Summary Report

1. Reading and understanding the data:

- First, we read the csv file and examined the data to understand it. Importing the necessary libraries to do so.

2. Data Cleaning and preparation:

- Checking total missing values for all columns in the data frame
- Dropping columns with more than 3000 missing values
- Dropping columns with highly imbalanced data that do not add information to our overall analysis (i.e., Search, Magazine, Newspaper Article, etc.)
- Dropping rows with missing rows for columns deemed as significant for analysis (i.e., Total Visits and Occupation columns).
- Converted null values in Specialization column to select so we can drop the dummy variable 'Specialization_Select' later on.

3. Dummy variable creation:

- Creating dummy variables for all categorical columns and dropping the duplicates.

4. Train Test Split:

- Assigning X and y variables and splitting the data set into train data and test data at a 70-30 split.

5. Feature Scaling:

- Importing and applying MinMax Scaler on the numerical features in the training data.

- Plotting a heatmap to check the correlation between all features of the data frame
- Finding the top 10 correlated features in the data frame

6. Model Building :

- Using RFE, we first selected the top 15 features for our model.
- We eliminated features from the model based on high p values and high vif values till we reached a model where all the features had p-value <0.05 and below a vif value <5
- With a cutoff point of 0.5, we checked the Converted vs Predicted values with the metrics (specificity, sensitivity and accuracy).
- We created an ROC AUC curve with an area 0f 0.87 below the curve.
- We plotted accuracy, sensitivity and specificity to find the optimal cutoff point for our model. Deciding on a cutoff point of 0.45.
- Plotted the Precision Recall curve to plot the tradeoff between both.

7. Make predictions on Test Data:

- Then we made predictions on test data with our current model and calculated the metrics (sensitivity, specificity and accuracy) on the test data.

Conclusions:

The final model shows an accuracy of 0.79 on training data and an accuracy of 0.78 on test data. It performs relatively on testing data as well as training data with the metrics used being relatively close on both data sets.

Training data: Sensitivity/Recall= 0.78 Specificity=0.81 Precision=0.79 Accuracy=0.79
Testing data: Sensitivity/Recall= 0.77 Specificity=0.79 Precision=0.78 Accuracy=0.78