Lead Score Case Study Summary Report

Objective:-

Build a model where you need to assign lead scores to all the leads who are visited page. The leads having higher lead score have higher conversion chance and lead having lower lead score have lower conversion chance. The model accuracy should be greater than 80%, so that 80% leads have higher conversion rate.

Analysis Methodology:-

1. **Loading Data:-**

   Load the required libraries and dataset for the analysis.

2. **Understanding The Data:-**
   Understand the dataset for the analysis according to the problem. According to the lead score dataset the dataset have 9240 data points and 37 features. Here the "Converted" column is the target variable as per our requirement. By considering lead score if converted value is "0" then not converted and "1" means converted.
3. **Data Cleaning:-**
   - i) By visualizing and check the dataset we get some missing values presented as "select". Convert them into NaN Values. Then remove the columns having higher missing values. After this remove the rows having higher missing values. Check the distribution of values of all the categorical values and drop the columns where one value has too high distribution from all the other variables.
   - ii) Check the spelling mistakes and rename values to reduce the distribution.
   - iii) Remove the outliers from "total Visits"
   - iv) Drop the columns having unique values and having same feature columns.
4. **Univariate Analysis:-**
   Visualized univariate analysis on data set by plotting graphs.
   - i) According lead origin feature "Landing page submission" have highest conversion rate
   - ii) According to Lead Source " Goggle" have highest conversion rate
   - iii) According to Occupation feature "Unemployed" have highest conversion rate
   - iv) According to Specialization "Finance Management" have highest conversion rate
   - v) According to Tags (Activity of leads after visited the page) "Will revert after reading the Email" have highest conversion rate
   - vi) According to Last Activity (Response of company) "SMS Sent" have highest conversion rate.
   - **vii)** According to city "Mumbai" have highest conversion rate

5. **Dummy Creation:-**
   For performing Logistic Regression we required categorical values so we have to convert the categorical string values to integer values like ("Yes": 1).
   Apart from this by using "Label Encoder" tools of pandas create columns for each categorical value.

6. **Train-Test-Split:-**

   For model building split the dataset into train and test dataset in the ratio of 70:30 train dataset have 70 percent of data and test dataset have 30 percent of data. Split the dataset by using train_test_split tool from sklearn.

7. **Scaling Data:-**

   Standardize the numerical valued columns of dataset "TotalVisits", "Total Time Spent on Website", "Page Views Per Visit" by using “StandardScaler”.

8. **Model Building:-**

   a. Build the Logistic Regression or classification model on the whole lead score dataset by using generalized linear model (GLM). After Building classification model for whole dataset analyze that some features have high p- values which will cause multicolinearity in the model.
   b. So remove those features by using VIF (Variance Inflation Factor). Eliminate the feature which has less importance.
   c. Predict the model.
   d. Plot the ROC curve (Receiver Operating Characteristic) curve and the cutoff point is 0.4 in this case.
   e. Fit the predicted model on the test dataset to get the final dataset.
   f. Concat the predicted dataset with the main the data frame.