Journal of Applied Sport Management

Volume 7 | Issue 2 Article 11

1-1-2015

Sport Analytics: Using Open Source Logistic Regression Software to Classify Upcoming Play Type in the NFL

Robert E. Baker

Ted Kwartler

Sport Analytics

Using Open Source Logistic Regression Software to Classify Upcoming Play Type in the NFL

Robert E. Baker
Ted Kwartler

Abstract

The purpose of this study was to utilize data analytics as means to classify National Football League offensive play types. The open source software R was employed to create a logistic regression based on data for the Cleveland Browns and Pittsburgh Steelers from 13 recent seasons. The regression is based on all first, second, and third downs within regulation play, totaling 26,310 data points. The initial algorithms classify rush or pass for each offense. Revealed through differing coefficients of the independent variables, each team shows a slightly different approach to play selections in response to in-game situations. Identifying the driving factors to play selection is possible by isolating each attribute within the regression. Further examination could yield improved precision to control for changes in head coach, offensive coordinators, player personnel and other factors such as weather because these may influence play type. Logistic regression shows promise as an in-game aid to determining opponent behavior. Specifically, Cleveland's offensive play selection algorithm was correct for 66.4% of plays versus 66.9% for Pittsburgh. Use of open source software and logistic regression of NFL play selection could be beneficial in aiding future game decisions. Further research is recommended to explore possible improvement of the algorithm accuracy.

Keywords: *sport analytics; sport management; data mining; NFL; regression*

Robert E. Baker is an associate professor in the Department of Sport and Recreation Studies at George Mason University.

Ted Kwartler is Director of Customer Success at DataRobot.

Please send correspondence to Robert Baker, rbaker2@gmu.edu

Introduction

Sport is big business. This top-10 segment of the global economy is estimated at $440 to $470 billion in North America alone (Fry & Ohlmann, 2012; Plunkett Research, 2014). Reflecting this economic position, sport organizations parallel business processes in any major economic segment. Decisions in sport organizations are increasingly informed by, and derived from data analysis (Andrew, Pedersen, & McEvoy, 2011). The use of statistical analysis is an essential component in data driven decision making. Sport is a rapidly growing arena for the application of analytics (Fry & Ohlmann, 2012). The essence of sport analytics includes managing data, using predictive analytics, and informing decision makers to provide a competitive advantage (Alamar, 2013). This reveals a continual analysis process in sport settings wherein situation-specific information informs analytical algorithms, which in turn guide data collection and analysis. Once analyzed, leaders employ data in decision making, which yields results that feed back into situational analyses (see Figure 1).

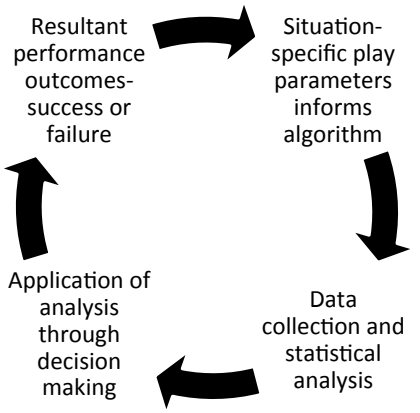

Figure 1. Continuous Analysis in Sport

Moneyball (Lewis, 2004) yielded perhaps the most visible example of the application of sport analytics. It is the story of the management strategy of Billy Beane, of the Oakland A's, who employed statistically driven, evidence-based practice to enhance efficiency in winning baseball games (Cullena, Myera, & Latessaa, 2009; Surendra & Denton, 2009). The book was a *New York Times* (2011) bestseller, and was developed into a successful movie grossing over $140 million (Nash Information Services, LLC, n.d.). While the *Moneyball* story obviously had entertainment value, the story of the evolution of sport analytics has become valued in many sport settings. The ability to generate and apply statistical interpretations to enhance efficiency is central to the use of data in sports (Hakes & Sauer, 2006). Beyond this application of Sabermetrics, the mathematical and statistical

analysis used in Major League Baseball (MLB), examples of the use of data analytics in sport abound (Costa, Huber, & Saccoman, 2008; Davenport & Harris, 2007). This developing phenomenon found its way to the English Premier League (EPL), the National Basketball Association (NBA), the Professional Golfers' Association (PGA), the National Hockey League (NHL), and the Rugby Union, among many other sport organizations (Alamar, 2013; Brugha, Freeman, & Treanor, 2013; Chu, 2013; Fry & Ohlmann, 2012; Gerrard, 2007; Goddard, 2005). Data analytics are used in a variety of capacities in sport organizations. Analytics can enhance productivity and efficiency in sport, whether evaluating prospective talent, individualizing customer service, or informing managerial and coaching decisions.

"Big Data," Data Mining, and Spot Analytics

Businesses, governments, communities, and individuals access huge collections of data, trillions of bytes known as "Big Data," which capture valuable information and provide them such benefits as improved efficiency and effectiveness, ultimately leading to a competitive advantage (Bryant, Katz, & Lazowska, 2008). To pursue this advantage, novel applications such as individualized customization will result from the capture and analysis of data in order to meet the demands to harness "Big Data" (Borkar, Carey, & Li, 2012; Brown, Chui, & Manyika, 2011). The rise of the phenomenon of "Big Data" may result in better tools, goods, and services; yet, it may yield invasions of individual privacy. Still, "Big Data" may help enhance our understanding of trends, communities, and individuals (Boyd & Crawford, 2012).

There is no specific size that determines how big "Big Data" has to be (Manyika et al., 2011). It varies by economic sector and intended purposes. Only by fueling innovation, competition, and productivity will "Big Data" manifest its value. As "Big Data" becomes more available and more usable in informing strategic decisions, the mining of relevant data is also being used to guide decisions in sport (Lohr, 2012).

Data mining involves the quest for and discovery of valuable structures in large datasets (Hand, 2007). While a global concern of data mining is in "Big Data," another aspect is in small-scale structures that can inform local decisions. While sport analytics has a broad concern with "Big Data," it is in the local arena where it can be essential to sport managers (Hand, 2007). For example, mining data to support the management of customer relationships can result in enhanced customer loyalty, specifically targeted prospects, and newly identified markets (Berry & Linoff, 1999). The array of techniques for data mining can be used to produce better decisions across many areas of any sport organization, from marketing and customer support strategies to player recruitment (Berry & Linoff, 1997). Basically, data mining can be used to achieve strategic results by formulating the problem, analyzing the data, interpreting the results, and utilizing this information (Berry & Linoff, 1999).

The discovery of new knowledge yields an increased pool of data to mine, and the mining of data yields new knowledge. The amount of accessible data has dramatically escalated through enhanced ability for data generation and collection (Han, Kamber, & Pei, 2006). The growth in the generation of data is due to more computerized transactions and increased use of digital cameras and bar codes. The collection of data is intensified by scanned text and images, satellite systems, and the vast amounts of data on the web. This growth in available data mandates that researchers and practitioners find ways to transform it into useful information, and ultimately competitive advantage and improved performance (Han, Kamber, & Pei, 2006). Therefore, the demand for intelligent tools that automatically assist in transforming data into useful knowledge is intense (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). There is an array of analytical tools for effective model building (Guazzelli, Lin, & Jena, 2012). One such data analysis tool is the open source interactive statistical software R. R is an open-sourced computing and graphics platform that comprises an integrated suite of software for "data manipulation, calculation, and graphical display" (R Foundation, 2014). Developed by the R Foundation, a nonprofit organization working in the public interest, R is available as free software that runs on a wide variety of UNIX platforms, Windows, and MacIntosh operating systems.

Data mining has a direct relation to statistical analysis (Hand, 2007). Aspects of data mining, such as classifying, clustering, trend and deviation analyses, and dependency modeling, intersect with the realm of statistics (Fayyad et al., 1996). Business intelligence (BI), the term used to describe analytic applications, essentially involves data input and output (Watson, & Wixom, 2007). While both input and output are important, the focus of many organizations is in the application of data output to be used in decision-making processes. This focus has led to increasing interest in data mining and the application of statistical algorithms (Guazzelli, Stathatos. & Zeller, 2009).

Data mining has created a need for statistical algorithms and open source software solutions that have prompted predictive analytics to become a standard approach to BI (Guazzelli et al., 2012). A challenge associated with data mining and analysis is the explosive increases in the volume of data (Guazzelli et al., 2012). If properly utilized, these knowledge management tools can inform strategic and tactical decision making despite the abundance of both relevant and irrelevant information. Creating competitive advantage in many arenas as knowledge discovery tools, data mining and predictive analytics allow sport managers to utilize the input to yield actionable outputs (McCue, 2005). Decision management systems, as they are called, utilize these knowledge management tools to yield effective decisions that produce competitive advantage (Taylor, 2012).

Pursuing competitive advantage is a fundamental principle in the realm of sport. The uses for data mining and subsequent predictive analytics in sport while evident, remains incomplete. There are many sport settings, both in the front of-

fice and on the field, where more efficient and effective decision-making systems, powered by data mining and predictive analyses, may enhance productivity. This study explores an on-field application of these techniques within the National Football League (NFL).

Purpose

The purpose of this study was to utilize data analytics as means to classify NFL offensive play types. This research more specifically examines the application of logistical regression analysis utilizing open source data in comparing two NFL teams' play selections between the 2000 and 2012 seasons.

Specifically, the Cleveland Browns and the Pittsburgh Steelers, the longest running rivalry in the American Football Conference (AFC), were selected for review. In this timeframe, the Browns have perennially been an unsuccessful team as measured by wins (71). In contrast, the Steelers have been extremely successful, not only winning 64 (135-71) more games than the Browns but also the 2006 and 2009 Super Bowl championships. However, in some respects, these teams, separated by 135 miles, are similar. Both teams play in the same conference, thereby sharing many opponents, have similar "blue-collar" work cultures, and represent similar-sized markets. This study had as an additional intent to compare a "bad" team to a "good" team while attempting to negate other factors. The primary objective is to demonstrate logistic regression as a viable methodology to help NFL defensive coordinators assess the probability of a rush or pass by an opposing offense.

Methodology

Data Collection

Data was acquired from www.armchairanalysis.com, and primarily came from the "CORE" data file. This file contains over 500,000 individual records. Each record represents a single NFL play and collected attributes. The data was reduced and segmented in this research using the following parameters. Fourth-down plays were removed because these are overwhelmingly punts and field goals, or otherwise extremely unique game situations. Play types such as kickoffs were removed. This resulted in a binary relationship, "rush" or "pass," being kept for analysis. The overtime was removed by excluding any number larger than four as a quarter. Overtimes were removed because they represent a unique game situation.

The resulting data set was then partitioned according to offense. This resulted in 12,187 Cleveland (CLE) play records and 14,123 Pittsburgh (PIT) play records. The models were built individually from each partition. Each record had 10 attributes that were selected for model building, along with a variable that changed play type to binaries: 0 (RUSH) or 1 (PASS). Table 1 shows the attribute name and definition. Table 2 shows example records with both used and unused attributes from the "CORE" data set.

Table 1

Attribute Dictionary Used in the Models

Attribute Name	*Stands for*	*Definition*
drive_seq	Drive Sequence	The number of the play in the drive sequence.
QTR	Quarter	The play takes place in 1 of 4 quarters of the regular game time so this is a classification of 1 to 4.
MIN	Minute	The exact minute of the play start within the quarter.
Off_pts	Offensive points	The points of the offensive team at the beginning of the play.
Def_pts	Defensive points	The points of the defensive team at the beginning of the play.
Off _TO	Offensive time outs	The remaining time outs of the offensive team.
Def_TO	Defensive time outs	The remaining time outs of the defensive team.
DWN	Down	The down of the play, due to prior data reduction this is limited to 1,2 or 3.
YTG	Yards to Go	The yards to go in order to get a first down or reach the goal line from the line of scrimmage.
YFOG	Yards from own goal	The distance between the line of scrimmage and the offensive team goal line.

Table 2

Example Records

GameID	Season	PlayID	Off	Def	Type	Drive _seq	Qtr	Min	Off _pts	Def_ pts	Off _TO	Def _TO	DWN	YTG	YFOG
12	2000	1795	PIT	BAL	1	1	1	15	0	0	3	3	1	10	38
12	2000	1796	PIT	BAL	0	2	1	15	0	0	3	3	2	6	42

The data was examined using a logistical regression analysis so as to provide the logistical odds of a binary event and the resulting probability. Logistic regression differs from linear regression in that it classifies the dependent variable in a binary relationship. A linear regression predicts continuous outcomes such as games scores. Instead of a game score, it classifies winning or losing. The general layout of the proposed logistic regression is as follows:

F(log odds) = intercept + coefficient*(attribute 1) + coefficient*(attribute 2)…+coefficient *(attribute n)+ e, where "n" is the total number of attributes and "e" is the error term.

Analysis

The open source R software was use for the logistical regression analysis. The R statistical software, freely available at http://www.r-project.org/, is an interactive programming language with extensive capabilities for quantitative analysis (Maindonald & Braun, 2009). The R scripts can be obtained by contacting the author. The R software was used to read the modified data set. Then, R created the coefficients for each attribute by fitting a logarithmic curve to the records. A visual comparison of the differences between linear and logistic regression is illustrated in Figures 2 and 3.

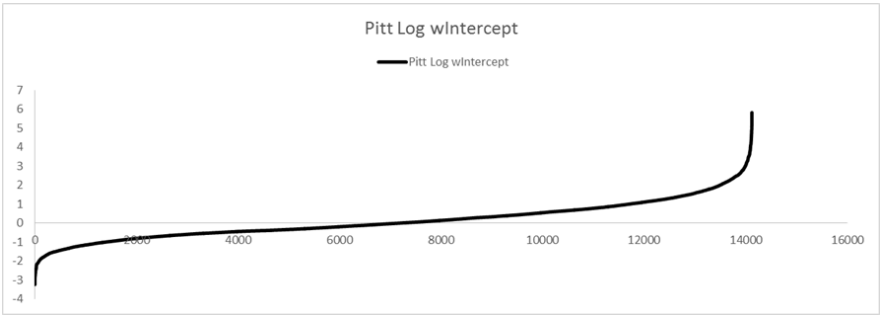

Figure 2. Logarithmic Odds

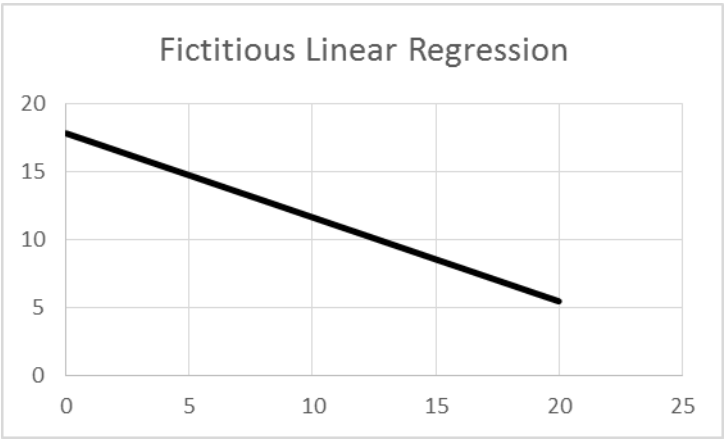

Figure 3. Fictitious Linear Regression

The logistic regression calculates the logistical odds (log odds) of an event occurring. However, in practical terms, this is not very helpful. Thus, once log odds have been calculated, it is used as input to get a more practical outcome, known as probability. Probability is the percentage likelihood of an event occurring or not occurring. The range of probability approaches 0% to 100%. The equation to move from log odds to probability is:

(e^log odds)/(1+(e^log odds), where e is the natural log or ~2.718.

When plotting the probability and log odds as a scatter plot, a curve from 0 to 1 occurs. This is illustrated in Figure 3. An event occurs somewhere on the probability curve and is then classified as a rush or pass based on a cutoff probability parameter. Commonly, the probability cutoff is .5 or 50%.

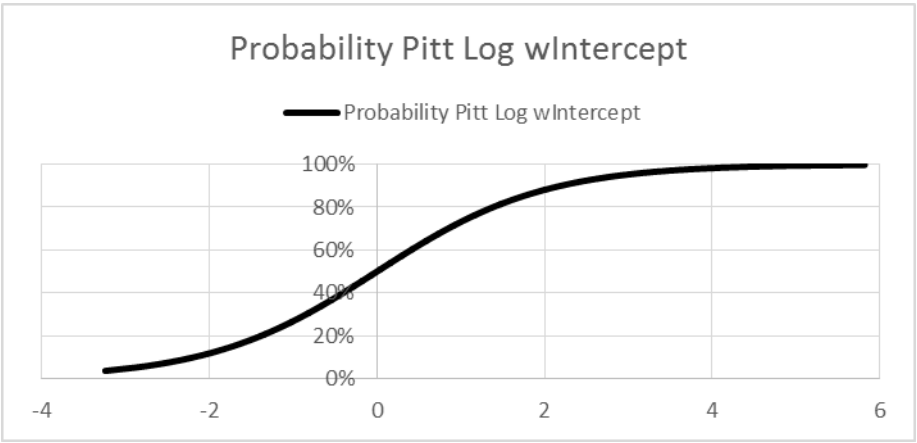

Figure 4. Probability Curve

Resulting Model

Using R, a coefficient table can quickly be calculated among the ~26,000 records. Table 3 shows the CLE and PIT logistic regressions based on the data set. Although this model calculates the log odds, which is of limited value, a comparison can still be made. All coefficient signs are the same with the exception of the quarter (QTR). This means that as the quarter increases the CLE model will show a higher log odds (and resulting probability) of passing in the model. Conversely the PIT model shows a diminishing desire to throw as the game quarters increase due to the negative sign. Further research could explore if this phenomenon is the result of the Browns trailing more often than the Steelers late in games. However this is out of scope of this research.

Model Evaluation

Classification models are often compared using a confusion matrix and applied to a holdout validation data set (see Tables 4 and 5). Since we have generated two models, two separate matrices are needed to compare the model classifications to the actual outcomes in the data set. Typically, classification models are partitioned a priori into training and validation sets. In this case, all records were used for the supervised learning of the model. While not as scientific, the additional records likely improve model development. The next truthful validation set could be obtained as the 2013 NFL season progresses. In the meantime, the confusion matrices are directionally sound to assess accuracy.

The Cleveland Browns the model was 66.4% accurate. The model classified 8,090 play types correctly and 4,097 incorrectly. The PIT model was 66.9% accurate. The model classified 9,442 play types correctly and 4,681 incorrectly.

Table 3

Logistical Regression Coefficients

Attribute Name	*CLE*	*PIT*
Intercept	-2.073	-2.008
drive_seq	0.025	0.017
QTR	0.065	-0.128
MIN	-0.009	-0.011
Off_pts	-0.046	-0.043
Def_pts	0.046	0.055
Off _TO	-0.261	-0.333
Def_TO	0.094	0.183
DWN	0.868	0.980
YTG	0.114	0.142
YFOG	-0.001	0.003

Table 4

CLE Confusion Matrix

CLE MODEL	Actual Play was Pass	Actual Play was Rush
Model Classified Pass	5,106	2,289
Model Classified Rush	1,808	2,984

Table 5

PIT Confusion Matrix

PIT MODEL	Actual Play was Pass	Actual Play was Rush
Model Classified Pass	4,765	2,140
Model Classified Rush	2,541	4,677

Example Calculations

In this section, a play for each team is examined and a third play is used to compare teams side by side. Pittsburgh's play identified as ID 3447 was against Baltimore Ravens' defense and resulted in a pass. It was the first play of the drive in the 14th minute of the fourth quarter. Pittsburgh was trailing Baltimore by a score of 10 to 13. Additionally, the Steelers only had a single timeout left versus the Ravens' three. Specifically, it was first-and-10 on the Pittsburgh 30-yard line.

The log odds calculate to .334 in this play scenario using the operators outlined in the Methodology section. To get to probability, one must still take the step using the natural log as outlined in the Methodology section:

(2.718^.334)/ (1+(2.718^.334)) = .582 or 58.2% probability of a PASS.

Cleveland's play identified as ID 531127 was against Cincinnati's defense and resulted in a rush. It was the second play of the drive in the sixth minute of the first quarter. Cleveland was losing to Cincinnati by a score of 0 to 7. Both teams had three timeouts. Also, it was second-and-3 on the Cleveland 23-yard line.

Table 6

Example PIT Play and Log Odds Calculation

Attribute Name	*PIT*	*PlayID3447*	*Log ODDS*
Intercept	-2.008		-2.008
drive_seq	0.017	1	0.017
QTR	-0.128	4	-0.512
MIN	-0.011	14	-0.154
Off_pts	-0.043	10	-0.430
Def_pts	0.055	13	0.715
Off _TO	-0.333	1	-0.333
Def_TO	0.183	3	.549
DWN	0.980	1	0.980
YTG	0.142	10	1.42
YFOG	0.003	30	0.09
TOTAL			**0.334**

Table 7

Example CLE Play and Log Odds Calculation

Attribute Name	*CLE*	*PlayID 531227*	*Log ODDS*
Intercept	-2.073		-2.073
drive_seq	0.025	2	0.05
QTR	0.065	1	0.065
MIN	-0.009	6	-0.054
Off_pts	-0.046	0	0
Def_pts	0.046	7	0.322
Off_TO	-0.261	3	-0.783
Def_TO	0.094	3	0.282
DWN	0.868	2	1.736
YTG	0.114	3	0.342
YFOG	-0.001	23	-0.023
TOTAL			**-0.136**

The log odds calculate to -0.136 in this play scenario. The next step is to use the log odds to create the probability of the Pass:

(2.718^-0.136) / (1+ (2.718^-0.136)) = .466 or 46.6% probability of a PASS.

A third example serves as a direct comparison between the teams. This is a real play from the Pittsburgh data set and is similar to plays in the Cleveland one as well. In this situation, and ones like it, the gap between the teams can be large. In this case, the result is a classification of a Pittsburgh Pass versus a Cleveland Rush.

Using the log odds for each team, PIT has a 50.1% probability of a pass while Cleveland had a 38.8%.

Discussion and Implications

The results of this study demonstrate that open source software can be effectively used in the analysis of situation-specific play selection in the NFL. It provides evidence that logistic regression has promise as a classification system for opposing NFL offenses. The analysis of large amounts of data regarding play selection can inform decisions by head coaches, defensive coordinators, position coaches, opposing coaches, and even player personnel managers. Head coaches, defensive coordinators, and position coaches can better develop a game plan and prepare athletes for opponents' anticipated play selection. Conversely, offensive coaches might choose to alter their documented play selection pattern simply by being informed of it. Player personnel managers may identify and pursue talent specifically based on an analysis of situation-specific play selection, both by their own team and by their opponents.

Table 8

Direct Team Comparison

Attribute Name	*PIT*	*CLE*	*Comparison Play*	*PITT Log ODDS*	*CLE Log ODDS*
Intercept	-2.008	-2.073		-2.008	-2.073
drive_seq	0.017	0.025	1	0.017	0.025
QTR	-0.128	0.065	1	-0.128	0.065
MIN	-0.011	-0.009	9	-0.099	-0.81
Off_pts	-0.043	-0.046	0	0	0
Def_pts	0.055	0.046	3	0.165	0.138
Off _TO	-0.333	-0.261	3	-0.999	-0.783
Def_TO	0.183	0.094	3	0.549	0.282
DWN	0.980	0.868	1	0.98	0.868
YTG	0.142	0.114	10	1.42	1.14
YFOG	0.003	-0.001	36	0.108	-0.036
TOTAL				**0.005**	**-0.455**

This research demonstrated that data analytics can add value in the decision making process in sport settings such as the NFL. It also demonstrated that humans are central to the effective implementation of sport analytics. Slaton (2013) noted that "…the entire sports organization, from the lowliest assistant coach and marketing employee to the most senior leader needs to adopt the analytics philosophy if it is to be truly effective" (p. 1). Coaches and sport managers are central in the collection of data, in the analysis of that data, and in the application of that data analysis.

Quantitative analysis has been shown to yield valuable information that sport managers and coaches can use to inform their decisions (Borrie, Jonsson, & Magnusson, 2002). There is an abundance of data available to support both on-the-field and in-the-front-office decisions in sport. It is incumbent on sport managers to obtain, analyze, and utilize appropriate data within their organization's knowledge management systems.

In the coming years, the United States will need up to 190,000 more expert analysts and an additional 1.5 million managers, including sport managers, who are prepared to use data to inform decisions (Manyika et al, 2011). The application of analytics in sport is no longer the niche pursuit of a few visionaries such as the Oakland A's Billy Beane, the Houston Rockets' Daryl Morey, or the Kraft Group's

Jessica Gelman, but rather it is a mainstream practice that sport organizations have embraced (Slaton, 2013). For example, over 2,700 interested researchers, analysts, and sport managers annually attend the Sloan Conference on the latest developments in sport analytics (Slaton, 2013).

As evidence of the successful application of sport analytics builds, a new style of sport manager is evolving in sport organizations (Fry & Ohlmann, 2012). These new sport managers and coaches are proficient in the application of data analytics. The organizations they work for are stimulating the expansion of data mining and predictive analysis in sport settings. The ongoing development of analytics in sport settings reflects the trend in other market segments and is inevitable (Alamar, 2013). The proper utilization of data management, predictive analyses, and data-driven decision-making by sport managers occur in pursuit of a competitive advantage (Alamar, 2013). Using the data analyzed in this study, coaches and managers seeking a competitive advantage can employ data-generated knowledge about opponent tendencies in their decisions on player personnel, formations, and play calls in these specific game situations.

This study supports the contention that the analysis of data in sport can be most useful in its application by qualified sport leaders (Lohr, 2012). In that context, the increasing effectiveness and acceptance of sport analytics, while informing decisions, does not negate the value of human insights and actions (McAfee & Brynjolfsson, 2012). Sport managers as decision makers, informed by data analytics' knowledge management tools, remain essential. They are crucial in collecting appropriate data and in analyzing that data accordingly. More so, effective coaches and sport managers of the future must be adept in the successful strategic application of the analysis to specific situations. For example, in this study, it remains incumbent on head coaches, offensive and defensive coordinators, position coaches, and player personnel managers to determine the best way to effectively employ the data analyses in the decision-making process. Additionally, their intended use of the data can inform the selection and analysis of algorithmic factors,

The process of continual analysis in sport is the cyclical progression through which analytics are effectively utilized in sport settings (see Figure 1). Initial situation-specific parameters (e. g. plays executed) are identified and inform the development of the analytical algorithm. Within the identified parameters, data are collected and statistically analyzed. Sport leaders (e. g. coaches) apply the analysis through the process of informed decision-making in similar situations. The resulting performance outcomes (i. e. success or failure) in like situations inform the pertinence of the identified parameters and influence the development or maintenance of the employed algorithms.

This study provided evidence of situation-specific tendencies for two teams in the NFL. The study was delimited to only the Cleveland Browns and Pittsburgh Steelers. All other teams were excluded, as were other professional sport leagues. Clearly, this analytical approach could be easily applied to any team, in any sport,

within any situation-specific parameters. In explicit game situations, such as time remaining or score, first-down plays, plays in the "red zone," special team plays could be similarly analyzed. Overtime plays were not analyzed, but could be analyzed to inform decisions on that specific situation. Likewise, fourth-down plays were excluded, but could be reviewed in a separate analysis.

To better inform the application needs of coaches and sport managers, a researcher could improve the utility and/or accuracy of the model by incorporating new attributes, thereby changing the data set. Changes in coaching staff, player personnel, and even opponents' style of play occur regularly in the NFL. Analyses incorporating algorithms that address these factors can be developed. For example, when Tim Tebow replaced Kyle Orton as the Denver Broncos quarterback, the number of rushes increased significantly. Judiciously excluding these plays as outlier situations may have a boosting effect. In addition, games with inclement weather may often have more rushes. So, adding an attribute for weather conditions may improve the analyses' utility. Additionally, a researcher could change the data structure to improve effectiveness. Coaches, offensive coordinators, and audible-calling quarterbacks, such as Peyton Manning, likely behave consistently in similar game situations despite the team. Thus, partitioning the data from team-based to a specific decision maker (e.g., quarterback) may further improve the predictive accuracy.

References

Alamar, B. C. (2013). *Sport analytics: A guide for coaches, managers, and other decision makers*. New York, NY: Columbia University Press.

Andrew, D., Pedersen, P., & McEvoy, C. (2011). *Research methods and design in sport management.* Champaign, IL: Human Kinetics.

Berry, M., & Linoff, G. (1999). *Mastering data mining: The art and science of customer relationship management*. New York, NY: John Wiley & Sons, Inc.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. New York, NY: John Wiley & Sons, Inc.

Borkar, V., Carey, M., & Li, C. (2012). *Inside "big data management": Ogres, onions, or parfaits*? New York, NY: Proceedings of the 15th International Conference on Extending Database Technology 3-14.

Borrie A., Jonsson, G. K., & Magnusson, M. S. (2002). Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. *Journal of Sports Sciences, 20*(10), 845–852.

Boyd, D., & Crawford, K. (2012). Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, and Society*, 15(5), 662–679.

Brown, B., Chui, M., & Manyika, J. (October, 2011). Are you ready for the era of big data? *McKinsey Quarterly*, pp. 24–35.

Brugha, C. M., Freeman, A., & Treanor, D. (2013). *Analytics for enabling strategy in sport.* OR55 Annual Conference Keynote Papers and Extended Abstracts, 138–150.

Bryant, R. E., Katz, R. H., & Lazowska, E. D. (2008). Big Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. Computing Research Initiatives for the 21st Century, Computing Research Association. Retrieved from http://www.cra.org/ccc/files/docs/init/Big_Data.pdf.

Chu, B. (February 25, 2013). Dr. Wayne Winston's work may do to the NBA what Sabermetrics did to MLB. *Yahoo!Sports*. Retrieved from http://sports.yahoo.com/news/dr-wayne-winstons-may-nba-sabermetrics-did-mlb-155600962–nba.html

Costa, G. B., Huber, M. R., & Saccoman, J. T. (2008). *Understanding Sabermetrics: An introduction to the science of baseball.* Jefferson, NC: McFarland & Company, Inc.

Cullena, F. T., Myera, A. J., &Latessaa , E. J. (2009). Eight lessons from *Moneyball*: The high cost of ignoring evidence-based corrections. *Victims and Offenders: An International Journal of Evidence-based Research, Policy, and Practice*, *4*(2), 197–213.

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Boston, MA: Harvard Business Review Press.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI Press.

Fry, M. J., & Ohlmann, J. W. (2012). Introduction to the special issue on analytics in sports, part I: General Sports Applications. *Interfaces, 42*(2), 105–108.

Gerrard, B. (2007). Is the *Moneyball* approach transferable to complex invasion team sports? *International Journal of Sport Finance, 2*(4), 214–230.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting, 21,* 331–340.

Guazzelli, A., Lin, W-C., & Jena, T. (2012). *PMML in action: Unleashing the power of open standards for data mining and predictive analytics*. Paramount, CA: CreateSpace.

Guazzelli A., Stathatos K., & Zeller, M. (2009). Efficient deployment of predictive analytics through open standards and cloud computing *ACM SIGKDD Explorations, 11*(1), 32–38.

Hakes, J. K., & Sauer, R. D. (2006). An economic evaluation of the *Moneyball* hypothesis. *The Journal of Economic Perspectives, 20*(3), 173–185.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco, CA: Morgan Kauffman Publishers.

Hand, D. J. (2007). Principles of data mining. *Drug Safety*, *30*(7), 621–622.

Lewis, M.(2004). *Moneyball: The art of winning an unfair game*. New York, NY: W. W. Norton & Company.

Lohr, S. (February 11, 2012). The age of big data. *New York Times*.

McAfee, A., & Brynjolfsson, E. (October, 2012). Big data: The management revolution. *Harvard Business Review*.

McCue, C. (2005), Data mining and predictive analytics: Battlespace awareness for the war on terrorism. *Defense Intelligence Journal, 13*(1&2), 47–63.

Maindonald, J. H., & Braun, J. (2009). *Data analysis and graphics using R: An example-based approach.* Cambridge, UK: Cambridge University Press

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute.*

Nash Information Services, LLC. (n.d.). The numbers. Retrieved from http://www.the-numbers.com/movie/Moneyball#tab=summary

New York Times. (2011). Best sellers. Retrieved from http://www.nytimes.com/best-sellers-books/2011-11-20/combined-print-and-e-book-nonfiction/list.html

Plunkett Research, Ltd. (2014). *Sports industry market research*. Retrieved from http://www.plunkettresearch.com/sports-recreation-leisure-market-research/industry-and-business-data

R Foundation. (2014). *The R project for statistical computing.* Retrieved from http://www.r-project.org/

Slaton, Z. (February 20, 2013). Why the Sloan Conference is the Super Bowl of sports analytics. *Forbes*. Retrieved at: http://www.forbes.com/sites/zachslaton/2013/02/20/why-the-sloan-conference-is-the-super-bowl-of-sports-analytics/

Surendra, N. C., & Denton, J. W. (2009). Designing IS curricula for practical relevance: Applying baseball's *Moneyball* theory. *Journal of Information Systems Education*, *20*(1), 77–86.

Taylor, J. (2012). *Decision management systems: A practical guide to using business rules and predictive analytics*. Boston: IBM Press.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer*, *40*(9), 96–99.