DISCOVERING THE DRIVERS OF F⚽⚽TBALL MATCH OUTCOMES WITH DATA MINING

Maurizio **C**arpita, **M**arco **S**andri,
Anna **S**imonetto, **P**aola **Z**uccolotto

DMS StatLab – University of Brescia

Paris - 2014, April 11th

2014 ENBIS-SFdS Spring Meeting

⚽UTLINE OF THE TALK

- The case study: data and goals
- Football Mining: the Data Mining process
- Focus: variable selection
- Main outcomes
- Concluding remarks

THE CASE STUDY: DATA AND GOALS

The top Italian professional football league «serie A»

20 teams
38 matches/season

4 seasons → 2008/2009 2009/2010 2010/2011 2011/2012

DATA → PANINI DIGITAL *DigitalScout*

1,300 recorded events for each match (e.g. free kicks and shots, action type, fouls, crosses, recovered balls, goal assists, average time of ball possession, saves, goals on free kicks, etc.)

482 variables

THE CASE STUDY: DATA AND GOALS

Indentify the factors which mostly affect the probability of winning the match

F⚽⚽TBALL MINING: THE DATA MINING PROCESS

Variable selection → Data reduction → Prediction

Random Forest
Variable Importance Measures
Bias correction

482 → **17**
(6+11)

Principal Component Analysis

17 → 6
(3+3)

Multinomial Logit

Game strategy

F⚽⚽TBALL MINING: THE DATA MINING PROCESS

Variable selection

Random Forest Variable Importance Measures

Bias correction

482 → **17** **(6+11)**

Data reduction

Principal Component Analysis

17 → **6** **(3+3)**

Prediction

Multinomial Logit

Game strategy

F⚽⚽TBALL MINING: THE DATA MINING PROCESS

Variable selection

Random Forest Variable Importance Measures

Bias correction

482 → **17** **(6+11)**

Data reduction

Principal Component Analysis

17 → **6** **(3+3)**

Prediction

Multinomial Logit

Game strategy

FOCUS: VARIABLE SELECTION

Variable selection → Data reduction → Prediction

Random Forest
Variable Importance
Measures
Bias correction

482 → **17**
(6+11)

Principal
Component Analysis

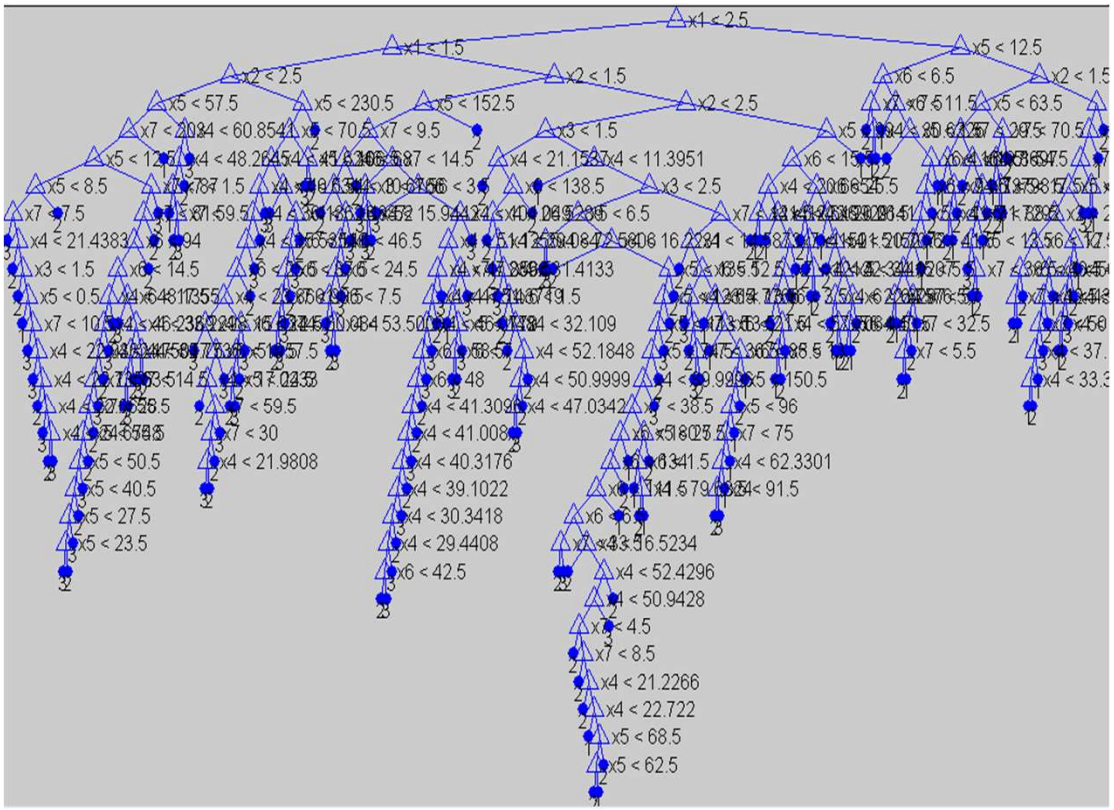
17 → 6
(3+3)

Multinomial Logit

Game strategy

FOCUS: VARIABLE SELECTION

Random Forest Variable Importance Measures (TDNI VIMs)



single tree

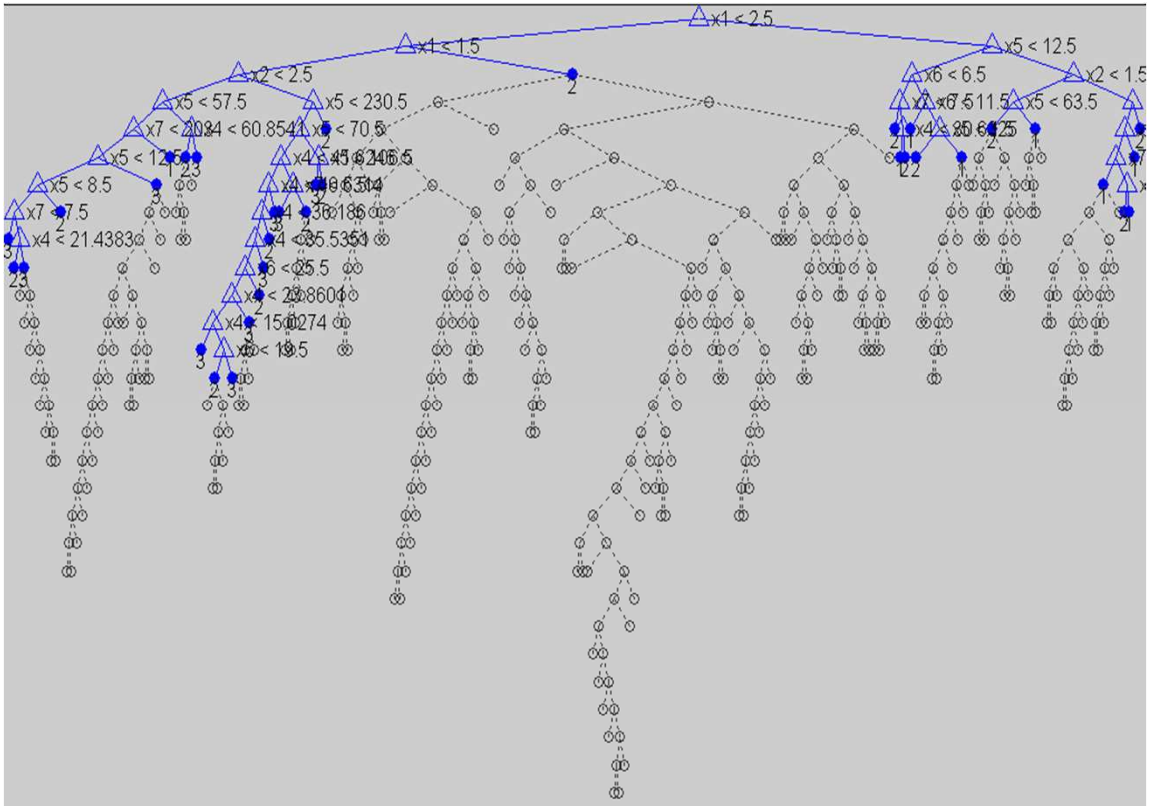
$$\widehat{\mathrm{VI}}_i(t) = \sum_{j \in J} \hat{d}_{ij} \cdot I_{ij}$$

RF

$$\widehat{\mathrm{VI}}_i = \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathrm{VI}}_i(t)$$

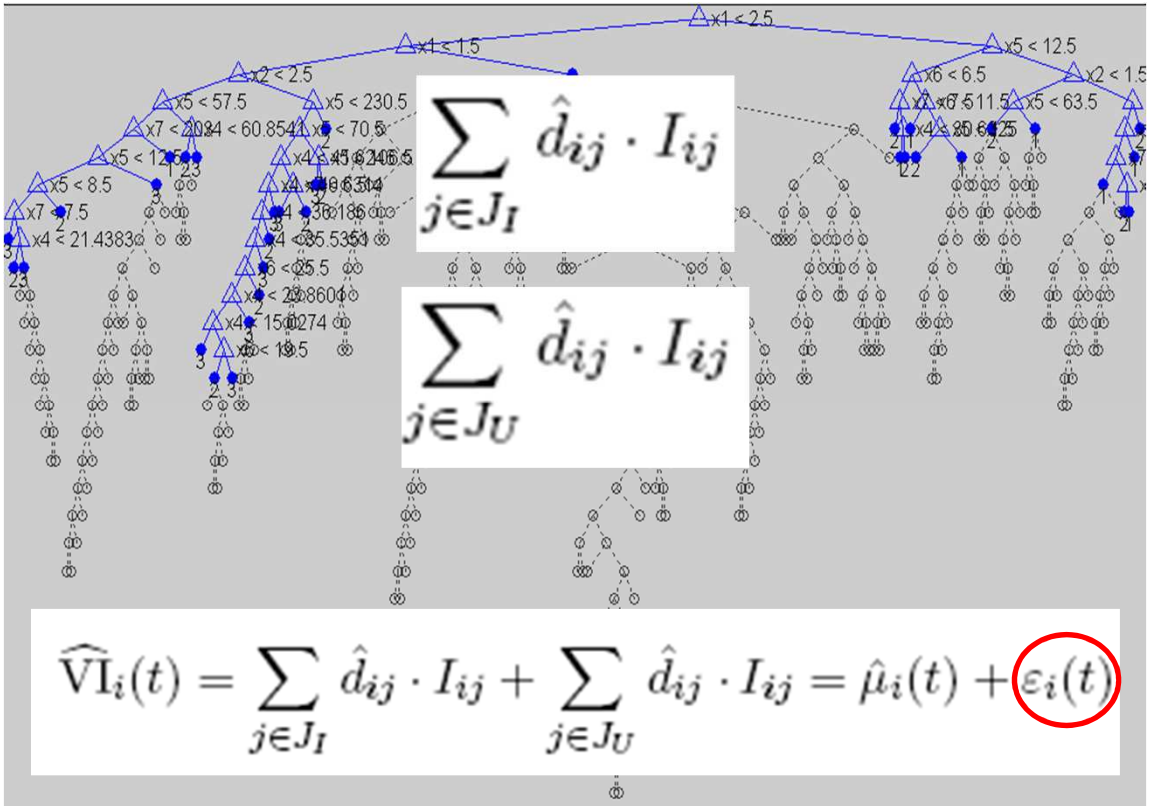
F⚽CUS: VARIABLE SELECTION

The source of **BIAS**: informative and uninformative splits (Sandri and Zuccolotto, 2008, 2010)



F⚽CUS: VARIABLE SELECTION

The source of **BIAS**: informative and uninformative splits (Sandri and Zuccolotto, 2008, 2010)



BIAS

FOCUS: VARIABLE SELECTION

Correction of **BIAS**: PSEUDO-COVARIATES METHOD (Sandri and Zuccolotto, 2008, 2010)

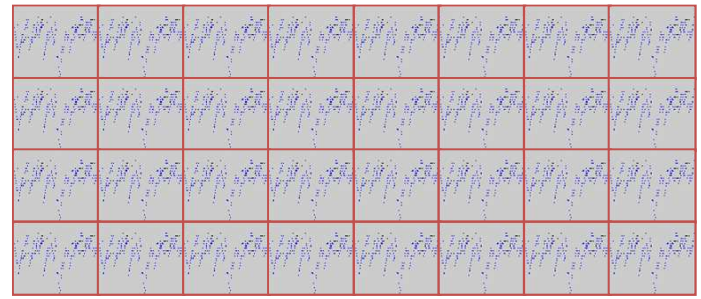
Original dataset

Soggetto	Y	X_1	...	X_s
1	y_1	x_{11}	...	x_{1s}
2	y_2	x_{21}	...	x_{2s}
...	...	...	...	...
N	y_N	x_{N1}	...	x_{Ns}

Pseudo-covariates

Z_1	...	Z_s
z_{11}	...	z_{1s}
z_{21}	...	z_{2s}
...	...	...
z_{N1}	...	z_{Ns}

Soggetto	Y	X_1	...	X_s	Z_1	...	Z_s
1	y_1	x_{11}	...	x_{1s}	z_{11}	...	z_{1s}
2	y_2	x_{21}	...	x_{2s}	z_{21}	...	z_{2s}
...	...	...	...	...	...	...	...
N	y_N	x_{N1}	...	x_{Ns}	z_{N1}	...	z_{Ns}



VIMs X

VIMs Z

S times

F⚽CUS: VARIABLE SELECTION

Correction of **BIAS**: PSEUDO-COVARIATES METHOD (Sandri and Zuccolotto, 2008, 2010)

Original dataset

Soggetto	Y	X_1	...	X_s
1	y_1	x_{11}	...	x_{1s}
2	y_2	x_{21}	...	x_{2s}
...	...	...	...	...
N	y_N	x_{N1}	...	x_{Ns}

Pseudo-covariates

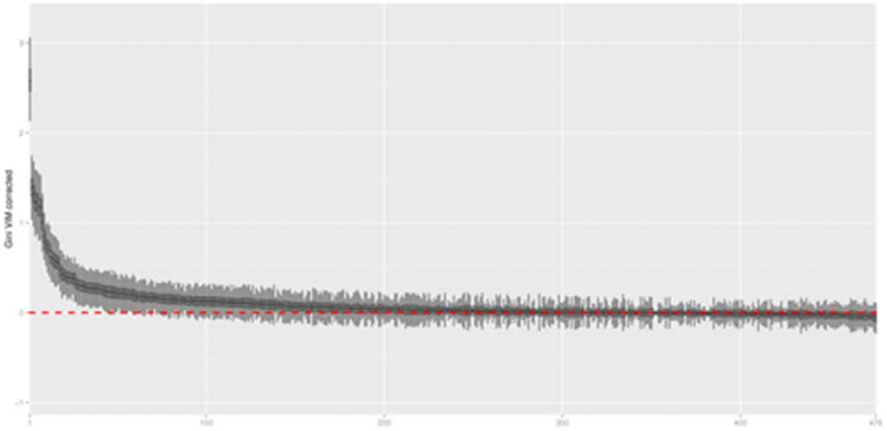
Z_1	...	Z_s
z_{11}	...	z_{1s}
z_{21}	...	z_{2s}
...	...	...
z_{N1}	...	z_{Ns}

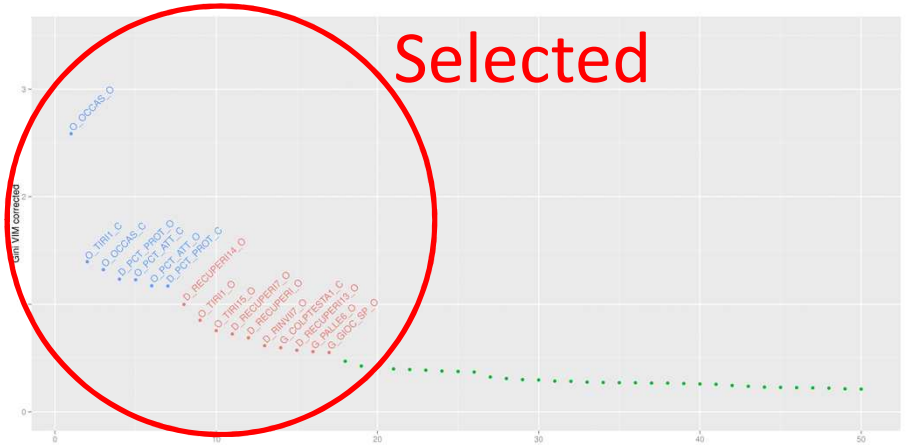
S times

$$\overline{\mathrm{VI}}_i = \frac{1}{S}\sum_{s=1}^{S}(\widehat{\mathrm{VI}}^s_{X_i} - \widehat{\mathrm{VI}}^s_{Z_i})$$

F⚽CUS: VARIABLE SELECTION

Correction of **BIAS**: PSEUDO-COVARIATES METHOD (Sandri and Zuccolotto, 2008, 2010)





MAIN ⚽ UTCOMES

Variable selection

Random Forest
Variable Importance Measures
Bias correction

482 → 17
(6+11)

Data reduction

Principal Component Analysis

17 → 6
(3+3)

Prediction

Multinomial Logit

Game strategy

MAIN ⚽UTCOMES

HOME TEAM

- shot.attack.home

ability to create opportunities to make shots on goal

- aerial.attack.home

aerial abilities (crosses and heading) when the team is on the attack

- defense.home

general defense abilities

MAIN ⚽UTCOMES

AWAY TEAM

- midfield-defense.counterattack.away

general defense abilities, long-range kicks and sudden counterattacks, with specific reference to actions in the midfield

- shot.attack.away

ability to create opportunities to make shots on goal

- area-defense.away

attitude to condense defense in the crucial penalty area

MAIN ⚽ UTCOMES

Variable selection → Data reduction → **Prediction**

Random Forest Variable Importance Measures

Bias correction

482 → **17** (6+11)

Principal Component Analysis

17 → **6** (3+3)

Multinomial Logit

Game strategy

MAIN ⚽ UTCOMES

- shot.attack.home
- aerial.attack.home
- defense.home

Indentify the factors which mostly affect the probability of winning the match

- midfield-defense.counterattack.away
- shot.attack.away
- area-defense.away

CONCLUDING REMARKS

- Data Mining tecnhiques offer interesting insights into sport strategies
- An effective variable selection technique is the starting point, as sport data are often big data
- The results of our analysis remain stable along the 4 examined seasons

MAIN REFERENCES

1. Agresti A. (2003). Logit Models for Multinomial Responses. In: Categorical Data Analysis, 2nd Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA.
2. Albert J., Koning R.H. (2008). Statistical Thinking in Sports, Chapman & Hall, Boca Raton.
3. Breiman L. (2001a). Random forests, Machine Learning, 45(1), 5-32.
4. Carpita M., Sandri M., Simonetto A., Zuccolotto P. (2014). Football Mining with R. In: Data Mining Applications with R (Edited by Y. Zhao, Y. Cen), Chapter 14. Elsevier.
5. Hastie T., Tibshirani R., and Friedman J.H. (2001). The elements of statistical learning: data mining, inference, and prediction. Springer, New York.
6. Jolliffe I.T. (2002). Principal Component Analysis. Springer Verlag, New York.
7. Sandri M., Zuccolotto P. (2008). A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees. Journal of Computational and Graphical Statistics, 17(3), 611-628.
8. Sandri M., Zuccolotto P. (2010). Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms. Statistics and Computing, 20, 393-407.