**Predicting Calorie Expenditure**

Elliot Blackstone

**Github** –

**Introduction –** What metrics are predictive of calories burned in a workout? This is the question that this project aims to clarify. This was also the topic of the Kaggle competition “[Predict Calorie Expenditure](https://www.kaggle.com/competitions/playground-series-s5e5/overview)”, which is the origination of the dataset. The dataset contains 1 million rows and 9 columns, containing 8 features (id, gender, age, height, weight, duration, heart rate, body temperature) and the target (calories burned).

**Preprocessing –** In the original dataset, the features height, weight, duration, heart rate, body temperature and the target calories are all saved as type float. It was clear that body temperature was correctly categorized as type float but the remaining features and calories appeared to be integers. Indeed, there was one height value (154.1) and one heart rate value (109.903) that was non-integer. Rounding these two values to the nearest integer and changing the columns from type float to type int reduced the file size by 16.5%. The id column was unnecessary and was deleted to save another 16.5%. In total, the file size was reduced by roughly 33%, which should improve model training speed.

**Model Selection and Results –** Models were judged based on their root mean squared log error (RMSLE). The following models were constructed (some with feature engineering and some without): linear regression, generalized additive model, XGBoost, LightGBM, CatBoost, an ensemble of XGBoost, LightGBM, CatBoost, and AutoGluon. Feature selection and hyperparameter tuning were performed using Optuna. The second-best performing model was the ensemble model, which achieved a RMSLE of 0.05879. The AutoGluon model, with a RMSLE of 0.05846, is my best performing model and ties for **4th place out of 4318 teams** in the Kaggle competition. Moreover, my AutoGluon model is by far the simplest of all top performing teams on Kaggle.

**Conclusions –** To answer the question “What metrics are predictive of calories burned in a workout?”, we examine the feature importance and SHAP values of each top performing model. Duration and heart rate are the most important features when predicting calorie expenditure, which, intuitively, makes perfect sense.