# Baseball Analytics: There's Math There?

Joe Datz, John Juozitis, Hugh McMurray, Elyssa Pollio, Sam Smallwood, Jeffrey Wheeler (Faculty)

Math 1103 - BIG Problems, Dept. of Mathematics, University of Pittsburgh

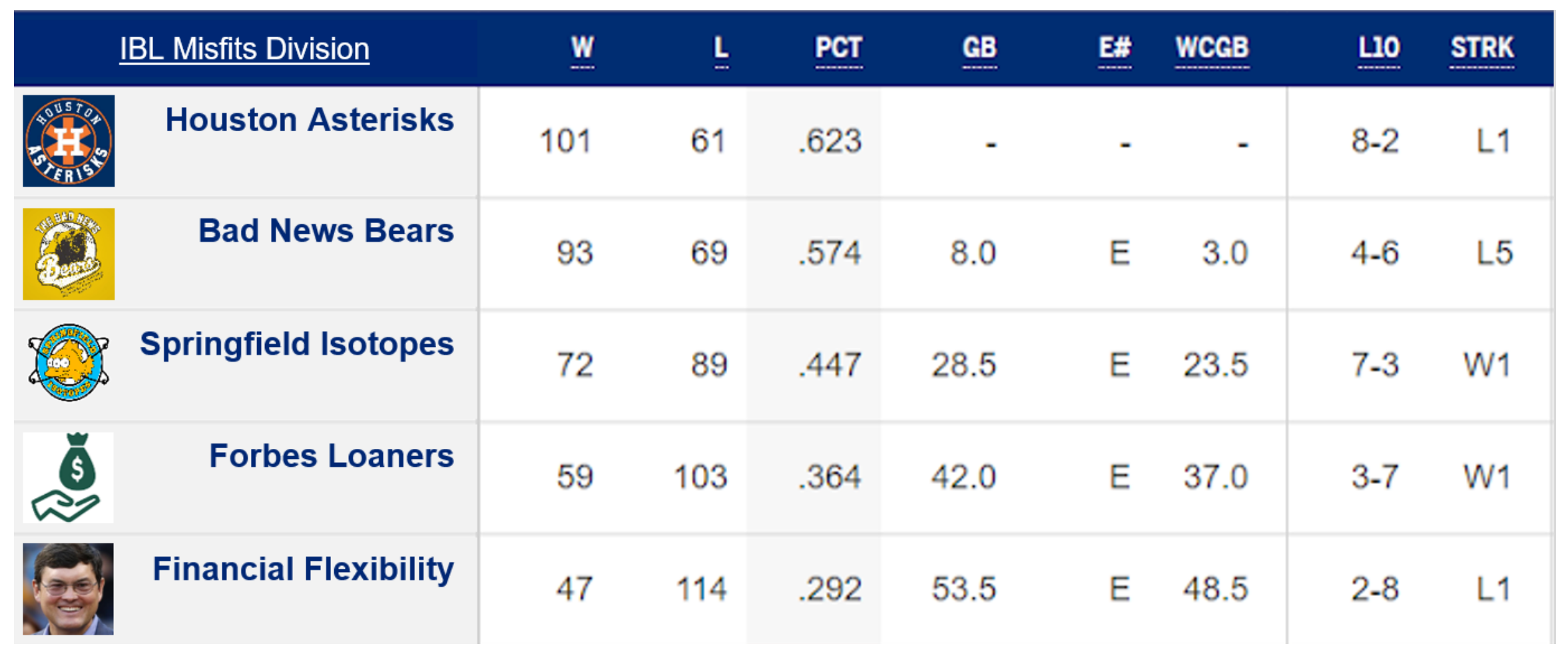
## The Problem

Is it possible to predict a player's ability to play at a particular position? Is it possible to predict when a player needs to switch to easier positions?

## The Forbes Loaners:

**Forbes Loaners** is a professional baseball team based in South Oakland, Pittsburgh, PA that is a part of the Imaginary Baseball League (IBL).

- Founded by Big Problems$^{TM}$
- Small-Market Team which is hoping to adopt data analytics into their decision-making process.
- Historically has been unable to win their division due to the fiscal advantage of other teams.
- Struggling financially due to their Coronavirus small business loan going to Shake Shack.
- Preference towards a financially flexible approach to avoid any long-term damage from future risks.



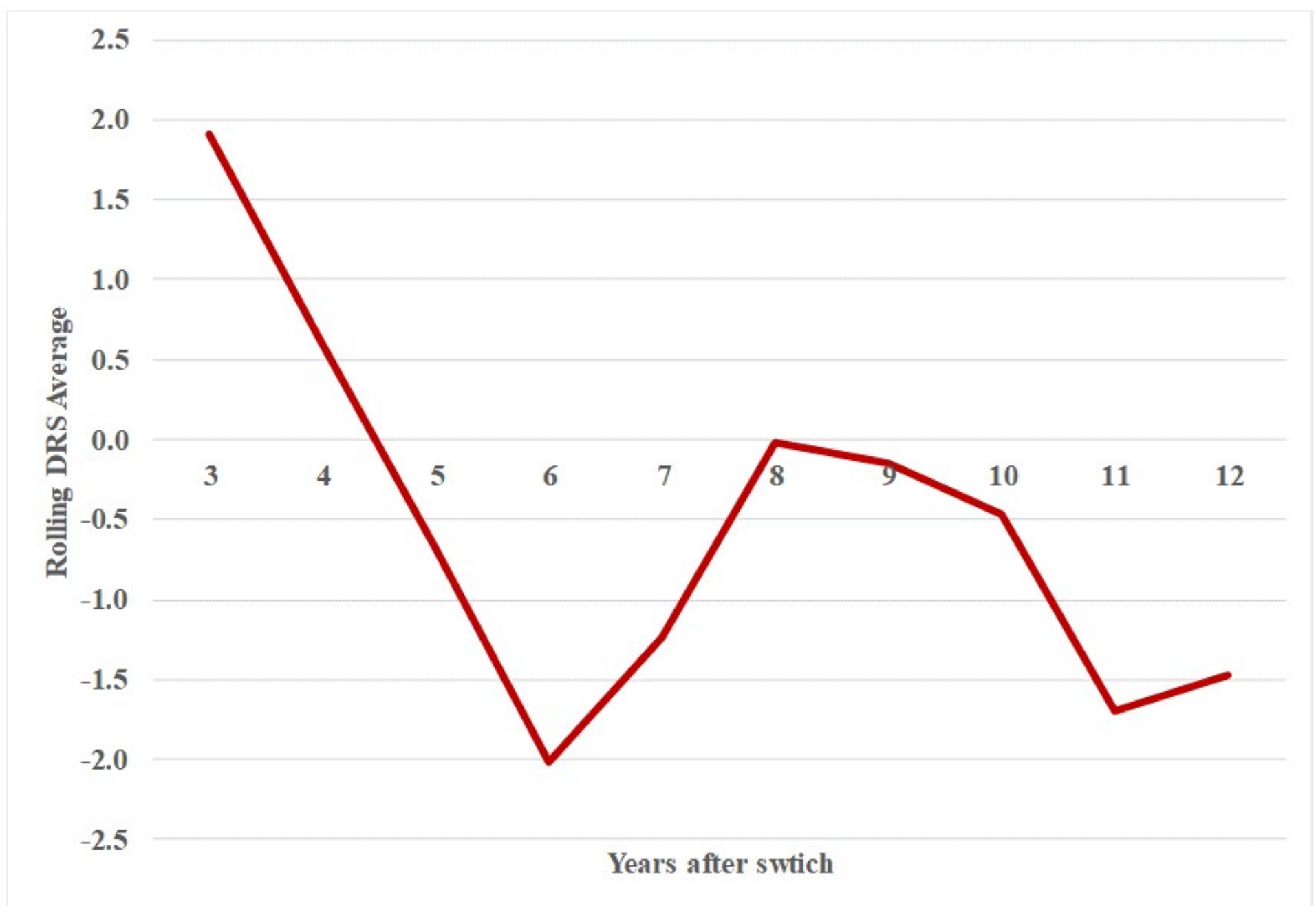
| IBL Misfits Division | W | L | PCT | GB | E# | WCGB | L10 | STRK |
|---|---|---|---|---|---|---|---|---|
| Houston Asterisks | 101 | 61 | .623 | - | - | - | 8-2 | L1 |
| Bad News Bears | 93 | 69 | .574 | 8.0 | E | 3.0 | 4-6 | L5 |
| Springfield Isotopes | 72 | 89 | .447 | 28.5 | E | 23.5 | 7-3 | W1 |
| Forbes Loaners | 59 | 103 | .364 | 42.0 | E | 37.0 | 3-7 | W1 |
| Financial Flexibility | 47 | 114 | .292 | 53.5 | E | 48.5 | 2-8 | L1 |

**About Sabermetrics** Although there were earlier writers, the start of the Sabermetrics movement is often accredited to Bill James' *Baseball Abstracts* first being released in 1977. Sabermetrics is defined as "the search for objective knowledge about baseball." The first teams to adapt it were small market teams in the early 200s. Now, all teams use Sabermetrics in one form or another in hopes to gain a competitive advantage.

## Methods Used

**Data Wrangling** The online baseball resources *Retrosheet*, *Lahman*, and *Fangraphs* were used to guide model development. Some of the following variables from *Fangraphs* were used:

- DRS
- Age
- Position
- Innings/Game



**DRS Modeling** The statistic **D**efensive **R**uns **S**aved was initially used for a single-variable prediction model. Java was used to calculate 3-year rolling averages of DRS after a player had switched positions (shown above). Similarly, Excel was used for a prediction model based on DRS that was adjusted for the frequency of player switches. The final probability model for outfielders consisted of:

$$Pr(w_{ab}, DRS) = (0.33)^{1-x}(0.66)^{x}w_{ab},$$

$$x = \begin{cases} 1 & DRS \leq -1.28 \\ 0 & DRS > -1.28 \end{cases}$$

Where $w_{ab}$ represents a frequency weight starting at position $a$ and moving to position $b$. Between both the Java and Excel models, the prediction rate for outfielders hovered around 66%.

## Tree Models

**Random Forests** Results on prediction accuracy of a player's need to switch positions were obtained using Random Forests and comparable machine learning techniques. Focusing on Random Forests allowed us to obtain results about the significance of variables in the dataset X.

## How it works

A **Decision Tree** is a series of if-then logical statements over a dataset X to classify points of data into distinct classes in Y. To decide how the if-then statements are constructed, we use two equations:

$$E(Y) = \sum_{i=1}^{c} p_i log_c p_i$$

$$G(Y, x_i) = E(Y) - \sum_{s \in values(x_i)} \frac{|Y_s|}{|Y|} E(Y_s)$$

$E(Y)$ represents **Entropy** in information theory and provides an informal sense of disorder in an information system given by classes $c$. $G(Y, x_i)$ represents **Information Gain** of a particular variable $x_i \in X$, and tells us how much each variable reduces the entropy in the system.

To pick the first if-else statement, Information Gain is computed for all variables. The first if-else statement is given to whichever variable has the highest value $G(Y, x_i)$. If this variable doesn't classify all data points, we continually use $G(Y, x_i)$ to add leaves to the tree until we can.

To reduce the problem of variance in decision tree models, **Random Forests** are used instead. This involves training many Trees on both random subsets and variables of X, and make a prediction based on majority vote of the trees.

## Results

Comparing additional models produced the following results on accuracy:

| Method | Acc. |
|---|---|
| Bagging | 71.7% |
| Random Forest | 73.0% |
| Logistic Regression | 71.1% |
| Linear Disc. Analysis | 72.0% |
| Support Vector Machine | 74.8% |

Additionally, the Random Forest model allowed us to find that the most important stats for predicting player positions were Putouts, Assists, the *Def* statistic from *Fangraphs*, and Age.

## Conclusion

After Modeling the Data using the machine learning process, we found a model with a 73% classification rate. We learned how to work with variables, how to clean data, and identify trends.

## Technologies Used/Skills Learned

• Python • SQL • R/Rstudio • Excel • Data Cleaning • Web Scraping • Random Forest • Support Vector Machines • LD Analysis

## Acknowledgements

Special thanks to Dr. Jeffrey Wheeler, Dr. Andrew Walsh of Health Monitoring Systems, and Dr. Sam Ventura of the Pittsburgh Penguins for their assistance and guidance. Additionally we extend our appreciation to the University of Pittsburgh Department of Mathematics for offering this course and providing the resources that aided in this project's success.

University of Pittsburgh