SUS^4 - The Markovians

Alessandro Arrigo Chiara Di Maria Stefania Martello Claudio Rubino

1 Obiettivo dell'analisi

L'azienda Findomestic dispone di un dataset composto da circa 67.000 osservazioni. Ogni osservazione rappresenta un cliente a cui è stata approvata una richiesta di finanziamento. Per 40.000 di essi è noto lo stato di pagamento, variabile denotata con *ClientStatus*, che presenta tre modalità: cliente regolare, cliente in recupero e cliente in contenzioso.

L'obiettivo dell'analisi è stabilire il valore di questa variabile, ignota per le restanti osservazioni, in base a trenta variabili esplicative che forniscono informazioni sui clienti. In altre parole, si vuole predire se un cliente sarà insolvente (in contenzioso), parzialmente insolvente (in recupero) o regolare sulla base di caratteristiche socio-demografiche, relative ad equipaggiamento del cliente, storico del cliente e comportamento del cliente.

2 Analisi esplorativa e Metodologia

L'analisi è stata svolta tramite il software statistico R, sul dataset di 40.000 osservazioni per cui è noto *ClientStatus*. Tale dataset, identificato da ora in poi col nome *Training*, presenta in totale 30 NA, di cui 3 nella variabile *REGIONE*, relativa alla regione di residenza del cliente, e 27 in *ANZ_BAN*, l'anzianità del conto corrente espressa in anni. È stata quindi effettuata un'imputazione: ai tre soggetti per cui *REGIONE* risulta mancante è stata attribuita la regione moda, ovvero il Trentino Alto-Adige (17). I *missing* in *ANZ_BAN* sono stati, invece, sostituiti con la mediana di tale variabile. Lo stesso è stato fatto sul secondo dataset in cui la variabile risposta è ignota, da ora in poi denominato *Validation*.

In entrambi i dataset le variabili regione, nazione di nascita, professione, canale di finanziamento, score comportamentale credit bureau sono state trasformate in fattori, data la loro natura qualitativa. La variabile risposta, le cui modalità sono codificate con 0 (cliente regolare), 1 (cliente con contenzioso), 2 (cliente in recupero) è stata anch'essa trattata come categoriale.

Essa risulta fortemente sbilanciata, come si può notare dalla Tabella 1. Quasi tutti i clienti risultano regolari, meno del 5% insolventi e appena lo 0,8% risulta in recupero.

In fase modellistica si è tenuto conto di tale problematica, identificata in letteratura come una potenziale fonte di errori di previsione.

Tabella 1: Descrizione sintetica di *ClientStatus* in Training.

Regolare	In contenzioso	In recupero
38.442	1.254	304
96,1%	3,1%	0,8%

È stata implementata una procedura di random forest di classificazione per la variabile dipendente *ClientStatus*. Tale metodo consiste nel ridurre la varianza delle previsioni mediante l'utilizzo di più alberi decisionali; ciascuno di essi è costruito utilizzando un diverso campione bootstrap come training set, circa due terzi delle osservazioni, e il terzo rimanente (*Out of Bag*, OOB) è utilizzato come test set per calcolare l'Out of Bag Error Rate , ovvero il tasso di errata classificazione. Dato che gli alberi sono costruiti in modo da essere quanto più profondi possibile, ogni albero ha un *bias* basso e il problema dell'alta variabilità delle previsioni viene risolto utilizzando come classe prevista quella che è stata ottenuta il maggior numero di volte considerando tutti gli alberi della foresta. Inoltre, ad ogni *split* viene selezionato casualmente un sottocampione di m<p predittori, in modo da rendere meno simili tra loro gli alberi della foresta e, di conseguenza, ridurre il problema dell'*overfitting*. Nell'analisi è stata utilizzata la regola empirica che consiste nello scegliere $m = \sqrt{p}$ predittori ad ogni split (James et al, 2013). L'algoritmo della random forest, pur portando a un notevole miglioramento nell'accuratezza delle previsioni rispetto all'utilizzo di un albero singolo, ha lo svantaggio di non essere facilmente interpretabile, non essendo più rappresentabile una chiara regola decisionale. Per ovviare a tale problema è possibile sintetizzare l'importanza di ciascuna variabile esplicativa all'interno del processo decisionale sommando i decrementi nell'indice di impurità di Gini occorsi ogni volta che una certa variabile è stata usata in uno split. L'entità di tale misura (*importance*) per ciascuna variabile è mostrata nel grafico in Figura 1. Le tre variabili più importanti sono regione, reddito richiedente e reddito familiare.

Considerata la natura del problema in esame, gli errori di classificazione hanno diversi livelli di gravità. In particolare, il più grave consiste nel classificare un cliente in contenzioso come regolare. Per tenere conto dei diversi costi di errata classificazione si è optato per un campionamento stratificato all'interno della procedura di Random Forest. Ogni strato è una categoria di *ClientStatus* e per le categorie 1 e 2 è stata scelta una dimensione campionaria maggiore rispetto alla categoria 0, in modo da compensare lo sbilanciamento nella risposta e contemporaneamente tenere conto dei costi. Tale modello, per costruzione, restituisce una previsione in cui nessun cliente in contenzioso risulta regolare, ma, al contempo,

presenta un elevato tasso di errata classificazione dei clienti regolari, errori che, tuttavia, avendo un peso minore, non influenzano fortemente la previsione finale.

Un altro modello è stato stimato, senza considerare costi di errata classificazione differenti. Questo ha consentito di ottenere un errore di classificazione complessivo più basso rispetto al modello precedente, sebbene gli errori gravi sopracitati siano questa volta presenti.

In Figura 2 è rappresentata la stabilizzazione dell'OOB error rate per il secondo modello.

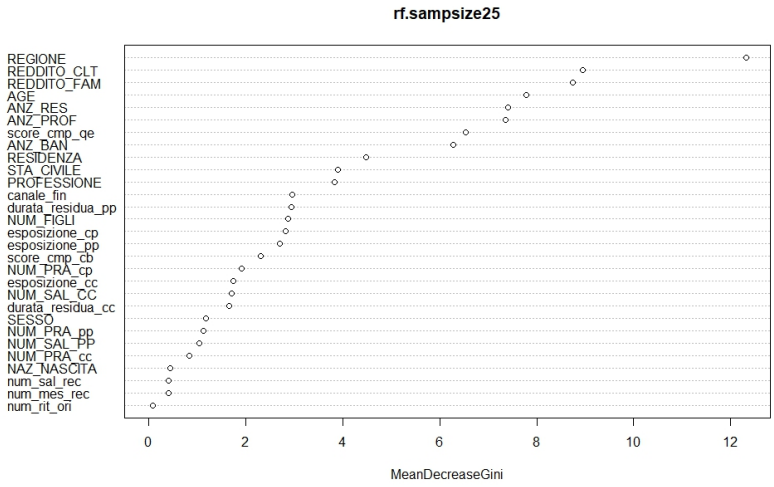

Figura 1: Importanza delle variabili.

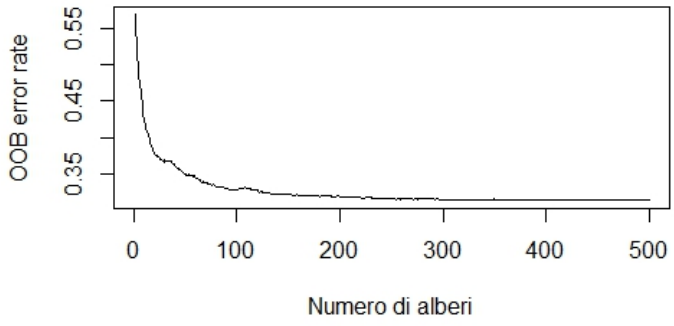

Figura 2: Stabilizzazione della procedura.