Tampa Rays' R&D Intern Candidate Data Project

Bosan Hsu

Introduction

The following methodology was used to project the true average speed-off-bat in a roster of baseball players. The code was written in Python on the Jupyter platform. Attached to this report are the procedures, including data cleaning, building machine learning models, etc., and graphics that help illustrate the findings.

Process

1. Import the Data

2. Data Preprocessing
    a. Remove Hit Type: "U"
        i. After importing the data, I applied the group-by method to see if there was an additional hit type in the data. There are two additional "U"s, and I couldn't find their definitions, so I decided to remove them.
    b. Organize and Re-define Hit Type
        i. The data project instruction states the definition of different hit types, but there are no definitions of line drives.
        ii. Data cases' speeds and angles that are outliers in their own hit type are removed. The hit types of remaining data cases were re-defined by MLB official guidelines.
        iii. Note:
            Ground ball: Less than 10 degrees
            Line drive: 10-25 degrees
            Fly ball: 25-50 degrees
            Pop-up: Greater than 50 degrees
    c. Remove NA (Null Value)
        i. Data cases without values in System A and B were removed.
        ii. Two data frames were created. One filtered out NA for System A (speed A and vertical angle A), and one filtered out NA for System B (speed B and vertical angle B.)
    d. Bootstrapping the Data
        i. First, batters with less than 30 data cases were filtered out.
        ii. Second, the batter with the greatest number of data cases was selected, and his/her number of data cases was set as a target number.
        iii. Third, all batters' data cases were resampled, and their numbers were all equal to the previous greatest number of data cases.

iv. Also, I compared the distribution of speed and angle between the bootstrapping data and the initial data, ensuring that the distribution was still the same.

v. Lastly, I graphed a scatter plot to see the relationship between speed and angle and find the angle that generates the highest speed.
(Note: System A data frame and System B data frame were resampled individually.)

3. Building Machine Learning Models
    a. Linear Regression Model
        i. Instead of building the model directly, a new column was added to the data frame. Since the scatter plot showed an inverted U-shape-like distribution, and, following baseball logic, the effect of hitting direction on speed is symmetrical, whether hitting from the upward or downward angle. Therefore, converting the angle to its absolute value can help the model better capture the relationship between speed and hitting direction.
        ii. After the calculation, the absolute value of the angle difference was set as an independent variable, and the speed was set as a dependent variable.
        iii. The data sets were split into the testing data (20%) and the training data (80%.) After the separation, the training data ran through linear regression to make a predicted formula for speed. Then, testing data was inputted into the formula.
        iv. Finally, I graphed the scatter plot of predicted and actual speed and showed the mean absolute error, mean squared error, and R squared value.
    b. Random Forest Model
        i. The angle was set as an independent variable, and the speed was set as a dependent variable.
        ii. The data sets were split into the testing data (20%) and the training data (80%.) After the separation, the training data ran through random forests, which consisted of 100 decision trees with a random state of 42, to make a predicted formula for speed. Then, testing data was inputted into the formula.
        iii. Finally, I graphed the scatter plot of predicted and actual speed and showed the mean absolute error, mean squared error, and R squared value.
    c. Neural Network
        i. The angle was set as an independent variable, and the speed was set as a dependent variable.
        ii. The data sets were split into the testing data (20%) and the training data (80%.) After the separation, the data was standardized to ensure that all

feature variables contribute evenly to the model's training, which can speed up the learning process and enhance model performance.

iii. the training data ran through neural networks, which consisted of 1 hidden layer with 100 neurons, with a random state of 42 for at most 1000 iterations in 0.001 learning rate to make a predicted formula for speed. Then, testing data was inputted into the formula.

iv. Finally, graph the scatter plot of predicted and actual speed and show the mean absolute error, mean squared error, and R squared value.

4. Calculating Posterior Mean

   Posterior Mean =
   (Prior Mean/Prior Variance + Sample Mean*Sample Size/Sample Variance) /
   (1/ Prior Variance + Sample Size/Sample Variance)

   a. Calculate the Posterior Mean of an Individual Batter's Speed
   b. Calculate the Posterior Mean of an Individual Batter's Angle

5. True Speed Prediction
   a. Prediction by Posterior Mean of an Individual Batter's Speed
   b. Prediction by Posterior Mean of an Individual Batter's Angle
      i. Input the posterior mean of an individual batter's angle into machine learning models to predict the true speed.
   c. A Function to Show the Result

Output

1. Initial Data
   a. There were 73,375 data cases, with 7,572 missing values in System A and 1,402 missing values in System B.
   b. The hit type variable contained 33,239 ground balls (45.30%), 18,166 line drives (24.76%), 16,722 fly balls (22.79%), 5,246 pop-ups (7.15%), and 2 Us (0.00%).

2. Data after Data Viewing, Cleaning, and Adjustments Before Bootstrapping
   a. There were 61,164 (83.34% of initial data) data cases in the System A data frame, containing 26,000 ground balls (42.51%), 16,973 line drives (27.75%), 15,662 fly balls (25.61%), and 2,529 pop-ups (4.14%).
   b. There were 66,402 (90.50% of initial data) data cases in the System B data frame, containing 28,897 ground balls (43.52%), 17,172 line drives (25.86%), 15,787 fly balls (23.77%), and 4,546 pop-ups (6.85%).
   c. The remaining number of data cases was considerable, and the distribution of different hit types was similar to the initial data. Thus, both data frames were representable for actual recorded data.

3. Data After Bootstrapping

    a. There were 181,030 (421 data cases * 430 batters) data cases in the System A data frame, containing 77,719 ground balls (42.93%), 49,372 line drives (27.27%), 46,297 fly balls (25.57%), and 7,642 pop-ups (4.22%).
    b. There were 205,530 (465 data cases * 442 batters) data cases in the System B data frame, containing 90,161 ground balls (43.87%), 52,241 line drives (25.41%), 48,809 fly balls (23.74%), and 14,309 pop-ups (6.96%).
    c. The distribution of different hit types in System A's data frame and System B's data frame were almost identical with previous data. Thus, both data frames after bootstrapping were representable for actual recorded data.
    d. The angle with the highest median speed in System A was 3.1164 and was 2.1434 in System B. The scatter plots of the two data frames were like an inverted U shape.

4. Machine Learning Model
    a. Linear Regression:
        i. For System A: The R-squared score was 0.0943.
           The Mean Absolute Error was 9.6544.
           The Mean Squared Error: 145.7091.
        ii. For System B: The R-squared score was 0.0181.
           The Mean Absolute Error was 13.7608.
           The Mean Squared Error: 277.5841.
    b. Random Forest:
        i. For System A: The R-squared score was 0.7374.
           The Mean Absolute Error was 3.0201.
           The Mean Squared Error: 42.2244.
        ii. For System B: The R-squared score was 0.8626.
           The Mean Absolute Error was 2.8451.
           The Mean Squared Error: 39.7546.
    c. Neural Network:
        i. For System A: The R-squared score was 0.1533.
           The Mean Absolute Error was 9.3087.
           The Mean Squared Error: 136.1431.
        ii. For System B: The R-squared score was 0.5188.
           The Mean Absolute Error was 9.2391.
           The Mean Squared Error: 139.2099.
    d. Best Model:
        i. For both System A and B, random forest generated the highest R-squared score, the lowest mean absolute error, and the mean squared error. This indicated that the random forest model may be the best model of the three models I used when predicting the speed if the dependent variable only contains angles.

        ii. Also, the values of the mean absolute error of random forest were small enough when it came to speed. An error within about 3 mph/h is acceptable.
        iii. By visual inspection, the curve generated by the neural network model closely resembles the actual distribution. When there are more independent variables, the model's ability to fit complex data distributions can be enhanced, improving its performance.

5. True Speed Prediction
    a. For batters with data cases measured by System A, their predicted performance only used data from System A for posterior mean calculation and machine learning to predict future performance.
    b. For batters without data cases measured by System A but with data cases measured by System B, their predicted performance only uses data from System B for posterior mean calculation and machine learning to predict future performance.
    c. For batters without any data cases or less than 30 data cases, their predicted performance was not estimated due to limited sources.

Findings and Thoughts

1. Observations by System A and B both showed a left-skewed distribution of speed. A reasonable explanation is that batters are capable of making the exit velocity fast but will ultimately hit a limit due to biological constraints.
2. The angle observations were normally distributed for System A but not for System B. This may support the suspicion that System A is much more accurate. The batters aim to hit the ball far, which may require aiming to hit at the right angle to generate more speed and power. Since they can't always hit the sweet spot, the angle would vary, and the distribution of angles should be normally distributed.
3. The graph of the random forest prediction showed a high degree of the actual speed dots' and the predicted speed dots' co-occurrence. Also, the random forest generated the highest R-squared score, the lowest mean absolute error, and the mean squared error of all three models. Lastly, the values of the mean absolute error of random forest were small enough when it came to speed. An error within about 3 mph/h is acceptable.
4. Despite the high mean squared error (MSE) of the neural network model, the predictions demonstrated a clear trend path that aligned well with the actual data. The reason might be that the direction of the predictions is correct, but the scale of the predicted values is off. Thus, when it comes to predicting the trend, using a neural network model would be great. When it comes to predicting a speed by a specific angle, using a random forest model would be great.
5. The predicted speed based on the posterior mean of speed and inputting the posterior mean of angle into the random forest model might not always be a close number. A potential reason for the circumstance was that a batter with an average strength might have a lesser speed difference using different predictions, and those who have a superior strength or a weaker strength might have a greater speed difference. For

example, a batter with superior strength can hit the ball faster than others at the same angle, so his posterior mean of speed may be higher than his predicted speed based on the posterior mean of angle. Thus, by comparing the two types of prediction, we may estimate the batter's strength.

Future Improvement

1. During the project, I attempted to resample and increase the data cases for each pitcher, similar to what I did for the batters. However, the system could not perform the bootstrapping function due to the increased size of the data. Had the function been successful, I would have been able to show how different pitchers affect the performance of batters. This is particularly relevant for quality pitchers who are known for inducing batters to hit balls with lower exit velocities or at extremely low or high angles. If a batter still manages to perform well against such pitchers, he/she can be considered a slugger.
2. I also tried to develop machine learning models specifically tailored to individual batters to predict their performance more accurately. Unfortunately, the system was unable to handle the large amount of data.
3. Also, predicting a batter's fluctuation is a must. Batters with a more consistent performance may benefit the team more.
4. Other than different pitchers, I'd like to collect data on the pitch types, pitcher's and batters' dominant hands, batters' swinging speed, pitchers' pitching spin rate, player injury status, and the wind direction. Models may perform better if there are more reliable independent variables.
5. A time series model might improve the accuracy of the prediction, as players would improve themselves or be aged.
6. Last, research on potential interaction variables would be necessary, and there are more machine learning models and deep learning models that I should learn about.

Summary

The data project aimed to project the true average bat speed in a list of baseball players using Python on the Jupyter platform. The methodology included importing the data, performing data viewing, cleaning, and adjustment, bootstrapping the data, building machine learning models (including linear regression, random forest, and neural network models), calculating the posterior mean, and making true speed predictions.

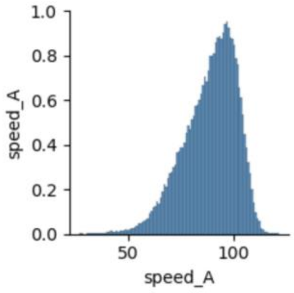
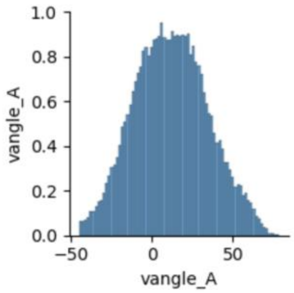
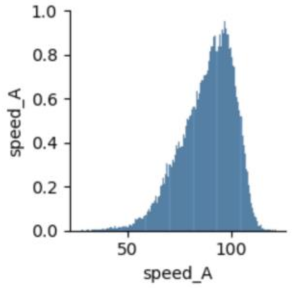
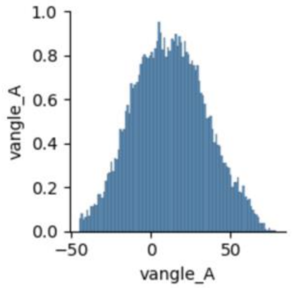
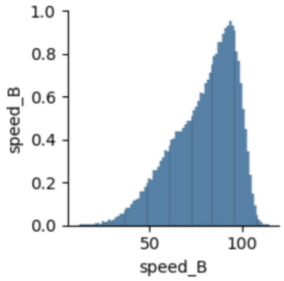
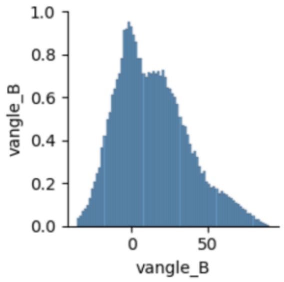
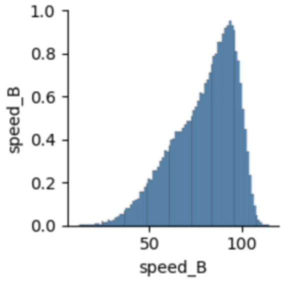
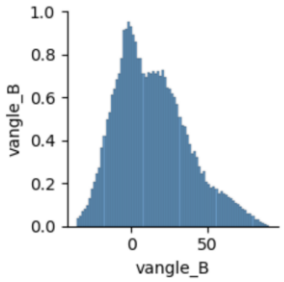
The initial data contained 73,375 data cases, which underwent various cleaning and adjustment processes. After bootstrapping, the data frames for System A and System B were representative of the actual recorded data. The best-performing model was the random forest, which produced the highest R-squared value and the lowest mean absolute error and mean squared error for both systems.

The results showed that an angle interval between 0 and 20 generates a faster exit velocity. Future improvements include exploring additional independent variables such as pitch types and player injury status, implementing time series models to improve prediction accuracy, and exploring potential interaction variables.

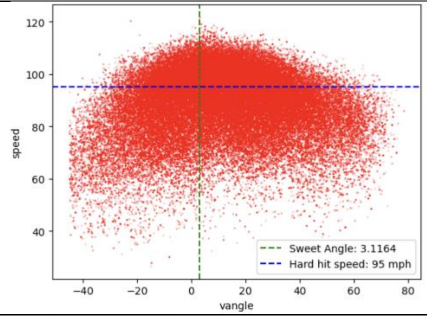
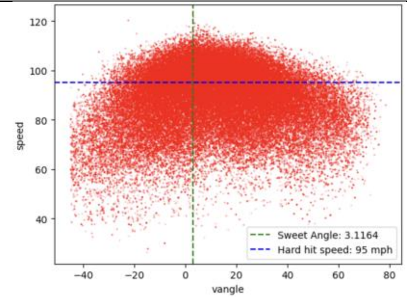
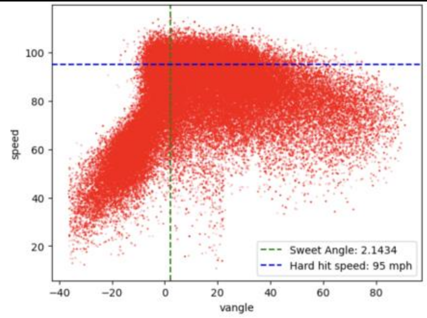
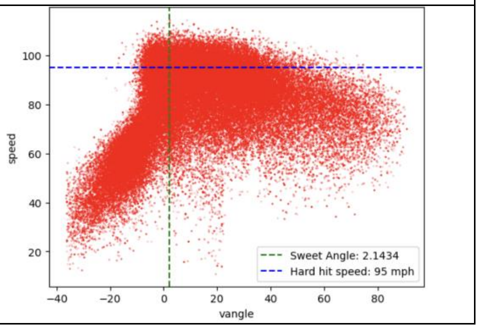
Overall, the project provided valuable insights into the prediction of baseball player performance and identified areas for further research and improvement.

Appendix:

Data Distribution Before & After Bootstrapping

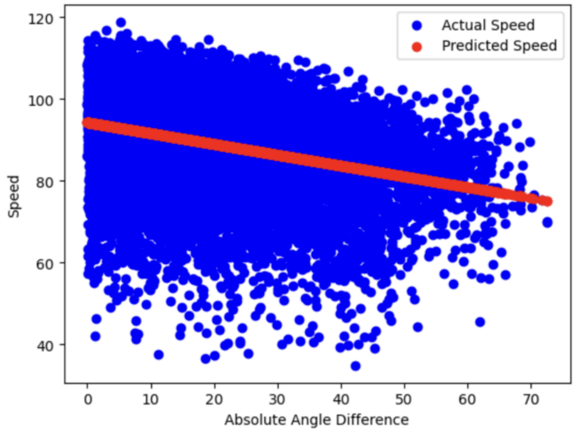
	Data Before Bootstrapping	Data After Bootstrapping
System A	 	 
System B	 	 

Scatter Plot – Speed x Angle Before & After Bootstrapping

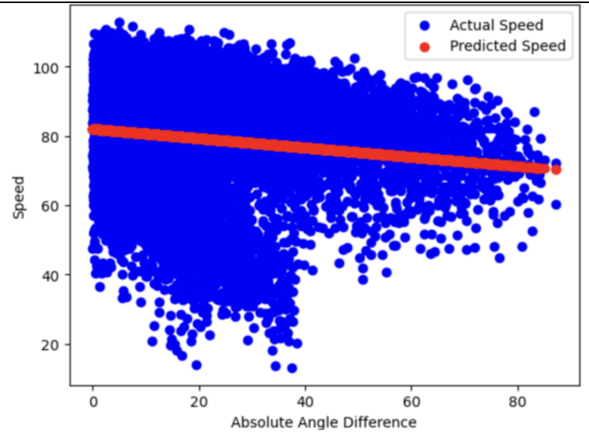
	Data Before Bootstrapping	Data After Bootstrapping
System A		
System B		

Actual Values x Predicted Values by Machine Learning

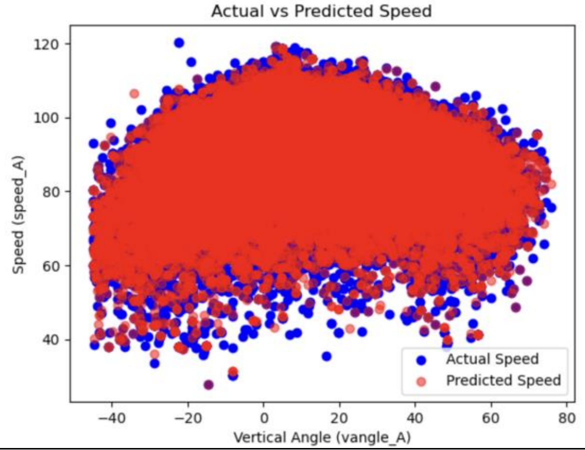
System A – Linear Regression



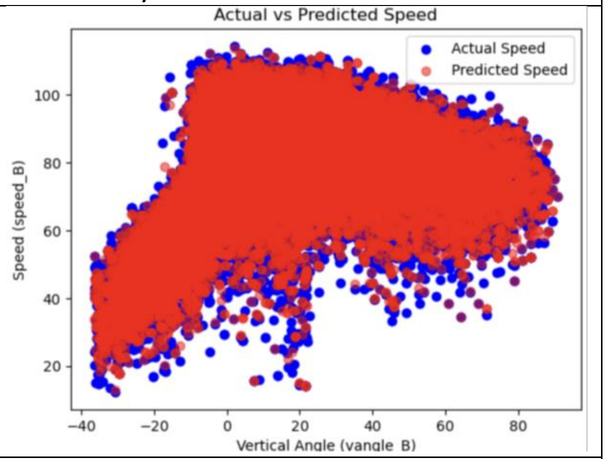
System B – Linear Regression



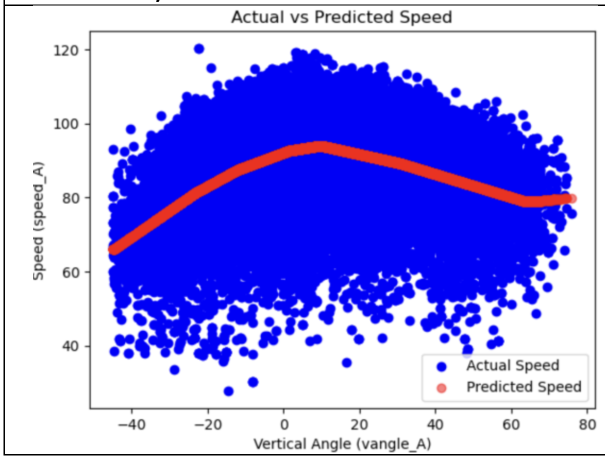
System A – Random Forest



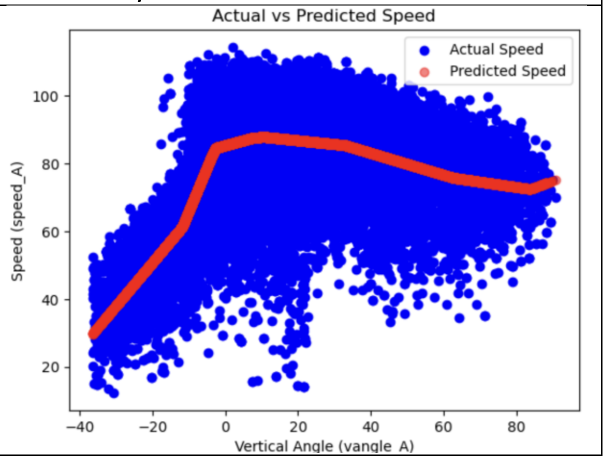
System B – Random Forest



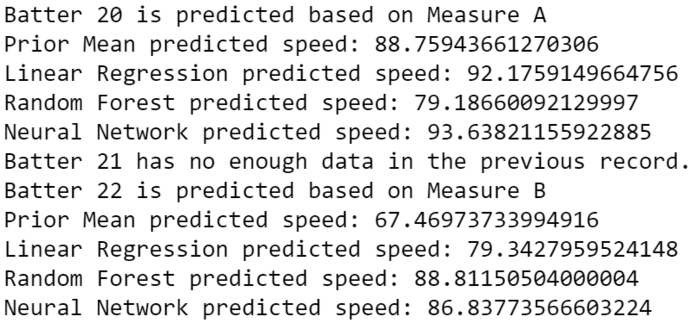
System A – Neural Network



System B – Neural Network

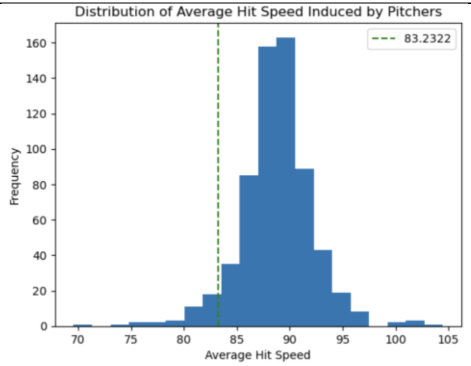
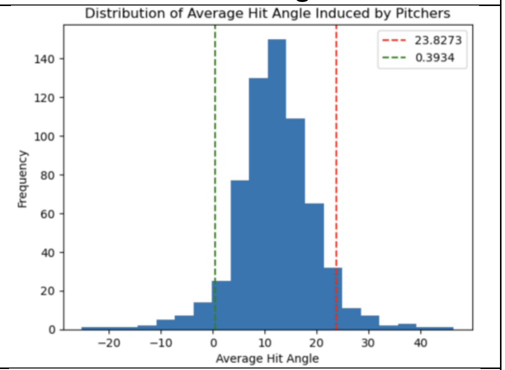
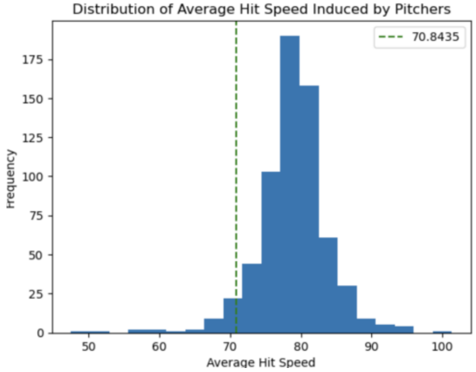
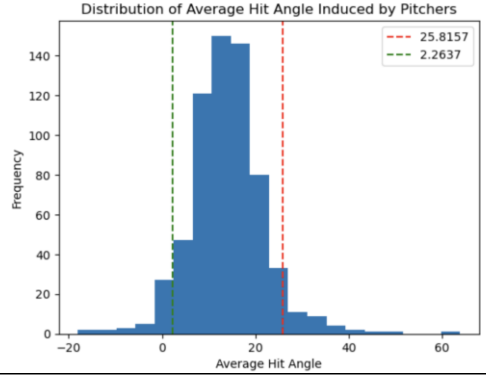


Predicted Outcome Example (Batter 20, 21, 22)



```
Batter 20 is predicted based on Measure A
Prior Mean predicted speed: 88.75943661270306
Linear Regression predicted speed: 92.1759149664756
Random Forest predicted speed: 79.18660092129997
Neural Network predicted speed: 93.63821155922885
Batter 21 has no enough data in the previous record.
Batter 22 is predicted based on Measure B
Prior Mean predicted speed: 67.46973733994916
Linear Regression predicted speed: 79.3427959524148
Random Forest predicted speed: 88.81150504000004
Neural Network predicted speed: 86.83773566603224
```

Quality Pitcher Example

	Lower Speed	Extreme Angles
System A		
System B	 Distribution of Average Hit Speed Induced by Pitchers 70.8435 Frequency 0 25 50 75 100 125 150 175 50 60 70 80 90 100 Average Hit Speed	 Distribution of Average Hit Angle Induced by Pitchers 25.8157 2.2637 Frequency 0 20 40 60 80 100 120 140 −20 0 20 40 60 Average Hit Angle