Predictive modelling of strikers' goals home/away in the Premier League

Saacid Mohamed Ali
150302698
Joshua Reiss
MSc Computing and Information Systems

***Abstract*—Association football is arguably the most popular sport in the world. Recently, data is playing a key part in the sport. Many research papers have been done in the past which aimed to make football predictions, but they mostly focused on predicting match results or goals. This paper used data on strikers to make machine learning algorithms to predict strikers' goals home and away. The predictive model which worked best was linear regression followed by random forest.**

***Index Terms*—Predictive Modelling, Machine Learning, Home Advantage, Strikers**

I. Introduction

Association football (henceforth referred to as football) is arguably the most popular sport in the world with approximately 250 million players in 200 countries and an unprecedented number of viewers in every corner of the world (Top 10 Most Popular Sports in The World - Sports Show, 2020). A staggering 1.12 billion people tuned in to watch the 2018 FIFA World Cup final (More than 3.5 billion people watched 2018 World Cup, says Fifa, 2018).

Football is a financially lucrative sport with teams receiving astronomical sums for winning competitions and players receiving substantial wages and bonuses for their participation. Aside from prize money, teams also generate revenue through other means such as deals with kit manufacturers, commercial sponsorships, matchday revenue and broadcasting revenue. For these reasons, winning matches is of paramount importance to teams and to win, teams must outscore their opponents.

A. Importance of Strikers

The main role of the striker in a team is to score goals and spearhead the team's attack. Team's often pin their hopes on their striker to score the winning goal in a tight match or to pull them level when they are losing. A team's tactic is often centred around the striker to create chances for them to score and to put them in positions to succeed. Due to the influence strikers have on matches and on tactics, teams are willing to spend audacious amounts to secure the services of a potent goal scorer in hopes of winning more games and in turn winning trophies. 3 of the top 5 most expensive signings of all time are strikers, which speaks to their importance.

B. Data in Football

In recent years, the use of data in football has increased heavily and now plays a part in the decision making process for football teams. Previously, the data being used in football was limited to stats on shots, goal, corners, passes etc (Sumpter, 2019.). As technology has evolved, so have the stats in football. Opta Sports was one of the pioneer's in the evolution of stats in football as they first began tracking the location of shots, passes, dribbles and tackles at first. Subsequently, they introduced expected goals – 'a system for calculating the likelihood of any shot being scored, based on its distance and angle from goal' (Burn-Murdoch, 2018). Before signing new players, scouts will examine the data on potential signings. Prior to facing an opponent, coaches will analyse the data on opposing teams and will take that into account when deciding on which tactic to employ. However, the data being recorded is not just limited to matches; many teams have their own unit of data analysts that analyse and track player performances in training via sensors. They record data on a player's vitals such as distance covered, fitness levels and any other metric that gives 'coaches, managers, and physios the information they need to improve performances and fine-tune the team's strategy' (Murray, Lacome and Lacome, 2019). The influence data now has on football cannot be understated, it plays a vital role in player recruitment, match analysis and athlete monitoring (Football Team Performance- Stats Perform, 2020).

C. Home Advantage in Football

Home Advantage has long been a talking point in football. When two teams of the same calibre face one another, the home team is likely to win more than 50% of the time (Home advantage in football – what can the data tell us? - Football Perspectives, 2020). The team that is home has the distinct advantage of playing in front of a partisan crowd, which applauds the efforts of the home team and is quick to drown the stadium in a chorus of boos at the slightest indiscretion by the opponent (Analysis: Why Is Home Advantage So Important In Football?, 2016). In the German Bundesliga, this season (19/20), home teams won 43.3% of games at their own ground prior to the stoppage in play caused by the COVID-19 virus. However, after the resumption of play, home teams were only victorious 21.7% of the time. This was a small sample size, but similar statistics were recorded in other European Leagues (Evans, 2020) Home crowds are said to influence how referees officiate matches and the mentality

of players, applying an added layer of pressure (Ahmed and Burn-Murdoch, 2020).

This paper aims to demonstrate whether home advantage has a positive/negative impact on the performance of strikers. Furthermore, different models will be used to predict the goal scoring rate of strikers and a comparison will be made between the models determining the most effective one.

The remainder of this paper is organized as follows. Section 2 explores the significant literature and themes relevant to this paper such as home advantage in sports, machine learning and statistical modelling and an insight into relevant studies. Section 3 outlines where the dataset was obtained, the methods involved in processing the data and the feature selection process. Section 4 details the results obtained by the machine learning algorithms. Section 5 provides a conclusion to the paper and outlines the limitations of the study and presents possible future extensions.

II. Literature Review

Machine learning is the use of computational methods to increase the accuracy of predictions or improve performance. It consists of creating accurate and efficient algorithms. These computational methods use previous information, usually in the form of electronic data, in their learning process. In most cases, the more data there is available, the easier the task and the more experienced the algorithm. However, the quality of the data is also important in machine learning. A low quality dataset will lead to an inaccurate machine learning algorithm. Due to the use of data, machine learning is closely linked to data analysis and statistics (Mohri, Talwalkar and Rostamizadeh, 2018). According to Mohri et al a few problems which can be faced with the use of machine learning include:

- Computer vision applications such as face recognition
- Natural language processing
- Speech processing applications such as speech recognition
- Text or document classification which includes spam detection

1) Supervised vs Unsupervised learning: In machine learning there are two models and they are supervised and unsupervised learning. Unsupervised learning is when the algorithm is provided data which has not been labelled and it has to make sense of the data. The input features X are provided but the labels Y are not. The algorithm aims to learn without the guide of a teacher. In supervised learning, the algorithm is fed labelled data and it makes a prediction based off this data (Mohri, Talwalkar and Rostamizadeh, 2018). Supervised learning is often used when a prediction is required from a given input (Müller and Guido, 2016). The data (the training data) which is fed to the algorithm is a set of inputs and it is the role of the algorithm to makes sense of this data to produce and accurate prediction. The algorithm should also be able to make rational predictions for inputs that were not met during the learning process. It should also be able to deal with small inaccuracies in the data which is known as noise. The goal of supervised learning is to predict a class or output. Supervised learning algorithms will be used in this paper as the data will be labelled. The two types of supervised learning algorithms are regression and classification algorithms (Marsland, 2014). Common supervised learning algorithms include some of the following:

- K-Nearest Neighbours
- Decision Trees and Random Forest
- Linear Regression
- Logistic Regression

2) Regression vs Classification: Supervised learning can be broken into two categories, regression, and classification algorithms. Classification algorithms are used when the prediction output required a discrete value such as true or false, male or female. Regression algorithms are used when the prediction output required is a continuous value such as price, salary, height, and weight. Examples of when regression algorithms are used is when house prices are being predicted. The model will be fed past data and will be trained on this data, and on the completion of the learning process, the model can easily predict house prices (Murphy, 2012).

A. Machine Learning Algorithms

In this paper, regression algorithms will be used as the predicted output, goals scored, is a continuous value. The following regression algorithms will be utilised in order to achieve the aims of this project:

- Decision Tree
- Random Forest
- Linear Regression

1) Decision Trees: Decision trees are a popular machine learning algorithm that can be used for both regression and classification. It is a tiered system of if/else questions which lead to a decision. The root node is at the top of the tired system and it is the first question which splits the sample into different sets. A node which splits into further sub-nodes is called a decision node and leaf/terminal nodes are ones which do not split. With each instance in a decision tree, we start at the root node and follow each decision node until we reach a leaf node. This process is repeated for each instance in a sample.

In most cases, when designing decision trees, it is best to make the decision tree as simple as possible as it will be easier to understand. Also, the complexity of the decision tree has an impact on its accuracy. When assessing the complexity of a tree, one of the following metrics are usually assessed: the total number of leaves, number of features used, tree depth and the total number of nodes. Whilst decision trees provide distinct advantages such as being unaltered by irrelevant features or missing values and being easily readable and easy to follow, it does have its drawbacks such as being susceptible to overfitting. Overfitting is when an algorithm fits the training data extremely well but falls victim to being unable to generalize to data which was not introduced during training. Instead of learning, the algorithm memorises the training data (Maimon and Rokach, 2008).

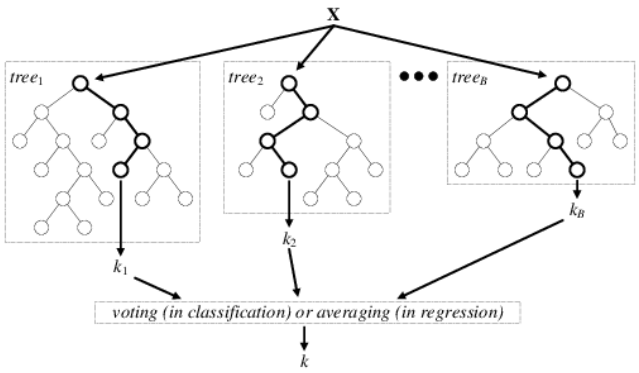

Fig. 1. Random Forest Algorithm

2) Random Forest: Random forest algorithm is similar to decision trees and can also be used for regression and classification. Random forest is the large generation of multiple decision trees which each use a random selection of features at each node. Once the trees are all created, the final output is obtained by using the mode of outputs in each decision tree for classification or the mean of the outputs in regression. The technique is illustrated in Fig. 1. Random features are used during the creation of the decision trees as the randomness improves the accuracy of the model and the creation of multiple trees helps minimise the danger of overfitting which is evident in the decision tree algorithm (Breiman, 2001).

3) Linear Regression: Linear regression is a model which aims to find a relationship between one or more dependent variables and the predictors (independent variables). Simple linear regression is used to find a relationship between a dependent variable x and an independent variable y. Multiple linear regression is used to find the relationship between one dependent variable and multiple independent variables.

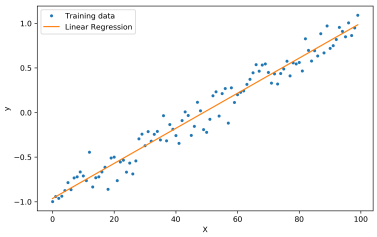

Fig. 2. Linear Regression Algorithm

As multiple linear regression maps the relationship between a dependent variable and multiple independent variables as opposed to simple linear regression, more issues have to be considered such as variance inflation, influential observation, collinearity, and detection of regression outlier. The three main purposes of linear regression are:

- Determining if there is a relationship between the dependent variable and the independent variables
- Predicting the value of the dependent variable based on the independent variables
- Analysing the independent variables to see which variables have a greater influence than others to explain the dependent variable so that the casual relationship can be determined more efficiently and accurately (Yan and Su, 2009)

B. Python

Python is general purpose programming language which is now the commonly adopted programming language used by data scientists. A number of things can be done with the use of Python such as data loading, visualisation, analysis, image processing and many more due to the diverse number of libraries that are available in Python. In this paper the following Python libraries will be used: Pandas, Matplotlib and Scikit-learn. Machine learning and data analysis are iterative process and therefore it is essential that the tools being used allow easy interaction and quick iteration. That is one of the many advantages of Python as it gives users the ability to interact with the code directly and provides instant feedback (Müller and Guido, 2016).

1) Pandas: Pandas is a library in python which makes working with data fast and easy as it provides which rich data structures and functions. It makes use of the data manipulation capacities of relational databases and spreadsheets and combines this with the array-computing capacities of NumPy, a library available in Python. Pandas smooths the process of selecting subsets of data, slicing and dicing data and performing aggregations due to the complex indexing functions it provides. The main Pandas object which will be used in this paper is the DataFrame, "a two-dimensional tabular, column-oriented data structure with both row and column labels" (McKinney, 2017).

2) Matplotlib: Matplotlib allows users to easy create simple visual plots such histograms, bar charts and scatter plots with a few commands. John D.Hunter originally created the Python library and it is now preserved by a team of developers. It provides a number of features for users (Barrett et al., 2005) such as:

- Interactive navigation which allows users to zoom in on sections of the plot
- Many predefined line styles and symbols
- Multiple axes and figures per page
- Date and financial plots

3) Scikit-learn: Scikit-learn is a popular machine learning library available in Python which allows users to utilise many well-known machine learning algorithms including regression, classification, clustering, and dimensionality reduction. It builds on the popular Python libraries NumPy and SciPy (Hackeling, 2014). There is a growing need amongst non-specialists in the software and web industries along with other industries outside of the computer-science, such as biology and physics, which require statistical data analysis more frequently. Scikit-learn makes it easier for these non-specialists to carry out statistical data analysis without having the prerequisite knowledge (Abraham et al., 2011).

C. Home Advantage in Sports

Home advantage has played a major role in sports such as football, basketball, and baseball for an incredibly long time. Courneya and Carron (1992) described home advantage as "the consistent finding that home teams win over 50% of the games played under a balanced home and away schedule". Over the first 20 seasons of the Premier League, a staggering 60.8% of the total points were won by home teams. Teams picked up an average of 32.0 points per season in home games and an average of 20.7 points per season in away games. This clearly shows that there is a home advantage in effect (Allen and Jones, 2012).

Home teams are said to have an advantage due to some of the following reasons:

- Crowd Effects
- Referee Bias
- Familiarity
- Special Tactics

1) Crowd Effects: A large home crowd can have a psychological effect on players and the way they perform. They can create a daunting/hostile environment for the away team making it difficult to play well. This atmosphere can also influence the decisions made by referees. A partisan home crowd can also motivate players to increase their energy and effort as the home team want to impress their fans and give them a sense of pride with a good performance and ultimately a win (Carmichael and Thomas, 2005).

2) Referee Bias: Research has found that a partisan home crowd has an effect on the decision made by a referee. A study by Nevill et al (2002) found that referees where more hesitant when viewing challenges and awarded fewer fouls to away teams due to the influence of the home crowd. They also found that when there is an ambiguous challenge which is difficult for the referee to assess, they may rely on the "salient yet potentially biased judgement of the crowd" when making their decision.

3) Familiarity: Home teams enjoy an advantage due to the familiarity aspect that comes with playing at home. They are playing at a familiar stadium in familiar conditions. When a team moves to a new stadium, there is a drop in their home advantage as they are not familiar with their new surroundings yet (Pollard, 2008).

4) Special Tactics: Home and away teams tend to use different approaches from a tactical standpoint when viewing upcoming games. Pollard (2008) found that if away teams were to set up with a more wary and defensive game plan, this would be handing the home team a psychological advantage before the game has even begun. However, there is no established evidence to prove that there is a link between tactics and home advantage.

D. Past Studies

The problem of football match prediction has been tackled by many in academia and industry due to its economic importance and interesting nature. Previous research falls into two major categories which are goal-based studies, which aim to predict the goals scored by each team, and results-based approaches, which aim to predict the results of matches (Baboota and Kaur, 2019).

Statistical modelling with the use of football data has been a theme since the middle of the 20th century with the focus initially being on predicting the number of goals scored in a game. Moroney (1956) used Poisson distribution in efforts to predict the number of goals scored by a team but found that this model was not well fitted to make such a prediction. He did suggest however that a "modified Poisson" (the negative binomial) could be used and it would provide a much better fit. Reep et al (1971) used the negative binomial distribution to analyse English Football League First Division data for four seasons and found that despite the improved fit of the model, they identified that 'chance dominates the game' and that it was impossible to avoid the ingrained noise in the observed data with the models they used.

Despite the conclusions reached by previous studies, Hill (1974) confirmed that predicting match results could be done with past data and statistical modelling. In his research, Hill compared the final league table of the 1971-72 football season with the predictions made by experts prior to the start of the season and found a significant positive correlation which presents the notion that match results are not merely based on chance and it is skill that dominates the game.

The first development came from Maher in 1982 where he used Poisson distribution which assessed the quality of teams' attacks and defences in home and away games and used this to predict the mean number of goals for each team. The findings were that this model gave a reasonably good fit to the data as there were only slight deviations from the model. Maher also used a bivariate Poisson model, and this improved the fit considerably. Dixon and Cole built on this model in 1997 and developed a method for estimating the probability of results and scores. The Dixon and Coles model is based on a Poisson regression model, which creates goal probabilities by transforming the expected goals for teams with Poisson distribution. The goal probabilities is in turn transformed into score probabilities and finally into match outcome probabilities.

Researchers began to move away from predicting match scores by using match outcome probabilities at the beginning of the 21st century and instead began to predict match results (win/draw/loss) directly. Forrest and Simmons (2000) were one of the first to predict match results directly using a classifier model instead of predicting the goals scored by teams. Due to the progressions in technology and the increased availability of football data, researchers began to adopt the use of modern machine learning algorithms when creating predictive models.

Goddard (2005) concluded that the forecasting performance of the two approaches that were used to model matches in the past, modelling the goals scored and conceded by teams and modelling match results, were never tested or there were never comparisons made between the two models. Goddard used or-

dered probit regression to predict match results with the use of 25 years of results data on English league football. This paper was one of the first to use other variables than match results, as Goddard also used few explanatory variables explanatory variables such as the geographical distance between teams and match significance.

More recently Prasetio and Harlili (2016) used logistic regression to predict the match results for Barclays Premier League matches. They also wanted to determine what the significant variables are to win matches. Instead of deciding what the significant variables are, they used they significant variables gathered from researches in the same field. Some of the significant variables they used included:

- Travelled Distance
- Ground Familiarisation
- Home Advantage
- Shots on goal
- Attack
- Defence

They concluded that the most significant variables are "Home Defence" and "Away Defence" but stated that prediction cannot be done with these two variables alone. Praestio and Harlili also found that using only significant variables can increase prediction accuracy.

Hucaljuk, and Rakipović (2011) examined multiple machine learning algorithms in order to achieve the best prediction results when predicting football scores. The optimal combination of features and classifiers were determined by the different algorithms. They concluded that feature selection can be tackled one of two ways. The first method implies that due to the lack of knowledge about the problem being faced, all features which could affect the outcome are selected and then those which are determined to have the greatest impact are selected. The second method implies that there is knowledge about the problem and the features selected are based on this knowledge and they are believed to affect the outcome the most. The following algorithms were used to predict football results:

- Naive Bayes
- Bayesian networks
- LogitBoost
- K-nearest neighbours algorithm
- Random forest
- Artificial neural networks

Hucaljuk, and Rakipović achieved their best results when using Artificial Neural Networks (ANN) with a prediction accuracy of up to 68%. This study is relevant to this paper as it aims to use different machine learning algorithms to predict striker's goals and identify which algorithm provides the best results.

E. Research Gaps

When reviewing the literature, several gaps were identified which were not previously studied. Although Courneya and Carron (1992) and Nevill et al (2002) have performed research on home advantage and the effect it has on teams, there has been a lack of studies which investigate the impact home advantage has on individual positions in different sports. This paper seeks to extend their research by focusing on the effect home advantage has on the performance of strikers. Another gap was that prior studies concentrated on either predicting the outcome of matches such as the studies by Prasetio and Harlili (2016) and Hucaljuk, and Rakipović (2011) or on predicting match scores such as the studies by Dixon and Cole and Maher. Previous studies have not focused heavily on predicting the output of individual positions in sports.

F. Research Aims

The gaps identified in the literature have to led to the following aims of this research:

- To critically assess whether home/away games have an impact on the performance of strikers
- To identify if referees officiate strikers differently in home/away games.
- To use different predictive modelling techniques to make predictions as to how many goals a striker will score home/away and testing the accuracy of these models.

III. Research Methodology

The dataset being analysed was collected from github and was available at the following link: https://github.com/vaastav/Fantasy-Premier-League. The Data was readily available online and was easy to download meaning that the data retrieval process was hassle free. The dataset used was a fantasy premier league library that provided all the basic stats for each player as well as gameweek specific data and seasons history for each player.

- 4 seasons of data ranging from 2016/2017 to 2019/2020
- 2424 player seasons data
- Detailed match events (goals, assists, key passes, errors leading to goals and offsides)

Only 3 seasons will be used as the current season 19/20 was not completed at the time this paper was written and the 19/20 was briefly interrupted by Covid-19 and since the restart of play was done behind closed doors without fans, home advantage has been significantly reduced thus skewing the data.

A. Data pre-processing

Data pre-processing is an important part of the data analysis process as prior to data being used for modelling and analysis, the data must be in a readable format and the data being used must be relevant. Data pre-processing includes removing duplicates, dealing with missing values, transforming the data, and removing redundant data.

1) Removing Redundant Data: Once the data was retrieved, it was realised that there was some redundant data present in the dataset. As the aim of this paper is to analyse the performance of strikers and to predict the goals scored by them, the data available on the other positions such as goalkeepers and defenders is redundant and must be removed from the dataset

prior to use. Also, the data available on fixtures such as the date of games and time of kick-off was removed as well as fantasy premier league data such as how many selections a player had, and the total points they accumulated.

2) Transforming Data: Prior to the data being used, the data has to be converted into the format which was required for the model to work and for analysis purposes. Below are the stages of data transformation process:

- Only the strikers that had played at least a minimum of 1710 minutes (19 games) were selected
- Once this was done, the strikers stats for each game week were totalled and were placed in one file.
- The totals for each striker were split into their totals for home and away games.
- Each striker's totals were converted into per 90 metrics
- Finally, the Home/Away columns were converted into an integer where 0 equals a home game and 1 equals an away game

Per 90 metrics were used in this study as it gives as a more accurate representation of players performances. When looking at surface level stats, it may be observed that striker A has scored the most goals, therefore they must be the best. However, this may not necessarily be the case as they may have scored the most goals due their superior number of minutes played. Another player, striker B, may have only scored a few less goals but they played considerably less minutes than striker A showing that in this case that striker A only scored more due to their increased minutes. For this reason, per 90 metrics will be used as it shows each strikers performance in relation to their minutes played.

3) Removing duplicates: The data was checked for duplicate entries as this would skew the data. Duplicates is a common human error when it comes to inserting data in datasets. Duplicate data was removed using the drop_duplicates() function in Python.

4) Dealing with missing values: Missing data can decrease the factual intensity of a study and can create one-sided estimates, prompting invalid assumptions and decisions. Missing values were checked for by using isnull() function in Python. There was no missing values present in the dataset.

B. Feature Selection

Prior to implementing each model, feature selection must be done. The quality of features used in the model directly influences the performances as shown by previous research . As stated by Hucaljuk, and Rakipović (2011), the features in this model will be selected based on knowledge, as there is not a lack of knowledge associated with the problems being tackled. The following features were selected:

- Was home
- Assists (per 90)
- Big chances created (per 90)
- Big chances missed (per 90)
- Key passes (per 90)
- Offside (per 90)
- Target missed (per 90)
- Winning goals (per 90)

IV. Findings

A. Does home advantage have an impact on the performance of strikers

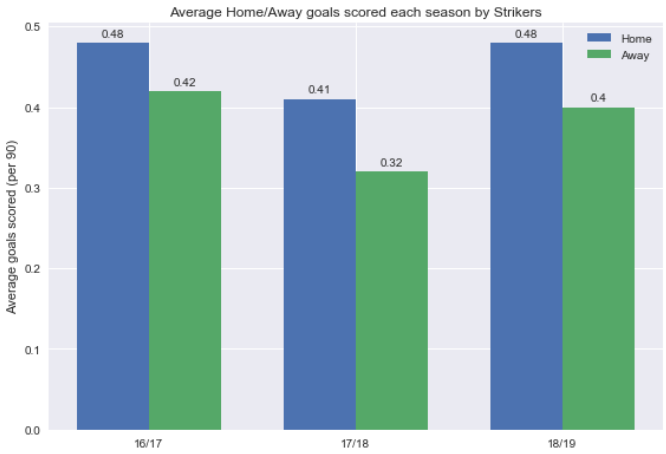

Fig. 3. Average Striker Goals Home/Away

To answer the first research question, does home advantage have an impact on the performance of strikers, we first looked at the goals scored by strikers as this is their primary job. As illustrated in Fig. 3, in each of the last 3 seasons strikers have scored more goals per 90 minutes at home compared to away games on average. Strikers did not just score more goals at home, they also scored more winning goals at home. On average, they scored 0.03 more winning goals per 90 minutes in home games. Whilst strikers did score more home goals in the each of the last 3 seasons, they did miss the target more often in home games and they also missed more big chances at home. They did however create more big chances in home games.

Although a striker's primary job is to score goals, they can still influence games in other ways which include setting up their teammates to score, retaining possession and winning possession for their team and not committing fouls. In each of their last 3 season, strikers provided more assists in home games, meaning their overall goal contribution was higher in home games. Strikers also had more key passes in home games.

Possession can be retained by making accurate passes. The pass completion rate of strikers, which is worked out by dividing completed passes by attempted passes, was nearly identical between home and away games in the last 3 seasons. Another way possession can be retained is by not being tackled by opposition defenders. In 2 of the last 3 seasons, on average strikers lost the ball 0.1 times more in home games but in the one season strikers lost the ball more times in away games, the difference was 0.4 which is much greater than the other 2 seasons.

B. To identify if referees officiate strikers differently in home/away games.

When it comes to assessing the impact of referees officiating of strikers, the 3 things we can assess from the data set are the number of offsides that are flagged and the number of yellow and red cards that are handed out. Strikers were caught offside far more often in home games compared to away games as illustrated in Fig. 4. It is also evident that peak number of times strikers were caught offside in away games is lower than the lowest number of times strikers were caught offside in the last 3 seasons. In each of the last 3 seasons, strikers received more yellow cards in away games but when it came to red cards, strikers received more red cards in home games in 2 of the last 3 seasons. However, there is not a great difference between the red cards received in home and away games.

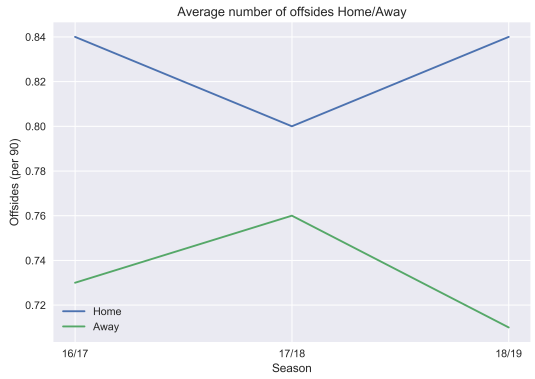

Fig. 4. Average Number of Offsides Home/Away

C. Model Performance

To evaluate the performance of the different models, the following metrics will be use:

- Model Score (R^2)
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

1) Model Score: Model score or coefficient of determination, as it is formally known, is "the proportion of variance explained by the regression model makes it useful as a measure of success of predicting the dependent variable from the independent variables". The output is a number between 0 and 1. The dependent variable can be predicted from the independent variable without error if the R^2 value is 1 and vice versa (Nagelkerke, 1991). The model score of each of the three models is shown in Fig. 5. The model with the best model score is linear regression with a model score of 0.702. Decision tree had the lowest model score with 0.389.

2) Mean Absolute Error: Mean absolute error along with the other 2 error metrics are used to describe the "average model-prediction error in the units of the variable of interest". MAE does not account for the direction of the error, be it positive or negative, and it is the mean of all absolute errors between the predicted values and actual values. Generally speaking, a low MAE number shows that the model making fairly accurate predictions (Willmott and Matsuura, 2005). The model with the lowest MAE number was linear regression with 0.119. Random forest came second with 0.138 and decision tree had the highest MAE number with 0.163.

	Model Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	0.702072	0.119495	0.022494	0.149979
Decision Tree	0.388863	0.163643	0.046141	0.214805
Random Forest	0.572619	0.138364	0.032267	0.179631

Fig. 5. Model Performance

3) Mean Squared Error: Mean squared error is the square of average differences between the predicted values and the actual values. It is calculated by working out the difference between each of the predicted and actual values which is the error. Each error is squared and then the sum is found. Lastly, the sum of squared errors is divided by the number of values to find the mean squared errors. It measures the quality of a model and MSE values which are close to 0 are better (Wang and Bovik, 2009). Again, in this case, linear regression had the lowest MSE number and decision tree had the highest MSE number as evident in Fig. 5.

4) Root Mean Squared Error: Root mean squared error is simply the square root of MSE and is commonly used as in some cases the MSE value may be too large for comparison purposes (Willmott and Matsuura, 2005). The lowest RMSE value of the 3 models was linear regression with a RMSE value of 0.15. Random forest had the second lowest RMSE value with 0.18 and decision tree had the highest RMSE value with 0.21.

5) Test Data Size: Each of the 3 models were first tested with a smaller sample size which was one season's data. All 3 of the models performed poorer with the smaller sample size. The model score dropped in each of the models meaning their predictive ability was worse. Furthermore, the MAE, MSE and RMSE was higher for each model in this case meaning that each of the 3 models were more error prone. With the smaller sample size, linear regression was still the best performing model with decision tree being the worst performing model. When comparing the performance of each of the models to themselves, linear regression and random forest improved considerably with the larger sample size. However, whilst there was in increase in performance in decision tree, the increase was not drastic.

	Model Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	0.546054	0.167732	0.037287	0.193097
Decision Tree	0.282613	0.176566	0.058925	0.242745
Random Forest	0.377017	0.174402	0.051171	0.226210

Fig. 6. Model Performance for One Season

V. Conclusion

As discussed, home advantage is prevalent in sports. Teams have a slight edge over their opponents when playing at home

and this is very much the case in football. However, there is no prior evidence which proves strikers perform better in home matches.

A good starting point for analysing the effect of home advantage on the performance of strikers is to look at the number of goals scored as this is their main responsibility. It is clearly evident that home advantage influences this as strikers scored more goals in home games compared to away games. They also scored more winning goals in home games. The reason for the increased number of goals in home games could be due to tactics used by the home/away team. Pollard (2008) detailed that home and away teams tend to use different approaches tactically in games and that could be the case in this scenario. The reason why strikers may score more goals in home games could be due to their team's tactic being more attack focused. This could also explain why strikers scored more winning goals in home games. When the game is a draw, away teams may change their tactic to a more defensive approach to secure the 1 point as they may view it as 1 point gained rather than 2 pts dropped. The home team may use a more attacking approach to secure all 3 points and they also may be motivated by the partisan home crowd to go and score the winning goal as they have an influence on home teams as conclude by Carmichael and Thomas (2005). Strikers did miss the target more in home games and missed more big chances. This could be due to them having more attempts and attempting more big chances but there is no evidence to prove this.

Apart from scoring goals, it is also important for strikers to retain possession for their team and to contribute to goals. Making accurate passes is one of the ways to retain possession. Playing at home does not seem to have an influence on the pass completion rate of strikers as this figure was practically identical in both home and away games. Furthermore, the number of attempted and completed passes was also practically identical in both home and away games. However, despite the pass completion rate and the number of completed passes and attempted passes being similar for strikers in home and away games, strikers provided more assists and more key passes in home games and created more big chances. An attack minded approach could be an explanation this as in home games a majority of the strikers' passes could be in and around the oppositions box.

Teams enjoy more lenient officiating from referees when playing at home as the referees may be more hesitant when viewing challenges due to the effect of the home crowd as found by Nevill et al (2002). When it comes to analysing how referees officiate strikers in home and away games, the 3 metrics that were assessed were offsides, red cards and yellow cards. Strikers were flagged for offsides far more often in home games compared to away games. This could be due to strikers attacking more in home games and making more runs, however there is no evidence to support this. Strikers did however receive more yellow cards in away games. Referees may have been influenced by the crowd to brandish more yellow cards. When it comes to red cards, the numbers were roughly similar in home and away games. Looking at these metrics, there is not enough concrete evidence to state that strikers are officiated differently in home and away games.

This paper aimed to use different predictive models to predict strikers' goals in home/away matches and to compare their performance. This aim was achieved as 3 models were used to make predictions which were: linear regression, decision tree and random forest. The best performing model was linear regression as it performed better than the other 2 models when looking at 4 different metrics.

A. Limitations and Challenges

The biggest challenge faced in this study was finding suitable data to analyse and to use to build the models. Much of the readily available data is on match outcomes and not on player stats. The datasets which had player stats that were available did not have home and away data. Many of the websites, which had extremely relevant data, had strict terms of use and required an official license. Methods such as web scraping were strictly prohibited.

Whilst the predictive models performed at an adequate level, they could have performed better with the use of more relevant features. Vital data on strikers such as the number of shots, shots on target and off target and the location of shots all could have potentially improved the performance of the models. Furthermore, when it comes to analysing the effect of home advantage on strikers, the use of stats like the ones mentioned above would have enhanced the analysis further.

B. Further Studies

There are many avenues for potential extensions of this research which can be done. One possible extension is to use more data in the predictive models as all the models used in this paper performed better with a larger training data set. The use of more detailed data would also improve the performance of the models. Another possible extension is to analyse the effect of home advantage on other positions in football such as midfielders and defenders or the focus can be placed on other positions in other sports. A third extension would be the inclusion of a team weighted metric. While good results were still achieved with this metric being ignored, future work should look to include a metric which considers the quality of the team a striker plays for to further enhance the performance of the models. It is quite clear that within this area of study, there are plenty of opportunities for further research.

ACKNOWLEDGMENTS

Firstly, I would like to thank Allah (swt) for giving me the strength and the ability to finish this project. Furthermore I would like to thank my supervisor for his guidance. I would like to especially thank my mother. She inspires me with her motivation, desire and dedication to excel in everything she does. My mother is my role model and I would not be in the position I am today without her unconditional love, support and guidance.

REFERENCES

[1] Sports Show. 2020.Top 10 Most Popular Sports In The World - Sports Show. [online] Available at: https://sportsshow.net/top-10-most-popular-sports-in-the-world/.

[2] Independent.co.uk. 2018. More Than 3.5 Billion People Watched 2018 World Cup, Says Fifa. [online] Available at: https://www.independent.co.uk/sport/football/premier-league/2018-russia-world-cup-england-france-croatia-final-fifa-viewing-figures-numbers-a8694261.html.

[3] Sumpter, D., 2019 WHAT DO YOU NEED TO LEARN TO WORK IN FOOTBALL ANALYTICS?. [online] Barça Innovation Hub. Available at: https://barcainnovationhub.com/what-do-you-need-to-learn-to-work-in-football-analytics/.

[4] Burn-Murdoch, J., 2018. How Data Analysis Helps Football Clubs Make Better Signings. [online] Ft.com. Available at: https://www.ft.com/content/84aa8b5e-c1a9-11e8-84cd-9e601db069b8.

[5] Murray, E., Lacome, E. and Lacome, M., 2019. What Difference Can Data Make To A Football Team? - Exasol. [online] Exasol. Available at: https://www.exasol.com/en/what-difference-can-data-make-for-a-football-team/.

[6] Stats Perform. 2020. Football Team Performance- Stats Perform. [online] Available at: https://www.statsperform.com/team-performance/football/.

[7] Football Perspectives. 2020. Home Advantage In Football – What Can The Data Tell Us? - Football Perspectives. [online] Available at: https://footballperspectives.org/home-advantage-football-what-can-data-tell-us/.

[8] Pundit Arena. 2016. Analysis: Why Is Home Advantage So Important In Football?. [online] Available at: https://punditarena.com/football/thepateam/why-is-home-advantage-so-important-in-football/.

[9] Evans, T., 2020. Will Home Advantage Still Be An Advantage?. [online] Independent.co.uk. Available at: https://www.independent.co.uk/sport/football/premier-league/premier-league-fixtures-restart-home-advantage-season-crowd-venues-a9567221.html.

[10] Ahmed, M. and Burn-Murdoch, J., 2020. Does Home Advantage Exist Without Football'S Partisan Fans?. [online] Ft.com. Available at: https://www.ft.com/content/0243c597-c32c-4cfa-8632-097e8393d229.

[11] Mohri, M., Talwalkar, A. and Rostamizadeh, A., 2018. Foundations Of Machine Learning, Second Edition. 2nd ed. [S.l.]: The MIT Press.

[12] Müller, A. and Guido, S., 2016. Introduction To Machine Learning With Python. O'Reilly Media, Inc.

[13] Marsland, S., 2014. "Machine Learning: An Algorithmic Perspective, Second Edition". 2nd ed. Chapman and Hall/CRC.

[14] Murphy, K., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

[15] Yan, X. and Su, X., 2009. Linear Regression Analysis. Hackensack, N.J.: World Scientific.

[16] Maimon, O. and Rokach, L., 2008. Data Mining With Decision Trees: Theory And Applications. World Scientific.

[17] Breiman, L., 2001. Journal search results - Cite This For Me. Machine Learning, 45(3), pp.261-277.

[18] McKinney, W., 2017. Python For Data Analysis: Data Wrangling With Pandas, Numpy, And Ipython. O'Reilly Media, Inc.

[19] Barrett, P., Hunter, J., Miller, T., Hsu, J. and Greenfield, P., 2005. matplotlib – A Portable Python Plotting Package. Astronomical Data Analysis Software and Systems XIV ASP Conference Series, 347.

[20] Hackeling, G., 2014. Mastering Machine Learning With Scikit-Learn. Birmingham, U.K.: Packt Publishing.

[21] Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B. and Varoquaux, G., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12.

[22] Courneya, K. and Carron, A., 1992. The Home Advantage In Sport Competitions: A Literature Review. Journal of Sport and Exercise Psychology, 14(1), pp.13-27.

[23] Allen, M. and Jones, M., 2012. The home advantage over the first 20 seasons of the English Premier League: Effects of shirt colour, team ability and time trends. International Journal of Sport and Exercise Psychology, 12(1), pp.10-18.

[24] Carmichael, F. and Thomas, D., 2005. Home-Field Effect and Team Performance. Journal of Sports Economics, 6(3), pp.264-281.

[25] Nevill, A., Balmer, N. and Mark Williams, A., 2002. The influence of crowd noise and experience upon refereeing decisions in football. Psychology of Sport and Exercise, 3(4), pp.261-272.

[26] Pollard, R., 2008. Home Advantage in Football: A Current Review of an Unsolved Puzzle. The Open Sports Sciences Journal, 1(1), pp.12-14.

[27] Baboota, R. and Kaur, H., 2019. Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), pp.741-755.

[28] Moroney, M., 1952. Facts from Figures. Applied Statistics, 1(1), p.80.

[29] Reep, C., Pollard, R. and Benjamin, B., 1971. Skill and Chance in Ball Games. Journal of the Royal Statistical Society. Series A (General), 134(4), p.623.

[30] Hill, I., 1974. Association Football and Statistical Inference. Applied Statistics, 23(2), p.203.

[31] Maher, M., 1982. Modelling association football scores. Statistica Neerlandica, 36(3), pp.109-118.

[32] Dixon, M. and Coles, S., 1997. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), pp.265-280.

[33] Forrest, D. and Simmons, R., 2000. Forecasting sport: the behaviour and performance of football tipsters. International Journal of Forecasting, 16(3), pp.317-331.

[34] Goddard, J., 2005. Regression models for forecasting goals and match results in association football. International Journal of Forecasting, 21(2), pp.331-340.

[35] Prasetio, D. and Harlili, D., 2016. Predicting football match results with logistic regression. International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)

[36] Hucaljuk, J. and Rakipović, A., 2011. Predicting football scores using machine learning techniques. 2011 Proceedings of the 34th International Convention MIPRO, pp.1623-1627.

[37] Nagelkerke, N., 1991. A note on a general definition of the coefficient of determination. Biometrika, 78(3), pp.691-692.

[38] Willmott, C. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, pp.79-82.

[39] Wang, Z. and Bovik, A., 2009. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. IEEE Signal Processing Magazine, 26(1), pp.98-117.

[40] Vaiciukynas, E., 2020. Figure 6. Architecture Of The Random Forest Model.. [online] ResearchGate. Available at: https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643.

[41] Hansen, C., 2019. How To Do Linear Regression And Logistic Regression In Machine Learning?. [online] Machine Learning From Scratch. Available at: https://mlfromscratch.com/machine-learning-introduction-8-linear-regression-and-logistic-regression//.