SUMMARY

This data analysis is performed to increase the sales of online courses for company called X-Education. Various attributes is used which may or may not be useful in ultimately deciding whether a lead will be converted or not.

The following process is used :-

1. **Data Cleaning & Processing** :-
    - Check duplicate/redundant data and remove from the dataset
    - Select value will be replaced with Nan
    - Drop the high percentage of null values
    - Categorical Column analysis
    - Some Imputations on the null values
    - Treated the missing values by imputing the favorable aggregate function like (Mean, Median, and Mode)
    - Detected the Outliers

2. **Exploratory Data Analysis** :-
    - Univariate Analysis with respect to target value of 'Converted'. Denoted the converted lead as 'one' and not converted as 'zero'
    - Used Heat Map for finding the Correlations between the features

3. **Data Preparations** :-
    - Dummy Variables creation: The dummy variables are created for all the categorical columns.
    - Scaling: Used Standard scalar to scale the data for Continuous variables.
    - Train -Test-Split: The Spit was done at 70% and 30% for train and test the data respectively.

4. **Model Building** :-

    Model Building takes place by logistic Regression. By using RFE with provided 15 variables. It gives top 15 relevant variables. Later the irrelevant features was removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value 0.05 were kept).

5. **Model Evaluation** :-
    - A confusion matrix was made.
    - Later find the optimum cut-off value by using ROC curve was used to find the accuracy, sensitivity and specificity.
    - Prediction: Prediction was done on the test data frame an optimum cut-off as around 0.30 with accuracy, sensitivity and Specificity of around 83%.

- Precision-Recall: The method was also used to recheck the data
- Check the model performance over the test data(confusion matrix, sensitivity)
- Generate Lead Score variable

6. Conclusions

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 76% and 77.5% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- And three important feature depending on the co-efficient of Final Model are a)Lead Origin_Lead Add Form ,b)Total Time Spent on website ,c)Total Visits.