**Lead Scoring Assignment Summary**

Problem Statement:

Creating a machine learning model for X Education Company , they have online platform for their

education courses which predicts and assigns a lead score to each lead based on different features

available from past data. This is a Logistic Regression problem as it has predicting the leads and

Probability of conversion.

The steps which we used are as follows:

1. Cleaning Data:

The data was almost clean except some null values and the by default option ‘Selected’ had to be replaced with null values as it was not giving any output. Few of the null values were dropped as their percentage of missing was very less.

1. EDA:

A EDA was done to thoroughly understand the data which was provided. With the help of visualization we found some interesting insights which were presented in presentation file also.

In categorical features there were many types of categories in some columns , if in columns there are many categories, so the categories which have very less data, we simply merge them and put it them to ‘Other’ category . In numerical features we handle outliers by capping them it to 90th percentile. We dropped some columns , those were not giving any insightful information.

1. Dummy Variables Creation:

Dummy variables were created on categorical features . For numeric features we used Min Max Scaler.

1. Train-Test split:

The split was done 70% on train data and 30% on test data.

1. Model Building:

After dummy variable creation there were many columns, so we decide to take top 15 columns using RFE . Later we dropped some columns manually by the criteria if the p value is greater than 0.05 and VIF greater than 4 .

1. Model Evaluation:

For model evaluation we used confusion matrix . Next for finding optimal curve we use ROC curve and find evaluation metrics such as accuracy ,sensitivity and specificity.

1. Prediction:

Prediction was made on test dataset with an optional cutoff of 0.3 and accuracy of 79% , sensitivity of 82% and specificity of 77% .

1. Precision and Recall:

Beyond above metrics we also calculated Precision and Recall on cutoff of 0.3 which came out to be 62% and 82% .