Comparing Machine Learning Techniques To Solve Prediction Dilemmas In Football

P S Abhiram*
spand086@uottawa.ca
University of Ottawa
Ottawa, Ontario, Canada

ABSTRACT

The aim of this paper is to maximize the utility of already processed data and incorporate into several machine learning models. For this multi class classification problem, an thorough examination of four machine learning models are done in order to determine how to get the best performance and counter some overfitting problems. The football datasets under consideration are from 2013/14 to 2019/20. 2020/21 season is omitted due to it being a season played under lockdown and with no fans. In addition to tacking problems, an exploratory data analysis is conducted to better understand the beautiful game. The code for this project can be found on GitHub [1].

CCS CONCEPTS

• >**Machine learning**; • **Football trends**; • **Big data management**;

KEYWORDS

Logistic Regression, Supervised learning, XgBoost, Support vector Machines

ACM Reference Format:
P S Abhiram. 2022. Comparing Machine Learning Techniques To Solve Prediction Dilemmas In Football. In *Proceedings of (COMP5118).* ACM, New York, NY, USA, 8 pages. https://doi.org/TrendsinBigdatamanagement

1 INTRODUCTION

Football is a ubiquitous sport. It is well known almost across the globe and almost everyone wants to involve in the spectacle. From one single game, there is a huge amount of data generated, which can be extracted for many purposes. For example, passes made per minute, progressive passes made per player, number of yellow cards and red cards, score from previous games, and many more in-depth analysis. Modern systems also consider data from many websites. Predicting the result of a football match is impossible, and can only be approximated based on the factors that could be considered. A team that is strong on paper might not be as strong on the pitch. Different factors like form, mentality, and play styles dictate a game. There is always an element of luck involved in the game too. Having a lenient referee over a strict one would affect how the game is played. An easy referee might give a soft penalty which could change the face of the game. It is unpredictable human factors like these which further amplify the difficulty in predicting the result. Although there are large amounts of large data to be processed. Unfortunately, the datasets available for predicting a football for research purposes lack depth. Some elements like progressive passes per 90, mentality, etc. can not be found. This information is present only with a proprietary organization like the football club and data analysts whose main task is to gather this kind of data for their organization. Neither of the above-mentioned organizations choose to release this data to the public. For example, in the 2021/22 premier league season, during live broadcast, Cognizant releases statistics of win percentage. Such data can only be processed by a well-funded team. These results usually undermine the research work done with minimal available data. Hence, we do not usually see a model with a perfect prediction.

Many people have tried to crack the code to predict the result of a soccer match with the utmost precision. During the 2010 world cup people even tried to use Paul the Octopus to predict the results. Surprisingly, Paul was doing so well until people decided that it was affecting the results by affecting the players' mentality and approach towards the game. Lets take this scenario. A person is predicting the results with the a 100 % (backslash percent) and there is no mistake made anytime. If this person said that a particular team will win the game ahead of the game, would that affect the results of the game? It is impossible to think someone predicting the future correctly all the time. But, if done so, it would definitely destroy the beauty of the game. Would we still enjoy the Cristiano Ronaldo's infamous Champions league performances? Would Manchester United's historic treble winning moment against Bayern Munich still be as exciting? Morally, it would seem like an unwise choice to let such a person keep predicting the games. But, that's where machine learning is evolving towards. It is up to the players to not get affected by such predictions and defy some improbable odds.

2 GOALS AND CONTRIBUTIONS

The goals and contributions of this project are as follows:

(1) To get a good understanding on how different machine learning models perform when trained under different scenarios.
(2) There is never a clear cut model that always outperforms other machine learning models. But in predicting football results, a model is suggested that does well over other ones.

3 CHALLENGES

Some of the existing challenges in the field of predictions of football results are as follows:

[1] https://github.com/PSatyadevAbhiram/COMP5118-Predicting-Football-Results

COMP5118, April 20, 2022, Ottawa, ON

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/TrendsinBigdatamanagement

(1) Do we get the same access to data as teams? One of the biggest problem in predicting football results are the datasets. No matter how far we go, we only get access to a limited amount of data. Exclusive data is often privatized by the football clubs/media companies.
(2) When considering sentiment analysis from experts, how much of an impact can humans deliver when predicting a result? One of the shortcomings is that all the pundits do not watch all the games, yet give their opinions on the game. How do we eliminate bias factor from human opinions?
(3) Another massive challenge is the decisions researchers have to make while performing machine learning tasks in sports. No matter how good we build our machine learning models, there are bound to b results where there is an unexpected winner. Such results are considered long shot results and even he bookmakers place higher odds on such a result occurring. So while building a model, do we focus on getting such results correctly? Or instead do we focus on getting the most obvious results correctly? How to find the balance between both worlds without the machine overfitting on either task?

4 RELATED WORK

As explained in Section 1, there is a lot of work conducted on predicting the results of football matches. Firstly, in [1], a baseline model is suggested in predicting results of football game. The authors offer a new application-focused benchmark dataset as well as findings from a collection of baseline Natural Language Processing and Machine Learning models for predicting match outcomes in football games in this work (soccer). By doing so, they provide a baseline for the forecast accuracy that may be attained by combining statistical match data with contextual pieces written by human sports journalists. The authors also cover an ensemble approach by combining different results from each approach.By doing so, the authors achieve a better precision percentage in predicting results that were unexpected. Machine Learning is not only about predicting results of a football match. It can be used to delve deep into the way the game is played. In [4], Michael Stöckl et. al use Graph Convolutional Neural Networks to make offensive play predictable. This paper focuses more on understanding how the game is played instead of predicting the result. The authors use Run maps of players and probabilistic analysis to understand player movements and expected areas the ball would be passed to. There have also been several sentiment analysis studies conducted by different researchers to extend this study. In [3] the authors attempt to determine if the sentiment included in tweets can be used as a relevant proxy to forecast match results, and if so, what magnitude of outcomes may be anticipated based on a degree of emotion. To answer these questions, the authors built the CentralSport system, which collects tweets connected to the English Premier League's twenty teams and analyses their emotion content, not just to anticipate match outcomes, but also to utilise as a wagering decision system. Despite the fact that sports events such as football matches generate a lot of public attention and a lot of online communication, social media analysis in general, and sentiment analysis in particular, are practically untapped tools in sports science so far. In [5] tests the feasibility of lexicon-based tools of sentiment analysis with regard to football-related textual data on the microblogging platform Twitter. When using machine learning, the next step for attaining better performance numbers is to use Deep learning algorithms on the required task. In [2] the authors proposed a deep neural network based model to automatically predict result of a football match. The model is trained on selective features and evaluated through experiment results. Furthermore, a comparison is made with the performance of feature-based classical machine learning algorithms.

5 METHODS

5.1 Logistic Regression

The problem to be solved is a multi class classification problem where the outcomes of a match are either the home team wins, loses or draws the game. For many prediction tasks, we utilize logistic regression as a baseline performance model. Multiclass classification is a simple extension of binary classification for logistic regression. As shown in Figure 1, we try to find the boundaries or the most optimal lines that separate our input data points. Logistic regression serves as a great baseline performance mainly because of it's flexibility in adapting to almost all problems. It might not be the best outright model, but gives a good performance when conducted on any dataset.

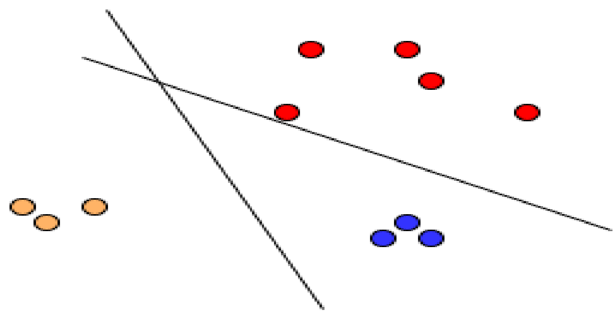

Figure 1: The functioning of a Logistic Regression model

The hypothesis for linear regression is given as:

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + ... + \theta_n x_n$$

The cost function of Logistic regression is given as a log loss function as below:

$$cost(h_\theta(x), y) = -log(h_\theta(x)) - log(1 - h_\theta(x))$$

The cost function of the model when we input the whole training dataset is given by its summation as follows:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x_i), y_i)$$

where m is the number of samples.
In case of multiclass classification, we do not use the sigmoid function since a sigmoid function only gives us the probabilities of a binary classification problem. Instead, we use a softmax activation function which gives us probabilities of three events occuring.

5.2 Support Vector Machines

A Support-vector machine (SVMs) is another supervised learning model which is used to analyze data associated for classification and regression analysis. An SVM training algorithm constructs a model that classifies samples to one of two categories in a binary classification problem or one of many in a multi class classification problem, given a collection of training examples, resulting in a non-probabilistic linear classifier. It translates training examples to points in space in order to widen the distance between the two or more categories as much as possible. New instances are then mapped into that same space and projected to belong to one of the categories based on which side of the gap they land on.
For the purpose of this project, an SVM is chosen mainly because of it's ability to perform well in non-linear classification problem. But we thus far have a linear classification model and we would like to test the hypothesis that SVM is not that good or maybe gives a similar performance in a linear classification problem. For non-linear classification, a kernel trick is used which implicitly maps inputs into higher dimensional feature space. One thing to observe

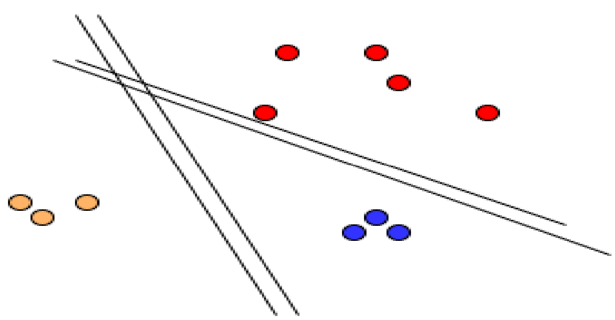

Figure 2: The functioning of a Support Vector Machine

in the figure is that SVMs, there is a double line used for separation, instead of a single line as previously observed in linear classification graph. This is mainly because of the previously discussed point that Support vector creates a sizeable gap between different classes. This can cause a hindrance especially when working with data that does not necessarily require this gap i.e linear data model. So, using SVMs is more to prove that it might not perform as well as Logistic Regression model which is our baseline. The loss function of SVM is again similar to that of Logistic regression.

5.3 Decision Trees

The next machine learning algorithms we use are decision trees. This is another approach in solving multi class classification problems. It consists, as the name suggest, a tree like structure that starts from rot and goes to leaf taking different decisions based on different features that are present while learning. A decision tree mainly consist of:

(1) Decision node: Indicates the node where a critical decision needs to be made.
(2) Chance nodes: This will emerge into two or more decisions from the decision nodes. It is like a connection between different different decision nodes.
(3) End nodes: These are more formally known as the leaf nodes, which represent the ultimat decision taken by the model after a series of decisions nodes.

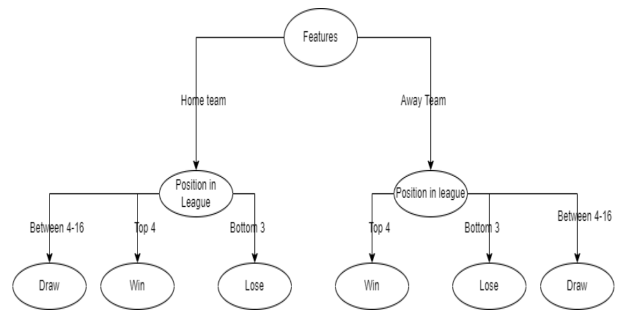

Figure 3: The functioning of Decision Tree algorithm

Some advantages and disadvantages of decision trees are as follows:

- Simple to implement.
- Work well with a rather poor input data as well.
- A small change in the dataset can result in a significant change in optimal decisions.
- They are more suitable as a baseline approach rather than an out right performance model.

5.4 XGBoost

Extreme Gradiesnt Boost (XGBoost) is a distributed gradient boosted decision tree and is the most popular machine learning model for popular machine learning algorithm for regression, classification and ranking problem. Why is it so popular? More recent trends according to Kaggle show that XGBoost delivers the most optimal performance. Hence, we test this hypothesis. XGBoost has its

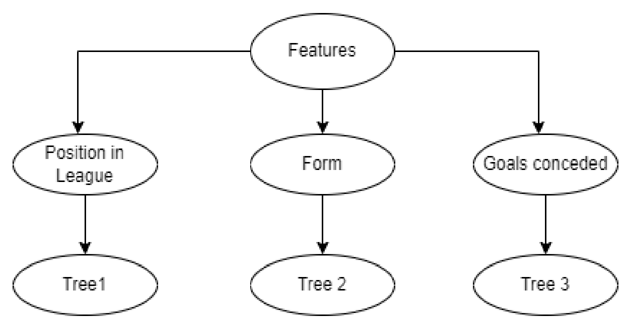

Figure 4: The functioning of XGBoost algorithm

own drawbacks. This will be further discussed in the experiments section where a lot of overfitting problems were faced and after multiple tweaks to the input parameters, we get a representative performance.

6 EXPERIMENTS

Although there were numerous experiments conducted using the above stated machine learning models, the experiments section in

this paper highlight the metrics chosen, performance over a different datasets and some cross validation techniques. The experiments were conducted on a GPU but as common with many machine learning algorithms, they can easily run on a simple system on a CPU as well. So, recreating the experiments section should not become a problem.
This data under consideration is taken from an open source data set that contains the results of a game between two teams along with the home team, away team, half time scores, yellow cards, red cards, shots on target by either team, referee name, etc. To enhance this dataset, a new column is added that signifies the overall rating of the home team given by FIFA. This dataset is also taken from an open source platform which keeps track of the teams' overall ratings holding with respect to the FIFA game. Furthermore, the referee name is replaced with referee score. To perform this operation, the Premier League official website is reference to check the statistics of the referee for a particular season. This dataset gives us clean details like the number of game a referee managed that season, red cards, yellow cards, penalties given by the referee for a particular season. Taking this data into consideration, Referee score(RS) is calculated as follows:

$$RS = \frac{RC + YC + Penalties}{N}$$

where RC denotes red card, YC denotes yellow card and N denotes the number of games the referee officiated till date. To further validate the data under consideration and if addition of these values actually do matter, we need to use the *scikit* package from Python.

6.1 Feature Selection

The following are the total features present in the dataset after making all the changes described above are Home Team, Away Team, Full Time Home Team Goals (FTHG),Full Time Away Team Goals (FTAG), Half time Home Team Goals (HTHG),Half time aWAY Team Goals (HTAG), Half time Result (HTR), Referee Score (RS), Home Team Shots (HS), Away Team Shots (AS), Home Team Shots on Target (HST), Away Team Shots on Target (AST), Home Team Fouls Committed (HF), Away Team Fouls Committed (AF), Home Team Corners (HC), Away Team Corners (AC), Home Team Yellow Cards (HY), Away Team Yellow Cards (AY), Home Team Red Cards (HR), Away Team Red Cards (AR), Home Team Overall (HomeTeamOVR), Away Team Overall (AwayTeam OVR), Home Team points (HTP), Away Team Points (ATP), Home Team Goals Scored (HTGS) , Away Team Goals Scored (ATGS), Home Team Goals Conceded (HTGC), Home Team Form Points (HTFormPts), Away Team Form Points(ATFormPts), Home Team Win streaks - 3,5, Away team Win Streaks - 3,5, Home Team Loss Streak - 3,5, Away Team Loss Streak 3,5, Matchweek (MW), Home Team Goal Difference (HTGD), Away Team Goals DIfference (ATGD), Difference in total points between home team and away team and finally, Difference in form points between Home Team and Away team.
There are many features added after several references from various GitHub pages and kaggle competitions. One main reference GitHub is https://github.com/llSourcell/Predicting_Winning_Teams/blob/master/Scraping%20and%20Cleaning.ipynb

6.2 Exploratory Data Analysis

With so many features, there is bound to be some overfitting concerns and a lot of the above features maybe be dependent on each other. To further understand this dependency, a mutual f-score is obtained using the features and the output variable - Full Time Result. In the interest of saving space this result is not showcased in this paper. However, this information can be found on the GitHub page. Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. Using this information, we are able to estimate how our features are related. Furthermore, an ablation study covers how different models perform by selecting only the top 15 features from our initial dataset. Please refer the Ablation study section to view the results of this study.
TO extract further information on how well our features co relate to the result and their significance in this process, an F score is extracted. Using the *scikit* feature selection package, we compute the ANOVA F-Value showcase the relevance of each feature. There are many more tests like the Chi-square test. But, in our case the dataset consists of negative values as well, for example, the goal difference, difference in form points, difference in total points. These values cannot be shifted to positive just for the sake of the test since a negative value signifies something here and converting it to a positive value might impact the study.

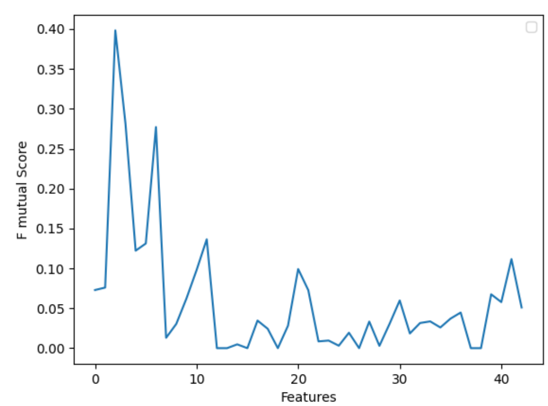

Figure 5: Features vs F mutual Score

From figure 5, we get to understand how much of a relation each feature has on itself. The features here are numbered based on the above listed features. The most highlighting points from this figure is that there is a heavy dependence of feature number 2 - Full Time Home Goals (FTHG). What this tells us is two things:

(1) Home Team goals matter more than away team goals.
(2) Generally, Home Team goals scored are more than away team goals scored.

To better understand the the features we are dealing with, different statistical analysis have been conducted. A scatter plot matrix is a

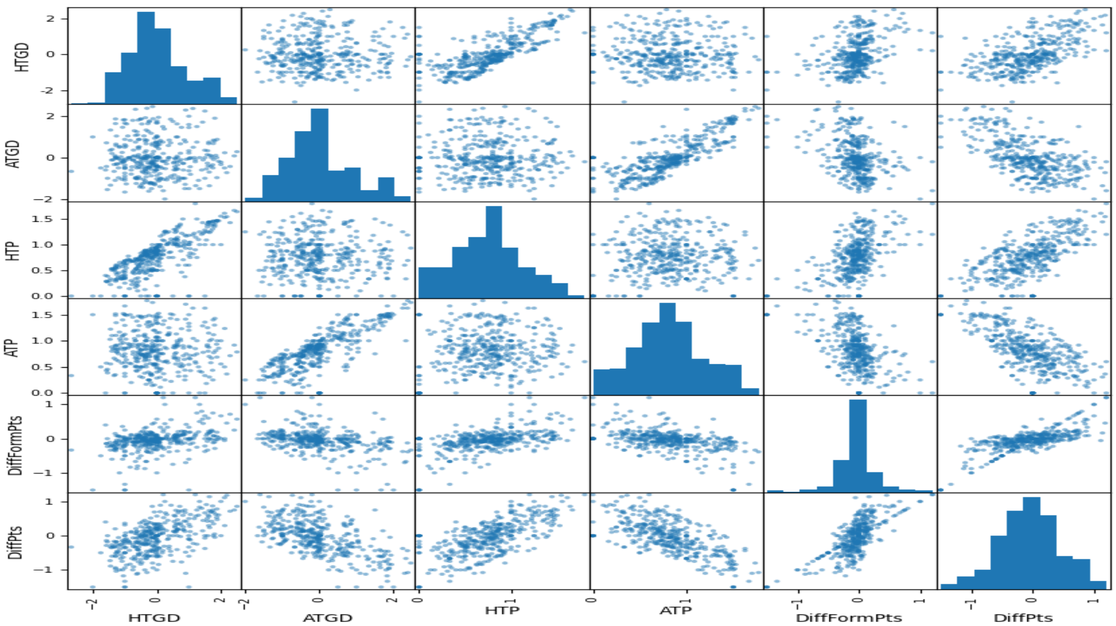

Figure 6: Scatter Matrix of 2018/19 season

grid (or matrix) of scatter plots used to depict bivariate connections between variable combinations. Each scatter plot in the matrix depicts the relationship between two variables, allowing several associations to be investigated in a single image. From figure 6, we can visualize these relationships. Some of the figures shown here have a positive co-relation and some have a negative co-relation. For example, if we consider the goals scored by the home team, it also means that the points are decreased for the away team.

It is a common assumption that home team performs better than the away team. Several factors affect this performance, like, familiarity with the pitch, the majority of fans support you and there is not that much of a travel time to play a game resulting in less fatigue. The bar graph below depicts the information for the 2011/12 season. As we can see, the number of home wins in each of the seasons is significantly higher than the total number of losses or draws. In such a crowd dependent sport, home teams always have an edge over the away team. They have something more to back them up. We have also seen UEFA competitions implement 'Away goals rules'. This is mainly done to give the away team a little bit more edge. However, this season this rule was scrapped. Not just in club football, even in international football, the host nation participating in the World Cup does better than they would have expected. This goes to show that each classifier will try to have some bias towards the home team. In figure 7, 11/12 depicts season 2011/12 and this applies for other similar values in the graph as well. Now that we have performed some data analysis to explore how the relationships between different features vary and how home teams get an advantage, we will now proceed to showcasing the results conducted using this experiment setup. The next section will cover the results.

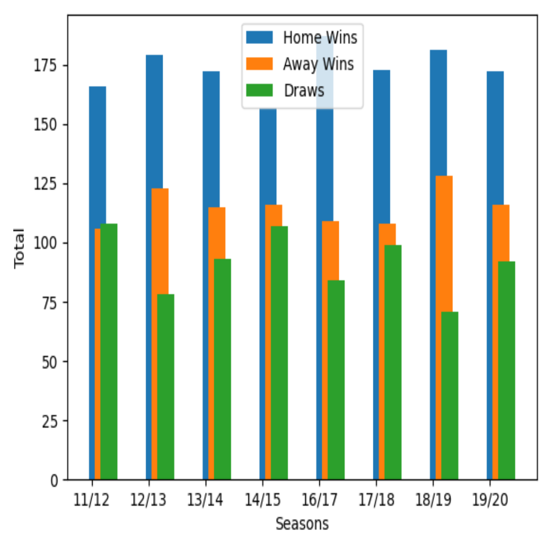

Figure 7: Total wins, losses and draws over 8 seasons

7 RESULTS

The results had to be carefully adjusted and several tries and parameter adjustments had to be made in order to curb overfitting mainly due to the number of features and the classifiers chosen.
For the purpose of this project, we have chosen three different seasons. Cross validations have been performed on these three seasons. the classifiers have been trained and tested using 2017/18, 2018/19 and 2019/ seasons as follows:

(1) M1 - train 2013/14 to 2018/19 and test 2019/20
(2) M2 - train 2013/14 to 2019/20 except 2018/19 and test 2018/19
(3) M3 - train 2013/14 to 2019/20 except 2017/18 and test 2017/18
(4) M4 - train 2013/14 to 2019/20 except 2016/17 and test 2016/17
(5) M5 - train 2013/14 to 2019/20 except 2015/16 and test 2015/16
(6) M6 - train 2013/14 to 2019/20 except 2014/15 and test 2014/15
(7) M7 - train 2014/15 to 2019/20 and test 2013/14

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.
Furthermore, parameter tuning of different models had to be done in order to prevent overfitting. Overfitting was the major problem with XgBoost and Decision trees. I suspect this was because of the many number of features and their dependencies. The ablation study covers the performance of models based on the select 15 best features. This will further help us get a more accurate data. For this section, the accuracy, F1 score and confusion matrices are extracted to showcase and validate the results. table 1 represents the results of data testing performed in the first scenario (Train: 2013/24 to 2018/19 Test:2019/20).

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.82	0.61	0.79	0.75
F1 score	0.80	0.46	0.79	0.63

Table 1: Results after using cross-validation M1.

These results, although look good on paper are not impressive for SVMs when we look at the confusion matrices. It is observed that SVMs are having a tough time predicting the class that is in minority. In our model, 0 - draw, 1 - away win, 2 - home win. So here the SVM classifier is cleverly predicting either 0 or 2. Although this gives a good accuracy number since there are more home wins and draws than away wins, SVM is not a good model to use. Furthermore, Decision trees surprisingly gives better performance compared to decision trees. But, even the confusion matrices of Decision trees give us a poor performance. Although all three classes are predicted, it sometimes favours home wins more. This is also the case with XgBoost.
The outright best classifier in this scenario is the most simple logistic regression. The confusion matrix for it shows that there is a good distribution among all classes and the model actually managed to learnt to predict all three scenarios.

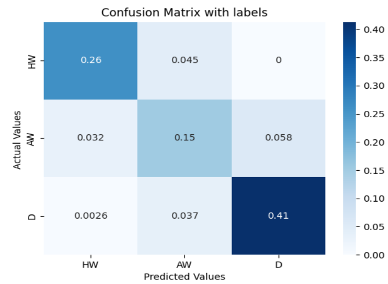

Figure 8: Confusion Matrix in Percentages using Logistic Regression obtained from M1

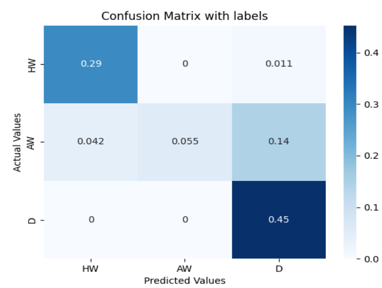

Figure 9: Confusion Matrix in Percentages using SVM obtained from M1

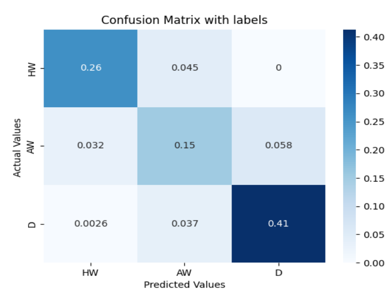

Figure 10: Confusion Matrix in Percentages using Decision Trees obtained from M11

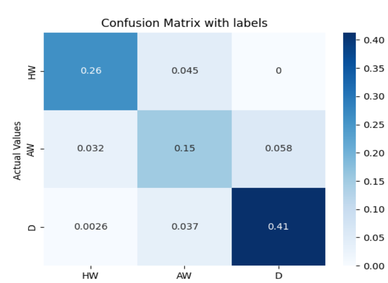

Figure 11: Confusion Matrix in Percentages using XgBoost obtained from M1

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.82	0.61	0.79	0.75
F1 score	0.80	0.46	0.79	0.63

Table 2: Results after using cross-validation M1.

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.87	0.69	0.78	0.80
F1 score	0.83	0.50	0.77	0.67

Table 3: Results after using cross-validation M2.

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.85	0.63	0.76	0.75
F1 score	0.82	0.48	0.76	0.66

Table 4: Results after using cross-validation M3.

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.85	0.68	0.77	0.78
F1 score	0.83	0.51	0.78	0.65

Table 5: Results after using cross-validation M4.

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.87	0.59	0.74	0.71
F1 score	0.86	0.45	0.74	0.64

Table 6: Results after using cross-validation M5.

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.82	0.61	0.73	0.72
F1 score	0.81	0.45	0.73	0.61

Table 7: Results after using cross-validation M6.

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.86	0.67	0.72	0.77
F1 score	0.83	0.49	0.72	0.63

Table 8: Results after using cross-validation M7.

8 ABLATION STUDY

The results of this study could be suspected as overfitting. One main reason this could be possible would be to reduce the total number of features and choose those features which give us a better "bang for your buck". More formally, we need to select those features which have less dependency and prevent the classifiers from overfitting. So, the best 15 features have been selected using the scikit's feature selection package. Decision tree and Logistic regression are the most affected classifier when the selected classifiers are chosen. Some of the solutions for this is explained in future work section. But, it is interesting to see the performance of XgBoost is reasonably good. Overall, we can conclude by saying that SVMs are not very good for this type of study and XgBoost can be a good option althoug requires a lot of parameter tuning in order to obtain the most optimal setup.

	Logistic	SVM	Decision Tree	XgBoost
Accuracy	0.48	0.30	0.64	0.77
F1 score	0.32	0.15	0.48	0.68

Table 9: Results after using cross-validation M1.

9 FUTURE WORK

The work above conducted signifies that making predictions in football is not as easy as it looks. Some future works are as follows:

(1) In this paper, we have only discussed how machine learning classifiers perform. However, this approach can be extended to other avenues like Deep Learning. What we can do is, to mix the Twitter Sentiment analysis, the features we have here and take an over all larger dataset. The current dataset consists of nearly 2.5k tuples. However, for this project to take place in a deep learning setup, we would require more than that. For example, the MNIST dataset consists of 60k training samples. One way to increase our dataset size is by combining data from 1993/94 season all the way to the current 2021/22 season. This gives around 15k samples, ad combined with Twitter dataset, we can get a good representative dataset to perform deep learning on this project.

(2) We have only looked at Decision trees up until now. An extension to this can be made by using the ensemble models like Random Forests. Furthermore, we have observed overfitting for Logistic Regression models. One way to deal with this is using a "penalty" setup model, like ridge regression which imposes a l_2 penalty. It would be interesting to see how such a model would perform in this approach.

This concludes the Future work section of this paper.

10 CONCLUSION

A detailed study on how different machine learning classifiers perform over a dataset is made. From the results obtained, we can say that XgBoost can be a good approach in performing this task, and we should not use SVM in this case as this is a linear model. Furthermore, the confusion matrices showcase how, even when the model puts up good numbers, it is actually not learning and being clever by predicting the obvious decision.

ACKNOWLEDGMENTS

I would like to thank Prof. Ahmed El-Roby, Professor at Carleton University, Canada for his constant support, motivation and guidance torught this project.

REFERENCES

[1] Ryan James Beal, Stuart Middleton, Timothy Norman, and Sarvapali Ramchurn. 2021. Combining machine learning and human experts to predict match outcomes in football: A baseline model. (2021).

[2] Dwijen Rudrapal, Sasank Boro, Jatin Srivastava, and Shyamu Singh. 2020. A deep learning approach to predict football match result. In *Computational Intelligence in Data Mining*. Springer, 93–99.

[3] Robert P Schumaker, A Tomasz Jarmoszko, and Chester S Labedz Jr. 2016. Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems* 88 (2016), 76–84.

[4] Michael Stöckl, Thomas Seidl, Daniel Marley, and Paul Power. 2021. Making offensive play predictable-using a graph convolutional network to understand defensive performance in soccer. In *Proceedings of the 15th MIT Sloan Sports Analytics Conference*.

[5] Fabian Wunderlich and Daniel Memmert. 2020. Innovative approaches in sports science—lexicon-based sentiment analysis as a tool to analyze sports-related Twitter communication. *Applied sciences* 10, 2 (2020), 431.