POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Tecniche di Machine Learning per la stima delle probabilità di default e applicazione ad un caso di pricing di una cartolarizzazione

Relatrice
prof. Patrizia Semeraro
Tutor Aziendale
Dott. Diego Giovannini
Dott. Alessandra Mazzone

Candidato
Carmen Frasca

Anno Accademico 2022-2023

Sommario

L'obiettivo di questa tesi è studiare il rischio di credito associato ad un portafoglio cartolarizzato, i cui rendimenti derivano da un pool di prestiti sottostanti che sono soggetti al rischio di insolvenza da parte del debitore. Poiché il ripiegamento delle note è in così stretto contatto con il pagamento delle rate dei prestiti sottostanti, è necessario analizzare e modellizzare la distribuzione di probabilità congiunta dei default dei debitori. Il modello standard che viene utilizzato per modellizzare default congiunti è il modello scambiabile di Bernoulli, che viene applicato ad un dataset di prestiti per stimare il rischio e la probabilità di default. Il dataset viene quindi separato mediante un algoritmo di clusterizzazione in modo da raggruppare i dati in maniera più omogenea e rientrare nelle ipotesi in cui il portafoglio è composto da pesi uniformi. Si confrontano diversi algoritmi di Machine Learning con l'algoritmo di Regressione Logistica, che viene classicamente usato per stimare la probabilità di default. Si vedrà che gli algoritmi di Machine Learning, a livello aggregato di portafoglio, ottengono prestazioni migliori nello stimare il rischio (misurato dal Value at Risk) associato al numero di clienti che vanno in insolvenza. Partendo dalle stime di questi algoritmi e dalle analisi dei prestiti contenuti all'interno di questi portafogli si stima il *fair price* di questo titolo e l'entità delle cedole da ripagare ai detentori del titolo e si confronta in che modo la forma funzionale della distribuzione del numero di default impatta il ripagamento delle note e la durata dell'investimento stesso.

Indice

Introduzione

Un ABS è uno strumento finanziario garantito da un insieme di crediti relativi a prestiti e mutui, realizzato a fronte di un'operazione di cartolarizzazione. Chi investe in un ABS viene ripagato sfruttando le rate sui prestiti che si trovano all'interno di un ABS. Il rischio relativo a questo investimento, dunque, deriva dal rischio di insolvenza del debitore che ha contratto un prestito presente all'interno di un ABS. Per prezzare uno strumento di questo tipo è fondamentale quindi avere una buona stima della probabilità di default individuale di ogni singolo debitore, e quella aggregata a livello di portafoglio. La metodologia classica per prezzare questo strumento è data dall'analisi del dato storico e all'analisi di alcune condizioni geografiche. In questa tesi si cerca di utilizzare degli strumenti alternativi e in particolare si cerca di sfruttare le informazioni economiche e sociali raccolte su ogni cliente per prevedere la probabilità di default attraverso degli algoritmi di Machine Learning e confrontando i risultati ottenuti. A partire dalle probabilità di default individuali stimate con questi algoritmi, si stima l'entità della perdita aggregata di un portafoglio omogeneo attraverso un modello scambiabile di Bernoulli, di cui viene calcolata un'approssimazione parametrica per poter lavorare in grandi dimensioni e viene confrontata con la distribuzione non parametrica (che può essere analizzata nel caso di un portafoglio di piccole dimensioni). In seguito, per ottenere dei risultati numerici, utilizzando una parte di questi dati, si costruiscono due titoli ABS, con scadenza rispettivamente di tre e cinque anni, per analizzare in che modo la probabilità di default stimata dagli algoritmi di Machine Learning impatta i risultati relativi al pricing della cartolarizzazione e al ripagamento delle note del titolo. Si prosegue poi analizzando l'impatto, sulla cartolarizzazione e sui fattori che incidono sul suo prezzo, derivante dalle differenti realizzazioni dei default nel tempo. Si vedrà che una maggiore concentrazione di default durante i mesi iniziali ha un impatto significativo sul ripagamento delle cedole e che invece, se si ha una concentrazione maggiore durante i mesi finali, quasi alla scadenza del prestito, l'impatto dei default è trascurabile. I risultati mostreranno che una cartolarizzazione è uno strumento molto complesso da analizzare ed giustifica l'interesse verso la ricerca di strade alternative per comprendere meglio il titolo e i rischi che si assumono nello scegliere un particolare modello piuttosto che un altro.

Capitolo 1

Strumenti finanziari

Gli ABS (*Asset Backed Securities*) sono strumenti finanziari garantiti da un insieme di crediti relativi a mutui e prestiti. Rappresentano dei titoli che conferiscono, a chi li possiede, il rimborso dell'investimento alla maturità del titolo e delle cedole periodiche come interesse. Sono simili a delle obbligazioni, con la differenza che, nel caso degli ABS, le note vengono ripagate tramite i flussi di cassa generati dai prestiti stessi. Vengono create a seguito di un'operazione di cartolarizzazione, che consiste nella cessione dei crediti, da parte di istituzioni finanziare come banche, a delle istituzioni speciali che li trasformano in titoli che possono essere venduti sul mercato. La cessione di questi crediti ha notevoli vantaggi per la banca, in quanto ottiene una maggiore liquidità, poiché i crediti derivanti da mutui e prestiti sono considerati dei beni non liquidi, mentre al contrario, i titoli creati a partire da questi crediti, essendo venduti sul mercato, creano liquidità. Inoltre, tramite questo processo, il rischio di insolvenza associato a questi crediti viene trasferito, almeno in parte, agli investitori, che verranno ricompensati per il rischio assunto. Per un investitore, scegliere di comprare un titolo di questo tipo può essere vantaggioso perché è garantito da un portafoglio formato da decine, centinaia o migliaia di prestiti (in base alla natura dei prestiti sottostanti, che possono essere personali o commerciali, e al loro importo) per cui, nonostante il rischio di insolvenza legato ad ogni cliente, la probabilità di inadempienza di tutti i sottostanti è estremamente remota (a meno di situazioni di stress economico, che porta ad un aumento del rischio sistematico), per cui un ABS è un investimento estremamente diversificato, e rispetto ad un titolo emesso da un singolo ente, è molto più tutelato dal punto di vista del rischio idiosincratico. Inoltre, nella struttura di un ABS vengono inclusi dei metodi di *credit enhancement*, ovvero delle tecniche che si utilizzano per diminuire il rischio associato all'insolvenza dei clienti ed è una misura della protezione che viene garantita agli investitori nel caso di insolvenza. Un altro grande vantaggio per gli investitori è la possibilità di scegliere tra un'ampia gamma di profili di rischio/rendimento, scegliendo il titolo più adatto alle proprie esigenze, poiché le note non vengono pagate allo stesso modo a tutti gli investitori, ma sono divise in *tranches* caratterizzati da rischi e rendimento differenti, e tempistiche di ripagamento separate. Il *pricing* di uno strumento cartolarizzato avviene tramite la metodologia del *discounted cashflow*, ovvero i flussi di cassa attesi, prodotti dal *pool* di prestiti (sulla base di determinate ipotesi di performance, tra cui tasso di *prepayment* e di default), vengono scontati utilizzando i tassi d'interesse

risk free ed una componente creditizia rappresentativa del rischio di credito associato agli asset sottostanti (*credit spread*). In questo lavoro si cerca una strada alternativa, in cui si cerca di trovare il *fair value* di un titolo cartolarizzato calcolando la probabilità di default congiunta dei prestiti contenuti al suo interno, e la correlazione tra i default.

1.1 Cartolarizzazioni

La creazione di questi titoli avviene in seguito ad un'operazione di cartolarizzazione. Una cartolarizzazione è un'operazione finanziaria che permette la trasformazione di un insieme di crediti derivanti da prestiti o mutui, che sono considerati illiquidi, in titoli finanziari collocabili sul mercato. È un'operazione molto complessa che si compone di più passi e prevede il coinvolgimento di più enti, che vengono elencati di seguito:

1. l'*originator*, solitamente una banca, che cede i crediti alla SPV per la creazione di questi titoli;
2. la SPV (*Special Purpose Vehicle*), che acquista i crediti ceduti dall'*originator* e provvede ad emettere i titoli garantiti da questi crediti. Inoltre è addetta a raccogliere i pagamenti dai prestiti sottostanti e trasferirli, sotto forma di interessi e principale, agli investitori.
3. l'agenzia di rating, che analizza e valuta la qualità dei crediti ceduti.
4. gli investitori, che in base alle loro esigenze, acquistano i titoli emessi dalla SPV.

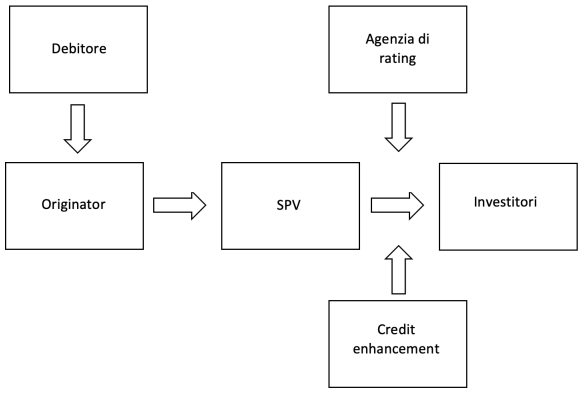

Figura 1.1: Schema generale di una cartolarizzazione.

1.1.1 Credit Enhancement

Le tecniche di *credit enhancement* sono delle garanzie che vengono applicate ai titoli cartolarizzati per ridurre il rischio legato all'insolvibilità di uno o più prestiti e garantire quindi una maggiore protezione nei confronti degli investitori, e vengono decise dalla SPV,

in accordo con le agenzie di rating. Il *credit enhancement* di un titolo cartolarizzato può modificare e migliorarne il rating, per cui è un elemento fondamentale nella struttura di ABS.
Le forme più comuni di *credit ehnancement* sono:

1. l'*overcollateralization*: è una tecnica in cui l'*originator* cede dei crediti di valore più alto rispetto a quello richiesto per ripagare le note agli investitori. In questo modo, chi investe in ABS ha una maggiore protezione in caso di adempienza di uno o più debitori.

2. il *credit tranching* è una tecnica utilizzata per separare un insieme di asset, come mutui o altri tipi di prestito, in categorie di rischio separate, caratterizzate da un livello di rischio e di rendimento differenti. In questo modo, gli investitori possono scegliere il profilo di rischio e rendimento più adatti alle loro esigenze. La tranche senior è quella che viene ripagata per prima, ma ha un tasso di interesse più basso, mentre le tranche più basse sono più rischiose, vengono ripagate dopo la senior, ma hanno un tasso di interesse più alto. Per la tranche senior, le tranche subordinate sono un ulteriore fonte di protezione, in quanto sono le tranche che assumono la maggior parte di rischio.

3. L'*excess spread* è la differenza tra il rendimento percentuale (*yield*)di un insieme di asset (come ad esempio il tasso di interesse pagato sul mutuo) e l'interesse pagato sui titoli garantiti da questi asset. È una misura della redditività di un'ABS e viene utilizzata per coprire le perdite che possono derivare dai sottostanti. Questa quantità è importante nel misurare la solvibilità di un titolo di questo tipo, in quanto è un cuscinetto contro eventuali perdite nei sottostanti. Un *excess spread* più alto indica una maggiore protezione nei confronti degli investitori in quanto il rischio associato al default dei clienti che hanno contratto un debito viene mitigato, appunto, da questo elemento.

4. Il *cash reserve* è un fondo monetario che viene messo da parte per coprire le perdite nei sottostanti. Viene utilizzato se i flussi di cassa non sono sufficienti per ripagare le note agli investitori.

1.1.2 Payoff di una cartolarizzazione

Il payoff di un titolo garantito da una cartolarizzazione è regolato da una struttura nota come *waterfall structure* (struttura a cascata) e descrive in che modo i flussi di cassa derivanti dai prestiti sottostanti sono distribuiti tra gli investitori, sotto forma di interessi e principale. La struttura a cascata deriva dalla divisione in tranches, in quanto vengono ripagate seguendo tempistiche e priorità differenti.
Prima di procedere con la descrizione della distribuzione del pagamento delle note ai vari investitori, è necessario quindi vedere in dettaglio come viene effettuato il *credit tranching*.
La divisione più classica comprende tre differenti classi di rischio:

1. la *tranche senior*;

2. la *tranche mezzanine*;

3. la *tranche equity*.

La tranche senior è caratterizzata da un rating molto alto e da un rischio basso, è la tranche con rendimento più basso, ma riceve i pagamenti prima delle altre due tranche.
La tranche mezzanine è subordinata alla tranche senior, in quanto gli investitori che hanno acquistato questi titoli ricevono il loro pagamento solo se la tranche senior è stata completamente rimborsata, ma ha un rendimento più alto rispetto alla senior.
La tranche equity è subordinata alle altre due, perciò viene rimborsata solo in seguito al pagamento integrale delle altre due tranches, con il denaro residuo. E' la tranche con il rischio più alto, perciò è anche quella con rendimento più alto. Solitamente questa tranche non viene venduta sul mercato, ma viene acquistata dalla banca stessa. Nel modello considerato in questa tesi si considera una struttura di tipo *full sequential*, in cui una tranche viene ripagata solo se la tranche precedente è stata ripagata del tutto, sia come quota capitale che come interesse.

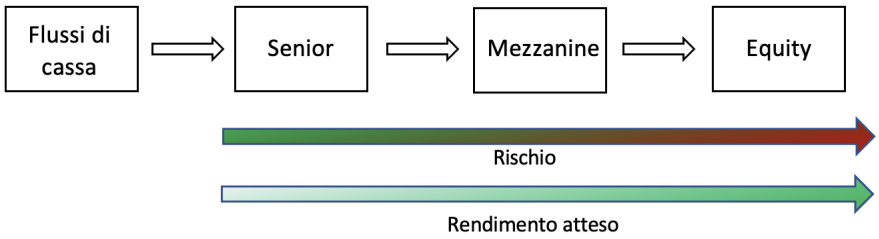

Figura 1.2: Struttura a cascata del payoff di un ABS e relativo rischio e rendimento atteso.

Attraverso questa descrizione è chiaro che, se uno o più prestiti sottostanti va in insolvenza, la classe che per prima subirà gli effetti di questa insolvenza è la classe che viene ripagata per ultima, in quanto il denaro raccolto tramite i flussi di cassa derivanti dai prestiti potrebbe esaurirsi prima che questa classe venga ripagata.
La struttura che regola il pagamento delle note dipende da alcuni fattori, che devono essere analizzati e valutati nel modo giusto. I fattori più importanti che incidono nel ripagamento delle note sono:

1. il *prepayment risk*, cioè il rischio che i prestiti inseriti nel portafoglio cartolarizzato vengano pagati in anticipo. In base alle condizioni esterne di mercato, se i tassi di interesse sono più alti rispetto al momento in cui è stato emesso il prestito, il tasso di prepayment osservato potrebbe essere più basso rispetto a quello stimato, se invece i tassi di interesse scendono, il tasso di prepayment osservato potrebbe essere più alto rispetto a quello stimato. In entrambi i casi questo fattore influisce sui flussi di cassa attesi.

2. i tassi di interesse: se il tasso di interesse sale, il valore di un ABS potrebbe diminuire, in quanto per un investitore sarebbe più conveniente investire in qualche asset *risk-free* sfruttando appunto gli elevati tassi di interesse, senza però dover correre i rischi di chi investe in ABS.

3. il tasso di default relativo ai prestiti sottostanti: se questo tasso di default risulta essere più elevato di quanto stimato (e tutelato tramite alcune forme di garanzie) , questo si traduce in perdite monetarie per gli investitori in ABS. è quindi fondamentale avere una stima accurata ed attendibile delle probabilità di default associate ai prestiti sottostanti una cartolarizzazione, per potersi tutelare attraverso forme di *credit enhancement* adeguate.

Il prezzo di un portafoglio cartolarizzato si stima tramite il metodo dei flussi di cassa scontati. Questo metodo consiste nello stimare i flussi di cassa che verranno generati dai sottostanti, derivanti dal pagamento delle rate sui prestiti contenuti all'interno dell'ABS e scontarli tramite un opportuno tasso di interesse, formato da due componenti: la somma tra il tasso di interesse *risk-free* e il tasso relativo al rischio di credito, chiamato anche *credit spread*, che sarà tanto più alto quanto più è rischiosa la tranche in cui si investe. Si aggiunge una componente relativa al valore nominale dell'investimento, che viene suddiviso in quote capitali che vengono ripagate insieme alle cedole.

$$P = \frac{c_1}{(1+y)} + \frac{c_2}{(1+y)^2} + \cdots + \frac{c_N}{(1+y)^N} \tag{1.1}$$

Dove $c_1, c_2, \ldots, c_N$ sono i flussi di cassa che ad ogni periodo $i = 1, \ldots, N$ vengono pagati agli investitori ed N è il numero totale di pagamenti. La formula (1.1) è una formula generale per il *pricing* di un ABS, però i parametri contenuti all'interno della formula dipendono molto dalla qualità dei prestiti, dalle dimensione del *credit enhancement* e dai rischi di pagamento anticipato e di default, oltre che dai sottostanti contenuti all'interno del portafoglio.
Poiché i flussi di cassa presenti nella formula derivano dal ripagamento delle rate sui prestiti, è chiaro che avere una buona analisi e una buona valutazione dei prestiti sottostanti, e in particolare della distribuzione del numero di default, è fondamentale per valutare correttamente un titolo cartolarizzato e capirne il rischio associato ed è la motivazione che spinge le analisi successive, in cui, tramite algoritmi di Machine Learning, si cerca di estrarre la probabilità di default di ogni cliente e la correlazione tra i vari clienti che hanno contratto dei debiti, attraverso l'analisi delle loro caratteristiche personali, come il reddito e l'impiego.

1.2 Modellizzazione

Sia dato il vettore aleatorio

$$Y = (Y_1, Y_2, \ldots, Y_d)$$

degli indicatori dei default di un insieme di d clienti, e sia dato il vettore $P = (w_1, w_2, \ldots, w_d)$ dei pesi di un portafoglio cartolarizzato associato ai d prestiti concessi ai d clienti, dove

$w_i \in (0,1]$ e $\sum_{i=1}^{d} w_i = 1$. Per modellizzare la perdita del portafoglio, si può considerare la somma L delle perdite individuali percentuali, data da:

$$L = \sum_{i=1}^{d} w_i Y_i \tag{1.2}$$

Se si considerano dei pesi omogenei $w_i = \frac{1}{d}$, allora la variabile aleatoria

$$S = \sum_{i=1}^{d} Y_i \tag{1.3}$$

che indica il numero di default, caratterizza in maniera univoca la perdita, in quanto $L = \frac{S}{d}$. Per rappresentare il vettore $\mathbf{Y}$ e la variabile S si considera un modello di mistura di Bernoulli, in cui si assume che il rischio di insolvenza di un debitore dipenda da un insieme di fattori economici comuni. Data una realizzazione di questi fattori, ogni default è indipendente. [(Modello di mistura di Bernoulli)] Dato $p < m$ e un vettore aleatorio $p-$dimensionale $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_p)'$, un vettore aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ è distribuito secondo un modello di mistura di Bernoulli con vettore di fattori $\mathbf{\Psi}$ se esistono funzioni $p_i : \mathbb{R}^p \to [0,1], \ \leq i \leq m$ tali che, condizionatamente a $\mathbf{\Psi}$, le componenti di $\mathbf{Y}$ sono variabili aleatorie indipendenti di Bernoulli e soddisfano $\mathbb{P}[Y_i = 1|\mathbf{\Psi} = \psi] = p_i(\psi)$. per $\mathbf{y} = (y_1, \ldots, y_d)' \in \{0,1\}^d$ si ha:

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}|\mathbf{\Psi} = \psi) = \prod_{i=1}^{d} p_i(\psi)^{y_i}(1 - p_i(\psi))^{1-y_i} \tag{1.4}$$

Questo modello assume che la dipendenza tra i default derivi da un insieme di fattori economici comuni, e che, condizionatamente a questi fattori, i default siano indipendenti.

1.2.1 Modello scambiabile di Bernoulli

È possibile semplificare il modello assumendo che il vettore degli indicatori di default $\mathbf{Y}$ sia *scambiabile*. Un vettore aleatorio si dice scambiabile se

$$(Y_1, \ldots, Y_m) \stackrel{d}{=} (Y_{\sigma(1)}, \ldots, Y_{\sigma(m)})$$

per ogni permutazione $(\sigma(1), \ldots, \sigma(m))$ di $(1, \ldots, m)$. Questa relazione modellizza la situazione di un portafoglio composto da un gruppo omogeneo di prestiti, per cui esiste una sola distribuzione di probabilità di insolvenza individuale, per cui $p_i(\mathbf{\Psi}) = p(\mathbf{\Psi})$. Inoltre, si suppone che questa distribuzione di probabilità dipenda da un insieme di covariate osservabili, comuni a tutti i clienti, perciò si può scrivere $p(\mathbf{X})$. Mettendo insieme tutte queste assunzioni, è possibile definire il modello di Bernoulli scambiabile ad un fattore. Data una variabile aleatoria Q, il vettore aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ segue un modello misto di Bernoulli scambiabile, con variabile mista Q e supporto su $[0,1]$, se condizionatamente a Q, l'indicatore di default $\mathbf{Y}$ è un vettore di variabili aleatorie indipendenti di Bernoulli con $\mathbb{P}(Y_i = 1|Q) = Q$.

La variabile mista è distribuita secondo una Beta, ovvero $Q \sim \beta(a, b)$. $Q = Q_h$ dipende dalle covariate osservabili $\mathbf{X}$, per cui le realizzazioni di Q_h sono funzioni delle realizzazioni di $\mathbf{X}$, ovvero $q_i = h(\mathbf{x}_i)$, per cui la probabilità di default condizionata è

$$\mathbb{P}(Y_i = 1|Q_h = h(\mathbf{x}_i)) = h(\mathbf{x}_i) \tag{1.5}$$

e la probabilità di default non condizionata si ottiene integrando sulla distribuzione $G_h(q)$ di Q

$$\mathbb{P}(Y_i = 1) = \int_0^1 q dG_h(q) \tag{1.6}$$

e la funzione di densità discreta non condizionata di $p_{\mathbf{Y}}((Y))$ si ottiene integrando su G_h:

$$p_{\mathbf{Y}}(\mathbf{y}) = \mathbb{P}(\mathbf{Y} = \mathbf{y}) = \int_0^1 \prod_{i=1}^{d} q^{y_i}(1-q)^{1-y_i} dG_h(q) \tag{1.7}$$

In questo contesto viene introdotta la seguente notazione per le probabilità di default individuali e congiunte:

$$\pi_k = \mathbb{P}(Y_{i1} = 1, \ldots, Y_{ik} = 1), \;\; \{i_1, \ldots, i_k\} \subset \{1, \ldots, m\}, \;\; 1 \leq k \leq m, \tag{1.8}$$

$$\pi := \pi_1 = p = \mathbb{P}(Y_i = 1), \;\; i \in \{1, \ldots, m\}. \tag{1.9}$$

Il momento k-esimo π_k è la probabilità congiunta che un arbitrario sottogruppo di k clienti vada in default. Sfruttando la scambiabilità del vettore $\mathbf{Y}$ e i risultati relativi ai momenti di una variabile aleatoria di Bernoulli, si possono ricavare delle formule semplici per alcune quantità di interesse[1]:

$$p = \mathbb{E}(Y_i), \tag{1.10}$$

$$Var(Y_i) = p(1-p), \tag{1.11}$$

$$Cov(Y_i, Y_j) = \pi_2 - p^2 \quad i \neq j, \tag{1.12}$$

$$\rho_Y = \rho(Y_i, Y_j) = \frac{\pi_2 - p^2}{p(1-p)}, \quad i \neq j. \tag{1.13}$$

Condizionatamente a $Q = q$, la variabile aleatoria che rappresenta il numero di default S è la somma di d variabili bernoulliane indipendenti, perciò è una distribuzione binomiale di parametri p e d e ha distribuzione $\mathbb{P}(S = k|Q = q) = \binom{d}{k} p^k (1-p)^{d-k}$, mentre la distribuzione non condizionata $p_S(k)$ del numero dei default S si ottiene integrando rispetto a q:

$$p_S(k) = \mathbb{P}(S = k) = \binom{d}{k} \int_0^1 q^k (1-q)^{d-k} dG(q) \tag{1.14}$$

È necessario fare una distinzione tra π_k e $\mathbb{P}(S = k)$. Si può dimostrare che la distribuzione di S può essere calcolata a partire dai π_k tramite la formula

[1] Si rimanda all'appendice A per ulteriori dettagli sulle formule qui riportate

$$P(S = k) = \sum_{i=0}^{d-k} (-1)^i \frac{d!}{i!k!(d-k-i)!} \pi_{k+i} \tag{1.15}$$

Da cui si vede come la distribuzione di S sia totalmente determinata dalla distribuzione congiunta degli indicatori di default. È possibile dimostrare[2] che in un modello di mistura di Bernoulli, i momenti incrociati di $\mathbf{Y}$ sono i momenti della variabile Q_h:

$$\pi_k = \mathbb{E}[Q_h^k]. \tag{1.16}$$

In particolare $\pi_1 = \mathbb{E}[Q_h] = p$ e, per $i \neq j$, $Cov(Y_i, Y_j) = Var(Q_h) \geq 0$, per cui in un modello di mistura di Bernoulli la correlazione ρ_Y tra i default è sempre nonnegativa. Mettendo insieme i risultati (1.8), (1.15) e (1.16) si capisce che i momenti della variabile mista Q determinano completamente la distribuzione congiunta dei default e quindi la distribuzione del numero di default. Usando i momenti campionari di Q_h, che si possono calcolare a partire dalle covariate osservabili, è possibile stimare la distribuzione di S. Purtroppo, la struttura della formula (1.15) presenta dei problemi di calcolo computazionale quando aumenta il numero di clienti ($d \approx 40$), in quanto all'interno della formula sono presenti dei fattoriali. Per poter lavorare in dimensioni più grandi, è necessario specificare una distribuzione parametrica per la variabile mista Q_h. La distribuzione di Q_h dipende dalla distribuzione congiunta delle covariate $\mathbf{X}$, che non è nota. Il rischio associato alla scelta di una distribuzione parametrica per Q_h è trascurabile, se si suppone che la probabilità di insolvenza p e la correlazione ρ_Y sono noti. Infatti si può dimostrare che la forma parametrica della variabile mista è meno importante rispetto alla stima corretta di p e ρ_Y, e che in particolare la coda di questa distribuzione è particolarmente sensibile a queste quantità. In questo caso si modellizza Q_h come una distribuzione Beta, $Q_h \sim \beta(a, b)$. Trovare i parametri a e b che descrivono una variabile aleatoria Beta in maniera univoca, è possibile, conoscendo almeno due tra le quantità p, π_2, ρ_Y, in quanto, per una distribuzione Beta, valgono le seguenti formule:

$$p = \frac{a}{a+b}, \quad \pi_2 = \frac{a}{a+b} \cdot \frac{a+1}{a+b+1}, \quad \rho_y = \frac{1}{a+b+1}. \tag{1.17}$$

Quando Q_h è distribuita come una Beta di parametri a e b, la distribuzione del numero di default S è distribuita come una variabile aleatoria Beta-Binomiale di parametri d, a e b, dove d è la dimensione del vettore degli indicatori di default $\mathbf{Y}$ o del numero di clienti. Una spiegazione più approfondita sulle variabili Beta e Beta-Binomiale e sulle formule (1.17) può essere trovata nell'appendice B.
Il modello classico per la stima di h e delle probabilità individuali di default è il modello di Regressione Logistica al primo ordine. Viene qui confrontato con dei modelli più complessi (La regressione logistica al secondo ordine (LR2), la Random Forest (RF), l'Ada Boost (AB) e il K-Nearest-Neighbors (KNN)) che riescono a cogliere interazioni non lineari tra le covariate e altre relazioni complesse tra le variabili che non si conoscono in anticipo.

[2]Si rimanda all'appendice A per ulteriori dettagli.

Si cerca di capire il rischio associato alla scelta di un modello di Machine Learning piuttosto che un modello di regressione logistica per stimare le quantità che caratterizzano Q_h e quindi S. Viene quindi analizzato il rischio associato alla distribuzione del numero di default che viene stimata da ognuno di questi modelli. La misura di rischio che viene utilizzata più frequentemente in ambito finanziario è il Value at Risk (VaR), la cui definizione viene qui riportata. Sia L una variabile aleatoria che rappresenta una perdita, con valore atteso finito, allora il VaRα al livello α è definito come

$$\mathrm{VaR}_\alpha(L) = \inf\{l \in \mathbb{R} : \mathbb{P}(L \leq l) \geq \alpha\}$$

Il VaR è quindi il quantile α per la distribuzione L. Solitamente si calcola il VaR ai livelli $\alpha = 0.90, 0.95, 0.99$, per cui il VaR dipende unicamente dalla coda della distribuzione, che dipende principalmente da p e ρ_Y, le cui stime variano in base alla funzione h scelta. Questa analisi viene condotta su un portafoglio di grandi dimensioni ($d = 10000$) e poi su portafogli più omogenei, di dimensioni più piccole ($d = 1000$). Si svolge anche un'analisi su un portafoglio di piccole dimensioni ($d = 25$) per confrontare il VaR che viene stimato in modo non parametrico, sfruttando la formula (1.15) con quello che viene stimato quando viene specificata una distribuzione parametrica Beta per Q_h, per capire quale sia il rischio associato alla scelta di una distribuzione per Q_h.

Capitolo 2

Risultati relative all'applicazione dei modelli di Machine Learning

Il dataset [1] analizzato proviene da una compagnia privata (la Grant Group Funding) che offre prestiti di natura commerciale, che ha raccolto le informazioni di 87500 clienti sulle seguenti 30 caratteristiche economiche e sociali

- ID: Identificativo del cliente.
- Asst_Reg: Il valore dei beni posseduti dal cliente.
- GGGrade: Il punteggio assegnato ad ogni prestito: indica la qualità del prestito stesso. (Assume sette valori, da 1 a 7),.
- Experience: Gli anni lavorativi del cliente.
- Validation: Indica se le informazioni del cliente sono verificate o no (Assume tre valori: Not Vfied = 1, Vfied = 2, Source Vfied = 3).
- Yearly_Income: Indica il reddito annuo del cliente.
- Home_Status: Indica lo stato abitativo del cliente e assume cinque valori (Rent = 1, Mortgage = 2, Own = 3, Other = 4, None = 5).
- Unpaid_2_Years: indica il numero di volte che il cliente ha fatto default negli ultimi due anni.
- Defaulted: Indica il numero di altri prestiti concessi al cliente che non è stato in grado di ripagare.

[1] `https://www.kaggle.com/datasets/marcbuji/loan-default-prediction`

- Designation: Indica l'impiego lavorativo del cliente.
- Debt_to_Income: indica il rapporto DTI (*debt-to-income*) che rappresenta il rapporto tra la rata mensile da pagare sul prestito, e il reddito mensile del cliente.
- Postal_Code: Indica il codice di avviamento postale del cliente e quindi identifica la città in cui è residente.
- Lend_Amount: Indica la quantità di credito concessa al cliente.
- Deprecatory_Records: indica il numero di record negativi registrati a nome del cliente.
- Interest_Charged: il tasso di interesse applicato al prestito.
- Usage_Rate: la quantità di denaro che viene usata per coprire i costi amministrativi relativi all'emissione del prestito.
- Inquiries: Numero di inchieste negli ultimi sei mesi.
- Present_Balance: Il saldo attuale nel conto del cliente.
- Gross Collection: l'importo lordo pagabile a titolo di richiesta di risarcimento.
- Sub_GGGrade: una specifica sul punteggio assegnato al prestito.
- File_Status: assume due valori: whole = 1, fully paid = 2
- State: Lo stato in cui è residente il cliente.
- Account_Open: Il numero di account aperti a nome del cliente.
- Total_Unpaid_CL: la quantità di rate non pagate su altri prestiti a nome del cliente.
- : Duration: La durata del prestito (assume due valori: 3 anni = 1 e 5 anni = 2).
- Unpaid_Amount: la quantità di denaro non ancora pagata su una carta di credito.
- Reason: la ragione per cui viene richiesto un prestito.
- Claim_Type: Se il conto viene aperto individualmente o in condivisione (I : individuale =1, J: condiviso = 2).
- Due Fee: una mora sul pagamento della rata se non viene effettuato entro la scadenza.
- Default: indica se il cliente ha fatto default o meno sul prestito in considerazione. (Variabile target: 0: non default, 1: default).

2.1 Preprocessing

Il primo passo per l'analisi di un dataset è il *preprocessing*, dove si pulisce il dataset attraverso l'eliminazione delle colonne del dataset che non spiegano la variabile *target* **default**, si eliminano le righe del dataset che contengono dei valori mancanti per far sì che tutti i clienti abbiano lo stesso numero di informazioni e si visualizzano le variabili più rilevanti per le analisi successive.

Andando a vedere la distribuzione della variabile target **default** si vede che è fortemente sbilanciata, in quanto più dell'80% di tutte le righe del dataset appartengono alla classe che è riuscita a ripagare il debito. Nella costruzione di un modello è fondamentale tenere in considerazione questo sbilanciamento all'interno della variabile che si cerca di spiegare. Si può lavorare a monte del problema, andando a fare oversampling della classe meno presente, o si può lavorare a valle del problema, usando come metriche di perfomance delle metriche che favoriscano la classificazione corretta della classe positiva piuttosto che quella negativa.

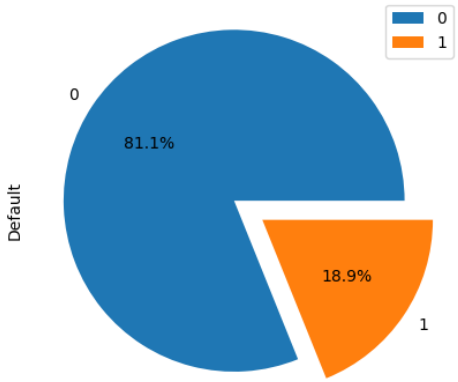

Figura 2.1: Distribuzione della variabile target **default**.

Per comprendere meglio il dataset e le informazioni che contiene, è possibile analizzare e visualizzare in dettaglio le covariate più significative, che spiegano meglio la variabile target.
La prima variabile che viene analizzata è **home_status**, una variabile categorica che descrive la condizione abitativa dei clienti: se hanno una casa di proprietà, vivono in affitto, hanno un mutuo aperto o hanno preferito altre soluzioni. Si osserva che la maggior parte dei clienti vive in una casa su cui ha un mutuo aperto o in affitto, mentre solo una piccola percentuale dei clienti che ha ottenuto un prestito da questa compagnia possiede una casa di proprietà. Dalla Fig. 2.3 è possibile vedere anche la distribuzione del numero di default per ognuno dei valori assunti dalla variabile. Chi vive in una casa di proprietà

ha il tasso di default più basso, che si attesta attorno al 16%, a fronte del 18% medio delle altre due classi.

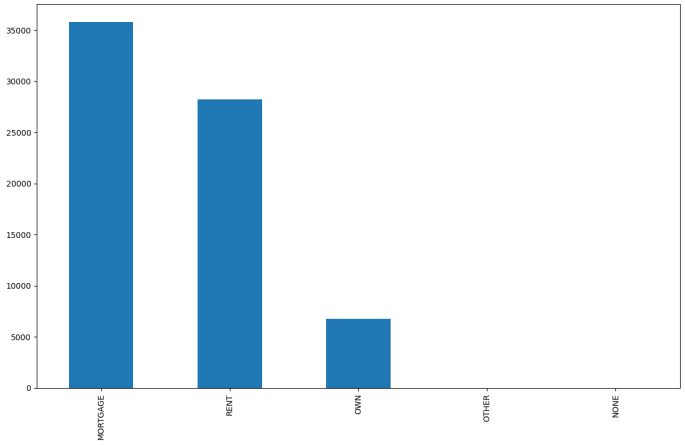

Figura 2.2: Distribuzione della variabile **home_status**.

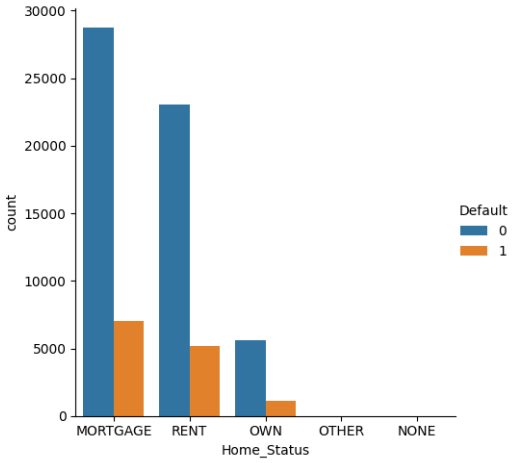

Figura 2.3: Distribuzione del numero di default per ogni valore assunto dalla variabile **home_status**.

La seconda variabile che viene analizzata è **Validation**, e indica se la fonte di reddito del cliente che ha richiesto il prestito è stato verificato o no. Si può vedere che la classe con reddito *source verified* è quella più frequente, ed è quella con tasso di default più basso (del 14.7%), a fronte del 21.5% medio delle altre due classi.

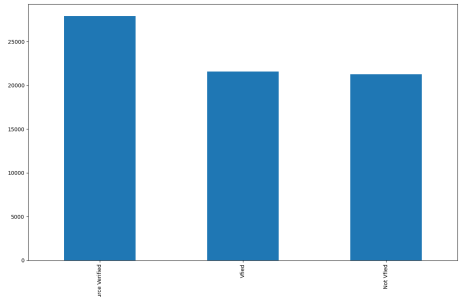

Figura 2.4: Distribuzione della variabile **Validation**.

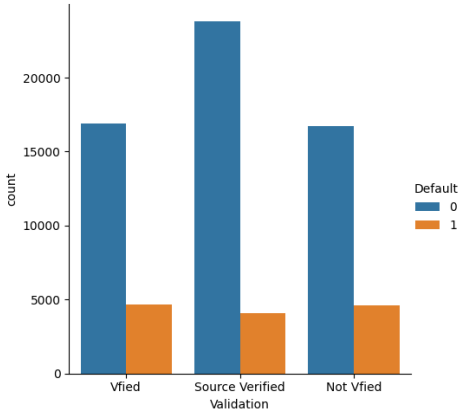

Figura 2.5: Distribuzione del numero di default per le tre classi della variabile **Validation**.

La terza variabile analizzata è la variabile **Duration**, che indica la durata totale del prestito, si osserva che la formula più frequente è quella che rateizza il prestito in 3 anni ma ha un tasso di default più alto rispetto al prestito che ha una durata di cinque anni, infatti il primo tipo ha un tasso di default del 21.5%, mentre il secondo del 12.9%.

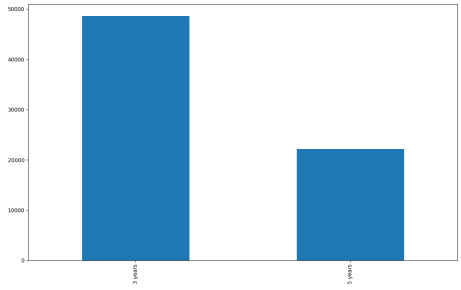

Figura 2.6: Frequenza delle due tipologie di prestito.

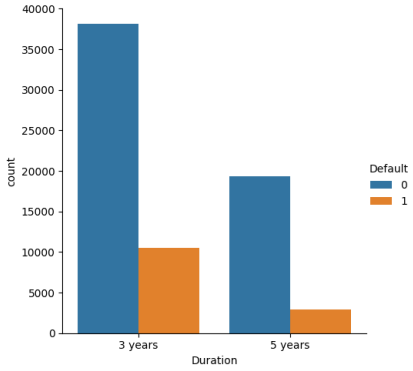

Figura 2.7: Distribuzione del numero di default per le due tipologie di prestito.

La covariata **Experience** indica invece la condizione lavorativa dei clienti, e in particolare il numero di anni di attività svolta da ogni cliente. Si nota che la maggior parte dei clienti inseriti nel dataset considerato ha un'esperienza lavorativa di oltre dieci anni, il che suggerisce che la compagnia tenda a favorire i clienti con una situazione lavorativa più stabile e prolungata come assegnatari di prestito. Questa caratteristica non garantisce però che il cliente considerato abbia una probabilità di default più bassa, in quanto la classe di clienti che lavora da oltre dieci anni ha un tasso di default simile a quello delle altre classi.

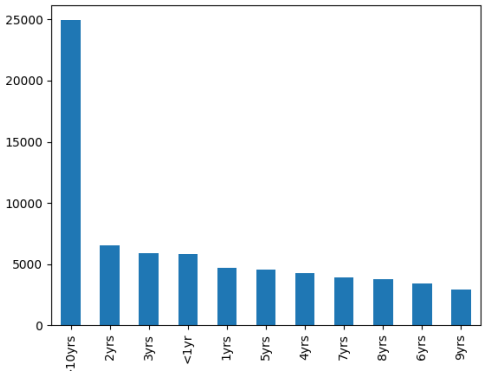

Figura 2.8: Distribuzione della variabile **Experience**.

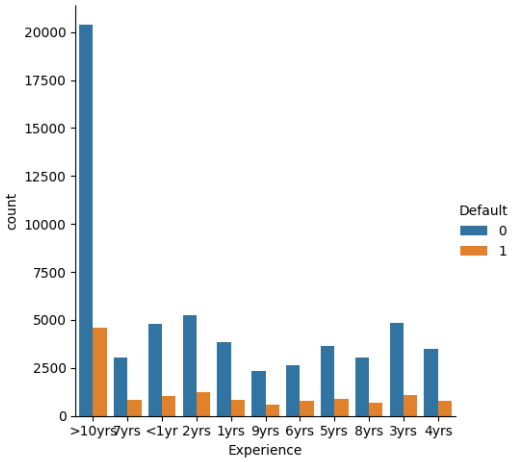

Figura 2.9: Distribuzione del numero di default per ogni valore della variabile **Experience**.

Dopo aver analizzato le variabile categoriche, che assumono un numero finito di valori, è possibile analizzare le variabili continue, per analizzare se sono presenti delle tendenze articolari nella densità di queste variabili. La prima variabile che si può indagare è **Asst_Reg**, che indica il valore di tutti i beni posseduti dal cliente. Si può notare che ci sono tre mode, che potrebbe identificare tre tipologie diverse di clienti in base alla loro situazione economica.

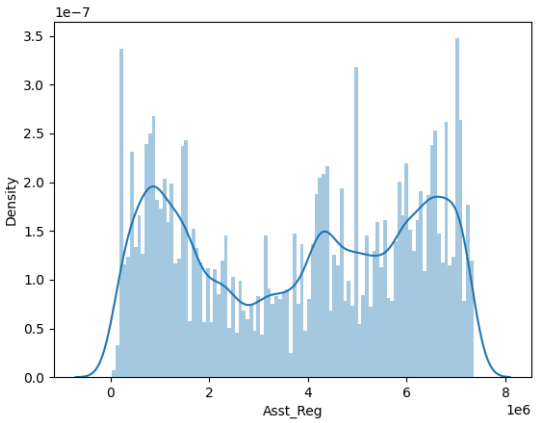

Figura 2.10: Distribuzione di densità della variabile **Asst_Reg**.

La seconda variabile che viene analizzata è quella relativa alla quantità di credito concessa per ogni prestito. Si nota che presenta molte mode, che potrebbe indicare che per ogni livello di prestito (da 1 a 7) si ha un valore medio di credito che viene concesso, in

quanto si supppone che solo per i prestiti ritenuti sufficientemente sicuri si conceda molto credito.

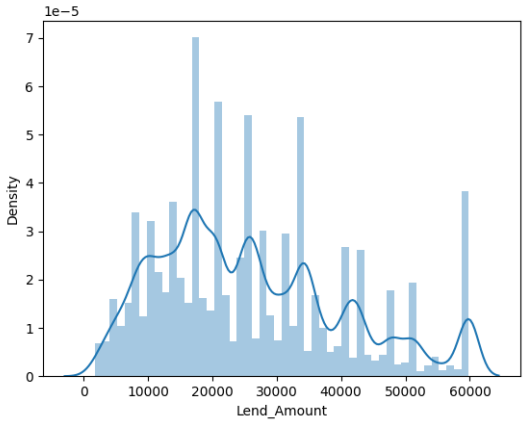

Figura 2.11: Distribuzione della variabile **Lend_Amount**.

Si osserva un andamento molto uniforme nel caso della variabile associata al tasso di interesse considerato per il prestito, che va da circa l'8% fino al 25%, , con un valore medio attorno al 16.51%.

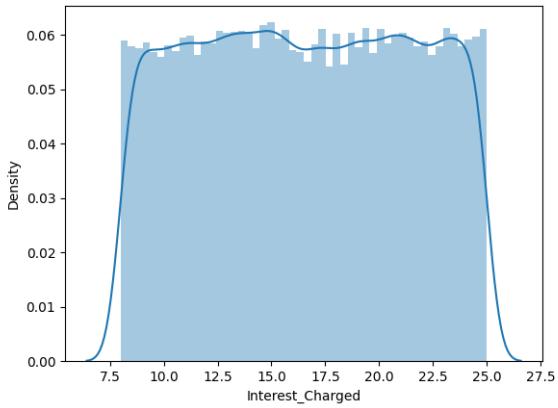

Figura 2.12: Distribuzione della variabile **Interest_Charged**.

La distribuzione della variabile **Yearly_Income** non è molto ben visualizzabile da questo grafico, perciò si applica una trasformazione logaritmica per avere una rappresentazione migliore della distribuzione del reddito annuo dei clienti. Infatti dal primo grafico non si coglie che esiste una variabilità molto grande all'interno di questo predittore, in quanto sono presenti clienti con un reddito medio annuo estremamente basso (con

un valore minimo di 8800) e altri con un reddito molto alto (con un valore massimo di 8264030.72). Il valore medio si aggira attorno a 134899.62.

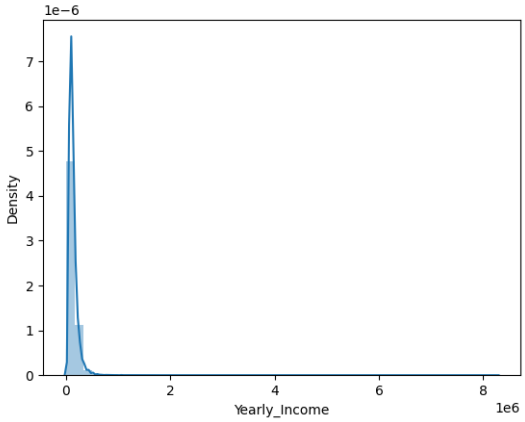

Figura 2.13: Distribuzione della variabile **Yearly_Income**.

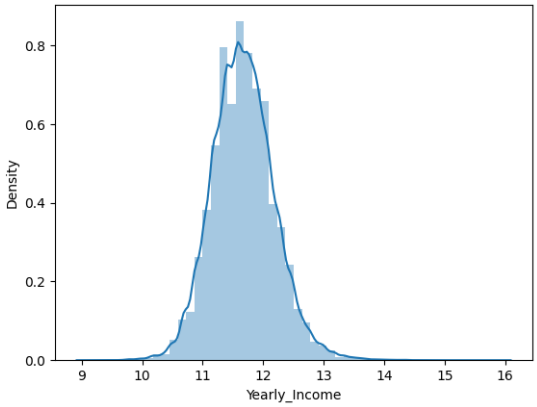

Figura 2.14: Distribuzione della variabile **Yearly_Income** dopo aver applicato una trasformazione logaritmica per una visualizzazione migliore.

L'ultima variabile continua che viene analizzata è **Debt_to_Income**, in cui si osserva una moda attorno al valore medio (che è sul 30.9%), ma si osserva una coda molto sottile e molto lunga con valori che toccano il 639%, perciò alcuni clienti hanno un reddito mensile che non è sufficiente a pagare le rate sul prestito che hanno contratto.

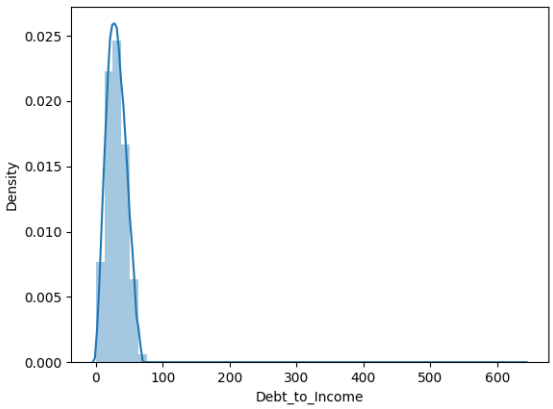

Figura 2.15: Distribuzione della variabile **Debt_to_Income**.

Infine si può andare a studiare la correlazione tra queste variabili, per vedere se presentano fenomeni di multicollinearità, perché la correlazione tra variabili in alcuni modelli può portare ad una stima affetta da varianza maggiore rispetto a quella che si potrebbe ottenere decorrelando le variabili. Non si notano situazioni particolarmente patologiche, si nota che le variabili **GGGrade** e **sub_GGGrade** sono molto correlate tra di loro, ma la cosa era intuibile visto che la seconda è solo una specifica della prima, e dalla seconda si può ottenere l'informazione contenuta nella prima variabile. Altre due variabili molto correlate tra di loro sono **Total_Unpaid_CL** e **Unpaid_Amount**, come era facilmente intuibile, in quanto un cliente che non è in grado di ripagare la rata su un prestito, probabilmente non sarà in grado di rimborsare i pagamenti effettuati con carta di credito.

2.2 Algoritmi di Classificazione

All'interno del dataset è specificato, pr ogni cliente, lo stato di solvenza del prestito, per cui è possibile costruire un modello supervisionato di classificazione che etichetti ogni cliente in base all'output del modello, se il cliente è sicuro o rischioso.
Gli algoritmi considerati per la costruzione del modello di classificazione sono la Regressione Logistica al primo ordine, che viene utilizzata come *benchmark*, la Regressione Logistica al secondo ordine, la Random Forest, l'Ada Boost e il K-Nearest Neighbors[2]. Gli algoritmi di Machine Learning vengono confrontanti con la Regressione Logistica perché riescono a cogliere meglio le relazioni tra le variabili, anche quelle più complesse, senza che queste vengano specificate in anticipo come parametri del modello. Le relazioni complesse tra le variabili si riflettono nei momenti della distribuzione Q_h, ovvero p e ρ. Il dataset è fortemente sbilanciato, poiché i clienti che appartengono alla classe positiva (quella che

[2]Si rimanda all'appendice C per una descrizione di questi algoritmi.

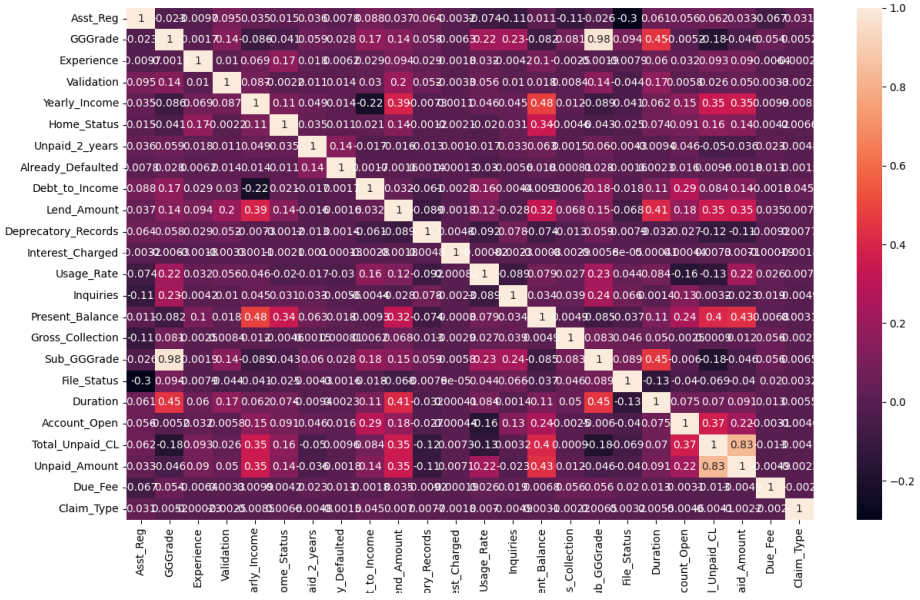

Figura 2.16: correlazione tra le variabili del dataset.

indica i clienti che fanno default, di maggior interesse per avere una buona classificazione) sono molto meno numerosi rispetto alla classe negativa, per cui è necessario considerare delle misure di perfomance che tengano in conto di questo sbilanciamento naturale presente all'interno del dataset. Le metriche che verranno utilizzate sono l'AUC, l'*f1-score*, il recupero e la precisione. Oltre a queste misure standard di perfomance di un algoritmo, si cerca anche di misurare il rischio associato alla scelta di un determinato algoritmo di machine learning nella determinazione dei parametri della distribuzione di probabilità del numero di default.

Modello	Precisione	Recupero	F1-score	AUC
LR	0.830	0.722	0.750	0.825
LR2	0.830	0.727	0.755	0.829
RF	0.837	0.837	0.791	0.826
AB	0.819	0.838	0.810	0.835
KNN	0.770	0.811	0.771	0.756

Tabella 2.1: Misure di perfomance dei cinque modelli considerati.

La tabella 2.1 mostra i risultati dei cinque algoritmi in termini di metriche. I modelli di regressione logistica al primo e al secondo ordine hanno valori alti di precisione e valori bassi di recupero, il che indica che questi modelli producono pochi falsi positivi ma molti falsi negativi, che può portare a delle decisioni rischiose perché non sono in

grado di classificare bene la classe dei clienti che va in default e potrebbe sottostimarne il numero. Il modello di KNN invece, al contrario, ha valori bassi di precisione e alti di recupero, perciò produce pochi falsi negativi e molti falsi positivi, perciò può portare a decisioni troppo conservative perché sovrastima il rischio di default dei clienti. Il modello di Random Forest e di Ada Boost invece hanno valori alti in entrambe le metriche, e perciò sembrano quelli più affidabili per avere una stima accurata del rischio associato ai clienti considerati.

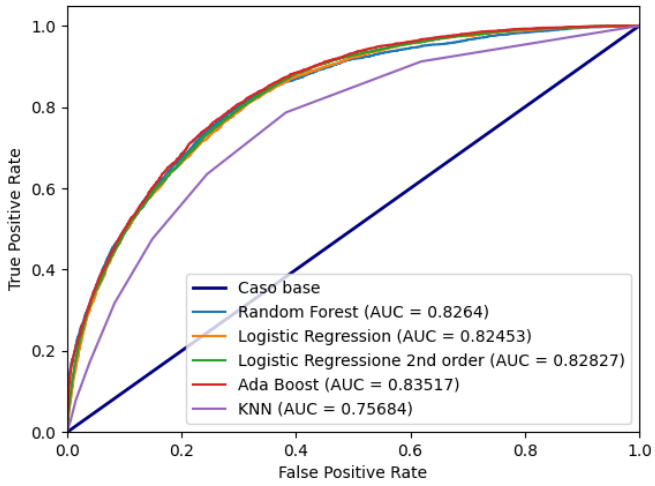

Figura 2.17: Curva ROC per ognuno dei cinque modelli considerati, confrontati con la bisettrice, che indica il caso di un classificatore randomico, che indovina, in media, solo metà delle volte.

Quando si hanno troppi predittori all'interno di un modello, questo diventa più complesso e rischia di andare in *overfitting*. Inoltre la correlazione tra variabili può aumentare la varianza, perciò, se è possibile, è necessario semplificare il modello, andando ad eliminare le variabili meno importanti per la spiegazione della variabile risposta. In tabella sono indicati i predittori che descrivono meglio il modello e quindi sono necessari e non possono essere eliminati.

Feature	Score
Asst_Reg	2304.64
File_Status	791.37
Duration	504.54
Validation	126.04
Inquiries	70.23
Lend_Amount	36.31
Gross_Collection	23.86
Claim_Type	8.15
Unpaid_2_years	7.52
Due_Fee	5.80
Debt_to_Income	5.58
GGGrade	4.42
Deprecatory_Records	3.36
Unpaid_Amount	3.31
Sub_GGGrade	3.29
Account_Open	2.23
Total_Unpaid_CL	1.64

Tabella 2.2: I predittori più importanti per spiegare la variabile risposta **default**.

Con le covariate contenute nella tabella 2.2 si costruiscono cinque modelli più semplici e si osserva se, in termini di metriche, si hanno perdite importanti, o se anche i modelli più piccoli sono in grado di descrivere bene la variabile risposta.

Modello	Precisione	Recupero	F1-score	AUC
LR	0.829	0.720	0.749	0.822
LR2	0.829	0.726	0.753	0.826
RF	0.828	0.838	0.798	0.829
AB	0.819	0.838	0.810	0.832
KNN	0.797	0.825	0.794	0.767

Tabella 2.3: Misure di prestazione dei cinque modelli più piccoli.

Dalla tabella 2.3 si può notare come i valori siano molto simili a quelli relativi ai modelli più grandi che considerano tutte le covariate, perciò anche i modelli più piccoli descrivono bene la variabile risposta e si possono quindi considerare solo questi predittori, per avere un modello più semplice.

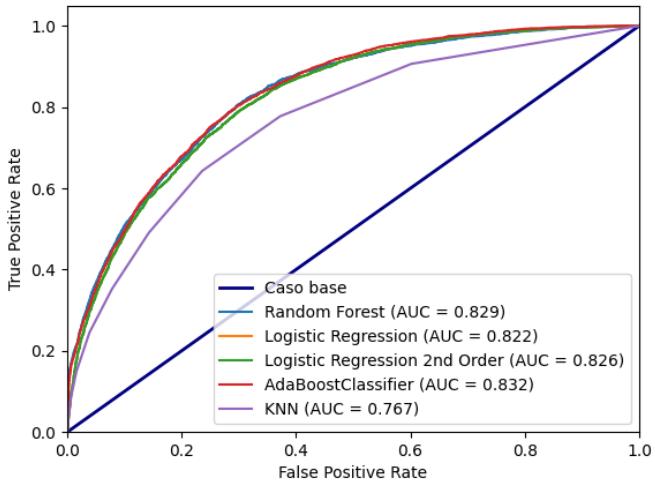

Figura 2.18: Curve ROC per i cinque modelli più piccoli.

2.2.1 Default di un portafoglio cartolarizzato.

La variabile aleatoria S rappresenta la somma dei default sui prestiti inseriti all'interno di un portafoglio cartolarizzato. La formula (1.15) mostra come calcolare la distribuzione di questa variabile aleatoria. Poiché la formula presenta dei fattoriali, è possibile utilizzare questa formula solo per un numero ridotto di prestiti, per cui si confronta il rischio stimato da questa formula con il rischio stimato da una distribuzione parametrica beta binomiale per un portafoglio composto da $d = 25$ clienti. Per l'analisi del rischio di un portafoglio cartolarizzato più grande è necessario utilizzare una distribuzione parametrica beta binomiale.

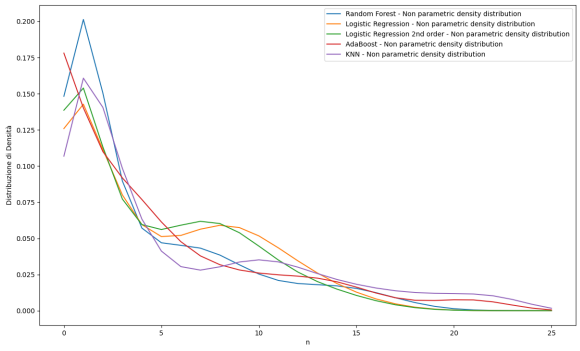

Figura 2.19: Distribuzione di densità del numero di default.

Momento	LR	LR2	RF	AB	KNN
p	0.1889	0.18864	0.1891	0.1905	0.1916
π_2	0.06663	0.06816	0.0892	0.0755	0.0831
ρ	0.2018	0.2128	0.3485	0.2545	0.2999

Tabella 2.4: Momenti empirici calcolati a partire da ognuno dei cinque modelli.

La tabella 2.4 fornisce i primi due momenti empirici della distribuzione di Q_h per ogni modello h considerato. L'algoritmo di regressione logistica, sia al primo che al secondo ordine, sottostima la probabilità marginale di default e il momento di ordine secondo. I tre modelli di Machine Learning, invece, poiché sono in grado di cogliere meglio le relazioni tra le variabili, hanno una correlazione più alta rispetto ai due modelli di regressione logistica. Questo parametro influenza molto il VaR della distribuzione S del numero di default. Per la costruzione di un portafoglio più realistico, che contiene centinaia o migliaia di prestiti differenti, è necessario specificare una distribuzione di probabilità per il numero di default all'interno di questo portafoglio. Bisogna quindi, inanzi tutto, ottenere i parametri della distribuzione beta della probabilità di default. Questi parametri si possono ottenere attraverso il metodo dei momenti, per ognuno dei cinque modelli considerati.

$\beta - parametri$	LR	LR2	RF	AdaBoost	KNN
a	0.7473254	0.6976505	0.3546806	0.5580564	0.4473574
b	3.207594	3.000621	1.516184	2.370058	1.886744
p	0.188961	0.1886423	0.1895811	0.1905856	0.1916615
ρ	0.2018196	0.2128442	0.3483271	0.2545751	0.2999309

Tabella 2.5: Parametri della distribuzione β per ogni modello.

La tabella 2.5 mostra i parametri della distribuzione Beta per ognuno dei cinque modelli considerati. Il test di Kolmgorov-Smirnov per le cinque distribuzioni β rifiuta l'ipotesi che spieghino la variabile aleatoria Q_h, però è possibile vedere i momenti empirici calcolati in modo non parametrico sono molto simili a quelli teorici calcolati tramite la distribuzione di probabilità, perciò ci si aspetta che, poiché il VaR dipende principalmente dai primi due momenti, il VaR del modello parametrico e di quello non parametrico siano simili.

	LR	LR2	RF	AB	KNN
distanza KL	0.04416	0.03333	0.07156	0.01433	0.07051

Tabella 2.6: distanza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica

La tabella 2.6 mostra la distanza di Kullback-Leibler tra la distribuzione non parametrica di Q_h, per ognuno dei modelli considerati, con la distribuzione beta stimata tramite il metodo dei momenti. Si può notare che i modelli non parametrici più simili a quelli parametrici sono il modello di Regressione Logistica al secondo ordine e quello di Ada Boost.

Stime VaR

α	LR	LR2	RF	AB	KNN
0.90	12	10	12	13	16
0.95	13	11	15	17	19
0.99	16	13	18	22	23

Tabella 2.7: VaR non parametrico dei cinque modelli.

La tabella 2.7 fornisce il VaR empirico per tre livelli di rischio ($\alpha = 0.90, 0.95, 0.99$) per la distribuzione di probabilità del portafoglio cartolarizzato che contiene 25 prestiti omogenei. Come si può vedere, i due algoritmi di regressione logistica sottostimano il rischio associato al portafoglio cartolarizzato, perché stimano una correlazione troppo bassa tra i clienti. I tre modelli di Machine Learning, invece, sono più realistici sul rischio di questo portafoglio.

α	LR	LR2	RF	AB	KNN
0.90	12	11	12	14	16
0.95	14	14	15	17	19
0.99	19	18	20	22	23

Tabella 2.8: VaR della distribuzione beta-binomiale S.

La tabella 2.8 mostra il VaR della distribuzione beta binomiale. Si può notare che nel caso della Regressione Logistica, sia al primo che al secondo ordine, il VaR non parametrico è più basso rispetto a quello parametrico, il che vuol dire che la regressione logistica non riesce a cogliere le code più grasse di una distribuzione. Nel caso dei tre modelli di Machine Learning invece, il VaR non parametrico e quello parametrico sono molto simili. La differenza più evidente si trova nel caso del quantile $\alpha = 0.99$ della Random Forest, in cui la distribuzione beta-binomiale stima un VaR più alto rispetto al modello non parametrico. Dall'analisi di queste tabelle si nota che, sia nel caso parametrico che in quello non parametrico, i due modelli di Regressione Logistica stimano dei valori di VaR più bassi rispetto ai tre modelli di Machine Learning, e quindi sottostimano il rischio della perdita. Le analisi svolte tramite tecniche di Machine Learning, quindi, sono necessarie per comprendere i rischi associati a dipendenze non lineari tra le variabili.

Si considera adesso la distribuzione beta-binomiale per un portafoglio più grande, composto da $d = 10000$ prestiti, utilizzando i parametri della 2.5.

α	LR	LR2	RF	AB	KNN
0.90	4654	4739	5609	5104	5285
0.95	5794	5916	7068	6409	6608
0.99	9088	9027	8927	9219	8489

Tabella 2.9: VaR della distribuzione beta-binomiale S per un portafoglio di 10000 clienti.

La tabella 2.9 mostra i valori di VaR per le distribuzioni beta-binomiali dei cinque modelli considerati. Al quantile $\alpha = 0.99$ i cinque modelli sono molto simili tra di loro, mostrando che le distribuzioni beta binomiali, sulla coda, hanno andamenti analoghi, mentre per i quantili $\alpha = 0.90, 0.95$, la Regressione Logistica al primo e al secondo ordine sottostimano il rischio della perdita. Nelle figure sottostanti è possibile guardare la funzione di densità discreta per ognuno dei cinque modelli considerati.

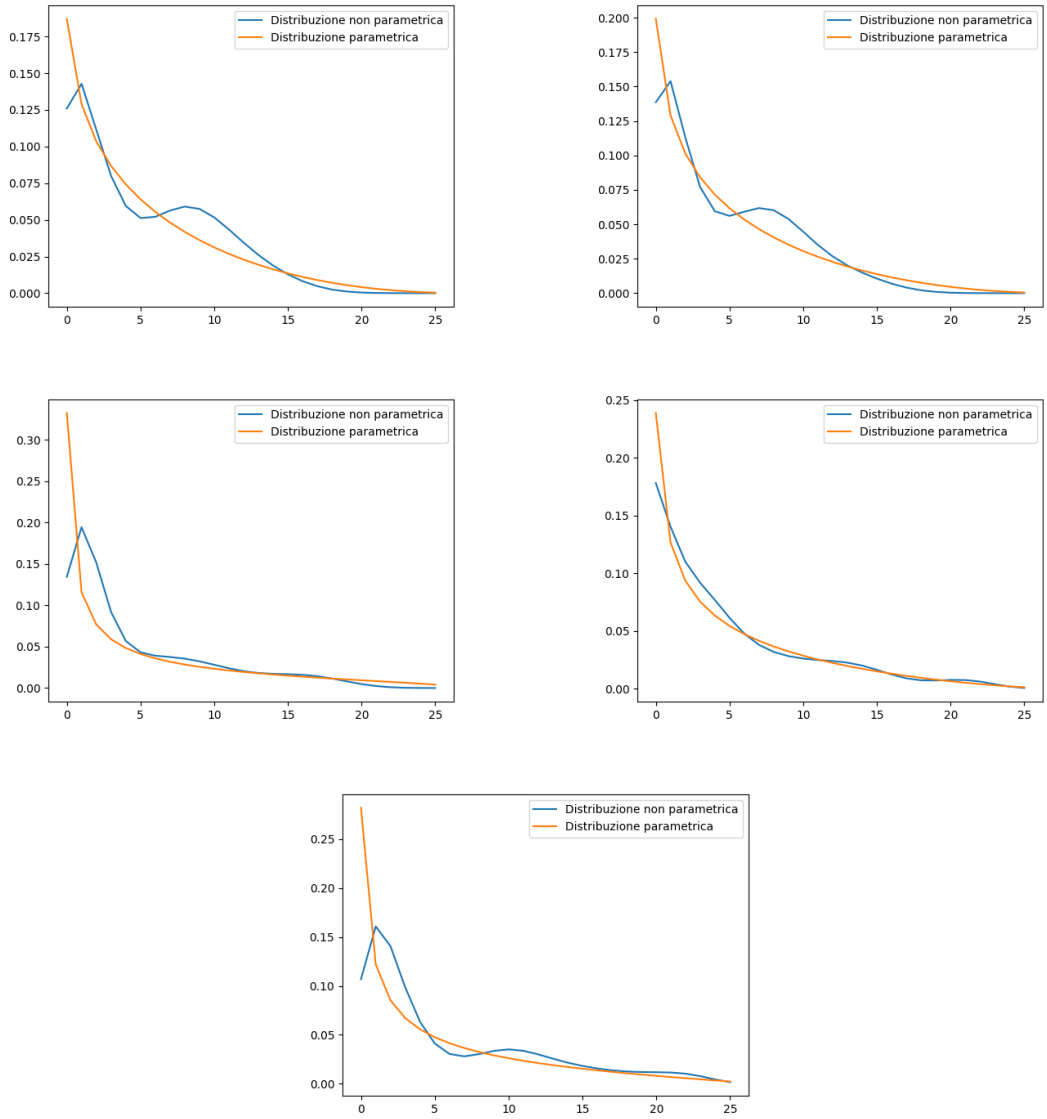

Figura 2.20: Distribuzione di densità della variabile S per i cinque modelli considerati (LR, LR2, RF, AB, KNN).

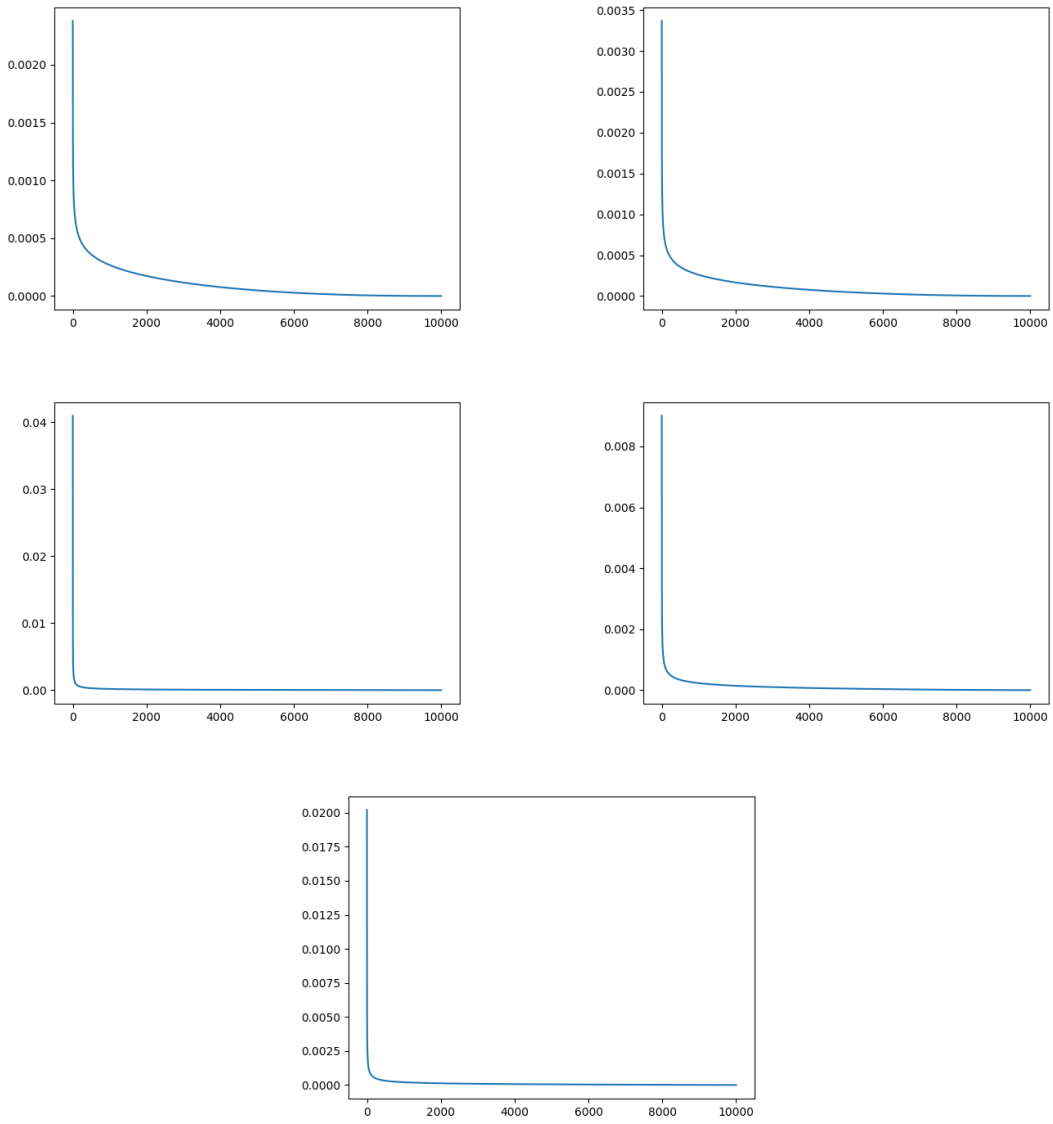

Figura 2.21: Distribuzione di densità della variabile S per i cinque modelli considerati (LR, LR2, RF, AB, KNN).

Capitolo 3

Costruzione e Analisi dei portafogli cartolarizzati.

Il dataset è molto grande e composto da clienti e prestiti molto eterogenei tra di loro, perciò si prosegue l'analisi del dataset suddividendolo in cluster più piccoli e omogenei, così che i prestiti contenuti all'interno di ogni cluster siano più simili e si possano costruire dei portafogli con pesi uniformi senza perdere di generalità.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	*Noise Points*
16442	9807	5715	17529	6	5	37

Tabella 3.1: Cluster identificati dall'algoritmo DBSCAN.

I cluster vengono generati mediante l'algoritmo di *Unsupervised Learning*, DBSCAN e nella tabella 3.1 sono riportati i risultati dell'algoritmo: sono stati identificati 4 cluster grandi e 2 cluster molto piccoli, che possono essere considerati dei punti rumorosi, perciò le analisi successive si focalizzeranno sui cluster più grandi.
Adesso si osserva nel dettaglio come sono distribuite le covariate più importanti all'interno dei quattro cluster identificati, per capire se ci sono dei *pattern* particolari all'interno dei dati e per comprendere meglio i dati presenti nei singoli cluster.
La prima covariata che viene analizzata è **Asst_Reg**, che indica l'entità del patrimonio posseduta dal cliente considerato. Dal grafico si può notare come i cluster 0 e 1 siano caratterizzati da un patrimonio mediamente più alto, mentre i cluster 2 e 3 siano caratterizzati da un patrimonio un po' più basso. Si potrebbe pensare quindi che i primi due cluster siano formati da clienti più abbienti, e che quindi all'emissione del prestito abbiano potuto dare più garanzie e siano quindi clienti più sicuri.

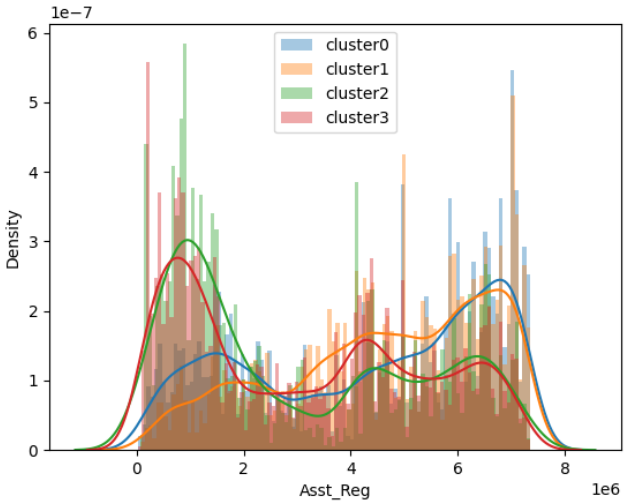

Figura 3.1: Distribuzione della variabile **Asst_Reg** per i quattro cluster identificati.

La seconda variabile considerata è **Yearly_Income**, da cui non si notano particolari differenze nei quattro cluster, perciò si può concludere che, in media, i clienti presenti all'interno dei quattro cluster percepiscano un reddito annuo simile.

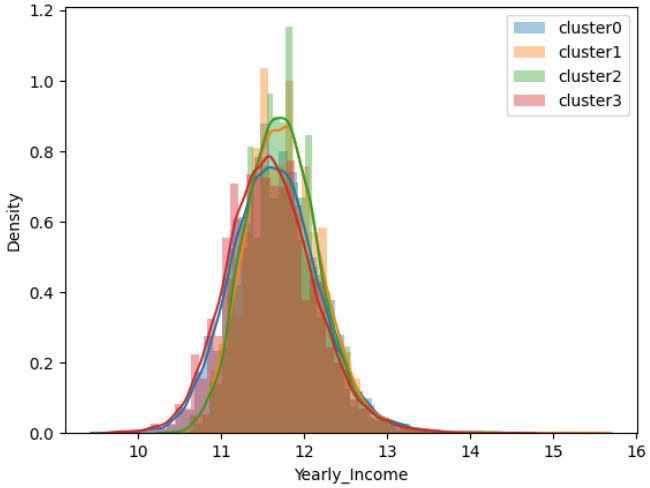

Figura 3.2: Distribuzione della variabile **Yearly_Income** per i quattro cluster identificati.

La terza variabile analizzata è **Lend__Amount**, ovvero la quantità di denaro concessa al cliente per il prestito richiesto. Si può notare un pattern molto evidente: i clienti appartanenti ai cluster 0 e 3 hanno ricevuto prestiti minori, mentre i clienti dei cluster 1 e 2 hanno ricevuto prestiti mediamente più alti.

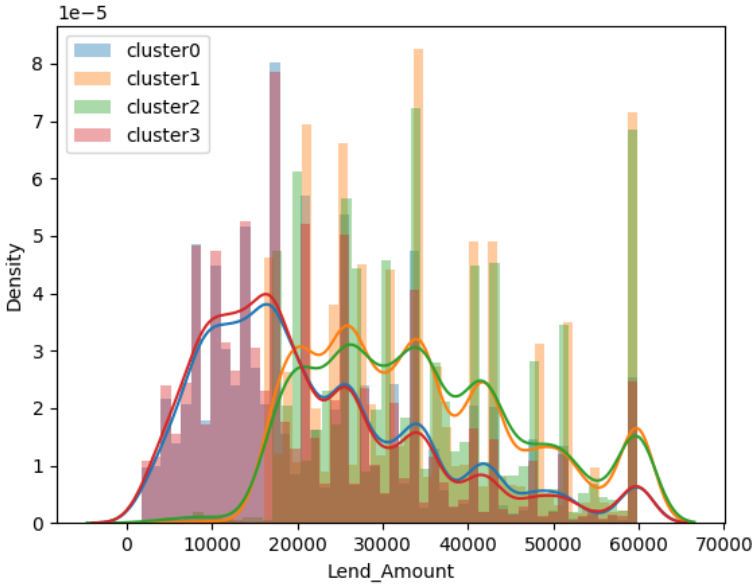

Figura 3.3: Distribuzione della variabile **Lend__Amount** per i quattro cluster identificati.

Anche l'analisi della variabile **Debt__to__Income** identifica una certa similitudine tra i cluster 0 e 3, e i cluster 1 e 2. In particolare, la moda per i cluster 0 e 3 cade prima di quella dei cluster 1 e 2, un fenomeno in linea con il fatto che i cluster 0 e 3 sono caratterizzati da prestiti più bassi rispetto agli altri due cluster.

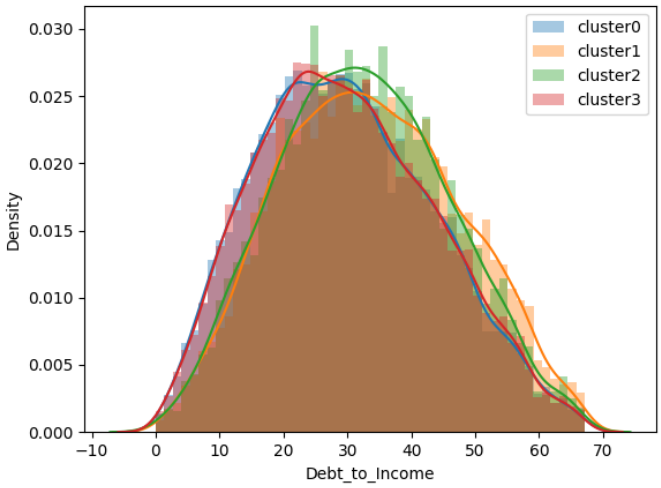

Figura 3.4: Distribuzione della variabile **Debt_to_Income**, dopo una trasformazione logaritmica, per i quattro cluster identificati.

La variabile **Interest_Charged** è distribuita in maniera uniforme per tutti e quattro i cluster, e non sembra che vi siano differenze sostanziali tra i vari cluster.

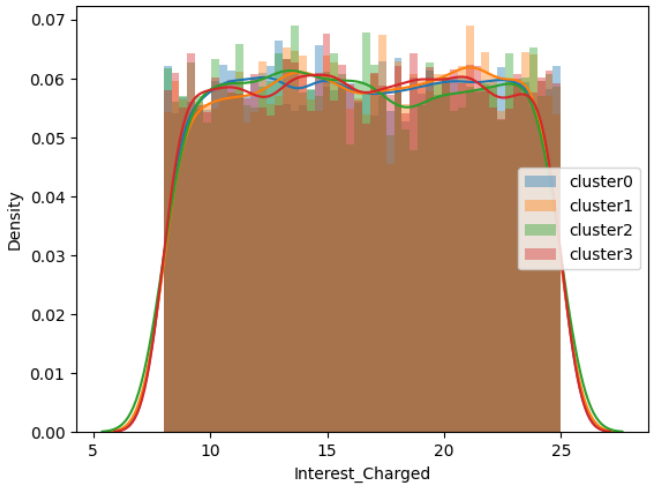

Figura 3.5: Distribuzione della variabile **Interest_Charged** per i quattro cluster identificati.

La variabile **Duration** indica la durata del prestito. Questa variabile è molto importante per la divisione in cluster, infatti ogni cluster è caratterizzato da un solo valore: 3

(36 mesi) 0 5 (60 mesi). In particolare i cluster 0 e 3 sono formati da prestiti della durata di 3 anni, e i cluster 1 e 2 sono formati da quelli che hanno una durata di cinque anni.

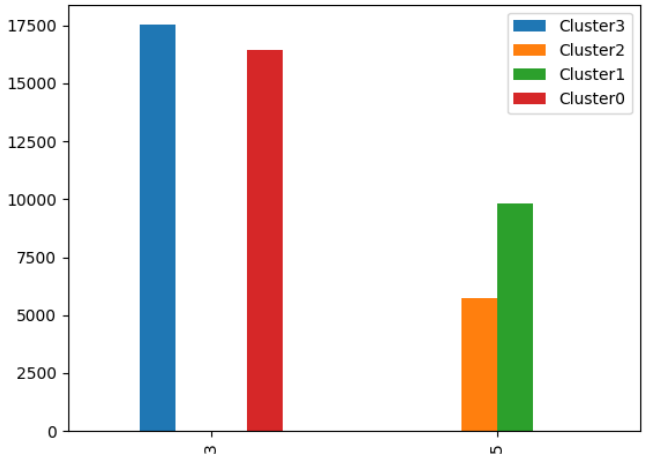

Figura 3.6: Distribuzione della variabile **Duration** per i quattro cluster identificati

La variabile **Experience** indica gli anni totali di impiego del cliente. Non sembrano esserci pattern particolari, ed è distribuita in maniera molto omogenea per tutti e quattro i cluster.

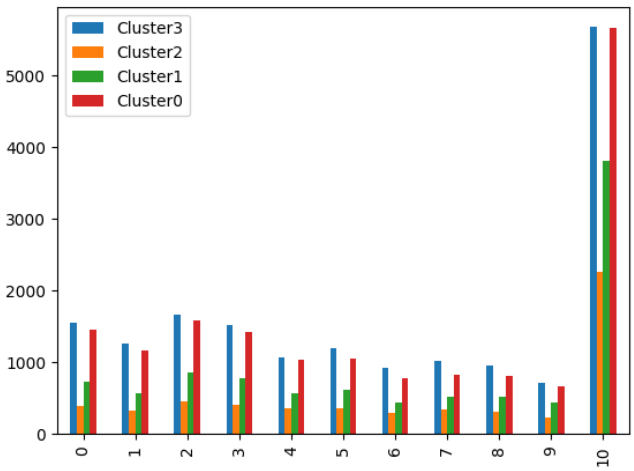

Figura 3.7: Distribuzione della variabile **Experience** per i quattro cluster identificati

La variabile **GGGrade** indica il rating assegnato al prestito: può assumere valori da 1 (migliore) a 7 (peggiore). Si può notare come i cluster 0 e 3 siano caratterizzati da rating mediamente più alti rispetto ai cluster 1 e 2.

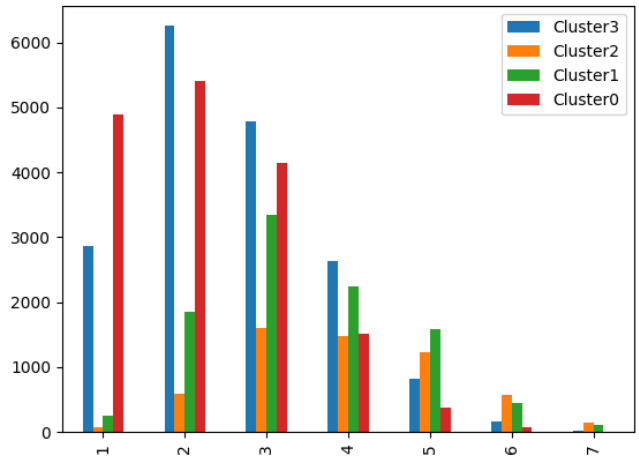

Figura 3.8: Distribuzione della variabile **GGGrade** per i quattro cluster identificati

Poiché una cartolarizzazione utilizza dei prestiti la cui probabilità di default è più bassa rispetto a quella che si osserva all'interno del dataset, si costruiscono quattro portafogli (uno per ogni cluster), prendendo i 1000 clienti che hanno la stima di default più bassa per l'algoritmo di Regressione Logistica, che viene appunto usato come modello di base con cui effettuare i confronti. Anche in questo caso si confrontano i risultati che si ottengono nella stima della variabile aleatoria S in maniera parametrica ed in maniera empirica, estraendo dai portafogli contenenti 1000 clienti un portofoglio più piccolo composto da $d = 50$ clienti (il numero aumenta rispetto al capitolo precedente perché avendo una probabilità di default più bassa, è necessario aumentare il numero di prestiti presenti all'interno del portafoglio per ottenere dei risultati significativi).

3.0.1 Portafoglio 0

Algoritmo	Precisione	Recupero	F1-score	AUC
RF	0.828	0.845	0.794	0.807
LR	0.8176	0.8455	0.8076	0.8138
LR2	0.8038	0.8392	0.79948	0.8129
AB	0.8175	0.8443	0.8175	0.8162
KNN	0.7918	0.8343	0.7910	0.7642

Tabella 3.2: Misure di perfomance dei cinque modelli per il cluster 0.

Prima di analizzare il portafoglio cartolarizzato formato da alcuni prestiti presenti all'interno di questo cluster, vengono riportate le perfomance degli algoritmi considerati fin'ora nel classificare i dati tra cui sono stati estratti i prestiti presenti all'interno del portafoglio e le loro probabilità di default. La tabella 3.2 mostra i risultati relativi alle metriche considerate, e si può notare che i valori sono molto alti, più alti rispetto a quelli che si ottengono quando si analizza il dataset globale.

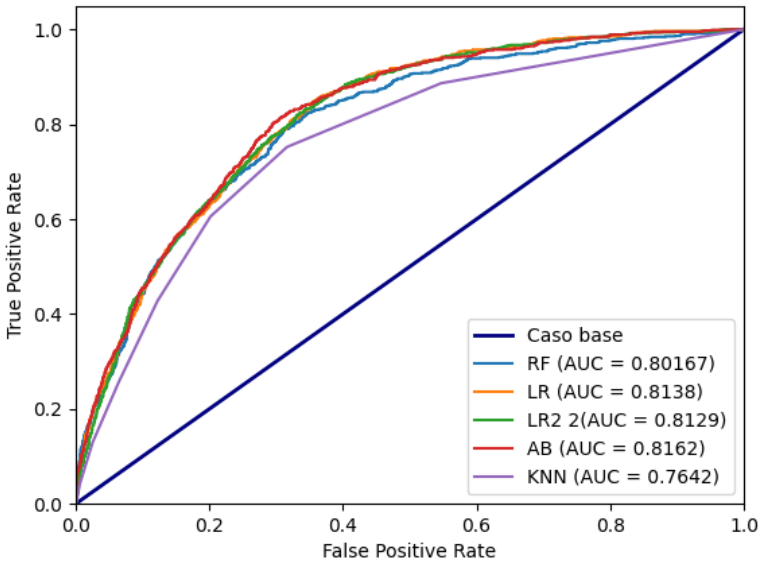

Figura 3.9: Curva ROC per i cinque modelli considerati, per il cluster 0.

Adesso è possibile costruire il primo portafoglio cartolarizzando 1000 prestiti contenuti all'interno di questo cluster.

Momento	LR	LR2	RF	AB	KNN
p	0.024810	0.02239	0.010406	0.013019	0.03583
π_2	0.000644	0.000548	0.0001096	0.000222	0.0017789
ρ	0.001191	0.002147	0.000133	0.004088	0.0009066

Tabella 3.3: Momenti stimati tramite i cinque modelli di Machine Learning.

La tabella 3.3 riassume i momenti empirici della distribuzione di probabilità del default individuale per il portafoglio ottenuto a partire dal cluster 0. La probabilità di default più alta viene stimata dall'algoritmo KNN, e si aggira attorno al 3.5%, mentre la Random Forest e l'Ada Boost stimano la probabilità di default più bassa, circa dell'1%. La Regressione Logistica al primo e al secondo ordine hanno comportamenti intermedi. La correlazione tra i default è stimata bassissima per tutti e cinque i modelli.

Per la costruzione della distribuzione di probabilità S della somma dei default, è necessario calcolare i parametri a e b della distribuzione Q_h per ogni h = LR, LR2, RF, AB, KNN, per questo portafoglio.

β-parametri	LR	LR2	RF	AB	KNN
a	20.794578	10.40842	78.114955	3.171095	38.13304
b	817.34249	454.2929	7428.1290	240.38799	1025.9328
p	0.0248104	0.022398	0.010406	0.0130198	0.0358371
ρ	0.0011917	0.0021472	0.004088	0.2546706	0.0009389

Tabella 3.4: Parametri della distribuzione Q_h stimati per ognuno dei cinque modelli.

Nella tabella 3.4 sono riportati i parametri della distribuzione Q_h stimati per ognuno dei modelli considerati. Il momento primo $E(Q_h) = p$ e la correlazione ρ stimati in maniera parametrica sono molto simili a quelli ottenuti in maniera empirica dai dati. Si è dimostrato che, sopratutto sulle code, non è importante tanto la forma funzionale quanto che i primi momenti siano stimati bene, perciò è possibile proseguire con le analisi di questo portafoglio.

	LR	LR2	RF	AB	KNN
distanza KL	0.09913	0.001843	$3.67*10^{(-5)}$	0.0003560	0.000520

Tabella 3.5: Distanza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica.

La tabella 3.5, mostra la divergenza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica per il portafoglio piccolo composto da $d = 50$ clienti. Si può vedere che la distribuzione parametrica approssima in maniera molto fedele quella empirica.

α	LR	LR2	RF	AB	KNN
0.90	4	3	1	2	4
0.95	4	3	2	2	4
0.99	5	4	3	3	5

Tabella 3.6: VaR della distribuzione empirica S relativa al primo portafoglio.

α	LR	LR2	RF	AB	KNN
0.90	3	3	1	2	4
0.95	3	3	2	2	4
0.99	5	4	3	3	5

Tabella 3.7: VaR della distribuzione beta-binomiale S relativo al primo portafoglio.

Le tabelle 3.6 e 3.7 mostrano il VaR calcolato per la distribuzione parametrica e quella empirica per il portafoglio composto da $d = 50$ clienti. Avendo ottenuto che per questo portafoglio sia i primi momenti, che il VaR che la distanza di Kullback-Leibler concordano nel dire che la distribuzione parametrica approssima in maniera sufficientemente buona i dati, è possibile proseguire le analisi su un portafoglio più grande formato da $d = 1000$ prestiti.

α	LR	LR2	RF	AB	KNN
0.90	35	34	16	25	48
0.95	39	38	17	29	51
0.99	45	46	20	39	58

Tabella 3.8: VaR della distribuzione beta-binomiale S per il portafoglio formato da 1000 prestiti.

La tabella 3.8 riporta i risultati relativi al VaR calcolato per i tre livelli standard considerati $\alpha = 0.9, 0.95, 0.99$. I due algoritmi di Regressione Logistica hanno delle stime molto simili che riflette la stima dei primi due momenti, che per questi due modelli erano, appunto, molto simili tra di loro. L'algoritmo di Random Forest ha i valori di VaR più bassi mentre l'algoritmo KNN quelli più alti.
Nelle figure sottostanti è possibile osservare le funzioni di densità relative alla distribuzione di probabilità S sia per il portafoglio piccolo (stimato sia in maniera empirica che in maniera parametrica) che per il portafoglio grande.

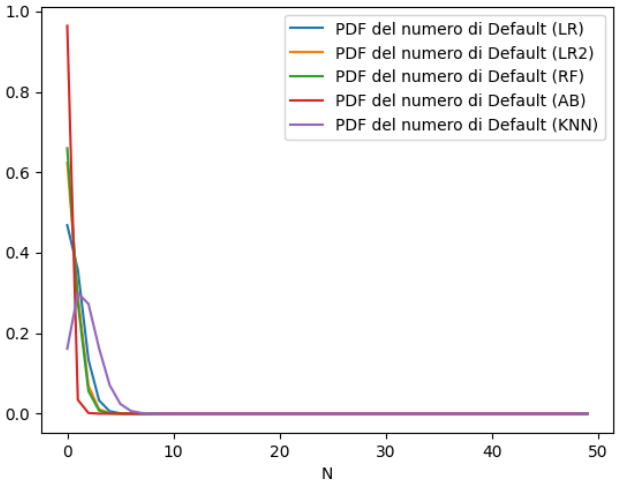

Figura 3.10: Distribuzione non parametrica della variabile S relativa al portafoglio piccolo, stimata da ognuno dei cinque modelli.

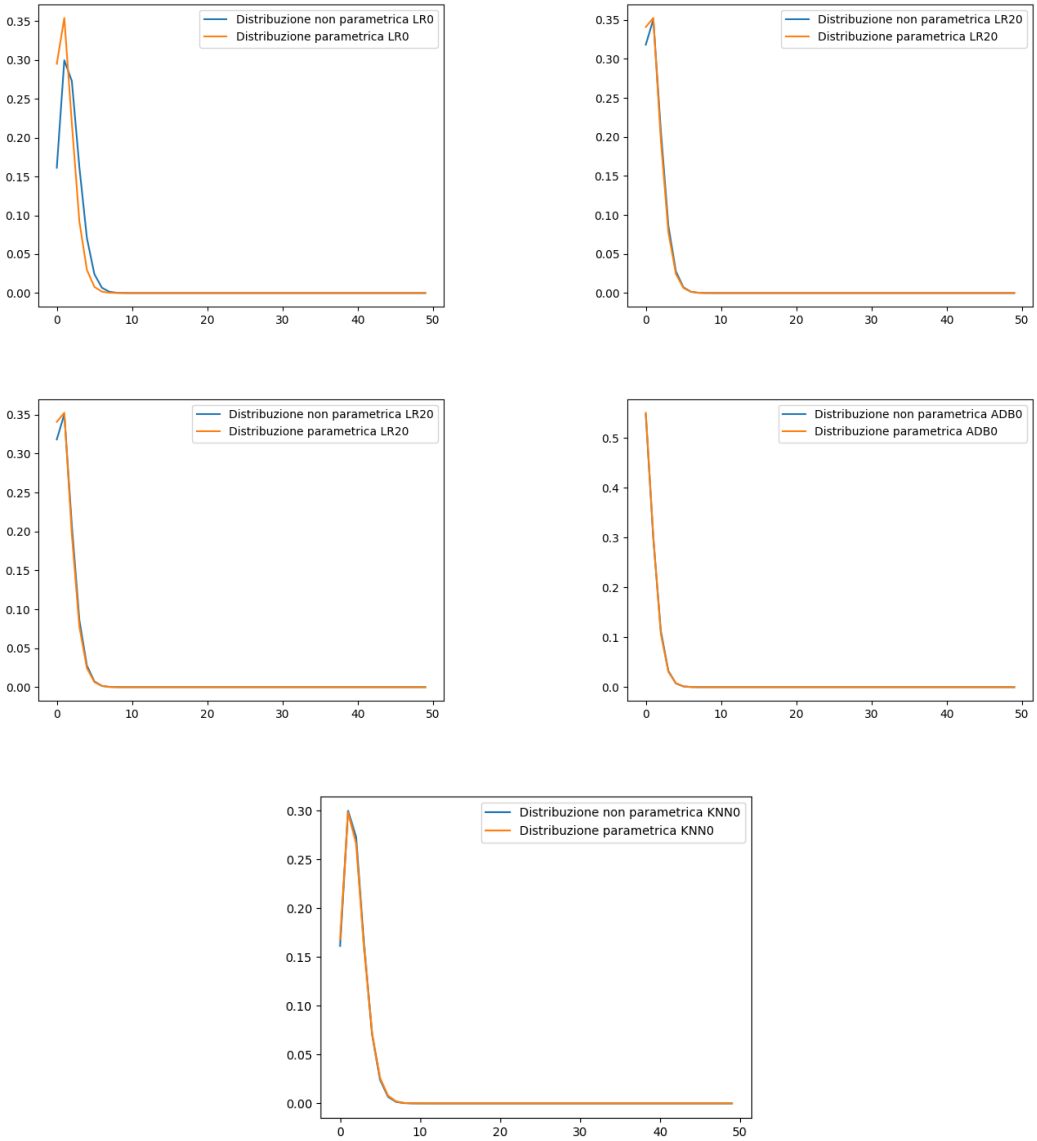

Figura 3.11: Confronto tra la distribuzione parametrica e quella non parametrica per la variabile S per il portafoglio piccolo, stimate da ognuno dei modelli considerati (LR, LR2, RF, AB, KNN).

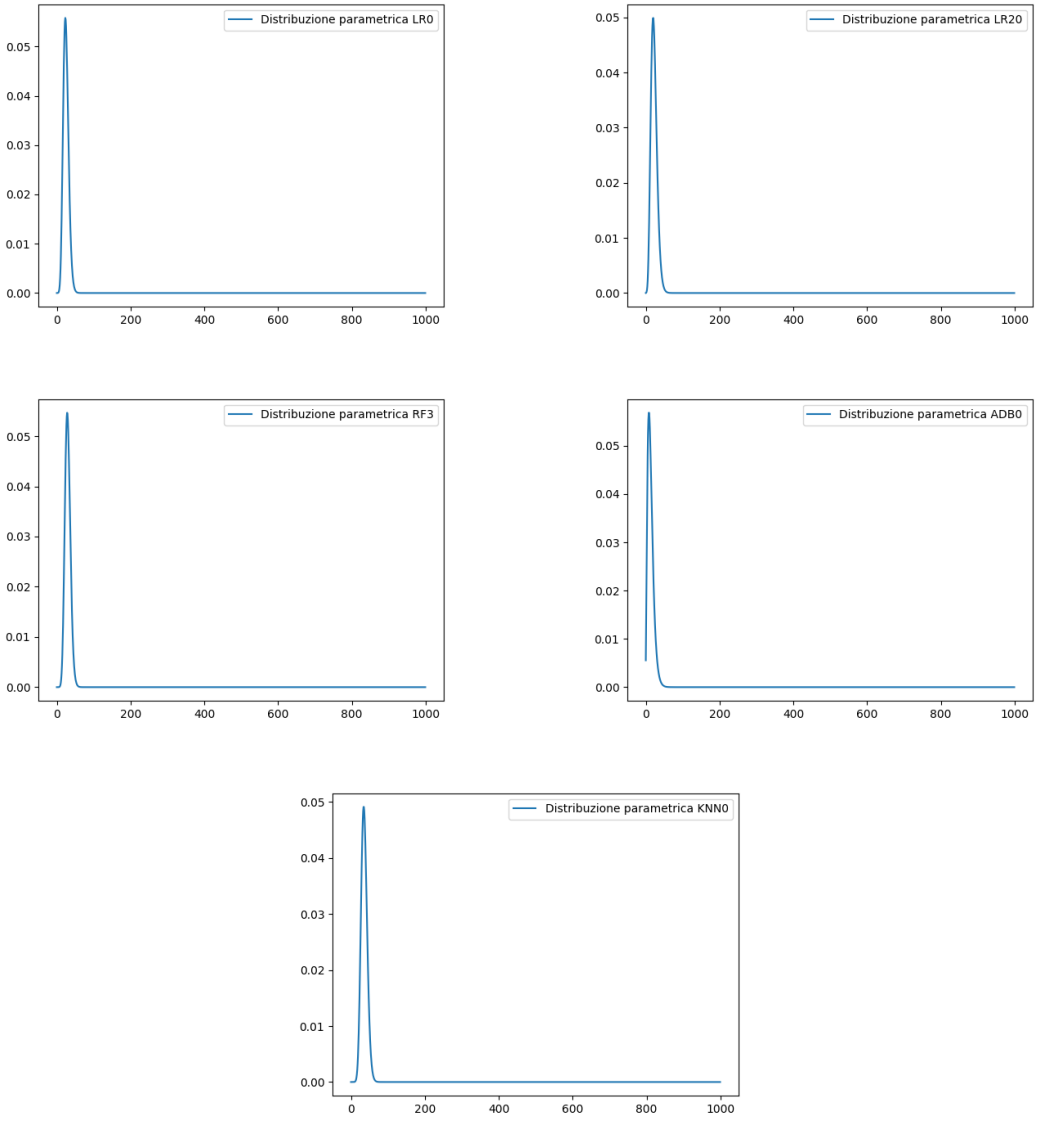

Figura 3.12: Distribuzione variabile S per il portafoglio grande, stimata dai cinque modelli considerati (LR, LR2, RF, AB, KNN).

3.0.2 Portafoglio 1

Algoritmo	Precisione	Recupero	F1-score	AUC
RF	0.8055	0.8972	0.8488	0.7363
LR	0.9081	0.8977	0.8496	0.7598
LR2	0.9081	0.8977	0.8496	0.7682
AB	0.8428	0.8943	0.8533	0.7542
KNN	0.8155	0.8892	0.84656	0.6053

Tabella 3.9: Misure di perfomance dei cinque modelli di Machine Learning per il cluster 1.

Prima di analizzare il portafoglio cartolarizzato formato da alcuni prestiti presenti all'interno di questo cluster, vengono riportate le perfomance degli algoritmi considerati fin'ora nel classificare i dati tra cui sono stati estratti i prestiti presenti all'interno del portafoglio e le loro probabilità di default. La tabella 3.9 mostra i risultati relativi alle metriche considerate, e si può notare che i valori sono molto alti, più alti rispetto a quelli che si ottengono quando si analizza il dataset globale.

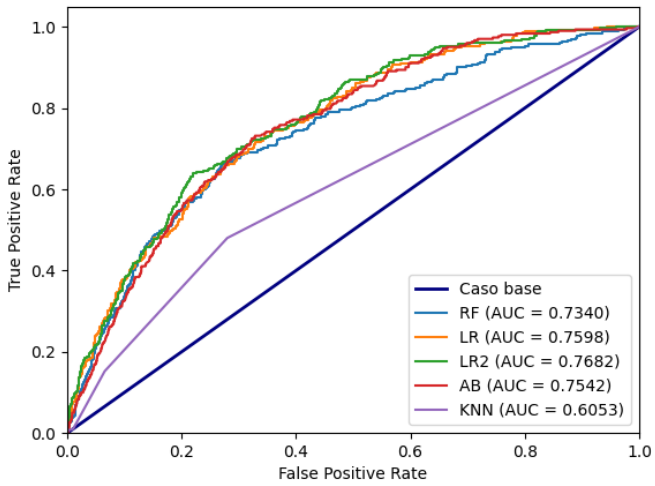

Figura 3.13: Curva ROC dei cinque modelli considerati, per il cluster 1.

Momento	LR	LR2	RF	AB	KNN
p	0.035350	0.032426	0.00313	0.017630	0.018751
π_2	0.001374	0.00120	$1.11*10^{(-5)}$	0.000500	0.0110473
ρ	0.003662	0.004951	0.0004380	0.010938	0.0017691

Tabella 3.10: Momenti stimati tramite i cinque modelli di Machine Learning.

La tabella 3.10 riassume i momenti empirici della distribuzione di probabilità del default individuale per il portafoglio ottenuto a partire dal cluster 1. La probabilità di default più alta viene stimata dai due algoritmi di Regressione Logistica al primo e al secondo ordine, e si aggira attorno al 3%, mentre la Random Forest stima la probabilità di default più bassa, circa dell'0.3%. L'Ada Boost e il KNN stimano tassi di default intermedi.

Per la costruzione della distribuzione di probabilità S della somma dei default, è necessario calcolare i parametri a e b della distribuzione Q_h per ogni $h =$ LR, LR2, RF, AB, KNN, per questo portafoglio.

$\beta - parametri$	LR	LR2	RF	AB	KNN
a	9.617672	0.9675108	7.147549	1.594244	10.61600
b	262.44757	9.387613	2274.9045	88.82948	555.5365
p	0.035350	0.032426	0.0031320	0.0176308	0.0187511
ρ	0.003662	0.08806597	0.0004380	0.0109380	0.0017631

Tabella 3.11: Parametri della distribuzione Q_h stimati per ognuno dei cinque modelli.

Nella tabella 3.11 sono riportati i parametri della distribuzione Q_h stimati per ognuno dei modelli considerati. Il momento primo $E(Q_h) = p$ e la correlazione ρ stimati in maniera parametrica sono molto simili a quelli ottenuti in maniera empirica dai dati. Si è dimostrato che, sopratutto sulle code, non è importante tanto la forma funzionale quanto che i primi momenti siano stimati bene, perciò è possibile proseguire con le analisi di questo portafoglio.

	LR	LR2	RF	AB	KNN
distanza KL	0.00069	0.00114	0.0007634	0.004871	0.0017534

Tabella 3.12: Distanza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica.

La tabella 3.5, mostra la divergenza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica per il portafoglio piccolo composto da $d = 50$ clienti. Si può vedere che la distribuzione parametrica approssima in maniera molto fedele quella empirica.

α	LR	LR2	RF	AB	KNN
0.90	4	4	1	2	2
0.95	4	4	1	3	3
0.99	6	6	2	4	4

Tabella 3.13: VaR della distribuzione empirica S relativa al secondo portafoglio.

α	LR	LR2	RF	AB	KNN
0.90	4	3	1	2	2
0.95	4	4	1	3	3
0.99	6	6	2	5	4

Tabella 3.14: VaR della distribuzione beta-binomiale S relativo al secondo portafoglio.

Le tabelle 3.13 e 3.14 mostrano il VaR calcolato per la distribuzione parametrica e quella empirica per il portafoglio composto da $d = 50$ clienti. Avendo ottenuto che per questo portafoglio sia i primi momenti, che il VaR che la distanza di Kullback-Leibler concordano nel dire che la distribuzione parametrica approssima in maniera sufficientemente buona i dati, è possibile proseguire le analisi su un portafoglio più grande formato da $d = 1000$ prestiti.

α	LR	LR2	RF	AB	KNN
0.90	53	52	7	38	29
0.95	59	58	8	47	33
0.99	71	72	10	67	39

Tabella 3.15: VaR della distribuzione beta-binomiale S per il portafoglio formato da 1000 prestiti.

La tabella 3.15 riporta i risultati relativi al VaR calcolato per i tre livelli standard considerati $\alpha = 0.9, 0.95, 0.99$. I due algoritmi di Regressione Logistica hanno delle stime molto simili che riflette la stima dei primi due momenti, che per questi due modelli erano, appunto, molto simili tra di loro. L'algoritmo di Random Forest ha i valori di VaR più bassi, poiché stimava una probabilità di default molto più bassa rispetto a quella degli altri modelli. Nelle figure sottostanti è possibile osservare le funzioni di densità relative alla distribuzione di probabilità S sia per il portafoglio piccolo (stimato sia in maniera empirica che in maniera parametrica) che per il portafoglio grande.

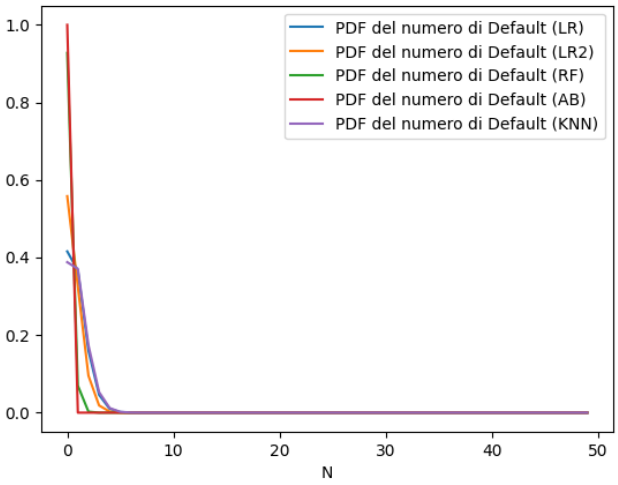

Figura 3.14: Distribuzione non parametrica della variabile S relativa al portafoglio piccolo, stimata da ognuno dei cinque modelli.

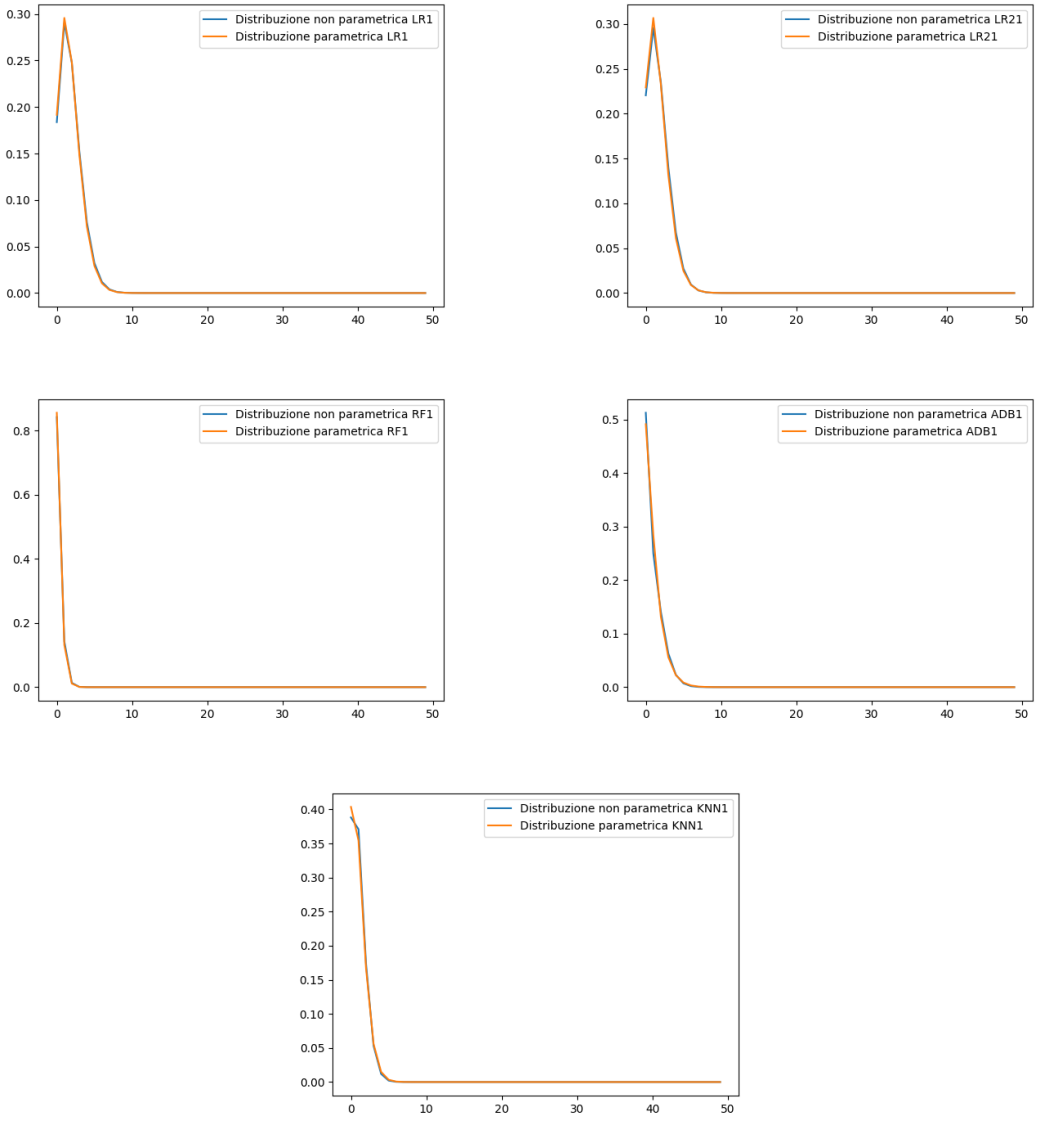

Figura 3.15: Confronto tra la distribuzione parametrica e quella non parametrica per la variabile S per il portafoglio piccolo, stimate da ognuno dei modelli considerati (LR, LR2, RF, AB, KNN).

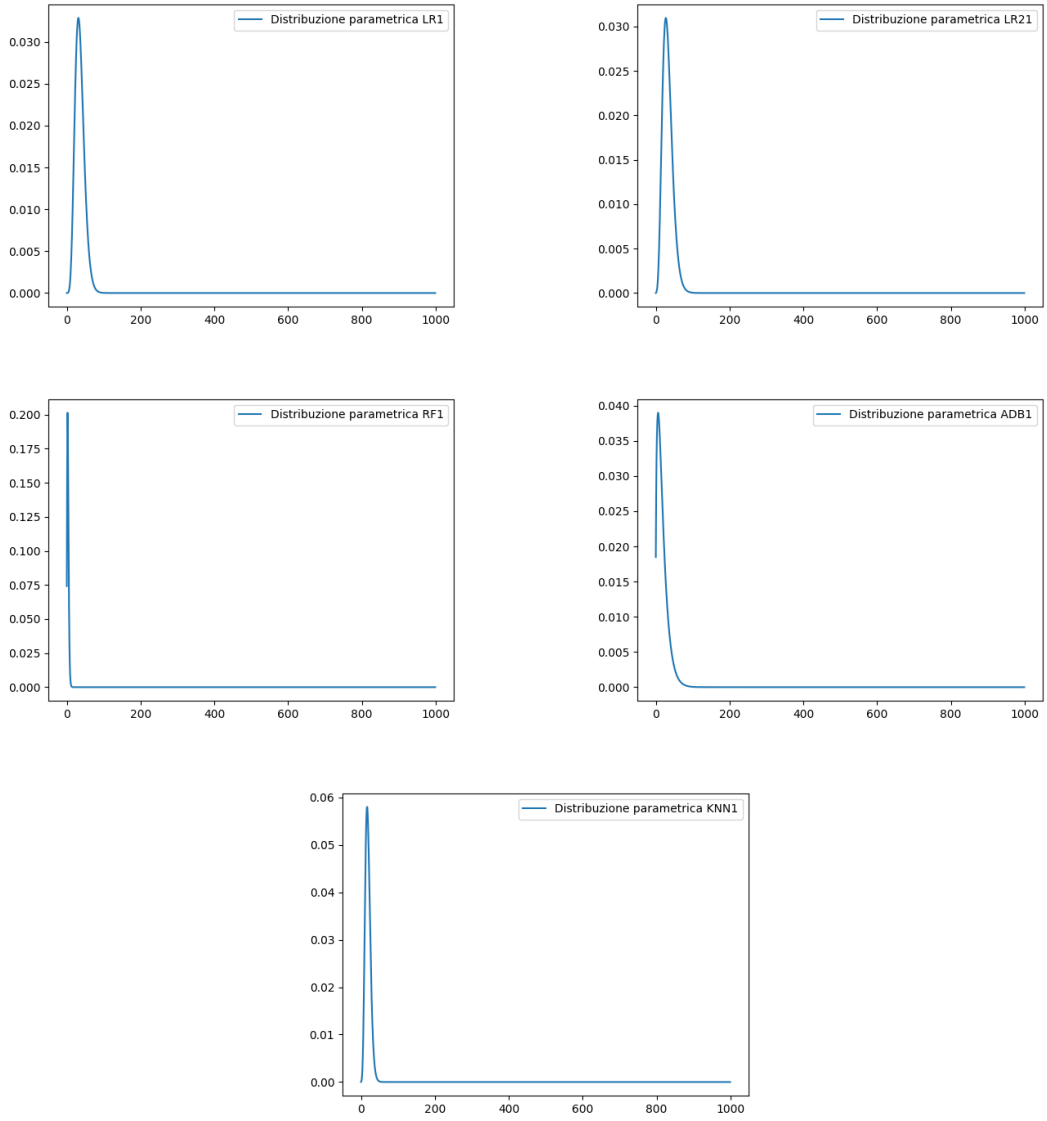

Figura 3.16: Distribuzione variabile S per il portafoglio grande, stimata dai cinque modelli considerati (LR, LR2, RF, AB, KNN).

3.0.3 Portafoglio 2

Algoritmo	Precisione	Recupero	F1-score	AUC
RF	0.72645	0.8087	0.72645	0.7217
LR	0.6555	0.80758	0.72369	0.7160
LR2	0.7286	0.8075	0.7300	0.72414
AB	0.7309	0.7947	0.7445	0.7245
KNN	0.7304	0.8034	0.7372	0.64544

Tabella 3.16: Misure di perfomance dei cinque modelli di Machine Learning per il cluster 2.

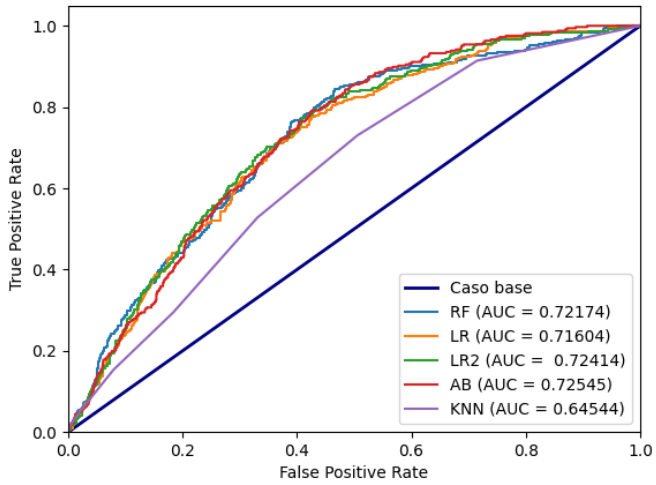

Figura 3.17: Curva ROC dei cinque modelli di classificazione, per il cluster 2.

Prima di analizzare il portafoglio cartolarizzato formato da alcuni prestiti presenti all'interno di questo cluster, vengono riportate le perfomance degli algoritmi considerati fin'ora nel classificare i dati tra cui sono stati estratti i prestiti presenti all'interno del portafoglio e le loro probabilità di default. La tabella 3.16 mostra i risultati relativi alle metriche considerate, e si può notare che i valori sono molto alti, più alti rispetto a quelli che si ottengono quando si analizza il dataset globale.

Momento	LR	LR2	RF	AB	KNN
p	0.159733	0.15447	0.062355	0.13328	0.126267
π_2	0.036419	0.035283	0.019452	0.030086	0.024045
ρ	0.081245	0.087439	0.266198	0.10666	0.073435

Tabella 3.17: Momenti stimati tramite i cinque modelli di Machine Learning.

La tabella 3.17 riassume i momenti empirici della distribuzione di probabilità del default individuale per il portafoglio ottenuto a partire dal cluster 2. Le probabilità di default stimate per questo portafoglio sono molto alte se confrontate con quelle degli altri portafogli. Solo la Random Forest stima una probabilità di default più bassa, un po' come avviene anche nel caso degli altri portafogli. Per la costruzione della distribuzione di probabilità S della somma dei default, è necessario calcolare i parametri a e b della distribuzione Q_h per ogni h = LR, LR2, RF, AB, KNN, per questo portafoglio.

$\beta - parametri$	LR	LR2	RF	AB	KNN
a	1.806315	1.612167	0.171890	1.11627	1.593157
b	9.502025	8.824258	2.584710	7.258928	11.0241474
p	0.159733	0.154475	0.06235	0.13328	0.1262676
ρ	0.081245	0.087439	0.266198	0.10666	0.073435

Tabella 3.18: Parametri della distribuzione Q_h stimati per ognuno dei cinque modelli.

Nella tabella 3.18 sono riportati i parametri della distribuzione Q_h stimati per ognuno dei modelli considerati. Il momento primo $E(Q_h) = p$ e la correlazione ρ stimati in maniera parametrica sono molto simili a quelli ottenuti in maniera empirica dai dati. Si è dimostrato che, sopratutto sulle code, non è importante tanto la forma funzionale quanto che i primi momenti siano stimati bene, perciò è possibile proseguire con le analisi di questo portafoglio.

	LR	LR2	RF	AB	KNN
distanza KL	0.048565	0.138345	0.12742	0.10902	0.03015

Tabella 3.19: Distanza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica.

La tabella 3.26, mostra la divergenza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica per il portafoglio piccolo composto da $d = 50$ clienti. Si può vedere che la distribuzione parametrica approssima in maniera molto fedele quella empirica.

α	LR	LR2	RF	AB	KNN
0.90	16	16	6	15	13
0.95	18	18	11	19	16
0.99	21	21	25	22	21

Tabella 3.20: VaR della distribuzione empirica S relativa al terzo portafoglio.

α	LR	LR2	RF	AB	KNN
0.90	16	16	11	13	14
0.95	19	18	17	15	17
0.99	24	21	31	19	23

Tabella 3.21: VaR della distribuzione beta-binomiale S relativo al terzo portafoglio.

Le tabelle 3.20 e 3.21 mostrano il VaR calcolato per la distribuzione parametrica e quella empirica per il portafoglio composto da $d = 50$ clienti. Avendo ottenuto che per questo portafoglio sia i primi momenti, che il VaR che la distanza di Kullback-Leibler concordano nel dire che la distribuzione parametrica approssima in maniera sufficientemente buona i dati, è possibile proseguire le analisi su un portafoglio più grande formato da $d = 1000$ prestiti.

α	LR	LR2	RF	AB	KNN
0.90	307	306	211	291	253
0.95	363	364	343	358	304
0.99	473	479	606	488	407

Tabella 3.22: VaR della distribuzione beta-binomiale S per il portafoglio formato da 1000 prestiti.

La tabella 3.22 riporta i risultati relativi al VaR calcolato per i tre livelli standard considerati $\alpha = 0.9, 0.95, 0.99$. I due algoritmi di Regressione Logistica hanno delle stime molto simili che riflette la stima dei primi due momenti, che per questi due modelli erano, appunto, molto simili tra di loro. L'algoritmo di Random Forest ha il valore di VaR più alto (nonostante stimasse la probabilità di default più alta) se si considera il livello $\alpha = 0.99$, dove il ruolo del secondo momento e della correlazione hanno un peso più importante. I valori di VaR più bassi si trovano invece in corrispondenza dell'algoritmo di KNN. Nelle figure sottostanti è possibile osservare le funzioni di densità relative alla distribuzione di probabilità S sia per il portafoglio piccolo (stimato sia in maniera empirica che in maniera parametrica) che per il portafoglio grande.

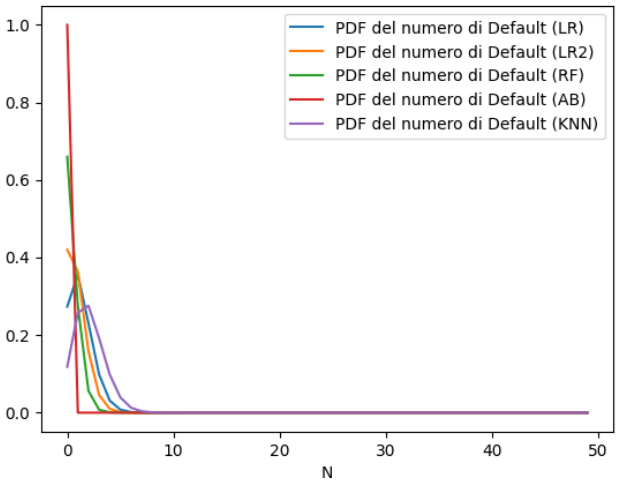

Figura 3.18: Distribuzione non parametrica della variabile S relativa al portafoglio piccolo, stimata da ognuno dei cinque modelli.

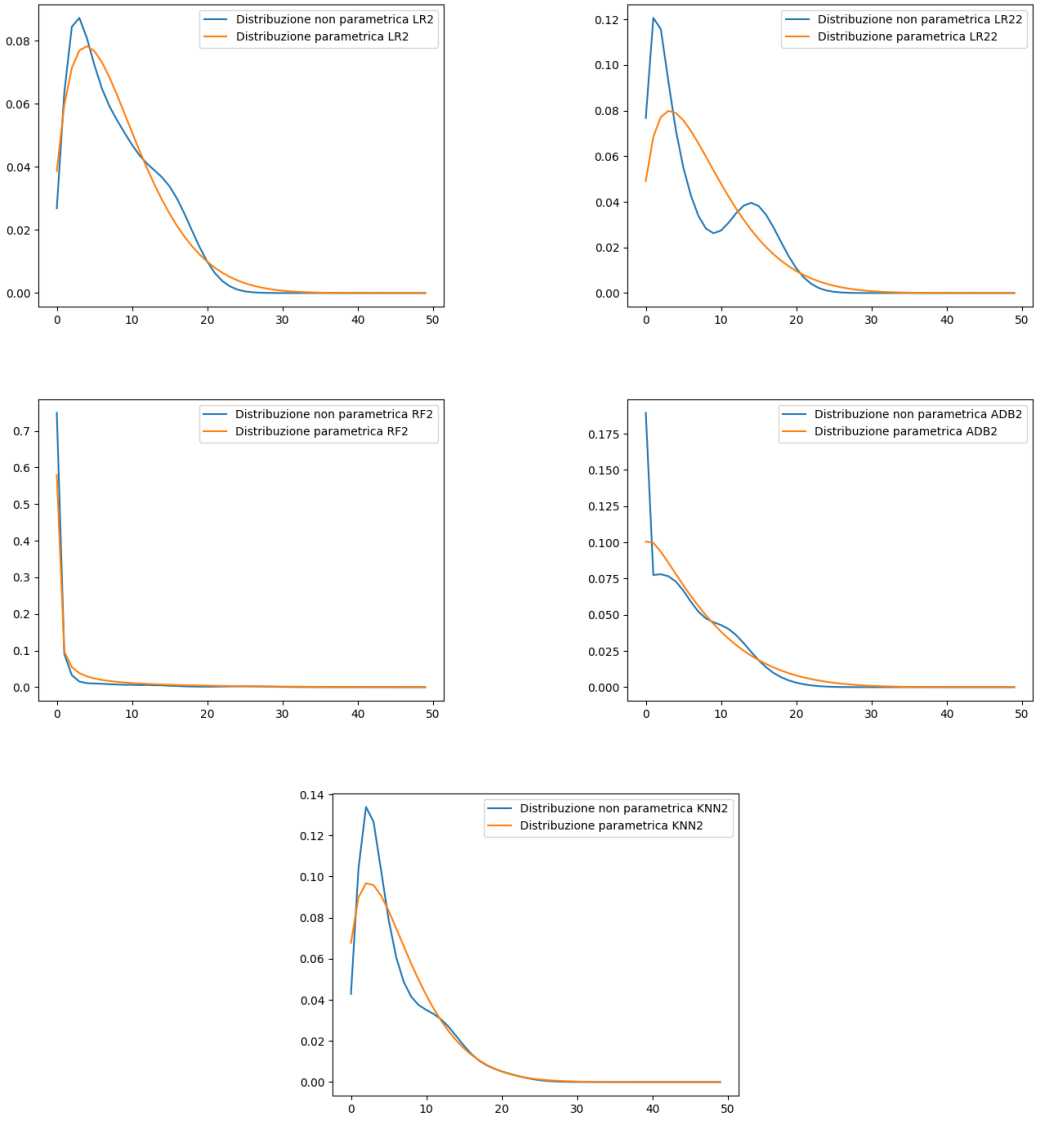

Figura 3.19: Confronto tra la distribuzione parametrica e quella non parametrica per la variabile S per il portafoglio piccolo, stimate da ognuno dei modelli considerati (LR, LR2, RF, AB, KNN).

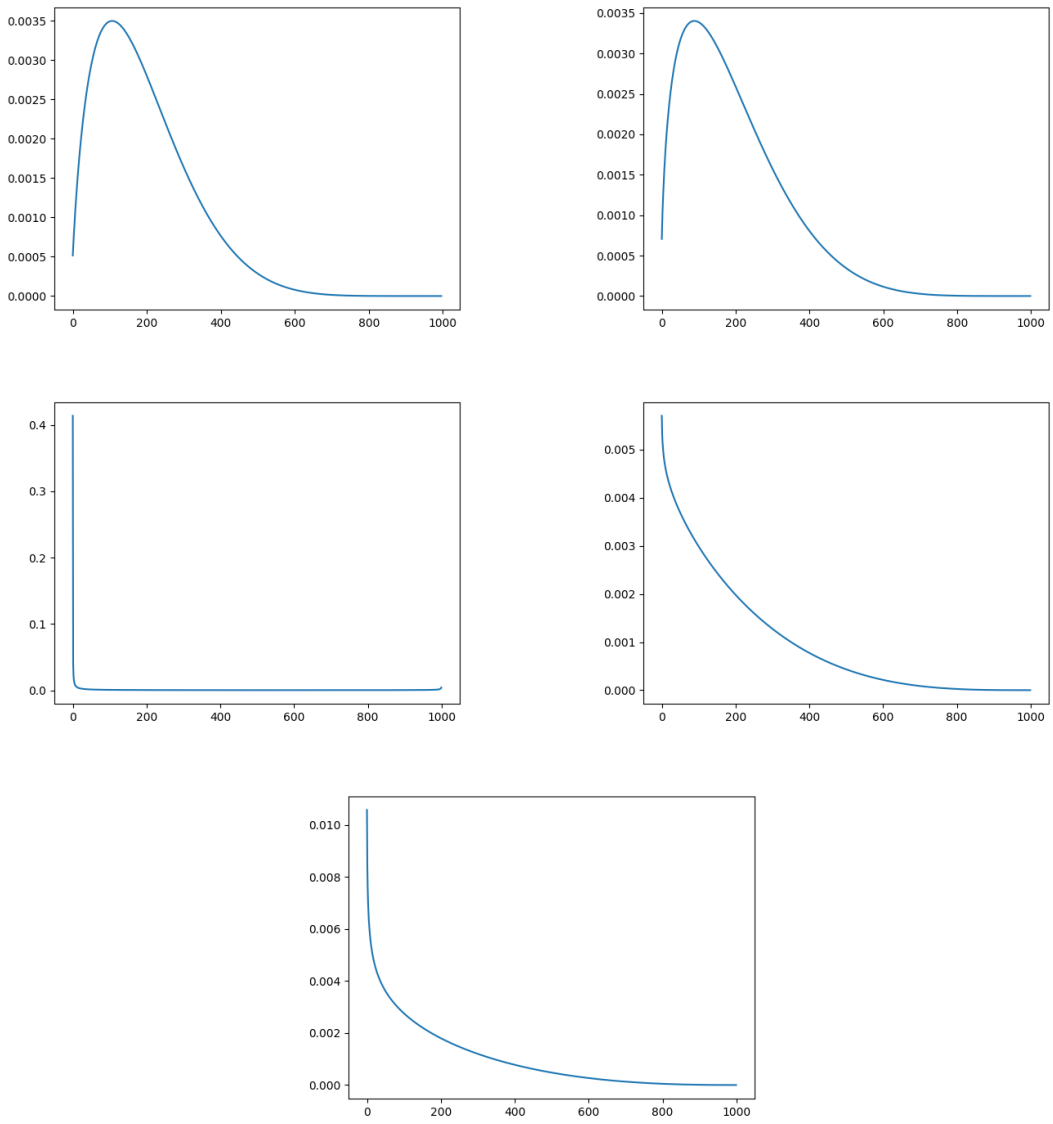

Figura 3.20: Distribuzione variabile S per il portafoglio grande, stimata dai cinque modelli considerati (LR, LR2, RF, AB, KNN).

3.0.4 Portafoglio 3

Algoritmo	Precisione	Recupero	F1-score	AUC
RF	0.7967	0.7961	0.7651	0.82744
LR	0.7835	0.7952	0.7795	0.81845
LR2	0.7821	0.7940	0.7805	0.8208
AB	0.7871	0.7976	0.7800	0.8309
KNN	0.7186	0.7465	0.7103	0.7688

Tabella 3.23: Misure di perfomance dei cinque modelli di classificazione per il cluster 3.

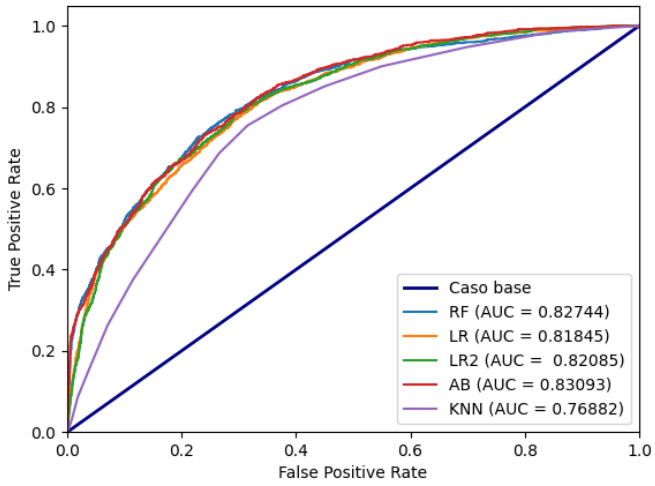

Figura 3.21: Curva ROC dei cinque modelli considerati, per il cluster 3.

Prima di analizzare il portafoglio cartolarizzato formato da alcuni prestiti presenti all'interno di questo cluster, vengono riportate le perfomance degli algoritmi considerati fin'ora nel classificare i dati tra cui sono stati estratti i prestiti presenti all'interno del portafoglio e le loro probabilità di default. La tabella 3.23 mostra i risultati relativi alle metriche considerate, e si può notare che i valori sono molto alti, più alti rispetto a quelli che si ottengono quando si analizza il dataset globale.

Momento	LR	LR2	RF	AB	KNN
p	0.04160	0.03937	0.029549	0.028484	0.060914
π_2	0.00202	0.00185	0.000898	0.00121	0.003880
ρ	0.00738	0.008136	0.000900	0.014587	0.002963

Tabella 3.24: Momenti stimati tramite i cinque modelli di Machine Learning.

La tabella 3.24 riassume i momenti empirici della distribuzione di probabilità del default individuale per il portafoglio ottenuto a partire dal cluster 3. Le probabilità di default stimate per questo portafoglio sono un po' più alte rispetto a quelle dei primi due, sopratutto per l'algoritmo KNN e i due modelli di Regressione Logistica al primo e al secondo ordine. La Random Forest e l'Ada Boost stimano invece delle probabilità di default molto simili tra di loro, cambia solo la correlazione, che è più alta per l'Ada Boost. Per la costruzione della distribuzione di probabilità S della somma dei default, è necessario calcolare i parametri a e b della distribuzione Q_h per ogni h = LR, LR2, RF, AB, KNN, per questo portafoglio.

$\beta - parametri$	LR	LR2	RF	AB	KNN
a	5.59215	4.79945	32.77942	1.92414	20.49600
b	128.812	117.10217	1076.54152	65.62581	315.9744
p	0.041606	0.0393715	0.029549	0.028484	0.060914
ρ	0.007385	0.008136	0.000900	0.0145879	0.0029632

Tabella 3.25: Parametri della distribuzione Q_h stimati per ognuno dei cinque modelli.

Nella tabella 3.25 sono riportati i parametri della distribuzione Q_h stimati per ognuno dei modelli considerati. Il momento primo $E(Q_h) = p$ e la correlazione ρ stimati in maniera parametrica sono molto simili a quelli ottenuti in maniera empirica dai dati. Si è dimostrato che, sopratutto sulle code, non è importante tanto la forma funzionale quanto che i primi momenti siano stimati bene, perciò è possibile proseguire con le analisi di questo portafoglio.

	LR	LR2	RF	AB	KNN
distanza KL	0.015744	0.00975	0.000118	0.01952	0.0004774

Tabella 3.26: Distanza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica.

La tabella 3.26, mostra la divergenza di Kullback-Leibler tra la distribuzione parametrica e quella non parametrica per il portafoglio piccolo composto da $d = 50$ clienti. Si può vedere che la distribuzione parametrica approssima in maniera molto fedele quella empirica.

α	LR	LR2	RF	AB	KNN
0.90	4	4	3	4	5
0.95	5	5	4	5	6
0.99	6	7	5	6	8

Tabella 3.27: VaR della distribuzione empirica S relativa al quarto portafoglio.

α	LR	LR2	RF	AB	KNN
0.90	4	4	3	3	5
0.95	5	5	3	4	6
0.99	7	7	5	6	8

Tabella 3.28: VaR della distribuzione beta-binomiale S relativo al secondo portafoglio.

Le tabelle 3.27 e 3.28 mostrano il VaR calcolato per la distribuzione parametrica e quella empirica per il portafoglio composto da $d = 50$ clienti. Avendo ottenuto che per questo portafoglio sia i primi momenti, che il VaR che la distanza di Kullback-Leibler concordano nel dire che la distribuzione parametrica approssima in maniera sufficientemente buona i dati, è possibile proseguire le analisi su un portafoglio più grande formato da $d = 1000$ prestiti.

α	LR	LR2	RF	AB	KNN
0.90	67	65	40	57	82
0.95	76	75	43	70	88
0.99	95	95	50	97	101

Tabella 3.29: VaR della distribuzione beta-binomiale S per il portafoglio formato da 1000 prestiti.

La tabella 3.29 riporta i risultati relativi al VaR calcolato per i tre livelli standard considerati $\alpha = 0.9, 0.95, 0.99$. I due algoritmi di Regressione Logistica hanno delle stime molto simili che riflette la stima dei primi due momenti, che per questi due modelli erano, appunto, molto simili tra di loro. L'algoritmo di Random Forest ha i valori di VaR più bassi, più bassi anche dell'algoritmo di Ada Boost, nonostante avessero la stessa probabilità di default (ma correlazione diversa). Quest'ultimo ha valori di VaR più simili ai due modelli di Regressione Logistica. Nelle figure sottostanti è possibile osservare le funzioni di densità relative alla distribuzione di probabilità S sia per il portafoglio piccolo (stimato sia in maniera empirica che in maniera parametrica) che per il portafoglio grande.

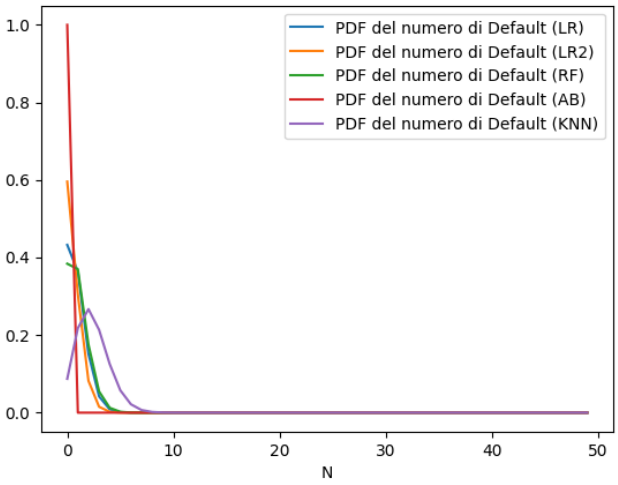

Figura 3.22: Distribuzione non parametrica della variabile S relativa al portafoglio piccolo, stimata da ognuno dei cinque modelli.

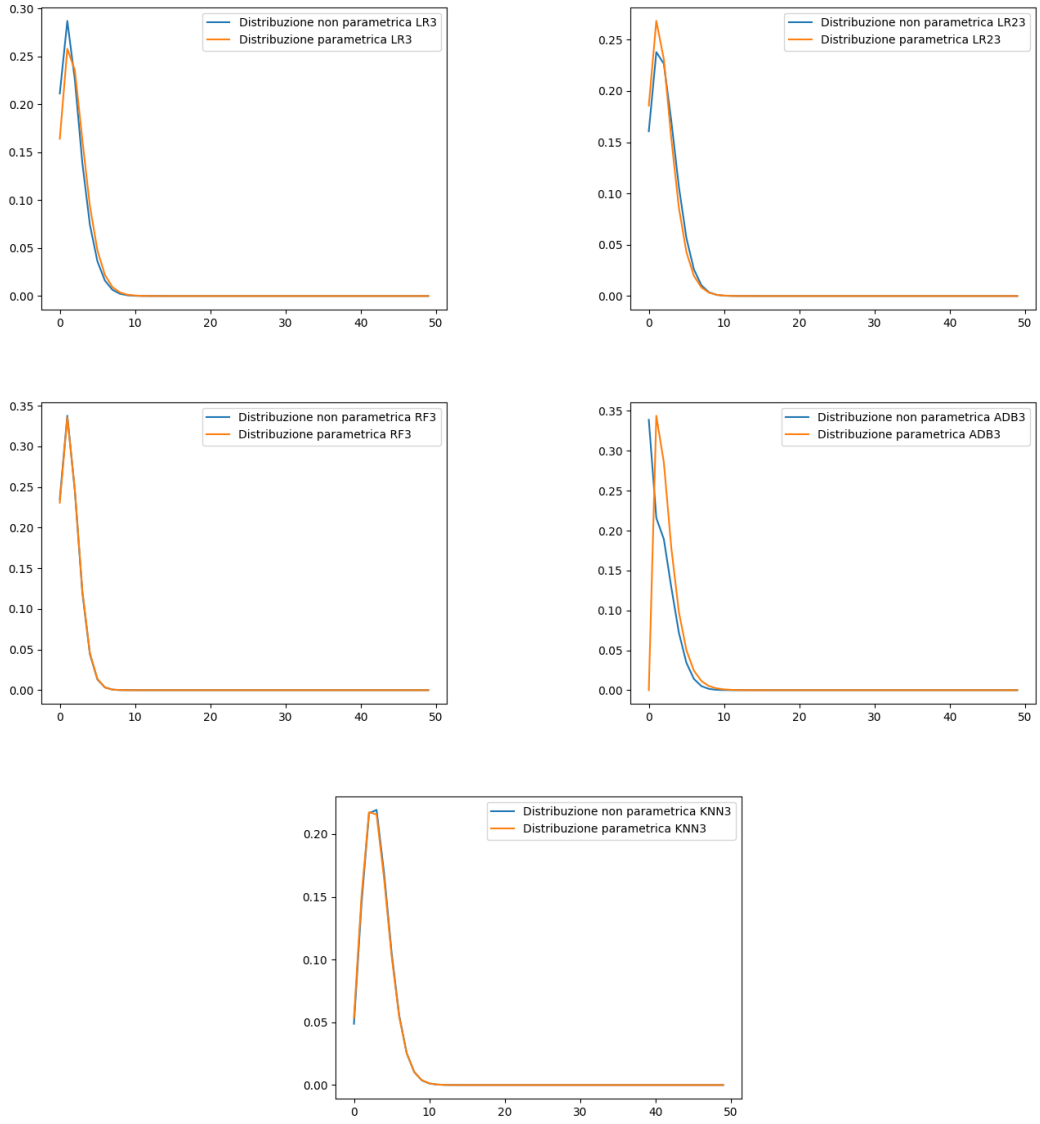

Figura 3.23: Distribuzione parametrica e non parametrica del numero di default per il cluster 3, per i cinque modelli di Machine Learning (LR, LR2, RF, AB, KNN).

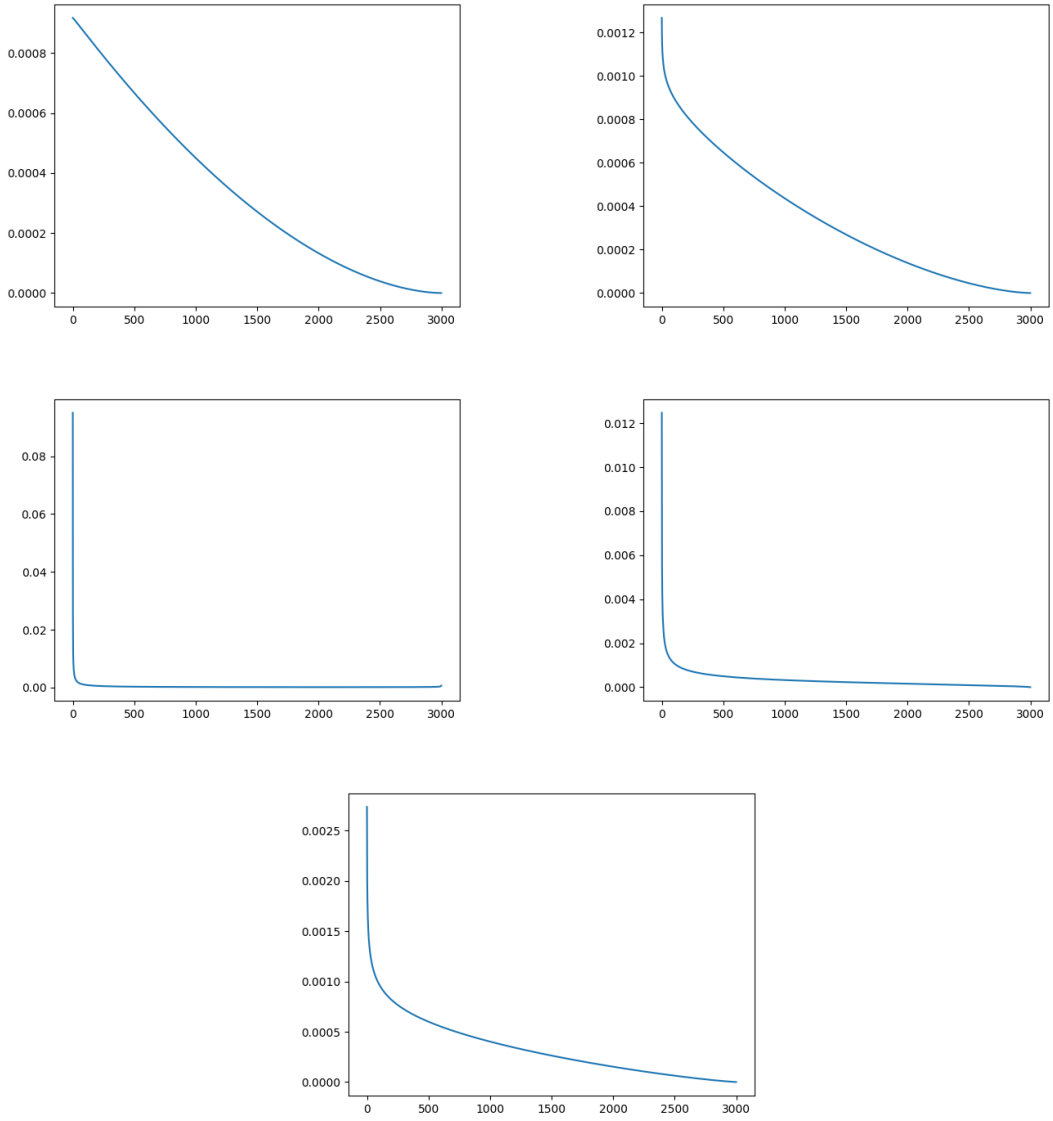

Figura 3.24: Distribuzione variabile S per il portafoglio grande, stimata dai cinque modelli considerati (LR, LR2, RF, AB, KNN).

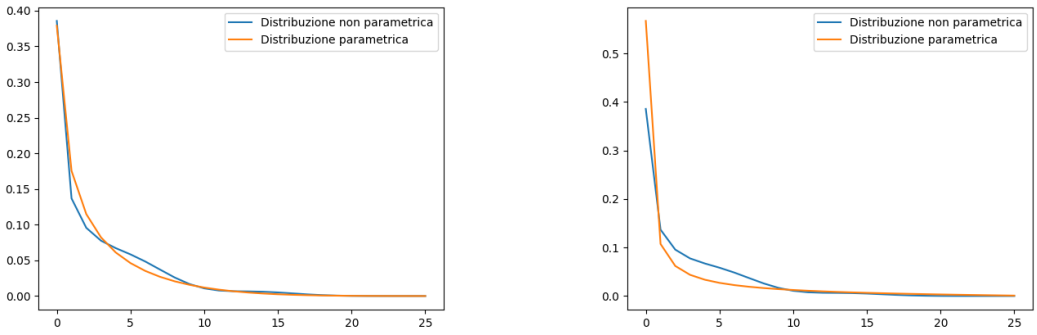

Figura 3.25: Distribuzione parametrica e non parametrica del numero di default per il cluster 1, per i cinque modelli di Machine Learning (LR, LR2, RF, AB, KNN).

Capitolo 4

Risultati relativi all'analisi del portafoglio cartolarizzato.

L'analisi della cartolarizzazione inizia partendo dalla struttura a cascata corrispondente, che in questo caso è stata scelta di tipo *full sequential*, in cui le tre tranche vengono ripagate in maniera sequenziale, per cui la tranche mezzanine non viene ripagata fino a quando non viene ripagata del tutto la tranche senior, e così per la tranche junior. Per questo motivo, la tranche senior sarà molto più stabile e meno influenzata dalla probabilità di default stimata, mentre le due tranche subordinate subiranno un'influenza maggiore.

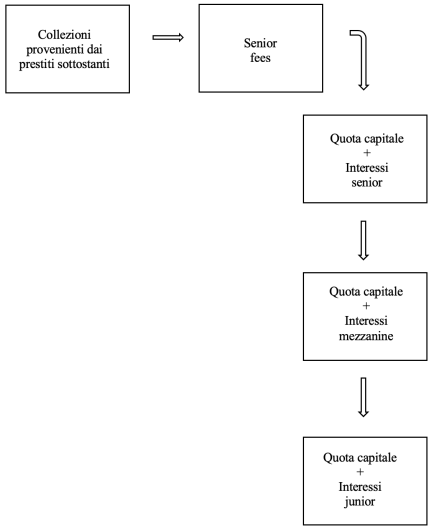

Figura 4.1: Struttura a cascata di tipo *full sequential*: a partire dai flussi di cassa generati dai prestiti, si pagano delle *fees* fisse e poi si ripagano in maniera sequenziale le tre tranches.

Nella struttura del ripagamento delle note è stata inserito anche un tasso di recupero

crediti, costante, che è il valore standard che si utilizza nel pricing di una cartolarizzazione. Per avere una modellizzazione accurata dei default e della loro influenza nel ripagamento delle note è necessario specificare non solo un tasso di default ma anche la distribuzione temporale dei default, cioè quando effettivamente avvengono i default. Le due forme funzionali più utilizzate prevedono una struttura decrescente, in quanto si suppone che un debitore che sia stato in grado di ripagare la maggior parte del mutuo, riesca a completare anche gli ultimi pagamenti, o una struttura uniforme, per cui i default totali stimati avvengono in maniera uniforme durante tutta la durata del prestito. Indicando con N il numero totale di anni di vita del prestito, la distribuzione decrescente è stata calcolata come:

$$\pi_{T,n} = \left(\frac{1}{N}\right)^2 \cdot ((N)^2 - (n + n + 1)),$$

Dove $\pi_{T,n}$ indica la frazione di default, rispetto al numero di default atteso totale, che avviene ogni singolo anno, con $n = 0, 1, \ldots, N$, Nei casi qui considerati, dove si analizza un portafoglio con scadenza a $N = 3$ anni e un portafoglio con scadenza a $N = 5$ anni, queste formule si traducono in:

$$\pi_{T,1} = \frac{5}{9}, \quad \pi_{T,2} = \frac{3}{9}, \quad \pi_{T,3} = \frac{5}{9},$$

per il portafoglio con scadenza a 3 anni, e

$$\pi_{T,1} = \frac{9}{25}, \quad \pi_{T,2} = \frac{7}{25}, \quad \pi_{T,3} = \frac{5}{25}, \quad \pi_{T,4} = \frac{3}{25}, \quad \pi_{T,5} = \frac{1}{25}$$

per il portafoglio con scadenza a 5 anni. La distribuzione uniforme invece precede semplicemente che, ogni anno ci siano $\frac{M}{N}$ default, dove M è il numero di default atteso. Per completare le analisi sono anche state considerate delle forme funzionali "limite" che considerano i default avvenire tutti all'inizio nel corso del primo anno, tutti alla fine nel corso dell'ultimo anno, o tutti durante gli anni intermedi della durata del prestito. Si vedrà che sia la forma funzionale scelta che la probabilità di default stimata influenzeranno il ripagamento delle note sia come tempistiche che come entità.

Per ottenere dei risultati numerici si è utilizzato un algoritmo in cui bisogna inserire come parametro la probabilità di default (indicata con π_D) e una distribuzione temporale per i default (indicata con π_T), a partire da questi valori, vengono calcolate le perdite monetarie dovute a questo tasso di default e il modo in cui queste perdite incidono sulla disponibilità effettiva di denaro che ad ogni istante temporale può essere utilizzato per ripagare le note.

L'algoritmo ha tra le ipotesi una distribuzione per le collezioni di denaro che ad ogni trimestre vengono raccolte dalle rate dei prestiti, e si è scelta una distribuzione decrescente, per cui il denaro raccolto nei primi anni è maggiore di quello raccolto negli ultimi anni. La scelta è motivata dal fatto che a volte un debitore riesce a chiudere il suo prestito in anticipo rispetto alla scadenza. Quando viene scelta una probabilità di default e viene indicata una distribuzione temporale per il numero di default, il denaro a disposizione ogni trimestre, derivante da queste collezioni, che viene utilizzato dalle banche per ripagare le note agli investitori, diminuisce. In formule, indicando con IAF_i il denaro a disposizione per ogni trimestre:

$$\mathrm{IAF}_i = \mathrm{CF}_i - C_{DLC,i} + R_i \tag{4.1}$$

dove il pedice i indica il trimeste considerato: la cartolarizzazione ha una maturità di dieci anni, perciò i varia da 1 a 40, ma può avere una durata più breve nel caso in cui si riesca a ripagare tutte le tranches prima della scadenza dei dieci anni. CF_i indica il denaro totale raccolto ad ogni istante temporale, $C_{DLC,i}$ il default lordo cumulato (cioè quanto denaro è stato perso dall'emissione del titolo per via dei default) e R_i la quanità cumulata di denaro recuperata in seguito ad ogni default. Il default lordo cumulato si calcola come:

$$C_{DLC,i} = C_{DLC,i-1} + \pi_D \cdot \pi_{T,n} \cdot V,$$
$$C_{DLC,1} = \pi_D \cdot \pi_{T,1} \cdot V.$$

e dove V indica il valore del portafoglio (che qui è assunto essere $V = 100.000.000$), $n = 1, \ldots, N$ indica la durata del prestito (3 0 5 anni) e il recupero cumulato si calcola come

$$R_i = R_{i-1} + \pi_R \cdot C_{DLC,i-1}$$

dove π_R indica il tasso di recupero credito (che è assunto costante). Il recupero avviene sempre un istante temporale successivo perché all'istante $i-1$ si osserva il default, e all'istante i se ne recupera una frazione.
L'equazione (4.1) può quindi essere riscritta mostrando la probabilità di default all'interno della formula

$$\mathrm{IAF}_i = \mathrm{CF}_i - \pi_D \cdot \pi_{T,n} \cdot C_{DLC,i} + \pi_{R,i} \cdot C_{DLC,i-1} \tag{4.2}$$

Il pagamento delle note avviene con cadenza trimestrale, utilizzando il denaro disponibile (IAF) ed è suddiviso in due parti: una parte di interessi (indicata con $\tilde{r}$) ed una parte capitale (indicata con q). In formule, dopo aver ripagato la quota fissa di commissioni, si ha:

$$\mathrm{IAF}_i = \tilde{r}_i + q_i \tag{4.3}$$

La parte di interesse viene ripagata per prima e paga una percentuale, chiamata margine e indicata con m, sul denaro che resta ancora da ripagare alle tre tranche come capitale, chiamato *outstanding* e viene indicato con O. La percentuale dipende dal rischio legato all'investimento, per cui sarà più alta per la tranche junior e più bassa per la tranche senior. L'interesse viene calcolato mediante le formule:

$$\tilde{r}_{S,i} = O_{S,i} \cdot m_S \cdot \tilde{t}, \tag{4.4}$$
$$\tilde{r}_{M,i} = O_{M,i} \cdot m_M \cdot \tilde{t}, \tag{4.5}$$
$$\tilde{r}_{J,i} = O_{J,i} m_J \cdot \tilde{t}. \tag{4.6}$$

Dove $\tilde{t}$ è relativo al numero di pagamenti che viene effettuato ogni anno: poiché si è scelta una cadenza trimestrale, sarà di $\frac{1}{4}$. Il margine si trova come quel valore di m che realizza

l'equazione

$$\sum_{i=1}^{N_S} e^{-rt_i}(q_i + \tilde{r}_{S,i}) = \mathrm{V}_S \tag{4.7}$$

$$\sum_{i=N_S}^{N_M} e^{-rt_i}(q_i + \tilde{r}_{M,i}) = \mathrm{V}_M \tag{4.8}$$

$$\sum_{i=N_M}^{N_J} e^{-rt_i}(q_i + \tilde{r}_{J,i}) = \mathrm{V}_J \tag{4.9}$$

Dove r indica il tasso di interesse *risk-free* ed è una stima del tasso euribor a 3 mesi all'emissione del titolo (non si hanno informazioni su quando è stato creato il dataset utilizzato, perciò non è stato possibile conoscere quale fosse il tasso euribor a 3 mesi corretto). t_i indica la distanza temporale dall'emissione del titolo ed è relativo al fattore di sconto dei flussi di cassa, $V_S = 85.000.000$, $V_M = 15.000.000$ e q $V_J = 5.000.000$, con $V_S + V_M + V_J = V$, mentre N_S, N_M, N_J indicano il numero totale di trimestri necessari a ripagare le tre tranches e, poiché la struttura è di tipo *full sequential*, il ripagamento di una tranche successiva inizia quando viene completato il pagamento di quella precedente. Non è noto a priori il numero di trimestri necessario ad esaurire ogni tranche, però i valori N_S, N_M, N_J hanno il vincolo ulteriore che

$$\sum_{i=1}^{N_S} q_i = V_S$$

$$\sum_{i=N_S}^{N_M} q_i = V_M$$

$$\sum_{i=N_M}^{N_J} q_i = V_J$$

Si può notare che questo sistema di equazioni non è lineare e non è possibile risolverlo in forma chiusa ma solo in forma numerica, utilizzando un algoritmo di ricerca esaustivo che va a cercare, in un range di valori, il valore di margine che rispetta l'equazione. Questa ricerca non può essere svolta in parallelo, ma in serie: quando si trova il margine m relativo alla tranche senior, si può proseguire con la ricerca del margine relativo alla tranche mezzanine e così per la tranche junior.
Il *pricing* di una cartolarizzazione in fase di emissione sul mercato primario viene effettuato poi sfruttando i margini ricavati nelle equazioni (4.7), (4.8), (4.9). In formule:

$$p_S = 100 = \frac{\sum_{i=1}^{N_S} e^{-r\cdot t_i}(q_{S,i} + \tilde{r}_{S,i}) \cdot 100}{V_S} \tag{4.10}$$

$$p_M = 100 = \frac{\sum_{i=N_S}^{N_M} e^{-r\cdot t_i}(q_{M,i} + \tilde{r}_{M,i}) \cdot 100}{V_M} \tag{4.11}$$

$$p_J = 100 = \frac{\sum_{i=N_M}^{N_J} e^{-r\cdot t_i}(q_{J,i} + \tilde{r}_{J,i}) \cdot 100}{V_J} \tag{4.12}$$

La quota capitale q invece, viene ripagata rimborsando alle tre tranches, tutto il denaro che viene raccolto dal pagamento delle rate dei prestiti sottostanti, che rimane al netto del pagamento degli interessi, viene ripagata ogni trimestre ed entra nella definizione di wal (*weighted average life*), che è una stima della velocità con cui il titolo ripaga i detentori e dipende da quando viene effettuato il pagamento principale tra quelli promessi ai detentori del titolo. In formule:

$$\text{wal}_{i,j} = \frac{t_i \cdot q_{j,i}}{V_j}$$

con $j = S, M, J$.

Nelle tabelle sottostanti sono riportati i risultati relativi al margine che si ottiene e al WAL corrispondente. PD indica la probabilità di default stimata, MS indica il margine della tranche senior mentre WS il wal corrispondente. Discorso analogo anche per MM (margine mezzanine) e MJ.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	2,48%	3,45%	0,970533	9,17%	2,4418	15,38%	3,01389
LR2	2,22%	3,45%	0,9655	9,84%	2,3765	18,79%	2,9071
RF	1,04%	3,48%	0,9451	8,51%	2,1542	19,53%	2,6031
AB	1,3%	3,47%	0,9483	12,38%	2,20	17,4%	2,65416
KNN	3,58%	3,41%	0,9969	9,22%	2,79249	18,36%	3,5059

Tabella 4.1: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il primo portafoglio analizzato utilizzando la forma funzionale decrescente.

La tabella 4.1 mostra i risultati, in termini di margine e wal, per il portafoglio cartolarizzato in cui i default sono distribuiti in maniera decrescente. Nel corso dei trimestri del primo anno si hanno più default che nei trimestri successivi. Si può osservare che la tranche senior è molto stabile e non è influenzata dalla probabilità di default più alta o più bassa. Le due tranche subordinate hanno una variabilità maggiore. Si può notare che l'algoritmo di RF, che stima la probabilità di default più bassa, ha il margine e il wal più bassi per la tranche mezzanine. Chi investe in questa tranche, poiché assume il rischio di default più basso, deve accettare un flusso di denaro più basso e di durata minore. Il rapporto migliore tra margine e wal, per questa tranche, è dato dall'algoritmo di AB che stima una probabilità di default abbastanza bassa, ma leggermente più alta della Random Forest, per cui viene premiato il rischio maggiore di default con un margine più alto (ma wal abbastanza basso). Gli altri tre algoritmi, che stimano probabilità di default maggiori hanno un margine più basso ma wal più alto, per cui chi detiene questa tranche ha un flusso di denaro stabile per un tempo più lungo, infatti il numero maggiore di default che viene stimato da questi tre modelli prolunga il tempo di vita della tranche, in quanto ci vorrà più tempo, in media, per poter ripagare completamente la tranche. La tranche junior, essendo la più rischiosa, è quella con i margini più alta, circa il doppio della tranche mezzanine. Si può notare che l'algoritmo KNN stima i wal più alti per tutte e tre le tranches, perciò è quello che impiega più tempo a ripagare tutte le note, che può essere un vantaggio per chi preferisce un investimento un po' più lungo e stabile.

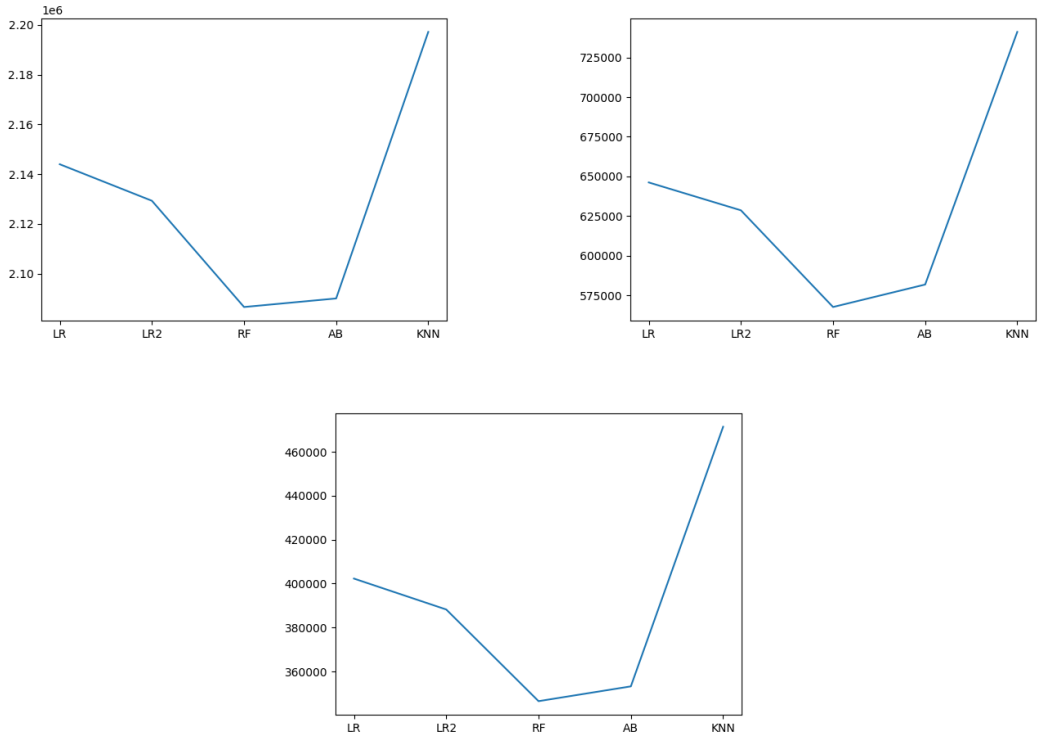

Figura 4.2: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale decrescente.

I grafici riportati in 4.2 mostrano gli interessi pagati sulle note per le tre tranches (senior, mezzanine e junior). Si nota che l'algoritmo di RF, che stima la probabilità di default più bassa, ha gli interessi minori, mentre l'algoritmo KNN, che stima la probabilità di default più alta, ha gli interessi più alti. I tre grafici mostrano la stessa tendenza, molto nota, per cui ad un rischio maggiore si associa un rendimento maggiore. Però le tre tranches hanno rischi differenti e quindi ci si aspetta che l'entità del rischio assunto (misurato dalla probabilità di default) abbia effetti quantitativi differenti nelle tre tranches, perciò per calcolare in maniera quantitativa l'effetto della probabilità di default maggiore o minore, si è calcolata la differenza percentuale tra l'interesse medio ($Barx$) che viene pagato alle tre tranches e l'interesse pagato (x_{ij}) con ognuna delle probabilità di default ottenute, mediante la formula

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}}{\bar{x}} \cdot 100.$$

dove $i \in \{\text{LR, LR2, RF, AB, KNN}\}, j \in \{\text{senior, mezzanine, junior}\}$.

Algoritmo	Senior	Mezzanine	Junior
LR	0.6837%	2.07271%	2.54076%
LR2	-0.00546%	-0.7045%	-1.04716%
RF	-2.01164%	-10.3416%	-11.6852%
AB	-1.8498%	-8.1019%	-9.9688%
KNN	3.1831%	17.0753%	20.16052%

Tabella 4.2: Differenza percentuale tra l'interesse pagato e l'interesse medio.

Dalla tabella 4.24 si può osservare come gli algoritmi di Regressione Logistica e di KNN ripagano degli interessi più alti della media, ma la tranche junior ripaga degli interessi molto più alti rispetto alla media, mentre la tranche senior ha una variazione decisamente più bassa.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	2,48%	3,46%	0,9527	11,53%	2,2566	13,07%	2,8198
LR2	2,22%	3,47%	0,9497	12%	2,23029	14,8%	2,7401
RF	1,04%	3,48%	0,94206	8,85%	2,11880	20,65%	2,5808
AB	1,3%	3,48%	0,94206	8,85%	2,11880	20,65%	2,5808
KNN	3,58%	3,45%	0,9655	9,19%	2,43995	19,14%	3,12142

Tabella 4.3: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il primo portafoglio analizzato per la forma funzionale uniforme.

La tabella 4.3 mostra i risultati relativi al caso in cui i default avvengono in maniera uniforme nel corso dei trimestri dei tre anni. Si può notare che il wal della tranche senior e mezzanine è molto simile al caso precedente, mentre il wal della tranche junior è più basso rispetto al caso precedente, per cui un portafoglio costruito in questo modo ha una durata più breve rispetto a quello precedente. Anche in questo caso la tranche senior è particolarmente stabile e non è influenzata dalla probabilità di default e dalla distribuzione differente dei default. Per la tranche mezzanine si può notare che gli algoritmi di RF, AB e di KNN hanno i margini più bassi ma, rispettivamente, il wal più basso (per RF e AB) e il wal più alto (per KNN), perciò con una probabilità di default bassa si ha un flusso di denaro minore per un tempo breve, mentre con un tasso di default più alto si ha un flusso di denaro minore ma per un tempo più lungo. Gli altri due casi sono intermedi. Per la tranche junior si ha invece un comportamento opposto: gli algoritmi di Random Forest, Ada Boost e KNN hanno margini più alti rispetto agli altri due modelli.

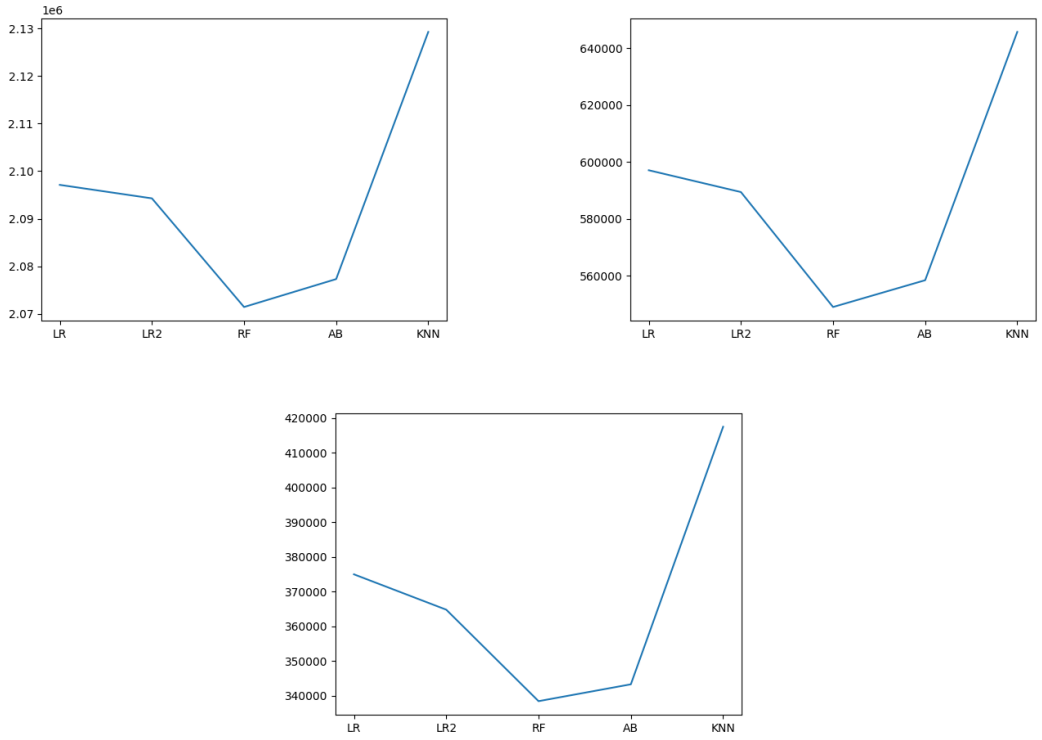

Figura 4.3: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale uniforme.

I grafici in figura 4.3 hanno lo stesso profilo del caso precedente, in cui gli interessi maggiori si osservano in corrispondenza dei modelli più rischiosi. Anche in questo caso possiamo dare delle stime quantitative su quanto maggiori siano gli interessi pagati per chi stima delle probabilità di default maggiori. Si può notare che quando i default avvengono in maniera uniforme, l'effetto della probabilità di default più alta è meno influente, in quanto la differenza percentuale è più bassa rispetto al caso precedente.

Algoritmo	Senior	Mezzanine	Junior
LR	0.15447%	1.5568%	1.9526%
LR2	0.0189%	0.25304%	-0.8174%
RF	-1.0730%	-6.619%	-7.9867%
AB	-0.7922%	-5.0176%	-6.6660%
KNN	1.69176%	9.82718%	13.51754%

Tabella 4.4: Differenza percentuale tra l'interesse pagato e l'interesse medio.

Adesso si considerano dei casi limite per avere dei *boundaries* su quanto si può essere guadagnare investendo in un titolo del genere. Il primo caso considerato è quello in cui

i default avvengono tutti durante il primo anno di vita del titolo, distribuiti in maniera uniforme nei quattro trimestri iniziali.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	2,48%	3,4%	1,0106	10,16%	2,69439	16,43%	3,27009
LR2	2,22%	3,41%	0,99910	11,48%	2,59036	21,33%	3,13330
RF	1,04%	3,46%	0,95730	11,97%	2,23176	16,65%	2,675611
AB	1,3%	3,45%	0,965653	11%	2,28887	13,73%	2,78718
KNN	3,58%	3,33%	1,07566	11,47%	3,242279	23,1%	3,95593

Tabella 4.5: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il primo portafoglio analizzato nel caso in cui tutti i default avvengono nei quattro trimestri del primo anno.

La tabella 4.5 mostra i risultati di questa casistica. Il wal aumenta per tutte e le tranches, anche se la differenza è più evidente nel caso della tranche junior piuttosto che in quella senior. Anche i margini delle due tranches subordinate aumentano rispetto ai due casi precedenti. Si può spiegare poiché in questo caso in cui i default avvengono all'inizio, è necessario più tempo per riuscire a ripagare le note.

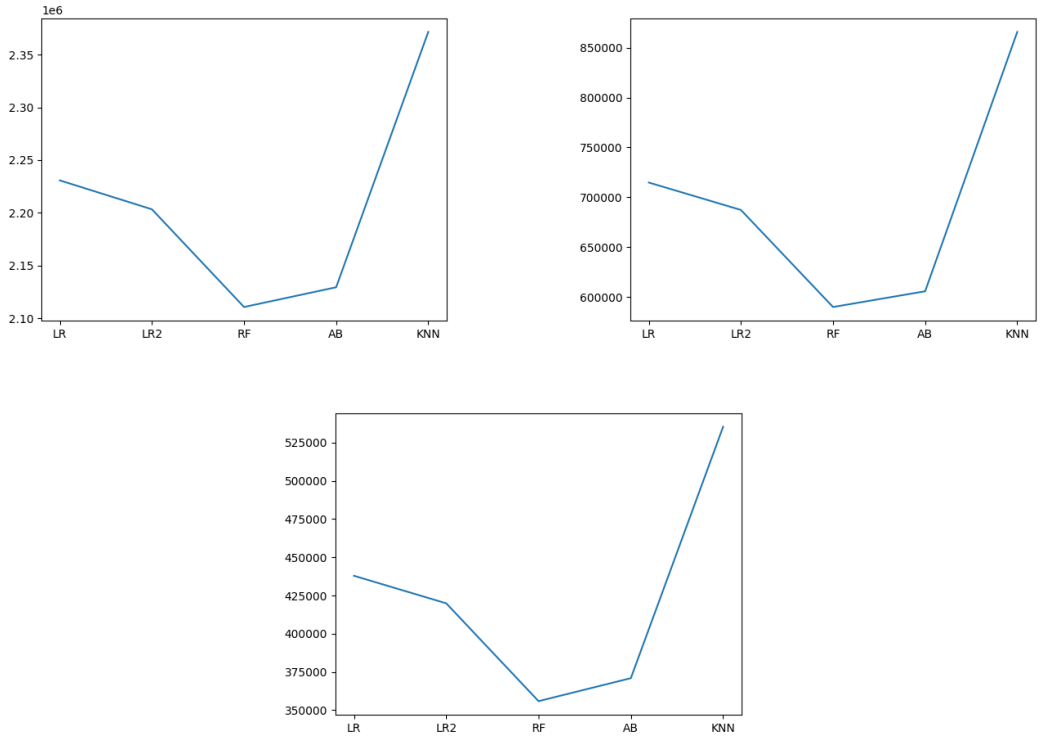

Figura 4.4: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui tutti i default avvengono durante il primo anno.

Algoritmo	Senior	Mezzanine	Junior
LR	0.97822%	3.1857%	3.2813%
LR2	-0.2630%	-0.7833%	-0.9703%
RF	-4.4567%	-14.837%	-16.0681%
AB	-3.6087%	-12.558%	-12.5206%
KNN	7.3503%	24.9937%	26.2779%

Tabella 4.6: Differenza percentuale tra l'interesse pagato e l'interesse medio.

I grafici in figura 4.4 mostrano lo stesso profilo degli altri due casi precedenti, ma in questo caso, dalla tabella 4.6 si nota come l'influenza della probabilità di default sia più alta che nei casi precedenti. In questo caso, infatti, la probabilità di default del 3.58% dell'algoritmo KNN arriva a pagare il 26.3% in più di interessi rispetto alla media per la tranche junior, ma anche il 7% in più per la tranche senior. Questo si spiega anche perché i pagamenti avvengono per un numero di mesi maggiori in quanto è necessario più tempo per ripagare tutte le note, e quindi, anche se la quota capitale è la stessa, vengono pagati degli interessi per più tempo.

Il caso successivo ipotizza che i default avvengano tutti durante il corso dei quattro trimestri del secondo anno (sui tre totali di vita del prestito)

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	2,48%	3,49%	0,93411	8,14%	2,19984	12,76%	2,83735
LR2	2,22%	3,49%	0,93392	8,37%	2,17260	14,42%	2,755471
RF	1,04%	3,49%	0,933051	9,62%	2,05113	17,97%	2,54715
AB	1,3%	3,49%	0,93324	9,32%	2,07648	20,52%	2,58320
KNN	3,58%	3,49%	0,93493	6,93%	2,39723	21,08%	3,13853

Tabella 4.7: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il primo portafoglio analizzato per la forma funzionale in cui tutti i default avvengono durante il secondo anno.

La tabella 4.7 mostra i risultati relativi al caso in cui i default avvengono durante il secondo anno. I margini della tranche senior sono più alti rispetto ai casi precedenti: in questo caso i margini aumentano per bilanciare il wal che è diminuito. I margini e il wal delle due tranche subordinate invece diminuiscono rispetto al caso in cui i default avvengono in maniera decrescente col passare dei tre anni o in cui avvengono tutti al primo anno, ed è invece molto simile al caso uniforme.

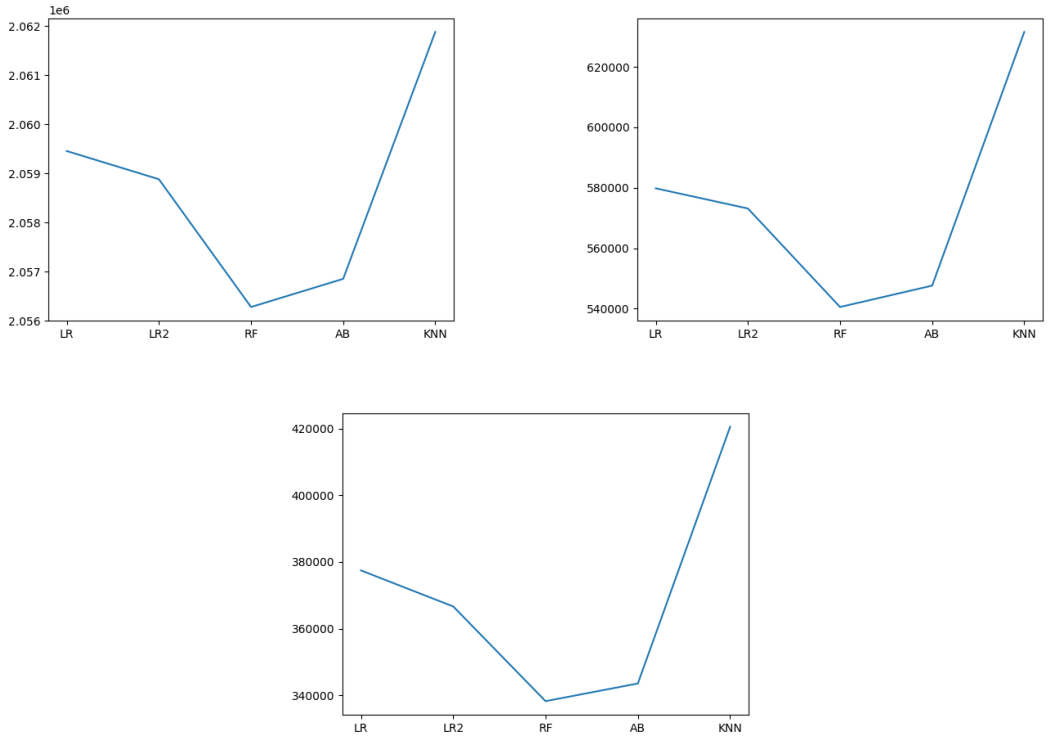

Figura 4.5: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui tutti i default avvengono durante il secondo anno.

Algoritmo	Senior	Mezzanine	Junior
LR	0.03815%	0.91924%	2.21385%
LR2	0.0102%	-0.2430%	-0.71594%
RF	-0.1161%	-5.9208%	-8.4091%
AB	-0.0882%	-4.6894%	-6.97496%
KNN	0.1559%	9.9341%	13.8861%

Tabella 4.8: Differenza percentuale tra l'interesse pagato e l'interesse medio.

I grafici in figura 4.5 mostrano, nuovamente, un profilo simile ai casi precedenti. Per avere una stima più quantitativa è possibile vedere i numeri presenti nella tabella 4.8. Si ha una variabilità inferiore rispetto agli altri casi considerati per la tranche senior, mentre per le tranche mezzanine e junior, che subiscono gli effetti dei default che avvengono proprio mentre le note di queste tranches vengono ripagate, si ha una variabilità in linea con quella osservata nei casi precedenti, sopratutto con il caso uniforme.

L'ultimo caso che viene considerato è un caso particolarmente accademico e poco realistico, in cui si suppone che i default avvengono tutti durante il terzo anno di vita del prestito. In realtà storicamente, chi riesce ad arrivare all'ultimo anno avendo fatto tutti i

pagamenti, è difficile che non riesca a ripagare le rate dell'ultimo anno, perciò è un caso da analizzare solo da un punto di vista accademico e non è qualcosa che si osserva realmente. La tabella 4.9 mostra i risultati relativi al caso in cui i default avvengono tutti durante

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	2,48%	3,49%	0,93228	10,38%	1,99825	15,9%	2,39093
LR2	2,22%	3,49%	0,93228	10,38%	1,99825	15,9%	2,39093
RF	1,04%	3,49%	0,93228	10,38%	1,99825	15,9%	2,39093
AB	1,3%	3,49%	0,93228	10,38%	1,99825	15,9%	2,39093
KNN	3,58%	3,49%	0,93228	10,38%	1,99825	15,9%	2,39093

Tabella 4.9: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il primo portafoglio analizzato per la forma funzionale in cui tutti i default avvengono durante il terzo anno.

l'ultimo anno. Si nota che i valori sono assolutamente identici e la probabilità di default non ha alcun impatto sui risultati. Questo comportamento anomalo si spiega guardando i wal della tabella, che sono tutti sotto i tre anni, per cui i detentori delle note vengono ripagati del tutto (interessi e valore nominale) prima che avvengano i primi default, rendendo questo portafoglio un'obbligazione standard con dei tassi di interesse più alti rispetto a quello *risk-free* di un'obbligazione.

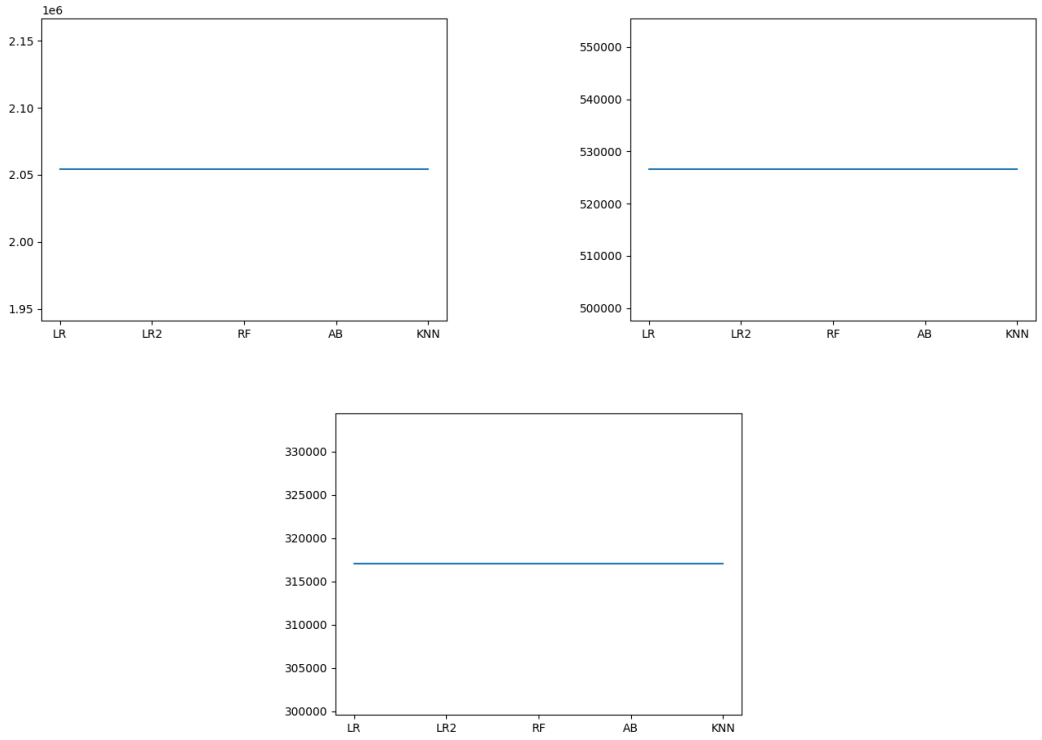

Figura 4.6: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui tutti i default avvengono durante il terzo anno.

Dopo aver analizzato le differenze quantitative tra i differenti algoritmi di Machine Learning nel ripagamento degli interessi, è possibile anche analizzare le differenze quantitative, per ogni singolo algoritmo, quando varia la forma funzionale considerata, per ognuna delle tre tranches.

	LR	LR2	RF	AB	KNN
decrescente	1.27156%	1.01258%	0.5143%	0.2765%	-3.7557%
uniforme	-0.9420%	-0.6489%	-0.2164%	-0.1699%	-0.2264%
primo anno	5.3709%	4.5250%	1.6760%	2.335%	11.1238%
secondo anno	-2.72104%	-2.32823%	-0.9179%	-1.152%	-3.38581%
terzo anno	-2.97942%	-2.5604%	-1.0560%	-1.2901%	-3.75579%

Tabella 4.10: Differenza percentuale tra l'interesse pagato per ogni casistica di forma funzionale e l'interesse medio di ogni modello per la tranche senior.

Nella tabella 4.10 sono riportati i risultati relativi alla tranche senior. Gli unici due casi in cui gli interessi pagati sono più alti della media sono quello decrescente e quello in cui i default avvengono durante il primo anno (tranne che nel caso del KNN, in cui

solamente quando i default avvengono al primo anno si ha un incremento positivo rispetto alla media). Si può notare quindi come, anche se la tranche senior si è rivelata essere la più stabile ai vari casi considerati, ha delle lievi differenze nel pagamento degli interessi.

	LR	LR2	RF	AB	KNN
decrescente	5.43141%	4.5908%	2.3185%	3.1468%	8.6444%
uniforme	-2.5845%	-1.9336%	-1.0357%	-0.9964%	-5.3523%
primo anno	16.6307%	14.3619%	6.3517%	7.3987%	26.9296%
secondo anno	-5.3929%	-4.6333%	-2.5580%	-2.9088%	-7.41006%
terzo anno	-14.0847%	-12.3858%	-5.07648%	-6.64027%	-22.8116%

Tabella 4.11: Differenza percentuale tra l'interesse pagato per ogni casistica di forma funzionale e l'interesse medio di ogni modello per la tranche mezzanine.

La tabella 4.11 mostra i risultati relativi alla tranche mezzanine. C'è molta più variabilita rispetto al caso precedente, dove più è alta la probabilità di default, maggiore è la variabilità corrispondente.

	LR	LR2	RF	AB	KNN
decrescente	5.33058%	4.5528%	2.14495%	2.20607%	9.0267%
uniforme	-1.8234%	-1.7586%	-0.2333%	-0.1699%	-3.4426%
primo anno	14.663%	13.0888%	4.91974%	2.335%	23.8355%
secondo anno	-1.1714%	-1.2580%	-0.2872%	-0.5973%	-2.7349%
terzo anno	-16.9994%	-14.6250%	-6.54411%	-8.2710%	-26.6847%

Tabella 4.12: Differenza percentuale tra l'interesse pagato per ogni casistica di forma funzionale e l'interesse medio di ogni modello per la tranche junior.

La tabella 4.12 mostra i risultati relativi alla variabilità della tranche junior. Si può notare che la variabilità è leggermente più bassa rispetto alla tranche mezzanine, che è quella che maggiormente subisce gli effetti della forma funzionale scelta nel prezzare la cartolarizzazione. L'unico caso in cui la tranche junior ha una variabilità maggiore della mezzanine è quello in cui i default avvengono tutti durante il terzo anno, in cui la tranche junior ha degli interessi molto più bassi rispetto alla media. Il fatto di non subire alcun default, infatti, impatta sul margine che diventa più basso rispetto a quello stimato negli altri casi e questo si ripercuote sugli interessi ripagati alla tranche junior, che diminuiscono.

Il secondo portafoglio che viene analizzato contiene 1000 prestiti che hanno una scadenza di cinque anni e quindi possono andare in default in un qualsiasi trimestre dei cinque anni considerati. Questo si traduce in wal un po' più alti rispetto al caso precedente perché la probabilità di default totale viene "spalmata" su più anni.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	3,5%	3%	1,68324	12,30%	4,62356	20,83%	5,7526
LR2	3,2%	3,01%	1,65888	13,69%	4,34708	23,71%	5,3606
RF	0,3%	3,06%	1,51182	12,9%	3,1359	27,3%	3,60427
AB	1,7%	3,04%	1,5698	17,33%	3,51658	26,25%	4,17038
KNN	1,8%	3,04%	1,57520	17,03%	3,55206	23,24%	4,2323

Tabella 4.13: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il secondo portafoglio analizzato utilizzando la forma funzionale decrescente.

La tabella 4.13 mostra i risultati relativi al caso in cui i default avvengono in maniera decrescente nel corso dei cinque anni: in questo caso nei primi anni ci saranno delle perdite monetarie più importanti che negli ultimi anni, e questo si ripercuote nel tempo che si impiega per ripagare le note. I due algoritmi di Regressione Logistica stimano una probabilità di default simile, un po' sopra il 3%.L'algoritmo di Random Forest stima una probabilità di default dello 0.3%, e gli algoritmi di Ada Boost e KNN una probabilità molto simile, sotto il 2%. La tranche senior è molto stabile ed è simile per tutti e cinque i modelli sia come margine che come wal, che si attesta intorno ad un anno e mezzo. La tranche mezzanine dipende di più dalla probabilità di default stimata: gli algoritmi di regressione logistica al primo e al secondo ordine stimano un wal di quattro anni e mezzo, mentre l'Ada Boost e il KNN di tre anni e mezzo. La Random Forest di circa tre anni. Il margine di Ada Boost e KNN è però più alto rispetto agli altri tre. Anche la tranhe junior ha lo stesso comportamento, con il modello di Random Forest che ha un wal molto basso, sotto i quattro anni, mentre gli algoritmi di Regressione Logistica al primo e al secondo ordine stimano un wal di pù di un anno superiore a quelli di Ada Boost e KNN.

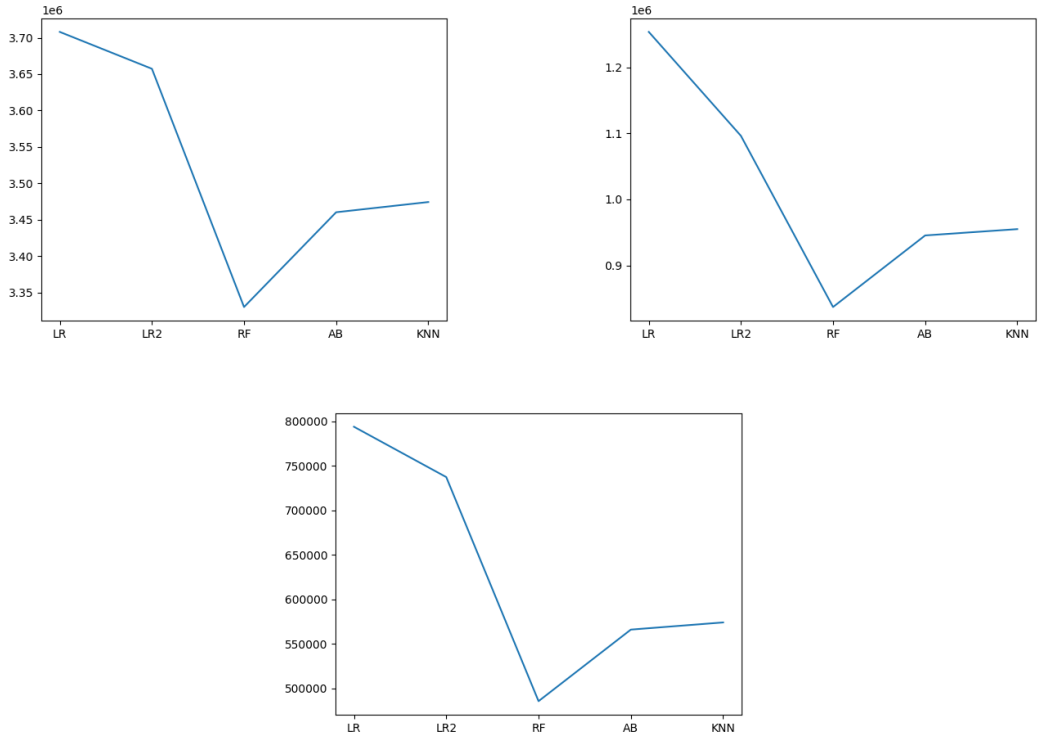

Figura 4.7: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale decrescente.

Algoritmo	Senior	Mezzanine	Junior
LR	5.16518%	23.2389%	25.70477%
LR2	3.72808%	7.7534%	16.7626%
RF	-5.5602%	-17.755%	-23.0577%
AB	-1.8652%	-7.0896%	-10.3414%
KNN	-1.4677%	-6.14765%	-9.0681%

Tabella 4.14: Differenza percentuale tra l'interesse pagato e l'interesse medio.

I grafici in figura 4.7 sono solo qualitativi e danno un'idea di come la probabilità di default stimata incide sugli interessi pagati sulle note: in tabella 4.24 si può vedere anche come incide in maniera quantitativa. La tranche senior ha delle differenze dell'ordine del 5%, mentre le due tranche subordinate hanno delle variazioni più significative e dipendono maggiormente dalla probabilità di default stimata. La tabella 4.15 riporta i risultati del caso in cui si utilizzi una probabilità di default uniforme nel corso dei cinque anni. In questo caso la tranche senior è sempre molto simile per tutti e cinque i modelli e simile anche al caso precedente: anche se cambia la forma funzionale dei default, il wal della tranche senior è sempre almeno di un anno e mezzo. I margini e il wal della tranche

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	3,5%	3,03%	1,5918	12,63%	3,7963	25,77%	4,74992
LR2	3,2%	3,04%	1,58227	14,15%	3,69152	22,45%	4,54378
RF	0,3%	3,065%	1,50779	13,42%	3,10573	29,1%	3,58003
AB	1,7%	3,05%	1,5404	15,33%	3,31803	25,79%	3,9034
KNN	1,8%	3,05%	1,5430	14,89%	3,3389	23,77%	3,94136

Tabella 4.15: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il secondo portafoglio analizzato per la forma funzionale uniforme.

mezzanine sono più simili rispetto al caso precedente: tutti i modelli hanno un wal di circa tre anni e il margine si aggira tra il 12 e il 15%. Per la tranche junior si notano delle differenze più significative: i due algoritmi di Regressione Logistica al primo e al secondo ordine stimano un wal di più di quattro anni, mentre gli altri tre sono tutti sotto i quattro anni. I margini sono tutti tra il 22 e il 29%.

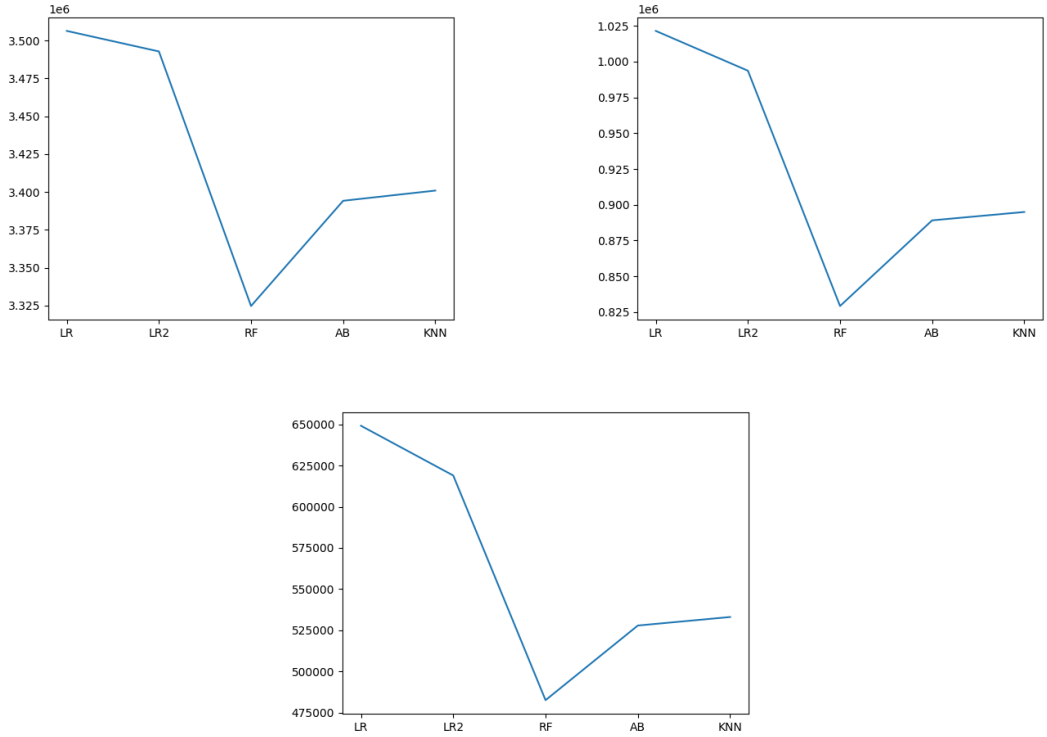

Figura 4.8: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale uniforme.

Algoritmo	Senior	Mezzanine	Junior
LR	2.4116%	10.3379%	15.4473%
LR2	2.0156%	7.3363%	10.0756%
RF	-2.8945%	-10.4087%	-14.1881%
AB	-0.8647%	-3.94748%	-6.12730%
KNN	-0.6680%	-3.31799%	-5.20753%

Tabella 4.16: Differenza percentuale tra l'interesse pagato e l'interesse medio.

In questo caso, le variazioni sono meno importanti rispetto al caso precedente, dove c'erano differenze anche del 20%. Anche in questo caso le tranches subordinate hanno una variabilità maggiore rispetto alla tranche senior.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	3,5%	2,915%	2,0486	14,16%	6,19234	21,75%	8,2932
LR2	3,2%	2,935%	1,9552	14,96%	5,5848	22,45%	6,84310
RF	0,3%	3,06%	1,52501	19,3%	3,18120	24,59%	3,64808
AB	1,7%	3,01%	1,67033	15,56%	3,9216	33,55%	4,6248
KNN	1,8%	3%	1,68343	14,11%	4,00689	27,53%	4,71497

Tabella 4.17: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il secondo portafoglio analizzato per la forma funzionale in cui tutti i default avvengono durante il primo anno.

La tabella 4.17 mostra i risultati relativi al caso in cui i default avvengano tutti nel corso del primo anno di vita dei prestiti. In questo caso la probabilità di default alta, stimata dai due algoritmi di Regressione Logistica, si ripercuote anche nella tranche senior, infatti il margine diminuisce ma il wal aumenta fino a diventare di circa 2 anni (a fronte dell'anno e mezzo dei casi precedenti). Anche l'Ada Boost e il KNN, che hanno una probabilità di default più bassa, hanno un wal più alto rispetto ai casi precedenti. La tranche mezzanine ha delle variazioni importanti tra i cinque modelli. Gli algoritmi di Regressione Logistica al primo e al secondo ordine hanno un wal di oltre 5 anni e mezzo, mentre gli algoritmi di Ada Boost e KNN si assestano attorno ai quattro anni. L'algoritmo di Random Forest, che stima una probabilità di default molto bassa, ha un wal di poco più di tre anni per questa tranche (anche se ha un margine molto alto). Anche la tranche junior è molto varia: l'algoritmo di Regressione Logistica un wal di più di 8 anni, mentre quello di Random Forest sotto i quattro anni.

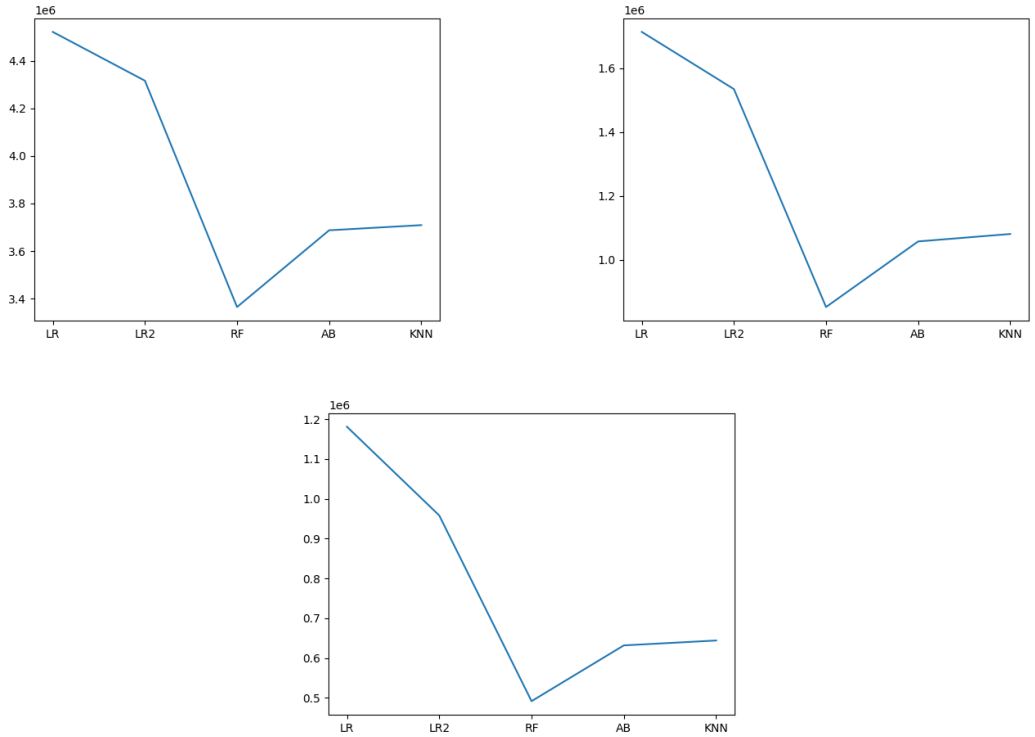

Figura 4.9: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui i default avvengono durante il primo anno.

Algoritmo	Senior	Mezzanine	Junior
LR	15.3469%	37.28976%	51.1520%
LR2	10.1138%	22.9564%	22.6358%
RF	-14.15768%	-31.6633%	-37.0737%
AB	-5.9274%	-15.2250%	-19.1414%
KNN	-5.37573%	-13.3578%	-17.5726%

Tabella 4.18: Differenza percentuale tra l'interesse pagato e l'interesse medio.

La tabella 4.24 mostra i risultati relativi al caso in cui i default avvengono nel corso del primo anno. In questo caso le variazioni sono molto grandi e si possono notare già a partire dalla tranche senior, dove i modelli che stimano la probabilità di default più alta ricevono degli interessi che sono circa il 15% in più rispetto alla media. Infatti in questo caso nel corso del primo anno si perdono circa 3.5 milioni, ed è necessario più tempo per riacquisire il denaro perso e questo si traduce in interessi ripagati per più mesi. Le due tranches subordinate hanno delle variazioni ancora più grandi, che toccano anche il 50% in più nel caso della tranche junior.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	3,5%	3,01%	1,66169	11,74%	4,85611	22,44%	5,97919
LR2	3,2%	3,02%	1,6353	11,6%	4,54413	27,77%	5,54376
RF	0,3%	3,065%	1,50851	12,8%	3,14296	26,86%	3,6106
AB	1,7%	3,05%	1,5533	16,65%	3,5673	22,6%	4,24812
KNN	1,8%	3,045%	1,5576	15,76%	3,60633	20,26%	4,31512

Tabella 4.19: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il secondo portafoglio analizzato per la forma funzionale in cui tutti i default avvengono durante il secondo anno.

La tabella 4.19 riporta i risultati relativi al caso, poco realistico, che tutti i default avvengano nel corso del secondo anno. Si può notare che la tranche senior è simile ai due casi iniziali (decrescente e uniforme) sia come margini che come wal. Anche la tranche mezzanine ha dei valori analoghi a quelli del caso uniforme, e anche per la tranche junior i due algoritmi di Regressione Logistica al primo e al secondo ordine hanno dei valori di wal più alti rispetto agli altri tre modelli.

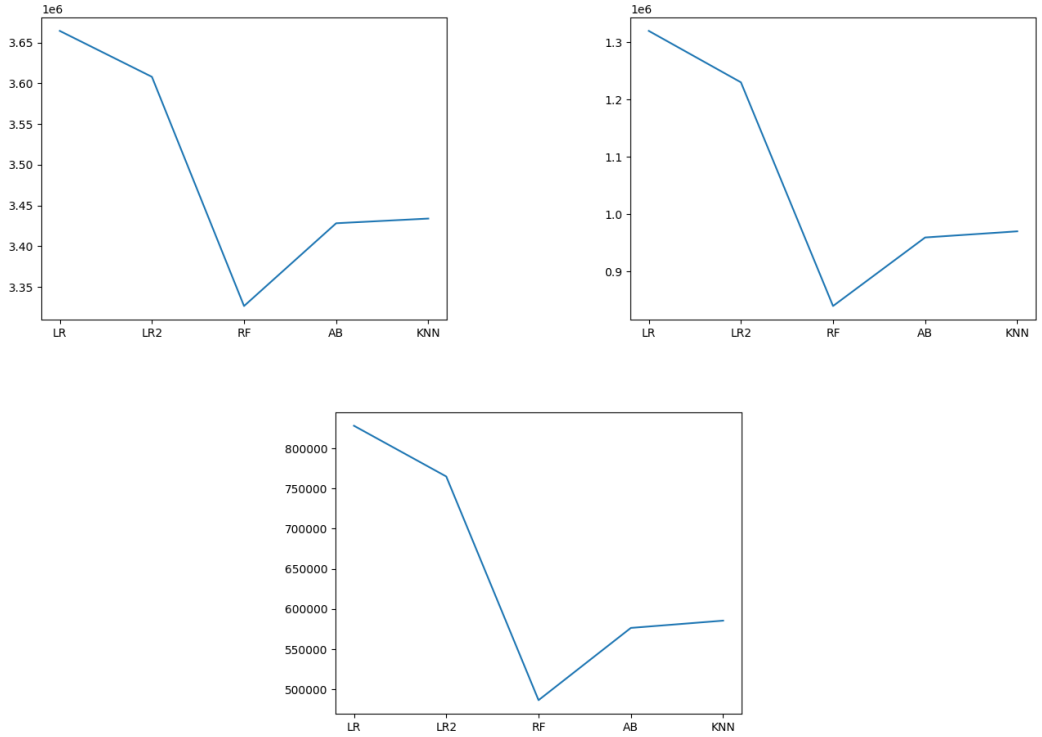

Figura 4.10: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui tutti i default avvengono durante il secondo anno.

Algoritmo	Senior	Mezzanine	Junior
LR	4.92955%	24.0954%	27.72162%
LR2	3.3152%	15.653%	17.9874%
RF	-4.74327%	-21.0918%	-24.9455%
AB	-1.8332%	-9.8321%	-11.0794%
KNN	-1.66830%	-8.8244%	-9.68411%

Tabella 4.20: Differenza percentuale tra l'interesse pagato e l'interesse medio.

In questo caso non si hanno le stesse variazioni del caso precedente, la tranche senior è abbastanza stabile, mentre le due tranches subordinate, come si è visto anche nei casi precedenti, hanno delle variazioni più importanti che toccano il 27% in più.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	3,5%	3,065%	1,50724	10,81%	3,63608	21,97%	4,75315
LR2	3,2%	3,065%	1,50638	11,88%	3,53456	22,85%	4,5323
RF	0,3%	3,065%	1,50215	13,8%	3,08525	23,58%	3,57499
AB	1,7%	3,065%	1,5031	11,59%	3,2314	27,18%	3,8807
KNN	1,8%	3,07%	1,5033	11,37%	3,24923	25,47%	3,9089

Tabella 4.21: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il secondo portafoglio analizzato per la forma funzionale in cui tutti i default avvengono durante il terzo anno.

La tabella 4.21 mostra i risultati relativi al caso in cui tutti i default avvengano nel corso del terzo anno. La tranche senior non è toccata dalla forma funzionale scelta, però le due tranche subordinate hanno un wal più basso rispetto al caso precedente, perché le quote capitali più alte vengono ripagate prima che avvengano i primi default nel corso del terzo anno e questo fa diminuire il wal.

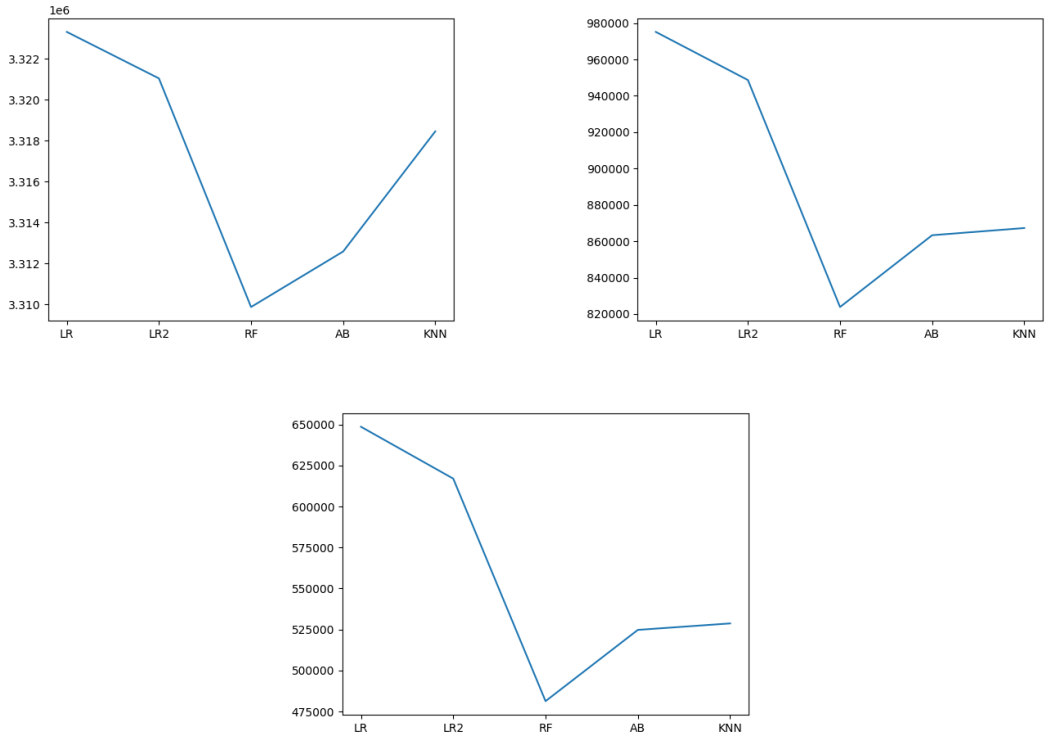

Figura 4.11: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui tutti i default avvengono durante il terzo anno.

Algoritmo	Senior	Mezzanine	Junior
LR	0.1886%	8.87566%	15.8054%
LR2	0.12025%	5.91959%	10.15959%
RF	-0.2164%	-8.0180%	-14.0612%
AB	-0.1347%	-3.60891%	-6.3032%
KNN	0.04223%	-3.16833%	-5.60049%

Tabella 4.22: Differenza percentuale tra l'interesse pagato e l'interesse medio.

Nel caso in cui i default avvengano nel corso del terzo anno le variazioni diventano meno evidenti. La tranche senior è molto stabile e le differenze sono trascurabili, mentre le due tranche subordinate non hanno le variazioni molto grandi che si sono osservate nei casi precedenti.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	3,5%	3,065%	1,5019	14,1%	3,0701	17,13%	3,54965
LR2	3,2%	3,065%	1,5019	14,1%	3,0701	17,22%	3,54603
RF	0,3%	3,065%	1,5019	14,1%	3,0701	18,2%	3,5113
AB	1,7%	3,065%	1,5019	14,1%	3,0701	17,7%	3,52803
KNN	1,8%	3,065%	1,5019	14,1%	3,0701	17,67%	3,5298

Tabella 4.23: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il secondo portafoglio analizzato per la forma funzionale in cui tutti i default avvengono durante il quarto anno.

La tabella 4.23 mostra i risultati del caso in cui i default avvengano nel corso del quarto anno. Si può notare che la tranche senior e la tranche mezzanine sono identiche per tutti i modelli, perché i default avvengono quando le due tranche sono già state ripagate e perciò la probabilità di default non è più rilevante. Diventa rilevante solo per la tranche junior, in cui si hanno delle differenze, ma molto meno marcate rispetto ai casi precedenti.

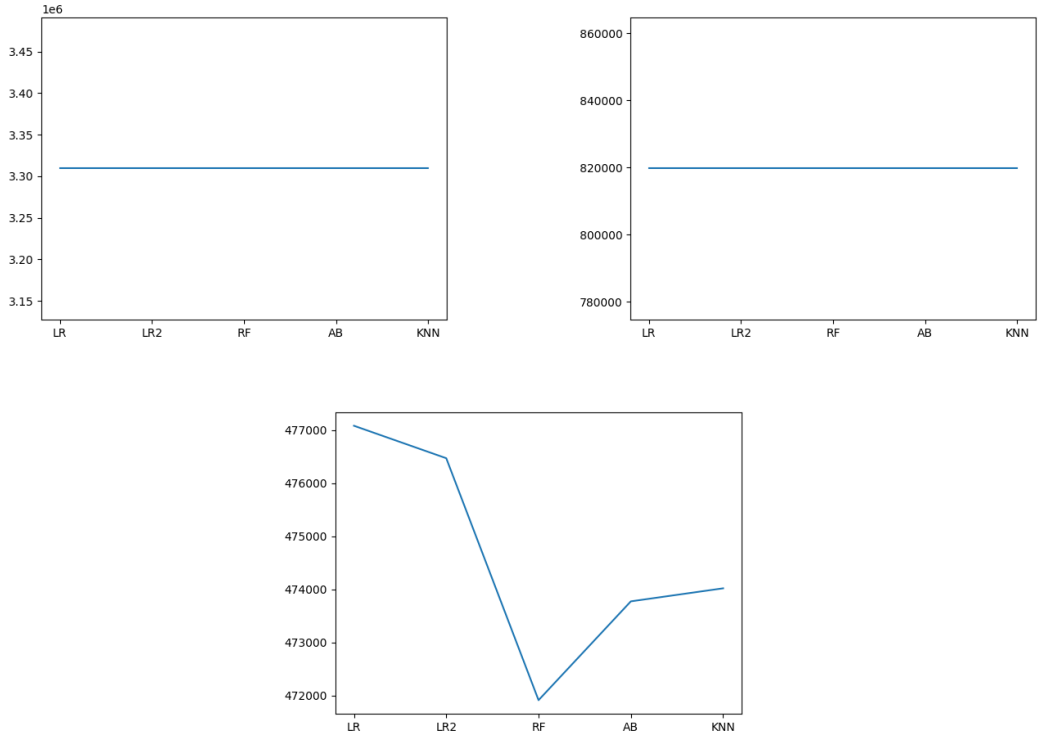

Figura 4.12: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui tutti i default avvengono durante il quarto anno.

Algoritmo	Junior
LR	0.51245%
LR2	0.38352%
RF	-0.5776%
AB	-0.18510%
KNN	-0.1332%

Tabella 4.24: Differenza percentuale tra l'interesse pagato e l'interesse medio.

Nel caso in cui i default avvengano tutti nel corso del quinto anno invece non ci sono proprio differenze perché le tranches vengono tutte ripagate prima che avvengano i default.

Algoritmo	PD	MS	WS	MM	WM	MJ	WJ
LR	3,5%	3,065%	1,5019	14,1%	3,0701	18,3%	3,5077
LR2	3,2%	3,065%	1,5019	14,1%	3,0701	18,3%	3,5077
RF	0,3%	3,065%	1,5019	14,1%	3,0701	18,3%	3,5077
AB	1,7%	3,065%	1,5019	14,1%	3,0701	18,3%	3,5077
KNN	1,8%	3,065%	1,5019	14,1%	3,0701	18,3%	3,5077

Tabella 4.25: Risultati relativi al margine di ogni tranche e al *weighted average life* corrispondente, per il secondo portafoglio analizzato per la forma funzionale in cui tutti i default avvengono durante il quinto anno.

La tabella 4.25 mostra i risultati relativi al caso in cui i default avvengano tutti nel corso del quinto anno. È un caso irrealistico ma interessante da analizzare. Si può notare infatti che le tre tranche per tutti i modelli sono assolutamente identiche, perché i default avvengono quando le note sono state tutte ripagate, per cui la probabilità di default stimata non è più rilevante.

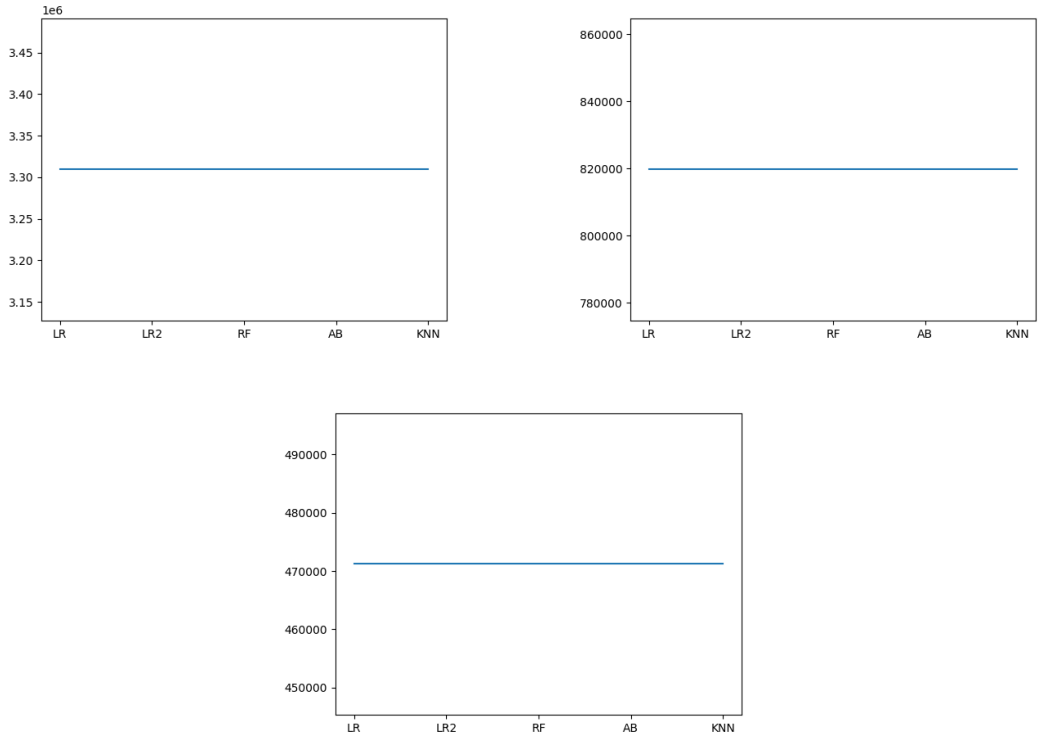

Figura 4.13: Interessi pagati ai detentori del titolo cartolarizzato per ognuno dei cinque modelli per la forma funzionale in cui tutti i default avvengono durante il cinque anno.

Nelle tabelle sottostanti sono invece riportati i risultati relativi alle differenze che si notano se si tiene fissa la probabilità di default e si varia solo la forma funzionale scelta. Per analizzare bene uno strumento del genere non è infatti sufficiente stimare una probabilità di default ma anche stimare bene i momenti temporali in cui avvengono i default, perché il fatto che avvengano all'inizio o alla fine ha un impatto significativo.

	LR	LR2	RF	AB	KNN
decrescente	2.4249%	2.3499%	0.14979%	1.3394%	1.5205%
uniforme	-3.1449%	-2.2544%	-0.0051%	-0.5917%	-0.6190%
primo anno	24.882%	20.7801%	1.1949%	7.9893%	8.3777%
secondo anno	1.2182%	0.9678%	0.0502%	0.40311%	0.34510%
terzo anno	-8.2021%	-7.0619%	-0.4516%	-2.9824%	-3.0296%
quarto anno	-8.5893%	-7.3907%	-0.46908%	-3.0788%	-3.29735%
quinto anno	-8.5893%	-7.3907%	-0.46908%	-3.0788%	-3.29735%

Tabella 4.26: Differenza percentuale tra l'interesse pagato per ogni casistica di forma funzionale e l'interesse medio di ogni modello per la tranche senior.

La tabella 4.26 mostra i risultati relativi alla tranche senior, che è quella più stabile. In realtà delle differenze ci sono e si può notare ad esempio che nel caso in cui i default avvengano tutti nel corso del primo anno, allora gli interessi ripagati sono più alti rispetto alla media. Si può notare anche che le variazioni più importanti sono relative ai modelli che stimano una probabilità di default più alta, in cui l'entità della perdita di denaro ha un impatto maggiore sul ripagamento delle note e questo influenza gli interessi che vengono ripagati ai dententori del titolo.

	LR	LR2	RF	AB	KNN
decrescente	10.78837%	9.57420%	0.6260%	4.1451%	4.32477%
uniforme	-9.7616%	-7.5597%	-0.2788%	-2.0500%	-2.2294%
primo anno	51.3417%	42.7316%	2.5250%	16.5244%	18.0989%
secondo anno	16.6341%	14.4657%	0.9362%	5.6698%	5.96062%
terzo anno	-13.8455%	-11.7385%	-0.9384%	-4.8929%	-5.25411%
quarto anno	-27.5785%	-23.7366%	-1.4349%	-9.6981%	-10.4503%
quinto anno	-27.5785%	-23.7366%	-1.4349%	-9.6981%	-10.4503%

Tabella 4.27: Differenza percentuale tra l'interesse pagato per ogni casistica di forma funzionale e l'interesse medio di ogni modello per la tranche mezzanine.

La tabella 4.27 mostra i risultati relativi alla tranche mezzanine. Anche in questo caso più alta è la probabilità di default stimata, maggiori sono le variazioni. In questo caso sono molto maggiori rispetto al caso della tranche senior.

	LR	LR2	RF	AB	KNN
decrescente	10.0559%	11.1379%	0.8907%	5.0684%	5.47809%
uniforme	-9.9956%	-6.7030%	0.19634%	-2.042%	-2.0875%
primo anno	63.7348%	44.4246%	2.09009%	17.2391%	18.299%
secondo anno	14.8046%	15.2995%	1.04064%	6.9834%	7.5581%
terzo anno	-10.0750%	-7.0026%	-0.0540%	-2.6143%	-2.8807%
quarto anno	-33.8547%	-28.181%	-2.0086%	-12.078%	-12.9236%
quinto anno	-34.66987%	-28.9751%	-2.1550%	-12.5554%	-13.4438%

Tabella 4.28: Differenza percentuale tra l'interesse pagato per ogni casistica di forma funzionale e l'interesse medio di ogni modello per la tranche junior.

La tabella 4.28 mostra i risultati relativi alla tranche junior. Le variazioni sono anche in questo caso molto maggiori nel caso dei modelli che stimano una probabilità di default più alta (I due algoritmi di Regressione Logistica) e molto minore nel caso dell'algoritmo di Random Forest, che stima una probabilità di default più bassa. Le variazioni sono simili a quelle della tranche mezzanine e più alte rispetto a quelle della tranche senior.
Si può concludere quindi che non è sufficiente assegnare una probabilità di default al portafoglio cartolarizzato per avere un'idea chiara del titolo su cui si sta facendo un'investimento, ma è necessario anche capire come vengono modellizzati i default nel corso

degli anni e la durata dei prestiti usati come sottostante. La variabilità dei risultati ottenuti giustifica anche l'interesse per le analisi svolte e fa riflettere sul fatto che rispetto ad altri titoli finanziari che hanno delle formule e delle analisi molto più semplici e lineari, l'analisi di una cartolarizzazione necessita di indagini più approfondite e non è immediato riconoscere quali siano i fattori che maggiormente impattano il ripagamento delle note e in che modo ne impattino i risultati.

Conclusioni

Questa tesi introduce lo strumento finanzario delle cartolarizzazioni e una possibile strada alternativa per calcolarne il prezzo di mercato. Utilizzando alcune tecniche di Machine Learning infatti si cerca di stimare la probabilità di default individuale di ogni cliente, sfruttando le informazioni economiche e sociali raccolte per ogni cliente presente all'interno del dataset. Poiché la probabilità marginale di default è funzione di queste covariate osservabili, è possibile modellizzare un portafoglio cartolarizzato con un modello misto scambiabile di Bernoulli e avere una stima delle perdite a livello aggregato di portafoglio. A partire dalla stima del numero dei default per ogni portafoglio, è possibile calcolare il valore e il prezzo di questa cartolarizzazione e l'entità delle cedole da assegnare ai detentori del titolo. Per un'analisi più completa si procede poi ad analizzare differenti forme funzionali per modellizzare la realizzazione dei default, ipotizzando che avvengano in forma decrescente nel corso degli anni di vita del prestito, in forma uniforme e analizzando dei casi limite. La variabilità dei risultati che si ottengono mostrerà che non è sufficiente trovare una probabilità di default o stimare bene gli istanti temporali in cui avvengono i default, ma fare delle analisi congiunte che tengano conto di entrambi i fattori per avere un modello completo e giustifica l'interesse per quest'analisi. L'utilizzo di tecniche di Machine Learning, che sempre più spesso trova applicazioni nel mondo della finanza, si è mostrato utile e innovativo per affrontare il problema del pricing di una cartolarizzazione, che, rispetto ad altri strumenti finanziari, ha una complessità che giustifica il ricorso a questo tipo di tecniche di apprendimento e classificazione.

Capitolo 5

Derivazione delle formule relative al modello scambiabile di Bernoulli

In questa sezione vengono svolti i calcoli relativi alle formule riportate in (1.10), (1.11), (1.12), (1.13) ed (1.16).

$$p = \mathbb{E}(Y_i) = \mathbb{E}(\mathbb{E}(Y_i|Q)) = \mathbb{E}(1\cdot(\mathbb{P}(Y_i = 1|Q))+0\cdot(\mathbb{P}(Y_i = 0|Q))) = \mathbb{E}((\mathbb{P}(Y_i = 1|Q)) = \mathbb{E}(Q),$$

$$Var(Y_i) = \mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2 = \mathbb{E}(\mathbb{E}(Y_i^2|Q)) - p^2 = \mathbb{E}(1^2 \cdot (\mathbb{P}(Y_i = 1|Q)) + 0^2 \cdot (\mathbb{P}(Y_i = 0|Q))) = \mathbb{E}(Q) - p^2 = p - p^2 = p(1-p),$$

$$Cov(Y_i, Y_j) = \mathbb{E}(Y_iY_j) - \mathbf{E}(Y_i)\mathbf{E}(Y_j) = \pi_2 - p^2 \quad i \neq j,$$

$$\rho_Y = \rho(Y_i, Y_j) = \frac{Cov(Y_i, Y_j)}{\sqrt{Var(Y_i)Var(Y_j)}} = \frac{\pi_2 - p^2}{\sqrt{(p^2(1-p)^2)}} = \frac{\pi_2 - p^2}{p(1-p)}, \quad i \neq j.$$

$$\pi_k = \mathbb{E}(Y_{i1} \cdot \cdots \cdot Y_{ik}) = \mathbb{E}(\mathbb{E}(Y_{i1} \cdot \cdots \cdot Y_{ik})|Q) = \mathbb{E}(\mathbb{E}(Y_{i1}|Q) \cdot \cdots \cdot \mathbb{E}(Y_{ik}|Q)) = \mathbb{E}(\mathbb{P}(Y_{i1} = 1|Q) \cdot \ldots \mathbb{P}(Y_{ik} = 1|Q)) = \mathbb{E}(Q \cdot \cdots \cdot Q) = \mathbb{E}(Q_h^k)$$

Capitolo 6

Distribuzioni Beta e Beta Binomiale

Sia Q una distribuzione di probabilità Beta di parametri a e b, ovvero $Q \sim Beta(a, b)$. La densità di una distribuzione Beta è data da:

$$f_Q(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \qquad a, b > 0, \quad x \in [0,1] \tag{6.1}$$

dove

$$B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}dy \tag{6.2}$$

Si ha

$$\mu_Q = \frac{a}{a+b}, \qquad Var(Q) = \frac{ab}{(a+b)^2(a+b+1)}, \qquad \rho = \frac{1}{a+b+1}$$

A partire dal primo momento della distribuzione è possibile calcolare i momenti successivi tramite la formula iterativa

$$\mu_{k+1} = \frac{a+k}{a+b+k}\mu_k \tag{6.3}$$

La distribuzione beta binomiale è una distribuzione di probabilità discreta e dipende da tre parametri a, b, d e descrive la distribuzione del numero di successi su un numero totale d di prove indipendenti, in cui la probabilità di successo non è un parametro noto a priori, ma è distribuita come una variabile aleatoria Beta. Sia $S \sim B - Binom(a, b, d)$, la sua funzione di probabilità è data da:

$$f_S(k) = \mathbb{P}[S = k] = \binom{d}{k} \frac{B(a+k, b+d-k)}{B(a,b)} \tag{6.4}$$

Capitolo 7

Algoritmi di Machine Learning

In questa sezione vengono descritti gli algoritmi di Machine Learning utilizzati per la stima delle probabilità individuali di default. In particolare, verranno descritti la Regressione Logistica, la Regressione Logistica al secondo ordine, la Random Forest, L'Ada Boost e il KNN.

7.0.1 Regressione Logistica

La regressione logistica (LR) modella la probabilità di appartenere ad una classe, in un problema in cui vi sono solamente due classi possibili. È un'estensione del modello di regressione lineare per i problemi di classificazione. Nel caso della regressione lineare, la relazione tra l'output y e le covariate x_i viene modellata come

$$\tilde{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i} + \cdots + \beta_p x_p^{(i)} \tag{7.1}$$

La regressione logistica sfrutta la funzione logistica per comprimere l'output del modello lineare 7.1 in modo tale da essere compreso tra 0 e 1. La funzione logistica ha la forma

$$l(x) = \frac{1}{1 + e^{-x}}$$

e quindi, mettendo insieme le due cose, si ottiene

$$\mathbb{P}[y^{(i)} = 1] = p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i} + \cdots + \beta_p x_p^{(i)})}} \tag{7.2}$$

Risolvendo l'equazione rispetto all'esponenziale, si ottiene:

$$e^{(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i} + \cdots + \beta_p x_p^{(i)})} = \frac{p_i}{1 - p_i}$$

e, applicando la trasformazione logaritmica ad ambo i membri, si ottiene

$$\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i} + \cdots + \beta_p x_p^{(i)} = \log\left(\frac{p_i}{1 - p_i}\right) = \log\left(\frac{\mathbb{P}[y^{(i)} = 1]}{\mathbb{P}[y^{(i)} = 0]}\right)$$

Il rapporto $\frac{p_i}{1-p_i}$ è chiamato *odds-ratio*, e la funzione $\log(\frac{p_i}{1-p_i})$ è chiamata funzione *logit*. Per calcolare i coefficienti $\tilde{\beta}$ si utilizza il metodo della massima verosimiglianza. La variabile obiettivo Y_i può assumere valori nell'insieme $\{0, 1\}$ e quindi può essere considerata la realizzazione di una variabile di Bernoulli di parametro p_i.

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

dove i p_i dipendono dalle covariate x_i e dai coefficienti β Poiché le osservazioni sono indipendenti, se si assume che gli errori siano indipendenti, allora la funzione di verosimiglianza è il prodotto delle probabilità individuali:

$$L = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\beta^T \mathbf{x}_i}}\right)^{y_i} \left(\frac{e^{-\beta^T \mathbf{x}_i}}{1 + e^{-\beta^T \mathbf{x}_i}}\right)^{1-y_i} \tag{7.3}$$

Avendo scritto in forma matriciale $\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i} + \cdots + \beta_p x_p^{(i)} = \beta^T \mathbf{x}_i$. L'equazione (7.3) può essere semplificata applicando la funzione logaritmica e si ottiene quindi

$$\begin{aligned} \mathcal{L} = \log L &= \sum_{i=1}^{n} \log\left(\frac{1}{1 + e^{-\beta^T \mathbf{x}_i}}\right) y_i + \sum_{i=1}^{n} \log\left(\frac{e^{-\beta^T \mathbf{x}_i}}{1 + e^{-\beta^T \mathbf{x}_i}}\right) \\ &= \sum_{i=1}^{n} y_i \beta^T \mathbf{x}_i - log(1 + e^{\beta^T \mathbf{x}_i}) \end{aligned} \tag{7.4}$$

Infine, per trovare i coefficienti β che massimizzano la funzione di verosimiglianza, è necessario risolvere il problema

$$\beta = arg \max_{\beta} \mathcal{L} \tag{7.5}$$

7.0.2 Regressione Logistica al secondo ordine

L'algoritmo di regressione logistica al secondo ordine estende l'algoritmo classico di regressione logistica considerando anche dei termini quadratici e le i coefficienti di interazione. In questo caso però la funzione logistica non riceve in input una relazione lineare, ma una relazione quadratica della forma

$$\tilde{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots \beta_p x_p^{(i)} + \beta_{p+1}(x_1^{(i)})^2 + \ldots \beta_{p+p}(x_p^{(i)})^2 + \sum_{k=1}^{p} \sum_{j \neq k} \beta_{2p+k+j} x_k^{(i)} x_j^{(i)} \tag{7.6}$$

. E i coefficienti β vengono stimati allo stesso modo che nel caso della regressione logistica classica, utilizzando il metodo della massima verosimiglianza. La regressione logistica al secondo ordine riesce a cogliere relazioni più complesse tra le variabili e spiegare meglio la variabile risposta, rispetto alla classica regressione logistica.

7.0.3 Random Forest

La *Random Forest* è un algoritmo di classificazione che si basa sull'aggregazione di più alberi decisionali. Un albero decisionale è un predittore

$$h : \mathcal{X} \to \{0, 1\}$$

che predice l'etichetta associata ad un insieme di istanze $\mathbf{x_i}$ passando dal nodo radice al nodo foglia. Ad ogni nodo del percorso, il nodo successivo viene scelto tramite una separazione dello spazio. La separazione è effettuata su una delle covariate di $\mathbf{x}_i$ o attraverso un insieme di regole predefinite. Ogni foglia contiene un'etichetta specifica. Gli alberi decisionali hanno molti vantaggi, infatti sono facilmente interpretabili, possono essere rappresentati graficamente e possono gestire anche variabili categoriche. Gli alberi decisionali però hanno generalmente una varianza molto alta, il che vuol dire che il modello può variare molto se variano, anche di poco, i dati iniziali. Per ridurre la varianza di un albero decisionale esistono molte tecniche, tra cui il *bagging*. L'idea dietro questa procedura è che si hanno n osservazioni indipendenti $Z_1, Z_2, \dots Z_n$ e ogni osservazione ha varianza σ^2, la varianza della media $\bar{Z}$ delle osservazioni è $\frac{\sigma^2}{n}$. Fare una media sulle osservazioni, perciò, riduce la varianza. Prendendo ripetutamente dei campioni dal dataset di *training*, si possono generare B dataset di *training* e si possono quindi costruire B alberi decisionali, ognuno dei quali produrrà una predizione $\tilde{h}_b(\mathbf{x}_i)$. Avendo a disposizione le predizioni dei B alberi, si può poi calcolare la media di queste predizioni

$$\tilde{h}(\mathbf{x}_i) = \frac{1}{B}\sum_{b=1}^{B} \tilde{h}_b(\mathbf{x}_i)$$

e quindi decidere la classe dell'istanza $\mathbf{x}_i$ attraverso un voto di "maggioranza". La *Random Forest* è un miglioramento rispetto al *bagging* perché decorrela gli alberi generati. Nella Random Forest vengono costruiti più alberi decisionali su dei campioni ripetuti, ma quando bisogna effettuare una separazione in un albero, si scelgono m predittori dall'insieme dei p predittori e la separazione può essere effettuata solo su uno di questi m predittori. Ad ogni taglio viene generato un nuovo insieme di m predittori. Solitamente si sceglie $m = \sqrt{p}$. La motivazione dietro questa procedura è che esistono dei predittori più importanti di altri, i B alberi considereranno questi predittori prima degli altri, e quindi gli alberi saranno tutti simili tra di loro e quindi saranno molto correlati tra di loro, e fare una media tra valori correlati tra di loro non diminuisce la varianza. La Random Forest, forzando gli alberi a separarsi solo su un sottoinsieme casuale di covariate (dove potrebbero non comparire i predittori più importanti) permette la costruzione di alberi decorrelati, la cui media è più affidabile e meno variabile.

7.0.4 Ada Boost

L'algoritmo *Ada Boost* (*Adaptive Boosting*) combina più algoritmi di apprendimento "deboli" per costruire un algoritmo di apprendimento "forte". Nel dettaglio, l'AdaBoost riceve in input un insieme di N dati $S = \{(x_1, y_1), \dots, (x_N, y_N)\}$, dove $y_i \in \{0, 1\}$. L'algoritmo procede in maniera iterativa, ed ad ogni iterazione vengono assegnati dei pesi w_i ai dati. Un peso maggiore indica che il dato avrà un'influenza maggiore nel modello di apprendimento. Allo step iniziale, i dati di training hanno tutti lo stesso peso:

$$w_{i,0} = \frac{1}{N}$$

I pesi sono sempre compresi tra 0 e 1 e i dati pesati sommano sempre ad 1. Quindi questi dati vengono passati all'algoritmo di apprendimento debole, che all'iterazione k restituisce

un modello di classificazione $h_k(x)$ e si definisce l'errore associato a questo modello come

$$\epsilon_k = \sum_{i=1}^{N} w_{i,k} 1_{[h_k(x_i) \neq y_i]}$$

e si calcola il parametro

$$\alpha = \frac{1}{2} \log \left(\frac{1 - \epsilon_k}{\epsilon_k} \right)$$

Quindi viene aggiornato il valore dei pesi attraverso la formula

$$w_{i,k} = w_{i,k-1} \cdot e^{+\alpha}$$

nel caso in cui il campione x_i è stato classificato bene dall'algoritmo di apprendimento, e viene usata la formula

$$w_{i,k} = w_{i,k-1} \cdot e^{-\alpha}$$

nel caso in cui il campione x_i sia stato classificato male. Questo perché se un campione viene classificato bene, si può diminuire il peso associato al campione, perché il modello riesce già a classificarlo bene. Se un campione viene classificato in maniera errata, invece, il suo peso aumenta in quanto l'algoritmo si dovrà focalizzare di più su questo dato problematico e dovrà cercare di classificarlo in maniera corretta. Questo algoritmo viene iterato fino a quando non viene raggiunto un criterio di stop.

7.0.5 K Nearest Neighbors

L'algoritmo *K Nearest Neighbors* si basa sulla densità spaziale dei dati. Cerca di predirre la classe a cui appartiene un dato x_i in base alla classe a cui appartengono i k dati più vicini al dato che si vuole classificare. La motivazione dietro questo algoritmo è l'assunzione che dati vicini nello spazio $\mathcal{X}$ dei predittori, apparteranno alla stessa classe. Per capire quali dati siano vicini tra di loro è necessario assegnare una metrica allo spazio dei predittori. La metrica è una funzione

$$\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

e viene scelta in base alle caratteristiche dello spazio dei predittori, in quanto i predittori possono essere categorici o continui. In definitiva, dato un numero intero positivo k e dato un punto x^*, l'algoritmo KNN identifica l'insieme $\mathcal{N}^*$ formato dai k punti del training set più vicini al dato x^* e stima la probabilità di appartenere ad ognuna delle due classi $y, y = 0{,}1$ come la frazione dei punti che appartengono a quella classe, sul totale dei punti considerati, in formule:

$$\mathbb{P}[Y_i = y | X = x^*] = \frac{1}{k} \sum_{i \in \mathcal{N}^*} 1_{[y_i = y]} \tag{7.7}$$

e assegna il punto x^* alla classe che ha la probabilità più alta.

7.0.6 DBSCAN

Il DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) è un algoritmo di clustering che raggruppa punti che si trovano vicini tra di loro a formare delle regioni dense nello spazio. Non è necessario specificare in anticipo il numero di cluster da identificare, poiché vengono trovati dall'algoritmo tramite i due parametri che vengono forniti:

1. il raggio ϵ, che indica la distanza massima a cui possono trovarsi due punti per potersi considerare vicini
2. *MinPoints*, il numero minimo di punti che devono trovarsi vicini tra di loro per poter formare un cluster.

Ogni punto nello spazio è classificato come:

1. *core point* p se almeno *MinPoints* punti si trovano a distanza minore di ϵ da p.
2. un punto direttamente raggiungibile q_d se si trova ad una distanza minore di ϵ da p.
3. un punto raggiungibile q_r se esiste un cammino formato da punti $q_1, q_2, \ldots, q_n$ tale che $\mathrm{q}_{i+1}\mathrm{q}_{i+1}$ è direttamente raggiungibile da q_i.
4. un punto *noise point* se è distante più di ϵ da qualsiasi altro punto del dataset.

Capitolo 8

Divergenza di Kullback-Leibler

La divergenza di Kullback-Leibler è una misura non simmetrica della distanza tra due distribuzioni di probabilità P e Q, in particolare misura la perdita di informazione che si realizza quando la distribuzione Q viene utilizzata per approssimare P. Questa misura non è una metrica vera e propria in quanto non rispetta le proprietà matematiche delle metriche. Per una distribuzione discreta è definita come:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \cdot ln\left(\frac{P(x)}{Q(x)}\right)$$

Capitolo 9

Metriche

Valutare la bontà e la perfomance di un algoritmo di Machine Learning è necessario per poter fare confronti tra diversi algoritmi e avere una stima della qualità dei dati di output. Esistono molti metodi per misurare la bontà di un algoritmo di Machine learning, con finalità differenti, che possono essere combinati insieme per avere una panoramica più ampia sulle perfomance globali del modello. La maggior parte delle metriche utilizzate per misurare la perfomance di un algoritmo di classificazione sfrutta le informazioni contenute all'interno della *confusion matrix*, una matrice associata ad ogni algoritmo di classificazione, che sulla diagonale contiene, per le due classi, i dati correttamente classificati, e nelle due celle extradiagonali contiene i falsi positivi (errore del I tipo) e i falsi negativi (errore del II tipo).

TP	FP
FN	TN

Le metriche che sono state utilizzate per valutare gli algoritmi di Machine Learning scelti sono la precisione, il recupero, l'*f1-score* e l'AUC.
La precisione è definita come:

$$p = \frac{TP}{TP + FP} \tag{9.1}$$

Il recupero è invece definito come:

$$r = \frac{TP}{TP + FN} \tag{9.2}$$

L'F1-score è invece la media armonica tra queste due metriche, ed è quindi definito come:

$$f_1 = 2 * \frac{p \cdot r}{p + r} \tag{9.3}$$

È possibile calcolare questi tre valori per ognuna delle due classi (andando a considerarla come classe positiva) e poi farne una media pesata sul numero di osservazioni che cadono all'interno di ogni classe e avere quindi delle metriche "sbilanciate", che rispecchino lo sbilanciamento presente all'interno del dataset.
Un altro importante strumento per misurare la bontà del modello in esame è la curva ROC (*Receiver operating characteristic*), in cui sulle ascisse si trova il tasso dei falsi positivi e sulle ordinate si trova il tasso dei veri positivi, ed è quindi una misura del rapporto tra gli allarmi veri e i falsi allarmi. L'area al di sotto di questa curva, chiamata AUC (*Area Under Curve*) è un'altra misura utilizzata per la valutazione del modello. Tanto più vicino questo valore ad 1, tanto migliore è il modello in quanto il numero di falsi positivi è tanto più vicino a zero. Lo studio della curva ROC può essere utile anche per determinare delle soglie diverse da quella standard (del 50%), nel caso in cui si preferisca avere un numero maggiore di positivi (sia veri che falsi) se si preferisce che la classe positiva sia, in generale, meglio classificata di quella negativa, che può essere di minor interesse. Nello studio della curva ROC spesso è presente anche la curva base, che è semplicemente la bisettrice del primo quadrante, e indica la curva ROC che avrebbe un modello che classifica in maniera randomica (lanciando una moneta bilanciata) i dati, e ha quindi un'AUC di 0.5. Un buon algoritmo di classificazione ha una curva che si trova sopra questa curva base e ha un'area che sia strettamente maggiore di 0.5.

Capitolo 10

Algoritmi

L'algoritmo Beta mostra come stimare i parametri a, b relativi ad un vettore di dati che si vogliono modellizzare tramite una distribuzione beta di cui si conoscono media e varianza. Gli input della funzione sono m (la media) e v (la varianza) e gli output della funzione sono a, b.
L'algoritmo VaR_num mostra come stimare il VaR di un vettore i cui elementi sono dati dalla formula (1.15). Prende in input il vettore *list* e il quantile α che identifica il livello del VaR e restituisce il valore di VaR a quel livello.
L'algoritmo VaR_par mostra come stimare il VaR di una variabile aleatoria distribuita come una Beta Binomiale di parametri n, a, b. Prende in input i parametri n, a, b che caratterizzano la distribuzione e il quantile α che identifica il livello del VaR, e restituisce il valore di VaR a quel livello.

[H]

Figura 10.1: Stima dei parametri di una distribuzione Beta.

Betam, v
$a \leftarrow \left(\frac{1-m}{v-\frac{1}{m}}\right) \cdot m^2$ $b \leftarrow a \cdot (\frac{1}{m} - 1)$ a, b [H]

Figura 10.2: Stima del VaR della distribuzione non parametrica.

VaR_NUMlist, quantile
$N \leftarrow len(list)$ i = 0:N sum[i] = cumsum(list[0:i])
cumsum è la funzione che fa la somma cumulata di un vettore $sum[i] \geq quantile$ i [H]

Figura 10.3: Stima del VaR della distribuzione Betabinomiale.

VaR_PARn, a, b, quantile list = betabinomialpmf(n, a, b)
betabinomialpmf è la funzione che calcola la pmf di una distribuzione betabinomiale i = 0:n sum[i] = cumsum(list[0:i])
cumsum è la funzione che fa la somma cumulata di un vettore $sum[i] \geq quantile$ i

Bibliografia

Angelos Delivorias. Understanding securitisation, October 2015.

Semeraro P. Doria M, Luciano E. Machine learning techniques in joint default assessment. May 2022.

A. Embrechts P, Frey R., McNeil. *Quantitative Risk Management*. 2005.

F.J. Fabozzi. *The Handbook of Fixed Income Securities*. Handbook of Fixed Income Securities, 7th Ed. McGraw-Hill, 2005.

Guggenheim Investments. The abcs of asset-backed securities. URL `https://www.guggenheiminvestments.com/GuggenheimInvestments/media/PDF/The-ABCs-of-Asset-Backed-Securities-2022.pdf`.

Helge Münkel. *Asset-Backed Securities: It's as easy as this! A Practical Factbook*, Febraury 2006.

Standard & Poor's. The basics of credit enhancement in securitizations. URL `https://fcic-static.law.stanford.edu/cdn_media/fcic-docs/2008-06-24%20S&P%20Basics%20of%20Credit%20Enhancement%20in%20Securitizations.pdf`.