## 5.2 model 2

足球队中成功团队合作有许多指标，通过数据分析和实际经验，我们主要考虑以下indicators：静态指标和动态指标。首先，我们使用 评价一场比赛的球队整体发挥，作为单场比赛表现标签，定义：

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 :

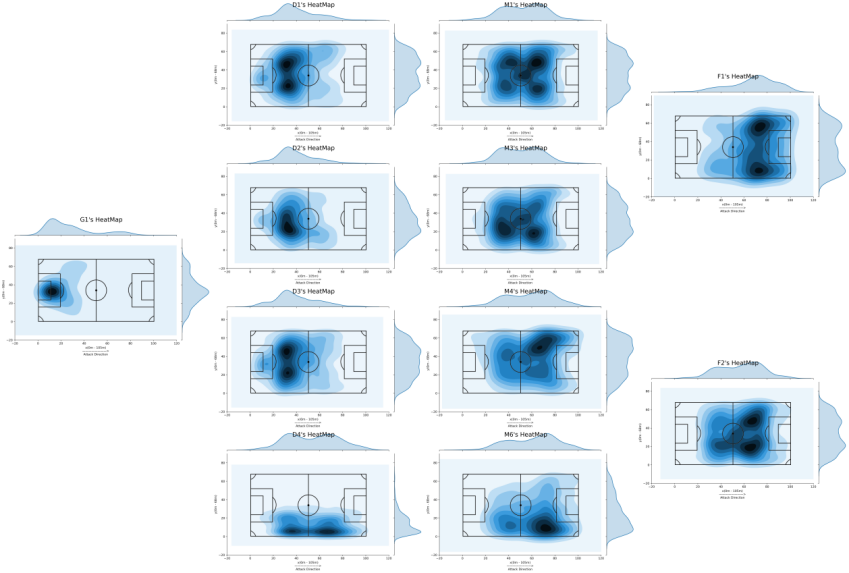
### 5.2.1静态指标

为了考虑球员位置分布，我们采出每个球员在整个赛季中的位置坐标，做出球员运动位置的热点图，热力图每个点的值定义如下：

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:

颜色越深则表示出现在此处的频率较大，越浅表示出现的频率越小。经过的计算(经过计算…)，主力11人的位置热点图如下：

The darker the color is, the more active the player is in this position(, and the lighter the less the frequency.) (没有必要)After calculating , the position heat map of the main 11 people is got as follows:



球员位置分布热点图

在一场球赛中，球队的阵型对团队协作起到重要作用，我们考虑在一场球赛中球员阵型，我们采取每一场比赛中每一位球员的运动坐标，采用坐标对时间积分的方法，找出每场球赛中，每一位球员平均坐标。将在数据中可以获取（球员出现在Origin/Destination）的时间点作为新的横坐标，X或Y坐标作为新的纵坐标，得出函数。我们近似认为在任意两个有记录的时间点，球员在X或Y方向上匀速移动，这样就将离散型的数据集转换为了连续性的数据集（每个）。因此平均坐标，以X坐标为例，Y坐标同理：

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 obtained 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 data set (each). So the average coordinate, taking the X coordinate as an example (the Y coordinate is the same), is:

将这11位球员的位置标在图中绘制出每场球赛的阵型图，部分阵型图如下：

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



Match 1 and Match 11球员阵型图

### 5.2.2动态指标

动态指标包括了球队人为影响因素和在比赛里产生的技术数据：人为影响因素包括了教练、对手水平、主客场，技术数据包括了射门、传球、解围在内的各种events统计。原始的数据以单个事件作为样本的单位，而我们将其分类统计为以一场比赛为单位的动态类型数据，通过观察以新结构存储的数据，提取出其中的若干特征信息。

Dynamic indicators include the team's artificial 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.

#### 5.2.2.1数据清洗和特征工程

In feature engineering，为了降低特征的维度，不仅使用PCA筛选并剔除影响不显著的特征，还可以使用ChiMerge这一特征分箱的方法，将EventSubTypes分为传球，进攻，防守和Fail四个方面，与教练、主客场、对手水平一起作为一场比赛的特征。通过标准化、哑变量、结合分析等方法处理统计后的数据来量化比赛的特征：

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's feature binning method to divide EventSubTypes into four aspects: passing, offense, defense, and fail. These aspects along with coaches, home or away, and opponent’s levels is considered to be the characteristics of a match, and use standardized, dumb variables, combined analysis and other methods to process the statistical data to quantify it:

（1）统计型数据 Statistical data

（2）多事件结合分析型数据 Multi-event combined analysis data

（3）One-Hot编码哑变量数据 One-Hot encoded dummy variable data

#### 5.2.2.2 可视化分析

分析对于影响：

Analyze the effect of on and :



主客场与得分的关系图

时的分布更多，分布更高，因此主场表现结果整体上比客场要好。

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

分析不同Coach的执教水平以及对于球队的指导成效：\

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

从boxen图我们可以看出，在Coach 3指导下，球队等数据较好，其次是Coach 2和Coach 1。我们还可以得出教练们的执教风格，例如：教练1更具侵略性，防守就显得平庸；教练2强调强硬防守；教练3则较为平衡，战绩最佳。

From the boxen chart, we can find that under the guidance of Coach 3, the team's and other data are better, followed by Coach 2 and Coach 1. We can also draw their coaching styles, for example: coach 1 is more aggressive, the defense ??? is mediocre; coach 2 emphasizes tough defense; coach 3 is more balanced and has the best record.

分析、对于的贡献：

Analysis of and 's contribution to :



进攻、传球与净胜球之间关系图

从图中我们可以看出，在不同净胜球数下，进攻和传球大体上为线性相关，斜率为正。

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

与呈正相关，且分布越集中，的方差较小。我们可以得出结论：在一场球赛乃至整个赛季，越多，大概率有着更高的。

is positively correlated with and , and the more concentrated the distribution, 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 high and is.

分析、对于的贡献：

Analysis of and 's contribution to :



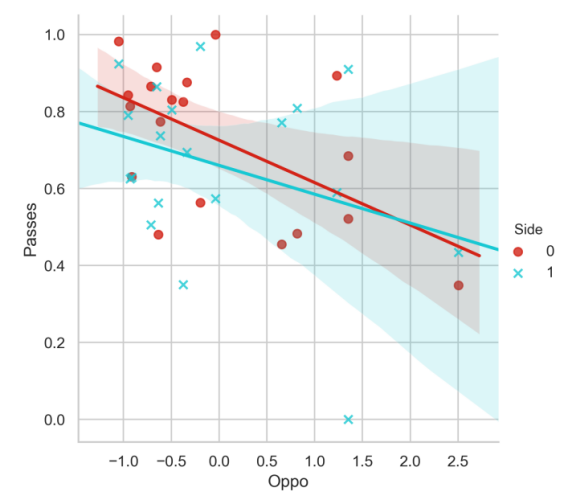
防守、失败与净胜球之间关系图

与呈正相关，与呈负相关，且分布越集中，的方差较小。观察发现：图2左1的点分布在下方，因此防守不好会导致输球；右1左半边没有点，因此期望赢球则失误不能多。

以作为考察球队整体表现的positive指标，结合指标进行多角度分析：

is positively correlated with , negatively correlated with , and the more concentrated the distribution, the smaller the variance of and . Observations: The points on the left of Figure 2 are distributed below, 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.

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

从左图中我们可以看出数据重心分布在右下角，认为整个赛季上（进攻表现）显著优于（防守表现）。从右图中我们可以看出不论是在主场还是客场，，但主场更可能有较小提升；结论是对手水平越高，我方传球率越低。

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 (offensive 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 rate of our pass.

综合所有处理得到的特征，通过Pearson相关系数的计算来估计出变量间两两特征相关性。

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

令矩阵，得：

Let the matrix , get:



动态指标整合的特征相关系数矩阵

#### 5.2.2.3 model建立and训练

我们以作为每场比赛评价标签，希望学习后的模型能够基于处理后的数据对比赛进行分类，对应到的标签。由于个特征数量较多，且与标签相关性不一，不宜采用线性模型进行分类；且样本数据数量极少，在尝试一些深度学习算法时容易有较大偏差。综上，我们选择随机森林模型建立标签分类器。

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 trying some deep learning algorithms. In summary, we choose a random forest model to build a label classifier.

随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定。对于很多种资料，它可以产生高准确度的分类器；它可以在决定类别时，评估变数的重要性；在建造森林时，它可以在内部对于一般化后的误差产生不偏差的估计。建立随机森林分类器Random Forest Classifier的方法如下：

Random forest is a classifier containing multiple decision trees, and the output category is determined by the mode of the categories 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. 输入特征数目，用于确定决策树上一个节点的决策结果；

Input the number of features m, which is used to determine whether the decision result of a node on the decision tree meets m <√ (2 & M)

1. 利用Bootstrap取样，从个训练用例中以有放回抽样的方式，取样次，形成一个训练集，并用未抽到的用例作预测，评估其误差；

Use Bootstrap sampling to sample N times from the N training use cases with a sampling method to form a training set, and use the unselected use cases as predictions to evaluate their errors

1. 对于每一个节点，随机选择m个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这m个特征，计算其最佳的分裂方式；

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;

1. 每棵树都会完整成长而不会剪枝，这有可能在建完一棵正常树状分类器后会被采用。

Each tree will grow completely without pruning. This may be used after building a normal tree classifier.

随机森林分类器的训练后，使用网格搜索grid search进行参数调优，选定

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

作为参数，利用K折交叉验证验计算其accuracy score，用于评估模型准确率。

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.

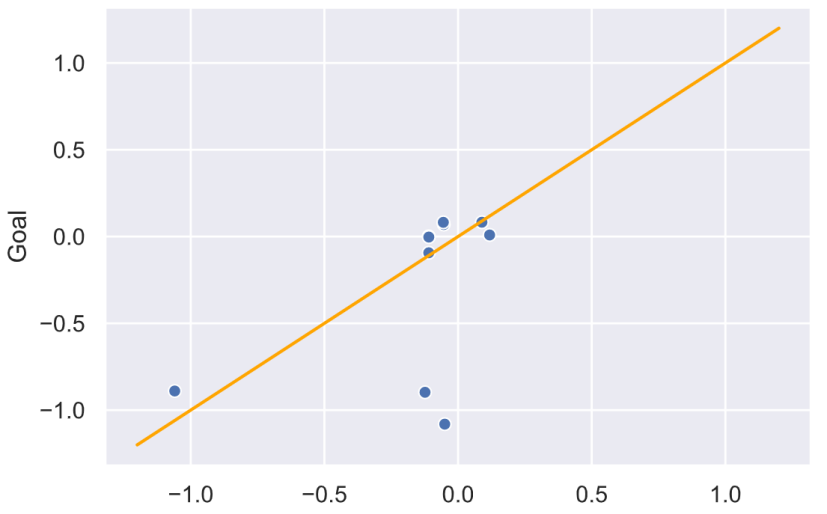


图 1模型预测交叉验证情况

经过一定的数据调整和多次模拟结果，平均情况下得分为，最好的数据情况下可以达到的得分，在样本规模仅有的情况下，我们可以接受这一模型通过动态指标对比赛净胜球情况进行预测的准确率。

After certain data adjustments and multiple simulation results, the average score is 65.8%, and the best data 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 through dynamic indicators.