Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. **Cleaning data:**
   The data was partially clean except for a few null values and the option 'select' had to be replaced with a null value since it did not give us much information.
   Many features have some categories which are very margin, so they are clubbed in a group to see the clear data representation.

2. **Exploratory Data Analysis:**
   A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

3. **Dummy Variables Ceartion:**
   The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the StandardScaler.

4. **Train-Test split:**
   The split was done at 70% and 30% for train and test data respectively.

5. **Model Building (Logistic Regression):**
   Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. **Model Evaluation:**
   A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came as below

PREDICTED / ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	3573	392
CONVERTED	296	2181

ATTRIBUTES	PERCENTAGES
ACCURACY	89.32%
PRECISION	84.76%
SENSITIVITY	88.05%
SPECIFICITY	90.11%

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum cut off as 0.28 with accuracy, sensitivity and specificity as below-

PREDICTED / ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	1540	167
CONVERTED	131	924

ATTRIBUTES	PERCENTAGES
ACCURACY	89.21%
PRECISION	84.69%
SENSITIVITY	87.58%
SPECIFICITY	90.21%

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
4. When the last activity was:
    a. SMS
    b. Olark chat conversation

Conclusion

```
1) Optimum cut off is 0.27 (as per both ROC curve &  Precision Recall curve )

2) Our final model have 12 features -
    const                                          -1.316424
    Lead Source_Welingak Website                    4.968015
    Last Activity_SMS Sent                          1.913047
    Tags_Busy                                       4.422293
    Tags_Closed by Horizzon                         9.905369
    Tags_Lost to EINS                               9.816265
    Tags_Ringing                                   -1.493942
    Tags_Will revert after reading the email        4.042608
    Tags_switched off                              -2.207254
    Lead Quality_Not Sure                          -3.736397
    Lead Quality_Worst                             -3.581532
    Last Notable Activity_Modified                 -1.667835
    Last Notable Activity_Olark Chat Conversation  -1.698337

3) Top 3 features in our model
    1) Tags_Closed by Horizzon      = 9.905369
    2) Tags_Lost to EINS            = 9.816265
    3) Lead Source_Welingak Website = 4.968015

4) Our Model below attributes-
    1) Accuracy on test set   = 89.2%
    2) Sensitivity on test set= 87.5%
    3) Specitivity on test set= 90.2%
    4) Precision on test set   = 84.6%
```