Antics of Statistics

Strategia

I vini rivedibili rappresentano solamente poco più dell'1% del training set. In un problema di classificazione così sbilanciata, per ogni modello, abbiamo deciso di sovra-campionare i casi rivedibili con un campionamento del training set che segue questo schema:

1. 500 ripetizioni;
2. estrazione casuale senza reinserimento del 50% dei casi rivedibili;
3. bagging (estrazione con reinserimento) di 1000 casi rivedibili precedentemente estratti (punto 2);
4. estrazione casuale senza reinserimento di 7000 casi idonei;
5. composizione del training set basata sulle osservazioni estratte al punto 3 e 4.

In questo sistema la dimensione delle osservazioni della popolazione che non vengono estratte (out of sample) è molto simile a quella dello scoring set (osservazioni da prevedere).

Per ognuna di queste 500 iterazioni, al training set così estratto, è stato applicato un modello gradient boosting tree (package R xgboost), pesando al 50% i casi idonei e al 50% i casi rivedibili. La previsione di ognuno di questi modelli è la media delle 500 previsioni (classi 1/0) di ogni singola iterazione.

Complessivamente, abbiamo scelto 3 modelli gradient boosting tree. Due di questi utilizzavano un dataset che definiva una classe della variabile "Articolo", costituita da tutte le categorie che avevano meno di 15 osservazioni e una media di casi rivedibili maggiore del 10%. L'altro modello non presentava questa categoria.

La stima finale si è basata su una media pesata del 50% per il modello

senza categoria e del 25% per ognuno dei due modelli con la categorizzazione descritta.
In generale, per i modelli gradient boosting tree, l'errore è stato definito come quello proposto dalla competizione. Lo shrinkage è stato posto pari a 0.2. Sono stati elaborati 500 alberi, limitati ognuno a 5 nodi, il numero minimo di osservazioni per ogni nodo è stato posto a 3 ed ogni albero era calcolato estraendo casualmente il 70% delle osservazioni ed il 70% delle variabili.
La classificazione finale tra idoneo e rivedibile si è basata su un cutoff determinato dal f-score pesando 60/61 l'inverso della Precision e 1/61 l'inverso della Recall. Questo per rappresentare meglio la metrica di accuratezza proposta nella competizione.
Durante la competizione abbiamo provato una selezione delle variabili più rilevanti, definendone un ranking attraverso la differenza con la devianza del modello nullo per ognuna di queste, che però non ha dato buoni risultati. È verosimile che esistano numerose interazioni tra i predittori, cosa che supporta la nostra scelta di utilizzare un modello ad albero.

Trattamento delle variabili

Per ridurre il rischio di overfitting, mantenendo i gradi di libertà del training set, si è deciso di trasformare le variabili categoriali in variabili ordinali (numeriche), ordinando le categorie in base alla media di casi rivedibili. In questo processo, sono state inoltre accorpate in una categoria "altro" tutte quelle categorie che presentavano meno di 15 osservazioni nel training set.

Variabili più discriminanti nella determinazione dei casi idonei/rivedibili

Variabile	Capacità discriminante
Articolo	0.747
Qta_effettiva	0.031
Anidride_Solforosa	0.027
Acidità_Volatile	0.020
Acidità_Totale	0.019
Titolo_AlcolometricoTotale	0.018
Sfuso	0.017
Titolo_Alcolometrico_Effettivo	0.017
Data	0.016
Estratto_Non_Riduttore	0.015

Principali interazioni

- Articolo – Acidità_Volatile – Data
- Articolo – Sfuso – Titolo_AlcolometricoTotale – Annata – Acidità_Totale
- Articolo – Titolo_AlcolometricoTotale – Zuccheri_Riduttori – Acidità_Volatile
- Articolo – Tipo – Titolo_AlcolometricoTotale – Annata – Acidità_Volatile
- Articolo – Sovrapressione – Acidità_Totale – Estratto_Non_Riduttore – Data