Summary Report

Problem Statement:

X Education company provides Online course which receive request from professional against the course which is classified as Lead.

Current lead conversion ratio is 30 % where as CEO of company given target of lead conversion rate is 80%.

Data Preparation:

- **Data Set**: Data set having 37 feature variables and 9240 observations, out of that almost 80 % feature variables are qualitative variables with one or more categories.
- **Data Issues**: More than 50 % variables have missing data, Apart from missing data (None), few categorical variables have value "Select" which does not have any meaning. So will treat tha as a missing data.
- **Data Cleaning**: Features like How did you hear about X Education, Lead Profile, etc. have more than 70% missing data, apart from that few other features have more than 30 % missing data. Will drop all such features as dropping more than 30% rows is not a good idea.

  Finally dropped approx. 2 % observations/rows because of missing data.
- **Derived Categories & Variables**: Lead Source feature variable have 21 different categorical values, but almost 80 % of data variance explained by top 4 categories, so will merge all other categories in new "Other Lead Source" category. Apply similar logic to some other features and Convert all categorical variables into dummy variables.

EDA:

- **Quantitative Variable**: Pair Plot, Heat Map and Distribution plot
  - **Outlier Analysis**: Box plot & Distribution plot shows clear outlier in **Total Visits & Page Views Per Visit** features, will drop outlier data rows
  - **Correlation**: Total Visits & Page Views Per Visit variable are highly correlated with each other (0.77), so drop one of the variable.
- **Qualitative Variable**:
  - Category wise data distribution analysis shows, some categories have high numbers of leads good conversion ratio like Lead Source (Google, Olark Chat, Direct Traffic), Lead as well as Origin (Leading Page Submission), Last Activity (Email Opened, SMS Sent), etc.
  - **Correlation**: **LS Reference, LS Lead Add Form** and **NA SMS Sent, LA SMS Sent** are highly correlated with each other, so drop one variable from each group.

Model Building:

- **Data Scaling**: Split entire data into train & test set also Scale using Standard Scaler.

- **Feature Selection**: After data cleanup and dummy variable creation activity we end up with 17 features. Using **RFE (Recursive Feature Elimination)** technique will choose top 13 features and build Logistic Regression model.
- **Logistic Regression:**
    1. Train Logistic Regression Model using **GLM** method of **statsmodels**, which provide detail summary.
    2. Drop insignificant variable by using P-Value score.
    3. If there are more than one insignificant variable, drop one by one and repeat step 1 & 2
    4. Check VIF score: It shows multicollinearity between features
    5. If VIF score greater than 5 than drop that feature and Repeat all steps again till we have Low VIF score & High Significant variable

Model Evaluation:

- **Train Set Prediction:** Predict conversion probability of all leads. Using this value generate Lead Score and using specific cut off value mark that Lead as Hot or Cold Lead (1 or 0).
- **Evaluation:** Some model evaluation metrics used like model accuracy, ROC-AUC curve and other metrics like Precision-Recall or Sensitivity-Specificity based on business problem.
- **Evaluation Based on Problem statement:** Our core problem is to increase conversion rate and target is 80 %. Here we are targeting Precision (how accurately we predict actual True values out of all predicted True values), which should be more than 80 %.
- **Cut off Score**: Based on Precision score and target conversion ratio we set lead score cut off to 60.