Summary of 'A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives'

Kahurangi Burkitt, Annie Foote, Abia John and Paige Martin

Introduction

This study used supervised machine learning to holistically predict basketball performance at three levels. Individual players were assessed based on their readiness and fatigue, team performance was evaluated through each player's game statistics, and at the conference level, performance was measured by analysing player contribution per minute. Unlike existing literature that focussed on either individual athletes or teams, this study integrated data sources collected over a full Division 1 Women's basketball season. The primary aim of the study was to help coaches to monitor athlete readiness, optimise training, and make informed decisions by combining machine learning outputs with coaching expertise. This multi-level, data-driven approach to coaching supports performance prediction, injury risk assessment and strategic planning.

Data Collection

To provide a holistic analysis and to attempt to predict Player Readiness (RSI), Game Score (GS), and Player Efficiency Rating (PER), a multi-faceted dataset was collected with both factual and subjective measures in mind. Objective data included detailed training workload information which tracked the volume and intensity of practices, conditioning, and strength training sessions. Biometric data was gathered non-invasively using WHOOP straps, which monitored sleep patterns, recovery metrics, heart rate, and heart rate variability. During competition, in-game performance was quantified using Polar Team Pro monitors which captured real-time statistics such as player speed, distance covered, and acceleration. Furthermore, athlete explosiveness and readiness were assessed weekly through counter-movement jump tests performed on force plates.

Subjective data was also collected from questionnaires looking at results related to stress and recovery. This provided insights into the athletes' psychological and emotional states which are known to influence physical output.

Data cleaning

To address missing sleep and recovery entries in the questionnaire data, the study applied the Multiple Imputation by Chained Equations (MICE) technique. This method inputs missing values by conditionally modelling each missing feature with respect to the other observed features in the dataset.

Factor analysis and Feature importance

Factor analysis

Initially from the Polar Band data 40 features were identified under the following categories of:

- Heart rate
- Distance
- Speed zone
- Recovery time

- Acceleration

Factor analysis discovered latent factors for obtaining a lossless and compact feature representation, resulting in eight compact factors (see Fig. 1).

Factors	Feature
Speed and Total Acceleration Zones (F0)	Number of accelerations: 2.99–2.00 (m/s^2)
	Number of accelerations: 1.99–1.00 (m/s^2)
	Number of accelerations: 0.99–0.50 (m/s^2)
	Number of accelerations: 0.50–0.99 (m/s^2)
	Number of accelerations: 1.00–1.99 (m/s^2)
	Distance in Speed zone 1: 1.00–4.99 (km/h)
	Distance in Speed zone 2: 5.00–6.99 (km/h)
	Distance in Speed zone 3: 7.00–10.99 (km/h)
	Total distance (m)
Average Speed and Distance (F1)	Average speed (km/h), Distance (m/min), HR avg (bpm)
Average Speed and Acceleration Zone (F2)	Number of accelerations: 2.00–2.99 (m/s^2)
	Distance in Speed zone 4: 11.00–14.99 (km/h)
	Distance in Speed zone 5: 15.00 (km/h)
	Number of accelerations: 50.00–3.00 (m/s^2)
	Sprints
Minimum Heart Rate (F3)	HR min (bpm)
Maximum Heart Rate (F4)	HR max (bpm)
Recovery Time (F5)	Recovery time (h)
Maximum Speed (F6)	Maximum speed (km/h)
High Intensity Acceleration Zone (F7)	Number of accelerations: 3.00–50.00 (m/s^2)

Figure 1: Grouping of observed features into eight latent factors following factor analysis

Feature importance

Feature importance revealed the input variables that are most useful for predicting a target variable. The study used a combination of methods, outlined below, to predict these important features:

- **Random Forest (RF) and XGBoost (XGB):** Uses a Gini Index to measure how important a feature is.
- **Correlation (CORR):** Measures how strongly each feature is related to the target value, generating a value ranging between -1 to 1.
- **Nonlinear weighted averaging:** Combines the scores from RF, XGB and CORR using a custom weighting scheme to produce one final score per feature.

Player

RSI was best predicted using training, sleep, recovery, and stress data, with the top features being Training Strain, Resistance Training Volume Load, Total Weekly Load, Heart Rate Variability, and Mental Performance Capability. These top features were mostly related to the training modality, which showed that training data gives us the strongest predictors of player performance.

Team

Game Score, a team-level key performance indicator (KPI), was predicted using data from reactive strength, in-game statistics, sleep, recovery, stress, and training modalities. The top features mostly came from the in-game modality, showing that these in-game metrics are the strongest predictors of team performance.

Conference

PER, a conference-level KPI, was predicted using data from the reactive strength, in-game statistics, sleep, recovery, subjective stress, and training modalities. The top features were spread across reactive strength, in-game statistics, sleep, recovery, and stress, highlighting a balanced set of predictors with no single dominant modality.

Prediction

Player Level

The prediction model at the player level predicted an athlete's RSI value, an indicator of the player's fatigue caused by their training and competition load. This value was predicted using results of players' counter-movement jump tests and data on their sleep and recovery patterns, training workload and cognitive state as inputs. Data from the previous week was used to predict values for the current week.

Values were then split into quartile ranges, and the synthetic minority oversampling technique (SMOTE) was used to balance the dataset across the upper and lower quartile ranges. This allowed an XGB classifier model to be used effectively to predict the RSI values.

Team Level

At the team level, the model was trained to predict a player's GS statistic. This is a value that quantifies the impact a player has on the overall team's performance, factoring in both positive (points, assists, steals, etc.) and negative (fouls, missed shots, turnovers, etc.) contributions a player makes within a game. The model was trained to predict this value using data gathered from players' sleep, training, questionnaire, and jump records, along with their previous in-game statistics measured by the Polar units.

K-means clustering was used to divide the GS data set into three clusters of bad, average, and good, and a combination of SMOTE and undersampling (ENN) techniques were used to balance the data across the three clusters. The data was then split into training and test sets at a 70:30 ratio and an XGB classifier was used to predict the GS values.

Conference Level

At the conference level, the model predicted players' PER statistics. This value can be used to measure a player's efficiency relative to the average of all other players in the conference. The model took sleep, recovery, training, subjective stress, reactive stress and in-game statistics as inputs to make the predictions. A 70:30 split of training and testing data was used for development of an XGB regressor. PER was able to be represented as a continuous variable output, so no clustering techniques were needed.

Results

- **Player Readiness (RSI) - 98.67% accuracy** - Using the previous week's training, sleep and stress data, the model could predict which performance category (high readiness to low readiness) an athlete would be in for the next week with very high accuracy.
- **Game Score (GS) - 94.20% accuracy** - The model was also highly successful at predicting how well a player would perform in a specific game (categorised as "bad," "average," or "good").
- **Player Efficiency Rating (PER) - Mean Squared Error (MSE) of 0.026 and R^2 of 0.68**
  - *MSE*: The MSE value was extremely low (perfect score is 0), which means the model's predictions for a player's season-long efficiency rating were, on average, extremely close to their actual rating.
  - R^2: An R^2 of 0.68 (where 1 is a perfect fit) represents the variation in the outcome that can be explained by the input data collected.

Strengths and Limitations

Strengths

- Access to full-team data across different training periods across a season enabled detailed analysis.

- Internal and external training load metrics allowed athlete's responses to training and competitive stress to be evaluated.
- Multi-level analysis provided insights at a player, team and conference level.

Limitations

- The current study did not generalise across multiple sports limiting the scope of the model.
- The model also only encompassed one year of data. More data would be needed for the model to work across seasons.
- Analysis at each level was also based on a single metric, limiting the representation of performance at each level.

Our takeaways

Our analysis identified several limitations of the study. A limitation that we identified is the lack of consideration for the human and situational elements that define a basketball game. Factors such as court vision, on-court team chemistry, the unpredictable nature of emotions during a game, and even "luck", were not captured by the model. These factors can significantly influence performance outcomes.

The model also didn't consider the opposition and their players' attributes, or situational moments within a game, when making predictions. These considerations would still be up to the coaching staff. Based on this, it seems it may be more likely that the results of the predictive modeling are used to reinforce existing coaching decisions, rather than to drive a new tactical direction within the team.

Future work

While this study identified KPIs, future work will test how well these KPIs improve athletic performance. Along with this, key metrics should also be quantified for other sports at various levels, to build a model that fits that sport's requirements. Future work also includes refining the model to offer interpretable and applicable insights to coaches and players, using familiar sports and science terminology.