Insegnamento di Analisi dei dati (Data mining)
Prova d'esame del 4 luglio 2016 - parte pratica

Daniele Cugnigni

2023-02-21

Testo d'esame

Nel dataset frodi.csv sono presenti 91557 transazioni finanziarie, effettuate attraverso carte di credito, relative a un periodo di un mese. Oltre ai dati della transazione, sono già stati calcolati alcuni indicatori che esperienze passate hanno mostrato essere utili per identificare eventuali frodi. In particolare sono disponibili degli indicatori di anomalia della transazione, relativamente all'importo speso e a precedenti modi d'uso della medesima carta. Il contenuto specifico di ciascuno di questi indicatori non è disponibile per motivi legati alla proprietà intellettuale degli stessi.

L'obiettivo dell'analisi consiste nello scoprire le operazioni fraudolente in relazione alle caratteristiche delle transazioni, in modo da prevedere le prime in funzione delle seconde.

Il dataset è composto dai seguenti campi:

- Id: Identificativo della transazione
- Id_carta: Identificativo della carta di credito.
- Importo: L'importo della transazione.
- 9 indicatori di anomalia sull'importo della transazione. Esempi: confronto con il mese precedente, confronto con il semestre precedente.
- 8 indicatori di anomalia comportamentale. Esempi: anomalia rispetto ai posti in cui la carta ha operato, anomalia rispetto alla frequenza delle transazioni.
- 8 indicatori di confronto della carta con le carte ad essa simili.
- frode: variabile indicatrice. Assume valore 1 per le transazioni fraudolente; 0 per le transazioni non fraudolente.

Pulizia del dataset

Il file "frodi.csv" è composto da 91557 unità statistiche (le transazioni) sulle quali sono state rilevate complessivamente 30 variabili, con la variabile *Frode* che rappresenta la variabile risposta.
Prima di procedere all'analisi del dataset, è opportuno effettuare delle operazioni di pulizia. In primo luogo si nota la presenza delle variabili *Id* e *Id_carta*, le quali non sono altro che l'identificativo della transazione e l'identificativo della carta di credito e pertanto vengono escluse dall'analisi. Inoltre la variabile *ora_GMT* dà informazione su anno, mese, giorno ed ora della transazione: avendo il dataset composto da transazioni rilevate nell'arco di un mese, non è possibile valutare un eventuale effetto di anno, mese e giorno della transazione, mentre è possibile valutare l'effetto dell'ora. A tal proposito, sembra ragionevole focalizzare l'attenzione non tanto sull'ora esatta ma sul momento della giornata in cui avviene una transazione, di conseguenza si crea la variabile *momento* avente quattro modalità: *mattina* (6-12), *pomeriggio* (13-18), *sera* (19-23) e *notte* (0-5).

Successivamente si analizza l'eventuale presenza di valori mancanti nel dataset e si nota che il 24.74% delle osservazioni non ha un valore relativamente alla variabile *Anomaly_importo9*. Poichè si hanno a disposizione altri 8 indicatori di anomalia sull'importo delle transazioni, si decide di eliminare la variabile *Anomaly_importo9*.

In seguito a queste operazioni, il dataset è composto da 91557 unità statistiche e 27 variabili. A questo punto, prima di procedere con la modellazione dei dati:

- per tenere in considerazione il compromesso tra varianza e distorsione, si procede con la divisione del dataset in insieme di stima (75%) e insieme di verifica (25%), ottenendo un insieme di stima con 68668 osservazioni ed un insieme di verifica con 22889 osservazioni;
- si verifica se, nell'insieme di stima, le classi della variabile risposta siano bilanciate. A tal riguardo, si nota come le classi della variabile risposta risultano essere fortemente sbilanciate: 68582 osservazioni (99.87%) sono transazioni non fraudolente mentre 86 (0.13%) risultano essere fraudolente. Una possibile soluzione per tenere in considerazione questo aspetto è sottocampionare (senza reinserimento) le osservazioni che fanno riferimento ad operazioni non fraudolente. In questo caso, poichè il perfetto bilanciamento non è possibile in quanto porterebbe ad una perdita d'informazione troppo elevata, si decide di ricampionare 850 osservazioni relative ad operazioni fraudolente, in modo da avere un insieme di stima di 936 osservazioni e composto per il 10% da operazioni fraudolente e per il 90% da operazioni non fraudolente. E' importante far presente che questa scelta fa perdere molta dell'informazione a disposizione riguardo le operazioni non fraudolente ma, allo stesso tempo, permette di adattare i modelli su un dataset con classi meno sbilanciate della variabile risposta, portando quindi ad una maggiore attenzione, in fase di stima, verso le operazioni fraudolente;
- si verifica l'assenza di variabili esplicative degeneri (e quindi inutili per l'analisi) nell'insieme di stima meno sbilanciato, rilevando l'assenza della modalità "1" nella variabile dicotomica *Behaviour_Anomaly8*, che pertanto viene eliminata.

In seguito a queste operazioni, l'insieme di stima è composto da 936 osservazioni e 26 variabili, mentre l'insieme di verifica è composto da 22889 osservazioni e il medesimo numero di variabili.
Concluse le operazioni di pulizia del dataset, si può procedere con l'analisi esplorativa sull'insieme di stima.

Analisi esplorativa

Tenendo in considerazione che la variabile risposta è una variabile categoriale con due modalità e le variabili esplicative risultano essere in parte quantitative e in parte qualitative, un'analisi esplorativa (abbastanza) completa ed adeguata si avrebbe con l'analisi della distribuzione della variabile dipendente al variare delle singole variabili indipendenti. Poichè l'obiettivo primario non è quello di effettuare l'analisi esplorativa ma di adattare i modelli, si valuta la distribuzione della risposta solamente per alcune variabili esogene.

I barplot in Figura 1 danno indicazione di un possibile effetto significativo del momento della giornata in cui avviene la transazione (*momento*), dell'importo della transazione (*Importo*), del primo indicatore dell'anomalia comportamentale (*Behaviour_Anomaly1*) e del primo (*Population_Anomaly1*) e settimo (*Population_Anomaly7*) indicatore di confronto della carta con le carte ad essa simili. In particolare, si nota che la mattina e il pomeriggio vengono quasi esclusivamente compiute transazioni non fraudolente, mentre la sera e la notte la percentuale di operazioni fraudolente e non fraudolente è praticamente la stessa. Per quanto riguarda l'importo della transazione, emerge la quasi assenza di operazioni fraudolente per importi inferiori a €250. Infine, si nota un andamento decrescente della proporzione di operazioni fraudolente all'aumentare dei valori assunti dal primo e settimo indicatore di confronto della carta con le altre carte.

Conclusa l'analisi esplorativa nell'insieme di stima, si può procedere alla modellazione dei dati.

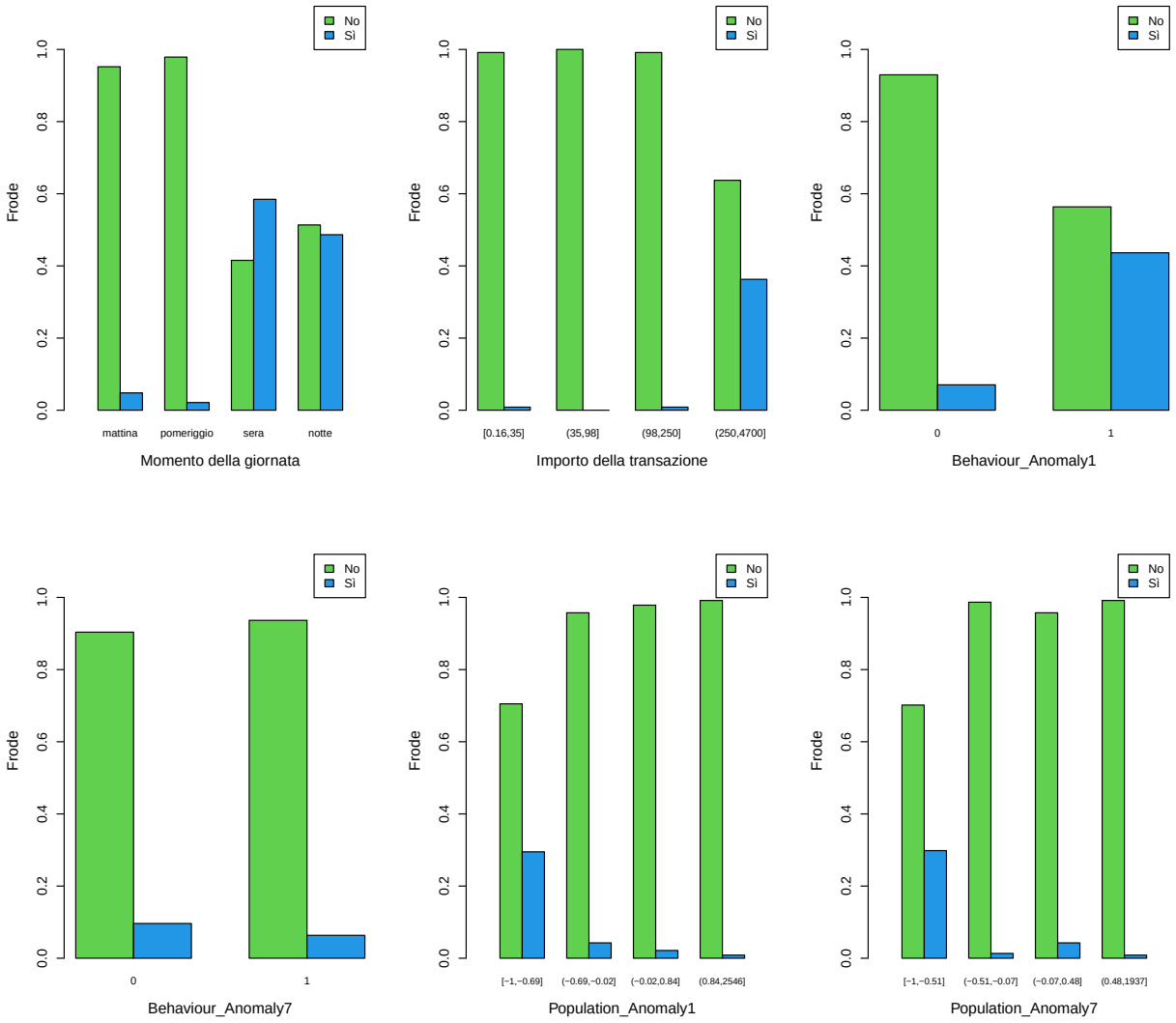

Figure 1: Barplot della variabile risposta rispetto ad alcune variabili esplicative

Modellazione dei dati

In questo contesto, è importante tenere conto del differente peso degli errori di previsione, in quanto è molto più grave prevedere come non fraudolenta un'operazione che lo è piuttosto che prevedere come fraudolenta un'operazione che non lo è. In altri termini, è più importante minimizzare il numero di falsi negativi rispetto al numero di falsi positivi, pertanto si fissa il valore della soglia pari a 0.10, ovvero la proporzione di operazioni fraudolente presenti nell'insieme di stima, e si valuterà la performance dei modelli sia in termini di tasso di errata classificazione sia in termini di percentuale di falsi negativi.

Modello logistico stepwise

Il primo modello che si adatta è il modello di regressione logistica stepwise basato sulla minimizzazione dell'AIC, con ricerca in entrambe le direzioni e a partire dal modello con la sola intercetta.

Nel modello finale sono incluse 14 delle 26 variabili esplicative: l'importo della transazione, il momento della giornata in cui avviene la transazione, cinque indicatori relativi all'anomalia dell'importo, tre indicatori relativi all'anomalia comportamentale e quattro indicatori di confronto della carta con carte ad essa simili. Ad un livello di significatività del 10%, gli effetti dell'importo della transazione e del secondo indicatore di anomalia sull'importo della transazione risultano non essere significativi.
Il tasso di errata classificazione e la percentuale di falsi negativi nell'insieme di verifica sono pari rispettivamente al **2.00%** e allo **0.00%**.

Albero di classificazione

Si prosegue la fase di modellazione con l'adattamento di un albero di classificazione, con l'entropia come funzione da minimizzare. Poichè questo modello prevede la selezione del numero di foglie ottimale, si fa crescere l'albero nell'insieme di stima e si effettua la fase di potatura tramite convalida incrociata con 5 fold. Nella fase di crescita dell'albero viene impostata una numerosità minima di osservazioni per foglia pari a 2 e una diminuzione dell'entropia per consentire uno split pari a 0.0000005, in modo da far diventare l'albero il più profondo possibile. Nella fase di potatura viene valutata la devianza in convalida incrociata al variare del numero di foglie dell'albero.
Il grafico in Figura 2 mostra come il minimo si ottenga con un albero con 11 foglie.

Uno dei pregi di questo modello è la facile interpretabilità nel caso in cui l'albero sia poco profondo. A tal riguardo, il grafico in Figura 3 mostra che l'importo, il momento della giornata e il quinto indicatore di anomalia sull'importo della transazione sono le variabili esplicative che determinano i primi split. E' comunque importante ricordare che con questo modello la misura di importanza delle variabili è condizionata alle suddivisioni effettuate in precedenza.

Nell'insieme di verifica il tasso di errata classificazione e la percentuale di falsi negativi risultano essere pari rispettivamente al **2.21%** e al **10.71%**.

Analisi discriminante lineare

Si adatta un modello di analisi discriminante lineare, utilizzando per la stima tutte le variabili esplicative disponibili. Il modello ottiene, nell'insieme di verifica, un tasso di errata classificazione pari all' **1.00%** e una percentuale di falsi negativi pari al **10.71%**.

Data la presenza di variabili qualitative, non risulta particolarmente sensato stimare un modello di analisi discriminante quadratica, in quanto quest'ultima si appoggia sull'ipotesi di normalità delle covariate.

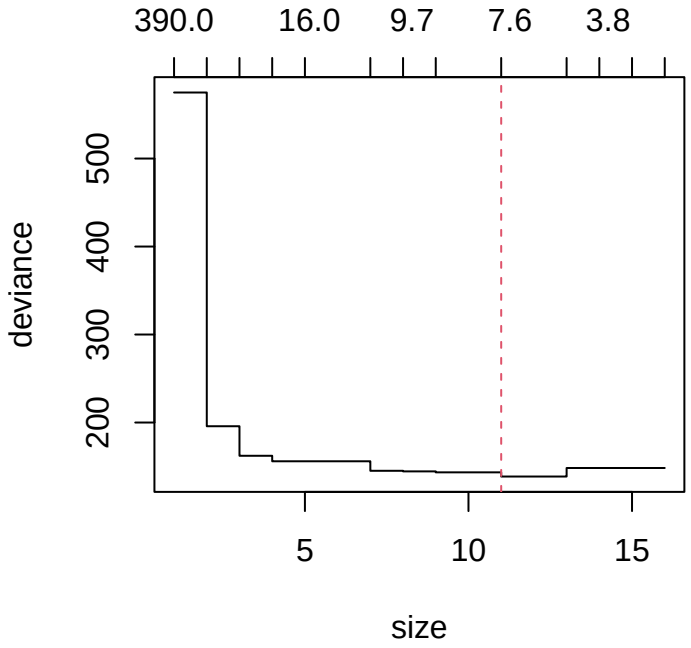

Figure 2: Errore in convalida incrociata in funzione del numero di foglie

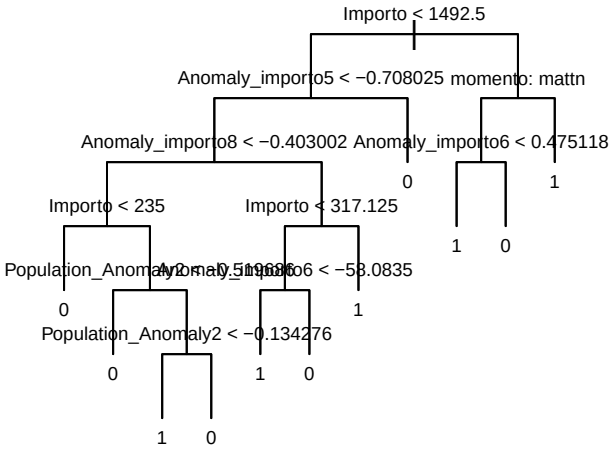

Figure 3: Albero di classificazione selezionato

Random forest

Si procede con l'adattamento del *random forest*. Il parametro di regolazione del modello è il numero di covariate da considerare ad ogni suddivisione dell'albero. A tal riguardo, l'insieme di stima viene diviso in un insieme di stima ridotto e uno di convalida e viene adattato il *random forest* con 500 alberi in corrispondenza di ognuno dei possibili valori del numero di covariate considerate. Il numero di covariate selezionato è il valore corrispondente al modello con tasso di errata classificazione minore nell'insieme di convalida. La Figura 4 mostra che, con tale procedura, si sceglie un numero di colonne da campionare per ogni albero pari a 4.

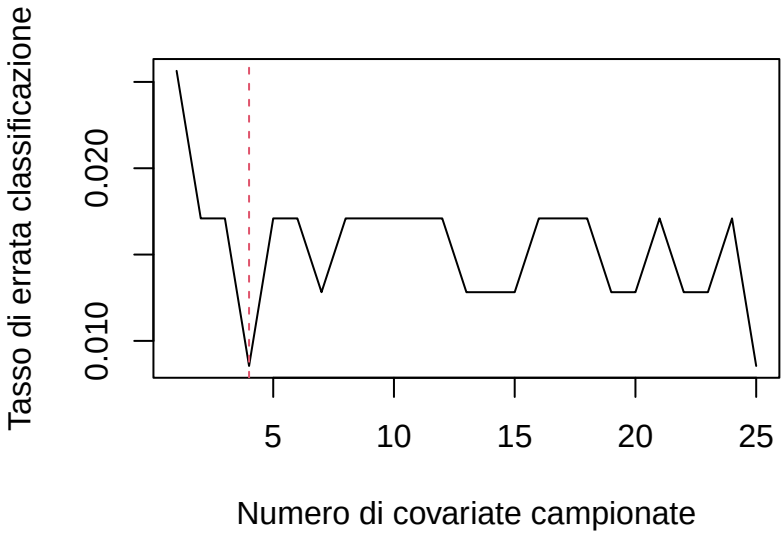

Figure 4: Errore nell'insieme di convalida in funzione del numero di covariate campionate

Successivamente il modello selezionato (4 variabili esplicative considerate in ogni albero) è adattato sull'intero insieme di stima e permette di ottenere un tasso di errata classificazione e una percentuale di falsi negativi nell'insieme di verifica pari rispettivamente al **3.36%** e allo **0.00%**.
Questo modello permette di ottenere una misura di importanza delle variabili esplicative, senza però avere indicazione sulla direzione dell'effetto di esse sulla risposta. In queso caso, la Figura 5 mette in luce che le variabili più importanti in termini di diminuzione dell'errore di previsione risultano essere l'importo della transazione, il secondo, il terzo, il quarto, il quinto e il sesto indicatore di confronto della carta con carte ad essa simili.

Bagging

Si adatta un *bagging* con alberi di classificazione. Viene calcolato l'errore OOB per diversi valori del numero di campioni boostrap (e quindi di alberi) utilizzato dal modello, scegliendo il valore per cui l'errore OOB è minore. In questo caso è pari a 230, come si evince dalla Figura 6, in cui si riporta il grafico dell'errore OOB in funzione del numero di campioni bootstrap.

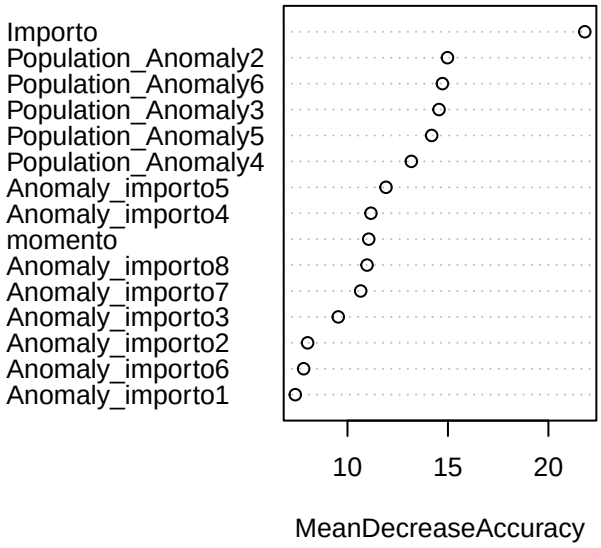

Figure 5: Importanza delle variabili nel random forest

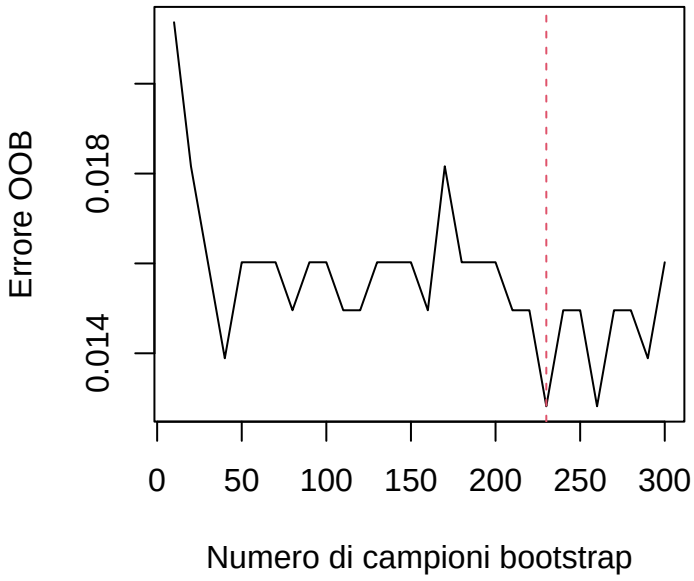

Figure 6: Errore OOB (Out-Of-Bag) nell'insieme di stima in funzione del numero di campioni bootstrap

Il modello selezionato ottiene sull'insieme di verifica un tasso di errata classificazione pari al **3.36%** e una percentuale di falsi negativi pari allo **0.00%**.

Boosting

Si adatta un *boosting* con alberi di classificazione. Per individuare il numero di alberi necessari a stabilizzare l'errore di previsione, si divide l'insieme di stima in un insieme di stima ridotto e uno di convalida. La Figura 7 mostra l'errore di previsione nell'insieme di convalida in funzione del numero di iterazioni dell'algoritmo, facendo notare che l'errore è minimo e costante dopo 130 iterazioni.

Figure 7: Errore di previsione nell'insieme di convalida in funzione del numero di iterazioni

Il modello selezionato, e riadattato sull'intero insieme di stima, ottiene un tasso di errata classificazione e una percentuale di falsi negativi nell'insieme di verifica pari rispettivamente allo **0.83%** e allo **0.00%**.
Anche questo modello ha il pregio di portare informazione sull'importanza delle variabili esplicative. La Figura 8 permette di far notare che le variabili maggiormente presenti negli stumps risultano essere il secondo, il terzo, il quarto, il quinto e il sesto indicatore di confronto della carta con carte ad essa simili.

Support Vector Machine

L'ultimo modello adattato è una *Support Vector Machine* con nucleo radiale. Il parametro di regolazione è il costo relativo alle errate classificazioni e, per selezionarlo, si suddivide l'insieme di stima in un insieme di stima ridotto e uno di convalida e si considera una griglia di valori interi da 1 a 30. Il valore scelto è quello che corrisponde al modello con minor tasso di errata classificazione nell'insieme di convalida ed è pari a 27, come si può vedere dalla Figura 9. Il modello selezionato è adattato su tutto l'insieme di stima e ottiene, nell'insieme di verifica, un tasso di errata classificazione pari all' **1.74%** ed una percentuale di falsi negativi pari allo **0.00%**.

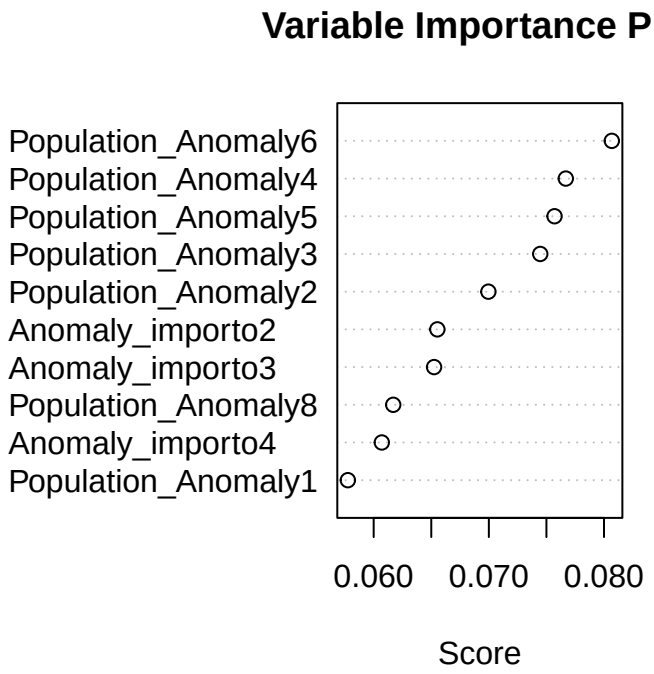

Figure 8: Importanza delle variabili nel boosting

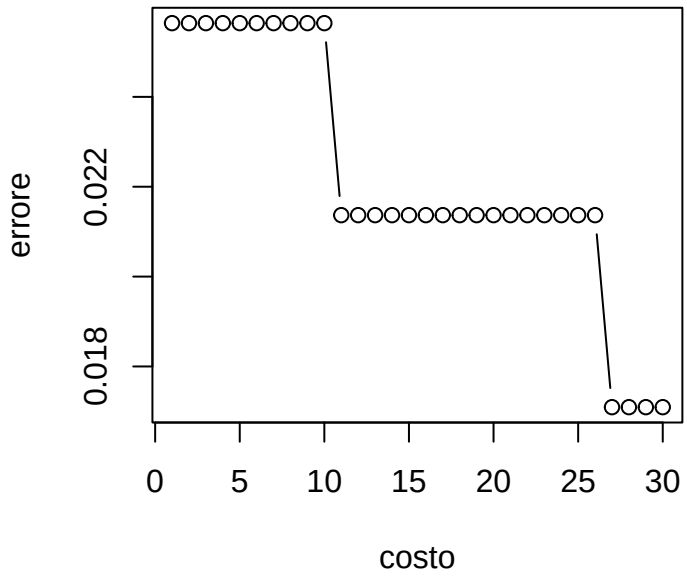

Figure 9: Tasso di errata classificazione nell'insieme di convalida in funzione del costo dell'errore

Risultati

Nella Tabella 1 si riportano i risultati ottenuti coi diversi modelli adattati in termini di tasso di accuratezza (ovvero il complemento ad 1 del tasso di errata classificazione) e di proporzione di falsi negativi.

Table 1: Misure di performance dei modelli adattati

	Falsi negativi	Accuratezza
Boosting	0.0000	0.9917
Analisi discriminante lineare	0.1071	0.9900
Support Vector Machine	0.0000	0.9826
Logistico stepwise	0.0000	0.9800
Albero	0.1071	0.9779
Bagging	0.0000	0.9664
Random Forest	0.0000	0.9634

Si nota che, ad eccezione dell'*albero di classificazione* e dell'*analisi discriminante lineare*, tutti i modelli non prevedono mai un'operazione fraudolenta come non fraudolenta. Alla luce di questa considerazione, sembra ragionevole scegliere il modello che permette di ottenere il tasso di accuratezza più elevato tra tutti i modelli che non prevedono falsi negativi. La Tabella 1 mostra come questa misura sia più elevata nel *boosting* (99.17%), con a seguire la *Support Vector Machine* (98.26%), il modello *logistico stepwise* (98.00%), il *bagging* (96.64%) e il *random forest* (96.34%).
Focalizzando l'attenzione sul *boosting*, come già è stato detto in precedenza, questo modello permette di avere una misura di importanza delle variabili esplicative, senza però avere indicazione sulla direzione dell'effetto di queste variabili sulla risposta. In questo caso, le variabili maggiormente importanti risultano essere il secondo, il terzo, il quarto, il quinto e il sesto indicatore di confronto della carta con carte ad essa simili.