Lead Scoring Case Study Summary Report

We followed these steps while working on our case study:

- **Understanding the data**: Imported the data and checked for shape, info & describe to understand the give dataset.
- **Cleaning the data**:
    1) We observed a lot of values as 'Select', since this was equivalent to null values - we replaced the 'Select' with NaN.
    2) We dropped the columns with more than 35% missing data.
    3) We analyzed the rest of the columns with missing values & since they were within 1.5 %, we dropped the rows with missing value.
- **Data Preparation**:
    1) We've grouped the values with lesser counts in Last Activity, Lead Source & Last Notable Activity as 'Others'
    2) We identified the candidate columns for that requires dummy variables to be created. Once the dummies are concatenated with the original dataframe – we dropped the original columns.
    3) We've dropped the column which were highly skewed.
    4) We capped the outliers at 99 percentiles.
- **Exploratory data analysis:**
    1) Did value counts and combined the insignificant values to a separate group 'Others'
    2) Compared Lead Origin vs Converted and observed Landing Page and API generate most of the leads
    3) Compared Lead Source vs Converted and observed Google/Direct Traffic generates most leads but References/Wellingak Website had better conversion rates even with lesser leads.
    4) We dropped the columns which were highly skewed.
    5) We plotted heatmap to analyse the highly co-related columns after creating dummies and chose to use RFE for feature elimination.
- **Building the Model:**
    1) We split the data into train and test (70:30) and built logistic regression model
    2) Performed feature scaling
    3) Dropped the columns with high-p value and VIF value (multicollinearity) and re-built the model.
    4) Performed feature selection using RFE
    5) Plotted ROC and Precision-Recall curves & found the optimal cut-off point
- **Model Evaluation:**

  Calculated Accuracy, Sensitivity, Specificity, Precision and Recall

Training Dataset	Test Dataset
Accuracy: 81.33%	Accuracy: 80.20%
Sensitivity: 80.33%	Sensitivity: 79.07%
Specificity: 81.94%	Specificity: 80.87 %
Precision: 73.18%	Precision: 70.93%
Recall: 80.33%	Recall: 79.07%