Does Real Madrid Overperform in the UCL knockout stages

Harvard Stat 143 – Spring 2021

1 Introduction

Real Madrid has won the UEFA Champions League 4 times in the last 10 years. In three of these seasons, they didn't win the Spanish La Liga and they arguably weren't the top competitors for the UCL. However, a common thing to see in the last years is that Real Madrid would start the season poorly, lose against a lot of small teams, but in the big matches in the Champions League towards the end of the season, they would end up performing really well and win these matches and possibly the title.

Some fans described many of Real Madrid's wins in that era by luck. Take as an example Real Madrid's penalty at minute 90+7' against Juventus in the quarter finals of the 2018 season that lead them to the semi finals, or their equalizer goal against At. Madrid at minute 90+3' in the final of the 2014 season.

But to find out if these titles could be by luck or not, we want to do a formal study. To do so, we want to gather data about Real Madrid's observed and expected results in the knockout stages in the UCL and in other matches in the UCL and the Spanish League. Then, we want to compare these to find if there is a significant difference between them.

2 Data

2.1 Gathering the Data

We gathered data about the top 7 leagues in Europe: English, Spanish, Italian, German, French, Portuguese, and the Dutch League, and data about the UEFA Champions League and the UEFA Europa League, for the last 10 seasons (seasons from 2011 to 2020). Each of these data sets contain data about each match of the season. For this project, we only care about the two teams playing and the goals scored.

We got the data for the leagues from `https://www.football-data.co.uk/data.php`, the data for the UCL and UEL for the seasons from 2011 to 2017 from a private dataset (provided by professor Shaw), and from 2018 to 2020 from `https://fbref.com/en/`.

The reason for that is because the last website has the data as a text and in a format that needs cleaning, so we wanted to use other dataset for the data whenever possible.

While we might only need the Spanish League and UCL data for Real Madrid, we gathered all of this data to make sure we can predict the results for any team, so that if they play against Real Madrid, we would have a good expectation of the result of the game. It should be noted that all of the teams that Real Madrid faced in the UCL knockouts between 2016 and 2020 are in these top 7 leagues. This period will be the most important when we later do our hypothesis testing.

2.2 Data Cleaning

Because the data came from different sources, we did some data cleaning to make sure it is all compaitable. For the data from *fbref*, the scores were recorded in the same column (such as "2-1" for a two one win for the home team), so we separated those. The team names also came with a prefix or suffix to indicate the country of the club (such as "Chelsea eng"), so we removed that too. The private dataset for european matches had all the UCL and UEL matches together, so we separated those and indicated whether they are in the knockout stages or the group stages.

In addition, some team name differs a little bit between datasets, such as using "Manchester United" or "Man Utd". We manually checked some of the names and fixed their names in all of the datasets. While we can't do this for all teams, we made sure all the teams that Real Madrid faced in the UCL knockouts between 2016 and 2020 have consistent names.

3 Methodology

3.1 ELO Ratings

After getting the data, we want to build a model for predicting match results. To do so, we build an Elo system based on the World Football Elo Ratings [1]. In the Elo system, each team has a rating, and the difference between two teams' ratings is an indicator for the expected result of the match between them. After each game, the ratings of the two teams get updated according to the result of the game to better reflect the strengths of the teams.

For our system, all the teams start with an initial rating of 1500. Then, when two teams play, the ratings of each team get updated by the rule:

$$R_n = R_0 + K(W - W_e)$$

Where:

- R_n is the new rating, R_0 is the old rating.
- K is the weight of the competition. It is 50 for European competitions, and 30 for local leagues. K is then adjusted for the goal difference in the game. It is increased by a factor of half if a game is won by two goals, by 3/4 if a game is won by three goals, and by $3/4 + (N - 3)/8$ if the game is won by four or more goals, where N is the goal difference.
- W is the result of the match, 1 for a win, 1/2 for a draw, and 0 for a loss.
- W_e is the expected result (win expectancy). It is calculated using the formula:

$$W_e = \frac{1}{1 + 10^{-dr/400}}$$

  Where dr is the difference in rating before the match starts, and we add 100 to it for the home team.

If we ever encounter a team for the first time, we give it a rating equal to the mean of all the teams that were relegated from that competition from all previous years. This means that we take the means of all the teams that are in our system for this competition but aren't in that year's competition. This happens when a team gets promoted to one of the top 7 leagues, or when a team from outside of the top 7 leagues qualify to the UCL or the UEL. We think this mean is a good estimation of the team's strength. Luckily, all the teams that Real Madrid faced in the UCL knockout stages between 2016 and 2020 are from the top 7 leagues.

According to the World Football Elo Ratings [1], the strengths of the teams converge to the true relative strengths after about 30 matches. For local leagues, this is equal to around one and a half or two seasons. However, we note that there are very few inter-league matches, namely, the UCL and UEL matches. Half of the teams participating in these two competitions end up only playing the group stage, which is 6 matches, and the rest can play up to a max of 13 matches if they reach the finals, though they typically play much less.

It should also be noted that the leagues differ significantly in the strengths of their teams. The 4th or 5th best teams in Leagues like the Spanish and English Leagues might be stronger than the top teams of the Portuguese and Dutch Leagues. So, inter-league matches are very important to calibrate the strengths of the teams in the different leagues.

So, to make sure the teams converge to their true relative ratings, we ran the system for the first 5 available seasons, from 2011 to 2015. So, we use all of these seasons just to train the model. At the end of these ten seasons, the top ten teams in Europe were, in order:

1. Barcelona
2. Bayern Munich
3. Real Madrid

4. Sevilla
5. Juventus
6. Porto
7. Chelsea
8. Ath Madrid
9. Paris SG
10. Arsenal

We note that in the 2014-2015 season, Barcelona won all of its three possble titles, and Bayern Munich, Real Madrid, and Juventus, were the other three sides of the UCL semi finals, so these showing up in the top 5 is reasonable. While fans wouldn't usually expect to see Sevilla and Porto in the top 10, we note that Porto made it to the top 8 in UCL that year and Sevilla won the UEL, which might be why they are showing up here.

Now that we got reasonable ratings for the teams, we start the recording part. We run the system for the next 5 seasons, 2016 to 2020. For these seasons, in addition to updating the ratings after every match, we also record the difference $W - W_e$ for every match Real Madrid plays (the difference between the observed and expected outcome). We record those in two sets, one for the UCL knockouts, and the other for all other matches.

Now we have these two sets, S_{ucl} and S_{other}. The mean of these two sets should reflect the change in Real Madrid's strength between 2015 and 2020. However, we saw that:

$$Mean(S_{ucl}) \approx 0.083$$

$$Mean(S_{other}) \approx -0.061$$

This already indicates that Real Madrid is performing better than expected in the UCL knockouts and worse in the other games. However, we can't yet tell if this small difference could be due to luck or not.

3.2 Hypothesis Testing: T-test

To check whether the differences between the sets S_{ucl} and S_{other} could be due to chance or not, we use Hypothesis testing. Assume both sets are coming from the same distribution (a null hypothesis), what is the probability of seeing such differences?

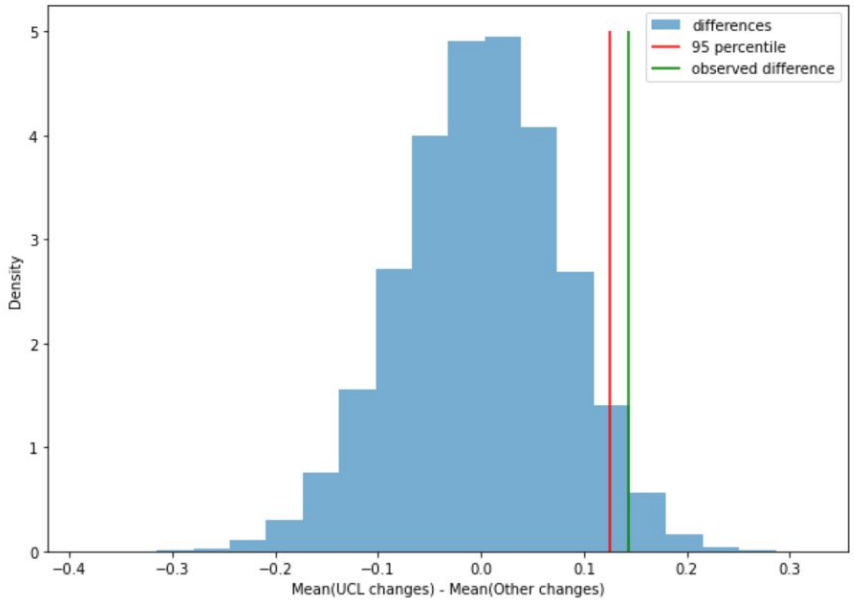
To calculate this probability, we first run a one sided t-test. In general, the t-test works by assuming that the mean of each of the sets will be normally distributed, and finds the probability that one of the means of the sets is in the confidence interval of the other mean. The reason we run a one-sided test is because if we end up rejecting the null hypothesis, we want to be able to conclude that Real Madrid is over-performing, not just that the two sets are not the same.

Running that, we get a value of $p = 0.0323$. This shows that there is only a 3.23% chance that the difference between the two sets happened by luck or chance.

3.3 Hypothesis Testing: Permutation tests

To ensure that our results are robust, we also run a permutation test. For this test, we combine the two sets into one set S. Then, for a big number of iterations, we randomly split the set into two sets, S_{ucl} and S_{other}, meaning that we randomly relabel the data. For each iteration, we record the difference between $Mean(S_{ucl})$ and $Mean(S_{other})$ and record that. This gives us a distribution of this value. Then, we can check for how many iterations was the generated difference equal or bigger than our observed difference. The fractions of these iterations gives us the probability that we observed a difference that big by chance or luck.

Figure 1: The distribution of the differences between the mean of the UCL matches and other matches for randomly relabeled data.



We ran this method for a million iteration. The distribution of the differences is shown in figure 1. We can see that the observed difference between the means (the green line) is higher than the 95th percentile (red line). Finding the number of iterations with equal or higher difference than observed gives a value of $p = 0.0282$. This means that there is only a 2.82% chance that the big observed difference is due to chance or luck.

4 Conclusion

When running the hypothesis tests, we get p-values of 0.0323 and 0.0282, both below the standard thershold of 0.05. This means that we should reject our hypothesis, which stated that Real Madrid performs equally in the UCL knockouts and in other matches. Because the tests we did are one sided, we can conclude that Real Madrid is over-performing in the UCL knockouts, and the results are not merely due to luck.

While this project affirms that Real Madrid weren't just lucky, it doesn't show any reason of this over-performance. It could be that the team and the coach prepare more and focus more on these matches, or it could be because their playing style doesn't work well against the small teams in the Spanish League, or for other reasons. This prompts possible next steps beyond this project.

5 Data Sources

[1] `https://www.football-data.co.uk/data.php`

[2] `https://fbref.com/en/`.

[3] Professor Laurie Shaw

6 References

[1] "About World Football Elo Ratings". World Football Elo Ratings. `http://eloratings.net/about`