Calorie Burn Prediction

SOHINI MANDAL

Introduction

Estimating caloric expenditure during physical activity is an important problem in fitness analytics and exercise monitoring. Direct measurement of calories burned typically requires specialized metabolic equipment and controlled laboratory environments, which limits its feasibility for routine or large-scale use. Consequently, caloric expenditure is commonly estimated indirectly using demographic characteristics, anthropometric measures, and exercise-related indicators.

The objective of this study is to develop and compare supervised regression models for predicting calories burned during structured exercise sessions. The focus is on understanding the structure of the data, evaluating linear and nonlinear modeling approaches, validating predictive performance rigorously, and interpreting results critically in light of dataset limitations. The study also demonstrates practical usability through deployment of a prediction application.

Data in Hand

The dataset used in this study is publicly available on Kaggle (Calories.csv) and consists of 15,000 observations with nine variables describing individual exercise sessions. The variables include user id (which was later dropped), gender, age, height, weight, exercise duration, heart rate, body temperature, and the target variable representing calories burned.

The dataset contains one categorical variable (Gender) and multiple numerical variables. Most physiological and exercise-related features are recorded as integers, indicating discretized or rounded measurements, while body temperature is recorded as a continuous variable. No missing values were present in the dataset.

The dataset source does not provide explicit documentation regarding measurement units or data collection procedures. Variable interpretation therefore relies on physiologically plausible ranges commonly reported in exercise science literature, indicating that the dataset is primarily suitable for methodological demonstration rather than real-world physiological inference.

The data was divided into training and test subsets using an 80–20 split, resulting in 12,000 training observations and 3,000 test observations. The training data was verified to contain no duplicate records. Exploratory analysis and model development were conducted using the training data, while the test data was reserved for final evaluation.

Exploratory Data Analysis

After performing the train–test split, all exploratory data analysis was carried out using the training dataset. The target variable, calories burned, exhibits a strongly right-skewed distribution. Most observations fall below 150 calories, with a prominent mode between approximately 25 and 50 calories and a long upper tail extending beyond 300 calories. This pattern reflects substantial heterogeneity in exercise intensity and session duration.

Exercise duration and heart rate display moderate variability consistent with light-to-moderate activity levels. Weight shows slight positive skewness, while height follows an approximately normal distribution with no prominent outliers. Age exhibits positive skewness, with most individuals concentrated in young to middle adulthood. Gender is nearly evenly distributed, indicating no significant class imbalance.

Body temperature displays unusually low variability around a high central value (approximately 40 °C). This pattern is physiologically implausible and suggests that the data may be constrained or synthetic in nature.

Pairwise analysis reveals strong positive relationships between exercise duration and calories burned, as well as between heart rate and calories burned. Height and weight are also strongly positively correlated. Several of these relationships are non-linear, motivating the inclusion of nonlinear regression models alongside linear approaches.

Methodology

Preprocessing steps included transforming the categorical gender variable using one-hot encoding with a reference category dropped and standardizing numerical features to ensure comparability across models.

Multiple supervised regression models were trained and compared, including linear regression, ridge regression, lasso regression, random forest regression, gradient boosting regression, XGBoost regression, and support vector regression (SVR). Hyperparameter tuning was conducted using grid-based search combined with five-fold cross-validation. This approach reduces dependence on a single data split and provides a reliable estimate of model generalization performance while maintaining reasonable computational cost.

Model performance was evaluated using the coefficient of determination (R^2) as the primary metric, supplemented by RMSE and MAE to assess prediction error magnitude.

Results and Model Comparison

Linear models achieved strong performance, with R^2 values close to 0.97, indicating that a substantial proportion of the variation in caloric expenditure is explained by linear relationships among the predictors. However, nonlinear models consistently outperformed linear approaches.

Among all evaluated models, support vector regression demonstrated the highest predictive accuracy. The SVR model achieved near-perfect performance on the test data, with an R^2 of 0.9996 and very low RMSE and MAE values. Training and test performance were closely aligned, indicating minimal overfitting.

Critical Evaluation and Limitations

Despite excellent predictive performance, several limitations must be acknowledged. The constrained distributions of certain variables—particularly body temperature—suggest that the dataset may encode simplified or synthetic relationships. In realistic physiological settings, greater variability and measurement noise would be expected.

Additionally, near-perfect predictive accuracy across multiple modeling approaches is uncommon in real-world exercise data, indicating that the dataset likely exhibits strong deterministic structure. Therefore, the results should be interpreted as a demonstration of modeling methodology rather than evidence of real-world predictive reliability.

Conclusion

This study demonstrates the application of supervised regression techniques for predicting caloric expenditure using demographic, anthropometric, and exercise-related variables. Through systematic exploratory analysis, rigorous cross-validation, and model comparison, nonlinear models—particularly support vector regression—were shown to achieve superior predictive performance on the given dataset.

The primary contribution of this work lies in illustrating a complete analytical pipeline, from data exploration and model evaluation to deployment through a Streamlit-based prediction application. While the results highlight the effectiveness of advanced regression models on structured data, they also emphasize the importance of critical interpretation when evaluating predictive performance.
