Features impact on the prediction of football matches results

Genís Ruiz , Universitat Autònoma de Barcelona, December 2023

Abstract

This project investigates the machine learning prediction of football match outcomes in the Spanish league. After optimizing the dataset through feature engineering, a Random Forest Classifier is trained. The study analyzes model accuracy across seasons, rounds, and teams, emphasizing the importance of features like team points and previous standings. The conclusion suggests that while certain features impact predictions, accurately forecasting football match outcomes remains challenging, especially for teams with a high draw rate. The project is available on GitHub for detailed code and visualizations.

Key Concepts

Machine learning, Feature Engineering, Encoding, Normalizing, SKlearn, Logistic Regression, Accuracy, Random Forest Classifier, Feature Importance and football match streak.

1 Introduction

The introduction acknowledges the inherent unpredictability of forecasting football match results. Despite this unpredictability, the paper aims to identify key features or events that may influence outcomes. The focus is on analyzing these features to predict whether the result will be a defeat, draw, or win for the local team in La Liga, the 1st Spanish Football division.

2 Original Dataset

The original dataset was downloaded from:
https://www.kaggle.com/datasets/ricardomoya/football-matches-of-spanish-league
The dataset contains all match results from 1970-71 to 2017-18 season. The columns of the DataFrame are:

- *season* : string value from 1970-71 to 2017-18
- division : 1st or 2nd Spanish League
- *round* : The round of the game played. Integer number from 1 to 38 or 42
- *localTeam* & *visitorTeam* : string name of the playing teams
- *localGoals* & *visitorGoals* : number of goals scored of each team
- *date* : DD/MM/YYYY of the game
- *timestamp* : *date* converted to a numerical value

3 Dataset Optimization

3.1 Data Cleaning

The initial procedure involved the systematic exclusion of undesired rows from the dataset, with a specific focus on La Liga predictions, leading to the removal of second-division data. Standardization of comparisons across seasons was achieved by omitting those seasons deviating from the typical 38 rounds.
The dataset, spanning the years 1987 to 2018 (excluding 1986-87, 1995-96, and 1996-97), exhibited no missing values (NaN), obviating the need for additional row removal.

3.2 Feature Engineering

From the few original columns lots of features have been created.

3.2.1 Team Statistics

New features have been developed for either the *localTeam* and *visitorTeam.*

- *result* : target variable = 0,1,2 if *localTeam* gets a defeat, draw or victory in that order.
- *T1PliS*, *T2PliS*, *T1PviS* & *T2PviS* : T1 and T2 correspond to *localTeam* and *visitorTeam* respectively. It describes the number of **P**oints as **l**ocal **i**n **S**eason and **P**oints as **v**isitor **i**n **S**eason each team has before that match. These points are normalized (float between 0.0 and 1.0).
- *GoalsliS*, *GoalsviS*: Number of Goals the local and visitor teams have scored in that league before the match.
- *Rk_loc* & *Rk_vis*: Current ranking position in that season before the match.
- *local_3streak*, *visitor_3streak*, *local_5streak* & *visitor_5streak*: Number of league points obtained in the last 3 and 5 matches of the *localTeam* and *visitorTeam.*
- *last_1ra_loc* & *last_1ra_vis*: Boolean value (0 or 1) which is 1 if the local/*visitorTeam* took part the previous season in La Liga.
- *last_top6_loc* & *last_top6_vis*: Boolean value (0 or 1) which is 1 if the local/*visitorTeam* ended up in the Top6 positions in the classification the previous season.

3.2.2 Temporal features

Originally, the dataset contained *date* and *timestamp*. From that point, some implementations have been carried out.

- *month*: Integer value from 1 to 12 selected from *date*. It is 1 in August, at the beginning of the league, and so on.
- *day*: Integer value from 1 to 7 extracted from *date*.

3.3 Features Removal

- *timestamp* & *date*: *timestamp* did not have hour information. Both features are no longer necessary.
- *localGoals* & *visitorGoals*: Two columns unknown in the prediction.
- *localTeam* & *visitorTeam*: No need to be included in the model as its stats are featured.

4 Target analysis

The match result is the variable to predict. There are three options; *localTeam* defeat (0), draw (1) or victory (2).

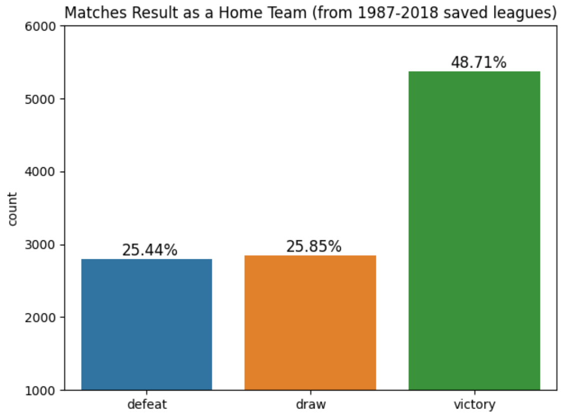


Figure 1:

The number of victories is almost the double of defeats and draws. That result means the main result for a *localTeam* in a football match is a victory.

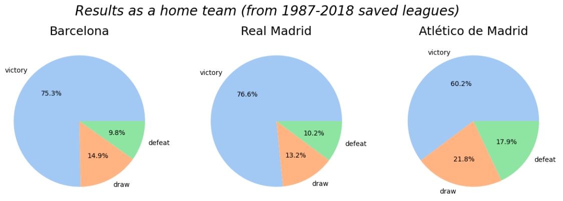


Figure 2: Example of Top3 local teams results

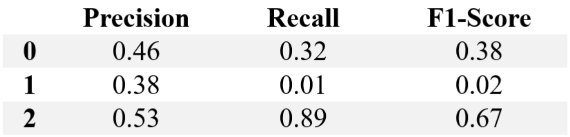
5 Encoding & normalizing

The dataset undergoes a stratified split into training(X_train,y_train) and testing (X_test,y_test) sets, ensuring balanced representation across target classes. Variables such as *season* and *round* are encoded using OrdinalEncoder() from *SKlearn*, considering their ordered nature. Data normalization via StandardScaler(),from *SKlearn*, prepares the dataset for subsequent modeling.

6 Modeling and Metric Selection

6.1 Train basic model

The first model computed was a simple logistic rergression. Let's see its classification report:



	Precision	Recall	F1-Score
0	0.46	0.32	0.38
1	0.38	0.01	0.02
2	0.53	0.89	0.67

Figure 3: Metrics in train model for 0 (defeat), 1 (draw) and 2 (victory)

The model exhibits a tendency to overpredict victories, with a notable recall value of 0.89 indicating a low number of False Negatives in predicting victories. However, the precision for victories is around 50%, indicating an inclination to assign more victories than actual occurrences.

In terms of draws, the recall is minimal, leading to numerous actual draws being inaccurately classified.

For defeats, both recall and precision fall short, with 64% of defeats considered as victories or draws, and almost half of predicted defeats not aligning with actual outcomes.

6.2 Metric selection

The training model achieves an accuracy of 0.519, indicating suboptimal performance. Despite this, the main project objective focuses on analyzing the influence of diverse features, so the project proceeds without introducing additional features. The chosen metric for optimization is **accuracy**, emphasizing the aim of predicting as many results as possible.

6.3 Complex models

In the pursuit of an improved classification model, various models undergo a hyperparameter search. Some examples are the Random Forest Classifier, Gradient Boosting (XGBoost), and logistic regression,

considering both One vs One and One vs All classification strategies.
Following the identification of optimal hyperparameters for the training model, the test matrix is meticulously prepared and subjected to testing using the aforementioned models. Among them, the Random Forest Classifier stands out, featuring an accuracy value of 0.517 on the test set, closely aligning with the training accuracy of 0.522. This proximity suggests the absence of overfitting. Henceforth, this Random Forest Classifier will be referred to as ”the model,” signifying its selection as the preferred model.

7 Results Analysis

Once the model has undergone computation, a detailed exploration of cases where the model excels in prediction is undertaken, examining the impact of features on the predictions.

7.1 Feature Importances

Initially, an analysis is conducted to identify the most influential features for the model. Columns exhibiting higher feature importance include *T1PliS/PviS* and *T2PliS/PviS*, *T2PliS*, *last_top6_vis/loc*, *GoalsliS/viS*, *season* and *Rk_loc/vis*. Although different models produced varying feature importance values for each feature during testing, these attributes consistently ranked high.

The significance of predicting outcomes is underscored by the importance of considering normalized points accumulated by each team in the match, as well as factors such as goals scored in the current league, team ranking, and their Top6 position in the previous season.

However, an intriguing aspect is exploring the relevance of league points accumulated in the last matches preceding the game and whether they contribute to enhanced prediction accuracy. Therefore, an in-depth analysis is carried out to assess how accuracy fluctuates when the team’s point streaks are incorporated into the model.

7.2 Streak analysis

To analyze the impact of team streaks on predictions, four distinct cases were examined. The accuracy of the Random Forest model was compared with that of a model relying solely on the last 3 matches’ streak, another considering the last 5 matches, and one without incorporating any streak information. The accompanying graph illustrates how the model’s accuracy is influenced by the inclusion of these streaks.

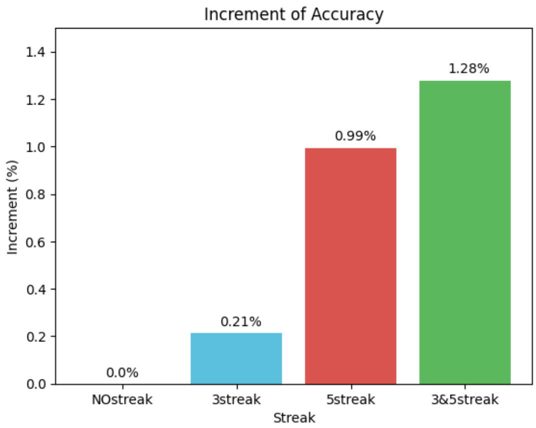


Figure 4: Increment of Accuracy depending on the streaks considered

Considering the last 5 matches streak has a more significant impact on accuracy compared to considering only the last 3 matches. The model achieving the highest accuracy incorporates both streaks, resulting in a mere 1.28% improvement over the model without streak information.

In conclusion, each team’s streak positively influences match result predictions, particularly when considering longer streaks. Interestingly, these features did not exhibit the highest feature importance, indicating that other factors play a more crucial role in prediction.

7.3 Model errors

Let us delve into instances where the model fails to make accurate predictions.

7.3.1 Accuracy per *season*

The accuracy of the model has been calculated individually for each *season*.

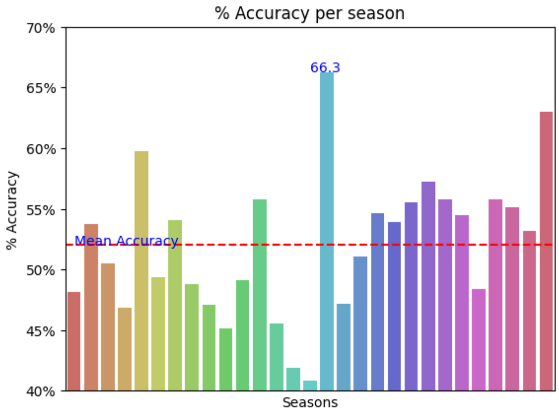


Figure 5:

Recent seasons exhibit above-average accuracy, indicating the model improves in recent data.

7.3.2 Accuracy per *round*

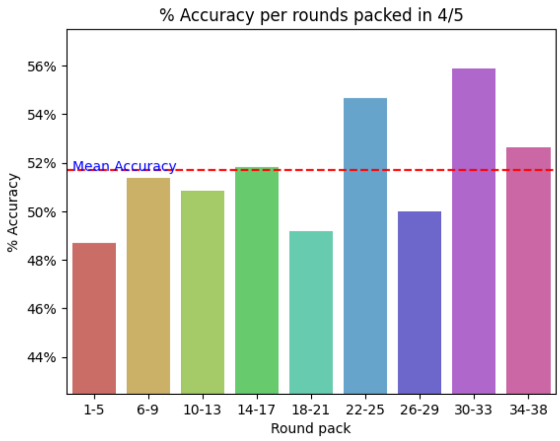


Figure 6:

It is observable that there is not a clear pattern in the rounds' accuracy. Nonetheless, the half-last rounds have a slightly higher accuracy caused by stronger acumulative feature values.

7.3.3 Accuracy per team

Firstly, the accuracy of each team as a *localTeam* has been calculated. Its values oscillate from 0.0 to 0.8. The teams with higher accuracy are Mérida and FC Barcelona. The ones with the lowest are Castellon and Lleida.

- Mérida (Acc: 0.80): It only appears in one season in the dataset, and most of the predictions are defeat, which is correct. The only error is a predicted loss against Madrid, which resulted in a draw.
- FC Barcelona (Acc: 0.78): Most of the predictions were victory, which was true.
- Lleida (Acc: 0:00): It only appears in one season and only had 3 instances of local matches in the test matrix. (Irrelevant)
- Castellon (Acc: 0.00): It had lots of defeat predictions which resulted to be a draw. But no draw was predicted.

The confusion matrix for Castellon as a *localTeam* reveals an absence of draw predictions, prompting the exploration of whether teams with a high proportion of draws are unpredictable.

7.4 Draw Accuracy Analysis

At this point, the accuracy of the model in a certain team depending on how many matches it has drawn will be studied. From this point, we no longer distinguish between a local or a visitor team, the number of draws is the entire one.

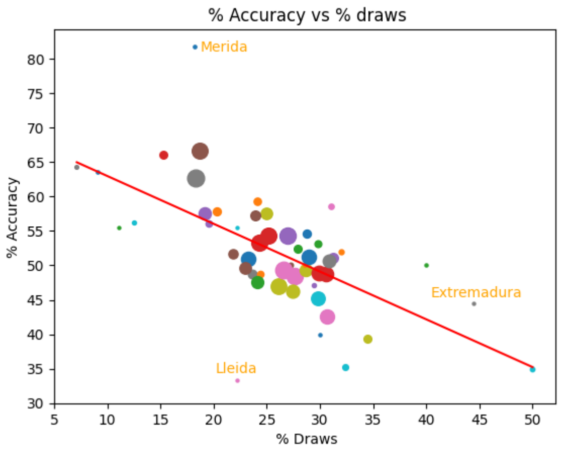


Figure 7:

This graph illustrates the correlation between the percentage of draws for a specific team and its accuracy. A notable trend emerges, indicating a decrease in accuracy for teams with a higher proportion of draws. A linear regression is employed to elucidate this tendency, with larger points denoting teams with more seasons played in La Liga. Interestingly, a second linear regression, focusing solely on these larger points, exhibits a steeper negative slope. This supports the hypothesis that teams with more draws are inherently more unpredictable, especially among established 1st division clubs.

Outlying teams in the graph, such as Merida or Lleida, are deemed as outliers due to their limited samples, having participated in only a single season in the dataset, rendering them less reliable.

8 Conclusion

In conclusion, creating a precise model for predicting football match outcomes is challenging due to inherent difficulties. Key features, such as the proportion of points earned and the seasonal ranking, significantly impact predictions. Additionally, team streaks play a noteworthy role. However, accurately forecasting teams with frequent draws proves to be a more substantial challenge. Despite the importance of various features, the persistent unpredictability of specific situations remains a constant hurdle in this predictive endeavor.

9 Code and Graphics

For code revision or additional analysis of my project, please scan the QR code below, which will direct you to my GitHub repository.