Investigating the "home-advantage" in football using different machine learning models

Kaan Centoglu[1], Vlad-Andrei Cursaru[2], Laura Duits[3], Anas Kabboul[4], and Isha Kashyap[5]

[1]k.centoglu@student.vu.nl (2663242)
[2]v.cursaru@student.vu.nl (2669719)
[3]l.b.m.duits@student.vu.nl (2661221)
[4]a.kabboul@student.vu.nl (2669759)
[5]i.kashyap@student.vu.nl (2664411)

April 1, 2022

Abstract

1 Abstract

This paper analyses whether there is a significant home advantage in football. Firstly, we conduct a literature review regarding previous research about this topic. We then take our dataset and start inspecting the framework and prepare it for our models. Three machine learning models are chosen and using a validation set, the hyperparameters are picked. To investigate the home advantage, the models are trained with and without features that indicate the home and away team. We then analyse our results which indicate there is no significant home advantage in football.

Keywords— machine learning, home advantage, football, match prediction

2 Introduction

A home advantage can be described as the benefit a home team has during the match over the away team (or visiting team). This advantage has sparked many questions as to what attributes to it. This could be either physical effects, such as players from the away team experiencing jet lag, or it could be psychological effects, such as the influence of the crowd in the stadium. According to papers studied in our literature review, the home advantage has been declining. As of 2021/2022, UEFA has decided to remove the 'away-goal rule' from all its competitions due to this decline. According to the UEFA president, "home advantage is nowadays no longer as significant as it once was" [1]. In this research, we utilize various machine learning methods to investigate the home advantage in football. We approach the problem by training three different machine learning models on a dataset of historical football matches. The three models selected are the kNN classifier, artificial neural network and random forests classifier. To investigate the relevance of home advantage, each model will be trained twice, one time the features indicating which team was playing at home

were included, and the other time they were withheld. By analysing the difference in predictive performance between the two versions, we attempt to answer the following research question: Is there a significant home advantage in football?

In section 3, we present our literature review based on several research papers. These papers are based on research conducted regarding similar topics to the one discussed in this paper. In section 4, the data is first inspected. Then, the data is pre-processed and split into a train and test set, which can be used to train the different models. The chosen methods are described in section 5. Section 6 discusses the hyperparameter selection for the methods. The final models are chosen and used to run the test set to determine the performance. In section 7, the performance of the models belonging to the different methods are compared and the research question is answered. Finally, section 8 discusses further research that can be conducted.

3 Literature Review

There have not been many approaches to investigate home advantages in football using machine learning techniques; however, there are many research papers available analysing home advantage considering several parameters and comparing previous data. Also, the prediction of football results using machine learning algorithms and models is very common research material. For our research paper, the information in the following articles was useful and influential to us.

Thomas et al. [2] examined home advantage in the English Football Premiership using and extending a previously done research by Pollard, who also examined the same topic. In his paper, Pollard uses the one-sided chi-square test to find home advantage in the English Football Premiership between 1888-1984 [3]. Thomas et al. investigated the data from 1984-2003 and found that, compared to the results of Pollard, the mean home average is significantly lower than their findings. However, they also state that there is no evidence to support that this reduction would continue over time.

Pollard continues his previously mentioned research that he conducted in 1986. In this paper, he states that "home advantage in football has long been established as an important factor in determining the result of a game" [4]. However, he also mentions that home advantage varies from country to country and it has declined in the major leagues in Europe over the last 15 years. Some of the factors he considers in the paper are crowd effects, referee bias, rule factors, travel effects, familiarity, and more. Lastly, he states that 20 years after his paper "Home advantage in soccer: a retrospective analysis" [3], his conclusion of "there is still much to be learned about the complex mechanism that causes home advantage" still holds.

In a study by [5], the home advantage in Australian football between 2005-2012 is investigated. Poisson regression analysis is used to calculate the effect of home-team crowd support and away-team travel distance. The researchers found that home advantage increases with increased crowd size up to 20,000 persons. Also, in competitions where time zones are greater, jet lag of the away team could be a greater home advantage than crowd support for the home team.

Lgeaz-Arrese et al. [6] mention that previous studies identify five main reasons for home advantages in sports: crowd, familiarity, travels, rules, and territory. While conducting this research, they have a physiological and behavioural point of view. Researchers conclude that not a single factor but a number of factors combined and interacting with each other, such as influence of the crowd, travel fatigue, territoriality, behavioural states of coaches, referees, and athletes, are effective on home advantage. Even some less explored factors, such as competition pressure, athletes' salaries, and even ticket prices could impact the home advantages according to Legaz-Arrese et al.

The research Carron Paradis [7] conducted confirms the factors that affect home advantage

mentioned by Pollard [4], such as crowd support, game location, psychological states, and territorial effects. However, as Carron Paradis mention in their paper, there are certain factors such as the team quality that moderate the effect of the home advantage. Also, they state that they have found mixed results after analysing the crowd support factor. To be more specific, absolute crowd size and crowd behaviour (such as booing or cheering) are generally unrelated to the home advantage, whereas, crowd density is positively related to home advantage. Carron Paradis also mention that laboratory studies show that home crowds have a positive influence on officiating decisions, meaning that home teams receive more favourable calls than away teams.

There are also several research papers where machine learning is used to predict football match outcomes. Ganesan Murugan [8] uses data from football.co.uk to predict English Premier League games using three different Machine Learning algorithms: Logistic Regression, Support Vector Machines, and XGBoost. Then they select the best performing algorithm that gives them the target label, which they refer to as "Full-time Match Result (FTR)". The authors state that they got inspired by Microsoft's search engine, Bing, which accurately predicted each and every knockout stage match outcome of FIFA's 2014 world cup with an accuracy of 100% [9]. In the pre-processing stage, to simplify the dataset, the authors calculate the Scatter Matrix to see how much an attribute affects the others. The data can then be split and used to build the different models. Finally, even though the authors state that the accuracy of their model is fairly accurate, they do not give much more detail or any other results of the research.

Another research that uses Machine Learning for football prediction is conducted by Tax Jousta [10]. The researchers use a self-made dataset containing 13 seasons of the Dutch Eredivisie league. On this dataset, researchers test several reduction techniques and classification techniques. The best results were achieved with PCA (15% variance) with a Naive Bayes or Multilayer Perceptron classifier. Predicting the number of goals is less accurate than predicting win/lose/draw. Also, the researchers state that even though home advantage plays a significant role in predicting, the effect of home advantage depends on the leagues. In conclusion, the highest accuracy is obtained by using Naive-Bayes and Multilayer Perceptrons (Neural Nets), which achieved an accuracy of 54.7%. Also, a combination of LogitBoost and ReliefF on the hybrid model, which makes use of public data and betting odds features, increases the accuracy to 56%.

4 Data Inspection & Preparation

4.1 Data Inspection

We used a database from Kaggle [1] for our paper. As the research only uses supervised machine learning methods, the dataset needed to have the labels for each instance as well. Therefore, only the 'test' set provided by Kaggle is used in this paper. The dataset contains more than 150,000 world football matches between 2019 and 2021. These matches are taken from more than 860 leagues and 9,500 teams. The provided features are divided into two parts: descriptive and historical features. While historical features contain information of the ten previously played games from the home and the away team, descriptive features contain descriptions of a match in a particular moment.

There are three target results in the dataset, which are: 'home', 'draw', and 'lose', as demonstrated in Figure 1. In the dataset, there was a small class imbalance: 43.4% of the instances are labelled 'home', 31.7% are labelled 'away' and 24.9% are labelled 'draw'.

[1] https://www.kaggle.com/c/football-match-probability-prediction/data

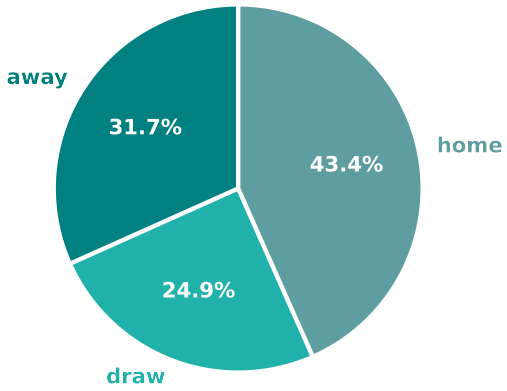


Figure 1: Division of the classes in the data set.

4.2 Data Preparation

The first step in the data preparation phase was the removal of irrelevant features which include:

- Team_history_match_date(i) (features 1 and 2 apply for home team and away team. (i) represents the (i)th match which ranges from 1 till 10)
- Team_history_league_id(i)
- Team names
- Match date
- League name and League_id

The reason for omitting these features is that they do not have an important impact on the final results. Furthermore, removing the unwanted data yields a smaller yet useful dataset. The remaining features contained missing values. The decision was made to replace these missing values by meaningful ones while they are not expected to occur in production. In order to tackle this problem, the features were split into two parts: categorical features and numerical features. Because the word Messing does not occur in the dataset, it was chosen to replace the null values for all the categorical features. On the other hand, the numerical features were approached in a different manner and that is:

- For opponent rating related features, the median of all the opponent ratings (the median of all the medians of the features) replaces the null values. The same approach was used for home team goal related features. This value was chosen because it considers only opponents' ratings/goals regardless of the team's ratings/goals.
- For home team rating related features, the median of home team's ratings per row was used. The same approach was used for home team goal related features. This value was chosen because it considers only the home team's ratings/goals.

- For away team rating related features, the median of away team's ratings per row was used. The same approach was used for away team goal related features. This value was chosen because it considers only the away team's ratings/goals.
- All the rows that contain only missing values were dropped from the numerical and categorical feature. This method was chosen because such rows do not provide enough information to conclude the row-based medians such as the median for home team rating related features.

After replacing the missing values, the categorical features were turned into numerical features using one-hot-coding and then merged with the numerical ones. This approach was, however, not applied to the coach_id related features. The reason for this is that applying one-hot-coding on coach_id is problematic because it contains many unique values. In the next step, the data was split into labels and features. Furthermore, the features were fitted into a Min Max scaler and then transformed into normalized features. Since testing multiple times on the test dataset is considered as a bad practice, the processed dataset was split into a training, a test and a validation set. Although the data has been cleaned and split, the training data (after splitting) was still suffering from class imbalance. To solve this problem, the minority classes in the target feature were over sampled in such a way that the occurrence of all classes is equal. In the last step PCA with 4 components was applied to the numeric features, and the resulting components were normalized. Note that PCA was not applied to the data that was fed to the artificial neural network, since this model type is generally powerful enough to make good use of the full dataset.

5 Methods

5.1 kNN

K-Nearest Neighbours (kNN) is a supervised machine learning algorithm that can be used to solve classification and regression problems. The symbol 'k' denotes the number of nearest neighbours to an unknown variable that has to be predicted. When running the kNN algorithm, different values for 'k' must be tried out. This might lead to high computation cost, but it is a fairly simple algorithm. In addition to this, the similarity measure between two data points can also be seen as a drawback [11].

5.2 Random Forests

The random forests method is an example of a supervised machine learning algorithm that uses the ensemble technique to make predictions. The method applies bagging (i.e., bootstrap aggregating) to train an ensemble of decision trees and then making the prediction based on the majority vote. The method can be used for both regression and classification. For classification, the majority vote is determined using the mode and for regression, the mean is used. Where decision trees are known to be prone to overfitting, this is mitigated by the random forest algorithm because of its bootstrapping and ensemble scheme [12].

5.3 Neural Networks

Artificial neural networks (ANNs) are a machine learning technique loosely modelled on the human brain. The basic building blocks of these networks are neurons and the weighted connections between them. Neurons are most commonly arranged in layers, where members of one layer are not connected to one another, but only to members of the previous and following layers. The first layer of a network

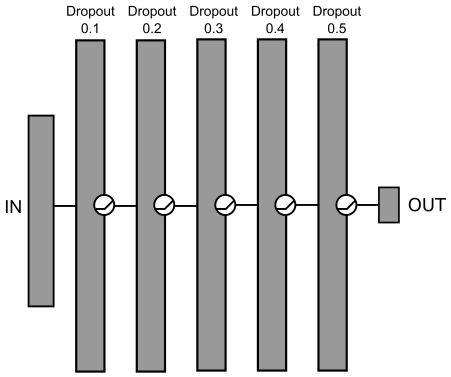


Figure 2: Layer structure of the artificial neural network.

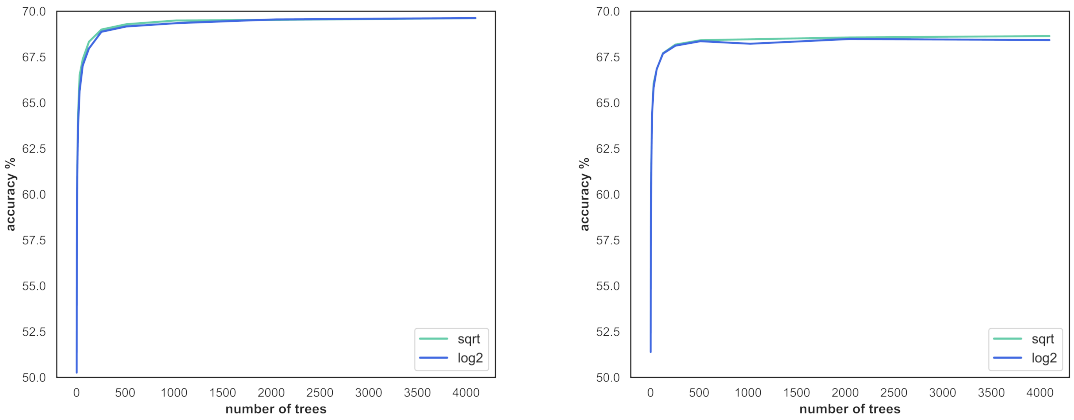
is known as the input layer, the final layer is named the output layer, while all layers between the former and the latter are commonly known as hidden layers. ANNs in which information flows in one direction, from the input layer, through the hidden layers and towards the output layer are known as feedforward networks. Layers within which each neuron is connected to all neurons of the previous layer are named fully-connected layers. ANNs were chosen because of their ability to make proper use of high-dimensional data in machine learning tasks, including classification [13]. The structure of the network can be seen in the diagram below. It consists of 5 fully-connected layers of 1,000 neurons each. The hidden layers use ReLu activations and the output layers use a softmax activation. The optimizer used is Nadam, a version of the Adam optimizer that implements Nesterov momentum. After each layer, an increasing amount of dropout is applied, in an attempt to limit overfitting. This starts as 10% after the first layer and grows in steps of 10% each to 50% after the fifth layer. For the loss function, categorical cross-entropy was used.

6 Results

As a baseline for comparing our results, a decision tree classifier model was used. Such a model consists of nodes that carry a condition. Based on that condition, the dataset is split into several sections. This is repeatedly done to each section until a stop condition is reached. In our case, we limited the maximum depth of the tree to 6 in order to combat overfitting, as that worked the best on our validation set. This model achieved an accuracy of 0.4011 on the test data when running with the "play home" features, and an accuracy of 0.4178 when those features were withheld.

6.1 kNN

The kNN method has one major hyperparameter, namely the n_neighbors. This allows for playing with the number of k's to use for training. When looping through 100 values for k, we found that the accuracy on the validation set was the highest at k=1, with an accuracy score of approximately 0.63. One reason for this could be that the training data and test data are the same. However, this was not the case for us. On the test set the accuracy was 0.35 at k = 1. When we get rid of the features regarding home-play, the accuracy on the validation set was the highest at k = 1, with an



(a) Home features were included.

(b) Home features were withheld.

Figure 3: The accuracy of the Random Forests models on the validation set.

accuracy of approximately 0.64. On the test set, the accuracy score was 0.37. However, the accuracy scores stayed the same after a certain amount k's and did not fluctuate much. Therefore, it is safe to say the kNN method is too underpowered for the complexity of our task.

6.2 Random Forests

The random forests method has several hyperparameters, but for this paper only the number of trees and the maximum number of features are taken into consideration. For the maximum number of features, the square root and the binary logarithm are considered. Similar to Oshiro et al. [14], the number of trees considered are from 2 to 4096, doubling the number at every iteration. In their research, Oshiro et al. found that "a larger number of trees in a forest only increases its computational cost, and has no significant performance gain". The results on the validation set confirm this conclusion as Figure 3 shows that there is no significant increase of the accuracy when the number of trees is larger than 256. Furthermore, Figure [fig] also shows that there is no significant difference between the usage of the binary logarithm (log2) or the square root (sqt) for considering data with and without home features. However, Figure 3a shows that the square root as indication for the maximum numbers of features gives a slightly better accuracy than the binary logarithm. Therefore, the choice was made to use 256 for the number of trees and the square root for the maximum number of features. This resulted in an accuracy of approximately 0.3931 for the model that takes the 'home' features into consideration and 0.3848 for the model where these 'home' features were withheld.

6.3 Neural Networks

The ANN described in section 4.3, after 5 epochs of training, displayed a categorical accuracy of 0.4667 when provided with the full feature set. Withholding the features regarding home-play resulted in the same model performing with a very similar categorical accuracy of 0.4420. For a 3-class classification with low class imbalance, both of these values are better than random guessing,

but they are quite far off the performance achieved by other models [15]. The difference between the two runs, while non-zero, is not pronounced enough to be significant.

7 Conclusion

	kNN	**Random Forests**	**Neural Networks**
Home features included	0.3525	0.3931	0.4667
Home features excluded	0.3704	0.3848	0.4420

Table 1: Accuracy of the models on the test data.

After choosing the individual models, the test data was run on the final models chosen for the three methods. From Table [table], it is visible that the neural networks performed slightly better than both kNN and random forests. For each of the models, however, the accuracy was higher than when the class would have been predicted by random guessing. Before the different models were trained, the baseline was set using a decision tree with a maximum depth of 6. This gave an accuracy of 0.4011 when running with the "play home" features, and 0.4178 without these features. Only the models that were trained using the neural networks method had better accuracy. The research question stated in the introduction of the paper was whether there is a significant home advantage in football. To investigate home advantage, the decision was made to train models with the features regarding which team plays at home and which team plays away included as well as models where these features were excluded. As Table 1 shows, the accuracy on the models where these features regarding playing at home were included was slightly better than where these features were withheld for both random forests and neural networks. For kNN, on the other hand, the model where the features were excluded gave a better accuracy. The difference between the accuracy of the two models for each of the methods, however, was at most 0.0247 (neural networks). Based on the current results, it is not possible to conclude that there is a significant home advantage in football.

8 Further Research

As previously discussed and as can be seen in 1, the differences in performance were quite small when the "play home" features were withheld. In some cases, the model performs better without the features, which was not expected, although not by a large margin. Since our overall accuracy was not high, we speculate that the model was not able to infer the meaning of the features. We do not know whether this is caused by an underpowered model or by a lacklustre dataset and finding out would involve sampling another dataset. Therefore, there is room for further research into this topic, preferably starting from a model with high predictive performance. A possible research question would be: Does a model with high predictive accuracy for football outcomes experience worse performance when information about playing at home is withheld?

References

[1] UEFA. (2021, Jun.) Abolition of the away goals rule in all uefa club competitions. UEFA. [Online]. Available: https://www.uefa.com/returntoplay/news/026a-1298aeb73a7a-5b64cb68d920-1000–abolition-of-away-goals-rule-in-all-uefa-club-competitions/

[2] R. C. Thomas, S. and S. Davies, "An analysis of home advantage in the english football premiership," *Perceptual and Motor Skills*, vol. 99, p. 1212–1216, Dec. 2004.

[3] R. Pollard, "Home advantage in soccer: a retrospective analysis," *Journal of Sport Sciences*, vol. 4, pp. 237–246, 1986.

[4] ——, "Home advantage in football: A current review of an unsolved puzzle," *The Open Sports Sciences Journal*, vol. 1, pp. 12–14, Jun. 2008.

[5] C. Goumas, "Home advantage in australian soccer," *Journal of Science and Medicine in Sport*, vol. 17, pp. 119–123, Jan. 2013.

[6] A. Legaz-Arrese, D. Moliner-Urdiales, and D. Munguía-Izquierdo, "Home advantage and sports performance: Evidence, causes and psychological implications," *Universitas Psychologica*, vol. 12, pp. 933–943, Aug. 2013.

[7] A. Carron and K. Paradis, *The Home Advantage*, 01 2014, pp. 351–355.

[8] A. Ganesan and H. M, "English football prediction using machine learning classifiers," *International Journal of Pure and Applied Mathematics*, vol. 118, pp. 533–536, 2011. [Online]. Available: https://acadpubl.eu/hub/2018-118-22/articles/22a/79.pdf

[9] W. Sun. (2014, Jul.) Predicting the beautiful game: How bing predicts did. Microsoft Bing Blogs. [Online]. Available: https://blogs.bing.com/search/2014/07/13/predictsfinal

[10] N. Tax and Y. Joustra, *Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach*, 09 2015.

[11] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for knn classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, jan 2017. [Online]. Available: https://doi-org.vu-nl.idm.oclc.org/10.1145/2990508

[12] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues(IJCSI)*, vol. 9, pp. 272–278, Sep. 2012.

[13] M. Zekić-Sušac, S. Pfeifer, and N. Šarlija, "A comparison of machine learning methods in a high-dimensional classification problem," *Business Systems Research Journal*, vol. 5, no. 3, pp. 82–96, 2014. [Online]. Available: https://doi.org/10.2478/bsrj-2014-0021

[14] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 154–168.

[15] S. G. Hubacek, O. and F. Zelezny, "Exploiting sports-betting market using machine learning," *International Journal of Forecasting*, vol. 35, pp. 783–796, 2019.