# Customer Lifestyle Modelling Project

## Background

In recent months, our health tracker company has noticed that many customers drop out of the sign-up process when they have to self-identify their exercise lifestyle (ht\_users.lifestyle) – this is especially true for those with a "Sedentary" lifestyle. As a result, the company is considering removing this step from the sign-up process. However, the company knows this data is valuable for targeting introductory exercises and they don't want to lose it for customers that sign up after the step is removed.

In this data science project, our business stakeholders are interested in identifying which customers have a sedentary lifestyle – specifically, they want to know if we can correctly identify whether somebody has a "Sedentary" lifestyle at least 95 percent of the time. If we can meet this objective, the company will be able to remove the lifestyle-specification step of the sign-up process without losing the valuable information provided by the data.

## Project Objective

The objective of the project is to develop a model using customer health data that can predict whether or not a customer has a sedentary lifestyle with at least a 95% accuracy.

## Solution

The dataset used for the project was customer-level health data, including average active and resting heartrates, vo2, bmi, steps, workout time, country and pre-existing lifestyle labels. The lifestyle label was converted into a binary indicator column where a value of 1 signified that the customer had a lifestyle entry of ‘Sedentary’ and 0 otherwise.

All possible solutions were assessed on the basis of accuracy, F1 score (the harmonic mean of precision and recall) and recall. The dataset was also divided into training and test sets (80:20 split) to simulate solution performance on new customers.

The first baseline solution was a very naïve approach in which every single customer was predicted to have a sedentary lifestyle. This produced universally very poor performance and was immediately discarded.

A closer look at the training data using descriptive statistics indicated that on two of the input features, customers with a sedentary lifestyle could be separated with 100% confidence from non-sedentary customers. Hence, the second solution was a simple rule that labelled customers with less than 6,000 average steps as sedentary and non-sedentary otherwise. This produced 100% accuracy across all metrics on both the training and test sets.

To double-check the solution, a logistic regression model was developed using the same input variables (scaled to ensure that variables with higher magnitudes didn’t incorrectly influence model results). The logistic regression correctly identified the same two features as being most important in identifying customers with a sedentary lifestyle.

## Solution Maintenance and Drift

Whilst current data indicates that a simple business rule can cleanly and 100% accurately identify customers with a sedentary lifestyle, this may not always be the case in the future. Customer data should be periodically reviewed to double-check whether the relationship still holds and a more sophisticated model be developed/used (such as a logistic regression) if there are signs that the classes are starting to overlap.

## Project Improvement Recommendations

The analyst was unsure as to exactly what data is provided by customers at sign-up. It would be useful to have a clearer indication of this fact, and additional models be developed to incorporate this potentially richer dataset, especially in the event that a customer doesn’t provide information about a key variable like average number of daily steps (for example they may just not know).