Predicting Sports Readiness in Collegiate Female Soccer Players

Loyola Marymount University

Gavin Butts*, Shaye O'Beirne*, John Brasher+, Junyuan Lin*
*Department of Mathematics, +Sports Performance Program
Loyola Marymount University

Background

Only 13% of research in sports science focuses on women athletes and nearly **65% do not study women at all**. [1]

The primary goal is understanding an athlete's **readiness to perform or train**. These metrics can be used to later predict and prevent injury.

Athletes on **LMU's Women Soccer team** (33 individuals) respond to daily questionnaires, the data is then used to **train machine learning algorithms**.

Questions

- What variables are significant predictors of readiness?
- How much data is enough to converge on model accuracy?
- How can individual athletes better their readiness score?

Methods

Athletes surveyed daily using Google Forms. Statistics can be found in Table 1.

Removed outlying data using Cook's Distance and analysis of the hat matrix.

Variable	Mean	Std	Min	25%	50%	75%	Max
Stress	7.47	2.06	1.00	6.00	8.00	9.00	10.00
Sleep Quality	7.46	1.66	1.00	7.00	8.00	8.00	10.00
Hours Slept	7.62	1.25	3.00	7.00	8.00	8.00	10.00
Soreness	6.37	1.83	1.00	5.00	6.00	8.00	10.00
Hydration	7.58	1.70	2.00	7.00	8.00	9.00	10.00
Consumption	8.01	1.57	2.00	7.00	8.00	9.00	10.00
No Injury	8.83	3.22	0.00	10.00	10.00	10.00	10.00
Some Injury	1.01	3.011	0.00	0.00	0.00	0.00	10.00
Full Injury	0.16	1.27	0.00	0.00	0.00	0.00	10.00
Readiness Score	79.22	14.66	1.00	71.00	80.00	90.00	100.00

Table 1. Descriptive statistics used in survey and model creation

Trained **Ridge and Linear Regression models** using all possible combinations of predictor variables. Using significant predictors, **trained Neural Networks** (ReLU and softmax activation functions, trained with the Adam optimizer using sparse categorical cross-entropy loss). Then, predicted ready/not ready using **Logistic Regression**.

The highest response rate is 47, so individual models for athletes with >40 responses were created.

Results

The accuracy of the models begins to converge when n = 60 (Figure 1).

Linear Regression outperforms Ridge (R^2 = 0.553) and Neural Networks (R^2 = 0.484) with R^2 = 0.554. Linear Regression resulted with the following **significant predictors**:

- 'How stressed are you?'
- 'How well did you sleep?'
- 'How many hours did you sleep?'
- 'How sore are you?'
- 'How well did you fuel?'
- 'No Injury'

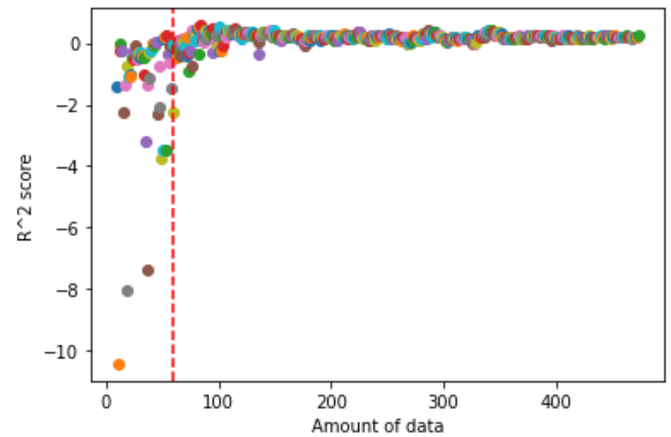

Figure 1. Convergence of accuracy given size of data set

If the self-reported readiness score is below 75, player is classified as 'not ready'. **Logistic regression finished with R^2 = 0.833**.

Individual profiles were created, finding significant predictors for each athlete and training a Linear Regression model. **R^2 varies between 0.881 and 0.977**.

Discussion

The **most significant predictor is 'No Injury'.** Given all other variables are held constant, **an athlete with no injury will have an average 52.13 point increase** in readiness compared to an injured athlete.

Future work:

- Assessing correlation between injury and readiness score.
- Using soreness data in model training.

References

[1] Paul, R. W. *et al. Am. J. Sports Med.* https://doi.org/10.1177/03635465221131281 (2022).