Stats Under the Stars 3

Pierfrancesco Alaimo Di Loro Riccardo Giubilei
Hugo Maldini Marco Mingione Giuseppe Serra

28 giugno 2017

Sommario

Findomestic, banca che opera nel campo del credito al consumo delle famiglie, è interessata a individuare i clienti che con maggiore probabilità potrebbero sottoscrivere uno dei prodotti di prestito personale. Ciò permetterebbe di contattare solo tali clienti, in maniera tale da ottimizzare i costi di pubblicizzazione delle offerte.

A tale scopo, è stato messo a disposizione un dataset composto da 72 variabili che possono essere utilizzate per stimare la probabilità di sottoscrizione. Le unità che compongono il dataset di *train* sono 40000, mentre quelle che costituiscono il dataset di *test* sono 160000. L'obiettivo è individuare, tra le unità appartenenti al dataset di *test*, le 10000 che hanno maggiore probabilità di accettare una delle proposte di prestito personale.

La presente analisi utilizza un metodo *ensemble* composto da una *Random Forest*, un *Extreme Gradient Boosting* e una *Neural Network*. Tale metodo ha permesso di raggiungere uno score di 52.75% nella classifica parziale.

1 I dati

I dataset utilizzati per l'analisi sono due: uno di *train*, composto da 40000 osservazioni e corredato della variabile *target*, e uno di *test*, costituito da 160000 osservazioni per le quali la *target* non è nota. Ognuno di essi presenta 72 variabili (più una variabile corrispondente al codice identificativo del cliente) che possono essere utilizzate come predittori per la variabile risposta binaria, i cui livelli sono *sottoscrive* e *non sottoscrive* (codificati rispettivamente con 1 e 0). Il dataset di *train* è sbilanciato rispetto alla variabile risposta, in quanto contiene 37647 clienti che non hanno sottoscritto una proposta di prestito personale e soli 2353 che invece lo hanno fatto (rispettivamente il 94.12% ed l'5.88%).

Le variabili esplicative possono essere ricondotte a tre macro categorie:

- variabili socio-demografiche
- variabili di equipaggiamento
- variabili storico-comportamentali

2 L'analisi

2.1 Preprocessing

Innanzitutto si è effettuato un controllo di coerenza delle variabili e sono stati considerati come *missing values* i valori che presentavano incongruenze. In particolare, ciò è stato riscontrato per le variabili ANZ_BAN, ANZ_RES e ANZ_PROF. I valori mancanti così ottenuti sono stati successivamente imputati tramite il valore medio.

La procedura di imputazione è stata effettuata anche per le variabili che presentavano valori mancanti nel dataset originari. Nella fattispecie, i *missing values* delle variabili IMP_RED e IMP_FAM sono stati sostituiti con le rispettive medie. Al contrario, quelli in corrispondenza delle variabili PPQ_18_IMP_FIN, PPQ_NUM_MEN_RES, FIND_PPQ18_IMP_FIN e FIND_NUM_MEN_RES sono stati imputati con la mediana, in quanto le rispettive distribuzioni risultavano essere particolarmente asimmetriche.

Si è inoltre notata un'incongruenza in relazione alle variabili CRT_REV18_NUM_FIN_REV e CRT_REV18_NUM_FIN_OCF, che risultavano avere valori negativi, i quali sono stati sostituiti con il valore 0.

Successivamente, si è deciso di ricodificare le modalità della variabile FIND_PPQ18SS_MONTH_DAT_MAX_FIN, che risultava avere 39134 valori mancanti. Tali valori mancanti sono stati posti pari a 1, e tutti le altre modalità pari a 0 sulla base della distribuzione condizionata di tali modalità rispetto alla variabile *target*.

Inoltre, si è creata una nuova variabile che conta il numero di zero in corrispondenza di ogni unità.

Infine, si è proceduti alla simmetrizzazione e standardizzazione dei dati.

2.2 I modelli

2.2.1 Random Forest

L'algoritmo *Random Forest* consiste nel *bagging* di alberi decisionali in cui ogni split è effettuato utilizzando solo un sottoinsieme casuale delle variabili a disposizione. I parametri ottimali dell'algoritmo, ottenuti tramite CV sul dataset di *train*, sono riportati di seguito:

- **Numero massimo di alberi**: 200;
- **Criterio per effettuare gli split**: Gini;
- **Variabili considerate ad ogni split**: radice quadrata del numero di variabili originali;
- **Profondità massima del singolo albero**: 50;
- **Numero minimo di unità in una singola foglia**: 5.

2.2.2 Neural Network

Una *Neural Network* (o Rete Neurale) è un algoritmo di *Machine Learning* che si fonda sulla modellizzazione matematica dei processi logici che avvengono all'interno del nostro cervello. In particolare, le *Feed-Forward Neural Networks* sono delle reti che apprendono dai dati mediante una procedura nota come *Back-Propagation*. I parametri selezionati per l'implementazione del modello sono i seguenti:

- **Numero di *layers***: 2;
- **Numero di neuroni**: 100 nel primo strato, 50 nel secondo;
- **Funzione di attivazione**: funzione logistica.

2.2.3 Extreme Gradient Boosting

L'algoritmo *Extreme Gradient Boosting* è un metodo ensemble sequenziale che prevede, ad ogni step, la minimizzazione di una funzione di perdita. Il modello è stato implementato utilizzando alberi decisionali come *base learners* e impostando i relativi parametri nel seguente modo:

- **Numero di alberi**: 2000;
- ***Learning rate***: 0.01;
- **Profondità del singolo albero**: 3.

2.2.4 Metodo ensemble

I metodi *ensemble* sintetizzano i risultati di due o più algoritmi diversi col fine di ottenere modelli più robusti, stabili e accurati. Nel caso in analisi, la stima finale delle probabilità di sottoscrizione è stata ottenuta combinando i risultati degli algoritmi descritti nelle sezioni precedenti attraverso la media aritmetica. Poiché è stato questo il modello selezionato per la sottomissione dei risultati, le probabilità stimate sul *test* sono state ottenute mediante la combinazione dei suddetti modelli dopo essere stati ristimati sull'intero dataset di *train*, in modo da sfruttare integralmente l'informazione presente nei dati.