CASE STUDY: LEADS SCORING

Summary Report

1. Data Cleaning
    - To begin with, we dropped the redundant columns which did not value to the information. (City, Country, I agree to pay the amount through check and A free copy of Mastering the Interview)
    - Some columns had a "Select" value so, we replaced all the Select values will null (NaN in python)
    - We checked for missing values or NaN for each column. Any column with missing values greater than 30% were dropped. Any column with missing values less than 30% were imputed with the most occurring value in that column.
    - On the leads source, we had duplication of categorical value "Google" which had a lowercase "google". So, we capitalized the entire column.

2. Data Transformation
    - We started with replacing all the columns (categorical) with Yes/No values to 0s and 1s.
    - We created dummy variables for the categorical columns and dropped the first column for each dummy variable (dummy variable trap)
    - We joined the dummy variables to the original cleaned dataset.
    - We checked for outliers in the numerical variables and found outliers in TotalVisits and Page Views Per visit column after observing percentile wise distribution and Boxplots.

3. Data Preparation
    - We split the dataset into independent (X) and dependent variables(y)
    - We checked the correlation and the heatmap between the independent variables.
    - We dropped column "Lead Source_Olark chat" which had high correlation with multiple columns
    - We split the data into train and test set and scaled the numerical columns using MinMax Scaler, so columns are in the same 0-1 range for a logistic regression modelling.
    - We checked for class imbalance and found the target variable to be 1s: 39% and 0s: 61%
4. Model Building / Modelling
    - Using statsmodel.api we created a logistic model and checked the summary of the model.
    - We used RFE to select the 20 best features and we created a model with the 20 best features from RFE and checked it summary.
    - We dropped column whose P-value was greater than 0.05 (α=0.05 or 5%) and then rebuild model using a function to minimize manual work.
    - After the final model was ready, we proceeded towards model evaluation

5. **Model Evaluation**
    - **We predicted the train set on the model**
    - **We plotted the ROC curve, and it was 0.88 which was good enough**
    - **We found the optimal cut-off of 0.35 using a line plot of Accuracy, sensitivity and specificity**
    - **We calculated Precision and recall on the train set and found it to be around 72% and 81% respectively.**
    - **We finally predicted the test set on the model and set the cut off for the predicted probability.**
    - **Then we again calculated Precision and recall on the test set and found it to be around 74% and 80% respectively which was close to the train set, and it meant the model was robust.**
    - **We finally concluded the evaluation and used the probability as a score by using x100 multiplier which gave us a probability percent score for each lead.**