Predicting Liverpool FC’s upcoming 2021-2022 English Premier league season using Bayesian statistics and Machine Learning

Harry Agyemang
200112345
Prof. Thomas Roelleke
MSc Electronic Engineering Big Data Science

Abstract*— This project focuses on two popular prediction methods, Bayesian statistics and machine learning, while applying this to Football data. This paper will investigate the techniques used in these two areas to predict Liverpool FC’s new season on their points and position in the league and the outcome of their upcoming matches. The projects aim to predict the points that each team should have by the end of the season. It will also create a new dataset of Liverpool FC’s upcoming games is applied to a Bayesian statistics model and a machine learning model to predict if Liverpool's future matches will be a win, draw or away win. The critical aspect is evaluating the model's performance and finding the most suitable methods to approach this subject. The results from this project will benefit bookmakers as the predictions can produce confidence that these outcomes can happen in real life.

I. Introduction

Prediction is mainly concerned with estimating outcomes while forecasting predictions based on time-series data.[1] The project aims to focus on various aspects of football, from the match outcomes to who are the title challengers for this season by points accumulated from games. It will also go into depth analysis of the team performance using a model that can predict the future outcome of matches for Liverpool FC. The challenge of researching this topic is finding the best predictive modelling techniques to address the problem. This project will involve statistical models, such as Bayesian inference, Poisson distribution, machine learning and probabilistic models. One aim is to compare Machine learning algorithms with football data and see how models preforms and outputs predictions on future matches while evaluating each model against a metric such as confusion matrix and f1 score. The models will predict the full-time results from the Dataset containing Liverpool's past matches and other teams in the English premier league. The historical Dataset from Football data provides rows of Liverpool's home and away games against teams in the league, scores at full time and halftime, shots on target and many more. The experiments introduce the data, methods of prediction and results of predictive accuracy. This will then be applied to forecasting football signals and events in the future, such as Liverpool vs Chelsea.

II. Background

A. Bayesian Probability

Bayesian involves statistical methods that assign probabilities or distributions of events based on guesses before experimentation and data collection. We apply the Bayes theorem to revise the probabilities after obtaining data. There are two interpretations of Bayesian probability. The research will be exploring Objectivists that interpret probability as an extension of logic. Bayesian is a systematic study of valid inference rules intending to use variables relating to football, e.g., wins, losses and draws. The probability assigned to the hypothesis can range between 0 to 1. Bayesian inference is a method of statistical inference in the Bayes theorem. It updates the probability of the theory as more information becomes available. The Bayesian inference computes the posterior probability according to the Bayes theorem. Frequency probability is associated with random physical systems, e.g., roulette wheels. In this system, a given type of event tends to occur at a constant rate in the long run of trials—an alternative account of probability in predicting future observations based on past observations. [2]

Figure 1: Formula for Bayesian probability: Bayesian inference for parameter estimation.[4]

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}, \tag{1.1}$$

- A & B are events
- P(A|B) is a conditional probability that A occurs given the event of B has also occurred.
- P(B|A) same as P(A|B) but revered.
- P(A) & P(B) are marginal probabilities of both events a and b.

B. BM25

In Information retrieval, Okapi BM25 is a ranking function used by search engines to estimate the relevance of documents when given a search query. A retrieval function ranks a set of papers based on the query terms that appear in each record. BM25 is a family of scoring functions with different components and parameters. Based on the probabilistic retrieval framework. A probabilistic model is a statistical model which uses probability distributions and is used in such areas, for example, machine learning and data analysis.[3]

C. Machine learning techniques

Machine Learning is the ability to acquire knowledge from raw data using modelling tools to extract it. There are two stages of machine learning. The first stage is learning where the model is built, and the second stage is a deployment where the model is used. A machine learning model completes the task of training a model, which is a model created using data and a quality metric and then fitting the model to the Dataset. The testing stage is the model's performance during deployment to assess how well it does against new unseen data. The goal is to estimate the missing value from a collection of known items, for example, the outcome of Liverpool's future games. In this case, the challenge is to build a model with x predictors, e.g., teams, score and other factors, and y label, e.g., W/D/L or win/ not win. Supervised learning techniques make predictions based on evidence, taking known data inputs and response data, then training the model to generate predictions of the response to the new unseen data. A vital aspect of the project is making predictions on points as each team's previous results will impact next seasons predictions. It develops a predictive model based on both input and output features. For example, this task is predicting a discrete response. [7]

D. *Random Forest Classifer*

Random forest is a flexible and easy to use in machine learning. It does not necessarily require hyperparameters to obtain a result. Random forest operates by building decision trees during training, this is done by taking different parts of the dataset as the training set for each tree. Random forest is versatile and is a popular predictive modelling approach that uses statistics, data mining and machine learning. This includes a decision tree that uses a proactive model which comes from an observation about the item. The branches represent the results and the leaves represent the target values . A supervised learning technique builds a forest that stems from the decision tree and is trained with a bagging method. This method can fit multiple models on different subsets and then combine the predictions from all the models used. Random forest builds numerous decisions trees, merging them to create a more accurate forecast. The model is highly accurate and does not suffer from overfitting problems as it takes the average of all predictions. In addition to this, it can find the relative feature importance, helping to select the essential features for the classifier. Knowing the importance of an element can help decide if it contributes to the prediction process. It's default hyperparameters produce good results such as n_estimators because an increase in trees causes an increase in performance, making the prediction stage more stable. The Random Forest classifier uses a voting system and the prediction with the most votes is selected as the outcome. The classifier forecasts the result with an accuracy score. This model can be used for predicting points per team as it takes the average of scores from previous games and produces an output W/L/D for each game based on historical data.[12]

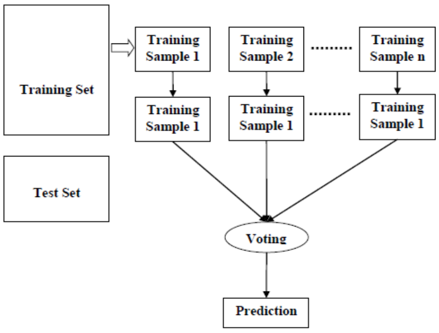


Figure 2: Random Forest[19]

E. *Multi-Layer percepton*

Neutral networks formed from artificial neurons have weighted input signals and produce an output signal using an activation function. Each neuron has a bias and a weighted value. The input layer takes data in more detail and then passes it to hidden layers whereas the output layer is responsible for outputting the value. For binary, it will have a single output neuron and use the activation function to output a value between 0 and 1 and represent this as the probability of predicting a value for a class. In contrast, multi-class classification will have multiple neurons in the output layer for each class for this project e.g., W/D/L. The model works by providing input and forward passing, which gives an output that produces predictions. The neural network can prepare data and relate it to its output variable, which it's trying to predict. The predictive capability comes from the hierarchy structure of its networks. This model is therefore suitable to use alongside football data as it has the capacity to predict the outcome of matches. It is also very robust and easy to use, it includes multiple layers which makes its predictions very accurate. [13]

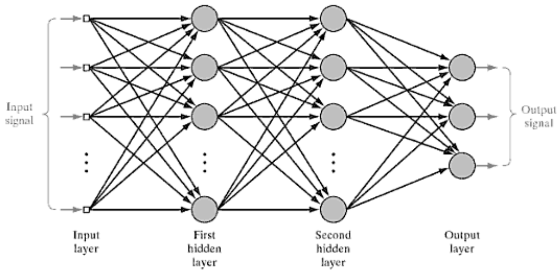


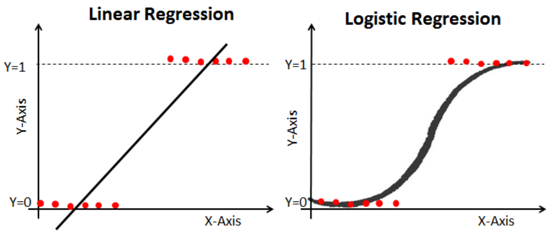
Figure 3: Multi-layer perceptron[20]

F. *Logistic regression*

Logistic regression is another machine learning algorithm applied to the dataset. Logistic regression predicts the probability of a target variable. Typically, this model aims to predict binary target variables for data. The dataset Full-time result column can be converted into binary format, the result has three possible outcomes from the matches. In addition, the logistic function maps the predicted values to their intended probabilities, it uses a sigmoid function which maps values between 0 to 1. Decision bounty determines the outcome of a set of outputs based on probability, the method used is imported through the prediction function. It then produces an output by returning a probability score. For example, win, draw, and away win are decided through a threshold value and classify each output value.

Another example is if there were two classes, win 1 and not win 0. To decide how to classify the values into class 1 with a threshold value and the value below the threshold classified as class two. If the threshold is at 0.5 and the prediction function returns a value of 0.8, then the class of this observation is a class win, but if they prediction returns 0.2, then it would classify these observations as not win. The cost function represents the model's objective. It was created and minimized to develop an accurate model with minimal errors by limiting the cost function between 0 to 1. So, Gradient descent can be used to reduce to reduce the cost value. [14]

Figure 4:Linear and logistic regression [18]



G. *Optimisation function*

Optimisation function is finding a set of inputs to an objective function either maximum or minimum function. For example, to optimise logistic regression a local search optimisation algorithm is used, this finds a set of coefficients for the model which falls into a minimum of prediction error or a maximum of classification accuracy. [21]

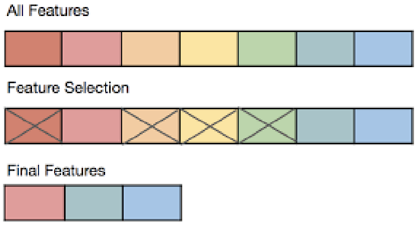
H. *Grid Search for Hyperparameters tuning*

Grid Search is a library function which loops through already predefined hyperparameters and fits the model on to a training set, so it ends up with the selected best parameters from the list of hyperparameters. Grid search is important as its used to find the optimal hyperparameters for a chosen model which results in the most accurate predictions. It builds the best model from every option associated to the model and evaluates it in addition to using the evaluation metric. [22]

I. *Feature selection*

Feature selection helps with creating an accurate predictive model by choosing the features which contribute to the model and will give better accuracy requesting less data. During the processes it can be used to identify and remove any unnecessary attributes that do not contribute to the accuracy of a predictive model which causes a decrease in the accuracy overall. The reason being is fewer attributes the more it reduces the complexity of the model. The three classes of feature selection are filter, wrapper and embedded. Filter selection method applies statistical measure and assigns a scoring to each feature. These are ranked by the score and are either selected to be kept or removed. Wrapper method selects a set of follow the process of being prepared, evaluated, and compared to the other feature combination. Embedded method leans the features which best contributes to the accuracy of the model while its created. One example of this is the regularization method.[23]

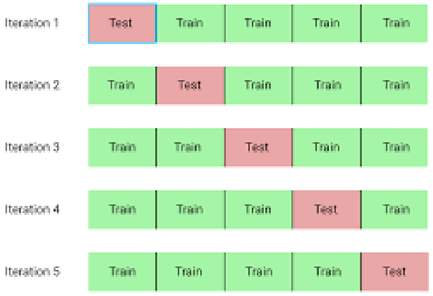
Figure 5: Feature selection[25]



J. *K-Fold Cross-Validation*

Cross validation is a resampling mechanism used to evaluate machine learning models on data samples. This calls a single parameter named k which refers to the number of groups that the data sample is split into. The k value can be a specific such as k=10 meaning 10-fold of cross validation applied to the model. Cross validation is used to estimate the machine learning model on unseen data. It uses a limited sample to estimate how the model is expected to perform when it's used to make predicts on unseen data during the training phase. The football dataset can benefit from cross validation and to test if the model is fit for purpose.[24]

Figure 6 : K-Fold Cross-Validation [26]



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

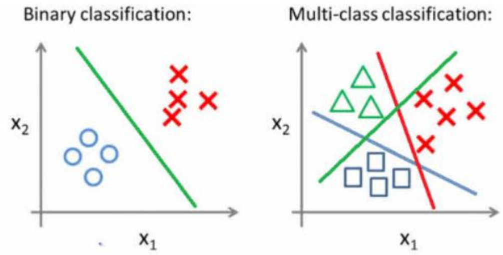
K. *Bayesian prediction*

For Bayesian, it uses posterior predictive distribution and predictive inference on the data. For example, this predicts the distribution of an unseen sample. Many prediction methods would use a fixed point as a prediction but, the Bayesian method does it by distributing the new selection and optimizing predictors in forecasting events and classification problems. Bayesian probability uses probability to predict the likelihood of a future event happening. This measure of percentage shows the confidence in believing that the event will go this way, e.g., Liverpool having the higher rate of winning against Chelsea. Bayes' theorem can use prior knowledge to help compute the probability of winning, losing or drawing [4]

L. *Multi-label classification vs binary classification*

Multi-label classification relates to multi-class. Multi-label consists of classes the that the instances are assigned to, e.g., W/D/L. Classification is a predictive modelling problem that involves class labels with some input. Classification requires training data with an example of inputs and am outputs that it learns from. One form of classification metric is accuracy. The matric evaluates the performance of a model based on the predicted class labels. The aim is to predict a label from the input data e.g., a win/lose or draw given the predictor variables. In comparison, binary refers to a classification task that has two possible outcomes two class labels. In this case, we have won/not win. Not win includes draw or loss. For this project the Full-time results column will be converted to either Multi label or binary. For example, win = -1 and lose/draw = 1. Another would be to class Each W/L/D in a binary format e.g., -1 = win, 0 = draw and 1 = lose.   [5]

Figure 7: Binary and Multi-class classification[17]



M. Objectives

- Produce predictive outcome for Liverpool’s 2020-2021 season results against all teams in the premier league. Use various techniques to predict different other outcomes related to team performance in games and points accumulated over the season to overall judge if Liverpool will be being title contenders.
- Use the approach of Bayesian statistics to predict the outcome of matches whether its W/D/L or win/not win. Explore Bayesian probability and apply it to the scenario. use this method on Liverpool vs Manchester united and display the results.
- To evaluate different machine learning algorithms able to predict Football results and to use F1 scoring, mean squared error and accuracy of the model to measure the correctness of the predictions.
- Discuss the important of both multi label and binary classification for match results.

N. Performance metric - Classification Accuracy

Classification accuracy is calculated by the number of correct predictions divided by all the predictors. The best way to use it is when the number of observations in each class is equivalent, if not it's probably best to use another metric that gives a false sense of high accuracy. [18]

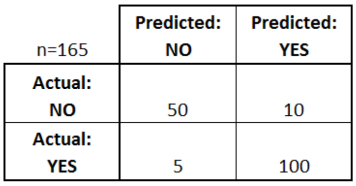
Accuracy formula:

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.2}$$

O. Performance metric - Confusion matrix

A binary classier with two classes belong to either a yes or no. Figure 9 represents the accuracy of the model either being a binary or multi-label. The correct predictions are located on the diagonal line top to bottom.

Figure 8: Example Confusion matrix [16]



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

P. Performance metric - F1 score

F1 score is used for test accuracy, and it’s based on the average between precision and recall. Generally, the score is between 0-1, giving a value on how precise the classifier is and its robustness. The better the F1 score, the better the performance of the model as it tries to find the balance between precision and recall.

F1 score equation[16]

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \tag{1.3}$$

III. Related Research

A. Machine Learning Research

This paper investigates predicting match outcomes in the premier league 06-07. It goes into detail about the performance of past football matches and observing the essential attributes in the game. The algorithms used in this paper are Support vector machines, XGBoost and logistic regression. It then discusses the uses of prediction systems in sport and how Machine Learning algorithms could be implemented in sports betting in the sport. The data is pre-processed, split into training and test and then machine learning classifiers are applied to make a prediction. The model which performed the best was the Support vector machine The target variable is full-time match results, and the labels contain home win, away win and draws, making it a multi-label classifier. For the best performing classier, the hyperparameters are chanced to enhance performance and accuracy of the model and then finally obtain a prediction for the outcome of the match being the target variable full-time result optimizing the classier with the best results. [8]

This paper applies three different Machine learning algorithms, for instance, Random Forest, Neural network, and Support vector machine, to predict match results and then compare their accuracy metric against each model. The focus is on the machine learning approaches because of their applicability in predicting. The aim is to predict expected goals in matches. The Dataset is of the English premier league season of 2005-2006. It follows a process of retrieving data, clearing it, selecting, transforming, performing data mining, and then evaluating it and extracting knowledge. The model uses cross-validation to test its capability on unseen data. For evaluation metrics, it uses accuracy, precision, and recall comparing each Machine learning algorithm classification report results. The decision forest provided the best solution for predicting football results. It exhibited faster calculations comparing it to the support vector machine and neural networks. [9]

B. Bayesian Research

This paper investigates football match outcome prediction. There are three types of approaches to determine the outcome of a football match: Win, Draw, Lose. This paper proses the Bayesian approach within machine learning such as naive Bayes, decision tree, and general Bayesian network to predict the match outcome using premier league match results in three different seasons between 2014-2017. The result showed naïve Bayes achieved the highest predictive accuracy of 90% compared to Bayesian approaches and decision tree. It was successful in predicting the Premier League season 2015-2016.[10]

This paper discusses how modelling football data has become popular; different models proposed aim to estimate the game's outcome or predict the score. This paper discusses using the Bayesian network to predict the result of future matches by performing a Bayesian network to predict the games involving Barcelona FC 2008-2009 season. It also discusses how the results of a football match are dependent on many factors such as team morel, skill, and current score. The Bayesian network approach involves factors that affect the results, such as weather and games. The main factors were split into two groups psychological such as, results against teams and home games and non-psychological, which is the performance of leading players and average goal for home. Overall, it was sucssful in predicting Barcelona' upcoming matches. [11]

IV. Methodology

A. Dataset

This section will discuss the dataset used in more detail and the factors which will help identify the predictors and target features for the project. Factors such as shots on target, goals conceded and more impact the game as they can give possible expectations for the match outcome. For every win, a team scores 3 points, draw 1 point and loss 0 points. The dataset for this project is of the premier league data from 2000-2021; it's publicly available for anyone to use. The database format is a CSV excel file.

Match Information

Name	**Description**
Home Team	Name of the Home team
Away Team	Name of the Away team
FTHG and HG	Full Time Home Team Goals
FTAG and AG	Full Time Away Team Goals
FTR	Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result (H=Home Win, D=Draw, A=Away Win) [6]

Match Statistics

Name	**Description**
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HHW	Home Team Hit Woodwork
AHW	Away Team Hit Woodwork
Referee	Name of the referee for this game
HC	Home Team Corners
AC	Away Team Corners
AF	Away Team Fouls Committed
HF	Home Team Fouls Committed
HFKC	Home Team Free Kicks Conceded
AFKC	Away Team Free Kicks Conceded
HO	Home Team Offsides
AO	Away Team Offsides[6]

Figure 9: Segments which represent each variable of Full-time result for Liverpool FC.(Home team)

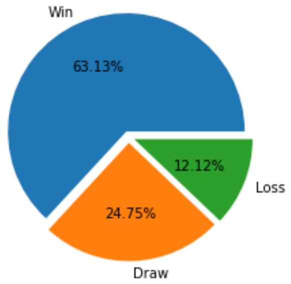
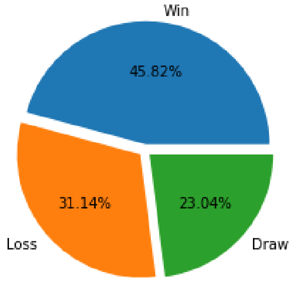


Figure 10: Liverpool FC (Away Team)



B. Data inspection

The first insights are after opening the dataset imported the CSV file using pandas, then explored more of Liverpool FC stats. There are 22 columns defining each match played and the dataset contains 7600 games. Liverpool has 242 wins as the home team, implying Liverpool dominate their home matches and have a greater probability of winning at home. The 2020-2021 season data contains referees and attendance. For future projects adding a home advantage to the predictor variable is essential as home teams tend to perform better with their crowd can affect the predictions. The top referee who has appeared in over 400+ games is Mike Dean. His experience in big games can be considered as a target variable that affects the match outcome.

Name	**Liverpool(Home Team)**
Home Wins (W)	242
Draw (D)	95
Away Wins (L)	43
	Liverpool(Away Team)
Away Wins (W)	118
Home Wins (L)	178
Draws (D)	84

V. Evaluation

A. Predicting the league outcome and points for each team.

Data preprocessing

The first part of the prediction is investigating data from last season to predict the upcoming season as previous seasons would not help with the prediction. The match statistics from last season would provide the best estimate of the strength of the teams now. Initial preprocessing involves columns for each team’s attack and defense and calculating the average of goals scored and conceded. For example, for each score and conceded for home and away team this is divided by 19 e.g., 19 teams in the league. The statistics taken form the dataset to build this model was FTHG and FTAG. Next, the values of each column are scaled between [0 ,1] to be processed by the models, creating four new columns of HMS, HMC, AMS, AMC derived from the initial match statistics. An additional column was created home attacking strength and away attacking strength. Using the inputs home attacking strength was calculated by multiplying HMS with AMC and for Away attack strength was calculated by AMS multiplied by HMC. A new column results is the target variable to predict. The logic behind results is if Home Goal is more than Away Goal it equals -1 if it's equal return 0 and if Away Goal is more than Home Goal return 1. This would eliminate the FTR column as the new column results would be our target variable.

Model inputs

Name	Description
HMS	Home Mean Score
HMC	Home Mean Conceded
AMS	Away Mean Scored
AMC	Away Mean Conceded
HAts	Home Attacking Strength
AAts	Away Attacking Strength

The data is then split into X values and Y values 20% test and 80% training before this data is trained and tested. Firstly, the data is fitted to each model and an accuracy score is printed. Overall Random Forest Classifier preformed the best and then its split into training and test and fitted. The cross-validation test accuracy for the random forest is 53%.To analysis the performance a classification report is used. Wins had the highest precision and f1 score while losses with the highest recall. Prediction was made on the test data produced output. A new column is created and populated with the predicted value against the match results. For example, it correctly predicted the first five games in Figure 13.

B. Evlaution

Random Forest Classifier performed the best and made predictions on the premier league 2021- 2022 season. The model provides the best prediction on where each team will stand in the league and the total points they will receive after each game. A function is created where a table was made to count the number of predicted wins equal to 3 points, predicted draw equal to 1 point and loss would equal to 0 points. Each team would then have a predicted point against their name and the highest would be the champions for the season. From the result, Liverpool would come second in the league behind Manchester City. Figure 14. The parameters used for Random Forest Classier were n_estimators= 10 and criterion="entropy". The random Hyperparameters grid gave these as the best to use. Adjusting the algorithms to optimise its performance turning the model to get the best possible outcome.

Model results

Model	**Training**	**Test**
Multi-Class results		
Random Forest Classifier	97%	53%
Logistic regression	58%	51%
Multi-Layer perceptron	57%	49%
Binary Results		
Random Forest Classifier	98%	68%
Logistic regression	73%	61%
Multi-Layer perceptron	74%	65%

C. Binary classification on prediction the league points

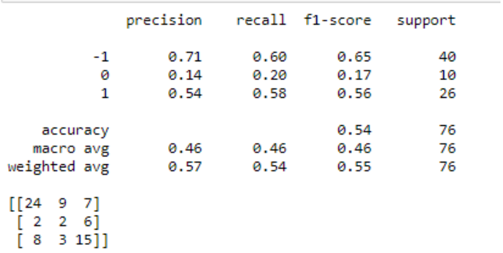
Pre-processing

With the output being W/D/L, to convert this into a binary in this example, win = 1 and D and L = 0. This would form the result column. This follows the same initial pre-processes as multi class, the same model inputs are used but the target variables are binary. The first model trained on the data set is Logistic regression and this achieved an accuracy of 73%.Random Forest achieved 98% and multi-layer perceptron achieved 74%. Overall Random Forest has the highest accuracy rate and is put forward for testing against the unseen data. The data is then split into X and Y values 20% test and 80% training. Using Cross validation, the model produces a test accuracy of 68% then the data was used to make predictions on the new season. The idea behind using cross validation is to split the data into K folds of equal bin sizes to give better estimates on the model overall and how it will perform on unseen data.

D. Evaluation

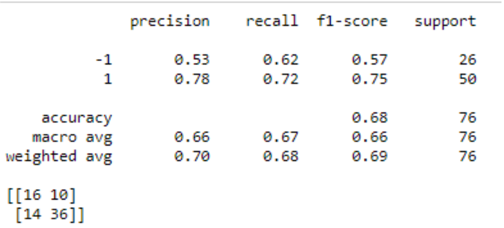
To evaluate the performance a classification report is used on the accuracy of the model gave details on precision, f1 score and recall. Win had the highest precision while win/not win had the highest f1 score and recall. Then created a predictions column and added it to the data frame against the actual results. In figure 14 on the right Liverpool comes second with 87 points, but overall Manchester City are the champions with 96. The results show binary prediction on the upcoming season as it calculates Liverpool winning a lot of games than losing.

Figure 11: Multi label classification report



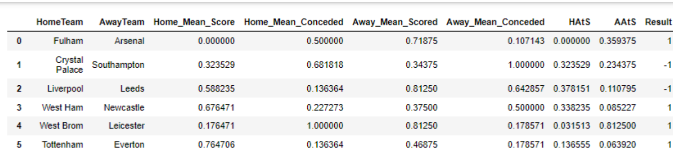
```
              precision    recall  f1-score   support

          -1       0.71      0.60      0.65        40
           0       0.14      0.20      0.17        10
           1       0.54      0.58      0.56        26

    accuracy                           0.54        76
   macro avg       0.46      0.46      0.46        76
weighted avg       0.57      0.54      0.55        76

[[24  9  7]
 [ 2  2  6]
 [ 8  3 15]]
```

Figure 12: Binary classification report



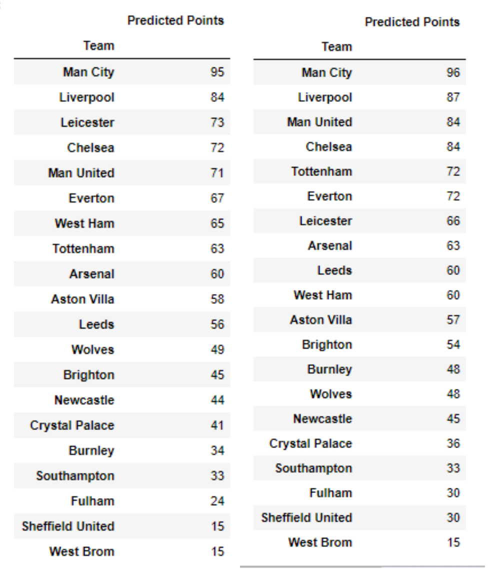
```
              precision    recall  f1-score   support

          -1       0.53      0.62      0.57        26
           1       0.78      0.72      0.75        50

    accuracy                           0.68        76
   macro avg       0.66      0.67      0.66        76
weighted avg       0.70      0.68      0.69        76

[[16 10]
 [14 36]]
```

Figure 13: Merging Predicted to the Data Frame.



	HomeTeam	AwayTeam	Home_Mean_Score	Home_Mean_Conceded	Away_Mean_Scored	Away_Mean_Conceded	HAtS	AAtS	Result
0	Fulham	Arsenal	0.000000	0.500000	0.71875	0.107143	0.000000	0.359375	1
1	Crystal Palace	Southampton	0.323529	0.681818	0.34375	1.000000	0.323529	0.234375	-1
2	Liverpool	Leeds	0.588235	0.136364	0.81250	0.642857	0.378151	0.110795	-1
3	West Ham	Newcastle	0.676471	0.227273	0.37500	0.500000	0.338235	0.085227	1
4	West Brom	Leicester	0.176471	1.000000	0.81250	0.178571	0.031513	0.812500	1
5	Tottenham	Everton	0.764706	0.136364	0.46875	0.178571	0.136555	0.063920	1

Figure 14: Results from the predicted points of each team. Multi-class (Left). Binary predicted points (Right)



	Predicted Points
Team	
Man City	95
Liverpool	84
Leicester	73
Chelsea	72
Man United	71
Everton	67
West Ham	65
Tottenham	63
Arsenal	60
Aston Villa	58
Leeds	56
Wolves	49
Brighton	45
Newcastle	44
Crystal Palace	41
Burnley	34
Southampton	33
Fulham	24
Sheffield United	15
West Brom	15

	Predicted Points
Team	
Man City	96
Liverpool	87
Man United	84
Chelsea	84
Tottenham	72
Everton	72
Leicester	66
Arsenal	63
Leeds	60
West Ham	60
Aston Villa	57
Brighton	54
Burnley	48
Wolves	48
Newcastle	45
Crystal Palace	36
Southampton	33
Fulham	30
Sheffield United	30
West Brom	15

E. *Predicting football match outcomes Machine Learning*

To predict the outcome, the data is collected from previous match results and match statistics from season 2000-2021. The next stage is to pick out features which describe Liverpool's, home, draw and losses over this period, they have a 0.34 per cent of winning at home.

Feature extraction

The first steps included separating the dataset into feature set and target variable. X_all contained all the columns but not Full-time result. But y_all only contained Full time results e.g. Win/Draw/Away Win. The X_all is then standardised. Next, to obtain continuous variables that are integers for the input data, the removal of any categorical variable is processed in a pre-process feature function following the steps of converting the categorical variables into dummy variables, initialising the new output data frame, investigating each feature column, then collects the revised columns and returns an output. After the feature extraction , the next step uses a function to train and test the models while having a performance metric to assess the model.

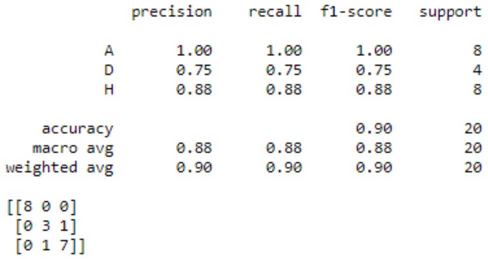
F. *Evaluation*

The model which preformed the best was multi-layer perceptron. Using a test size of 360 to evaluate the model, it was trained and tested return an output of the model performances. A new test data set was used called Liverpool 2021-2021 with their upcoming fixtures. The new dataset was fitted to the model and produce the outputs seen in Figure 24 and provides predictions for Liverpool's upcoming season also provide a probability rate of win, draw and away win for each game. The parameters used for multi-layer were the basic such as activation = relu, alpha = 0.0001 and hidden layer size was 100. These affected the training set as it performed very well. To get the best out of the model hyperparameter tuning was preformed and the model achieved better results from both training, testing and f1 score.

Evaluation results

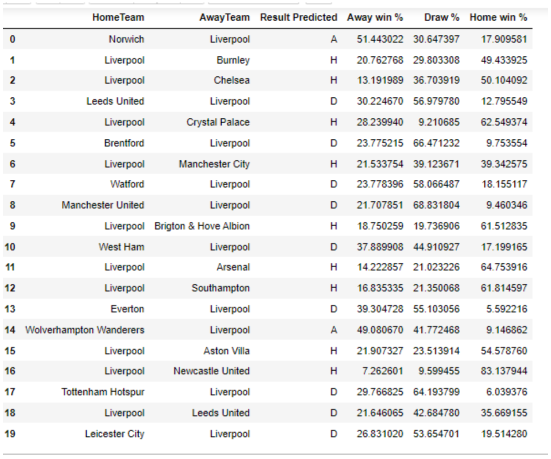
Model	Training	F1 Score	Test	F1 Score
Logistic Regression	95%	96%	93%	95%
Random Forest classifier	99%	99%	85%	74%
Multilayer perceptron Classifier	99%	99%	90%	79%

Figure 15: Confusing matrix for Multilayer perceptron



```
              precision    recall  f1-score   support

           A       1.00      1.00      1.00         8
           D       0.75      0.75      0.75         4
           H       0.88      0.88      0.88         8

    accuracy                           0.90        20
   macro avg       0.88      0.88      0.88        20
weighted avg       0.90      0.90      0.90        20

[[8 0 0]
 [0 3 1]
 [0 1 7]]
```

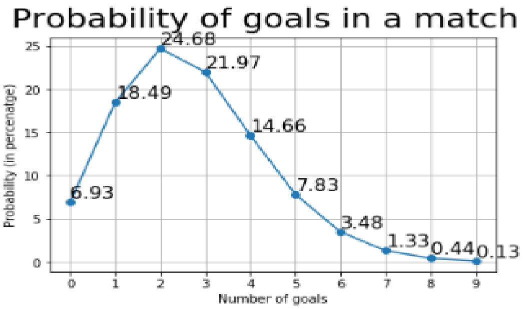
Figure 16 : Results from predicting match outcome with probability.



	HomeTeam	AwayTeam	Result Predicted	Away win %	Draw %	Home win %
0	Norwich	Liverpool	A	51.443022	30.647397	17.909581
1	Liverpool	Burnley	H	20.762768	29.803308	49.433925
2	Liverpool	Chelsea	H	13.191989	36.703919	50.104092
3	Leeds United	Liverpool	D	30.224670	56.979780	12.795549
4	Liverpool	Crystal Palace	H	28.239940	9.210685	62.549374
5	Brentford	Liverpool	D	23.775215	66.471232	9.753554
6	Liverpool	Manchester City	H	21.533754	39.123671	39.342575
7	Watford	Liverpool	D	23.778396	58.066487	18.155117
8	Manchester United	Liverpool	D	21.707851	68.831804	9.460346
9	Liverpool	Brigton & Hove Albion	H	18.750259	19.736906	61.512835
10	West Ham	Liverpool	D	37.889908	44.910927	17.199165
11	Liverpool	Arsenal	H	14.222857	21.023226	64.753916
12	Liverpool	Southampton	H	16.835335	21.350068	61.814597
13	Everton	Liverpool	D	39.304728	55.103056	5.592216
14	Wolverhampton Wanderers	Liverpool	A	49.080670	41.772468	9.146862
15	Liverpool	Aston Villa	H	21.907327	23.513914	54.578760
16	Liverpool	Newcastle United	H	7.262601	9.599455	83.137944
17	Tottenham Hotspur	Liverpool	D	29.766825	64.193799	6.039376
18	Liverpool	Leeds United	D	21.646065	42.684780	35.669155
19	Leicester City	Liverpool	D	26.831020	53.654701	19.514280

G. *Predicting the Score line of a Football Match using Poisson Distribution*

Figure 17: The probability of goals in a match for Liverpool FC.



The dataset used was of the premier league from 2000-2021. To finding the probability of goals in a match, a function is created using the total number of goals in each match. The expected mean goal for matches is 2.67. Using the formular in figure 29 created a visual for finding the probability of k goals in 90 minutes. In figure 17 it's a 24% chance of 2 goals. The Poisson model is a distribution which is used to help find the probability of a randomly occurring event for example matches between Liverpool and Manchester united end in a draw, to find this out if it is true the probability of the observation reads n events in a 90-minute match and expectation of the event occurring within that 90 mins gives a probability value of that chance happening. The Poisson model imports the length of both home and away teams from the dataset. A function named predict score takes the mean home score as the expected goals for home team and the mean away score as the expected goals for the away team and predict the score line. Home team's expected score is calculated by HS + AC / 2 and Away team's expected score is calculated by AS + HC / 2. The mean score for both is then implemented into the Poisson model and simulates the expected score line. The output is displayed in figure 18. [15]

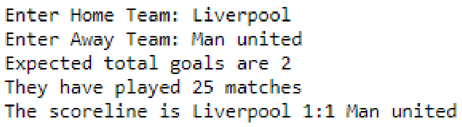
Inputs for calculating predicted Score line

Name	Description
HS	Mean home goals scored
AS	Mean Away goals scored
HC	Mean Home gaols conceded
AC	Mean Away goals conceded

Equation for Poisson distribution

$$P(x) = \frac{e^{-\lambda} * \lambda^{x}}{x\,!} \qquad (1.4)$$

Figure 18: Predicting two teams scores using Poisson distribution.



```
Enter Home Team: Liverpool
Enter Away Team: Man united
Expected total goals are 2
They have played 25 matches
The scoreline is Liverpool 1:1 Man united
```

H. *Conclusion*

This paper has explored Machine learning and Bayesian statistics to predict Liverpool FC's upcoming games. They both provided an in-depth study of prediction methods that are already available. Machine learning techniques focused on three aspects of the problem, predicting points/league position both multi label and binary, also match outcome. While Bayesian helped to discover ways of predicting the scores for each match as well as the outcome. From the results, evaluating the model performance for the machine learning algorithm was important as this determined if the model was fit for purpose. The best model for predicting points using multi label and binary was Random Forest Classifier, this model achieved better results than Logistic regression and multi-layer perceptron. Binary was the better option than multiclass because the performance metrics used to evaluate the model had better accuracy rates than the multiclass.

This project achieved the aim of understanding the match statistics which influence the game and played in a huge role predicting outcome. One negative point is data for teams such as Leeds United was hard to predict as they haven't been in the premier league since 1999 so it was difficult to calculate their features compared to teams such as Arsenal who have constantly been in this league.

Given more time this project could be improved, future developments mentioned here should be considered. A future recommendation would be to apply this to a different league such as Italian and Spanish leagues and to other teams for example Barcelona or Juventus as the model can be applied to any domain. Another suggestion would be to exploring other models such as Deep learning and using different types of distributions e.g., Turkey-lambda and compare each one against a performance metric. Additional changing the target variable to predicting the score would be one area to focus on as this would support the match outcome as there would be predicted scores for both home and away teams as well as a W/D/L outcome. Another recommendation would be to apply different methodological approaches for example feature selection which is reducing the number of input selection when developing a model and improving the overall performance. For example, calculating other features for input such as considering bet 365 odds would improve the inputs for the model and provide a strong process to predict a target variable. Applying feature selections such as the embedded method and implementing regularization to find the best features could improve this project. A final recommendation could be to implementing an optimization function for logistic regression as it can minimize error when fitting a model. reducing error can impact the learning rate and allow the model to learn even faster. This stage is performed during the model selection, and hyperparameter tuning to prepare for predictive modeling.

References

[1] “Prediction vs Forecasting,” Datascienceblog.net, 09-Dec-2018. [Online]. Available: https://www.datascienceblog.net/post/machine-learning/forecasting_vs_prediction/. [Accessed: 16-Jul-2021].

[2] Wikipedia contributors, “Bayesian probability,” Wikipedia, The Free Encyclopedia, 06-May-2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bayesian_probability&oldid=1021832646. [Accessed: 17-Jun-2021].

[3] Wikipedia contributors, “Okapi BM25,” Wikipedia, The Free Encyclopedia, 24-Feb-2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Okapi_BM25&oldid=1008742667. [Accessed: 17-Jun-2021].

[4] J. Brooks-Bartlett, “Probability concepts explained: Bayesian inference for parameter estimation,” Towards Data Science, 05-Jan-2018. [Online]. Available: https://towardsdatascience.com/probability-concepts-explained-bayesian-inference-for-parameter-estimation-90e8930e5348. [Accessed: 16-Jul-2021].

[5] J. Brownlee, “4 types of classification tasks in machine learning,” Machinelearningmastery.com, 07-Apr-2020. [Online]. Available:https://machinelearningmastery.com/types-of-classification-in-machine-learning/. [Accessed: 17-Jun-2021].I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[6] “Football-Data.co.uk,” Football-data.co.uk. [Online]. Available: https://www.football-data.co.uk/. [Accessed: 17-Jun-2021].

[7] Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Edition, Cambridge University Press, March 2020. ISBN: 978-1108473989.

[8] A. Ganesan and H. Murugan, “English football prediction using Machine Learning classifiers,” Int. J. Pure Appl. Math., vol. 118, no. 22, pp. 533–536, 2020.

[9] A. A. Azeman, A. Mustapha, S. A. Mostafa, S. W. Abu Salim, M. A. Jubair, and M. H. Hassan, “Football match outcome prediction by applying three machine learning algorithms.”

[10] M. H. A. Abdul Rahman et al., “Bayesian approach to classification of football match outcome,” Int. J. Integr. Eng., vol. 10, no. 6, 2018.

[11] F. Owramipur, P. Eskandarian, and F. S. Mozneb, “Football result prediction with Bayesian network in Spanish league-Barcelona team,” Int. j. comput. theory eng., pp. 812–815, 2013.

[12] “A complete guide to the random forest algorithm,” Builtin.com. [Online]. Available: https://builtin.com/data-science/random-forest-algorithm. [Accessed: 20-Jun-2021].

[13] J. Brownlee, “Crash course on multi-layer perceptron neural networks,” Machinelearningmastery.com, 16-May-2016. [Online]. Available: https://machinelearningmastery.com/neural-networks-crash-course/. [Accessed: 30-Jun-2021].

[14] A. Pant, “Introduction to Logistic Regression - towards data science,” Towards Data Science, 22-Jan-2019. [Online]. Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148. [Accessed: 01-Jul-2021].

[15] Adrish, “Predict the scoreline of a football match using Poisson distribution!,” Analyticsvidhya.com, 24-Oct-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/predicting-the-scoreline-of-a-football-match-using-poisson-distribution/. [Accessed: 11-Jul-2021].

[16] A. Mishra, “Metrics to Evaluate your Machine Learning Algorithm,” Towards Data Science, 24-Feb-2018. [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. [Accessed: 18-Jul-2021].

[17] A. Band, “Multi-class classification — one-vs-all & one-vs-one,” Towards Data Science, 09-May-2020. [Online]. Available: https://towardsdatascience.com/multi-class-classification-one-vs-all-one-vs-one-94daed32a87b. [Accessed: 28-Jul-2021].

[18] Datacamp.com. [Online]. Available: https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python. [Accessed: 28-Jul-2021].

[19] “Classification Algorithms - Random Forest,” Tutorialspoint.com. [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm. [Accessed: 01-Aug-2021].

[20] Ogail, “Multilayer Perceptron,” Wordpress.com, 10-May-2010. [Online]. Available: https://elogeel.wordpress.com/2010/05/10/multilayer-perceptron-2/. [Accessed: 05-Aug-2021].

[21] J. Brownlee, “How to use optimization algorithms to manually fit regression models,” *Machinelearningmastery.com*, 09-Feb-2021. [Online]. Available: https://machinelearningmastery.com/optimize-regression-models/. [Accessed: 10-Aug-2021].

[22] M. Sharma, “Grid search for hyperparameter tuning,” *Towards Data Science*, 20-Mar-2020. [Online]. Available: https://towardsdatascience.com/grid-search-for-hyperparameter-tuning-9f63945e8fec. [Accessed: 10-Aug-2021].

[23] J. Brownlee, “An introduction to feature selection,” *Machinelearningmastery.com*, 05-Oct-2014. [Online]. Available: https://machinelearningmastery.com/an-introduction-to-feature-selection/. [Accessed: 14-Aug-2021].

[24] J. Brownlee, “A Gentle Introduction to k-fold Cross-Validation,” *Machinelearningmastery.com*, 22-May-2018. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/. [Accessed: 14-Aug-2021].

[25] “Feature Selection: Beyond feature importance?,” *Kdnuggets.com*. [Online]. Available: https://www.kdnuggets.com/2019/10/feature-selection-beyond-feature-importance.html. [Accessed: 14-Aug-2021].

[26] R. Shaikh, “Cross Validation Explained: Evaluating estimator performance,” *Towards Data Science*, 26-Nov-2018. [Online]. Available: https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85. [Accessed: 14-Aug-2021].

VI. Appendix

Figure 19: Correlation matrix demonstrating the relationship between each other.

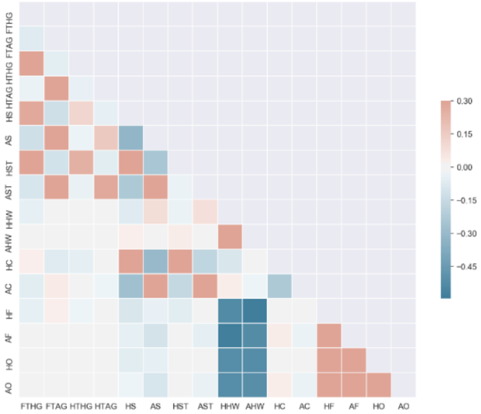


Figure 20: table on confusion matrix

Name	**Description**

True positive	Predicted Yes ad actual output was also Yes
True negative	Predicted No and actual output was No
False positive	Predicted Yes but the actual output was No
False negative	Predicted No but the actual output was Yes[16]

MSc Project - Reflective Essay

Project Title:	Predicting Liverpool FC’s upcoming 2021-2022 English Premier league season using Bayesian statistics and Machine Learning
Student Name:	Harry Agyemang
Student Number:	200112345
Supervisor Name:	Professor Thomas Roelleke
Programme of Study:	MSc Electronic Engineering Big Data Science

Overview

The project was successful as I made it in line with my plan and got the outcome from my milestones, predicting the outcome and league tables using Bayesian and machine learning prediction techniques. I used Jupyter notebook to write all my code in python and produce an outcome where I recorded my findings and displayed them in the project thesis. The critical part was that the data I used match results and statistics, extracting the essential information to build a new data frame. Designed to have every team's average attack vs defence and used this information to predict the outcome in the model. In the Jupyter Notebook cells, I also conducted data analysis but primarily focused on Liverpool, creating visualisation plots on the team.

Strengths and weaknesses

Strengths

- Critically analysed both prediction methods and gained a better understanding of football.
- Used Poisson distribution to predict Football match outcomes and scores for future researchers.
- Used machine learning techniques on football data, can apply this to different seasons, and similar projects may follow this.
- Added an optimisation function to improve the model's performance during training and testing to achieve the best results.

Weaknesses

- I couldn't add more match statistics that affected the game, and the focus was on each team's attack and defence stats, but more should be on other features, e.g., shots, possession, and Free kicks.
- The method should have been training, validate and testing as this is the correct option.
- I should have used hyperparameters turning as I should have done more research into it and improve/optimise the model performance enable to get the best possible result.

- Adjusted the train/test split to different values, e.g., 30/70 and saw how the data performed from there.

Future work

- If I had more time, create a dashboard showcasing my predictions for Liverpool FC for their upcoming season, a visual showing the expected outcome and the league table
- I believe this was a good idea to focus on one team. The project can have different domains, such as Manchester City and predicting how many points they will get in their upcoming games for the new season.
- Using either machine learning or Bayesian to predict in-game stats from players, e.g., how many passes will Paul Pogba make in a game.

Critical analysis of the relationship between theory and practical

During this project, I have spent hours studying prediction methods and different ways of classifying match outcomes. One area I found to increase more in was multi-class vs binary, and my project showed both sides of this argument. In terms of prediction methods, the focus was on using Bayesian statistics to predict a future event. Creating a model was tricky as there was not much research on football match outcomes, but there was for Medicine. I used what I've learnt from Bayesian for this project as medicine uses a similar binary classification approach, but for football, I would have to class win as one but draw and lose as zero. The theory behind Bayesian was interesting as I learnt more about the wide range of Bayesian statistics and how Bayesian can predict future events. In this project, I wanted to implement this. Still, practically it wasn't easy because existing models for predicting football matches and the available ones used a python library named pymc3, which wasn't easy to learn. I had tried many ways to adapt this model to the project but to not much success.

Compared to Machine learning, there are hundreds of materials on the internet that supported my project, and researchers had done this before but by using four leagues or on different teams. The theory behind machine learning is exciting and took me a bit more time to understand how to implement machine learning to predict these outputs. Because of my previous knowledge of machine learning, I was already equipped and had ideas for approaching this. Knowing this was a classification problem, I would need to investigate supervised learning techniques to predict a label from my input. Practically this was very difficult as, firstly, I struggled to know which inputs I needed for this task. I studied a bit more on what others had done and found a way to solve the problem of my project by calculating each team's attack and defence, then scaling for the model.

Predicting the premier league table task went well as I produced predictions on both multi-class and binary. I utilised the input parameters associated with each team, giving their best value in assessing their mean score and conceded value and calculating attack and defence. I also changed the FTR categorical feature into a numerical one. The target variables were either Win/Lose/Draw or Win/Not Win. I attached the prediction to the data frame and created a function to count the number of times a team had won/lose/ draw and then developed a standing table of all teams and their points. I would change,hyper-tuning, the models to get the best performance by changing the parameters as they affect the model dramatically. The results showed that Liverpool would come second behind Manchester city but secure top four champions league positions in another way.

After this period, I trained the model on three machine learning algorithms and computed the f1 score for the training and test set. Later the model was used on the data, and the model made predictions on last season's dataset, which was good results. But the main task for this project was to see how Liverpool Fc would do in the new season. I found a way of doing this by manually putting Liverpool’s upcoming fixtures for the new season in a test.csv file and ran the model on the file to the predictions. The model was a success as it predicted the new unseen data and gave probability rates on the likelihood of a win, lose draw outcome for Liverpool. I then exported the results and attached them to the dissertation. The only way to see if my predictions were correct is to watch Liverpool's new season and compare it against the forecast to see how close the model was to the real thing and compare to a bookmaker’s predictions.

If given more time for this project, I would have done it differently by discovering more online resources with Bayesian predictions. I would also change the input and output parameters to predict scores and predict player performances in tournaments such as the Euros or the World Cup. I believe this task can be adapted to different scenarios in football and come out with different results which may benefit its target audience, e.g., bookmakers. The timescale affected my project as I balanced four modules each term and had exam periods, which made me pause the process. Having more time on this project would help me spend more time implementing as I ran out of time and started writing my report.

Personal development

I am thankful for this opportunity to study for a postgraduate degree at this university; I have achieved my goals and overcome challenges while doing my masters. Coming from a computer science background, I knew I'd have to maintain my standards and work twice as hard. Over the year, I've made tremendous progress, become more familiar with new technologies, and widen my knowledge with the courses I've learned. This project is particularly challenging not having a Mathematics background didn't keep me from succeeding as I was able to learn new areas which I would never have thought I would. This project has improved my research, programming, and critical thinking skills. Lastly, I'd like to thank my supervisor Thomas for his support over this past year.

Legal, social, and ethical issues and sustainability

Legal issue: For legal issues, Clubs often use devices and data analysis to improve the performance of their players and achieve better results. The recent involvement of technology poses a risk in the use of player data. Football clubs collect large amounts of data this records players movements, runs and stats during the 90 minutes. There are legal issues that may happen with technology monitoring these players. Clubs can watch players, heart rate, distance and sprints made in a game; data collected over a season then it is analysed. This technology gives players the chance to perform better in games. Another analysis is performance-based for scouts to make better decisions on signing a player who has excellent potential. One example of a Legal issue is GDPR, which regulates personal data and can lead to harsh penalties if found to violate. Sports teams will have to obtain consent through player contracts, and clubs must state their purpose of using the data, e.g., for performance-related reasons. Clubs analyse players biometric and health data, thus requiring consent to process this. Other legal issues include when a player retires or transfers to another club to play. [2]

Ethical issue: Companies like Zone7 have developed a data-driven AI system that works with football teams. The difficulty is players not allowing them to store and collect data on their data. Many football organisations are investing in data; the reasoning includes improving the player's performance, reducing injuries, and assessing the player's talent. Additional data is collected, e.g., during matches and biometrics such as blood pressure. Clubs are trying to create revenue streams by improving their connection with fans and are not considering more about their data and how private it is. Another ethical issue is that clubs put in players contacts that they will have to wear a wrist strap to track their performance, health, and welfare while away from games and training grounds. If clubs are collecting sleeping patterns from players, there is a potential for a line crossed when collecting private data on players. The data presents an ethical issue as clubs could write contracts where it allows them to track biometrics on plays while they are away from the sport.[1]

Social Issue: The existence and accessibility of football data could be argued to encourage social issues such as gambling. This issue affects finances, interpersonal relations, health and wellbeing, and impact vulnerable people the most. For example, this article states that the Scottish premier league made up to three million selling their data to betting firms. With the recent rise in sports betting, and big firms buying data from clubs to fuel this: People can bet on the next goal scorer to goals scored in a game. Betting is a concern as there is a growing trend in encouraging gamblers to continue betting. Charities spread increased awareness of the problems with football betting and its damage to people's finances and lives. While betting companies promote their brands and increase adverts. Betting companies such as Ladbrooks are paying to be sponsored by professional football clubs, generating extra revenue streams. A company named FootballDataCo sells data on teams' behalf. They sell on data rights of clubs in England and Scotland. The betting companies would buy football stats from FootballDataCo to use. Their high-quality data is sold to the big betting firms for turnover. Its two main markets are one supplying stats to media and bookies, and the other collects information using scouts around the world.

Sustainability issue: For sustainability, football clubs follow strict regulations and provide evidence of waste management, plant-based and providing sustaining transport for fans. This data is essential as clubs need to follow guidelines and give this information. It works by clubs being awarded points per category if they have achieved this, whether in their stadium, training grounds or offices. Each club in the premier league is ranked against this. At the top is Arsenal, Manchester City, Manchester United and Tottenham. The aim is to help reduce the impact on the environment and spread more awareness about these problems and the solutions to combat them. [4]

Conclusion

Overall, I am delighted with how my project went. The success of the project came from my way of planning and setting milestones and achieving them. The project helped me a lot as I created mini-goals and worked towards them to achieve tremendous results in my project. One successful outcome was using a machine learning algorithm that performed well on the data to predict Liverpool's upcoming games, and hopefully, my predictors were suitable. The project includes Bayesian statistics and machine learning which influences predictions with any domain. Still, these techniques are more than adequate for this project and have given good results in predicting match outcomes and points accumulated by each team to see who wins at the end of the season. I have offered a new way of studying a team's upcoming results and using prediction methods such as Machine learning to do it. The only way to determine if my predictions were correct is to watch Liverpool matches in the new 2021-2022 season and compare the results against my forecasts.

References

[1] G. Blackstock, “Football bosses making millions selling the match data used by live bet bookies,” Sundaypost.com, 04-Nov-2018. [Online]. Available: https://www.sundaypost.com/fp/the-numbers-game-how-football-bosses-are-making-millions-selling-the-match-data-used-by-live-bet-bookies/. [Accessed: 17-Aug-2021].

[2] A. Nolan and L. Steele, “Sports technology and the GDPR: data privacy concerns in sports analysis,” Lexology.com, 24-Jun-2020. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=38355c57-547b-4e76-b575-e600d9519d2a. [Accessed: 17-Aug-2021].

[3] O. Daskal, “Data tracking of athletes tests ethical boundaries,” Ctech, 31-Jan-2021. [Online]. Available: https://www.calcalistech.com/ctech/articles/0,7340,L-3890711,00.html. [Accessed: 17-Aug-2021].

[4] D. Lockwood, “How green are Premier League clubs? Tottenham top sustainability table,” BBC.com,25-Jan --2021.[Online].Available:https://www.bbc.co.uk/sport/football/50317760. [Accessed: 17-Aug-2021].