PITCH-PREDICT: ENGLISH PREMIER LEAGUE MATCH RESULTS
ENEL 645 - WINTER 2024

Henry Zhao, Paolo Geronimo, Carrie Chan, Israel Robles, Jon Fulford
https://github.com/z11ru/premier-league-match-result-prediction

Department of Electrical and Computer Engineering, University of Calgary

ABSTRACT

This project focuses on creating a predictive model for English Premier League (EPL) matches, leveraging deep learning techniques to forecast match outcomes based on historical data. It involved the development of a comprehensive dataset of player and match statistics, followed by experimentation with various neural network architectures. Enhancements in model performance are achieved through tuning of hyperparameters and the application of optimization strategies. The culmination of these efforts is a predictive tool that not only forecasts match outcomes but also provides insights into the key factors influencing these predictions. This project demonstrates the potential of deep learning in transforming sports analytics and enhancing predictive accuracy in real-world scenarios.

1. INTRODUCTION

The English Premier League (EPL) stands as the most prominent and widely followed football league globally, renowned for its intense competition and high level of play. The league's popularity extends beyond the UK, reaching 643 million viewers per game and a worldwide television audience of 3.2 billion people [1]. This large following leads to substantial financial interest in sports betting. According to Darren Small, director of integrity at betting and sports data analysts at Sportradar, "The current estimations, which include both the illegal markets and the legal markets, suggest the sports match-betting industry is worth anywhere between $700 billion and $1 trillion a year. About 70% of that trade has been estimated to come from trading on football" [2].

This immense market highlights the importance of integrity and the use of robust predictive analytics in sports. This can provide a way to safeguard against manipulation by alerting officials to outcomes that are unexpected and potentially fixed. This can also support a more ethical betting environment where decisions by consumers are made with access to reliable and comprehensive data. These predictive insights are not only valuable to the betting market but can provide value for team management and coaching strategies to improve team performance.

Due to limitations on the data available (since access to things like player heatmaps, real-time player positions, detailed statistics like shot speed, and player speed and distance travelled, etc. is not readily available), the goal is to produce a predictive system with only limited information as input. Specifically, this study will create a predictive system that predicts the outcome of an EPL match based only on the team sheets (list of participating players, including substitutes) as the input.

This is a challenging task given the league's multi-faceted nature. A random prediction equates to a 33.3% accuracy rate for three potential outcomes (Home Win, Draw, Home Loss). Therefore, the project goal is to achieve an accuracy rate of around 50-60%, which would significantly surpass the baseline of random guessing. This goal was set due to the complexity and inherent randomness of match outcomes as investigated in similar works seen in index two.

The report will detail the comprehensive methodology, encompassing the data collection, preprocessing, and the reasoning behind selecting specific neural network architectures. The subsequent sections will present the results of the model, delving into its performance metrics and analyzing the outcomes. A thorough discussion will follow, interpreting the results and exploring the implications of the findings. Finally, the report will conclude with a summary of the key findings and the limitations of the current study.

2. RELATED WORK

In the evolving field of sports analytics, there are many articles that have been dedicated to the development of predictive models in sports. This section aims to provide an overview of the existing literature, highlighting different methodologies and approaches that have been explored with a specific focus on the use of machine and deep learning.

A significant portion of recent research in sports analytics appear to gravitate towards using Long short-term memory (LSTM), a type of Recurrent Neural Network

(RNN). As this network was not covered in the course, it was researched and observed as a type of artificial neural network (ANN) which uses sequential, or time series data and the output of recurrent neural networks depends on the prior elements within the sequence [3]. However, RNN's can have issues with vanishing or exploding gradients, resulting in a gradient that is too large or too small potentially leading to a model that is no longer learning or unstable [3]. The LSTM architecture addresses the issues of the vanishing gradient and long-term dependencies by using cells, which have gates controlling the input, output or forgetting of the information that is needed to make the prediction [3].

One study [4] applying LSTM with a softmax classifier reported predicting 63.3% matches correctly from the 2018 World Cup group stages, though this accuracy appeared to decline in subsequent advanced matches. Another research [5] implemented LSTM with an attention mechanism and a sliding time window, enhancing its focus on pivotal match outcomes and allowing the model to focus on the team's short term game status. This achieved an accuracy range of 60-80% in predicting the outcome of 5 games for a given team. Similarly, research [6] demonstrated that LSTM outperformed traditional ANNs, achieving a test accuracy of 80.75%. These studies collectively highlight LSTM's strengths in sports prediction, particularly for its capability in understanding long-term dependencies in complex, sequential data.

Contrasting with RNN based approaches, other studies have explored different neural network architectures and machine learning models. Research [7] investigated convolutional neural networks, traditionally used in image recognition, for predicting sports events. By focusing on player capability levels, they achieved accuracy rates close to 80%, showing potential in a non-RNN approach. In a different approach, study [8] employed conventional machine learning models like Linear Regression, Gradient Boosted Decision Trees, and Random Forest on Premier League and La Liga data, achieving a best accuracy of 63.8% on the test set.

Additionally, from a betting perspective, this study [9] noted that despite technological advancements, predicting sports events remains highly challenging due to numerous unpredictable factors. This insight highlights that while neural networks show promise, the unpredictability inherent in sports continues to pose a significant challenge to any predictive model.

3. MATERIALS AND METHODS

3.1. Dataset Creation

The first step in creating the model is to process and compile the datasets into a more concise and usable format. There are two datasets used in this model, the first is composed of individual player statistics, and the second is composed of match statistics.

The dataset containing player statistics [10] covers the 1992/93 up to the 2022/23 seasons. There are 40 CSV files in the dataset, which each describe an individual statistic for all players. For example, there are files for goals, assists, minutes played, etc. Some of the more intricate statistics in the dataset include aerial battles won/lost, errors leading to goal, and woodwork (post) hits. A Python script was created to iterate through each of the 40 files and compile them into a single CSV.

The dataset containing match statistics [11] contains statistics for several leagues, including the year, goals scored by the home team and away team, and unique identifiers for every player who played in that match. Another Python script was written to create a subset of this dataset that only contains matches from the English league. This dataset was then exported into its own CSV. The dataset is composed of 1900 matches from the 2017 – 2021 seasons. Since the player IDs are used in the match data, there is another file in this dataset that includes the first and last names of the players assigned to each player ID to better understand which players played in each match.

Note that this project includes both starting and substitute players. According to "Early Prediction of Physical Performance in Elite Soccer Matches—A Machine Learning Approach to Support Substitutions" [12], substitutions are a critical tool for a coach to influence a game, initiated by factors such as player injury, tactical changes, or player underperformance. This project therefore includes the substitute players for the matches, even though in many cases, these players do not see game time.

The first step in merging the two datasets is to map each player's name from the first dataset to their ID from the second dataset. This results in essentially converting the "Player" column in the player dataset from names to IDs. This allows for easier merging between the datasets since the match dataset contains the IDs of the players in each match.

Next, the two datasets are merged. A Python script goes through each match, collecting previous season statistics from each player in the match and determines the mean for each team. The final dataset used to train the model is composed of the average statistics of each team, such as average goals, assists, passes, shots, etc. There are a total of 78 parameters. The final dataset also includes the result of the match, where a home team win, draw, and away team win are represented as a 1, 0, and -1 respectively. Finally, the dataset includes the goal differential from the home team's perspective. The script for this final step was computationally intensive, as it iterated over each of the 40 statistics for 36 players in 1900 matches (which multiplies to 2.7 million computation steps) and took approximately 10 minutes to complete.

3.2. Classification Models

To preprocess the data, the “Match ID” and “Home Goal Difference” columns were dropped, since the Match ID is an identifier and not a feature, and the Home Goal Difference is the target vector for the regression model, not the classification model. The data was scaled using a standard scaler, and the class labels were converted to 0, 1, and 2, instead of -1, 0, and 1. Five folds are used for cross validation.

Three classification models were created, and their results will be compared. The first model is a Fully Connected Neural network, and the second model is a Convolutional Neural Network. Finally, a Support Vector Machine was created.

The Fully Connected Neural Network has three layers. The first layer has an input size of 78 and an output size of 64. The next layer has an input size of 64 and output size of 32. The final layer has an input size of 32, and an output size of 3 since there are 3 possible classes. The activation function used in the model is ReLU.

The Convolutional Neural Network has two convolutional layers followed by one max pooling layer. The first convolutional layer has 1 input channel and 16 output channels. The next convolutional layer has 16 input channels and 32 output channels. Both convolutional layers have a kernel size of 3, a stride of 1, and padding of 1. The max pooling layer has a kernel size of 2, and a stride of 2. After the max pooling layer, there are two fully connected layers. The first fully connected layer has an input size of 78, and an output size of 64. The second fully connected layer has an input size of 64 and output size of 3.

Both the Fully Connected Neural Network and Convolutional Neural Network use cross entropy loss as their loss function, and both use an Adam optimizer with a learning rate of 0.001. 100 epochs were used during the training process for each model.

The Support Vector Machine used a linear kernel, and a grid search was used to find the best regularization parameter, C. 10 potential C values were used during the training and validation stages of the grid search.

3.3. Regression Models

Preprocessing the data for the regression portion of the project was done in a similar fashion for the classification models. The difference being that the regression models use the “Home Goal Difference” metric as the target vector, instead of the “Result” vector. While the Home Goal Difference metric also indicates who won, it also provides the extent to which the winning team won. This metric is calculated by subtracting the away team’s goals from the home team’s goals. A positive result indicates a home team win, a zero indicates a draw, and a negative result indicates an away team win.

The first regression model is a Fully Connected Neural Network with three layers. The first layer has an input size of 78, which is the number of features, and an output size of 64. The next layer has an input size of 64 and output of 32. The final layer has an input size of 32 and an output size of 1 since it is a regression model. ReLU is used at the model’s activation function.

A Convolutional Neural Network analysis was not performed. However, if it was, it would have followed a similar structure to that in the classification section, with two convolutional layers. The model was skipped as the Tabular CNN in the classification section did not meaningfully outperform the Fully Connected model in the same section.

The final regression model is a Ridge Regression model. A grid search was performed to find the best alpha parameter for the model. A total of 20 alpha values were used in the grid search. Both models use mean square error as the loss function.

4. RESULTS AND DISCUSSION

The objective of this project was to develop models that could predict the outcomes of English Premier League matches based on the average performance statistics of participating players from the previous season. Various machine learning and deep learning models were explored to obtain two primary models: a classification model to predict the match outcome as win, draw, or loss, and a regression model to predict home goal difference.

4.1. Classification Results

The three classification models that were explored and compared were: a Fully Connected Neural Network (FCNN), a Convolutional Neural Network (CNN), and a Support Vector Machine (SVM) with a linear kernel. The FCNN model achieved the highest validation accuracy during the training phase, with a 55.53% validation accuracy. However, in the test set evaluation, the SVM model performed the best, with a test accuracy of 52.89%, in contrast to the CNN model’s test accuracy of 42.37% and FCNN model’s test accuracy of 41.84%. Compared to the baseline accuracy of 33.33%, which is equivalent to randomly guessing among the three outcomes, the classification model's performance results slightly surpass it.

The following three figures show the confusion matrix results of their corresponding model. All three models showed strength in predicting home team wins (label 2), with the SVM model showing the highest number of correct predictions in this category. However, SVM showed significant weakness in identifying draws (label 1) and it failed to correctly classify any matches as such. All three models exhibit challenges in classifying matches with a high degree of accuracy as the models confuse home wins (label 0) with the other two possible outcomes.

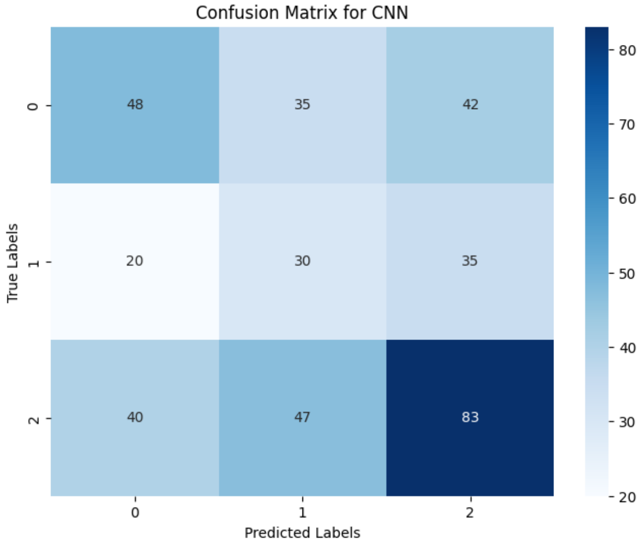

Figure 1. Confusion Matrix for CNN

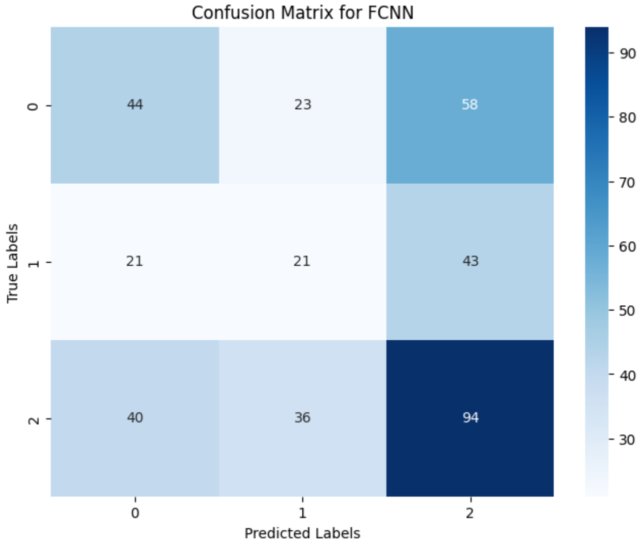

Figure 2. Confusion Matrix for FCNN

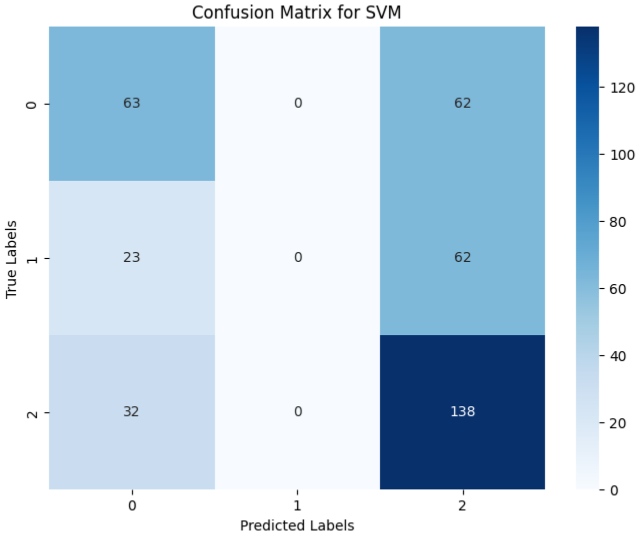

Figure 3. Confusion Matrix for SVM

Table 1 below summarizes the performance metrics obtained from each of the models. The CNN model has the highest Precision and F1 Score, indicating that a larger portion of predictions for the match outcomes were correct and it has a better balance between precision and recall compared to the other models. The SVM model had the lowest precision but the highest recall, indicating that the model identified a higher portion of actual positive outcomes, but also made more false positive predictions, leading to a lower precision value. FCNN had slightly lower scores across all metrics in comparison to the other models.

	Metrics		
Model	**Precision**	**Recall**	**F1 Score**
CNN	0.4104	0.4084	0.4065
FCNN	0.3879	0.384	0.3841
SVM	0.3535	0.4386	0.3858

Table 1. Comparison of performance metrics between models

4.2. Regression Results

Two regression models were explored to predict the goal difference for the home team. The two models that were evaluated were: FCNN and Ridge Regression models. The R^2 score and mean absolute error for the Ridge Regression model was calculated to be 0.1809 and 1.3042 respectively. The Ridge Regression model outperformed the FCNN model in terms of Root Mean Square Error (RMSE) on the test set, achieving a RMSE of 1.686, compared to FCNN's RMSE of 2.314.

4.3. Discussion

The errors for the regression models are quite high at 1.686 and 2.314, as compared to the typical goal difference expected in a game, which is usually around 0 to 2 goals, and rarely ever more than 4. The classification models achieved test accuracies in the 40% to 50% range, which is also not very high. However, in terms of usability, the classification models are much better, as they do significantly exceed the random guessing threshold of 33.3%, indicating that some amount of predictive quality exists in just the player lists.

Both neural network models struggled with overall accuracy and specifically with distinguishing between the three outcomes, as evidenced by their confusion matrix results and lower performance metrics.

Overall, the SVM model demonstrated the best accuracy, achieving the target of between 50 – 60% at 52.89%. However, it did suffer greatly in precision, especially because it was simply unable to predict draws as an outcome. This suggests that the model overfitted to the most common outcome of home wins. This does however give some insight, that there is a quantitative reason for the concept of “home advantage”.

FCNN was used for its theoretical capability of modelling non-linear relationships, and its flexibility in working with any data type. However, it likely also easily overfitted. The quantity of data may also not be sufficient at 1900 matches. It is also very difficult to interpret and adjust the FCNN, as it is a ‘black box’.

The CNN was used to capture any spatial hierarchies in data, to find if there was any combination of player statistics that more readily determine the outcome of the match than through a simple FCNN alone. It did perform marginally better than FCNN, but the results are rather inconclusive.

For regression, the results are rather unusable. Although theoretically they should capture more nuanced results other than just a win, draw, or loss, the performance was just too low. It is likely that the data just was not linear and thus cannot be modelled this way.

For ridge regression, the regularization should have helped with reducing complexity and overfitting, but this comes at the drawback of increased bias. For the FCNN used, it shares the strengths and drawbacks as it had in the classification section. It overfits easily and requires sufficient data.

The low accuracies and high errors indicate that there are significant factors outside of the list of players involved that affect the outcome of a match. However, the classification models do show promise that there is some predictive ability of just the player lists.

5. CONCLUSIONS

In the exploration of machine learning and deep learning models to predict English Premier League match outcomes from player performance statistics, a dual approach was utilized targeting both classification and regression outcomes. The objective was to find predictive insights from aggregated previous season performance data of players, normalized by minutes played. This study was also aimed to contribute to practical applications such as betting and fantasy sports, where rapid and accurate predictions based on limited, readily available information, are valuable.

The classification models, encompassing a Fully Connected Neural Network (FCNN), a Convolutional Neural Network (CNN), and a Support Vector Machine (SVM) with a linear kernel, despite not achieving high test accuracies—with figures hovering between 40% and 52.89%, demonstrated a capacity to surpass the baseline random guessing accuracy of 33.3%. This indicates a degree of predictive quality within player lists. Interestingly, the SVM's had no ability to predict draws, suffering from overfitting towards home wins.

In regression, both the FCNN and Ridge Regression models struggled to deliver usable insights, with error metrics roughly equal in magnitude to the metric they are supposed to represent, likely due to the non-linear nature of the data at hand. Before this study was conducted, linearity was not known, but it now conclusively can be said that the relationship between player stats and outcome is not linear.

Data quality may also have been an issue. 1900 matches may not be enough to train deep learning models. It was also quite time consuming to investigate each player stat individually (there were 40 for each player), so feature engineering was not done effectively. The data was only up to 2021, so it may not accurately reflect the nature of the game today in 2024.

In conclusion, the classification approach, particularly through SVM, seems to be promising. There is evidence that player list alone has some predictive quality over the outcome of a match, since all the models, despite being quite different in architecture, significantly exceeded the random guessing threshold. It can also be said that the relationship between player stats and outcome is not linear, due to the poor performance of regression models. Future work may explore more sophisticated data representations, alternative modeling techniques, or better datasets.

6. REFERENCES

[1] P. Moore, "Top 5 of the Most Watched Football Leagues in the Word today!," The Entertainment Engine, [Online]. Available: https://medium.com/the-entertainment-engine/top-5-of-the-most-watched-football-leagues-in-the-world-today-2b21007237db. [Accessed 01 April 2024].

[2] F. Keogh and G. Rose, "Football betting - the global gambling industry worth billions," BBC Sport, [Online]. Available: https://www.bbc.com/sport/football/24354124. [Accessed 01 April 2024].

[3] IBM, "What are recurrent neural networks?," [Online]. Available: https://www.ibm.com/topics/recurrent-neural-networks. [Accessed 01 04 2024].

[4] M. Rahman, "A deep learning framework for football match prediction," *SN Appl. Sci,* vol. 2, no. 165, 2020.

[5] Q. Zhang, X. Zhang, H. Hu, C. Li, Y. Lin and R. Ma, "Sports match prediction model for training and exercise using attention-based LSTM network," *Digital Communications and Networks,* vol. 8, no. 4, pp. 508-515, 2022.

[6] E. Tiwari, P. Sardar and S. Jain, "Football Match Result Prediction Using Neural Networks and Deep Learning," *in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization,* pp. 229-231, 2020.

[7] S.-H. Lin, M.-Y. Chen and H.-S. Chiang, "Forecasting Results of Sport Events Through Deep Learning," *2018 International Conference on Machine Learning and Cybernetics,* pp. 501-506, 2018.

[8] S. Hu and M. Fu, "Football Match Results Predicting by Machine Learning Techniques," *2022 International Conference on Data Analytics, Computing and Artificial Intelligence,* pp. 72-76, 2022.

[9] T. Korotyeyeva, R. Tushnytskyy and V. Kulyk, "Applying Neural Networks to Football Matches Results Forecasting," *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies,* pp. 278-282, 2018.

[10] D. A. Teixeira, "Premier League Player Statistics (1992/93 - 22/23)," Kaggle, January 2024. [Online]. Available: https://www.kaggle.com/datasets/davidantonioteixeira/premier-league-player-statistics-1992-2022. [Accessed 8 March 2024].

[11] H. Mathien, "European Soccer Database," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/hugomathien/soccer/data. [Accessed 8 March 2024].

[12] T. B. Dijkhuis, M. Kempe and K. A. P. M. Lemmink, "Early Prediction of Physical Performance in Elite Soccer Matches—A Machine Learning Approach to Support Substitutions," *Data Analytics in Sports Sciences: Changing the Game,* vol. 23, no. 952, 2021.