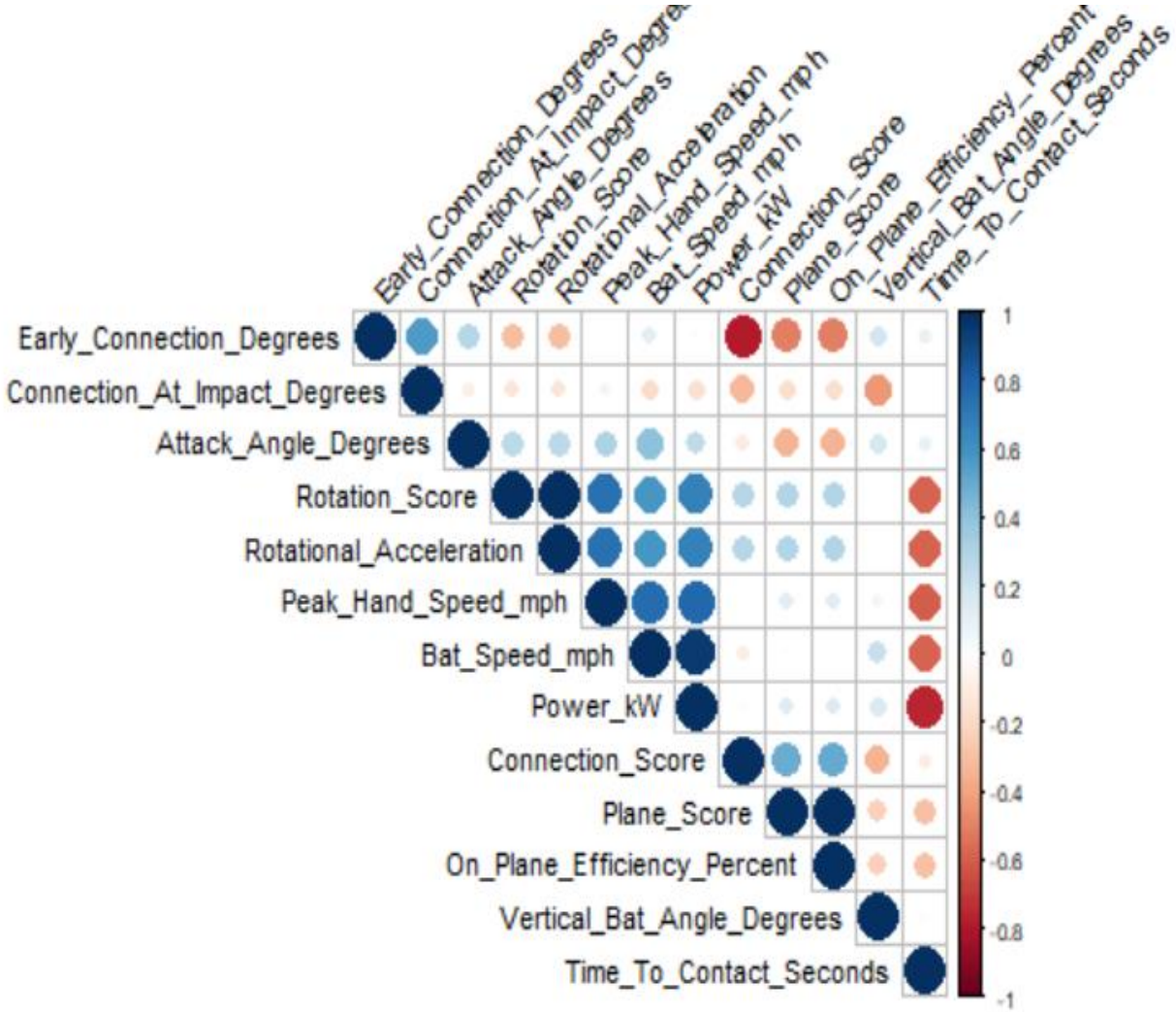
Examining a Player's Blast Motion Plane, Connection, and Rotation Score Sensitivities and Relationships After Removing Substantial Independent Variables with R's Caret Package

Zachariah Zanger M.S.

(Correlation Matrix of Blast Motion Data)



Introduction and Specifications

Blast Motion data points form a framework for understanding of a Baseball swing and the movements involved. Blast Motion data is headlined by Plane, Connection, and Rotation scores. This writing and research is not a deep dive of what these metrics mean and what they stand for, but instead the examination of the change in these scores for different Blast Motion data fields. If one can understand the sensitivities of scores, one could train the independent data fields and observe the scores improve. Even better, if one can understand the deeper dependencies and movements of all Blast Data fields, one could form a more robust understanding than what the brochures say. For instance, Blast publishes that, "On-Plane Efficiency is how consistently a player maintains their arm and wrist throughout the swing, or the percentage of the swing where the bat is on-plane. The score shows you where you fall within your skill level, so both matter. The value of the On-Plane Efficiency metric is obviously going to drive your score; The higher it is, the higher your score." (Blast Motion Blog, 2019). This analysis is to find what other data fields other than On-Plane Efficiency move with Plane Score. Given the interconnectedness of the swing, from stride to follow through, it would be foolish not to examine deeper relationships with Blast data.

The data used for this exercise is from a player's swings throughout a Summer season. The LM (Ordinary Least Squares Regression), Ranger (Random Forest), and GLMnet (Generalized Linear Model) methods will be utilized to dial up predictive models for the three scores. An 80/20 split will be used to train and test data, in addition to a cross validation with 10 folds.

An observation base of 841 swings will be used. The findings of this analysis are unique to this player, who cannot be identified for non-disclosure purposes. For each score, models will be comprised of independent variables that are not what Blast Motion publishes as the leading independent variables for the different scores (dependent variable). Specifically:

Dependent Variable	Independent Variables
Plane Score	Connection Score, Rotation Score, Bat Speed MPH, Rotational Acceleration, Attack Angle Degrees, Early Connection Degrees, Connection at Impact Degrees, Vertical Bat Angle Degrees, Power kW, Time to Contact Seconds, Peak Hand Speed MPH
Connection Score	Plane Score, Rotation Score, Bat Speed MPH, Rotational Acceleration, Attack Angle Degrees, Power kW, Peak Hand Speed MPH, On Plane Efficiency Percent
Rotation Score	Connection Score, Plane Score, Bat Speed MPH, Attack Angle Degrees, Early Connection Degrees, Connection at Impact Degrees, Vertical Bat Angle Degrees, On Plane Efficiency Percent, Peak Hand Speed MPH

It should be noted that the three algorithms used in this analysis perform differently and have unique strengths and weaknesses for use. The LM method in the Ranger package performs vanilla OLS (Ordinary Least Squares) regression that creates a line of best fit that minimizes the distance of the residuals. Unfortunately, OLS regression is often susceptible to overfitting or is unable to handle non-linear relationships that well.

The Ranger method in the Ranger package runs the Random Forest algorithm on the data. Random Forests are primarily used for classification purposes with decision trees and information gain but could also be used for prediction for continuous data. Random Forests are robust to overfitting and work well with non-linear relationships. The problem is that Random Forests require hyperparameter tuning before fitting a model, most importantly *mtry*, or the number of selected variables used on each split.

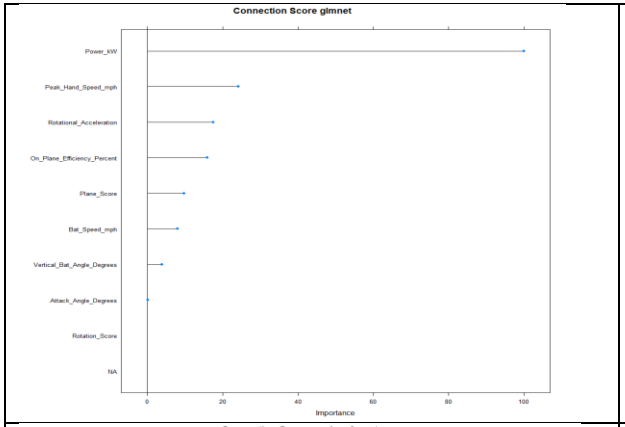
Lastly, the GLM method utilizes Generalized Linear Models that helps Linear Regression models that struggle with multi-collinearity (multiple independent variables that are correlated) and overfitting. Generalized Linear Models takes two primary forms: Lasso Regression that penalizes the number of non-zero coefficients, and Ridge Regression that penalizes the absolute magnitude of coefficients. GLM models sort of give OLS regression a push towards greater accuracy. The end user must specify the alpha and lambda parameters that determine what combination of Lasso and Ridge regression to utilize.

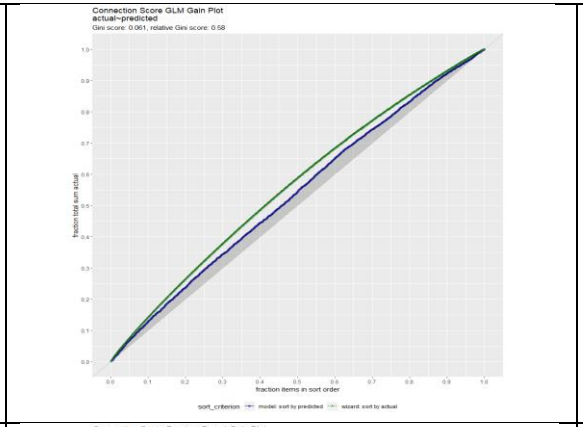
Models Strength

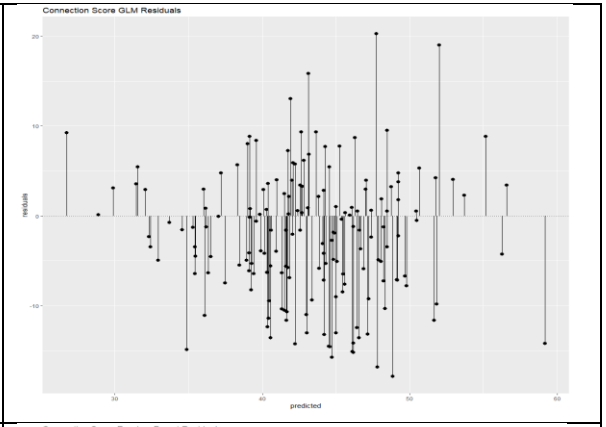
Interpreting the strength and fit of the three algorithms for the three different scores for the given independent variables, a summary table is compiled below:

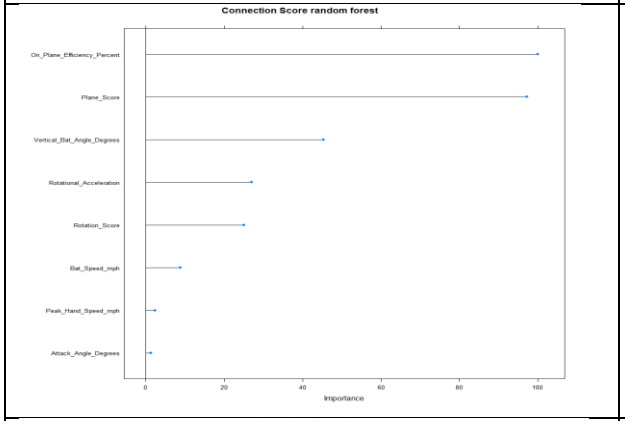
	Plane Score	Rotation Score	Connection Score
LM R^2	.45	.71	.42
LM RMSE	4.77	4.71	7.43
LM MAE	3.72	3.67	5.70
Random Forest R^2	.49	.69	.39
Random Forest RMSE	4.58	4.86	7.65
Random Forest MAE	3.57	3.73	5.89
GLM R^2	.39	.70	.39
GLM RMSE	4.90	4.71	7.49
GLM MAE	3.84	3.71	5.75

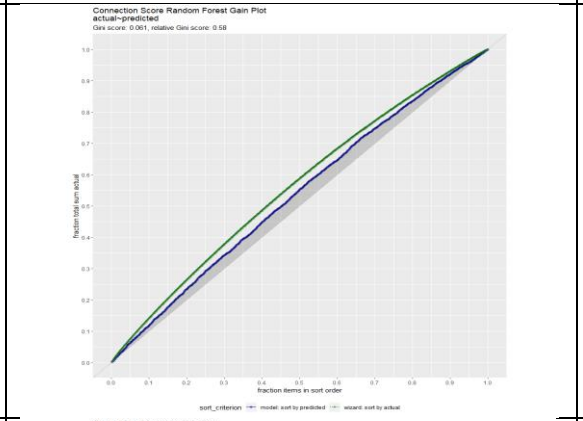
These models took substantial hits when the important independent variables were removed. For instance, the R^2 term represents the variance explained by the models and is generally interpreted as how well a model fits the data. R^2 is on a scale in between 0 and 1, and the models for Plane Score and Connection fall closer to 0. In addition, the RMSE term represents the standard deviation of the residuals and a measure for how spread out model predictions are from actual observations. These models miss typically from 4-7 points and on a 20-80, 4-7 points should not be ignored. From this table, it can be inferred that the models do not fit the data all that well, but still have some value if safely interpreted. Connection Score is fit particularly poor.

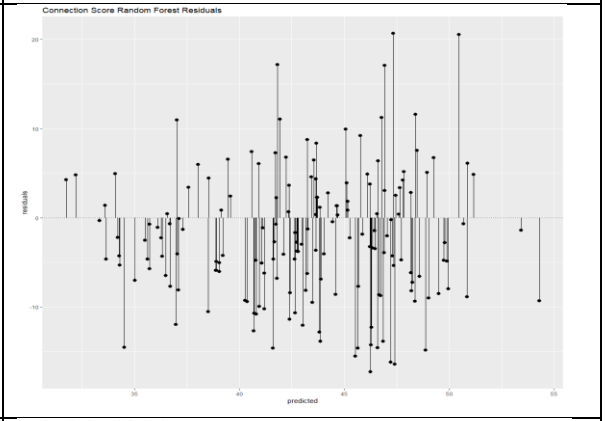

Connection Score glmnet
Power_kW
Peak_Hand_Speed_mph
Rotational_Acceleration
On_Plane_Efficiency_Percent
Plane_Score
Bat_Speed_mph
Vertical_Bat_Angle_Degrees
Attack_Angle_Degrees
Rotation_Score
NA
0
20
40
60
80
100
Importance

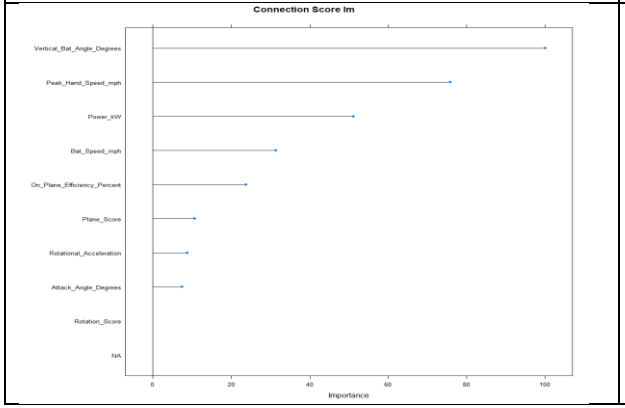

Connection Score GLM Gain Plot
actual~predicted
Gini score: 0.061, relative Gini score: 0.58
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual

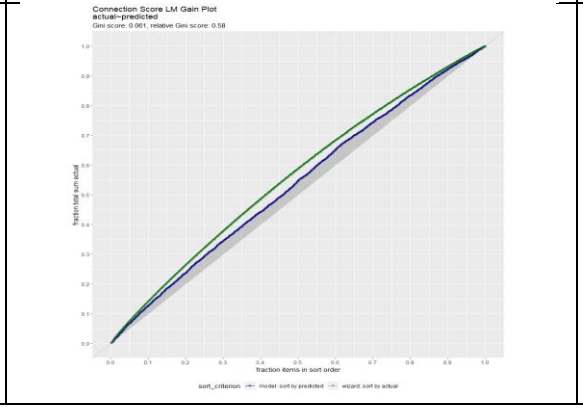

Connection Score GLM Residuals
residuals
predicted

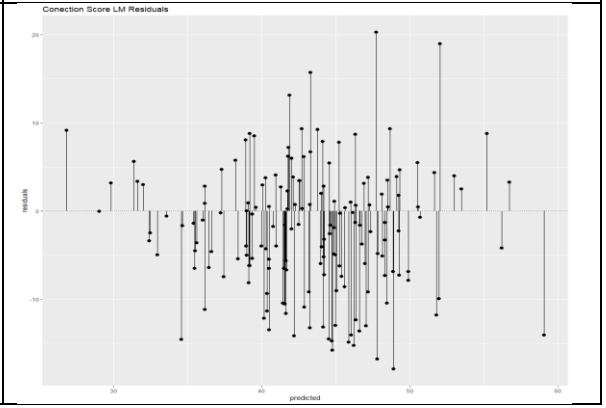

Connection Score random forest
On_Plane_Efficiency_Percent
Plane_Score
Vertical_Bat_Angle_Degrees
Rotational_Acceleration
Rotation_Score
Bat_Speed_mph
Peak_Hand_Speed_mph
Attack_Angle_Degrees
0
20
40
60
80
100
Importance

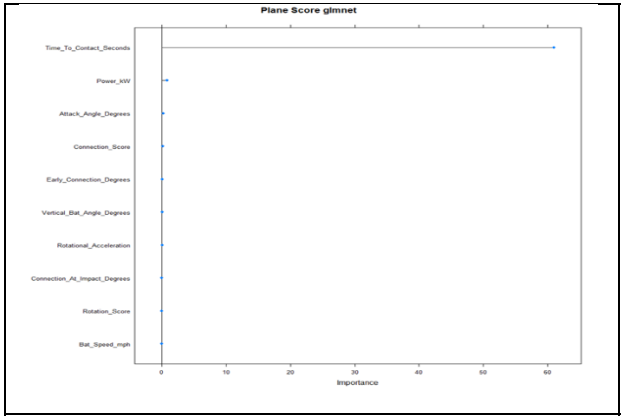

Connection Score Random Forest Gain Plot
actual~predicted
Gini score: 0.061, relative Gini score: 0.58
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual

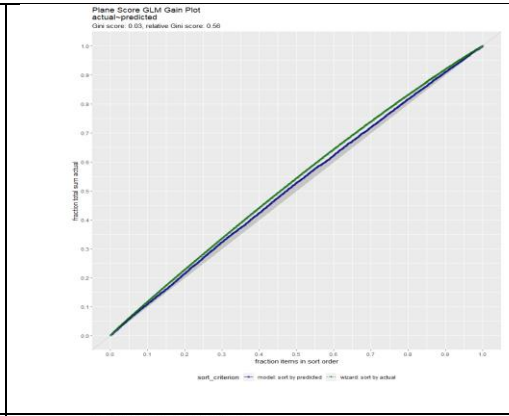

Connection Score Random Forest Residuals
residuals
predicted

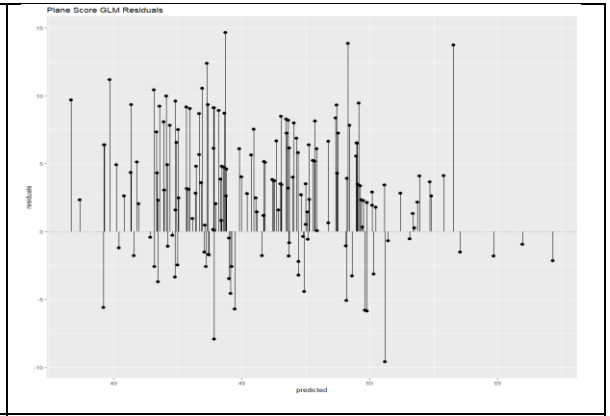

Connection Score lm
Vertical_Bat_Angle_Degrees
Peak_Hand_Speed_mph
Power_kW
Bat_Speed_mph
On_Plane_Efficiency_Percent
Plane_Score
Rotational_Acceleration
Attack_Angle_Degrees
Rotation_Score
NA
0
20
40
60
80
100
Importance

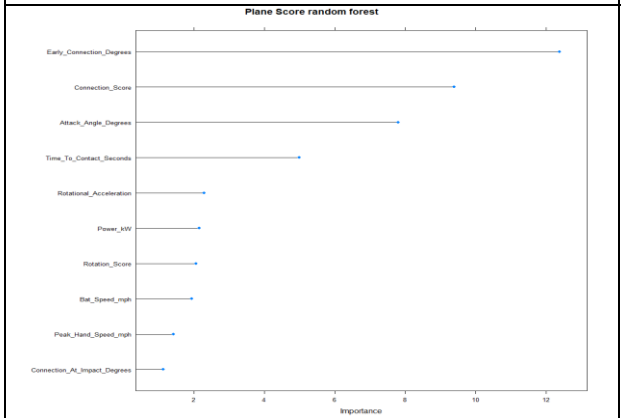

Connection Score LM Gain Plot
actual~predicted
Gini score: 0.061, relative Gini score: 0.58
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual

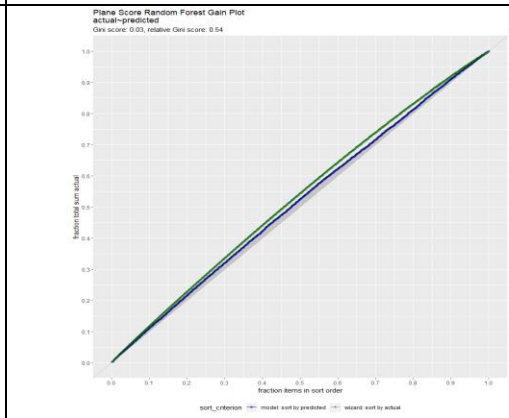

Conection Score LM Residuals
residuals
predicted

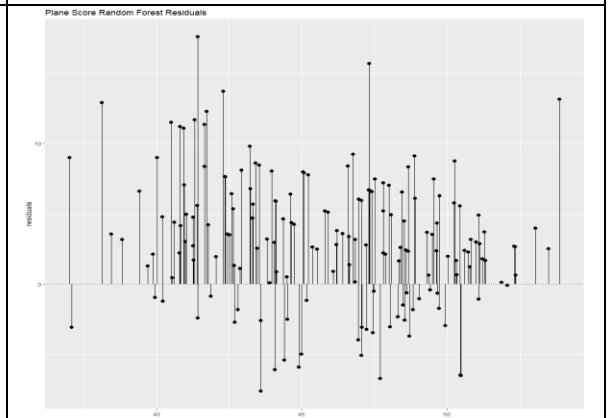

Plane Score glmnet
Time_To_Contact_Seconds
Power_kW
Attack_Angle_Degrees
Connection_Score
Early_Connection_Degrees
Vertical_Bat_Angle_Degrees
Rotational_Acceleration
Connection_At_Impact_Degrees
Rotation_Score
Bat_Speed_mph
0
10
20
30
40
50
60
Importance

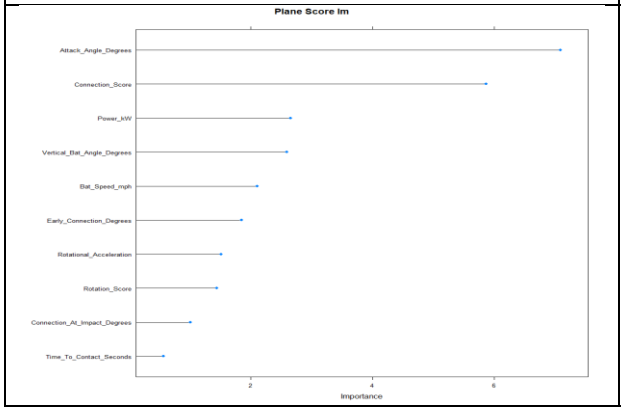

Plane Score GLM Gain Plot
actual~predicted
Gini score: 0.03, relative Gini score: 0.56
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual

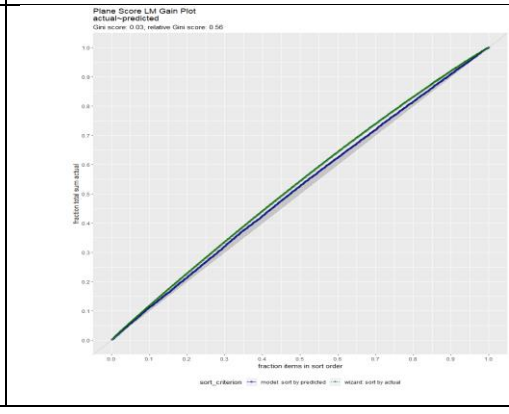

Plane Score GLM Residuals
residuals
predicted

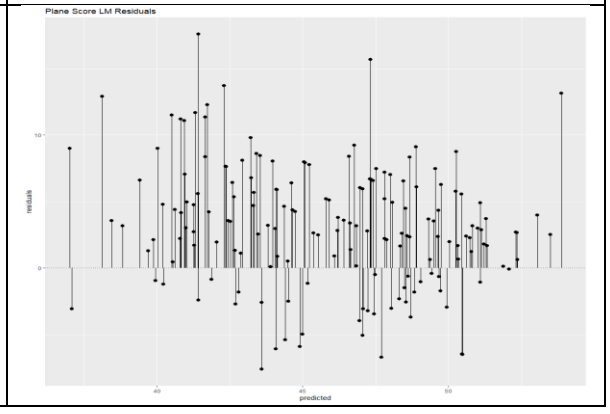

Plane Score random forest
Early_Connection_Degrees
Connection_Score
Attack_Angle_Degrees
Time_To_Contact_Seconds
Rotational_Acceleration
Power_kW
Rotation_Score
Bat_Speed_mph
Peak_Hand_Speed_mph
Connection_At_Impact_Degrees
2
4
6
8
10
12
Importance

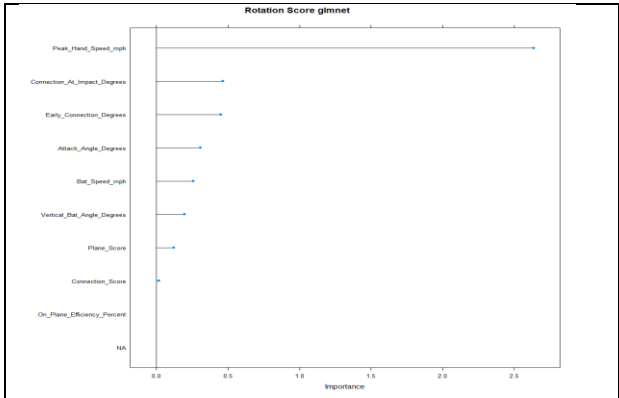

Plane Score Random Forest Gain Plot
actual~predicted
Gini score: 0.03, relative Gini score: 0.54
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual

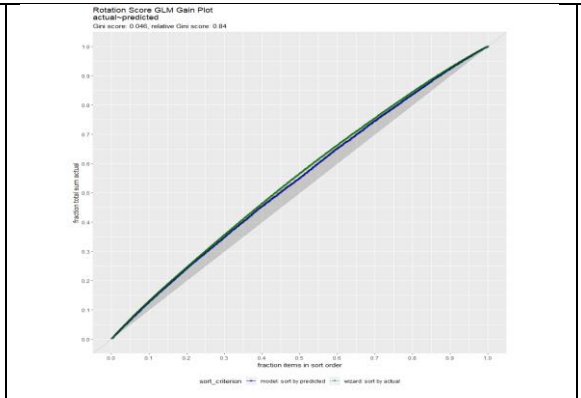

Plane Score Random Forest Residuals
residuals
predicted

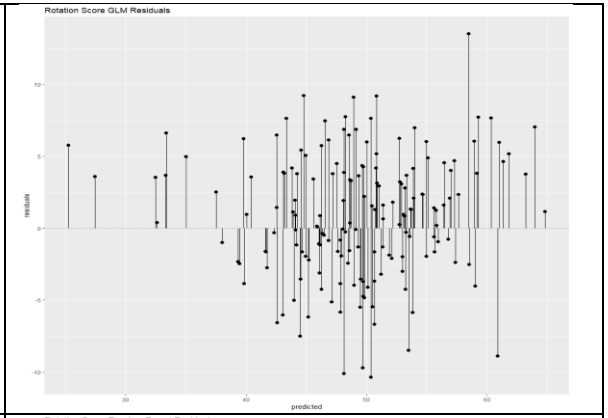

Plane Score lm
Attack_Angle_Degrees
Connection_Score
Power_kW
Vertical_Bat_Angle_Degrees
Bat_Speed_mph
Early_Connection_Degrees
Rotational_Acceleration
Rotation_Score
Connection_At_Impact_Degrees
Time_To_Contact_Seconds
2
4
6
Importance

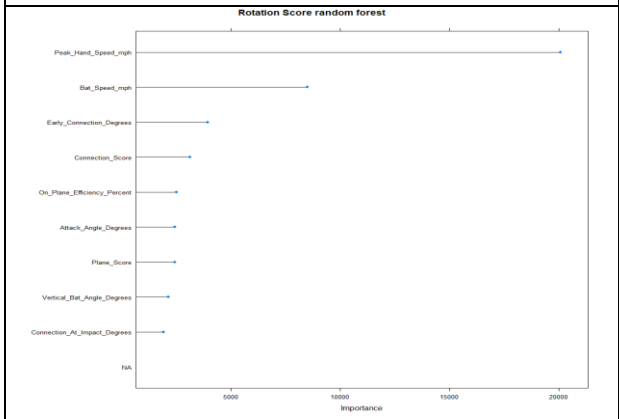

Plane Score LM Gain Plot
actual~predicted
Gini score: 0.03, relative Gini score: 0.56
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual

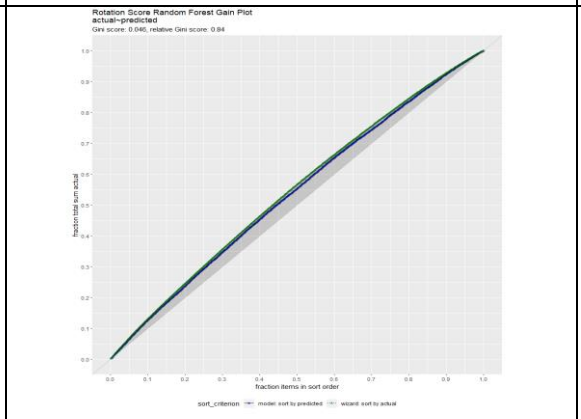

Plane Score LM Residuals
residuals
predicted

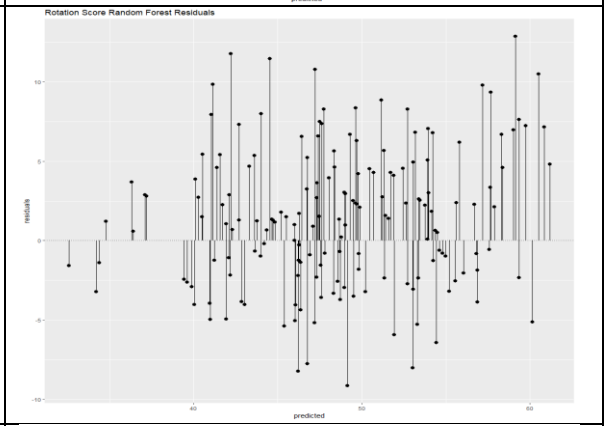

Rotation Score glmnet
Peak_Hand_Speed_mph
Connection_At_Impact_Degrees
Early_Connection_Degrees
Attack_Angle_Degrees
Bat_Speed_mph
Vertical_Bat_Angle_Degrees
Plane_Score
Connection_Score
On_Plane_Efficiency_Percent
NA
0.0
0.5
1.0
1.5
2.0
2.5
Importance

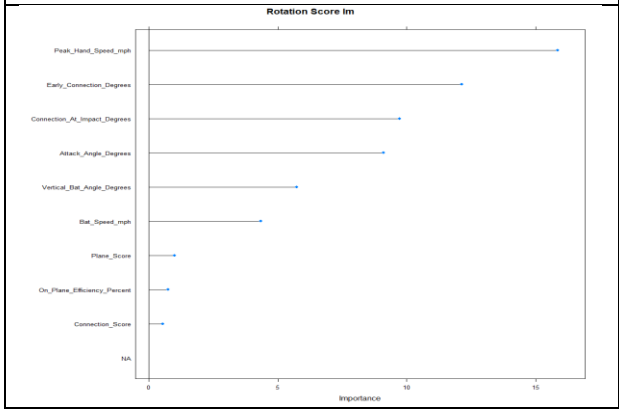

Rotation Score GLM Gain Plot
actual~predicted
Gini score: 0.040, relative Gini score: 0.84
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual

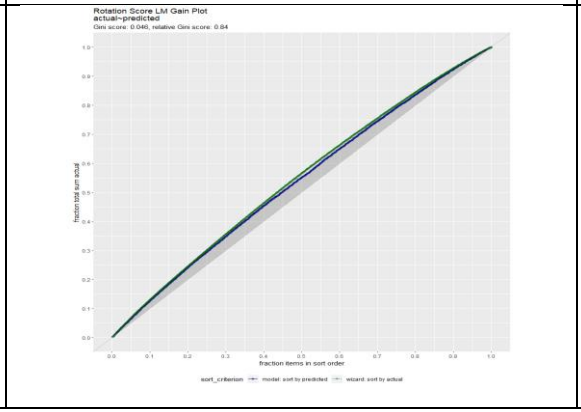

Rotation Score GLM Residuals
residuals
predicted

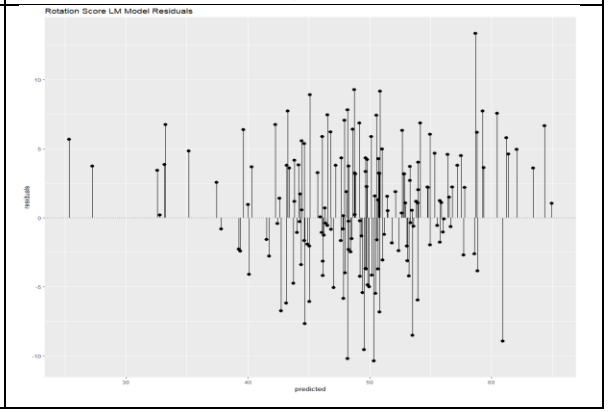

Rotation Score random forest
Peak_Hand_Speed_mph
Bat_Speed_mph
Early_Connection_Degrees
Connection_Score
On_Plane_Efficiency_Percent
Attack_Angle_Degrees
Plane_Score
Vertical_Bat_Angle_Degrees
Connection_At_Impact_Degrees
NA
5000
10000
15000
20000
Importance


Rotation Score Random Forest Gain Plot
actual~predicted
Gini score: 0.040, relative Gini score: 0.84
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual


Rotation Score Random Forest Residuals
residuals
predicted


Rotation Score lm
Peak_Hand_Speed_mph
Early_Connection_Degrees
Connection_At_Impact_Degrees
Attack_Angle_Degrees
Vertical_Bat_Angle_Degrees
Bat_Speed_mph
Plane_Score
On_Plane_Efficiency_Percent
Connection_Score
NA
0
5
10
15
Importance


Rotation Score LM Gain Plot
actual~predicted
Gini score: 0.040, relative Gini score: 0.84
fraction total sum actual
fraction items in sort order
sort_criterion
model: sort by predicted
wizard: sort by actual


Rotation Score LM Model Residuals
residuals
predicted

Connection Score

Examining the importance plot in the left most column of the Connection Score page, Power kW is one field that both the LM and GLM models agree on as important. Examining the Gain Curves, seemingly all models perform poorly in the observations of data clustered around the middle. This is because the Wizard Curve (green curve) separates from the curve derived from the specific model (navy curve), therefore the two do not match that well in the middle. In addition, examining the Residual plots in the Connection Score table, all three contain a higher proportion of higher residuals, such that the models are more often overestimating what true connection score is. Although both GLM and LM models agreed on Power kW generation (Bat Mass X Avg Acceleration X Bat Speed At Impact), these models are pretty poor and should not be used for a complete examination of Connection Score and sensitivities outside of Early Connection and Connection at Impact.

Plane Score

It is interesting that the Random Forest model and the LM model both agree that Connection Score is important. Something happened with the GLM model, and would require a deeper dive to see why the model only liked Time to Contact as an important variable for Plane Score. Both Gain Curves for the Random Forest and LM models look pretty good, and residuals are spread out evenly both positive and negative, although maybe a bit more toward positive. Examining model output and model diagnostic, it can be inferred that Plane Score is likely to be sensitive to Connection Score.

Rotation Score

The output from the three models for Rotation Score are the clearest. All models state that Peak Hand Speed is the most important variable for Rotation Score, and all Gain Curves look the cleanest. Residuals are evenly spread out and even better, the model diagnostic terms R^2, RMSE, and MAE are the best. From the three models, it is safer to state that Rotation Score is highly sensitive to Peak Hand Speed.