SUMMARY

(Lead scoring case study-Group study)

The data provided to us gave a lot of information, about how the customers visit site, how they reach there, how they spend time there and the conversion rate of customers

The following steps are used:

1) **Data Cleaning/ Data Preprocessing:**
   - There are few null values, except those values data is partially clean
   - Select option has been replaced with NaN values, means no any option is selected, because it does not give us much information.
   - Columns with only one unique value have been dropped because it will not affect our analysis; also the variables having null values more than 35% are also dropped.
   - We also dropped ‘country’ and ‘city’ columns, because we think they does not affect much on analysis
   - We don’t want to lose much data so the NaN values of few variables are replaced by ‘not provided’.

2) **Data Visualization (EDA)**
   - It includes categorical variables analysis and numerical variable analysis
   - Few of the categorical variables needs to be dropped because there is not much of relevant information in those
   - Few of the imbalanced variables are also need to be dropped
   - Data of the numerical variables seems good, there are some outliers in few variables
   - Outlier treatment had been done by caping the outliers to 95% value

3) **Data Preparation**
   - Dummy variables were created
   - Later on, dummies with ‘not provided’, ‘Others’ elements were removed

4) **Test-Train split and Feature Scaling**
   - The split were done as: 70% for train data and 30% for test data
   - For numerical features we used StandardScaler

5) **Model Building**

- First running the Recursive feature Elimination(RFE) with 15 variables as output
- Later, some of the variables were removed based on their p-values and Variance Inflation Factor (VIF) value. (The variables with VIF <5 and p-value <0.05 were kept only)

6) **Model Evaluation and Predictions**

- Initially, the confusion matrix was made
- Later, various probability cut-off values were used to find out 'Accuracy', 'Sensitivity' and 'Specificity'
- 'Accuracy', 'Sensitivity' and 'Specificity' values are came around 80%
- For predictions first scaling of test data set was done
- After that, we choose arbitrary cut-off value as 0.5
- Using arbitrary cut-off value we found out best cut-off equal to 0.34
- So, 0.34 is the optimum point to take it as a cut-off probability
- With best cut-off probability we found 'Accuracy', 'Sensitivity' and 'Specificity'

7) **Precision and Recall**

- Precision recall curve used to recheck cut-off value, which was found to be ~0.4, with precision 72% and recall 80%

List of the variables in descending order which are mattered most in the potential buyers to buy courses:

1) Total Time Spent on Website
2) Lead origin as lead add form
3) When the current occupation as working professional
4) When the lead source was-
   a) Google
   b) Direct traffic
   c) Organic search
   d) Referral sites
   e) Welingak website
5) When the last activity was-
   a) Converted to lead
   b) Olark chat conversation

Keeping these all variables in mind X Education can progress, as they have high chance to get almost all the potential buyers to change their mind and buy the courses