**Context**

The project aims to provide a list of individuals who are likely not following an active lifestyle but are likely to adopt an active one if contacted. There are two data sets available. The first data set is a training data set that includes an ID, the individual’s age category (1 = under 30, 2 = 31-45, 3 = 46-60, 4 = over 60), gender (0=male, 1 = female), marital status (0 = Unmarried, 1 = married), choice of breakfast food (1 = breakfast bar, 2 = oatmeal, 3 = cereal), income, and if the individual lives an active lifestyle (0= no, 1= yes). The scoring data set has all the attributes of the training data set but does not include the active lifestyle attribute.

**Data Preparation**

In data preparation, the datasets are checked for missing values, inconsistent data entries, outliers, and value ranges. After checking for missing values, no missing values are found in either data set. The ranges of each attribute are compared in both data sets. The scoring data set is filtered for values that fit within the ranges of the training data set. The specific values that are filtered in the scoring data set are age values greater than or equal to 21 and less than or equal to 55. Income is filtered for values greater than or equal to 14 and less than or equal to 176. The training data set is checked for outliers. The outliers are detected using box plots and IQR testing. The detected outliers are replaced with the mean of the outlier’s respective attribute. Inconsistent data is checked, and no inconsistent data is found in either set.

**Building the Model**

The first step removes the ID attribute because ID does not predict an active lifestyle. Then a logistic regression model is run using active lifestyle as the dependent variable and all other attributes as the independent variables. The results of the first model found that marital status, choice of breakfast, and income are insignificant and not predictive of an active lifestyle. Therefore, those attributes are removed from the model. After running the model with the attributed removed, the age category is significant to the 99.9% confidence level, gender to the 95% confidence level, and age to the 90% confidence level. The model is checked for multicollinearity, and no variable has a variance inflation factor of 1.05 or more. Indicating there is little correlation between the variables. Additionally, a correlation matrix is created that corroborates the results of little correlation between the independent variables.

**Applying the Model**

With a log model, a prediction can be made for everyone in the scoring data set. The prediction is the confidence that someone lives an active lifestyle. The prediction values are in percentages. The resulting prediction list is filtered for individuals 50% to 65% likely to live an active lifestyle. This specific range is selected because these individuals are likely not living an active lifestyle; however, they are likely to live an active lifestyle if contacted or pushed to live an active one. Therefore, the project's objective is accomplished, and a list of individuals is available to direct the campaign.

**Supporting Code**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

**Supporting Visualizations**

Text

Description automatically generated

\*Results from log model

-------------------------------------------------------------------------------------------------------------------------------

Text

Description automatically generated with low confidence

\*Variance inflation factors

-------------------------------------------------------------------------------------------------------------------------------



\*Correlation matrix

-------------------------------------------------------------------------------------------------------------------------------

Table

Description automatically generated

\*Head of the list of individuals