SUMMARY

An analysis has been performed for the X Education company to infer its potential leads and to find ways to get more industry professionals to join their courses. The data that was provided gave adequate information such as the number of visits to the site, the time spent there, how the users reached the site and the conversion rate.

To perform the analysis, the following methods were used:

- **Importing Libraries** - All relevant libraries were imported in the python file in order to load data for numerical computation, ignoring warnings, visualization, plotting, feature selection, building and evaluating model.
- **Loading and reading the data** - This helped to check the number of columns and rows using shape and checked the stats of the numerical column using the describe and info functions.
- **Understanding and cleaning the data**
    - Verified the null values and removed columns having less than 20% null values
    - The label with "Select" values was used to identify null and missing values.
    - Using value count identified columns with single value and accordingly removed as it will not train the model.
- **EDA - Univariate and Bivariate Analysis**
    - Performed univariate and bivariate analysis.
    - Verified the outliers using box plot
    - Verified the correlation using heatmap
- **Preparing data for modelling**
    - Replaced categorical columns that had less value in comparison to others
    - Created dummy variables for the categorical column
    - Split the dataset into train and test set
    - Applied standard scaler to the train dataset (fit_transform)
    - Applied standard scaler to test data (transform)
- **Model Building** - Added the constant in the X train and fitted with the y train. Applied this first model to the whole dataset and checked the statistics for the model.
- **Feature selection using RFE**
    - Used logistic regression to check for feature selection and selected 15 features from the data
    - Applied the stats model to these 15 features
    - Checked the VIF and P value based on the summary
    - Repeated the feature until the p value was less than 0.5 and the VIF was less than 5, thereby finalizing on the fourth model

- **Final Model Evaluation**
    - The default value of 0.5 was used for the cut off and then converted the predicted value in 0 and 1
    - The overall accuracy is 0.79 and got printed in the confusion matrix
- **Checking other accuracies**
    - Sensitivity = 0.6625
    - Specificity = 0.8747
    - False positive rate = 0.125
    - Positive predictive value = 0.7634
    - Negative predictive value = 0.8094
- **ROC Curve**
    - Plotted the ROC curve using sklearn.metrics class
    - Final curve between the converted value of 0, 1 and probability float between 0 and 1.
- **Optimal Cut Off**
    - A column was created using different dataframe with an interval of 01 from 0 to 0.9
    - Accuracy, sensitivity and specificity were inferred. Accuracy = 0.7914, Sensitivity = 0.7932, and Specificity = 0.7903.
    - Performed plotting and derived the optimal cutoff value, which is 0.35
    - The false positive rate is 0.2096
    - The positive predictive value is 0.6977 and the negative predictive value is 0.8623
- **Calculating Precision and Recall**
    - Imported sklearn.metrics to derive the precision, recall scores
    - The precision score is 0.6977
    - The recall score is 0.7932
- **Predictions on the test set**
    - Constant was added to the test data set
    - Predicted the X test
    - Mapped the probability value in 0 and 1 with a cutoff point as 0.35
    - Accuracy = 0.7837, Sensitivity = 0.7740, and Specificity = 0.7897.
- **Conclusion**

  These inferences were made based on the analysis:
    - The following sources impacted the potential leads:
        - Google
        - Direct traffic
        - Organic Search
    - The last activity that affected the leads were:
        - Opened emails
        - SMS
        - Olark Chat Conversation
    - The total time spent on the website and the total number of visits had their share of impact

- Working professions contributed to the lead
- Final Model (res) res = logm4.fit()
- The cut off probability is 0.35
- More than 0.35 were converted as lead
- Less than 0.35 will not be converted as lead
- Accuracy of the train data 0.791
- Accuracy of the test data 0.784
- When the lead is increased or decreased, the Cut off can be adjusted
- The lead score targeting can be done from the top.