Un modello predittivo per l'analisi del rischo di credito

Team 40: Roberto Berlucchi, Andrea Caciolo, Nicolò Ghioldi, Fedrico Manenti, Ivan Mera

Abstract

Le banche affrontano tutti i giorni diversi tipi di rischi, tra i quali quello di credito. L'ufficio di Risk Management è quel dipartimento volto a limitare le perdite legate a questi fenomeni. Nell'ambito del rischio di credito prevedere se un determinato cliente non ripagherà un debito offre agli istituti finanziari un notevole vantaggio per evitare future perdite. Nel nostro progetto ci siamo focalizzati sulla situazione taiwanese intorno al 2006, periodo in cui si è verificato un aumento dei cattivi pagatori. L'obiettivo di questo lavoro è stato, quindi, quello di scegliere il modello migliore per prevedere se un cliente farà fronte ai suoi impegni oppure se non sarà in grado di rimborsare l'ammontare di quanto dovuto. Abbiamo utilizzato tecniche di feature engineering per superare diversi probelmi riscontrati nel dataset utilizzato, quale per esempio la class imbalance. Per portare a termine il nostro obiettivo ci siamo avvalsi di cinque diversi tipi di classificatori: Logistic, Random Forest, J48, Naive Bayes e Multi Layer Perceptron. Successivamente li abbiamo allenati e testati con due metodologie differenti per la classificazione binaria: Holdout e Cross Validation. Abbiamo proceduto all'analisi dei costi ed infine, dopo la valutazione finale, abbiamo ottenuto il miglior classificatore per questi dati: Multi Layer Perceptron.

CONTENTS

1 INTRODUZIONE

Attorno al 2006, gli emittenti di carte di credito in Taiwan hanno affrontato una crisi che ha portato a raggiungere un tasso di insolvenza elevato nel terzo trimestre dell'anno. Le banche, con il fine di aumentare la loro quota di mercato, hanno concesso prestiti ed emesso carte di credito a clienti che non rispettavano i requisiti ideali e che in condizioni normali non avrebbero superato la fase di valutazione del rischio di credito. Dall'altra parte, i possessori di carte di credito non hanno saputo valutare la loro capacità di rimborso finendo con consumare più del dovuto e accumulando pesanti debiti che non avrebbero potuto rimborsare.

La crisi ha causato una diminuzione della fiducia verso i consumatori nel mercato bancario, evento che può essere pericoloso per l'economia reale. Infatti, se le banche soffrono ingenti perdite sono costrette ad applicare una stretta del credito e a richiedere il rimborso anticipato, anche a quei clienti che sono in grado di far fronte ai loro impegni, propagando la crisi all'economia reale.
Viene meno anche la fiducia tra le banche, elemento su cui si basa l'intero sistema finanziario. In mancanza di essa, gli istutuiti di credito non utilizzano più il mercato interbancario dei prestiti $overnight$ per paura di non essere rimborsati non avendo informazioni sulle controparti.
Con il fine di evitare tutto ciò le banche dovrebbero svolgere correttamente la loro funzione principale, ovvero valutare il rischio di credito. Nell'eseguire questo compito le banche si affidano di norma a modelli di rating interni utilizzando informazioni finanziarie come il reddito, il numero e l'ammontare delle transazioni, lo storico dei pagamenti ecc. per ridurre il più possibile il rischio di insolvenza.
Nella nostra analisi cercheremo di determinare quale sia il modello migliore, tra quelli scelti, per classificare chi pagherà e chi no utilizzando metodi statistici. In questo modo una banca può capire se un cliente non rimborserà il dovuto e quindi procedere ad applicare un blocco preventivo sulla carta di credito per evitare perdite.

Dataset

Per la nostra analisi abbiamo scelto il dataset $UCI\_Credit\_Card.csv$ presente sulla piattaforma Kaggle reso disponibile dal professore I-Cheng Yen. Il dataset contiene informazioni sui pagamenti dei titolari di una carta di credito ed è composto da $30\,000$

osservazioni e 25 variabili.
Abbiamo deciso di nascondere il 10% dei dati nella nostra analisi al fine di simulare uno scenario di business in cui, dopo l'addestramento dei modelli, la valutazione avviene su dati sconosciuti e perciò teoricamente indipendenti.
Riportiamo una breve descrizione delle variabili presenti nel dataset:

- $LIMIT\_BALL$ è il plafond che la banca concede al cliente espresso in NT (dollari taiwanesi).
- SEX è il sesso e assume valori 1 per uomo e 2 per donna.
- $EDUCATION$ indica il grado di istruzione (1 = scuola secondaria di primo grado; 2 = università; 3 = scuola secondaria di secondo grado; 0, 4, 5, 6 = altri).
- $MARRIAGE$ (1 = sposato; 2 = single; 3 = divoziato; 0 = altro).
- AGE è l'età del cliente.
- $PAY\_$(1-6) indica lo storico dello stato dei pagamenti da Aprile (PAY_6) a Settembre (PAY_1) 2005 (-2 = nessun uso della carta; -1 = pagato per intero; 0 = ha scelto di usare il revolving credit; 1-9-... = numero di mesi scelti per posticipare il pagameto).
- $BILL\_AMT\_$(1-6) è l'estratto conto (come sopra 1 settembre ... 6 aprile) che può assumere valori negativi nel caso in cui il cliente abbia pagato più del dovuto.
- $PAY\_AMT\_$(1-6) è l'ammontare dei pagamenti in riferimento al mese precedente (come sopra 1 settembre ... 6 aprile).
- $default.payment.next.month$ indica il comportamento del cliente (1 = non ha pagato; 0 = ha pagato).

La variabile obiettivo, che i nostri modelli dovranno prevedere, è $default.payment.next.month$.

2 Preprocessing e feature engineering

Da una prima analisi esplorativa abbiamo riscontrato l'asssenza di valori mancati, l'assimetria delle variabili continue *BILL_AMT 1-6* e *PAY_AMT_1-6* e una forte correlazione tra le variabili *BILL_AMT_1-6*. Quest'ultima è presumibilmente causata dal fatto che il totale dell'estratto conto varia di mese in mese in modo non significativo.

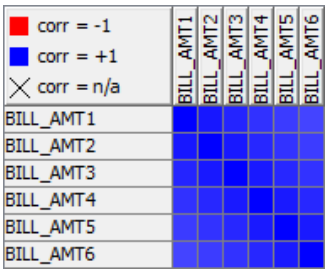

Figure 1. Matrice Correlazione

- Per affrontare il problema dell'assimetria abbiamo deciso di sostituire i valori negativi degli attributi con 0, non essendo interessati a considerare le posizioni in credito verso la banca, e di applicare successivamente una trasformazione logaritmica del tipo $\ln(x+1)$ dove x rappresenta la variabile in esame.

- Per quanto riguarda la correlazione abbiamo deciso di considerare al posto delle singole variabili *BILL_AMT_1-6* un nuovo attributo chiamato *avg_bill*, media delle precedenti, in modo tale da ridurre la dimensione del dataset e migliorare le prestazioni dei classificatori.

Con lo stesso obbiettivo abbiamo raggruppato la variabile *AGE* in 8 intervalli e proceduto con la binarizzazione delle variabili nominali, avendo prima raggruppato i 4 valori distinti 0, 4, 5, 6 della variabile *EDUCATION*.
Vogliamo inoltre sottolineare, che abbiamo preferito evitare di procedere con una selezione delle variabili. Infatti, nella nostre prove, i selettori usati andavano a scegliere un certo mese non essendo sempre concordi sulla sua determinazione. Abbiamo deciso di non dare maggior peso ad un mese specifico, ma di considerare tutte le variabili. Questa cosiderazione deriva dal fatto che il nostro dataset contiene informazioni temporali.

3 Modelli e Misure di valutazione

I modelli utilizzati nel corso della nostra analisi, implementati con i nodi di $WEKA$, sono i seguenti:

- *Albero di regressione* ($J48$).
- *Random Forest* con 100 alberi, 5 random features e con profondità massima di 5 per iterazione.
- *Regressione Logistica Semplice (RLS).*
- *Multi Layer Perceptron (MLP)* con 100 epoche e 11 strati nascosti.
- *Naive Bayes (NB).*

Prima di affrontare la discussione dei modelli va notato che il dataset preso in esame presenta uno sbilanciamento nella variabile di classe: i clienti che non pagheranno sono solo circa il 22%. Questa minoranza della classe di intesse potrebbe causare problemi e vengono presentate alcune possibili soluzioni.

3.1 Misure alternative all'accuracy

Utilizzare l'accuracy come misura di $performance$ non è consigliabile nel nostro caso, quidi abbiamo preferito considerare anche le seguenti misure alternative:

$$Recall = \frac{TP}{TP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Fmeasure = 2\frac{Precision * Recall}{Precision + Recall} \tag{3}$$

Visto che il nostro interesse è prevedere tutti i clienti che non pagheranno, abbiamo deciso di dare maggior peso alla *Recall*. Ciò deriva dal fatto che è più grave classificare un cliente che non pagherà come un cliente che lo farà piuttosto che commettere l'errore inverso, perchè per la banca significherebbe una perdita sicura di denaro.

3.2 Analisi dei costi

Un'altra tecnica per dar maggior peso all'errore di predizione a noi interessato è quella di addestrare i classificatori mediante una matrice dei costi.
La scelta dei pesi è stata puramente soggettiva e posta a priori, favorendo l'errore di sbagliare a prevedere un cattivo pagatore rispetto alla mancata previsione di esso.

		Predicted 0	Predicted 1
Actual	0	$C_{--} = 0$	$C_{-+} = 1$
	1	$C_{+-} = 2$	$C_{++} = 0$

La decisione di utilizzare questo tipo di analisi dovrebbe arrivare dall'esperto di settore che inoltre dovrebbe fornirne i pesi. Abbiamo deciso di metterci nello scenario di business in cui ci venga passata questa matrice confidando che la scelta dei pesi da noi utilizzata sia, se non ottimale, almeno sensata.

3.3 Down Sampling

Il modo più semplice per ribilanciare la variabile obiettivo è quello di procedere con un ricampionamento.
Dopo le opportune prove abbiamo preferito l'utilizzo del Down Sampling rispetto all'Up Sampling, poichè il primo otteneva risultati migliori. Abbiamo fatto ciò mediante il nodo *Equal Size Sampling* che mantiene tutte le istanze contenenti la classe minoritaria e campiona in modo casuale quelle di maggioranza ottendo un rapporto $50/50$ tra le due.

4 Holdout e Cross Validation

Seguendo l'idea presentata nell'introduzione, abbiamo preventivamente diviso, attraverso un $Stratified\ Sampling$, il nostro dataset in due parti. La scelta di utilizzare questo tipo di campionamento è dovuta alla volontà di cercare di mantere la stessa distribuzione nei due sottoinsiemi di dati.
Presentiamo le metodologie di validazione utilizzate.

4.1 Holdout

Abbiamo suddiviso il nostro dataset in $2/3$ e $1/3$ rispettivamente per il Train Set e Test Set attraverso un campionamento stratificato sulla variabile obiettivo. Prima di procedere con l'addestramento dei classificatori abbiamo utilizzato la tecnica precedentemente esposta del Down Sampling.
Di seguito riportiamo i valori delle misure calcolate attraverso i nodi Scorer, prestando particolare attenzione alla $Recall$.

	Recall	Precision	F-measure	Accuracy
J48	0.66	0.39	0.49	0.69
Random Forest	0.63	0.47	0.54	0.76
Logistic	0.61	0.47	0.53	0.76
MLP	0.63	0.46	0.53	0.76
Naive Bayes	0.61	0.42	0.50	0.73

Riportiamo inoltre la curva ROC.

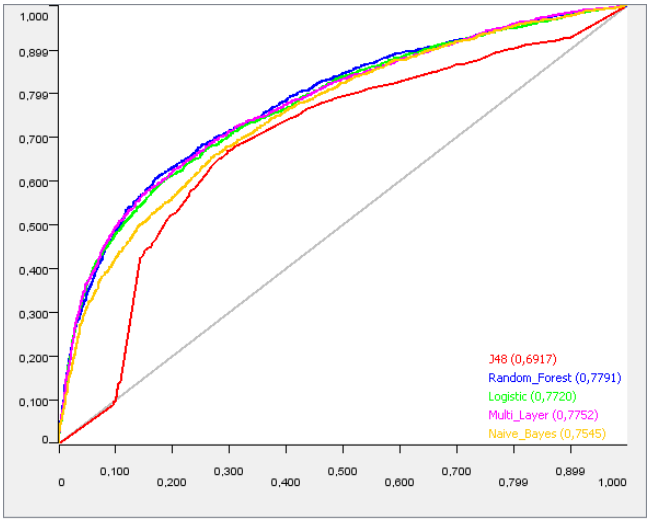

Figure 2. Curva ROC

I modelli che risultano più performanti sono: *Random Forest, Regressione Logistica e Multi Layer Perceptron*. Questi, infatti, ottengono il valore di AUC più alto.
Successivamente abbiamo valutato i modelli migliori con quei "dati nascosti" per valutarne l'efficenza. Riportiamo i valori ottenuti:

	Recall	Precision	F-measure	Accuracy
J48	0.67	0.38	0.48	0.69
Random Forest	0.61	0.46	0.52	0.75
Logist	0.60	0.47	0.52	0.76
MLP	0.62	0.45	0.52	0.75
Naive Bayes	0.63	0.44	0.52	0.74

4.2 Cross Validation

Abbiamo deciso di utilizzare anche la metodologia della $Cross\ Validation$ in aggiunta alla semplice tecnica dell'$Holdout$.
Questa scelta segue dal fatto che il metodo permette di addestrare il modello su tutte le osservazioni di partenza. Infatti, esso suddivide il dataset in k gruppi di egual numerosità e ad ogni giro ne sceglie uno come Test Set mentre tutti gli altri saranno il Training Set. Questo permette, nei dataset con una buona numerosità, di risolvere problemi di overfitting e cattivo campionamento mentre allo stesso tempo, utilizzando tutte le osservazioni, introduce una dipendenza da quest'ultime impedendo un buona valutazione solo attravereso il Test Set. Dopo alcune prove abbiamo deciso di fissare il numero di gruppi a 10 e applicare la tecnica del Down Sampling al Training Set.
Per calcolare gli intervalli di confidenza per $Recall$, $Precision$ e $F\_measure$ dobbiamo fare affidamento sulla distribuzione $Beta$ di parametri a e b. Scriviamo allora

le probabilità a posteriori per Precision e Recall che sono rispettivamente:

$$Recall|D \sim Beta(TP + \lambda, FN + \lambda) \quad (4)$$

$$Precision|D \sim Beta(TP + \lambda, FP + \lambda) \quad (5)$$

La distribuzione β ci garantisce molta flessibilità e, in particolare, la scelta di porre il parametro λ = 1/2, (Jeffrey's non-informative prior) ci garantisce la proprietà di invarianza rispetto alle riparametrizzazioni.
Infine, essendo nelle ipotesi del Teorema del limite centrale, possiamo approssimare la distrubuzione con quella Normale. Quanto affermato può essere consultato nello studio effettuato da Cyril Goutte and Eric Gaussier dal nome "Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation".
Riportiamo i grafici con gli intervalli di confideza.

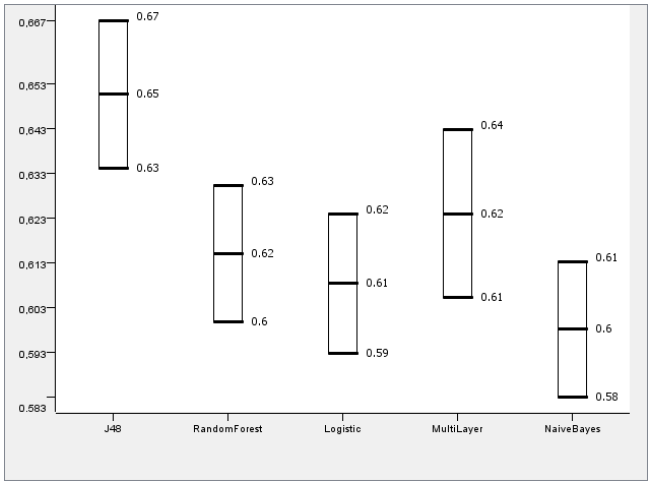

Figure 3. Recall

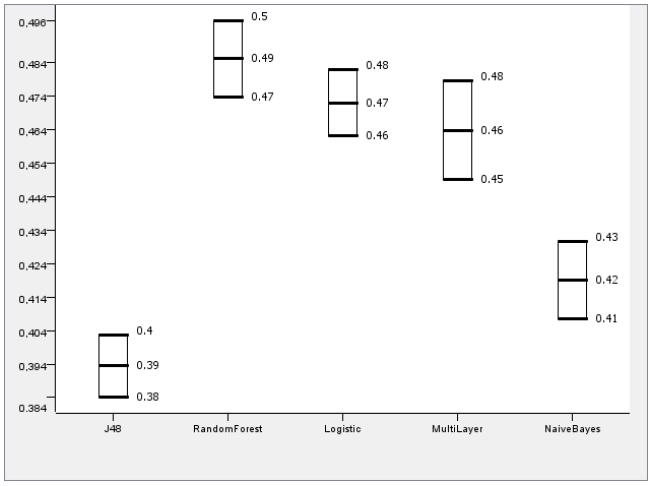

Figure 4. Precision

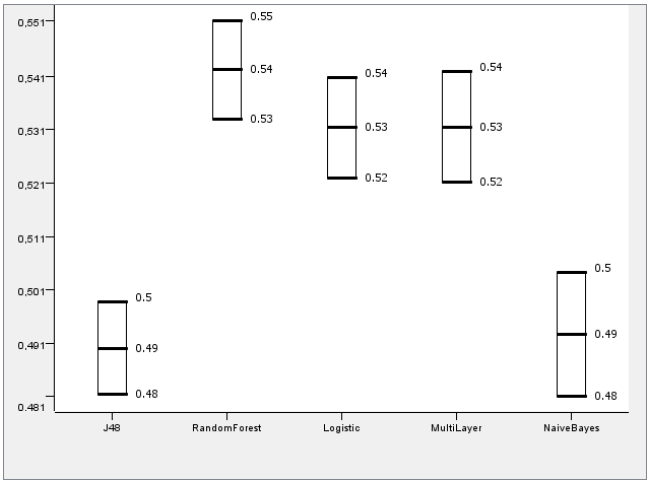

Figure 5. F-measure

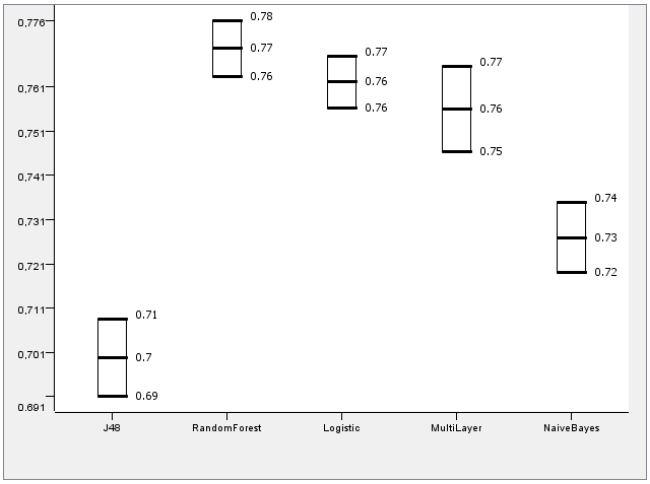

Figure 6. Accuracy

Quest'ultimi sono stati costruiti calcolando l'intervallo di confidenza a livello del 95%.
I modelli migliori rimangono ancora *Logistic*, *Random Forest* e *MLP* .
Come nel caso dell'Holdout riportiamo i valori ottenuti dopo aver valutato i modelli sui "dati nascosti".

	Recall	Precision	F-measure	Accuracy
J48	0.67	0.38	0.48	0.68
Random Forest	0.62	0.47	0.53	0.76
Logist	0.62	0.48	0.54	0.76
MLP	0.65	0.46	0.54	0.75
Naive Bayes	0.61	0.43	0.51	0.73

4.3 Confronto classificatori

Avendo constatato che questi ultimi tre modelli risultano i più performanti per la nostra analisi, abbiamo deciso di confrontarli sulla base di una misura di errore alternativa rispetto alla classica *1 - Accuracy*. Visto e considerato il nostro obiettivo di classificare correttamente quanti più possibili cattivi pagatori, abbiamo deciso di utilizzare come misura il *FNR = 1 - Recall*.
Riportiamo i grafici del confronto ottenuto come differenza degli errori tra i tre classificatori.

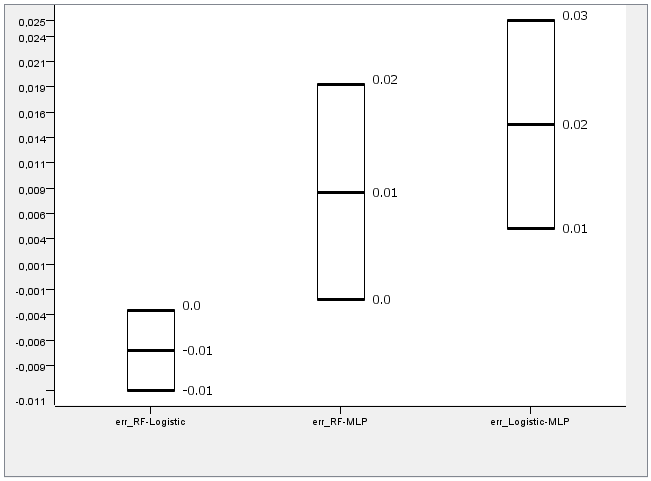

Figure 7. Confronto FNR

Siccome alcuni risultati sono prossimi allo 0, riportiamo di seguito in tabella i valori più precisi.

	RF-Logistic	Logistic-MLP	RF-MLP
min	-0.011	0.005	-0.002
media	-0.007	0.016	0.009
max	-0.003	0.026	0.019

Dal confronto effetuato si può evincere che Il $RandomForest$ e MLP risultano clasifcatori migliori rispetto al $Logistic$ al livello del 95%, mentre tra $RandomForest$ e MLP la differenza non risulta statisticamente significativa.

5 Costo

Abbiamo deciso di applicare l'analisi dei costi sopra descritta ai modelli selezionati precedentemente per cercare di ottenere un risultato migliore in termini di $Recall$.
Riportiamo i valori ottenuti dopo aver applicato la matrice dei costi, facendo notare il netto miglioramento in termini di $Recall$.

	Recall	Precision	F-measure	Accuracy
MLP	0.81	0.32	0.46	0.57
Logistic	0.86	0.31	0.46	0.55
Random Forest	0.88	0.31	0.45	0.53

Tutti e tre i modelli migliorano sensibilmente la loro $Recall$ mantendo comunque un valore di $Precision$ sopra il 30% e di $Accuracy$ sopra il 50%.
Abbiamo, inoltre, calcolato il costo per ognuno dei tre modelli per avere un ulteriore sicurezza metodologica su quale scegliere. Il calcolo è stato eseguito con la seguente formula, considerando perciò il migliore quello con il costo più basso:

$$costo = C_{--}TN + C_{-+}FP + C_{+-}FN + C_{++}TP \quad (6)$$

	MLP	Logistic	Random Forest
Costo	3827	3924	4002

6 Conclusioni

Nella nostra analisi sono stati presentati diversi modelli utilizzati per profilare i possessori di carte di credito. Il dataset utilizzato presentava uno sbilanciamento e abbiamo applicato alcune tecniche per superare questa problematica. Successivamente abbiamo condotto una analisi dei costi per migliore l'efficacia dei classificatori.
Presentiamo adesso le nostre conclusioni per rispondere alla domanda che ci eravamo posti su quale fosse il modello statistico più adatto alle nostre esigenze.
Il miglior modello, sotto le nostre ipotesi, è risultato essere il *Multi Layer Perceptron*, infatti esso ottiene il risultato migliore nell'analisi dei costi. Con essa il valore di $Recall$ è passato da 0.65 a 0.81 ottenendo un sensibile miglioramento dato il nostro fine di evitare la mancata previsione di un cattivo pagatore.
Vogliamo comunque far notare che anche il *Random Forest* ottiene buoni valori, nonostante risulti leggermente peggiore nel test finale. Una considerazione ulteriore va affrontata analizzando i valori ottenuti con il test svolto sui "dati nascosti". Otteniamo, infatti, valori simili e leggermente inferiori a quelli ottenuti nel Test, dandoci sicurezza sulla bontà della nostra analisi.
Un prossimo lavoro potrebbe essere indirizzato all'ottimizzazione del modello scelto per ottenere risulati migliori in termini di $Precision$, che si rivela bassa nei nostri risulati. Inoltre, data la soggettività delle scelte fatte, andrebbero esse discusse con un esperto di dominio al fine di valutarne la bontà e in caso contrario di rivalutare l'analisi sotto le nuove codizioni.

7 Riferimenti

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation
https://pdfs.semanticscholar.org/e399/9a46cb8aaf71131a77670da5c5c113aad01d.pdf
https://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf
https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf
http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf