RISK MANAGEMENT MAGAZINE

Vol. 16, Issue 1
January – April 2021

EXCERPT

https://www.aifirm.it/newsletter/progetto-editoriale/

Artificial Intelligence: the Application of Machine Learning and Predictive Analytics in Credit Risk

Stefano Bonini and Giuliana Caivano

Artificial Intelligence: the Application of Machine Learning and Predictive Analytics in Credit Risk

Stefano Bonini (Università di Bologna - Università degli Studi di Milano), Giuliana Caivano (Accenture Strategy & Consulting)[1]

Article submitted to double-blind peer review, received on 4th September 2019 and accepted on 20th January 2021

Abstract

In the last years Machine Learning (and the Artificial Intelligence), is experiencing a new rush thanks to the growth of volume and kind of data, the presence of tools / software with higher computational power and cheaper data storage size (e.g. *cloud*). In Credit Risk Management, the PD (Probability of Default) estimation has attracted lots of research interests in the past literature and recent studies have shown that advanced Artificial Intelligence (AI) methods achieved better performance than traditional statistical methods tied to simplified Machine Learning techniques. The study empirically investigates the results of applying different advanced machine learning techniques in estimation and calibration of Probability of Default. The study has been done on big data sample with more than 800,000 Retail customers of a panel European Banks under ECB Supervision, with 10 years of historical information (2006 - 2016) and 300 variables to be analyzed for each customer. The study shows that *neural network* produces a higher population riskiness ranking accuracy, with 71% of Accuracy Ratio. However, the authors' idea is that *classification tree* is more interpretable from an economic and credit point of view. In terms of model calibration, *cluster analysis* produces rating classes more stable and with a predicted risk probability aligned with the observed default rate.

***

Negli ultimi anni il Machine Learning e, più in generale, il mondo dell'Intelligenza Artificiale, sta acquisendo un nuovo slancio grazie alla crescita del volume e della varietà dei dati, a processi di elaborazione / strumenti con elevata potenza computazionale oltre agli spazi per l'archiviazione dei dati sempre più a buon mercato (es. *cloud*). Nell'ambito del Credit Risk Management, la modellizzazione della PD (Probabilità di Default) ha attirato l'interesse accademico nella letteratura passata e recenti studi analizzati dagli autori hanno mostrato che l'applicazione di tecniche di Intelligenza artificiale (IA) avanzate permette di ottenere performance migliori rispetto alla statistica tradizionale legata a tecniche di Machine Learning più semplificate. In questo paper si analizzano empiricamente i risultati derivanti dall'applicazione di diverse tecniche avanzate di machine learning nella stima e calibrazione del parametro di Probabilità di Default. Lo studio è stato condotto su un campione contenente oltre 800.000 clienti Retail di un panel di banche europee sotto supervisione della BCE, con 10 anni di informazioni storiche (2006 - 2016) e 300 variabili da analizzare per ciascun cliente. I risultati mostrano una maggiore accuratezza del ranking della popolazione (in termini di rischiosità) ottenuto attraverso l'applicazione di *reti neurali*, con un valore di Accuracy Ratio (AR) del 71%. È idea degli autori, tuttavia, che al di là delle prestazioni ottenute, l'*albero di classificazione* risulti essere maggiormente interpretabile da un punto di vista economico e creditizio. In termini di calibrazione del modello, l'applicazione della *cluster analysis* genera classi di rating stabili e con una rischiosità stimata allineata alla rischiosità empirica osservata.

Key Words:
Risk Management, Credit Risk, Machine Learning, Big Data, Data Analysis, Advanced Predictive Analytics

DOI 10.47473/2020rmm0081

1 Introduzione

Il tema dell'Intelligenza Artificiale è la buzz-word del momento e il mercato bancario ha iniziato a scoprirla solo negli ultimi anni, ma affonda le sue radici nel passato: basti pensare a tutti quei ricercatori che iniziarono a indagare se fosse possibile l'apprendimento dei computer a partire dai dati. È questo l'assunto alla base del Machine Learning (o apprendimento automatico), ossia che i computer possano imparare ad eseguire dei task semplicemente osservando le relazioni esistenti tra i dati, imparando dai dati con una efficacia tanto maggiore quanto maggiore è la disponibilità di informazioni.

Gli ultimi anni sono stati caratterizzati da una rivoluzione tecnologica e digitale che offre nuove opportunità per il miglioramento e l'efficientamento delle prassi operative e l'adozione di approcci metodologici più avanzati in diversi campi di ricerca. In un contesto sempre più competitivo con riduzione dei margini di profitto il Machine Learning e, più in generale, il mondo dell'Intelligenza Artificiale, sta acquisendo un nuovo slancio grazie alla crescita del volume e della varietà dei dati, a processi di elaborazione / strumenti con elevata potenza computazionale oltre agli spazi per l'archiviazione dei dati sempre più a buon mercato (es. *cloud*). Il Machine Learning, in particolare, può ricoprire un ruolo chiave sia in ambito tecnologico sia di business, consentendo alle istituzioni finanziarie di gestire al meglio grandi moli di dati e facilitare l'adattamento e la ricalibrazione dei modelli.

Negli ultimi anni sono state molte le tecniche di Machine Learning pensate per la stima di variabili binarie, in molti campi della scienza. Nell'ambito del Credit Risk Management, in particolare, la modellizzazione della PD (Probabilità di Default) ha attirato l'interesse accademico nella letteratura passata e recenti studi analizzati dagli autori hanno mostrato che l'applicazione

[1] Le opinioni espresse rappresentano esclusivamente il punto di vista degli autori e non riflettono necessariamente quello dell'Istituto/Azienda d'appartenenza.

di tecniche di Intelligenza artificiale permette di ottenere performance migliori rispetto alla statistica tradizionale sia nell'applicazione ai problemi di credit scoring ([16], [27]) sia nella stima della Probabilità di Default (cfr. [3], [4], [7], [8], [20]).

Obiettivo principale di questo studio è evidenziare la rilevanza della scelta degli algoritmi, dei parametri, la selezione delle variabili (caratteristiche) rilevanti, il ruolo dei criteri di valutazione e l'importanza del contributo esperto nella definizione della Probabilità di Default di un portafoglio di prestiti. Lo studio presentato nel paper, inoltre, cerca di rispondere anche ad una serie di quesiti tipici che emergono quando si ha a che fare con l'applicazione di algoritmi e tecniche statistiche avanzate: come risolvere o raggiungere un determinato obiettivo?

Dalla letteratura analizzata emergono diversi paper che applicano tecniche di Machine Learning a un campione di dati con elevato numero di osservazioni e relativo ad un solo intermediario finanziario. Il nostro paper si differenzia dalla letteratura esistente per una serie di motivazioni: mentre la letteratura accademica analizzata si focalizza sulla creazione di una misura cardinale dell'evento default (cfr. [25]) utilizzando algoritmi di classificazione generalizzati e tecniche alberi di classificazione, il nostro studio è focalizzato, da una parte, sulla capacità di ordinamento / ranking, in particolare su come diversi algoritmi di machine learning e deep learning prevedono l'evento default. In aggiunta, il paper si focalizza anche sulle variabili selezionate da ciascun algoritmo come drivers di rischio e, infine, analizza il potere di calibrazione delle stime ottenute tramite l'utilizzo di tecniche di tipo non supervisionato per la definizione delle classi di rating.

Il nostro studio si basa sull'applicazione di algoritmi già utilizzati in letteratura, ad es. in [10] gli autori utilizzano alberi decisionali, regressione logistica e random forest per l'analisi del livello di delinquency dei consumatori utilizzando dati relativi a sei differenti banche. In [23] gli autori applicano alberi decisionali ad un portafoglio mutui per prevedere l'evento default e confrontano i risultati con tecniche k-nearest neighbors (KNN), reti neurali artificiali (ANN) e modelli probit.

L'utilizzo di modelli classici di regressione (logistica e lineare, cfr. [12]) è ben noto nel mondo bancario, pertanto nel nostro studio abbiamo utilizzato la regressione logistica come benchmark e comparato la sua capacità di fitting (in termini di *Accuracy Ratio* e *Tasso di corretta classificazione*) con quella di altri modelli non parametrici annoverati tra le tecniche di machine learning e deep learning nella letteratura più recente. Abbiamo utilizzato tre approcci: alberi di classificazione, random forest e deep learning (rete neurale) applicandoli ad un campione contenente oltre 800.000 clienti Retail di un panel di banche europee sotto supervisione della BCE, con 10 anni di informazioni storiche (2006 - 2016) per valutare non solo la capacità di fitting di ciascun modello rispetto alla regressione logistica ma anche la combinazione di variabili selezionate da ciascun modello.

Il paper è cosi strutturato: nella Sezione 2 è presentata una descrizione delle principali logiche metodologiche sottostanti agli algoritmi utilizzati nell'ottica di illustrarne il funzionamento rispetto all'obiettivo / variabile target da modellizzare. La Sezione 3 illustra i principali criteri utilizzati per la classificazione e il confronto dei risultati mentre la Sezione 4 descrive il dataset utilizzato per lo studio empirico e i principali risultati ottenuti. Infine, la Sezione 5 fornisce le conclusioni dello studio e una vista delle possibili evoluzioni della ricerca.

2 Aspetti teorici delle metodologie più diffuse in letteratura[2]

In questa sezione sono descritti gli algoritmi di machine learning utilizzati per:

a) costruire un modello di ranking / scoring della popolazione utilizzata (*apprendimento supervisionato*);

b) calibrare lo score e definire quale probabilità di default associare a ciascuna classe di rating (*apprendimento non supervisionato*).

2.1 Apprendimento supervisionato per la definizione dello scoring

Il primo obiettivo – costruire un modello di ranking della popolazione – si presta ad essere formulato come un problema di apprendimento supervisionato che rappresenta una delle tecniche di machine learning più utilizzate in letteratura.

Nel framework del *supervised learning*, un *"learner"* si presenta con coppie di input / output dai dati storici in cui i dati di input rappresentano attributi pre-identificati per essere utilizzati nel definire il valore dell'output. I dati di input sono comunemente rappresentati come vettori e, in funzione dell'algoritmo di apprendimento scelto, possono consistere in valori continui e/o discreti con o senza dati mancanti. L'apprendimento supervisionato risolve un problema di tipo *regressivo* quando l'output è una variabile continua, di *classificazione* quando l'output è una variabile discreta.

Una volta presentati i dati di input / output, il compito del *learner* è trovare una funzione che mappi correttamente i vettori di input verso i valori di output, ad esempio memorizzando tutte le precedenti coppie di valori input / output. Anche se questo metodo mappa correttamente le coppie di valori di input / output nel campione di training, è improbabile che il modello funzioni nel prevedere i valori di output se i valori di input sono diversi da quelli contenuti nel dataset di training o se il dataset di training contiene *"noise"*. In questo contesto, la sfida dell'apprendimento supervisionato è trovare una funzione che generalizzi oltre il dataset di training, in modo che la stessa sia in grado di mappare accuratamente input verso output *out-of-sample*.

[2] Parte dei dettagli metodologici qui illustrati è estratta dal position paper AIFIRM #14 "*Intelligenza Artificiale: l'applicazione di Machine Learning e Predictive Analytics nel Risk Management*"

Ad esempio, nel caso specifico del nostro studio, l'output del modello è una variabile continua tra 0 e 1 che può essere interpretata (sotto certe condizioni) come una stima della probabilità di un cliente di andare in default entro i 12 mesi successivi date certe caratteristiche del cliente stesso e/o del prodotto in oggetto.

Nel nostro studio abbiamo costruito diversi modelli di ranking / scoring della popolazione utilizzando e confrontando tra loro, in particolare, quattro approcci di apprendimento supervisionato:

1. Regressione logistica;
2. Albero decisionale (CART);
3. Random Forest;
4. Rete neurale (*deep learning*).

La regressione logistica è un'estensione del modello di regressione lineare in cui la relazione lineare alla base di quest'ultimo modello è aggiustata attraverso una trasformazione esponenziale, chiamata trasformazione logistica. In particolare, la regressione logistica analizza la relazione tra multiple variabili indipendenti e una singola variabile dipendente dicotomica - nel caso dello studio in oggetto la variabile "*good*" /"*bad*" – tramite la stima di un punteggio di probabilità e con l'obiettivo di discriminare al massimo i due gruppi individuati dalla variabile dicotomica.

$$y_i = f(w_i) = \frac{1}{1+e^{-w_i}} \quad (1)$$

Dove la variabile indipendente w_i è data dalla funzione lineare degli indicatori selezionati:

$$w_i = \alpha + \sum_{j=1}^{m} \beta_j x_{i,j} \quad (2)$$

Combinando le equazioni definite e aggiungendo il termine di errore, si ottiene il modello logit come:

$$y_i = \frac{1}{1+e^{-\alpha-\sum_j \beta_j x_{i,j}}} + \varepsilon_i \quad (3)$$

Il campo di valori generati dalla funzione logistica (il " codominio " della funzione) è ora limitato all'intervallo (0,1). Ciò garantisce che la variabile dipendente y_i sia sempre compresa tra 0 e 100% e può pertanto essere correttamente interpretata come una probabilità di default.
Modelli lineari come la regressione logistica hanno l'indubbio vantaggio di produrre buoni risultati laddove l'aspettativa sia di avere funzioni lineari e di essere comprensibile e spiegabile. Di contro, tale modello non gestisce in maniera efficiente le variabili categoriche e la presenza di elevata correlazione tra le variabili può generare problemi; inoltre le performance possono risentire in caso di variabili non lineari e spesso si osserva una propensione all'*underfitting*.

Logiche diverse sono invece alla base dei modelli *CART (Classification and Regression Trees)*, ossia un insieme di tecniche di stima molto utilizzate nell'ambito del Machine Learning, applicabili sia a problemi di classificazione sia di regressione (cfr. [31]) e in cui una variabile dipendente o output (discreta o continua) è legata ad un insieme di variabili indipendenti (o di input) attraverso una sequenza ricorsiva di semplici relazioni binarie (da qui il riferimento ad "albero"). L'insieme delle relazioni ricorsive divide lo spazio multidimensionale delle variabili indipendenti in distinte "aree" in cui la variabile dipendente è tipicamente assunta, nel caso di un albero regressivo come nello studio in oggetto, come legata linearmente alle variabili indipendenti con parametri univoci per ciascuna "area".

In *Figura 1* è rappresentato un modello CART con due variabili indipendenti non negative (x_1, x_2) anche note come "*feature vector*" e una variabile dipendente discreta che assume due valori, "*good*" e "*bad*". La sequenza di relazioni ricorsive binarie rappresentate nell'albero in *Figura 1* suddivide il dominio di (x_1, x_2) in cinque aree distinte determinate dai parametri L_1 ….L_4.

In particolare, questo modello implica che tutti i valori di (x_1, x_2) con $x_1 < L_1$ e $x_2 < L_2$ sono associati a un outcome "bad" e tutti i valori di (x_1, x_2) con $x_1 < L_1$ e $x_2 >= L_2$ sono associati a un outcome "good".

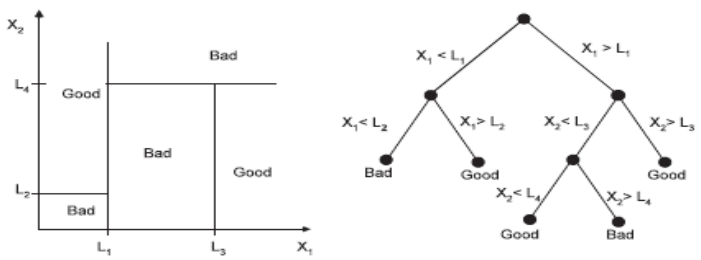


Figura 1 – Esempio di albero regressivo con variabile target binaria

I parametri (Lj) sono scelti per minimizzare, in ciascun passaggio, la distanza tra la variabile dipendente e i valori fittati all'interno di ciascuna categoria e massimizzare invece quella tra le diverse categorie. Questi due vincoli vengono incorporati nella formula della funzione obiettivo:

$$D = \sum_{i=1}^{K} \left\{ \sum_{j \in S_i} \left(y_i - \widehat{\beta}_i \right)^2 \right\} = \sum_i D_i, \tag{4}$$

dove y_i e $\widehat{\beta}_i$ sono rispettivamente i valori della variabile target (nel caso dello studio in oggetto 0 e 1) e il parametro ad essi associato presenti all'interno di uno dei K sottospazi dei dati S_i. In ciascun passaggio il processo si ripete, cercando tra i dati dei sottoinsiemi il valore di soglia per la variabile in grado di ottenere la divisione ottimale. Questo processo viene iterato fino a quando non si verificano determinate condizioni che ne determinano l'arresto. Una di queste cause può essere, ad esempio, la creazione di un sottospazio di dati aventi la stessa categoria della variabile target (nel *Classification Tree*) o che possiedono gli stessi valori (nel *Regression Tree*).

Quando l'algoritmo genera un albero particolarmente fitto e complesso, composto da molti rami e foglie, il risultato può risultare scarsamente interpretabile, per l'elevato numero di tagli e per la tendenza al sovradattamento dei dati (*overfitting*). È pertanto necessario ridurre l'albero tramite una procedura automatica chiamata "potatura": una tecnica che, partendo dal modello completamente sviluppato, elimina sequenzialmente i rami non utili ai fini della stima o con la minore carica informativa. Definendo la funzione di perdita come segue:

$$C_\alpha(K) = \sum_{i=1}^{K} D_i + \alpha K \tag{5}$$

dove K è la dimensione dell'albero considerato in ogni singolo passo e α il parametro associato al costo computazionale del modello, verrà ad ogni passo rimossa la foglia la cui eliminazione comporta il minore aumento della funzione obiettivo $\sum_{i=1}^{K} D_i$. La procedura continua fino a quando il valore di $C_\alpha(K)$ sarà stabilizzato.

Quello degli alberi decisionali è un algoritmo molto utilizzato nella pratica e spesso citato nella letteratura poiché presenta grandi vantaggi rispetto alle altre tecniche di Machine Learning. Risulta infatti uno dei modelli più informativi, grazie alla sua alta semplicità logica che permette di comunicare facilmente le regole alla base della sua struttura, mettendo in evidenza quali sono i principali driver implicati nella stima. Correlato a questo fatto, gli alberi risultano essere un ottimo metodo automatico di riduzione della dimensionalità dei dati, selezionando soltanto le variabili più importanti ai fini dell'approssimazione dei dati. Un altro vantaggio di questo modello è la sua ridotta complessità computazionale anche quando la mole di osservazioni e il numero di variabili è molto alto: proprio per questo, i CART sono spesso utilizzati come strumento base di varie tecniche di combinazione di stimatori. In generale, gli alberi decisionali sono facili da interpretare e da costruire ma uno degli svantaggi principali è la loro tendenza all'overfitting e sono fortemente dipendenti dalle caratteristiche del dataset di training.

Una diretta evoluzione dei modelli CART sono le tecniche di *Random Forest*. Una Random Forest è uno speciale classificatore formato da un insieme di classificatori semplici (Alberi Decisionali), rappresentati come vettori random indipendenti e identicamente distribuiti, dove ognuno di essi contribuisce per la classe più popolare in input (cfr. [34]. Ciascun albero all'interno di una Random Forest è costruito e addestrato a partire da un sottoinsieme casuale dei dati presenti nel training set: gli alberi pertanto non utilizzano quindi il set completo, e ad ogni nodo non viene più selezionato l'attributo migliore in assoluto, ma viene scelto l'attributo migliore tra un set di attributi selezionati casualmente. La casualità è un fattore che entra quindi a far parte della costruzione dei classificatori e ha lo scopo di accrescere la loro diversità e diminuirne così la correlazione. Il risultato finale restituito dalla Random Forest è che la media dei risultati di ciascun albero (nel caso di utilizzo per il *forecasting*), o la classe restituita dal maggior numero di alberi nel caso la Random Forest sia utilizzata a fini di *clustering*.

In letteratura le Random Forest ottengono risultati estremamente consistenti nelle stime probabilistiche (cfr. [9], [27], [28], [32]) e sono spesso state oggetto di confronto con i metodi parametrici classici [cfr. [21] testandoli su diversi tipi di dati. Rispetto al singolo albero decisionale, tuttavia, risulta meno intuitivo e facile da spiegare e può risultare complicata la calibrazione dei parametri nel tempo.

A rappresentare, infine, un ampio insieme di tecniche *machine learning* sono le reti neurali (*neural network*). Il termine *neural network* nasce come modellizzazione matematica di quello che in passato si riteneva essere il meccanismo di funzionamento del cervello animale (cfr. [23]). Una rete neurale è sostanzialmente uno schema di regressione non lineare (cfr. [25]) a due o più stadi ([1], [47]) costituito da strati di neuroni che, collegati tra loro da ideali bottoni sinaptici, mettono in relazione le variabili di input con quelle di output. Il neurone, in sostanza, è interpretabile come una funzione matematica (definita funzione primitiva) delle variabili esplicative ([10],[12], [18]). Il processo di apprendimento della rete neurale consiste nell'identificare i coefficienti delle funzioni di rete – sigmoidi - ([22], [17],[14]) che legano tra loro i neuroni (ed esprimono pertanto le relazioni che intercorrono tra le variabili di input a quelle di output) attraverso la minimizzazione di una funzione obiettivo ([13], [11]) espressa come scarto quadratico medio tra il valore reale dell'output ed il valore calcolato ([33], [6]).

Pur riuscendo a catturare le relazioni non-lineari e non-monotòne che intercorrono tra la PD e le variabili esplicative, tali modelli presentano numerosi inconvenienti: arbitrarietà nella scelta di molti parametri e soprattutto difficoltà di interpretazione dei risultati (spesso vengono indicati come black box).

2.2 Apprendimento non supervisionato

A differenza dell'apprendimento supervisionato, quello non supervisionato prevede l'utilizzo di dati non strutturati o senza etichetta. Le tecniche di apprendimento non supervisionato, in particolare, consentono di osservare la struttura dei dati e di estrapolare informazioni cariche di significato. In queste tecniche non esiste però una variabile o una funzione obiettivo note a priori, a differenza di quanto accade invece nell'apprendimento supervisionato. Nel nostro studio abbiamo utilizzato logiche di clusterizzazione basate su un algoritmo *K-means*. Tale algoritmo (cfr. [30]) è una metodologia di clustering che permette di suddividere N osservazioni in K cluster, nei quali ciascuna osservazione appartiene al cluster avente il punto medio a questa più prossimo: tale obiettivo viene raggiunto dalla metodologia minimizzando la varianza totale intra-cluster. Esprimendo il concetto in termini formali: dato un insieme di osservazioni $(x_1, x_2, \dots, x_N)$, dove ciascun elemento può essere rappresentato da un vettore reale a d dimensioni, il K-means clustering ha lo scopo di partizionare le N osservazioni in K ($\leq$ N) insiemi $S = \{S_1, S_2, \dots, S_K\}$ in modo da minimizzare la varianza espressa dalla WCSS (Within-Cluster Sum of Squares).

In termini matematici, l'obiettivo è il seguente:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2 = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{K} |S_i| \operatorname{Var} S_i \qquad (6)$$

Dove μ_i è la media dei punti in S_i.
L'algoritmo standard impiega una tecnica iterativa di aggiustamento: dato un insieme iniziale di K medie $m_1^{(1)}, \dots, m_k^{(1)}$, la procedura evolve alternando le due fasi seguenti:

I. *Fase di Assegnazione (Assignment step):* viene assegnata ciascuna osservazione al cluster la cui media è caratterizzata dalla distanza euclidea minima. Matematicamente significa partizionare le osservazioni impiegando un diagramma di Voronoi (Voronoi diagram) generato dalle medie.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \right\} \forall j, i \qquad (7)$$

Dove ciascun x_p è assegnato ad uno ed un solo $S^{(t)}$

II. *Fase di aggiornamento (Update step):* sono calcolate le nuove medie che costituiranno i centroidi delle osservazioni nel nuovo cluster:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \qquad (8)$$

L'algoritmo converge quando non avvengono più cambiamenti significativi alla configurazione trovata. Rispetto alla formulazione base, sono state proposte numerose varianti atte ad incrementare le performance di ricerca del metodo. Tra le più popolari si ricorda il *K-medians* clustering, il *K-means* ++ e la soft *k-means* (detta anche *Fuzzy C-means*).

3 Applicazione del Machine Learning al Rischio di Credito: stima della Probabilità di Default

L'idea alla base del nostro studio è stata quella di utilizzare le tecniche multivariate di machine learning supervisionato sopra citate per arrivare alla quantificazione del merito creditizio (Probabilità di default – PD) della clientela a fini di erogazione / concessione di nuovo credito, sfruttando poi l'efficacia della cluster analysis per giungere ad una rappresentazione discreta (scala di rating) del merito creditizio di ciascun cliente.

3.1 Campione di dati utilizzato

Lo studio in oggetto è stato condotto su un campione contenente oltre 800.000 clienti Retail di un panel di banche europee vigilate BCE, con 10 anni di informazioni storiche (da gennaio 2006 a dicembre 2016). A livello di cliente si è studiato il potere informativo e predittivo di un set di dati riconducibile alle seguenti fonti informative:

- *Credit Bureau:* sono state analizzate informazioni afferenti alle seguenti categorie di prodotti:
  - *Carte:* Importo residuo utilizzo carta, numero di contratti attivi, numero di carte in possesso;
  - *Prodotti non rateali:* numero di contratti attivi, importo accordato, importo utilizzato, importo sconfinato;

- *Prodotti rateali:* importo rata mensile, importo rate residue, importo rate scadute e non pagate, numero totale di contratti attivi;
- *Dati complessivi (banca e sistema):* numero di banche affidatarie a sistema, numero contratti attivi presso l'istituto, numero di contratti attivi a sistema, totale importi scaduti non pagati, Score Credit Bureau, presenza sofferenze a sistema, presenza di protesti a sistema, presenza di pregiudizievoli a sistema, presenza di note negative.

- *Prodotto*: a livello di prodotto sono state utilizzate informazioni relative a importo rata mensile, rapporto rata / reddito, importo richiesto, valore dell'immobile, grado dell'ipoteca e tipologia di immobile (in caso di mutuo), durata e scopo del finanziamento;
- *Informazioni socio-demografiche:* Nazionalità, Area geografica di residenza, Anni di residenza presso l'indirizzo attuale, Anzianità bancaria, Anzianità lavorativa, Età, Tipo contratto di lavoro del richiedente (tempo determinato, indeterminato, etc.), Tipo controparte (persona fisica, cointestazione, garante affidato, etc.), Professione, SAE, Situazione abitativa (es. proprietà, affitto…), Stato civile, Possesso carta di credito (aggiuntiva), reddito netto annuo da lavoro, reddito netto annuo (comprensivo di altri redditi), possesso di immobili.

3.2 Costruzione vettori di input

Per ciascuna mese di riferimento del campione (nel periodo compreso tra gennaio 2006 e dicembre 2016) è stata analizzata la dinamica dei passaggi a default (past-due a 90 giorni, inadempienze probabili e sofferenza) delle pratiche erogate in ciascun mese nei 12 mesi successivi, costruendo in tal modo la variabile target.

Dato l'obiettivo principale del nostro studio, ossia di trovare – attraverso metodologie di Machine Learning diverse – le migliori combinazioni tra le informazioni sopra citate nel prevedere l'evento default, si riportano di seguito alcune analisi grafiche finalizzate a mostrare la relazione esistente tra l'andamento delle singole variabili e il tasso di default sull'intero campione utilizzato. Tali analisi sono riportate – a titolo esemplificativo - solo per le variabili ritenute più esplicative per ciascuna area informativa considerata.

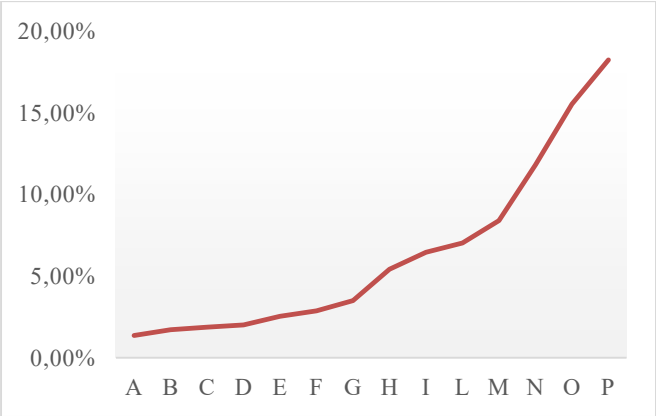


Figura 2 – Distribuzione CBScore vs Tasso di default

Come evidenziato dall'analisi grafica, lo score Credit Bureau mostra – in linea con le aspettative, un trend positivo rispetto al tasso di default: all'aumentare della classe di score / rating, aumenta la rischiosità osservata.

Figura 3 – Distribuzione importi scaduti vs Tasso di default

Anche nel caso dell'importo totale dei rapporti scaduti.

Figura 4 – Distribuzione rapporto rata / reddito vs Tasso di default

Per selezionare il set di variabili da sottoporre alle tecniche multivariate di machine learning supervisionato sopra menzionate, le variabili iniziali a disposizione sono state sottoposte ai "classici" trattamenti di normalizzazione legati alla gestione dei valori mancanti e al trattamento di valori anomali. Nello specifico tutte le variabili con percentuale di valori mancanti superiore al 15% sono state escluse a priori dal processo.

La selezione delle singole variabili è stata poi effettuata combinando l'analisi grafica sopra riportata con una regressione logistica e l'imposizione dei seguenti vincoli predefiniti per ciascuna variabile analizzata:

- Coerenza del segno del coefficiente con il senso economico atteso tra la variabile e il tasso di default;
- Significatività statistica del coefficiente (*p-value* inferiore al 5%);
- Capacità predittiva di ciascuna variabile misurata attraverso un Accuracy Ratio superiore al 10%.

Le variabili cosi selezionate sono state poi sottoposte ad un'analisi di correlazione, eliminando pertanto quelle con correlazione superiore a |0.5|.

La tabella seguente riporta le variabili che sono state utilizzate ai fini della costruzione del modello multivariato.

Tabella 1 – Input finali utilizzati per la selezione dei modelli multivariati di Machine Learning

Input del modello	
Sociodemografici	***Credit Bureau***
Nazionalità	**Carte**
Area geografica di residenza	Importo residuo utilizzo carta
Anni di residenza presso l'indirizzo attuale	Numero contratti attivi
Anzianità bancaria	Numero di carte in possesso
Anzianità lavorativa	**Prodotti non rateali**
Età	Numero contratti attivi
Numero garanti collegati al rapporto principale	Importo accordato
Numero componenti della pratica (co-obbligati e garanti)	Importo sconfinato
Tipo contratto di lavoro del richiedente (tempo determinato, indeterminato, etc.)	Importo utilizzato
Tipo NDG (persona fisica, cointestazione, garante affidato, etc.)	**Prodotti rateali**
Professione	Importo rate mensilizzate
SAE	Importo rate residue
Situazione abitativa (es. proprietà, affitto…)	Importo rate scadute e non pagate
Stato civile	Numero totale contratti attivi
Carta di credito - carta aggiuntiva	**Dati di sistema**
Reddito netto annuo da lavoro	Numero di banche affidatarie a sistema

Reddito netto annuo (comprensivo di altri redditi)	Numero contratti attivi presso l'istituto
Possesso immobili	Numero contratti attivi a sistema
Informazioni di prodotto	Totale importi scaduti non pagati
Importo rata mensile	Score Credit Bureau
Rapporto rata / reddito	Presenza Sofferenza a sistema
Importo richiesto	Presenza protesti a sistema
Valore dell'immobile	Presenza pregiudizievoli a sistema
Durata del finanziamento	Presenza di note negative
Grado dell'ipoteca	
Scopo del finanziamento	
Tipologia immobile	

3.3 Modelli multivariati di Machine Learning: principali evidenze

In questa sezione descriviamo le evidenze derivanti dall'applicazione degli algoritmi di machine learning utilizzate per costruire i modelli di previsione del default sul nostro campione di pratiche erogate tra gennaio 2006 e dicembre 2016 su un portafoglio di controparti Retail derivanti da un panel di banche vigilate ECB.
Come già illustrato nella sezione 2.1, la costruzione di modelli di previsione probabilità di default è un tipico problema di apprendimento supervisionato, che rappresenta una delle tecniche di machine learning più utilizzate. Nel framework di apprendimento supervisionato, un "*learner*" è rappresentato da coppie di valori di input / output sui dati storici dove i dati di input rappresentano gli attributi predefiniti da utilizzare per determinare il valore di output. I dati di input sono comunemente rappresentati come un vettore e, in funzione dell'algoritmo di apprendimento, possono consistere in valori continui e/o discreti con o senza dati mancanti. Il problema dell'apprendimento supervisionato è un tipico problema "regressivo" quando l'output è continuo, di "classificazione" quando l'output invece è di natura discreta. Obiettivo del "*learner*" è trovare una funzione che mappi correttamente i vettori di input rispetto ai valori di output. Un possibile approccio per questo mapping è memorizzare tutti i precedenti valori di coppie di input / output. Anche se questo approccio mappa correttamente le coppie di valori input / output nel dataset di training, è poco probabile che funzioni nella previsione dei valori di output se i valori di input sono diversi da quelli presenti nel dataset di training o quando quest'ultimo contiene "*noise*". Pertanto, l'obiettivo dell'apprendimento supervisionato è trovare una funzione che generalizzi al di là del dataset di training, cosi che la funzione trovata possa mappare accuratamente coppie di input / output anche su campioni out-of sample.
L'output del nostro modello è una variabile continua con valori tra 0 e 1 che può essere interpretata (sotto certe condizioni) come una stima della probabilità di andare in default nei successivi 12 mesi di vita di un contratto, date specifiche variabili di input.

Definizione del modello multivariato

Per la costruzione del modello previsivo abbiamo costruito e confrontato tra loro tre algoritmi Machine Learning:

- *Rete neurale* a tre strati, basata sull'algoritmo di *backpropagation*, completamente connessa e feed-forward;
- *Modello CART* che utilizza nella "fase di potatura" l'indice di Gini come funzione obiettivo per la riduzione dell'albero;
- *Modello Random Forest*.

Tali metodologie sono state scelte in quanto maggiormente diffuse in letteratura e i risultati a cui si è pervenuti, in termini di performance e capacità predittive sono stati confrontati con quanto invece ottenuto con il tradizionale approccio di regressione logistica (anch'esso, ricordiamo, annoverabile tra le tecniche di Machine Learning di tipo supervisionato).
Si riportano di seguito le performance ottenute:

Tabella 2 – Performance metodologie di Machine Learning

Metodologia	Accuracy Ratio	CCR
Rete Neurale	71%	86%
Random Forest	68%	81%
Albero di classificazione	66%	79%
Regressione Logistica	66%	77%

Dal confronto tra i modelli, eseguito in termini di Accuracy Ratio (AR) e Correct Classification Rate (CCR), il ranking prodotto dalle reti neurali rappresenta il modello migliore, con un valore di AR pari al 71% e di CCR pari all' 86%.

Le Random Forest hanno prodotto performance leggermente inferiori e paragonabili a quelle degli alberi di classificazione, con valori di AR rispettivamente pari a 68% e 66% e di CCR del 81% e 79%.

È nostra idea tuttavia che le reti neurali abbiano dato un risultato così poco superiore perché siamo partiti da indicatori standard. Un algoritmo di rete riuscirebbe a fare la differenza nel momento in cui andassimo ad aggiungere nuove informazioni con dati anche meno strutturati.

Molto importante è comunque sottolineare, al di là delle prestazioni ottenute, che l'albero di classificazione risulta essere, tra i tre approcci di machine learning "avanzati", l'approccio maggiormente interpretabile da un punto di vista economico e creditizio, mentre gli altri due non permettono una buona e diretta comprensione dei risultati e dei legami tra le variabili di input e quella di output.

È questo il motivo per cui si è privilegiato il ranking ottenuto con gli alberi di classificazione ai fini della calibrazione delle PD.

3.4 Calibrazione del modello

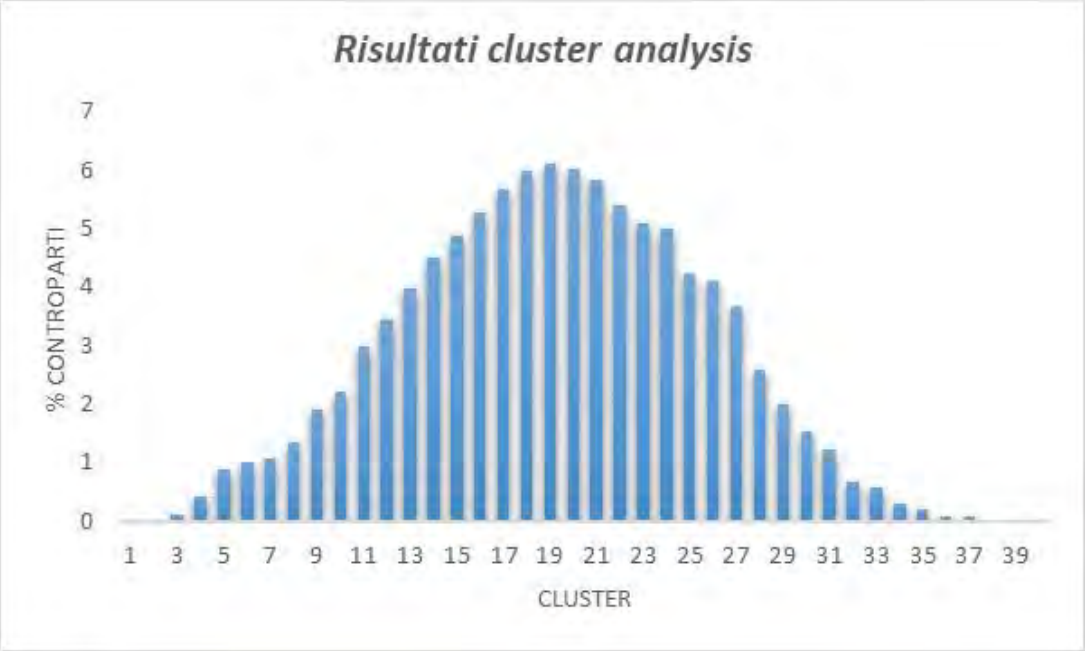
L'ultimo step di un modello di stima di PD è rappresentato dalla calibrazione degli score ai fini della loro trasformazione in PD passando attraverso la creazione di classi di rating.

Abbiamo pertanto sottoposto gli scores derivanti dal modello multivariato identificato dall'albero di classificazione a una calibrazione di tipo bayesiano, ancorando gli scores alla Central Tendency di lungo periodo e infine identificando la rating *scale* più adeguata e compliant con i requisiti normativi.

La creazione delle scale di rating è stata fatta ricorrendo ad un approccio di machine learning di tipo *unsupervised*, in particolare la **clusterizzazione *k-means*** con l'applicazione dei seguenti parametri:

- *Set iniziale dei parametri:* identificazione di 40 cluster iniziali, scala finale con massimo 11 classi di rating;
- *Split dei cluster:* concentrazione superiore al 15%.

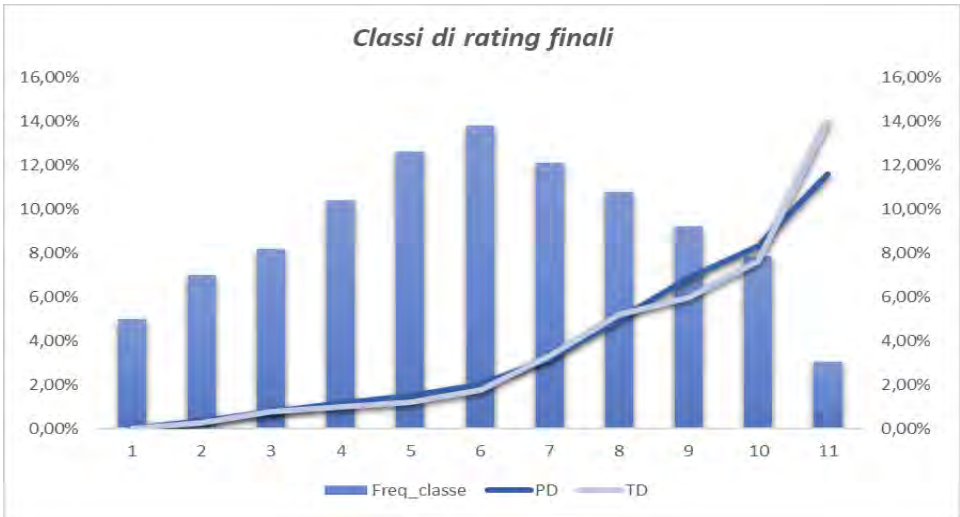
Figura 5 – Distribuzione popolazione per cluster



I cluster sopra identificati sono stati sottoposti ad un algoritmo combinatorio iterativo finalizzato a valutare:

- La forma corretta della scala di rating (simmetria e forma della campana);
- La presenza di concentrazione in ciascuna classe di concentrazione non elevata;
- Risultato test di calibrazione (Binomiale e *chi-square*);
- Monotonicità del trend di PD / Tasso di default.

Figura 6 – Classi di rating finale con applicazione Machine Learning



4 Conclusioni

Il paper ha mostrato in che modo diverse metodologie di Machine Learning possono essere applicate all'interno del framework complessivo di stima della Probabilità di default. Abbiamo in particolare comparato le tecniche più comunemente utilizzate per la modellizzazione della PD con tecniche più avanzate (Rete Neurale, Albero di classificazione e Random Forest). Per far questo abbiamo utilizzato un campione di dati molto ampio (2006-2016) basato su dati panel di diverse banche europee sotto supervisione della BCE, composto da oltre 800.000 clienti Retail e un rilevante numero di indicatori da analizzare per ciascun cliente. L'albero di classificazione, pur mostrando capacità predittiva leggermente inferiore a Random Forest e Reti Neurali, è stato considerato molto più interpretabile e pertanto utilizzato per l'ultimo step dell'applicazione: la creazione di classi di rating attraverso un algoritmo di *k-means*. È nostra idea che le reti neurali abbiano dato un risultato così poco superiore perché siamo partiti da indicatori standard. Un algoritmo di rete riuscirebbe a fare la differenza nel momento in cui andassimo ad aggiungere nuove informazioni con dati anche meno strutturati.
Un ulteriore sviluppo di questa ricerca è rappresentato dall'applicazione di ulteriori tecniche di machine learning al campione, eventualmente estendendo l'analisi anche a un portafoglio Corporate.

Stefano Bonini and Giuliana Caivano

Bibliography

[1] ALTMAN, E., BARBOZA, F., KIMURA, H. (2017). *Machine learning models and bankruptcy prediction*. Expert Systems with Applications 83: 405-417

[2] ADDO P., GUEGAN D., HASSANI B. (2018). *Credit Risk Analysis using Machine and Deep Learning*. Documents de travail du Centre d' Economie de la Sorbonne

[3] ANTUNES F., RIBEIROA B., PEREIRA F. (2017). *Probabilistic modeling and visualization for bankruptcy prediction.* Applied Soft Computing 60: 831–843.

[4] ARMINGER G., KRUPPA J., SCHWARZ A., ZIEGLER, A. (2013). *Consumer credit risk: Individual probability estimates using machine learning.* Expert Systems with Applications.

[5] AUGUSTIN L., BOULESTEIX A., T STROBL C. (2007). *Unbiased split selection for classification trees based on the Gini index.* Computational Statistics & Data Analysis.

[6] BHAVSAR H., PANCHAL M. H. (2012*). A review on support vector machine for data classification.* International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(10), pp-185.

[7] BONINI S., CAIVANO G. (2013). *Survival analysis approach in Basel2 Credit Risk Management: modelling Danger Rates in Loss Given Default parameter.* Journal of Credit Risk, 9 (1).

[8] BONINI S., CAIVANO G. (2016). *Estimating loss-given default through advanced credibility theory.* The European Journal of Finance 22 (13).

[9] BREIMAN L. (2001). *Random forests.* Machine Learning 45 (1): 5-32.

[10] BUTARU F., QUINGQUING C., BRIAN C., SANMAY D., ANDREW W. L., and AKTARE S. (2016). *Risk and risk management in the credit card industry.* Journal of Banking and Finance 72: 218–39.

[11] CHAUDHURI A., DE K. (2011). *Fuzzy support vector machine for bankruptcy prediction.* Applied Soft Computing 11: 2472–2486.

[12] CHEN C., SCHWENDER H., KEITH J., NUNKESSER R., MENGERSEN K., MACROSSAN P. (2011). *Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression.* IEEE/ACM Trans Computer Biol Bioinform 8: 1580–1591.

[13] CHEN S., HARDLE W., MORO R.A. (2006). *Estimation of Default Probabilities with Support Vector Machines.* SFB 649, Economic Risk, Berlin, SFB 649 discussion paper, No. 2006-077.
[14] CHEN M.L., TSAI C.F. (2010). *Credit rating by hybrid machine learning techniques.* Applied Soft Computing
[15] DE GROOT J. (2016). *Credit risk modeling using a weighted support vector machine.* Mathematical Institute, UTRECHT University, Master Thesis
[16] DWYER D.W., STEIN R.M. (2006). *Inferring the default rate in a population by comparing two incomplete default databases.* Journal of Banking & Finance 30: 797–810.
[17] FONSECA P., LOPES H. (2017). *Calibration of Machine Learning Classifiers for Probability of Default Modelling.* James Finance, Crowd Process Inc.
[18] GHODSELAHI A. (2011). *A hybrid support vector machine ensemble model for credit scoring.* International Journal of Computer Applications, 17(5), 1-5.
[19] HARDLE W., MORO R.A., SCHAFER D. (2005). *Predicting Bankruptcy with Support Vector Machines.* SFB 649, Economic Risk, Berlin, SFB 649 discussion paper, No. 2005-009.
[20] HARDLE W., MORO R.A., HOFFMANN L. (2010). *Learning Machines Supporting Bankruptcy Prediction*. SFB 649, Economic Risk, Berlin, SFB 649 discussion paper, No. 2010-032
[21] HERNÁNDEZ-LOBATO D., SHARMANSKA V., KERSTING K., LAMPERT C.H., QUADRIANTO N. (2014). *Mind the nuisance: Gaussian process classification using privileged noise.* Advances in Neural Information Processing Systems 27: 837–845.
[22] HOSMER D. W., LEMESHOW S. (2000). *Applied logistic regression.* 2nd ed. New York: Wiley.
[23] HUANG Z., CHEN H., HSU C.J., CHEN W.B., WU S. (2004). *Credit rating analysis with support vector machines and neural networks: a market comparative study.* Decision Support Systems 37: 543-558.
[24] KABIR M.J., KANG B.H., LIU Y., WASINGER R., ZHAO Z., XU S. (2015). *Investigation and improvement of multi-layer perception neural networks for credit scoring.* Expert Systems with Applications.
[25] KHANDANI A.E., KIM J., LO A.W. (2010). *Consumer credit-risk models via machine-learning algorithms.* Journal of Banking & Finance.
[26] KHASHMAN A. (2010). *Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes.* Expert Systems with Applications.
[27] LESSMANN S., BAESENS B., SEOW H. V., THOMAS L.C. (2015). *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.* European Journal of Operational Research 247 (1): 124–136.
[28] LIAW A., WIENER M. (2002). *Classification and regression by random forest.* R News 2 (3): 18-22.
[29] LOH W.Y. (2011). *Classification and regression trees.* WIREs Data Mining Knowledge Discovering 1: 14–23.
[30] MIN J. H., LEE Y. (2005). *Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters.* Expert Systems with Applications, Vol. 28, Issue 4, pp.603-614.
[31] STEINBACH M., TAN P.N. (2009). kNN: *k-nearest neighbors.* In: Wu, X., Kumar, V. (eds.). The top ten algorithms in data mining. Chapman & Hall/CRC: 151–162.
[32] STEINBERG D. (2009). *CART: classification and regression trees*. In: Wu, X., Kumar, V. (eds.). The top ten algorithms in data mining. Chapman & Hall/CRC: 180–201.
[33] STROBL C., BOULESTEIX A., AUGUSTIN L. T. (2007). *Unbiased split selection for classification trees based on the Gini index.* Computational Statistics & Data Analysis 52 (1): 483-501.
[34] TSAI C. F., CHEN M. L. (2010). *Credit rating by hybrid machine learning techniques.* Applied soft computing, 10(2), 374-380.
[35] WRIGHT M. N., ZIEGLER A. (2017). *Ranger: A fast implementation of random forests for high dimensional data in C++ and R.* Journal of Statistical Software 77:1-17