Summary of Lead Scoring Modelling Assignment

The aim of the study is to identify the leads that are most likely to convert into paying customers. Model will provide a lead score to each of lead in range from 0 to 100, where higher score indicates higher chances of lead to be converted. Steps taken during the assignment and learning at each step are as follows: -

- **Exploratory Data Analysis (EDA):**
  - *Data Cleaning – Missing Value Handling:*
    - 6 columns (including all Asymmetrique score/index and Lead profile/quality) are observed with missing values > 40% and marked inadequate for modelling.
    - Columns 'City' and 'Country' are dropped due to very high clubbed missing values and single category option like 'Mumbai' and 'India'.
    - For 'Specialization', 'Tags', and 'What is your current occupation' categories options are restructured based on their distribution and are kept due to their good correlation with target variable.
    - Missing values are kept as 'Unknown' for 'Specialization', merged with one category for 'Tags', and proportionally distributed to all options for 'What is your current occupation'.
    - Mean and Mode is used for columns with low missing values.
  - *Data Cleaning – Understanding and Manipulation:*
    - 'Last Activity' and 'Last Notable Activity' in dataset represent relatively similar information and later looks more relevant due to its notable factor and due to last activity (action) may be temporary & exactly opposite to lead's sentiments.
  - *Univariant analysis:*
    - Categorical Flags (Yes/No) Variables
      - Dropping columns where single value has > 99% occurrences.
    - Categorical Variables:
      - Clubbing low frequency categories in 'Lead Source', 'Last Notable Activity'.
    - Numerical Variables:
      - Outliers handling for columns 'TotalVisits', 'Page Views Per Visit'.
  - *Multivariant Analysis:*
    - Dropping column 'TotalVisits' due to its high correlation with 'Page Views Per Visit'.
- **Data Preparation:**
  - Replacing categorical variables with dummy variables and dropping selective one column.
  - Making column 'Lead Number' as index and dropping as column.

- Splitting the Data into Training and Testing Sets in 70-30 ratio.
- Rescaling the Numerical Features using Standard Scalar.
- Finding correlation/Heatmap between all variables, and dropping variables with high correlation > 70%.

- **Building Model:**
  - Final dataset after cleaning and preparation consists of 19 variables.
  - In first logistic regression model using statsmodels, it is found that there exist variables with high p-value.
  - With multiple experiments using automatic feature selection (RFE), adequate 11 features have been selected, where all variables have p-value < 0.05 and VIF in range.
  - ROC plot show ROC curve area is 0.95, which is pretty good.
- **Model Evaluation:**

	Accuracy, sensitivity, and specificity balance [Train Data]	**Precision, and Recall Balance [Train Data] (Finally selected)**	**On Test Data**
Optimal Cut-off	0.27	0.36	
Confusion Matrix • True Positives (TP) • True Negatives (TN) • False Positives (FP) • False Negatives (FN)	 • 2181 • 3571 • 431 • 285	 • 2134 • 3669 • 333 • 332	 • 949 • 1547 • 130 • 146
Accuracy	0.8893	0.8972	0.9004
Sensitivity	0.8844	0.8654	0.8667
Specificity	0.8923	0.9168	0.9225
Precision	0.8350	0.8650	0.8795
Recall	0.8844	0.8654	0.8667

- **Factor affecting the Conversion Probability:**
  - Tags_Conversion_Chances_HIGH_CONVERSION_CHANCES
  - Last_Notable_Activity_SMS_Sent
  - Lead_Source_Direct_Traffic
  - Lead_Source_Organic_Search
  - Lead_Origin_Lead_Import
  - Lead_Source_Google
  - Lead_Source_Reference
  - Lead_Source_Olark_Chat
  - Total_Time_Spent_on_Website
  - Last_Notable_Activity_Email_Opened
  - Lead_Origin_Landing_Page_Submission