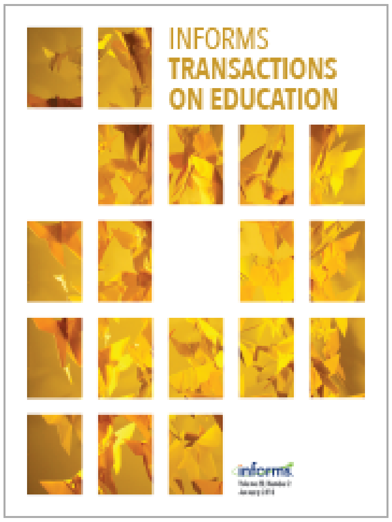

INFORMS Transactions on Education

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Case Article—Baseball Analytics: Advancing to Prescriptive Analytics in the Major League Baseball Front Office

Sean L. Barnes, Margrét V. Bjarnadóttir

To cite this article:
Sean L. Barnes, Margrét V. Bjarnadóttir (2019) Case Article—Baseball Analytics: Advancing to Prescriptive Analytics in the Major League Baseball Front Office. INFORMS Transactions on Education 19(3):143-145. https://doi.org/10.1287/ited.2018.0201ca

informs
http://pubsonline.informs.org/journal/ited/

INFORMS TRANSACTIONS ON EDUCATION
Vol. 19, No. 3, May 2019, pp. 143–145
ISSN 1532-0545 (online)

Case Article—Baseball Analytics: Advancing to Prescriptive Analytics in the Major League Baseball Front Office

Sean L. Barnes,[a] **Margrét V. Bjarnadóttir**[a]

[a] Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742-1815

Contact: sbarnes@rhsmith.umd.edu, http://orcid.org/0000-0001-5497-6277 (SLB); margret@rhsmith.umd.edu, http://orcid.org/0000-0003-2955-1992 (MVB)

Received: October 27, 2017
Accepted: May 28, 2018
Published Online in Articles in Advance: May 20, 2019

https://doi.org/10.1287/ited.2018.0201ca

Abstract. This case uses player evaluation and personnel decision making in Major League Baseball (MLB) to introduce many of the key steps of data analytics projects. The data analytics process is a unique combination of art and science, and teaching the art of data analytics is challenging to do in a standard classroom setting with small data sets. The goal of this case is to move beyond the simple "cookie cutter" data sets and introduce students to the challenges of dealing with real data to answer important questions, as well as introduce or reinforce multiple data-mining/machine-learning methods. The case builds on a very rich data set collected by the authors, which allows for students or groups of students to arrive at different answers to the same question.

Supplemental Material: An extensive teaching note with sample questions and the associated R code and output are at https://doi.org/10.1287/ited.2018.0201ca.

Keywords: data analytics • prediction modeling • machine learning • sports analytics • baseball • descriptive analytics • predictive analytics • prescriptive analytics

Introduction

The data analytics process includes multiple steps, from defining the purpose of the project and acquiring the appropriate data, through data cleaning, data exploring, and the application of data-mining and machine-learning methods. In the end, if the analysis is successful, the process can lead to useful insight and knowledge or the deployment of the analytical model to support decision making in practice. Although many of these concepts can be introduced through lecture and with tidy, small-scale data sets, real-world data offer a potentially richer learning experience. A successful data analytics project requires decisions to be made about variable selection, data modeling, and cleaning. For each of these cases, there is no one right answer; therefore, decisions must be made on the basis of appropriate data analysis principles. The goal of this case is to introduce some of these real-world data analytics lessons to students that are difficult to teach through lecture and are better taught through hands-on exercises.

This case is flexible enough to be tailored to any program level in business, economics, computer science, mathematics, or engineering (including operations research and management science), although it is best suited for undergraduate and masters levels (both MBA and specialty MS). The case emphasizes the data analytics process, and the depth of the analysis can be adjusted (i.e., from very structured to open-ended) to reflect the capabilities of the students and the desired time to be allotted to the case.

Overview of the Case

Sports analytics, and baseball analytics in particular, have gained popularity over the past two decades but go as far back as the 19th century, when newspapers began publishing team and player statistics. In recent years analytics have been introduced to all aspects of the game, including roster and lineup construction, defensive positioning, pitcher management, and increased emphasis on more nuanced aspects, such as launch angles and route efficiencies. This case focuses on the analysis of individual player data and how it can be used to improve player acquisitions in the free agent market. It provides a roadmap for students to navigate the process of exploring data, developing explanatory and prediction models, and leveraging these models to support decision making. The data contain individual player and team statistics, as well as contract information for all MLB free agent batters from the 1998 season through 2013.

The case is divided into four parts: part A, data exploration and visualization; part B, explanatory modeling; part C, predictive modeling; and part D, prescriptive analytics. Parts A through C of the case can be assigned independently, whereas part D requires students to have built at least one model to explain/predict salaries. Each part can be assigned either as an individual or team exercise or assignment and offers a variety of opportunities for analysis. It has also been used to introduce specific methods during class. The latter parts can be assigned without the students working through part A, although it is the authors' hope that part A will be emphasized owing to the importance of data exploration for any data analytics project.

Depending on the level of the students, the case lends itself very nicely to unstructured study (e.g., develop the best prediction model), or the analysis can be guided through more structured study questions (e.g., fit a linear regression model using backward elimination). The different parts of the case could be assigned at different points throughout a course—eliminating the time it takes for students to understand a new data set for each assignment—or the case can be used as a final project.

Teaching Objectives

The goal of the case is to provide an opportunity for students to analyze real-world data, which includes the following specific aims:

- Emphasize data cleaning and exploration, bringing to the forefront the time this process requires, the value of developing in-depth familiarity with the data, and the importance of how data exploration can support model building
- Clarify the difference between explanation and prediction and emphasize why the same model may not be the right model for both purposes
- Focus on thoroughly analyzing the model output—for example, extracting insights and understanding variable importance—and translating new knowledge to relevant stakeholders
- Demonstrate how to advance from predictive to prescriptive analytics, that is, how we apply our models to improve decision making

In addition, this case can be used to

- Discuss interpretation of variable coefficients in a high-dimensional environment in which many variables are highly correlated, and in addition, discuss statistical significance versus variable importance
- Introduce variable selection methods, their importance when data are strongly correlated, and their limitations
- Introduce or practice different methods for training prediction models (e.g., linear regression, regression trees)
- Address the question of what makes for a good model and explain why there may not be a single model that is the best
- Introduce data modeling and visualization with R (or another suitable programming language or software package)

After completing the case, the students will be familiar with the key steps of a data analytics project and will have been exposed to data analysis functionality in R (or the software/language of the instructor choice).

Teaching Suggestions

The case can be used either as an assessment for credit or as an in-class activity. Students can work on the case independently, in groups, along with the instructor, or a mixture of these options.

If used as in-class exercise, the case can be scaled down depending on the available time. Because of the richness of the case, we suggest devoting at least one class session to each selected part of the case, especially if the associated data-mining methods are being introduced concurrently. Alternatively, part of the case can be assigned as homework, and key parts of the assignment could be discussed in class. For example, instructors could require students to build their best predictive model for homework (as required for part C) and provide predictions for a held-out test set. Then, the instructor could report error measures in class and discuss the best approaches along with other relevant insights.

The case can also be assigned in its entirety as a team project in a data analytics class, and devote the last class session to student presentations and discussion. This approach offers an opportunity for students to present their approach to an audience familiar with the problem and will also highlight the fact that different teams will take different approaches. At the instructor's discretion, all teams could be required to present, or with advance submissions, a representative (and diverse) sample of projects could be selected to be presented during class.

Preparing for In-Class Use

If the case is assigned as an in-class exercise, we recommend posting the data before class and requiring students to complete some initial analysis in advance, so that they are better prepared to participate, which will improve the teaching experience. For example, in part A—which focuses on data exploration—students can be asked to calculate a correlation matrix and/or create a simple scatterplot (depending on whether these concepts have been introduced) as a preclass activity. If the students are required to complete this exercise before class, it will ensure a minimal level of familiarity with the data and result in a more fruitful discussion in class.

"Warm calling"—for which a specific student is notified in advance that he or she will have to present their analysis—could also be utilized to enhance the experience. Leveraging online discussion boards or other collaborative technologies also facilitates higher student engagement, for example, by requiring students to post their best visualization or discuss relevant insight gleaned from data visualization or model building. Similarly, one could require students to run a simple linear regression model before discussing part B, and so on.

Classroom Experience

The case has been used in its entirety over two class periods in a data analytics course for master's level engineering students. The data file was posted before class, and in-class time was provided for the students to conduct their analysis. Because of only using two class periods, the complete extent and depth of the case was not fully explored. The instructor identified the broad applicability of the case as its strength—that is, it provides an opportunity for students to navigate the entire analytical process from exploring the data to interpreting the results—commenting that it was the "perfect transition between the classroom and the real world." Some of the strengths of the case are the variety of techniques for the students to explore, and the case helped the students gain experience with making nontrivial modeling decisions. The students especially enjoyed learning more advanced features in R and running R-scripts, which are beyond what you find in basic text books.

Parts of the case have also been used in other courses. We have used part A of the case extensively to introduce data visualization in MBA and specialty MS courses. The students have enjoyed learning features in R that facilitate data visualization. For example, we introduce the *identify* function that allows the user to select and annotate individual data points in a visualization, which is difficult to perform in Excel. Part C of the case has also been used in an MBA course to teach students how to build effective predictive models.