PERFORMANCE ANALYSIS OF SPORTS PERSONS USING DATA MINING BASED RANKING AND CLASSIFICATION METHODS
Rajeswari C*
School of Information Technology & Engineering, VIT University, Vellore, India.
*Email: rajeswari.c@vit.ac.in

Received on 09-08-2016
Accepted on 05-09-2016

Abstract

Data Mining has become a vital component in business development today due to its importance in assisting in decision making that can subsequently result in organisational growth as well as preparation for facing various market scenarios. The sports industry is one such where the application of various mining, analytical and statistical techniques has been used to perform analyses of performance and training, medical studies to assist in player selection and recruitment. This study proposes a method where statistical analyses are used in order to effectively analyze player performance and determine solutions to player selection and recruitment in Football.

Various sports have already seen the implementation of data science or analytics techniques with various reference material, research and publications regarding statistical approaches contributing to expanding its outreach. Football, however, has been slow in applying similar approaches. An attempt to bridge the gap between a data scientific approach and the sport by introducing an approach along the lines of the similar techniques used in various sports is proposed. Certain techniques of analyzing team and player performance statistics and utilizes the results to devise an effective method of player selection and recruitment based on historical performance data are introduced. A number of scenarios are considered in order to explain each method efficiently. Also, a ranking methodology is proposed which uses the performance data in order to effectively rank the players of the specific team based on their previous performances.

Keywords: Sports Performance Analysis, Statistical Mining, Player Ranking, Data Mining, Classification Techniques.

1. Introduction

Data Analytics is an industry that today has become one the most thriving domains around primarily thanks to the potential and the power that data brings to businesses and organisations. Analytics can be applied to more or less any
domain around in order to study various types of data generated and to analyze it in such a way that the results provide the decision makers with key insight on ways to improve and develop their businesses so as to attain profit and sustainability and at the same time growth. The sports industry is no different in the way that sporting organisations are very similar to business. Organisations want to achieve a way of sustaining their organisation while achieving results in their respective sports and growing within the process.

1.1 Background

Various approaches have been made to use statistical techniques into analyzing sports performance and using the results to make calculated and informed decisions regarding player selection and recruitment. Baseball initially saw activity involving the use of statistics even though it was not without opposition from the traditionalists and sports purists who remarked that sports should be left to those who were experienced in the game, knowledgeable and instinctive. Bill James (1) published various articles and his work introduced a system of using certain key statistics to determine and achieve targets. These ideas were advocated by Billy Beane (2) where his implementation of the statistical approach won him much appreciation and widespread acclaim. The approach gained popularity and now we see various number of teams ranging from baseball to basketball to even American football employing data analysts to study and provide detailed evaluations of their teams' data in order to remain competitive.

1.2 Motivation

Finding a suitable way of identifying the best team to select and players to recruit using performance statistics is a method that has been adopted by various football organisations in order to build teams and sustain a competitive nature. With the advent of technology and in-game tactics and strategies constantly evolving, it has become imperative that clubs develop their approaches into recruiting and selecting their players in order to keep up with the times.

Statistical techniques have been used by teams for a while now and many football clubs employ data analysts and executives to work in tandem with their coaches and scouting departments in order to optimize team selections and find the best players that can contribute to their team in the best possible manner. Soccernomics (3) details methods and techniques that can be followed in football which deals with the study of statistical measures in order to achieve target objectives set by football clubs and franchises.

The primary issue lies within the fact that sports such as baseball or cricket are fundamentally different from football or basketball in the manner that each player's performance is less reliant on his or her teammates' performance. In football, there are various parameters as well as the fact the all players on the pitch continuously interact with each
other throughout the game thereby making each others' contributions interdependent. Hence the need for a more refined model using the data and statistics is required.

Certain hypotheses are devised in this study and tested from the data using this information can be drawn regarding potential team selections and player recruitment which can, in the long term, make the difference to a team's achievements.

2. Data Mining and Sports

Data Mining is defined as the process of analyzing huge amounts of data into identifying patterns and converting them into useful information which can be used to make informed business decisions by the organisation or individual for organisation growth and development. This field extends to various industries apart from the information technology domain, such as varying from medicine to physics, business to sports. Gullo (4) explores this area in a more detailed manner and provides a technical review about Data Mining and it's methods and techniques. The concept of Knowledge Discovery in Databases (KDD) is also elaborated along with the applications of data mining and the various steps or processes that form the iterative and interactive sequence of data mining techniques. Gandomi and Haider (5) summarizes the need for Big Data Analytics and emphasizes and how various techniques and tools need to be developed and implemented in order to leverage the full power of big data.

Data Mining has become increasingly relevant in the sports industry today. Today a large number of coaches or managers employ analysts to study vast amounts of historical data and find key patterns or identify methods to form relations between the data and potential future performance or even predict future results. Leung and Joseph (6) described this approach by analyzing historical data of team and players in the National Collegiate Athletic Association(NCAA) in order to predict future results. Li et al (7) describe an approach where an individual player's sport physiological data was collected via tests and the results were analyzed using clustering techniques. These results were then used to plan training sessions for the players by grouping together the players within the same or nearby clusters in order to keep players of similar abilities within the same group or cluster. Foehrenbach and Fokoue (8) used performance statistics and related variable in order to determine the calibre of team competing in the National Football League (NFL). Various parameters were analyzed to form a distinction between the statistic indicators and the actual results to determine how good a team was. Tavana et al (9) presented a novel method of applying the fuzzy set theory into developing a ranking system to determine which players merit selection for the team. The assessment was based on analyzing certain technical and mental skills of the individual players. Following this, a system was proposed to evaluate various combinations or formations that could be built using the
selected players. This proposed system was called the Fuzzy Inference System (FIS). McGee and Burkett (10), studied the performance results of players in the National Football League Combine, a testing camp where rookie athletes from collegiate football are assessed over various performance measurements. The experiment attempted to study the relation of these performance results on the final draft or selection status of the players as they were inducted into the National Football League.

Performance statistics are studied carefully by analysts and certain key statistics are identified that can be vital in gaining sporting advantage. Implementing scientific methods into evaluation of players and determining their in game value, selection and other parameters pertaining to their sports skills has become a regular feature these days for team and franchises from all types of sports. FC Midtjylland (11-13), a Danish top tier club that recently won their first league championship in their history used an analytical approach with statistical studies.

This study introduces certain approaches towards utilising data analysis in sport. First, a method of statistical analysis which focuses on employing a mining approach to determine team based targets and how this target can be used to identify recruitment and selections of players to fit the manager's tactical system. The second, is a method of rating players in a sports team based on studying their past performance data and developing an effective method of ranking these players based on their respective tactical roles. The principle used to develop the ranking methodology is inspired by Google's PageRank which was introduced by Larry Page and Sergey Brin (14). The PageRank algorithm was developed as a way to rate the importance of web pages in the search engine that was developed by Google. Ed Feng (15) developed an algorithm based out of PageRank to rank sports teams based on a limited number of key statistics from historical data. Much of the data used for this study is derived from the Football Manager (16) game series. Davenport (17) published a report which highlighted the emergence of analytics in sport and how this helped teams find ways to remain competitive by utilising statistics.

For any business intelligence operation, data warehousing or data mining application, an effective provision to extract or accumulate the data into the application's data store needs to be in place so that this data can be converted into the relevant format needed for analysis.

To elaborate, this process can be performed by following steps: Data Collection, Data Integration, and Data Analysis. Data Collection involves the collection of the data from the relevant sources. Currently, sporting organisations use various techniques for collecting live data during official matches and training sessions in order to provide them with the basic repository in order to perform their analyses operations. This technology can be seen during these matches via certain match day experience features such as player heat maps, passing graph, shots graph, and possession.
statistics amongst others in football. Certain organisations (18,19) contribute heavily in this field where the primary importance of recording and transforming the data is focussed upon. This data is later produced as data feeds or raw materials for team performing their analyses. We have seen a large number of implementations of data mining in various industries which show the increasing importance in today's world. By using various data mining analysis, Rajeswari et all (24-29) detailed approaches that were applied across the mechanical domain in analysing machine performance data. This showed the potential of data mining in its capacity to provide insights in the respective domains.

3. Detailed Design of the Project

3.1 System Design and Architecture

The players are sorted into three categories namely Defenders, Midfielders and Forwards, and the key attributes relevant to each role are identified. Now the data can be studied to find the relevant patterns that need to be used to achieve the desired result. Different coaches or managers would have varying requirements as they would try to recruit players or select players based on a certain approach as per their individual ideology. As a result of this, the key performance indicator required for analysis would also vary depending on end user. Various scenarios are elaborated later on in this report.

The method proposed in this study identifies certain key statistics, otherwise called Key Performance Indicators (KPIs) in order to initiate the analyses. The primary data has been acquired from the sources (20,23) and are basically forming two categories – Overall team statistics accumulated over more than 10 seasons of the Barclays Premier League (21) and individual player data that have been synthesized for use in this report. Using this data, target objectives are determined to serve as a base for the results. The experiment is divided into two steps – Using the performance data to calculate the Player Performance Score (PPS) and using the scores to calculate PlayerRank (PR) in relation to the group of players. The PPS is further utilized to demonstrate methods of player selection and recruitment. Based on the player selections, each selection's performance statistics are tabulated and projected against the defined objectives. Multiple scenarios are considered and the results are analyzed by way of graphs and charts in order to provide relevant insight on the proposed methods.

3.2 Statistical Analysis

Statistical analysis is the part of data analytics which, in the context of business intelligence (BI), comprises of accumulating and analyzing every data sample in a set from which samples can be drawn. A sample can be defined as a representative selection drawn from a total population. The goal of statistical analysis is to identify patterns. In
sports, the application of statistical approaches have increased phenomenally over time and nowadays, various sports franchises, organisations and teams employ techniques and personnel for collect and analyze the data. Certain key performance indicators are first determined. This may vary from manager to manager based on his or her tendencies to play the game, but the philosophy is almost always the same. Once the KPIs are identified, it is used to project targets that will be set as team objectives for the upcoming season. This in turn will be used to identify individuals who can complement the team. Sæther (22) describes in detail about how coaches across the various leagues in Norway determine the concepts of talent identification in order to study the potential of various young players. The coaches use this knowledge to identify key performance indicators (KPIs) which would in turn be evaluated for each player before the final decision is made regarding whether to bring the player into the team or not. A simple architectural view of the system shows how each data set is grouped with the related data sets to derive further parameters which are later used for the analyses. The view of the two data sets and how they are linked together to form the underlying data framework is displayed below in fig.1 and fig.2.

**Fig 1**: Team Database – Data for all the recorded attributes for each team over the n seasons as the cumulative team data.

**Fig 2**: Player Database – Data for all the recorded attributes for the players in the team and cumulative player data.
All the data including the derived data from the two primary datasets can be stored as a cumulative datasets. The dataset contains two categories – Player and Team. Each of the two will contain certain common attributes. The modular view of the proposed method is shown in fig. 3.

**Fig. 3: Visual representation of the proposed model.**

### 3.3 Experimental Setup

There are certain principles that are considered in designing the methodology and algorithms. These principles are mentioned as follows:

1. Each player's performance data is used to identify certain Key Performance Indicators(KPIs) that can best describe the pivotal factors that can ultimately determine the calibre of a player in a team.
2. It is considered that certain attributes can be of importance to more than one tactical role i.e. it is not necessary that a certain attribute is relevant or important to one specific role only. With this in consideration, each attribute is assigned a weight factor to define it’s relative importance to each role.
3. A damping factor is also considered to include the possibility that a certain percentage of players may have multi-role suitable attributes.
4. The team average for each statistic is calculated and used for comparison. This aids the calculation in a way that it helps determine the calibre of the team in general as well as provide a degree of relation between the player and the team or group of players.
5. It is also considered that a player may individually possess attributes that might be suitable for other roles. This is introduced as the second damping factor

### 3.3.1 Player Performance data

The primary input for this experiment will be the historical performance data for each player within a group of players. Using the performance data, certain key performance indicators (KPIs) are defined. These are defined by determining those important statistics that are required to form the base of future calculations for the scores. Table 3.1 shows a list of attributes or key statistics and corresponding values of a certain players that will be used as the primary data to calculate player scores and subsequently the ranks.

#### Table 3.1 Sample Data for 18 players and their performance statistics.

<table>
<thead>
<tr>
<th>Player</th>
<th>Position</th>
<th>Tackles</th>
<th>Chances Created</th>
<th>Goals</th>
<th>Shots</th>
<th>Shots on Target</th>
<th>Interceptions</th>
<th>Dribbles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarke</td>
<td>D</td>
<td>5.60</td>
<td>1.00</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>2.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Kaluba</td>
<td>D</td>
<td>3.20</td>
<td>0.30</td>
<td>0.10</td>
<td>0.60</td>
<td>0.40</td>
<td>1.80</td>
<td>1.10</td>
</tr>
<tr>
<td>Jones</td>
<td>D</td>
<td>3.60</td>
<td>0.40</td>
<td>0.10</td>
<td>0.50</td>
<td>0.10</td>
<td>2.00</td>
<td>1.30</td>
</tr>
<tr>
<td>Shaw</td>
<td>D</td>
<td>2.30</td>
<td>0.50</td>
<td>0.20</td>
<td>0.30</td>
<td>0.20</td>
<td>1.40</td>
<td>1.20</td>
</tr>
<tr>
<td>Rothery</td>
<td>D</td>
<td>1.80</td>
<td>0.40</td>
<td>0.00</td>
<td>0.30</td>
<td>0.10</td>
<td>2.40</td>
<td>0.80</td>
</tr>
<tr>
<td>Turner</td>
<td>D</td>
<td>1.80</td>
<td>0.40</td>
<td>0.00</td>
<td>0.30</td>
<td>0.20</td>
<td>1.60</td>
<td>0.90</td>
</tr>
<tr>
<td>Thomas</td>
<td>D</td>
<td>3.10</td>
<td>3.60</td>
<td>0.20</td>
<td>0.50</td>
<td>0.40</td>
<td>2.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Phillips</td>
<td>F</td>
<td>1.50</td>
<td>0.90</td>
<td>0.90</td>
<td>1.90</td>
<td>1.10</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>Hare</td>
<td>F</td>
<td>1.80</td>
<td>0.90</td>
<td>1.30</td>
<td>1.80</td>
<td>1.40</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>Walmsley</td>
<td>F</td>
<td>1.20</td>
<td>1.00</td>
<td>0.00</td>
<td>0.30</td>
<td>0.30</td>
<td>0.60</td>
<td>0.90</td>
</tr>
<tr>
<td>Parker</td>
<td>M</td>
<td>3.10</td>
<td>3.00</td>
<td>0.20</td>
<td>0.50</td>
<td>0.20</td>
<td>1.90</td>
<td>1.40</td>
</tr>
<tr>
<td>Lloyd</td>
<td>M</td>
<td>1.50</td>
<td>3.60</td>
<td>0.30</td>
<td>0.80</td>
<td>0.40</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>Sullivan</td>
<td>M</td>
<td>0.80</td>
<td>4.20</td>
<td>0.20</td>
<td>1.80</td>
<td>1.20</td>
<td>1.20</td>
<td>0.90</td>
</tr>
<tr>
<td>Roberts</td>
<td>M</td>
<td>1.90</td>
<td>3.70</td>
<td>0.20</td>
<td>0.70</td>
<td>0.60</td>
<td>1.20</td>
<td>0.80</td>
</tr>
<tr>
<td>Roberts</td>
<td>M</td>
<td>1.30</td>
<td>2.50</td>
<td>0.00</td>
<td>0.50</td>
<td>0.40</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>Stewart</td>
<td>M</td>
<td>5.70</td>
<td>3.60</td>
<td>0.00</td>
<td>0.40</td>
<td>0.40</td>
<td>1.90</td>
<td>1.00</td>
</tr>
<tr>
<td>Williams</td>
<td>M</td>
<td>2.20</td>
<td>0.40</td>
<td>0.20</td>
<td>0.40</td>
<td>0.30</td>
<td>1.80</td>
<td>1.00</td>
</tr>
<tr>
<td>P Jones</td>
<td>M</td>
<td>1.70</td>
<td>0.50</td>
<td>0.00</td>
<td>0.30</td>
<td>0.30</td>
<td>1.30</td>
<td>0.90</td>
</tr>
</tbody>
</table>
3.3.2 Attribute Weightage

In order to satisfy the condition that each attribute is of importance to each of the three outfield tactical roles, weights are determined. These weights will provide an indication of the relevant importance of each attribute with respect to each role. Table 3.2 displays the weighting table. The weights used in this case vary between the three values 0.20, 0.35 and 0.45. These weights will provide the relation between each attribute and role such that three distinct performance scores can be calculated.

Table 3.2 Table describing weightage assigned to each attribute.

<table>
<thead>
<tr>
<th>Stat</th>
<th>D</th>
<th>M</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passes</td>
<td>0.35</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>Pass Completion</td>
<td>0.35</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>Key Passes</td>
<td>0.35</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>Tackles</td>
<td>0.45</td>
<td>0.35</td>
<td>0.20</td>
</tr>
<tr>
<td>Tackles Won</td>
<td>0.45</td>
<td>0.35</td>
<td>0.20</td>
</tr>
<tr>
<td>Key Tackles</td>
<td>0.45</td>
<td>0.35</td>
<td>0.20</td>
</tr>
<tr>
<td>Aerial Challenges</td>
<td>0.45</td>
<td>0.20</td>
<td>0.35</td>
</tr>
<tr>
<td>Aerial Challenges Won</td>
<td>0.45</td>
<td>0.20</td>
<td>0.35</td>
</tr>
<tr>
<td>Key Aerial Challenges</td>
<td>0.45</td>
<td>0.20</td>
<td>0.35</td>
</tr>
<tr>
<td>Chances Created</td>
<td>0.20</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>Goals</td>
<td>0.20</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>Shots</td>
<td>0.20</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>Shots on Target</td>
<td>0.20</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>Interceptions</td>
<td>0.45</td>
<td>0.35</td>
<td>0.20</td>
</tr>
<tr>
<td>Dribbles Attempted</td>
<td>0.20</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>Dribbles Completed</td>
<td>0.20</td>
<td>0.45</td>
<td>0.35</td>
</tr>
</tbody>
</table>

3.3.3 Multi-Role Suitability

It is taken into account the fact that players will almost always have impressive statistics in areas that are not his or her primary role. This is considered as a damping factor in the calculations. By analyzing the input data across the attributes, it can be observed that about 80% of the players possess attributes that can be suited to their non-primary role. Hence, the damping factor will be considered as 0.8.
3.3.4 KPI Team Average

The average value for each KPI or attribute is calculated. This average value is the team average for that particular statistic. This will provide the relation between each individual player and the team or group of player with whom this player is being ranked against. Utilising the average values for the attributes as a base for comparison and analysis satisfied the following principles:

a. As the team average increases, so does the standard of the group of players

b. As the team average increases, the player's individual scores shall also increase

3.3.5 Multi-Role Attributes

Individually, players may possess skill sets that can be suited for other roles. In this player set, an analysis to identify the percentage of attributes a player possesses outside of their primary role attributes i.e. of all the attributes that are used to determine the player's score resulted in an observation that produced a percentage of those attributes that satisfy the condition that a player can play more than one role. This was determined by comparing each attribute or KPI with the role that holds maximum weightage. The individual value for each player was compared with the team average. If the attribute held a higher value than team average, it was considered that the player is suitable for that role. Using this method, each non primary attribute was compared and analyzed and a final percentage was determined that defined the relevance of the player's attributes (including primary) towards non-primary roles for each player. This value was considered to be a second damping factor which was used in the scoring and ranking calculations in the later stages of the experiment. The value used is 0.6 as it was observed that about 60% of a player's attributes could be effective for non-primary tactical roles.

3.3.6 Team performance data

This data is an accumulation of statistics for the Barclays Premier League (21) over an extended period of time. The time frame used here from 2003-2004 to 2014-2015 is 12 seasons (years) worth of data that show the various team positions over the years. The final derived data set seen here is taken as an average of each statistic over the time period across the positions 1 to 4.

4. Implementation of the methodology

It is imperative to find a relation between the performance data of an individual player and the team average for each attribute and considering the importance of an attribute to a tactical role, introduce a weight factor for each attribute based on the relative importance of each to the role (Defence, Midfield and Forward) and use this to calculate a Player Performance Score (PPS). These scores can be utilised by the coaches or team managers for analysis before
selecting the players for the next game based on their preferred tactical operation. The score itself is turn used to
develop a ranking system which can be used to determine the importance of players based on their past performance.
Certain assumptions will be elaborated in the Results and Discussion section for ease of understanding the
calculations.

Table 4.1 League data showing the final positions of the 20 teams in the 2014-2015 seasons.

<table>
<thead>
<tr>
<th>Position</th>
<th>Team</th>
<th>Played</th>
<th>W</th>
<th>D</th>
<th>L</th>
<th>F</th>
<th>A</th>
<th>GD</th>
<th>Pts</th>
<th>Fl</th>
<th>Y</th>
<th>R</th>
<th>Sh</th>
<th>ShT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chelsea</td>
<td>38</td>
<td>26</td>
<td>9</td>
<td>3</td>
<td>73</td>
<td>32</td>
<td>41</td>
<td>87</td>
<td>382</td>
<td>77</td>
<td>4</td>
<td>563</td>
<td>210</td>
</tr>
<tr>
<td>2</td>
<td>Man City</td>
<td>38</td>
<td>24</td>
<td>7</td>
<td>7</td>
<td>83</td>
<td>38</td>
<td>45</td>
<td>79</td>
<td>445</td>
<td>77</td>
<td>2</td>
<td>668</td>
<td>228</td>
</tr>
<tr>
<td>3</td>
<td>Arsenal</td>
<td>38</td>
<td>22</td>
<td>9</td>
<td>7</td>
<td>71</td>
<td>36</td>
<td>35</td>
<td>75</td>
<td>377</td>
<td>68</td>
<td>2</td>
<td>612</td>
<td>228</td>
</tr>
<tr>
<td>4</td>
<td>Man United</td>
<td>38</td>
<td>20</td>
<td>10</td>
<td>8</td>
<td>62</td>
<td>37</td>
<td>25</td>
<td>70</td>
<td>453</td>
<td>64</td>
<td>5</td>
<td>509</td>
<td>179</td>
</tr>
<tr>
<td>5</td>
<td>Tottenham</td>
<td>38</td>
<td>19</td>
<td>7</td>
<td>12</td>
<td>58</td>
<td>53</td>
<td>5</td>
<td>64</td>
<td>441</td>
<td>79</td>
<td>4</td>
<td>524</td>
<td>172</td>
</tr>
<tr>
<td>6</td>
<td>Liverpool</td>
<td>38</td>
<td>18</td>
<td>8</td>
<td>12</td>
<td>52</td>
<td>48</td>
<td>4</td>
<td>62</td>
<td>385</td>
<td>66</td>
<td>3</td>
<td>590</td>
<td>194</td>
</tr>
<tr>
<td>7</td>
<td>Southampton</td>
<td>38</td>
<td>18</td>
<td>6</td>
<td>14</td>
<td>54</td>
<td>33</td>
<td>21</td>
<td>60</td>
<td>468</td>
<td>57</td>
<td>3</td>
<td>506</td>
<td>170</td>
</tr>
<tr>
<td>8</td>
<td>Swansea</td>
<td>38</td>
<td>16</td>
<td>8</td>
<td>14</td>
<td>46</td>
<td>49</td>
<td>-3</td>
<td>56</td>
<td>398</td>
<td>48</td>
<td>5</td>
<td>422</td>
<td>147</td>
</tr>
<tr>
<td>9</td>
<td>Stoke</td>
<td>38</td>
<td>15</td>
<td>9</td>
<td>14</td>
<td>48</td>
<td>45</td>
<td>3</td>
<td>54</td>
<td>486</td>
<td>82</td>
<td>1</td>
<td>501</td>
<td>136</td>
</tr>
<tr>
<td>10</td>
<td>Crystal</td>
<td>38</td>
<td>13</td>
<td>9</td>
<td>16</td>
<td>47</td>
<td>51</td>
<td>-4</td>
<td>48</td>
<td>527</td>
<td>63</td>
<td>4</td>
<td>438</td>
<td>139</td>
</tr>
<tr>
<td>11</td>
<td>Everton</td>
<td>38</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td>48</td>
<td>50</td>
<td>-2</td>
<td>47</td>
<td>388</td>
<td>66</td>
<td>2</td>
<td>484</td>
<td>166</td>
</tr>
<tr>
<td>12</td>
<td>West Ham</td>
<td>38</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td>44</td>
<td>47</td>
<td>-3</td>
<td>47</td>
<td>420</td>
<td>64</td>
<td>2</td>
<td>485</td>
<td>152</td>
</tr>
<tr>
<td>13</td>
<td>West Brom</td>
<td>38</td>
<td>11</td>
<td>11</td>
<td>16</td>
<td>38</td>
<td>51</td>
<td>-13</td>
<td>44</td>
<td>423</td>
<td>64</td>
<td>3</td>
<td>411</td>
<td>130</td>
</tr>
<tr>
<td>14</td>
<td>Leicester</td>
<td>38</td>
<td>11</td>
<td>8</td>
<td>19</td>
<td>46</td>
<td>55</td>
<td>-9</td>
<td>41</td>
<td>457</td>
<td>49</td>
<td>4</td>
<td>457</td>
<td>137</td>
</tr>
<tr>
<td>15</td>
<td>Newcastle</td>
<td>38</td>
<td>10</td>
<td>9</td>
<td>19</td>
<td>40</td>
<td>63</td>
<td>-23</td>
<td>39</td>
<td>433</td>
<td>66</td>
<td>7</td>
<td>467</td>
<td>143</td>
</tr>
<tr>
<td>16</td>
<td>Sunderland</td>
<td>38</td>
<td>7</td>
<td>17</td>
<td>14</td>
<td>31</td>
<td>53</td>
<td>-22</td>
<td>38</td>
<td>435</td>
<td>94</td>
<td>3</td>
<td>404</td>
<td>129</td>
</tr>
<tr>
<td>17</td>
<td>Aston Villa</td>
<td>38</td>
<td>10</td>
<td>8</td>
<td>20</td>
<td>31</td>
<td>57</td>
<td>-26</td>
<td>38</td>
<td>400</td>
<td>70</td>
<td>7</td>
<td>414</td>
<td>127</td>
</tr>
<tr>
<td>18</td>
<td>Hull</td>
<td>38</td>
<td>8</td>
<td>11</td>
<td>19</td>
<td>33</td>
<td>51</td>
<td>-18</td>
<td>35</td>
<td>451</td>
<td>71</td>
<td>6</td>
<td>429</td>
<td>131</td>
</tr>
<tr>
<td>19</td>
<td>Burnley</td>
<td>38</td>
<td>7</td>
<td>12</td>
<td>19</td>
<td>28</td>
<td>53</td>
<td>-25</td>
<td>33</td>
<td>406</td>
<td>64</td>
<td>2</td>
<td>429</td>
<td>125</td>
</tr>
<tr>
<td>20</td>
<td>QPR</td>
<td>38</td>
<td>8</td>
<td>6</td>
<td>24</td>
<td>42</td>
<td>73</td>
<td>-31</td>
<td>30</td>
<td>447</td>
<td>75</td>
<td>3</td>
<td>533</td>
<td>151</td>
</tr>
</tbody>
</table>

Table 4.1 shows data from the Barclays Premier League (21) season 2014-2015. Similar to the table seen, data for 12 seasons of the tournament has been accumulated and a new table containing averages for all the statistics has been derived. Tables 3.3.1, 3.3.2 and 3.3.3 show the averages across the first four positions over the 12 seasons.

From this data, there have been certain patterns found and observations made. Key mentions as follows:

1. It was found that the winning team averaged 88.25 points across the 12 seasons

2. It was observed that the winning team won an average of 27.17 games out of 38 every season
3. On average, the top four teams average a goal to shots on target conversion rate between 22 and 26 percent. This meant that they scored goals on about one-fourth of their shots they took on target.

4. The champions almost always averages 82 goals scored per season and 29 goals conceded.

These observations mentioned above, amongst others form the basic setup for the experiment wherein two sets of calculations will be performed. These calculations are based on two formulae that take into account all the relevant parameters based on the task ahead and the sporting attributes to be evaluated. The calculation steps are divided into two parts – Player Performance Score (PPS) with which a method of player selection and recruitment is proposed and PlayerRank which can be used to define a ranking system for the players which acts as a comparison or evaluation of the players relative to each other.

4.1 Player Performance Score

The PPS is proposed here as a means to devise a method of evaluating a player's performance over a certain number of games or seasons. Three values are proposed here as part of the PPS – Defensive Score, Midfield Score and Forward Score – one for each tactical role per player.

The PPS is based on the principles that were defined earlier as part of the KPI principles. By calculating a score for each role by using each attribute of the player and attaining it's product with the corresponding weightage across each of the three tactical roles, the final results will be as follows:

The calculation of the PPS from the attributes is carried out using the following factors:

\( a \) – attribute value

\( w \) – Weight value for attribute (\( w_d \), \( w_m \), \( w_f \) – weight for each of the three roles)

Score Calculation:

Using these variables, the equation to calculate the player score can be defined as:

\[
Score_{\text{defensive}} = w_{\text{defensive}} \sum_{i=a_{d}}^{a} a_i \\
Score_{\text{midfield}} = w_{\text{midfield}} \sum_{i=a_{m}}^{a} a_i \\
Score_{\text{forward}} = w_{\text{forward}} \sum_{i=a_{f}}^{a} a_i
\]

This derives a data set that is classified as the player scores across the three tactical roles. It is using these scores with which:

a. Player Selection and Recruitment will be proposed
b. PlayerRank will be calculated for each player

The results will be seen in later section with the use cases.

The Player Performance Score can be calculated in two ways – for each match as well as over a certain set of matches. In this calculation, the total over a set of matches is considered.

The PPS for each player is considered while devising the player selection for the future games. Players for each position are selected based on the highest PPS for each tactical role or position. Following this, the potential line ups are made.

The PlayerRank is calculated using the formula:

$$ RankPoint_{Player} = m \times p \left( \frac{Score_{Player} - Score_{AverageRem}}{n} \right) $$

Where:

- m – Damping factor 1 (players suited to more than one role) – value taken as 0.8
- p – Damping factor 2 (attributes of a player from total attributes that possess higher than team average) – value taken as 0.6
- Rank Points – Contribution of each player's score relative to remaining players. This is used to determine the Rank of the player
- $Score_{Player}$ – Score of the player
- $Score_{AverageRem}$ - Average Team Score for the remaining players
- n – Number of players in team

The PlayerRank is then calculated after sorting the Rank Points for each player

The principle of the calculation here follows the lines of the one used in the PageRank algorithm. In the PageRank (14) algorithm, the principle behind calculating the relevant importance of each page lies in the fact that it is dependant on the rank of the pages that are part of its linkage (i.e. pages that are linked to it either by way of outgoing or incoming links). The PlayerRank also takes flight from the PageRank principle such that each player's rank is dependent on the other players present in the group or team.

**Table 4.3 Table showing the PlayerRank calculated for each player.**

<table>
<thead>
<tr>
<th>Player Name</th>
<th>Ranking Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones</td>
<td>0.13</td>
<td>1</td>
</tr>
<tr>
<td>Stewart</td>
<td>0.09</td>
<td>2</td>
</tr>
</tbody>
</table>
Seen in table 4.3 is a list of the players ranked relative to each other based on the Ranking Scores that was calculated from their individual PPS. Identifying the KPIs is an important task in this process. A key aspect of the KPIs is that it can vary between users. For example, a coach or a team manager might prioritize certain attributes over others and consider them as KPIs for his or her management philosophies as compared to other coaches or managers. It is important to note that these attributes have been considered for this experiment and its use cases and that attributes may change with different end users. Using these statistics, analyses were done and the following target objectives were determined:

1. For any team targeting to win the tournament, it would be required to achieve a minimum of 27 wins
2. 82 goals would need to be the target number of goals to be scored and less than 30 conceded in order to maintain a healthy goal difference which would reflect in points
3. The title winners over the analyzed 12 seasons averaged at 88 points per season. Hence, it would be imperative that any team harbouring title hopes would need to achieve close to the same number of points.

The individual statistics for each player is used to calculate a team total for the same attribute. Using these, it is projected over the entire season's worth of games. Following the analyses of the performance data, a correlation between two parameters – Goals and Wins – is identified. While a simplistic approach would be to declare that the
more number of goals scored would obviously result in more wins, there are certain other sub factors that must be considered – Goals conceded and the number of matches that do not end up as wins.

For this purpose, the following four KPIs are derived and used:

1. Goals Scored
2. Goals Conceded
3. Wins
4. Non-Wins

To correlate these four KPIs, the following relation is introduced:

\[ \text{Wins per goal} = \frac{\text{Wins}}{\text{Goals Scored}} \]

\[ \text{Non-Wins per goal} = \frac{\text{Non-Wins}}{\text{Goals Conceded}} \]

As part of the methodology proposed, the primary target for any team irrespective of the position is considered to be the number of wins. Using the above mentioned KPIs, the overall team performance is measured against the targets set and using this, the number of wins is projected. This will also lead to projecting the number of points a team can achieve.

Following the calculation of the scores, the manager can select the team based on their application of strategy. Each selection made will be measured across the KPIs defined and compared against the target objectives that were determined prior.

5. Results and Discussions

The application of the results of the performed calculation is one that defines the objective of developing this Player Performance Score and Ranking methodology. The results are explained in detail using use cases. The applications of the results are proposed in the following areas:

1. Player Selection
2. Player Recruitment
3. Player Ranking

5.1 Team Selection

Every coach will have a tendency to approach every match based on his or her tactical preferences and the opposition's tactics and form factors. In football, the most commonly used formation is 4-4-2 comprising of 4 defenders, 4 midfielders and 2 forwards and 4-3-3 which has 4 defenders, 3 midfielders and 3 forwards. Fig 4 depicts these formations. Four use cases are presented in this section in order to provide a clearer picture regarding how
player selection for various formations can affect team performance. As a base for the analysis, target objectives have been derived from studying the historical data. The player data serves as a basis for selecting the players following which, each player's achievements are used in order to forecast or project potential team achievement for the upcoming season.

![Image of football formations](image)

**Fig. 4. The 4-4-2 formation and the 4-4-3 Formation.**

The use cases discussed are four team formations and selection the best possible line up based on the PPS.

1. 4-4-2 Formation
2. 4-3-3 Formation
3. 4-5-1 Formation
4. 3-5-2 Formation

The selection of players in tactical positions will follow the pattern Defender-Midfielder-Forward i.e. 4-4-2 implies that the selection will consist of 4 defenders, 4 midfielders and 2 forwards and so on. Using the PPS, the players are first sorted into their tactical roles (defenders, midfielders and forwards) and then sorted on PPS. Players with the highest scores per role are selected for their respective roles.

An example is shown below showing how the players are sorted and selected for the 4-4-2 formation. Tables 5.1, 5.2, 5.3 depict the players after sorting their scores.

Note: These formations are considered assuming that the end user (coach or manager) considers Goals Scored as the defining factor towards achieving wins. At times, it is possible that a defensive minded coach may consider Goals Conceded as the defining factor. In such cases, Goals Conceded will be taken for calculation.

Step 1: Sorting the defenders based on the D-Score. The top four defenders are selected to take the four defender positions.

Step 2: Sorting the midfielders based on the M-Score. The best four midfielders are taken for the midfield positions.
Step 3: Sorting the forwards on their F-Score. The top two forwards are considered for selection.

Hence, based on this selection, the 4-4-2 formation can be fulfilled with the following line-up as shown in table 5.4.

Similarly, the remaining three formations, when selected in the same method proposed, will provide the following selections as seen in tables 5.5, 5.6 and 5.7. Here, the following KPIs are considered as mentioned in earlier section: Goals Scored, Goals Conceded, Wins and Non-Wins. The wins and non-wins are calculated using the historical data as discussed in the previous section.

**Table 5.4 Selection for 4-4-2 formation.**

<table>
<thead>
<tr>
<th>Formations</th>
<th>Player</th>
</tr>
</thead>
<tbody>
<tr>
<td>4_4_2</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Thomas</td>
</tr>
<tr>
<td>D</td>
<td>Jones</td>
</tr>
<tr>
<td>D</td>
<td>Kaluba</td>
</tr>
<tr>
<td>D</td>
<td>Shaw</td>
</tr>
<tr>
<td>M</td>
<td>Parker</td>
</tr>
<tr>
<td>M</td>
<td>Stewart</td>
</tr>
<tr>
<td>M</td>
<td>Roberts</td>
</tr>
<tr>
<td>M</td>
<td>Sullivan</td>
</tr>
<tr>
<td>F</td>
<td>Phillips</td>
</tr>
<tr>
<td>F</td>
<td>Hare</td>
</tr>
</tbody>
</table>

For a team competing to win the tournament, it would need to target the following:

a. 27 wins

b. 82 goals scored (attack-minded manager)

c. 29 goals conceded (defence-minded manager)

Using each of the selections derived above, the following forecasts or projections are made for each of the four selections as shown in table 5.8.

**Table 5.8 Comparison between selections for each formation.**

<table>
<thead>
<tr>
<th>Comparative Analysis</th>
<th>Goals Scored</th>
<th>Goals Conceded</th>
<th>Wins</th>
<th>Non Wins</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>81.92</td>
<td>28.83</td>
<td>27.17</td>
<td>10.83</td>
<td>88.25</td>
</tr>
<tr>
<td>4_4_2</td>
<td>129.20</td>
<td>42.56</td>
<td>26.92</td>
<td>11.08</td>
<td>86.29</td>
</tr>
<tr>
<td>4_3_3</td>
<td>121.60</td>
<td>43.32</td>
<td>24.02</td>
<td>13.98</td>
<td>79.06</td>
</tr>
<tr>
<td>4_5_1</td>
<td>91.20</td>
<td>41.42</td>
<td>14.66</td>
<td>23.34</td>
<td>55.65</td>
</tr>
<tr>
<td>3_5_2</td>
<td>133.00</td>
<td>44.08</td>
<td>27.54</td>
<td>10.46</td>
<td>87.86</td>
</tr>
</tbody>
</table>

The forecasting can be summarized as follows:
1. For each of the selections, the corresponding projections calculated as to how many goals will be scored is shown in the table. This is calculated by taking the sum of each of the individual players' contribution to Goals Scored per game and multiplies across 38 games (one season is 38 games).

2. Based on the wins per goal scored value calculated earlier, the number of goals scored would contribute to the corresponding values in the wins column.

3. Based on the wins projected, the points achieved is also forecasted as shown in the points column.

The following assumptions are made in relation to goals conceded:

1. From the tackles, the difference between tackles and tackles won is considered as Tackles Lost. It is assumed that on an average, 50% of the lost tackles lead to chances being created by the opposition.

2. Furthermore, from this 50%, it is assumed that 25% of them are shots on target taken by the opposition.

3. From this 25%, 25% of the shots on target are scored and this amounts to the goals conceded.

4. From the Non-Wins, it is considered that 50% of them are draws hence earning the team 1 point per draw.

![Fig 5 Projection for goals scored for each selection.](image)

As seen in the fig 5, the projected number of goals scored for each formation is weighed along with the target goals to be scored. From this, it clearly shows that the 3-5-2 formation works best.

From fig 6, the goals conceded are a very close call but it shows that all of these formations seem to be well in excess of the objective defined. This can be modified by choosing defensive KPIs over attacking ones. Despite the excess, the 4-5-1 formation is seen to have the better defensive record.

![Fig 6 Projection for the Goals Conceded for each selection.](image)
The fig 7 suggests that the 3-5-2 formation shows the best return for points going narrowly close to the target defined, even though it is slightly less than the target.

![Fig 7 Comparison graph between each selection measured across projected point’s total.](image)

**Fig 7** Comparison graph between each selection measured across projected point’s total.

Fig. 8 shows the percentage of wins to the total games played. From this, the 3-5-2 formation once again seems most suitable.

These use cases have been designed with a manager who tends to be more attack minded in developing the strategy for his or her team to follow. For defence minded managers, the KPIs can be altered depending on their preferences.

![Fig 8 Win to Non-Win ratio for each selection.](image)

**Fig 8** Win to Non-Win ratio for each selection.

To put it simply, the purpose of this analysis is to propose a selection methodology which would allow the manager to view each selection dynamically based on his or her preferences and assist in the decision making process, much like a business intelligence application.

After analysing the data for the past 12 seasons and making the observations, the technique used for projecting the achievements where tested against the actual current season data in order to check the accuracy of the forecasts. At the time of writing this report, the number of games played this season was around 28 games per team. The following attributes were compared - Goals Scored and Goals Conceded with respect to the number of Wins and Non-Wins respectively. When analyzed, the results turned out to match six of the eight parameters within an accuracy of 0.9 i.e.
for the teams placed in the first through fourth positions in the current season, the values for the number of goals each of the four teams scored and allowed were used and multiplied with the Win/Non-Win per goal values in order to project how many matches each team would win. The results were compared with the actual season statistics for this season and it was found that they matched for six of the eight results within accuracy of 0.9.

Table 5.9 Projection comparison with actual data from current season.

<table>
<thead>
<tr>
<th>Position</th>
<th>Projections Wins</th>
<th>Actual Wins</th>
<th>Deviation Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.91</td>
<td>16</td>
<td>-0.91</td>
</tr>
<tr>
<td>2</td>
<td>15.72</td>
<td>15</td>
<td>-0.72</td>
</tr>
<tr>
<td>3</td>
<td>14.79</td>
<td>15</td>
<td>0.21</td>
</tr>
<tr>
<td>4</td>
<td>14.84</td>
<td>14</td>
<td>-0.84</td>
</tr>
</tbody>
</table>

The following observations were made with reference to Table 5.9:

1. For the teams placed in the top four positions across the analyzed data for 12 seasons of the tournament, it was seen that the average number of wins achieved with every goal scored was 0.33, 0.32, 0.34 and 0.31 for the teams positioned one, two, three and four respectively.

2. Using the model proposed in this thesis, these Win values were used to project the number of wins and points that teams occupying the first four positions in the current seasons' standings were likely to have. These projections were compared with the actual values attained this season.

3. The projections gave the following results for the teams in the first four positions – 16.91, 15.72, 14.79 and 14.84 Wins respectively. These values were within 0.9 wins of the actual values for this season. The number of wins was calculated by multiplying the number of goals scored by the team in each position with the Wins/per goal scored value for each position.

4. The same process was done for the goals conceded and multiplied with the value for Non-Wins. The results were projected as 11.65, 8.90, 13.04 and 14.82 for the positions one, two, three and four respectively. It was seen that except for the second placed team, the remaining teams projected to within 1 Non-Win of the actual value.

5. With the assumption taken that 50% of Non-Wins are Draws that earn a point each, the number of points were calculated for each team in the top four positions. The projected values show that the results were within 5 points of the actual (5 points deviation is considered for this theory). The projected values were 56.56, 51.60,
50.90 and 51.93 points respectively for the top four positions which were 0.44, 2.40, 0.10 and 4.93 points from the actual points the teams have accomplished at the time of considering the calculations. Table 5.9 displays the projection comparison.

Note: 5 points deviation is considered since it has been assumed that 50% of the Non-Wins are considered draws worth one point each to support ease of calculation.

These results go to show that the projection model proposed in this system were relatively accurate to within 75% of the actual results. With further work and sufficient amounts of data, it would be possible to find more substantial patterns that could help bring the gap closer.

5.2 Recruitment of players

Depending on the requirement of the manager and the coach of the team, they might look to recruit certain players to fill out their team roster. The manager would analyze the past team performance over a certain time period to determine the projection for the targets i.e. how many points would the team need to win in order to win the trophy this season or how many goals would the team need to score in order to put themselves in a winning positions. Based on the manager's need, the questions that are asked can vary inadvertently. Based on these questions, the scouting and recruitment team will search for available players who can meet the individual targets and help the team. To elaborate this, a scenario is explained below.

Consider that after studying the data for over 10 seasons, the analyst has observed that the trophy winning team has averaged 75 points per season. Assuming that the number of teams competing in this league is 20 and that each team plays every team twice –home and away, the total number of games played by a team is 38. Hence, this averages to 1.97 points per match.

The analyst also observes that the average number of goals scored by teams that finished in the top three for each of those 10 seasons totals to about 90 goals per season which in turn is 2.36 goals per match. Using these observations, it is projected that the team would need to average about 2 points per game and well over 2 goals per game in order to stand a chance of winning the trophy.

Analyzing the existing players' data and past seasons' data for the team in question, consider that the analyst observed that the team scored an average of 1.6 points per game and scored 1.85 goals per game, falling short of almost 0.4 points and 0.51 goals per game. Table 5.10 shows data for two players across certain KPIs – Goals scored and Shots on Target. These two players are not of the current team in question. They are potential recruitment targets for the manager in order to find a solution to his goal scoring issues.
Based on this, the scouting team can analyze the vast amounts of player data in order to find players that can fit these criteria i.e. they can look for players who can help score 0.5 goals per game in order to make up the deficit that the team currently is having. This would be one approach of recruiting a forward player. However, if the team has forward players of a good calibre, the manager might use this information to find out other KPIs such as chances created per game for instance and he could use this to find players that create more chances per game in order to help his forward players score more. This approach can be customized across the various roles and preferences of the manager.

Consider a scenario where a manager is looking to recruit a player for the midfield role. This scenario is one where his team is creating an average of 5 goal scoring chances per game. Based on the historical analysis, a championship winning team needs to score about 2.1 goals per game. It is also seen that the past data throws out the information that the winning team has on an average converted 25% of their shots on target into goals. Let us assume that of all the chances created per match, 25% of the chances lead to shots on target.

Based on the player data studied, the players contribute to 3.5 goals per game. This happens to be a good return. However, in the case that a team that scored less than the required goals, but the forwards have a good return on shot conversion; it would be wise to look into midfielders that create more chances. Using this KPI, the manager can identify and analyze players based on chances created per game attribute and use this to find the best fit for his team.

The scope for this approach to be customized according to preferences is vast which can result in a details study of each attribute and determining the KPIs as required.

### 5.3 Ranking the players

The ranking algorithm is derived from the player scores that have been calculated earlier. Instead of using the individual scores of players in order to rank them, a slightly different approach is used such that calculation of a
player's rank takes into account the variables of individual player scores, average team score and the damping factors $m$ and $p$. Using these variables, the following conditions are satisfied:

a. the rank of a player is dependant on his or her individual score

b. the rank of a player is dependant on the scores of his or her team mates

c. the rank of a player is dependant on the average performance score for the team

Points b and c introduce the fact that each player's rank is relative to the group of players he or she is part of. It is understood that as the team average for any attribute or score increases, the calibre of the team also increases. Creating a ranking method is extremely useful in determining the calibre of the players relative to the group of players that the testing sample comprises of. This group can be of varied types – a single team of players, or groups of players within the same team sorted based on positions or even groups of players from multiple team. This report considers players from a single team. The rank is developed using a simple algorithm to calculate the same from the player's performance score and can be an effective way to rank the players over a larger time period (for example, a season or multiple consecutive seasons) instead of solely using the player's performance scores that can vary after each match. Table 5.11 shows a sample ranking list.

**Table 5.11 PlayerRank for each player.**

<table>
<thead>
<tr>
<th>Player Name</th>
<th>Ranking Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones</td>
<td>0.13</td>
<td>1</td>
</tr>
<tr>
<td>Stewart</td>
<td>0.09</td>
<td>2</td>
</tr>
<tr>
<td>Shaw</td>
<td>0.08</td>
<td>3</td>
</tr>
<tr>
<td>Kaluba</td>
<td>0.08</td>
<td>4</td>
</tr>
<tr>
<td>Parker</td>
<td>0.08</td>
<td>5</td>
</tr>
<tr>
<td>Clarke</td>
<td>0.05</td>
<td>6</td>
</tr>
<tr>
<td>Thomas</td>
<td>0.04</td>
<td>7</td>
</tr>
<tr>
<td>Rothery</td>
<td>0.03</td>
<td>8</td>
</tr>
<tr>
<td>Roberts</td>
<td>0.03</td>
<td>9</td>
</tr>
<tr>
<td>Phillips</td>
<td>0.02</td>
<td>10</td>
</tr>
<tr>
<td>Turner</td>
<td>0.01</td>
<td>11</td>
</tr>
<tr>
<td>Williams</td>
<td>-0.01</td>
<td>12</td>
</tr>
<tr>
<td>Sullivan</td>
<td>-0.02</td>
<td>13</td>
</tr>
<tr>
<td>Lloyd</td>
<td>-0.02</td>
<td>14</td>
</tr>
<tr>
<td>Hare</td>
<td>-0.03</td>
<td>15</td>
</tr>
<tr>
<td>Roberts</td>
<td>-0.15</td>
<td>16</td>
</tr>
</tbody>
</table>
6. Conclusions and Future Work

6.1 Conclusions

This study hypothesizes certain techniques that attempt to bridge the gap between implementing a Moneyball (2) inspired method of statistical analysis and football. The analyses also introduces a statistical method of selecting teams based on various tactical strategies based on the manager's preferences and attempts to use each selections' in-game statistics to project the KPIs for each selection over the course of the number of games that will be played during the season. The creation of the PPS enabled implementing this concept into the task of player selection. Using the projections and comparing them to the defined target objectives that were derived from analysing the historical data; this method assists the managers in planning their team selection and strategy for the future games.

This study allowed building on a method of recruiting players for the team by studying the historical data for the current team and identifying weak areas that would need to be improved. Using this information, it is applied into the list of players that may be available for recruitment to identify which of the players would be most suitable to fill in the gaps observed by the prior analyses. The study enabled the development of a formula that can be used for ranking the players in among the group in question whether they are part of the same team or sub groups of players within the team or even a group of randomly assembled players. The ranking algorithm used that PPS and certain key factors to derive a player rank for each player relative the players in the same group.

However, it is to be noted that there are drawbacks to the methodologies expressed above. They will be elaborated in the next sections.

6.2 Limitations of the proposed approach

Despite the potential this proposed methodology offers, it has serious limitations which would need to be addressed.

The key limitations are explained below:

6.2.1 Real-Time Implementation

The lack of a real time scenario to effectively test this hypothesis hinders the potential to evaluate it and ultimately refine the process. For any theory, concept, method, application, etc. to successfully work, it is important to release it or test it in a live environment. With this, it is possible to observe the results as it occurs and identify areas to strengthen, modify or improvise which will result in an effective refinement of the process en route to ensuring that a more effective method can be developed out of it.
6.2.2 Data Collection
This can be done in the old fashioned manner, such as with a pen and a pad of paper and watching a live match and noting down manually all the relevant parameters, or by using various technological means such as recording the data on computers by watching live matches or replays, or setting up a wireless sensor network so that the sensors would be programmed to track and collect the required data and store the same in the central database network. There are various organisations that perform this type of tasks (18,19). Acquiring this data as per the specific requirement can also be done by contacting the various organisations that perform this type of work and sharing the relevant requirements as to what type of data is needed.

6.2.3 Real Time Data
The Premier League (21) data that was acquired from (20) was factual and provided the platform to perform a historical study to identify patterns and hidden information in order to define targets.

The player data used was fabricated for the project due the lack of feasible sources and financial constraints. The format was adopted from the popular game (16). It would have been very useful if real time player data could be acquired in order to perform the analyses. A minimum of ten seasons worth of data would be required in order to perform detailed analyses and identify firm patterns and information.

6.2.4 Varied understanding of the game
These decisions may vary from personnel depending on their understanding of the technicalities of the sport. However, a brainstorming session with those in question would be an effective way of identifying the best possible manner to develop the design.

6.3 Future scope and application
This approach to ranking players in teams has the capability to be implemented across various team sports. The concept of calculating the scores and ranks is an effective way of determining the calibre of the players in a team with respect to the remaining players in the same team or group of players. The player pool can be extended to beyond a single team such that it can be used among a set of players irrespective of team or it can even be used amongst groups of players within the same team like determining ranks of players within the same player roles.

Further work in this field can enable the development of analytics applications or business intelligence systems that can be implemented across various sports in order to assist teams gain a competitive edge over their rivals. The method used can be customized according to the needs to the end user - manager or coach - such that based on the
collective team objectives; it can be decided whether to focus on a group of KPIs or a single or few KPIs in order to determine the value of each player for the team.

One of the key aims of this study is to shed light on the fact that analytics in sport is a growing industry and is one with much potential. The fact that assigning numbers is an effective method of determining the calibre and value of players and teams goes to show that using simple statistics can help derive various implications that can be extremely useful in providing teams with the competitive edge required to maintain a sustainable model and achieve their targets.

7. References


**Corresponding Author:**

**Rajeswari C**

**Email:** rajeswari.c@vit.ac.in