SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)

EPRA International Journal of Research and Development (IJRD)

Volume: 5 | Issue: 10 | October 2020 - Peer Reviewed Journal

MACHINE LEARNING BASED PREDICTIVE MODEL FOR SOCCER

Ajayi Adebowale

Ajayi Oluwabukola

ABSTRACT

Advancements in data crunching technologies have hugely impacted sports science especially soccer with most football teams having a data science department aimed at providing data driven solutions for various issues not limited to : Match strategy, tactics, and analysis, Identifying players' playing styles, Player acquisition, player valuation, and team spending. Identifying the right features to be incorporated in predictive models for soccer matches however remains a viable research area.

This study presents a machine learning based predictive model for soccer outcomes using historical data. Data preprocessing algorithms are presented as well as a framework for computing expected goals and incorporating individual player performance in predicting soccer outcomes.

KEYWORDS: *Machine Learning, Predictive model, preprocessing, sports science, data driven.*

MACHINE LEARNING BASED PREDICTIVE MODEL FOR SOCCER

Soccer is one of the most widely followed sports with hundreds of millions of soccer fans all over the world. The financial implications of the strong soccer fan base is clearly evidenced in the huge sums of money soccer clubs pay for recruiting players. Against the backdrop of huge financial influx into soccer, the need for predictive systems have never been higher in the soccer industry. Football clubs desire forecasts on player performance to enable them make informed decisions as to retention or procurement of player services. Soccer betting companies are also in need of predictive models to set appropriate betting odds based on past observations in order to stay in business.

Soccer fans who stake on the matches also demand informed suggestions to decide what odds to bet on. Everyone wants a piece of the action and informed decision making on soccer matters have not been more pertinent. Soccer pundits give their opinions as to the direction of soccer games or the expected performance of soccer teams through the course of a season. However, their predictions are more of a hit and miss as soccer proves to be quite difficult to predict using intuition and personal football knowledge alone. This limitation has led to increased demand for statistical models of predicting soccer outcomes. Advancements in digital technology has led to the collation of vast amounts of vital soccer statistics requiring advanced techniques for processing and usage. This study presents a novel soccer prediction framework using an ensemble of machine learning algorithms and historical soccer data.

METHODOLOGY

The quasi experimental research design was adopted for this research project as several experiments were carried out using machine learning algorithms in an attempt to get a baseline model for optimization.

Dataset

We have obtained a dataset from the Kaggle Data Science website called the 'Kaggle European Soccer Database'. This database has been made publicly available and regroups data from three different sources, which have been scraped and collected in a usable database:

- Match scores, lineups, events: http://football-data.mx-api.enetscores.com/
- Betting odds: http://www.football-data.co.uk/

- Players and team attributes from EA Sports FIFA games: http://sofifa.com/ It includes the following:
- Data from more than 25,000 men's professional football games
- More than 10,000 players
- From the main football championships of 11 European countries
- From 2008 to 2016
- Betting odds from various popular bookmakers
- Team lineups and formations
- Detailed match events (goals, possession, corners, crosses, fouls, shots, etc.) with additional information to extract such as event location on the pitch (with coordinates) and event time during the match.

We used only 5 leagues over two seasons as they possess geographical data for match events that we needed to build our expected goals models:

English Premier League French Ligue 1 German Bundesliga ,Spanish Liga and Italian Serie A

We only used data from the 2016/2017 as well as 2017/2018 seasons as they are the most recent seasons available in the database and the only ones containing the data that we need.

This gives us usable dataset of:

- 3,800 matches from the top 5 European leagues
- 88,340 shots to analyse
- More than 100 different teams

Data pre-processing

An important step before building our model is to analyse and pre-process the data to make sure that it is in a usable format for us to use when training and testing different models.

Three pre-processing steps were taken in order to achieve this:

- Part of the data that we needed, namely all match events such as goals, possession, corners, etc. was originally in XML format in the database. We therefore built a script in R to extract this data and store it in new tables, linked to the 'Matches' table thanks to a foreign key mapping to the match ID. An extract of the XML for a goal in one match is presented below:

```
<goal>
<value>
<comment>n</comment>
<stats>
<goals>1</goals>
<shoton>1</shoton>
</stats>
<event_incident_typefk>406</event_incident_typefk>
<coordinates>
<value>18</value>
<value>67</value>
</coordinates>
<elapsed>35</elapsed>
<player2>35345</player2>
<subtype>header</subtype>
<player1>26777</player1>
<sortorder>1</sortorder>
<team>9826</team>
<id>3647567</id>
<n>200</n>
<type>goal</type>
<goal_type>n</goal_type>
</value>
</goal>
```

- Some data elements were set to NULL, which led to us deleting some unusable rows and in other cases to us imputing values to be able to use the maximum possible amount of data. For instance, the possession value was missing for some games, so we entered a balanced value of 50% possession for each team in this case.
- Finally, having extracted the geographical coordinates for each shot in the dataset, we generated distance and angle to goal values which we added to our goals and shots database tables. The following formula was used to generate the distance to goal of a shot, where the coordinates of the goal are (lat=0, lon=23):

$$D(lat, lon) = \sqrt{lat^2 + (lon - 23)^2}$$

The following formula was used to generate the angle of a shot:

$$A(lat, lon) = tan^{-1}(\frac{|lon-23|}{lat})$$

Data features

A simplified diagram of the database structure and features is presented in Fig.1.

We will now present the different tables and features that we have in our database
and that we can use in our models:

• Matches table
– ID
– League ID
– Season
– Date
– Home team ID
– Away team ID

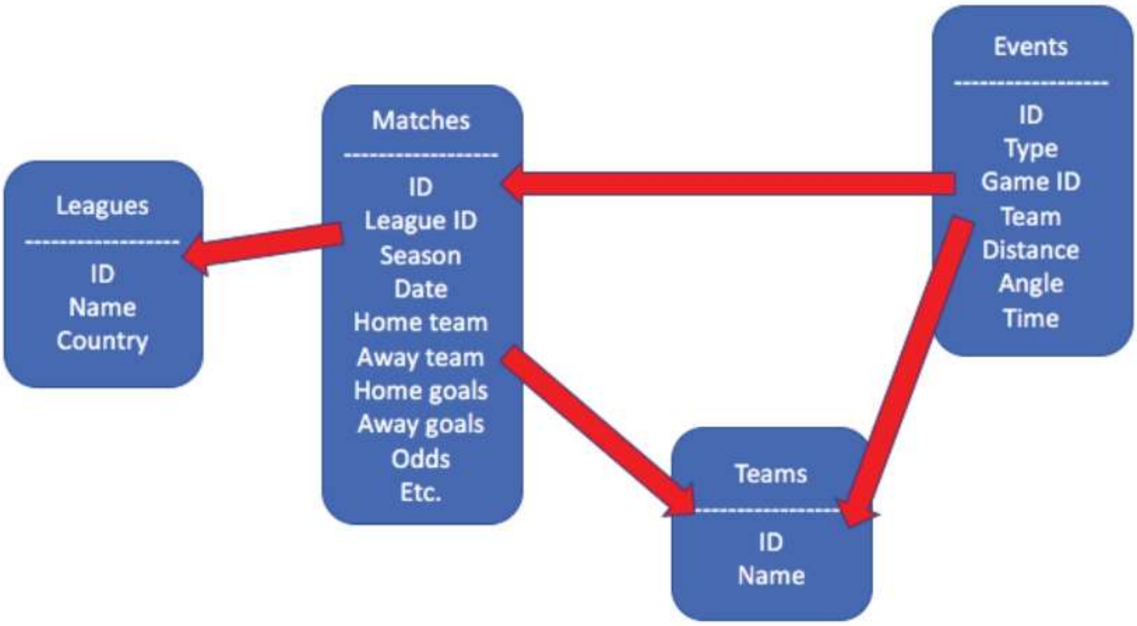

Figure 1: Structure of the Database

– Home team goals scored
– Away team goals scored
– Home team possession
– Away team possession
– Home win odds
– Draw odds
– Away win odds
• Events tables:
Here is a list of the different match events tables that we have extracted:
– Goals
– Shots on target
– Shots off target
– Corners
– Crosses
– Cards
– Fouls
For each of these match event tables, we have the following features:
– ID
– Type
– Subtype
– Game ID
– Team ID
– Player ID
– Distance to goal (only for goals and shots)
– Angle to goal (only for goals and shots)
– Time elapsed

SYSTEM DESIGN AND ANALYSIS

This section presents the general design of our model and the choices we have made. Our model is a mixture of multiple regression and classification algorithms used to generate different metrics that are finally used as inputs for our classification model (for match outcomes) and regression model (for match scores).

Model components

In this section, we will introduce the different components of our model and explain their role. A diagram of our different components and how they are linked is presented in Fig 2.

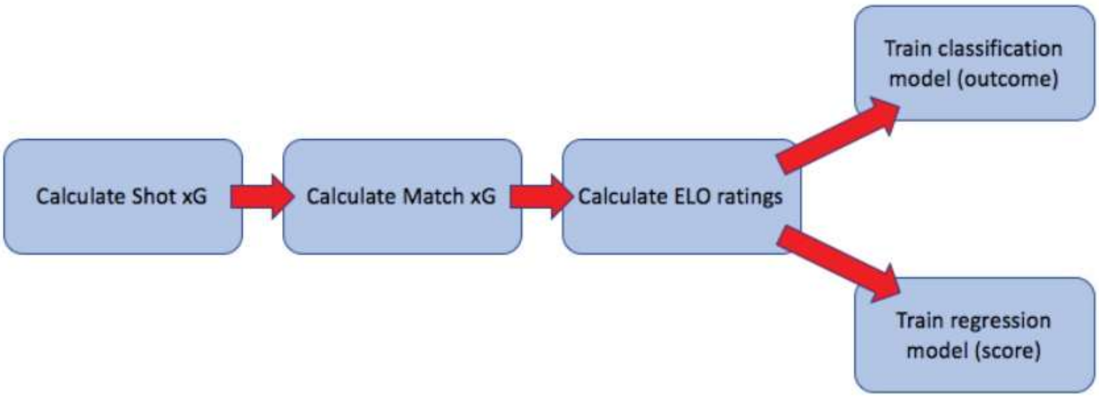
Calculate Shot xG
Calculate Match xG
Calculate ELO ratings
Train classification model (outcome)
Train regression model (score)

Figure 2 : Diagram of model components

We have five main model components which we will present one by one:

Shot xG generation

This component's objective is to generate an expected goal value for each shot representing the probability that the shot results in a goal, with some adjustments to reflect specific match situations.

Match xG generation

This component's objective is to generate a shot-based expected goals value for each match by looking at the expected goals values for each shot during that match. In addition to this, a non-shot-based expected goals value is generated using match information other than shots.

ELO calculation

This component's objective is to generate offensive and defensive team ELO ratings after each match using expected goals values and the actual performance. ELO ratings are recalculated after each match and the team ratings are stored for use in our predictive classification and regression models.

Classification model training

This component's objective is to train and test a classification model capable of taking two teams' ELO ratings and generating a prediction for a match between these two teams between a home team win, a draw and an away team win.

Regression model training

This component's objective is to train and test a regression model capable of taking two teams' ELO ratings and generating a prediction for the expected number of goals each team will score. These values are then used to generate a prediction for the match outcome.

CONCLUSION

Our main objective of building an expected goals model by exploring different Machine Learning techniques has been accomplished. Indeed, we used modern Machine Learning algorithms such as Neural Networks, Random Forest and Support Vector Machines techniques to generate match outcome and match score predictions. We managed to find and improve a database containing enough information to generate expected goals metrics, through both shots and other in-game statistics, and ELO team ratings.

REFERENCES

1. *Anthony Costa Constantinou, Norman Elliott Fenton. 2013 Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. Journal of Quantitative Analysis in Sports.*
2. *Brian Macdonald 2012. An Expected Goals Model for Evaluating NHL Teams and Players. MIT Sloan Sports Analytics Conference.*
3. *Harm Eggels, Ruud van Elk, Mykola Pechenizkiy. 2016 Explaining soccer match outcomes with goal scoring opportunities predictive analytics. Eindhoven University of Technology.*
4. *Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, Iain Matthews 2015. Quality vs Quantity: Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data. MIT Sloan Sports Analytics Conference.*
5. *FiveThirtyEight football predictions [https://projects.fivethirtyeight.com/soccer-predictions/11].*
6. *Kaggle European Soccer Database [https://www.kaggle.com/hugomathien/soccer12] . pages 23 [43] Pandas Python library [https://pandas.pydata.org/13].*
7. *Scikit-Learn Python library [http://scikit-learn.org/stable/index.html14].*
8. *Luigi: Python module to build batch jobs pipeline [https://github.com/spotify/luigi15].*

9. *openMOLE parameter optimisation method diagram [https://next.openmole. org/Calibration.html17].*

10. *WhoScored.com football statistics website [https://www.whoscored.com/18].*

11. *Opta Sport data provider [http://www.optasports.com/19]*

12. *[Augur: Decentralized prediction markets [http://www.augur.net/20]*