Summary:

Title: Logistic Regression-Lead scoring case study.

Objective:

Read, evaluate, and clean the given data, as well as visualize the available patterns in the given datasets to know lead conversion ratio. Splitting the data into Train and Test data set and prepare the data set for Model building. Calculate model Accuracy, Specificity, Sensitivity and plotting the ROC curve to know Area Under the Curve. Find out the Optimal Cut-Off point of the model to filter out the Hot leads. Assign lead score to each lead according to the lead conversion probability and categories Hot leads to increase the lead conversion ratio from 30% to 80%.

Methodology:

I)   Understand the various steps of Logistic Regression with the help of Recorded videos and given in course contents.

II)   Import and read the given data set.

III)   Then inspect the data-frame for it's dimension, null values, and the summary of various numeric columns.

IV)   Data Cleaning and Preparation by doing null value treatment (removing or imputing the values), and elimination of useless columns.

V)   Then did EDA as learn before. With the help of EDA Data visualization (Univariate, Bi/Multivariate analysis, Outlier treatment, Box-Plot and Heat-maps to see correlation) is possible and it helps to see the various available pattern in the data. Which further helps to take business decision.

VI)   After EDA preparation of the cleaned data is required for Model building step. Therefor we prepared the data (Yes/No to 0/1 and create Dummy variables of the categorical data).

VII)   Once data preparation has been done, I splitted the data set into Train and Test data.

VIII)   After data preparation I did feature rescaling. For Feature rescaling I have used Standardization method as data are very large. It makes all the data set on the same scale.

IX)   Model building is the next step we took to analyze the P value and VIF values and remove those variables which has very high P-value and/or VIF values.

X)   Once model has been built plot the ROC curve and check the Accuracy, Sensitivity, Specificity of the system.

XI)   Recalculate the Cut-off point and find out the optimum cut-off point to increase the lead conversion by narrowing the criteria to find out Hot leads to increase the conversion.

XII)   After doing all the operations on Train data, same thing I did for Test data and compare the results, whether our model is stable or not.

XIII)   I found that proposed model is pretty good and much stable and having higher Accuracy, Sensitivity and Specificity.

XIV)   At the end of the process I assign Lead Score to each lead to differentiate the leads on the basis of conversion probability which helps to differentiate Hot leads from other leads increase the lead conversion.

Outcomes/Learnings:

I have learned about various important steps of Logistic Regression.
I found that the selection of columns and imputing the proper values for Null values and when proper information is not mentioned is the most important step to find the best fit model.

****