**Predicting Calories Burned from Gym Member Data**

**Research Question:** Can we accurately predict the number of calories burned during a workout session based on a gym member's physical attributes, workout details, and fitness experience?

**Data Source**

The dataset will be sourced from [Kaggle’s Gym Members Exercise Dataset](https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset/data), which includes 973 records with variables related to demographics, physiological metrics, workout types, and session data.

**Process:**

**Data Loading and Initial Checks:** The exercise tracking data was successfully loaded. Initial checks revealed no missing data in the target variable, although other columns might have had NaNs which were handled later.

**Feature Engineering:** New features like BMI, BPM\_Difference, and BPM\_Increase\_from\_Rest were created. Categorical features 'Exercise\_Type' and 'Gender' were processed ('Exercise\_Type' one-hot encoded, 'Gender' numerically encoded), and a 'BMI\_Category' was derived.

**Exploratory Data Analysis (EDA):** Visualizations showed the distributions of numerical features, highlighted gender-based differences in metrics like weight, height, session duration, calories burned, and heart rates. Outliers were identified, particularly in 'Calories\_Burned', and removed using a gender-specific IQR method. Correlation analysis revealed strong positive relationships between 'Calories\_Burned' and features like 'Session\_Duration', 'Weight', and heart rate metrics.

**Modelling:** Several regression models (Linear Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM) were trained and evaluated using user-specified features including Session Duration, Avg BPM, Weight, Water Intake, Max BPM, and Gender, along with one-hot encoded categorical features.

**Hyperparameter Tuning:** GridSearchCV and RandomizedSearchCV were applied to Random Forest and XGBoost models to find optimal parameters.

**Model Performance:** Comparing the baseline models and tuned models on the test set showed that tree-based models generally outperformed Linear Regression. Among the tuned models, one of the tuned ensemble models (likely Random Forest or XGBoost, depending on the specific tuning results) achieved the best performance, exhibiting the lowest RMSE and highest R-squared value on the test set. This indicates that the best model is most accurate in predicting calories burned based on the selected features and effectively explains the variance in the target variable.

**Summary of Findings:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE (Test)** | **RMSE (Test)** | **R-squared (Test)** |
| **XGBoost** | **361.2629** | **19.0069** | **0.9957** |
| XGBoost (GridSearchCV) | 563.8088 | 23.7447 | 0.9932 |
| LightGBM | 584.2029 | 24.1703 | 0.9930 |
| XGBoost (RandomizedSearchCV) | 658.6306 | 25.6638 | 0.9921 |
| Linear Regression | 1570.1159 | 39.6247 | 0.9812 |
| Random Forest (GridSearchCV) | 1835.2196 | 42.8395 | 0.9780 |
| Random Forest | 1836.4584 | 42.8539 | 0.9780 |
| Random Forest (RandomizedSearchCV) | 1850.2728 | 43.0148 | 0.9778 |

Based on the test set performance, the best performing model is: **XGBoost**

**Reasoning:**

The XGBoost achieved the lowest Root Mean Squared Error (RMSE) and the highest R-squared value on the test set

- RMSE measures the average magnitude of the errors. A lower RMSE indicates that the model's predictions are, on average, closer to the actual 'Calories\_Burned' values.

- R-squared (0.9957) represents the proportion of the variance in the dependent variable ('Calories\_Burned') that is

predictable from the independent variables (the features used). A higher R-squared indicates a better fit of the model to the data.

Therefore, the XGBoost is performing best because it minimizes the prediction error (lowest RMSE) and explains

the largest proportion of the variance in calories burned (highest R-squared) on unseen data.

Link to Jupyter Notebook: [Click here](https://github.com/nabiharaza/Predicting-Calories-Burned-Capstone_UCBerkeley/blob/main/main_notebook.ipynb)