

Predicting Pokémon Battle Outcomes Using Timeline Features

Sapienza University of Rome – Master’s in Data Science
Course: Fundamentals of Data Science

Gaitanaru Gabriel, Montuori Tommaso, Autieri Marco

11/11/2025

1. Overview

This project develops a model for the Kaggle competition of *Fundamentals of Data Science*, predicting whether Player 1 wins a Pokémon battle. The dataset, provided in JSONL format, includes detailed *timelines* of up to 30 turns per match, describing HP percentages, moves, and status conditions.

2. Methodology

Only official data are used: `train.jsonl` for training and `test.jsonl` for prediction. No external knowledge, type charts, or additional datasets are employed. All features are extracted directly from the timeline, including: average/min HP, number of moves, base power means, inflicted status counts, KOs, cumulative damage and recovery, and early HP differences (first 15–30 turns). A synthetic **momentum index** summarizes early advantage and move activity. Missing values are filled with zero, and constant features are removed.

3. Model and Evaluation

The model uses either **XGBoost** (preferred) or **Random Forest** when unavailable. Training uses 5-fold stratified cross-validation with fixed seed (42) for reproducibility. Average validation accuracy is about **0.81**, showing robust predictive power without overfitting. After validation, the model is retrained on the full dataset and applied to the test set.

4. Submission and Compliance

Predictions are saved to `submission.csv` with columns `battle_id` and `player_won` (`true/false`). The code is fully deterministic, offline, and rule-compliant:

- Only data provided by the organizers are used;
- No type-based or manual knowledge is introduced;
- Same results are obtained when rerun.

This implementation represents the final, official submission for evaluation.