Summary:

An education company named X Education, which sells online courses to industry professionals, wants to identify hot leads from initial pool of leads. Then nurture those hot leads and improve their lead conversion rate. For this purpose, a logistic model is built that assigns a lead score to each lead, by the following steps.

1. Reading & Understanding the Data:
   In this step structure of the data set, basic info of the features, and basic stats of the numerical features is understood.
2. Data Cleaning:
   a) Dealing with Missing values: The features with over 35% missing values are dropped, features with less than 5% are imputed with median value (numerical) and with mode values (categorical), classes of a feature, less than 1% count are grouped together.
   b) Dealing with Outliers: Outliers are dealt by caping the numerical features between 1% and 99% percentiles.
3. EDA:
   a) Numerical Features: Box plot is plotted for each numerical feature with 'Converted' as X variable. Heat map is plotted using correlation matrix.
   b) Categorical Features: Bar plot is plotted for each categorical feature with 'Converted' as X variable. Conversion rate for each class of a feature is highlighted.
4. Data Preparation:
   - A binary variable is mapped with 0/1 and one hot encoding is done for categorical variables with multiple levels. These features are integrated with original data set and redundant features are dropped.
   - Data set is split into train and test set in 70/30 ratio. Non binary train set features are fit transformed using standard scaler.
   - Customized functions are created to simplify the model building process.
5. Model Building:
   - Top 15 feature variables are selected through RFE method.
   - Statistically insignificant features and features with high multicollinearity are dropped, classes with low weightages are also dropped to make the model lighter and eventually 10 parameters are selected as the final model parameters.
6. Checking the Model Metrics:
   Accuracy, sensitivity and specificity of the model is calculated with a random cut-off value and ROC Curve is plotted which had AUC score of 0.89.
7. Finding the Optimal Cut-off value:
   - Optimal cut-off value is calculated by plotting sensitivity, specificity & accuracy against different thresholds, probability at the intersection is chosen, which was 0.35.
   - Model metrics using the optimal cut-off value: accuracy = 80.84%, sensitivity = 81.55% & specificity = 80.41%

8. Making predictions on the test set using the Logistic Model:
    - Non-binary numerical features are transformed using the mean and standard deviation values of the train set.
    - Predicted probabilities of the test set is calculated using the model 6.
    - Predicted values are calculated using the optimal cut-off probability, I.e., 0.35.
    - Lead score value for each lead is calculated using probabilities predicted by the model.
    - Test set metrics values are calculated. Metrics: accuracy = 79.98%, sensitivity = 80.04%, specificity = 79.94%
9. Conclusion:
    - Coefficients of model parameters are interpreted using odds.
    - Train set and test set metrics are highlighted.
    - Lead score of each Hot Lead is highlighted.