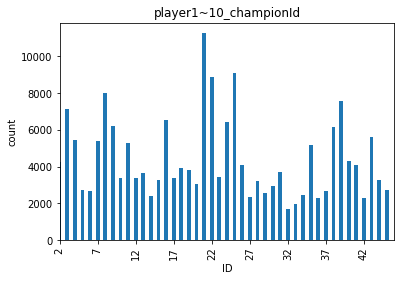
**LOL Feature Extraction**

**PB20000156 徐亦昶**

**Part 1 The exploration of data distribution**

At the beginning,I’d like to explore the distribution of the data.The first thing is team1\_win.As the result shows,38525 teams won and 41425 teams lost.

In the picture above,1 and 0 indicates whether a team wins.

Then championId.I explored the distribution of player1 to player 5,and then player 1 to 10 totally.I found that in every distribution,21 is the most popular ID.

Then I made a deeper observation on ID 21,to see if that ID has any relationship with the game’s result.



result:

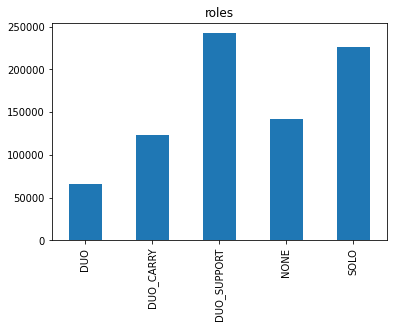
win 2854

loss 2875

accuracy: 0.5018327805899808

Unfortunately,the champion ID has nothing to do with the result.

Another thing is roles.Here I only process the data of the 10 players *totally*:



It’s easy to find that DUO\_SUPPORT is the most popular one.I used the data of the 5 players in team1 and explored it as well:

result:

win 36704

loss 39576

accuracy: 0.5188253801782905

DUO\_SUPPORT is also a useless feature.

**Part2 Result Prediction**

Firstly,I would explore the relationship between team1\_firstBlood to team1\_firstRiftHerald and the result.From the result,I discovered that team1\_firstInhibitor has the strongest relationship with team1\_win.I computed the accuracy and entropy of every classifier,and found that team1\_firstInhibitor has the lowest entropy and the highest accuracy.Here I list the result:

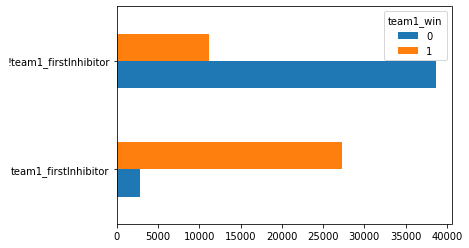
team1\_win 0 1 sum

team1\_firstInhibitor 2849 27336 30185

!team1\_firstInhibitor 38626 11189 49815

weighted entropy= 0.6486788709672755

weighted accuracy= 0.824525



So it’s a nice classifier.

And the computed accuracy of each classifier is no less than 0.6,so in the further study,I may build a decision tree to predict the result.

Now it’s time to analyze numeric data.At first,I processed kills,deaths and assists of each player in team1.I added up the numbers from player1 to player10,and found the best threshold of each item.The best classifier is deaths,which has an accuracy of 0.73835.In this case,the team is judged to loss if the total deaths reaches 22.In the database,34561 teams whose total deaths>=22 finally lost,while only 14018 teams won.

**Part 3 Relations Between Each Item**

Here I only explored the relationship among the boolean columns of team1.I wrote a 2-layered for loops,trying each combination of these features.I found that all of them have relationships with each other.The relationship between team1\_firstBaron and team1\_firstInhibitor is the strongest,which has a weighted entropy of 0.681 and an accuracy of 0.7674.

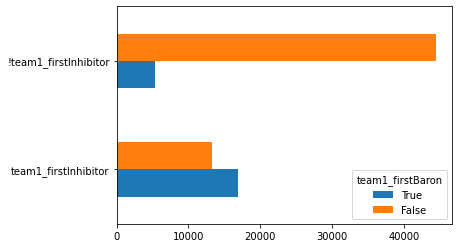
team1\_firstBaron True False sum

team1\_firstInhibitor 16968 13217 30185

!team1\_firstInhibitor 5391 44424 49815

weighted entropy= 0.6810330129849608

weighted accuracy= 0.7674000000000001



From the graph above,we can also know that once the first inhibitor is FALSE,the first baron is almost FALSE.However,if the first inhibitor is true,it won’t matter a lot.