LEAD SCORING CASE STUDY

BY ANUSHA KULKARNI & SATYAJEETA UGILE

Contents

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

BUSINESS OBJECTIVE

This case study aims us to identify and help X Education in selecting 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

DATA UNDERSTANDING

Leads.csv' file consists all the 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. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

READING AND INSPECTION

After reading the Leads.csv below are the findings:

- There are 37 columns and 9240 rows.
- There are select values for many columns. Because customer did not select any option form the list, hence it shows select.
- Select values are as good as NULL. Hence, converting ‘select’ values to NAN.

DATA CLEANING

1. Dealing with Null values more than 45 %:

- Number of columns having missing values more than 45% : 9 Columns
- After dropping 9 columns we are left with 28 columns

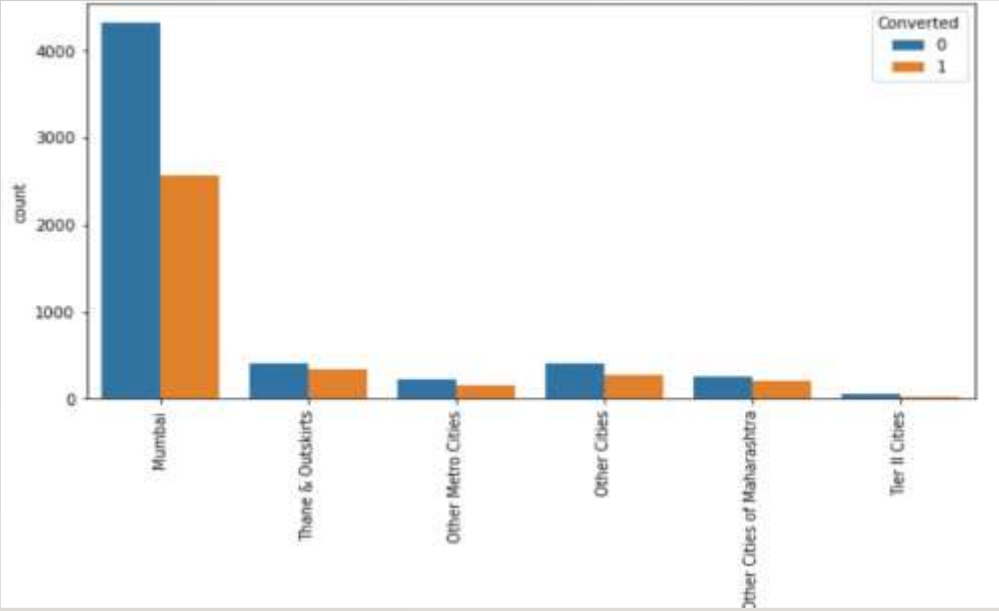
2. DEALING WITH NULL VALUES IN EACH COLUMN ONE BY ONE:

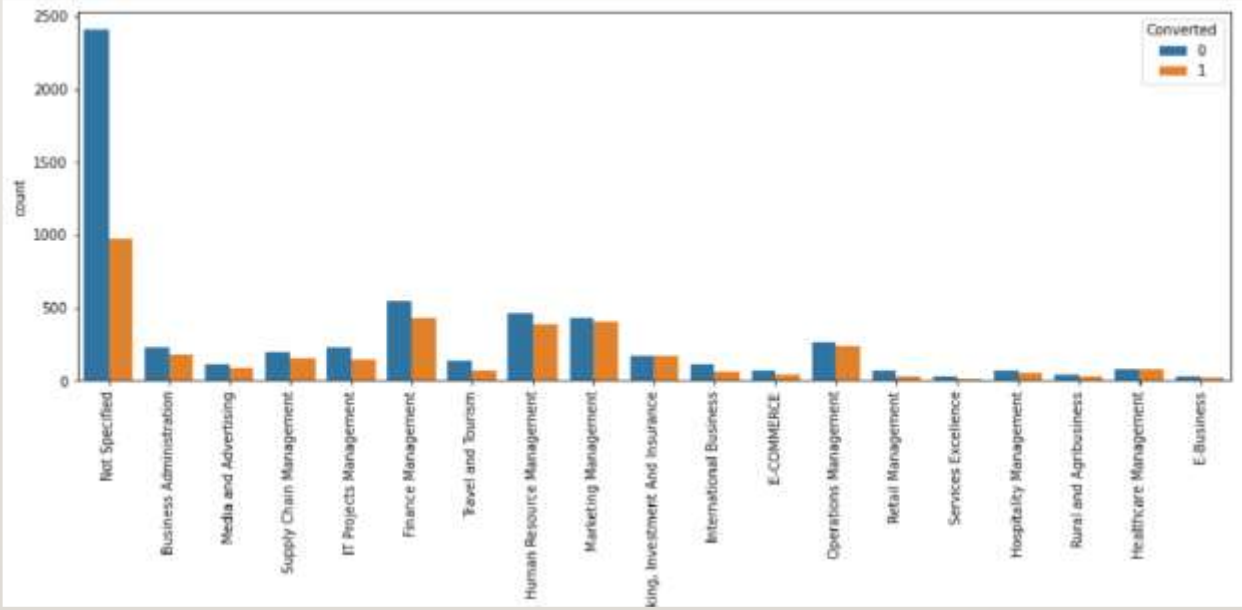
A. CITY:

1. AROUND 60% OF THE DATA IS MUMBAI. SO WE CAN IMPUTE MUMBAI IN THE MISSING VALUES.

B. SPECIALIZATION:

1. IT MAYBE THE CASE THAT LEAD HAS NOT ENTERED ANY SPECIALIZATION IF HIS/HER OPTION IS NOT AVAILABLE ON THE LIST, MAY NOT HAVE ANY SPECIALIZATION OR IS A STUDENT.
2. HENCE WE CAN MAKE A CATEGORY "OTHERS" FOR MISSING VALUES.





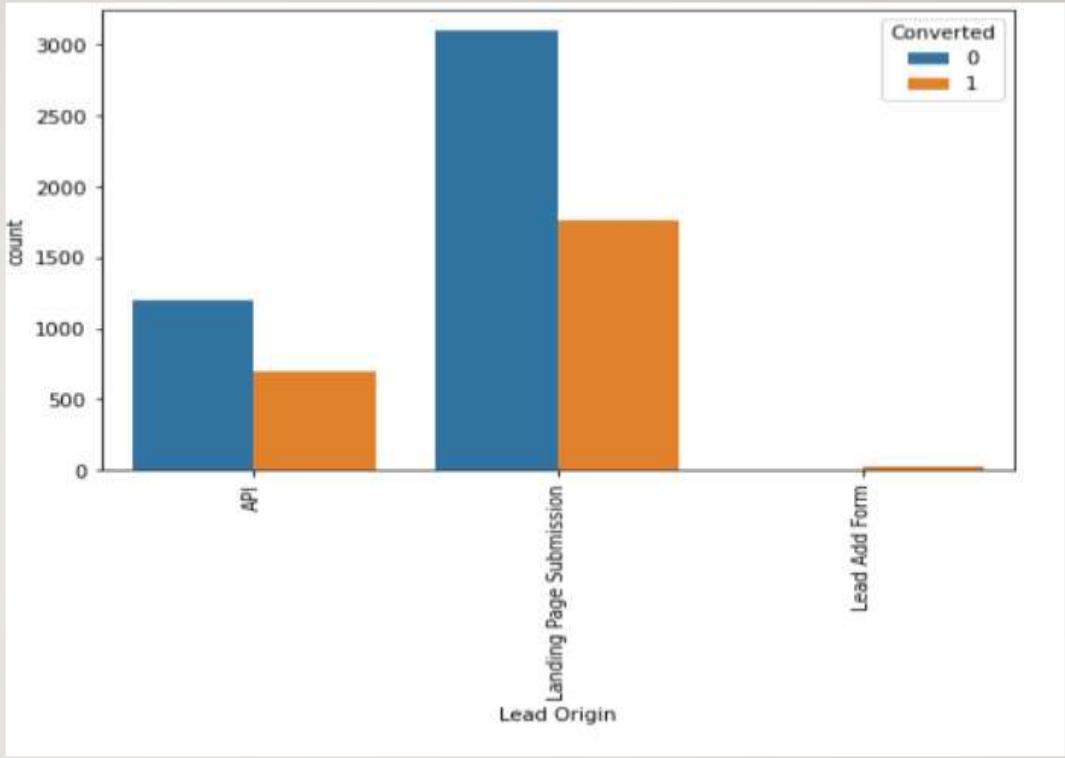
EXPLORATORY DATA ANALYTICS

Univariate Analysis

- Converted: Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0).
- Lead Origin

Insights:

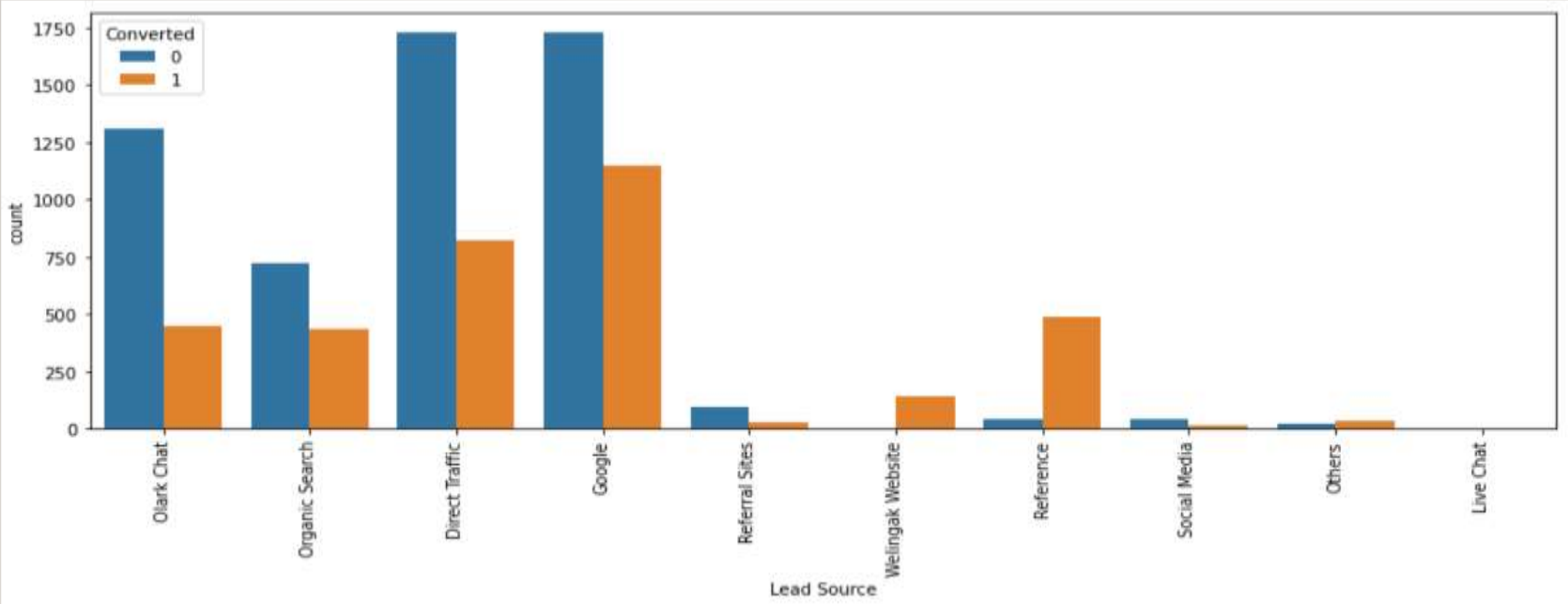
- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form



Lead Source:

Insights:

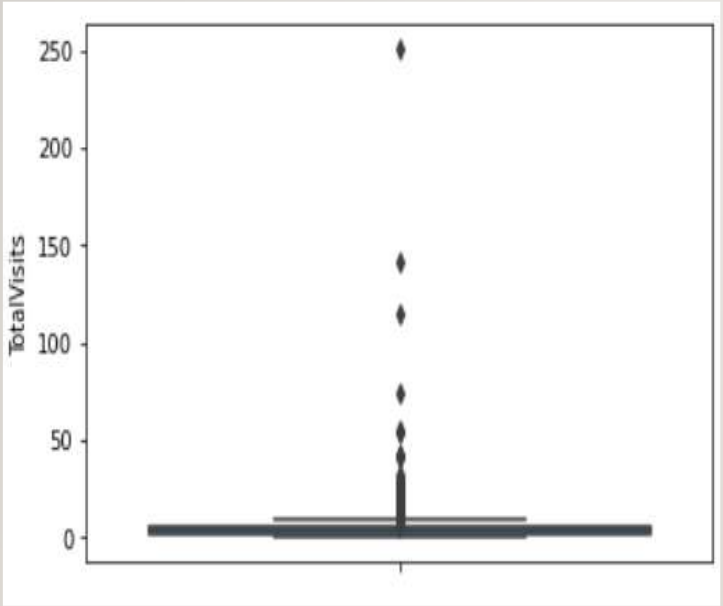
- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

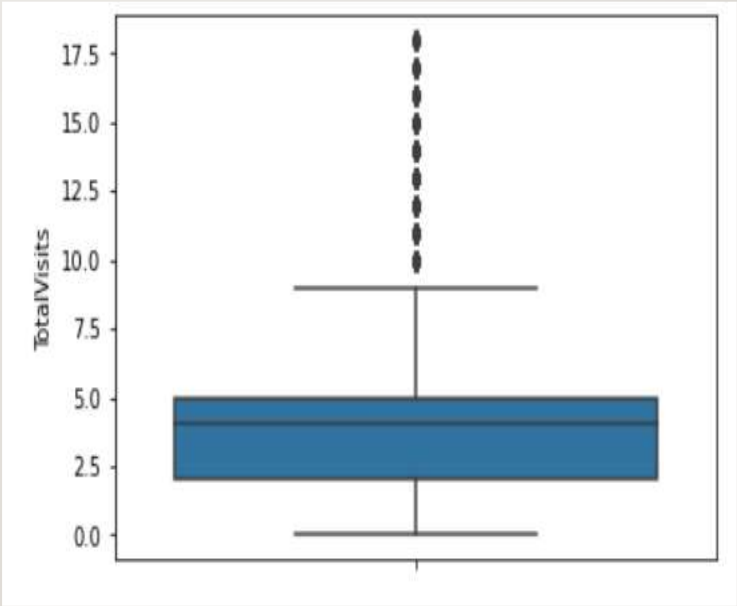


Total Visits:

Insights:

- There are outliers which needs to be dealt and sorted (Fig A before handling outliers)
- Nothing conclusive can be said on the basis of Total Visits.(Fig B after handling outliers)

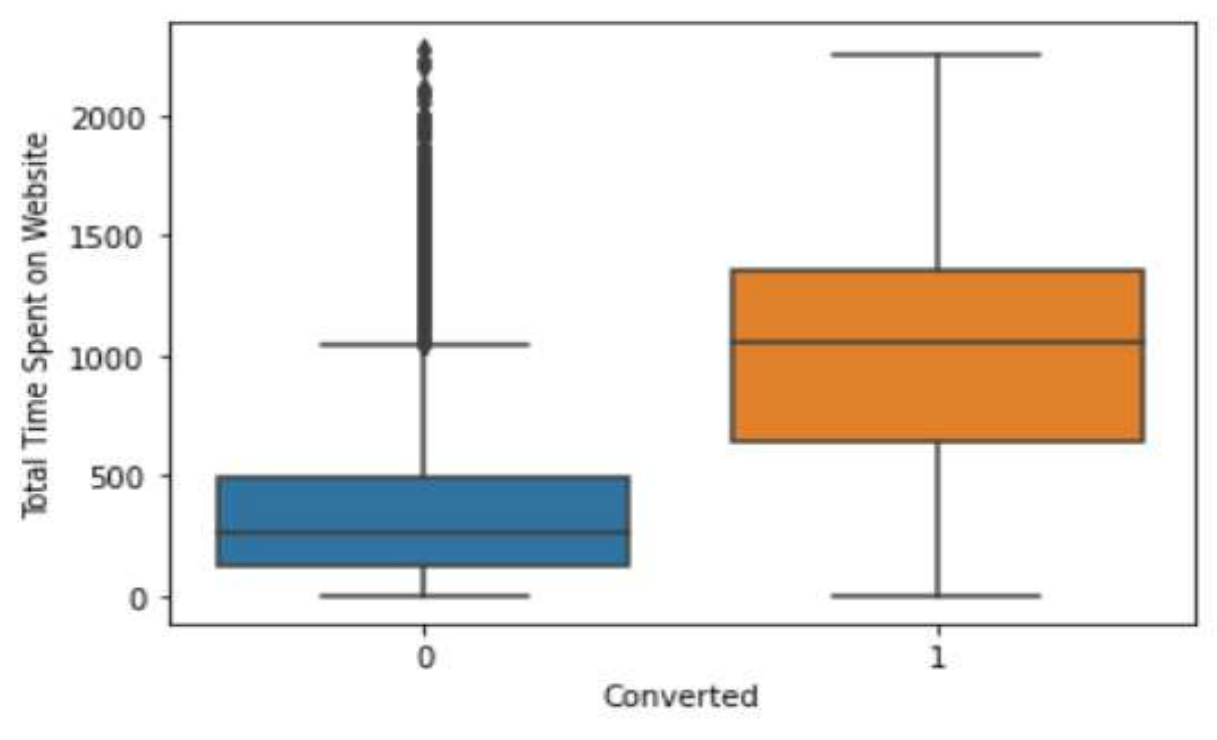




Total time spent on website:

Insights:

- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.

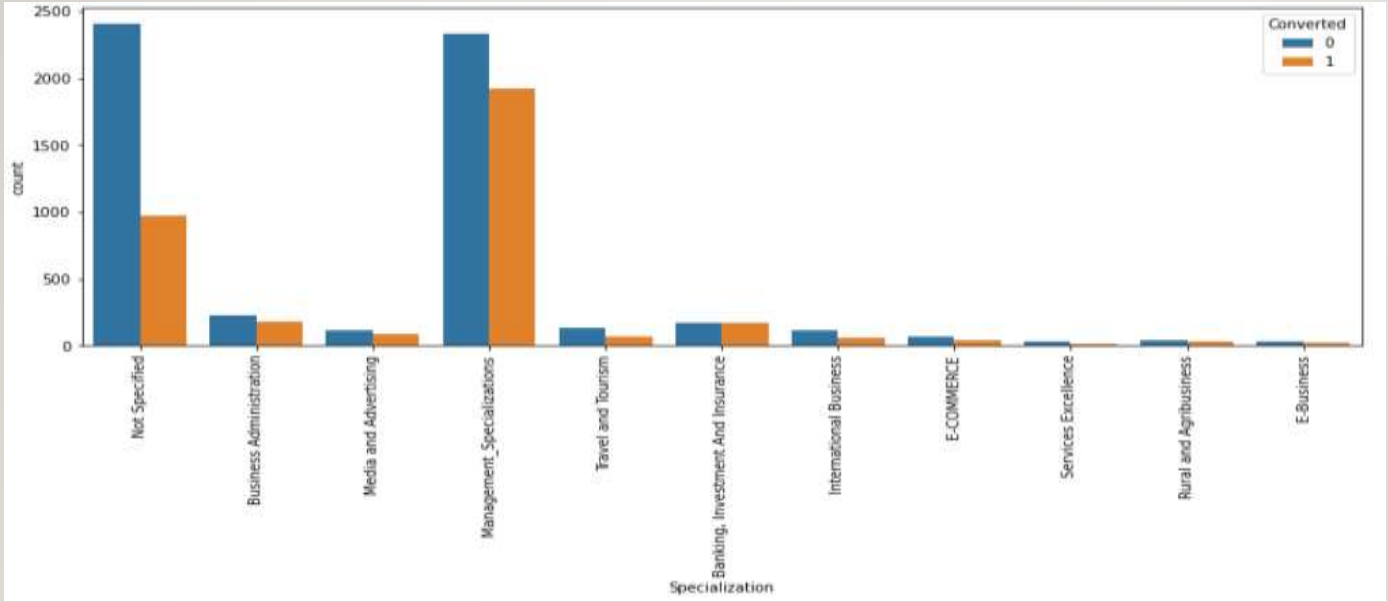


- Country:

Insights:
Most values are 'India' no such inference can be drawn

- Specialization:

Insights:
Focus should be more on the Specialization with high conversion rate.



Data Preparation

Data Preparation:

1. Converting some binary variables (Yes/No) to 1/0
2. For categorical variables with multiple levels, create dummy features (one-hot encoded):

Creating a dummy variable for some of the categorical variables and dropping the first one

Model Building

Running Your First Training Model

Model 1.

```
# Logistic regression model
X_train_sm = sm.add_constant(X_train[col])
logm1  = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm1.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4648
Model:	GLM	**Df Residuals:**	4632
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-873.41
Date:	Tue, 18 Oct 2022	**Deviance:**	1746.8
Time:	12:40:01	**Pearson chi2:**	6.21e+03
No. Iterations:	24		
Covariance Type:	nonrobust		

Model 2

```
X_train_sm = sm.add_constant(X_train[col])
logm2 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4648
Model:	GLM	**Df Residuals:**	4633
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-952.57
Date:	Tue, 18 Oct 2022	**Deviance:**	1905.1
Time:	12:48:32	**Pearson chi2:**	5.85e+03
No. Iterations:	24		
Covariance Type:	nonrobust		

Model 3

```
X_train_sm = sm.add_constant(X_train[col])
logm3 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm3.fit()
res.summary()
```

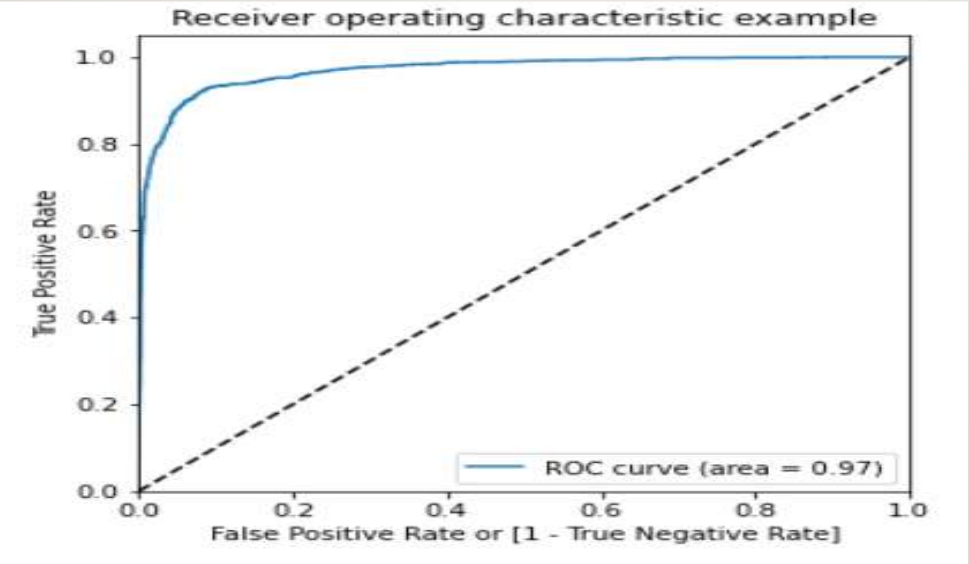
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4648
Model:	GLM	**Df Residuals:**	4633
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-952.57
Date:	Tue, 18 Oct 2022	**Deviance:**	1905.1
Time:	12:40:56	**Pearson chi2:**	5.85e+03
No. Iterations:	24		
Covariance Type:	nonrobust		

Plotting the ROC curve

Plotting the ROC Curve:

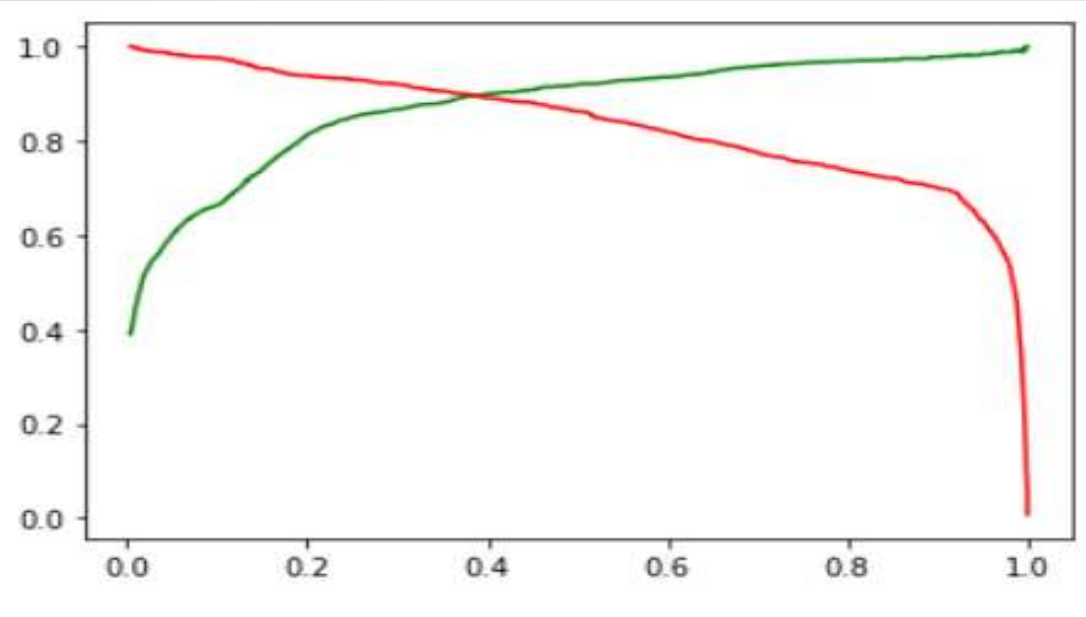
- An ROC curve demonstrates several things:
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Precision and Recall Curve

❖ Precision and Recall Curve:

```
plt.plot(thresholds, p[:-1], "g-")
plt.plot(thresholds, r[:-1], "r-")
plt.show()
```



Making Predictions on Test Set

Steps followed in making predictions

```
y_test_pred = res.predict(X_test_sm)
```

1. Converting y_pred to a dataframe which is an array

```
y_pred_1 = pd.DataFrame(y_test_pred)
```

2. Converting y_test to dataframe

```
y_test_df = pd.DataFrame(y_test)
```

3. Putting CustID to index

```
y_test_df['Prospect ID'] = y_test_df.index
```

4. Removing index for both dataframes to append them side by side

```
y_pred_1.reset_index(drop=True, inplace=True)
y_test_df.reset_index(drop=True, inplace=True)
```

5. Appending y_test_df and y_pred_1

```
y_pred_final = pd.concat([y_test_df, y_pred_1],axis=1)
y_pred_final.head()
```

6. Renaming the column

```
y_pred_final= y_pred_final.rename(columns={ 0 : 'Converted_prob'})
```

7. Rearranging the columns

```
y_pred_final = y_pred_final[['Prospect ID','Converted','Converted_prob']]
y_pred_final['Lead_Score'] = y_pred_final.Converted_prob.map( lambda x: round(x*100))
```

Final Predictions

Overall accuracy.
0.9195352839931153

sensitivity of our logistic regression model
0.918790752815649

specificity
0.919959473150962 5

precision_score
0.8673754896474538

recall_score
0.918790752815649

Conclusion

After running the model on the Test Data these are the figures we obtain:
Accuracy : 92.78%
Sensitivity : 91.98%
Specificity : 93.26%

Final Observation: Let us compare the values obtained for Train & Test: Train Data:
Accuracy : 92.15%
Sensitivity : 91.98%
Specificity : 91.99%

Test Data:
Accuracy : 92.78%
Sensitivity : 91.98%
Specificity : 93.26%

The Model seems to predict the Conversion Rate very well.

THANK YOU