Stats Under the Stars 3

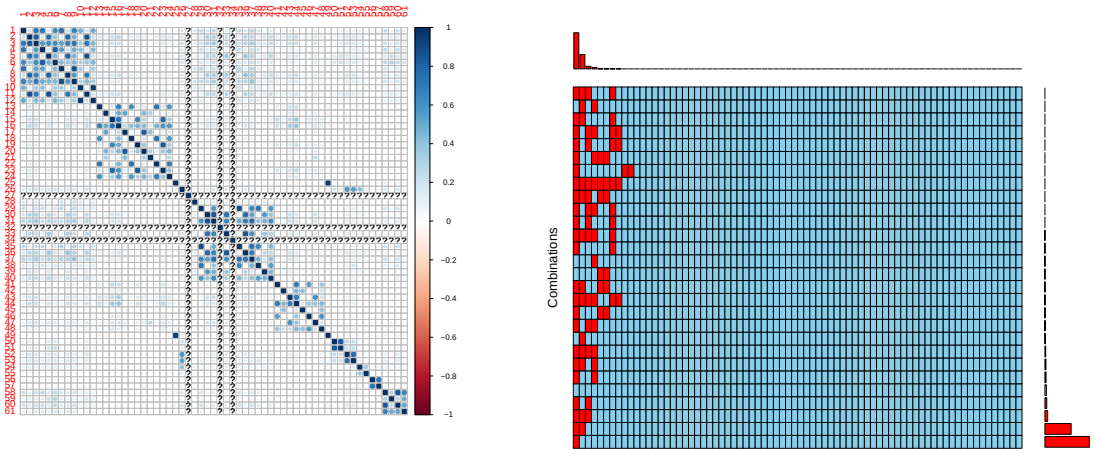
Bertarelli, G. Cappozzo, A. Cavicchia, C. Denti, F.

27-28 Giugno 2017

1 Obiettivo e Dati

Findomestic è una banca che opera nel Credito al Consumo delle famiglie. Attraverso il canale B2C, Findomestic colloca i suoi prodotti di Prestito Personale. Tramite un'indagine preliminare su un campione di n=40000 nominativi estratti casualmente da una popolazione di N=200000 clienti, si vuole creare uno scoring di propensione all'accettazione dei prodotti di Prestito Personale da applicare ai clienti restanti in modo da selezionare i 10000 clienti con la più alta probabilità di accettare la proposta di Prestito Personale.

Oltre alla variabile di interesse TARG TOT, dicotomica, rappresentante l'accettazione del Prestito Personale, i dati forniti presentano 73 variabili che possono essere ricondotte a tre macro categorie: variabili socio-demografiche, variabili di equipaggiamento e variabili storico-comportamentali. Una prima importante informazione relativa alla variabile TARG TOT nel dataset di stima è la bassa frequenza di accettazione del Prestito Personale telefonico sul numero di osservazioni complessivo (0.06% (2353 osservazioni)). In presenza di una distribuzione della variabile di risposta estremamente sbilanciata il processo di apprendimento può essere distorto, perché il modello tende a focalizzarsi sulla classe prevalente e ignorare gli eventi rari, nonostante la classe minoritaria di solito rappresenti il concetto di interesse, come nel caso in esame. La regressione logistica, per esempio, nota come uno dei metodi parametrici tradizionali più utilizzati per



(a) Correlazione fra i dati (b) Missing pattern

Figura 1: Descrittive Dati

la classificazione binaria, non è consigliabile quando le classi sono sbilanciate, perché la probabilità condizionale della classe rara è sottostimata. Si è scelto dunque di utilizzare SMOTE (Synthetic Minority Over-sampling Technique), un metodo di campionamento consistente nella modifica di un set di dati sbilanciati, in modo da fornire una distribuzione equilibrata tramite un lavoro di pre-processing sui dati. Si è dunque prodotto un nuovo dataset training di 96473 osservazioni in cui la variabile d'interesse ha una frequenza del 0.51 %. Il secondo problema affrontato è stato relativo alla selezione delle variabili da utilizzare nel modello. Si nota infatti una correlazione a blocchi tra alcune variabili (fig. 1a). A seguito di questa osservazione si è quindi ridotto a 39 il numero di covariate considerate nel modello finale. Infine, la figura 1b, mostra tutte le esistenti combinazioni di dati mancanti (rosso) e osservati (azzurro). Inoltre, le frequenze delle differenze combinazioni sono visualizzate tramite un piccolo diagramma a barre sulla destra del grafico. Si è quindi proceduto all'imputazione dei pattern mancanti tramite il pacchetto mice di R.

2 Risultati

Sul dataset nel quale sono state applicate le operazioni di preprocessing precedentemente descritte, si è poi scelto di applicare un classico modello logistico. Uno dei punti di forza della metodologia è l'interpretabilità dei risultati immediata che si andrebbe a perdere con metodi più complessi. Fra le variabili più significative l'età e il totale dei prestiti saldati negli ultimi 18 mesi. Poiché diversi autori concordano sul fatto che il random oversampling può aumentare la probabilità che si verifichino problemi di overfitting, è stato utilizzato anche un modello logistico modificato che tenesse conto della peculiare situazione in cui i successi osservati fossero in proporzione di molto minore rispetto agli insuccessi. Nel dettaglio, è stato utilizzato un Rare Events Logistic Regression for Dichotomous Dependent Variables (relogit). Nonostante ciò la proporzione di clienti correttamente classificati è rimasta pressoché invariata e si è dunque scelto di mantenere il modello sui dati soggetti a SMOTE.

Bibliografia Essenziale

Buuren, Stef, and Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R." Journal of statistical software 45.3 (2011).

Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

Imai, Kosuke, Gary King, and Olivia Lau. "Zelig: Everyone's statistical software." R package version 3.5 (2009).