**Predicting Calorie Expenditure**

Elliot Blackstone

**Github** – <https://github.com/ElliotBlackstone/S25_Predict_Calories>

**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 –** Our preprocessing is comprised of the following steps:

* Log transform Calories
* One-hot encoding for gender and recast as categorical type
* Delete id column
* Change height, weight, duration, heart rate from float to int

Notably, in the last step, only 1 value of heart rate and height was non-integer, so these were rounded to the nearest integer. These preprocessing steps reduced file size by roughly 10.5%.

**Stakeholders –** Telemetry device manufacturers, such as FitBit or smartwatch manufacturers, and exercise scientists.

**Model Selection and Results –** Models were judged based on their root mean squared log error (RMSLE) with the testing set. 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 lastly 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 the 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.