Report della squadra **"Completely Random Methods"**

SUS[4], Palermo, 19-06-2018

Marta Galvani, Chiara Ghiringhelli, Amir Khorrami Chokami, Giovanni Rebaudo e Tommaso Rigon

1 Operazioni preliminari

Complessivamente, l'azienda Findomestic ha predisposto un dataset di 66.521 osservazioni. Di queste, 40.000 sono utilizzate per la fase di stima mentre le rimanenti per la valutazione finale. L'obiettivo è prevedere la variabile qualitativa `ClientStatus`, composta da 3 categorie:

1. Cliente regolare (`Regolare`), circa il 96.1% del totale,
2. Cliente contenzioso (`Contenzioso`), circa il 3.1% del totale,
3. Cliente recupero (`Recupero`), circa il 0.8%, del totale.

Il dataset consta di 29 covariate. Molte di queste sono state riorganizzate per facilitare la procedura di stima. Le medesime operazioni sono state effettuate in maniera totalmente speculare anche nell'insieme di valutazione. In breve, le operazioni preliminari che sono state effettuate sono:

1. Tutte le variabili che nella documentazione ufficiale erano indicate come "qualitative" sono state codificate come tali.
2. Solamente due variabili presentavano valori mancanti: `ANZ_BAN` e `REGIONE`. In questi casi, i valori mancanti sono stati imputati con il valore mediano (`ANZ_BAN`) e con il valore modale (`REGIONE`). Ad ogni modo, il numero di valori mancanti era solo una piccola percentuale rispetto al totale dei dati.
3. Le categorie con bassa frequenze di alcune variabili qualitative sono state accorpate alla categoria modale. Ad esempio, i valori `X` in `STA_CIVILE`—ovvero stato civile sconosciuto—sono stati trasformati nel valore modale `C`, e pertanto trattati come se fossero valori mancanti.
4. Infine, si è considerata la trasformazione logaritmica di alcune variabili (e.g. `REDDITO_CLT`) per ridurre l'effetto dei valori anomali in fase di previsione.

2 Ottimizzazione della funzione di perdita

La matrice di perdita utilizzata per la valutazione penalizza fortemente i valori previsti come `Regolare`, che però appartengono in realtà alla categoria `Contenzioso`.

Per ridurre questo tipo di errore, evidentemente particolarmente oneroso per Findomestic, si è proceduto come segue: sia $\hat{\pi}_i = (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \hat{\pi}_{i3})$ un vettore di probabilità previste tramite un qualche metodo di classificazione per l'osservazione i-esima. Ciascuna osservazione è stata allocata aggiustando le probabilità $\hat{\pi}_i$ tramite dei pesi (c_1, c_2, c_3). Più precisamente, l'unità i-esima è stata classificata nella categoria tale che massimizza la probabilità riscalata, ovvero

$$\max\{c_1\hat{\pi}_{i1}, c_2\hat{\pi}_{i2}, c_3\hat{\pi}_{i3}\},$$

dove i pesi c_1, c_2, c_3 sono stati scelti tramite **convalida incrociata**.

Sebbene scelte opportune di questi pesi c_1, c_2, c_3 consentano di minimizzare efficacemente la funzione di perdita, ci preme far notare che la classificazione da noi ottenuta tende ad allocare la maggior parte delle osservazioni nella classe `Recupero`, essendo quella meno penalizzata. Sarà quindi cura di Findomestic valutare se questo comportamento è auspicabile o meno ed è, fondamentalmente, una diretta conseguenza della matrice di perdita scelta per la valutazione della competizione.

L'algoritmo utilizzato per prevedere le probabilità $(\hat{\pi}_{i1}, \hat{\pi}_{i2}, \hat{\pi}_{i3})$ è una random Forest, anche se una semplice LDA consente di ottenere—stando alle nostre valutazioni— performance analoghe anche se leggermente inferiori.

Per quanto consapevoli che questo non aumenterà in nessuna maniera le nostre possibilità di vittoria, alleghiamo una nostra fotografia, a testimonianza del fatto che ci siamo divertiti :-)

I Completely Random Methods