Summary Report

Lead Scoring is binary classification ML problem implemented using Logistic Regression Algorithm. Below steps summarize the implementation.

Reading & Understanding Data (EDA):

The dataset contains 9k records with 30 categorical and 7 numeric features out of which 17 features had null values. No duplicate records found.

Data Cleaning:

- Variables causing class Imbalance and variables with single binary category were dropped as they are not significant to infer any pattern for model building.
- Few variables come with a level 'Select' which is as good as null, so converted as "nan".
- Features having >35% missing data were dropped.
- Filled missing values with median for numeric valuables.
- Created a separate category named 'Missing' for couple of categorical features because of high % of missing data.

Univariate Analysis:

- Categorical features plotted using count plot
- Identified variables that contribute to high conversion rate.
- Grouped all values having < 1% of whole dataset into single column to restrict creation of unnecessary dummy variables.

Bivariate Analysis:

- Numeric variables plotted using pair plot
- Visualized correlation between independent variables and found “TotalVisits” and "Page Views Per Visit" highly correlated.

Outlier Handling:

- Numeric variables plotted using histogram and box plot.
- Capped outliers to below 0.99 percentile.

Data Preparation for Model Building:

- Performed binary encoding for binary categorical variables.
- Dummy variables created for multi-level categorical features and dropped original features.
- Train and Test split: 70: 30
- Used standard scalar to scale numerical variables.
- Identified highly correlated features using correlation matrix
- Dropped them to avoid multicollinearity

Model Building:

- Feature Engg begin with total 31 features
- Applied Automated Feature Selection using RFE, drilled down to top 15 features.
- Fitted GLM Stats model from binomial model family.
- Total 3 Models fitted iteratively and p- value and VIF monitored
- Model 3 selected as final model after dropping features one-by one having p-value > 0.05. Finally, Retained 13 features.
- For Model 3, no multicollinearity observed, VIF values for all features < 3.

- Predicted values calculated using 0.5 lead score probability cut-off initially.

Model Evaluation Model 3:

Selected 0.35 probability as optimal cut-off point where accuracy, sensitivity and specificity metrics converge. Model re-evaluated after calculating predicted values using the optimal cut-off.
The following model performance metrics:
Accuracy: 80.57
Sensitivity/Recall: 80.54
Specificity: 80.58
Precision: 71.88
ROC curve: AUC = 0.89
F1-Score: 75.96

With same cut-off probability, predicted leads on test data and observed metrics performance as:
Accuracy: 81.93
Sensitivity/Recall: 81.55
Specificity: 82.17
Precision: 74.92
ROC curve: AUC = 0.89
F1-Score: 78.09

Inferences:

- Observed just 1% difference in accuracy between training and test data sets implies model not overfitting.
- High Sensitivity or Recall shows almost all leads who are likely to convert are correctly predicted
- High Specificity shows that the leads who are not likely to convert are correctly predicted.
- Lead scoring is done between 0 and 100 using Model 3.

- Identified 708 Hot Leads using Lead Score threshold 90
- Predictor features:
    - Lead Source_Welingak Website
    - What is your current occupation_Working Proessional
    - Lead Source_Reference
    - Last Notable Activity_SMS Sent
    - What is your current occupation_Student