Problem Statement

A company called X-Education in the education industry produces leads from numerous sources and gives them to the sales staff. After contacting the leads, the sales staff attempts to convert them, however the current rate is just about 30%. Building a logistic regression model is necessary to provide a score of 0-100 for each lead, and the conversion rate must grow to about 80%.

STEP BY STEP APPROACH

1. Reading and Understanding the Data

2. Cleaning the Data
    a. Missing Value treatment
       We deleted the columns with missing values larger than 30% after replacing "select" values with NaN values, and we imputed the remaining missing values in categorical variables by using the mode or generating new values. Numerical columns similarly contain fewer values. So, we removed those rows.
    b. Outlier Value treatment
       We eliminated the two numerical columns' outliers by capping the upper and lower quantiles at 75% and 25%, respectively.

3. Exploratory Data analysis
    a. Categorical Variables Analysis
       We performed EDA to analyse categorical factors and found that lead source, lead origin, last activity, and current occupation all had values that influenced lead conversion.
    b. Numerical Variables Analysis
       Using EDA on numerical variables, we discovered that if a lead spends more time on the website, the likelihood of conversion increases.

4. Preparing the Data for Modelling
    a. Creating Dummy Variables (Categorical Variables)
       All categorical variables with more than two values were given dummy variables.
    b. Feature Scaling (Numerical Variables)
       We have scaled the numerical variables using standard scaler.
    c. Splitting the Data into Train-Test Data sets
       We have split the data into train-test sets at 70%-30% respectively.

5. Creating Logistic Regression model
   Initially, we constructed a raw model containing all 28 variables, and then we began optimising it. At the beginning, we built another model with 15 variables after using RFE to eliminate features. Finally, we improved our model by deleting features with high multicollinearity(VIF) and significance (p-value) values.

6. Evaluating the Model on the Train Data Set
   For our final model, we first assumed that leads would be converted with a probability of 0.5; however, by plotting accuracy, sensitivity, and specificity curves at various probabilities, we discovered the best cut-off point to be 0.35. Using 0.35 we calculate accuracy, sensitivity, and specificity at 80.5%, 79.8% and 80.8% respectively.
   We also calculated precision and recall scores at 71.8% and 79.8% respectively.

7. Evaluating the Model on the Test Data Set
   Evaluating the model on test data set and predicting values at 0.35 probability the accuracy, sensitivity, and specificity are 80.4%, 79.3% and 81.2% respectively.

Conclusion:-

1. The predictions made on the Train and Test Data Sets are approximately same.
2. Hence we can say that our model is good at distinguishing positive and negative classes.
3. Finally, most important features effecting our model are
    a. Lead Origin_Lead Add Form
    b. What is your current occupation_Working Professional
    c. Last Activity_SMS Sent
4. Followed by these Features
    a. Lead Source_Welingak Website
    b. Last Activity_Other
    c. Last Activity_Email Opened