Knowledge Management for UEFA Champions League

MIS 580: Knowledge Management

Harsha Gunnam
Hetal Mehta
Nargis Memon
Manish Wadhwa

Spring 2008

Table of Contents

Introduction

The **UEFA Champions League** is a seasonal club football competition organized by one of the Union of European Football Associations (UEFA) since 1955 for the most successful football clubs in Europe. The prize, the European Champion Clubs' Cup, is considered the most prestigious club trophy in the sport. The UEFA Champions League is separate from the less prestigious UEFA Cup and the defunct Cup Winners' Cup. The current holders of the UEFA Champions League trophy are AC Milan, who beat Liverpool FC 2-1 at the Olympic Stadium in Athens, Greece on 23 May 2007.

Game Format

Table 1: Game Format

UEFA Champions League Competition System	
1st qualifying round	**24**
2nd qualifying round	**16+12**
3rd qualifying round	**18+14**
Group stage	**16+16**
First knock-out round	**16**
Quarter finals	**8**
Semi-finals	**4**
Final	**2**

The tournament consists of several stages as shown in the above table and begins with three preliminary knockout qualifying rounds. Different teams start in different rounds, according to their position in domestic league and the UEFA coefficients of their league, while the sixteen top ranked teams spread across the biggest domestic leagues qualify directly. Eight winners and eight runners up teams join each other in first knock out round and thereafter half of the teams start getting eliminated with each round till two teams meet in the final.

Research Objectives

Using the concept of Data Mining in sports, we aim to generate knowledge from the raw data available for football matches. Specifically, we aim to identify the top three leagues for 2008-2009 on the basis of previous 5 seasons' data; Identify top 4 clubs for these 3 associations on the basis of previous years' data. We will also calculate the home and away advantages of these 12 clubs against every other opponent played in last five seasons and also calculate their head to head probability. Using these two findings, we will find a possible array of winners for 2008-2009.

Research Process

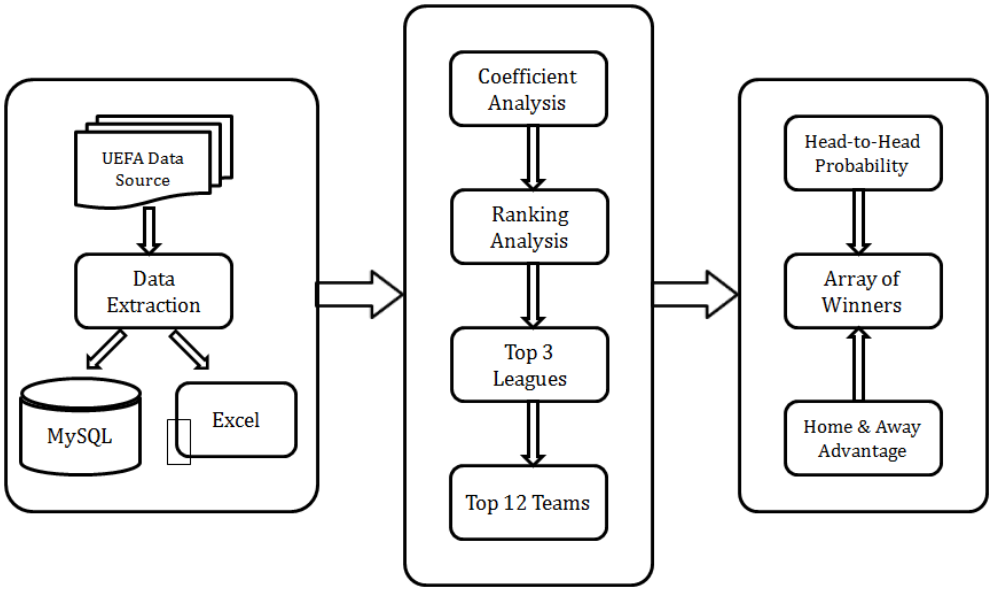


Figure 1: Research Process

We collected team statistics data into an excel sheet and organized that data. Then we created tables in MySQL and after filtering the data (removal of white spaces) in the excel sheet, we dumped it (excel data) into the MySQL tables. Then we wrote queries to perform coefficient analysis and ranking analysis on it. Then we found the top three leagues and the top twelve clubs of those teams. We then collected matches data into an excel sheet and calculated home and away advantages of these twelve clubs

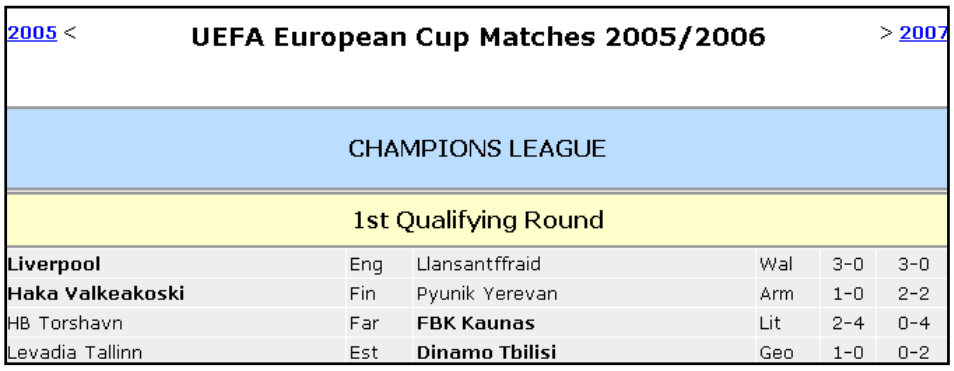
against every other club played in last six seasons and also calculated their head to head probability. These two results were then used to find array of winners for coming season.

Data Sources

The data source which provided a comprehensive dataset for UEFA Champions league is listed below and it is the source we used throughout the project:

http://www.xs4all.nl/~kassiesa/bert/uefa/

This website has matches data and team statistics from 1955 to current season. Following is the sample screen print of the matches' data available on this site:



2005 < **UEFA European Cup Matches 2005/2006** > 2007

CHAMPIONS LEAGUE

1st Qualifying Round

Liverpool	Eng	Llansantffraid	Wal	3-0	3-0
Haka Valkeakoski	Fin	Pyunik Yerevan	Arm	1-0	2-2
HB Torshavn	Far	**FBK Kaunas**	Lit	2-4	0-4
Levadia Tallinn	Est	**Dinamo Tbilisi**	Geo	1-0	0-2

Figure 2: Sample of Data

We collected the data for six seasons starting 2002 to 2008 for our research. The following table shows the detail of the data that we collected for our project:

Table 2: Collected Data Attributes

Table Name	Team Statistics	Team Country	Team Coefficient	Country Coefficient	Matches
Fields	**Team**	**Team**	**Team**	**Country**	**Year**
	Cup	**Country**	**Team Coefficient**	**Country Coefficient**	**Round**
	Qualifying Wins		**Year**	**Year**	**Team**
	Qualifying Draws				**Goals**
	Qualifying Losses				
	Number of Wins				
	Number of Draws				
	Number of Losses				
	Bonus				
	Points				
	Year				
Cardinality	1185	449	788	310	637

The above table shows the attributes of the data that we collected Team Statistics, Team Country, Coefficient and Matches data. The fields highlighted (bold) are the attributes that we used for our project. The cardinality of each of our table is also shown e.g. the number of rows in our matches table were 637 and number of rows in our team statistics table were 1185.

Literature Review

Very little research is done in the field of soccer, and Champions league in particular. We were able to find only two papers which analyzed the team performance in European soccer. These papers and their brief outline are listed below:

- Papahristodoulou, Christos, "Team Performance in UEFA Champions League 2005-06." Munich Personal RePEc Archive (2006) Unpublished, Paper #138

  This study uses a multi-output multi-input Data Efficiency Analysis (DEA) to estimate the performance of all thirty-two participated football teams in the UEFA Champions League (CL) tournament 2005-06. The estimates are based on official match statistics from all 125 matches.

- Barros, Carlos Pestana, Leach, Stephanie, “Performance evaluation of the English Premier Football League with data envelopment analysis.” Applied Economics Vol. 38 No. 12 (2006): 1449-1458

  This paper uses data envelopment analysis (DEA) to evaluate the performance of English Premier League football clubs from 1998/99 to 2002/03 combining sport and financial variables. The paper evaluates how close the clubs are relative to the frontier of best practices, analyzing how they manage sport as well as financial results.

Basic Findings

In our initial analysis we found the top three leagues and top four teams from each of these 3 leagues. For this we used standard UEFA calculations of country and team coefficients.

Coefficient Analysis

UEFA country and team coefficients are calculated every year. We followed the same method to calculate the country and team rankings.

A) Country Coefficients

UEFA country coefficients are determined by the aggregated number of points of the teams under that country, divided by the number of participating teams of that country.

cteams = number of teams per country
tpoints = number of points per teams
cpoints = sum(team points) per country
country coefficient =cpoints/cteams

The table below shows a sample of the country coefficient calculations.

Table 3: Country Coefficient Calculations

Country	Year	Number of Teams	Total Points	Country_Coefficient
England	2005	7	109.0000	15.57142857
England	2006	7	101.0000	14.42857143
England	2007	8	133.0000	16.62500000
England	2008	8	102.0000	12.75000000
Estonia	2004	3	1.0000	0.33333333

B) Team Coefficients

Team coefficient is calculated using the formula:

Team coefficient= tpoints+33% of country coefficient

The points scored by that team are based on number of wins, losses and draws it has during that season. The table below shows sample team coefficient calculations.

Table 4: Team Coefficient Calculations

ct_country	ct_team	year	ts_points	Team_Coef
England	Chelsea	2006	13	17.76142857
England	Chelsea	2007	23	28.48625000
England	Chelsea	2008	14	18.20750000
England	Everton	2006	2	6.76142857
England	Everton	2008	15	19.20750000

Ranking Analysis

Country ranking is calculated by summing five years of country coefficients and team ranking is calculated by summing five years of team coefficients. Using the country and team coefficients from 2003-2008, we got the following rankings for the coming season of 2008-2009 as shown below.

Table 5: Country Rankings, Team Rankings-2009

Country	Country_Ranking_2009
Spain	72.41964286
England	70.62500000
Italy	58.91071429
France	52.38392857
Germany	46.29464286
Russia	40.75000000
Romania	40.60000000
Portugal	38.92857143
Netherlands	37.54761905
Turkey	30.97500000

ct_country	ct_team	Team_Ranking_2009
Italy	AC Milan	121.44053571
England	Liverpool	118.80625000
England	Chelsea	114.30625000
England	Arsenal	109.30625000
Spain	FC Barcelona	107.89848214
Italy	Internazionale	99.44053571
Spain	Sevilla	97.37276786
England	Manchester United	97.30625000
Spain	Real Madrid	94.89848214
Spain	Villarreal	85.62848214
Italy	Juventus	80.73633929
Italy	AS Roma	78.44053571

By using these results, the top three leagues for the coming season of 2008-2009 were identified to be Spain, England and Italy and the top four teams for each of these three leagues were identified using the team rankings and are shown in the above table.

Country Ranking: 2003-2008

For the purpose of analyzing the ranking of the leagues and teams, we calculated the rankings of countries for the seasons of 2003-2008 and also plotted a graph for the top 10 teams which is shown below.

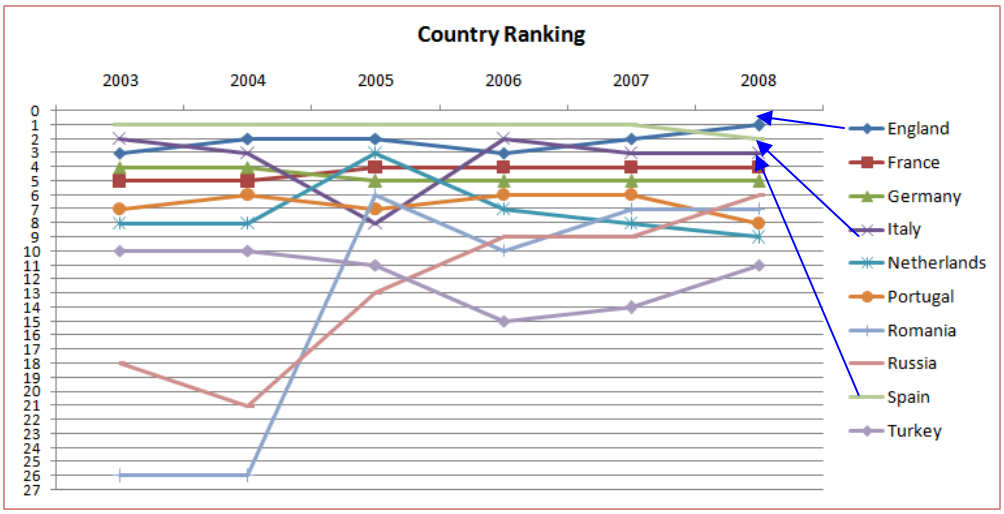


Figure 3: Country Rankings

One key observation in this analysis was that Spain, England and Italy have been the top three leagues over the past six seasons.

Team Ranking: 2003-2008

We also calculated the team rankings of top 12 teams over the seasons of 2003-2008 based on team coefficients and plotted a graph which is shown below.

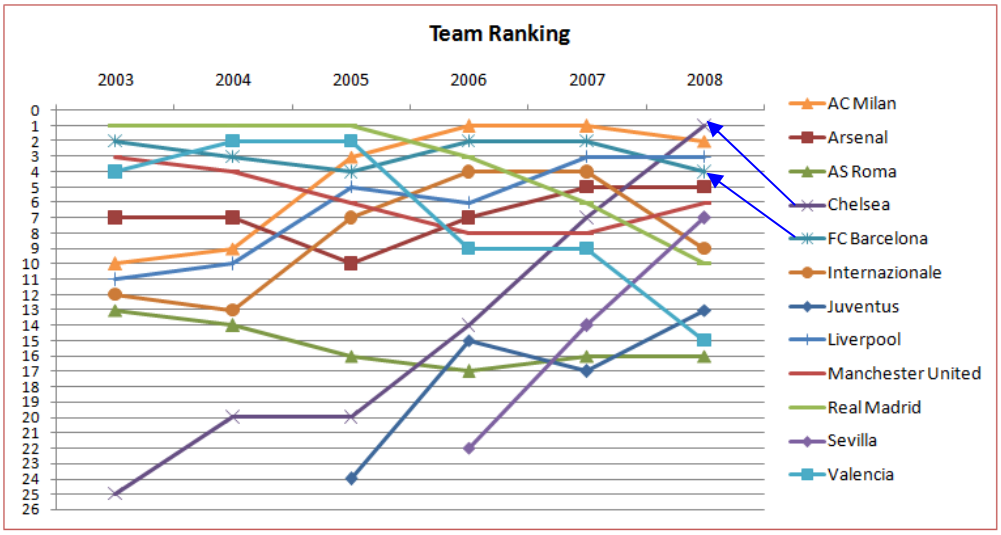


Figure 4: Team Rankings

Some key observations that have been made during this analysis are that FC Barcelona has been consistent between ranks two and four and that Chelsea has shown as rapid improvement from rank 25 to being among the top two teams.

From all these basic research and analysis, we were able to identify the top three leagues for the coming season of 2008-2009 which are Spain, England and Italy. Then we also identified the top four teams of each of these top three leagues. They are shown in the following table.

Table 6: Top Leagues & Teams

Leagues	Teams
Spain	FC Barcelona
	Real Madrid
	Sevilla
	Valencia
England	Arsenal
	Chelsea
	Liverpool
	Manchester United
Italy	AC Milan
	AS Roma
	Internazionale
	Juventus

Interesting Findings

Once we got the top 12 teams through initial analysis, we proceeded with our interesting analysis of finding head-to-head probability which provided us with the odds of each of those top 12 teams winning against others for all the matches played between 2003 and 2007.

Head-To-Head Probability

To come up with the head-to-head probability, we firstly gathered the data for 5 different seasons i.e. from 2003-2007. We then sorted this dataset to get the chunk of data relevant to top 12 teams as identified in our initial analysis. We then grouped this data based on the 2 teams that play in each of

those matches. We used the Naïve Bayes theorem to find out the conditional probability of one team winning against another. We simplified the formula as:

Probability of team A winning against team B = ((No. of wins against B * 2) + (No. of draws against B * 1))/ (Total matches between A & B) * 2

To find out the losing probability between the teams, we used the simple fact that sum of probability of all the results is 1. Thus,

Losing Probability of team A against team B = 1 – Winning probability of team A against team B

We repeated these procedures for each of the other 11 teams in the pool of top 12 teams. Most of this work was done with the help of MS excel sheets. The use of excel sheets eased the process of doing the trend analysis by availing wide range of graphs which simplified the process of analysis. We used 2-Dimensional column graphs to locate the trends in the dataset. The figure below shows the head-to-head probability of winning and losing for Arsenal against all the teams that they played against in UEFA Champions League during 2003-07. The trend found out from this analysis was that Arsenal is strong against 6 teams (indicated by the tall green bars reaching the probability level of 1) and weak against FC Barcelona (indicated by the tall purple bar reaching the probability level of 1). Similar trend analysis was done for other 11 teams and the results were factored into final conclusion.

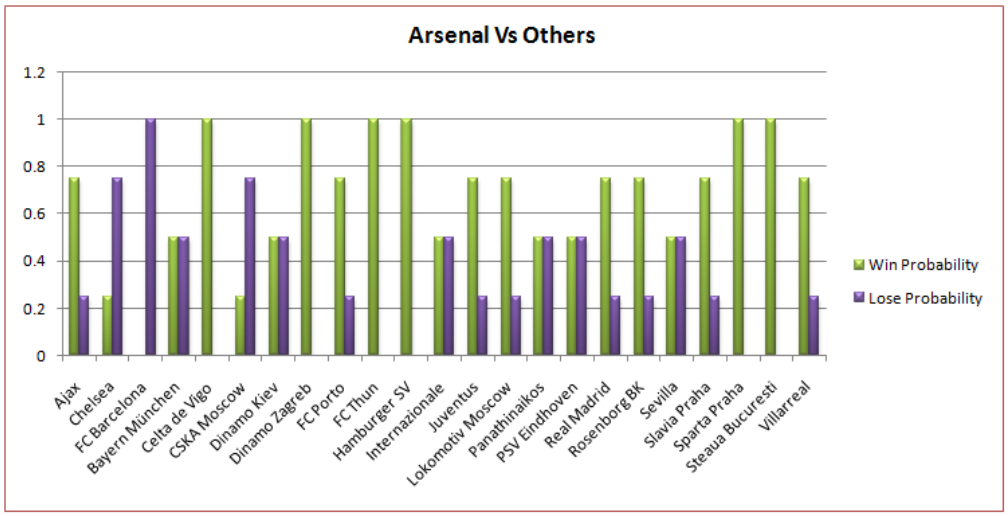


Figure 5: Probability Analysis of Arsenal

Testing of Head-To-Head Probability:

Interesting findings were done mainly to assist in the final objective of finding array of 4 winners for the next season. Thus every time some results were drawn from the analysis, those were testified to ascertain that the team was heading in the correct direction. Hence we tested the head-to-head probability results from dataset 2003-07 on the dataset available for 2008. We followed the exact same method for finding the probability on dataset for 2008 as the one used for 2003-07. We then,

1. Extracted the match records where we found two same teams for dataset 2003-07 and 2008
2. Took the difference of winning probability and losing probability for each record set
3. Compared the signs of the resultant for both the datasets:
    a. If the signs matched, that meant the prediction was correct, and
    b. If the signs differed, that meant the prediction was incorrect.

We made one of the key assumptions in this process. If the resultant (win probability – lose probability) results in zero, the next match can favor any of the two teams. Using this methodology, we tested the accuracy of 80% for the analysis that was done for head-to-head probability.

Home-Away Analysis

Once we finished head-to-head probability, team brainstormed to determine other elements that can impact the winning of a team in the UEFA Champions League. After much of the consideration, it was evident that the ground on which the match is being played is also one of the key drivers for winning and losing. We thus did the home-away analysis for each of the top 12 teams to find how each team is fairing on home grounds and away grounds. For this analysis, we again used the 5 season's data from 2003 to 2007. We then pulled out the matches' records where at least one of the top 12 teams had played. We used scaled approach for this kind of analysis. We defined 4 different scales as:

1. Very Strong – Team winning with a difference of 2 or more goals
2. Strong – Team winning with a difference of 1 goal

3. Weak – Team drawing or losing with a difference of 1 goal
4. Very Weak – Team losing with a difference of 2 or more goals

Since the analysis was based on home ground and away ground, we separated the dataset into home data and away data. We then counted the matches in each of the above 4 categories for top 12 teams on home grounds. The resultant graph was as shown below.

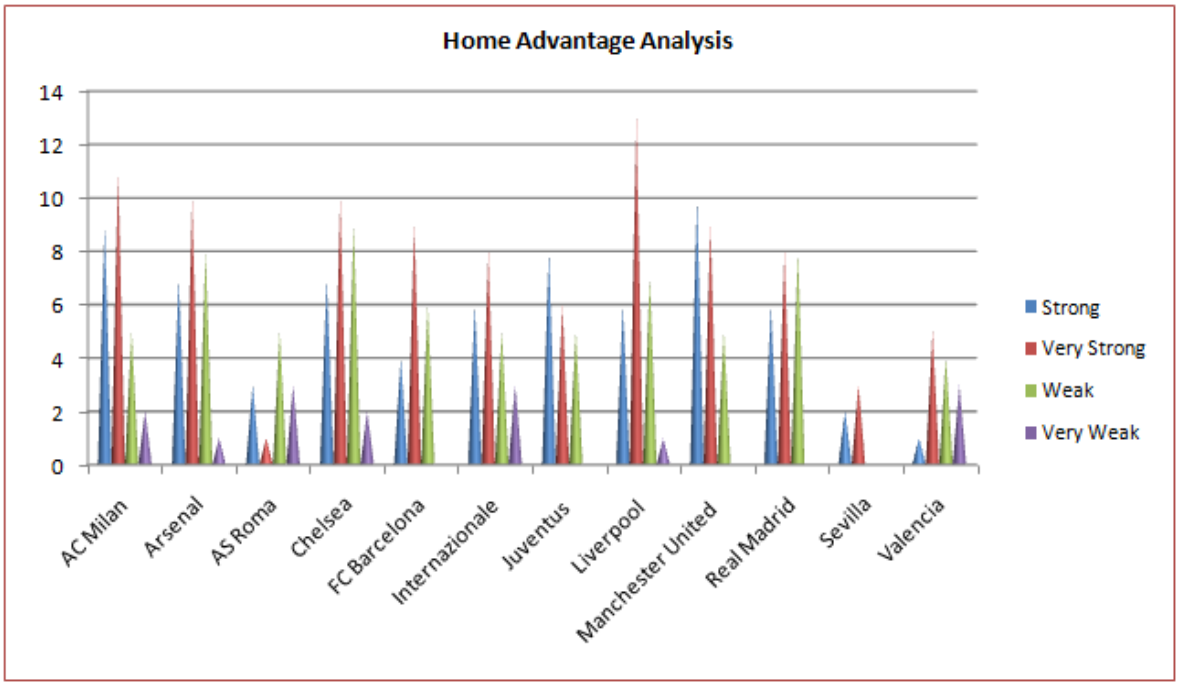


Figure 6: Home Advantage Analysis

We followed the same process for doing the analysis for the matches played on the away grounds and the resultant graph was as shown below.

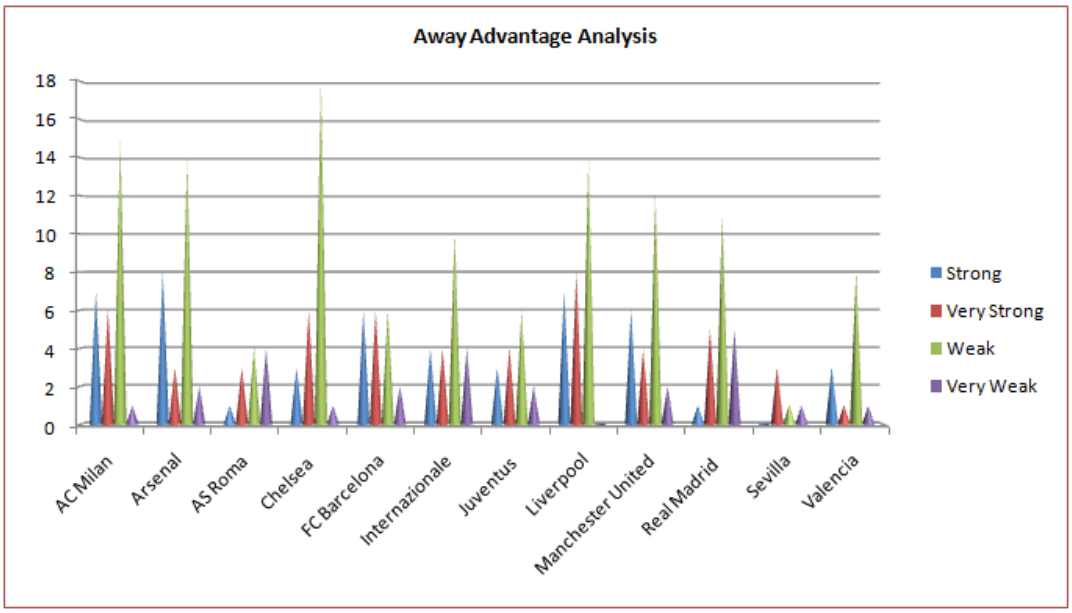


Figure 7: Away Advantage Analysis

The close examination of these two graphs helped us with following intermediary conclusions:

- Strongest Home Team: AC Milan
- Weakest Home Team: Real Madrid
- Strongest Visiting Team: Liverpool
- Weakest Visiting Team: Chelsea

At this stage, it became crucial for the team to decide whether to consider more factors that affect the final results or to move ahead with these two factors in hand. After much of the consideration, we realized that these two are the major elements that impact victory or loss of a team and decided to continue with our final prediction with these results in hand.

Winners Analysis

After doing the head-to-head analysis and the home-away analysis, we proceeded forward with the winners' analysis. The objective of this analysis is to predict an array of winners for the coming season of 2008-2009. For this analysis, we took the data set for the seasons of 2003-2008.

For this analysis, we combined the results of the head-to-head analysis and the home-away analysis. Then the probabilities of a team winning at home and away are calculated. Then the probabilities at home and away are aggregated and the top four teams are identified as the array of winners for the coming season of 2008-2009. The procedure that was followed for doing this analysis is as follows.

After defining different levels of strengths to the level of performance of teams in a match, we assigned values to these strength levels. The values assigned to the strength levels are shown in the following table.

Table 7: Strength Levels

Strength Level	Value
Very Strong	1.00
Strong	0.75
Weak	0.50
Very Weak	0.25

After assigning values, we aggregated the values obtained against the same two teams and calculated the average. For example, if AC Milan and Manchester United played between themselves on two different times and got the values of 0.5 and 0.75, both the values are added and the average is calculated to be 0.625.

After the final values are calculated, these values are multiplied with the probability of a team winning or losing in the coming season of 2008-2009 and the final value of probability for a team winning or losing is calculated for the coming season of 2008-2009.

Then, for each particular team, all their probabilities for the whole data set are added and their average is calculated to find their probability of the team to be qualified for the array of winners. Hence the final formula we used for this analysis is:

Team Win Probability = (Strength Value*Probability)/ Number of Matches Played

So the results obtained for home and away win probabilities for the top 12 teams for the season of 2008-2009 are shown in the following tables.

Table 8: Home Win Probability-2009

Home Win Probability	
Team	**Probability**
FC Barcelona	0.7115625
Manchester United	0.61601563
Liverpool	0.60256579
Real Madrid	0.58089286
Valencia	0.578125
Internazionale	0.56819444
Arsenal	0.52309783
AC Milan	0.50989318
Chelsea	0.5002425
Juventus	0.45089286
Sevilla	0.38125
AS Roma	0.29888889

Table 9: Away Win Probability-2009

Away Win Probability	
Team	**Probability**
FC Barcelona	0.62822917
Liverpool	0.52776316
Manchester United	0.50664063
Real Madrid	0.48178571
Chelsea	0.4796875
AC Milan	0.47955909
Valencia	0.4609375
Juventus	0.45089286
Internazionale	0.445
Arsenal	0.42663043
Sevilla	0.42663043
AS Roma	0.295511111

Hence by looking at these two tables, the top four teams that could be identified into the array of winners are FC Barcelona, Liverpool, Manchester United and Real Madrid.

For the purpose of adding value and strength to our prediction, we tested these results by taking the data set of the seasons of 2002-2007 and identifying the array of winners for the season of 2007-2008 using the same technique. The results obtained are shown in the following tables.

Table 10: Home Win Probability-2008

Home Win Probability	
Team	**Probability**
FC Barcelona	0.6476647
Manchester United	0.58220109
Valencia	0.57211538
Liverpool	0.52321023
Real Madrid	0.5140625
Juventus	0.51171875
AC Milan	0.503115
Internazionale	0.4919325
Arsenal	0.4880475
Chelsea	0.4759375
AS Roma	0.21875

Table 11: Away Win Probability-2008

Away Win Probability	
Team	**Probability**
FC Barcelona	0.58799342
Liverpool	0.48579545
Chelsea	0.48263889
Manchester United	0.46535326
AC Milan	0.45249
Valencia	0.44711538
Real Madrid	0.44375
Juventus	0.4390625
Internazionale	0.43638205
Arsenal	0.43074604
AS Roma	0.27083333

The teams that entered the semi-finals of this year's competition are FC Barcelona, Liverpool, Manchester United and Chelsea. The results obtained for the away win probability that we calculated are the same teams that entered the semi-finals of this year. Hence, we achieved an accuracy rate of 100%. Whereas for the home win probability, three teams among the top four teams identified are the ones that entered the semi-finals. Hence, our accuracy rate is 90%. After doing all the analysis, we were able to identify the final array of winners for the coming season of 2008-2009 and they are shown in the following table.

Table 12: Array of Winners-2009

Array of Winners
FC Barcelona
Liverpool
Chelsea
Manchester United

Conclusion

After all the observations and analysis, one key observation is that the teams that perform better when they are visiting have a better chance of winning the tournament due to the importance of away goals and also due to the fact that the number of away matches are greater than the home matches.

The most important benefit for a study like this is to the transfer the tacit knowledge to explicit knowledge. There is a lot of tacit knowledge in the field of sports and a study like this could be used to make this knowledge explicit. It was also surprising to us that there has been very little research and very few papers published in this domain. The field of sports is an area with a very wide scope for research and is also very useful for many sporting teams and betting organizations. The sport of soccer in particular offers a very wide scope for data mining and knowledge management.

References

- Papahristodoulou, Christos, "Team Performance in UEFA Champions League 2005-06." Munich Personal RePEc Archive (2006) Unpublished, Paper #138

- Barros, Carlos Pestana, Leach, Stephanie, “Performance evaluation of the English Premier Football League with data envelopment analysis.” Applied Economics Vol. 38 No. 12 (2006): 1449-1458

- Websites:
  - http://www.betinf.com/champ.htm
  - http://www.betstudy.com/soccer-stats/c/europe/uefa-champions-league/
  - http://en.uclpredictor.uefa.com/
  - http://www.uefa.com/competitions/ucl/index.html
  - http://en.wikipedia.org/wiki/Uefa_Champions_League
  - http://en.wikipedia.org/wiki/Bayes_theorem
  - http://www.xs4all.nl/~kassiesa/bert/uefa/
  - http://www.soccerbase.com/
  - http://europeancups.altervista.org/