Machine Learning Report: Data Cleaning and RSI Prediction

Urjit Mehta, Krina Khakhariya, Brijesh Munjiyasara

Ahmedabad University

Team: Trinity (Github link)

WEEK NUMBER: 3

I. INTRODUCTION

The focus of this week was on refining the Athlete Readiness dataset and training machine learning models to predict the RSI (Readiness Score Index). This involved cleaning missing data, understanding feature relationships, and evaluating model performance.

II. DATA PREPROCESSING

A. Dataset Overview

The dataset contains 3111 records with 28 attributes covering athlete performance, sleep quality, and recovery.

B. Handling Missing Data

- About 1130 entries had missing values.
- Used Mean, Median, and Mode imputation methods.
- Used interpolation method with forward and backward filling.
- Implemented K-Nearest Neighbors (KNN) Imputation for numerical values.
- Categorical values were imputed using Mode.

C. Feature Engineering

- Computed statistical insights (mean, median, mode) for key attributes.
- Visualized feature distributions to check data patterns.
- Identified the strongest correlating features for RSI prediction.

III. Exploratory Data Analysis

- **Key Findings:**
  - Sleep consistency showed a positive impact on RSI.
  - Unexpected negative correlation between Sleep Score and RSI.
  - Light sleep had a negative effect on RSI readiness.
- **Visualizations:**
  - Created heatmaps and scatter plots.
  - Analyzed key features like Sleep Score and REM Sleep.

IV. Machine Learning Model Training

A. Models Evaluated

- Linear Regression (Baseline).
- Ridge and Lasso Regression.
- Random Forest Regressor.
- Gradient Boosting (XGBoost, LightGBM).

B. Model Performance

- Linear Regression:R^2: 0.1663, MAE: 0.0204
- Ridge Regression:R^2: 0.1645, MAE: 0.0204
- Lasso Regression:R^2: 0.0297, MAE: 0.0215
- Random Forest:R^2: 0.4761, MAE: 0.0139
- XGBoost:R^2: 0.4887, MAE: 0.0144
- LightGBM:R^2: 0.4988, MAE: 0.0141

V. Next Steps

- If we get more data then we can explore deep learning models for better prediction.
- Optimize hyperparameters for boosting models.
- Apply additional feature engineering techniques.

VI. Conclusion

This week, we successfully cleaned the dataset and built baseline models. While initial results show room for improvement, the next steps involve fine-tuning and testing more advanced models.