Sam Denomme | Bio-Mech Velocity Model | Analysis

1. Summary

Developed a Bayesian predictive model using marker-less motion capture data to estimate fastball velocity with high precision. Leveraging a Bayesian Generalized Additive Mixed Model framework, the model quantifies the contributions of specific kinematic and kinetic variables—such as shoulder internal rotational velocity and hip-shoulder separation—to pitch speed. By incorporating random effects to capture pitcher-specific variability and using robust standardization techniques, the model explains over 73% of the variance in fastball velocity.

The analytical framework employs rigorous variable selection techniques including Random Forest importance and Maximal Information Coefficient analysis to detect non-linear relationships, while Variance Inflation Factor filtering ensures multicollinearity is minimized. This approach allows the model to function as both a diagnostic tool for assessing the efficiency of a pitcher's kinetic chain and a strategic instrument for pinpointing mechanical inefficiencies—insights that can be directly translated into targeted strength training and player development interventions.

2. Purpose

While pitch velocity is a well-established performance driver—reducing hitter reaction time and narrowing decision windows—less is known quantitatively about the specific biomechanical factors that underpin it. This project aims to bridge that gap by systematically identifying which movement patterns and force metrics most strongly correlate with pitch speed. By isolating high-impact kinematic and kinetic variables, the model creates a data-driven foundation for individualized performance evaluation, longitudinal tracking, and forecasting future velocity development.

3. Data Cleaning

Pitch-level biomechanical data was merged with player metadata (e.g., height, mass, playing level) using the session-level identifier. The resulting dataset was cleaned and validated through the following steps:

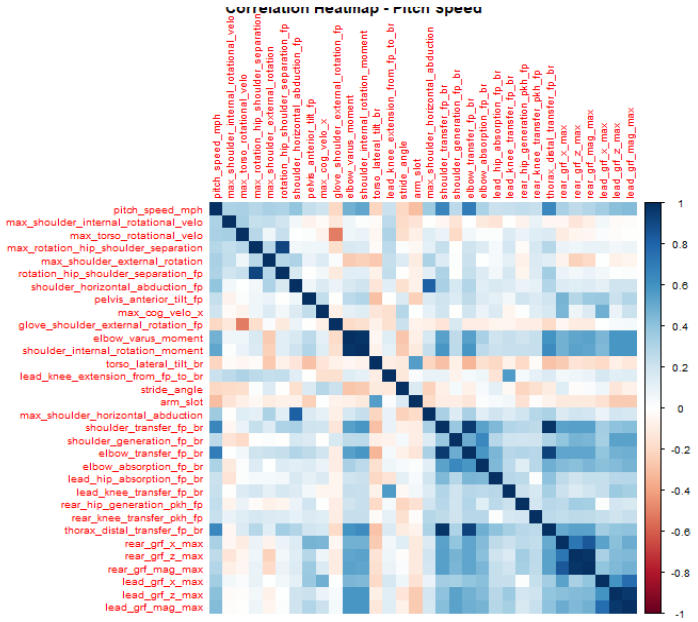
1. **Missing Value Imputation**
    - Random Forest imputation (missRanger) with 100 trees and predictive mean matching (pmm.k = 3) was applied.
    - This method preserved non-linear and interaction effects among biomechanical features.
    - All missing values were successfully resolved post-imputation.
2. **Distributional Testing**
    - A Shapiro-Wilk test was performed on each numeric variable to assess normality.

- This guided whether z-score (for Gaussian distributions) or IQR-based (for non-Gaussian distributions) thresholds were used for outlier detection.

3. **Outlier Detection and Flagging**
    - Rather than removing outliers, the script appended _outlier flags for each numeric column.
    - This allows for flexible downstream filtering, exploratory visualization, or subgroup analysis.
4. **Output**
    - A fully imputed and flagged dataset was saved to outputs/pitch_data_cleaned.csv.
    - Normality results and outlier counts were logged and printed for auditability.

4. EDA

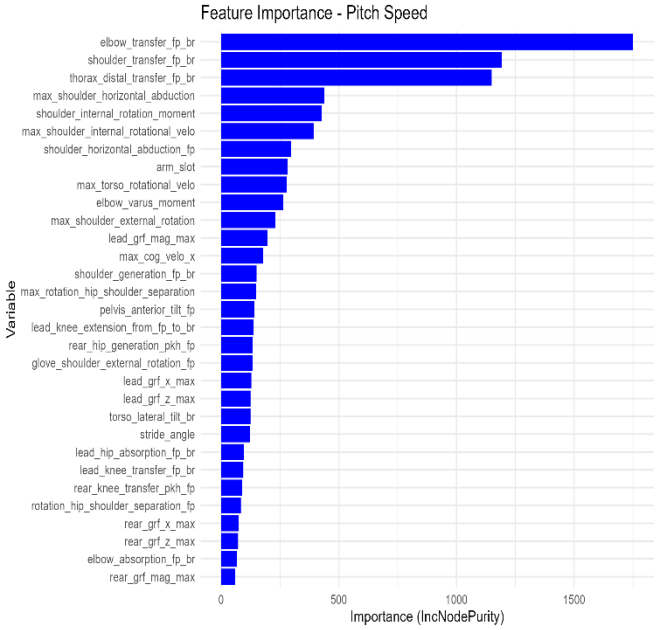
Exploratory data analysis focused on identifying the biomechanical variables most strongly associated with fastball velocity. A Pearson correlation matrix was computed across all numeric features, and a filtered heatmap (|r| > 0.20) was generated to visualize the most meaningful relationships. The strongest positive correlations with pitch velocity were observed in shoulder internal rotational velocity, torso rotational velocity, and hip-shoulder separation, reinforcing established principles of energy transfer through the kinetic chain. Additional high-impact variables included elbow varus moment, shoulder transfer force, and lead knee extension, which reflect both proximal-to-distal sequencing and force generation efficiency. These insights directly informed feature prioritization and the specification of interaction terms in the final velocity model.



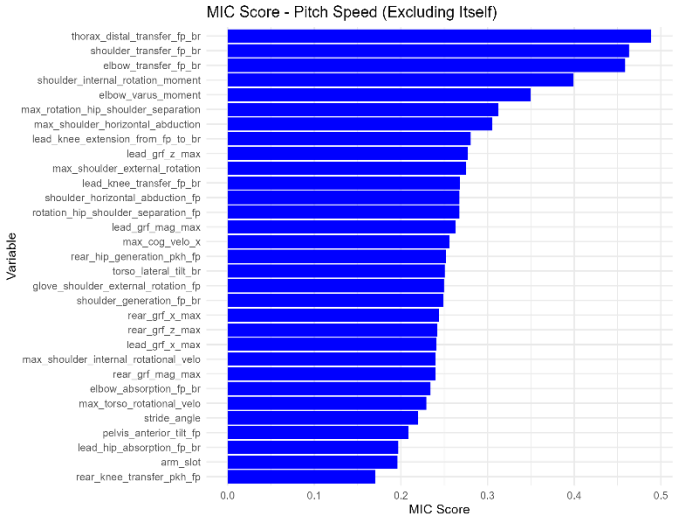
5. Feature Selection

Feature selection incorporated multiple complementary techniques to identify the most biomechanically informative predictors of fastball velocity, balancing both statistical rigor and domain interpretability.

- **Random Forest Variable Importance**
  A Random Forest regression model (2,500 trees) was trained using variables correlated with pitch speed ($|r| > 0.2$). Variable importance scores, based on increase in node purity, highlighted elbow transfer force, shoulder transfer force, and thorax distal transfer as the top contributors. These metrics are key to understanding energy transfer through the upper kinetic chain.



- **Maximal Information Coefficient (MIC)**
  To uncover non-linear relationships missed by standard correlation, MIC analysis was conducted. Variables with MIC ≥ 0.2 were retained, reinforcing the predictive value of rotational mechanics, lead knee extension, and ground reaction forces, in addition to the upper-body transfer metrics already identified.



- **Variance Inflation Factor (VIF) Filtering**
  To mitigate multicollinearity, predictors with VIF > 5 were excluded. This ensured model interpretability and reduced redundancy from closely related movement metrics (e.g., overlapping joint velocities).

The final feature set retained high-impact predictors across both the lower and upper kinetic chain, with emphasis on:

- Energy generation (e.g., hip and shoulder rotational velocities)

- Energy transfer (e.g., thorax and elbow forces)
- Sequencing efficiency (e.g., lead knee extension, GRF timing)

These features served as the foundation for the model's fixed effects structure and were also used to construct meaningful biomechanical interaction terms.

6. Model

To quantify biomechanical contributors to fastball velocity, a series of Bayesian Generalized Additive Mixed Models (GAMMs) were implemented using rstanarm::stan_gamm4(). This framework was selected to accommodate the non-linear and hierarchical nature of the data, where each pitcher may contribute multiple observations across sessions.

Bayesian GAMMs offer several advantages:

- Model non-linear relationships using smooth terms
- Account for within-subject variability via random effects
- Maintain interpretability through additive model structure

All numeric predictors were z-score standardized to ensure comparability across biomechanical metrics with differing units. Scaling parameters were saved for consistent downstream use. Each model was trained with 4,000 iterations (1,500 warmup) across 2 chains, using adapt_delta = 0.95 and max_treedepth = 12 to ensure convergence and accommodate complex interactions.

Model	**Predictors**
Force	rear_grf_z_max, lead_grf_z_max, p_throws, user (random effect)
Lower Body	max_torso_rotational_velo, max_rotation_hip_shoulder_separation, pelvis_anterior_tilt_fp, max_cog_velo_x, lead_knee_extension_from_fp_to_br, lead_knee_transfer_fp_br, rear_hip_generation_pkh_fp
Upper Body	max_shoulder_internal_rotational_velo, max_shoulder_external_rotation, max_shoulder_horizontal_abduction, elbow_varus_moment, elbow_transfer_fp_br
Full	All variables above plus interactions: • max_rotation_hip_shoulder_separation × lead_knee_extension_from_fp_to_br • elbow_varus_moment × elbow_transfer_fp_br • rear_grf_z_max × lead_grf_z_max

Smooth terms (s()) were applied to variables such as **max_torso_rotational_velo**, **rear_grf_z_max**, and **max_cog_velo_x** to flexibly capture nonlinear effects without overfitting. The Full Model included biomechanically informed interactions to capture synergistic effects across body segments—for example, the coordination between hip rotation and lead knee extension, or bilateral timing of ground force application.

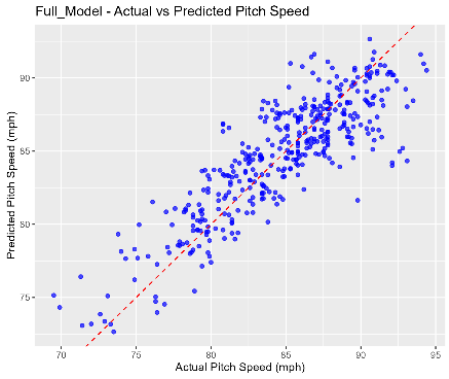
7. Model Evaluation

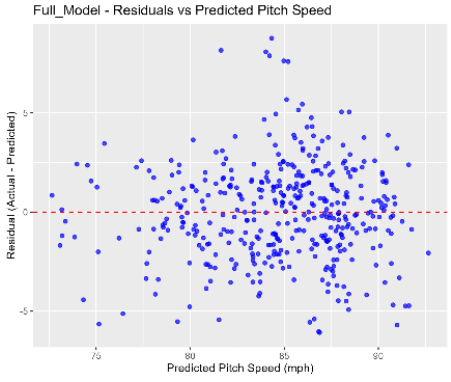
Model performance was assessed on a held-out test set using three standard regression metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 (coefficient of determination). These metrics quantify both average prediction error and overall model fit.

Model	RMSE	MAE	R^2
Force	4.22	3.42	.2
Lower Body	3.75	3.01	.37
Upper Body	2.85	2.24	.63
Full	2.44	1.9	.73

The Full Model, which combined all biomechanical domains along with targeted interaction terms, achieved the highest predictive accuracy—explaining over 73% of the variance in fastball velocity. This outperformed models focused solely on ground reaction forces, lower body mechanics, or upper body movements, reinforcing the importance of kinetic chain integration in pitch velocity generation. Two visualizations were produced for each model to assess model calibration and residual structure:

- Actual vs. Predicted Plot
  Displays how closely predictions align with true pitch speeds. A tighter clustering around the identity line (y = x) indicates stronger predictive accuracy.
- Residuals vs. Predicted Plot
  Evaluates systematic bias or heteroscedasticity in model residuals. Ideally, residuals should be randomly dispersed around zero.





For the Full Model, both plots indicated a strong and well-calibrated fit:

- Predicted values tracked closely with actual fastball velocities.
- Residuals showed no obvious trends, suggesting no major violations of model assumptions.

The Full Model's performance confirms that multi-segment biomechanical integration, including both upper- and lower-body contributors and their interactions, provides the most reliable foundation for pitch velocity estimation. The approach demonstrates strong potential for use in performance profiling, mechanical assessment, and development tracking in pitcher evaluation systems.