Football Match Result Prediction

Student Name: Connor Faulkner Student ID: D00185637

***Abstract*—**
Football is an international business in which passionate fans and companies put hundreds of millions of euros, in the forms of bets, sponsorships, sale of jerseys, etc. The topic of predicting football outcomes has been talked about as lots of companies have interests in the football business. Presently, predicting the results of football matches is done by betting companies, football experts and advanced machines. However, the introduction of new techniques that enable anyone to try and predict football outcomes has increased the potential of beating the betting companies. Football produces a plethora of statistical information on the matches that are played, teams that play each other, players from each team and events that lead to goal scoring opportunities, etc. This data can be used by machine learning algorithms to predict information related to a particular football match, such as, the result of the match, performance of a player in the match and scouting new talents in the matches, etc. In this proposed project, I would be using different algorithms and methods, on a dataset of football matches and then compare the algorithms to see which of the methods is the most suitable to predict football match results.

***Key Terms*- Machine Learning Algorithms, Prediction**

I. INTRODUCTION

Football is one of the most popular sports around the world. It has a huge fan base that follows the day to day events happening within this sport. It is an activity requiring physical prowess and great skill to compete at the highest levels. Football has become one of the largest businesses in the world, having thousands of companies with their primary revenue streams coming from the professional game. Therefore, it has brought on great interests in building predictive and statistical models to find out the results of games, to further drive up profits for these companies. Also, these techniques in question can be used by management of the football associations, in getting their teams prepared for the upcoming matches, to maximize the chance of success in the fixtures [1]. In the professional game, there is an immense number of statistics collected on fixtures and players, from this it is possible to generate several metrics for both team performance and or the individual.

However, to find that advantage in a game, Football has turned to machine learning to transform the huge amount of data into usable insights, such as the goals scored in previous matches by a team, the winning streak of a team and the current form of the players in a team, etc. There have been several attempts that have tried to predict outcomes of football matches, but humans have always been more accurate than machines at predictions. This is because, humans consider the emotional factors and the technicalities that would affect the result of a game, this increases the accuracy in predicting football. The downside is that humans let their emotions cloud their predictions and that can lead to the wrong answer. Machine learning algorithms identify complex patterns that are not perceptible to humans, due to the apparent random nature of the data and do this by reviewing the data by itself. The type of models that are used for this kind of prediction is called multi-class classification, as it has three classes win, loss and draw [2]. Win and loss are easier to classify, however draws are difficult to classify as an outcome.

This project is challenged with developing models that can predict the final outcomes of football matches, using the dataset of football matches in the English Premier League. I will firstly create a dataset of the recent Premier League seasons with features such as home and away goals scored, the forms of the teams either winning or losing forms and the differences of the features. This dataset would then be used by various classifiers (algorithms or methods) such as random forests, k-Nearest Neighbours, Naïve Bayes, Neural Networks, AdaBoost and Support Vector Machine (SVM). The accuracy of the system will vary from classifier to classifier and feature to feature. These features are the parameters that will affect the outcome of a football match. For example, the form of a team in the last few games is especially important feature as for example a team has lost its last four games consecutively, there is a high probability that the team will lose the next game and vice versa. Consideration is needed to choose the correct features as these important features are essential to our prediction model. There is some limitation, for example, I will be using only the English premier league as the training dataset, so tournaments and other leagues would need additional work to support different leagues. Another limitation would be also the selection and classification of features, this will be further developed below. Also, the training of the Classifiers would need to be carefully validated using training datasets. The main contribution of the project is from [3], however special mentions too [2,16].

The remainder of this paper is organized as follows. Section II will be Review of Literature. Section III will give the Method. Section IV will be Results & Discussions, Section V Comparisons of Related Works and finally Section VI The Conclusion.

Problem Formulation

The problems of football match predictions involve selecting factors of the game that contribute to the outcome of the match and how these factors will allow different classifiers to predict the outcomes of matches with greater accuracy. I will use several classifiers which involve machine learning algorithms and data mining techniques to find the more significant features from the data set and then train the different classifiers to predict the outcomes. thus, the set of features are extracted, inputted into the classifiers, the classifiers formulate a pattern from the features by learning from a training set of datapoints. The models are fully trained and ready to predict the winning outcome of the fixture.

Motivation

Football match prediction systems differ in their selections of features and classifiers. By referring the set of features known in literature, I decided upon the features I believe to be most significant. However, considering the classifiers, many researchers have tried different classifiers, So I have used some commonly implemented classifiers for such a problem of predicting football matches to see which would be the best choice when considering accuracy of outcomes.

Proposed Solution

I will be discussing what I believe to be a solution that will compare the accuracy of different classifiers when predicting football match outcomes. I would focus on selecting features and computing accurate algorithms on the features, then run the features through several classifiers. The prediction model would be based features of goal difference, current form and difference in points of the teams.

The current form is calculated, by grouping the teams that are at home or away and finding out the outcome of either team's result. So, for example if a team with the last three games at home are all wins and the away team has lost all its previous away games, then the home team would have a higher probability to win. Also, if two teams with both their current forms all wins, this could impact the prediction differently, so current form is a measure how meaningful the outcomes of current form are. Also, the goal difference is also especially important feature that is used for this project, it is calculated in the same way the current form is, separated into home and away goal difference for each team.

After selection of features and computation, the features are used on the different classifiers. From review of literatures, different classifiers give different accuracies, however, which of the classifiers should be considered the best of the lot.

Scope of the project

This project provides parameters of two teams (Home & Away) within the English premier league and the outcome of the match. These parameters include forms, goal difference, points gained, etc. Analysis on these parameters, which can be used for several purposes such as improvement of the teams in the parameters described. The predicted outcomes can be used for betting, coaching and journalism. Also, further parameter inclusion could increase the accuracy of prediction. Such as individual players attributes or impact of formation, etc, could lead to improvements.

II. REVIEW OF LITERATURE

Since football is popular worldwide and generates data every day, it is not surprising to find large amount of research in football prediction. Most academic research is focused on developing models for specific leagues or tournaments. These different techniques are separated into machine learning, statistical model and a rating system.

Research classified as machine learning are as follows: Laura Hervert-Escobar proposed a procedure that is supported by two techniques, the first is a Bayesian model based on the rank of each team in the English Premier league, the second is based on the fixtures shared by the teams that are competing. The ranking is computed using a combination of total goals scored for and against throughout the season and FIFA procedure. The Bayesian model uses the rank of each team to obtain a probability of success and failure. The analysis of the historical fixtures between the teams are included and found that there is high correlation between the historical fixtures and the results of the teams in the tournament. This is used in order to predict the 2018 world cup [2]. Another researcher, Danisik proposed a design that was based on player attributes data and match history. The attributes of the players were gained from FIFA and combined with real football matches. This prediction model would be beneficial for management of football teams in picking there starting eleven, as it could bring insights into dependence of each player on specific games and thus optimise the starting team [3]. This researcher Rahul Baboota proposed using features such as attack, midfield, defence and overall ratings for the respective teams. Also used goal difference, corners, shots on target, form and goals as features to see if these features are good selections, the problems associated with them and how the different features could determine the most important factors for the gradient boosting technique used to predict the matches [4]. Rotshtein created a model that underlies the method of identifying nonlinear factors by fuzzy knowledge bases. This model analysed the previous results of football teams in a tournament for the championship in Finland. By tuning the parameters using a combination of genetic and neural networks to improve the accuracy of their results [5]. Koning used a Monte-Carlo method with a Bayesian network to see if the results of a game is determined by home ground advantage and difference in quality between two teams. This model confirmed that the balance in competition in the Dutch league has not changed from the mid-70s [6].

Another machine learning technique was proposed by Joseph, their focus was on the creation of a Bayesian model and

compare it with machine learning techniques such as k-nearest neighbour, MC4-decision tree and a naïve Bayesian learner. This comparison was done to see if this Bayesian model would outperform the other techniques. This predictive model focused only on one team during a specific time, which was Tottenham Hotspurs during the period of 1995-1997 [7].

Research classified as statistical models are:
Firstly, Dixon was motivated to exploit the inefficiencies in the football betting market during the period of 1995-1996. The techniques used was bivariate Poisson regression model with the features of looking at the number of goals scored by each team. This model was shown to return a positive return based on betting information.[8]. Built upon the work of Dixon, Crowder decided to refine the original stochastic process model by changing Dixon approximation of a constant for attack and defence with an attack and defence that evolves with time. This estimation was implemented with an approximation that is computationally fast and compares favourably with results. Focussing on the probabilities of home and away wins, draws because of the betting market, this was modelled during the years 1992-1997 for 92 English teams.[9]. Koopman developed a statistical model using a bivariate Poisson distribution with coefficients that change over time. This model uses statistical time series analysis of football results and is based on state space using the intensity coefficients for attacking and defencing teams that changed over time. This demonstrated that the model produces a significant positive return over the betting market [10]. Another statistical approach was done by Boshnakov, he proposed the use of a count process that is derived when the inter-arrival times are followed by an independent and identically distributed Weibull count distribution. They allow the dependence between the goals scored by the teams during a match by employing a copula to generate a bivariate distribution with positive and negative dependence, using the English Premier League [11]. Rue took an applied statistics approach in which they suggest using a Bayesian dynamic generalised linear model to estimate the time-dependent skills of football teams in the English Premier League, but they need to estimate the skills of all teams simultaneously as they are dependent, and they solved this by using Markov chain Monte-Carlo iterative technique [12].

Research classified as rating system: Rating systems applications are mainly based on the ELO rating system, which was developed for ranking the strength of chess players. Leitner suggested a simulation that allows to approximate abilities rating of teams to winning probabilities from the betting market and apply them to the 2008 EURO's [13]. Constantinou proposed a novel rating system called pi - rating, that is computational efficient with low complexity. This can be used to formulate both score-based and result based match predictions and can be incorporated into more sophisticated models to enhance predicting capability. This provides relative measure of superiority between football teams based on the discrepancies in scores between rivals [14]. Halicioglu proposed a model that statistically analyses and attempts to predict the winner of EURO 2000 tournament based on seasonal coefficients of variations of end of season points from the top leagues of the participating countries in the EURO 2000. These coefficients ranked the countries to determine the most likely winner of the tournament.[15]. As I can see there is lots of academic research done into this field, however just a observation, many of the research is driven mostly to 'beat the bookmakers' as the bookmakers predictive models always seem to be the most accurate, but I would need to compare myself to confirm this observation.

III. METHOD

Data collection

In order to do any analysis, I will need information on football games, such as number of goals, number of red cards, etc. searching through several databases found the dataset that would be used. The dataset for this project would be taken from https://www.football-data.co.uk/englandm.php. This dataset contains metrics of the English Premier League from seasons 2010-11 to 2020-21. The benefits with the dataset are the number of games per season is fixed as in during the season three hundred and eighty matches are played during a premier league season. Also, I would be able to remove data that would not be required. Within the datasets there are lots of metrics that are helpful for the model such as "FTHG" (Full time home goals) and "FTAG" (Full time away goals) and other metrics such as referee names, cards produced and betting odd from certain betting companies.

Data selection

From this dataset, I have decided to collect seven years of data of the English Premier League from 2013–2020 season. From each season the features from the datasets that was required for this project was the date, so I can use the dataset as a sequential set of data. The home and away teams of the match in question, the number of goals scored by the home and away team in separate column in the match in question and finally the result of the matches, this will be particularly important feature as this will determine the accuracy of my predictions models later in the project.

Feature engineering

In this project, there is lots of feature engineering occurring, as the features extracted need to be given mathematical values that can be used for comparison. Firstly, from this dataset the summation of all the rows of home goals and away goals scored is done to then generate several other features from these two features. For example, the goals conceded if a team is home or away, etc. Also, from the previous features, other features are created such as creating aggregated points for each team, creating match weeks and this feature is important for scaling the other features as all the features will be divided by the match week, this is done for the classifiers as normalising or standardising data inputs for classifiers is necessary as some of the models will struggle with unnormalized datapoints. The reason for this is because some features will have different ranges and could take longer for

gradients to converge. All the teams will have their form calculated by finding the teams with five games won in a row and three games in a row and their loss forms are also calculated. Other feature engineering in this project is pre-processing features, for example, the all the form features are in categorical form and for the models to interpret these features will need dummy variables, which will convert the form features into separate features than can be converted into numerical values (i.e., HM1_W = 1, which means home team first game from five games ago was a win). Some of the features generated are not used for further computation but were given a mathematical value.

.

Modelling

The extracted features are used to train several classifiers in order to design a pattern from the dataset and predict a winner. I would have to consider how to deal with draws in matches as certain classifiers have difficulties dealing with draws (three outputs), so special consideration would be needed on implementing my solution. After comparing the respective teams, I get an output which will determine if the winner is either home or away. This output is now compared to the actual outcome of that match, so I can calculate the accuracy of my prediction. Other non-functional requirements for this project to consider are, if the number of features in use are high, it could take longer to train the classifier to form a pattern and or overfitting the model and it could map noise and reduce the accuracy of the output. Generally, the accuracy would be considered better than randomly guessing the output should it be over 50%. Each classifier is detailed below

A. K-Nearest Neighbours Classifier

The principle behind nearest neighbour (KNN) method is to find a defined number of training samples closest in distance to the new point and predict the label. KNN classification is non-generalising learning, as in it does not construct a general internal model but keeps instances of the training data. In this project, the number of neighbours to find will be determined by the data, as a large k will suppress the effects of noise, but the downside is the boundaries are less distinct.

B. Neural Networks Classifier

Looking at related works within the field of football outcomes, many researchers have attempted to predict the outcomes of football matched. Many have attempt to use statistical modelling such as Dixon [8], where he used a bivariate Poisson regression model to predict the English premier league. Also, Boshnakov [11] proposed using a Weibull count distribution which allowed the dependencies between goals scored to have both positive and negative dependencies. Other ways such as using ranking systems such as in Chess called ELO rating, where it measures the strength of the chess player, researchers have created similar ranking systems such as the pi-rating by Constantinou [14]. This system could be used to help more sophisticated model improve the accuracy of their predictions. However, machine learning seem to the choice of more recent research in the area as more research is done into optimising neural networks and further improvements. The proposed solution for the neural network is to use the scikit model called MLPClassifier (Multi-layer Perceptron classifier), This model is a class of feedforward artificial neural network that uses stochastic gradient descent, which is an iterative method for optimising functions. This method is further optimised with Adam (Adaptive Moment Estimation), in this algorithm the running averages of both the gradient and the second moment of the gradients are used. This neural network has parameters that need to be specified such as the amount of hidden layer or neurons and the shape required for the model, also the learning rate is needed.

C. Support-Vector Machine Classifier

Support vector machine (SVM) is given a set of training data, that are marked as belonging to one of two categories and the SVM will build a model that assigns new examples to one of the categories. SVM constructs a hyper-plane in high dimensional space. This hyperplane separates the two categories, and a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any category. SVM is scale invariant, so scaling input data is highly recommended. For this project, the C-supported vector classification and this multiclass support is handled to a one vs one scheme.

D. Gaussian Naïve Bayes Classifier

This classifier works by using the Bayes Theorem in which calculating the posterior probability from the prior probabilities of a hypothesis. For this classifier rather than calculating the values of each attribute value (i.e., P(D1,D2,D3|H)), they are conditionally independent given the target value and calculated as P(D1|H)* P(D2|H), etc. The classifier uses Gaussian distribution and to make predictions The features can be inputted into the Gaussian PDF with a new input for the variable and it returns the probability of that new input value.

E. AdaBoost Classifier

The principle of AdaBoost is to fit models that are only slightly better than guessing on repeated modified versions of the data. The prediction from all of them are then combined through a weighed sum to produce the final prediction. The modified data at each boosting iteration consist of applying weights to each training sample. For each succussive iteration, the sample weights are modified, and the learning algorithm is reapplied to the reweighted data. The training examples that were incorrectly predicted by the boosted model will have their weights increased, whereas the correct prediction will have the weights decreased. [14]

F. Random Forest Classifier

A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the data and uses averaging to improve the predictive accuracy and it controls overfitting. When splitting each node during the construction of a tree, the best split is found either from all features or random subset of the max features. The purpose of these two

sources of randomness is to decrease the variance of the forest the randomness in the forest yield decision trees with decoupled prediction errors, and by taking the average of those predictions, some of the error can cancel out. [14]

Features

From the literature review, different models had their own different set of features and classifiers. The accuracy of the models I have selected depends on the feature selections, dataset manipulation and the classifier used. To see which classifier will produce the best accuracy, the dataset needs to be standardised and scaled before any classifiers are introduced and all manipulation need to prepare correctly so that all the inputs conditions for each classifier is satisfied. The main features used in the dataset are goal difference, home, and away form, winning and losing forms and difference in points. Below is a sample of the final dataset with all the data manipulation done.

HTP	ATP	HM1_D	..	AM3_W	HTGD	ATGD	DiffFormPts
-1.0379	-0.5733	0	.	0	0.0146	-1.0124	-0.25
-1.0379	-0.9122	1	.	1	0.0146	0.9821	-1.00

Table 1: Final Dataset

The above table is used to feed into the classifiers, all the columns were divided by the match week with all 20 teams and there home and away fixtures.

Timeline Chart

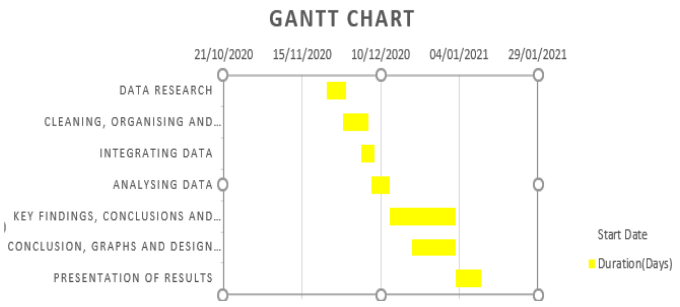

Figure 1: Timeline Chart

The above figure shows the Gantt chart of the project. It shows the tasks involved in the project; the time duration required to complete in each task. This timeline was accurately followed for the duration of the project. The project was expected to complete by the 9th of January.

IV. Result & Discussions

Finally, using the dataset and then having the dataset was divided into training and test of the ratio 85:15, the testing data is of the length three hundred and eighty, which is the number of fixtures of The English Premier League. Looking at the data within the dataset, I can see that the final data types are all correctly changed into numerical values or representation that is needed for input into the classifiers. All the columns were of the data type of Float64 or uint8(integer) and below is a figure of some the basic statistics for all the features in the final dataset.

```
In [102]: print(X_all.describe().transpose())
              count          mean       std  ...       50%       75%       max
HTP          2450.0 -2.998055e-16  1.000204  ... -0.167467  0.693816  3.061789
ATP          2450.0 -5.691592e-17  1.000204  ... -0.177173  0.700051  3.034598
HM1_D        2450.0  2.379592e-01  0.425921  ...  0.000000  0.000000  1.000000
HM1_L        2450.0  4.293878e-01  0.495090  ...  0.000000  1.000000  1.000000
HM1_W        2450.0  3.326531e-01  0.471260  ...  0.000000  1.000000  1.000000
HM2_D        2450.0  2.293878e-01  0.420525  ...  0.000000  0.000000  1.000000
HM2_L        2450.0  3.710204e-01  0.483177  ...  0.000000  1.000000  1.000000
HM2_W        2450.0  3.995918e-01  0.489914  ...  0.000000  1.000000  1.000000
HM3_D        2450.0  2.497959e-01  0.432983  ...  0.000000  0.000000  1.000000
HM3_L        2450.0  3.734694e-01  0.483824  ...  0.000000  1.000000  1.000000
HM3_W        2450.0  3.767347e-01  0.484666  ...  0.000000  1.000000  1.000000
AM1_D        2450.0  2.334694e-01  0.423125  ...  0.000000  0.000000  1.000000
AM1_L        2450.0  3.416327e-01  0.474354  ...  0.000000  1.000000  1.000000
AM1_W        2450.0  4.248980e-01  0.494428  ...  0.000000  1.000000  1.000000
AM2_D        2450.0  2.485714e-01  0.432273  ...  0.000000  0.000000  1.000000
AM2_L        2450.0  3.869388e-01  0.487149  ...  0.000000  1.000000  1.000000
AM2_W        2450.0  3.644898e-01  0.481385  ...  0.000000  1.000000  1.000000
AM3_D        2450.0  2.228571e-01  0.416248  ...  0.000000  0.000000  1.000000
AM3_L        2450.0  3.853061e-01  0.486767  ...  0.000000  1.000000  1.000000
AM3_W        2450.0  3.918367e-01  0.488260  ...  0.000000  1.000000  1.000000
HTGD         2450.0 -2.197788e-18  1.000204  ... -0.199006  0.605661  3.708599
ATGD         2450.0  2.184194e-17  1.000204  ... -0.205132  0.649678  3.707999
DiffFormPts  2450.0 -1.524013e-02  0.373318  ...  0.000000  0.156250  2.166667

[23 rows x 8 columns]
```

Figure 2: Basic Statistics

From the figure above, I can see statistics for all the features, also can see the dummy columns used for the ease of use in the classifier to predict a pattern from the forms of the football team. From the figure, I can see the mean of all the columns are close to zero, either negative or positive, for example the form columns of 'HM1_W' seem to prefer the positive side of zero with a value of 0.332653. I can see that the standard deviation is 1.000204 for the features of HTP, ATP, HTGD, ATGD and DiffFormPts. The standard deviation for all the form columns is roughly around ~0.45. These two standard deviations are good for using these columns as a good training features to put into the classifier. Below is a Figure that shows how some of the features in the dataset ranges, mean is represented.

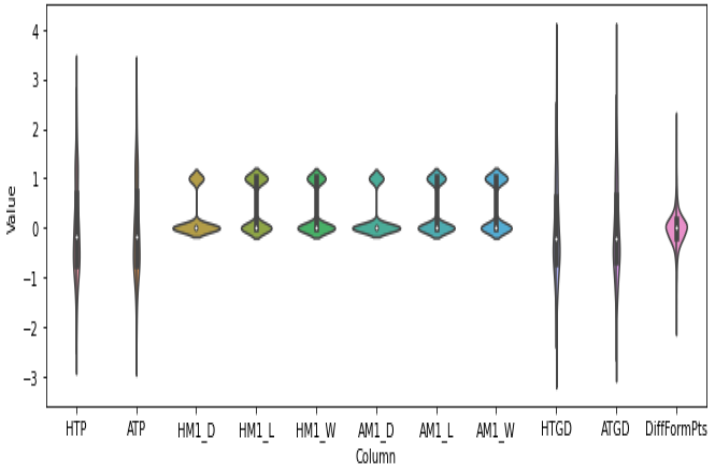

Figure 3: Ranges of Features

As I can see all the features have most of their datapoints either around the zero value or in the case of the form features its between one and zero. Also looking at the form features can see that the draws in home and away have more of their datapoints at one, whereas the loss and win features seem to get a more even spread of the datapoints. Between -3 and 3. Further investigation in future works. Before looking at the

results of the classifiers on the dataset, some statistical inference was investigated. The plot below shows all the continuous data features plotted against each other and their distributions.

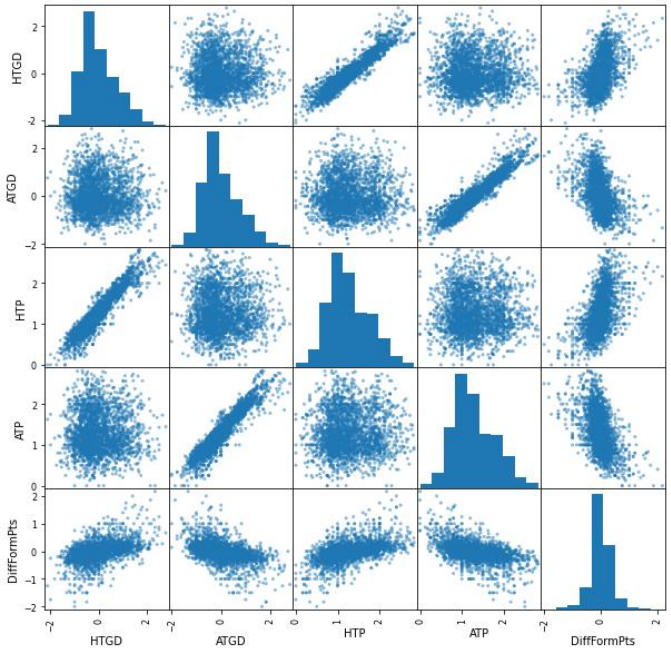

Figure 4: Pair-plot

In the figure above, I firstly look to the plots with the distributions of the features and I can say that they seem to follow a Normal Distribution, however ever so slightly skewed to the left in this dataset. However, I will assume it follows the Normal Distribution, which helps me to investigate any linear relationships. From the Figure, I can only see two well defined linear relationship which are of the features, HTP vs HTGD and ATP vs ATGD. Another look at the figure, I could say maybe the feature DiffFormPts has a weak linear relationship with the other features, but further investigation would be required. Using a Pearson correlation test on the features HTP vs HTGD, I found that there is a strong relationship with 0.93613 value that is remarkably close to 1 meaning a strong positive linear relationship and a p-value stated here of 0.0. This is also the case for the feature of away points and goal difference.

OLS Regression Results

Dep. Variable:	HTGD	R-squared:	0.876
Model:	OLS	Adj. R-squared:	0.876
Method:	Least Squares	F-statistic:	1.735e+04
Date:	Fri, 08 Jan 2021	Prob (F-statistic):	0.00
Time:	17:44:04	Log-Likelihood:	-217.34
No. Observations:	2450	AIC:	438.7
Df Residuals:	2448	BIC:	450.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.7866	0.015	-123.209	0.000	-1.815	-1.758
HTP	1.3933	0.011	131.723	0.000	1.373	1.414

Omnibus:	306.413	Durbin-Watson:	1.742
Prob(Omnibus):	0.000	Jarque-Bera (JB):	603.460
Skew:	0.782	Prob(JB):	9.13e-132
Kurtosis:	4.861	Cond. No.	5.52

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 5: Model One

Doing an Ordinary Least Squares regression model on the two features (HTP vs HTGD) and checked if all the assumptions are met to perform this regression model, the two values of R-squared and Adjusted R-squared came to be 0.876, meaning 87% of the variability in the data is explained in the linear regression model with p-values of 0.000 which is well below the conventional threshold of the 95% confidence intervals.

OLS Regression Results

Dep. Variable:	DiffFormPts	R-squared:	0.264
Model:	OLS	Adj. R-squared:	0.263
Method:	Least Squares	F-statistic:	875.9
Date:	Fri, 08 Jan 2021	Prob (F-statistic):	8.03e-165
Time:	17:45:07	Log-Likelihood:	-687.17
No. Observations:	2450	AIC:	1378.
Df Residuals:	2448	BIC:	1390.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4985	0.018	-28.380	0.000	-0.533	-0.464
HTP	0.3792	0.013	29.596	0.000	0.354	0.404

Omnibus:	214.974	Durbin-Watson:	2.109
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1335.305
Skew:	0.101	Prob(JB):	1.10e-290
Kurtosis:	6.611	Cond. No.	5.52

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 6: Model Two

Considering the same linear regression model but for the features of HTP and DiffFormPts, the R-squared result was 0.264 and the Adjusted R-squared result was 0.263, this suggests that 26% of the variability in the data is explained. This is not a good model for attempting explain the relationship between the difference in form and the home points.

Considering my project and its outcomes, I have classified the problem as being a classification problem rather than a regression problem. The difference in the two is that the classification model will try to predict the class, or probability distribution for the classes, whereas the regression models will try to predict real numerical values. In my project the outputs that I am looking for are Wins, Draws and Loses, so regression models are not to be used to predict my outcomes. Firstly, below is a figure that will show how over the years, the frequency in which the result has been classified as home win, away win or a draw.

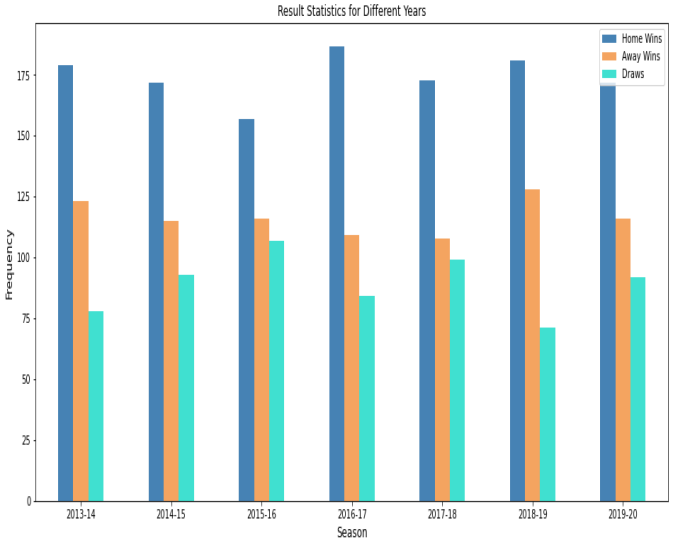

Figure 7: Frequencies over the years

As I can see, over the seven years most of the results have been classified as home wins and the least is draws, so from my interpretation, my prediction model should have more home wins than draws, also highlights a bias towards home teams winning their home games more.

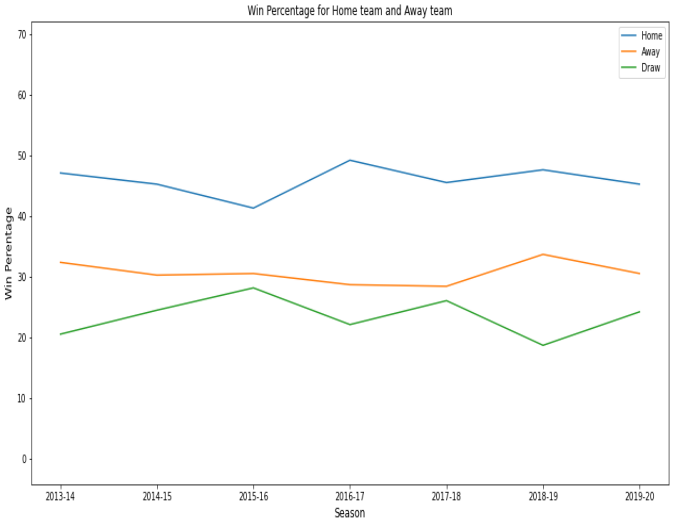

Figure 8: Win rate over the years

From the above figure, I can see that the home win rate is higher than the away win and draws which is below the 40%-win rate, this clearly shows that when a team is playing at home, they are much more likely to win that game than if they were the away team.

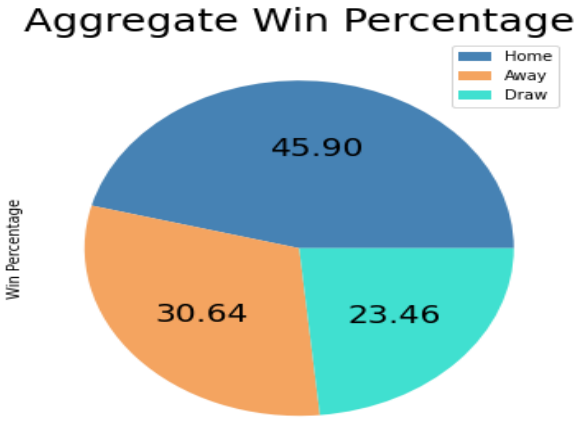

Figure 9: The aggregate win percentage

And as I can see from the figure above, when all the home wins, away wins and draws where aggregated over the years, it was found to have the above percentages for each class of outcomes. Onto predicting the outcome of a football match, I have the accuracies for both the training set and the test set to see how accurate each of the readily available classifiers are.

The results are as below in the Figure 10.

```
TRAINING
Accuracy for AdaBoost: 0.5449275362318841
Accuracy for RandomForest: 0.5280193236714976
Accuracy for KNN: 0.7053140096618358
Accuracy for svm: 0.5937198067632851
Accuracy for Neural Nets: 0.5449275362318841
Accuracy for Naive Bayes: 0.5193236714975845
TESTING
Accuracy for AdaBoost: 0.5263157894736842
Accuracy for RandomForest: 0.5684210526315789
Accuracy for KNN: 0.3657894736842106
Accuracy for svm: 0.5631578947368421
Accuracy for Neural Nets: 0.5710526315789474
Accuracy for Naive Bayes: 0.5052631578947369
```

Figure 10: The Accuracies of the Classifiers

From the figure 10, I can see the accuracies of the six classifiers that were used to predict the football match outcomes. The accuracies are in decimal form, so from now on will refer them as percentages (i.e., 0.54492 = 54.49%). Firstly, looking at the training set accuracies, I can see that the K-Nearest Neighbour Classifier has the highest percentage with 70.53% and the lowest percentage being the Naïve Bayes Classifier with 50.52%. The lowest being higher than 50% I Feel is a successful outcome, however it is quite low but with the KNN classifier being 70.53%, this is a huge leap in accuracy. Whereas looking at the testing set of the data, I can see rather opposite result for the KNN with an accuracy of 36.57%, which is a considerable drop in accuracies, perhaps can be explained due to the modelling perhaps focussing on the wrong feature to base its model off or the testing set having datapoints too close or too far for the predefined K value to impact the generated model to predict. The highest percentage in the testing set was the Neural Network with 57.10%, this is an improvement from the training set. Also, I looked at the time it takes these classifiers to predict the outcomes and found that even though the Neural network had the highest accuracy it had the longest time 4.39 seconds, but all the other classifiers are in the milliseconds. Another solution to improve the accuracy of the models is to append the draws to the lose class that need to be classified and I can see that the highest accuracies are around now the 62%, however I would lose the ability to predict draws, also did not include these accuracies due to that lose in prediction of draws.

V. Comparison with Related Works

Comparing other academic research is not the most straightforward, as other researchers have trained on different datasets such as [13] used the EURO 2008 as the dataset or [6] using the Dutch League. Some researchers have used the same league but have used it at different times such as [8] and [9]. Considering accuracies as the metric for comparison, from [3] the highest accuracy was 52.479% and this journal would be consider close to our relevant topic and in this journal, they also compared it with the bookmakers which had a accuracy of 52.946% which is still the highest accuracy of that journal.

Also, this accuracy considers draws, whereas some academic journals will just consider win and lose. Another academic journal [16] like our proposed work stated a test accuracy of 80.75%, however its not stated if that includes draws, as if that was the case this model predicts outcomes of games very well. I believe my final highest accuracy on the testing data was my Neural Network with a 57.10%. Considering this value and other values, I consider this an acceptable accuracy by comparison to other literature.

VI. Conclusion & Future works

Concluding remarks on my proposed work of trying to predict the outcomes of football matches in the English premier league, using match histories and several Classifiers. Firstly, the highest accuracy achieved was by a Neural Network with 57.10%, however it took considerably longer to compute the model compared to the other classifiers. I believe if I remove the draw classifier, I will obtain a higher accuracy on finding the winning team. Also, other researcher has discussed how the training phase of their datasets was struggling to implement the first week of football outcomes and the conditions needed to accurately say from the confusion matrix what a draw is. The researcher also suggested solutions that I have tried to improve on my work. My approach to the work would lead to a predictive model that uses minimal information, in terms of match history features such as form and goal difference and various classifiers to produce the best performing model. From this work would like to have a positive impact on future works using this method, and find elusive predictive attributes not considered before. Future works would be using different features that would incorporate match day features like formation, attacking and defensive form, starting eleven. These features could increase accuracies for our selected model of Neural Network. Probable uses of the model could be used in journalism to compared real results and predicted results or could be used to further progress the field in prediction football outcomes.

References

[1] Ryan Beall, “Optimising Game Tactics for Football”, Accessed:23/11/2020, [https://arxiv.org/pdf/2003.10294.pdf]

[2] Laura Hervert-Escobar, “Prediction learning model for soccer matches outcomes”, Accessed23/11/2020, DOI:10.1109/MICAI46078.2018.00018 [https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9046485]

[3] N. Danisik, P. Lacko and M. Farkas, "Football Match Prediction Using Players Attributes," 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Kosice, 2018, pp. 201-206, DOI: 10.1109/DISA.2018.8490613.

[4] Rahul Baboota, Harleen Kaur, “Predictive analysis and modelling football results using machine learning approach for English Premier League”, International Journal of Forecasting Volume 35, Issue 2, 2019, Pages 741-755, ISSN 2070, [https://doi.org/10.1016/j.ijforecast.2018.01.003.] (http://www.sciencedirect.com/science/article/pii/S0169207018300116)

[5] A. P. Rotshtein, M. Posner, and A. B. Rakityanskaya, “Football predictions based on a fuzzy model with genetic and neural tuning, “Cybernetics and Systems Analysis, vol. 41, no. 4, pp. 619–630, 2005. [http://dx.doi.org/10.1007/s10559-005-0098-4]

[6] R. H. Koning, “Balance in competition in Dutch soccer, “Journal of the Royal Statistical Society: Series D (The Statistician), vol. 49, no. 3, pp. 419–431, 2000. [Online]. Available: http://dx.doi.org/10.1111/1467-9884.00244

[7] Joseph A., Fenton N.E. and Neil M., “Predicting football results using Bayesian nets and other machine learning techniques”, Knowledge-Based Systems, 19 (2006), pp.544-553, DOI:10.1016/j.knosys.2006.04.011, [https://www.sciencedirect.com/sdfe/reader/pii/S0950705106000724/pdf]

[8] Dixon M.J. and Coles S.G, “Modelling association football scores and inefficiencies in the football betting market”, Journal of the Royal Statistical Society. Series C. Applied Statistics, 46 (1997), pp. 265-280, [http://refhub.elsevier.com/S0169-2070(18)30011-6/sb12]

[9] M. Crowder, M. Dixon, A. Ledford, and M. Robinson, Journal of the Royal Statistical Society. Series D the Statistician, number = 2, Pages = 157-168, publisher = Royal Statistical Society, Wiley, title= Dynamic Modelling and Prediction of English Football League Matches for Betting, volume = 51, 2002, [http://www.jstor.org/stable/3650316]

[10] Koopman S.J. and Lit R, “A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League”, Journal of the Royal Statistical Society. Series A. Statistics in Society, 178 (2015), pp. 167-186, [https://doi.org/10.1111/rssa.12042]

[11] Boshnakov G., Kharrat T. and McHale I.G., “A bivariate Weibull count model for forecasting association football scores”, International Journal of Forecasting, 33 (2017), pp. 458-466, [https://www.sciencedirect.com/science/article/pii/S0169207017300018]

[12] Rue, H, Salvesen Ø, “Prediction and retrospective analysis of soccer matches in a league”, Journal of the Royal Statistical Society Series D: The Statistician Volume 49, Issue 3, 2000, Pages 399-418, [https://www.scopus.com/record/display.uri?eid=2-s2.0-0001134174&origin=inward&txGid=be47c667b49b1e5b2ce26265a9e1e6b9]

[13] Christoph Leitner, Achim Zeileis, Kurt Hornik, “Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008”, International Journal of Forecasting, Volume 26, Issue 3, 2010, Pages 471-481, ISSN 0169-2070, [https://doi.org/10.1016/j.ijforecast.2009.10.001.],

(http://www.sciencedirect.com/science/article/pii/S0169207009001459)

[14] Constantinou, Anthony & Fenton, Norman. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. Journal of Quantitative Analysis in Sports. 9. 37-50, DOI: 10.1515/jqas-2012-0036, [https://www.researchgate.net/publication/236944354_Determining_the_level_of_ability_of_football_teams_by_dynamic_ratings_based_on_the_relative_discrepancies_in_scores_between_adversaries]

[15] Halicioglu, Ferda, "Can We Predict the Outcome of the International Football Tournaments? the Case of Euro 2000", Doguş Üniversitesi Dergisi. 6, 2005, [DOI: 10.31671/dogus.2019.288]

[16] E. Tiwari, P. Sardar and S. Jain, "Football Match Result Prediction Using Neural Networks and Deep Learning," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 229-231, DOI: 10.1109/ICRITO48877.2020.9197811.

[17] Scikit Learn API Reference, [https://scikit-learn.org/stable/supervised_learning.html#supervised-learning]