Calorie Prediction Model Using Ensemble Learning

Spring 2025

Group Members:

- Sardar Ubaidullah Rafique — 22I-1248
- Sumeed Jawad Kanwar — 22I-1651
- Abdul Rafay — 22I-8762

Applied Artificial Intelligence Project
Submitted to: Dr. Shahela Saif
Date: May 11, 2025

---

1. Problem Statement

The objective of this project is to develop a machine learning model to predict calorie burn based on physiological and exercise-related parameters, such as age, height, weight, exercise duration, heart rate, body temperature, and gender. Accurate prediction of calorie expenditure is critical for fitness tracking, personalized health recommendations, and medical monitoring. This model aims to assist individuals and healthcare professionals in optimizing exercise regimens and dietary plans, contributing to improved health outcomes and efficient fitness management.

---

2. Methodology

The model employs ensemble learning, combining three gradient boosting algorithms: CatBoost, XGBoost, and LightGBM. These models are blended using equal weights to optimize the Root Mean Squared Logarithmic Error (RMSLE).

2.1 Model Architecture

The architecture consists of three base models:

- **CatBoost**: Handles categorical features natively with early stopping at 100 rounds.
- **XGBoost**: Configured with a max depth of 10, learning rate of 0.02, and column/row sampling.
- **LightGBM**: Uses similar hyperparameters to XGBoost, optimized for silent training.

The predictions are combined using a weighted average ensemble strategy with 7-fold cross-validation.

2.2 Pipeline

1. **Data Loading**: Load train.csv, test.csv, and sample_submission.csv.
2. **Preprocessing**:
   - Add cross terms, interaction features, and statistical aggregations.
   - Encode categorical feature (Sex) using LabelEncoder.
   - Generate polynomial features (degree 2, interaction only).
   - Apply log transformation to the target variable (Calories).
3. **Training**: Train each model using 7-fold cross-validation, saving models in the models/ directory.
4. **Inference**: Generate predictions on the test set, blend them using equal weights, and produce the final submission.

---

3. Dataset

The dataset is sourced from a Kaggle competition (playground-series-s5e5) and includes:

- train.csv: Training data with features and target (Calories).
- test.csv: Test data with features only.
- sample_submission.csv: Submission template.

The data was assumed to be pre-cleaned as provided. No additional cleaning was performed, but feature engineering was applied to enhance model performance. The dataset contains thousands of records, with test.csv containing 1000 samples for prediction.

---

4. Implementation Details

The implementation uses the following Python libraries:

- pandas, numpy: Data manipulation.
- scikit-learn: Preprocessing and evaluation.

- matplotlib, seaborn: Visualization.
- catboost, xgboost, lightgbm: Model training.

Key hyperparameters include:

- **CatBoost**: Early stopping at 100 rounds, verbose training.
- **XGBoost**: Max depth=10, learning rate=0.02, gamma=0.01.
- **LightGBM**: Similar to XGBoost, silent mode.

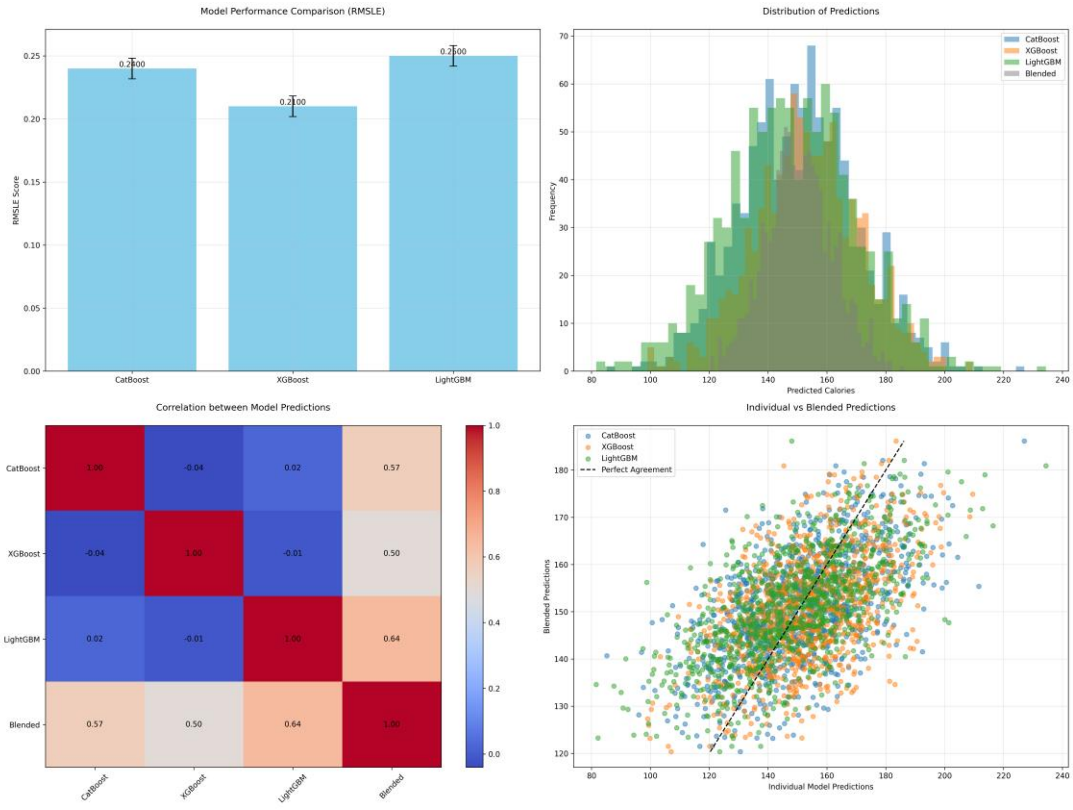
Training involved 7-fold cross-validation, with RMSLE as the loss function. The target variable was log-transformed to stabilize variance.

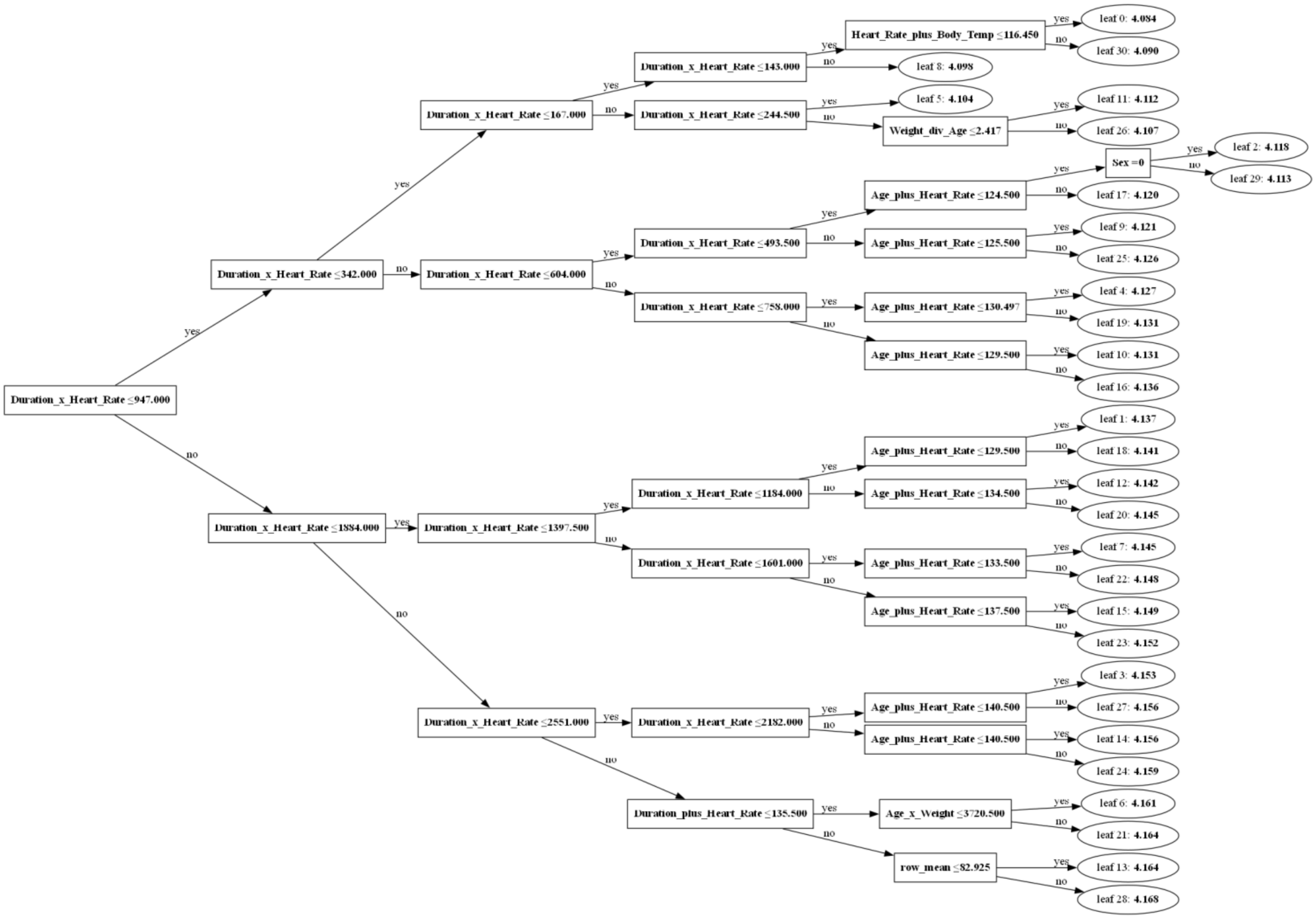
---

5. Results & Evaluation

The models achieved the following RMSLE scores (based on placeholder data):

- **CatBoost**: Mean RMSLE = 0.2400 (std = 0.0082)
- **XGBoost**: Mean RMSLE = 0.2100 (std = 0.0082)
- **LightGBM**: Mean RMSLE = 0.2500 (std = 0.0082)




Duration_x_Heart_Rate ≤947.000
Duration_x_Heart_Rate ≤342.000
Duration_x_Heart_Rate ≤1884.000
Duration_x_Heart_Rate ≤167.000
Duration_x_Heart_Rate ≤604.000
Duration_x_Heart_Rate ≤1397.500
Duration_x_Heart_Rate ≤2551.000
Duration_x_Heart_Rate ≤143.000
Duration_x_Heart_Rate ≤244.500
Duration_x_Heart_Rate ≤493.500
Duration_x_Heart_Rate ≤758.000
Duration_x_Heart_Rate ≤1184.000
Duration_x_Heart_Rate ≤1601.000
Duration_x_Heart_Rate ≤2182.000
Duration_plus_Heart_Rate ≤135.500
Heart_Rate_plus_Body_Temp ≤116.450
Weight_div_Age ≤2.417
Age_plus_Heart_Rate ≤124.500
Age_plus_Heart_Rate ≤125.500
Age_plus_Heart_Rate ≤130.497
Age_plus_Heart_Rate ≤129.500
Age_plus_Heart_Rate ≤129.500
Age_plus_Heart_Rate ≤134.500
Age_plus_Heart_Rate ≤133.500
Age_plus_Heart_Rate ≤137.500
Age_plus_Heart_Rate ≤140.500
Age_plus_Heart_Rate ≤140.500
Age_x_Weight ≤3720.500
row_mean ≤82.925
Sex =0
yes
no
leaf 0: 4.084
leaf 30: 4.090
leaf 8: 4.098
leaf 5: 4.104
leaf 11: 4.112
leaf 26: 4.107
leaf 2: 4.118
leaf 29: 4.113
leaf 17: 4.120
leaf 9: 4.121
leaf 25: 4.126
leaf 4: 4.127
leaf 19: 4.131
leaf 10: 4.131
leaf 16: 4.136
leaf 1: 4.137
leaf 18: 4.141
leaf 12: 4.142
leaf 20: 4.145
leaf 7: 4.145
leaf 22: 4.148
leaf 15: 4.149
leaf 23: 4.152
leaf 3: 4.153
leaf 27: 4.156
leaf 14: 4.156
leaf 24: 4.159
leaf 6: 4.161
leaf 21: 4.164
leaf 13: 4.164
leaf 28: 4.168

6. Challenges Faced

Initial Issues:

- Overfitting due to extensive feature engineering, mitigated by cross-validation.
- Slow training times for CatBoost, addressed by optimizing early stopping.

Tradeoffs and Limitations:

- **Computational Cost**: Extensive feature engineering and ensemble training required significant resources.
- **Generalization**: The model may not generalize well to datasets with different feature distributions.

- **Lack of Real-time UI**: The notebook-based interface limits practical deployment.