UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

WORKSHOP

BDSports: BIG DATA ANALYTICS IN SPORTS

bodai.unibs.it/BDSports/

November 16, 2017
Aula Magna, Via S. Faustino 74/b, Brescia

Session 1: 9.30-12.30

Chair: P. Zuccolotto, University of Brescia, Italy

AN ALTERNATIVE RANKING FOR NATIONAL SOCCER TEAMS BASED ON STRENGTH PARAMETERS
C. Ley, Universiteit Gent, Belgium

MODELING AND PREDICTION OF INTERNATIONAL SOCCER MATCHES
A. Groll, LMU München, Germany

DRIVERS OF FOOTBALL MATCH OUTCOMES WITH DATA MINING: A BENCHMARKING STUDY OF ITALIAN AND ISRAELI STATISTICS
R. S. Kenett, KPA Ltd., Israel

Session 2: 15.00-16.30

Chair: M. Manisera, University of Brescia, Italy

TEACHING STATISTICS IN UPPER SECONDARY SCHOOL: THE POSITIVE PLS (PIANO LAUREE SCIENTIFICHE) EXPERIENCE
S. Bianconcini, A. Lubisco, S. Mignani, A. Macaluso, University of Bologna, Italy

TENNIS ANALYTICS: A REVIEW
F. Lisi, University of Padova, Italy

SPORTS RANKING: IDENTIFYING CLUSTERS OF EXTREME PREFERENCES
M. Iannario, R. Simone, University of Napoli Federico II, Italy

Participation is free. To register, please send an email to marica.manisera@unibs.it

BODAI – LAB
BIG&OPEN DATA INNOVATION LABORATORY

Via S. Faustino 74/b, 25122 Brescia – Italy
Tel +390302988742 Fax +390302988703
email: paqricerca.dem@unibs.it

With the financial support of «Fondo di Ateneo per attività a carattere internazionale», Università di Brescia

ABSTRACTS

AN ALTERNATIVE RANKING FOR NATIONAL SOCCER TEAMS BASED ON STRENGTH PARAMETERS

C. Ley, Universiteit Gent, Belgium

The FIFA ranking of national (male) soccer teams suffers from various drawbacks that cause heated debates, all-the-more as this ranking governs the seating in the international tournaments and the associated qualification rounds. In this talk, I shall discuss and compare five alternative models based on strength parameters that are determined by maximum likelihood estimation as well as the ELO model used for the women's FIFA ranking. Predictive performance will be the main measure of comparison since, contrary to regular league rankings, the FIFA ranking is designed to measure the current strength of teams, hence should allow accurate predictions of future matches. The best performing model is used to build the alternative FIFA ranking and to post-analyze the 2016 European Championship (EURO2016).

MODELING AND PREDICTION OF INTERNATIONAL SOCCER MATCHES

A. Groll, Lmu München, Germany

When analyzing and modeling the results of soccer matches, one important aspect is to account for the correct dependence of the scores of two competing teams. A common approach is to use (univariate) Poisson regression, treating the number of goals scored by the competing teams as independent random variables given the covariate information of both teams. In this talk we first illustrate such a (regularized) Poisson regression model, which includes various potentially influential covariates describing the national teams' success in previous FIFA World Cups. It has been used in Groll et al. (2015) for the modeling and prediction of the FIFA World Cup 2014. In this context, the information contained in bookmakers' odds turned out to be of particular importance. In order to analyze if this type of modeling is appropriate or if additional explicit modeling of the dependence structure for the joint score of a soccer match needs to be taken into account we present a second approach. There, the number of goals a team scores against a specific opponent is modeled by a joint bivariate Poisson model, including covariate information of both competing teams (Groll et al., 2017). The model was estimated using the R-package gamboostLSS (Hofner et al., 2016; Mayr et al., 2012). With gamboostLSS the model family of GAMLSS (Generalized Additive Models for Location, Scale and Shape) is combined with a gradient boosting estimation technique. It allows to use multi-parametric distributions in regression models in combination with implicit variable selection. Based on all matches from the three previous UEFA European football championships a sparse model has been obtained: from a set of potential covariates already used in Groll and Abedieh (2013) only three covariates, namely the bookmakers' odds (odds for winning the title before the tournament), the market value and the UEFA points were chosen. This model was then used to repeatedly simulated (1,000,000 times) all match outcomes of the UEFA European football championship 2016 in France, resulting in winning probabilities for all participating national teams.

ABSTRACTS

DRIVERS OF FOOTBALL MATCH OUTCOMES WITH DATA MINING: A BENCHMARKING STUDY OF ITALIAN AND ISRAELI STATISTICS

R. S. Kenett, Kpa Ltd., Israel

In recent years, the role of data and statistics in the world of professional sports in general, and specifically professional football, is becoming exceedingly significant. Thousands of data points, in hundreds of parameters and factors, are recorded for every top level match and can be used to consolidate outcome driving parameters and thresholds, along side finding parameters that characterize playing style, formations, teams, leagues and competitions. Carpita et al. (2014) apply several data science and machine learning tools to identify a relatively small set of factors that have a significant effect in determining match outcome. Specifically, they apply random forest algorithms and principal component analysis (PCA) for feature selection and parameter reduction. The original data set consisted of 482 variables from 4 seasons of Italian Serie A football and they generate match outcome probability predictions using different methods whose accuracy is compared. Specifically, they show that the amount of goal scoring opportunities, defensive actions near own goal, lost balls and crosses and attack success rate, are drivers of match outcome for the home team. Moreover, the amount of goal scoring opportunities, headers in own penalty area, general defensive actions, ball touches in midfield and more, were found to determine match outcome for the away team. Furthermore, the data was used to examine the similarities between different seasons of Italian top tier football.

In this study, we expanded the Italian study to similar data from the football league in Israel. This comparison provides a powerful benchmark evaluation with insights that can be further generalized. We model the relationship between outcomes of a football match (win, lose or draw) and a set of variables describing the game across time, by analyzing data from consecutive yearly championships in Israeli and Italian leagues. Differences and similarities are highlighted and interpreted. Characteristics of the two leagues are used to interpret the results.

TEACHING STATISTICS IN UPPER SECONDARY SCHOOL: THE POSITIVE PLS (PIANO LAUREE SCIENTIFICHE) EXPERIENCE

S. Bianconcini, A. Lubisco, S. Mignani, A. Macaluso, University of Bologna, Italy

In 2010 the Italian Ministry of Education introduced statistics and probability contents into the mathematics curriculum as a fundamental topic at all school levels under the domain named “Uncertainty and data”. In the Italian school system, few mathematics teachers have been trained in or are familiar with statistical concepts and methods, and most of them are not even competent in selecting relevant, useful, and meaningful real examples. As a consequence, many Italian school students consider statistics as a boring discipline. Trying to face these critical issues, a team of professors of the Department of Statistical Sciences (University of Bologna) supports training activities for teacher and laboratories for high-school students to enhance statistical thinking and reasoning through a data-oriented approach. These activities are included in the national project

ABSTRACTS

“Piano Lauree Scientifiche (PLS)”, of Ministry of Education. It started in 2005 in order to increase enrollments to scientific University degrees, and it consists in joint initiatives between high schools and universities.

In this work we present our last scholastic year experience in teaching Statistics through meaningful real examples using sport data. The project involved students who worked together as a team to solve assigned tasks or to reach a common goal. Students and teachers were encouraged to analyze and discuss the proposed problems and to send solutions, suggestions, and remarks. In this way, a learning community has been created. This represents a virtual environment where individuals and groups can exchange ideas and find statistical tools to solve problems.

This experience represents/have been a good practice to the Statistics learning for increasing the synergy between University and high school.

TENNIS ANALYTICS: A REVIEW

F. Lisi, University of Padova, Italy

The talk gives a short review of the applications of sport analytics in tennis and shows how quantitative analyses and statistical methods have been applied to tennis and main results.

SPORTS RANKING: IDENTIFYING CLUSTERS OF EXTREME PREFERENCES

M. Iannario, R. Simone, University of Napoli Federico II, Italy

The talk is addressing theory on models for ranked preference data based on the inflation of extreme categories. The three models which will be in focus are Generalized CUB models (GeCUB, Iannario and Piccolo, 2016), the nested CUSH models presented in Simone and Piccolo, 2017), and a variation of the mixture of IHG distributions (Simone and Iannario, 2017). The predominant body of the literature introduces covariates for taking the inflation at the extremes into account. Here an alternative method will be presented. The approach deals with a two-component mixture of IHG distributions to determine groups with opposite preferences. A theoretical description, inferential issues and some notation for identification of the new mixture will be presented. A discussion on a survey on sports ranking collected thanks to the Big&Open Data Innovation Laboratory concludes the talk.