MLT

Emanuele Noventa, Manuela Scioni, Enrico Tonini (C)

September 9, 2015

Dopo un'attenta esamina delle variabili presenti nel dataset si è notato che la principale variabile predittiva per la variabile `Risposta` sembra essere la variabile nominale `Articolo`: in particolare si è osservato che la non idoneità del vino è fortemente legata alla presenza della parola "ATTO" nella descrizione dell'articolo. Partendo da questa considerazione, è stata creata una variabile dicotomica, indicatrice della presenza del termine "ATTO" nella variabile `Articolo`.
Si è scelto di partire da un modello logistico completo con tutte le variabili esplicative che poteva aver senso inserire, insieme alle loro interazioni con la variabile indicatrice di cui sopra; scremando le variabili non significative si è giunti a un modello logistico finale di cui riportiamo di seguito l'output (ottenuto in R):

```
##
## Call:
## glm(formula = Risposta ~ Atto + Tirato_Imbottigliato, family = binomial,
##     data = ds1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9447  -0.0388  -0.0388  -0.0388   3.7927
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -8.2247     0.5174 -15.897  < 2e-16 ***
## AttoTRUE                      8.9189     0.4034  22.108  < 2e-16 ***
## Tirato_ImbottigliatoTIRATO    1.0331     0.3936   2.625  0.00866 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1973.35  on 11639  degrees of freedom
## Residual deviance:  340.07  on 11637  degrees of freedom
## AIC: 346.07
##
## Number of Fisher Scoring iterations: 10
```

Le uniche variabili esplicative risultate significative sono l'indicatrice dell'"ATTO" e se il prodotto è stato tirato o imbottigliato.
Si sono testate altre tecniche alternative al modello logistico quali le foreste casuali, gli alberi di classificazione, le reti neurali, le support vector machine, senza tuttavia riuscire a migliorare la capacità previsiva del modello logistico. La capacità previsiva è stata testata sia sul dataset messo a disposizione per la previsione (guardando il "punteggio"), sia sul dataset messo a disposizione per la stima, dividendolo casualmente in dataset di stima e dataset di verifica.
Di seguito si riporta anche la matrice di classificazione applicata ai dati di verifica:

```
##         osservati
## previsti IDONEO RIVEDIBILE
##    FALSE   5693          3
```

```
##    TRUE      21        103
## errore totale:  0.004123711
## falsi positivi & falsi negativi:  0.1693548387 0.0005266854
```

Punteggio ottenuto (sulle 5820 osservazioni del dataset di verifica): 201.