Summary Report

I am very grateful to Upgrade/IIIT-B to give a chance to work on such type of good assignment.

We have proceeded assignment with very first step of load data and gather information about data. Problem is **classification problem**, so we used **logistic regression** to solve this problem. Problem comes under **supervised learning**. *We* divided assignment into phases.

Phase 1: Gather information about data

We collected data, inspect data at the level of get to know basic information about data. Purpose of assignment, requirement of business, type problem.

Phase 2: Inspection of data and infer information dataset

Read data to know any hidden pattern in data set. Size of dataset, how much missing values in datasets. In lead assignment, we found "select" value which have no meaning so must change in Nan.

Phase 3: Clean data / Wrangling data

We dropped columns which have more than 70% missing values, even columns also dropped which are not adding any information in model.

Some columns are having "select" value in columns, no sense of it in dataset so we have changed it in nan.

Phase 4: Handel outliers

Very few numerical columns have outlier. We analysis them and remove outliers.

Phase 5: Create dummy variables and Feature scaling

Convert all categorical variables into dummy variables.

Phase 6: Feature selections

Use RFE to select most relevant to predict target variable.

Phase 7: Split data in train & test and Model Preparation

Make model using logistic regression method. Make model perfect with select relevant feature in model. We do iteration to train model with best features. We removed and add features to select best fitted model. Check **VIF to find multicollinearity** and remove features which has very high multicollinearity.

Phase 8: Assign leads 0 to 100

We assign leads 0 to 100 according to probability to conversion of any lead. 0 is very poor chance to conversion and 100 is very high probability to conversion. We will educate X Education team to call all who have good probability to convert.

Phase 9: Check model accuracy score, Roc Curve and best cutoff

Check model accuracy. I have 92% Accuracy in my model. Check **sensitivity** and **specificity** of model. Check precision and recall score. ROC curve to check model to how much data covered.

In out model, 0.3 is best cutoff. At the 0.3 cutoff, we got balanced score of sensitivity and specificity.

Phase 10: Test data with trained model

In this phase, we will fit test data in train model and check accuracy of model with test data set.

Now we must notice the accuracy of trained model and accuracy of test dataset approximately near about each other.

Learning of assignment:

We learned in this assignment, how to check sensitivity and specificity. As we got X Education requirements then we used how to increase and decrease score of sensitivity and specificity as per business.

We found that coefficient value is weight of variables that help lead toward convert customer and best variables which most relevant to find hot leads.

Thank you