## Summary

X Education tries 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. Data Cleaning:

Initially duplicate values are handled. Then NA values and their percentages are taken into consideration. There are several values called 'Select' which are formed due to customers not filling the required fields. There are considered as NA and was allotted to relevant categories in each column. Columns with higher percentage of missing values are dropped and outliers are handled.

### 2. Exploratory Data Analysis:

EDA is done to check the relatability of the data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seemed relevant and no outliers were found.

### 3. Dummy Variables:

The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values MinMaxScaler is used.

### 4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

### 5. Model Building:

Recursive feature elimination is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated

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 formed. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

## 7. Accuracy

The overall accuracy was found to be 89%

## 8. ROC

ROC shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

Optimal cut off probability is that probability where we get balanced sensitivity and specificity. Optimal cut off was found to be 0.35.

## Conclusion:

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

- The total time spend on the Website.
- Total number of visits.
- When the lead source was Google, Direct traffic, Organic search
- When the last activity was SMS, Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

X education has to focus on the above factors to convert the website visitors into potential students of the course.