Summary Report
Submitted By - Sakshi Manchalwar & Shreepriya Dogra

Steps of Analysis and Approach

0. Importing Libraries and Checking the Data
1. Missing Value Treatment
2. Outlier Treatment
3. EDA
    1. Univariate Analysis
    2. Bivariate Analysis
4. Data Preparation
5. Train Test Split
6. Feature Scaling
7. Feature Selection Using RFE & Further by Manual Inspection
8. Finding Optimal Cutoff Point Using ROC
9. Metrics beyond simply accuracy & Plotting the ROC Curve
10. Precision and Recall
11. Making predictions on the test set

Description

Step 0 - This involves importing the necessary libraries and understanding the data, its dimensions, and the data types of the columns. Identifying the target variable and the variables out of which feature selection will be done at a later stage. Duplicates were also checked.

Step 1 - Missing value treatment was done in a number of ways. The missing values in the case of different columns were replaced by either mean, median, or mode depending on the type of data in the column, i.e., numerical or categorical, and keeping in mind their business significance.

Step 2 - After looking at the percentile values in the columns outliers were removed from the TotalVisits and Page Views Per Visit columns. Boxplots were used to confirm the data after cleaning.

Note - After cleaning data 98% of data had been retained.

Step 3 - EDA

Step 3.1 -Data were analyzed after grouping according to the target variable. The relation of some of the variables with the Converted data was seen using countplots and catplots.

Step 3.2 - A pairplot was generated to understand the correlation between the numerical variables. Columns with high correlation can be dropped as they will not influence the target variable independently.

Step 4 - Based on the EDA analysis it is seen that many columns are not adding any information to the model, hence we drop them before further analysis. After dropping the unnecessary columns for categorical variables with multiple levels, dummy features using one-hot encoding was done.

Step 5 & 6 - Data is divided into Train and Test. Features that require scaling are done so using the Standard Scaler.

Step 7 - Feature Selection is done using the RFE method and the number of variables is reduced to 18. Then manually the model summary is analyzed and variables having a p-value of more than 0.05, are dropped one by one. After dropping each variable the model summary is again analyzed.

Step 8 - The confusion matrix is created and accuracy, sensitivity, and specificity for various probability cutoffs are calculated. The value of 0.4 is selected as a cutoff.

Step 9 - The fpr, tpr, and threshold values are calculated. Area under the ROC curve is calculated as 0.97

Step 10 - Using sklearn utilities the precision and recall score were calculated. A precision and recall tradeoff is done by plotting both of them and finding out where they intersect.

Step 11 - Using the cut-off of 0.4 the predicted and actual scores an accuracy of 93% is achieved using the model.

Learnings