Bayesian Learning from Top European Professional Soccer Leagues about Attendance Influencers: An Examination at the Season Level

(Authors' names blinded for peer review)

We collected season-long performance data via data scraping from the ESPN FC website.

Key words: European Professional Soccer Leagues, Machine Learning, Frequentist Statistics, Bayesian Inference, Aggregated Attendance, Bayesian Network

---

Introduction

The popular frequentist statistical inference process starts with the formulation of an alternative research hypothesis(Ha), such as "people with higher income live happier than low income earners", which is typically set up against a null non-effect hypothesis (Ho), such as "income level has no effect on happiness". Then researchers collect relevant data (each subject's perceived happiness and income), and conduct a statistical significance test (t test) to see how likely such results would hold if chance (noise) alone were at work (testing against the null hypothesis). The illustrated example of the popular null-hypothesis significance test (NHST) will eventually compare the p value associated with our sample test statistic against the golden standard of 0.05 as the threshold for significance.

"p value is influenced both by effect size and by sample size"(Wagenmakers 2007, pp. 787) For this reason, sooner or later, you are guaranteed to get a significant result if you run subjects long enough and stop when you get the p value you want [Wagenmakers, 2007].

"facility to sample from the prior or posterior is a very informative feature of the Bayesian paradigm"Tipping (2004) "The Bayes factor pits one theory against anotherfor example, Theory1 against Theory2."(Dienes 2011, p. 277) "Typically, this means one should use a power calculation to plan in advance how many subjects to run. Running subjects until a significant result is obtained is forbidden, because this will always succeed, given sufficient time, even if the null is true"(Dienes 2011, p. 278)

"the goal of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with current data to produce a posterior probability distribution for the parameter that contains less uncertainty"(Lynch 2007, p. 50)

"The problem of knowing the sampling plan is even more prominent when NHST is applied to data that present themselves in the real world (e.g., court cases or economic and social phenomena), for which no experimenter was present to guide the data collection process."(Wagenmakers 2007, pp. 784)

"in the NHST framework, every null hypothesis that is not exactly true will eventually be rejected as the number of observations grows large. Much less appreciated is the fact that, even when a null hypothesis is exactly true, it can always be rejected, at any desired significance level that is greater than 0 (e.g., 5 .05 or 5 .00001). The method to achieve this is to calculate a p value after every new observation or set of observations comes in, and to stop the experiment as soon as the p value first drops below .(Wagenmakers 2007, pp. 784)

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.(Bell et al. 2009, Hey et al. 2009)

Context and Data

Soccer Ststistics

According to ESPN FC(www.espnfc.us), eight season-long performance metrics are used to characterize a professional soccer team's regular league season. Below, we define those statistics using the 2015/16 La Liga season of Real Madrid C.F. as an example.

- Most Home Goals (MHG) = maximum goals scored in a single match played at home. For the season 2015/2016, Real Madrids MHG is 10. They beat Rayo Vallecano by 10-2 at Santiago Bernabu Stadium on 12/20/2015.

- Most Away Goals (MAG) = maximum goals scored in a single away match. For the season 2015/2016, Real Madrids MAG is 6. They defeated Espanyol 6-0 on 9/12/2015 at RCDE stadium.

- Largest Margin of Victory (LMV) = the largest difference betweem the number of goals scored and the number of goals surrendered by the winning team in a single Liga regular season match. Real Madrid achieved a LMV of 8, when they won against Rayo Vallecano 10-2 on 12/20/2015.

- Largest Margin of Defeat (LMD) = the largest difference betweem the number of goals surrendered and the number of goals scored by the losing team in a single Liga regular season match. Real Madrid's 2015/16 LMD is 4, when they lost to Barcelona 0-4 on 11/21/2015.

- Longest Winning Streak (LWS) = the maximum number of wins in succession, or the maximum number of wins in a row. For the 2015/2016 season, Real Madrid enjoyed a LWS of 12 games between 3/2/2016 and 5/14/2016.

- Longest Unbeaten Streak (LUBS) = the maximum number of matches in succession played without being defeated (win or draw). Between 3/2/2016 and 5/14/2016, Real Madrid played 12 La Liga matches without suffering a single loss.

- Longest Losing Streak (LLS) = longest series of losses by a team. Real Madrid is considered one of the best teams in La Liga and in the world, evident from their LLS being only 2 games between 11/8/2015 and 11/21/2015.

- Longest Winless Streak = most matches without a win, their either draw or loose, Real Madrid had a winless streak of only 2 games in the same season.

Data Source

The statistics we use in the present paper are freely available to the public; we develop our own R-based data scraper (program) and use it to extract our data from the website ESPN FC. Our data set covers all of the Big Five (EPL, La Liga, Bundesliga, League 1, Serie A) and spans from seasons 2001/2 - 2015/16. in addition to the eight performance metrics we defined in earlier section, we also collect our response values of aggregated attendance for each team-season unit.

References

Bell G, Hey T, Szalay A (2009) Beyond the data deluge. *Science* 323(5919):1297–1298.

Dienes Z (2011) Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science* 6(3):274–290.

Hey T, Tansley S, Tolle KM, et al. (2009) *The fourth paradigm: data-intensive scientific discovery*, volume 1 (Microsoft research Redmond, WA).

Lynch SM (2007) *Introduction to applied Bayesian statistics and estimation for social scientists* (Springer Science & Business Media).

Tipping ME (2004) Bayesian inference: An introduction to principles and practice in machine learning. *Lecture notes in computer science* 3176:41–62.

Wagenmakers EJ (2007) A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review* 14(5):779–804.

Table 1 Descriptive Statistics

	Mean	Median	Std. Dev.	Min.	Max.	Interquartile Range
MHG	3.634	4	1.676	0	9	2
MAG	2.884	3	1.676	0	10	2
LMV	4.319	4	1.409	1	10	2
LMD	3.588	3	1.186	1	8	1
LWS	4.303	4	2.254	1	22	2
LUBS	8.844	8	5.213	2	45	6
LLS	2.881	3	1.283	1	13	2
LDDS	5.578	5	2.741	1	21	3
AATT	705808.736	608990.5	451624.726	4048	2477095	528828

all performance variables including attendance

Table 2 Correlation Matrix

	LLS	LMD	LMV	LUBS	LWLSS	LWS	MAG	MHG	AATT
LLS	1.000	0.308	-0.255	-0.412	0.539	-0.379	-0.172	-0.274	-0.247
LMD	0.308	1.000	-0.222	-0.347	0.296	-0.292	-0.168	-0.207	-0.134
LMV	-0.255	-0.222	1.000	0.415	-0.396	0.436	0.558	0.768	0.417
LUBS	-0.412	-0.347	0.415	1.000	-0.452	0.612	0.314	0.362	0.409
LWLSS	0.539	0.296	-0.396	-0.452	1.000	-0.416	-0.279	-0.355	-0.335
LWS	-0.379	-0.292	0.435	0.612	-0.416	1.000	0.336	0.382	0.478
MAG	-0.172	-0.168	0.558	0.314	-0.279	0.336	1.000	0.185	0.260
MHG	-0.274	-0.207	0.768	0.362	-0.355	0.382	0.185	1.000	0.383
AATT	-0.247	-0.134	0.417	0.409	-0.335	0.478	0.260	0.383	1.000

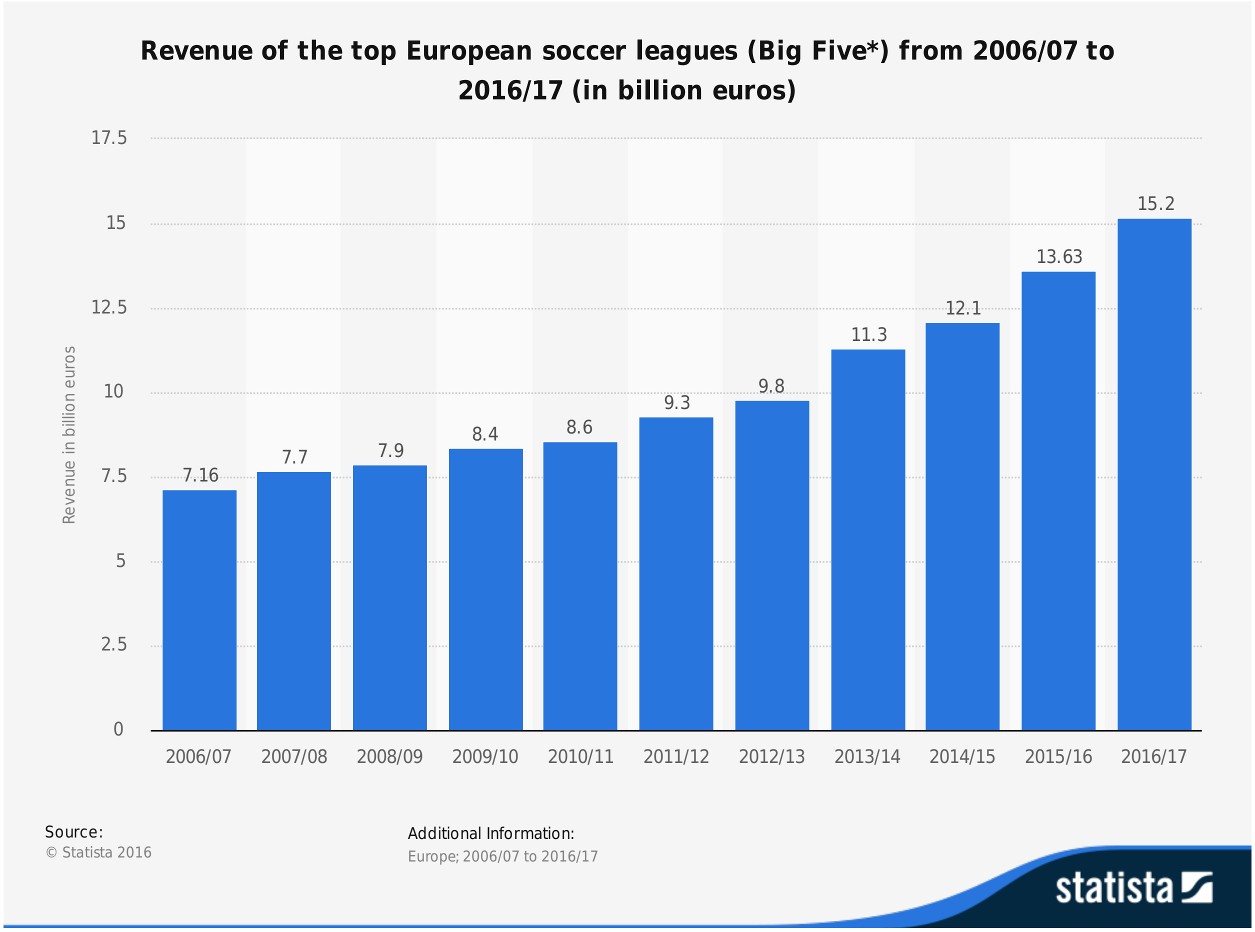
all coefficients are significant at the p value of 0.001 level

Table 3 Model Results

Variable Name	OLS	CV-LASSO	CV-Elastic Net	CV-Ridge Regression
MHG	0.165 (***)	0.149	0.152	0.159
MAG	0.029 (***)	0.019	0.021	0.039
LMV	0.234 (**)	0.247	0.243	0.222
LMD	0.139 (NS)	0.109	0.112	0.103
LWS	0.346 (***)	0.341	0.339	0.294
LUBS	0.132 (***)	0.125	0.127	0.131
LLS	0.004 (NS)			-0.014
LDDS	-0.114 (*)	-0.105	-0.106	-0.106
CV-MSE	0.291	0.290	0.277	0.291

Tex of notes

Figure 1 **Revenue of the top European soccer leagues (Big Five*) from 2006/07 to 2016/17 (in billion euros)**



Note. Notes

Figure 2 **Average per Game Attendance of the Biggest European Soccer Leagues from 96/97 t0 2015/16 (in thousands)**

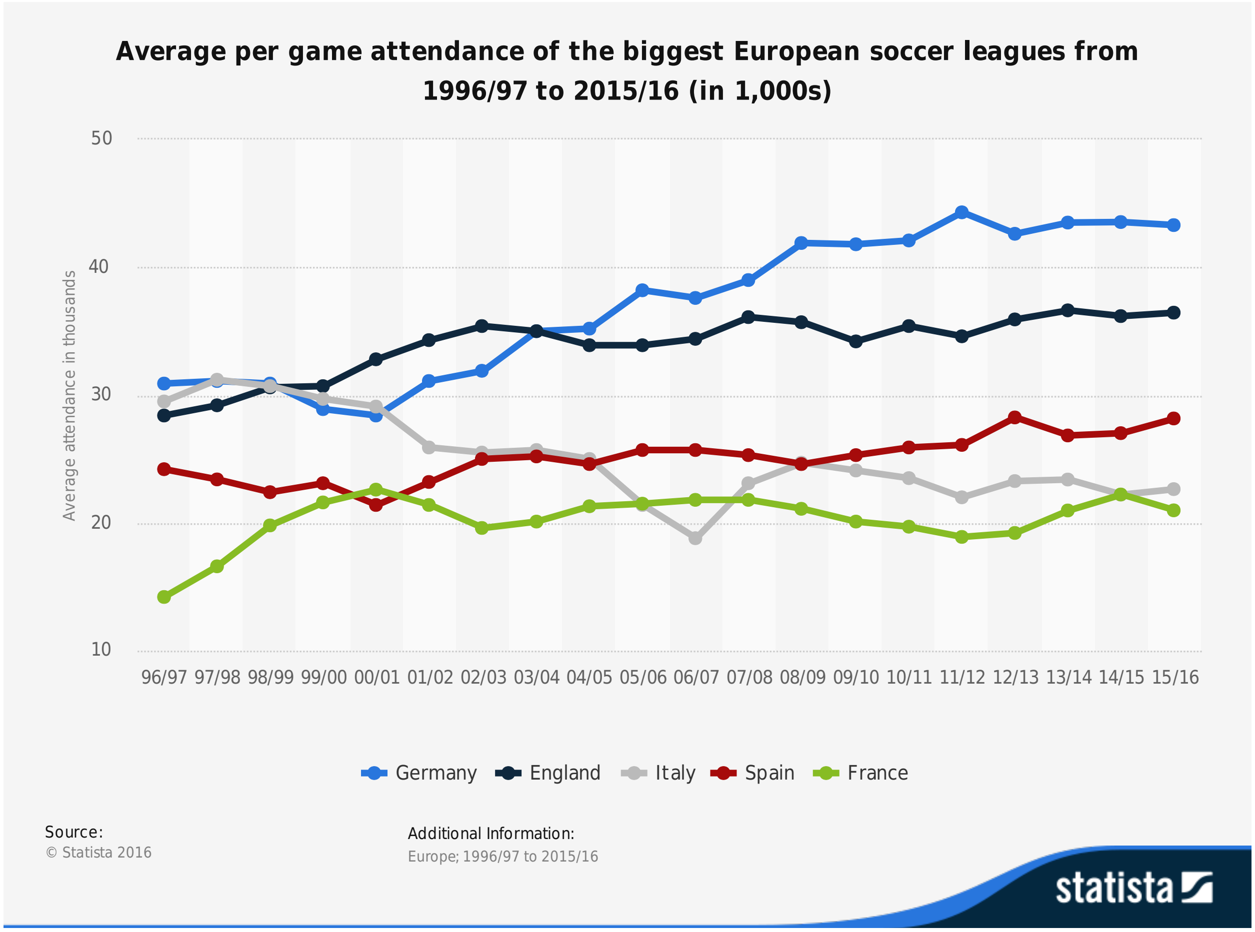


Figure 3 Most Valuable Soccer Brands in 2017 (in million U.S. $)

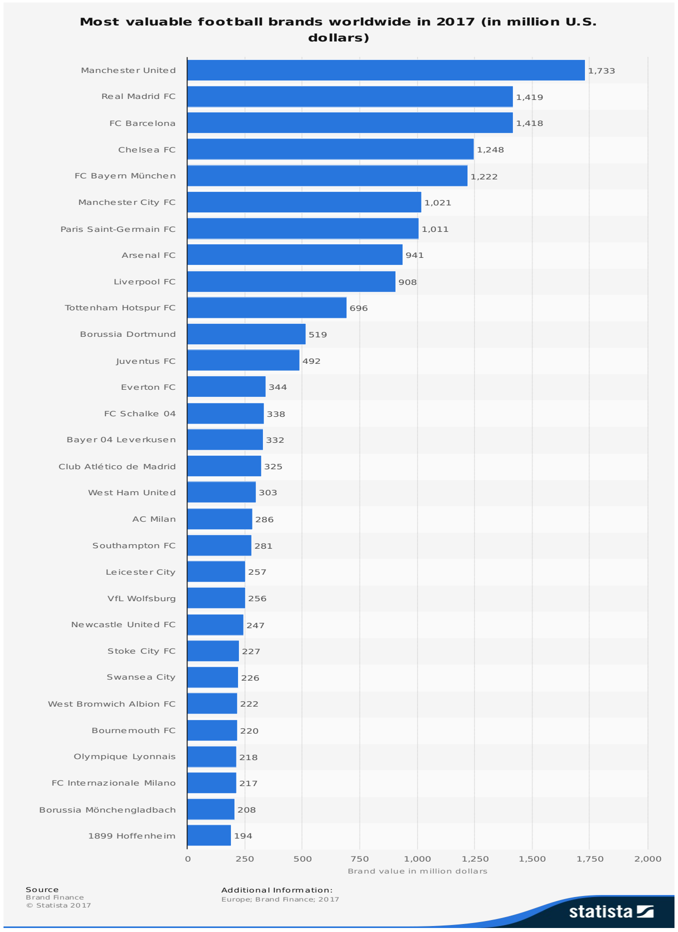


Figure 4 Club-Seasons by League

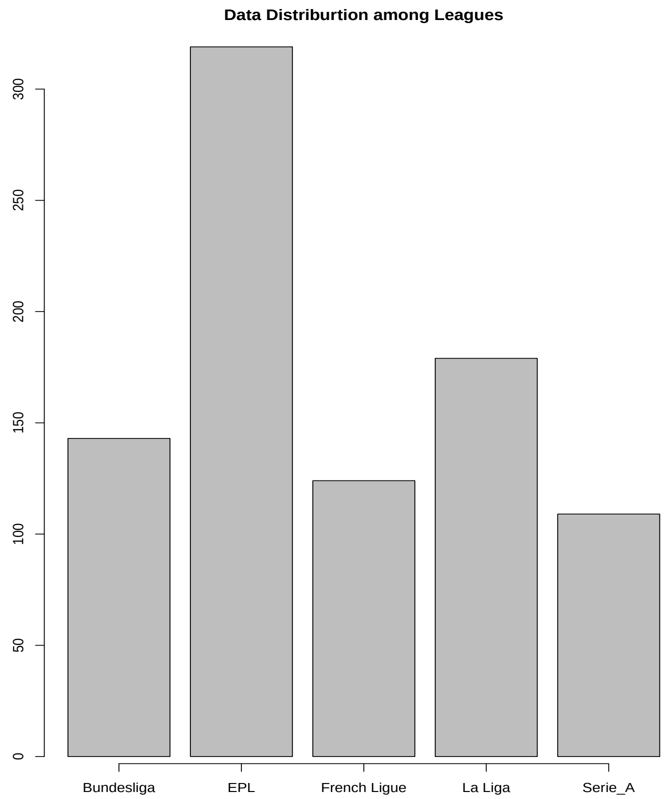


Figure 5 Relative Importance by Team Performance Metrics

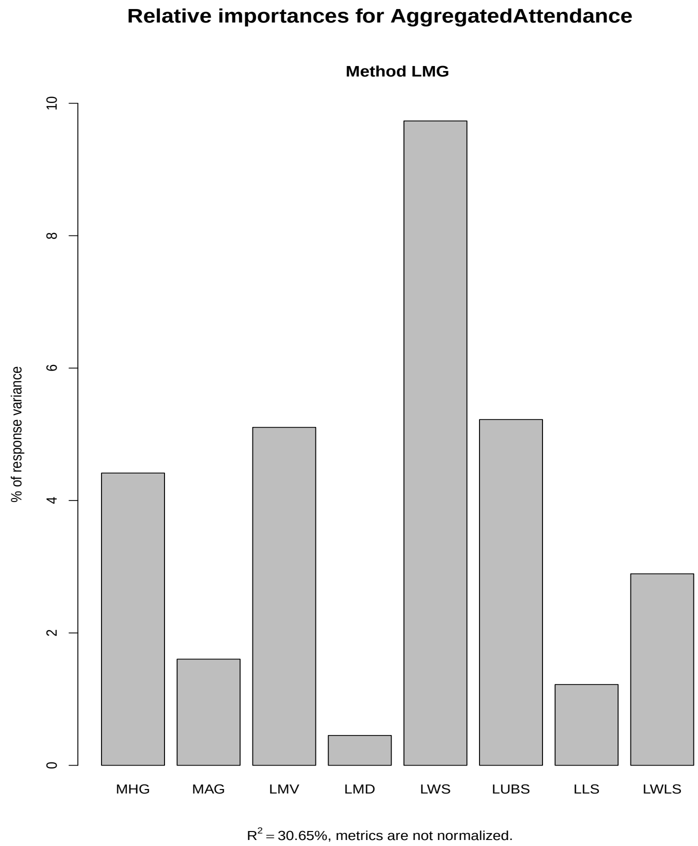


Figure 6 Bayesian Network Graphical Model