Football Match Prediction Using Machine Learning and Statistical Techniques with a Focus on Formations.

Varun Singh
220054876
Fahad Ahmed
MSc Big Data Science

Abstract— **In his influential book "Inverting the Pyramid," Jonathan Wilson emphasized the significance of formations in football: "Football isn't mainly about the players, it is about the shape, spaces, intelligent deployment of players, and the movement of players in those deployments.". With a game now populated with a plethora of tactical systems and formations, this dissertation aims to find how much impact a formation has statistically, additionally how different formations match up against each other. By adopting various germane metrics such as the formation score and the past form of the teams playing, we dive further into this investigation. Descriptive statistical analysis provides insights into the distribution and trends of different formations and player metrics. Advanced Machine learning algorithms, including regression analysis and decision trees are adopted to uncover meaningful relationships between these metrics and forecast future performances. With accuracies of approximately 80% we are able to ascertain the validity of these models and techniques. The key findings of this dissertation contribute to an ever-growing repository of football statistics providing new insights into formations and will help players coaches and various professionals in the field shed light on football formations and how they match up against each other, this paper also aims to help professionals in the field determine what they can expect from teams and players in the future, in terms of opposition strength, player scouting and player development.**

Keywords—Football Formations, Descriptive Statistical Analysis, Machine Learning Models, Regression, Decision Trees.

I. INTRODUCTION

Football as a sport has always been an exceedingly captivating subject matter for enthusiasts, pundits, coaches, and analysts. Today football has truly become nothing less than a global phenomenon. This explosive growth in the sport is accompanied by various aspects to analyse and dive in to understand the impact on the game itself. Once such aspect critical to the game are the formations deployed by various teams and coaches. The study of formations in regard to predictive analysis is a relatively unexplored concept, however, has gained a lot of traction in recent years, offering a promising avenue for professionals in the field and fans of the sport to help better grasp another extremely crucial aspect of the sport, and help delve into the overlap of sports, data and analytics.

The evolution of football over the years and the trajectory it is on has observed and will continue to observe the continuous refinement of formations and tactics to exploit strengths and counteract opposition strategies, accommodate players of particular play style, which is yet another ever changing aspect of the game. This dawn of data driven methods has assisted the uprise of a new era of football analytics. From understanding key metrics that help players fit into a coaches play style, to making sense of ball distribution patterns, formations hold vital information on guiding a team to success and how various formations match up against each other. There have been various studies carried out to be able to predict the match outcome such as a paper published in 2019 [19] that looks at predicting match outcomes based on opponents faced. However, there have been no such notable explorations of the impact of formations on match outcomes.

This dissertation intends to bridge the gap between the hypothesis that formations impact the outcome of a football match and real match data, using various machine learning models and statistical techniques. It also aims to determine various statistics regarding the sport such as how formations match up against each other and trend of formations over the years. The primary driving force behind this paper is a comprehensive exploration of the diverse tactics and formations to aid in prediction of match outcomes and finally not necessarily reshape but add to the already expansive norms governing football strategies and tactics.

Helping coaches and professionals understand various statistics with regard to formation will have substantial implications off the pitch, allowing them to compare their formation and tactics to an opposing team before a game and can further the information a team has in order to prepare in a sounder fashion. The findings of this study will also help create more opportunities for overlap between sports and analytics. By diving into the relationship that connects formations and match outcomes this study aspires to reveal the intricacies that lie below the synergy between tactics and results.

The task at hand however is a challenging one, as formations are dynamic and makes it difficult to just begin by using the starting formations for this study, however, to cover for this complexity a multidimensional approach is taken into consideration. The crux of this paper is dependent on being able to first distinguish between an offensive formation and a defensive formation. Various online repositories outline various methods of doing so such as the guidetofootball.com [8] the method used for this study is to develop a weighted average, more weight being assigned to positions further up the pitch implied that the greater the value of this formula the more offensive a team is.

Previously used methods for predicting match results include Bayesian Networks used to predict the performances of the team Tottenham Hotspur between 1995-1997, which performed well however fell short as it only helped make predictions for 1 particular team, Gradient Boosting also was

used and had promising results however fell short as it was not able to outperform the bookmaker's predictions [1] and an 'Elo' rating method used to help derive covariates which then are used in ordered logit regression models [2], this study was conducted in (2010) and had successfully performed better than 4 out of 6 benchmark predicting methods. However the main drawback in both studies proved to be the lack of data and the highly competitive nature of the English Premier League. Where smaller and weaker teams have a relatively high frequency of upsets.

The contents of the paper from here onwards are in the order as follows, the next section looks into various studies similar to this paper and related work, followed by the methodology used to arrive at the conclusions that have been drawn out. Finally the empirical findings are presented through a discerning lens and shine light on the impacts of these revelations within the vast world of football, and future work is discussed.

II. Related Work

This section explores the vast world of football and various studies similar to this. We look at how various studies have developed a foundation for analytical learning and led to a plethora of discoveries and findings in the domain of football.

The first study we look at is one conducted by Rahul Baboota, Harleen Kaur titled "Predictive analysis and modelling football results using machine learning approach for English Premier League"[1], this paper explains the authors' process and pipeline which entails exploratory data analysis and feature engineering. Based on various features vital to the game they create a prediction model to predict English Premier League match results. The study focused on gradient boosting classifier and makes use of a rank probability score metric to determine performance. Based on the findings the authors state that week's 6 to 38 across two seasons (2014-2015 and 2015-2016), their top-performing model attained an RPS of 0.2156. However, they contrast this performance with that of reputable bookmakers (Bet365 and Pinnacle Sports), who for the same time period attained an RPS value of 0.2012. Admitting that the model did not perform better than the bookmaker's model, however this paper helped build a rigid and strong foundation on which further study can be conducted.

Another study looked at as a launch vehicle for this study was the paper titled "The Effect of Team Formation on Defensive Performance in Australian Football" [14] this study looks at the various visual data pertaining to football such as GPS data for player movement and video files from 22 matches in a season of Australian Football and aims to investigate whether there is a relation between team formation and team defensive performance. Notably, the study did not find any meaningful univariate differences between favorable and unfavorable defensive outcomes in team formation variables. The research found a crucial team formation characteristic using multivariate modelling techniques, including logistic regression (LR) and decision tree (DT) approaches: larger team breadth compared to length emerged as a key component associated with obtaining successful defensive outcomes. In addition, certain other variables, such as the average distance between players and the centroid of the squad, showed shaky correlations with defensive effectiveness.

Another study that made use of formations was published in (2011) and is titled "The effect of playing formation on high intensity running and technical profiles in English FA Premier League soccer matches" [15]. This paper aimed to examine the effect of deployed formations on high intensity actions and performance during football matches. This paper dives deep into the performance of players in a formation and how they perform at various points in a match, hence providing great insight to the intricacies and working of tactics deployed in a football match. Analysis of three popular formations—4-4-2, 4-3-3, and 4-5-1—shows that overall ball possession is constant in each of them. Players in a 4-5-1 formation engage in less very high intensity running while in possession and more when they don't. High intensity running doesn't differ significantly between formations. Attackers using a 4-3-3 formation run at a high energy level about 30% more than those using a 4-4-2 or 4-5-1 formation. There are differences in passing frequency, with the 4-4-2 formation showing the highest percentage of completed passes. Ball possession and the impact of formation on physical performance show unique defensive and attacking characteristics. Attackers in a 4-5-1 formation display a decline in high intensity play in the second half, presumably as a result of the demands of the formation. Although technical performance is generally stable, passing frequency and technical actions show variations. These key findings help understand formations better and facilitate studies such as ours to further explore the vast and complex niche of football formation analysis.

A study that explores the relation between players and requirements of a formation perfectly is one by Toni Modric, Sime Versic, and Damir Sekulic [16]. The paper aimed to explore the impact of formational tactics on running performances of players. The players this paper studied were Croatian professionals and mainly explores the differences in running performances between a back three defense and a back four defenses.

To compare the running performances for each playing position in the two tactical formations, the authors used statistical techniques, such as analysis of variance and discriminant canonical analysis. Based on tactical formations for various positions, the study discovered considerable variances in running performances. For midfielders, the amount of accelerations and decelerations varied dramatically between the back 3 and back 4 tactical solutions, with the former having more occurrences. Total running distance and high-intensity running were greater for central defenders in a back 3 tactical formations. The running performance of the wide defenders and attackers did not significantly differ between the two formations.

III. Data Selection, preparation and Understanding

In this section, we illustrate the process of data acquisition and preparation vital to the study of our predictive analysis framework for match results. To overcome various inherent data shortcomings and improve the scale of the study a consolidation of two secondary datasets was carried out.

The first dataset used for this study covers all matches held from the 2016-17 season to the 2021-22 season and is an easily

accessible Premier League dataset found in CSV (comma separated variable) format on Kaggle.com [5].

The data set is categorized into various fields containing match statistics for each fixture, some of which include number of goals score by home team, number of goals scored by away team, full time result, fouls, cards, corners and so on. However one key attribute missing from the dataset was the starting formations deployed by the home and away teams in each fixture in the given time period.

Similar datasets to the first dataset have been used for various studies aiming to make predictions about varied aspects of football. One such study [7] aimed to predict the match outcome based on two different models and compare the two, the first being a goal-based model which has a richer dataset and the second being a direct result-based model. The models were built using Bivariate Poisson Regression for the goal-based model and Ordered Probit regression for the latter. The study found that both models were relatively similar in accuracy and the data intensive approach for the goal-based approach had no significant advantage over the other model.

To obtain this information, the only possible solution was to web scrape the online football repository Fbref.com, known for its well organised repository of painstakingly curated football information. Fbref also happened to be one of the few online repositories that contain formation data for premier league games in previous years. Fbref is also a trusted source for the Athletic [18], which is a renowned football news platform. They also provide in depth analysis of teams and games played. The website presented the lineups and formation used by opposing teams for each fixture as shown below.

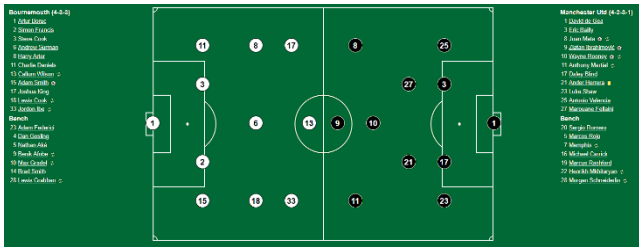

Fig. 1. Fbref.com data representation

The information for the secondary dataset was harvested leveraging web scraping techniques in python, the package beautiful soup was made use of. This process however was an extremely time-consuming task as it meant scraping each individual web page containing information about a single fixture. By undertaking this task, the distillation and production of the secondary dataset was made possible.

Post the web scraping we were left with information that looked something like "<th colspan="2">Brentford (3-5-2)</th>……". To clean and extract the information pertaining to the study we made use of regular expressions in python. The information we were looking for was the home team, away team and their respective formations.

The process involves iteratively traversing the dataset and deploying a set of regular expressions to extract the valuable data. The regular expressions help by removing the unnecessary information in a string. The expression that we were looking for was stored between '>' and '(' symbols.

The union of these two datasets was yet another challenging task, ensuring that the data was joined seamlessly, and each formation of home and away team scraped in the second dataset matched the corresponding fixture in the first dataset was of utmost importance. The process by which this was carried out is explained in detail in the methodology.

This amalgamation of datasets has allowed us to create the dataset that gives us the ability to carry out the study at hand. This section has highlighted the rationale behind selecting the two datasets and why it was necessary to combine them, hence emphasizing the vitality of supplementary data in enriching the analytical landscape. With the required information collected and prepared the following section delves into the in-depth explanation of the methodologies, advanced machine learning model selection, feature engineering, feature extraction and model deployment use to carry out the study and how the conclusions were arrived at.

The two datasets individually contain a vast amount of information that can be used for various studies allowing us to explore various intricacies of the footballing world. The first provides a detailed dive into a football match with various information such as shots on target, corners and fouls for both the home and away team giving us a new perspective on a game of football. The second dataset paints a picture of the types of formation used, how they changed over the years, what teams tend to use what formations, what formations prove to be more successful. However both datasets combined creates the perfect blend required to carry out this study. Details on how this combination was created will be found under data preprocessing in the methodology.

IV. METHODOLOGY

This section of the research explains in great detail the necessary steps required to carry out this study. Covering data pre-processing and augmentation, which entails cleaning, outlier handling and increasing the coverage of the data by means of additional information added as new columns. The second being feature engineering which talks about what features are relevant to the study and in fact have an impact on the outcomes, since correlation does not imply causation hence this step is of utmost importance. Following the feature engineering is the modelling itself, going over what models are used and why and a comparison between models.

A. Data Pre-Processing & Augmentation

During this assimilation phase, preceding the integration the two datasets in this study, extensive data cleaning steps were carried out to ensure that the highest benchmark for data quality and comprehensibility were met. This vital preparatory step helped build the foundation upon which further analysis and modelling took place in this study. We begin by importing all the essential libraries. Missing values and inconsistencies in data were targeted. For any missing values in home formation, a specific strategy of replacing the void values with the mode formation for that particular team in that particular year was employed, this same approach was mirrored for the away formation. In order to tackle the outliers present in the first dataset methodologies such as winsorization and data transformation were made use of as explained in the paper by Kwak S.K. and Kim J.H. [6], done with the intention to mitigate the influence of outliers on the data and any further analysis that took place. Again this process was carried out for both datasets. These data pre-processing strategies were carried out in such a way that the overall integrity of the data remained unscathed.

The second dataset consisting of formation data posed a set of unique challenges with regard to the storage of football formation data in a comprehensible manner. In instances where the second dataset contained some formations having four key positions exemplified by a "4-3-2-1" no change was necessary however, if a formation was comprised of just three key positions for example a "4-3-3" a small augmentation was required. With the addition of a zero-value added to the end of the string, hence making it "4-3-3-0", the data was represented in a standard fashion, simplifying further augmentation and feature engineering necessary for the latter stages of this study.

The combination of the two datasets was a complex task and was of utmost importance that the correct formations for home and away team are matched with the correct fixture. To do so, the csv files were first read into colab.google and converted into pandas data frames, then both the data frames were ordered twice first based on the date such that all fixtures were arranged in chronological order, the second ordering was done by home team. Both sorts were performed in ascending order. This allowed a match to be created between the two datasets, which were then joined.

Once the combined dataset was prepared the data augmentation could be carried out. This step involved creating quantitative columns that allowed for further modelling to take place. Essentially the pre-processing that took place in the data augmentation and pre-processing phase encapsulated the vital steps necessary to bring the datasets up to the required standard for a study of this calibre. This section explored data cleaning and restructuring; the following section looks at the augmentation procedures.

The data augmentation section looks at the additional columns and fields added to the dataset that gives way to further analysis. The first field that we look at is:

Formation Score:

An innovative and fundamental idea for this study was to incorporate the starting formations of football teams in this study to help predict the match outcomes and observe whether there is any significant impact that the aforementioned has on the outcome.

Since the formations are stored as strings after the web scraping procedures, the initiation to the data augmentation and exploration mandated the conversion of the formation data into a quantifiable metrics that can be used further. Given that the starting formations were encoded as strings the first and foundational step involved to first set up a python environment and followed by importing the combined dataset into a pandas data frame.

Subsequently the formation strings were deconstructed and split into the fundamental positions involved in a football team, which are defense, midfield, attacking midfields and forwards. The formations were hence split into four columns, for instance if we have a formation "4-3-2-1" then the resultant split would be 4 in column1, 3 in column 2 and so on, it is also important to note that during this split the formation is typecast to integer to help quantify the data. The key benefit of splitting the formation column lay in its ability to provide a comprehensive critique of the arrangement of players in a formation.

As a result of the split the dataset now contained 8 new columns which were, home-defense, home-mid, home-attack-mid, home-forward, away-defense, away-mid, away-attack-mid, away-forward. The reason behind splitting the columns in such fashion is to be able to determine the formation score of the team, the formation score for this study is defined by how attacking a team is, implying how many players does a formation deploy in attacking position and to quantify this we use a weighted average giving increasing weights to more forward positions, the weights used for this study are as follows, a weight of 0.25 for defenders, a weight of 0.5 for midfielders, a weight of 0.75 to attacking midfielders and a weight of 1 to forwards. The formula can be mathematically represented as:

$$\frac{0.25 * (\text{homedefense}) + 0.5 * (\text{homemid}) + 0.75 * (\text{homeattackmid}) + 1 * (\text{homeforward})}{0.25 + 0.5 + 0.75 + 1}$$

This process was repeated for the away formations as well and we were left with two new columns for the away and home formation scores, these scores help mathematically quantify and differentiate the various formations in the dataset, by creating a scale for how attacking a formation makes a football team, the greater the value of the home and away formation scores, the greater the affinity of a particular team to tend to play attacking and free flowing football. It was observed that bigger teams, i.e. teams in the top six of the Premier League are more likely to adopt an attacking style of play. This also shines light on the stark difference in not only spending power but also competitive ability of the top 6 Premier League clubs as compared to the others in the same league. These trends can be seen in the graphical representation below.

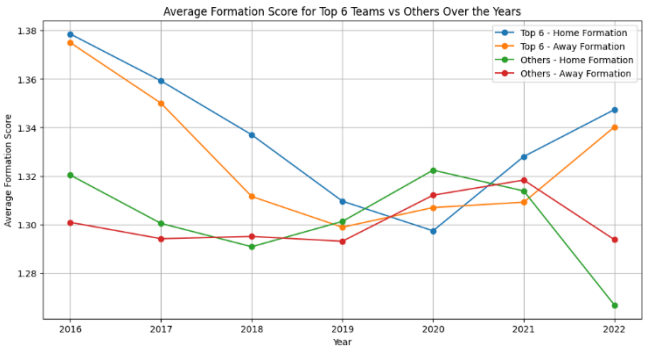

Fig. 2. Graph representing Formation scores over time

The graphs depicts the trend of the average formation scores both home and away for the top premier league teams vs the rest of the league. It is observed that for all years bar one, i.e. (2020) the average formation scores for the top 6 are higher. This outlier in the year (2020) can be attributed to the unprecedented COVID pandemic that left football stadiums empty with no fans, which may have led to a significant decrease in advantage to home teams. A study published in September (2021) [9] takes a deeper look at the home advantage in the year (2020) and found that in stadiums void of an audience, referees were more frequent in penalizing the home team. With regards to team performances, teams scored more points when playing at home compared to away, but the difference was less noticeable when there was no audience. As per the study, prior to COVID, teams gained 0.39 points per game on average more at home than on the road, but after the absence of spectators, this HA was reduced to just 0.22 points.

To summarize, the steps carried out in the first part of data augmentation have been done so to allow the formations of both home and away teams be represented in a quantitative manner, allowing them to be used not only for further analysis

but also be able to include the crux of this study to our models which will be described in detail in the modelling section.

Home team and Away team form:

The form of the teams playing can be defined for this study is the match results for the past five fixtures of that particular team. The form of a team quantified not only aids in developing the model as a key predictor but also serves as an illuminating gateway into the recent historical trajectory of Premier League teams competing with one another. This new column also undergoes a series of meticulous steps to bring the field up to a standard that can be used further in the modelling section.

To retrieve the results of the previous five games of the team under consideration it was essential to perform an in-depth traversal of the dataset. For this study we have considered the form to be calculated based off five most recent matches to follow the standard presented of information depicted on the official Premier League website "premierleague.com/tables" [17]. It is important to note that the formatting of these match results is done to encode the match results of every match whether it be a win, draw or loss. The steps followed to reach the final result, which is a single field that contains the record of the past five matches in the format, example "WWWLD", the first letter from the left being the most recent fixture are illustrated below.

Encoding of the Full Time Results Field:

The first step is to encode the results in a manner that is firstly easy to concatenate and secondly is easily understandable. The encoding that is carried out is as follows, a win for the home team is encoded as a H and a win for the away team is an A, a draw as a D for both home and away and finally a loss for the home team is A and a loss for the away team is encoded as H. This process may seem relatively unimportant however greatly impacts latter stages of this study by reducing the overall complexity of this study, making processes such as modelling and feature engineering far simpler.

Concatenation:

Post the encoding our target is to be able to concatenate the fixtures into a single string. This is done yet again to simplify the dataset by reducing the overall number of columns that we would be working with. Instead of having the ten columns first "fixture home team", "first fixture away team" and so on, post the concatenation we have only two columns which are "home team form" and "away team form". Yet again having an essential impact to the complexity of this study, facilitating more efforts to have a comprehensible study.

Quantification and Transformation:

A crucial step in reaching our final goal in the data augmentation phase is the complex process of quantification and transformation. The two columns of match results that we obtained in the last step had to now undergo this step to reach the benchmark that was necessary for this study.

The first step in this stage was to create distinct dictionaries to store the changing form sequences for the home and away teams. The historical results of games are painstakingly recorded as we iteratively traversed the dataset. For both home and away teams, a sequence is created by appending the outcomes of previous matches to the existing sequence. For example, at the second game week the only past match would be a single game hence the sequence will only contain one match result, however in game week three the result of game week two is appended to the sequence. This created a comprehensive representation of performance with respect to time.

Two distinct sequences, each depicting a team's recent performance trend, are formed at the culmination of this stage. After making sure that the sequences and formatting are securely in place, we carried out a series of disciplined replacements in order to guarantee uniformity across the dataset. The focus is primarily on the letters used to represent wins and losses with respect to home and away teams. The current syntax of H for a home win and A for an away win is transformed to a simple W representing a win. On the other hand, a loss is now transformed into an L from H representing an away loss and A representing a home loss. This step is crucial in creating a set format and helps avoid confusion in the latter stages.

The mathematical quantification of these form strings serves as the culmination of the data augmentation process as a whole. We quantify this sequence by means of a cleverly designed weighted average process, which serves to give greater weightage to more recent matches as compares to matches played further back in recent past. For this study the values of results are defined as follows, a win is set to have a value of three, a draw is 1 and a loss is 0. The formulae used to thereby calculate the form score for home and away teams are:

$$\frac{5 * (g5) + 4 * (g4) + 3 * (g3) + 2 * (g2) + 1 * (g1)}{5 + 4 + 3 + 2 + 1}$$

For this formula g5 represents the most recent game and g1 is the oldest game in a record of five games.

To conclude the addition of home_form, away_form, home_formation_score and away_formation_score have helped create the right fields necessary for this study. These additions make use of the immense power the python programming language and environment has to offer but also incorporates a knowledge of football dynamics. Thanks to this multifaced process the study was ready for the next phase.

B. Feature Engineering

Feature engineering by definition is a vital process carried out in the pipeline of developing a machine learning model, essentially it refers to the process by which new features are created. It also encompasses transforming existing features in a dataset to amplify the performance of a machine learning model. Feature engineering is an extremely vital step as it is directly proportional to the machine learning model's ability to learn from patterns in the data. Hence carrying out this stage in a meticulous and careful fashion can have a substantial impact on the effectiveness of the model itself. Key aspects of feature engineering include, Feature selection, Feature transformation, feature creation and feature

scaling. However, for this study our main focus will be feature selection.

We initiate this process by first loading the combined dataset with the additional fields created in the data augmentation phase into a python environment specifically a pandas data frame. Following this our first order of business it to identify the non-numeric columns that can be used. Upon inspection of that data frame, these columns are found to be the “Home team” and “Away Team” columns which contain the names of the teams that played the fixtures, these teams are then converted into numerical fields using One Hot encoding, which is a process by which each unique string is given a unique numerical value. The resultant fields are a matrix of encoded features.

The original dataset is then combined with the encoded features using concatenation to create a composite DataFrame known as "final_df." The categorical columns "HomeTeam" and "AwayTeam" are then dropped from the "final_df" since their encoded equivalents now appropriately capture the necessary information.

Following this a correlation matrix is then built using the features found in the "final_df" as a foundation. The pairwise associations between the attributes are captured in this correlation matrix and are expressed as absolute correlation coefficients. The set threshold value of 0.65 is used as a standard to gauge how closely two features are correlated. The matrix is scanned to find pairs of highly linked characteristics if it meets these criteria. Particularly, columns with correlation coefficients over the set threshold are identified as being highly associated.

We also make use of the SelectKBest and f_classif libraries in python to allow us to carry out feature selection based on the ANOVA f-value score. We create a selector object using f_classif as the evaluation metric and then proceed to fit the selector object with the training data. The ANOVA f-value score is used in this study as it has better performance in classification tasks such as the task being performed in this study.

This process allows us to find the exact fields that have the greatest impact on the full-time result of a fixture. This process allows us to finally move on to the modelling stage in the methodology.

C. Machine Learning Modelling

The final part of the methodology involves the machine learning modelling phase, here we look at the specific techniques made use of to develop the model which will help us predict the full-time result based on the highly corelated features we found in the last phase. The machine learning models used in this study are the following

1) Logistic Regression

The logistic regression model is an extremely common statistical method used in modelling the relation between a dependant and an independent variable. The models output is particularly well suited for predicting the probability of an event or the likelihood of an outcome. We make use of this model in this study as the logistic regression model is well suited to binary classification tasks, which is the aim of this study, to be able to predict match outcomes.

There have been multiple studies where Logistic Regression has been made use of for the subject matter of predicting football match outcomes. One such study published in (2017) [10], looks at using match prediction with regards to four different fields, namely Home defense, Home attack, Away defense and Away attack. The model developed in this study was trained on years in batches of 5 from 2010-2015 and 2011-2016. The models greatest accuracy was 69% and was also affected by the season 2015-2016 where there were a great number of unexpected results.

2) Random Forest Classifier:

Random Forest Classifier is another model used frequently in the domain of football match prediction; it is classified under ensemble learning implying it combines multiple base models to produce a more refined model. These base models are decision tree algorithms. Some advantages to Random Forests are that they are less prone to overfitting and generally produces higher accuracy. Since the random forest has these advantages and is suitable for this task, we have selected it to compare to the other models to see if there is any improvement in accuracy.

A particular study where random forests were used published in (2019) [11], looked to be able to predict match results such that investors and other higher ups have a more informed way to make decisions regarding a football club. The model used in this study achieved an accuracy of 70%.

3) SVC

Unlike the two models described previously, the support vector classifier us an ML algorithm that can be used for both binary and multiclass classification problems, whereas the previous models can only be used for binary classification. This model is classified under supervised learning that aims to find a decision boundary that divides the classes. An SVC model performs better the more the number of features are used and is quite effective with regards to outliers. Having said this a major drawback of the SVC model is that it is computationally expensive for large datasets and has been included in this study to serve as a comparative model.

SVC have been used in football prediction, in a paper published in (2017) [12] SVC’s were made use of for the classification of passes using temporal data. The study aimed to classify a pass as good, bad or okay and produce an automated system that is able to do so. The model has an accuracy of 90.2%.

4) Gradient Boosting Classifier

Like the random forest classifier a Gradient Boosting classifier also belongs to the ensemble ML algorithms and can be used for multiclass and binary classification. This model makes use of gradient descent to minimize the loss function. This is the model that gave us the highest accuracy in this study and has been used in multiple studies pertaining to football as well such as, a paper published in (2019) [13] that aims to predict match results with a focus on tree-based model classification. This study makes use of various features already present in the dataset such as half-time home goals, home and away shots. The study achieves an accuracy of 64.87% with gradient boosting and we look to build on this. The gradient boosting classifier has been used in this study to not only be able best capture the relationship between formations and match results, but also serve as a benchmark for the other models and see how they compare.

The target variable “y” and the features “X” are the two categorisations of the dataset. An assortment of crucial elements, such as half-time goals, formation scores, team formations, shots on target, and corner kicks, were included in the carefully selected features. The feature selection procedure trains the model on elements thought to be important for forecasting the goal variable, Full-Time Result “FTR”.

Data segmentation was the primary component of this algorithm, making sure that model performance is checked with respect to an independent test set. This was achieved by making use of the 'train_test_split' function, which separates the data into training and test subsets while maintaining the integrity of the underlying patterns and ensures they were valid.

Regarding feature selection, the code uses the SelectKBest algorithm and the f_classif “ANOVA F-value” score metric. Using this method, it was possible to accurately pinpoint the traits that have the most potential for prediction. The number of top features to choose was set by the 'k' parameter, further focusing the model's attention.

A Logistic Regression model “model 1”, a Random Forest Classifier “model_2”, an SVC model “model_3” and a gradient boosting classifier model “model 4” were created, to make use of the predictive power of the chosen features. Resulting in a complex alignment between the model architecture and the data. The model with the greatest accuracy was selected. In this study the Gradient Boosting Classifier model proved to be more accurate by achieving an accuracy of 81%. The model performances are measured using the four primary indicators, which are, accuracy, f1 score, precision and recall.

	Accuracy	F1	Precision	Recall
Logistic Regression	0.77193	0.75662	0.75575	0.77193
Random Forest	0.78216	0.77736	0.77645	0.7821
SVC:	0.77193	0.74959	0.75604	0.7719
Gradient Boosting	0.81725	0.81097	0.81106	0.77193

Table. 1. Table representing model performances

V. Results & Discussion

The table shown in the previous section outlines the performances of all the models used in this study to predict the match outcomes with a focus on the formation a team adopts. From the table it can be noted that the gradient boosting classifier has the greatest performance predicting the correct outcome 81% of the time.

This model was used to predict the outcomes of fixtures for the 2022-23 season. The graph show below:

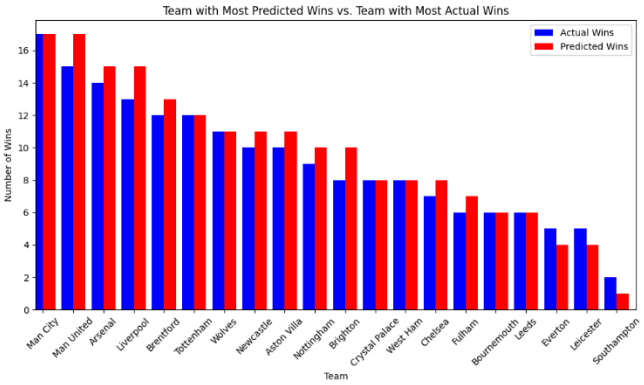

Fig. 3. Graph showing Actual vs Predicted results

Represents the predicted verses actual number of wins per team, in the Premier League for the season 2022-23. For most teams the model predicts a higher number of wins than what is actually observed.

The model is able to closely predict premier league outcomes and this study has helped build a foundation to incorporate formations as metric in the prediction of football fixture outcomes. The model developed in this study can be used by footballing professionals to help make informed decisions regarding things as formations deployed before an upcoming fixture. This model can also be used by bookmakers to predict wins or losses.

This study has not only built a model to predict match results but also provides vital statistics regarding football formations such as what are the most used formations in the topflight of British football. It also provides an in depth as to what teams deploy what kind of formation the most. As a result of this study we aim to publish this dataset on Kaggle for future collaborations and work. This study has allowed us to be able to discover information such as how successful home and away formation have been over the years; this is depicted in the graph shown below:

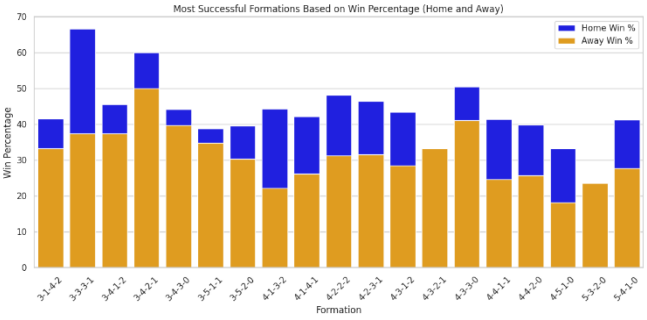

Fig. 4. Graph showing number of wins for each formation

The graph shows us the number of wins for each formation deployed in the premier league and we observe that the 3-3-3-1 formation has been the most successful with and almost even home win percentage and away win percentage.

The heatmap below allows us to understand how formations match up against each other. Based on the percentage of wins this study has allowed us to uncover vital information regarding formations that will greatly impact how teams decide strategies before an upcoming fixture.

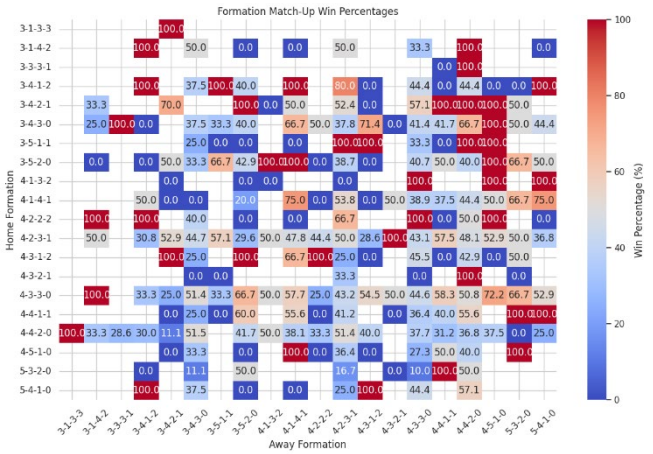

Fig. 5. Heatmap depicting formation head-to-head

Having significantly improved accuracies achieved in the past for binary predictions of football match results, a big advantage that this model brings forth is that it is able to make multiclass predictions, unlike models in the past this model is not limited to making predictions based on just wins or loses but also considers draws.

VI. Conclusion & Future Work

This paper serves as foundation for future work and facilitates a deeper dive into football match prediction. Based on the study carried out we have arrived at the conclusion that formations do in fact impact the model in predicting the match results of premier league fixtures. This can be observed by the improved accuracy in predicting the same. This study enables us to dive deeper into understanding football formations. We have been able to not only improve fixture result prediction accuracies but also been able to create a readily available football formations dataset, display that starting formations do indeed have an impact on the match outcome and also provide valuable insights into the facet of formations such as, how formations compare to one another in terms of win percentages and how they match up against one another.

However this paper barely scratches the surface as there remains a vast number of factors that can be explored. There are a few limitations to this model as well, the first being the lack of data, football being the ever changing and dynamic sport that it is as the years go by historical data becomes increasingly irrelevant. The types of data being captured should be increased and diversified with progression in ML models and analytical techniques there is no limit to what can be analysed and studied. This brings us to the second point.

The Data considered for this study is not dynamic and to match the sport as best as possible it should be. For future projects the data collected to model Football formations as best as possible should contain a mix of data collected for this study and a variety of dynamic data such as video data of matches played, GPS data of player movement. This will allow us to be able to consider how a formation changes over the course of a match. Making the paper more well-rounded as it would then encompass the entire match played.

Currently deployed on Premier League data the next step would be to test the model on the various other topflight leagues in the world such as the Bundesliga and La Liga. Doing so would be able to help us in highlighting the key differences and similarities between the various leagues in the world.

A study that can greatly impact the transfer market in the footballing world is to understand what players fit what formations and what are the key factors that make a player suitable for a particular style. Going of this study we have already been able to rate a formations offensive and defensive affinity, if we are able to highlight a player that fits a particular system this will greatly help clubs make informed decisions when investing in new talent.

VII. References

[1] Baboota , R. et al. (2018) Predictive analysis and modelling football results using Machine Learning Approach for English Premier League, International Journal of Forecasting. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0169207018300116#preview-section-references (Accessed: 06 August 2023).

[2] Hvattum, L.M. et al. (2009) Using elo ratings for Match Result Prediction in association football, International Journal of Forecasting. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0169207009001708 (Accessed: 06 August 2023).

[3] Mesoudi, A. (2020) Cultural evolution of football tactics: Strategic Social Learning in Managers' choice of formation: Evolutionary Human Sciences, Cambridge Core. Available at: https://www.cambridge.org/core/journals/evolutionary-human-sciences/article/cultural-evolution-of-football-tactics-strategic-social-learning-in-managers-choice-of-formation/D5482D76F9355FC314EA1811EF96BC7D (Accessed: 05 August 2023).

[4] "Football statistics and history," FBref.com, https://fbref.com/en/ (accessed Jun. 19, 2023).

[5] Alvin (2022) English Premier League (EPL) results, Kaggle. Available at: https://www.kaggle.com/datasets/irkaal/english-premier-league-results (Accessed: 08 August 2023).

[6] Kwak, S.K. and Kim, J.H. (2017) Statistical Data Preparation: Management of missing values and outliers, Korean Journal of Anesthesiology. Available at: https://synapse.koreamed.org/articles/1156715 (Accessed: 08 August 2023).

[7] Goddard, J. et al. (2004) Regression models for forecasting goals and match results in association football, International Journal of Forecasting. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0169207004000676 (Accessed: 09 August 2023).

[8] *Formations | Guide to Football*. (n.d.). Guide to Football. https://www.guidetofootball.com/tactics/formations/#:~:text=An%20attacking%20formation%20is%20a,relatively%20fewer%20advanced%20playing%20positions.

[9] D. McCarrick, M. Bilalić, N. Neave, and S. Wolfson, "Home advantage during the COVID-19 pandemic: Analyses of European football leagues," *Psychology of Sport and Exercise, vol. 56, p. 102013, Sep. 2021, doi: 10.1016/j.psychsport.2021.102013.*

[10] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), Penang, Malaysia, 2016, pp. 1-5, doi: 10.1109/ICAICTA.2016.7803111.

[11] Pugsee, P., & Pattawong, P. (2019, August 28). *Football Match Result Prediction Using the Random Forest Classifier | Proceedings of the 2nd International Conference on Big Data Technologies*. ACM Other conferences. https://dl.acm.org/doi/abs/10.1145/3358528.3358593

[12] Chawla, S., Estephan, J., Gudmundsson, J., & Horton, M. (2017, August 10). *Classification of Passes in Football Matches Using Spatiotemporal Data | ACM Transactions on Spatial Algorithms and Systems*. ACM Transactions on Spatial Algorithms and Systems. https://dl.acm.org/doi/abs/10.1145/3105576

[13] Yoel F. Alfredo, Sani M. Isa, "Football Match Prediction with Tree Based Model Classification", International Journal of Intelligent Systems and Applications(IJISA), Vol.11, No.7, pp.20-28, 2019. DOI: 10.5815/ijisa.2019.07.03

[14] M. F. Aarons, C. M. Young, L. Bruce, and D. B. Dwyer, "The effect of team formation on defensive performance in Australian football," *Journal of Science and Medicine in Sport*, vol. 25, no. 2, pp. 178–182, Feb. 2022, doi: https://doi.org/10.1016/j.jsams.2021.09.002.

[15] P. S. Bradley *et al.*, "The effect of playing formation on high-intensity running and technical profiles in English FA Premier League soccer matches," *Journal of Sports Sciences*, vol. 29, no. 8, pp. 821–830, May 2011, doi: https://doi.org/10.1080/02640414.2011.561868.

[16] Modric, T., Versic, S., & Sekulic, D. (2020, December 10). *Position Specific Running Performances in Professional Football (Soccer): Influence of Different Tactical Formations*. MDPI.

[17] *Premier League Table, Form Guide & Season Archives*. (n.d.). Premier League Football News, Fixtures, Scores & Results. https://www.premierleague.com/tables?co=1&se=489&ha=-1

[18] Spiers, T. (2023, August 25). *Can Richarlison's problems be blamed on Spurs' system?* The Athletic.

[19] Bilek, G., & Ulas, E. (2019, October 28). *Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators*. www.tandfonline.com. https://www.tandfonline.com/doi/abs/10.1080/24748668.2019.1684773