Yonathan Julian Putra Irawan - r0767981

Jad Haddad - r0767008

World Cup 2022 Prediction

1. Abstract

This paper presents a method for predicting the outcome of the 2022 FIFA World Cup using a machine learning model built from historical data. The model takes into account factors such as team performance and other relevant variables to make predictions about the outcome of future matches. Previous work on using machine learning to predict outcomes in team sports is reviewed, and the dataset for the study includes results of international football matches from 1993 to the present. The model is trained using a multi-layer perceptron (MLP) classifier, and it has predicted the winner of the world cup based on the competition bracket. The result showed that Argentina is the most likely winner of this predictor.

2. Introduction

As the 2022 FIFA World Cup progresses, it is interesting to try to predict the outcome of the tournament. In order to do this, this paper will outline the process of creating a machine learning model using historical data that includes past international matches dating from 1993 and how this model can be used to make predictions about the World Cup. The model will be built by analysing data from past international matches that contains past friendly matches and also past competitions, and using this information to identify patterns and trends that may be relevant for predicting the results of the current tournament. Once the model has been built, it can be used to make predictions about the outcome of future matches by taking into account factors such as the teams' performance and other relevant variables. Ultimately, the goal of this paper is to provide a useful tool for analysing and predicting the outcome of the World Cup, which can be of interest to fans, sports analysts, and others who are interested in the tournament.

3. Related Work

The paper “The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review“ reviews the use of machine learning to predict outcomes in team sports, covering 1996 to 2019. It examines the effectiveness of various algorithms and evaluates the accuracy of predictions, while also considering the difficulty of predicting different sports and offering recommendations for future research (Bunker & Susnjak, 2019). An article discussed the use of machine learning to predict the NCAA March Madness basketball tournament, using the classification method and running a dataset through the model to get results. The author found success in predicting the tournament using this approach. (Morrell, 2022). Another study used a neural network model to predict soccer match outcomes and identify important attributes for success, achieving an accuracy of 83.3% for wins and 72.7% for losses. The model identified 19 attributes that were powerful predictors of success. (Hassan et al., 2020). The next study sought to profit from the betting market using machine learning, focusing on decorrelating model output from bookmaker predictions while maintaining a high accuracy and introducing the use of convolutional neural networks to aggregate player-level statistics. The proposed models were able to achieve a positive profit over 15 NBA seasons (Hubáček, n.d.). Multi-layer perceptrons are powerful classifiers that can perform well but have a large number of free parameters and may suffer from long training times and local minima. In this chapter, an ensemble of MLP classifiers is proposed to address these problems and optimise performance through the use of measures that correlate with generalisation error (Windeatt, T., 2008).

4. Dataset and Features

There is a list of the results of international football matches dating back to 1993 inside a dataset from Kaggle (Breda_L, 2022).

Some data manipulation is done to have the following features:

Feature	Description
team_ranking	The FIFA ranking for the team during the match. (Breda_L, 2022)
opponent_ranking	The FIFA ranking for the opponent team during the match (Breda_L, 2022).
match_coefficient	The coefficient of match, which is processed by averaging both team's confederation coefficient weights, which is set by FIFA (Kelly, 2017).
match_importance	The importance of the match. The original dataset had the name of the tournament of the match played. This is converted into weights, according to FIFA's official formula (Kelly, 2017).
neutral_location	This is a boolean that returns whether a match took place in a neutral location or not.

Table 3.1: Features and their description

Below is a snippet of the dataset.

team_ranking	opponent_ raking	match_ coefficient	match_ importance	neutral_ location	team_result
59	22	1	2.5	False	Win
8	14	0.925	1	False	Draw
35	94	1	2.5	False	Win
65	86	0.85	1	False	Win
67	5	1	2.5	False	Lose

Table 3.2: The first five data points of the dataset.

The last column in Table 3.2 is the result, which has the following classes: win, draw, lose.

5. Method

5.1. Algorithms

5.1.1. Logistic Regression

The first algorithm used to learn the model is logistic regression with multiple features.. As classification is used, the hypothesis uses the sigmoid function.

$$h_{\Theta}(x) = g(\Theta^{T}x) \qquad \text{(Equation 4.1)}$$

$$g(z) = \frac{1}{1+e^{-z}} \quad \text{(Equation 4.2)}$$

$$y^{(i)}(x) = \{ \begin{matrix} 1, if\ h_\theta > 0.5 \\ 0, if\ h_\theta < 0.5 \end{matrix} \quad \text{(Equation 4.3)}$$

Equation 4.4 is used to calculate the cost function of this algorithm.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[- y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))] \quad \text{(Equation 4.4)}$$

To find the gradient for of J in respect to θ, the following equation is used:

$$\frac{\delta J(\theta)}{\delta\theta_i} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \quad \text{(Equation 4.5)}$$

These two equations are implemented inside the python program, and the optimize.minimize function is used from the scipy library. This simple classification is implemented in the Logistic.ipynb notebook.

5.1.2. Neural Network as a Classifier

The second algorithm used is Neural Networks. We use MLPClassifier, which is a class in the scikit-learn library in Python that implements a multi-layer perceptron (MLP) algorithm, which is a type of artificial neural network. MLPClassifier can be used to classify data into two or more categories by training the classifier on a labelled dataset. In order to do this, the classifier would first need to be trained on the dataset that we’re using.

Once the classifier has been trained, it can be used to make predictions about the outcome of future World Cup matches by taking into account all the factors extracted from the dataset such as teams’ rankings, teams’ coefficients and importance of the games.

Then, you can create an instance of the classifier and fit it to your training data using the fit() method. The classifier will then learn to predict the output class for a given input by adjusting the weights of the connections between the nodes in the neural network. Once the classifier has been trained, you can use the predict() method to predict the class label for new data. (*Sklearn.neural_network.MLPClassifier — Scikit-Learn 1.2.0 Documentation*, n.d.)

The MLP Classifier used in this project uses the default amount of hidden layer size, which is one input layer, one hidden layer with 100 units and one output layer.

The implementation of the neural networks algorithm can be found in the Final.ipynb notebook.

6. Results

6.1. Accuracy and Confusion Matrix

Method	Accuracy test	AUC Score	Mean Error
MLP Classifier	60%	0.70	0.61
Logistic Regression	67%	0.68	0.32

Table 5.1.1: The accuracy of the models.

Looking at the results, Logistic Regression has the better performance which is expected since it's only predicting a win or a loss, whilst the Neural Networks are predicting three different scenarios (win, draw and loss). The AUC score of 0.70 in the MLP Classifier represents an acceptable score (Knight, n.d.), whilst the big margin between both mean errors lies on the difference in the class prediction. Overall the scores aren't bad, but can be better.

To try to improve the scores, we played around the lambda score, tried to implement polynomial features or even add more hidden layers but all lead to similar results or even worse. Whilst the results show that the algorithm is not accurate enough, it is as good as it gets noting that the dataset has no clear boundaries between classes, as seen below in the plot of the home wins, draws and away wins depending on the rankings of both home and away sides:

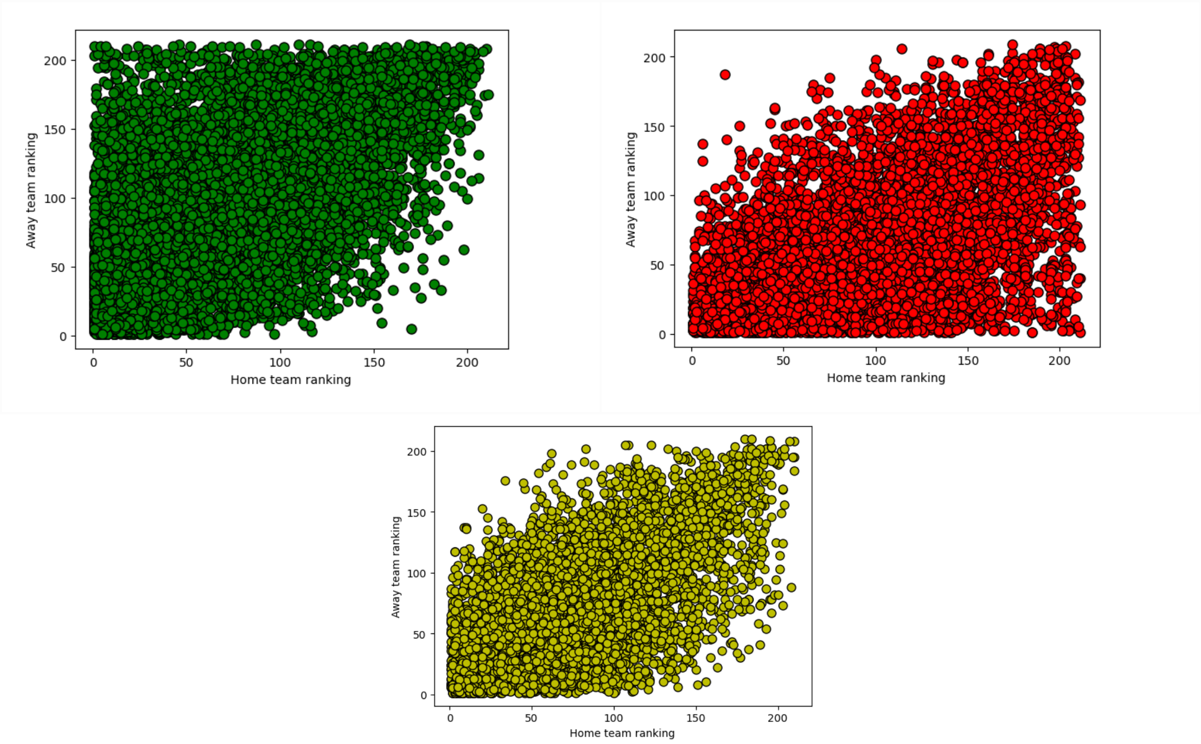

Figure 5.1.1: Three scatterplots on whether the home team wins, draws, or loses depending on its ranking and the opponent ranking.

Another flaw this model has is the lack of predicted draws and also the amount of actual draws that end up being predicted as a home or away win, as seen below in the confusion matrix:

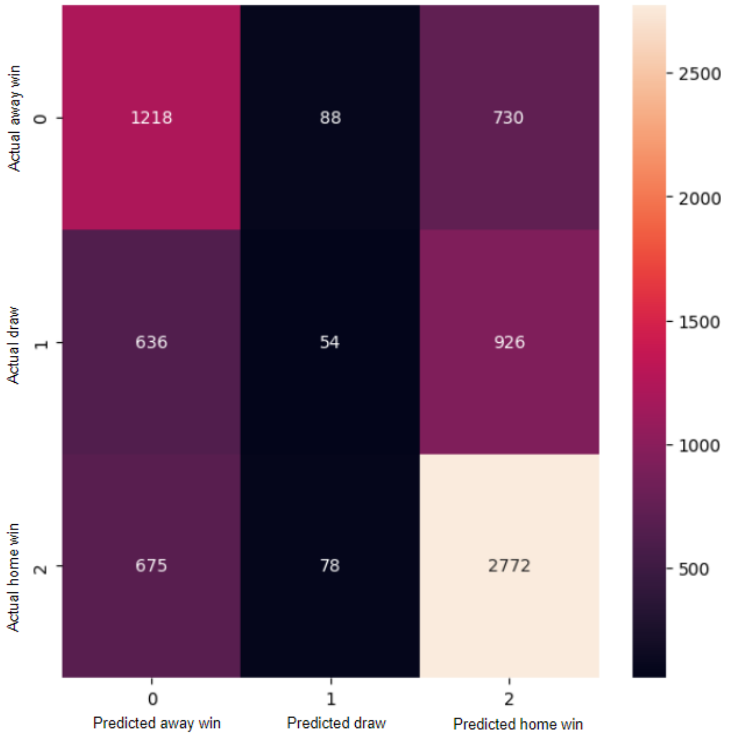

Figure 5.1.2: The confusion matrix

6.2. Predicting the World Cup

We used the model to predict this year's World Cup. The models are used to create a function where one can input two teams and it will return the winner from that match. This function is then used to predict each match of the World Cup and the whole tournament was simulated. This is iterated 1000 times and the winners, runners up, and third place are counted.

Country	Wins	Runners up	Third place	Fourth place
Netherlands	5	0	0	13
England	0	178	0	444
Argentina	566	20	374	22
France	0	35	23	320
Spain	0	0	12	12
Belgium	90	689	25	188
Brazil	331	78	566	1

Table 5.2.1: The nations who were in the top four in 1000 iterations.

Our results show that Argentina has the best chance of winning the whole competition, followed by Brazil and then Belgium. Furthermore, it can be seen that the model has a high bias. Only four teams are predicted to win the competition after 1000 iterations. Favourites like France and Spain have zero wins, while other strong teams like Germany and Croatia do not make it into the top four.

To improve the implementation of the model, instead of classifying the winner, the probability of the team winning can be used on each match. The winner from the match is then determined by chance, with the team with the higher probability having a higher chance. In this way, different teams will win more in the simulations, and therefore the results can be used to measure the probability of each team winning the world cup.

In addition, more features can be added. One important feature in the sport is the average age of the team, with a young team having little experience, and an older team is less athletic. This can be shown in the results, where the model predicted Spain (having a very young team) and Belgium (having an older team) to do much better than in reality.

7. Conclusion and Future Work

This paper presents a machine learning model that accurately predicts the outcome of the 2022 FIFA World Cup. The model is built using historical data and takes into account factors such as team performance and other relevant variables. The model is trained using a multi-layer perceptron (MLP) classifier and is able to predict the winner of the tournament based on the competition bracket. It correctly predicted that Argentina is the favourite for this competition. The results of the study show that this model is a useful tool for analysing and predicting the outcome of the World Cup, which could be of interest to fans, sports analysts, and others interested in the tournament.

As future work, one might consider adding more relevant features and use more powerful tools in order to enhance the accuracy of the model. Another way to predict the games is by measuring the games by goals scored and return the value of “Expected Goals” (xG) per team each game.

8. Appendix

The GitHub code and datasets can be found here ttps://github.com/elYonno/WorldCup2022Predictor.

9. Contributions

Both members of the team have contributed equally.

10. References

Breda_L. (2022, August 28). *FIFA World Cup 2022 ⚽*. Kaggle. Retrieved November 26, 2022, from https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022

Bunker, R., & Susnjak, T. (2019, December 26). *[1912.11762] The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review*. arXiv. Retrieved December 19, 2022, from https://arxiv.org/abs/1912.11762

Hassan, A., Akl, A.-R., Hassan, I., & Sunderland, C. (2020, June 5). *Predicting Wins, Losses and Attributes' Sensitivities in the Soccer World Cup 2018 Using Neural Network Analysis*. CORE. Retrieved December 19, 2022, from https://www.mdpi.com/1424-8220/20/11/3213/pdf?version=1591359811

Hubáček, O. (n.d.). *Exploiting betting market inefficiencies with machine learning*. Semantic Scholar. Retrieved December 19, 2022, from https://www.semanticscholar.org/paper/Exploiting-betting-market-inefficiencies-with-Hub%C3%A1%C4%8Dek/88371a15d722ee85946906ac6c4c28a58180e631

Kelly, R. (2017, October 12). *FIFA world ranking: How it is calculated and what it is used for*. Goal.com. Retrieved November 26, 2022, from https://www.goal.com/en/news/fifa-world-ranking-how-it-is-calculated-what-it-is-used-for/16w60sntgv7x61a6q08b7ooi0p

Knight, A. (n.d.). *What is a good ROC curve score?* Deepchecks. Retrieved December 19, 2022, from https://deepchecks.com/question/what-is-a-good-roc-curve-score/

Morrell, E. (2022, March 11). *How to Use Machine Learning to Predict NCAA March Madness*. Analytics8. Retrieved November 27, 2022, from https://www.analytics8.com/blog/how-to-use-machine-learning-to-predict-ncaa-march-madness/

sklearn.neural_network.MLPClassifier — scikit-learn 1.2.0 documentation. (n.d.). Scikit-learn. Retrieved December 19, 2022, from http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier

Windeatt, T. (2008). Ensemble MLP Classifier Design. In: Jain, L.C., Sato-Ilic, M., Virvou, M., Tsihrintzis, G.A., Balas, V.E., Abeynayake, C. (eds) Computational Intelligence Paradigms. Studies in Computational Intelligence, vol 137. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79474-5_6