X Education company

Problem Statement:

X Education gets a lot of leads, its lead conversion rate is very poor. The current conversion rate is at 30%. To make this process more efficient, the company wishes to identify the most potential leads. If they successfully identify this set of leads, the lead conversion rate should go up. The company requires to build a model o assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Goal:

To increase the current lead conversion rate from 30% to 80%

Model:

Logistical Regression Case Study

Data Understanding

Data Description:

Rows – 9420
Columns – 37

Numeric Variables:

Total Visits
Total Time Spent on Website
Page Views Per Visit

Data Cleansing:

Dropping 9 variables due to missing value and distinct value observation

- How did you hear about X Education 7250
- What matters most to you in choosing a course 2709
- Receive More Updates About Our Courses 0
- Lead Quality 4767
- Lead Profile 6855
- Asymmetrique Activity Index 4218
- Asymmetrique Profile Index 4218
- Asymmetrique Activity Score 4218
- Asymmetrique Profile Score 4218

Percentage of Rows Retained

98.21%

EDA – Vital Variables Showcased

Univariate Analysis and Bivariate Analysis:

The lead conversion rate is 38%

Lead Origin:

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 Source:

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.

Total Visits

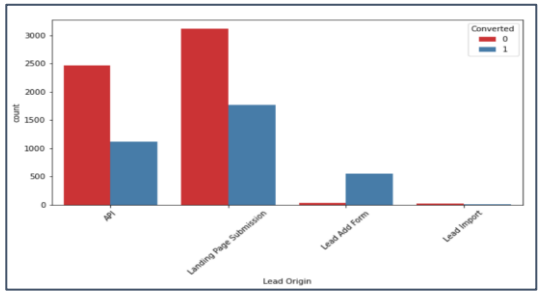
- Median for converted and not converted leads are the same.
- Nothing can be concluded on the basis of Total Visits.

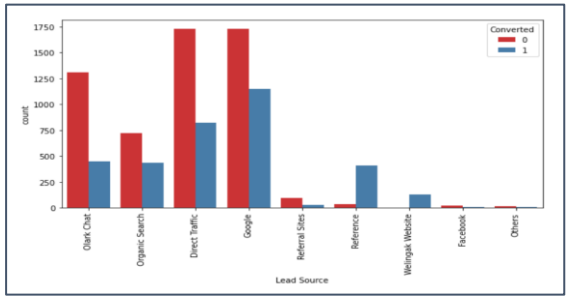
Total Time Spent on Website

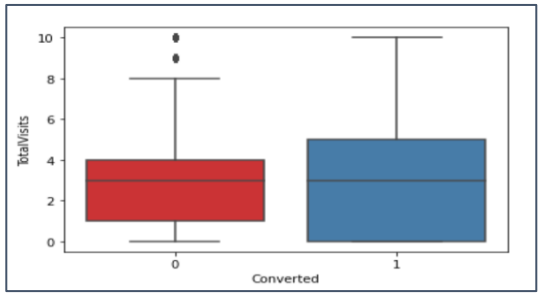
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.

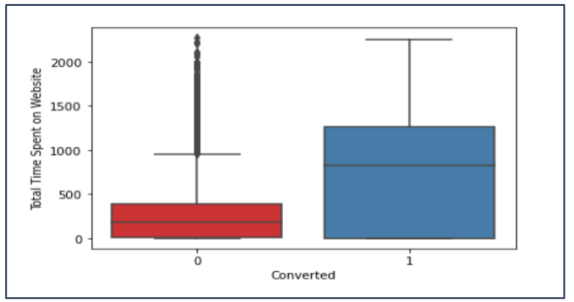
Last Activity

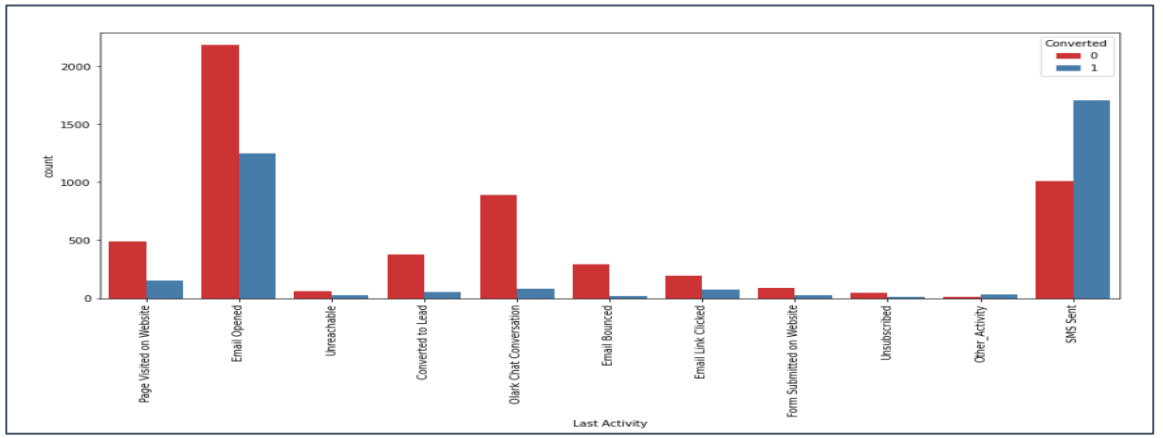
- Most of the lead have their email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost 60%.











Predictive Modeling

VIF Scores

	Features	VIF
1	Lead Origin_Lead Add Form	inf
2	Lead Origin_Lead Add Form	inf
3	Lead Source_Google	inf
4	Lead Source_Google	inf
5	Lead Origin_Lead Add Form	inf
6	Lead Origin_Lead Add Form	inf
7	Lead Source_Direct Traffic	inf
8	Lead Source_Direct Traffic	inf
9	Lead Source_Organic Search	inf
10	Lead Source_Organic Search	inf
11	Lead Source_Referral Sites	inf
12	Lead Source_Referral Sites	inf
14	What is your current occupation_Unemployed	2.57
15	What is your current occupation_Working Profes...	1.26
0	Total Time Spent on Website	1.11
13	What is your current occupation_Student	1.05

All variables have a good value of VIF. So we need not drop any more variables and we can proceed with making predictions using this model only

Optimal Cutoff Training

Finding Optimal Cutoff Point

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

```
# Let's create columns with different probability cutoffs
numbers = [float(x)/10 for x in range(10)]
for i in numbers:
    y_train_pred_final[i]= y_train_pred_final.Converted_prob.map(lambda x: 1 if x > i else 0)
y_train_pred_final.head()
```

	Converted	Converted_prob	Prospect ID	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.723862	7962	1	1	1	1	1	1	1	1	1	0	0
1	0	0.135471	5520	0	1	1	0	0	0	0	0	0	0	0
2	0	0.215136	1962	0	1	1	1	0	0	0	0	0	0	0
3	1	0.966340	1566	1	1	1	1	1	1	1	1	1	1	1
4	0	0.297109	9170	0	1	1	1	0	0	0	0	0	0	0

Predictive Modeling

Inference Train Data

```
y_train_pred_final['Lead_Score'] = y_train_pred_final.Converted_prob.map( lambda x: round(x*100))

y_train_pred_final[['Converted','Converted_prob','Prospect ID','final_Predicted','Lead_Score']].head()
```

	Converted	Converted_prob	Prospect ID	final_Predicted	Lead_Score
0	0	0.723862	7962	1	72
1	0	0.135471	5520	0	14
2	0	0.215136	1962	0	22
3	1	0.966340	1566	1	97
4	0	0.297109	9170	0	30

```
# checking if 80% cases are correctly predicted based on the converted column.

# get the total of final predicted conversion / non conversion counts from the actual converted rates

checking_df = y_train_pred_final.loc[y_train_pred_final['Converted']==1,['Converted','final_Predicted']]
checking_df['final_Predicted'].value_counts()
```

```
1    1805
0     614
Name: final_Predicted, dtype: int64
```

```
# check the precentage of final_predicted conversions

2005/float(2005+414)
```

```
0.8288548987184787
```

Hence, we can see that the final prediction of conversions have a target of 83% conversion as per the X Educations CEO's requirement . Hence, we can say that this is a good model.

Inference Test Data

```
y_pred_final['final_Predicted'] = y_pred_final.Converted_prob.map(lambda x: 1 if x > 0.3 else 0)
```

```
y_pred_final.head()
```

	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	3504	0	0.286339	29	0
1	4050	1	0.914510	91	1
2	7201	0	0.368443	37	1
3	1196	0	0.285925	29	0
4	8219	1	0.185175	19	0

```
# checking if 80% cases are correctly predicted based on the converted column.

# get the total of final predicted conversion or non conversion counts from the actual converted rates

checking_test_df = y_pred_final.loc[y_pred_final['Converted']==1,['Converted','final_Predicted']]
checking_test_df['final_Predicted'].value_counts()
```

```
1    774
0    268
Name: final_Predicted, dtype: int64
```

```
# check the precentage of final_predicted conversions on test data

865/float(865+177)
```

```
0.8301343570057581
```

Hence we can see that the final prediction of conversions have a target rate of 83% (same as predictions made on training data set)

Predictive Modeling

Precision & Recall Metrics Train Data

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_Predicted)
```

```
0.7862523540489642
```

```
confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_Predicted )
confusion2
```

```
array([[3205,  748],
       [ 614, 1805]], dtype=int64)
```

```
TP = confusion2[1,1] # true positive
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.7461761058288549
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8107766253478371
```

Inference: So as we can see above the model seems to be performing well. The ROC curve has a value of 0.86, which is very good. We have the following values for the Train Data:

- Accuracy : 78.62%
- Sensitivity :74.61%
- Specificity : 81.07%

Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value,Negative Predictive Values, Precision & Recall.

Precision & Recall Metrics Test Data

Precision and Recall metrics for the test set

```
precision_score(y_pred_final.Converted , y_pred_final.final_Predicted)
```

```
0.7133640552995392
```

```
recall_score(y_pred_final.Converted, y_pred_final.final_Predicted)
```

```
0.7428023032629558
```

Inference: After running the model on the Test Data these are the figures we obtain:

- Accuracy : 78.79%
- Sensitivity :74.28%
- Specificity : 81.15%

Conclusion:

- As we have checked both the Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 78%, 74% and 81% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Total Time Spent on Website

Our Models are working as per the Goal indicated by the CEO