Lead Scoring Case Study

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Answer: The mentioning steps below is used in the assignment.

1) **Data Collection and Data Cleaning**:

   (a) Importing the data then cleaning it, checking if there are any null values.
   (b) We found out that some columns were in lead score data have null values so, we corrected them using the correct formulas.
   (c) Removed columns having more than 45% null values.
   (d) Rest of the missing values have imputed with the maximum items in the columns.

2) **EDA (Data Visualizations):** 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.
   (a) Univariate data analysis: value count, distribution of variable etc.

   (b) Bivariate data analysis: correlation coefficients and pattern between the variables etc.

3) Data Transformation:
   (a) The dummy variables were created with the categorical columns and binary variables into '0' and '1'.
   (b) After that Removed all the redundant and repeated columns.

4) Data Preparation:
   (a) Train-Test split: The split was done at 70% and 30% for train and test data respectively.
   (b) Feature Scaling: Scaling will be done with the Standard Scaler.

5) Model Building:
   (a) Use RFE for Feature Selection
   (b) Running RFE with 15 variables as output

(c) Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5

6) **Model Evaluation:**

(a) A confusion matrix was made with the help of metrics.

(b) Later on, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 90% each.

7) **Predictions:** Prediction was done on the test data frame and with an optimum cut off as 0.30 with accuracy, sensitivity and specificity of 90%.

8) **Precision – Recall:** This method was also used to recheck and a cut off of 0.3 was found with Precision around 89% and recall around 92% on the test data frame.

9) **Final Observation:**

Let us compare the values obtained for Train & Test:

Train Data:

- Accuracy: 92.29%
- Sensitivity: 91.38%
- Specificity: 92.79%

Test Data:

- Accuracy: 92.89%
- Sensitivity: 92.32%
- Specificity: 93.25%

10) **Conclusion:**

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

- Lead Origin_Lead Add Form

- Lead Source_Direct Traffic
- Lead Source_Welingak Website
- Last Activity_SMS Sent
- Last Notable Activity_Modified

We got Accuracy, Precision and Recall for the both Training and Test set.

We got high recall score than precision score which we were exactly looking for.

The Model seems to predict the Conversion Rate very well.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.