**Summary Report – Lead Scoring Case Study**

This analysis focuses on X Education and aims to discover strategies to attract more industry professionals to their courses. The provided data offers valuable insights into potential customers’ behavior, including their site visits, time spent on the site, methods of reaching the site, and the conversion rate.

These are the following steps used for the analysis and making the model

Step 1: Import libraries and Read data

Step 2: Inspecting the dataframe

We came to know there are 9240 rows and 37 columns. There

Step 3: EDA

Missing Value Analysis: There were missing values in few columns. The columns were dropped which has missing values > 35%.

Plotting different graphs to understand the behaviour of target variables against the predicted variable.

Checking types and number of categories in categorical variables.

Dropping the missing values which are having less number in the columns.

Frequency plotting of each variables for Converted. Mostly columns graphs are created

And box plot is done.

Then normalizing the numeric variables

Step 4: Creating dummy variables

Dummy variables are created with get dummies to convert the categorical variables into numerical.

Step 5: Train-Test Split

Train-Test split is done in the data with 70:30 ratio

Step 6: Scaling and Transforming

Trained data is scaled and transformed with MinMaxScaler

‘Receive More Updates About Our Courses’, ‘Update me on Supply Chain Content’, ‘Get updates on DM Content’, ‘I agree to pay the amount through cheque’ and ‘Magazine’ are he columns which have only one unique values. Since these will not help in analysis hence dropped.

Correlation is checked with all the variables

Step 7: Model Building

Firstly, RFE was done to attain the top 30 relevant variables to filter out from 43 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).

Step 8: Model Evaluation

A confusion matrix was made.

The optimum cut off value (using ROC curve) was used to find accuracy, sensitivity and specificity. The area under the curve was 79% which is quite good value.

Step 9: Predicting class for the test data

Step 10: Assigning the lead score

Step 11: Evaluating the model

ROC curve is created to understand the model and we found the area under the curve is 79% which is good to go for the model.

Confusion matrix is created and values are calculated.

When calculating we get

* + - Accuracy is 80.67%,
    - Sensitivity is 86.44%
    - Specificity is 71.13%
    - FPR is 28.87%
    - Precision is 83.2%

Conclusion:

The dataset is analysed and overall model is build and evaluated and checked with test data with good accuracy, sensitivity and specificity.

The main features to be focussed are:

* Total Time on Website
* Total Visits
* Lead Source with elements Olark Chat
* Last Activity with elements SMS Sent
* Last Activity Others.

More focus on making website attractive, investing on marketing on Lead source, mechanism for tracking of last notable activity etc. to be done