Citadel Data Open Report

Raghav Gupta, Chang Chuan Hong, Joshua Gei, Yuen Tat Li

1. Topic Question

Association (club) football is one of largest and most lucrative sports in the world, boasting a market size of over US$34 billion [1] and top european leagues averaging over 70 million unique viewers per game [2]

With the enormous amount of money and attention of football, many clubs strive to find a 'secret sauce' behind winning matches - be it offloading US$263 million for a world-class striker (the current world record), or trying to acquire the services of top coaches to implement the best tactical playing styles for the team.

With this in mind, we pose the following topic question in analyzing the dataset:

Topic Question: What factors contribute to the performance of football teams in matches?

We will be contacting a deep dive analysis into 2 areas in particular:

1. What are the optimal team attributes and playstyle of a football club to achieve best performance?
2. What attributes of players affect a football team's performance in a given match?

By answering these questions, we hope to gain deeper insight into the effects of various factors in affecting team performance, and will be able to devise optimal strategies (e.g. optimal transfer strategies) in order to best improve teams' performance.

2. Executive Summary

Through our analysis on the data, we obtained the following insights:

Firstly, we successfully managed to clean the given data of missing values and errors, enriched it with additional data (historical football ELO ratings), and performed feature engineering to obtain a clean, functional dataset. We then validated that with the features engineered we could obtain substantive accuracy by training our model on the team attributes data to predict match results, and found this was preliminarily most accurate with a logistic regression model.

Thereafter, we proceeded to perform an investigation on the **intrinsic factors** (team composition, player attributes, and team attributes) as well as **extrinsic factors** (game of season, location of match) affecting team performance.

For intrinsic factors, we performed statistical analysis, such as logistic regression coefficients, to identify team attributes that were predictive of match outcomes. Furthermore, association rule mining was carried out and an association rule graph generated, showing **stronger association using team attributes than player attributes**. We then looked into the correlation between team and player attributes using the Pearson method, and determined a weak linear correlation between team attributes, providing legitimacy that team attributes generate better association rules as two datasets are nearly independent, thus showing the **importance of playing strategy and team playing style over individual player attributes. Teams that were aggressive and better at creating opportunities were rewarded.**

For extrinsic factors, we were able to obtain strong evidence showing a **clear home-team advantage**, and applied a T-test verification to further validate our results. Finally, we also showed that the **stage of the season influenced match results**, with decisive results (wins/losses) occurring much more often towards the end of seasons.

3. Technical Exposition

We began our data analysis by firstly cleaning the dataset of missing values and anomalies (Section 3.1). We also searched for potentially related factors that could impact performance to further enrich the datasets provided. To this end, we supplemented the datasets with historical football ELO ratings to provide an objective measure of teams' performance (Section 3.2).

The datasets were then reconciled together, and exploratory data analysis and feature engineering were carried out (Section 3.3). To this prepared dataset, we first determined the predictive accuracy of various algorithms on the target variable (home team winning), and then examined various team attributes in predicting team wins, both with and without inclusion of historical team ELO rating data. (Section 3.4)

Thereafter, we performed a thorough analysis on the 'intrinsic factors' that affect team's performance - the player attributes and team play styles, by use of association rule (Section 3.5).

Lastly, we performed further analysis on the 'extrinsic attributes' that affect team performance, focussing in particular the effect of the "stage" of the match on teams' performance, and provided a verification of home-team advantage using a T-test (Section 3.6).

3.1 Data Cleaning

We placed a heavy emphasis on dealing with the errors and missing values in the given datasets. This is because highly quality datasets are essential for drawing valid conclusions even before analysis can be done.

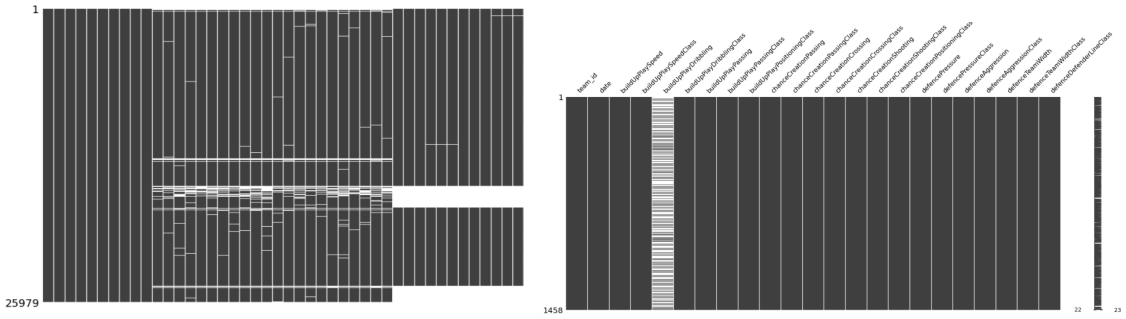

Fig 1.1: Missing values in 'Matches' (left) and 'Team Attributes' (right) datasets

Fig. 1.1 shows a visual representation of the missing values in the given match.csv (excluding the betting data) and team_attributes.csv datasets, where the white rows indicate absence of data.

3.1.1 Missing Data Imputation

7 separate datasets were provided, but only three datasets - the 'Team Attributes', 'Player Attributes' and 'Matches' datasets had missing values.

We chose to drop rows with missing values as opposed to inputting with mean or median values due to the risk of biasing the dataset 'look-ahead bias'[3]. This is because of the time series nature of the match, team attributes and player attributes datasets. Furthermore, the missing rows amount to less than 10% of the given data which is illustrated visually in Fig 1.1, allowing sufficient remaining data even after the removal of these rows.

3.1.2 Combining Datasets

When joining the matches and attributes datasets, we selected **latest attribute values on the date of the match**. This also works to avoid biasing the dataset with future data.

3.1.3 Dealing with Categoricals
Categorical variables were one hot encoded with the different classes as new columns.

3.1.4 Dealing with errors
Several errors in the datasets were identified, with noteworthy ones as follows:

- **Player Attributes**: In the attacking_work_rate and defensive_work_rate categorical variables, other than the intuitive “high”, “medium”, “low”, there were minor categories such as “_0”. However, what led us to conclude they are likely data entry errors were the presence of some categories, such as “norm” in attacking_work_rate and “ormal” in defensive_work_rate with the *same frequencies*. Intuitively, they make up the word “normal” and together with their co-occurrence lead us to conclude the above. Furthermore, they are uninterpretable. As such, we chose to set these categories to null.

- **Player**: The height of the player dataset is likely binarized and not of their actual heights. For example, for player name “zezinho”, there are two players which on research reveals are separate players (from Brazil and Biassau-Guinea). However, they have the exact same height of 175.26. This is highly unlikely. Furthermore, this is not isolated. For example, out of 11060 players, 1954 of them have the exact same height of 182.88. This is highly unnatural. We chose not to use any features from this dataset.

3.2. Data Scraping - Historical ELO ratings

To evaluate the performance of teams objectively, we used the World Football Elo Ratings system, as this provides a better measure of teams’ quality rather than a percentage win rate, which would be higher for dominant teams in smaller leagues than average teams in more competitive leagues.

ELO ratings of football clubs are calculated according to the formula below [4], which is similar to standard ELO systems except that ‘Goal Difference’ is weighted into the calculation, and an additional weight index of the match - highest for intercontinental championships (e.g. Champions League) and lowest for friendly matches - is included into the scoring.

$$R_n = R_o + P$$

where

$$P = KG(W - W_e)$$

Where;

R_n = The new team rating
R_o = The old team rating
K = Weight index regarding the tournament of the match
G = A number from the index of goal differences
W = The result of the match
W_e = The expected result
P = Points Change

We obtained historical ELO rating data by scraping data from http://clubelo.com/, and found the full ELO data (2008/09 - 2015/16), updated after every match, for 269 teams (out of the 299 in the dataset). Thereafter, we reconciled this data with the matches.csv dataset given.

Elo ratings were merged with the match dataset. As before, elo ratings on the particular match day for the team in question was used to avoid biasing the dataset.

3.3. Initial Exploration and Feature Engineering

3.3.1 Feature Engineering

The target for analysis was the result of each match. There are three categories; home team winning, away team winning and draw. These were one hot encoded.

The features used for analysis are a mix of one-hot-encoded categoricals and numeric variables. We will use the team attributes to examine their relationship with the teams winning a tournament. The team attributes for home team (prefix 'h_') and away team ('a_') are used for this study.

Further to this:

- All categorical variables were one hot encoded using their original variable names appended with _Class,
- Numerical variables are scaled to take a range of [0,1] to avoid biasing the models used.
- Null values are dropped, as mentioned in Section 3.1.

Furthermore, we engineer several more features:

1. Elo ratings for home and away teams. (From Section 3.2)
2. Ratios of the numerical variables for home and away using the formula home_attr/away_attr: *['r_buildUpPlaySpeed', 'r_buildUpPlayDribbling', 'r_buildUpPlayPassing', 'r_chanceCreationPassing', 'r_chanceCreationCrossing', 'r_chanceCreationShooting', 'r_defencePressure', 'r_defenceAggression', 'r_defenceTeamWidth', 'r_elo']*
    - This tells us the magnitude of difference.
3. We include the **stage** of the season as well (i.e. Stage 'n' referring to the nth game of the season), as it may provide information on how different team attributes may contribute as a league progresses.

3.3.2 Initial Exploration - determining the predictive accuracy of various algorithms

We first determined the predictive accuracy of various algorithms on the target variable (home team result), to ascertain that the features were predictive.

To do so, we preliminarily tested three algorithms that can be set up without excessive hyperparameter tuning - Naive Bayes, Random Forest (10 estimators) and Logistic Regression (newton). We performed the testing on the cleaned and prepared 'Match' dataset, using a 60-40 train-test split.

The results obtained, shown in Table 1, were sufficient to show that the models were predictive, with Logistic Regression giving the best performance.

Algorithm	**Test Set Accuracy**
Random forest	0.595
Naive Bayes	0.597
Logistic Regression	0.646

Table 1. Test Set Accuracy of algorithms

3.4 Examining various team_attributes in predicting team wins

We will use logistic regression coefficients to determine the contributory effect of each team attribute to the respective teams winning the match.

The coefficient estimates below gives the relationship between the independent variables and the dependent variable. These estimates tell us the amount of increase in the predicted log odds of h)win = 1 that would be predicted by a 1 unit increase in the predictor, holding all other predictors constant.

For the independent variables which are not significant, the coefficients are not significantly different from 0. The P > |Z| column provide the z-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0. [5]

3.4.1 Effect of Team Attributes in Predicting Team Wins - including ELO data

The results are obtained as follows for predicting each result category

Home team win	Away team win	Draw
0.642	0.740	0.748

Table 2. Accuracy of result prediction using logistic regression

Let us assume that p values < 0.05 are significant for the logistic regression coefficients. The following findings are then identified:

- Home team win: ratio of defenceAggression, the elo ratings, stage of match, away and home team defence aggression.
- away team win: elo ratings, ratio of buildUpPlayDribbling and home team's buildUpPlayDribbling attribute.
- draw: elo ratings, stage and ratio of defence pressure.

Elo ratings contributed a significant degree of predictive value, with comparatively large coefficient magnitudes that goes in both ways. For example, for home team win, the coefficient for home team elo is 6.34 while away team elo is -6.00; in comparison, another significant attribute r_defenseAggression was -1.97.

3.4.2 Effect of Team Attributes in Predicting Team Wins - without ELO data

Given the outsized impact of elo ratings in the earlier analysis, we also studied the effects of team attributes without elo ratings. The following attributes have been identified as significant.

- home win: r_chanceCreationPassing, a_chanceCreationShooting, h_defenceAggression, h_buildUpPlayDribbling
- away win: r_chanceCreationCrossing, r_chanceCreationPassing, r_buildUpPlayDribbling, a_defencePressure, a_chanceCreationCrossing, a_chanceCreationPassing, a_buildUpPlayDribbling, a_buildUpPlaySpeed,h_defencePressure, h_chanceCreationCrossing, h_chanceCreationPassing, h_buildUpPlayDribbling
- draw: ratio of defense pressure, stage

There are several common factors such as r_chanceCreationPassing, h_defencePressure and h_buildUpPlayDribbling. Furthermore, their coefficients are of different directions (i.e. positive for hom win, negative for away etc.). This indicates that these are factors that are strongly predictive of match outcomes. From these results we can conclude that effective teams are able to create opportunities for scoring by passing while simultaneously building pressure on opposing teams.

3.4.3 Attributes that make a good team

Another perspective that we can take to examine the effects of team attributes on match outcomes is to consider the question of what differentiates a top team from a bottom team by number of matches won. We will use the top and bottom 20% to obtain sufficient data points for this statistical analysis. To identify factors, T-test is again used with the null hypothesis that the top and bottom teams do not differ significantly in mean for each attribute.

The following attributes have been identified as the differing factor between the top and bottom teams: `buildUpPlayPassing, chanceCreationCrossing, chanceCreationShooting, defencePressure, buildUpPlayDribblingClass, buildUpPlayPassingClass, buildUpPlayPositioningClass,chanceCreationPassingClass, chanceCreationShootingClass, chanceCreationPositioningClass and defencePressureClass.`

We can also visualize the results to illustrate the differences, shown in Fig. 2 below.

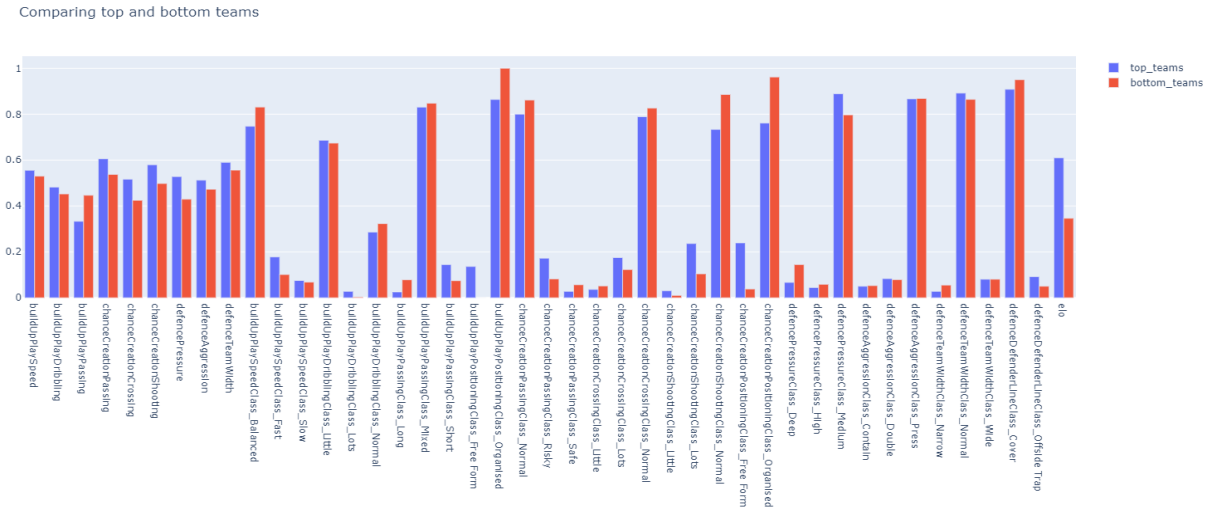

Figure 2: Comparison of Top and Bottom teams

In general, top teams are better at creating opportunities for themselves, putting pressure on the opposing defense, and pass the ball around more often during build up play. Furthermore, top teams are more aggressive as can be seen from their play style classification.

3.5 Internal Factors affecting match performance

3.5.1 Association Rule Analysis

In the previous section, logistic regression was adopted to examine the predictability of team attributes toward match results. To take a step further, we tried to adopt association rule learning in mining key factors (both team attributes and player's attributes) associated with match results.

In association rule learning, three figures would be used to illustrate the association between players' performance and match result, namely Support, Confident and Lift.

$$Rule: X \Rightarrow Y \longrightarrow \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\ Confidence = \dfrac{frq(X,Y)}{frq(X)} \\ Lift = \dfrac{Support}{Supp(X) \times Supp(Y)} \end{cases}$$

3.5.3 Contribution factor analysis of player's performance toward match result

Concerning players performance, data from player_attributes.csv could be used as an index representing isolated performance of players.

To perform association rule analysis of players' performance, combined data, containing categorical data and numerical data, must be transformed into categorical data, followed by one-hot encoding.

3.5.3.1 Data Transformation of Players Attributes

Holistic Performance of players within the same will be represented as a team rating, by taking the mean, maximum or mode of the subset. Most of the numerical ratings (except ratings of goalkeeper) of players will be averaged as a team average. For rating of goalkeepers in a team, since it is assumed that the player with highest rating will be the goalkeeper, the highest rating of goalkeeping will be taken as the team goalkeeping rating. For categorical data, such as *attacking_work_rate* and *defensive_work_rate*, mode of the team will be taken to describe the situation in a team.

Table 3 describes the feature engineering method of attributes.

Type of feature engineering	**Attributes**	
Mean	● overall_rating ● potential ● crossing ● finishing ● heading_accuracy ● short_passing ● volleys ● dribbling ● curve ● free_kick_accuracy ● long_passing ● ball_control ● acceleration ● sprint_speed ● agility	● reactions ● balance ● shot_power ● jumping ● stamina ● strength ● long_shots ● aggression ● interceptions ● positioning ● vision ● penalties ● marking ● standing_tackle ● sliding_tackle
Maximum	● gk_diving ● gk_handling ● gk_kicking	● gk_positioning ● gk_reflexes
Mode	● attacking_work_rate	● defensive_work_rate

Table 3. Type of feature engineering performed on attributes

After creating a holistic rating of players, it is important to turn numerical data into categorical data. To align with other categorical data which separated players' performance into "Low", "Medium" and "High", same classes will be applied in the classification.

Class	**Standard**
Low	<25th percentile of that attribute
Medium	>= 25th and <=75th percentile of that attribute

High	>75th percentile of that attribute

Table 4. Classification of players' ratings

The categorical data will then be transformed by one-hot encoding to binary data

3.5.3.2 Apriori algorithm

Apriori algorithm is one way to generate the support of each association. While setting the support threshold as 0.1, larger numbers of subsets would be created before scanning the qualified subset. By adopting the breadth-first traversal algorithm, problems encountered with large dataset are that the efficiency is low and memory requirement is high. We then turned in using **FP growth algorithm** which was more efficient.

3.5.3.3 Result of Association Rule

Results are sorted according to confidence, and the top 5 association rules are listed below:

Antecedents	Consequents	Support	Confidence	Lift
frozenset({'away_defensive_work_rate_<medium>', 'away_overall_rating_<medium>', 'away_attacking_work_rate_<medium>', 'away_ball_control_<medium>'})	frozenset({'Home_Result_<Win>'})	0.2011	0.4981	1.0857
frozenset({'away_overall_rating_<medium>', 'away_attacking_work_rate_<medium>', 'away_ball_control_<medium>'})	frozenset({'Home_Result_<Win>'})	0.2048	0.4981	1.0856
frozenset({'away_short_passing_<medium>', 'away_attacking_work_rate_<medium>', 'away_ball_control_<medium>'})	frozenset({'Home_Result_<Win>'})	0.2005	0.4974	1.0841
frozenset({'away_overall_rating_<medium>', 'away_short_passing_<medium>', 'away_attacking_work_rate_<medium>'})	frozenset({'Home_Result_<Win>'})	0.2007	0.4969	1.0830
frozenset({'away_overall_rating_<medium>', 'away_attacking_work_rate_<medium>', 'away_potential_<medium>'})	frozenset({'Home_Result_<Win>'})	0.2087	0.4968	1.0829

Table 5. Result of association rule learning on players' attributes toward Home team winning

3.5.3.4 Interpretation of Results of association rule

1. Antecedents listed are all related to class "medium"

This might be a result from the standard adopted for classifying players' rating is problematic, whose rating fall within 25th and 75th percentile is classified as "medium". Possible improvement might be classifying ratings in more classes.

2. Confidence" of association rule are generally not satisfying

It could be seen that all association rules result in a "confidence" lower 0.5, showing that these association rules might not be good rules to rely on.

3.5.4 Contribution factor analysis of team performance toward match result

We analysed how the team attributes effect's their probability of winning a match using association rule mining.

3.5.4.1 Data Transformation of Team Attributes

The team attribute data consists of several categorical as well as integer data. Since, we require categorical data for association rule mining we transformed the integral data as following

Class	Standard
Low	**0 to 25**
Medium	**25 to 50**
High	**50 to 75**
Very High	**75 to 100**

Table 6. Classification of teams' ratings

3.5.4.2 Result of Association Rule Mining

We used visNetwork, an R package, to plot an interactive association rule mining graph, which could be accessed via https://raghav0307.github.io/armines2.html, shown in Figure 3 below.

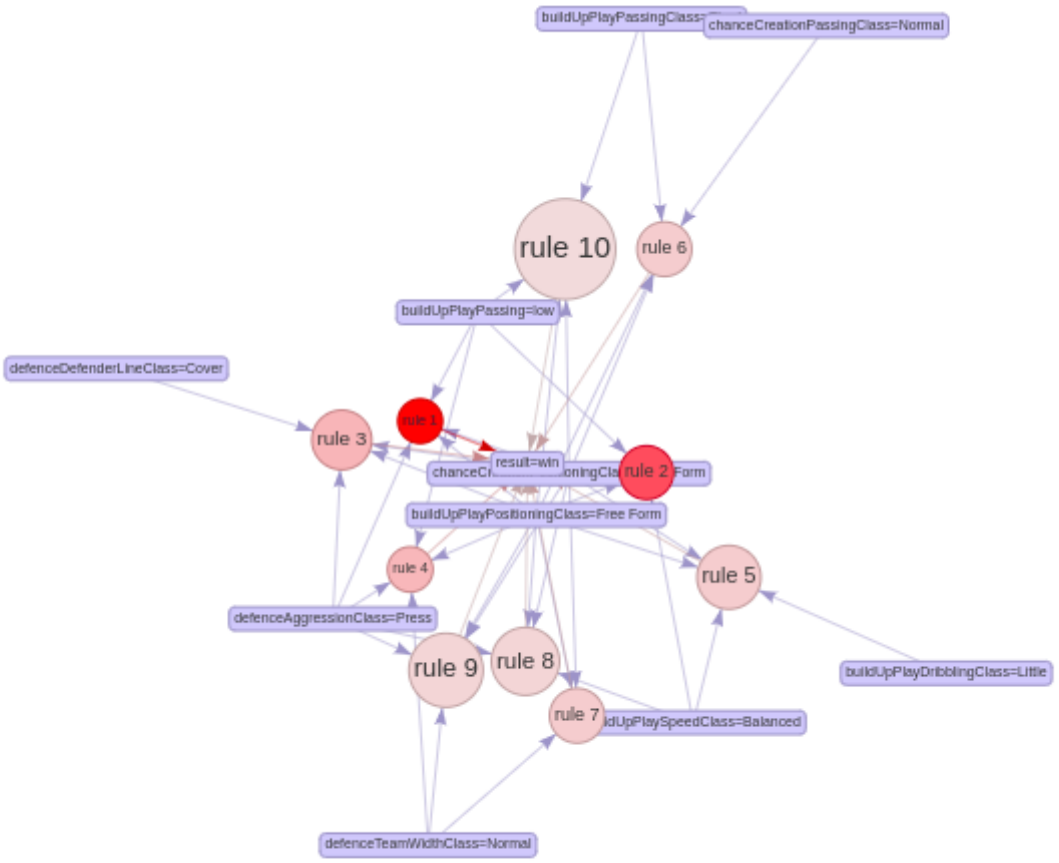

Figure 3: Association Rule Graph of Attributes influencing match performance

The top 10 association rule mining rules, from the graph above, sorted on the basis of confidence are as follows in Table7:

Antecedents	Consequents	Support	Confidence	Lift
{buildUpPlayPassing=low, buildUpPlayPositioningClass=Free Form, chanceCreationPositioningClass=Free Form, defenceAggressionClass=Press}	{result=win}	0.0104	0.623	1.67
{buildUpPlaySpeedClass=Balanced, buildUpPlayPassing=low, buildUpPlayPositioningClass=Free Form, chanceCreationPositioningClass=Free Form}	{result=win}	0.0119	0.612	1.64
{buildUpPlayPositioningClass = Free Form, chanceCreationPositioningClass = Free Form, defenceAggressionClass = Press, defenceDefenderLineClass=Cover}	{result=win}	0.0129	0.596	1.59
{buildUpPlayPassing = low, buildUpPlayPositioningClass = Free Form, defenceAggressionClass=Press, defenceTeamWidthClass=Normal}	{result=win}	0.0105	0.595	1.59
{buildUpPlaySpeedClass=Balanced, buildUpPlayDribblingClass=Little, buildUpPlayPositioningClass=Free Form, chanceCreationPositioningClass=Free Form}	{result=win}	0.0136	0.59	1.58
{buildUpPlayPassingClass=Short, buildUpPlayPositioningClass=Free Form, chanceCreationPassingClass=Normal, chanceCreationPositioningClass=Free Form}	{result=win}	0.012	0.59	1.58
{buildUpPlaySpeedClass=Balanced, buildUpPlayPositioningClass=Free Form, chanceCreationPositioningClass = Free Form, defenceTeamWidthClass=Normal}	{result=win}	0.0121	0.59	1.58
{buildUpPlaySpeedClass=Balanced, buildUpPlayPositioningClass=Free Form, chanceCreationPositioningClass = Free Form, defenceAggressionClass=Press}	{result=win}	0.0142	0.588	1.57
{buildUpPlayPositioningClass=Free Form, chanceCreationPositioningClass = Free Form, defenceAggressionClass = Press, defenceTeamWidthClass=Normal}	{result=win}	0.0153	0.587	1.57
{buildUpPlayPassing=low, buildUpPlayPassingClass=Short, buildUpPlayPositioningClass=Free Form, chanceCreationPositioningClass=Free Form}	{result=win}	0.0169	0.586	1.57

Table 7. Result of association rule learning on team attributes toward team winning

3.5.5 Further Investigation

It is found that using *team_attributes.csv* generates stronger association rules compared to using player_attributes.csv. At the first glance, it seems that player attributes might be strongly related to team attributes, leading to a question of why team attributes generate stronger association rules then player attributes. We then turn to investigate what causes the difference of result.

3.5.6 Correlation matrix

Correlation between team attributes and player attributes are investigated. The Pearson method is adopted to find the linear correlation between numerical attributes of two datasets. It is found that the correlation between team attributes and player attributes range from 0.25335 to -0.22629, which shows a weak linear correlation between team attributes. It provides the legitimacy of team attributes that generate better association rules, since two datasets are nearly independent.

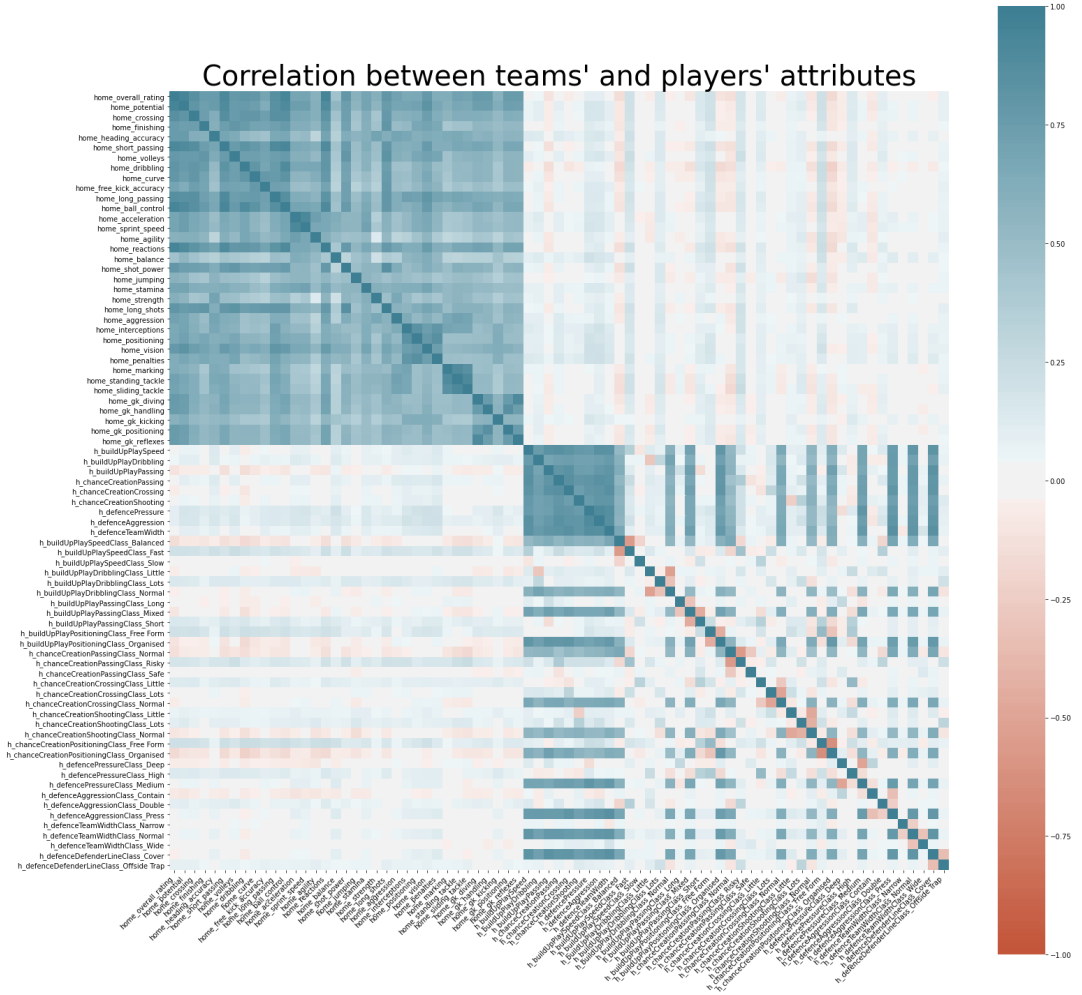

Figure 4: Correlation between teams' and players' attributes

On top of it, it is important to point out that team attributes provide categorical data which describe the playing strategy of a team (e.g buildUpPlayPositioningClass). Since such data do not exist in player attributes, it illustrates the importance of taking playing strategy into consideration while forming an optimal football team.

3.6 External Factors affecting match performance

Several additional factors were identified as having an impact on predicting match outcomes during our analysis. To complete our analysis, we will study them here.

3.6.1 Effect of stage of season on match result

One thing that stood out from the previous analysis was the strong effect of stage of the season on predicting draws. We shall examine the percentage of drawn matches with respect to stage.

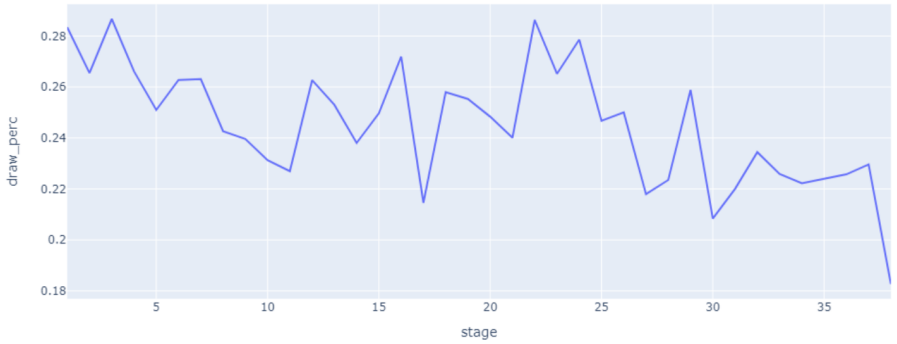

Figure 5: Graph of draw percentage against stage of season

As can be seen, there is a downward trend in the percentage of matches that end in draws as stages progress.

An intuitive explanation of this is that teams could be more risk averse at the start of the season, but start to play more aggressively towards the end, as they require wins (3 points) to either qualify for international competitions, or avoid relegation, which inadvertently results in fewer draws.

3.6.2 Effect of home-team advantage on match result

By plotting the percentages of various match outcomes together, one thing that stands out is the vast difference in percentage of home team wins as compared to the other outcomes, showing a strong **home team advantage**.

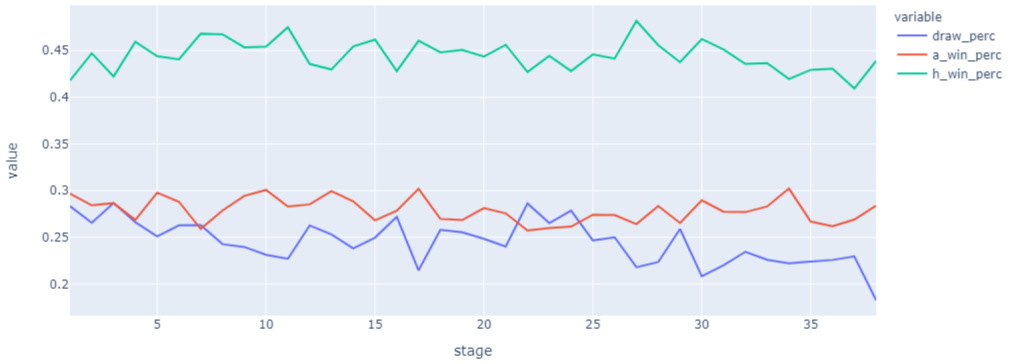

Figure 6: Graphs of Home Win (green), Draw (blue) and Home Loss (orange) against stage of season

The percentage of home team wins range between 40-50%, which is significantly higher than the percentage of away team wins and draws. However, to verify that such an observation is statistically significant, we can use the t-test.

3.6.3 Validation of home-team advantage using T-test

To verify our visual observation of a home team advantage, we can use the T-statistic test.We run a two-sided test for the null hypothesis that 2 **related** percentages, (home vs away) have identical average values which should be expected if there is no home advantage.

Tests are run on for percentages across seasons as well as across the various stages.

The p-values are 1.79e-07 and 1.08e-34 respectively, showing that there is a very low probability that the mean of home team winning and away team winning are identical, rejecting the null hypothesis. This is indeed a home ground advantage present in European soccer.

4. Conclusions and Further Study

In this study, we have investigated the internal (team composition and style) and external (stage, home-team advantage) factors affecting the match performance of football teams in club football.

It was found that the team had a greater contribution to match outcomes than the players, and teams that are more aggressive and better at creating chances were rewarded by having better probability of match outcomes. Besides the individual factors, we also identified external factors as contributing to match outcome probabilities, such as the home ground advantage and game progression during a season.

Further study could be conducted to evaluate in more granular detail the precise contribution of each team attribute, as well as a comparison of different playing styles, and ultimately allow teams to devise appropriate transfer strategies within a given budget to bring in appropriate managers, training staff, and players, to realise such a team.

References

[1] Lange, D. (2020, November 26). European football market size 2006-2019. Retrieved March 24, 2021, from https://www.statista.com/statistics/261223/european-soccer-market-total-revenue/

[2] Statistically ranking the biggest football leagues in the world today. (2020, May 18). Retrieved March 28, 2021, from https://www.theexeterdaily.co.uk/news/sport/statistically-ranking-biggest-football-leagues-world-today

[3] Smigel, L. (2019, October 09). Look-Ahead bias: What it is & how to avoid. Retrieved March 28, 2021, from https://analyzingalpha.com/look-ahead-bias

[4] World football elo ratings. (2021, March 27). Retrieved March 28, 2021, from https://en.wikipedia.org/wiki/World_Football_Elo_Ratings

[5] Home. (n.d.). Retrieved March 28, 2021, from https://stats.idre.ucla.edu/stata/output/logistic-regression-analysis/