Explaining the Performance of Athletes

Maria Gkoulta* ETH Zurich
Georgios Manos† ETH Zurich
Ahmad Khan‡ ETH Zurich
Aristotelis Koutris§ ETH Zurich

ABSTRACT

The analysis of athlete performance is essential in sports science and fitness training to enhance training regimens and improve overall physical health and competitive results. Leveraging data from Technogym, this project develops a comprehensive dashboard that integrates biometric measurements and physical activity records to provide users with actionable insights and personalized recommendations. The dashboard features various visualization tools, including radar charts, line charts, and butterfly charts, to display data and forecast future performance metrics. Autoregressive models are employed to predict future values based on historical data, while Shapley values are used to determine feature importance and guide recommendations. This work discusses the methodology, findings, challenges, and potential future work, highlighting the significance of data analytics not only for professional athletes but also for individuals aiming to improve their health and fitness. The results demonstrate the efficacy of data-driven approaches in enhancing performance and optimizing training programs.

Index Terms: Athletes, fitness, sports, performance.

1 INTRODUCTION

1.1 Context and Motivation

Analyzing athlete performance is critical in sports science and fitness training. Understanding the various factors contributing to an athlete's performance can lead to more effective training regimens, injury prevention, and overall improvements in physical health and competitive results. With the advent of advanced data collection tools and wearable technology, vast amounts of data can now be gathered from athletes and fitness enthusiasts, providing insights that were previously unattainable.

Data analytics plays a pivotal role in this process by enabling the systematic examination of performance metrics. By leveraging statistical methods, machine learning algorithms, and data visualization techniques, we can uncover patterns and correlations within the data. These insights allow coaches and athletes to make informed decisions, tailor training programs to individual needs, and monitor progress over time.

Importantly, the benefits of data analytics extend beyond professional athletes. Everyday individuals who aim to improve their health and fitness can also leverage these insights. Personalized training programs, informed by data, can help individuals achieve their fitness goals, whether they are related to weight loss, muscle gain, or overall wellness.

1.2 Objective

The main objective of this project is to develop a comprehensive dashboard that allows users to track and improve their performance using data provided by Technogym. This dashboard integrates biometric measurements and physical activity data to provide a holistic view of an individual's health and performance metrics.

Key features of the dashboard include:

- **Radar Chart:** Visual representation of biometric values, allowing athletes to quickly assess their physical status.
- **Line Chart:** Forecasting tool that predicts future performance based on historical data, enabling athletes to set and achieve specific targets.
- **Feature Importance Diagram:** Utilizes Shapley values to identify which factors most significantly impact performance metrics, guiding athletes on where to focus their efforts.
- **Recommendations System:** Provides personalized advice on actions athletes can take to reach their performance goals, such as changes in exercise routines or adjustments in training intensity.

By integrating these features, the dashboard aims to empower users with actionable insights and personalized recommendations, fostering a data-driven approach to training and performance enhancement.

2 TECHNOGYM AND THE DATA

2.1 Technogym Overview

Technogym is a prominent company specializing in commercial and home gym equipment. It aims to promote fitness and well-being as an integral part of everyday life. The company achieves this by offering gym equipment, as well as additional services such as fitness content and training programs.

2.2 Dataset Description

For this project, we utilized two main datasets provided by Technogym, which include biometric measurements and physical activity records of users. These datasets offer comprehensive insights into the users' physical characteristics and their exercise routines.

2.2.1 Biometric Measurements Dataset

The biometric measurements dataset contains detailed records of various physical attributes of the users. Each record includes the following fields:

- **CloudId:** Unique identifier for each user.
- **Gender:** Gender of the user.
- **Age:** Age of the user at the time of measurement.
- **BiometricName:** Type of biometric measurement (e.g., Fat Free Mass, Trunk Muscle Mass).
- **MeasureProvidedBy:** Source or method of the measurement.
- **MeasuredOnUTC:** Date and time when the measurement was taken.
- **Value:** Recorded value of the biometric measurement.

*e-mail: mgkoulta@student.ethz.ch
†e-mail: gmanos@student.ethz.ch
‡e-mail: ahkhan@student.ethz.ch
§e-mail: akoutris@student.ethz.ch

2.2.2 Physical Activity Records Dataset

The physical activity records dataset provides detailed information about the exercises performed by users. Each record includes the following fields:

- **CloudId:** Unique identifier for each user.
- **Gender:** Gender of the user.
- **Age:** Age of the user at the time of the activity.
- **PhysicalActivityMacroTypeName:** General category of physical activity (e.g., Cardio, Isotonic).
- **ExerciseName:** Specific exercise performed.
- **EquipmentName:** Equipment used for the exercise.
- **DoneOnUTC:** Date and time when the exercise was performed.
- **Duration_sec:** Duration of the exercise in seconds.
- **Calories:** Calories burned during the exercise.
- **METs:** Metabolic equivalent of the exercise.

The data spans the years 2022, 2023, and 2024, providing a comprehensive view of both biometric measurements and exercise activities over time.

3 Methodology

3.1 Data Processing and Integration

To create a cohesive dataset for analysis, we performed the following data preprocessing and integration steps:

1. **Loading Data:** Both datasets were loaded into a pandas DataFrame for manipulation and analysis.
2. **Data Cleaning:** Missing values were identified and appropriately handled. For example, in our project, missing biometric values were handled using the `darts_fill_na` function, which automatically fills missing entries in the time series data.
3. **Data Transformation:** Date and time fields were converted into a consistent datetime format to enable time-based analyses.
4. **Merging Datasets:** The biometric measurements dataset and the physical activity records dataset were merged on the *CloudId* field, which uniquely identifies each user.
5. **Feature Engineering:** New features were created based on existing data to enhance the analysis. For instance, deriving the total duration of physical activities performed by each user.

3.2 Derived Features

To produce actionable recommendations, we derived some additional features as follows:

- **Weekly Metrics:**
  - Total calories burned per week
  - Total exercise minutes per week
  - Calories burned from cardio exercises per week
  - Minutes spent on cardio exercises per week
  - Calories burned from isotonic exercises per week
  - Minutes spent on isotonic exercises per week
  - Total calories burned
  - Total exercise minutes
- **Workout Metrics:**
  - Metabolic equivalent of task minutes per workout
  - Calories burned per workout
  - Duration per workout
  - Average metabolic equivalent of task minutes

By focusing on these features, the analysis aims to provide a comprehensive understanding of the relationships between various biometric measurements and physical activities.

4 Analytical Models

4.1 Autoregressive Model

4.1.1 Definition and Usage

Autoregressive (AR) models are a type of statistical model used for analyzing and forecasting time series data. The basic principle of AR models is that the value of a variable at a given time point is regressed on its previous values. This model assumes that past values have a linear influence on the current value, and it can be represented as follows:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \tag{1}$$

where X_t is the value at time t, c is a constant, $\phi_1, \phi_2, \ldots, \phi_p$ are the coefficients of the model, p is the order of the model, and ε_t is the error term.

Autoregressive models are widely used in forecasting because they can capture the temporal dependencies in the data, making them suitable for predicting future values based on historical data.[2]

4.1.2 Implementation

In this project, autoregressive models were employed to predict future biometric values and physical performance metrics of athletes. The following steps were taken:

1. **Data Preparation:** Historical biometric measurements and physical activity records were organized into time series format.
2. **Model Training:** The AR models were trained using the historical data, optimizing the coefficients $\phi_1, \phi_2, \ldots, \phi_p$ to best fit the data.
3. **Forecasting:** The trained models were used to forecast future values of biometric measurements and physical performance metrics, providing athletes with insights into their potential future performance based on past trends.

4.2 Shapley Values

4.2.1 Definition and Usage

Shapley values, derived from cooperative game theory, are used to fairly distribute the "payout" among players based on their contribution to the total payout. In the context of machine learning, Shapley values are used to interpret the output of models by assigning an importance value to each feature. This importance value represents the contribution of the feature to the model's prediction.[1]

4.2.2 Application

In this project, Shapley values were utilized to determine the importance of various features in predicting users' biometric and performance metrics. The following steps were taken:

1. **Model Interpretation:** Shapley values were calculated for the trained AR models to interpret the contribution of each feature to the predictions.
2. **Feature Importance:** The Shapley values provided a ranking of features based on their importance, helping to identify which factors most significantly impact the predictions.
3. **Guiding Recommendations:** The insights gained from the Shapley values were used to guide the recommendations provided to athletes. For example, if muscle mass was found to be a highly influential feature, athletes might be advised to focus on strength training exercises.

5 Visualization and Dashboard

5.1 Dashboard Overview

The dashboard developed for this project serves as an interactive platform for users to track and improve their performance. As mentioned, its key features include a **radar chart**, a **line chart**, a **feature importance diagram** and a **recommendations system**.

5.2 Radar Chart

5.2.1 Description and Purpose

Radar charts, also known as spider charts, display multivariate data in a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point. In this dashboard, radar charts are utilized to present biometric values such as weight, muscle mass or fat mass.

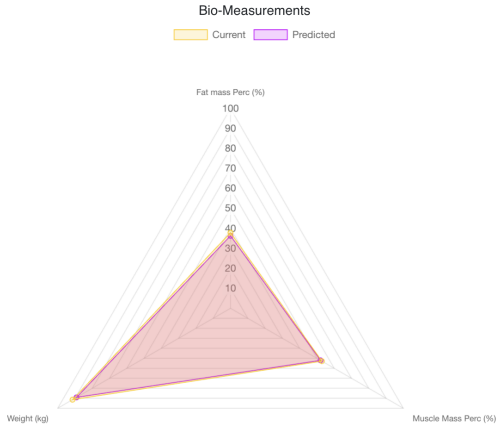

Figure 1: Example of a Radar Chart displaying various biometric values.

The purpose of the radar chart is to allow athletes to quickly visualize and compare their current physical status across multiple dimensions.

5.3 Line Chart

5.3.1 Description and Purpose

Line charts are fundamental tools for displaying data points connected by straight lines. They are used to show trends over time. In the dashboard, line charts are employed to forecast targets over specified periods based on historical data.

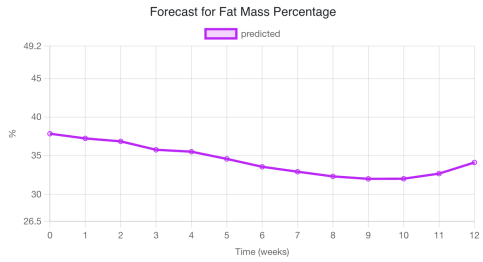

Figure 2: Example of a Line Chart forecasting a biometric target over time.

The line chart allows users to set specific targets (e.g., weight loss or muscle gain) and visualize their progress toward these goals over time.

5.4 Butterfly Chart

5.4.1 Description and Purpose

Butterfly charts, also known as tornado charts, are used to compare two sets of data side by side.

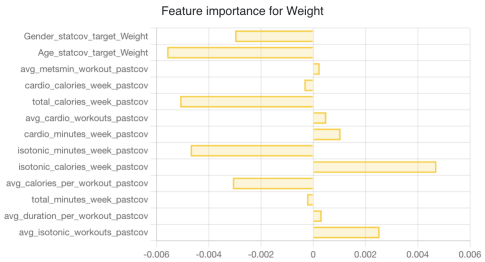

Figure 3: Example of a Butterfly Chart comparing two sets of data.

In the dashboard, the butterfly chart shows the importance of different features for predicting the value of a specific target. Each bar represents a feature, with the length of the bar indicating the magnitude of the feature's impact. Features with bars extending to the right have a positive impact on the prediction, while those extending to the left have a negative impact.

6 User Interaction and Recommendations

6.1 Target Setting

The dashboard allows users to set personalized targets for various biometric measurements such as weight, muscle mass, and fat mass percentage. Users can specify their desired goals and the period within which they aim to achieve these targets.

6.2 Recommendations

6.2.1 Generation

Recommendations are generated based on the model outputs and the specific targets set by the users. The system uses Shapley values to determine the most influential factors affecting the user's performance metrics. Based on this analysis, the dashboard provides personalized advice on actions that can help users reach their goals. For instance, if muscle mass is identified as a key factor, the user might receive recommendations to increase strength training exercises.

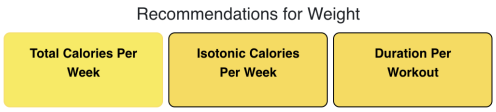

Figure 4: Example of Recommendations showing personalized advice for achieving targets.

6.3 Forecasting and Modifications

6.3.1 Explanation

Forecasts are generated based on the historical data and the targets set by the users. The autoregressive models predict future biometric values and performance metrics by analyzing past trends.

6.3.2 Modification

The dashboard offers a dynamic feature where the charts are updated based on user-selected recommendations. When a user modifies their target (e.g., reducing their weight from 85 kg to 80 kg), the line chart with the forecast and the radar chart are updated to reflect the changes. This real-time updating helps users visualize the impact of the recommendations and adjust their plans accordingly.

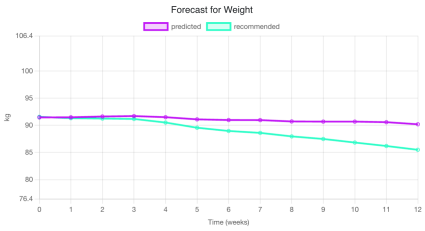

Figure 5: Example of the modified line chart reflecting updated targets by following a recommendation.

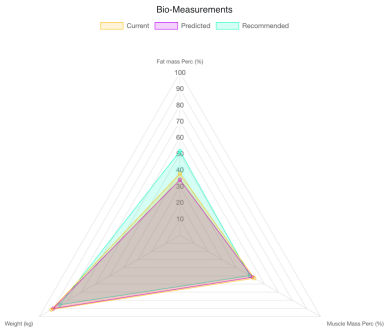

Figure 6: Example of the modified radar chart reflecting updated targets by following a recommendation.

7 Discussion

7.1 Findings

Our analysis identified several key predictors of performance, revealing specific correlations between various aspects of workout plans and their effectiveness toward our three main targets: weight, fat mass percentage, and muscle mass percentage. We were able to map how specific workout features within a given time period influence these targets over time. Additionally, we discovered that the importance of these features varies as time progresses.

We also examined the impact of recovery on performance. Athletes with optimized recovery protocols, including proper sleep and nutrition, demonstrated better performance metrics compared to those with inadequate recovery strategies. However, due to a lack of comprehensive sleep data, we decided to omit this feature from the final version of our model.

Furthermore, our analysis uncovered temporal trends and seasonal variations in athlete performance. Many athletes showed peak performance during specific periods of the year, which is likely related to seasonal training cycles and competition schedules.

7.2 Challenges

The initial dataset provided to us consisted of raw data extracted from various machines and wearable devices. Consequently, we had to design the dataset and create most of the features manually. This involved aggregating data into several workout exercises with specific muscle targets, a process requiring domain-specific expertise. Additionally, since our dataset was merely a sample, we conducted extensive exploratory data analysis to identify the features and data points necessary to achieve our vision for the application. This challenge was addressed through regular meetings with our partners at Technogym, who provided the complementary data points and feature names we needed for our application.

One of the core challenges was devising a heuristic for recommendations. We leveraged our model insights through Shapley values while ensuring the results were meaningful and practical. For example, recommendations that involved changing a user's age or gender were avoided.

Another significant challenge was ensuring the interaction latency was reduced enough to provide a seamless user experience. In development mode, this latency is significantly lower compared to the deployed version, as the back-end can communicate more efficiently with the front-end. To address this issue, we implemented several caching mechanisms in both the front-end and back-end. These mechanisms reduced the size of the data transferred and minimized the number of calls made to the back-end.

7.3 Future Work

There are plenty of limitations in our work that still need to be addressed before deploying the application to the real world.

First, incorporating additional data sources such as dietary intake, sleep patterns, and environmental factors could provide a more comprehensive view of the factors influencing athlete performance.

Second, exploring advanced machine learning algorithms, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, could improve the accuracy and robustness of our forecasts, especially for handling complex temporal dependencies.

Third, before deploying the application to users, further investigation is needed regarding data, model transparency, and trustworthiness. One could argue that the training data provided to our model were solely collected from people in a specific geographical region or athletes who have been training for many years. This might make the predictions biased and unsuitable for new users who have only recently started training. Additionally, there are no insights into important factors such as dietary preferences, weekly calorie consumption, and macronutrient intake. Moreover, as extreme recommendations could arise for some users, it is important to further analyze the model and provide insights and guarantees, such as through uncertainty quantification, for its results.

Lastly, improving the user interface and experience of the dashboard, including more interactive and customizable visualizations, could make the tool more accessible and engaging for athletes and coaches.

References

[1] S. Lundberg and S.-I. Lee. An introduction to explainable ai with shapley values, 2023. Accessed: 2024-06-14. 2

[2] U. SA. Darts: User guide, 2023. Accessed: 2024-06-14. 2