PREDICTING COLLEGE BASKETBALL PERFORMANCE WITH SUPERVISED MACHINE LEARNING

AN ARTICLE BY TABER, C.B.,
SHARMA, S., RAVAL, M.S. *ET AL.*

SUMMARY

ISSUE

Coaches lack a holistic, data-driven system to predict player performance and manage **athlete readiness**. Existing methods focus on either **individuals or teams**, but not **both**.

GOALS

Use a **supervised machine learning model** to holistically predict basketball performance at the **Player**, **Team** and **Conference** levels.

METHOD

Use a **multi-level eXtreme Gradient Boosting (XGB)** classifier model to analyse a comprehensive dataset from a **full Division 1 Women's basketball season.**

DATA GATHERED

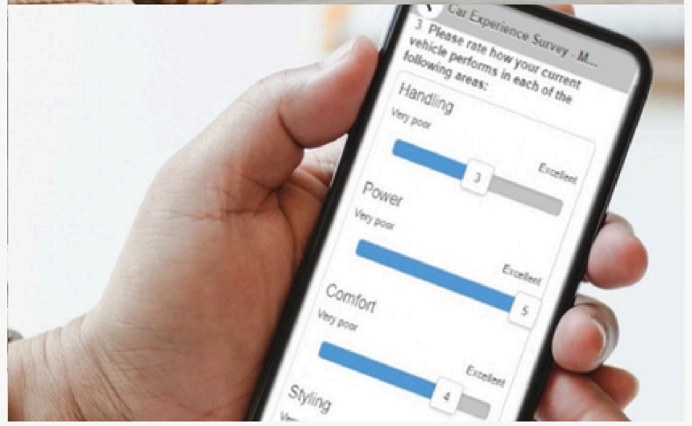
Mix of Objective, Subjective and Biometric data

Training Load

Volume and intensity of practices and strength training.

Physical Readiness

Countermovement jump tests to assess athlete explosiveness.



Questionnaires

Provides insights into athlete stress and recovery.

WHOOP Straps

Monitored sleep patterns, recovery, and heart rate.

Polar Team Pro Monitors

Quantified in-game performance (speed, distance, acceleration).

FACTOR ANALYSIS

The process of simplifying the dataset by reducing many observed variables into hidden factors.

40 Features

8 lossless factors:

1. Speed and total acceleration zones
2. Average speed and distance
3. Average speed and acceleration zones
4. Minimum heart rate
5. Maximum heart rate
6. Recovery time
7. Maximum speed
8. High intensity acceleration zone

FEATURE IMPORTANCE

Random Forest (RF), XGB and Correlation (CORR) models were used to analyse feature importance. The scores from RF, XGB and CORR were averaged using a custom weighting scheme to produce one final score per feature.

PLAYER

- Training Strain
- Resistance Training
- Volume Load, Total
- Weekly Load,
- Heart Rate
- Variability
- Mental Performance Capability

TEAM

- Average Speed
- Distance
- Recovery Time
- Daily Average
- Speed and Total Acceleration Zones
- High-Intensity Acceleration Zones

CONFERENCE

- Peak Power
- Maximum Speed
- Sleep Consistency
- Deep Sleep Hours
- Emotional Balance

DATA PREPROCESSING

Before applying the XGB model, the data was preprocessed using the following techniques:

Missing Data

K-means clustering

Oversampling - SMOTE

Undersampling - ENN

LEARNING MODEL USED

eXtreme Gradient Boosting (XGBoost)

TL;DR: Lots of small trees + each fixes the last one's mistakes + smart maths to guide improvements.

At a High Level

1. **Start With Weak Model (eg, a small tree)**
2. **Calulate Error**
3. **Grow a tree to fix mistakes**
4. **Add new tree**
5. **Repeat (2-4)**
6. **Use "Gradient Descent" to Guide It** (looks at the steepest slope to reduce error)

*XGBoost includes extra controls to prevent overfitting

Thank you Gemini Imagen

HOW IT WAS USED

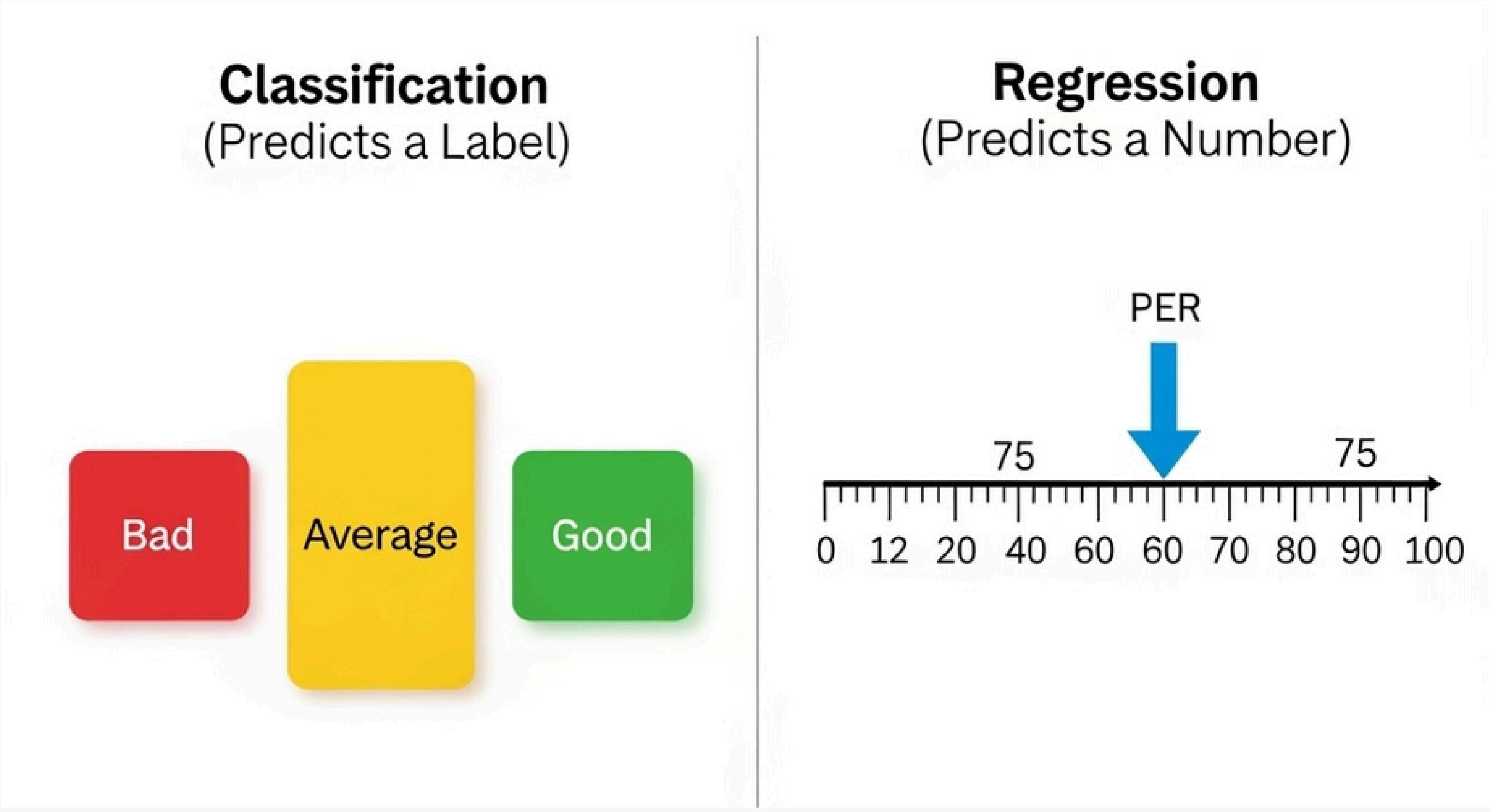
70% used to train the model / 30% saved for final test (results)

Classifier - (predicting group/label)

- Players' readiness (RSI) - Lower/Lower-Middle/Upper-Middle/Upper
- Game Score (GS) - Bad/Average/Good

Regressor - (predicting a number)

- Player Efficiency Rating (PER)



RESULTS

PLAYER READINESS (RSI)

Using the previous week's training, sleep, and stress data, the model **accurately predicted players' readiness** category (high to low readiness) **for the next week**.

GAME SCORE (GS)

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)

A **very low MSE** means the model's predictions are very **close to the actual values** (on average). The **R^2 value** indicates that **68% of the variation** in the outcome is explained by the model's input.

STRENGTHS

Access to full-team data across different training periods during the season allows for details analysis

1

Internal and external metrics allows for analysis of athlete's responses off and on the field.

2

Multi-level analysis provided insights at a player, team and conference level.

3

LIMITATIONS

Non-generalised model

Trained model only includes data from *one* year

Single metric analysis limits the representation of performance at each level

FUTURE WORK

Build a model that considers other sports at various levels to fit multiple different sport requirements.

Test how well these KPIs improve athletic performance

Offer clear and interpretable feedback to coaches and players

OUR TAKEAWAYS

Limitations of the study

- Lack of consideration for the human and situational side of a basketball game

Possible applications of the model

- Used as a tool to re-enforce existing coaching, rather than as a tool to drive team tactics.

ANY QUESTIONS?

THANK YOU

FOR LISTENING