Analyzing Catapult GPS Data to Predict High Speed Running

4.29.24

Prepared for the University of Notre Dame Athletics Department by
Jack Arbuckle, Daniel Huang, Mason Marrone, Ian Pezzella, & Kaylin Slattery

The Lineup: Meet our Team

Jack
Arbuckle

Daniel
Huang

Mason
Marrone

Ian
Pezzella

Kaylin
Slattery

The Game Plan: Overview of Key Presentation Topics

- Executive Summary
- Methodology
- Preliminary Modeling Insights
- Final Models & Results
- Conclusion & Implications

Image from Notre Dame Athletics

Executive Summary

How & Why Notre Dame Athletics Leverages Catapult Data

Data Background

- Men's & Women's Soccer & Lacrosse
- Catapult system monitors training & practice loads

Lose Key Metrics Indoors

- No Duration-*, Distance-, or Velocity-based metrics
- ~700 → 200 variables

Image from Notre Dame Athletics

* Except Total Duration

Plan to Capture Additional Catapult Metric Value Indoors

<u>Critical Issue</u>

- Cannot track ***High Speed Running (HSR)*** indoors
  - Key for Performance Staff

GOAL

Develop models to predict HSR based on variables measured indoors

Top Takeaways: Key Findings & Most Accurate Model

Overall Findings

- Important predictors differ by model
- Most accurate models reflect component of player-individuality

Best Model: Composite Model

Composite Model % Residuals within ±150m		
Men's Lax	**Women's Lax**	**Soccer**
91.7%	86.8%	77.5%

What is the "Coaching Corner" and Why is it Included?

Purpose

- Communicate with various stakeholders
- Varying levels of understanding and comfortability with data from Performance Staff to Team Coaches
- Integral component of Sports Performance Presentation
- Integrated throughout the various presentation sections

Methodology

Creating Connections by Analyzing Correlated Variables

Exploring Relationships with HSR

- Analyzed the correlation of indoor variables with HSR
- **Goal:** identify potentially important relationships

8/12 Related to Player Load

Top 10 Indoor Variables Correlated to HSR

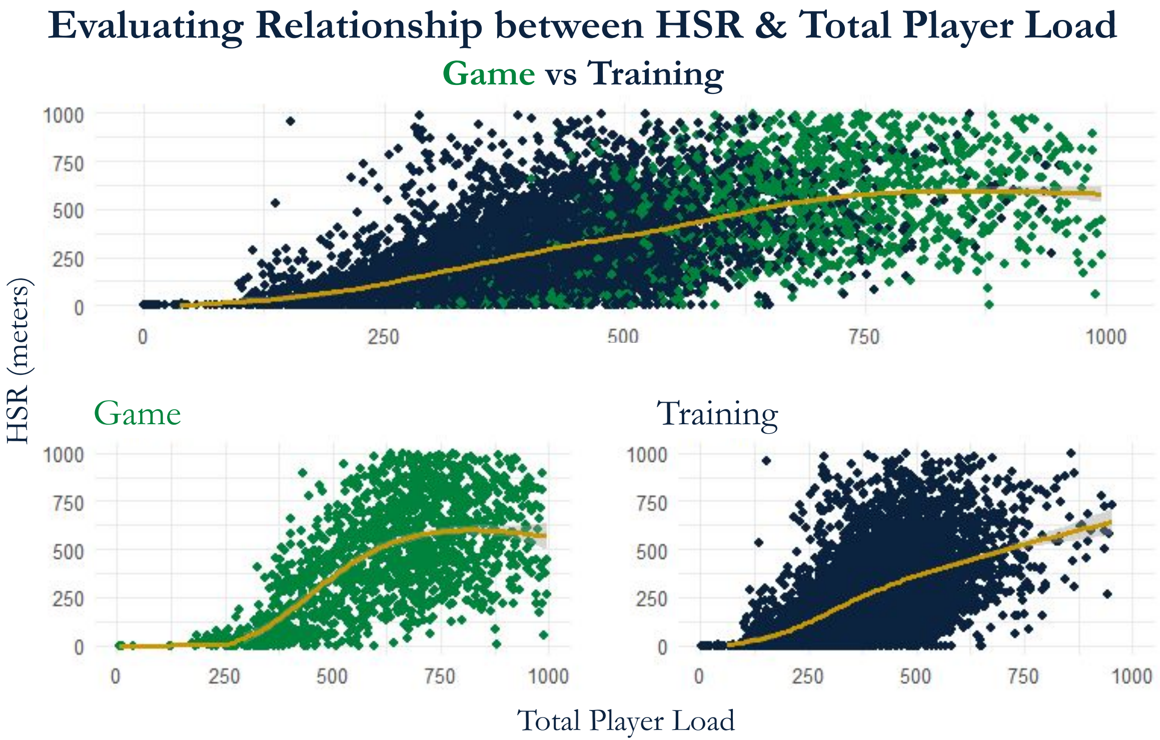
Descending by absolute value

Variable	Correlation
player_load_band_7_total_player_load	0.754
player_load_band_8_average_effort_count	0.741
player_load_band_7_average_effort_count	0.740
player_load_band_8_average_effort_count_session	0.740
player_load_band_7_average_effort_count_session	0.696
player_load_band_6_average_effort_count	0.651
high_intensity_load	0.645
high_intensity_load_avg	0.645
high_intensity_player_load	0.614
player_load_band_6_average_effort_count_session	0.562
work_rest_ratio	0.550
field_load	0.518

Evaluating Outliers: Player-Related… Not Game-Related

Evaluating Outliers:

- Assessing effect of training vs game
- Exhibit a **non-linear** polynomial relationship
- Support outliers were player specific and not related to games



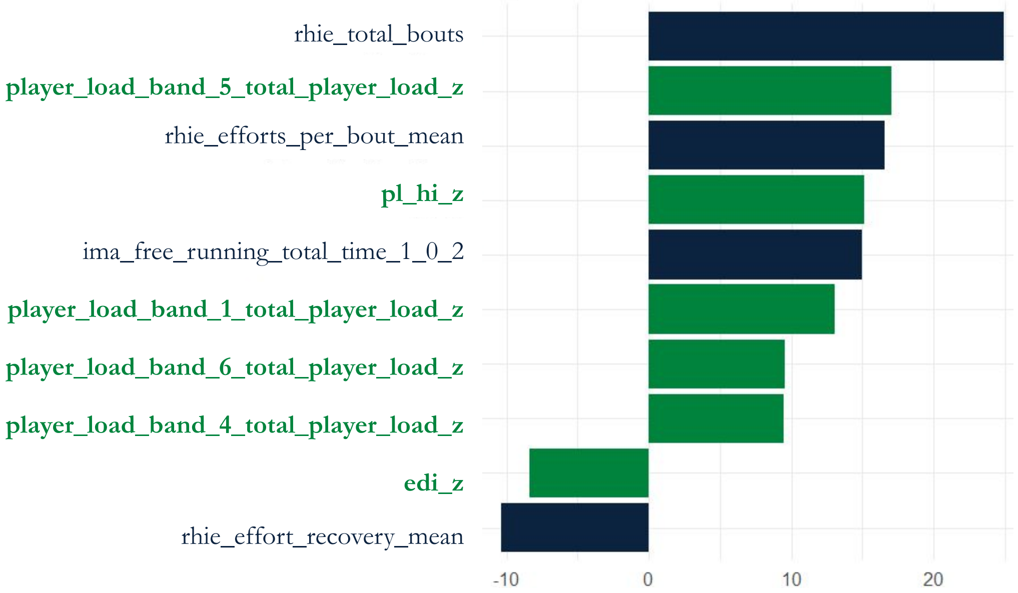
Incorporating Individuality: Importance of Player Z-scores

Creating Player Z-Scores

- Based on outlier-analysis
- Created z-scores for all numeric indoor metrics
- Many z-scores in top 10 predictors across XGBoost models
- Common practice within the athletics department

Top 10 Predictors from Preliminary XGBoost

Combined Soccer Data



Z-Score Variable

Separating Lacrosse Models Due to Predictor Differences

Lacrosse Differences

- Share **1/10** top predictors
- Tactics
- Rules of the Game

Result: Proceeding with Separate Lacrosse Models

Top 10 Variables from Tuned XGBoost Models

Lacrosse Data

Men	Women
player_load_1d_up_2_z	*player_load_band_3_total_player_load_z*
rhie_efforts_per_bout_mean	edi_z
high_intensity_load_avg_z	average_player_load_slow_session_z
high_intensity_load_z	player_load_slow_z
player_load_per_minute_z	player_load_1d_up_z
player_load_band_3_total_player_load_z	average_player_load_slow_z
player_load_band_4_total_player_load_z	player_load_band_2_average_effort_count_session
high_intensity_player_load_z	ima_7_o_clock_low_1_0_z
rhie_efforts_per_bout_mean	pl_hi_sess_z
peak_player_load	player_load_band_2_total_player_load_z

In Both

Combining Soccer Models Given Shared Predictors

Soccer Similarities

- Share **8/10** top predictors
- Tactics & Rules

Result: Create one soccer model

- Add gender variable

Top 10 Variables from Preliminary XGBoost Models

Soccer Data

Men	Women
ima_free_running_total_time_1_0	*rhie_effort_recovery_mean*
rhie_total_bouts	*rhie_total_bouts*
rhie_effort_recovery_mean	*activity_type_binary*
activity_type_binary	*rhie_efforts_per_bout_mean*
rhie_efforts_per_bout_mean	*ima_free_running_total_time_1_0*
ima_free_running_total_time_1_0_2	*ima_free_running_total_time_1_0_2*
ima_free_running_band_3_event_count_1_0	peak_player_load
ima_free_running_band_2_event_count_1_0	*ima_free_running_total_time*
ima_free_running_band_1_average_stride_rate_1_0	ima_free_running_band_2_event_count
ima_free_running_total_time	*ima_free_running_band_2_event_count_1_0*

In Both

Model Line-Up: Initial & Advanced Models for HSR

Preliminary Models

- Stepwise Regression
- Forward Subselection
- Backward Elimination
- XGBoost

Advanced Models

- Difference from Player's Average HSR
- Duration-Adjusted HSR
- Principal Components Analysis
- Composite Model

Image from Notre Dame Athletics

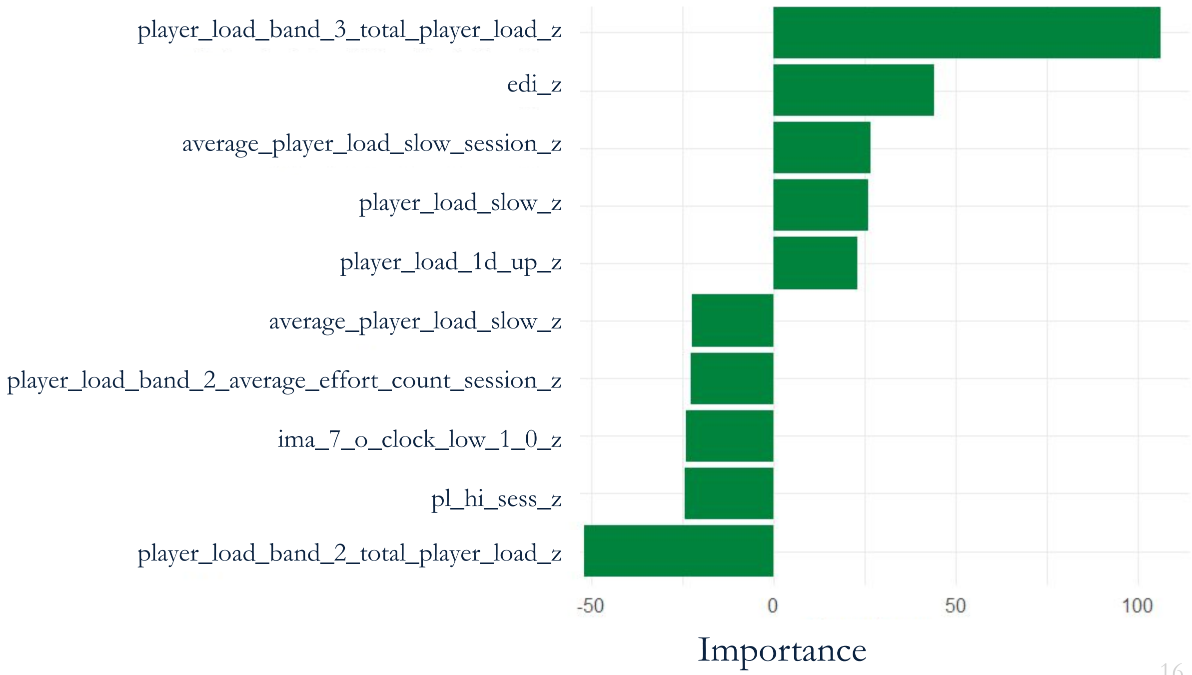
Performance Review: Approaches to Assessing Models

Evaluating Predictors

- Importance plots to identify top predictors
- Demonstrate importance of player-specific metrics (z-scores)

Preliminary Model Accuracy Metrics

Women's Lacrosse Data



Performance Review: Approaches to Assessing Models

Evaluating Predictors

- Importance plots to identify top predictors
- Demonstrate importance of player-specific metrics (z-scores)

Comparing Initial Models

- RMSE and R2

Preliminary Model Accuracy Metrics

Men's Lacrosse Data

	RMSE	MAE	R2
Forward	3.656	2.089	0.401
Backward	3.656	2.089	0.401
Stepwise	3.676	2.099	0.395
XGBoost	3.457	1.606	0.465

Performance Review: Approaches to Assessing Models

Evaluating Predictors

- Importance plots to identify top predictors
- Demonstrate importance of player-specific metrics (z-scores)

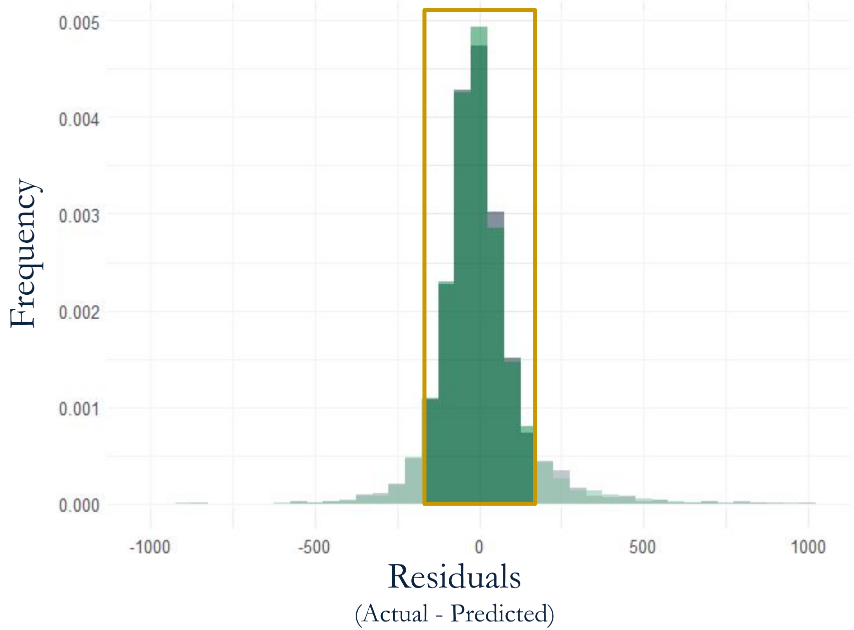
Comparing Initial Models

- RMSE and R2

Assessing Advanced Models

- Practical measure to understand model error
- Percent of residuals within **±150m**

Women's Lacrosse XGBoost Prediction Residuals Pre-Tune vs Post-Tune



Histogram bar width of 50 meters

High-Level Overview of the Project Methods

Coaching Points

- Used metrics available indoors to predict HSR value
- Compared important metrics to identify key Catapult variables
- Combined Soccer into one model based on similarities
- Separated Lacrosse models due to tactical and data-informed differences

Image from Notre Dame Athletics

Preliminary Modeling Insights

Starting Small: Modeling with Hypothesized Predictors

13 Hypothesized Predictors

- rhie_total_bouts
- ima_accel_low
- field_load_avg
- high_intensity_load
- explosive_efforts
- ima_9_o_clock_high_1_0
- ima_12_o_clock_high_1_0
- ima_3_o_clock_high_1_0
- ima_6_o_clock_high_1_0
- player_load_2d
- peak_player_load
- total_ima
- activity_type

Forward Selection

Stepwise Regression

Backward Elimination

XGBoost

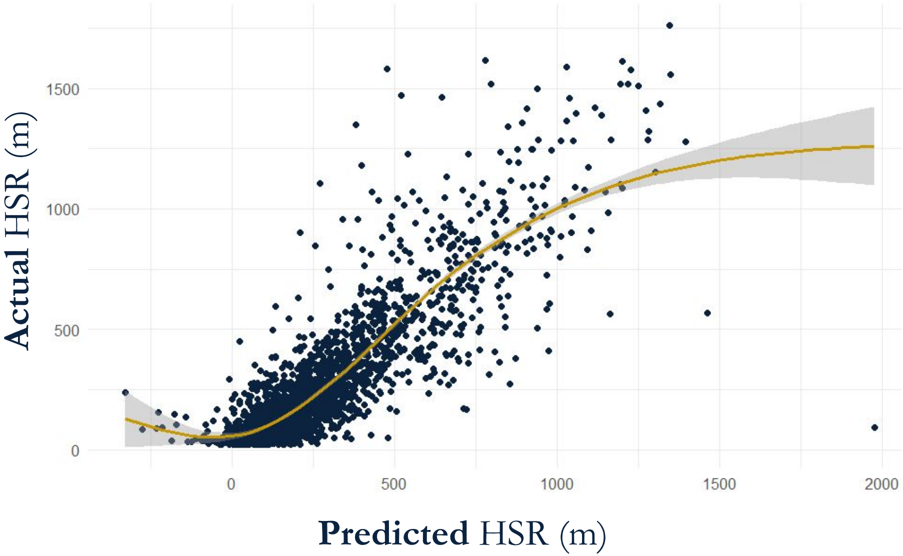
Initial Models Left Room for Improved Accuracy

Key Findings from Initial Models

- Z-score variables allowed the model to incorporate individual player differences
- *Consideration*: Require outdoor instances to provide baseline values

Initial XGBoost Model Accuracy		
Men's Lax	**Women's Lax**	**Soccer**
78.0%	84.2%	74.5%

Predicted vs Actual Values Tuned XGBoost
Women's Lacrosse



Summary of Preliminary Modeling Insights

<u>Coaching Points</u>

- Began process using a small set of the variables to predict HSR
- Non-linear trend for HSR
- Best performing models reflected a player-specific component (Z-scores)
- Model tuning process improved the model accuracy, but left room for improvement

Image from Notre Dame Athletics

Final Models & Results

Down to the Details: Duration-Adjusted Model

- Predicts *high_speed_distance_per_min*
- **Goal**: control for time across instances
- Estimate HSR (meters) by multiplying predictions by duration

Image from Notre Dame Athletics

Down to the Details: Player HSR Ratio Model

- Predicts difference from a player's average HSR:

$$HSR \,/\, \overline{HSR}$$

- **Goal**: capture more player-specific features given observed importance of player-specific variables

- Interpretability
  - Value 1 → HSR = Player's Avg
  - Value > 1 → More Intense than Avg
  - Value < 1 → Less Intense than Avg

Image from Notre Dame Athletics

Down to the Details: Principal Components Analysis

- Common approach in sports science
- **Goal**: Eliminate multicollinearity and retain a large amount of the variance
- Reduce dimensionality of dataset

Image from Notre Dame Athletics

Evaluating ***Tradeoff*** Between
Accuracy & Explainability

Comparing Final Models

- Women's Lacrosse HSR generally most predictable
- Soccer HSR least predictable
  - Greatest HSR variability
    - ***Soccer: avg 301m | max 11625m***
    - MLax: avg 166m| max 3032m
    - WLax: avg 132m| max 1771m
- **Composite Model most accurate** HSR predictions
  - Loses predictor interpretability

Model	Team	% Residuals within ±150m
Original XGBoost	Men's Lax	78.0%
	Women's Lax	84.2%
	Soccer	74.5%
Duration-Adjusted XGBoost with Z-scores	Men's Lax	71.5%
	Women's Lax	81.1%
	Soccer	72.2%
Difference from Player's Average HSR XGBoost with Z-scores	**Men's Lax**	**78.2%**
	Women's Lax	**85.1%**
	Soccer	65.8%
Principal Components (Linear Model)	Men's Lax	87.3%
	Women's Lax	85.8%
	Soccer	76.3%
Composite Model (Linear Model with Z-scores & Principal Components)	**Men's Lax**	**91.7%**
	Women's Lax	**86.8%**
	Soccer	**77.5%**

Green: Best model regardless of predictor interpretability
Bold: Best sport-specific model maintaining interpretability

Principal Components: MLAX 32 | WLAX 32 | Soc 29

Model Insights and Summaries

Coaching Points

- Created advanced models to account for shortcomings of initial models
- Interpretability of predictors a key difference between some models
- Must evaluate the tradeoff between accuracy and model interpretability

Image from Notre Dame Athletics

Conclusion & Implications

Bringing it Together: Key Findings & Conclusions

- Important predictors differ by model
- Composite model lacks interpretability
- Other models reduce accuracy

Compare model accuracy within ±150m vs investment in indoor Catapult system

- Consider potential error accumulation over multiple trainings/games

Image from Notre Dame Athletics

Taking it Further: Next Steps & Future Work

Incorporate Spring 24 Data as Validation Data

Consult with Performance Staff Regarding the Value of HSR metric

Stay Up to Date with Newly-Released Catapult Metrics

Image from Notre Dame Athletics

Thank you!
Go Irish!

APPENDIX

Appendix

Comparing Final Models

- Women's Lacrosse HSR generally most predictable
- Soccer HSR least predictable
  - Greatest HSR variability
    - ***Soccer: avg 301m | max 11625m***
    - MLax: avg 166m| max 3032m
    - WLax: avg 132m| max 1771m
- **Composite Model most accurate** HSR predictions
  - Loses predictor interpretability

Principal Components: MLAX 32 | WLAX 32 | Soc 29

Model	Team	% Residuals within ±150m
Original XGBoost	Men's Lax	78.0%
	Women's Lax	84.2%
	Soccer	74.5%
Duration-Adjusted XGBoost	Men's Lax	74.0%
	Women's Lax	81.2%
	Soccer	**76.4%**
Duration-Adjusted XGBoost with Z-scores	Men's Lax	71.5%
	Women's Lax	81.1%
	Soccer	72.2%
Difference from Player's Average HSR XGBoost	Men's Lax	74.5%
	Women's Lax	80.7%
	Soccer	65.3%
Difference from Player's Average HSR XGBoost with Z-scores	**Men's Lax**	**78.2%**
	Women's Lax	**85.1%**
	Soccer	65.8%
Principal Components (Linear Model)	Men's Lax	87.3%
	Women's Lax	85.8%
	Soccer	76.3%
Composite Model (Linear Model with Z-scores & Principal Components)	**Men's Lax**	**91.7%**
	Women's Lax	**86.8%**
	Soccer	**77.5%**

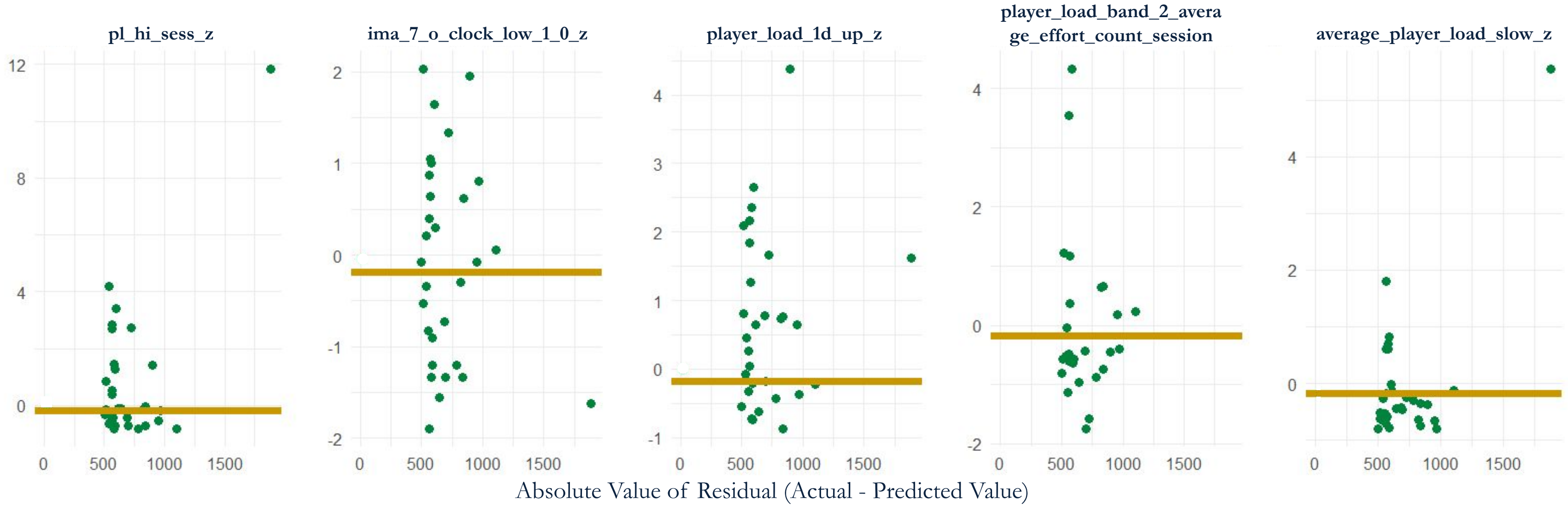
Green: Best model regardless of predictor interpretability
Bold: Best sport-specific model maintaining interpretability

Data Cleaning

- Identify variables not measured indoors
- Removed columns with over 95% missing values
- Scraped game schedules to add ***activity_type***
- Updated ***position_name*** column
- Modified format of all duration-based columns
- Added binary ***team_gender*** column

XGBoost Outlier Instances – Women's Lacrosse

Comparing Outlier Instances with Top Variable Average Values (II/II)



Insight: Variables have distributed values that are above and below the respective average lines, but a few key outliers exist on the graph, especially with the highest residuals (top right corners)

Variable of Interest: ***player_load_1d_up_z*** variation in values may be a result from variation in picking up more ground balls

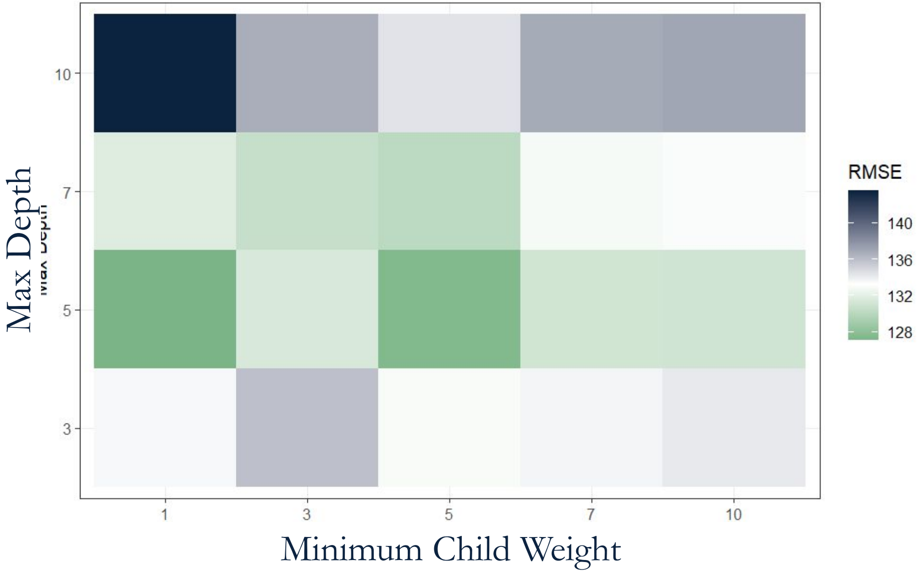
Parameter Tuning Process to Improve Model Accuracy

XGBoost Parameters

- Cross-validation to identify best parameters
  - RMSE measures accuracy
- Tuning Parameters
  - Minimum Child Weight
  - Max Depth
  - Gamma
  - Column Sample
  - Subsample
  - Alpha
  - Lambda

Example Tuning for Max Depth & Min Child Weight

Combined Soccer Data



Lowest RMSE: 1 min Child Weight, 5 Max Depth

Determining Accuracy Evaluation Meters Cutoff

Evaluation Cutoff Analysis

- Cutoff ranges differ by 25m
- Greater accuracy with larger cutoff range

Accuracy by Different Cutoff Values
Composite Model

Cutoff Distance	Team	% Residuals within ± Cutoff
100m	Men's Lax	76.3%
	Women's Lax	81.4%
	Soccer	65.8%
125m	Men's Lax	85.7%
	Women's Lax	81.4%
	Soccer	71.5%
150m	Men's Lax	91.7%
	Women's Lax	86.8%
	Soccer	77.5%

Modeling Updates

Evaluating Prediction Errors: Combined Soccer XGBoost

- Outliers especially over 2500 meters High Speed Distance
- ***Negative Instances***
  - Model predicting some instances as negative high speed distance – want to investigate such instances