Data Mining in Sport Management: Random Forests and the Hockey Hall of Fame

Brian Mills, University of Florida
Steve Salaga, Florida Institute of Technology

Research/statistical methodology | **Saturday, June 1, 2013** | **20-minute oral presentation (including questions)**
Abstract 2013-281 | **1:35 PM** | **(Room 416)**

Background & Justification

Rapid advances in computing power and network integration have allowed the storage of enormous data sets within the sports industry and beyond. Data mining and analytics—the attempt at finding hidden patterns or irregularities in data—have become commonplace within many sectors of the sports industry. Websites like Kaggle.com have shown the increasing popularity of these techniques for use in a number of problems, with the Netflix Prize perhaps the most prominent data science competition. Many of these problems are relevant to sport management practitioners. Data science practices are most leveraged within marketing and consumer relationship management (CRM), looking to make use of large databases containing information on customers. In this pursuit, analysts may use previous purchasing behavior, responses to surveys and opt-ins for newsletters, and even social media data to predict future purchase behavior and optimize the experience for their customers. Additionally, with the publication of Moneyball (Lewis, 2003), the world of sport has seen an explosion in the use of simulation and machine learning methods for evaluating on-the-field performance.

The permeation of quantitative analysis makes clear the importance in preparing both undergraduate and graduate students for understanding its use in a variety of careers within the sports industry. While the responsibility of analyzing data and building databases usually falls upon those with computer science and statistics degrees, equipping sport management students with a basic understanding of the concepts involved could prepare them to use these techniques in making important management decisions. Sports employers in the age of data may even expect sales representatives to efficiently use information produced from advanced data analysis in order to sell tickets and products to targeted customers.

While data mining practices can be somewhat controversial in academic research—namely, with respect to interpretation of statistical significance and lack of guiding theory—we note that these techniques can still be useful for both academics and practitioners alike. In this presentation, we highlight the use of Random Forests for an interesting classification problem. While Random Forests are often used in biological research—for example, gene classification—the lessons from this literature can be applied to classification problems in sports as well. Random Forests implement boosting, bagging and bootstrap procedures in order to make classification predictions from training data. The method consists of building an ensemble of decision trees in order to predict the class of individual observations, hence the name "forest." One important advantage of the method is that it allows for an evaluation of the relative impacts of each input variable on the classification decision. In the case of data mining, this is not always possible with neural network techniques.

Our Application

We apply the Random Forests classification algorithm to a topic of substantial interest in the sports literature – namely, have French-speaking players been discriminated against in the professional hockey labor market? The existing literature is ripe with contributions which both support (Lavoie, Gernier & Coulombe, 1987; Longley, 1995; 2003) and fail to support (Jones & Walsh, 1988; McLean & Veall, 1992) the existence of discrimination against French-speaking players. Based on these mixed and inconclusive results, we extend this research to the subjective voting involved in Hockey Hall of Fame inductions. While Hall of Fame inductions differ from the traditional hockey labor market, induction often brings opportunity of new employment and sponsorship.

We predict National Hockey League (NHL) player induction using training data from historical NHL statistical records and previous inductions. Our approach closely follows that of Frieman (2010) and Mills and Salaga (2011). However, we expand on the existing literature by taking a longitudinal approach to evaluating possible

discrimination as opposed simply evaluating the short-run potential of the behavior. We use the statistical package "randomForest" (Liaw & Weiner, 2002) within the R statistical environment in order to complete the classification task.

Utilizing NHL performance data for forwards and defenseman dating back to 1968, our Random Forests algorithm produces an out-of-box error rate of 2.04%. In other words, the methodology correctly predicted 97.96% of Hall of Fame inductees within the training data set. When evaluating these results for potentially discriminatory voting practices, we find no evidence of language-based discrimination. This is confirmed through the utilization of multi-dimensional scaling techniques which allow for the ability to visually inspect players that are on the margin for Hall of Fame induction. These longitudinal results support the previous work of Jones and Walsh (1998) and McLean and Veall (1992). This result may also provide support that the discriminatory practice could be coming from customers (Longley, 2003), rather than those in positions of power in the world of hockey. Finally, we use the decision rules to make predictions about players that will be eligible for induction in the future.

Limitations and Future Work

While data mining methods are often used for large data sets, Random Forest requires a significant amount of computing power in order to run its procedure. This is, of course, dependent on the number of trees grown and size of the data set. While the R statistical environment is an invaluable resource for statistical computing, it does have limitations with data storage and speed. We will address further shortcomings of the method and software in this talk, and expand on other options available to the sport management researcher within the realm of data mining. These include, but are not limited to, cluster analysis, discriminant analysis, machine learning, support vector machines, and neural networks.

We hope to encourage sport management researchers to explore the world of data mining and neural networks in their research. While the methodology is constantly changing, many of these techniques are both free and easily accessible in the R statistical computing environment. Faculty experience with these tools can also help to relay to sport management graduates an understanding of the techniques available, and will prepare them in providing significant value to their employers as data science practices continues to permeate the industry.