SUMMARY

The analysis is performed for X Education and to find ways to get more industry professionals to join the courses. The dataset provided gave us a lot of information about how the potentials customers visit the site, the time they spend over there, then how they reached the site and the conversion rate.

The following steps are used in building the Logistic Regression

1. **Data Cleaning:**
    - Check for the duplicates to remove the redundant observations. However, we didn't find any duplicates.
    - The data set is partially clean except for a few null values and the option "select" has to replace with a null value since it didn't provide valuable information.
    - Dropped the features which are having high percentage of Null values, i.e., greater than 40 percent.
    - Checked for number of unique categories for all categorical columns.
    - Treated the missing values by imputing the favorable aggregate functions like Median and Mode.

2. **Exploratory Data Analysis**
    - Processed with further analysis to check the conditions of data. It was encountered that a lot of elements in the categorical variables are irrelevant. The numeric values are seeming good but found the outliers.
    - Performed Univariate analysis and Bivariate analysis for both continuous and categorical features.

3. **Dummy Variables**
    - The dummy variables are created for all the Categorical columns.

4. **Scaling**
    - Used Standard scalar to scale the data for continuous variables.

5. **Train-Test Split**
    - The split was done on 70% and 30% for train and test the data respectively.

6. **Model building**
    - By using RFE with Cross Validation score, figured out the optimal number of features. With the optimal features, we built a GLM model by using the stats method. Later the irrelevant features are removed manually depending on the P-Value and VIF values. The features with VIF < 5 and P-Value < 0.05 are considered in the final model.

7. Model Evaluation

- A confusion Matrix was prepared and checked the accuracy, sensitivity and specificity by using the ROC curve optimal point. We had achieved the blow scores for Training data set.

a.	Accuracy	: 84%
b.	Sensitivity	: 83%
c.	Specificity	: 86%

8. Prediction

- Prediction was done on the test data an optimum cut-off as 0.38. We had achieved the blow scores for Test data set.

a.	Accuracy	: 84%
b.	Sensitivity	: 84%
c.	Specificity	: 85%

9. Precision-Recall

- The method was also used to recheck and a cut-off of 0.42

10. Conclusion

- Accuracy, Sensitivity and Specificity values of both train and test sets are around 85 percent
- We have noted that the features that important the most in the potential buyers are:
  - The total time spend on the Website.
  - When the lead origin is Lead add Form
  - When the lead source is Olark Chat
  - When the lead activity is Email Opened
  - When the last notable activity is SMS sent