Summary report of Lead Scoring Case Study

Objective:

● To help X-education select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

● We need to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance

Process followed:

Step1:Data Cleaning

- Understand the dataset by inspecting it and referring to provided data dictionary
- Look for missing values. If missing value count is more than 40%, then drop those columns(variables). Otherwise, missing values will be imputed accordingly with Mean/Media/Mode.
- `Select' category should also be considered as missing value.
- Check for Outliers and deal with them accordingly.
- Drop the Sales team generated variables as these are not required for Model building**.**

Step2: Data Preparation

- Perform EDA on the dataset to get useful insights on numeric and categorical variables.
- Create dummies for all the categorical columns. This dataset will be used for model building.
- Perform train-test split assigning 70% to Training set and remaining 30% to test set.
- Perform scaling of numeric variables using Standard Scaler method to get all values in the range of appx. -3 to +3.

Step3: Modeling

- Use techniques like RFE to perform feature selection. Select top 15 features.
- Build the first Logistic Regression model and check for the statistics such as VIF and p-values.
- Identify the variables with largest VIF value and drop this column.
- Follow the above 2 steps until we get all the variables with VIF value less than 5
- Take random cut-off such as 0.5 and make the Confusion matrix and calculate Accuracy, Sensitivity and Specificity. We found Sensitivity value was 62% which is less and has to be above 75% to predict Potential leads
- Calculate the above 3 metrices at each cut-off from 0.1 to 0.9. We found that at 0.3 cut-off, these 3 values are coinciding and are above 75%.
- Check the model performance over the test data and found the above metrices are close to that of Training set.
- Generate lead score for the test set.