LEAD SCORING CASE STUDY

Vikas Bhartiya Batch ds-43

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.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. 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 typical lead conversion rate at X education is around 30%.

PROBLEM STATEMENT CONTINUED….

- Now, although 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. 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

OBJECTIVE FOR PROBLEM

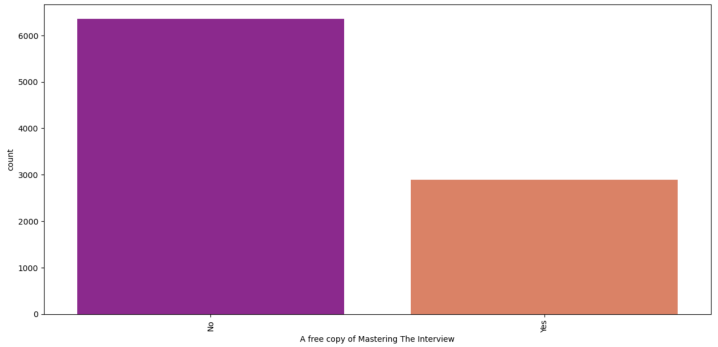
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have 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%.
- 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.

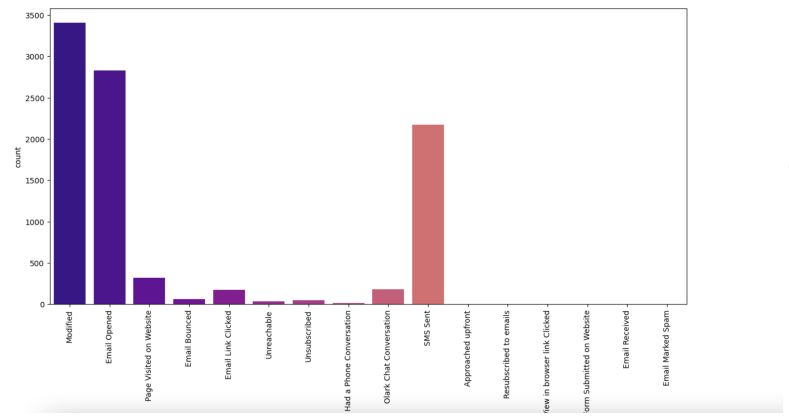
Missing Value

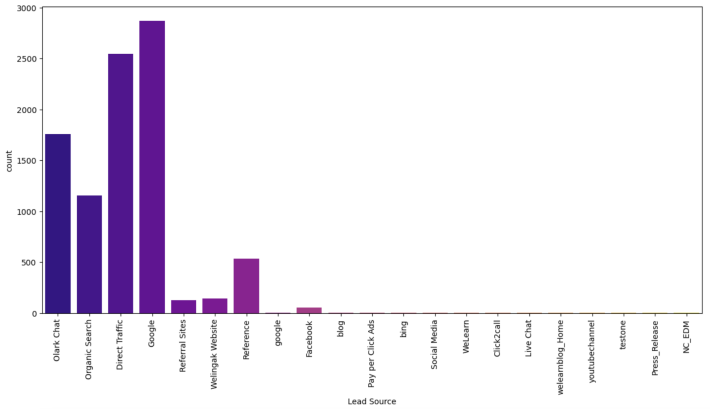
Apart from the null value in the column we have few value as “SELECT” which means that person has not provided that data. So that is also part of the missing value.

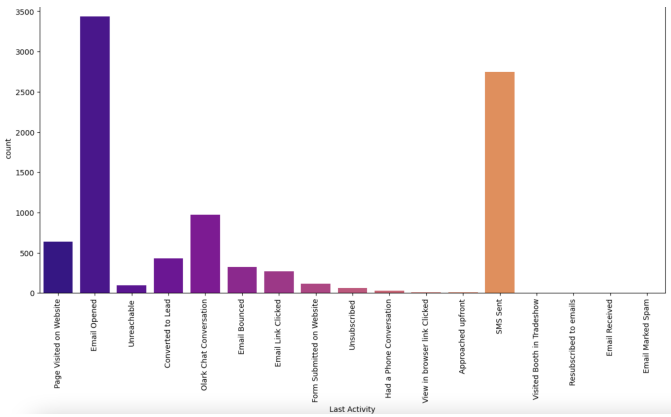
Lead Quality	51.590909
Asymmetrique Activity Index	45.649351
Asymmetrique Profile Score	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Index	45.649351
Tags	36.287879
Lead Profile	29.318182
What matters most to you in choosing a course	29.318182
What is your current occupation	29.112554
Country	26.634199
How did you hear about X Education	23.885281
Specialization	15.562771
City	15.367965
Page Views Per Visit	1.482684
TotalVisits	1.482684
Last Activity	1.114719
Lead Source	0.389610
Receive More Updates About Our Courses	0.000000

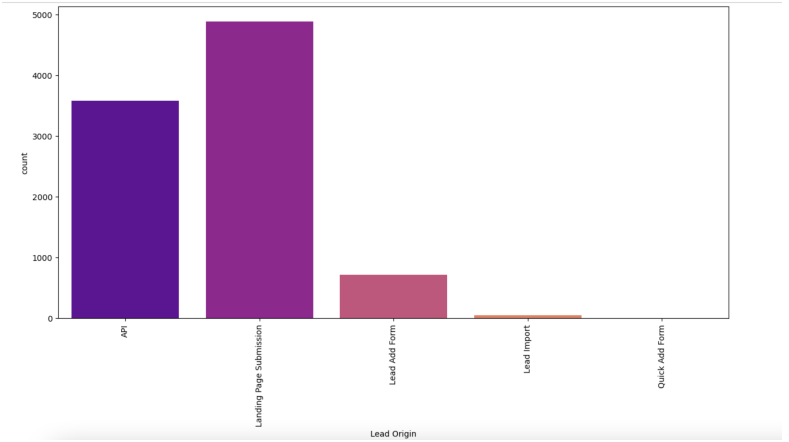
EDA-Categorical Column

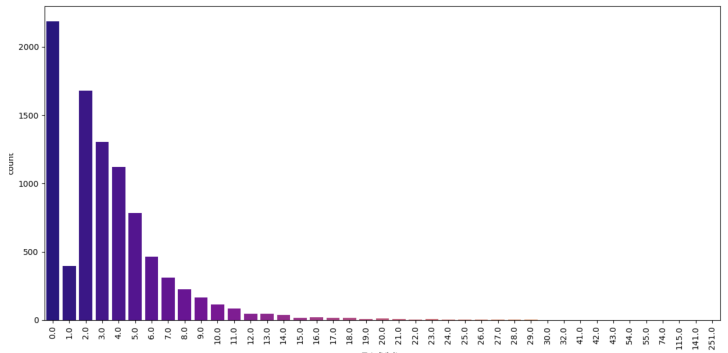












Data Preprocessing

Creating dummy variable

```
dummy1 = pd.get_dummies(df_cleaned[['Lead Origin', 'Lead Source', 'Last Activity', 'Last Notable Activity']], drop_first=True)
```

```
# Adding the results to the master dataframe
df_cleaned = pd.concat([df_cleaned, dummy1], axis=1)
```

Splitting the data set in train and test

[54]:
```
X=df_cleaned.drop('Converted',axis=1)
y=df_cleaned['Converted']
```

[55]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

[56]:
```
X_train.shape,X_test.shape,y_train.shape,y_test.shape
```

[56]:
```
((7249, 36), (1813, 36), (7249,), (1813,))
```

Applying standard scaler to the train dataset

[57]:
```
scaler = StandardScaler()
X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']] = scaler.fit_transform(X_train[['
```

Feature Selecting Using RFE

SELECTED TOP 15 FEATURE USING THE RFE METHOD

```
[66]: X_train.columns[~rfe.support_]
```

```
[66]: Index(['TotalVisits', 'Page Views Per Visit',
              'A free copy of Mastering The Interview',
              'Lead Origin_Landing Page Submission', 'Lead Source_Facebook',
              'Lead Source_Google', 'Lead Source_Organic Search',
              'Lead Source_Referral Sites', 'Lead Source_other_LeadSource',
              'Last Activity_Email Link Clicked',
              'Last Activity_Form Submitted on Website',
              'Last Activity_Page Visited on Website', 'Last Activity_Unreachable',
              'Last Activity_Unsubscribed',
              'Last Notable Activity_Email Link Clicked',
              'Last Notable Activity_Email Opened', 'Last Notable Activity_Modified',
              'Last Notable Activity_Olark Chat Conversation',
              'Last Notable Activity_Page Visited on Website',
              'Last Notable Activity_Unsubscribed',
              'Last Notable Activity_other_Last Notable Activity'],
             dtype='object')
```

MODEL SELECTION

FINAL MODEL SELECTED USING THE FEATURE HAVING P VALUE LESS THAN 0.05 AND VIF LESS THAN 5.
FINAL IMPORTANT FETURES ARE 13.

5]:

	Features	VIF
9	Last Activity_SMS Sent	4.98
11	Last Notable Activity_SMS Sent	4.89
2	Lead Source_Olark Chat	1.78
8	Last Activity_Olark Chat Conversation	1.40
0	Total Time Spent on Website	1.29
6	Last Activity_Email Opened	1.16
3	Lead Source_Reference	1.15
4	Lead Source_Welingak Website	1.05
1	Lead Origin_Lead Import	1.01
5	Last Activity_Email Bounced	1.01
7	Last Activity_Had a Phone Conversation	1.00
10	Last Activity_other_LastActivity	1.00
12	Last Notable Activity_Unreachable	1.00

z	P>|z|	[0.(
0	0.000	-2.
)1	0.000	1.(
6	0.001	0.5
4	0.000	1.(
51	0.000	3.7
4	0.000	4.4
6	0.000	-1.9
4	0.000	0.5
0	0.000	1.
4	0.000	-1.
6	0.000	0.5
8	0.001	0.8
9	0.000	1.(
4	0.000	1.6

EVALUATION OF THE MODEL

EVALUATION OF THE MODEL USING THE DEFAULT CUTOFF 0.05

ACCURACY >> 0.7943
SENSTIVITY >> 0.6625
SPECIFICITY >> 0.8747

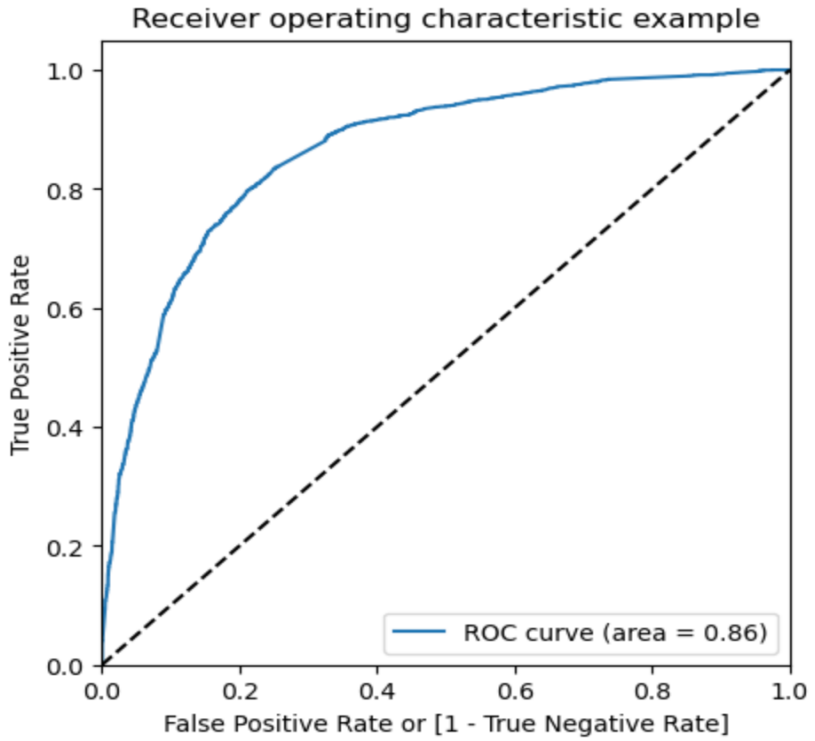
[82]:

	Converted	Conv_Prob	Conv_ID	Predicted
0	0	0.208428	433	0
1	1	0.862163	3132	1
2	1	0.279103	8475	0
3	0	0.078953	6068	0
4	0	0.032016	7581	0
...	...	...	...	...
7244	0	0.146849	5859	0
7245	0	0.831957	5306	1
7246	0	0.119233	5507	0
7247	0	0.185906	897	0
7248	1	0.466334	7424	0

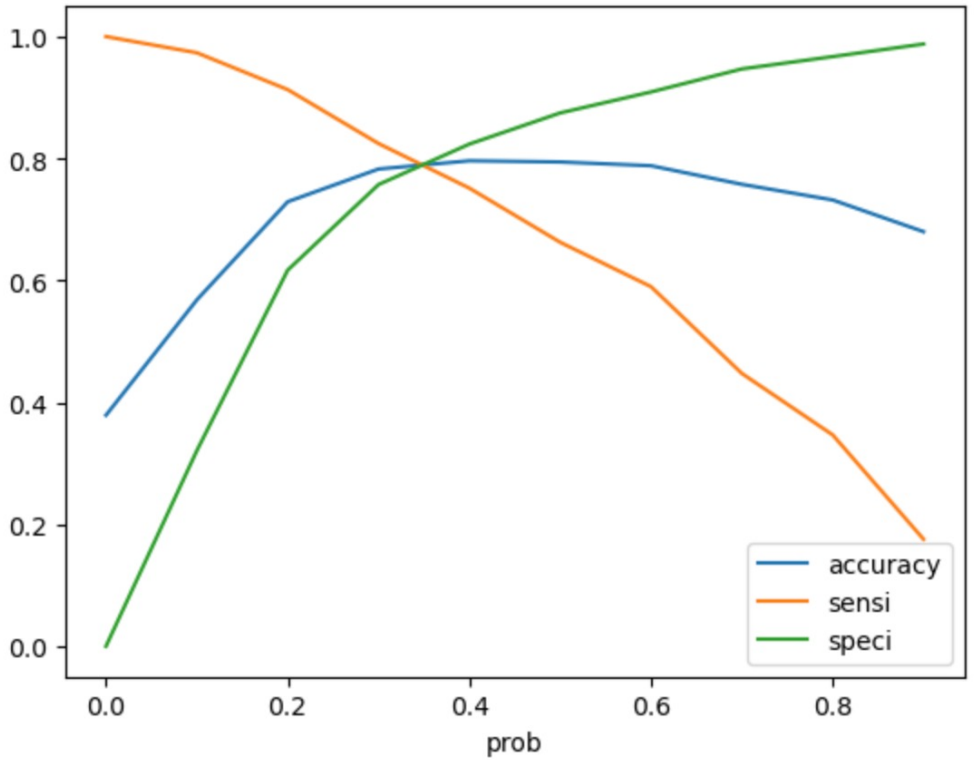
7249 rows × 4 columns

CHECKING THE CUT OFF PROBABILTY

ROC CURVE



ACCURACY,SENSTIVITY AND SPECIFICITY GRAPH



EVALUATION AFTER CUT POINT 0.35

ACCURACY -- 0.7914

SENSTIVITY– 0.7932

SPECIFICITY – 0.7903

FALSE POSITIVE RATE– 0.2096

POSITIVE PREDICTED – 0.6977

NEGATIVE PREDICTEVE – 0.8623

PRECESION – 0.6977

RECALL – 0.79322

EVALUATION OF THE TEST DATA SET

Accuracy test data

[118]:
```
# Let's check the overall accuracy.
metrics.accuracy_score(y_test, y_test_pred_1.Final_prediction)
```

[118]: 0.7837837837837838

Confusion Matrix

[119]:
```
conf_testset = metrics.confusion_matrix(y_test, y_test_pred_1.Final_prediction
conf_testset
```

[119]:
```
array([[890, 237],
       [155, 531]])
```

Sensitivity of test set

[121]:
```
# Sensitivity of the test set
TP / float(TP+FN)
```

[121]: 0.7740524781341108

Sepcificity of test set

[123]:
```
# Specificity for the test set
TN / float(TN+FP)
```

[123]: 0.7897071872227152

FINAL CONCLUSSION

- Final Model (res) res = logm4.fit()
- Can be converted to pickle file for further use
- Cut off probability 0.35
- Greater than 0.35 converted as lead
- Less than 0.35 will not converted as lead
- Accuracy of the train data 0.791
- Accuracy of the test data 0.784
- For increasing or decreasing number of the Lead Cut off prob can be adjusted
- We can also follow the lead score targeting from the top.

THANK YOU

PRESENTED BY- VIKAS BHARTIYA

EMAIL – vikas6050@gmail.com

BATCH NO. DS43