Lead Scoring Case Study

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) The presence of “Select” value in many of the categorical columns are handled by converting them to “NaN”(null) values
   (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):** We begin by studying the dataset provided and do exploratory data analysis using statistical and visualization methods to both the continuous and categorical data. The dataset has a larger section of categorical data which points us to apply more of a Logistic regression model post our Exploratory Data Analysis(EDA).

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) Checks are made on the Sensitivity, Specificity, Accuracy, False positive rate and True positive rate in the model evaluation phase using the Confusion Matrix created
   (b) An ROC curve is also plotted to find the area under it for an optimal prediction and find the probability cut off by plotting the Specificity, Sensitivity and Accuracy.
7) **Predictions:** Prediction was done on the test data frame and with an optimum cut off as 0.36 with accuracy, sensitivity and specificity.

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: 80.76%
- Sensitivity: 80.43%
- Specificity: 80.97%

Test Data:

- Accuracy: 80.34%
- Sensitivity: 81.04%
- Specificity: 79.91%

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_Welingak Website
- What is your current occupation_Working Professional
- Last Notable Activity_SMS Sent
- Do Not Emai

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.