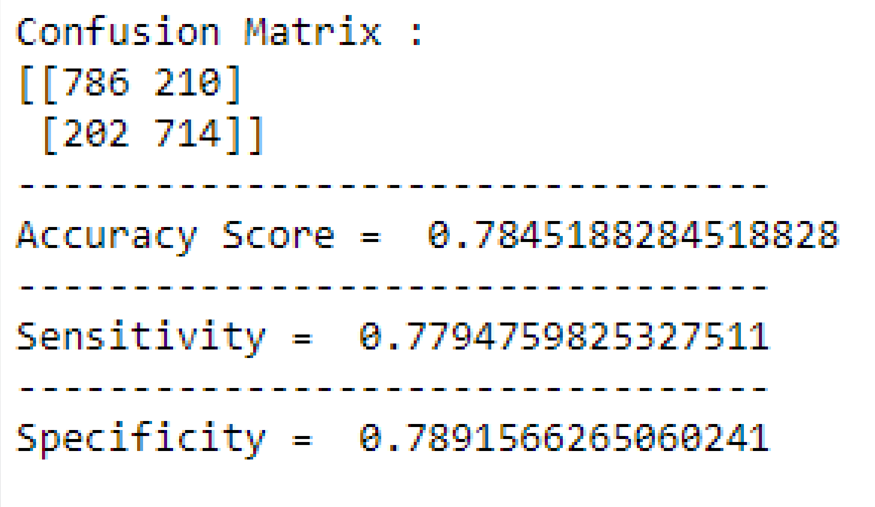


Here, the cut-off is 0.42.



```
Confusion Matrix :
[[1823  489]
 [ 444 1705]]
-----------------------------------
Accuracy Score =  0.7908540685944856
-----------------------------------
Sensitivity =  0.793392275476966
-----------------------------------
Specificity =  0.7884948096885813
```

Model Evaluation: Precision and Recall on Train Dataset



```
Confusion Matrix :
[[786 210]
 [202 714]]
---------------------------------
Accuracy Score =  0.7845188284518828
---------------------------------
Sensitivity =  0.7794759825327511
---------------------------------
Specificity =  0.7891566265060241
```

The cut-off is 0.42.

Conclusion

- The accuracy, sensitivity and specificity are 78.45%, 77.94% and 78.91% which is almost close to what we got in the train dataset.
- Also, the lead score calculated in the trained dataset shows the conversion rate of the predicted model to be around 80%.
- Hence, the model seems good.