Progetto di Modelli e Metodi per l'Inferenza Statistica

Manuale per allenatori NCAA

Riccardo Cadei, Riccardo Fradiani

Politecnico di Milano

riccardo1.cadei@mail.polimi.it
riccardo.fradiani@mail.polimi.it

8 luglio 2020

Overview

Table of Contents

Presentazione del problema

- **Attacco vs. Difesa** (Manuale per allenatori NCAA): *Cosa rende un team vincente? Come organizzare al meglio gli allenamenti?* Confronto dell'impatto sulla percentuale di vittorie stagionali delle statistiche offensive e difensive di un team.
- **Previsione della qualificazione alla fase finale (March Madness)**: Confronto tra Regressione Logistica e modelli di Machine Learning (*Support Vector Machine, Random Forest*)

Il Dataset

Dati riguardanti le stagioni 2015, 2016, 2017, 2018, 2019 delle 351 squadre appartenenti alla prima divisione di college basketball americano.

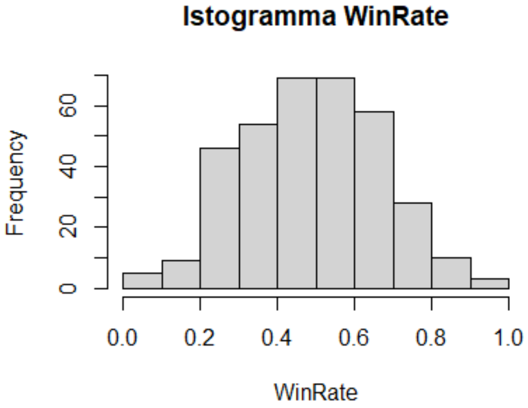
- **Features**: *TEAM*, *G* (Partite giocate), *W* (Partite vinte), *ADJOE* (Efficienza offensiva), *ADJDE* (Efficienza difensiva), *EFGO* (Percentuale di canestri realizzati), *EFGD* (Percentuale di canestri concessi), *TOR* (Frequenza Turnover), *TORD* (Frequaenza Steal), *ORB* (Percentuale di rimbalzi offensivi), *DRB* (Percentuale di rimbalzi difensivi), *FTR* (Frequenza tiri liberi), *FTRD* (Frequenza tiri liberi concessi), *2PO* (Percentuale canestri da due punti), *2PD* (Percentuale canestri da due punti concessi), *3PO* (Percentuale canestri da tre punti), *3PD* (Percentuale canestri da tre punti concessi), *POSTSEASON* (Round della MM in cui il team è stato eliminato).

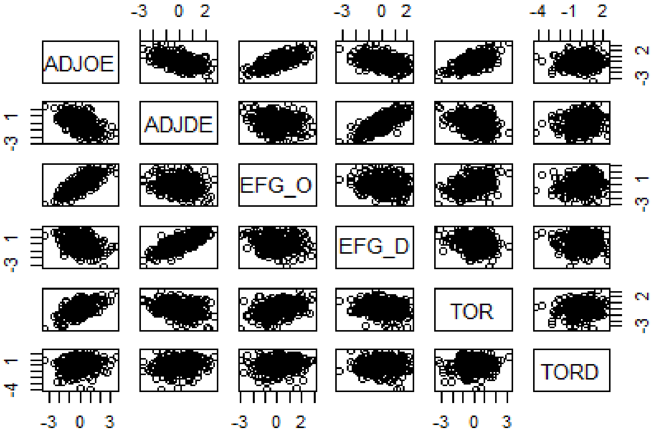
Perchè questo Dataset?

- **Normalità dei dati**: La suddivisione per squadre e i dati stagionali favoriscono la distribuzione normale dei dati.
- **Dati qualitativi**: La presenza di statistiche qualitative permette di realizzare un' analisi più coerente con il problema posto e più indipendente da fattori non controllabili dagli allenatori (*i.e. calendario, avversari*).

Table of Contents

- **Standardizzazione** dei dati
- WinRate = W/G
- **Analisi di normalità**: Shapiro test su WinRate (*p-value* = *0.3634*)





Osservazione

Alcune variabili potrebbero risultare eccessivamente correlate.

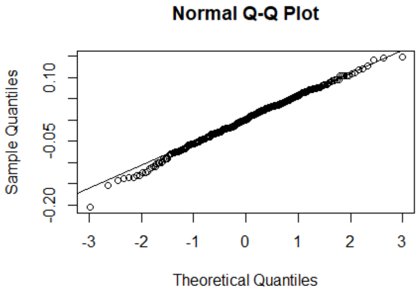
Regressione esplorativa

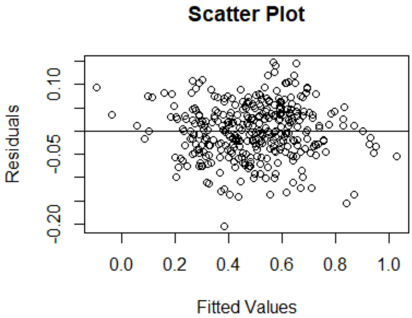
Come primo approccio, abbiamo realizzato un modello di regressione lineare ponendo il WinRate come risposta aleatoria e utilizzando tutti i predittori a nostra disposizione.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.204023 -0.034928  0.000551  0.037778  0.147118

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.902e-01  3.135e-03 156.350  < 2e-16 ***
ADJOE       -6.234e-03  1.186e-02  -0.525  0.59962
ADJDE        4.528e-05  1.055e-02   0.004  0.99658
EFG_O        1.327e-01  4.328e-02   3.066  0.00234 **
EFG_D       -1.287e-01  4.369e-02  -2.946  0.00344 **
TOR          5.466e-02  5.648e-03   9.679  < 2e-16 ***
TORD        -4.803e-02  5.489e-03  -8.750  < 2e-16 ***
ORB          4.455e-02  5.194e-03   8.577 3.64e-16 ***
DRB         -2.546e-02  4.773e-03  -5.335 1.76e-07 ***
FTR          1.828e-02  3.544e-03   5.158 4.27e-07 ***
FTRD        -2.771e-02  3.969e-03  -6.981 1.58e-11 ***
X2P_O       -4.073e-02  2.889e-02  -1.410  0.15958
X2P_D        4.751e-02  3.215e-02   1.478  0.14044
X3P_O       -1.592e-02  2.062e-02  -0.772  0.44065
X3P_D        2.257e-02  1.909e-02   1.182  0.23797
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05874 on 336 degrees of freedom
Multiple R-squared:  0.8984,    Adjusted R-squared:  0.8942
F-statistic: 212.2 on 14 and 336 DF,  p-value: < 2.2e-16
```

Dall'**analisi dei residui** di tale modello abbiamo ottenuto i seguenti risultati:

- *Shapiro test*: p-value = 0.3669





Feature selection

Sebbene l'ipotesi di normalità sui residui sia rispettata e gli indicatori di bontà complessiva del modello siano positivi, andando ad analizzare la correlazione tra predittori attraverso il **VIF** ci siamo accorti che alcune features risultavano eccessivamente correlate.

ADJOE	ADJDE	EFG_O	EFG_D	TOR	TORD	ORB	DRB	FTR	FTRD	X2P_O	X2P_D	X3P_O	X3P_D
14.279082	11.290201	189.994769	193.594188	3.235355	3.055884	2.736662	2.310729	1.273887	1.598167	84.679355	104.878447	43.129006	36.970128

A questo punto, abbiamo proceduto rimuovendo i predittori che generavano questa eccessiva correlazione. In particolare, abbiamo rimosso *ADJOE, ADJDE* (poichè, in quanto indicatori generali sull'efficienza, risultavano eccessivamente correlati agli indicatori più specifici) e anche *X2PO, X2PD, X3PO, X3PD* (mantenendo solo gli indicatori più generali sull'efficienza dei tiri).

Nuovo Modello di Regressione

Il nuovo modello presenta il seguente VIF

```
   EFG_O    EFG_D      TOR     TORD      ORB      DRB      FTR     FTRD
1.324264 1.310637 1.348600 1.285066 1.242574 1.244946 1.137287 1.404905
```

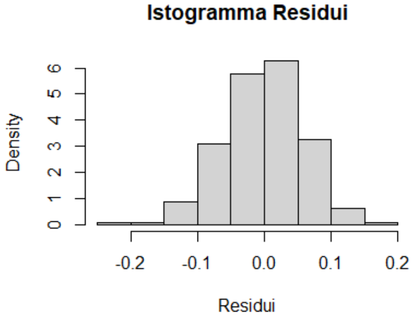
Inoltre, il coefficiente di determinazione risulta quasi inalterato mentre i p-value associati all'ipotesi di nullità dei coefficienti dei singoli predittori risultano tutti migliorati (e prossimi a 0).

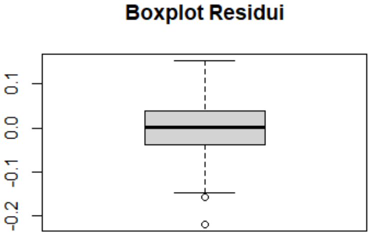
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.219304 -0.037391  0.001685  0.039411  0.153077

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.490209   0.003160 155.136  < 2e-16 ***
EFG_O        0.079139   0.003641  21.733  < 2e-16 ***
EFG_D       -0.067297   0.003623 -18.577  < 2e-16 ***
TOR          0.053168   0.003675  14.468  < 2e-16 ***
TORD        -0.046090   0.003587 -12.849  < 2e-16 ***
ORB          0.040815   0.003527  11.571  < 2e-16 ***
DRB         -0.025034   0.003531  -7.090 7.71e-12 ***
FTR          0.017217   0.003375   5.102 5.59e-07 ***
FTRD        -0.025826   0.003751  -6.886 2.77e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0592 on 342 degrees of freedom
Multiple R-squared:  0.895,     Adjusted R-squared:  0.8925
F-statistic: 364.3 on 8 and 342 DF,  p-value: < 2.2e-16
```

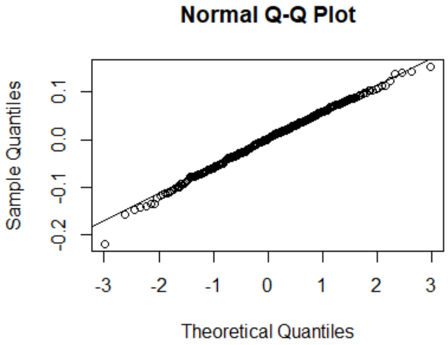
Analisi dei residui

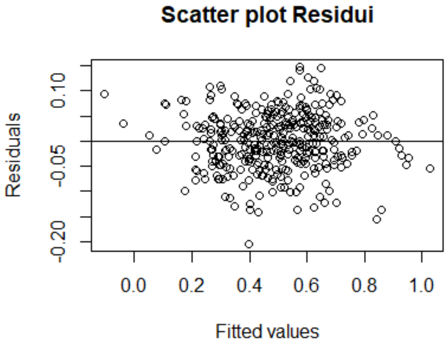
Dall'analisi dei residui del nuovo modello abbiamo ottenuto i seguenti risultati:

- *Shapiro Test*: p-value = 0.5045







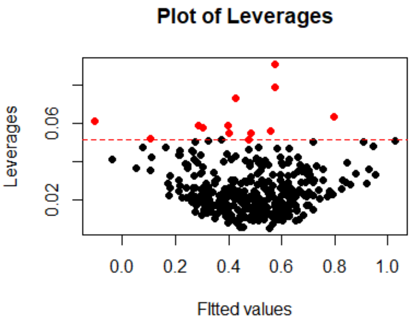


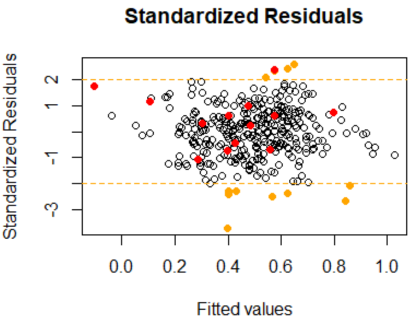
Osservazione

I risultati ottenuti suggeriscono che l'ipotesi di normalità dei residui continua ad essere rispettata anche per il nuovo modello.

Analisi Leverages e Outliers

Nonostante la presenza di alcuni punti di leverage e outliers, la bontà complessiva del modello e il rispetto delle ipotesi di normalità ci permettono di mantenere i dati interessati per tutelare la rappresentatività del campione e la scalabilità del modello. Le stesse considerazioni valgono anche per i punti con distanza di Cook elevata.





Separazione features: ATT vs. DIF

Confermata la bontà del modello, abbiamo proceduto analizzando separatamente i predittori legati all'attacco e quelli legati alla difesa. In particolare, abbiamo realizzato un indice offensivo e uno difensivo da associare a ciascun team:

$$ind_{att} = \frac{\beta_{EFGO} \cdot EFGO + \beta_{TOR} \cdot TOR + \beta_{ORB} \cdot ORB + \beta_{FTR} \cdot FTR}{\beta_{EFGO} + \beta_{TOR} + \beta_{ORB} + \beta_{FTR}}$$

$$ind_{dif} = \frac{\beta_{EFGD} \cdot EFGD + \beta_{TORD} \cdot TORD + \beta_{DRB} \cdot DRB + \beta_{FTRD} \cdot FTRD}{\beta_{EFGD} + \beta_{TORD} + \beta_{DRB} + \beta_{FTRD}}$$

Osservazione

Gli indicatori offensivi influiscono maggiormente sulla definizione del modello, infatti:

$$\beta_{tot,att} > \beta_{tot,dif}$$

Confronto modelli: ATT vs. DIF

A questo punto, abbiamo realizzato due diversi modelli di regressione: uno utilizzando l'indice offensivo trovato come unico regressore e uno utilizzando l'indice difensivo.

Modello Offensivo:

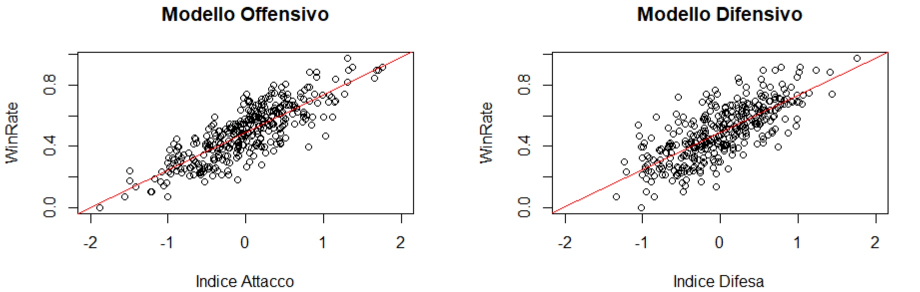
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.297891 -0.073722  0.001301  0.068822  0.249829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.490209   0.005447   90.00   <2e-16 ***
ind_att     0.247093   0.009041   27.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.102 on 349 degrees of freedom
Multiple R-squared:  0.6815,    Adjusted R-squared:  0.6806
F-statistic: 746.9 on 1 and 349 DF,  p-value: < 2.2e-16
```

- *Shapiro test:* p-value = 0.3426

Modello Difensivo:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.34296 -0.07887 -0.00672  0.07187  0.36231

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.490209   0.006483   75.62   <2e-16 ***
ind_dif     0.243664   0.011825   20.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1215 on 349 degrees of freedom
Multiple R-squared:  0.5489,    Adjusted R-squared:  0.5476
F-statistic: 424.6 on 1 and 349 DF,  p-value: < 2.2e-16
```

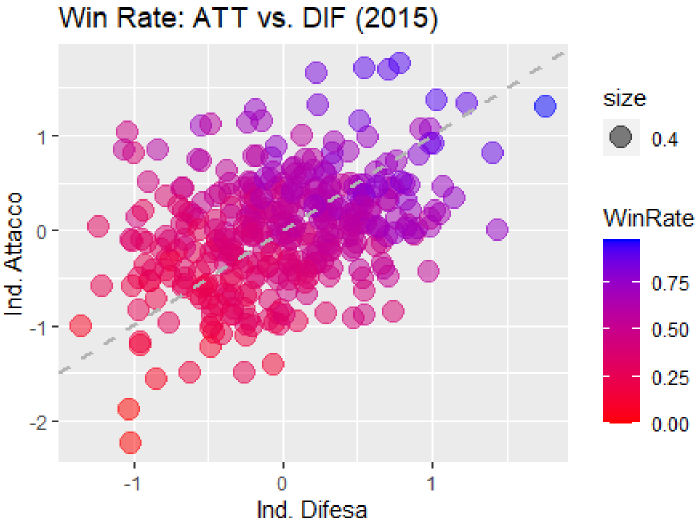
- *Shapiro test:* p-value = 0.4496

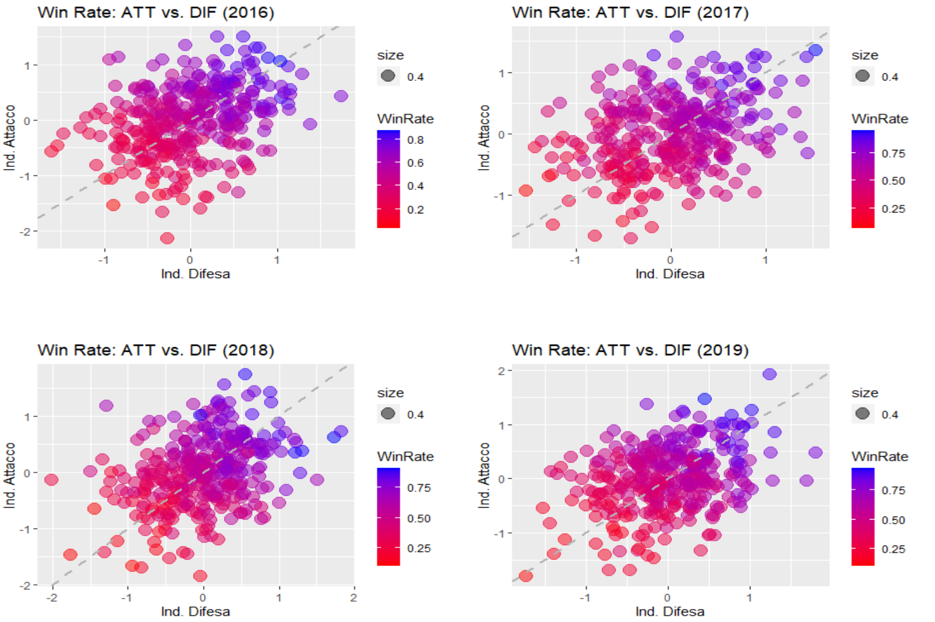


Osservazione

Sebbene entrambi i modelli risultino validi e rispettino le ipotesi di normalità dei residui, il modello basato sull'indice offensivo presenta un *coefficiente di determinazione più elevato* (0.68 *vs.* 0.55). Tale risultato corrobora l'osservazione precedente sulla maggiore incidenza dell'attacco nel determinare la performance stagionale di un team.

Bubble Plot: WinRate

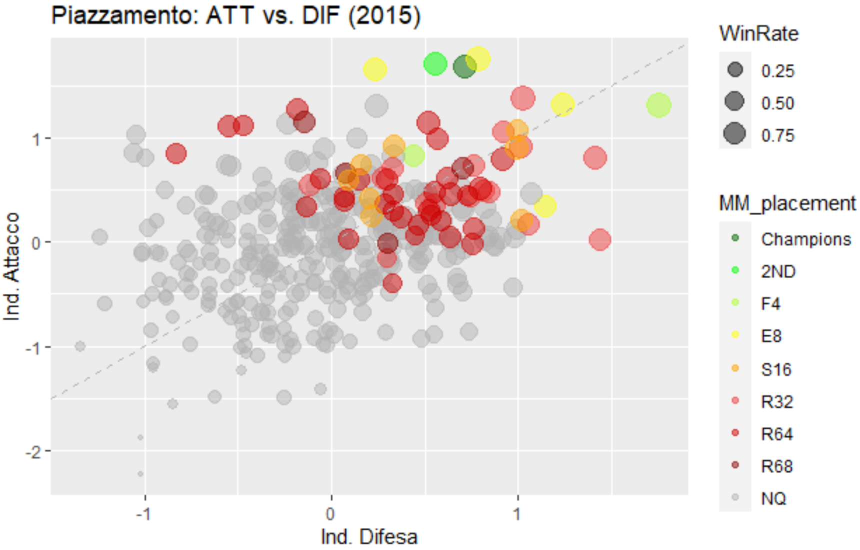


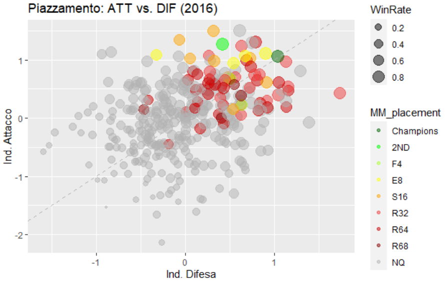

Win Rate: ATT vs. DIF (2016)
Win Rate: ATT vs. DIF (2017)
Win Rate: ATT vs. DIF (2018)
Win Rate: ATT vs. DIF (2019)
Ind. Attacco
Ind. Difesa
size
0.4
WinRate
0.8
0.6
0.4
0.2
0.75
0.50
0.25

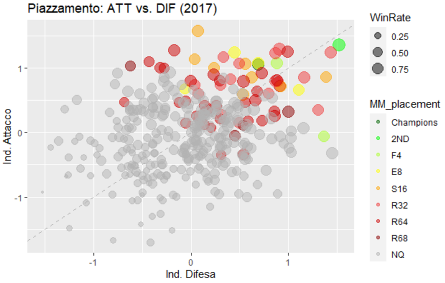
Conclusioni

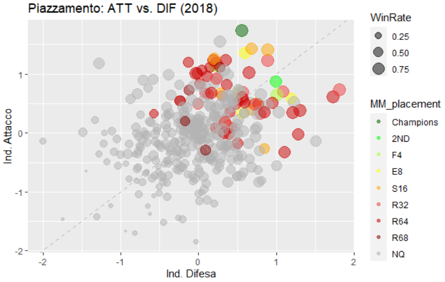
- Sono i team più bilanciati ad ottenere i win rate più elevati. Raramente i team con attacco molto migliore della difesa (o viceversa) ottengono buoni risultati.
- Analizzando i dati sui 5 anni, i team con *ind. offensivo* $>$ *ind. difensivo* ottengono in media un win rate più elevato (**+2.7%**)

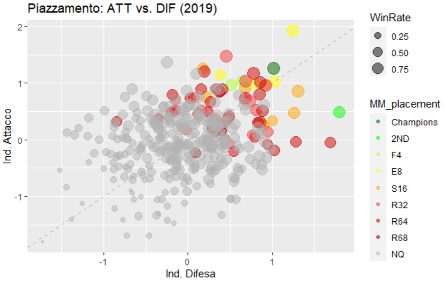
Bubble Plot: Piazzamento




Piazzamento: ATT vs. DIF (2016)
Ind. Attacco
Ind. Difesa
1
0
-1
-2
-1
0
1
WinRate
0.2
0.4
0.6
0.8
MM_placement
Champions
2ND
F4
E8
S16
R32
R64
R68
NQ


Piazzamento: ATT vs. DIF (2017)
Ind. Attacco
Ind. Difesa
1
0
-1
-1
0
1
WinRate
0.25
0.50
0.75
MM_placement
Champions
2ND
F4
E8
S16
R32
R64
R68
NQ


Piazzamento: ATT vs. DIF (2018)
Ind. Attacco
Ind. Difesa
1
0
-1
-2
-2
-1
0
1
2
WinRate
0.25
0.50
0.75
MM_placement
Champions
2ND
F4
E8
S16
R32
R64
R68
NQ


Piazzamento: ATT vs. DIF (2019)
Ind. Attacco
Ind. Difesa
2
1
0
-1
-1
0
1
WinRate
0.25
0.50
0.75
MM_placement
Champions
2ND
F4
E8
S16
R32
R64
R68
NQ

Osservazione

Nonostante la modalità ad eliminazione diretta del torneo, i risultati ottenuti sul piazzamento risultano in linea con quelli ottenuti sul win rate. Ciò ci fornisce un'ottima base di partenza nell'implementazione dei modelli predittivi.

Table of Contents

Problema 2: Previsione

Obiettivo

Predire le ammissioni alla March Madness a partire dalle statistiche di gioco annuali selezionate nel *Problema 1*

3 classificatori a confronto:

- *Logistic Regression* (classifier)
- *Support Vector Machine*
- *Random Forest*

Training Set - Test Set

Dividiamo il Dataset in due parti:

- **Training Set** (80%): a partire dal quale determiniamo i parametri dei 3 classificatori
- **Test Set** (20%): sul quale valutiamo l'accuracy dei modelli

Unbalanced Dataset

Il numero di training example per ciascuna classe (qualificato, non qualificato) non è bilanciato (Imbalance Ratio = 0.24028). Si potrebbe ricorrere a tecniche di:

- **Over Sampling**: es. *Majority Weighted Minority Oversampling TEchnique, RApidly COnverging Gibbs, Random Walk OverSampling*
- **Under Sampling**: es. *Random Under Sampling*

Tuttavia nessun algoritmo sembra migliorare significativamente la previsione.

Accuracy

$$Sensibilit\grave{a} = \frac{TP}{TP + FN}$$

$$Specificit\grave{a} = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1_{score} = \frac{2TP}{2TP + FP + FN}$$

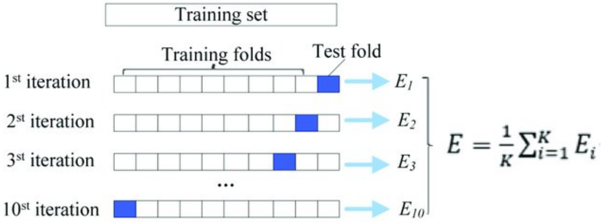
		Predicted class	
		P	**N**
Actual class	**P**	**TP**	FN
	N	FP	**TN**

Figura: Confusion Matrix

k-Fold Cross Validation

Per valutare la bontà di ciascun classificatore:

- **A priori**: Stimiamo l'Accuracy dividendo il Training Set in k parti applicando Cross Validation.



- **A posteriori**: Valutiamo l'Accuracy della previsione sul Test Set.

Logistic Regression Classifier

Modelliaziamo $logit(p_i) = log(\frac{p_i}{1-p_i}) = b + w^T x$:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.1444     0.4147  -7.582 3.41e-14 ***
EFG_O         1.8062     0.3466   5.211 1.88e-07 ***
EFG_D        -1.0992     0.2967  -3.705 0.000211 ***
TOR           0.9516     0.2666   3.570 0.000358 ***
TORD         -0.8813     0.2936  -3.002 0.002682 **
ORB           1.0662     0.2800   3.808 0.000140 ***
DRB           0.1236     0.2744   0.450 0.652450
FTR           0.4952     0.2376   2.084 0.037166 *
FTRD         -0.6634     0.3040  -2.182 0.029106 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 274.59  on 279  degrees of freedom
Residual deviance: 130.42  on 271  degrees of freedom
AIC: 148.42
```

Logistic Regression Classifier

$$CutOff \approx 0.25$$

$$Sensibilit\grave{a} = 0.9122807$$

$$Specificit\grave{a} = 0.9285714$$

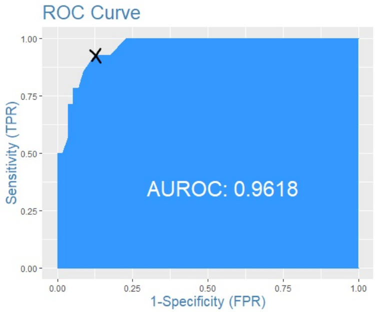
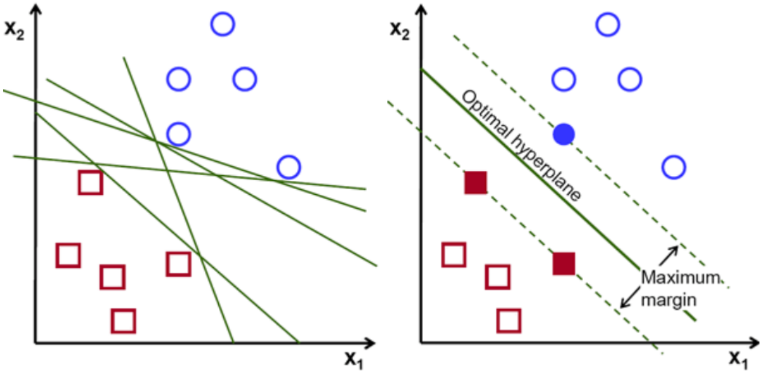


Figura: Receiver Operating Characteristic

Cerchiamo l'iperpiano separatore che massimizzi il margine geometrico



Support Vector Machine

Nell'ipotesi che i dati siano linearmente separabili:

$$\text{(P1)} \qquad \max_{\gamma,w,b} \quad \gamma$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma \qquad i = 1, ..., m$$
$$\|w\| = 1$$

sostituendo $\gamma = \hat{\gamma}/\|w\|$ otteniamo:

$$\text{(P2)} \qquad \min_{\gamma,w,b} \quad \frac{1}{2}\|w\|^2$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 \qquad i = 1, ..., m$$

risolviamo determinando il problema duale (condizioni di Karush Kuhn Tucker) e risolvendo attraverso metodi numerici (es. *Sequential Minimal Optimization*,J.Plat).

Support Vector Machine

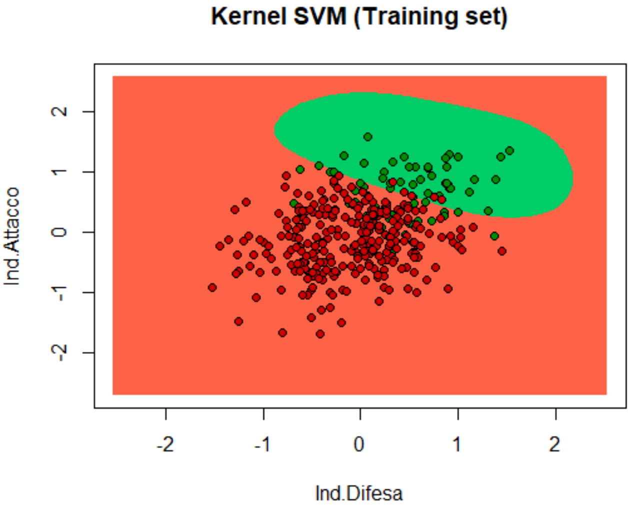
Nell'ipotesi che i dati non siano linearmente separabili:

$$
\begin{aligned}
\text{(P3)} \qquad \min_{\gamma,w,b} \qquad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i & \\
\text{s.t.} \qquad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i & i = 1, ..., m \\
& \xi_i \geq 0 & i = 1, ..., m
\end{aligned}
$$

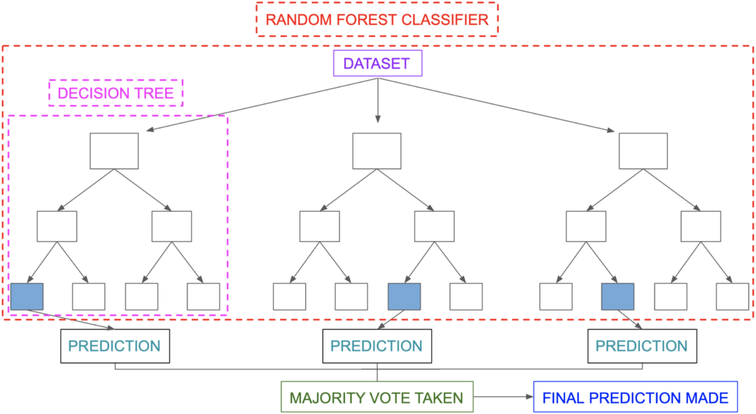
Kernel Trick

Feature Mapping: $K(x, z) = \phi(x)^T\phi(z)$
Attributes $\longrightarrow$ *Features* $\longrightarrow$ *Attributes*

Support Vector Machine



Random Forest (1995)



Risultati (Accuracy)

		Cross-Validation	Prediction
8 Features	LR	0.925873	0.915493
	SVM	0.9002381	0.8873239
	RF	0.9088095	0.9295775
2 Features	LR	0.934444	0.943662
	SVM	0.903015	0.8732394
	RF	0.8578571	0.9014085

Tabella: Accuracy

Risultati ($F1_{score}$)

		Cross-Validation	Prediction
8 Features	LR	0.9586419	0.9473684
	SVM	0.9376023	0.9344262
	RF	0.9120946	0.9401709
2 Features	LR	0.956389	0.943662
	SVM	0.9406801	0.9243697
	RF	0.9217391	0.9217391

Tabella: $F1_{score}$

Conclusioni

- Le statistiche qualitative di gioco sono buoni predittori della possibilità di qualificarsi alla March Madness.
- Le performance dei tre classificatori sono ottime e molto simili tra loro.
- La concordanza dell'accuracy stimata attraverso Cross-Validation e l'accuracy di predizione scartano la possibilità di overfitting.

Attenzione!

Machine Learning is not magic!

Cosa sarebbe successo se avessimo applicato un algoritmo di Machine Learning senza a priori un analisi esplorativa (preprocessing, feature selection,...)?

$$Accuracy < 80\%$$