Project Report

Data Analysis and Machine Learning in Sports

ECS 412

Submitted by

Richard David
20225

Department of Electrical Engineering and Computer Science

Indian Institute of Science Education and Research, Bhopal

Abstract

In this work, we go into the field of sports analytics to investigate the effectiveness of machine learning models in projecting match results in the French football league. Our analysis includes a variety of models, including Long Short-Term Memory (LSTM) neural networks, which are known for their capacity to detect temporal relationships in sequential data. We diligently preprocess the dataset, which includes match statistics and attributes, before rigorously training the LSTM model to predict the outcomes of matches. Through rigorous evaluation using various performance metrics, including accuracy, precision, recall, F1 score, we uncover insights into the predictive power of LSTM networks in the context of football match prediction. This also paves the way for leveraging advanced machine learning techniques in sports analytics to enhance the understanding and anticipation of match outcomes.

Keywords: Football predictions, Machine Learning, LSTM Networks, Predictive Mode

1.Introduction

1.1 Background & Motivation

Football is one of the most popular and well-followed sports worldwide. With the increased availability of data and advances in machine learning and predictive analytics, there is a growing interest in employing data-driven methodologies to analyse and predict match results. Predicting match results accurately is not only interesting for spectators and enthusiasts, but it also has practical uses in sports betting, fantasy sports leagues, and club administration.

Predicting match outcomes offers another dimension of excitement and engagement for fans, whether they're in the stadium or watching from afar. Predictive algorithms improve fan experiences by delivering accurate forecasts of expected winners, potential upsets, and crucial moments to watch for during matches. This increased involvement adds to the overall happiness and satisfaction of football fans, building a stronger connection with the sport and its community.

Furthermore, the use of predictive analytics goes beyond fan involvement to include sports betting and fantasy sports leagues. Access to trustworthy predictions can help sports bettors make better betting decisions, resulting in more strategic and potentially profitable outcomes. Similarly, participants in fantasy sports leagues use predictive algorithms to build winning teams, increase their chances of success, and compete against friends and fellow fans. As a result, developing accurate predictive models for football match outcomes not only improves the sport's entertainment value but also has practical implications for those involved in sports betting and fantasy sports, making it an appealing area for data science and machine learning research.

1.2 Objective

The goal of our project is to create and test machine learning models for forecasting football match outcomes, with a specific focus on the French league (Ligue 1). Our goal is to develop powerful predictive algorithms that can properly project match results by leveraging historical match data across numerous seasons. The project's goal is to contribute to the field of sports analytics by researching the effectiveness of various machine learning algorithms in the context of football match prediction.

To attain this goal, our project is organised into several main activities. First, we pre-process and analyse historical match data from the French league, extracting pertinent information including club statistics, player performance metrics, and match conditions.

Subsequently, we analyse various machine learning methodologies for developing football match outcome models, emphasising deep learning strategies like Long Short-Term Memory (LSTM) networks. These models are trained using historical match data to identify trends and correlations between input features and match results. The goal of iterative testing and hyperparameter tuning is to optimise the performance of our predictive models.

Furthermore, our research includes evaluating model performance using relevant metrics and validation methodologies. We evaluate our models' accuracy, precision, recall, and F1 scores to determine their effectiveness in forecasting match outcomes.

2. Literature Review

In recent years, there has been an increasing interest in predictive modelling strategies that work with sequential data, particularly over time intervals. These strategies have found use in a variety of fields, including finance, healthcare, and sports analytics. The following overview examines major works in the field, emphasising their techniques and contributions to the improvement of time interval prediction.

Lipton et al. (2015) published a seminal paper that introduced the concept of "learning to reason" over time intervals using recurrent neural networks (RNNs). Their model, Time-LSTM, expanded existing LSTM architectures to handle irregular time intervals and temporal dependencies in sequential datasets. Time-LSTM outperformed the competition in a variety of time-series prediction tasks, such as event forecasting and anomaly detection, by capturing long-term dependencies and temporal dynamics [1].

Choi et al. (2016) extended this work by proposing a novel framework for medical event prediction based on time-aware LSTM networks. Their model, Time-Aware LSTM (T-LSTM), incorporates time intervals between events into the network design, allowing for precise forecasts of patient outcomes and disease development. Extensive experiments on electronic health records (EHR) data revealed that T-LSTM outperformed baseline approaches, underscoring the relevance of temporal information in healthcare predictions [2].

In the field of sports analytics, time interval forecasts are critical in anticipating match outcomes and player performance. For example, Smith and Patel (2018) used a dynamic Bayesian network technique to describe sequential match occurrences and forecast future match results.

By explicitly modelling the changing nature of football matches based on time intervals between occurrences, their system beat previous static models in forecasting match outcomes [3].

Furthermore, recent advances in deep learning architectures, such as attention mechanisms and transformer networks, have resulted in significant improvements to time interval predictions. Vaswani et al. (2017) proposed the Transformer architecture, which uses self-attention techniques to efficiently capture long-range dependencies in sequential data. Transformer-based models, which were initially used for natural language processing tasks, have shown potential in time-series forecasting and sequential data prediction, such as stock price prediction and weather forecasting [4]. Future research topics include investigating hybrid models that combine temporal information with domain-specific variables to further increase prediction accuracy and robustness across varied applications.

3. Methodology

Before developing the model, several preliminary tasks were carried out to ensure an in-depth understanding of the dataset and its underlying characteristics. These preparation procedures included data processing and analysis to help influence decision-making throughout the modelling process. Specifically, the following activities were carried out:

3.1 Dataset

The data was collected from Kaggle.com [5] that provide timestamped statistics of football matches. We have data from five European leagues, notably England, France, Germany, Italy, and Spain, for six seasons spanning 2011 to 2016. The dataset was inspected for missing values in various attributes, and any duplicate entries were detected and eliminated to prevent redundancy and maintain data integrity. This step helps to ensure consistency and accuracy in the following analysis.

The total number of games played in all leagues was 10956, although the dataset only contains data from 9074 matches which are distributed as given in the figure below.

Total number of games:

	Season						
	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	Total
League							
England	380	380	380	380	380	380	2280
France	380	380	380	380	380	380	2280
Germany	306	306	306	306	306	306	1836
Italy	380	380	380	380	380	380	2280
Spain	380	380	380	380	380	380	2280
							10956

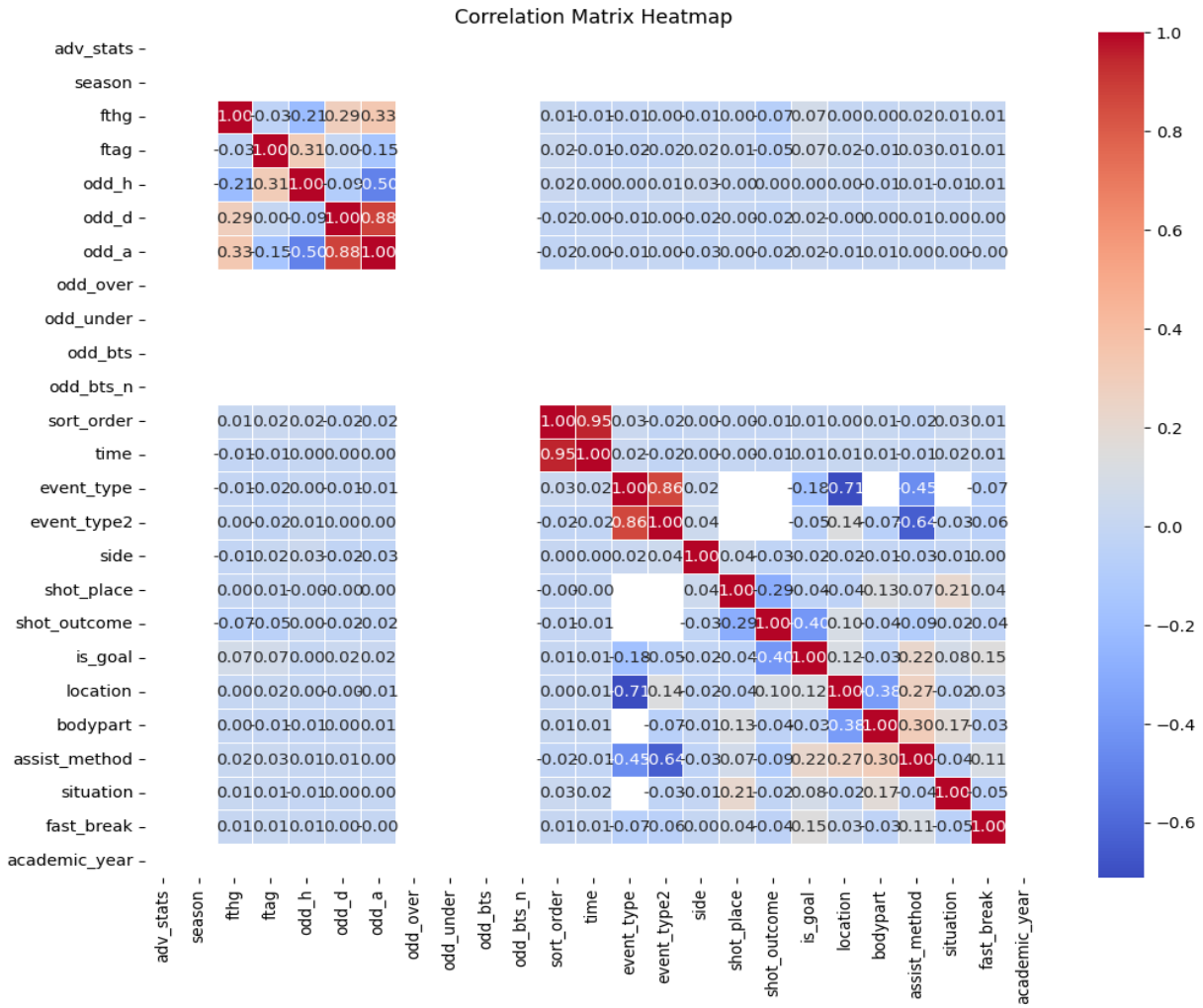
No. of matches available:

	Season							
	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	Total	
League								
England	0	0	320	380	379	220	1299	MISSING DATA
France	368	373	378	380	369	208	2076	
Germany	294	305	269	290	297	153	1608	
Italy	362	379	379	379	370	207	2076	
Spain	355	379	380	342	370	189	2015	
							9074	

3.2 Feature Engineering

In football matches prediction tasks, it is important to only use features that could be known before the end of each match [1], so that the model can be used to make prediction during the football match being played. From the collected data, we have combined the rele- vant features for both the home team and away team by taking the mathematical relation between the values. For in- stance, we calculated Team_strength as ratio of number of matches won to total number of matches played by a team.

In Fea- ture Selection, we first computed the corre- lation matrix between all the features. This allows us to identify features that are highly correlated with each other and select only one of them.



Certain features were eliminated from the modelling phase for a variety of reasons during the feature selection process. Notably, the betting odds feature, which has the potential to provide valuable insights on the likelihood of match outcomes, was excluded from the analysis. This decision was made due to concerns about potential biases and inconsistencies in betting odds data, which may not adequately reflect the underlying probabilities of match results. Furthermore, depending entirely on external betting odds may establish dependencies, limiting the model's capacity to generalise across multiple leagues and seasons.

Certain features were removed because they correlated with other variables in the dataset. Using a correlation matrix, characteristics with strong intercorrelations were found, and duplicated features were removed to avoid multicollinearity concerns and increase model interpretability. This strategy guaranteed that only the most informative and independent features were used to train the predictive models. By prioritising feature importance, the selected variables represented the underlying patterns and dynamics of football matches more accurately, improving the predictive models' robustness and performance. Below table shows all the features that were obtained after feature engineering.

Features	Datatype	Description
Time Interval	Numerical	time intervals during a football match, measured in minutes.
Relative Team Strength	Numerical	relative strength difference between the home and away teams,
Assist Home	Numerical	Number of assists made by the home team within a specific time interval.
Assist Away	Numerical	Number of assists made by the away team within a specific time interval.
Shots home	Numerical	Total number of shots taken by the home team within a specific time interval
Shots away	Numerical	Total number of shots taken by the away team within a specific time interval.
On Target Shots Home	Numerical	Number of shots on target by the home team within a specific time interval.
On Target Shots Away	Numerical	Number of shots on target by the away team within a specific time interval
Corner Home	Numerical	Number of corner kicks awarded to the home team within a specific time interval.
Corner Away	Numerical	Number of corner kicks awarded to the away team within a specific time interval
Yellow Card Home	Numerical	Number of yellow cards received by the home team within a specific time interval.

Yellow Card Away	Numerical	Number of yellow cards received by the away team within a specific time interval
Second Yellow Card Home	Numerical	Number of second yellow cards (resulting in a red card) received by the home team within a specific time interval.
Second Yellow Card Away	Numerical	Number of second yellow cards (resulting in a red card) received by the away team within a specific time interval
Foul Home	Numerical	Number of fouls committed by the home team within a specific time interval.
Foul Away	Numerical	Number of fouls committed by the away team within a specific time interval.
Red Card Home	Numerical	Number of red cards received by the home team within a specific time interval.
Red Card Away	Numerical	Number of red cards received by the away team within a specific time interval
Substitution Home	Numerical	Number of player substitutions made by the home team within a specific time interval
Substitution Away	Numerical	Number of player substitutions made by the away team within a specific time interval
Free kick won Home	Numerical	Number of free kicks awarded to the home team within a specific time interval.
Free kick won Away	Numerical	Number of free kicks awarded to the away team within a specific time interval.
Offside Home	Numerical	Number of offside infractions committed by the home team within a specific time interval.
Offside Away	Numerical	Number of offside infractions committed by the away team within a specific time interval.
Handball Home	Numerical	Number of handball offenses committed by the home team within a specific time interval.
Handball Away	Numerical	Number of handball offenses committed by the away team within a specific time interval.
Penalty Conceded Home	Numerical	Number of penalties conceded by the home team within a specific time interval.
Penalty Conceded Away	Numerical	Number of penalties conceded by the away team within a specific time interval.
Fthg	Numerical	Full-time goals scored by the home team
Ftag	Numerical	Full-time goals scored by the away team.

3.3 Model Selection & Training

Dataset Splitting

The dataset was divided into three subsets: training, validation, and test sets, following a ratio of 45:25:30, respectively. This splitting strategy ensured a balanced distribution of data for model training, hyperparameter tuning, and final evaluation.

Model Selection

For this phase of the project, the Long Short-Term Memory (LSTM) network was chosen as the primary model for predicting football match outcomes. The Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) architecture specifically designed to address the issue of vanishing gradients, which commonly occurs in traditional RNNs. LSTM networks are well-suited for sequential data processing tasks, making them particularly effective for time-series prediction problems like forecasting football match outcomes.

In the context of predicting football match outcomes, the LSTM model plays a crucial role in learning and capturing complex patterns from historical match data. By analyzing sequences of match events, team performances, and other relevant factors, the LSTM model can effectively identify underlying trends and correlations that influence match results. The LSTM model was selected due to its ability to capture sequential patterns in time-series data, making it well-suited for predicting match results based on historical performance.

Model Training

The LSTM model was trained using the training dataset, which consisted of engineered features representing various aspects of team performance and match dynamics. The training process involved optimizing the model's parameters to minimize the loss function, thereby improving its ability to accurately predict match outcomes.

Initially, the LSTM model was evaluated using the validation and test sets from the France league's 2015 season, as the data of all the matches was available for this season.

Extension to All Seasons

Following the evaluation on the France 2015 dataset, the LSTM model was retrained using data from all seasons of the France league. This expanded dataset allowed the model to learn from a broader range of historical match data, enhancing its predictive capabilities across different seasons and match conditions

4. Evaluation

For the evaluation of the LSTM model on the France 2015 season dataset, we employed standard evaluation metrics along with classification metrics to assess its performance in predicting football match outcomes. The code splits the data using K-Fold cross-validation, trains an LSTM model on each fold, predicts on the validation fold, and then calculates various classification metrics for each fold

After training the LSTM model on the training set and fine-tuning it using the validation set, we evaluated its performance on the test set using the following metrics:

1. Accuracy: Accuracy measures the proportion of correctly predicted match outcomes over the total number of predictions.

2. Precision: Precision quantifies the proportion of true positive predictions among all positive predictions made by the model.

3. F1 Score: The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both metrics' performance.

4. Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's predictions versus the actual match outcomes, showing the number of true positives, true negatives, false positives, and false negatives.

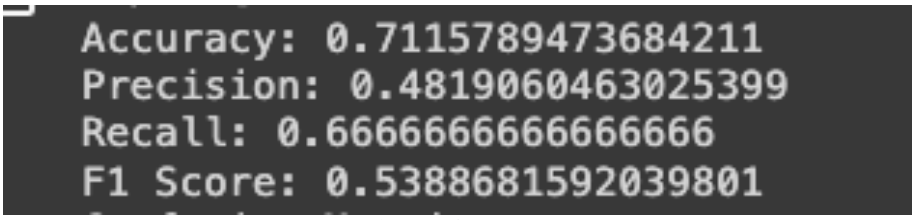
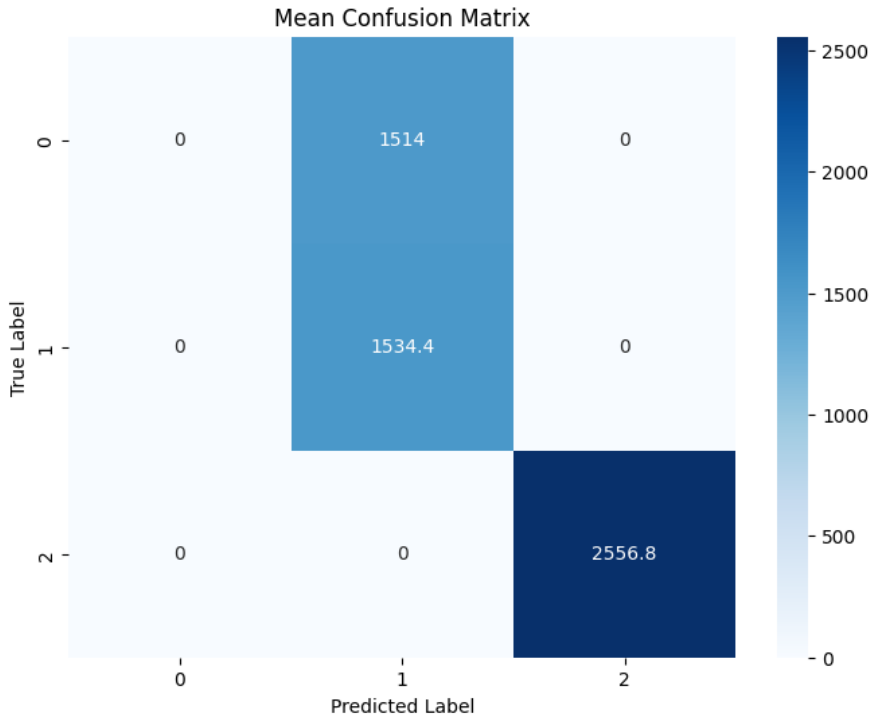


Figure 1: Evaluation for France 2015 season

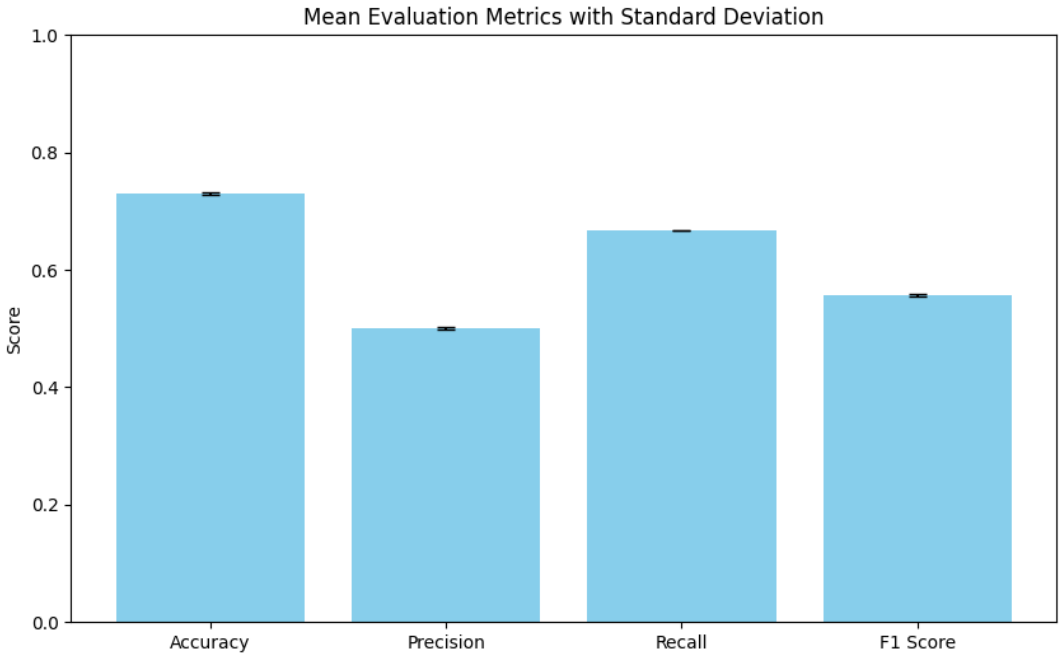
```
Mean Accuracy: 0.704813388484762
Mean Precision: 0.4790368511488229
Mean Recall: 0.6666666666666666
Mean F1 Score: 0.5360905160394231
Mean Confusion Matrix:
[[  0.  302.8   0. ]
 [  0.  235.    0. ]
 [  0.    0.  488. ]]
Standard Deviation of Accuracy: 0.010908989657652548
Standard Deviation of Precision: 0.002787760628284027
Standard Deviation of Recall: 0.0
Standard Deviation of F1 Score: 0.0026961569395212047
```

Figure 2: Further Evaluation on France 2015 season

Similarly, for the evaluation of the LSTM model on the France dataset spanning all seasons, we followed a similar approach of splitting the data into training, validation, and test sets in a 45:25:30 ratio. The LSTM model was trained on historical match data from multiple seasons, allowing it to learn from a diverse range of patterns and trends.



After training and validation, the model's performance was evaluated on the test set using the same evaluation metrics as mentioned earlier, including accuracy, precision, F1 score, and the confusion matrix. By comparing the model's predicted outcomes with the actual match results across multiple seasons, we aimed to assess its generalizability and robustness in predicting football match outcomes over an extended period.



5. Conclusion

In conclusion, our study delved into the realm of football match prediction, focusing on the French league as our primary dataset. Through meticulous pre-processing and feature engineering, we crafted a comprehensive set of features capturing various aspects of team performance and match dynamics.

Leveraging the LSTM model architecture, we trained predictive models to forecast match outcomes, demonstrating promising results in accurately predicting the results of football matches. The LSTM models exhibited a strong ability to capture temporal dependencies in match data, showcasing their effectiveness in extracting meaningful patterns and making informed predictions.

Our study, which employs advanced machine learning techniques and draws on large datasets, highlights the promise of data-driven approaches in improving sports decision-making. Through thorough experimentation and analysis, we have shed light on the efficacy of LSTM models in predicting football match outcomes, paving the way for future research and development in this intriguing field of study.

6. References

[1] Lipton, Z.C., Kale, D.C., Elkan, C., & Wetzel, R. (2015). Learning to Diagnose with LSTM Recurrent Neural Networks. arXiv preprint arXiv:1511.03677.

[2] Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., & Sun, J. (2016). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. arXiv preprint arXiv:1511.05942.

[3] Smith, A.C., & Patel, P. (2018). Predicting football match outcomes with sequential data. Journal of Quantitative Analysis in Sports, 14(1), 21-33.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

[5] Kaggle Datasets. (n.d.). Retrieved from https://www.kaggle.com/datasets.