Analysis of Football Match Data to Predict Goals

Kushagra Dwivedi

Table of Contents

Analysis of Football Match Data to Predict Goals

The purpose of this Research study is to explore how Machine Learning and Data Analytics can be applied to forecast football goals and player metrics. This study focuses on the development and assessment of predictive models that can give predictive insights and help generate new analytics in sports. The study uses the dataset of 900,000 soccer event over 9,074 European games and develop a model that can predict the probability of the goal being scored irrespective to the player. This predictive modelling can be used to create new analytics to access the performance of the player and better team creation.

Aim and Objective

- The study aims to investigate the role of Machine Learning techniques in predictive forecasting of soccer goals. With the increased interest in sports analytics and the availability of massive amounts of match data, this study focuses on the development and evaluation of prediction models that can provide insights the football matches.
- The goal of the study is to analyze the data that is available and use it to create an AI model that can give the likelihood of goal being scored based on number of factors like the position of the player, location on the field, situation of the match to output a fuzzy number that represents the likelihood of a goal being scored.
- The study's secondary goal is to use the fuzzy results of the predictive models that were developed to establish a standard basis for players' performance that is only dependent on their physical circumstances and situations. This standard scoreability of goals can then be

used as a benchmark for analyzing player performance, removing luck and other abstract concepts from the equation, and determining the players' skill levels.

- The study can be separated into milestone objectives-
  - Exploration and Understanding of the dataset.
  - Cleaning the dataset, extracting, and generating features according to study's needs.
  - Selection, training, and evaluation of predictive ML models.
  - Develop the benchmark scorability metric and design features based on it to analyze a player's performance.

Literature Review

Sports industry is a multibillion-dollar industry with audience around the world. The football industry only is worth 3.2 billion US Dollars with football clubs spending 100's of million dollars per year on the team. And in the last few years it has experienced fastest growth among the gambling markets. The unpredictability of football and high investment density have given rise to development of prediction models to support gambling, analytical models to monitor player performance and help decision making for the clubs scouting staff and coaches.

Germany's national team successfully applied Machine Learning to support their scouting team in analyzing opposing teams and monitor their players and help their coach in decision making, in the 2014 World Cup. The 2021 Journal of Sports Sciences article "Machine learning for predicting football results: a systematic review" by Fabio Calefato et al. is a systematic review that examined the findings of 50 studies on machine learning models for football match prediction. Numerous machine learning algorithms are covered in the review, such as random

forests, deep learning models, and Bayesian networks. And the different features that these models make use of, like player information, match statistics, and meteorological conditions.

Image Retrieval and Feature Extraction Based on Content: A Comprehensive Review presented a comprehensive literature review on various CBIR and image representation techniques. This study provides an overview of various techniques used in various research models over the last 12-15 years. Following this review, it is summarized that image features are represented by using low-level visual features such as color, texture, spatial layout, and shape.

Football-Data.co.uk is website that provides predictions over various major and minor leagues using Machine Learning algorithms like Random Forests, they utilize vast amounts of historical data that includes factors like recent results, player performance, meteorological conditions to train their machine learning algorithms and then use these models to get accurate odd for betting on the games to increase their profits.

Three-dimensional face recognition in the presence of facial expressions: an annotated deformable model approach addresses the main challenges of a 3D field deployable face recognition system. It has developed a fully automated system that achieves complete pose-invariance by registering 3D facial scans with a 3D facial model via a composite alignment algorithm. This system measures the difference between the facial scan and the model in a way that achieves a high degree of expression invariance and, consequently, high accuracy. It does this by fitting the 3D facial model to the aligned 3D facial scans using a deformable model framework.

Methodology

Method Overview

The Fig. 1 gives the methodical overview of the study and all the steps involved.

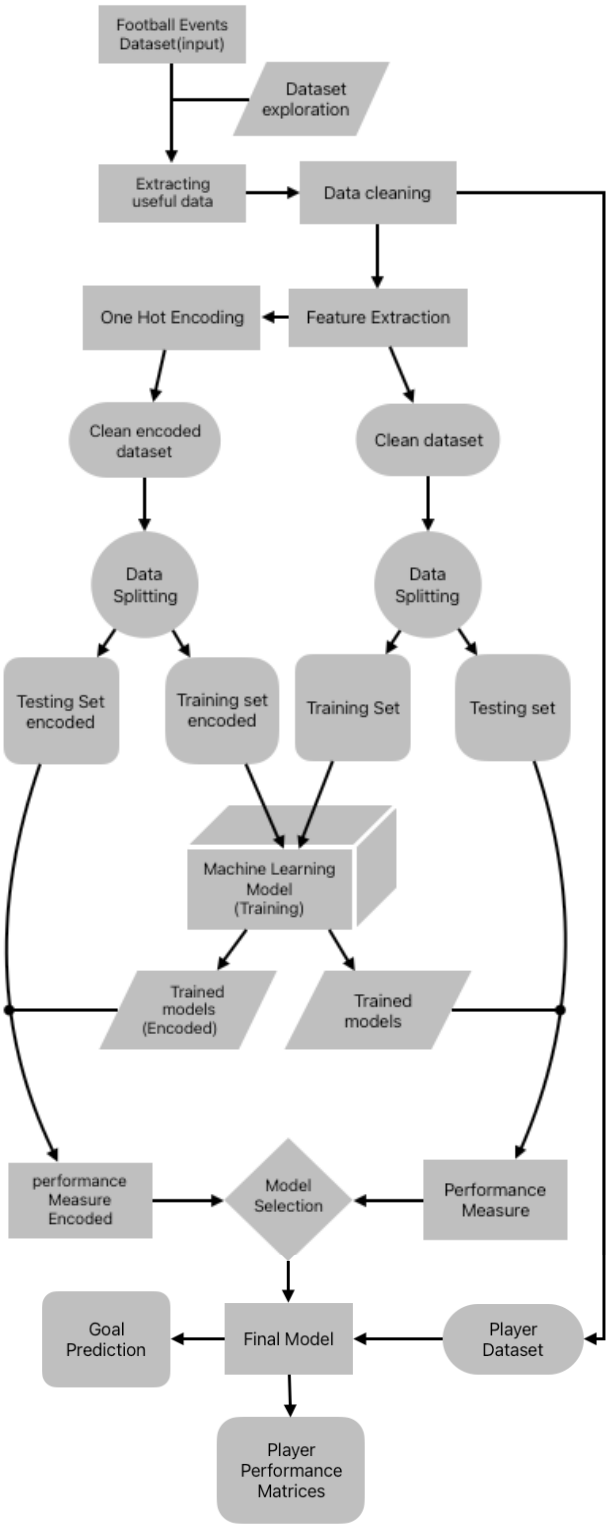

Fig. 1 Methodology Block Diagram

- **Dataset Selection and Exploration**

The Football Events dataset from Kaggle was used for the study. From the 2011–2012 season to the 2016–2017 season as of January 25, 2017, the dataset offers a detailed perspective of 9,074 games, totaling 941,009 events, from the top 5 European football (soccer) leagues: England, Spain, Germany, Italy, and France. The dataset is a result of web scraping and data extraction and contains 3 files :-

- **events.csv** contains event data about each game. Text commentary was scraped from: bbc.com, espn.com and onefootball.com
- **ginf.csv -** contains metadata and market odds about each game. odds were collected from oddsportal.com
- **dictionary.txt** contains a dictionary with the textual description of each categorical variable coded with integers

Events.csv contains information about an array of events that can happen during the match, The events are encoded in the following format given below

```
0     Announcement
1     Attempt
2     Corner
3     Foul
4     Yellow card
5     Second yellow card
6     Red card
7     Substitution
8     Free kick won
9     Offside
10    Hand ball
11    Penalty conceded
```

From this data the events containing Attempt (1) event can be used for training a Machine Learning Model to predict likelihood of a goal being scored for an attempt. The file has multiple columns - 'id_odsp', 'id_event', 'time', 'text', 'event_type', 'event_type2', 'side', 'is_goal', 'player', 'opponent', 'player2', 'player_in', 'player_out', 'shot_place', 'shot_outcome', 'location', 'bodypart', 'assist_method', 'situation', 'fast_break', 'sort_order', 'event_team'

- **Data Cleaning and Feature Selection**

The extracted data from event.csv was cleaned, as there were multiple unknown values in the location marker, which is one of the important features. There were also null values in

EVENT	No. of Null Values
event_type2	59826
player	1
player2	59887
player_in	227685
player_out	227685
shot_place	1008

Table.1- Null values

Most of these features are not used as a feature in the study so, only location marker had to be dealt with for which the unknown values were dropped, the dropped values included 1438 attempts that were goal and 12 attempts that were misses.

Total 7 features were extracted from the data to train the model. The features are 'time', 'side', 'bodypart', 'location', 'situation', 'assist_method', 'fast_break'. Another dataset was also created by One Hot encoding the categorical features. Finally the 2 feature sets were test train split in 0.8 train ratio with stratification.

- **Selection, training, and evaluation of predictive ML models**

In the study an ensemble of machine learning models was trained and tested to find the best performing models. The models were separately trained on both sets of training data encoded and non-encoded. The models that were used include Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier, GaussianNB, Linear Discriminant Analysis, Quadratic Discriminant Analysis, KNeighbors Classifier, LinearSVC along with custom deep neural network that was separate for both training sets.

After training of ML models, they were evaluated over the test set with test accuracy as the metric for evaluation, as it is accuracy we are looking to improve. Most of the models

performed well, Gradient Boosting Classifier was the best performing model and AdaBoost Classifier followed closely behind. The result can be seen in Fig. 2-

```
==============================
Neural Net non encoded
****Results****
1424/1424 [==============================
Accuracy: 90.9239%
==============================
DecisionTreeClassifier
****Results****
Accuracy: 89.8742%
==============================
RandomForestClassifier
****Results****
Accuracy: 89.8917%
==============================
AdaBoostClassifier
****Results****
Accuracy: 90.9788%
==============================
GradientBoostingClassifier
****Results****
Accuracy: 90.9854%
==============================
GaussianNB
****Results****
Accuracy: 89.2637%
==============================
LinearDiscriminantAnalysis
****Results****
Accuracy: 89.4152%
==============================
QuadraticDiscriminantAnalysis
****Results****
Accuracy: 88.9123%
==============================
KNeighborsClassifier
****Results****
Accuracy: 89.8237%
==============================
LinearSVC
****Results****
Accuracy: 89.8961%
==============================
```

```
==============================
Neural Net encoded
****Results****
1424/1424 [==============================
Accuracy: 90.9348%
==============================
DecisionTreeClassifier
****Results****
Accuracy: 89.8720%
==============================
RandomForestClassifier
****Results****
Accuracy: 89.7248%
==============================
AdaBoostClassifier
****Results****
Accuracy: 90.9195%
==============================
GradientBoostingClassifier
****Results****
Accuracy: 90.9436%
==============================
GaussianNB
****Results****
Accuracy: 62.6743%
==============================
LinearDiscriminantAnalysis
****Results****
Accuracy: 90.5044%
==============================
QuadraticDiscriminantAnalysis
****Results****
Accuracy: 14.4652%
==============================
KNeighborsClassifier
****Results****
Accuracy: 89.8742%
==============================
LinearSVC
****Results****
Accuracy: 90.8338%
==============================
```

Fig. 2 – Accuracy scores of non-encoded set (left) and encoded set (right)

Thus, for the study Gradient Boosting Classifier is selected as the best model, with 90.9853% accuracy score that can go over 91% in some cases. It is also used for further derivation of expected goal metric and create other new metrics for player analysis.

- **Player performance Metric Inference**

To create player performance metrics the data after cleaning just before feature extraction is used and a column of expected goal likeliness is added for each instance in the dataset. Expected goal likeness is the likelihood of goal being scored once the shot is taken depending on the conditions at the time of goal. The selected model developed in the study is used to find this likelihood.

After adding the expected goal column, a list of all the players in the dataset is created. For each player following metrics are derived from the data-

- nb_shots – Number of shots taken by the player
- expected_goals – Number of goals predicted by the model that could be scored
- goals_scored – Number of goals scored by the player
- xg_diff – Difference of goals scored and goals predicted to be scored
- nb_headers – Number of headers attempted by the player
- expected_head_goals – Number of headers that scored according to prediction
- head_goals_scored – Number of headers scored by the player
- head_xg_dif – Difference between headers that scored and ones predicted to score
- xg_per_shot – The average predicted chance of goal for a player
- pct_goals_counter – Percentage of counter goals by the player

The table now contains every player from 9,074 matches with parameters adjustable according to the coach's requirement. Based on the circumstances they faced, the xG difference reveals which players perform better than average. The xG difference in the air provides us with the same information, but only for headers, allowing us to identify players who outperform in the air. The xG per shot indicator displays which players attempt shots with the highest likelihood of being

made, as well as which players frequently miss opportunities and take shots that are hard to make. The percentage of goals scored during counterattacks identifies players who are accustomed to counterattacking tactics.

The Fig.3 shows the first few rows of the dataset inferred

	player	nb_shots	expected_goals	goals_scored	xg_dif	nb_headers	expected_head_goals	head_goals_scored	head_xg_dif	xg_per_shot	pct_goals_counter
0	mladen petric	62	8.128841	5	-3.128841	8	1.079347	0	-1.079347	0.131110	20.000000
1	shinji kagawa	151	17.205218	25	7.794782	13	1.317911	2	0.682089	0.113942	20.000000
2	kevin grosskreutz	110	10.789076	9	-1.789076	10	0.853300	0	-0.853300	0.098083	11.111111
3	mats hummels	120	17.970696	9	-8.970696	80	11.126770	5	-6.126770	0.149756	0.000000
4	tomas rincon	98	5.358366	3	-2.358366	2	0.121599	0	-0.121599	0.054677	0.000000

Fig. 3 – Player performance Metric

Conclusion

The study explored the use of Machine Learning models for analyzing the situation and predicting likelihood of a goal by analyzing the data. Several key objectives were covered during the study, we worked with the data available and developed and evaluated multiple ML models that took in features, such as the player's position, field location, and the match situation to output the probability of goal being scored.

Another objective was to use the probability output of the models developed to create various player performance metrics that could be used to accurately represent the performance of a player.

While the models developed show good amount of accuracy and the metrics inferred are fairely accurate there is a lot of scope for further development, currently only 7 basic features were used, in the future with increasing availability of data multiple different features can be included to further improve the prediction. Also with the development in the field of NLP the model can be synced with the commentry that can be analyzed and processed to act as a feature. For Player some important metrics like height, personality, market value, etc. are missing That are an important factor in player analytics that could be added in the future.

References

1. Rodrigues, F. and Pinto, Â. (n.d.). Prediction of football match results with Machine Learning. *ScienceDirect*, Procedia Computer Science 204 (2022) 463–470.
2. football-data.co.uk. (n.d.). *Football Betting | Football Results | Free Bets | Betting Odds*. [online] Available at: https://football-data.co.uk.
3. Patel, S., Kate, A., Wavare, K., Gujar, M. and Bachav, G. (2023). Predicting Football Match Results using Machine Learning. *IJCRT*, 11(4).
4. Rahman, M.A. A deep learning framework for football match prediction. *SN Appl. Sci.* **2**, 165 (2020). https://doi.org/10.1007/s42452-019-1821-5
5. Grunz A, Memmert D, Perl J (2012) Tactical pattern recognition in soccer games by means of special self-organizing maps. Hum Move Sci 31(2):334–343
6. Gomes, J., Portela, P. and Santos, M. F., (2015). Decision Support System for predicting Football Game result. 348-353
7. Latif A, Rasheed A, Sajid U et al (2019) Content-based image retrieval and feature extraction: a comprehensive review. Math Probl Eng 2019:9658350
8. Kakadiaris IA, Passalis G, Toderici G, Murtuza MN, Lu Y, Karampatziakis N, Theoharis T (2007) Three-dimensional face recognition in the presence of facial expressions: an annotated deformable model approach. IEEE Trans Pattern Anal Mach Intell 29(4):640–649