Università degli Studi di Roma "La Sapienza"
Dipartimento di Scienze di Base e Applicate per l'Ingegneria
Professoressa Barbara Vantaggi

Probabilistic Models for Finance
2019

Report finale:

Il modello logistico nell'ambito del Credit Scoring

Marco Minici

Abstract

In questo documento viene descritto il progetto finale sul modello logistico applicato al credit scoring studiato nel corso di Probabilistic Models for Finance. Nella sezione 1, verrà fatta una introduzione dell'argomento spiegando cosa sia il credit scoring e i motivi del suo esteso uso negli istituti di credito. Nella sezione 2, verrà presentato il dataset analizzato e le varie fasi di pre-processing dei dati. Nell'ultima sezione invece verranno presentati i risultati ottenuti, valutati secondo varie metriche e confrontandoli con i risultati ottenuti in letteratura.

1 Introduzione

La recente crisi economica, originata da un numero inatteso di insolvenze nei mutui ipotecari del mercato immobiliare americano, ha accresciuto la richiesta di competenze statistiche, note come *credit scoring*, per la misurazione *ex ante* del rischio di credito. Nel momento in cui riceve una richiesta di finanziamento, la banca o l'intermediario finanziario deve valutare il rischio del soggetto che contrae il debito di non riuscire a fare fronte agli impegni contrattuali.

In passato, i finanziamenti venivano concessi sulla base di un carattere soggettivo come il legame personale fra il richiedente e l'analista del credito dell'ente erogatore. Negli anni recenti però, a causa dell'elevato numero di richieste e della varietà di prodotti finanziari offerti, una sempre maggiore spersonalizzazione si è affermata nel rapporto fra le due controparti. Allo stesso tempo, la crescente competizione nel mercato richiede una tempestività nella decisione di concedere un finanziamento visto che la clientela, se meritevole, può ottenere il finanziamento da un ente concorrente. In sintesi, questi sono stati gli elementi che hanno portato alla gestione automatizzata delle informazioni relative al potenziale cliente ed alla elaborazione di un punteggio, detto *score*, che ne rifletta l'affidabilità creditizia. Le tecniche di *scoring* sono adoperate in due diverse tipologie di operazioni, quando si tratta di concedere del credito ad un cliente si parla di *application scoring*, mentre se invece l'operazione riguarda la gestione di un cliente già affidato allora si parla di scoring comportamentale, *behavioural scoring*. In questo progetto ci focalizzeremo su un caso di *application scoring*, in cui vogliamo prevedere se un potenziale cliente sia solvibile sulla base delle informazioni in possesso al momento della richiesta del finanziamento. Di conseguenza, la variabile che definisce l'evento da prevedere è binaria atta a suddividere la potenziale clientela in due insiemi: "buoni" e "cattivi" pagatori. In maniera più formale, lo scoring di accettazione è il processo attraverso cui alcune informazioni relative ad un richiedente credito vengono combinate e convertite in un punteggio, detto *score*, costruito in modo tale da essere proporzionale alla probabilità stimata che il richiedente sia solvibile. Lo score del potenziale cliente è confrontato con un apposito valore di soglia: se lo score è superiore al valore di soglia, il richiedente è classificato come "solvibile", altrimenti è classificato come "non solvibile" e non accede al finanziamento.

2 Dataset

La base di dati consiste, per ogni unità, nelle informazioni rilevate al momento della richiesta di finanziamento che descrivono sia il tipo di finanziamento come la durata e l'ammontare sia il profilo socio-demografico del cliente. Sono presenti anche le informazioni di natura finanziaria in possesso dei *credit bureaux* e che riguardano la storia creditizia del cliente.

Il dataset usato nel progetto è stato rilasciato da Findomestic Banca nell'ambito dell'hackaton SUS4 [2]. Nella tabella 1 vengono mostrate tutti gli attributi del dataset, suddivisi per macrocategorie e con la relativa descrizione. Per questioni di spazio la tabella è stata riportata nell'appendice del report.

2.1 Pre-processing dei dati

Il pre-processing è uno step importante nel processo di estrarre conoscenza dai dati. I metodi di raccolta dei dati sono spesso imprecisi e poco controllati, per questo possono risultare in dati fuori dal dominio della variabile(redditi negativi), combinazioni impossibili(sesso: maschio, incinta: si), valori mancanti e altro. Fare uso di dati che prima non sono stati controllati può condurre a risultati fuorvianti. La rappresentazione e la qualità dei dati è un requisito principale per poter analizzare i dati.

L'output finale del pre-processing è quello della produzione del training set da usare in fase di analisi.

Eliminazione variabili discriminatorie Come espresso a pagina 2 in [1], le procedure automatiche di supporto alla decisione di concessione del credito sono state sottoposte ad una rigorosa legislazione per prevenire comportamenti discriminatori. Il primo esempio di regolazione statale è l'Equal Credit Opportunity Act [3] emanato dal governo USA nel 1974. Seguendo quanto scritto in [3] vengono **eliminate** dal dataset le seguenti informazioni:

- "Sesso": il genere del richiedente.
- "Età": l'età del richiedente.
- "Stato Civile": lo stato civile del richiedente.
- "Nazione di nascita": la nazione di nascita del richiedente

Trasformazione variabile target Come descritto in [1], l'obiettivo del credit scoring è quello di prevedere una variabile binaria che discerne la potenziale clientela tra "buoni" e "cattivi" pagatori.

Il nostro dataset identifica la clientela in tre diverse categorie: "regolare", "contenzioso" e "recupero". Per quanto detto precedentemente, **verrano ricondotti i due casi negativi in uno unico**, il cosiddetto *cattivo*. Questo passaggio potrebbe sembrare artificioso, ma segue il principio di definizione della variabile di classificazione definito in [1] come *l'evento a partire dal quale si ingenera un reale disagio per l'ente che eroga il credito.*

Generazione variabili dummy Quando il dataset è caratterizzato da variabili qualitative, un modo di trattarle è quello di trasformarle in una serie di variabili binarie attraverso il processo di **dummization** [1]. Una variabile dummy assume valore 0 o 1 per indicare l'assenza o la presenza di un qualche effetto categorico che ci si aspetta potrebbe influire sull'output del modello. Il processo di dummization consiste nel trasformare le variabili qualitative in tante variabili dummy quanti livelli sono espressi dalle variabili qualitative.

Le variabili trasformate in dummy nel nostro dataset sono le seguenti:

- "score_cmp_cb": che è lo score comportamentale del credit bureau.
- "RESIDENZA": rappresenta la tipologia di residenza "Proprietario", "Locatario", etc.
- "CANALE_FIN": il canale di comunicazione da cui si è attivata la procedura di richiesta di finanziamento(telefono, filiale fisica, etc.).
- "REGIONE": che è la regione di residenza del richiedente.

Gestione valori nulli Può succedere che nei nostri dati compaiano dei valori nulli/NAs. Questi valori possono essere gestiti in vari modi in base alla loro interpretazione. Possibili strategie sono:

- Eliminazione NAs.
- Imputazione tramite media, mediana, moda o modelli predittivi.

Le ragioni, i vantaggi e gli svantaggi di ogni strategia non verrano espressi qui, visto che non sono argomento principale di questo lavoro.

Sono presenti 27 records con valore NA della variabile ANZ_BAN e 3 records con valore NA della variabile REGIONE. Visto il ridotto numero si è deciso di eliminarli.

Divisione in campione di sviluppo e campione di convalida Come evidenziato in [1], per evitare di utilizzare le unità due volte, è buona prassi dividere il campione in due sotto-campioni. Il campione di sviluppo per costruire il sistema di scoring e il campione di convalida per valutarlo. Il secondo è generalmente di inferiore numerosità e in questo caso verrà usata la divisione training-test 70-30.

La divisione "$\frac{1}{3}$-$\frac{2}{3}$" del campione è stata usata in letteratura da molti studi (Lee et al, 2002; Desai et al, 1996; Boritz & Kennedy, 1995; Dutta et al, 1994) come mostrato nell'ampia review sul credit scoring in [4].

Standardizzazione delle variabili predittive Una pratica usata nell'ambito del Machine Learning è la standardizzazione delle variabili predittive. Questa trasformazione dei dati fonda il suo uso su diverse ragioni:

1. Migliore interpretazione dei coefficienti di regressione. Interpretare il valore dei coefficienti di regressione tra variabili che operano su scale diverse(euro vs kg, mm vs km, etc.) è più complesso rispetto a variabili che operano sulla stessa scala.

2. Corretto uso di algoritmi basati su distanza. Algoritmi come il K-Means basati sulla distanza tra oggetti possono interpretare fallacemente variabili su scale diverse e produrre risultati non desiderabili.

3. Sinergia con procedure di regolarizzazione. I metodi di regolarizzazione come LASSO richiedono una standardizzazione iniziale dei regressori, in modo che la penalizzazione sia equa con tutti i regressori. [9]

4. Altri motivi sempre relativi alla varianza di scala di algoritmi, procedure di ottimizzazione(gradiente discendente), etc.

La standardizzazione usata in questo lavoro segue la Formula 1 per ogni variabile x del dataset:

$$z = \frac{x - \mu_x}{\sigma_x} \tag{1}$$

dove μ_x è il valore medio della variabile x e σ_x è la deviazione standard della variabile x.

Sotto-campionamento per avere un campione bilanciato. Come riportato in [1], un problema del credit scoring è quello di avere una distribuzione sbilanciata sulla variabile di risposta Y rispetto alle unità "buone". Ricordiamo che i dati ottenuti sono disponibili soltanto per clienti che inizialmente avevano ottenuto l'erogazione del finanziamento, quindi è auspicabile che il numero dei clienti insolventi sia molto minore di quelli regolari.

Il rischio di un dataset sbilanciato è quello di oscurare la relazione fra le variabili esplicative e la variabile risposta e porta ad una percentuale elevata di errori di classificazione a sfavore delle unità rare, sulle quali vi è meno informazione.

Per ovviare a questo problema in [1], viene presentato un campionamento stratificato rispetto alla variabile di risposta Y che consiste in:

- Estrarre in modo casuale una frazione delle unità sane tale da riportare il campione ad essere bilanciato.

- Spostare le unità sane non estratte dal campione di stima al campione di convalida.

3 Analisi e valutazione dei risultati

Per l'implementazione del modello logistico, è stato fatto uso del linguaggio di programmazione Python e della libreria sklearn.

Come descritto nella review in [5], non c'è uno studio che abbia dimostrato l'esistenza di un criterio di valutazione ottimale in ambito credit scoring. Gli autori di [5] presentano quelli più usati in letteratura che sono: Average Correct Classification(ACC) rate, il costo atteso di errata classificazione, l'errore quadratico medio(MSE), la radice dell'errore quadratico medio(RMSE), errore medio assoluto(MAE), la Receiver Operating Characteristic curve(ROC) ed il coefficiente di Gini.

Per una questione di sintesi efficace, in questo lavoro verrano adoperati due criteri: quello ottenuto tramite matrice di confusione(ACC) e ROC. Il primo perché nella review in [5] è dimostrato essere quello più diffuso ed il secondo perché è proposto nella fonte principale di questo progetto [1].

Curva ROC La curva ROC, chiamata anche diagramma di Lorenz, è un grafico bidimensionale che confronta la sensibilità sulle ordinate con 1-specificità sulle ascisse. La sensibilità rappresenta la percentuale di veri positivi cioè la percentuale di buoni creditori classificati correttamente, mentre la specificità corrisponde alla percentuale di veri negativi cioè la percentuale di cattivi creditori classificati correttamente. La curva è ottenuta tramite interpolazione delle coppie (1-specificità,sensitività) risultanti dall'uso di diverse soglie critiche c, si ricorda infatti che la regione di accettazione di un cliente è $A_1 = \{x | \frac{P(Y=1|x)}{P(Y=0|x)} > c\}$ [1].

Per poter effettuare la comparazione tra modelli, generalmente, viene calcolata l'area sotto la curva ROC, abbreviata come AUC(Area Under Curve). Visto che l'AUC è la porzione di un'area unitaria, il suo valore sta in un range tra 0 ed 1. La curva diagonale rappresenta la strategia di classificare ogni unità in maniera casuale quindi l'obiettivo minimo di un classificatore è quello di avere un AUC maggiore almeno di 0.5, perché significherebbe aver estratto informazione dai dati per effettuare la predizione.

I risultati ottenuti dal modello sono mostrati in Figura 1. Il modello è superiore alle performance di un classificatore random raggiungendo un valore di AUC uguale a 0.7.

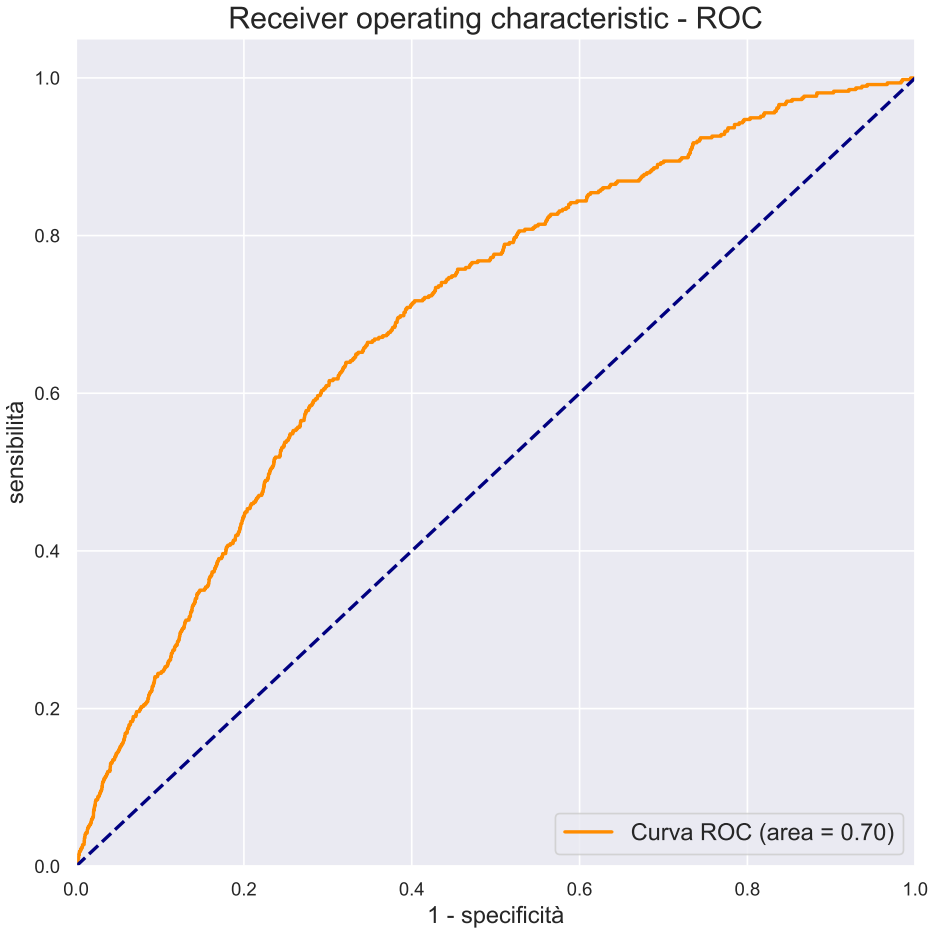

Figure 1: Curva ROC

Per le valutazioni che seguono è stata utilizzata la soglia critica c che massimizza la differenza tra falsi positivi e veri positivi sulla curva ROC.

Matrice di confusione La matrice di confusione è uno dei criteri più usati in ambito di contabilità e finanza, ma anche in altre aree come nel marketing e nella sanità. L'average correct classification(ACC) rate misura la proporzione di casi correttamente classificati come buoni e cattivi creditori. L'ACC è un criterio importante nel valutare le capacità di classificazione dei modelli di scoring e può essere estrapolato da una matrice, detta *matrice di confusione*, che presenta le combinazioni di osservazioni reali e predette in un certo dataset.

Nonostante l'ACC sia il criterio più usato nel credit scoring perché sottolinea l'accuratezza delle previsioni questo non coglie la differenza di costi che deriva dai diversi tipi di errore. La differenza tra errore di classificazione per un buon creditore valutato cattivo e viceversa è ignorata. Molti lavori affermano che i costi associati con errori di tipo II siano molto più alti rispetto ad errori di tipo I. Nel lavoro di West [4], viene messo in risalto che il creatore del dataset tedesco ha considerato gli errori di tipo II cinque volte più costosi degli errori di tipo I.

La matrice di confusione ottenuta dal modello logistico si trova in Figura 2 da cui deriva un valore di $ACC = \frac{VeriPositivi+VeriNegativi}{Totali} = 0.678$

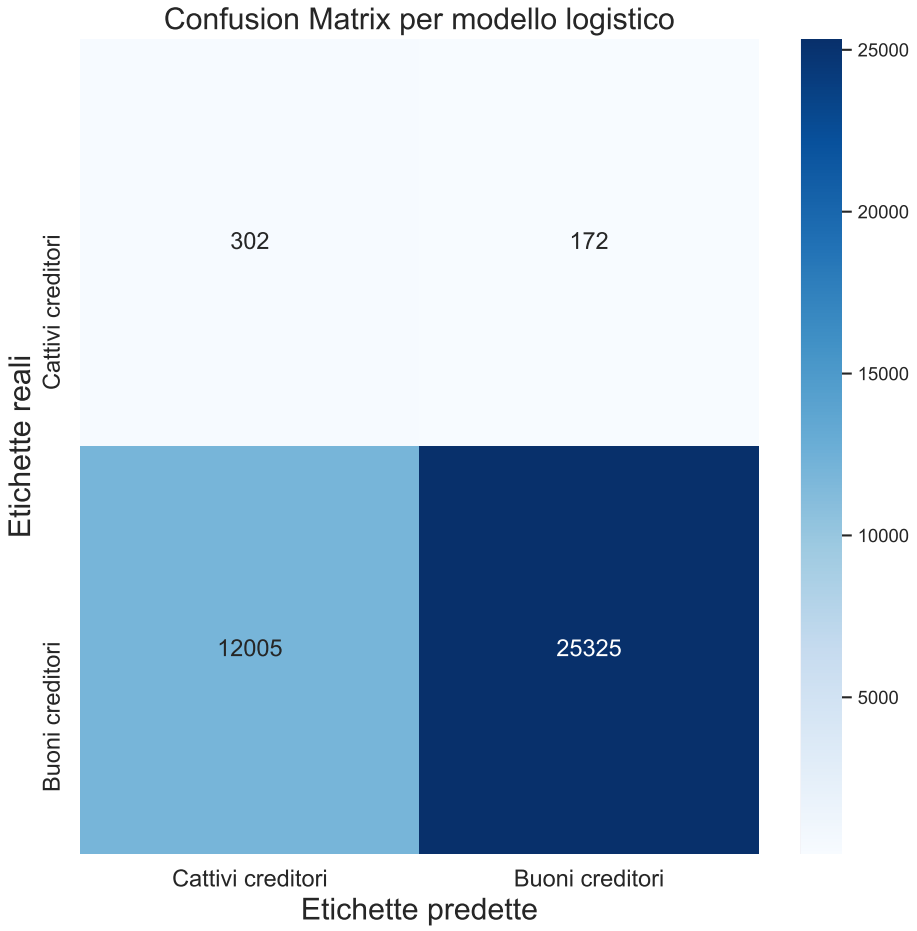

Figure 2: Matrice di Confusione

Confronto con i risultati in letteratura I criteri di valutazione delle capacità predittive usati precedentemente possono avere poco significato se presi in maniera a se stante. Per giudicare la bontà dei risultati del nostro modello può essere utile considerare i risultati

di altri modelli di credit scoring implementati in letteratura. Per effettuare una comparazione equa, viene considerata la review di West [4] che fa uso del modello logistico su diversi dataset, utilizzando gli indicatori AUC, sensitività e specificità.

In Figura 3 vengono riportati i risultati ottenuti da West con Logistic Regression su tre diversi dataset e dal nostro modello:

Autore+Modello+Dataset	AUC	Sensitivity	Specificity
West+LogisticRegr+German	0.7325	0.7750	0.6900
West+LogisticRegr+Japanese	0.8307	0.8280	0.8333
West+LogisticRegr+England	0.6357	0.6546	0.6169
Minici+LogisticRegr+SUS	0.6980	0.6392	0.6784

Figure 3: Confronto con modelli in letteratura

Si può osservare come i valori cambino in maniera abbastanza significativa in base al dataset usato, tuttavia il nostro modello risulta essere migliore per due dei tre parametri rispetto al modello implementato da West sul dataset England. Ciò suggerisce che il nostro modello possa essere un buon punto di partenza per poter costruire un modello di credit scoring basato su Logistic Regression.

Interpretabilità e diritto alla spiegazione Oggigiorno l'incredibile alta accuratezza predittiva dei modelli di Machine Learning ha reso il loro uso indispensabile in una vasta gamma di applicazioni. Tuttavia la loro struttura non lineare li rende poco trasparenti ed è difficile capire quale informazione nell'input guidi la decisione del modello [6]. Questo aspetto risulta rilevante in ambiti come la diagnosi medica, in cui sarebbe poco cauto affidarsi ciecamente alla decisione di un modello, o il credit scoring in cui il cosiddetto "diritto alla spiegazione" obbliga gli istituti di concessione di credito a riportare una lista di ragioni per cui un finanziamento non possa essere erogato [7].

Per stabilire l'importanza predittiva di un modello di Machine Learning è stato proposto il permutation feature importance algorithm da [8].

L'algoritmo prende in input i seguenti elementi: Modello trainato f, matrice di dati X, vettore etichette y, funzione di perdita $L(y, f)$. La procedura è la seguente:

1. Stima l'errore originale del modello $e_{orig} = L(y, f(X))$, in questo caso L è $1 - AUC$.
2. Per ogni feature $j = 1, \ldots, p$:
   - Genera la matrice di dati X_{perm} permutando in maniera casuale la feature j nei dati X. Questo spezza l'associazione tra la feature j e la vera etichetta y.
   - Stima l'errore $e_{perm} = L(y, f(X_{perm}))$ basando le predizioni sui dati permutati.
   - Calcola il permutation feature importance $FI_j = \frac{e_{perm}}{e_{orig}}$
3. Ordina le features in maniera discendente rispetto a FI.

Per avere una stima più accurata, il punto 2 viene ripetuto per ogni feature j un numero $N = 100$ di volte e la media e la deviazione standard del campione di N permutazioni sono mostrate in Figura 4.

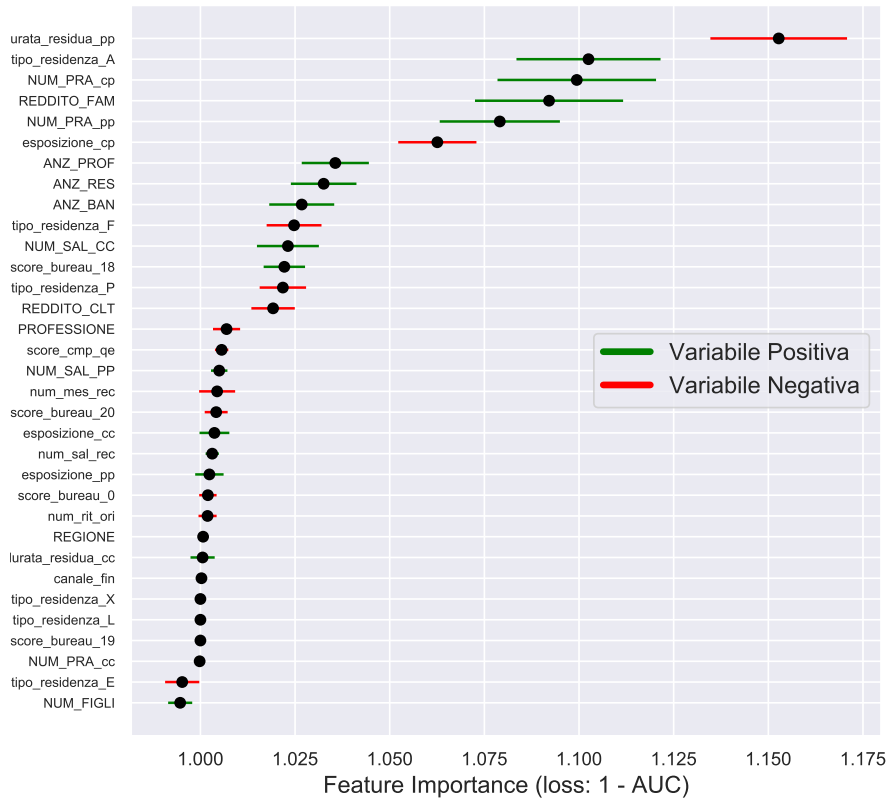

Figure 4: Importanza delle variabili nella predizione

È possibile osservare che tra le variabili più importanti ci sono:
1) Durata residua al saldo dei prestiti personali in corso.
2) Se il cliente è al momento residente sotto contratto di affitto.
3) Numero di prestiti personali in corso.
4) Reddito familiare.

La variabile 1 ha una correlazione negativa con la variabile target(buon pagatore). Il modello logistico pone maggiore importanza sulla durata residua dei prestiti personali del cliente, pare essere sensato visto che più tempo manca alla restituzione del prestito più si potrebbe avere interesse a non ripagarlo ed al contrario quando manca poco tempo risulta più conveniente pagare per non rischiare.

La variabile 2 ha una correlazione positiva con la variabile target. Questo fatto non è facilmente spiegabile, forse ci si aspetterebbe che i clienti con una casa di proprietà siano dei bravi pagatori proprio perché forse il finanziamento è servito all'acquisto della stessa.

La variabile 3 ha una correlazione positiva e anche questo non sembra essere di immediata interpretazione. Forse questa caratteristica deriva da come il dataset è raccolto, generalmente molteplici prestiti vengono prestati solamente ad aziende o imprenditori che, essendo più abili nella loro gestione finanziaria, riescono a saldare il debito.

La variabile 4 ha una correlazione positiva ed è facilmente spiegabile visto che più alto è il reddito più facilmente il finanziamento verrà restituito dal cliente.

References

[1] Stanghellini, E. (2009). Introduzione ai metodi statistici per il credit scoring. Springer Science & Business Media, .

[2] http://www.bee-viva.com/competitions/____sus4

[3] https://en.wikipedia.org/wiki/Equal_Credit_Opportunity_Act

[4] West, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27, 1131-1152.

[5] Abdou, H. & Pointon, J. (2011) 'Credit scoring, statistical techniques and evaluation criteria: a review of the literature ', Intelligent Systems in Accounting, Finance & Management, 18 (2-3), pp. 59-88.

[6] Samech, W., Wiegand, T., & Muller, K. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. Arxiv prePrint:arXiv:1708.08296, .

[7] https://en.wikipedia.org/wiki/Right_to_explanation

[8] Fisher, A., Rudin, C., & Dominici, F. (2018). All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. Arxiv prePrint:arXiv:1801.01489, .

[9] Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. Statistics in Medicine, 16, 185-195.

4 Appendice

Ambito	Variabile	Descrizione	Tipologia
Caratteristiche socio demografiche	AGE	Eta'	Discreta
	REGIONE	Regione di residenza	Qualitativa
	ANZ_BAN	Anzianità del conto corrente	Discreta
	RESIDENZA	Tipologia di residenza	Qualitativa
	ANZ_RES	Anzianità di residenza dell'attuale dimora	Discreta
	STA_CIVILE	Stato civile	Qualitativa
	NUM_FIGLI	Numero figli	Discreta
	SESSO	Sesso	Qualitativa
	REDDITO_CLT	Reddito richiedente	Continua
	REDDITO_FAM	Reddito Famiglia	Continua
	PROFESSIONE	Professione	Qualitativa
	NAZ_NASCITA	Nazione di nascita	Qualitativa
	ANZ_PROF	Anzianità lavorativa	Qualitativa
Equipaggiamento del cliente	CANALE_FIN	Canale di finanziamento	Qualitativa
	NUM_PRA_PP	Prestiti Personali in corso - numero pratiche	Discreta
	esposizione_PP	Prestiti Personali in corso - importo residuo al saldo	Continua
	durata_residua_pp	Prestiti Personali in corso - durata residua al saldo	Continua
	NUM_PRA_CC	Totale prestiti finalizzati in corso - numero pratiche	Discreta
	esposizione_CC	Totale prestiti finalizzati in corso - importo residuo al saldo	Continua
	durata_residua_CC	Totale prestiti finalizzati in corso - durata residua al saldo	Continua
	NUM_PRA_CP	Carta - Cliente in possesso di carta	Discreta
	esposizione_CP	Carta - Esposizione Carta di credito	Continua
Storico del cliente	NUM_SAL_PP	Prestiti personali saldati negli ultimi 24 mesi - numero pratiche	Discreta
	NUM_SAL_CC	Prestiti finalizzati saldati negli ultimi 24 mesi - numero pratiche	Discreta
Comportamento del cliente	num_men_rit	numero ritardi nei pagamenti dell'origine	Discreta
	score_cmp_qe	score comportamentale interno	Continua
	score_cmp_cb	score comportamentale credit bureau	Qualitativa
	num_sal_rec	numero di salite al recupero negli ultimi 12 mesi	Discreta
	num_mes_rec	numero di mesi al recupero negli ultimi 12 mesi	Discreta
TARGET	ClientStatus	1=contenzioso 2=recupero 0=regolare	

Table 1: Descrizione del dataset