**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. **Importing and Reading the data**: All the required libraries were imported and the data was pulled from the csv file. All the sanity checks were performed to understand the data.
2. **EDA:** A quick EDA was done to check the condition of our data. It was helpful for understanding the categorical variables. The numeric values seem good.
3. **Cleaning data:** The data was cleaned for null values and the option select had to be replaced with a null value since it did not give us much information. In our analysis, few missing values were imputed with null values and few were removed since they had more than 30% missing values. There were few columns which were biased to one part for up to 90%. So those columns were removed.
4. **Dummy Variables:** The dummy variables were created and the original elements were removed. For numeric values we used the Standard Scaler.
5. **Train-Test split:** The split was done at 70% and 30% for train and test data, respectively.
6. **Correlation:** The correlation of the train data is performed to check the multicollinearity.
7. **Model Building:** Firstly, RFE was done to attain the top 20 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).
8. **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. Precision and Recall are also calculated which were around 72% and 80% respectively.
9. **Prediction Analysis:** Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity which was around 81%. Also, Precision and Recall are also calculated for test data which were around 72% and 81% respectively.
10. **Conclusion:**

It was found that the variables that mattered the most with top 3 highest co – efficient values are

1. Lead Origin\_Lead Add Form,
2. Last Notable Activity\_Had a Phone Conversation and
3. Occupation\_Working Professional

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.