Analisi di un insieme di dati di Findomestic

Rosti Ma Robusti*

28 Giugno 2017

1 Introduzione al problema

Il problema di interesse riguarda l'ottimizzazione nella scelta dei clienti a cui sottoporre proposte telefoniche.

In particolare, si vuole prevedere con maggior precisione quali clienti hanno più probabilità di rispondere affermativamente alla proposta di finanziamento dell'azienda. Per la risoluzione del problema si hanno a disposizione 40000 osservazioni supervisionate, ovvero le informazioni di 40000 clienti (rappresentativi dell'intero portafoglio clienti) a cui si è già presentata la proposta e di cui si conosce la risposta. Le informazioni disponibili per ogni cliente sono riconducibili a 3 macro categorie:

- variabili sociodemografiche
- variabili di equipaggiamento
- variabili storico-comportamentali

1.1 Esplorazione Dataset

La variabile di interesse è definita come "TARG_TOT" ed è descivibile come una categoriale binaria. Inizialmente l'insieme di dati presenta quindi 40000 unità statistiche (righe) e 73 variabili (colonne) oltre alla variabile risposta. Di queste 68 sono inizialmente valutate come variabili quantitative e le restanti 5 come categoriali.

La variabile "codice.identificativo" considera un codice di identificazione per ogni cliente e quindi per ogni singola osservazione, pertanto non è assolutamente informativa per la previsione e non viene perciò considerata nella costruzione dei modelli. Le variabili "NUM_PPQ_C", "CC_C_NUM", "PPQ_18_NUM_PRA", "PPQ_18_IMP_FIN", "CC_18_NUM" rappresentano dei totali di altre variabili, risultandone quindi collineari. Pertanto, poichè l'informazione contenuta in queste covariate è equivalente a quella presente in altre covariate, si è deciso di non considerarle.

Per alcuni variabili quantitative di conteggio definite da un massimo di 6 valori, si è preferito ricodificarle come variabili qualitative per garantire una maggior flessibilità senza perdere troppo in parametrizzazione. Correlato a ciò, nonostante non fossero presenti variabili qualitative con dimensione particolarmente elevata (ovvero con molte modalità), si è studiato come accorpare le modalità con frequenze molto basse. Questa scelta è conseguenza sia di una valutazione di possibili problemi computazionali, sia di valutazioni più puramente statistiche. Infatti, se una modalità risulta poco presente nell'insieme di dati, non è più permesso suddividere il dataset in insiemi di stima e di verifica, perchè potrebbe risultare presente solo nel secondo di essi e pertanto rendere impossibile la previsione. Inoltre una modalità che caraterizza pochissime osservazioni non può essere determinante nel discriminare gli 1 dagli 0, o meglio non si è ritenuto sufficientemente alto, tale da giustificare l'aumento dell'onere computazionale, il contenuto informativo presente nelle modalità raggruppate. Si è pertanto scelto di classificare come un'unica modalità 'altro' tutte le categorie con una frequenza assoluta nell'insieme di stima inferiore a 500.

A seguito di una veloce analisi preliminare si osserva come possa risultare utile una trasformazione quadratica dell'età, elemento che trova spesso riscontro in letteratura. Pertanto si è aggiunta la covariata "AGE2" che rappresenta il quadrato dell'età e che, al pari delle altre variabili, verrà valutata di volta in volta per ogni modello.

*Universitá degli Studi di Padova

1.2 Valori mancanti

Un'analisi preliminare dei dati mostra che sono presenti dei valori mancanti. Questi sono principalmente definiti come NA, ma per le variabili "ANZ_PROF" e "ANZ_BAN" si deve considerare alla stregua di valori mancanti anche i valori "98" e "99" in quanto rappresentano una codifica manuale di informazioni sconosciute per l'operatore che ha inserito i dati. Questa operazione è particolarmente importante poichè altrimenti quei valori rappresenterebbero degli outliers per variabili che misurano in anni l'anzianità delle relazioni tra il cliente e le banche.

L'imputazione degli NA è stata differenziata a seconda della codifica della variabile. Si è provveduto ad aggiungere una modalità denominata 'n.a.' in tutte le variabili categoriali con valori mancanti. Per le variabili numeriche si è preferito adottare un metodo di imputazione che sostituisse i valori mancanti con delle previsioni basate sulle altre covariate. Si è selezionato un sottoinsieme dei dati con osservazioni complete (4000 osservazioni corrispondenti al 10% dei dati) sul quale si è fatto crescere un albero di regressione usando come variabile risposta una delle variabili quantitative con valori mancanti e come esplicative tutte le altre variabili dell'insieme di dati (esclusa la variabile risposta). L'albero è stato poi potato attraverso una procedura automatica basata su convalida incrociata. Questa procedura è stata iterata per ogni variabile numerica con valori mancanti. In ogni cella vuota e' stata quindi inserita la previsione dell'albero dati i valori delle altre variabili, alla quale si è aggiunta una distorsione normale di media zero e varianza pari alla varianza della foglia da cui si è tratta la previsione. L'insieme di dati completi utilizzato per la stima degli alberi e' stato quindi escluso dalle successive operazioni di analisi dati. Questo è stato fatto anche per variabili con una quota di NA molto alta, poichè su un insieme di dati così sbilanciato potrebbe esser molto informativa proprio la presenza di valori in contesti dove vi sono quasi solo NA. Per tener conto di questa informazione sono state inserite nel dataset variabili dummy che indicassero la presenza di NA per ogni variabile numerica che li conteneva prima delle procedure di imputazione.

La matrice di dati così modificata conta 36000 righe e 73 colonne, per un totale di 52 variabili numeriche e 20 variabili categoriali, oltre alla variabile risposta.

1.3 Divisione del dataset

Data la grande mole di osservazioni a disposizione si è stabilito di dividere la matrice di dati in due insiemi disgiunti, denominati di *stima* e di *verifica*, rispettivamente contenenti il 75% e il 25% dei dati rimanenti (ovvero 36000 osservazioni tolta l'imputazione degli NA). Il primo insieme è stato impiegato per la stima dei diversi modelli ed in particolare è stato ulteriormente suddiviso in due parti contenenti il 50% e il 25% delle osservazioni totali. La prima parte, che definiremo di *apprendimento*, è stata utilizzata per la stima del modello; la seconda, definita come insieme di *convalida*, ha permesso la determinazione dei parametri di regolazione o la selezione del miglior modello in una classe di modelli simili (ad esempio, per selezionare il miglior modello lineare tra tutti i modelli lineari provati). Dopo aver scelto il parametro di regolazione o le variabili di interesse per il modello con l'approccio stima e verifica nelle parti apprendimento-convalida, si è stimato nuovamente il modello finale (quindi con il parametro di regolazione fissato) su tutto l'insieme di stima.

Questa procedura è stata evitata nei metodi di regolarizzazione e negli alberi di regressione e MARS, in quanto il parametro di regolazione influenza maggiormente e più direttamente la struttura del modello stimato, che risulterebbe quindi instabile rispetto ad una nuova stima su un numero maggiore di dati (si pensi a come la scelta della profondità di potatura sia strettamente legata alla struttura dell'albero). In questi modelli si è quindi preferito stimare il modello finale, dopo aver fissato il parametro di regolazione e/o la profondità di potatura, solo sulla parte di apprendimento dell'insieme di stima. Infine l'insieme di verifica è stato utilizzato per il confronto di modelli appartenenti a classi diverse (ad esempio, per confrontare il miglior modello additivo con il miglior albero) e per la valutazione finale dell'errore di previsione stimato. Si noti che per alcuni modelli dove non era necessario individuare parametri di regolazione o compiere altri confronti (come il modello lineare completo o il GAM completo), l'insieme di convalida e di apprendimento hanno formato un unico insieme di stima contenente il 75% dei dati, utilizzato interamente nella stima del modello.

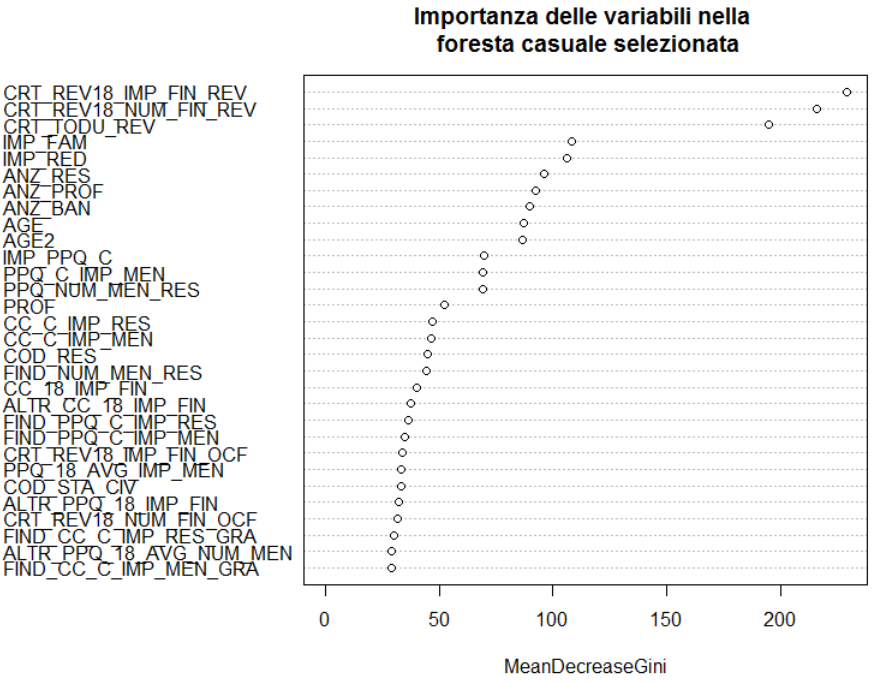
Si è osservato come le frequenze delle modalità in cui è suddivisia la variabile risposta sono fortemente sbilanciate. In particolare la modalità indicata come 1, compare nel 5.88% delle osservazioni. Ciò nonostante non si è ritenuto necessario bilanciare il campione o provvedere in altre forme, poiché l'interesse è principalmente rivolto all'ordinamento delle probabilità previste. In questo modo si è evitato di distorcere l'informazione originara cambiando il dataset di stima e ciò si è rivelato come una scelta produttiva anche al momento della valutazione numerica.

2 Modello

Per prima cosa si è definita la funzione obiettivo su cui ottimizzare i modelli di interesse. La funzione calcola la quota di osservazioni non presenti tra le n con le previsioni più alte, ma appartenenti alle n osservazioni con risposta affermativa dell'insieme di *verifica*. In questo modo si definisce una misura d'errore, che pertanto deve essere minimizzata. Tra i differenti modelli stimati abbiamo selezionato nell'insieme di *convalida* i migliori per ogni famiglia, i quali sono stati a loro volta confrontati nell'insieme di *verifica*.

Tra questi è risultato con le migliori performance la **Foresta Casuale**. Pertanto infine, si è deciso di ristimarlo su tutto l'insieme di *stima*, scegliendo il parametro di regolazione nell'insieme di *verifica*.

Il modello è definito come combinatore di 500 (nell'utilizzo in questione) alberi di classificazione, ciò permette un miglior potere previsivo in un insieme di dati non supervisionato, poichè mantiene limitato il sovradattamento. Questo ha, tuttavia, un lato negativo: ovvero una scarsa interpretabilità delle variabili in gioco. Ciò nonostante è possibilie ottenere una misura dell'importanza delle variabili valutando la loro efficacia previsiva: si confrontano i risultati del *modello completo* con quello che considera un contributo nullo della variabile in oggetto.



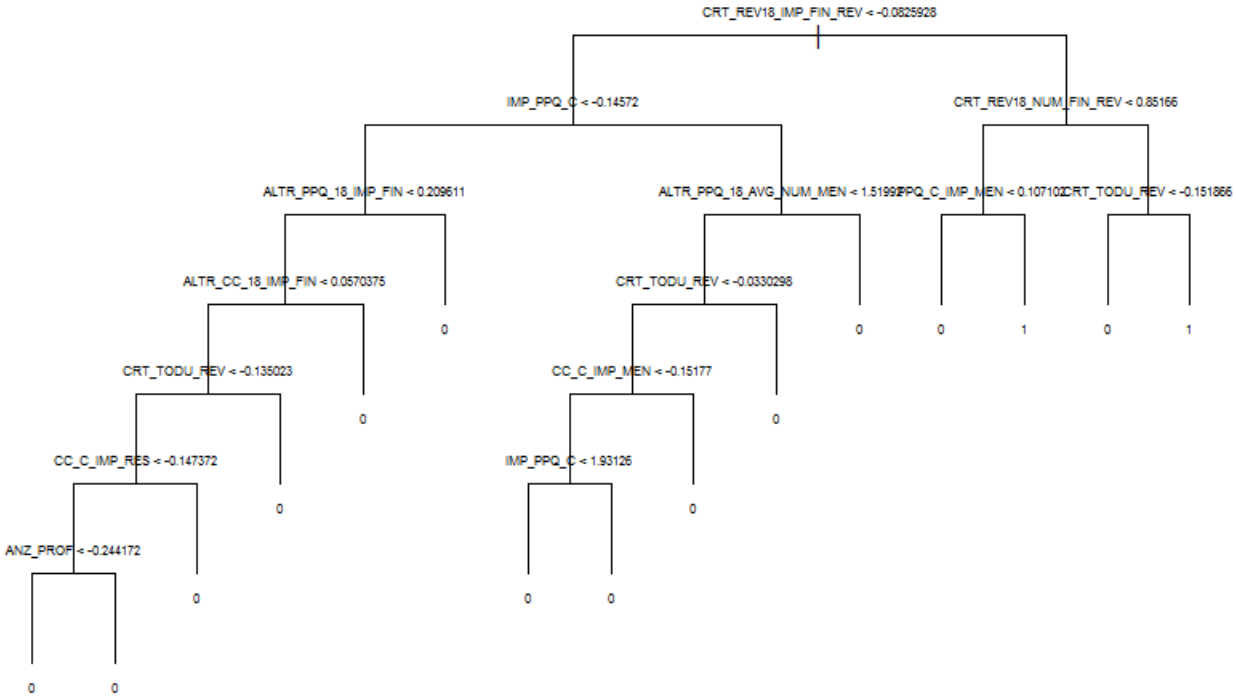
Le variabili più importanti per prevedere le persone che risponderanno affermativamente alla proposta telefonica risultano essre:

- CRT_REV_18_NUM_FIN_REV
- CRT_REV_18_IMP_FIN_REV
- CRT_TODU_REV

Tutte e tre le variabili sono relative all'utilizzo della carta di credito, che pertanto sembra giocare un ruolo fondamentale nella segmentazione dei clienti per il criterio desiderato.

Non potendo ottenere dal modello, per costruzione, un' indicazione sul tipo di relazione che hanno queste variabili con la variabile risposta, si mostrano i risultati relativi ad un altro modello che permettono di visualizzare facilmente la direzione di tali relazioni. Il modello esplicativo scelto è un albero di classificazione, poichè permette una facile interpretazione ed è fortemente legato alla Foresta Casuale.

In particolare emerge da una prima analisi grafica come i clienti di interesse, ovvero quelli con risposta affermativa, siano caratterizzati da valori alti delle variabili elencate in precedenza.


CRT_REV18_IMP_FIN_REV < -0.0825928
IMP_PPQ_C < -0.14572
CRT_REV18_NUM_FIN_REV < 0.85166
ALTR_PPQ_18_IMP_FIN < 0.209611
ALTR_PPQ_18_AVG_NUM_MEN < 1.51992
PPQ_C_IMP_MEN < 0.107102
CRT_TODU_REV < -0.151866
ALTR_CC_18_IMP_FIN < 0.0570375
CRT_TODU_REV < -0.0330298
CRT_TODU_REV < -0.135023
CC_C_IMP_MEN < -0.15177
CC_C_IMP_RES < -0.147372
IMP_PPQ_C < 1.93126
ANZ_PROF < -0.244172
0
1