Forecasting the winner of the UEFA Champions League 2008/09

Christoph Leitner*, Achim Zeileis**, Kurt Hornik***

∗ Department of Statistics and Mathematics, WU Wirtschaftsuniversität Wien, Augasse 2–6, 1090 Wien, Austria, Christoph.Leitner@wu-wien.ac.at
∗∗ Department of Statistics and Mathematics, WU Wirtschaftsuniversität Wien, Augasse 2–6, 1090 Wien, Austria, Achim.Zeileis@wu-wien.ac.at
∗∗∗ Department of Statistics and Mathematics, WU Wirtschaftsuniversität Wien, Augasse 2–6, 1090 Wien, Austria, Kurt.Hornik@wu-wien.ac.at

Abstract. The UEFA Champions League is the most prestigious football club competition in Europe and hence there is major interest, among fans and experts alike, in forecasting the winner of this tournament. To investigate this issue, a class of linear mixed-effects models for quoted winning odds from various bookmakers is explored. Based on this "prospective" data reflecting the expectations of the bookmakers (as opposed to past performances used in many other forecasting methods) different models for the "true" odds of winning the tournament can be established, capturing both team-specific effects (along with effects for the team's tournament group and national association) and bookmaker-specific variations. A selection among various model specifications yields a model with a fixed team effect plus a random bookmaker-specific deviation. It forecasts team FC BATE Borisov with a probability of 0.10% as the winner of the tournament; the runner-up is Anaethosis Famagusta FC with a winning probability of 0.11%. In addition to the forecast of the winning probability, information about the groups of the preliminaries and the European football associations can be obtained from the model.

1. Introduction

The *UEFA Champions League* is the most prestigious club competition of the Union of European Football Associations (UEFA) and so one of the most popular annual sports tournaments all over the world. It was originally created as the European Champion Clubs' Cup for the 1955/56 season, before its format and name were changed in 1992 (Union of European Football Associations, 2009).

Millions of football supporters in Europe and throughout the world are interested in the games and the title winner. Here, we introduce a general mixed-effects model framework that extends the ideas of Leitner *et al.* (2008) for forecasting the winner of such a tournament. Unlike many other sports prediction methods it is not based on historical data (see e.g., Dyte and Clarke, 2000; Goddard and Asimakopoulos, 2004) but designed for bookmakers odds for winning the UEFA Champions League, i.e., reflecting current expectations. The motivation for using bookmakers odds is that (a) they incorporate expectations about a specific tournament, (b) bookmakers have financial incentives to rate teams correctly, (c) other empirical studies have shown that odds provide an efficient forecasting instrument for the outcomes of single games (see e.g., Forrest *et al.*, 2005; Dixon and Pope, 2004). Based on these ideas, Leitner *et al.* (2008) use quoted odds to forecast the outcome of a whole tournament, the EURO 2008. Their study performed successfully, in particular predicting the final correctly. Here, we adapt this method into a more general model framework for the log-odds capturing different effects associated with the participants, the bookmakers, the groups of the preliminaries and the team's associations leading to a variety of mixed-effects models. After establishing the general modeling approach a subsequent model selection yields the final model upon which our forecasts for the tournament are based. For this study we use the quoted long-term odds for winning the UEFA Champions League 2008/09 of 31 international bookmakers which were published online after the group draw but before the tournament started (accessed on 2008-09-01 from the bookmakers' websites). The 31 bookmakers are all bookmakers who offer odds for this event out of 50 European top-selling online sports bookmakers.

The UEFA Champions League is an annual tournament where a selection of European teams compete in a multi-stage modus (qualification, group and knockout stage) to determine the best team. First, 32 teams are determined via three qualification rounds for the group stage and drawn into eight groups (A–H). The number of eligible teams is determined by UEFA's Coefficient Ranking System for its member associations (for more details see Union of European Football Associations, 2009). In the 2008/09 season, teams from 17 associations

This is a preprint of an article published in R Koning and P Scarf (eds.), *Proceedings of the 2nd International Conference on Mathematics in Sport – IMA Sport 2009*, 94–99.

out of UEFA's 53 members qualified for the group stage. The four teams of a group play a round-robin—every team plays every other team twice (one home and one away match), for a total of twelve games within the group—and the group winners and runners-up advance to the knockout stages. In the knock-out stage, each round pairings are determined by means of a draw and played under the cup (knock-out) system, on a home-and-away basis, where the winners conntest to the next round until two teams remain. The two teams play the final as one single match at a neutral venue yielding the winner of the UEFA Champions League (Union of European Football Associations, 2009).

Using the quoted long-term odds for winning the UEFA Champions League 2008/09 of all 32 participating teams from 31 international bookmakers, our approach predicts FC BATE Borisov as the winner of the tournament with probability 0.10%; the second best team is Anaethosis Famagusta FC with a winning probability of 0.11%.

The paper is organized into four sections: Section 2 gives a description of our method which is applied in Section 3. Section 4 concludes the paper with a brief discussion.

2. Method

Our general model class assumes a relationship between the quoted odds and the unknown "true" odds for winning the tournament. To estimate these latent true odds, we first adjust the quoted odds $\widetilde{odds}_{i,j}$ of bookmaker j for team i into the adjusted odds $odds_{i,j}$ (reflecting the underlying beliefs of bookmaker j for team i) by subtracting one, the stake, and removing the over-round (Section 2.1). Subsequently, in Section 2.2, we discuss various approaches for capturing the dependence of the true odds $odds_i$ on the team (as well as its group or association) in a mixed-effects model (e.g., Pinheiro and Bates, 2000), which additionally allows for bookmaker-specific deviations of the adjusted odds from the true odds. An additional model selection step then yields the final model: its estimates $\widehat{odds}_i$ for the latent true odds will be analyzed in detail in Section 3.

2.1 Pre-processing

The quoted odds of the bookmakers do not represent the true chances that a team will win the tournament, because they include the stake and a profit margin, better known as the "over-round" on the "book" (for further details see e.g., Wikipedia, 2009). To recover the underlying beliefs of the bookmakers, we have to adjust the quoted odds. We first reduce the quoted odds by one, the stake, to get the profit for the successful punter and then we adjust it by the profit of the bookmaker (the over-round). We assume that the over-round is constant for each bookmaker across all teams, i.e.,

$$\widetilde{odds}_{i,j} - 1 = \delta_j \, odds_{i,j}, \tag{1}$$

where $\widetilde{odds}_{i,j}$ are the quoted odds, $odds_{i,j}$ are the adjusted odds of bookmaker j for team i, and δ_j is the proportion that bookmaker j pays in case of a win (i.e., the reciprocal value of the over-round) in addition to the stake. These odds can be transformed to a probability scale via

$$p_{i,j} = 1 - \frac{odds_{i,j}}{1 + odds_{i,j}}, \tag{2}$$

and then the pay-out proportion δ_j for each bookmaker j can be computed by constraining the sum of all probabilities per bookmakers to be one: $\sum_i p_{i,j} = 1$ (for all bookmakers j). Note that we have to use complementary probabilities in Equation 2 to obtain the winning probability $p_{i,j}$ of bookmaker j for team i.

2.2 Modeling

Using the adjusted winning odds $odds_{ij}$ for each of the $i = 1,\ldots,32$ teams of the $j = 1,\ldots,31$ international bookmakers, we propose a stochastic model capturing the underlying odds distribution on a log-scale. The model relates the adjusted log-odds $\log(odds_{ij})$ to the (unobservable) "true" log-odds $\log(odds_i)$ plus an additional (unobservable) error term $\varepsilon_{i,j}$ of bookmaker j for team i. To capture these latent quantities by a linear mixed-effects model, we allow the true log-odds to depend on a team effect α_i, a tournament group effect $\beta_{g(i)}$ for group g of team i, an association effect $\gamma_{a(i)}$ for association a of team i, as well as an overall intercept ν. The error can additionally depend on μ_j, the effect of bookmaker j. In summary, this can be written as

$$\log(odds_{ij}) \quad = \quad \log(odds_i) + \varepsilon_{ij} \tag{3}$$

Table 1. Mixed-effects models for $\log(odds_{ij})$ of team i by bookmaker j with different fixed and random effects, where ν is the intercept, μ_j is the effect of bookmaker j, α_i is the effect of team i, $\beta_{g(i)}$ is the effect of group g of team i, $\gamma_{a(i)}$ is the effect of association a of team i, and Z_{ij} is a standardized error. Each model is evaluated by the log-likelihood value (logLik), the number of estimated parameters (df), and the BIC.

	Model	Fixed effects	Random effects	logLik	df	BIC
1	$\nu + \sigma Z_{ij}$			-1796.65	2	3607.09
2	$\nu + \mu_j + \sigma Z_{ij}$	μ_j		-1793.11	32	3807.01
3	$\nu + \mu_j + \sigma Z_{ij}$		μ_j	-1796.68	3	3614.06
4	$\nu + \mu_j + \beta_{g(i)} + \sigma Z_{ij}$	$\beta_{g(i)}$	μ_j	-1759.04	10	3587.08
5	$\nu + \mu_j + \gamma_{a(i)} + \sigma Z_{ij}$	$\gamma_{a(i)}$	μ_j	-1655.39	19	3441.87
6	$\nu + \mu_j + \beta_{g(i)} + \gamma_{a(i)} + \sigma Z_{ij}$	$\beta_{g(i)}, \gamma_{a(i)}$	μ_j	-1338.66	26	2856.71
7	$\nu + \alpha_i + \sigma Z_{ij}$	α_i		-118.87	33	465.42
8	$\nu + \alpha_i + \mu_j + \sigma Z_{ij}$	α_i, μ_j		-2.51	63	439.70
9	$\nu + \alpha_i + \mu_j + \sigma Z_{ij}$	α_i	μ_j	-50.86	34	336.32
10	$\nu + \alpha_i + \sigma Z_{ij}$		α_i	-243.12	3	506.93
11	$\nu + \alpha_i + \mu_j + \sigma Z_{ij}$	μ_j	α_i	-130.51	33	488.72
12	$\nu + \alpha_i + \mu_j + \beta_{g(i)} + \sigma Z_{ij}$	$\mu_j, \beta_{g(i)}$	α_i	-129.26	40	534.50
13	$\nu + \alpha_i + \mu_j + \gamma_{a(i)} + \sigma Z_{ij}$	$\mu_j, \gamma_{a(i)}$	α_i	-125.77	49	589.63
14	$\nu + \alpha_i + \mu_j + \beta_{g(i)} + \gamma_{a(i)} + \sigma Z_{ij}$	$\mu_j, \beta_{g(i)}, \gamma_{a(i)}$	α_i	-114.86	56	616.11
15	$\nu + \alpha_i + \mu_j + \sigma Z_{ij}$		α_i, μ_j	-1796.65	4	3620.91

$$= \quad \nu + \alpha_i + \beta_{g(i)} + \gamma_{a(i)} + \mu_j + \sigma Z_{ij} \tag{4}$$

where Z_{ij} is a standardized error and σ is the standard deviation. Even if contrasts are employed, this model is overspecified when all four effects α_i, $\beta_{g(i)}$, $\gamma_{a(i)}$, and μ_j are included as fixed effects due to the the dependence of group $g(i)$ and association $a(i)$ on the team i. To overcome this methodological issue, there are various conceivable solutions which can also be motivated by subject-matter considerations: (a) The team effect could be omitted reflecting the assumption that it can be captured by group and/or association differences. (b) Conversely, the group and/or association effects could be omitted signalling that all teams are sufficiently different. Note that the full team effect then still captures group and association differences. (c) Alternatively, α_i could be specified as a random effect (with zero mean) conveying that the log-odds for each team deviate randomly from the mean as captured by the remaining effects. (d) Finally, a random effect for the bookmakers would be conceivable implying that the bookmakers' odds deviate randomly from the mean as captured by the remaining effects. Combinations of these ideas lead to 15 different mixed-effects models which are shown in the first column of Table 1. To find a parsimonious model which still gives a good approximation of the underlying data-generating process, standard model selection methods can be employed. Below, we use the Bayesian Information Criterion (BIC).

3. Results

Based on the modeling approach discussed above, we first choose the final mixed-effects model (Section 3.1) from which the associated probabilities $\widehat{p}_i$ for winning the UEFA Champions League 2008/09 for all teams are derived (Section 3.2). In addition some more information about the groups of the preliminaries (Section 3.3) and the European associations of the participating teams (Section 3.4) can be extracted.

3.1 Model selection

Fitting all 15 conceivable mixed-effects models discussed in the previous sections yields the results in Table 1 which provides the log-likelihood, number of parameters, and associated BIC.

The best model emerging from the BIC selection is Model 9 (BIC = 336.32), containing a fixed team effect (and hence no additional group or association effect) and a random bookmaker effect. Moreover, the three best models (7–9) all have a fixed team effect, followed by Models 10–14 which have a random team

effect and perform slightly worse. Finally, all models which have no team effect at all (or just try to capture it by group and/or association effects) perform clearly worse. In summary, this conveys that the bookmakers employ knowledge about each individual team when fixing their odds (rather than being mainly determined by group or association considerations). Furthermore, the fact that the bookmaker effect can be captured well as a random effect suggests that there are no large systematic deviations between the bookmakers. In retrospect, this model probably comes at no surprise because its interpretation is so intuitive: All teams are expected to perform differently and the bookmakers' expectations just vary randomly around some common overall latent log-odds $\log(\widehat{odds}_i) = \hat{\nu} + \hat{\alpha}_i$. It is reassuring that this intuitive model has been selected from a more general class of models, where some of the alternatives would have also had appealing interpretations.

3.2 Probability of winning the UEFA Champions League 2008/09

The estimated log-odds $\log(\widehat{odds}_i) = \hat{\nu} + \hat{\alpha}_i$ from Model 9 can easily be transformed to the associated estimated odds $\widehat{odds}_i$ and probabilities $\widehat{p}_i$ of winning the tournament for all participating teams (see Table 2). Additionally, the team code, the eight origin groups of the preliminaries, and the football association of the teams are shown. Chelsea FC is the best team of the 32 teams and has the highest probability (13.71%) of winning the tournament. The expected runner-up (if the tournament schedule allows such a final) comes also from England, Manchester United FC (winning probability: 12.13%). The top two are followed by the champion of the "Serie A" FC Internationale Milano (10.16%) and the champion of the "Primera Division" FC Barcelona (10.11%). The last four teams are participating for the first time in the tournament and have just a winning probability of 0.20% or less. Four teams out of the first seven ranked teams are from England which implies that England is the strongest European association. Three teams out of the first eleven are members of group H, but only two of them can advance to the next round. Using this informaiton in combination with the winning probabilities of the participating teams (Table 2) the following 16 teams (eight group-winners and eight runners-up) are expected to play the first knock-out round: CHL, ROM (group A), INT, BRM (B), BAR, SHA (C), LIV, ATL (D), MU, VIL (E), BAY, LYO (F), ARS, POR (G), RM, and JUV (H). These 16 teams are not the 16 participants with the highest winning probabilities implying that the group drawn has an effect to the tournament outcome. In this paper we focus on predicting the winner of the tournament. Nevertheless, if someone is interessted in the dynamic of the tournament we suggest to use a simulation approach (see Leitner *et al.*, 2008) to determine, e.g., the four teams playing the semi-finals.

3.3 Which is the strongest group of the preliminaries?

The forecast of the expected 16 teams qualifying for the first knock-out round implies a group effect. Although our model contains a fixed team effect and hence no group effect (redundant information), we can answer the question: "Which is the strongest group of the preliminaries?". The estimated team effects $\widehat{\alpha}_i$ imply a "group effect". To derive this group effect we calculate the difference in log-odds between the average team effect in group g and the overall mean ν of all 32 participating teams of the tournament. Table 3 shows these group effects for all eight groups (A–H). The group with the best chance to include the winner is group H (-0.30), followed by group D (-0.21). Despite the fact that group A includes the bookmakers' favorite of winning the Champions League (Chelsea FC), group A follows just on the third position (-0.18). Group G can be interpreted as the average group (0.02). The smallest chance to include the winner has group B (0.32).

3.4 Which is the strongest European association?

In addition to the group effect the estimated team effects $\widehat{\alpha}_i$ also imply a "association effect". We derive this association effect by computing the difference in log-odds between association a and the overall mean ν of all 32 participating teams of the tournament. This result can be used to rank the 17 associations of the participating teams and give an answer to the interesting question: "Which is the strongest European association?". Table 4 shows this association effects and the number of qualified teams for the 17 associations. There is a strong correlation between the estimated association effects on the log-odds scale and the number of qualified teams (-0.75). England, Spain and Italy have the maximum number of qualified teams (four), but England with the lowest association effect (-1.99) is the strongest European association. Russia with only one team (FC Zenit St. Petersburg) is rated better than Germany (two teams), France (three teams) and Portugal (two teams). Not

Table 2. Estimated log-odds $\log(\widehat{odds}_i)$, odds $\widehat{odds}_i$ and probabilities $\widehat{p}_i$ for all 32 teams.

	$\log(\widehat{odds}_i)$	$\widehat{odds}_i$	$\widehat{p}_i(\%)$	Code	Group	Association
FC BATE Borisov	6.89	977.75	0.10	BAT	H	Belarus (BEL)
Anaethosis Famagusta FC	6.84	936.19	0.11	ANO	B	Cyprus (CYP)
Aalborg BK	6.58	717.89	0.14	AAB	E	Denmark (DEN)
CFR 1907 Cluj	6.20	493.10	0.20	CLU	A	Romania (ROU)
FC Basel 1893	6.16	473.94	0.21	BSL	C	Switzerland (SUI)
FC Steaua Bucuresti	5.75	315.71	0.32	STE	F	Romania (ROU)
Celtic FC	5.33	207.42	0.48	CEL	E	Scotland (SCO)
FC Dynamo Kyiv	5.31	202.51	0.49	DYN	G	Ukraine (UKR)
Panathinaikos FC	5.17	176.09	0.56	PAN	B	Greece (GRE)
Sporting Clube de Portugal	5.16	173.40	0.57	SCP	C	Portugal (POR)
FC Shakhtar Donetsk	5.11	164.95	0.60	SHA	C	Ukraine (UKR)
FC Girondins de Bordeaux	5.09	162.11	0.61	BDX	A	France (FRA)
PSV Eindhoven	4.99	147.51	0.67	PSV	D	Netherlands (NED)
Fenerbahce SK	4.89	133.23	0.75	FEN	G	Turkey (TUR)
Olympique Marseille	4.81	123.05	0.81	MAR	D	France (FRA)
FC Porto	4.43	83.98	1.18	POR	G	Portugal (POR)
Werder Bremen	4.33	76.20	1.30	BRM	B	Germany (GER)
ACF Fiorentina	4.20	66.93	1.47	FIO	F	Italy (ITA)
Villarreal CF	3.93	50.69	1.93	VIL	E	Spain (ESP)
Club Atletico de Madrid	3.81	45.16	2.17	ATL	D	Spain (ESP)
Olympique Lyonnais	3.68	39.63	2.46	LYO	F	France (FRA)
FC Zenit St. Petersburg	3.67	39.28	2.48	ZNT	H	Russia (RUS)
AS Roma	3.38	29.50	3.28	ROM	A	Italy (ITA)
Juventus	3.22	25.02	3.84	JUV	H	Italy (ITA)
FC Bayern München	3.04	20.82	4.58	BAY	F	Germany (GER)
Liverpool FC	2.78	16.15	5.83	LIV	D	England (ENG)
Arsenal FC	2.68	14.65	6.39	ARS	G	England (ENG)
Real Madrid CF	2.26	9.59	9.44	RM	H	Spain (ESP)
FC Barcelona	2.19	8.89	10.11	BAR	C	Spain (ESP)
FC Internationale Milano	2.18	8.84	10.16	INT	B	Italy (ITA)
Manchester United FC	1.98	7.24	12.13	MU	E	England (ENG)
Chelsea FC	1.84	6.29	13.71	CHL	A	England (ENG)

Table 3. Group effects for the eight groups of the preliminaries A–H.

A	B	C	D	E	F	G	H
−0.181	0.322	0.343	−0.209	0.145	−0.141	0.020	−0.300

surprisingly, the team with the lowest probability to win the Champions League (FC BATE Borisov) comes from the weakest rated association (Belarus).

4. Discussion

This paper investigates a general model class for the "unknown" true log-odds for winning a sports tournament based on quoted bookmakers odds. The flexible model framework allows for capturing different effects (e.g., team, group, and association). The variety of possible linear mixed-effects models can be derived to a parsimonious model which still gives a good approximation of the underlying data-generating process by a

Table 4. Association effects and number of qualified teams for the 17 team's associations.

ENG	ESP	ITA	RUS	GER	FRA	POR	TUR	NED
−1.99	−1.26	−1.06	−0.64	−0.62	0.22	0.48	0.58	0.68
4	4	4	1	2	3	2	1	1
GRE	UKR	SCO	ROU	SUI	DEN	CYP	BEL	
0.86	0.90	1.03	1.67	1.85	2.27	2.53	2.58	
1	2	1	2	1	1	1	1	

standard model selection approach (BIC). This model is then used to forecast the outcome of a sports tournament. Here we apply this method to forecast the winner of the UEFA Champions League 2008/09. The model selection approach yields a model with a fixed team effect plus a random bookmaker-specific deviation forecasting Team Chelsea FC as the winner (winning probability: 13.71%). Furthermore, we give answers to the questions: "Which is the strongest group of the preliminaries?" (Answer: Group H) and "Which is the strongest European football association?" (Answer: England). Luckily for all football supporters, football is, like all other sports, a game and cannot be truly predicted using rational strategies and statistical methods.

Computational details

All computations were carried out in the R system (version 2.9.0) for statistical computing (R Development Core Team, 2008). In particular, the R package nlme version 3.1-91 (**?**) was used for the estimation of the mixed-effects models (see Pinheiro and Bates, 2000).

References

Dixon M. and Pope P. (2004) The value of statistical forecasts in the uk association football betting market. *International Journal of Forecasting* 20, 697–711.

Dyte D. and Clarke S. (2000) A rating based poisson model for world cup soccer. *The Journal of the Operational Research Society* 51, 993–998.

Forrest D., Goddard J. and Simmons R. (2005) Odds-setters as forecasters: The case of english football. *International Journal of Forecasting* 21, 551–564.

Goddard J. and Asimakopoulos I. (2004) Modelling football match results and the efficiency of fixed-odds betting. *Journal of Forecasting* 23, 51–66.

Leitner C., Zeileis A. and Hornik K. (2008) Who is going to win the EURO 2008? (A statistical investigation of bookmakers odds), Report 65, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, URL `http://epub.wu.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_d5f`.

Pinheiro J. and Bates D. (2000) *Mixed-Effects Models in S and S-PLUS*, Statistics and Comupting, Springer-Verlag New York, New York, USA, ISBN 0-387-98957-9.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

Union of European Football Associations (2009) UEFA Champions League, URL `http://en.euro2008.uefa.com/tournament/index.html`, [Online; accessed 2009-04-23].

Wikipedia (2009) Sports rating system — wikipedia, the free encyclopedia, URL `http://en.wikipedia.org/w/index.php?title=Sports_rating_system&oldid=290898733`, [Online; accessed 20-May-2009].