An Intelligent System to Predict Future Performance of Youth Football Players using Machine Learning

Wenjie Fu¹, Yu Sun², Fangyan Zhang³, Bo Guo⁴

¹American Heritage School of Boca Delray, Delray Beach, FL 33484
²California State Polytechnic University, Pomona, CA 91768
³Mississippi State University, Mississippi State, MS 39759
⁴Meizu Inc., Beijing, China 100871

Abstract—As the most popular sport in the world, soccer has become more and more competitive year by year. More teenagers join soccer programs and prepare to become professional players. Therefore, it is challenging for soccer clubs to select young players that will contribute to the clubs in the future. In this project, we address this problem by using mobile computing and machine learning. A mobile app has been developed to collect the soccer players’ data. Once a sufficient amount of data has been acquired, the system will be able to train the machine learning model to predict the players’ performance in the future based on data. Experiments show that our prediction can reach to 91% accuracy when the young players are under routine training.

Index Terms—Football Statistics, Machine Learning, Mobile Computing.

I. INTRODUCTION

The sports industry plays an important role in culture and economy because it contributes to over 1 percent of the GDP almost every year. The utilization of Artificial intelligence in sports may have been rare five years ago. [1] However, now AI and machine learning have entered plenty of sports industry applications such as data analysis and predictions.

Soccer, [2] which is called football in most countries outside of the United States, is played by about 265 million people all over the world, according to the Federation International de Football Association (FIFA), the international soccer governing organization. There is an influential soccer event referred as World Cup, which is held every four years. Also, many other soccer leagues and championships are popular year-round among both adults and children.

Because the English Football Association was the first to establish rules to the game, modern soccer was officially created in 1863 together with the establishment of the English Football Association, [3] although there's evidence indicating that activities like soccer existed in China about 2,000 years ago, and it is also said that ancient Greeks and Romans played similar games. In 1882, all the organizations related to soccer in England collaborated to create a united set of officially recognized rules, and the International Football Association Board (IFAB) was developed to oversee these rules. FIFA was founded in Paris in 1904, and in 1913, became a member of the IFAB. At that time, it only had seven members including Belgium, Denmark, France, Netherlands, Spain, Sweden and Switzerland. Today, FIFA has 208 members.

Soccer, or football, or “the beautiful game” is the world’s most popular sport. It is actually easy to prove this to a fan of North American sports by global TV audience numbers. In contrast with the fact that the 2014 Super Bowl had audiences of about 160 million viewers all over the world, the same year, the FIFA World Cup final had a global audience of around 1 billion. [4]

The collection of data from soccer games and the analysis of basic soccer statistics [5] establish the basis for soccer player profile-keeping, soccer team records generated over time, and the numerical evaluation of player performance. More information such as advanced statistics, analytics, metrics, and modeling of soccer players and their performance are following this initial capture of soccer information. Then coaches may choose the soccer data they wish to collect and the statistics they wish to calculate. Many of the items have “initialisms” which may be used as shorthand for the item. [6] The shorthand is noted in parenthesis after the item; some items have more than one. Most of the definitions may be found in the Master Glossary of American Soccer Terms. The recent sports analytics movement, however, didn’t start from the world’s most popular sport. Most people are likely to agree that it began with baseball and then spread to other North American sports such as hockey, basketball, and American football. Compared to these sports, the use of advanced statistics and technology such as machine learning in soccer is still in the early stages. [7]

It is very challenging to predict a player's potential performance in future once he or she joins the professional league. Coaches, clubs, and parents tend to predict performance of young soccer players to make better plans. However, because there is a large amount of data on many aspects for every player, it is great challenging for human brains to make accurate predictions. In many cases, players’ potentials are not completely explored due to improper estimation and insufficient advices.

To address this challenge, data collection, machine learning, data training, and data prediction implements automated prediction of the players’ performance. In this paper, we present a new approach for young soccer players to predict their performance. Since most of the young soccer players follow routine training schedules, their performance each year tent to follow a pattern that is based on styles and talents of themselves. Summarizing the pattern will allow us to predict their performance in the future. Thus, we have developed a mobile application that enables the young soccer players to input their personal data on many aspects. Once the application collects enough data, it trains a model using machine learning algorithms which could be used to predict the blood sugar level based on the given input. The mobile
application is supported by a backend running in the cloud that collects data, trains the model, and makes predictions. As the soccer players input more data with real test results, the accuracy of the tests will improve accordingly.

The rest of the paper is organized as follows: Section 2 details the challenges in this research project; one solution is presented in Section 3, followed by showing the experimental results in Section 4; we compare the related works in Section 5, and Section 6 offers the conclusion remarks and future work directions.

II. CHALLENGES

A few challenges exist in training the athlete statistics model and making the accurate predictions.

**Challenge 1. Deciding the exact set of factors to include in the training model is difficult.** The athlete’s statistics includes a number of data such as age, height, weight, goals per year, assists per year, and average scores every year. Getting the right data may require some statistical analysis and model selection techniques, and a lot of experiments. While too many data sets may lead to overly complicated models that have less precise predictions, only using a few data sets would lead to overlooking potential lurking variables and confounding variables. Therefore, first listing all data may be helpful, as we can eliminate the futile factors in the process of training the machine.

**Challenge 2. The statistics and analysis are very specific to individuals, which means that the prediction has to be personalized based on each athlete’s situation.** The model and the training approach might be adapted to the different individual athlete. Through collecting data, we can categorize athletes according to their positions including forward, midfield, and defender so as to have higher accuracy when predicting the outcome. We plan to provide more options for users to select as they input their data, subsequently they can see their predictions using people in similar circumstances as reference. As the system continues to take in more data, statistics and analysis of different positions can be categorized to provide higher confidence in the predicted outcome and lower the range of the predictions.

**Challenge 3. The model training needs to be updated frequently with the new data.** As the athletes grow up, their personal statistics change from time to time. The system needs to provide a way to easily collect new data and keeps improving the training model by itself using a technique known as reinforcement learning [8]. There may be significant outliers as everyone has vastly different conditions; therefore, the system’s ability to react immediately to new data and decipher the significance is crucial to its development. We intend to give users an easy and efficient way of utilizing the app for more people to be willing to spend time in putting their data.

III. SOLUTION

A. Overview of the Solution

An overview of the system is presented in Figure 1. Athletes use the mobile client to input their data such as age, height, weight, goals per year, assists per year, and average scores every year. These data build the training dataset, so each time a new record is inputted by an athlete, it triggers the training process and build the updated prediction model. Once the dataset grows over a certain threshold, athletes can start to input the information and get the predicted performance.

B. Machine Learning Model and Feature Selection

Table 1 shows the training features selected in this machine learning process, which are described as follows:

- **Age**: the age of the athlete
- **Position**: the athlete’s position such as forward, midfield, and defenders.
- **Height**: the height of the athlete in centimeters
- **Weight**: the weight of the athlete in kilograms
- **Goal**: goals scored by the athlete every year
- **Assist**: assists of the athlete every year
- **Dribble**: dribbles through opponents per game
- **Pass Accuracy**: the percentage of successful passes out of total passes per game
- **Intercept**: intercepts towards opponents’ passes per game
- **Score**: the evaluation of the performance every year
- **Score Four Years Later**: the evaluation of the performance four years later

C. Training and Prediction

We used the standard machine learning library scikit-learn [9] to training and prediction the model. Scikit-learn provides different machine learning procedures such as classification, regression, cluster dimensionality reduction. It is built on strong mathematical packages such as NumPy and SciPy. To get the most accurate approach, we have used 3 different training algorithms to test the results, followed by some comparisons. The approaches we have chosen are: Linear regression, Polynomial regression, and Gaussian Process
Regression. Linear regression attempts to minimize the sum square of residuals, which is a simple approach for single modal problems [10]. Polynomial regression is a form of linear regression with polynomial terms of features, offering more complexity [11]. The Gaussian Process Regression interpolates values between data points with Gaussian process with prior covariance information [12].

D. Mobile Application
The frontend mobile application is developed using MIT App Inventor [13]. MIT App Inventor is a web-based integrated development environment for developing Android mobile applications. MIT App Inventor provides an interactive and real-time debugging and testing environment that accelerates building simple applications.

As shown in Figure 2, the app has one dedicated screen that shows a beautiful stadium and enables users to go to the next screen.

Users can start to make predictions of the performance four years later by triggering the other screen of the app as shown in Figure 3. In this screen, users need to input all the data except the performance four years later, which will be predicted using the trained model in the cloud.

IV. EXPERIMENTS
To evaluate the accuracy of our approach, we have collected 50 copies real dataset from 10 professional soccer players. To get the compare the approaches, we conducted experiments to verify two aspects: the accuracy of using different machine learning models, and the different selection of the training features.

A. Comparison of Different Machine Learning Algorithms
Figure 4 illustrates the accuracy of the prediction by using three different machine learning algorithms. As can be seen, the polynomial regression has the best result, due to its capability of handling the non-linear factors in the training process. By comparison, the linear repression has the lowest accuracy, which proves that performance prediction in this case is not a linear situation.

B. Comparison of Different Sets of Training Features
We also investigated how the selection of the different training set will affect the accuracy. It turns out that the selection does have an impact on the accuracy. For instance, the Feature Set 1 and Feature Set 2 both have more features,
but the accuracy is lower, which indicates that both sets have the problem of over training. The Feature 3 is the set we showed in this project as shown in Table 1.

![Figure 5. The Accuracy of Prediction using Different Feature Set](image)

<table>
<thead>
<tr>
<th>Age</th>
<th>Position</th>
<th>Height(cm)</th>
<th>Weight(kg)</th>
<th>Goal</th>
<th>Assist</th>
<th>Dribble</th>
<th>Pass Accuracy</th>
<th>Interception</th>
<th>Score</th>
<th>Score 4 Years Later</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>1</td>
<td>170</td>
<td>72</td>
<td>34</td>
<td>10</td>
<td>4.1</td>
<td>85.7</td>
<td>1.2</td>
<td>8.67</td>
<td>8.34</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>170</td>
<td>72</td>
<td>31</td>
<td>18</td>
<td>5.6</td>
<td>86.2</td>
<td>0.6</td>
<td>8.76</td>
<td>8.84</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>170</td>
<td>72</td>
<td>50</td>
<td>16</td>
<td>4.8</td>
<td>84.2</td>
<td>0.6</td>
<td>8.88</td>
<td>8.46</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>183</td>
<td>76</td>
<td>9</td>
<td>9</td>
<td>1.2</td>
<td>82.2</td>
<td>1</td>
<td>7.59</td>
<td>7.34</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>183</td>
<td>76</td>
<td>1</td>
<td>4</td>
<td>0.6</td>
<td>86.3</td>
<td>0.4</td>
<td>6.82</td>
<td>7.42</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>183</td>
<td>76</td>
<td>4</td>
<td>9</td>
<td>1.4</td>
<td>87.6</td>
<td>1.4</td>
<td>7.52</td>
<td>7.33</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>180</td>
<td>76</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>83.6</td>
<td>1</td>
<td>6.54</td>
<td>7.55</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>180</td>
<td>76</td>
<td>0</td>
<td>0</td>
<td>1.1</td>
<td>82.9</td>
<td>1.3</td>
<td>6.58</td>
<td>7.24</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>180</td>
<td>76</td>
<td>2</td>
<td>2</td>
<td>1.4</td>
<td>86.8</td>
<td>1.3</td>
<td>6.97</td>
<td>6.94</td>
</tr>
</tbody>
</table>

Table 1. A Sample of the Training Dataset

E. Rampinini [14] aims to examine the validity of a statistics approach to predict soccer players’ physical performance using data from selected field test records such as mass, height, total running distances, and sprinting speed. In contrast to this approach, my work employs both statistics and machine learning to predict performance in the future based on mathematical models. M. Buchheit [15] presents analysis towards the performance of youth soccer players based on their positions, ages, and physical conditions by statistics from a large amount of games. However, this approach also does not illustrate the utilization of machine learning on the statistics. Bradley [16] aims to predict soccer player performance at different playing positions by using computational systems to quantify of acceleration and maximal sprinting speed profiles of professional soccer players. Therefore, based on the predictions acquired from models, coaches can develop suitable training for specific soccer players.

V. RELATED WORK

In this project, we proposed an intelligent approach to address the problem of performance predictions for young soccer players using mobile computing and machine learning. A mobile app has been developed to collect the players’ statistics. Once the sufficient amount of data has been collected, the system is able to train the machine learning model and predict the athletes’ performance based on the statistics. Experiments show that our prediction can reach to 89% accuracy when the athletes are under regular training and do not experience major events such as serious injuries.

As for the future work, we will investigate other machine learning algorithms to keep improving the accuracy. We also would like to explore the possibility of applying deep learning in this problem domain.

In addition, one limitation related with the app is that it does not suggest the sufficient threshold of the training dataset. One feature we plan to add in the next version of the app is to automatically evaluate the accuracy of the training process and prompts the sufficiency of the dataset to users automatically.

REFERENCES