REPORT ON LEAD SCORING CASE STUDY:X EDUCATION

This analysis is based on an education company named X Education sells online courses to industry professionals have enough leads but the conversion rate of these leads is very poor. To speed up this process of finding potential leads we are doing this analysis. The data provided here gave us enough information like how potential customer reach the site, conversion rate etc.

ANALYSIS:

- Import the data and understand it.

- **DATA CLEANING:** Data was needed to be clean as it doesn't have any duplicates but needed to replace apply with NaN values, handling missing values, dropped the columns having missing value percentage>35%, missing value percentage between 35% to 20%, is replaced with others or highest mode. Those columns with less than 2% missing values we dropped the rows.

- **OUTLIERS TREATMENT & EDA:** Outliers are treated by removing 1% top & bottom outlier values. EDA is performed over categorical and numerical variables, where we have dropped highly imbalanced columns. In some columns we have clubbed the low value count columns to one single group.

- **DUMMY VARIABLES:** Created the dummy variables for non-binary categorical variables with multiple levels and later dropped the repeated variables.

- **TEST-TRAIN SPLIT:** Split takes place with Train 70%, Test 30% respectively.

- **LOGISTIC REGRESSION MODEL BUILDING:** First we used RFE to select the most important 15 variables and then we used stats model and VIF**.** We considered final features with p-value<.005 and VIF< 5, so all features are significant and there is very less multicollinearity present in the data set. The area under the ROC Curve should have a value close to 1. We are getting a good value of 0.88 indicating a good predictive model & 0.4 is the optimum point to take it as a cutoff probability. By working on confusion matrix of train data, we get Accuracy: 80.88%, Sensitivity: 76.52% and

Specificity: 83.57%. The precision and recall tradeoff come out to be 0.4.

- **MAKING PREDICTIONS:** The overall accuracy in test data comes out to be 80.67% whereas, sensitivity 77.72% and specificity is 82.46%

- **FINAL OBSERVATION**: The hot lead customers should have high lead score, say lead score value >80. They are mostly the following customers:
    - Leads with add form
    - Customer who spent highest time on website
    - Working Professional
    - Customer Active in Olark Chat Conversation