LEAD SCORING CASE STUDY SUMMARY

Step-1 – EDA – Exploratory data analysis

- Dropped the variables from **leads_df** which are having greater then 40% of missing/null values.
- Checked the remaining missing value percentage and went for imputation
- In case of imputation, I used “Mode” to impute as majority of the variables with missing values are Categorical variables.
- Numerical varibles like 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit' had missing values below 2%. Instead of dropping the variables, I went for dropping the rows with null/missing values. 98% of rows retained from the original dataframe.
- **Outlier treatment** – Checked outliers in numerical variables, noticed outliers and then performed treatment for outliers.Treatment done to cap the outliers by setting values below the 5th percentile to the value at the 5th percentile (Percentiles_1[0]) and values above the 95th percentile to the value at the 95th percentile (Percentiles_1[1]).
- **Univariate and Bivariate analysis** – Then used a for loop code to display the countplot between all the variables and the target variable “Converted”. After observing the generated countplots, clumped the lower frequency values of some columns to a separate columns .**Eg. “Management_Sector”** for “Specialization”.
- **Multivariate analysis** – Generated heatmap to check correlation among numerical variables.

Step 2 – Feature Engineering

- Dropped some unsignificant variables from the leads_df dataframe.
- Mapped binary variables with values “Yes”, “No” with 1,0.

- Created dummy variables for categorical variables which has more than 2 attributes.Then, dropped the original variables from which the dummy variables created.

Step -3 –Splitting the data into train and test set and Scaling

- Splitted the data into train and test set. Train_size = 0.7 ; test_size = 0.3
- Scaled all the numerical variables by StandardScaler().

Step 4 – Model Building

- By using statsmodels, the logistic regression model was built.
- Used statsmodels to get statistical significant parameters about the model.
- Model building process were repeated until the **p-values** & **Variance inflation factor (VIF)** were stabilized.(p-value <0.05 % VIF - <5).
- “result_7” model resulted in good stabilised p-value and VIF
- Optimised the cut-off as 0.3
- Exhibited ROC as 0.89
- Calculated evaluation metrics like accuracy ,sensitivity, specificity, false positive rate, positive predicted value ,confusion matrix.

Step 5 – Making predictions on Test dataset

- Scaled the numerical variables and made predictions.
- Evaluation metrics like accuracy , sensitivity ,specificity etc., were calculated

Step-6 Observation

- **TRAIN DATA :**Accuracy : 80.68 %, Sensitivity :82.37% , Specificity :79.56%
- **TEST DATA :**Accuracy : 80.53 % ,Sensitivity : 81.90 % ,Specificity : 79.75%
- Overall we built a good model
- Focus on these features "Lead Source_Welingak Website",”Lead Source_Reference","What is your current occupation_Working

Professional", "Last Notable Activity_SMS Sent","Last Notable Activity_Unreachable","Total Time Spent on Website" ,"Lead Source_Olark Chat".