# 2 Preparation of the Models

## 2.1 Data Cleaning

|  |  |  |
| --- | --- | --- |
| **Data Name** | **Processing Type** | **Feature Name** |
| **Side** | Map + Dummy | Side\_1, Side\_0 |
| **Coach** | Dummy | Coach\_1, Coach\_2, Coach\_3 |
| **Opponent Strength** | Analysis | Oppo |
| **Shots** | Count | Attack |
| **Dribbles** |
| **Touch** |
| **Corner** |
| **Offside** |
| **Tackle** | Count | Defence |
| **Dispossess** |
| **Aerial Won** |
| **Interception** |
| **Clearance** |
| **Blocks** |
| **Saves** |
| **Passes** | Count | Pass |
| **Possession** | Search + Integrate |
| **Pass Success** | Calculate |
| **Foul** | Count | Fail |
| **Loss of Possession** | Search + Count |

## （图名）

## 2.2 Processing Tools

|  |  |
| --- | --- |
| **Tool** | **Uses** |
| **Visual Studio Code 1.42** | Coding, Visualization |
| **IPython 3.6.8** | Run Code |
| **Visio** | Design Flowchart |
| **Excel** | Arrange Dataset |
| **GitHub** | Synchronization, Storing |

## （图名）

# 3 Establishment of PNM and Analysis of Influence Factors

In order to construct a structured passing network, which is used to analyze the tacit understanding of passing between players, it should be analyzed in different dimensions and states. For example, from the behavior between two players at the micro level, to the behavior between multiple players at the macro level; and the time scale from the unit time in the game to the entire season.

## 3.1 Pass Evaluation Index（PEI）

The evaluation index of pass between two players is used to evaluate the degree of cooperation between them. In a game, from a macro perspective, players can be regarded as nodes, the field can be considered as a network, and each pass can be considered as the connection between the nodes. We define as the pass evaluation index for each pass. In a multiplayer pass evaluation system, three nodes are connected into a closed loop, and the sum of edge weights is the 3-player pass evaluation index.

According to the experience of life and the rules discovered by data mining, a PEI calculation model can be constructed as follows:

(1) Weight table of pass types:

（图名）

(2) Calculate pressure from defenders when passing or receiving

For x is the abscissa from the player to the opponent's gate, and it is negatively related to the pressure from defenders.

（3）Single pass evaluation, , is the weight of the pass type multiplied by weighted average of the pressure from defenders, quantified as the following formula:

According to this pass evaluation index model, an adjacency matrix of all N players participating in the game within a certain time range is calculated. From , we can get the sum of all values of pass evaluations between each two players:（value蓝色要改成黑色）

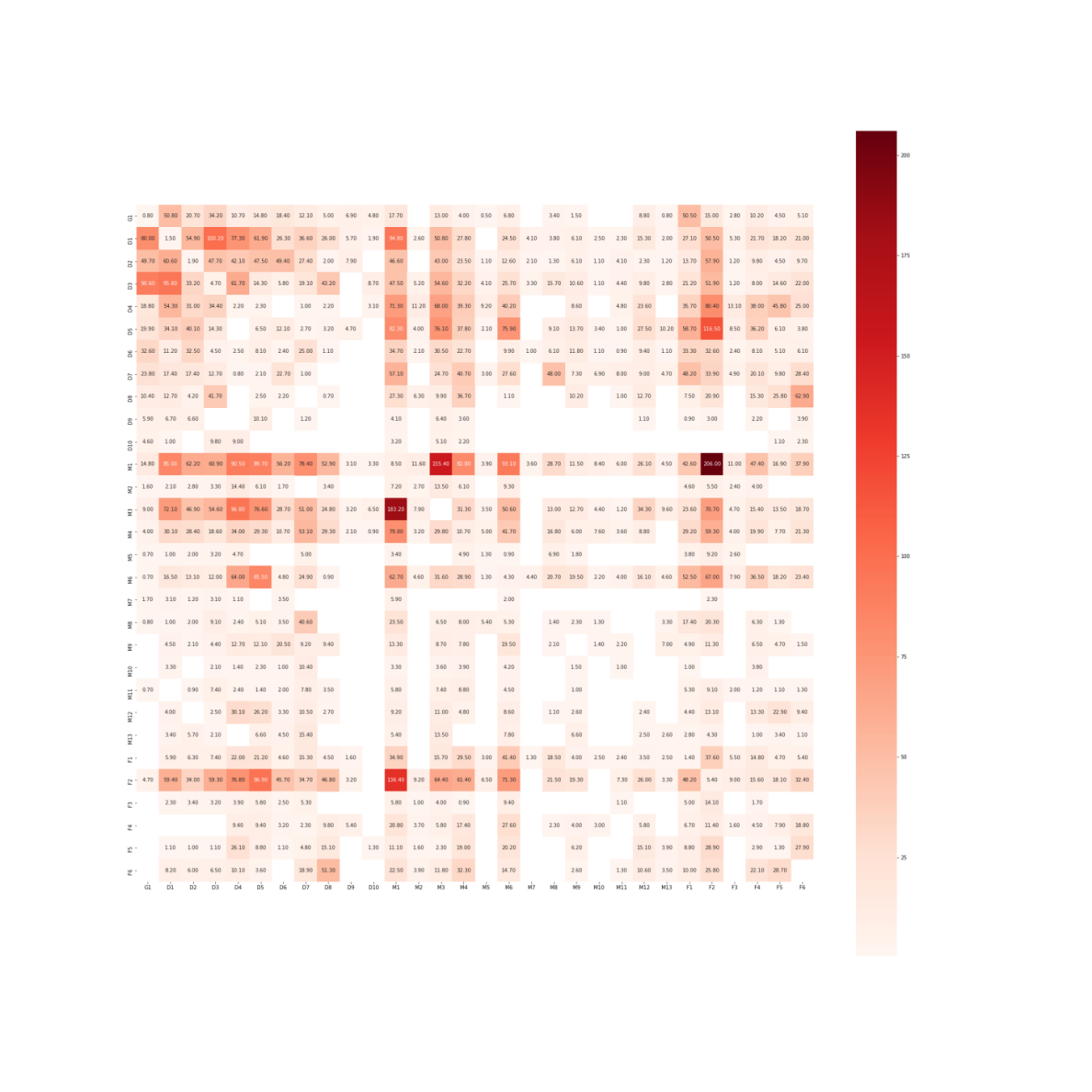


Figure 1 PassValue(i) for whole season

## 3.2 Pass Network Model (PNM) Establishment and Recognition of Network Pattern

The connection between two players on the network, Macroscopically, it is the sum of the evaluation of passing between players. Screening the edges whose pass evaluation exceeds a certain threshold, using graph theory methods to selectively remove the crossing edges, and visualize the line-up passing network constructed based on the pass evaluation index, which is expressed by the shades of the line.

：



Figure 2 Pass network in one game

From the visualization of this model, we can intuitively analyze the players who pass frequently and tacitly, and can also see the combination of multiple passing cooperation among the main players intuitively.

（字体）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **N** | **Players** | | | | **Score** |
| **2** | M1 | F2 |  |  | 342.4 |
| M1 | M3 |  |  | 338.6 |
| D5 | F2 |  |  | 213.4 |
| **3** | M1 | M3 | F2 |  | 816.1 |
|  | D5 | M1 | F2 |  | 727.8 |
| **4** | D5 | M1 | M3 | F2 | 1113.5 |

## （图名）3.3 Fluctuation of Passing State at the time

Passing frequency fluctuate on the time scale.

And is used as an indicator of the team's real-time status. At the beginning of the game, the players' bodies were not warmed up, resulting in a low probability of passing. After 5-10 minutes, Passing efficiency gradually improved and generally stabilized, that is:

As the time goes, the players' physical strength and the pass density decrease, that is, the increase in the number of passes slows down (although the number of successful passes in a game is still increasing, the frequency of pass failure begins to increase), after that the pass density Showing a downward trend, that is:

Looking at the pass frequency density of players in 38 games throughout the season, one the whole the trend of frequency density is the same as that showes in a single game. Plot with time as the abscissa and successful pass density as the ordinate:

Figure 3 Whole Season Pass Density Figure 4 No.1 Match Pass Density

Generally, the density of passes is relatively stable in a time scale. If we use the Monte Carlo method to simulate each pass, the time probability distribution for the next pass after the last pass is set to obey, where is the statistical average of interval time of pass. Then when the sample size N satisfies , it will approximate the distribution on the left graph; as the sample size increases, when is satisfied, it will approximate the distribution on the right graph. Therefore, we can consider that the probability of passing events at each time point obeys .

# 4 Soccer Team Indexes and Performance Prediction Based on ML

There are many indicators for successful teamwork in a football team. Through data analysis and practical experience, we mainly consider the following indicators: static indicators and dynamic indicators. First, we use to evaluate the overall performance of a team in a game. define :（蓝色改）

## 4.1 Static Index (SI)

In order to consider the distribution of players’ positions, we took the position coordinates of each player throughout the season and made a heat map. The value of each point in the heat map is defined as follows:

The darker the color is, the more active the player is in this position. After calculating , the position heat map of the main 11 players is got as follows:

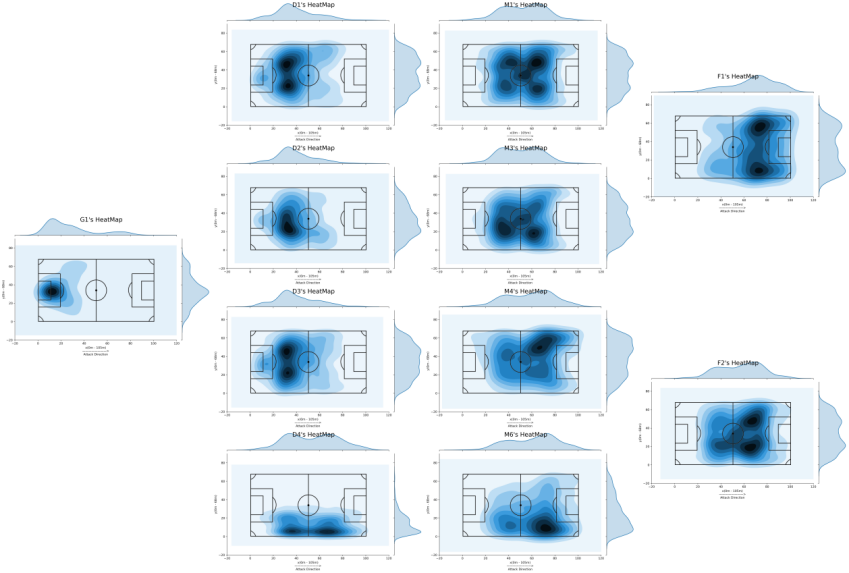


Figure 5 Players Heatmaps

In a match, team formation plays an important role in collaboration. We want to find out what the formation is like. We take the coordinates of each player in each match and integrate the coordinates over time to find out the average coordinates. We take the time which can be got from data (the player is Origin / Destination Player) as the new abscissa, and the X or Y coordinate as the new ordinate, so we got functions X (t) and Y (t). Approximately, we thought that between any two closest recorded time points, the player moves at a constant speed in the X or Y direction, so that the discrete data set is converted into a continuous dataset for each match. So the average coordinate, taking the X coordinate as an example (the Y coordinate is the same), is:

Plot these 11 players’ average coordinate on the map, we got the formation graph of each match. Some of them are as follows:



Figure 6 Average Co-ordinate (Approximate Formation)

## 4.2 Dynamic Index (DI)

Dynamic indicators include the team's man-made influence factors and technical data generated in the match: artificial influence factors include coaches, opponents' levels, home or away, and technical data include statistics on various events including shooting, passing, clearance etc. The original data uses a single event as a sample unit, and we classify it as dynamic data in units of one match. By observing the data stored in the new structure, we can extract some of the feature information.

### 4.2.1 Data Cleaning and Feature Engineering

In feature engineering, in order to reduce the dimensionality of features, we can not only use PCA to screen and remove features that have no significant impact, but also use ChiMerge feature binning method to divide EventSubTypes into four aspects: passing, defense, and fail. These aspects along with coaches, home or away, and opponent’s levels is considered to be the features of a match, and use standardized, dummy variables, combined analysis and other methods to process the statistical data to quantify it:

1. Statistical data
2. Multi-event combined analysis data
3. One-Hot encoded dummy variable data

### 4.2.2 Visualization Analysis

Analyze the effect of on and :

（图名）

主客场与得分的关系图

When , has more distribution, and has more distribution. Therefore, the overall performance at home is better than away.

Analyze the coaching levels of different coaches and the effectiveness of coaching for the team :





不同教练指导下球队4种表现数据和净胜球对比图（建议用黑色不要用灰色）

From the figures, we can find that under the guidance of Coach 3, the team's and other features are better, followed by Coach 2 and Coach 1. We can also show their coaching styles, for example, coach 1 is more aggressive, while his defense is mediocre; coach 2 emphasizes violent defense; coach 3 is more balanced and has the best performance.

Analysis of and 's contribution to :



进攻、传球与净胜球之间关系图

From the figure we can find that under different goal difference numbers, the attack and the pass are generally linearly related, with a positive slope.

is positively correlated with and , and the more the distribution concentrated, the smaller the variance of and . We can conclude that the more in a game or even the entire season, the higher the probability of better and is.

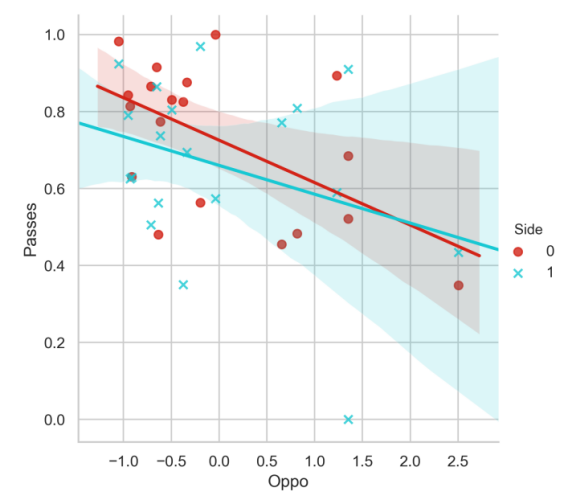
Analysis of and 's contribution to :



防守、失败与净胜球之间关系图

is positively correlated with , negatively correlated with , and the more the distribution concentrated, the smaller the variance of and . Observations: The points on the left of Figure 2 are distributed at the bottom of square, so a bad defense will lead to a loss; there is no point on the left half of the right, so you can't make too many mistakes if you want to win.（蓝色）

Take as the positive indicators for examining the overall performance of the team, along with , for multi-angle analysis:



左：进攻和防守之间关系图 右：对手水平与传球关系图

From the picture on the left, we can find that the center of gravity of the data is distributed in the lower right corner. It is believed that (attacking performance) is significantly better than (defensive performance) throughout the season. From the picture on the right, we can find that , whether at home or away, but it is more likely to have a small improvement at home. The conclusion is that the higher the opponent's level, the lower the relative value of our pass.

Synthesize all the processed features, and estimate the correlation of the pairwise features among the variables by calculating the Pearson correlation coefficient.

Let the matrix :



Figure 7 Correlation Coefficient Matrix From DI

### 4.2.3 RFC Establishment, Optimization, and Training

We use as the evaluation label for each match. We hope that the learned model can classify the game based on the processed data and correspond to the label. Due to M = 10 features are too many and their correlations with labels are different, it is not appropriate to use a linear model for classification; and the number of sample data N = 38 is very small, so it is easy to have large deviations when using some deep learning algorithms. In summary, we choose a random forest model to build a label classifier.

Random forest is a classifier containing multiple decision trees, and the output is determined by the mode of the output by individual trees. For many kinds of data, it can generate a high-accuracy classifier; it can evaluate the importance of variables when determining categories; and when it builds forests, it can internally produce unbiased estimates of generalized errors. The method for establishing a Random Forest Classifier is as follows:

1. Input the number of features m, which is used to determine whether the decision result of a node on the decision tree meets
2. Use Bootstrap sampling to sample N times from the N samples with a sampling method to form a training set, and use the unselected samples as predictions to evaluate their errors
3. For each node, randomly select m features. The decision of each node in the decision tree is determined based on these features. Calculate the best splitting method based on these m characteristics;
4. Each tree will grow completely without pruning. This may be used after building a normal tree classifier.

After training the random forest classifier, use grid search to optimize the parameters and select

as parameters, the K-fold cross validation test was used to calculate its accuracy score, which was used to evaluate the accuracy of the model.

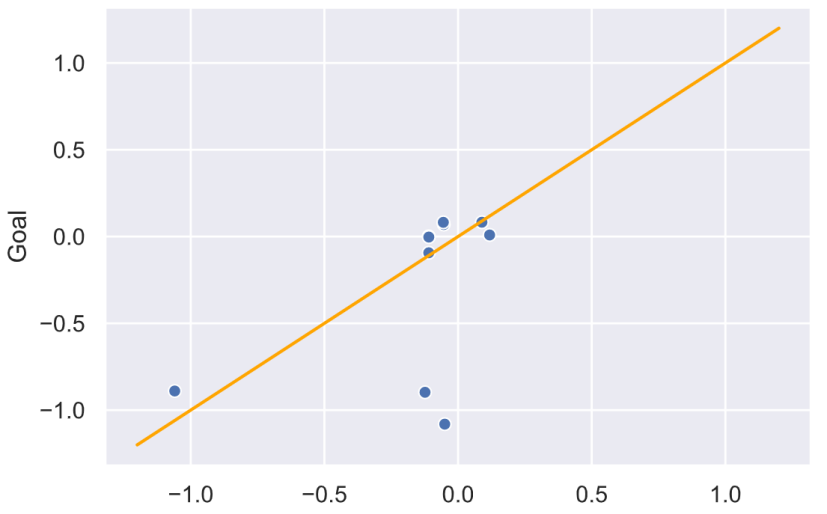


Figure 8 Result of Cross Validation

After certain data adjustments and simulation results, the average score is 65.8%, and the best situation can reach a score of 80-90%. When the sample size is only N = 38, we can accept the accuracy of this model to predict the goal difference by dynamic indicators.

没什么大问题，建议：图名用黑色，图名要简练不要一长串。（这部分大都机翻，有疑问处已确认）没啥问题

哦还有，没有支撑材料，我看我们还没到22页，可以加一两个表格