Elaborating Role of Big Data Analytics in Sports

Gagandeep Kaur[1] Dr. Vijay Laxmi[2]
[1]Research Scholar, Ph.D. Comp. Sc. Engg., Guru Kashi University, Talwandi Sabo, Punjab, India.
[2]Professor, Dept. of Comp. Appl., Guru Kashi University. Talwandi Sabo, Punjab, India.

Abstract - This paper aims to investigate the role of data analytics and technology in the sports industry. The paper elaborates on what the current practices are and identify their utilization levels to inform decision-making within the sports industry. Examination of the literature shows that there is a lean towards the use of data analytics on-pitch as opposed to data analytics use off the pitch. The paper focuses on different types of data and data mining techniques. The paper discusses the role played by big data analytics in the sports sector from enhancing performance to increasing revenues.

Keywords – Big Data, Data mining, Information, Sports, Technology.

I. INTRODUCTION

Analysis of sports person behavior is the first idea that pops up in everyone’s mind when it comes to big data in sports. Each game can be represented as a colossal array of digital data and the number of conclusions and deductions drawn from this data is nearly limitless. Coaches can validate the correctness of their strategic decisions, match the team’s performance against historic values, track individual progress, identify trends in opponents’ playing styles to come up with counter-tactics, and do a lot more things based on actual data, not hunches and guesswork. Some of this is still being done using more conventional methods, but companies are already offering state-of-the-art solutions that take sports analytics to a whole new level, both in terms of the scope of collected data, its detail, quality of visualization and depth of analysis. The most serious challenge of leveraging big data in sports is not its collection, but processing and further application, including monetization. To that end, companies can use readily available BI products or build custom BI solutions and Big Data analytics tools. This software enables users to consolidate and normalize heterogeneous data sets, add complex mapping and visualize the results so as to obtain meaningful and actionable insights. The latter, in turn, can be used for a variety of purposes – from preventing cumulative injuries among athletes or adjusting their training programs to changing the overall strategy/composition of a team based on competitor analysis.

II. TYPES OF DATA BASED ON SOURCE OF ORIGIN

The data can be broadly classified into three categories based on the source of origin which are as under.

- **Human-sourced information**: All information ultimately originates from people. This information is a subjective record of human experiences, previously recorded in books and works of art, and later in photographs, audio, and video. Human-sourced information is now almost entirely digitized and electronically stored everywhere from tweets to movies. Structuring and standardization—for example, modeling—defines a common version of the truth that allows the business to convert human-sourced information to more reliable process-mediated data. This starts with data entry and validation in operational systems and continues with the cleansing and reconciliation processes as data moves to Business Intelligence.
- **Process-mediated data:** Business processes to record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions, reference tables, and relationships, as well as the metadata that sets its context. Process-mediated data has long been the vast majority of what IT managed and processed, in both operational and BI systems.
- **Machine-generated data**: The output of sensors and machines employed to measure and record the events and situations in the physical world is machine-generated data, and from simple sensor records to complex computer

logs, it is well structured and considered to be highly reliable. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is amenable to computer processing, but its size and speed are often beyond traditional approaches for handling process-mediated data.

III. TECHNIQUES FOR DATA MINING

There are several techniques which are used for mining data depending upon the type of data. The prominent techniques available for data mining are briefly explained as under.

- **Association:** Association is one of the best and straightforward data mining techniques. It refers to the strategy that may assist you to recognize the interesting relations between many different variables in a huge set of databases. In other words, it is a discovery of patterns based upon the correlation of the items in the same transaction. The association technique is mainly used in the market basket analysis to identify the products that are frequently purchased by the customer.
- **Classification:** The classification technique is based on machine learning. Basically, classification is used in the set of data to classify each item into a predefined set of classes. The classification method uses mathematical techniques like a neural network, linear programming, statistics, and decision trees. In classification, one tends to develop the software which would find out how to classify the data items into groups.
- **Dependency modeling:** The identification of a model that holds information related to the dependencies amongst variables is known as the dependency modeling method. There are two levels of dependency models: (a) The structural level of the model on the first level. This level shows (frequently in graphic frame) the variables that are locally dependent. (b) The quantitative level of the model is the second level. This method shows the strengths of the dependencies utilizing some numeric scale.
- **Clustering:** It is a process to recognize the data items that are similar to each other. Cluster is a collection of objects that belongs to a similar class. Clustering is the process of making a group of abstract objects into classes of similar objects. This means in clustering the objects which are similar to each other are put in the same group and they are dissimilar or different from the objects of other groups. Clustering analysis is the method of determining clusters or groups in such a way that the highest degree of association is attained between two objects if they belong to the same group. The core advantage of clustering over-classification is that it is adaptable to changes and assist in selecting the useful features that differentiate among groups.
- **Prediction:** As the name implies, the prediction is the technique used to discover the relationship between independent variables and the relationship between dependent and independent variables. It could be a wide topic and used for fraud detection, predict the failure of the machinery or components, and even for the prediction of a company's profit. Prediction can be used with other data mining techniques like pattern matching, classification, analyzing trends and relation. By analyzing the events of the past, one can make a prediction about the future.
- **Sequential patterns:** The analysis of sequential patterns is used to discover and identify some similar patterns or regular events in the data of transactions over a business period. In sales, with the use of historic transactional data, companies can find out the set of items that customers buy together at different times in a year. Then companies use this information for marketing strategy by recommending the customers to buy those products at better deals based upon their past frequency of purchase.
- **Summarization:** Summarization is the method to identify the description of the subset of data. For instance, the classification of a mean, as well as standard deviation for all the areas, can be provided. There are also applications that involve the summary rules, multivariate visualization methods and the identification of functional relationships amongst various variables which can help in providing assistance to further steps. These methods can be provided for interactive exploratory data analysis and the creation of the automatic generation within various applications.

- **Decision trees:** Decision trees is one of the most commonly used data mining technique as its model is easily understandable by the users. In this technique, the root of the decision tree is a condition or a question that has multiple answers. Based upon each answer a set of conditions or questions are further raised that assists in determining the data so that the final decision could be taken.
- **Regression:** A function that provides the mapping of data items into a real-valued prediction variable is known as the regression learning method. It provides the study which is related to the assumptions made from the previously present methods. The assumptions provided here are helpful is providing the other assumptions made for the future. The assumptions can be made on the basis of different parameters and the selection of parameters depends on the type of application in which the method is being utilized.

IV. BIG DATA ANALYTICS IN SPORTS

Today, professional sports are characterized by the search for new paths concerning how a sportsperson or a team may apply new technologies and sports performance data to gain the cutting-edge competitive ability that will elevate them to the top of the podium or to win the major league or championship titles. Therefore, professional sports properties are left with massive data pools that (without giving away competitive sporting advantages) can be utilized to assist athletes and teams in monetizing their relationships with commercial stakeholders. In reflections over why the application of technology and data is important in sports, there are clear benefits in terms of optimizing decision-making processes on and off the playing field and thus sporting quality. Better sporting performances may be supported by technology and data and can help to enhance the entire business model and the backing from fans and other stakeholders in relation to revenue generation from various sources, e.g., ticket sales, broadcasting contracts, merchandise sales, good sponsorship activation and the increased value of fan engaging content production. Technology and data have become manifested elements of professional sports in the hunt for enhanced and elevated performance platforms. Additionally, their application reaches beyond that of professional sports and even targets fans and consumers outside the spot-lighted playing fields in professional sports. In the aftermath of this development, actors in the professional sports industry underline the importance of technology and data as being directly associated with winning titles. However, the vital role of technology and data in influencing sporting performances is clear but there is reasonable meaning in acknowledging that technology and data are only tools and not a universal quick fix to performance challenges in professional sports.

This illustrates the value of technology and data in lifting sporting performances, but with the hint that sporting tracking and positional data must be ‘qualified’ to be applied intelligently and effectively in order to positively reinforce sporting performances in practice. For instance, it holds essential meaning that the application of technology and data is based on a solid understanding of the interplay between various contextual factors and the importance of timing when executing performance-based decisions, e.g., the technical and tactical capabilities of the individual players in contrast to the team’s tactical game plan, the strength and weaknesses of the opponent or the fact that the coach has only a short window to receive, interpret and execute on data during the game. It is imperative to have a platform of knowledge before decisions are made. This knowledge is often mixed with intuition and passion, e.g., from a coach, in reality in professional sports, which helps to determine the perception of decision-making and therefore adds an extra complexity level in dealing with the intersection between technology, data and sports.

From a critical perspective, the quality of the application of technology and data in professional sports is also subject to bias as there is always a person behind the data. Taking the qualification of data towards a higher degree of organizational, economic and commercial sense-making, data management works as a vehicle of positive strategic change that sometimes may be associated with some extent of risk aversion. No matter what you’re doing with a sports team, unless you are winning, there’s going to be upset fans. Therefore, applying technology and data may sometimes risk being subject to ‘reverse research’ in which sport organizations know the desired conclusions beforehand and aim for evidence that will back these conclusions or look to apply conclusions that are not fully backed by the empirical data collection. Despite this managerial complexity, the application of technology and data adds value to sporting performances in team sports, e.g., football, by offering tracking opportunities concerning the positioning of the athletes and how that changes dynamically in the game and how that affects the possibility of scoring or improving one's team position over another. This inspires the potential to optimize sporting performances via a positive influence on the sporting quality, on the outcome of the game, and the associated learning.

V. LITERATURE SURVEY

Table 1 summarizes different research papers having relevance with Data Mining and Big Data Analytics.

Table 1: Summarized Literature Survey

Title of Research paper	Author's & Publication year	Discussions & findings	Tools used	Conclusion
Need and Application of Data Mining	Dhaka and Kumar, 2018	Detailed about different sectors making use of big data. Discussed tools like WEKA, RapidMiner, R, Orange, NLTK. The paper suggested that only after proper analysis, the conclusion can be made so that futuristic decision is taken with a higher degree of surety.	WEKA, RapidMiner, R, Orange, NLTK.	This paper provided a review of data analysis methods and tools for the data mining process
A Guided FP-growth algorithm for multitude-targeted mining of Big Data	Lior Shabtay, Rami Yaari, and Itai Dattner, 2018	The paper presents the Guided FP-growth algorithm for multitude-targeted mining. GFP-growth algorithm yields the exact frequency-counts for the required itemsets.	FP-Growth algorithm and Guided FP-Growth algorithm.	The research work conducted in the paper is the development of the Minority-Report Algorithm that uses the GFP-growth for boosting performance when generating the minority-class rules from imbalanced data
Big Data Analysis Of Indian Premier League using Hadoop and MapReduce	Rajdeep Paul, 2017	The paper elaborated on the popularity of IPL and how Twitter was flooded with the number of tweets.	MapReduce and Hadoop	Hadoop based MapReduce algorithm is well equipped to handle unstructured data

				like tweets and mine enormous data to reach a perfect conclusion.
A Data Mining Approach on Cluster Analysis of IPL	Pabita Kumar Dey et al., 2016	The paper worked on fuzzy grouping algorithm operating on N-Clusters to mine data	Matlab	The technique was capable of handling imprecise data effectively and efficiently
Hadoop-MapReduce: A Platform for Mining Large Datasets	Maedeh Afzali et al., 2016	The paper elaborated on the association of popular Apriori algorithm with MapReduce. The working of the technique is based on association rule.	Apriori algorithm, MapReduce algorithm, Hadoop framework	The Apriori in collaboration with MapReduce proved to be an effective mining algorithm in comparison with other conventional data mining techniques.
Big data analysis using Apache Hadoop	Padhy, Kumar, 2014	The paper suggests various methods for catering to the problems in hand through the Map-Reduce framework over HDFS.	Apache Hadoop, HDFS, MapReduce	Big Data analysis tools like MapReduce and Hadoop helps organizations better understand their customers and the marketplace, leading to better business decisions and competitive advantages
Application of Data Mining Techniques in Sports Training	Yingying Li et al., 2013	The paper emphasized on arranging training sessions using multiple clustering methods.	K-Means, DBSCAN, COBWEB	Hierarchical clustering and K-Means proved better than traditional techniques

JETIR

VI. CONCLUSION AND FUTURE SCOPE

Technology and data improve the commercial outcomes for teams, not only by raising the sporting performance of the sports person, but also by allowing the data to be commercialized. This includes technological innovation, biostatistics, movement data, and other game-based information, which improve how a club manages performance and enhances the circumstances for unfolding its talent on and off the pitch. This commercialization creates still more fandom above and beyond what's driven by the sporting outcomes and thereby opportunities for better comprehensive fan experiences and innovative commercial solutions that can support revenue generation. Currently, though, the sheer amount of data that is becoming available to sporting enterprises is difficult to harness and use effectively both for technical reasons, but also due to a lack of financial and human resources. However, what needs to happen is that the data needs to be qualified so that knowledge-sharing, individual and organizational learning co-exist along with the aim to apply technology and data to improve sporting and business performances. If there is a lack of resources, e.g., in terms of financial and human capital, one way to overcome this constraint is to strategically invest in technology and data to optimize the utilization of these forms of capital. In such, technology and data are potent vehicles that can change performance in the sport and business nexus and by catering to new audiences and improving the engagement with existing audiences, there is a good chance of increasing profitability.

REFERENCES

[1]. Spaaij R, Thiel A, "Big Data: Critical Questions for Sport and Society. European Journal for Sport and Society," Vol. 14(1), pp.1-4, 2017.

[2]. Fried G, Mumcu C, editors. Sport Analytics: A Data-Driven Approach to Sport Business and Management. 1st ed. London, New York: Taylor & Francis; 2016.

[3]. Yin RK. Case Study Research and Applications: Design and Methods. London: Sage Publications; 2017.

[4]. Cortsen K. Strategic sport branding at the personal, product and organizational level: Theory and practice for improving a Sports Brand's interactions [doctoral dissertation]. Aarhus, Denmark: Aarhus University; 2016.

[5]. Cortsen K. Data management is the new winning approach in the business of sports. 2015. Available from: http://kennethcortsen.com/data-management-is-the-new-winning-approach-in-the-business-of-sports/ [Accessed: Nov 20, 2019].

[6]. Watkins B, Lewis R., "Winning with apps: A case study of the current branding strategies employed on professional sport teams' mobile apps," International Journal of Sport Communication, vol. 7(3), pp. 399-416, 2014.

[7]. CNN. World Cup 2018: Hand-held devices to provide in-game analysis to coaches. 2018. Available from: https://edition.cnn.com/2018/06/12/sport/fifa-world-cup-2018-data-ana-lyst-spt-intl/index.html [Accessed: Nov 15, 2019].

[8]. Harrison CK, Bukstein S, editors. Sport Business Analytics: Using Data to Increase Revenue and Improve Operational Efficiency. Boca Raton, Florida, USA: CRC Press; 2017.

[9]. Funk D, Alexandris K, McDonald H. Sport Consumer Behaviour: Marketing Strategies. New York, NY: Routledge; 2016.

[10]. Krustrup P, Helge EW, Hansen PR, Aagaard P, Hagman M, Randers MB, et al., "Effects of recreational football on women's fitness and health: Adaptations and mechanisms," European Journal of Applied Physiology", vol. 118(1),pp. 11-32, 2018.

[11]. Foster G, O'Reilly N, Dávila A. Sports Business Management: Decision Making around the Globe. New York, London: Routledge; 2016.

[12]. N. J. Kimand J. K. Park, "Sports analytics & risk monitoring based on hana platform: Sports related big data & IoT trends by using HANA In-memory platform,"2015 International SoC Design Conference (ISOCC), Gyungju, 2015, pp. 221-222.

[13]. Irena Bojanova, "IT Enhances Football at World Cup 2014",IT Professional, vol. 16, no. , pp.12-17, July-Aug. 2014.s R. Brunelet al., "Supporting hierarchical data in SAP HANA,"2015 IEEE 31st InternationalConference on Data Engineering, Seoul, 2015, pp. 1280-1291.

[14]. Collins, R. (2016). Micro-sociology of sport: Interaction rituals of solidarity, emotional energy, and emotional domination. European Journal for Sport and Society, 13, 197–207. doi: 10.1080/16138171.2016.1226029.

[15]. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. Science, 343, 1203–1205. doi: 10.1126/science.1248506.

[16]. Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. Computers in Human Behavior, 48, 392–400. doi: 10.1016/j.chb.2015.01.075.