POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Tecniche di machine learning per la valutazione del rischio di default: previsione multi-periodale

Relatrice
Prof.ssa Patrizia Semeraro

Candidato
Alessandro Baronti

Anno Accademico 2024-2025

Sommario

Negli ultimi decenni, la valutazione del rischio di credito ha assunto un ruolo sempre più centrale nel settore finanziario, in particolare nelle istituzioni bancarie e società di credito. La capacità di prevedere correttamente il default di un cliente o di un gruppo di clienti riveste un'importanza cruciale per la gestione del rischio e per l'allocazione delle risorse. In questo contesto, la previsione per un cliente singolo viene solitamente effettuata utilizzando uno strumento statistico chiamato Regressione Logistica. Solo in tempi più recenti si è assistito a un crescente interesse verso l'utilizzo di tecniche di Machine Learning, che permettono una modellizzazione più flessibile e, in alcuni casi, più performante. La maggior parte della letteratura attuale si è concentrata proprio sull'identificazione del metodo di apprendimento più efficace per migliorare l'accuratezza delle previsioni di default.

Questa tesi si propone di analizzare il problema della previsione del default dei clienti di una banca attraverso l'utilizzo di tecniche di machine learning, adottando un modello comunemente utilizzato nella pratica, ovvero il modello misto scambiabile di Bernoulli. I principali parametri che prendiamo in considerazione sono la probabilità di default p, il coefficiente di equicorrelazione tra default ρ, e una misura di rischio, chiamata Value at Risk, associata ad un dato portfolio di clienti. A questo viene aggiunta un'analisi della dimensione temporale, per vedere come evolvono i parametri al variare del tempo e dell'orizzonte di previsione. Per le nostre analisi, è stato utilizzato il dataset *Default of Credit Card Clients*, reso disponibile dalla piattaforma Kaggle. Kaggle è una piattaforma online di professionisti e appassionati di data science e machine learning che consente di accedere a dataset pubblici, partecipare a competizioni, e condividere soluzioni. Il dataset utilizzato contiene informazioni su 30.000 clienti di una banca di Taiwan, raccolte tra aprile e settembre 2005, e comprende variabili demografiche, finanziarie e comportamentali, oltre alla storia dei pagamenti e dei saldi dei clienti nei mesi considerati. Le tecniche di classificazione supervisionata impiegate nelle

nostre analisi sono la regressione logistica (LR), il multi–layer–perceptron (MLP), k–nearest neighbors (KNN) e AdaBoost (AB). L'obiettivo del lavoro è duplice: da un lato, valutare quale tra i modelli considerati fornisce le migliori prestazioni in termini di previsione del default, secondo alcune metriche di valutazione standard (score F1, AUC e accuratezza), dall'altro analizzare l'evoluzione dei parametri di rischio al variare dell'orizzonte di previsione. La tesi è strutturata nei seguenti capitoli.

Il primo capitolo si suddivide in quattro sezioni. Nella prima, viene fornita una panoramica dei metodi di machine learning impiegati per la previsione dei default, oltre ad una descrizione delle metriche di performance che verranno utilizzate per valutare i modelli. Nella seconda sezione, definiremo il modello misto scambiabile di Bernoulli, e presenteremo alcuni risultati asintotici che si applicano a portafogli di grande dimensione. Nella terza sezione chiariremo come intendiamo estendere il modello per considerare anche la dimensione temporale. Infine, nella quarta sezione, definiremo i concetti di misura del rischio e di Value at Risk (VaR), strumenti che utilizzeremo per valutare il rischio associato a un determinato portafoglio.

Il secondo capitolo è dedicato alla descrizione del dataset impiegato per le analisi. In questa sezione, spiegheremo come il dataset verrà modificato e adattato per i nostri scopi.

Nel terzo e ultimo capitolo, presenteremo i risultati ottenuti attraverso l'applicazione dei modelli e delle tecniche descritti nei capitoli precedenti.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Descrizione del modello

1.1 Modelli di Machine Learning

In questa sezione descriviamo brevemente le tecniche di Machine Learning (ML) che utilizzeremo per prevedere i default. Per una trattazione più approfondita dei metodi presentati si veda Deisenroth, Faisal, Ong (2020) e James, Witten, Hastie, Tibshirani, (2021).

1.1.1 Regressione Logistica

La Regressione Logistica (LR) è un modello lineare generalizzato, in cui le variabili target $Y_1, \ldots, Y_n \in \{0,1\}$ sono categoriche e possono assumere due valori (nel nostro caso, non default e default). L'obiettivo del modello è di stabilire la probabilità con cui un'osservazione può generare uno o l'altro valore della variabile dipendente, e può essere anche utilizzato per classificare le osservazioni in base ai predittori $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Nel modello LR, la probabilità che una variabile Y_i assuma valore 1, dato il vettore di regressori $\mathbf{X}_i = \mathbf{x}_i$ è definita come:

$$p_i = \mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\beta^\top \mathbf{x}_i}}{1 + e^{\beta^\top \mathbf{x}_i}} = \frac{1}{1 + e^{-\beta^\top \mathbf{x}_i}}, \tag{1.1}$$

dove β è un vettore di parametri. Risolvendo l'equazione precedente rispetto all'esponenziale otteniamo una migliore comprensione del significato dei parametri β_j:

$$e^{\beta^\top \mathbf{x}} = \frac{p_i}{1-p_i}.$$

Applicando il logaritmo a entrambi i membri e aggiungendo un termine d'errore otteniamo il modello statistico:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta^\top \mathbf{x} + \epsilon.$$

Il logaritmo del rapporto di probabilità è noto come funzione logit.

A questo punto, al fine di stimare i coefficienti di regressione β dati un insieme di osservazioni $\mathbf{x_i}$ e $Y_i \in \{0,1\}$, per $i = 1, \ldots, n$, utilizziamo il metodo della massima verosimiglianza.

La variabile target Y_i, che può assumere valori y_i in $\{0,1\}$, può essere interpretata come una variabile casuale di Bernoulli, la cui funzione di densità è:

$$\mathbb{P}(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$

Dalla formula 1.1 sappiamo che p_i dipende dai regressori $\mathbf{x}_i$ e dal vettore di parametri β.

Assumendo indipendenza degli errori, le osservazioni sono indipendenti tra loro e la funzione di verosimiglianza è data dal prodotto delle funzioni di densità:

$$L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-\beta^\top \mathbf{x}_i}}\right)^{y_i} \left(\frac{e^{-\beta^\top \mathbf{x}_i}}{1+e^{-\beta^\top \mathbf{x}_i}}\right)^{1-y_i}.$$

La massimizzazione della funzione di verosimiglianza L rispetto al vettore di coefficienti β può essere semplificata calcolando il logaritmo della funzione stessa. Applicando le proprietà del logaritmo, otteniamo:

$$\begin{aligned}\mathcal{L} = \log L &= \sum_{i=1}^{n} \log\left(\frac{1}{1+e^{-\beta^\top \mathbf{x}_i}}\right) y_i + \sum_{i=1}^{n} \log\left(\frac{e^{-\beta^\top \mathbf{x}_i}}{1+e^{-\beta^\top \mathbf{x}_i}}\right)(1-y_i) \\ &= \sum_{i=1}^{n} y_i\, \beta^\top \mathbf{x}_i - \log(1+e^{\beta^\top \mathbf{x}_i}). \end{aligned} \tag{1.2}$$

L'obiettivo del modello LR è quindi quello di trovare il vettore di paramentri β che massimizzi la log–verosimiglianza (un metodo è quello di applicare l'algoritmo di discesa del gradiente).

1.1.2 Multi Layer Perceptron

Il Multi-Layer Perceptron è una rete neurale di tipo feedforward (senza ricorsioni o cicli) composta da una sequenza di strati di neuroni completamente connessi.

Un MLP composto da H strati nascosti aventi rispettivamente $n_1, \ldots, n_H$ neuroni, è rappresentato da una funzione $\hat{\boldsymbol{F}} : \mathbb{R}^{n_0} \to \mathbb{R}^m$ tale che

$$\hat{\boldsymbol{F}}(\mathbf{x}) : \mathbb{R}^{n_0} \xrightarrow{\mathcal{L}^{(1)}} \mathbb{R}^{n_1} \xrightarrow{\mathcal{L}^{(2)}} \cdots \xrightarrow{\mathcal{L}^{(H)}} \mathbb{R}^{n_H} \xrightarrow{\mathcal{L}^{(H+1)}} \mathbb{R}^m \tag{1.3}$$

e in particolare

$$\hat{\boldsymbol{F}}(\mathbf{x}) = \boldsymbol{\sigma}^{(H+1)}\left(W^{(H+1)}\boldsymbol{\sigma}^{(H)}\left(\cdots\left(W^{(2)}\boldsymbol{\sigma}^{(1)}\left(W^{(1)}\mathbf{x}+\boldsymbol{b}^{(1)}\right)+\boldsymbol{b}^{(2)}\right)\cdots\right)+\boldsymbol{b}^{(H+1)}\right), \tag{1.4}$$

dove

- $W^{(h)}$ è la matrice dei pesi del livello h
- $\boldsymbol{b}^{(h)} \in \mathbb{R}^m$ è il vettore dei bias del livello h
- $\boldsymbol{\sigma}^{(h)}$ è la funzione di attivazione del livello h, ed è una funzione vettoriale che applica elemento per elemento la funzione $\sigma^{(h)}$.

Ogni funzione $\mathcal{L}^{(h)}(\mathbf{x})$ è quindi definita come segue:

$$\mathcal{L}^{(h)}(\mathbf{x}) := \boldsymbol{\sigma}^{(h)}\left(W^{(h)}\mathbf{x}+\boldsymbol{b}^{(h)}\right).$$

E' utile notare come la funzione $\hat{\boldsymbol{F}}$ risulti essere la composizione di $\boldsymbol{\sigma}^{(h)}$ con funzioni affini del tipo $A\mathbf{x} + \mathbf{b}$, iterata più volte. Nelle applicazioni pratiche le funzioni di attivazione più utilizzate, a seconda anche dello scopo del MLP, sono le seguenti:

- Sigmoide: Definita come $\sigma(x) = \frac{1}{1+e^{-x}}$
- Tanh (Tangente iperbolica): Data da $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
- ReLU (Rectified Linear Unit): Definita come $\mathrm{ReLU}(x) = \max(0, x)$
- Softmax: definita come:

$$\mathrm{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}.$$

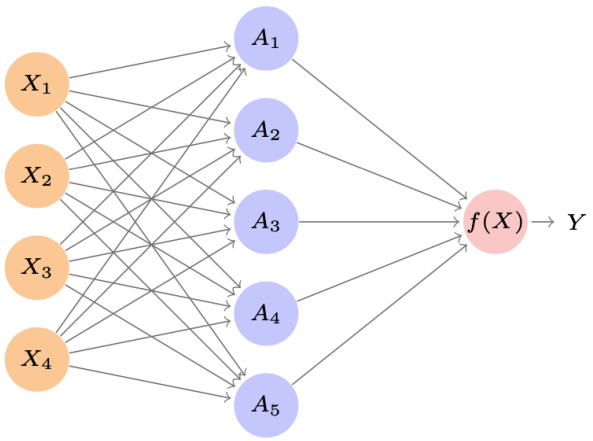

Figura 1.1. Esempio di Multi–layer perceptron con un solo strato nascosto.

Nella figura 1.1 riportiamo una possibile rappresentazione grafica di un Multi–layer perceptron a valori reali, con uno strato di input avente 4 neuroni e uno strato nascosto con 5 neuroni $A_1, \ldots, A_5$. Nello specifico, ogni neurone nascosto viene calcolato come segue:

$$A_k = \sigma^{(1)}(\sum_{j=1}^{4} w_{kj}^{(1)} X_j + b_k^{(1)});$$

l'output finale Y sarà poi dato da:

$$\sigma^{(2)}(\sum_{j=1}^{5} w_{kj}^{(2)} A_j + b^{(2)}).$$

Alla funzione F è poi associata poi una funzione di perdita (Loss), la quale misura l'errore che viene fatto nel classificare un dato $\mathbf{x}_i$. Una scelta comune per la funzione di Loss è la somma degli errori quadratici, data da Loss $:= \sum_{i=1}^{n} \left(y_i - \hat{F}(\mathbf{x}_i)\right)^2$ (dove n è il numero di osservazioni, e y_i è la variabile target associata all'osservazione $\mathbf{x}_i$). L'addestramento della rete neurale consiste nel cambiare i valori di pesi e bias in modo tale da minimizzare la funzione di Loss. Questo problema può essere risolto con l'algoritmo di discesa del gradiente, o con sue varianti.

1.1.3 K-Nearest Neighbors

L'algoritmo dei k-Nearest Neighbors (KNN) è tra i più semplici metodi di apprendimento automatico. L'idea alla base del modello KNN è che il modello memorizzi tutti gli esempi di addestramento e successivamente cerchi di prevedere la classe di una nuova osservazione sulla base delle classi dei suoi vicini più prossimi nel training set. Tale metodo si basa quindi sull'assunzione che le caratteristiche usate per descrivere i punti del dominio siano rilevanti per le classi, in modo tale che punti vicini tendano ad avere la stessa classe.

Supponiamo che il dominio delle osservazioni, $\mathcal{X}$, sia dotato di una metrica ρ. In particolare, $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ è una funzione che restituisce la distanza tra due elementi qualsiasi di $\mathcal{X}$. Per esempio, se $\mathcal{X} = \mathbb{R}^d$, allora ρ potrebbe essere la distanza euclidea:

$$\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\| = \sqrt{\sum_{i=1}^{d} (x_i - x_i')^2}.$$

A seconda della tipologia di dati a disposizione, esistono anche metriche molto utilizzate, come la distanza di Hamming (utile per vettori binari), la distanza di Manhattan (che misura la somma delle differenze in valore assoluto tra vettori di numeri reali), e la distanza di Minkowski, che generalizza le precedenti. In breve, il

concetto dell'algoritmo KNN è il seguente: dato un intero positivo K e una nuova osservazione $\mathbf{x}_0$, il classificatore KNN individua dapprima i K punti del training set più vicini a $\mathbf{x}_0$, rappresentati con $\mathcal{N}_0$. Successivamente, stima la probabilità condizionata della classe j come la frazione di punti in $\mathcal{N}_0$ la cui classe è uguale a j:

$$\mathbb{P}(Y = j \mid \mathbf{X} = \mathbf{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbb{I}(y_i = j), \tag{1.5}$$

dove $\mathbb{I}$ è la funzione indicatrice.

Infine, il KNN classifica l'osservazione $\mathbf{x}_0$ nella classe con probabilità stimata maggiore.

1.1.4 Ada Boost

L'algoritmo Ada Boost riceve in input un training set di osservazioni $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Il processo di boosing procede in una serie di iterazioni. Al passo t viene decisa una distribuzione di probabilità sull'insieme S, ovvero un vettore $D^{(t)} \in \mathbb{R}^n_+$ tale che $\sum_{i=1}^n D_i^{(t)} = 1$. A questo punto, viene addestrato un classificatore "debole" h_t, sul training set S, con lo scopo di minimizzare l'errore pesato ϵ_t:

$$\epsilon_t = \sum_{i=1}^{n} D_i^{(t)} \mathbb{I}_{[h_t(\mathbf{x}_i) \neq y_i]},$$

dove $\mathbb{I}$ è la funzione indicatrice.

L'algoritmo Ada Boost assegna al classificatore h_t un peso inversamente proporzionale al suo errore ϵ_t:

$$w_t = \frac{1}{2} \log(\frac{1}{\epsilon_t} - 1).$$

Alla fine dell'iterazione t, l'algoritmo aggiorna la distribuzione $D^{(t)}$ dando una massa di probabilità maggiore ai campioni classificati erroneamente da h_t, e una massa minore a quelli su cui h_t non sbaglia. Questo forzerà il classificatore debole dell'iterazione successiva a focalizzarsi sulle osservazioni problematiche.

Infine, l'output dell'algoritmo è un classificatore costruito con una somma pesata (tramite i pesi w_t) dei classificatori deboli.

1.1.5 Metriche di valutazione dei modelli

In questa sezione ricordiamo brevemente le principali metriche di valutazione di un classificatore. Se definiamo come positivo il caso di default e come negativo quello di non default, quando andiamo a valutare un modello sul test set le possibilità sono quattro:

- TP (true positives): clienti classificati correttamente come default, e che effettivamente sono andati in default.
- TN (true negatives): clienti classificati correttamente come non in default, e che effettivamente non sono andati in default.
- FP (false positives): clienti classificati come default, ma che in realtà non sono andati in default.
- FN (false negatives): clienti classificati come non default, ma che in realtà sono andati in default.

Una volta effettuata la classificazione, è possibile valutare la perfomance di un modello tramite diverse metriche. Tra le più comuni troviamo:

- Accuratezza: misura la proporzione di classificazioni corrette (sia positive che negative) sul totale delle osservazioni:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$
- Precisione: indica la proporzione di veri positivi sul totale delle osservazioni classificate come positive:
$$\text{Precision} = \frac{TP}{TP + FP}.$$
- Recall: rappresenta la proporzione di veri positivi sul totale dei positivi reali:
$$\text{Recall} = \frac{TP}{TP + FN}.$$
- Score F1: è la media armonica tra precisione e recall, utile per tenere conto di entrambe:
$$\text{Score F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Un altro strumento molto utilizzato per valutare le performance di un classificatore è la curva ROC (Receiver Operating Characteristic). Si tratta di un grafico che mostra come variano il tasso di veri positivi (TPR, o recall) e il tasso di falsi positivi (FPR) al variare della soglia di classificazione. Formalmente, tali quantità sono definite come:

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}.$$

La AUC (Area Under the Curve) è una metrica associata alla curva ROC che misura l'area sotto di essa. Il suo valore è compreso tra 0 e 1, dove un AUC pari a 0,5 indica una classificazione casuale, mentre un AUC pari a 1 corrisponde a un classificatore perfetto. In sintesi, maggiore è l'AUC, migliore è la capacità del modello di distinguere tra classi positive e negative.

1.2 Modelli misti di Bernoulli

In questa sezione descriviamo un possibile modello per gli eventi di default di un gruppo di $d \in \mathbb{N}$ debitori, che non considera il tempo.
Assumiamo l'esistenza di uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. Sia $\mathbf{Y} = (Y_1, \dots, Y_d)'$ il vettore aleatorio degli indicatori di default, i.e. per $i = 1, \dots, d$, $Y_i = 1$ se il cliente i va in default, altrimenti $Y_i = 0$. Il vettore $\mathbf{Y}$ si dice *scambiabile* se la distribuzione congiunta di $\mathbf{Y}$ è invariante per permutazioni, ovvero se per ogni σ permutazione di $\{1, \dots, d\}$ abbiamo che $(Y_1, \dots, Y_d)'$ e $(Y_{\sigma(1)}, \dots, Y_{\sigma(d)})'$ sono uguali in distribuzione. Questo implica in particolare che ogni cliente abbia la stessa probabilità di andare in default, e si presta quindi bene a modellare il concetto di portafoglio *omogeneo*.
Di seguito diamo la definizione di modello misto di Bernoulli:

Definizione 1 *Dato un intero $p < d$ e un vettore aleatorio p-dimensionale $\mathbf{\Psi} = (\Psi_1, \dots, \Psi_p)'$, il vettore aleatorio $\mathbf{Y} = (Y_1, \dots, Y_d)'$ segue un modello misto di Bernoulli con fattore $\mathbf{\Psi}$ se esistono funzioni $p_i : \mathbb{R}^p \to [0, 1]$, per $1 \leq i \leq d$, tali che, condizionatamente a $\mathbf{\Psi}$, le componenti di $\mathbf{Y}$ siano variabili casuali di Bernoulli indipendenti, che soddisfano*

$$\mathbb{P}(Y_i = 1 \mid \mathbf{\Psi} = \boldsymbol{\psi}) = p_i(\boldsymbol{\psi}).$$

Data questa definizione, per $\mathbf{y} = (y_1, \ldots, y_d)' \in \{0,1\}^d$, abbiamo che:

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{\Psi} = \boldsymbol{\psi}) = \prod_{i=1}^{m} p_i(\boldsymbol{\psi})^{y_i}(1 - p_i(\boldsymbol{\psi}))^{1-y_i}. \tag{1.6}$$

La distribuzione non condizionata del vettore degli indicatori di default $\mathbf{Y}$ si ottiene integrando rispetto alla distribuzione del vettore dei fattori $\mathbf{\Psi}$. In particolare, la probabilità di default del cliente i è data da

$$\bar{p}_i = \mathbb{P}(Y_i = 1) = \mathbb{E}(p_i(\mathbf{\Psi})).$$

In molte applicazioni pratiche è utile considerare un modello a un fattore singolo. L'informazione necessaria per calibrare un modello a più fattori può non essere sempre disponibile e i modelli a un fattore sono più veloci e facili da implementare. In questo caso Ψ è una variabile aleatoria a valori in $\mathbb{R}$, e le funzioni $p_i(\Psi) : \mathbb{R} \to [0,1]$ sono tali per cui, condizionatamente a Ψ, il vettore di indicatori di default $\mathbf{Y}$ è un vettore di variabili casuali di Bernoulli indipendenti, con $\mathbb{P}(Y_i = 1 \mid \Psi = \psi) = p_i(\boldsymbol{\psi})$.
Un'ulteriore semplificazione è quella di assumere che le funzioni p_i siano identiche. In questo caso il modello misto di Bernoulli si dice *scambiabile*, dato che il corrispondente vettore di indicatori di default è scambiabile.

Definizione 2 *Data una variabile aleatoria Q con supporto in* [0,1]*, il vettore aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ segue un modello misto di Bernoulli scambiabile con mixing variable Q se, condizionatamente a Q, il vettore di indicatori di default $\mathbf{Y}$ è un vettore di variabili di Bernoulli indipendenti e identicamente distribuite, con $\mathbb{P}(Y_i = 1|Q) = Q$*

Nella definizione abbiamo fondamentalmente assunto che $p_i(\Psi) = p(\Psi) = Q$, per ogni $1 \leq i \leq d$.

Se $\mathbf{Y}$ segue un modello misto di Bernoulli scambiabile, la probabilità di default del cliente i è data da:

$$\mathbb{P}(Y_i = 1) = \int_0^1 q dF(q),$$

dove $F(q)$ è la distribuzione di Q. Inoltre, la densità discreta $p_{\mathbf{Y}}(\mathbf{y})$ di $\mathbf{Y}$ è:

$$p_{\mathbf{Y}}(\mathbf{y}) = \mathbb{P}(\mathbf{Y} = \mathbf{y}) = \int_0^1 q^{\sum_{i=1}^d y_i}(1-q)^{d-\sum_{i=1}^d y_i} dF(q).$$

Possiamo definire, per ogni $1 \leq k \leq d$, il momento π_k come:

$$\pi_k := \mathbb{E}[Y_{i_1} \dots Y_{i_k}] = \mathbb{P}(Y_{i_1} = 1, \dots, Y_{i_k} = 1), \{i_1, \dots, i_k\} \subseteq \{1, \dots, d\}\,.$$

Se definiamo $p := \mathbb{P}(Y_i = 1)$ la probabilità marginale di default, abbiamo che $p = \pi_1$, mentre la correlazione ρ tra due indicatori di default Y_i, Y_j è data da

$$\rho = \frac{\pi_2 - p^2}{p(1-p)}.$$

Abbiamo inoltre che

$$\pi_k = \mathbb{E}[Y_1 \dots Y_k] = \mathbb{E}[\mathbb{E}[Y_1 \dots Y_k|Q]] = \mathbb{E}[Q^k], \tag{1.7}$$

ovvero le probabilità di default non condizionate possono essere viste come i momenti della mixing-variable Q.
Se definiamo $S := \sum_{i=1}^d Y_i$ il numero di default, in questo modello la distribuzione di S è

$$p_S(k) = \binom{d}{k} \int_0^1 q^k (1-q)^{d-k} dF(q). \tag{1.8}$$

Al fine di poter applicare gli algoritmi di machine learning, nel nostro modello assumeremo che Q sia una funzione h di delle variabili aleatorie predittive $\mathbf{X} = (X_1, \dots, X_n)$, che rappresentano le caratteristiche del cliente.
Avremo quindi che le realizzazioni di Q sono funzione delle realizzazioni di $\mathbf{X}$, ovvero $q = h(\mathbf{x})$, e quindi se due clienti i e j hanno covariate $\mathbf{x}_i, \mathbf{x}_j$, avranno probabilità condizionata di default rispettivamente $h(\mathbf{x}_i), h(\mathbf{x}_j)$.

1.2.1 Modello beta–mixing

Nel seguito delle nostre analisi assumeremo un modello misto di Bernoulli scambiabile utilizzato spesso nella pratica, ovvero il modello *beta–mixing*, in cui la variabile

Q segue una distribuzione Beta(a, b). In questo contesto è possibile calcolare analiticamente i valori di π_k e la distribuzione del numero di default S, mentre in altri modelli spesso è necessaria una approssimazione numerica degli integrali in 1.7 e 1.8.

Ricordiamo che, per definizione, la densità di una distribuzione Beta è data da

$$g(q) = \frac{1}{\beta(a,b)} q^{a-1}(1-q)^{b-1}, \quad a, b > 0,\ 0 < q < 1,$$

dove $\beta(a, b)$ denota la funzione beta. Ricordiamo inoltre il fatto che la funzione beta soddisfa la seguente formula di ricorrenza:

$$\beta(a+1, b) = \left(\frac{a}{a+b}\right) \beta(a, b).$$

Usando (1.7), otteniamo per i momenti π_k:

$$\pi_k = \frac{1}{\beta(a,b)} \int_0^1 q^k q^{a-1}(1-q)^{b-1}\, dq = \frac{\beta(a+k, b)}{\beta(a,b)}, \quad k = 1, 2, \ldots$$

La formula di ricorrenza per la funzione beta fornisce poi:

$$\pi_k = \prod_{j=0}^{k-1} \frac{a+j}{a+b+j};$$

in particolare:

$$p = \frac{a}{a+b}, \quad \pi_2 = p \cdot \frac{a+1}{a+b+1}, \quad \rho = \frac{1}{a+b+1}.$$

Abbiamo inoltre che il numero di default $S = \sum_{i=1}^{d} Y_i$ segue una cosiddetta distribuzione beta–binomiale. In particolare, otteniamo da 1.8 che:

$$\mathbb{P}(S = k) = \binom{d}{k} \cdot \frac{1}{\beta(a,b)} \int_0^1 q^{k+a-1}(1-q)^{d-k+b-1}\, dq = \binom{d}{k} \cdot \frac{\beta(a+k, b+d-k)}{\beta(a,b)}.$$

1.2.2 Comportamento asintotico per portafogli di grandi dimensioni

In questa sezione forniamo alcuni risultati asintotici per portafogli di grandi dimensioni nei modelli misti di Bernoulli.

In particolare, abbiamo che nei modelli a un solo fattore, la coda della distribuzione del numero di default S è essenzialmente determinata dalla coda della distribuzione della mixing–variable Q. La scelta di una specifica distribuzione parametrica per Q si riflette quindi sulla distribuzione di S. Tuttavia, come mostreremo in seguito, se fissiamo i primi due momenti di Q (o, equivalentemente, se fissiamo i valori di p e ρ), le differenze tra distribuzioni diverse diventano rilevanti solo oltre il 99–esimo percentile.

Questo risultato implica che la scelta specifica della distribuzione di Q sia di minore importanza rispetto alle stime di p e ρ.

Al fine di mostrare come la coda della distribuzione di S sia determinata dalla coda della distrubuzione di Q, consideriamo una successione infinita $(e_i)_{i\in\mathbb{N}}$ di esposizioni deterministiche positive, e la corrispondente sequenza $(Y_i)_{i\in\mathbb{N}}$ di indicatori di default. Sia inoltre $(\delta_i)_{i\in\mathbb{N}}$ una sequenza di variabili casuali con valori in (0,1] che rappresentano le perdite percentuali in caso di default. In questo contesto, la perdita (loss) per un portafoglio di dimensione d è data da:

$$L^{(d)} = \sum_{i=1}^{d} L_i, \qquad L_i = e_i \delta_i Y_i,$$

dove le L_i sono le perdite individuali.

Formuliamo ora alcune ipotesi tecniche per il nostro modello:

(I1) Esiste un vettore casuale $\mathbf{\Psi}$ p-dimensionale e funzioni $\ell_i : \mathbb{R}^p \to [0,1]$ tali che, condizionatamente a $\mathbf{\Psi}$, la sequenza $(L_i)_{i\in\mathbb{N}}$ è composta da variabili casuali indipendenti con media:

$$\ell_i(\boldsymbol{\psi}) = \mathbb{E}(L_i \mid \mathbf{\Psi} = \boldsymbol{\psi}).$$

Questa ipotesi estende la struttura di indipendenza condizionata dagli indicatori di default alle perdite L_i. Notiamo che non si assume che le perdite date il default δ_i e gli indicatori di default siano indipendenti, ipotesi che si fa in molti modelli standard. In particolare, (I1) consente che Y_i e δ_i siano solo condizionatamente indipendenti dati $\mathbf{\Psi}$, tali che $\ell_i(\boldsymbol{\psi}) = \delta_i(\boldsymbol{\psi}) p_i(\boldsymbol{\psi})$, dove $\delta_i(\boldsymbol{\psi})$ rappresenta la perdita

percentuale attesa dato il default, dato $\boldsymbol{\Psi} = \boldsymbol{\psi}$. Questa estensione è rilevante dal punto di vista empirico, poiché l'evidenza suggerisce che le perdite in caso di default dipendano dallo stato dell'economia sottostante.

(I2) Esiste una funzione $\bar{\ell} : \mathbb{R}^p \to \mathbb{R}^+$ tale che

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}(L^{(d)} \mid \boldsymbol{\Psi} = \boldsymbol{\psi}) = \lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \ell_i(\boldsymbol{\psi}) = \bar{\ell}(\boldsymbol{\psi})$$

per ogni $\boldsymbol{\psi} \in \mathbb{R}^p$. La funzione $\bar{\ell}(\boldsymbol{\psi})$ è detta perdita condizionata asintotica.

(I3) Esiste una costante $C < \infty$ tale che

$$\sum_{i=1}^{d}(e_i/i)^2 < C \quad \text{per ogni } d.$$

Questa assunzione impedisce che le esposizioni crescano sistematicamente con la dimensione del portafoglio.

La seguente proposizione mostra che, sotto queste ipotesi, la perdita media del portafoglio è essenzialmente determinata dalla funzione di perdita condizionata asintotica $\bar{\ell}$ e dalla realizzazione del vettore di fattori casuali $\boldsymbol{\Psi}$. La dimostrazione si basa su una versione della legge forte dei grandi numeri (si veda Frey e McNeil, 2003).

Proposizione 1 *Consideriamo una successione $L^{(d)} = \sum_{i=1}^{d} L_i$ che soddisfa le ipotesi (I1)–(I3). Indichiamo con $\mathbb{P}(\cdot \mid \boldsymbol{\Psi} = \boldsymbol{\psi})$ la distribuzione condizionata della successione $(L_i)_{i\in\mathbb{N}}$ dato $\boldsymbol{\Psi} = \boldsymbol{\psi}$. Allora*

$$\lim_{d\to\infty} \frac{1}{d}L^{(d)} = \bar{\ell}(\boldsymbol{\psi}), \quad \mathbb{P}(\cdot \mid \boldsymbol{\Psi} = \boldsymbol{\psi}) \text{ } q.o.$$

La proposizione 1 si applica ovviamente anche al numero di default $S = \sum_{i=1}^{d} Y_i$ se si pone $\delta_i = e_i = 1$. Per una successione data $(Y_i)_{i\in\mathbb{N}}$ che segue un modello misto di Bernoulli a p fattori con probabilità di default $p_i(\boldsymbol{\psi})$, le ipotesi (I1) e (I3) sono automaticamente soddisfatte, e l'ipotesi (I2) diventa:

$$\lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} p_i(\boldsymbol{\psi}) = \bar{p}(\boldsymbol{\psi}) \quad \text{per una certa funzione } \bar{p} : \mathbb{R}^p \to [0,1].$$

Per i modelli misti di Bernoulli a un singolo fattore, possiamo ottenere un risultato più forte che collega i quantili di $L^{(d)}$ ai quantili della mixing–variable. Come per il risultato precedente, possiamo trovare una dimostrazione in Frey e McNeil (2003).

Proposizione 2 *Consideriamo una sequenza $L^{(d)} = \sum_{i=1}^{d} L_i$ che soddisfa le ipotesi (I1)–(I3) con una mixing–variable unidimensionale Ψ e distribuzione F. Supponiamo che la funzione di perdita condizionata asintotica $\bar{\ell}(\psi)$ sia strettamente crescente e continua a destra, e che F sia strettamente crescente in $q_\alpha(\Psi)$, ovvero che $F(q_\alpha(\Psi) + \delta) > \alpha$ per ogni $\delta > 0$. Allora*

$$\lim_{d\to\infty} q_\alpha(L^{(d)}) = \bar{\ell}(q_\alpha(\Psi)).$$

L'assunzione che $\bar{\ell}$ sia strettamente crescente ha senso se si assume che bassi valori di Ψ corrispondano a buoni stati del mondo, con basse probabilità di default condizionate e basse perdite dato il default, mentre alti valori di ψ corrispondono a stati peggiori con probabilità e perdite più elevate.

Segue dalla proposizione 2 che la coda della perdita in un modello misto di Bernoulli a un fattore è essenzialmente determinata dalla coda della distribuzione della mixing–variable. Consideriamo in particolare due modelli misti di Bernoulli scambiabili con distribuzioni delle mixing–variable $F_i(q) = \mathbb{P}(Q_i \leq q)$ per $i = 1{,}2$. Supponiamo che la coda di F_1 sia maggiore di quella di F_2, cioè che $F_1(q) < F_2(q)$ per q vicino a 1. Allora, la proposizione 2 implica che, per d sufficientemente grande, la coda di S è maggiore nel modello 1 rispetto al modello 2.

1.2.3 Analisi del rischio di modello nei modelli misti di Bernoulli

In questa sezione esaminiamo brevemente un aspetto del rischio di modello nei modelli misti di Bernoulli. Consideriamo un modello misto scambiabile di Bernoulli per un portafoglio omogeneo e analizziamo il rischio associato alla scelta della distribuzione della mixing–variable, sotto il vincolo che la probabilità di default p e la correlazione ρ siano fissate.

Secondo la Proposizione 2, la coda della distribuzione del numero di default S è essenzialmente determinata dalla coda della variabile aleatoria Q. Nella figura riportiamo la funzione di sopravvivenza per diverse distribuzioni di Q:

- la distribuzione probit–normale, corrispondente al modello KMV/CreditMetrics a un fattore,
- la distribuzione logit–normale, corrispondente al modello CreditPortfolioView,
- la distribuzione Beta,
- e la distribuzione corrispondente alla copula di Clayton.

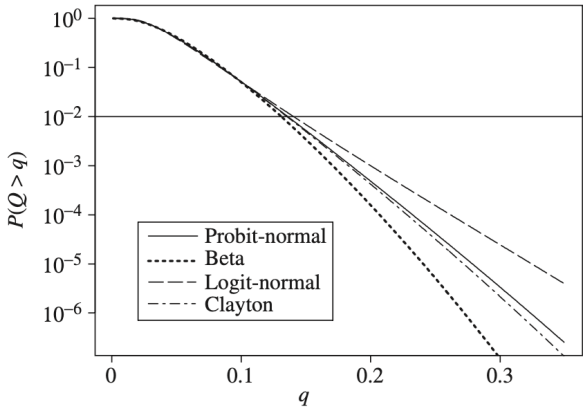

Figura 1.2. Coda della mixing–varibale Q in quattro diversi modelli misti scambiabili di Bernoulli (probit-normale, logit-normale, Beta, Clayton). In tutti i casi i primi due momenti sono fissati a $p = 0.049$ e $\pi_2 = 0.00313$, che corrispondono approssimativamente alla categoria B nel sistema di rating di Standard & Poor's

I grafici sono mostrati in scala logaritmica e, in tutti i casi, i primi due momenti di Q sono fissati ai valori $p = 0.049$ e $\pi_2 = 0.00313$, che corrispondono approssimativamente alla categoria B nel sistema di rating di Standard & Poor's.

La figura 1.2 mostra che le distribuzioni divergono in modo significativo solo oltre il 99–esimo percentile. Questo significa che, da un punto di vista pratico, la scelta della forma parametrica della distribuzione Q è di minore importanza, una volta che abbiamo fissato i suoi primi due momenti π_1 e π_2. Questo non implica

ovviamente che il rischio di modello nel caso dei modelli misti di Bernoulli a un fattore sia azzerato, in quanto la coda del numero di default S risulta comunque sensibile sia a p, che soprattutto a ρ.

1.3 Modello con la componente temporale

L'obiettivo di questa sezione consiste nel definire un modello che descriva i default di un gruppo di d clienti titolari di carte di credito, che includa la dimensione temporale. Adottiamo quindi una discretizzazione del tempo in intervalli della durata di un mese, e definiamo il default del cliente $i, i = 1, \dots, d$ relativo all'intervallo temporale $(t, t+j]$ con la variabile aleatoria

$$Y_i^{(t,t+j]} = \begin{cases} 1 & \text{se il cliente } i \text{ non paga almeno una rata nell'intervallo } (t, t+j], \\ 0 & \text{altrimenti.} \end{cases}$$

Utilizziamo la parentesi tonda in corrispondenza di t e quella quadra in corrispondenza di $t+j$ per intendere che i periodi di tempo che compongono l'intervallo $(t, t+j]$ sono quelli che vanno dal mese $t+1$ fino al mese $t+j$ (entrambi compresi).

Per estendere il modello misto di Bernoulli al caso multi–periodo, facciamo la seguente assunzione: fissato un istante di tempo t e un orizzonte di previsione j, data una variabile aleatoria Q_{tj}, assumiamo che

$$\mathbb{P}(Y_i^{(t,t+j]} = 1 | Q_{tj}) = Q_{tj}. \tag{1.9}$$

Assumiamo inoltre che

$$Q_{tj} = G_{tj}(\mathbf{X}_{it}), \tag{1.10}$$

dove nei predittori $\mathbf{X}_{it}$ è codificata tutta l'informazione disponibile sul cliente i al tempo t. L'assunzione è quindi che la probabilità di default del cliente i nell'intervallo $(t, t+j]$ sia una funzione deterministica, potenzialmente non lineare e complessa (ovvero G_{tj}) dei predittori $\mathbf{X}_{it}$. L'obiettivo delle nostre analisi è quello di stimare al meglio le funzioni G_{tj}, mediante tecniche di machine learning.

1.4 Misure di rischio e Value at Risk

Fissiamo uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ e denotiamo con $L^0(\Omega, \mathcal{F}, \mathbb{P})$ l'insieme delle variabili aleatorie quasi certamente finite definite su tale spazio. I rischi finanziari sono rappresentati da un *cono convesso* $\mathcal{M} \subseteq L^0(\Omega, \mathcal{F}, \mathbb{P})$ di variabili aleatorie. Qualsiasi variabile L appartenente a questo insieme sarà interpretata come una possibile perdita di un portafoglio di crediti. Ricordiamo che $\mathcal{M}$ è per definizione un cono convesso se $L_1 \in \mathcal{M}$ e $L_2 \in \mathcal{M}$ implica che $L_1 + L_2 \in \mathcal{M}$ e $\lambda L_1 \in \mathcal{M}$ per ogni $\lambda > 0$.

Definizione 3 *Dato un cono convesso $\mathcal{M}$ di variabili aleatorie, una misura di rischio con dominio $\mathcal{M}$ è una funzione*

$$\rho : \mathcal{M} \to \mathbb{R}.$$

In termini economici, interpretiamo $\rho(L)$ come la quantità di capitale che dovrebbe essere aggiunta a un portafoglio per coprire la perdita L, in modo che il portafoglio diventi "accettabile".

Indichiamo la distribuzione della perdita L con $F_L(l) = \mathbb{P}(L \leq l)$. Per le nostre analisi abbiamo scelto di utilizzare il Value-at-Risk (VaR) come principale misura di rischio per un portafoglio di credito. Richiamiamo ora la definizione di VaR per una generica variabile aleatoria $L \in \mathcal{M}$.

Definizione 4 *Dato un livello di confidenza $\alpha \in (0,1)$, il Value at Risk (VaR) del nostro portafoglio al livello di confidenza α è definito come il più piccolo numero l tale che la probabilità che la perdita L sia inferiore o uguale a l non sia minore di α. Formalmente,*

$$VaR_\alpha = \inf \{l \in \mathbb{R} \,|\, \mathbb{P}(L \leq l) \geq \alpha\}.$$

Questa definizione di VaR coincide con quella di quantile al livello α della distribuzione di L in termini dell'inversa generalizzata della funzione di distribuzione F_L.

Nel seguito delle nostre analisi, assumeremo un portafoglio *omogeneo* di d debitori, e sceglieremo come variabile aleatoria di perdita il numero di default $S = \sum_{i=1}^{d} Y_i$ associato a tale portafoglio.

Capitolo 2

Descrizione del dataset

Per le nostre analisi abbiamo deciso di utilizzare il dataset Kaggle "Default of Credit Card Clients" (https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data). Kaggle è una community online di professionisti, studenti e ricercatori, specializzata in competizioni di data science e machine learning, e rappresenta una delle principali fonti di dataset pubblici nel campo dell'analisi dei dati. Il dataset utilizzato riporta i dati relativi a 30000 clienti di una banca di Taiwan, da Aprile a Settembre 2005, e consiste in 24 covariate, tra le quali vi sono dati demografici e dati storici relativi a ogni cliente e ai suoi pagamenti. Di seguito una descrizione dettagliata delle covariate:

- **LIMIT_BAL**: Ammontare del credito concesso, in dollari taiwanesi (NT).
- **SEX**: Genere (1 = maschio, 2 = femmina).
- **EDUCATION**: Livello di istruzione (1 = dottorato, 2 = università, 3 = scuola superiore, 4 = altro, 5 = sconosciuto, 6 = sconosciuto).
- **MARRIAGE**: Stato civile (1 = sposato/a, 2 = single, 3 = altro).
- **AGE**: Età, in anni.
- **PAY_0**: Stato del rimborso a settembre 2005 (-1 = pagamento puntuale, 1 = ritardo di un mese, 2 = ritardo di due mesi, ..., 8 = ritardo di otto mesi, 9 = ritardo di nove mesi o più).

- **PAY_2**: Stato del rimborso ad agosto 2005 (stessa scala di PAY_0).
- **PAY_3**: Stato del rimborso a luglio 2005 (stessa scala di PAY_0).
- **PAY_4**: Stato del rimborso a giugno 2005 (stessa scala di PAY_0).
- **PAY_5**: Stato del rimborso a maggio 2005 (stessa scala di PAY_0).
- **PAY_6**: Stato del rimborso ad aprile 2005 (stessa scala di PAY_0).
- **BILL_AMT1**: Ammontare dell'estratto conto a settembre 2005 (in NT).
- **BILL_AMT2**: Ammontare dell'estratto conto ad agosto 2005 (in NT).
- **BILL_AMT3**: Ammontare dell'estratto conto a luglio 2005 (in NT).
- **BILL_AMT4**: Ammontare dell'estratto conto a giugno 2005 (in NT).
- **BILL_AMT5**: Ammontare dell'estratto conto a maggio 2005 (in NT).
- **BILL_AMT6**: Ammontare dell'estratto conto ad aprile 2005 (in NT).
- **PAY_AMT1**: Ammontare del pagamento effettuato a settembre 2005 (in NT).
- **PAY_AMT2**: Ammontare del pagamento effettuato ad agosto 2005 (in NT).
- **PAY_AMT3**: Ammontare del pagamento effettuato a luglio 2005 (in NT).
- **PAY_AMT4**: Ammontare del pagamento effettuato a giugno 2005 (in NT).
- **PAY_AMT5**: Ammontare del pagamento effettuato a maggio 2005 (in NT).
- **PAY_AMT6**: Ammontare del pagamento effettuato ad aprile 2005 (in NT).
- **default.payment.next.month**: Indicatore di default nel mese successivo (1 = sì, 0 = no).

Notiamo che la storia dei pagamenti dei clienti è descritta dalle variabili PAY_6, PAY_5, PAY_4, PAY_3, PAY_2, PAY_0 e default.payment.next.month, relative rispettivamente ai mesi da Aprile (PAY_6) a Ottobre (default.payment.next.month). Per comodità abbiamo rinomiato la variabile PAY_0 con PAY_1, e la variabile default.payment.next.month con PAY_0. Definiamo il default di un cliente, relativo ad un intervallo di tempo $(t, t+j]$ come il mancato pagamento di una rata nell'intervallo $(t, t+j]$. Questo avviene se almeno una delle variabili PAY_ relative a quell'intervallo temporale è strettamente maggiore di zero (ovvero se quel mese c'è un ritardo nel pagamento).
Le variabili BILL_AMTi, PAY_AMTi, con $i = 1, \ldots, 6$, sono indicizzate in maniera analoga.

Riprendendo le equazioni (1.9) e (1.10), otteniamo che

$$\mathbb{P}(Y_i^{(t,t+j]} = 1 | Q_{tj}) = Q_{tj} = G_{tj}(\mathbf{X}_{it}).$$

Il nostro obiettivo è stimare le funzioni G_{tj}, per ogni coppia (t, j), dove $t = 1, \ldots, 6$ è il mese corrente in cui facciamo previsioni (da Aprile a Ottobre) e $j = 1, \ldots, 7-t$ è l'orizzonte di previsione.
Occorre osservare che la struttura dei predittori $\mathbf{X}_{it}$ cambia al variare di t, in quanto cambiano al variare del tempo le informazioni disponibili sul cliente i; per esempio al tempo massimo $t = 6$, ovvero a Settembre 2005, abbiamo che tutte le 24 covariate sono osservabili, e l'orizzonte di previsione può essere solo di 1 mese (ovvero facciamo previsioni solo su Ottobre 2005); invece, al tempo $t = 1$, dobbiamo scartare tutte le variabili "future" BILL_AMTi, PAY_AMTi, PAY_i, con $i < 6$ (in quanto nel dataset sono indicizzate in ordine decrescente).
I metodi di machine learning utilizzati per stimare le funzioni G_{tj} sono la regressione logistica (LR), il multi–layer–perceptron (MLP), k–nearest–neighbors (KNN) e AdaBoost (AB).

2.1 Analisi esplorativa del dataset

Procediamo con un'analisi esplorativa del dataset. Nella seguente tabella, per ogni variabile è riportato il numero di valori, la media, la deviazione standard, minimo, massimo e i quantili ai livelli $25\%, 50\%, 75\%$.

Variabile	**Conteggio**	**Media**	**Std**	**Min**
ID	30000.0	15000.500000	8660.398374	1.0
LIMIT_BAL	30000.0	167484.322667	129747.661567	10000.0
SEX	30000.0	1.603733	0.489129	1.0
EDUCATION	30000.0	1.853133	0.790349	0.0
MARRIAGE	30000.0	1.551867	0.521970	0.0
AGE	30000.0	35.485500	9.217904	21.0
PAY_1	30000.0	-0.016700	1.123802	-2.0
PAY_2	30000.0	-0.133767	1.197186	-2.0
PAY_3	30000.0	-0.166200	1.196868	-2.0
PAY_4	30000.0	-0.220667	1.169139	-2.0
PAY_5	30000.0	-0.266200	1.133187	-2.0
PAY_6	30000.0	-0.291100	1.149988	-2.0
BILL_AMT1	30000.0	51223.330900	73635.860576	-165580.0
BILL_AMT2	30000.0	49179.075167	71173.768783	-69777.0
BILL_AMT3	30000.0	47013.154800	69349.387427	-157264.0
BILL_AMT4	30000.0	43262.948967	64332.856134	-170000.0
BILL_AMT5	30000.0	40311.400967	60797.155770	-81334.0
BILL_AMT6	30000.0	38871.760400	59554.107537	-339603.0
PAY_AMT1	30000.0	5663.580500	16563.280354	0.0
PAY_AMT2	30000.0	5921.163500	23040.870402	0.0
PAY_AMT3	30000.0	5225.681500	17006.961470	0.0
PAY_AMT4	30000.0	4826.076867	15666.159744	0.0
PAY_AMT5	30000.0	4799.387633	15278.305679	0.0
PAY_AMT6	30000.0	5215.502567	17777.465775	0.0
PAY_0	30000.0	0.221200	0.415062	0.0

Tabella 2.1. Statistiche sul dataset (prima parte).

Variabile	**25%**	**50%**	**75%**	**Max**
ID	7500.75	15000.0	22500.25	30000.0
LIMIT_BAL	50000.0	140000.0	240000.0	1000000.0
SEX	1.0	2.0	2.0	2.0
EDUCATION	1.0	2.0	2.0	6.0
MARRIAGE	1.0	2.0	2.0	3.0
AGE	28.0	34.0	41.0	79.0
PAY_1	-1.0	0.0	0.0	8.0
PAY_2	-1.0	0.0	0.0	8.0
PAY_3	-1.0	0.0	0.0	8.0
PAY_4	-1.0	0.0	0.0	8.0
PAY_5	-1.0	0.0	0.0	8.0
PAY_6	-1.0	0.0	0.0	8.0
BILL_AMT1	3558.75	22381.5	67091.0	964511.0
BILL_AMT2	2984.75	21200.0	64006.25	983931.0
BILL_AMT3	2666.25	20088.5	60164.75	1664089.0
BILL_AMT4	2326.75	19052.0	54506.0	891586.0
BILL_AMT5	1763.0	18104.5	51950.0	927171.0
BILL_AMT6	1256.0	17071.0	49198.25	961664.0
PAY_AMT1	1000.0	2100.0	5006.0	873552.0
PAY_AMT2	833.0	2009.0	5000.0	1684259.0
PAY_AMT3	390.0	1800.0	4505.0	896040.0
PAY_AMT4	296.0	1500.0	4013.25	621000.0
PAY_AMT5	252.5	1500.0	4031.5	426529.0
PAY_AMT6	117.75	1500.0	4000.0	528666.0
PAY_0	0.0	0.0	0.0	1.0

Tabella 2.2. Statistiche sul dataset (seconda parte).

Come possiamo osservare non sono presenti valori nulli o mancanti per nessuna variabile del dataset.

La visualizzazione dei dati è molto utile nell'analisi esplorativa, poiché aiuta a organizzare i dati in un modo più comprensibile e mettendo in evidenza tendenze e valori anomali. Per prima cosa osserviamo quanti clienti sono andati in default nel mese di Ottobre 2005 (variabile default_payment_next_month rinominata PAY_0). Come possiamo osservare dalla figura 2.1, vi sono 6636 clienti che non hanno pagato la rata del mese di Ottobre, e quindi sono contati come default, che ammontano al 22,12% del totale di 30000 clienti, e 23364 clienti che hanno pagato in tempo. Osserviamo che, per come abbiamo definito il concetto di default

relativo ad un intervallo di tempo, ovvero affermando che un cliente i è considerato in default relativamente all'intervallo di tempo $(t, t+j]$ se salta il pagamento di almeno una rata in quell'intervallo, la frazione di clienti considerati default non è costante nel tempo, e ovviamente non è la stessa se consideriamo orizzonti di previsione (e quindi intervalli) diversi.

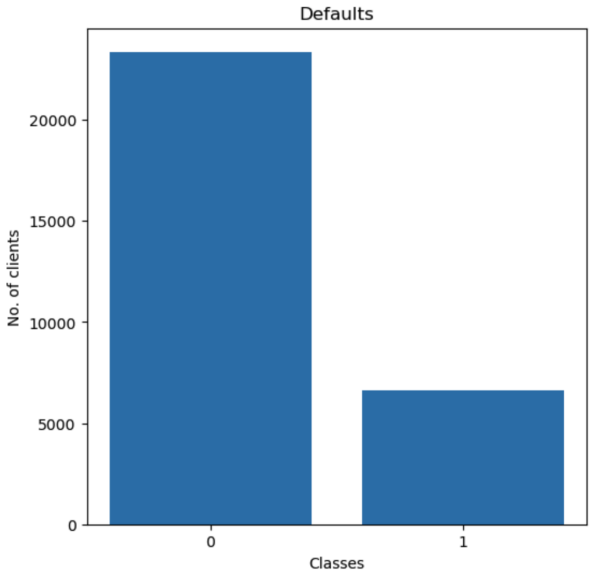

Figura 2.1. Numero di defaults nel mese di Ottobre

La figura 2.2 mostra il numero di clienti che sono andati in default negli ultimi due mesi, per ogni mese da Maggio a Ottobre.
Per quanto riguarda la distribuzione dell'attributo AGE nel dataset la figura 2.3 mostra come la fascia di età più presente sia quella tra i 25 e i 35 anni. La figura 2.4 invece riporta il numero di clienti considerati default per il mese di Ottobre, per varie fasce di età; non notiamo differenze significative tra fasce di età differenti.
La matrice di correlazione è uno strumento statistico importante in data science per analizzare e visualizzare le relazioni tra variabili quantitative all'interno di un dataset. Si tratta di una matrice quadrata A in cui l'entrata A_{ij} è la correlazione (campionaria) tra la i–esima e la j–esima variabile del dataset. Nella figura 2.5 è riportata la matrice di correlazione relativa alle 10 variabili più correlate con PAY_0 (ovvero con l'indicatore di default del mese di Ottobre). Notiamo come le

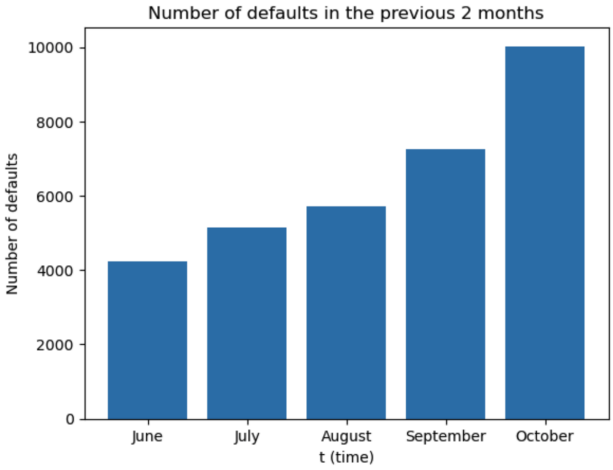

Figura 2.2. Numero di defaults per i 2 mesi precedenti.

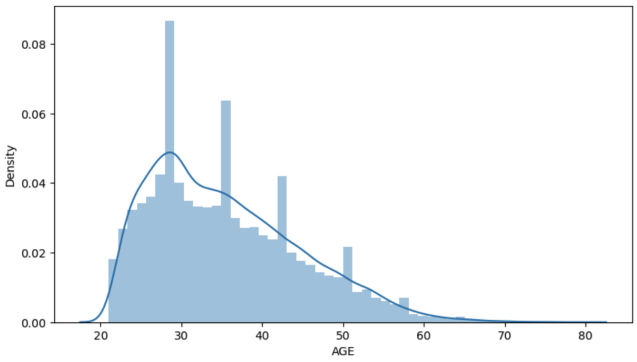

Figura 2.3. Distribuzione dell'età dei clienti nel dataset.

variabili più correlate con PAY_0 sono le altre variabili PAY_i, ovvero la storia dei pagamenti precedenti dei clienti.

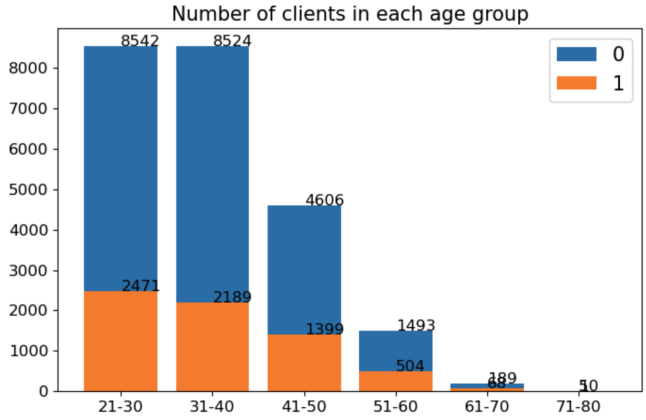

Figura 2.4. Numero di defaults nel mese di Ottobre per fascia di età.

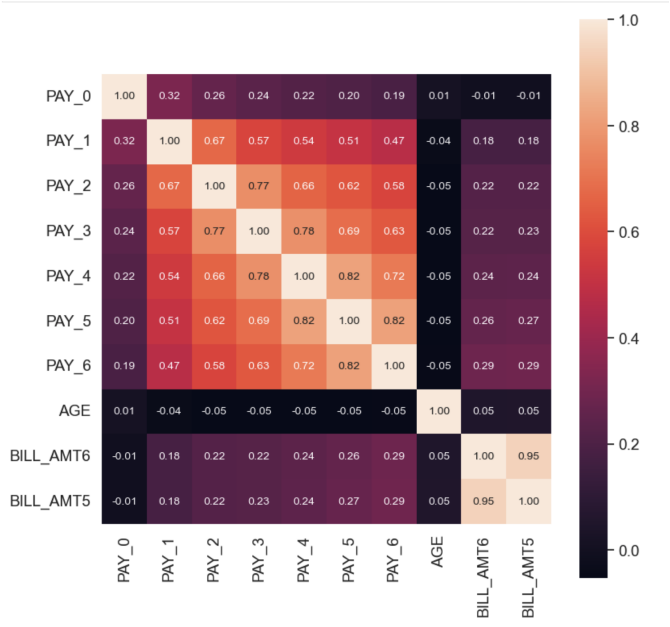

Figura 2.5. Matrice di correlazione per le 10 variabili più correlate con PAY_0.

Capitolo 3

Applicazione dei modelli e risultati

In questo capitolo presentiamo i risultati ottenuti applicando i vari modelli di Machine Learning. Il codice è stato sviluppato interamente in Python, facendo uso della libreria scikit–learn.

Ricordiamo che l'obiettivo delle nostre analisi è stimare le funzioni G_{tj}, per ogni coppia (t, j), dove $t = 1, \ldots, 6$ è il mese in cui facciamo previsioni (da Aprile a Settembre), $j = 1, \ldots, 7 - t$ è l'orizzonte di previsione e

$$\mathbb{P}(Y_i^{(t,t+j]} = 1 | Q_t) = Q_t = G_{tj}(\mathbf{X}_{it}),$$

ovvero $G_{tj}(\mathbf{x}_{it})$ è la probabilità di default di un cliente i caratterizzato dai predittori $\mathbf{x}_{it}$.

Per ogni coppia (t, j) abbiamo modificato il dataset di partenza mantenendo solo le variabili osservabili fino all'istante di tempo t, e considerando ogni cliente come default se è stato mancato almeno un pagamento nell'intervallo di tempo $(t, t + j]$. Abbiamo poi diviso ogni dataset in un training set, contentente l'80% delle osservazioni, su cui vengolo allenati i modelli, e in un test set, contenente il restante 20% delle osservazioni, sul quale vengono valutate le performance. Le metriche utilizzate per fare ciò sono l'accurateza, la precisione, il recall, l'AUC, e lo score F1. Come metrica principale di riferimento abbiamo scelto quest'ultima in quanto è la

media armonica di precisione e recall, e quindi si presta bene per una valutazione più generale del modello. Per ogni modello, al fine di selezionare i migliori iper-parametri, è stata effettuata una grid–search cross–validation in cui la metrica utilizzata è lo score F1. I modelli sono stati quindi valutati su ogni combinazione dei loro parametri, al fine di selezionare i migliori. Per il multi-layer-perceptron la griglia dei parametri è stata predisposta come segue:

- max_iter $\in \{1000,1500,2000\}$,
- hidden_layer_sizes $\in \{(10), (13), (17), (24), (13,13)\}$,

mentre per AdaBoost:

- n_estimators $\in \{50, 100, 500, 1000,2000\}$,
- learning_rate $\in \{(1,\ 1e-1,\ 1e-3,\ 1e-4,\ 1e-5\}$.

Per i k-nearest-neighbors, infine, sono stati valutati diversi valori di $k \in \{3, \dots, 15\}$.

Nel caso del MLP sono stati selezionati come migliori parametri un numero massimo di iterazioni max_iter $= 1000$ e due layer nascoti con 13 nodi. Per AdaBoost invece il numero di stimatori scelto è 2000, e il learning rate ottimale è 0.1. Per KNN invece il valore di k più performante è stato $k = 5$.

Per ogni modello, e per ogni coppia (t, j), abbiamo poi stimato i valori della probabilità di default $p := \pi_1 = \mathbb{E}[Q]$ e della correlazione $\rho = \text{corr}(Y_i, Y_k)$ presentati nella sezione 1.2. Queste quantità sono state stimate partendo dalle osservazioni sul test set delle probabilità di default previste da ogni modello, ovvero dalle osservazioni di Q; in particolare la nostra stima $\hat{p}$ di $p = \mathbb{E}[Q]$ è la media campionaria delle osservazioni, mentre la stima di $\rho = \frac{\pi_2 - p^2}{p(1-p)}$ è la seguente

$$\hat{\rho} = \frac{\hat{\sigma}_Q^2}{\hat{p}(1-\hat{p})},$$

dove $\hat{\sigma}_Q^2$ è la varianza campionaria delle osservazioni di Q.
Un metodo equivalente per stimare p e ρ, assumendo un modello beta-mixing, consiste nello stimare i parametri a e b della distribuzione Beta(a, b) di Q con il

metodo dei momenti, e ricavare p e ρ sapendo che, sotto le precedenti assunzioni, valgono

$$p = \frac{a}{a+b} \quad \text{e} \quad \rho = \frac{1}{a+b+1}.$$

Considerando il fatto che il dataset è fortemente sbilanciato a favore della classe 0 (non default), abbiamo inizialmente provato ad applicare un random oversampling nella fase di training dei modelli. Questo consiste fondamentalmente in un ricampionamento casuale dei punti del training set aventi classe 0. Nelle tabelle A.1–A.4, contenute nell'Appendice, riportiamo i risultati ottenuti, e le stime di p e ρ, per i quattro modelli considerati (LR, MLP, KNN, AB). Nel seguito, valori di t da 1 a 6 indicano rispettivamente i mesi da Aprile a Settembre.

Possiamo osservare come il modello con le performance migliori, in termini di score F1, sia il MLP (con uno score F1 medio pari a 0.5733, rispetto a 0.5393 di LR, a 0.5243 di AB e 0.5030 di KNN). L'oversampling fatto sul training set porta però i modelli a sovrastimare le probabilità previste di default; ad esempio, nel caso $t = 6, j = 1$, la media delle probabilità di default previste dovrebbe essere intorno al 22,12%, mentre p in questo caso è stimato a 0.4551 (per LR).
Nelle tabelle A.5–A.12 sono riportati invece i risultati relativi ai quattro modelli (LR, MLP, KNN, AB) senza utilizzare tecniche di oversampling. Anche in questo caso il modello più performante è MLP, con uno score F1 medio pari a 0.5534, rispetto a 0.5395 di KNN, 0.5220 di AB, e 0.5186 di LR. Le probabilità previste, ovvero le stime di p, in questo caso sono molto vicine alle percentuali di default osservate (per esempio, per $t = 6, j = 1$, LR ha una probabilità prevista di default pari a 0.2208, e la frazione di default osservati è il 22,12%). Notiamo che l'oversampling migliora solo leggermente le performance in termini di score F1, al prezzo di probabilità previste artificialmente maggiori. Per questo motivo decidiamo di concentrarci sui modelli senza oversampling.

Nelle tabelle A.5–A.12 abbiamo anche riportato i valori dei parametri a e b della distribuzione Beta, stimati con il metodo dei momenti, partendo dalle osservazioni di Q sul test set (ovvero dalle probabilità previste dai modelli). Abbiamo infine riportato le stime del Value at Risk (VaR) al 95% per il numero di defaut S di un

portfolio composto da $d = 5000$ clienti, calcolate analiticamente, assumendo un modello beta–mixing.

Possiamo notare inoltre che:

- come ci si poteva attendere, i valori stimati del parametro p risultano crescenti al crescere dell'orizzonte di previsione j, per un dato t fissato. Questo comportamento è dovuto al fatto che gli intervalli temporali considerati sono progressivamente inclusivi. Per esempio, nel caso $t = 3$, le stime di p variano da circa 0.14 per $j = 1$, fino a superare 0.37 per $j = 4$, come illustrato dalla figura 3.1.

- Al contrario, i valori stimati del parametro ρ tendono a decrescere all'aumentare di j, sempre per t fissato.

  Combinando queste due osservazioni, si può affermare che, su orizzonti temporali più lunghi, la probabilità di default tende ad aumentare in modo generalizzato ("tutti vanno in default"), mentre l'influenza delle correlazioni tra clienti perde progressivamente importanza.

- Le stime del parametro p ottenute dai diversi modelli (LR, MLP, KNN, AB) risultano complessivamente simili tra loro, a indicare una certa coerenza nella valutazione della probabilità di default. Diversa è la situazione per il parametro ρ, le cui stime variano in modo più marcato tra i modelli, come evidenziato dalla figura 3.2. In particolare, l'andamento di ρ risulta molto simile per i modelli AB, MLP e KNN, mentre si discosta leggermente nel caso di LR. Tra i modelli analizzati, KNN fornisce le stime più elevate di ρ, mentre AB e LR restituiscono i valori minori; le previsioni del modello MLP si collocano in una posizione intermedia.

- Stime differenti dei parametri p e ρ si riflettono in stime differenti del VaR. In particolare, valori più elevati di p e ρ si traducono in un aumento del VaR stimato. Come osservato in precedenza, mentre le stime di p risultano abbastanza coerenti tra i vari modelli, quelle di ρ mostrano una maggiore variabilità. Nella figura 3.3 sono riportati i valori del VaR ottenuti dai quattro modelli in funzione dell'orizzonte di previsione j, per $t = 2$ fissato. Si può

notare come i VaR stimati dal modello KNN siano i più elevati, partendo da circa 4000 per $j = 1$, fino ad arrivare a 4480, in corrispondenza di $j = 5$. Questo comportamento è in linea con le stime più alte di ρ fornite da KNN. Per quanto riguarda MLP, le stime del VaR partono da circa 3600 per arrivare a 3800 per $j = 5$, con un minimo in corrispondenza di $j = 2$, in cui il VaR stimato è 3285. I modelli AB e LR forniscono invece stime del VaR più contenute, comprese approssimativamente tra 3000 e 3400.

E' interessante notare come l'andamento del VaR risulti piuttosto simile per i modelli AB, MLP e KNN, mentre nel caso del modello LR si distingue: l'aumento di p al crescere dell'orizzonte di previsione j risulta compensato dalla diminuzione di ρ, determinando così valori del VaR complessivamente più stabili e inferiori rispetto agli altri modelli. In linea con quanto osservato nelle stime di ρ, il modello MLP si colloca in una posizione intermedia anche in termini di VaR.

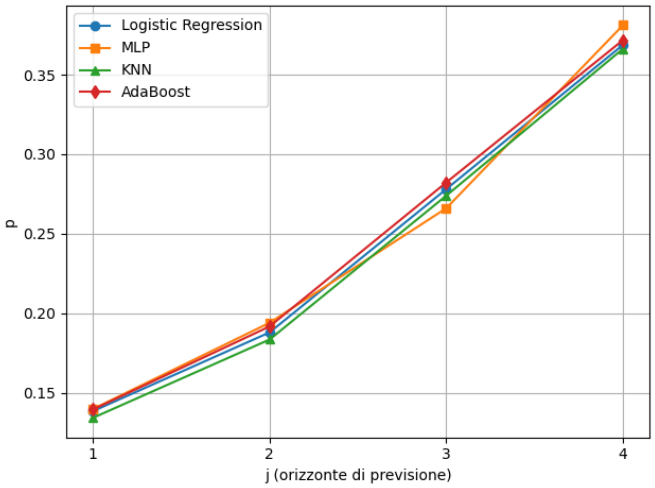

Figura 3.1. p in funzione di j (orizzonte di previsione), per $t = 3$ (Giugno)

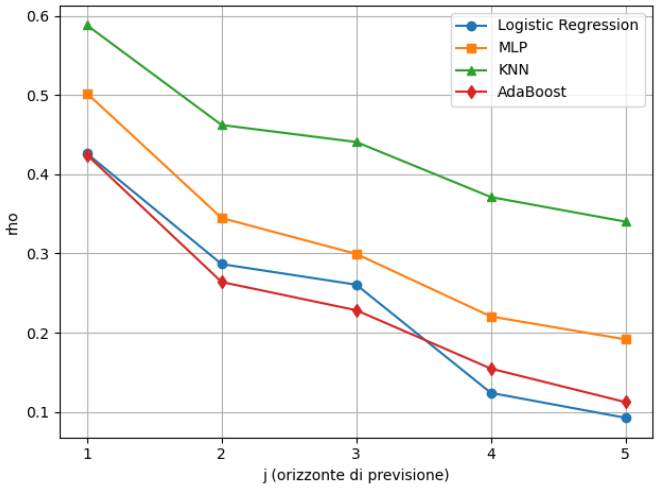

Figura 3.2. ρ in funzione di j (orizzonte di previsione), per $t = 2$ (Maggio)

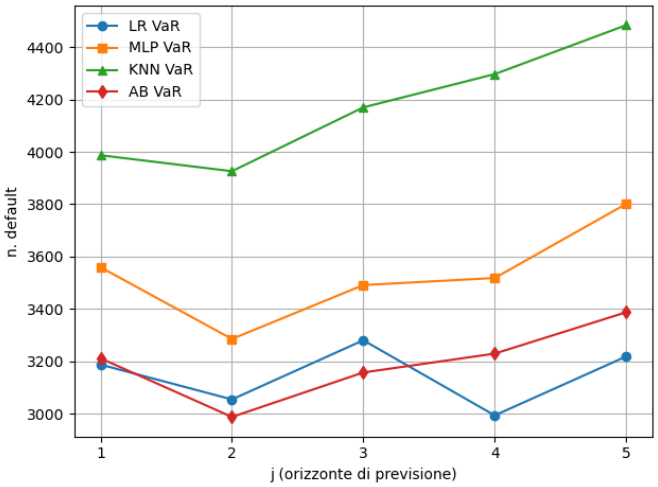

Figura 3.3. VaR in funzione di j (orizzonte di previsione), per $t = 2$ (Maggio)

- Le metriche di performance score F1 e AUC, per ogni modello, in generale sono decrescenti al crescere dell'orizzonte di previsione j, per t fissato, come evidenziato dalle figure 3.4 e 3.5.

  Un'interpretazione di questo andamento può essere data considerando che un modello che si riferisce ad un intervallo temporale più ampio (caratterizzato

da un orizzonte di previsione j maggiore, fissato t), è più incerto nel prevedere gli eventi di default, e quindi meno performante.

Una spiegazione più tecnica è la seguente. Se consideriamo due modelli G_{tj_1} e G_{tj_2} i quali si riferiscono rispettivamente agli intervalli temporali $(t, t + j_1], (t, t + j_2]$, con $j_2 > j_1$, abbiamo che il relativo dataset $\mathbf{X}_t$ è lo stesso per entrambi i modelli, in quanto contiene le variabili osservabili al tempo t, ma le variabili target $\mathbf{y}^{(t,t+j_1]}, \mathbf{y}^{(t,t+j_2]}$ contenenti gli indicatori di default per i due intervalli temporali contengono dati diversi. In particolare le relazioni tra i predittori $\mathbf{X}_t$ e $\mathbf{y}^{(t,t+j_2]}$ sono meno evidenti e quindi più difficili da cogliere per il modello G_{tj_2} essendo $t+j_2$ più lontano nel tempo, mentre le relazioni tra $\mathbf{X}_t$ e $\mathbf{y}^{(t,t+j_1]}$ sono più evidenti e permettono quindi al modello G_{tj_1} di ottenere performance migliori.

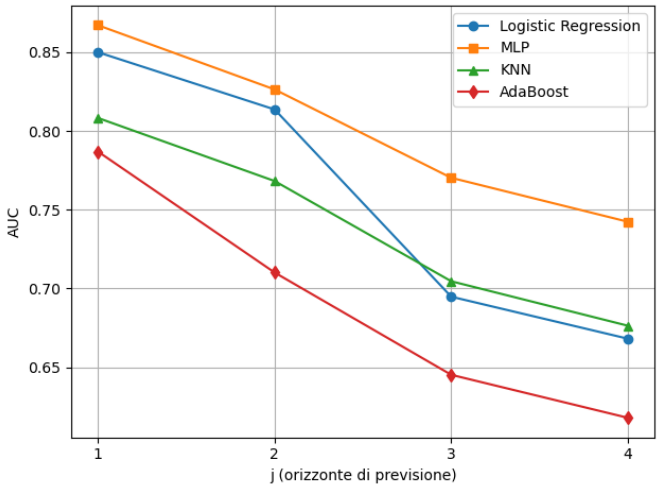

Figura 3.4. AUC in funzione di j (orizzonte di previsione), per $t = 3$ (Giugno)

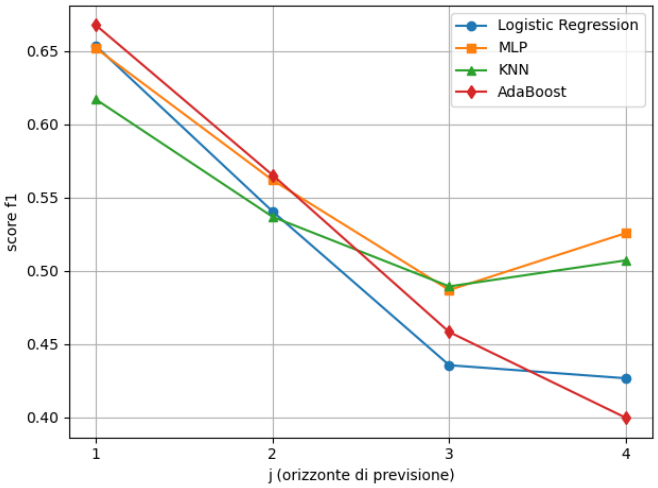

Figura 3.5. score F1 in funzione di j (orizzonte di previsione), per $t = 3$ (Giugno)

- Passando ad una valutazione delle performance dei modelli, in termini di score F1 il modello migliore è MLP, seguito da KNN, AB e LR, come evidenziato in precedenza.

Modello	**score F1**
MLP	0.5534
KNN	0.5395
AB	0.5220
LR	0.5186

Tabella 3.1. Score F1 medio

- Anche in termini di AUC il modello più performante è MLP, con un AUC medio pari a 0.8048, seguito da LR (0.7548), KNN (0.7398) e AB (0.6914).

Modello	**AUC media**
MLP	0.8048
LR	0.7548
KNN	0.7398
AB	0.6914

Tabella 3.2. AUC media

- Sebbene i valori differiscano di meno, anche per quanto concerne l'accuratezza il modello migliore è MLP, con un'accuratezza media di 0.8278, seguito da AB (0.8265), LR (0.8201) e KNN (0.8037).

Modello	**Accuratezza media**
MLP	0.8278
AB	0.8265
LR	0.8201
KNN	0.8037

Tabella 3.3. Accuratezza media

- Possiamo quindi concludere che il modello MLP risulta complessivamente il più performante, indipendentemente dalla metrica di valutazione scelta. Per quanto riguarda gli altri modelli, invece, non emerge una netta superiorità: le loro prestazioni variano in funzione della metrica considerata, senza che uno prevalga chiaramente sugli altri.

Capitolo 4

Conclusioni

L'analisi condotta ha permesso di approfondire il comportamento dei parametri fondamentali del modello misto di Bernoulli scambiabile (p e ρ), quando viene aggiunta la dimensione temporale, oltre alle implicazioni sulla stima del Value at Risk e alle prestazioni predittive dei diversi modelli di Machine Learning impiegati. Si è osservato che la probabilità di default p cresce all'aumentare dell'orizzonte di previsione j, mentre la correlazione tra default ρ tende a decrescere, suggerendo che su archi temporali più ampi il rischio di default diventa più diffuso, ma meno dipendente da dinamiche di gruppo.

Le differenze tra i modelli emergono più chiaramente nell'analisi delle performance predittive e nella stima di ρ. In particolare, il modello MLP si conferma il più efficace, sia in termini di score F1, AUC, che accuratezza, mostrando una maggiore capacità di catturare le relazioni tra le variabili predittive e gli eventi di default. Al contrario, gli altri modelli (KNN, AB, LR) mostrano prestazioni più variabili a seconda della metrica considerata, senza che uno di essi si distingua chiaramente sugli altri.

Infine, si è evidenziato come l'aumento dell'orizzonte di previsione j, a parità di istante iniziale t, tenda a peggiorare le performance predittive dei modelli, a causa della crescente difficoltà nel cogliere le relazioni tra i predittori $\mathbf{X}_t$ osservate al tempo t e gli eventi di default che si verificano più lontano nel tempo. Questo risultato suggerisce un trade-off tra ampiezza dell'orizzonte di previsione e capacità

predittiva.

Un possibile sviluppo futuro riguarda l'utilizzo di modelli più complessi e recenti, come le reti neurali convoluzionali (CNN). Tali modelli potrebbero risultare più efficaci nel cogliere la dinamica evolutiva del rischio di default nel tempo.

Inoltre, potrebbe essere utile estendere l'analisi facendo uso di un dataset che includa anche variabili macroeconomiche esterne (es. tassi d'interesse, PIL, indicatori di rischio sistemico), oltre alla storia dei pagamenti dei debitori, al fine di migliorare la capacità predittiva dei modelli e contestualizzare meglio i default osservati.

Dal punto di vista metodologico, si potrebbe esplorare l'uso di tecniche di interpretabilità dei modelli, per comprendere meglio l'importanza delle singole variabili nella previsione del default, o ancora di riduzione della dimensionalità (come la Principal Component Analysis).

Appendice A

Risultati dei modelli

A.1 Modelli con oversampling

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.9448	0.7161	0.7271	0.7216	0.9033	0.2962	0.3154
1	2	0.7618	0.3409	0.7428	0.4673	0.8373	0.3740	0.2543
1	3	0.7055	0.3734	0.7286	0.4938	0.7978	0.4133	0.2223
1	4	0.6886	0.4149	0.7224	0.5270	0.7838	0.4363	0.2031
1	5	0.6356	0.4537	0.6298	0.5274	0.6901	0.4767	0.1012
1	6	0.6148	0.5178	0.6125	0.5612	0.6654	0.4893	0.0796
2	1	0.9078	0.5947	0.6518	0.6220	0.8686	0.3251	0.2911
2	2	0.7276	0.3570	0.7092	0.4749	0.8078	0.3963	0.2397
2	3	0.7080	0.4121	0.7305	0.5269	0.7930	0.4208	0.2321
2	4	0.6366	0.4393	0.6350	0.5194	0.6926	0.4734	0.1067
2	5	0.6153	0.5073	0.6178	0.5571	0.6670	0.4882	0.0828
3	1	0.8870	0.5964	0.6573	0.6254	0.8492	0.3534	0.2765
3	2	0.7405	0.4036	0.7149	0.5160	0.8139	0.4033	0.2451
3	3	0.6431	0.4168	0.6139	0.4965	0.6977	0.4700	0.1088
3	4	0.6203	0.4938	0.6060	0.5442	0.6714	0.4860	0.0874
4	1	0.8813	0.5771	0.7035	0.6341	0.8658	0.3479	0.3174
4	2	0.6515	0.3646	0.5811	0.4481	0.6865	0.4641	0.1166
4	3	0.6206	0.4595	0.5902	0.5167	0.6664	0.4833	0.0926
5	1	0.7538	0.4723	0.6124	0.5333	0.7128	0.4391	0.1507
5	2	0.6420	0.4716	0.5796	0.5201	0.6775	0.4749	0.1161
6	1	0.6940	0.3811	0.6389	0.4775	0.7271	0.4550	0.1282

Tabella A.1. Tabella con i risultati del modello LR, con oversampling. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.9003	0.4955	0.7610	0.6002	0.8938	0.2743	0.3599
1	2	0.7763	0.3543	0.7180	0.4745	0.8329	0.3729	0.2844
1	3	0.7693	0.4413	0.6398	0.5224	0.7973	0.3930	0.2405
1	4	0.7400	0.4700	0.6481	0.5449	0.7886	0.4223	0.2313
1	5	0.6718	0.4935	0.6308	0.5538	0.7307	0.4691	0.1605
1	6	0.6571	0.5760	0.5590	0.5674	0.7106	0.4685	0.1516
2	1	0.8621	0.4438	0.7306	0.5522	0.8708	0.3024	0.3547
2	2	0.7961	0.4386	0.6209	0.5141	0.8111	0.3670	0.2778
2	3	0.7691	0.4858	0.6294	0.5484	0.7979	0.4007	0.2517
2	4	0.7006	0.5140	0.5805	0.5453	0.7440	0.4525	0.1922
2	5	0.6705	0.5771	0.5936	0.5852	0.7302	0.4710	0.1713
3	1	0.8670	0.5284	0.6794	0.5945	0.8652	0.3207	0.3496
3	2	0.7956	0.4793	0.6511	0.5522	0.8198	0.3768	0.3064
3	3	0.7286	0.5239	0.5848	0.5527	0.7643	0.4400	0.2162
3	4	0.7028	0.6070	0.5824	0.5944	0.7431	0.4549	0.1949
4	1	0.8583	0.5108	0.7274	0.6001	0.8706	0.3240	0.3968
4	2	0.7495	0.4893	0.6605	0.5621	0.7909	0.4306	0.2758
4	3	0.7146	0.5820	0.6018	0.5917	0.7549	0.4430	0.2366
5	1	0.8803	0.7129	0.8018	0.7547	0.9160	0.3383	0.4963
5	2	0.7786	0.6724	0.6603	0.6663	0.8286	0.4352	0.3362
6	1	0.7255	0.4180	0.6488	0.5085	0.7644	0.4235	0.2696

Tabella A.2. Tabella con i risultati del modello MLP, con oversampling. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.8651	0.3977	0.7220	0.5129	0.8343	0.1713	0.8072
1	2	0.7615	0.3213	0.6255	0.4246	0.7465	0.2588	0.7314
1	3	0.6900	0.3399	0.6077	0.4360	0.7082	0.3337	0.6521
1	4	0.6746	0.3876	0.6120	0.4747	0.7006	0.3681	0.5842
1	5	0.6190	0.4316	0.5689	0.4908	0.6473	0.4303	0.4572
1	6	0.6023	0.5049	0.5669	0.5341	0.6389	0.4649	0.3805
2	1	0.8306	0.3744	0.6790	0.4826	0.7922	0.2017	0.7988
2	2	0.7260	0.3430	0.6314	0.4445	0.7269	0.3041	0.6904
2	3	0.6960	0.3891	0.6414	0.4844	0.7211	0.3547	0.6219
2	4	0.6360	0.4345	0.5892	0.5002	0.6673	0.4250	0.4820
2	5	0.6150	0.5074	0.5821	0.5422	0.6555	0.4650	0.3960
3	1	0.8055	0.3980	0.6933	0.5057	0.7895	0.2377	0.7715
3	2	0.7280	0.3843	0.6744	0.4896	0.7501	0.3201	0.6710
3	3	0.6525	0.4254	0.6052	0.4996	0.6838	0.4061	0.5182
3	4	0.6310	0.5058	0.5802	0.5404	0.6662	0.4487	0.4222
4	1	0.8016	0.4008	0.7217	0.5154	0.8062	0.2490	0.7740
4	2	0.6766	0.3962	0.6262	0.4854	0.7064	0.3730	0.5994
4	3	0.6461	0.4879	0.5965	0.5367	0.6805	0.4329	0.4782
5	1	0.7968	0.5422	0.7409	0.6261	0.8320	0.3117	0.7152
5	2	0.7080	0.5542	0.6508	0.5987	0.7531	0.4131	0.5451
6	1	0.6615	0.3483	0.6283	0.4482	0.6897	0.3700	0.5952

Tabella A.3. Tabella con i risultati del modello KNN, con oversampling. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.9396	0.6866	0.6913	0.6890	0.8288	0.3050	0.2100
1	2	0.9001	0.6952	0.4909	0.5754	0.7282	0.3989	0.1253
1	3	0.8618	0.7688	0.3928	0.5199	0.6825	0.4340	0.1066
1	4	0.8306	0.8150	0.3439	0.4837	0.6602	0.4551	0.0954
1	5	0.7506	0.8356	0.2584	0.3948	0.6175	0.4756	0.0814
1	6	0.6763	0.8613	0.2127	0.3412	0.5952	0.4886	0.0731
2	1	0.9320	0.7224	0.6291	0.6725	0.7994	0.3387	0.1832
2	2	0.8800	0.7293	0.4290	0.5402	0.6988	0.4237	0.1161
2	3	0.8460	0.7931	0.3639	0.4989	0.6692	0.4454	0.1079
2	4	0.7663	0.8379	0.2709	0.4094	0.6242	0.4735	0.0909
2	5	0.6905	0.8741	0.2213	0.3531	0.6008	0.4874	0.0831
3	1	0.9180	0.7417	0.6068	0.6675	0.7868	0.3606	0.1821
3	2	0.8723	0.7477	0.4543	0.5652	0.7100	0.4232	0.1270
3	3	0.7898	0.8003	0.3208	0.4581	0.6451	0.4654	0.0964
3	4	0.7135	0.8588	0.2603	0.3995	0.6178	0.4823	0.0870
4	1	0.9130	0.7051	0.6705	0.6874	0.8119	0.3594	0.2172
4	2	0.8213	0.7235	0.4101	0.5235	0.6804	0.4526	0.1170
4	3	0.7418	0.7960	0.3190	0.4555	0.6386	0.4764	0.1030
5	1	0.8896	0.9042	0.5716	0.7004	0.7769	0.3958	0.2253
5	2	0.7950	0.9357	0.4053	0.5656	0.6958	0.4598	0.1707
6	1	0.7831	0.4985	0.5204	0.5092	0.6880	0.4582	0.1170

Tabella A.4. Tabella con i risultati del modello AB, con oversampling. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

A.2 Modelli senza oversampling

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.9446	0.7345	0.6847	0.7087	0.9035	0.0961	0.4114
1	2	0.9033	0.7382	0.4845	0.5851	0.8370	0.1379	0.2814
1	3	0.8635	0.8192	0.3947	0.5328	0.7978	0.1902	0.2375
1	4	0.8278	0.8541	0.3414	0.4878	0.7835	0.2327	0.2172
1	5	0.7450	0.8679	0.2478	0.3855	0.6888	0.3146	0.1117
1	6	0.6595	0.6431	0.3443	0.4485	0.6650	0.3963	0.0858
2	1	0.9328	0.7583	0.6203	0.6824	0.8676	0.1165	0.4260
2	2	0.8765	0.7701	0.4117	0.5365	0.8070	0.1688	0.2865
2	3	0.8405	0.8272	0.3585	0.5002	0.7928	0.2170	0.2605
2	4	0.7601	0.8687	0.2641	0.4051	0.6912	0.3015	0.1239
2	5	0.6815	0.7084	0.3174	0.4384	0.6672	0.3858	0.0925
3	1	0.9131	0.7647	0.5702	0.6533	0.8499	0.1384	0.3758
3	2	0.8636	0.7770	0.4142	0.5404	0.8136	0.1879	0.2798
3	3	0.7805	0.8287	0.2953	0.4354	0.6947	0.2780	0.1255
3	4	0.7065	0.7920	0.2918	0.4265	0.6680	0.3686	0.0954
4	1	0.9046	0.7097	0.5883	0.6433	0.8653	0.1497	0.3968
4	2	0.8061	0.7230	0.3305	0.4537	0.6833	0.2421	0.1379
4	3	0.7350	0.7864	0.3142	0.4490	0.6635	0.3429	0.1036
5	1	0.8793	0.8997	0.5341	0.6703	0.7092	0.2240	0.2278
5	2	0.7881	0.9360	0.3939	0.5545	0.6751	0.3314	0.1480
6	1	0.8098	0.6919	0.2361	0.3520	0.7269	0.2208	0.1339

Tabella A.5. Tabella con i risultati del modello LR, parte 1. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	a	b	VaR
1	1	0.1375	1.2924	2815
1	2	0.3519	2.2003	2748
1	3	0.6104	2.5978	2966
1	4	0.8381	2.7631	3148
1	5	2.5001	5.4467	2982
1	6	4.2203	6.4288	3220
2	1	0.1570	1.1897	3187
2	2	0.4203	2.0690	3054
2	3	0.6158	2.2210	3280
2	4	2.1298	4.9334	2993
2	5	3.7844	6.0235	3218
3	1	0.2299	1.4303	3212
3	2	0.4836	2.0895	3172
3	3	1.9369	5.0298	2869
3	4	3.4948	5.9843	3152
4	1	0.2276	1.2920	3431
4	2	1.5129	4.7360	2735
4	3	2.9645	5.6798	3077
5	1	0.7587	2.6285	3150
5	2	1.9076	3.8470	3311
6	1	1.4277	5.0373	2574

Tabella A.6. Tabella con i risultati del modello LR, parte 2. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.9460	0.7509	0.6745	0.7107	0.8998	0.0933	0.4507
1	2	0.9048	0.7609	0.4715	0.5822	0.8358	0.1349	0.3264
1	3	0.8640	0.8247	0.3939	0.5331	0.8001	0.1891	0.2810
1	4	0.8291	0.8466	0.3525	0.4977	0.7880	0.2354	0.2585
1	5	0.7500	0.8171	0.2906	0.4287	0.7337	0.3176	0.1858
1	6	0.6825	0.7432	0.3215	0.4489	0.7141	0.3963	0.1552
2	1	0.9330	0.7700	0.6045	0.6773	0.8841	0.1172	0.5021
2	2	0.8776	0.7841	0.4078	0.5366	0.8195	0.1616	0.3447
2	3	0.8415	0.8268	0.3645	0.5059	0.8053	0.2181	0.2994
2	4	0.7651	0.8420	0.2959	0.4379	0.7447	0.2921	0.2204
2	5	0.6953	0.7631	0.3221	0.4530	0.7270	0.3798	0.1915
3	1	0.9140	0.7777	0.5609	0.6518	0.8670	0.1398	0.4386
3	2	0.8658	0.7633	0.4444	0.5617	0.8263	0.1940	0.3373
3	3	0.7863	0.7814	0.3534	0.4867	0.7702	0.2660	0.2578
3	4	0.7185	0.7107	0.4171	0.5256	0.7423	0.3810	0.2090
4	1	0.9118	0.7612	0.5781	0.6571	0.8748	0.1459	0.4734
4	2	0.8170	0.7323	0.3915	0.5102	0.7913	0.2361	0.3105
4	3	0.7451	0.7245	0.4170	0.5293	0.7586	0.3416	0.2512
5	1	0.9083	0.9207	0.6574	0.7671	0.9205	0.2216	0.5988
5	2	0.8103	0.8782	0.5029	0.6396	0.8300	0.3245	0.3969
6	1	0.8191	0.6476	0.3808	0.4796	0.7680	0.2238	0.2382

Tabella A.7. Tabella con i risultati del modello MLP, parte 1. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	a	b	VaR
1	1	0.1137	1.1044	2933
1	2	0.2784	1.7842	2937
1	3	0.4836	2.0738	3187
1	4	0.6750	2.1917	3390
1	5	1.3907	2.9879	3463
1	6	2.1568	3.2847	3676
2	1	0.1162	0.8749	3559
2	2	0.3071	1.5933	3285
2	3	0.5102	1.8287	3491
2	4	1.0326	2.5022	3518
2	5	1.6031	2.6177	3801
3	1	0.1788	1.1002	3532
3	2	0.3811	1.5830	3513
3	3	0.7653	2.1118	3571
3	4	1.4414	2.3414	3897
4	1	0.1622	0.9496	3767
4	2	0.5241	1.6957	3663
4	3	1.0178	1.9610	3928
5	1	0.1484	0.5212	4764
5	2	0.4929	1.0257	4473
6	1	0.7155	2.4808	3206

Tabella A.8. Tabella con i risultati del modello MLP, parte 2. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.9393	0.7424	0.5864	0.6553	0.8517	0.0968	0.5554
1	2	0.8940	0.6984	0.4336	0.5350	0.7682	0.1375	0.4498
1	3	0.8475	0.6882	0.4142	0.5171	0.7330	0.1898	0.4216
1	4	0.8063	0.6590	0.4011	0.4987	0.7197	0.2306	0.4072
1	5	0.7125	0.5799	0.3970	0.4713	0.6626	0.3164	0.3476
1	6	0.6381	0.5617	0.4558	0.5033	0.6468	0.3990	0.3229
2	1	0.9281	0.7582	0.5616	0.6452	0.8118	0.1172	0.5880
2	2	0.8663	0.6967	0.4078	0.5145	0.7443	0.1690	0.4621
2	3	0.8206	0.6593	0.4026	0.5000	0.7410	0.2181	0.4406
2	4	0.7253	0.5842	0.3870	0.4656	0.6816	0.3044	0.3711
2	5	0.6495	0.5664	0.4480	0.5003	0.6615	0.3877	0.3401
3	1	0.9085	0.7727	0.5133	0.6168	0.8082	0.1342	0.5414
3	2	0.8561	0.7122	0.4306	0.5367	0.7681	0.1835	0.4515
3	3	0.7616	0.6342	0.3982	0.4892	0.7045	0.2739	0.3833
3	4	0.6743	0.5843	0.4478	0.5070	0.6762	0.3662	0.3591
4	1	0.9031	0.7120	0.5667	0.6311	0.8251	0.1497	0.5790
4	2	0.7990	0.6326	0.4161	0.5020	0.7247	0.2386	0.4633
4	3	0.7040	0.5910	0.4500	0.5110	0.6920	0.3399	0.3975
5	1	0.8798	0.8328	0.5965	0.6951	0.8439	0.2198	0.6201
5	2	0.7686	0.7080	0.5253	0.6032	0.7625	0.3301	0.4865
6	1	0.7950	0.5486	0.3564	0.4321	0.7077	0.2194	0.3705

Tabella A.9. Tabella con i risultati del modello KNN, parte 1. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	a	b	VaR
1	1	0.0774	0.7225	3473
1	2	0.1681	1.0542	3561
1	3	0.2602	1.1108	3895
1	4	0.3355	1.1196	4092
1	5	0.5937	1.2824	4251
1	6	0.8365	1.2597	4459
2	1	0.0821	0.6180	3987
2	2	0.1967	0.9667	3926
2	3	0.2767	0.9919	4170
2	4	0.5155	1.1781	4297
2	5	0.7520	1.1874	4484
3	1	0.1136	0.7330	3972
3	2	0.2228	0.9914	3990
3	3	0.4405	1.1677	4212
3	4	0.6533	1.1307	4480
4	1	0.1087	0.6177	4307
4	2	0.3084	0.9829	4264
4	3	0.5149	0.9999	4526
5	1	0.1345	0.4775	4812
5	2	0.3482	0.7066	4756
6	1	0.3725	1.3255	3853

Tabella A.10. Tabella con i risultati del modello KNN, parte 2. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	**accuracy**	**precision**	**recall**	**F1**	**AUC**	p	ρ
1	1	0.9396	0.6866	0.6913	0.6890	0.8288	0.0971	0.4165
1	2	0.9001	0.6952	0.4909	0.5754	0.7282	0.1410	0.2640
1	3	0.8618	0.7688	0.3928	0.5199	0.6825	0.1950	0.2188
1	4	0.8306	0.8150	0.3439	0.4837	0.6602	0.2385	0.1862
1	5	0.7506	0.8356	0.2584	0.3948	0.6175	0.3195	0.1277
1	6	0.6763	0.8613	0.2127	0.3412	0.5952	0.4000	0.0948
2	1	0.9320	0.7224	0.6291	0.6725	0.7994	0.1189	0.4243
2	2	0.8800	0.7293	0.4290	0.5402	0.6988	0.1749	0.2639
2	3	0.8460	0.7931	0.3639	0.4989	0.6692	0.2248	0.2283
2	4	0.7663	0.8379	0.2709	0.4094	0.6242	0.3087	0.1546
2	5	0.6905	0.8741	0.2213	0.3531	0.6008	0.3918	0.1123
3	1	0.9180	0.7417	0.6068	0.6675	0.7868	0.1395	0.3649
3	2	0.8723	0.7477	0.4543	0.5652	0.7100	0.1917	0.2441
3	3	0.7898	0.8003	0.3208	0.4581	0.6451	0.2822	0.1602
3	4	0.7135	0.8588	0.2603	0.3995	0.6178	0.3716	0.1186
4	1	0.9130	0.7051	0.6705	0.6874	0.8119	0.1481	0.3663
4	2	0.8213	0.7235	0.4101	0.5235	0.6804	0.2436	0.1899
4	3	0.7418	0.7960	0.3190	0.4555	0.6386	0.3448	0.1419
5	1	0.8896	0.9042	0.5716	0.7004	0.7769	0.2260	0.4232
5	2	0.7950	0.9357	0.4053	0.5656	0.6958	0.3346	0.2686
6	1	0.8295	0.7275	0.3377	0.4612	0.6514	0.2243	0.1352

Tabella A.11. Tabella con i risultati del modello AB, parte 1. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

t	j	a	b	VaR
1	1	0.1360	1.2640	2854
1	2	0.3928	2.3933	2692
1	3	0.6962	2.8728	2899
1	4	1.0420	3.3267	3008
1	5	2.1821	4.6469	3117
1	6	3.8154	5.7219	3306
2	1	0.1613	1.1950	3211
2	2	0.4876	2.3004	2987
2	3	0.7596	2.6193	3157
2	4	1.6877	3.7785	3229
2	5	3.0962	4.8044	3387
3	1	0.2427	1.4972	3170
3	2	0.5932	2.5012	3012
3	3	1.4787	3.7610	3115
3	4	2.7614	4.6682	3329
4	1	0.2562	1.4730	3266
4	2	1.0388	3.2249	3062
4	3	2.0836	3.9588	3343
5	1	0.3080	1.0543	4138
5	2	0.9109	1.8113	3979
6	1	1.4339	4.9574	2606

Tabella A.12. Tabella con i risultati del modello AB, parte 2. L'indice temporale t (da 1 a 6) rappresenta i mesi da aprile a settembre, mentre j indica l'orizzonte di previsione, espresso in mesi.

Bibliografia

McNeil, A. J., Frey, R., and Embrechts, P. *Quantitative Risk Management.* Volume 3, Princeton University Press, 2005.

Frey, R. and McNeil, A. J. *Modelling dependent defaults.* Technical report, ETH Zurich, 2001.

Frey, R. and McNeil, A. J. *Dependent Defaults in Models of Portfolio Credit Risk.* Journal of Risk, 2003.

Sigrist, F. and Leuenberger, N. *Machine Learning for Corporate Default Risk: Multi-period Prediction, Frailty Correlation, Loan Portfolios, and Tail Probabilities.* European Journal of Operational Research, 2022.

Deisenroth, M. P., Faisal, A. A., and Ong, C. S. *Mathematics for Machine Learning.* Cambridge University Press, 2020.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R.* Second Edition, Springer, 2021.