LEAD SCORING CASE STUDY
SUMMARY

- **PROBLEM STATEMENT:**

  X Education, an education company which sells online courses to industry professionals generates a lot of leads through their websites, marketing campaigns and past referral but its lead conversion rate(~30%) is very poor. The company requires to build a model to ensure that they achieve target lead conversion rate i.e. around 80%.

- **OBJECTIVE:**

  Building a logistic regression model wherein lead score need to be assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- **WORKFLOW:**

1. **Data Loading and Cleaning**
   a. Loaded the dataset which has 37 features of different data types and had 9240 data points.
   b. Replaced the 'select' values with NULL as the user did not select any option from the list.
   c. A threshold of 45% was set and features having missing values beyond it were dropped. The rest of the missing values were imputed with favourable aggregate function after analysis.
   d. Detected the outliers and used capping to handle them.
   e. Dropped irrelevant columns after thorough analysis.
   f. Fixed the incorrect values and clubbed low frequency values for the ease of analysis.

2. **Exploratory Data Analysis**
   a. Checked the lead conversion rate which stood as 38%.
   b. Performed Univariate Analysis on both numeric and categorical features and gained some useful insights.
   c. Plotted heatmap to gain insights on correlation among numeric features.

3. **Preparing the Data for Modelling**
   a. Mapped binary values to 0 and 1.
   b. Created dummy variables for the categorical features.
   c. Performed train-test split (70:30).
   d. Rescaled(fit-transform) the continuous variables of the train set.

4. **Training and Modelling the Data**
   a. Checked correlation coefficients to identify top-15 correlation pairs among independent variables.
   b. Performed X-y train splitting and selected features through feature ranking using Recursive Function Elimination (RFE).
   c. Manually eliminated independent variables on the basis of high p-value (>0.05) and high VIF (>5).
   d. Selected the final model on the basis of important statistics, significant variables and no multi-collinearity.

5. **Prediction and Model Evaluation**
   - ❖ Train Set
     a. Made predictions using the final model by setting random cut-off and evaluated it using various metrics (Accuracy, Sensitivity, Specificity, Precision, Recall etc.)
     b. Found the optimal cut-off as 0.3 using various probabilities and estimated the final predicted values.
     c. Revaluated metrics for final predicted values and assigned lead score.
   - ❖ Test Set
     a. Transformed the numeric features of the test data with Standard Scaler.
     b. Performed X-y test splitting and dropped features in X test to align with X train.
     c. Made predications on the test set using the final model set cut-off as 0.3(optimal cut-off) and evaluated the model using various evaluation metrics.

	Train Score	Test Score
Accuracy	92.981000	93.218000
Senstivity	91.160000	93.425000
Specificity	94.253000	93.083000
FPR	5.747000	6.917000
FNR	8.840000	6.575000
Recall	90.794809	93.424658
Precision	90.718000	89.816000
AUC	96.805288	97.316473

Evaluation Score Chart

- ➢ **LEARNINGS**:
  The scores of evaluation metrics for both train and test set are good and ensures the stability of the model hence can be used to achieve the business goal.