SUMMARY

Business Problem

With a lead conversion rate of only 30%, X Education is facing difficulties. The organization aims to find high-potential leads and raise the conversion rate to about 80% in order to maximize efficiency. Developing a logistic regression model with lead scoring is the aim in order to help the sales team prioritize communication and concentrate resources on prospects that have a better chance of converting.

The following steps have been used:

1. **Data Load:** Loaded the required libraries and packages and loaded the data into pandas data frame from the given csv and checked the shape, columns and datatype of the columns.

2. **Data Cleaning**: The data was cleaned by removing the columns which had only one type of value. Also, the columns having 'select' as value were replaced with null. Few columns which were insignificant were dropped and columns with high null values were imputed with 'not given' value. Then the rows were dropped for the remaining less amount of null values and finally, the columns containing heavily imbalanced data were dropped.

2. **EDA**: EDA was done on the cleaned data with uni-variate and bi-variate analysis of both the categorical as well as numerical data. Correlation was also analyzed using the heat map.

3. **Outlier Detection**: Outlier was checked using the statistical approach and no significant outlier was found.

4. **Dummy Variable Creation**: Dummy variable was created for all the categorical columns in the cleaned data set using one hot encoding.

5. **Train Test Split**: Data was split into train and test data set using 70:30 split.

6. **Scaling**: All the numerical column values were scaled using standard scaler.

7. **RFE**: RFE was used to get the top 15 columns to be used for model building.

8. **Model Building**: Model was built and columns were removed based on higher p-value or VIF, and was finalized once the p-value as well as VFI were under significant range.

9. **Prediction on train data set**: The model was evaluated on train data set and confusion matrix was created over different cut-offs, and optimal cut-

off was declared based on the accuracy, sensitivity, specificity curve and a data frame with actual, probability of conversion and predicted value was there and accuracy, sensitivity, precision and recall was evaluated based on the new confusion matrix.

10. **Precision Recall Trade-off**: Precision Recall curve was created and based on the curve, an updated cut-off threshold was made and the 'predicted' column was updated based on the cut-off for the 'conversion probability'.

11. **ROC Curve**: ROC curve was built and area under curve was calculated.

12. **Prediction on test data set**: Prediction was made on the test data set, and based on the cut-off the data frame was created with the actual converted, and predicted value and the confusion matrix was created and then accuracy, sensitivity, precision and recall was calculated.

13. **Assigning Lead score**: Based on the conversion probability, the lead score was calculated by multiplying it with 100 and adding to dat frame.