Statistics department, University of Wisconsin-Madison'

STAT 479 Project Overview

Richard Yang
Zhoujingpeng Wei
Jiaxin Hu

December 5, 2018

Content

Motivation

- In the World Cup 2018, France won the final champion, and it's their strategy that realy impresses the world—sometimes France even gives up the ball possession and use counter-attack as their only way to attack.
- This world cup let us recall some very common phenomenons in modern professional league—the high game domination or shot attempts usually leads to a bad outcome, and that's kind of counter-intuitive.

Goal and General idea

1. In this report, we want to use the Premier league data to analyze what kind of stratrgies tend to have positive effect when confronting different opponents.

Goal and General idea

1. In this report, we want to use the Premier league data to analyze what kind of stratrgies tend to have positive effect when confronting different opponents.
2. To finish this goal, we should first rate different aspects of a team, because the team's general performance is the most fundamental factor.

Goal and General idea

1. In this report, we want to use the Premier league data to analyze what kind of stratrgies tend to have positive effect when confronting different opponents.
2. To finish this goal, we should first rate different aspects of a team, because the team's general performance is the most fundamental factor.
3. Then we use above result, and find some proper algorithm to make classification or regression and analyze the model to get conclusion.

Step1:Rate team
PCA-Clustering

This part contains 3 steps

- Get the data from averaging performance over different games of each team, and then normalize the data
- Using PCA for different subgroup of features respectively
- Clustering using the transformed data

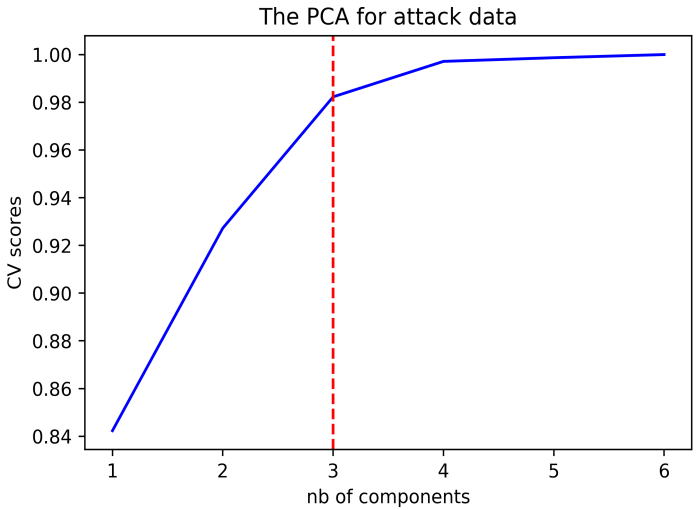
some explanation

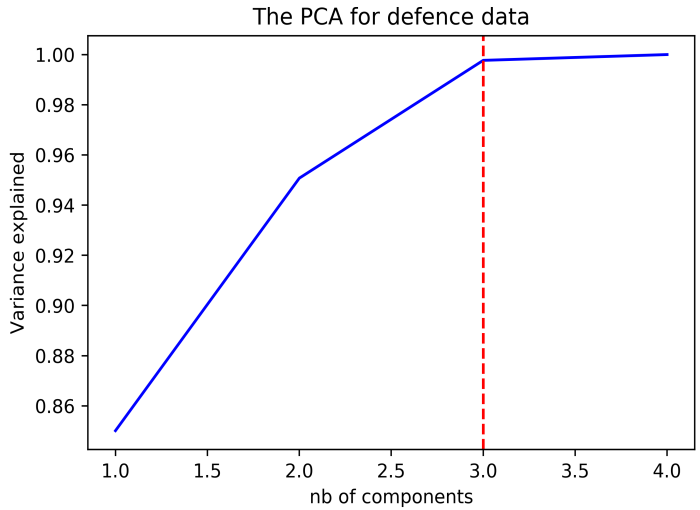
We divide our features into three sub-groups:

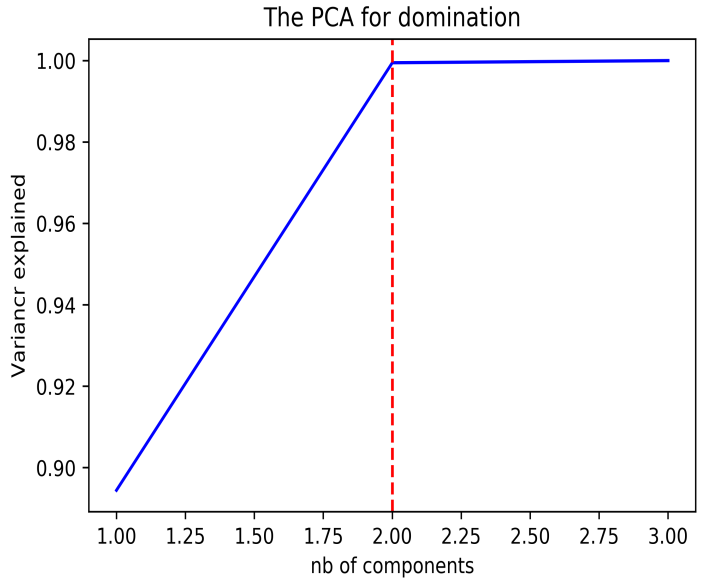
Attack	Shots, ShotsOT, Dribbles, Fouled ,Goals, Offsides
Defence	Shots conceded, Tackles ,Interceptions , Fouls
Dominating	Possession, Pass Accuracy, AerialsWon

- Features in each sub-group have strong muliticolinearity, and the reason is quite obvious.
- If we just use the raw data, it doesn't make sense to compute metrics like Euclidean distance.
- That's the reason why we use PCA, which can reduced features into several orthognal components.

The results of PCA







About clustering

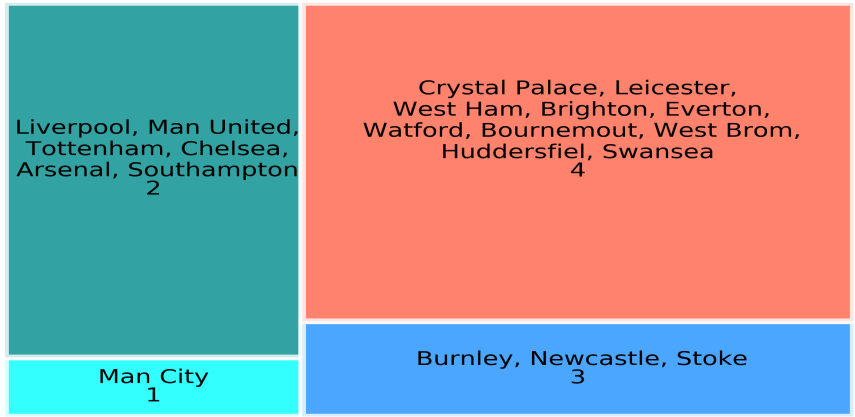
- So we extract 3, 3, 2 components for attack, defence, domination respectively, and want to rate each team using these components,give scores based on their attack, defence and domination for every team..
- However, these components are hard to explain, because they don't have a specific meaning(features can reflect both the positive and negative parts of a game), **so we choose to use clustering to cluster teams in terms of different aspects.**
- Although the higher the number of cluster k, the lower the square difference within group, we still cannot choose a big k value, for that doesn't make sense. And we finally choose $k = 4$, which can explain the major similarity and difference.

clustering result

Clustering result in terms of Attack

Clustering result in terms of defence

Clustering result in terms of domination



Step2: Statistical analysis
using softmax regression

Why we choose softmax:

Intuition

- Game results are not determined value given features, it should be a probability, and softmax is used to simulate the multinomial distribution, which corresponds to the distribution of our response(the game results).
- The probability of different responses varies according to features, which can explain the reality that different game statistics and different team and its opponent level can cause different probability to win or lose the game.

Implementation

- Because the softmax equation is $Pr(y = j|\boldsymbol{x}) = \frac{e^{\boldsymbol{x}^T w_j}}{\sum_k e^{\boldsymbol{x}^T w_k}}$, we can analyze the coefficient to determine the relationship between each feature and game results.

Softmax results

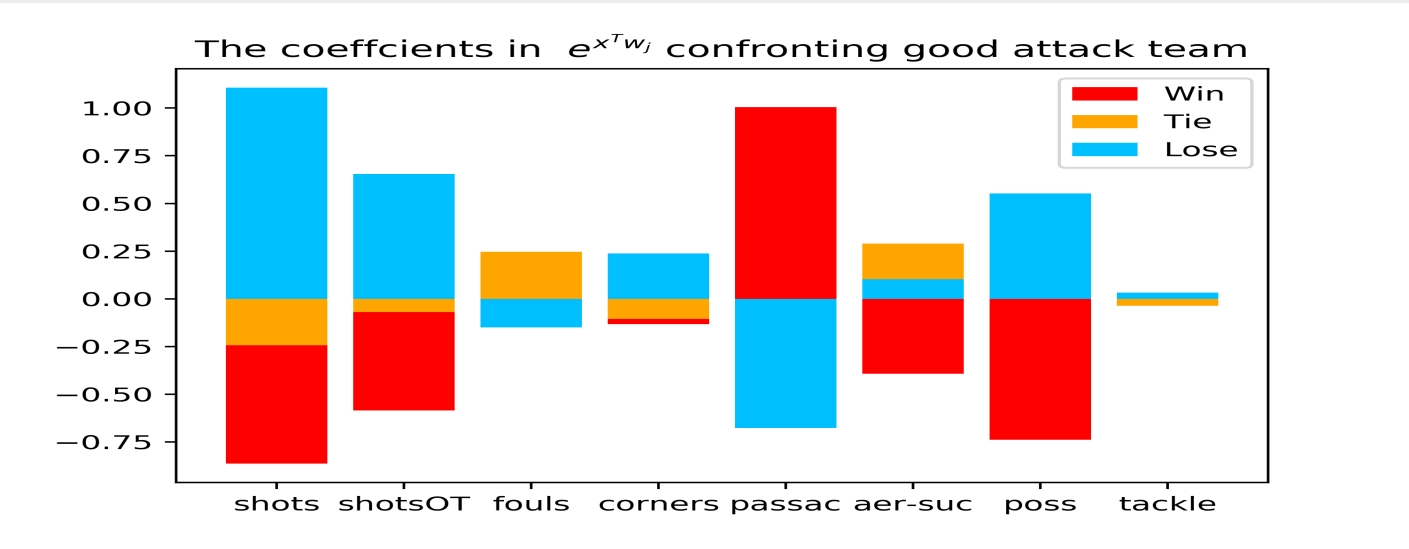
In order to proceed our ultimate goal, we first need to make sure that this model can explain the logic of the games.

- We use 10-folds stratified cross-validation, and choose the l_2 penalty to implement our algorithm, and average the accuracy to get the score.
- We first just consider plugging the label assigned in the step of clustering back in the corresponding game statistics (We treat the labels as nominal features), and fit the general model, **it achieve the accuracy of** 61% **in training data ,and** 58% **in cross validation**. Because the success for random guess is just $\frac{1}{3}$, and the games result is actually really tricky to predict, it's not a bad result.
- However, we think we would have a better result if considering interaction effects between the team level and the game statistics, such that game statistics would have different effect on teams with different levels , which coincides with our intuition.
- So we 'd better analyze diffrent team-opponent pairs respectively.

Analysis and inference

We just select some pairs with obvious interaction, and analyze the coefficients.

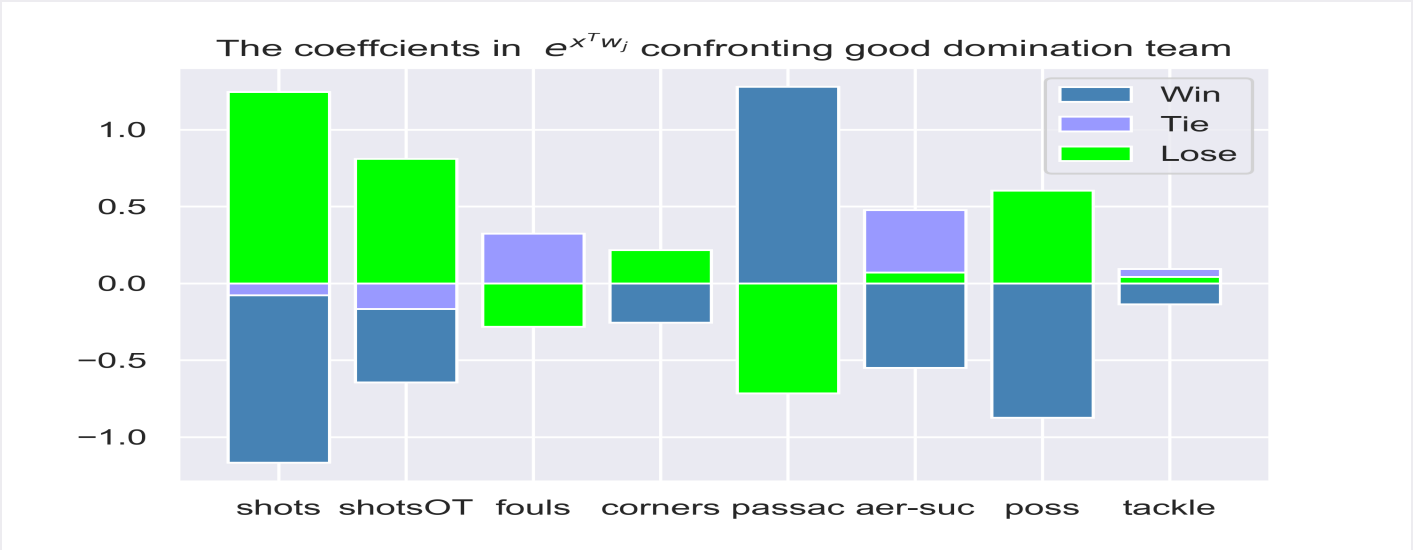
The games with high attack level opponents (The group1)



When confronting teams with strong attack level (group1), the probability of win or tie is negatively correlated with shots, shots on target and ball possession. And the accuracy for this part is 0.68, which explain the data better.

Analysis and inference

Games with high domination level opponents(group 1,2)

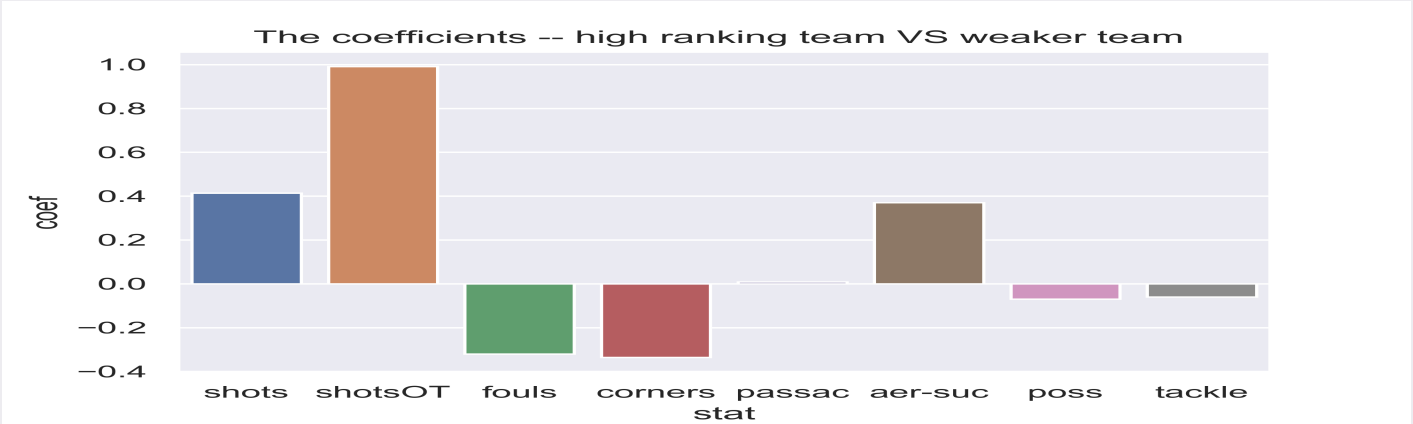


The results are similar as the previous one, because teams in these two kinds of groups have a strong overlap.
The two result shows that when the opponent is strong, we should use a conservative strategy and improve our pass success to gain points(win or draw).
And we guess the strategy statisfying the above data is counter-attack.

Analysis

When a stong team have games with a weaker one in terms of attack and domination. (Accuracy 0.75)

The probability of win



- For a high ranking team, having a tie with a weak team has the same effect as losing the game, so we just consider the probability to win (using regular logistic regression).
- According to the graph, we know the probability to win is positively correlated with shots, shots on target, aerial-ball success, which means the high ranking team should use an aggressive strategy to win the game.

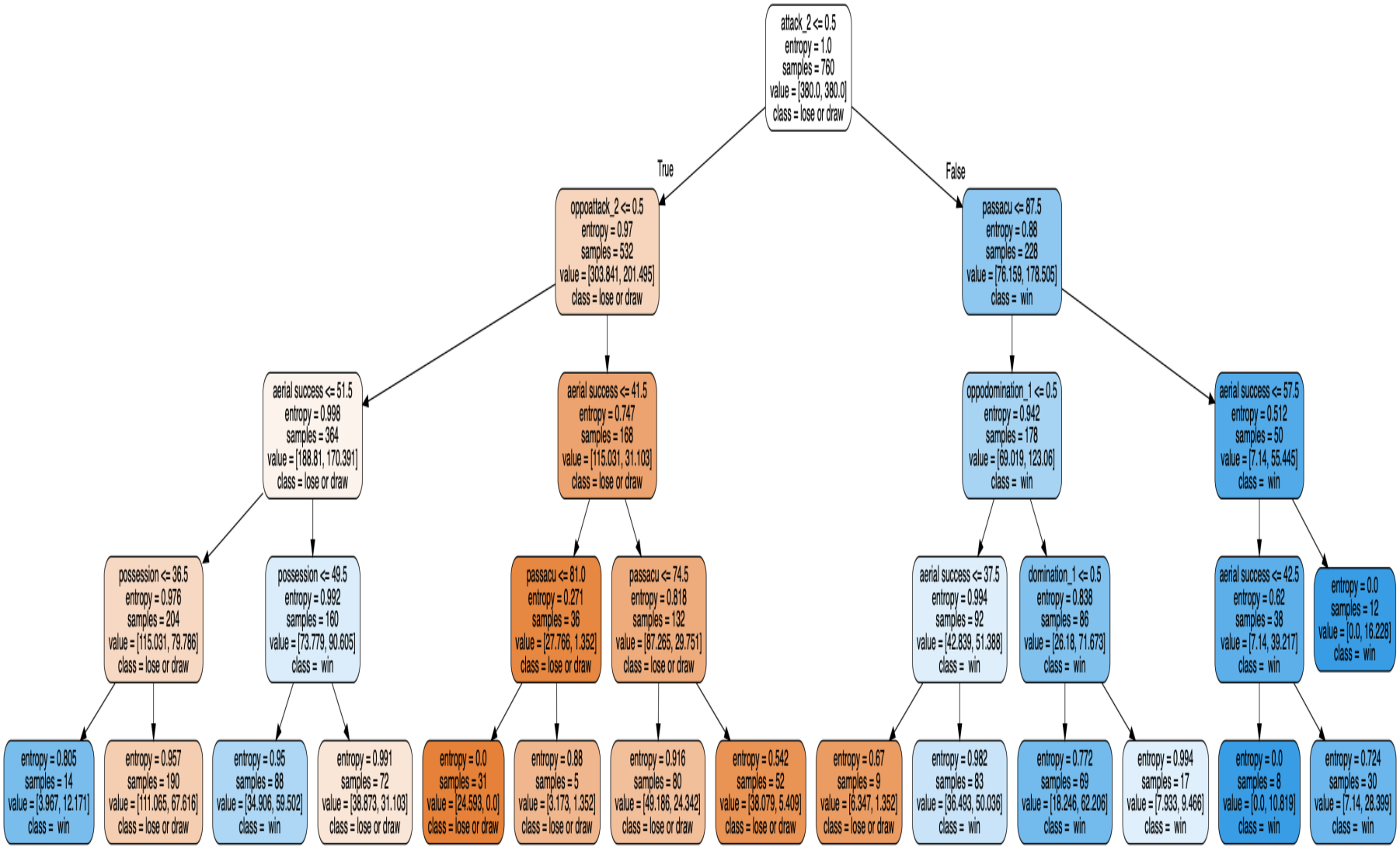
Step 3: Decison tree

Another less abstract way is to use decision tree, because the game results (win, lose, tie) would be too specific, we instead combine two of them together to form binary response.

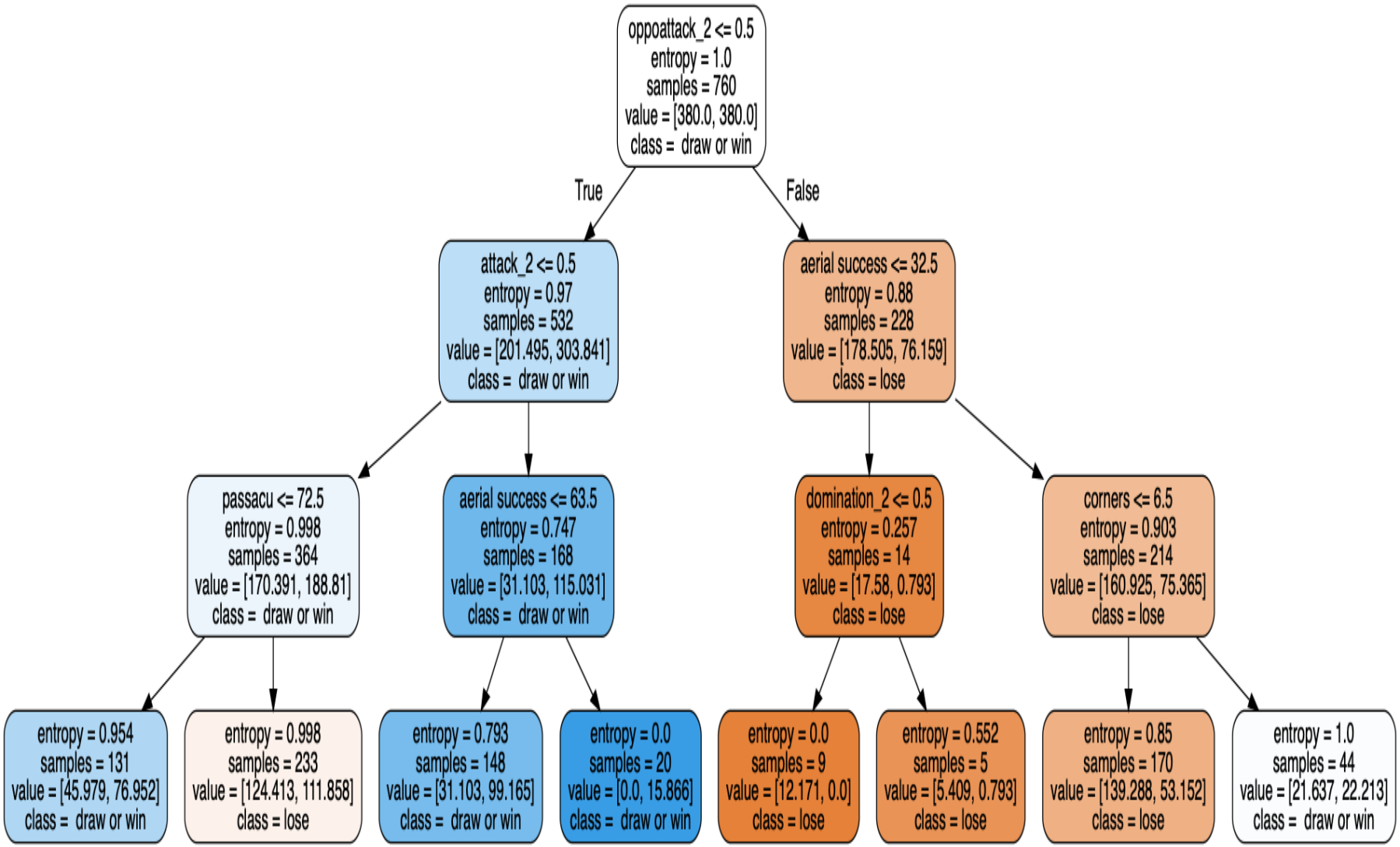
Because decision tree has the most intuitive way to do classification, and its leaf nodes correspond to rules, which can give us most clear explanation.

We try different models, and plot following two graphs that best explain the logic, and the accuracy is 0.74 and 0.76 respectively, which can well explain the potential information.

win∼ lose or draw



win or draw∼lose



Analyze the decision rule

In the decision tree, we know every leaf node represents a rule, so we can find some useful information from it.

In the first graph

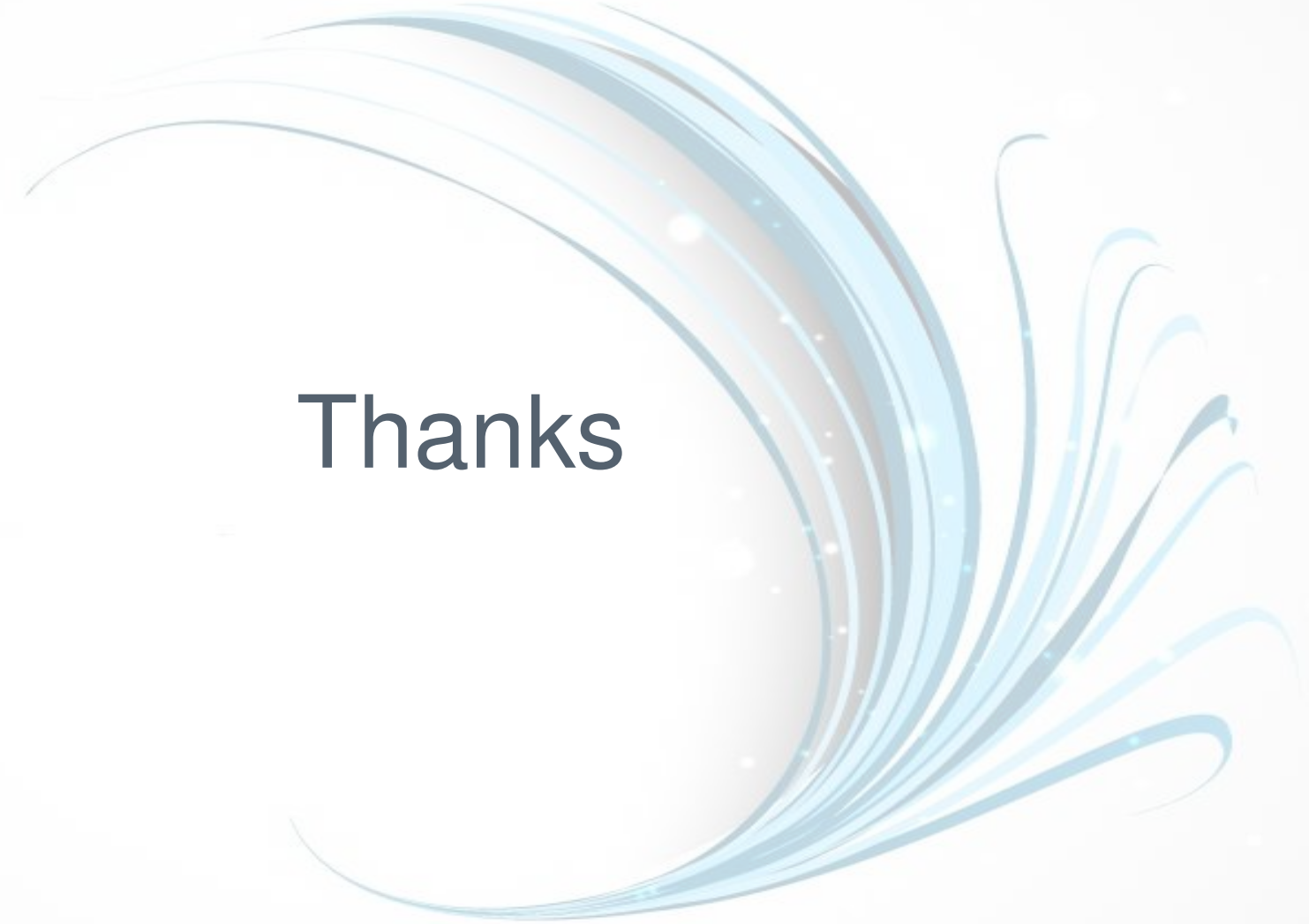
Strong attack team would lose or draw games only when **(1) its pass accuracy is not extremely high**$\rightarrow\rightarrow$**(2) but its domination level is high** $\rightarrow\rightarrow$ **(3) aerial success rate is low.** (which means these team don't play games in their own style and is forced to play in an unfamilar way.)

In the second graph

When confronting strong attack level opponents, the only predictable way to gain points (draw or win) is to **(1) have a high aerial-success** $\rightarrow\rightarrow$ **(2) have many corners** .

Conclusion

- After all analysis, we may draw conclusion that when having matches with high ranking teams, we'd better use a conservative strategies to gain points, like counter attack or improve the aerial-success.
- On the other hand, we should try some aggressive way to win the game when facing a relatively weak team, and avoid losing by enhancing the aerial ball success.


Thanks