Summary Report (Group: Ramakrishna and Karan)

The entire Lead scoring case study was carried out in partnership with my peer Karan, Sowjanya who was one of the other partners did not participate as she is currently on deferral. The 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 steps were used:

1. **Data Cleaning and Preparation**:
   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. We could see that there were a lot of columns which had high number of missing values. Clearly, these columns were not useful. Since, there were 9000 datapoints in our dataframe we eliminated the columns having greater than 3000 missing values as they were of no use to us. There are a few columns in which there was a level called '**Select**' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we identified the value counts of the level 'Select' in all the columns that it is present. Then we still have around 69% of the rows which seemed good enough.

2. **EDA**:
   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.

3. **Dummy Variables:**
   The dummy variables were created and later on the variables for which the dummy variables have been created were dropped.

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

5. **Scaling**: For numeric variables we used the MinMaxScaler.

6. **Model Building**:
   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-values (The variables with VIF <5 and p-Value <0.05 were kept). As there were a lot of variables present in the dataset, hence selected a small set of features from this pool of variables using RFE, we did feature selection using RFE. The area under the curve of the ROC is 0.86 which is quite good. The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. The optimal values of the three metrics (accuracy, sensitivity and specificity) were at 0.42.

7. **Making Predictions on the data set**

8. **Making predictions on the test set**

9. **Learnings:**
   The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost accurate. We had high recall score than precision score which is a sign of good model. In business terms, this model has an ability to adjust with the company's requirements in coming future. This concludes that the model is in stable state.