Data Open Competition East Coast 2021

Team Number 13

Non-Technical Executive Summary

Our Mission: Our research focused on football match prediction, betting, and power features for determining the win/lose of a match. The main objective of this research is to find the match win/lose/draw prediction that can help people "Predict With a Reason" and not just betting with bias(overrate our favorite team) all the time.

Question Solved: "Can we predict the football match win/draw/lose for the home/away team?" As stated above, this behavior bias will tend to overestimate the team they are betting on. This research presented the machine learning model prediction and compared it with the naive forecast. As a result, the best features that can be used to determine win/draw/lose were found. As more technical perspective, we will predict whether the home team will win/draw/lose comparing to away team.

Methodology and Summary of Research:

- Objective Defined: Win/Draw/Lose for Home/Away team
- Wrangling: cleaning and joining
- Data Exploratory Analysis: found the variables that have a significant impact in soccer match
- Feature Engineering: created a list of features used in this research, which consist of approximately 800 features for the football match prediction.
- Modeling: found the best machine learning model, best features, and reported the prediction result
- Investment Strategy: leveraged the prediction to perform a betting strategy
- Summary

The Main Result: The machine learning model presented improvement in accuracy for prediction with the naive Model. The Model achieves around 51% accuracy for predicting a game's outcome (win/draw/lose), which is much higher than a random picking(33.3%). It turns out that most of the vital features are features such as player win rate across the team. Few of the player statistics, such as ball control, potential score, and player synergies with the team, are also great features for determining the result. Finally, one can leverage the prediction model into making a profit from betting. A detailed investment strategy is also posted at the end of this report.

Technical Exposition

1. Data Wrangling

Even Though data-sets provided in this research are Pre-cleaned, there is still a need to clean some data. We described some of the essential cleaning procedures below.

1.1 Basic Cleaning

Match_id cleaning: The number of distinct matches is 25,979. However, some of the teams contain players less than seven people. To be in-lined with the international standard, we deleted those matches, resulting in approximately 25,221 games.

NaN imputation: The imputation of NaN was different across the features. we cannot simply apply mean / zero imputation for every feature. For instance, the player "win-rate", if the new player playing with zero matches played and we impute them with zero, it is equivalent to assume that they have zero win-rate, which is a very pessimistic assumption. The Mean imputation is appropriate for this case. However, for the "Match-Played," if we imputed a new player match-played with mean, it is equivalent to assume that the has average industry experience, which is too optimistic in the business sense. Therefore, zero imputation is better. We should decide which of these features will be imputation based on real-life logic.

Categorical Feature Cleaning: Some of the team and player attributes had unrecognized string names, for instance, "attacking_work_rate" in player_attributes.CSV contains values such as None, high, le, low, medium, norm, stoc, y. While most observations (97.4% of total rows) are high, low, and medium, the rest label had minority observations. Therefore, we create four features for the player attacking work rate 1 high, 2 medium, 3 low, 4 others to give us more clean insights and fewer dimensions from noisy data. This logic is also applied to every categorical column in the team and player attributes.

1.2 Wrangling Methodology

We have a total of 7 files. Our objective is to predict "whether the home team will win or not," therefore the Y variable should come from "match.csv". All data we used for modeling should fit the structure of "match.csv." The "country.csv" and "league.csv" are joined "match.csv" by country_id and league_id. To join "player.csv", "player_attributes.csv", "team.csv", and "team_attributes.csv", different joining logic are applied.

Join with ids: We spread the player_id from each position in the columns (position 1 home - position 11 away) into a single row, mark them as "player_id," and label them as either "Home or Away." One example of transformation was showing below: Table 1 is converted to Table 2.

Table 1: Example of match_csv

match id	home player 1	home player 2	home player 3	away player1	away player 2	away player 3	home team id	away team id
1	a	b	c	d	e	f	z1	x2

Table 2: Example of transformed match_csv

match id	player id	team id	home/away	position
1	a	z1	home	1
1	b	z1	home	2
1	c	z1	home	3
1	d	x2	away	1
1	e	x2	away	2
1	f	x2	away	3

Table 2 is used again to join the "player.csv" , "player_attribute" , "team_csv", and "team_attribute" by their player_id or team_id.

Join with date: Joining with ids is not enough; the player stats and team stats are constantly changing. If we ignored it, this would create a vast duplicated result. It is necessary to join the updated date (from player and team stats) with the particular match date. The method for joining the table is to use the latest updated day for either player or team stats and join them before the specific match date starts. After data cleaning and manipulation, the final table is called "match_master.csv" and was used for the feature creation, which provides the convenience for feature creation with only one master table.

The table joining techniques used in this case was only feasible for small data-sets, and it is not optimal for the firm with large data-sets.

2. Exploratory Analysis

In this section, we will go through some of the fundamental and essential exploratory analyses. This section will investigate the vital variables and the reason/assumption behind these variables before feature engineering.

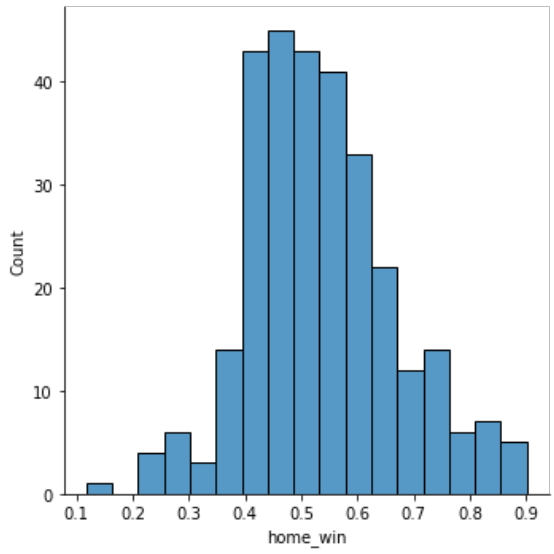
2.1 Win-Rate

Since the win-rate is a commonly used indicator to predict a team's future performance, a team's win rates and an individual player are vital to investigate. We construct three labels from the outcome of a game.

- If the Home team goal earned > Away team goal earned: label as 1 or Home Win
- If the Home team goal earned = Away team goal earned: label as 0.5 or Home Draw
- If the Home team goal earned < Away team goal earned: we label as 0 or Home Lose

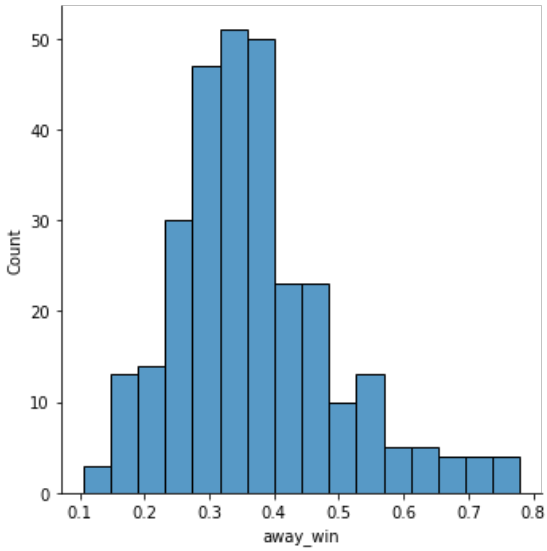
The distribution of our objective variable as followed:

Figure 1: The distribution of Home wins



As shown in Figure 1 and Figure 2, home-win occurs more often than Away-win. It is better to use Home-win as our objective variable (Y), because Home-Win has more observation clustered around the mean (0.5) than the Away-Win. The idea of

Figure 2: The distribution of Away wins



Home-Win is much easier to communicate with non-technical people. To predict Home-Win, we need to construct the features set to distinguish the features between the Home and Away teams. For instance, we could convert the feature "Team_score" into "Team_score_Home" for the Home team and "Team_score_Away" for the away team so that the machine learning model can use these feature to separate the class.

2.2 Player Attributes

There are 33 player attributes provided in the original data-sets. We construct a simple analysis of player attributes' correlation, distribution, and played times.

According to the correlation matrix for all the player attributes (Figure 3), it is clear to divide attributes into several categories. The matrix indicates multiple essential trends for this analysis. For example, all goalkeepers' attributes are highly correlated to each other. Thus we aggregated them into one single feature. Abilities that are not labeled as goalkeeper have almost no correlation with a goalkeeper's performance. Since the goalkeeper is a soccer game position, this finding hypothesized that all positions should have their major skill sets and created influential features based on that information. Besides, many skills, such as marking, standing_tackle, and sliding_tackle, are also highly correlated. This trend implies that we can categorize at-

tributes into different skill sets, which eliminating redundant features with similar statistical meaning.

According to the distribution of player attributes (Figure 4), it further confirmed the finding with the correlation matrix of player attributes. Only a few people have a high score in goalkeeping, which implies the goalkeepers. Most of the distribution graphs demonstrate a standard distribution curve. However, some of the charts have multiple peaks, especially for goalkeeper's skills, tackle, and marking. Since the trend found in goalkeepers' explained that the peak is caused by the difference in position (position goalkeeper and others), this finding might apply to other attributes. Each peak should represent a type of position, and positions should have different required skills.

Furthermore, the number of games each player has played also plays a vital role in this analysis, as shown by Figure 5. Since the player has changed all the time and the distribution of matches played indicates that most players only played a few games, the analysis should be based on each game, not individual players.

2.3 Team Attributes

There are 21 team attributes provided in the original data-sets. This section reported exploratory analysis of team attributes as following.

The team attributes' correlation matrix (Figure 6) indicates that attributes from each major category (buildup play, chance creation, and defense) are only correlated with features within the same categories. Therefore, we categorize the features within the same categories into groups, especially for the defense class.

3. Features Crteation

In this research, We have created approximately 800 features in total. The data cleaning process manipulated all the features in a structure that fits the "match.csv" file, where each row has a unique match_id. The feature contain four major categories: the player's statistics, player's synergies with teammates and opponent, team's statistics, the Elo system. The feature always include both sides: the home team data and the away team data. Since the objective is to predict whether the "Home team will [win/draw/lose]," we need to separate the feature based on home/away. Therefore, every feature has attributes corresponding to their Home/Away team.

3.1 Player's statistics

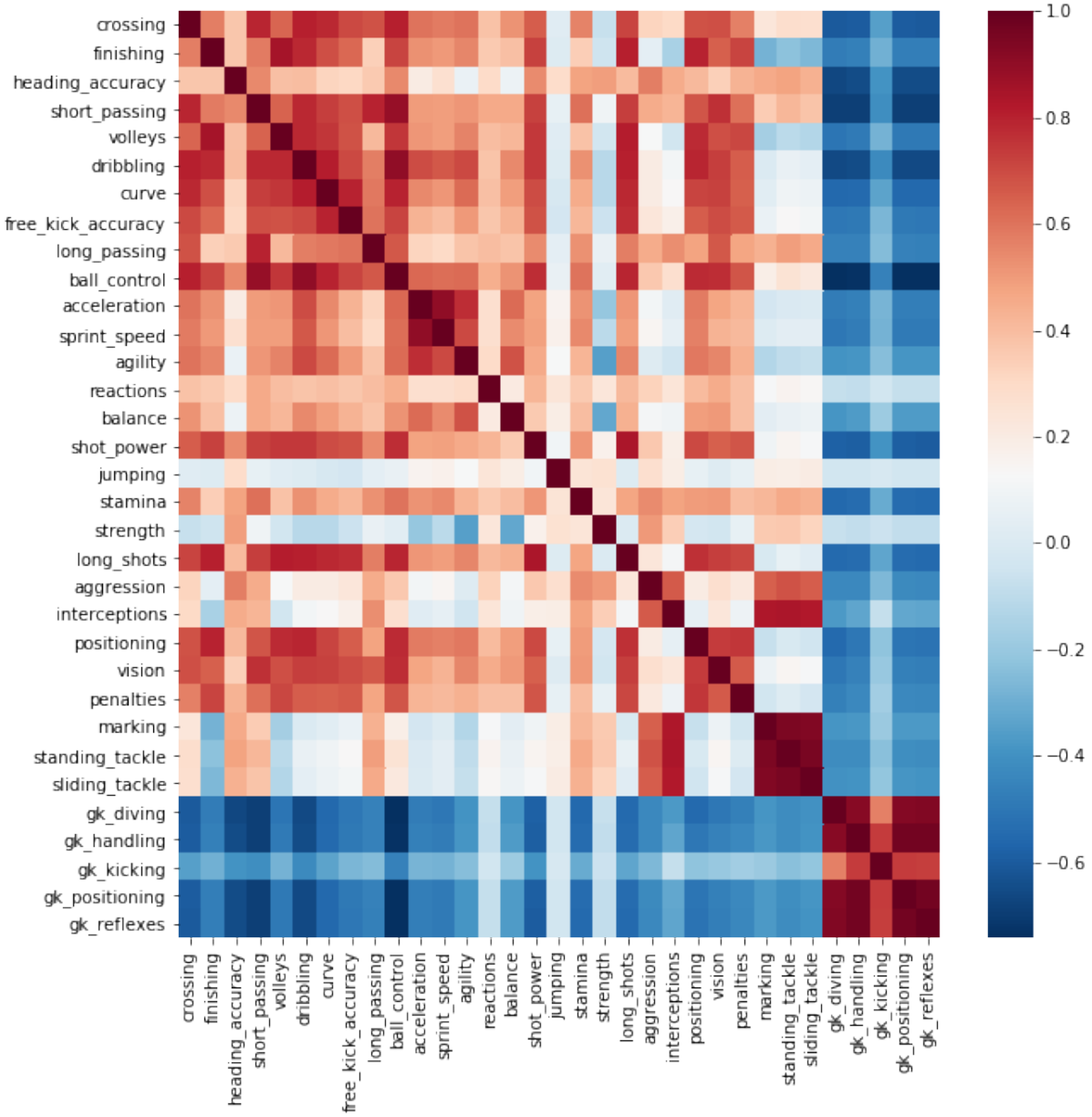
The player's statistics is collected through 3 approaches. First, a score to measure how talented a player is at a specific position is created. Second, the win-rate, ranging from past 1 to 3 years, of a player's both home and away games was calculated. Third, player attribute aggregation of the original data was calculated.

To calculate the talent score for each position, players' skills are classified into six main skill types (from the "player_attributes.csv" file): the goal keeper's skills, the shooting ability, the ball control ability, the ball passing ability, the physical strength, and agility in-game. A sample analysis of the skill sets was conducted based on each position's primary skills (ranging from positions 1 and 11). For example, the most critical skill for position 1 is the goalkeeper's skill, and the shooting ability is not essential for position 1. This paper's categorizing is conducted based on the most common and flexible soccer formation, 4-3-3, where positions 2,3,4, and 5 are the back positions, 6, 7, 8 are the midfield positions, and 9, 10, 11 are the front positions.

After removing goalkeepers' row and goalkeepers' ability columns from the data-frame, all other skills have a normal distribution. Therefore, the mean and standard deviation of each skillset are analyzed for each position. As a result, all positions have been assigned with their most important skills. For example, in Figure 7 means the shooting ability for front positions is much higher than in any other position. Similarly, according to Figure 8, control is most important for the back position. However, midfield positions tend to be skilled at all skills. For each game, every player's talent score is calculated based on their positions and required skills. The higher the score, the better the player can play his position in a specific game. Therefore, this approach hypothesizes that a team with more talented players at the position they played should win the game.

For the win-rate, each player's past performance was aggregated into two labels, home and away (from the "match.csv" file). Instead of looking at

Figure 3: Correlation of Player Attributes



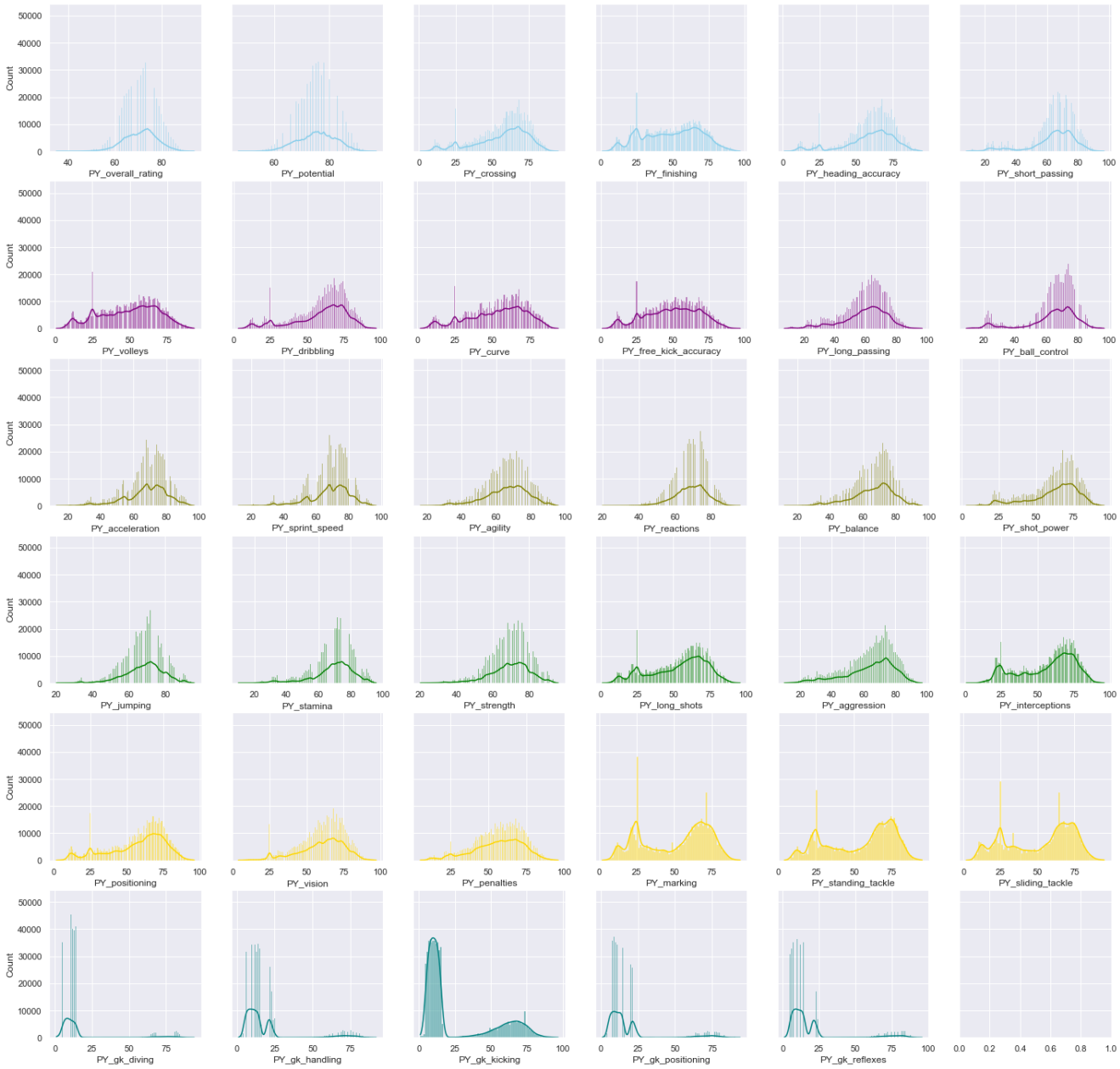
the team's win-rate, the research now looked at each player's win-rate across the team and aggregated them within a specified timeline (win-rate within 1 year, 2 year, 3year) and then aggregated across the player into the team. Since the team dynamic might change the player, this approach hypothesizes that a player's historical win-rate can help determine the win/lose of the team.

The original data contains around 33 player statistics (from the "player_attributes.csv"). However, each match has many players in one team. Then at each particular match, the research aggregated the latest player statistic across the team that presented before matched time. This approach hypothesized that some player attributes might influence the win/lose of a team.

3.2 Player's synergy

The player's synergies section contains two types of features: player synergies with their team and their opponent team. Many times, an individual player's performance might not be the only team performance indicator. The methodology applies to both the affiliated team and the opponent team.

Figure 4: Distribution of Player Attributes



For the player's synergy with the affiliate team, the synergy is defined as how comfortable the player is with other players in that particular team. The comfortability contains the historical win rate (within the last one year) and the number of matches these players played together. Then, we aggregated the comfortability across the team. For the player's synergy with the opponent team, the logic is the same but with players from the opponent team. These features then raised another hypothesis: does player' synergies in the affiliate/opponent team provide additional prediction power to the win/lose of the affiliate team.

3.3 Team's statistics

All the features about the team's statistics were collected from the team_attributes.csv, which contains 21 team statistics there. There are three types of features under this category, latest data before the match, players-team synergies, and play-styles.

First, we leveraged the information by selecting the latest team statistics before match time and presenting them as the predictor. Secondly, we believe

Figure 5: Number of Games Each Player Played

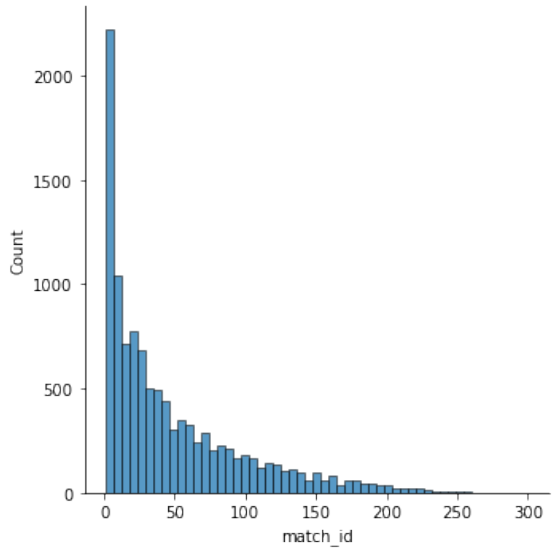
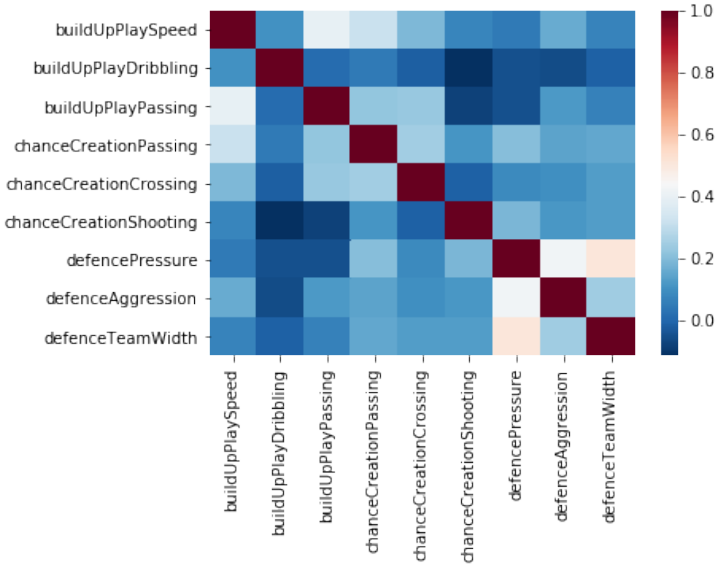


Figure 6: Correlation of Team Attributes



the interaction between the player and the current team might have historical synergies together. The win-rate represents the synergies between player and team for this player in this particular team. And finally, features with both the affiliate team interaction and opponent team interaction were created.

Thirdly, a team's play-style can be described into three categories: conservative, abnormal, and aggressive. If a team played aggressively, they would apply a lot of defensive pressure on opponents and create many chances for shooting. If a team played conservatively, they focused more on defending rather

Figure 7: Shooting Ability by Position

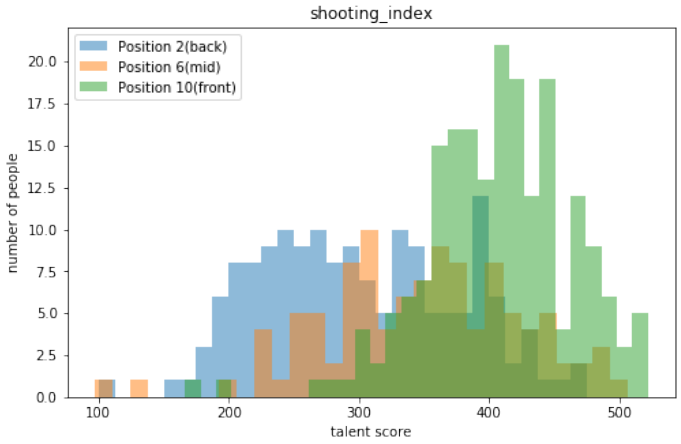


Figure 8: Ball-Control Ability by Position

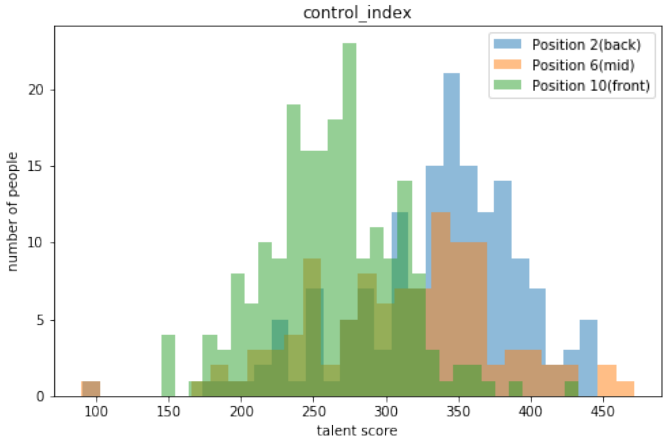
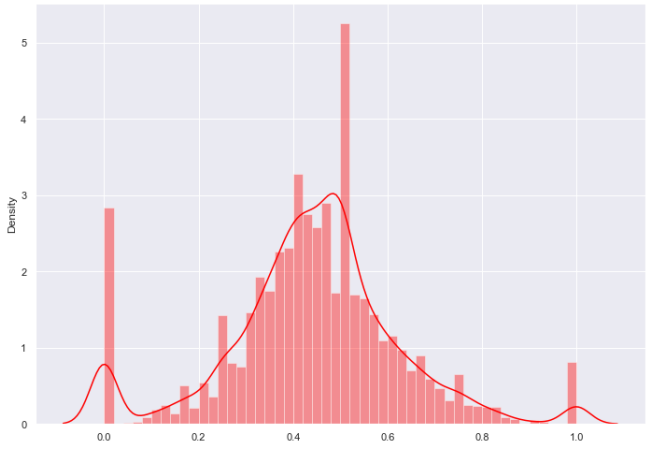


Figure 9: Distribution of Player Win-rate within 1 year



than attacking. Any distinctive play-style would be considered abnormal (risky, free form, etc.). This research calculated a score for each of the categories as

a feature. The higher the score, the more likely the team applied that play-style.

This section then raised another hypothesis: does the team attribute provide additional prediction power to the win/lose of the affiliate team.

3.4 Team Win-rate Analysis & ELO System Construction

Some more specific and more involved features we constructed involved observing that not all wins and losses were to mean the same. In football leagues, teams are ranked based on points (denoted *pt*), determined by +3 from a win, +1 from a draw, and +0 from a loss. Another statistic used during tiebreakers is the GD (Goal Difference) statistic during a season, the difference between the Goals For (goals scored) and the Goals Against (goals lost) accrued over matches in a season. These statistics, along with the win and loss streaks (normally over seven games), provide insight into how teams win within a season and how they stack up to one another. These metrics reset at the end of each season.

The motivation to keep track of these metrics is to note how a team's standings and to win (or loss) streak can affect their play-style and performance. Moreover, these metrics were also considered in measuring a team's win with the context of their opponent's strength: compete with lower-tier teams, very volatile team, or beat any other team?

With questions like these in mind, we decided to construct a metric, beyond the players' ratings which support the team's raw potential, which gauged a team's historical strength and hopefully could translate into a metric that could quantify the strength of each league without much direct interaction. Here, we implemented the ELO system rating borrowed from the FIDE World Chess Federation. Starting from the beginning of our observation period, we set all teams to have an ELO of 1500. The ELO of two teams is constructed because the spread of the ELOs provides a tentative expectation of how likely a team will win. If team A has ELO x and team B has ELO y, then team A has an expected win rate of:

$$E_x = \frac{1}{1+10^{((x-y)/400}}$$

The team's adjusted ELO due to the result of the match is dependent on E by the following equation:

$$x_{new} = x + (R - E_x) * k$$

Where k is an adjustment factor that we set to be at 40 to increase convergence of our model and we define R by:

$$R = \begin{cases} \text{Goals Scored - Goals Lost}, & \text{if win} \\ 0.5, & \text{if draw} \\ 0, & \text{if loss} \end{cases}$$

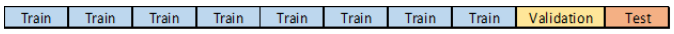
4. Models & Prediction

4.1 Train Test Split

Once the feature master table (includes all features) has been created, before putting in machine learning models, we split the train validation and test set to generalize the model. The split percentage is approximately (90% Training_preset 10% of Test set) and then split further from 90% training_preset into (90% Training set 10% validation set). Figure 10 demonstrates how the data has been split. The total observation of each set was approximate:

- Train 20,428 obs (used to develop ML Models)
- Validation 2,270 obs (used to evaluation and Tuning for ML Models)
- Test 2,523 obs (used to final evaluation and reporting of ML Models)

Figure 10: Test Training Validation Set



4.2 Model Selection

Due to the home win prediction data-set structure, the models with the cutoff features (tree-based model) are the best fit model in this scenario because of 2 reasons: cutoff-based feature and the hierarchical structure of feature. The nature of feature are suitable for a cut-off-based model, i.e., the number of matches-played should have several cutoffs matched-plays that define the player's experience and ability to win for the team. Besides, some of the cutoff criteria came after another criterion. For instance, if each

league's player score might be different, this presents the hierarchical structure that best suits the tree base model.

Therefore, we introduced 3 Tree models: A decision tree for the feature extraction, Random-forest for the feature explanation and accuracy improvement, and an Extreme Gradient Boosting tree for the optimal prediction accuracy. We also tuned each of the model parameters to get the best accuracy-precision-recall metric.

4.3 Prediction Result

After feature creation, we have approximately 800 features. Using Random Forests importance, we identified the significance of these features. We then ran the Random Forest algorithm iteratively by removing the redundant features. We determined that the prediction performance is independent of categorical variables like country_name, league_name, or season in our iterative process. Aggregation statistics (min, max, std of some features) created for the players and teams also did not contribute to better performance. The feature size is then reduced to 197 from approximately 800, which facilitated the convergence.

Draw predicting was consistently tricky for all models to predict. XGboost was the best classifier for this multi-class classification. Aiming to improve individual class prediction, we implement the One vs. Rest classifier on the XGboost model. However, this doesn't have a significant effect.

To achieve better results on our investment strategy, the model needs to maximize the precision scores at the expense of recall. As a result, Decision Trees performed better than the XGboost algorithm in precision for the winning classes. Results for all three models are reported in Table 3. The tree structure of the decision tree model for football match prediction is showed in Figure 11. We created Highest Odd Prediction by the prediction from the highest odd from betting website 365. The result from our model has a slight improvement from the betting website 365. Even though the accuracy improvement is not high, the monetary value from betting out of these website and our strategy are compared later(See section 5 for Investment strategy).

Table 3: Result from Modeling

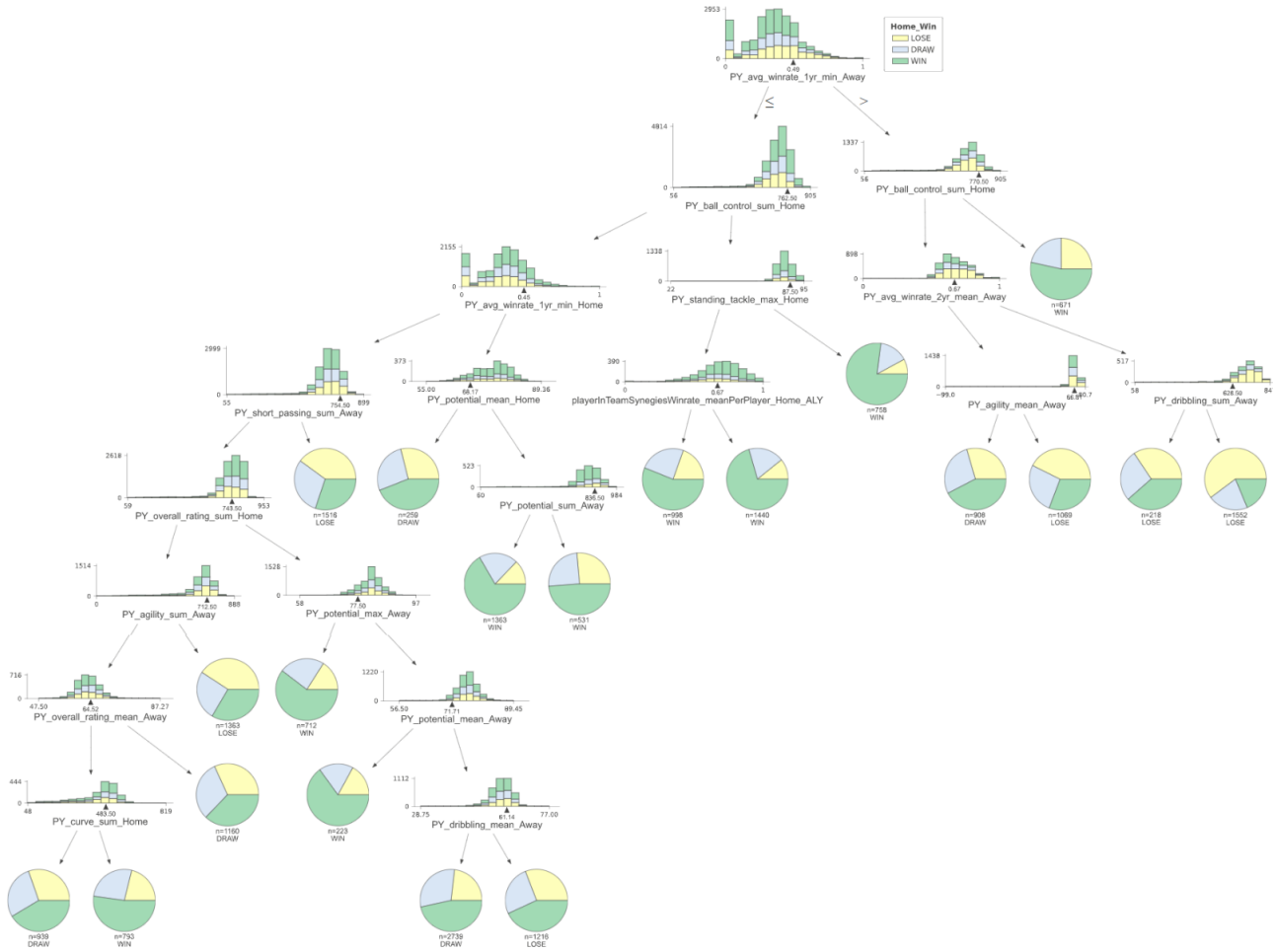
Model Name	Class Label	Precision	Recall	F1_Score	Accuracy
Highest-Odd-Prediction	Loss	0.52	0.41	0.46	0.5
	Draw	0.27	0.15	0.19	
	Win	0.54	0.76	0.63	
Decision Trees	Loss	0.39	0.52	0.45	0.45
	Draw	0.28	0.28	0.28	
	Win	0.63	0.49	0.56	
Random Forest	Loss	0.49	0.45	0.47	0.51
	Draw	0.36	0.07	0.12	
	Win	0.53	0.81	0.64	
XGBoost	Loss	0.47	0.39	0.43	0.52
	Draw	0.37	0.02	0.03	
	Win	0.53	0.88	0.66	
XGBoost-OneVsRest	Loss	0.48	0.42	0.45	0.52
	Draw	0.36	0.02	0.05	
	Win	0.53	0.86	0.66	

4.4 Optimal Features (From Decision Tree)

From the simple decision tree model with max_depth = 8. The most important features are:

- **Minimum of (1 year historical Average Win Rate) among Away players:** If the minimum player win rate in away is more than 49%, the Home team is likely to lose
- **Sum of Home Player ball control score:** The higher the score is, the more likely the home team will win.
- **Minimum of (1 year historical Average Win Rate) among Home players:** more similar features would provide more chances for home winning.
- **Maximum of (standing tackle) among Home players:** more similar features would provide more chances for home winning.
- **Average of (2 years historical Average Win Rate) among Home players:** more similar features would provide more chances for home winning.
- **The Sum of Home Player ball short passing score:** the higher the score is, the more likely the home team will lose.
- **The Average of the Home Player's potential score:** the higher the score is, the more likely the home team will win.

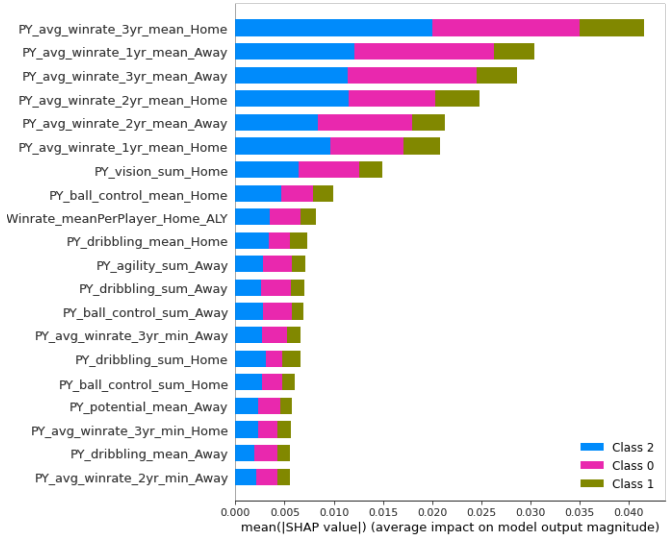
Figure 11: Tree structure for Football Match Prediction



- **The Synergies between player and team, represented by the historical win rate when the player played in this team:** the higher it is, the more likely the home team will win.
- **The Average of Away Players' ball agility score:** the higher the score is, the more likely the home team will Lose. (Lower value will yield more chance to draw)
- **The Sum of Away Player ball dribbling score:** the higher the score is, the more likely the home team will Lose.

Only the top 10 features are described above. For more essential features, refer to Figure 12. However, when we decide on betting, these factors in Figure 12 will be considered the essential factor for betting: Aggregate Win Rate of player, player stats of ball control, standing tackle, potential, short passing, the player with team synergies, agility, and dribbling score, etc... As reported above, the most significant features are not from the team statistic; the most critical factor in determining the win is most likely from player stats and player synergies. These results further enhance the assumption of win rate determination factors. Therefore, we should surely consider these factors with only human prediction before betting on their favorite team. If someone will predict "Team X will win the match or not" in the future, it is much better to look at the player stats and synergies between team/players.

Figure 12: Random Forest Feature Importance of Match Prediction



5. Investment strategy

5.1 Betting Strategy

We tried to create a betting strategy based on the predictions from our model. As seen from the results, our models are better at recognizing Home wins than other cases. Considering the high precision from the Decision Tree Model, we used to make predictions on the Home wins.

5.2 Methodology

The betting odds act as a proxy for the probability estimates of win, lose, draw by the odds provider. Our research identified that the implied probabilities from betting odds have good accuracy in predicting the match outcomes, making it difficult to beat the odds. Also, we need to note that implied distributions are skewed in favor of the odds providers, making it harder to generate profits in the long run. We will illustrate this idea with a simple experiment at the end of this section. In most betting platforms, you can only invest in one of the outcomes (win, loss, draw) in a match. Based on this, we initially tested some betting strategies for all the results one at a time. Due to the low precision for draw and loss, we incurred losses. Hence we pivoted on predicting the home win outcome.

- Based on the odds from **Bet365**, we calculate the implied probability of winning. Using our model's predict_proba_method, we generated probabilities for the outcomes.
- Post this, we created a decision rule and bet $100 on the winning outcome if decision criteria are satisfied.
- We experimented with multiple decision rules and found positive results for the decision rule: **"Model win probability $>$ 0.5 and Odds implied probability $<$ 0.5."**
- Even though our prediction rates are satisfactory for the model, the betting strategy typically has a lower expected value than the odds provider. The betting platform companies also collect extensive data regarding player injuries, health, and form, which aids in their odds calculation.
- To improve this strategy, we suggest improving the Decision Tree model's precision scores for win Class by adding more features indicative of the player's form and skill.

Table 4: Decision Rule and Their Result

Decision Rule	Capital Invested	Final Capital
Model win probability $>$ 0.6	7600	7320
Model win probability $>$ 0.5	27100	21213
Model win probability $>$ 0.55	27100	21213
Model win probability $>$ 0.5 and Odds implied win probability $<$ 0.4	200	180
Model win probability $>$ 0.5 and Odds implied win probability $<$ 0.5	600	730
Model win probability $>$ 0.5 and Odds implied win probability $>$ 0.5	26200	20483
Model win probability $>$ 0.5 and Odds implied win probability $>$ 0.4	26800	21133
Model win probability $>$ 0.4 and Odds implied win probability $>$ 0.75	21100	17700

To illustrate the idea of skewness in implied probability distributions from betting odd providers, we ran a small experiment. Using the odds from Bet365, we calculated the probabilities of win, loss, draw and used them to predict the match outcome. Based on this outcome, we bet $100, and results from the experiment are tabulated in Table 5. Even though the odds are generally indicative of the outcome, we were unable to make any accurate predictions in draw/loss outcomes. This strategy invests in every game and doesn't consider the probabilities' magnitude, but the results indicate skewness.

Table 5: Result if Betting on Bet365

Outcome	No.bets made	Capital Invested(in $)	Final Capital(in $)
Win	16,168	1,616,800.00	60,370.00
Draw	28	2,800.00	0.00
Loss	5,869	586,900.00	0.00

The strategy of our ML Model produces a better monetary value out of the sample environment. The betting odds from the website, however, lead us with a lower financial value. We simulated every scenario here, betting win/draw/lose. The betting outcome [in monetary value] from betting odd website was way worse than the one from our ML models. So we conclude that even though the prediction performance might have a little different value, our ML model generates better monetary value in the out-of-sample environment.

6. Conclusion and Summary

As a result of various data science methodologies and analysis, our paper finds a model for football match prediction using historical football and odd betting data. We find that despite somewhat accurate predictions solely relying on these statistics according to the betting-odd websites provided low earnings short term and negative profit in the long term.

Our analysis shows that if we want to predict the win/draw/lose by ourselves, we should also consider other important features such as the players' statistics and the aggregated statistic of the player's historical win rate for the team. The individual players' ball control, potential, dribbling, and agility as well player synergy with their teammates indicate more title success. Finally, we can see the improvement in the prediction results by comparing our Tree model and the website's betting odds. Even though the difference is not by a lot in terms of prediction evaluation, the profit value of betting improves drastically.