SUMMARY

We deduced from the problem statement that X- Education wants to target high-potential clients and increase their conversion rates.

Data Understanding – There are many different independent variables in the dataset, such as country, type of occupation and so on. We removed some features such as asymmetries scores/index. 'Converted' was chosen as the target variable.

Data Modification: We saw default values such as ('Select'), which we transformed to a null value and computed as percentages. We deleted the columns with more than 45% missing values. Because the data was skewed, the majority of the columns with significant missing values were eliminated some features eliminated asymmetric scores/ index. We also removed some features which hold only 1 unique value like magazine, get updates on DM content.

EDA – we shared with missing value imputation and filled the missing values with median for continuous data and filled the missing values of categorical data by making another category.

Then we did outlier treatment and fix all the outliers by capping the outliers. TotalVisits variable has outliers we treated the outliers. After that we performed data visualization using matplotlib and seaborn and got some meaning insights from the data.

Data Pre- processing – In data pre processing step we started with dummy variable creation then we did train test split on the X and Y data. Then we did the scaling of X_train part of the data using fit_transform and scaling of X_test using transform.

Modeling Building - Using RFE, we identified the top 15 factors that contributed to the model most, then after manual selection we finally got 12 variables. This was achieved by comparing the p value and VIFs.

Threshold and Prediction - In order to predict the target variable a correct cutoff should estimated. First we made a ROC curve through target variable and the probability came from model and got area under the curve 97.6% and it was also skewed to left top corner. Then, we took various cutoff values and find the accuracy, sensitivity and specificity on each of them and plot it. We got 0.35 as cutoff value where accuracy, sensitivity and specificity were 93.2%, 92.7% and 93.5% respectively. At last, we finally predicted the target variable by implementing the model and applying the cutoff. We also assigned the lead score by multiplying the probability with 100.

Inference- The leads can be categorized in further category such as cold leads, warm leads and hot leads. cold leads having lead_score less than 35, warm leads would be lead_score between 35 and 80 and hot leads having lead_score greater than 80. The company can mainly focus on the hot leads because they are most likely to get converted.