UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche

Tesi di Laurea Magistrale

Modelli di analisi di dati funzionali per la previsione del rischio di credito nel sistema bancario

Relatore Ch.mo Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Correlatore Dott. Angelo Basile

Laureando
Marianna Lax 2006658

Anno Accademico 2021/2022

Indice

Introduzione

Questa tesi si propone di analizzare, con metodologia statistica, un problema specifico degli istituti di credito: identificare tempestivamente i clienti a maggior rischio d'insolvenza. Nello specifico, viene elaborato un modello statistico da mettere a disposizione ad una banca che, in aggiunta ad altri strumenti di analisi tradizionali, permetta, sulla base delle transazioni trimestrali compiute dai suoi singoli clienti, una previsione più accurata del rischio di credito nel sistema bancario. In sostanza, la banca, al fine di prevedere gli eventuali default creditizi dei propri clienti, vuole sorvegliare la loro solvibilità finanziaria mediante le transazioni periodiche effettuate da ciascuno di essi. Prevedere con largo anticipo il default creditizio di un cliente è molto importate, poiché i default dei clienti comportano ripercussioni negative non solo nei rapporti tra cliente e banca, ma anche, se su larga scala, nell'equilibrio finanziario dell'istituto di credito, del sistema bancario e del sistema economico nel suo insieme, proprio per il ruolo di intermediazione che le banche svolgono tra risparmio e investimenti. Negli ultimi anni gli istituti di credito si sono dotati del cosiddetto Sistema di Allarme Preventivo per monitorare il 'ciclo di vita' dei crediti dei propri clienti, allo scopo di identificare tempestivamente i clienti a maggior rischio d'insolvenza. In questa tesi, Tale Sistema di Allarme Preventivo viene riesaminato mediante l'uso di tecniche statistiche. Nello specifico, vengono utilizzate le transazioni come covariate nei modelli adattati con lo scopo di ottenere previsioni più accurate dell'evento d'interesse oltre che una maggiore conoscenza del fenomeno in esame. Nel primo Capitolo, dopo aver presentato i dati grezzi a disposizione, il problema viene formalizzato da un punto di vista matematico-statistico. Nel secondo Capitolo vengono presentate le previsioni usando alcune tecniche di classificazione note in letteratura: i modelli lineari, i modelli lineari generalizzati, i modelli MARS e il *gradient boosting*. Nel terzo Capitolo, prima di procedere con la discussione di una classe di modelli alternativa e più sofisticata, ovvero i modelli per dati funzionali, vengono brevemente presentati

alcuni concetti teorici alla base di questi modelli. Sempre nel terzo Capitolo si propone un nuovo metodo di stima di modelli funzionali multipli con penalizzazione per selezionare automaticamente le variabili. Nello specifico, viene estesa al caso multivariato una metodologia presente in letteratura per il solo caso univariato. Questo nuovo metodo permette, come detto precedentemente, oltre alla stima dei coefficienti funzionali anche una selezione automatica delle variabili nel caso multivariato. Tutti gli approcci utilizzati sono stati confrontati sia da un punto di vista interpretativo che previsivo. Come spesso accade non esiste però una metodologia che prevale sull'altra, ma una combinazione di approcci, che sicuramente contribuisce ad avere una comprensione più accurata del fenomeno in esame. Nel quarto Capitolo, infine, viene condotta un'analisi della sopravvivenza dei clienti facenti parte dei tre segmenti: Grandi Imprese, Piccole e Medie Imprese e Privati, al fine di analizzare l'incidenza dell'evento d'interesse nell'arco temporale considerato.

Capitolo 1

Sistema di Allarme Preventivo per Rischio di Credito

1.1 Presentazione dell'azienda

Questo lavoro di tesi si basa su uno stage svolto presso l'azienda CRIF S.p.A, che ha messo a disposizione i dati per conto di un importante istituto di credito. CRIF è un'azienda con sede a Bologna, fondata nel 1988, è leader in Europa nel settore delle *credit information* bancarie ed è specializzata, inoltre, in analisi di dati e soluzioni in ambito digitale, offrendo a banche, società finanziarie, assicurazioni, società di telecomunicazioni e imprese un supporto qualificato in ogni fase della relazione con il cliente: dalla pianificazione delle strategie di sviluppo e di investimento, all'acquisizione di nuovi mercati e clienti, fino alla gestione del proprio portafoglio e degli eventuali crediti insoluti. I servizi offerti da CRIF consentono di anticipare l'evoluzione dei mercati, di ridurre i rischi di credito, di prevenire le frodi.

1.2 Presentazione del contesto

Un'importante banca del panorama italiano si è rivolta all'azienda di consulenza CRIF al fine di migliorare il proprio Sistema di Allarme Preventivo (*Early Warning System*), per prevenire con largo anticipo i default creditizi del proprio portafoglio clienti, identificati univocamente dal codice *CustomerID*, su tre segmenti: Privati, Piccole e Medie Imprese e Grandi Imprese.

L'*Early Warning System* è un sistema utilizzato dagli istituti di credito per identificare con tre mesi di anticipo i clienti a maggior rischio di insolvenza. Attraverso questo Sistema un cliente è categorizzato come in stato di 'allarme' ovvero in *Early Warning (EW)* quando, nei successivi tre mesi rispetto al momento di osservazione, i crediti deteriorati di quest'ultimo registrano almeno 30 giorni di insoluto rispetto all'ultima rata non pagata o al primo giorno di sconfino di conto corrente. Dove per sconfino si intende quando il cliente utilizza in via temporanea un importo superiore a quello precedentemente stabilito come ammontare massimo di fido.
Il Sistema di Allarme Preventivo si è sviluppato dopo la crisi finanziaria del 2008 allo scopo di tutelare le banche ed evitare ciò che accadde durante quel periodo storico, ovvero il fallimento di numerosi istituti di credito, causato in particolare da una sbagliata gestione dei prestiti verso i clienti.
Questo Sistema si evolve a partire da un modello bancario denominato *modello a tre stadi*, utilizzato per la valutazione del rischio di credito verso la clientela e che permette di monitorare il 'ciclo di vita' dei crediti di un cliente. Nel dettaglio, quest'ultimo prevede una suddivisione dei crediti in tre livelli:

1. *Stadio 1* con rischio creditizio basso: il cliente è in grado di adempiere ai suoi obblighi di pagamento di interessi e di rimborso del capitale;

2. *Stadio 2* con rischio creditizio intermedio: il cliente registra almeno 30 giorni di insoluto rispetto all'ultima rata non pagata o al primo giorno di sconfino;

3. *Stadio 3* con rischio creditizio alto: il cliente manifesta mancati pagamenti per un periodo superiore ai 90 giorni, in questo caso il cliente è passato in *default*.

Nel gergo del rischio di credito, l'obiettivo dell'*Early Warning System* è quindi quello di identificare con tre mesi di anticipo il passaggio di un cliente dallo Stadio 1 allo Stadio 2. In questo modo la banca riesce a percepire con largo anticipo anche eventuali default dei clienti (cioè il passaggio allo Stadio 3) e di porre in essere tutte le azioni cautelative al fine di evitare il deterioramento dei crediti bancari concessi ai clienti.
Una corretta identificazione dell'*Early Warning* da parte dello statistico è dunque essenziale per avere una corretta attuazione del Sistema di Allarme Preventivo ed è inoltre di supporto ad altre figure professionali per la definizione di una strategia tempestiva di gestione del recupero dei crediti scaduti.

Il problema in esame può quindi essere affrontato con un approccio statistico considerando come parametro d'interesse π_{ij}, ovvero la probabilità che il cliente i-esimo del segmento j-esimo venga categorizzato in stato di 'allarme' nel mese di osservazione:

$$\pi_{ij} = \mathbb{P}(\text{Cliente } i\text{-esimo è in } \textit{Early Warning}) \tag{1.1}$$

dove l'indice $j = 1, 2, 3$ indica i tre segmenti considerati e $i = 1, .., N_j$ dove N_j rappresenta il numero totale di clienti presenti in ciascuno dei tre segmenti.
Una stima di π_{ij} può essere ottenuta valutando alcune informazioni inerenti ai clienti. Il Sistema di Allarme Preventivo sinora adottato dalla banca prende in considerazione informazioni statiche nel tempo quali ad esempio dati socio-demografici, dati di bilancio, dati relativi al volume totale del transato o saldi statici mensili (come conto corrente, carte o fido). A differenza di tale approccio, in questa tesi sono state considerate informazioni dinamiche nel tempo, ovvero le transazioni periodiche (opportunamente categorizzate in costi e ricavi) effettuate da ciascun cliente dei tre segmenti, da dicembre 2018 a dicembre 2019. Il metodo proposto viene utilizzato non per sostituire il metodo tradizionale ma per integrare le informazioni della banca nella gestione del credito.
La banca si pone quindi un duplice obiettivo: identificare con tre mesi di anticipo i clienti con maggior rischio di insolvenza e verificare inoltre se le informazioni contenute nelle transazioni effettuate dai clienti siano significative da un punto di vista statistico per ottenere una corretta previsione dell'evento d'interesse.
Una stima di π_{ij} può perciò essere ottenuta attraverso:

$$\widehat{\pi_{ij}} = \widehat{f}(\text{Transazioni effettuate dal cliente } i\text{-esimo}) \tag{1.2}$$

Dove $\widehat{f}(\cdot)$ è una funzione che racchiude tutte le informazioni dinamiche relative al cliente i-esimo del segmento j-esimo. Una stima di $\widehat{f}(\cdot)$ può essere ricavata attraverso modelli parametrici o non parametrici. La scelta del modello migliore da parte dello statistico è per lo più soggettiva ed è talvolta un compromesso tra interpretabilità del modello e capacità predittiva di quest'ultimo.

1.3 Presentazione dei dataset

I tre dataset analizzati in questa tesi, forniti dall'azienda di consulenza CRIF per conto della banca, contengono informazioni relative rispettivamente a 323290, 134205 e 24072 clienti banca (identificati dal codice *CustomerID*) su tre segmenti: Privati, Piccole e Medie Imprese e Grandi Imprese. Per ogni cliente sono stati raccolti con frequenza *trimestrale*, da dicembre 2018 a dicembre 2019, dati relativi alle transazioni finanziarie effettuate in entrata e in uscita, consentendo così di avere a disposizione una 'fotografia' dei flussi bancari della persona nei mesi di osservazione considerati. Nel dettaglio, i dataset sono costituiti da un numero di covariate differenti per i tre gruppi.
Le variabili esplicative disponibili in ciascuno dei tre segmenti sono il risultato di una preliminare procedura di elaborazione condotta dall'azienda di consulenza CRIF sulla base dei dati grezzi messi a disposizione dalla banca. Di seguito si riassumono i passaggi:

(1) Un cliente può disporre di più Conti Correnti all'interno della stessa banca. Per un singolo Conto Corrente sono state estratte le variabili elementari, riportate nella Tabella 1.1 disponibili a livello giornaliero;

Variabile	**Descrizione**
Saldo C/C	Saldo giornaliero
Accordato	Accordato fido conto corrente
Utilizzato	Utilizzato giornaliero di conto corrente
Giacenza	Giacenza giornaliera
Data transazione	Data della transazione
Importo transazione	Importo di una singola transazione

Tabella 1.1: Variabili elementari di un Conto Corrente.

(2) Le variabili presenti nei vari Conti Correnti aperti dallo stesso cliente vengono aggregate in un unico *Conto* per *CostumerID*;

(3) Le informazioni disponibili a livello giornaliero del *Conto* ottenuto al punto precedente, vengono raggruppate con frequenza trimestrale nel periodo che va da dicembre 2018 a dicembre 2019;

(4) Le variabili ottenute al punto precedente vengono divise per tipologia di transazione: attiva o passiva codificata come *Dare* o *Avere* nel nome delle variabili;

(5) Viene sintetizzata l'informazione presente all'interno delle covariate per mezzo di indici, ad esempio la media o indici di variabilità (variazione rispetto al mese o trimestre precedente);

(6) Selezione delle variabili finali, eliminando quelle che presentavano una percentuale elevata di valori mancanti. I valori mancanti sono stati gestiti dall'azienda di consulenza sostituendoli con il valore mediano della variabile presa in considerazione.

In Tabella 1.2, 1.3 e 1.4 viene fornita una sintetica descrizione di alcune delle variabili presenti nel dataset, organizzato nel cosiddetto *'formato lungo'*, rispettivamente del segmento Grandi Imprese, Piccole e Medie Imprese e Privati, dopo l'eliminazione di alcune di esse che risultavano correlate tra loro (oltre il 40%). In presenza di due variabili correlate tra loro si è deciso, dopo aver adattato un modello logistico ai dati, di eliminare la variabile che risulta non significativa o che apporta un peggioramento in termini di *AIC* (*Akaike information criterion*) al modello.

Variabile	**Descrizione**	**Aggregazione temporale**
Importo_Min_Bon	Importo mensile minimo di bonifico per *CustomerID*	Mensile
Dare_Prestiti	Somma mensile nella sezione dare delle commissioni per l'avvio di prestiti per *CustomerID*	Mensile
Dare_Tasse	Somma mensile nella sezione dare per il pagamento di tasse per *CustomerID*	Mensile
Dare_MateriePrime	Somma mensile nella sezione dare dei pagamenti di forniture aziendali (materie prime, beni e servizi) per *CustomerID*	Mensile
Dare_Professionisti_Media_T	Media trimestrale sezione dare dei pagamenti di professionisti (notai, ingegneri o avvocati) per *CustomerID*	Trimestrale
Avere_TrasfDenaro_Media_T	Media trimestrale nella sezione avere dei trasferimenti di denaro per *CustomerID*	Trimestrale
Dare_TrasfDenaro_Media_T	Media trimestrale nella dare avere dei trasferimenti di denaro per *CustomerID*	Trimestrale
Avere_Vendite_Media_T	Media trimestrale nella sezione avere dei ricavi di vendite per *CustomerID*	Trimestrale
Dare_Salari_Media_T	Media trimestrale nella sezione dare dei pagamenti per i salari per *CustomerID*	Trimestrale
Dare_RateMutuo_Media_T	Media trimestrale nella sezione dare dei pagamenti per le rate del mutuo per *CustomerID*	Trimestrale
Dare_Contabilità_Media_T	Media trimestrale nella sezione dare dei pagamenti per spese di contabilità per *CustomerID*	Trimestrale

Tabella 1.2: Descrizione variabili segmento Grandi Imprese

Variabile	**Descrizione**	**Aggregazione temporale**
Importo_Min_Bon	Importo mensile minimo di bonifico per *CustomerID*	Mensile
Avere_Vendite	Somma mensile nella sezione avere dei ricavi di vendite per *CustomerID*	Mensile
Dare_Prestiti	Somma mensile nella sezione dare delle commissioni per l'avvio di prestiti per *CustomerID*	Mensile
Dare_Tasse	Somma mensile nella sezione dare per il pagamento di tasse per *CustomerID*	Mensile
Dare_MateriePrime	Somma mensile nella sezione dare dei pagamenti di forniture aziendali (materie prime, beni e servizi) per *CustomerID*	Mensile
Dare_RateMutuo	Somma mensile nella sezione dare dei pagamenti per le rate del mutuo per *CustomerID*	Mensile
Dare_Bollette_Media_T	Media trimestrale sezione dare dei pagamenti per bollette per *CustomerID*	Trimestrale
Avere_TrasfDenaro_Media_T	Media trimestrale nella sezione avere dei trasferimenti di denaro per *CustomerID*	Trimestrale
Dare_TrasfDenaro_Media_T	Media trimestrale nella dare avere dei trasferimenti di denaro per *CustomerID*	Trimestrale
Var_Salari_T	Variazione nella sezione dare rispetto al trimestre precedente dei pagamenti per i salari per *CustomerID*	Trimestrale

Tabella 1.3: Descrizione variabili segmento Piccole e Medie Imprese

Variabile	**Descrizione**	**Aggregazione temporale**
Importo_Max_Bon	Importo mensile massimo di bonifico per *CustomerID*	Mensile
Dare_Interessi	Somma mensile nella sezione dare dei pagamenti d'interessi per *CustomerID*	Mensile
Dare_Tasse	Somma mensile nella sezione dare per il pagamento di tasse per *CustomerID*	Mensile
Dare_Trasporti	Somma mensile nella sezione dare dei pagamenti per i trasporti pubblici per *CustomerID*	Mensile
Dare_Prestiti	Somma mensile nella sezione dare delle commissioni per l'avvio di prestiti per *CustomerID*	Mensile
Dare_Spesa_Media_T	Media trimestrale sezione dare dei pagamenti per la spesa per *CustomerID*	Trimestrale
Dare_Cinema_Media_T	Media trimestrale sezione dare dei pagamenti per cinema, teatri o musei per *CustomerID*	Trimestrale
Avere_TrasfDenaro_Media_T	Media trimestrale nella sezione avere dei trasferimenti di denaro per *CustomerID*	Trimestrale
Dare_Ristorante_Media_T	Media trimestrale nella dare dei pagamenti per ristoranti o bar per *CustomerID*	Trimestrale
Avere_Vendite_Media_T	Media trimestrale nella sezione avere dei ricavi di vendite per *CustomerID*	Trimestrale
Dare_Educazione_Media_T	Media trimestrale nella sezione dare dei pagamenti per l'istruzione (scuole, Università, Master) per *CustomerID*	Trimestrale
Var_Mediche_T	Variazione rispetto al trimestre precedente dei pagamenti per spese mediche *CustomerID*	Trimestrale

Tabella 1.4: Descrizione variabili segmento Privati

Dal momento che i dataset possiedono una struttura analoga, viene riportata a titolo di esempio nella Tabella 1.6 la struttura organizzata nel *formato corto* del dataset relativo al segmento Grandi Imprese.
La variabile dicotomica, denominata in ciascun dataset con il nome *Target*, assume valore 1 se il cliente è in stato di *Early Warning* nel mese di osservazione e 0 altrimenti. Di cruciale importanza è la corretta comprensione della variabile d'interesse. Per chiarezza, ad esempio, quando *Target* assume valore 1 per l'i-esima persona nel mese di riferimento (dicembre 2019) ciò vuol dire che il cliente al momento della rilevazione ha una situazione creditizia regolare, tuttavia nei *successivi tre mesi* (dopo dicembre 2019) passa in Stadio 2, registrando almeno 30 giorni di insoluto.

Dicembre 2019	*Gennaio 2020*	*Febbraio 2020*	*Marzo 2020*
Target: 1	·	Passaggio Stadio 2	·
Target: 0	·	·	·
Target: 1	Passaggio Stadio 2	·	·
Target: 1	·	·	Passaggio Stadio 2

Tabella 1.5: Esempio di attribuzione dei valori alla variabile *Target*

In Tab. 1.5 viene fornito un esempio di come sono attribuiti i valori alla variabile *Target*. Se nell'ultimo trimestre osservato (da gennaio 2020 a marzo 2020) un cliente passa allo Stadio 2, il Sistema categorizza il *CostumerID* come in stato di *Early Warning* ovvero $Target = 1$ nel primo mese che precede il trimestre considerato (in questo caso dicembre 2019).
Di particolare interesse in questa tesi, come precedentemente detto, è l'utilizzo di variabili dipendenti dal tempo per la previsione della variabile *Target*. L'utilizzo delle informazioni presenti nelle transazioni permette infatti di sorvegliare i flussi bancari dei clienti nei vari istanti temporali. L'idea in questo modo è quella di intercettare nel più breve tempo possibile i primi segnali di anomalia del credito monitorando i movimenti di denaro.

CustomerID	*Target_dic_2018*	...	*Target_dic_2019*	...	*Importo_Min_Bon_dic_2018*	...	...	*Importo_Min_Bon_dic_2019*
3417449	0	...	0	...	1.26	...	...	31.84
85441115	0	...	0	...	5360.00	...	...	60000.00
20441963	0	...	1	...	700.00	...	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
12473401	-	...	0	...	102.48	...	...	130.00
16264659	0	...	0	...	6100.00	...	...	4588.42
11412163	1	...	-	...	0	...	...	7000.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabella 1.6: Struttura parziale nel *formato corto* segmento grandi imprese

Viene riportata in Tab.1.6 la struttura parziale nel *formato corto* del dataset relativo al segmento Grandi Imprese. Si nota che i valori della variabile *Importo_Min_Bon* sono organizzati in colonna per ogni mese di rilevazione. *Importo_Min_Bon_dic_2018* indica ad esempio il valore della variabile nel mese di dicembre 2018. Analoghe considerazioni possono essere fatte per tutte le altre variabili. Per le analisi che seguono, vengono riportati i risultati del solo segmento Grandi Imprese. Le valutazioni e conclusioni tratte in questa tesi su tale segmento, si possono estendere alle Piccole e Medie Imprese e ai Privati. Per completezza, vengono riportati in Appendice A i risultati dei due segmenti non presi in considerazione.

1.4 Analisi descrittive

Viene riportata per i tre segmenti nel grafico della Figura 1.1, la distribuzione della variabile risposta *Target* nel mese di dicembre 2019, per i segmenti Grandi Imprese, Piccole e Medie Imprese e Privati. Come si può notare, nei tre gruppi le classi della risposta sono sbilanciate, mostrando infatti una percentuale molto bassa di coloro che sperimentano l'evento d'interesse. La caratteristica di sbilanciamento è presente anche nelle classi della variabile *Target* per gli altri mesi di osservazione (dicembre 2018, marzo 2019, giugno 2019 e settembre 2019).
In Figura 1.2 e 1.3 vengono messi a confronto i boxplot dei clienti che sono in stato di *Early Warning* (arancione) con coloro i quali non sono in stato di 'allarme' (azzurro) per le covariate *Importo_Min_Bonifico* e *Avere_TrasfDenaro_Media_T* per il segmento Grandi Imprese. Si precisa che, per semplicità di trattazione, vengono riportati i

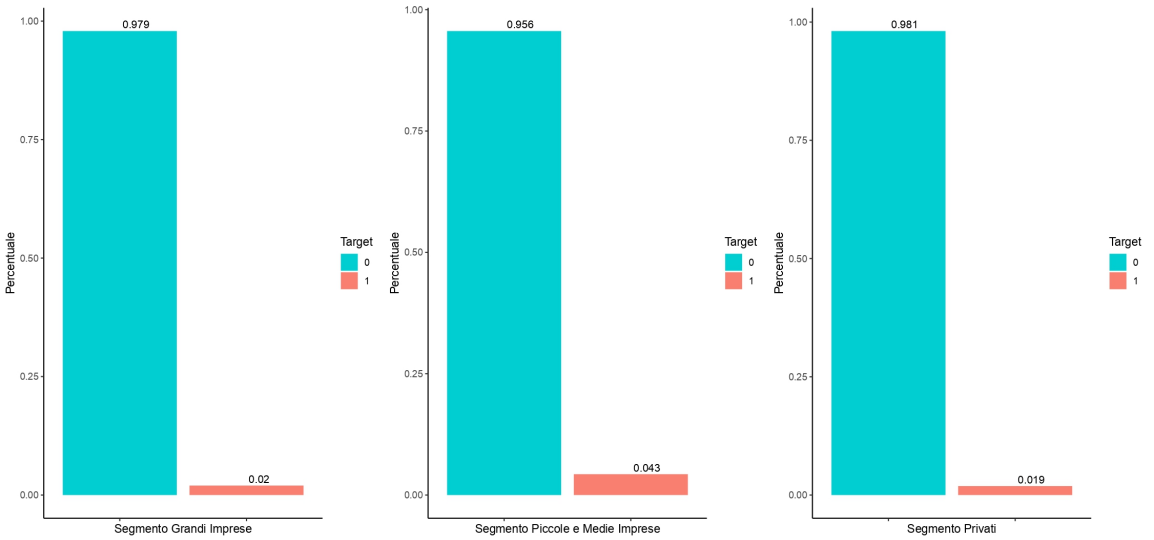

Figura 1.1: Distribuzione marginale della variabile risposta *Target* nel mese di dicembre 2019

boxplot di due sole covariate in quanto le conclusioni che si possono trarre dall'analisi dei relativi boxplot di tutte le variabili esplicative disponibili sono simili tra di loro. La distinzione tra le due classi è evidente nel periodo che precede il trimestre in cui il cliente passa allo Stadio 2. Questo può indicare un cambiamento nei movimenti bancari per i clienti in stato di 'allarme' in prossimità del trimestre nel quale sono insolventi. Questo trend viene notato per la maggioranza delle esplicative considerate nelle analisi per il segmento Grandi Imprese e per le covariate relative ai rimanenti segmenti.

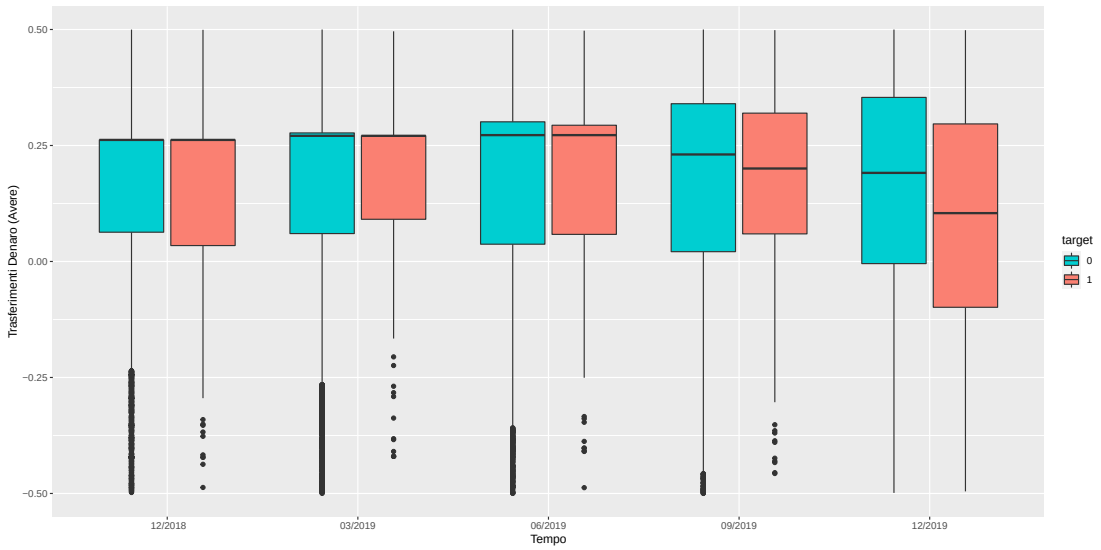

Figura 1.2: Boxplot della covariata standardizzata *Avere_TrasfDenaro_Media_T*

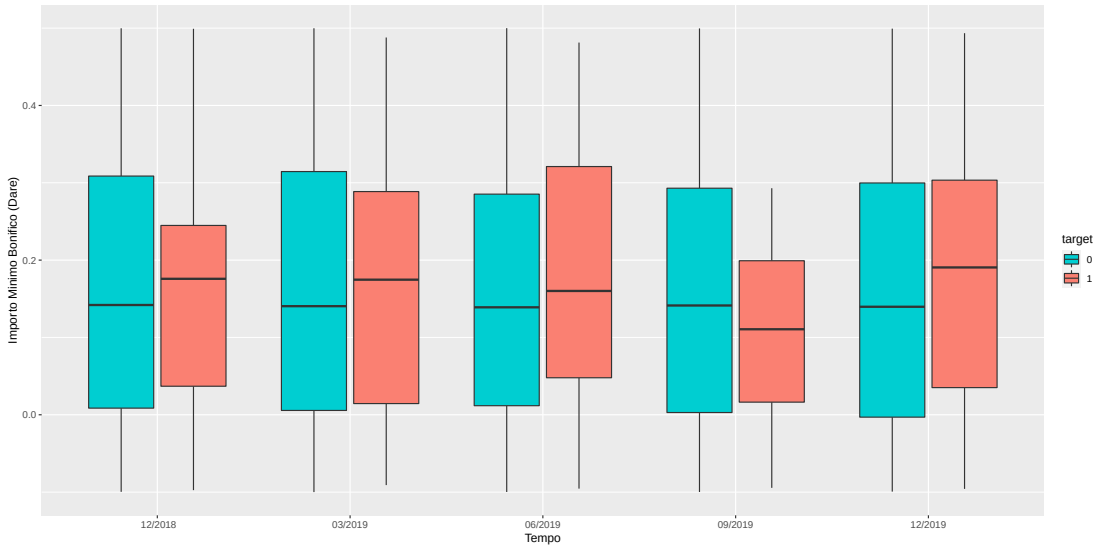

Figura 1.3: Boxplot della covariata standardizzata *Importo_Min_Bonifico*

Capitolo 2

Previsione del rischio di credito

2.1 Previsione della risposta

Si vuole prevedere, utilizzando dati inerenti alle transazioni periodiche, la probabilità che un cliente sia in stato di *Early Warning* a dicembre 2019. Inizialmente, nei modelli utilizzati, è stata trascurata la dipendenza temporale presente tra variabili relative allo stesso cliente e riferite ai diversi mesi di rilevazione. Questa è una pratica comune in numerosi istituti di credito, i quali ignorano nei modelli adottati tale caratteristica. Considerare invece la natura longitudinale dei dati può portare, come si vedrà nel Capitolo 3, ad avere una più completa comprensione del fenomeno d'interesse.
Formalmente si indica con la variabile discreta Y_i con supporto in $\{0, 1\}$ la variabile risposta $Target_i$ nel mese di dicembre 2019 per un cliente i-esimo. Si assume che i dati siano stati generati dalla legge:

$$\pi_i = \mathbb{P}(Y_i = 1) = f(\tilde{x}_i) \;\; i = 1, ..., N \tag{2.1}$$

Dove N è il numero di clienti mentre $\tilde{x}_i = [x_{i1}, ..., x_{ip}]^T$ è il vettore colonna corrispondente alla riga della matrice X contenente i valori delle p covariate. Si tratta di un problema di *classificazione* dove è d'interesse la stima di una regola $\widehat{\pi}_i = \widehat{f}(\tilde{x}_i)$ che consenta di prevedere π_i sulla base di $\tilde{x}_i$ anche in presenza di dati nuovi ovvero osservazioni precedentemente non osservate. Una volta ottenuta la stima $\widehat{\pi}_i$, un soggetto i-esimo viene classificato come in stato di *Early Warning* se $\widehat{\pi}_i > s$, dove s rappresenta una soglia fissata.
Per non *sovradattare* il modello ai dati, visto l'elevato numero di osservazioni dispo-

nibili, il dataset è stato diviso per il 60% in *insieme di stima* utilizzato per la stima dei modelli e il restante 40% in *insieme di verifica* impiegato per valutare le prestazione dei modelli. Per una trattazione completa della motivazione si veda Azzalini & Scarpa (2012).

2.1.1 Operazioni preliminari

Ai fini dell'analisi si conducono preliminarmente le operazioni utilizzando il software `R`, che possono essere riassunte nei seguenti punti:

(1) i tre dataset disponibili vengono organizzati nel *formato corto* (Tabella 1.6). In questo modo ogni riga del dataset corrisponde ad un singolo cliente che compare solo una volta all'interno dell'insieme di dati; quindi il numero delle osservazioni coincide con il numero di unità. I valori delle variabili esplicative sono organizzati per colonna: ciascuna ripetizione temporale dell'esplicativa diventa una variabile;

(2) eliminazione dei *CustomerID* per i quali non è disponibile l'ultima rilevazione di dicembre 2019, poichè la mancanza del valore della variabile risposta *Target* al tempo di osservazione rende impossibile verificare l'adeguatezza delle procedure di stima. Nella Tabella 2.1 è riportato uno schema dell'operazione effettuata dove in grassetto sono sottolineati alcuni dei *CustomerID* eliminati dal dataset per la mancanza del valore della variabile risposta, ovvero *Target* a dicembre 2019. Vengono eliminati 448 clienti pari all'1% del totale.

Dal momento che la banca rileva il fenomeno d'interesse nell'ultimo giorno del mese considerato, le informazioni delle covariate disponibili fino a dicembre 2019 possono essere utilizzate nella previsione.
Dopo le operazioni il numero di osservazioni risulta pari a 23624. La perdita di informazione, come precedentemente descritto, si verifica dall'operazione effettuata nel punto (2). Si sono infine standardizzate le variabili esplicative.

CustomerID	*Target_dic_2018*	*Target_mar_2019*	*Target_giu_2019*	*Target_sett_2019*	*Target_dic_2019*	Eliminazione
3417449	0	0	0	0	0	No
15149335	**1**	**-**	**-**	**-**	**-**	**Si**
85441115	0	-	1	0	0	No
12473401	0	0	0	0	0	No
14252578	**0**	**0**	**0**	**0**	**-**	**Si**
⋮	⋮	⋮	⋮	⋮	⋮	

Tabella 2.1: Schema dell'operazione effettuata

2.1.2 Regressione Logistica

Per le risposte binarie il modello statistico più semplice per l'i-esimo soggetto è un binomiale $Y_i \sim Bi(1, \pi_i)$, dove il predittore lineare è definito nel seguente modo:

$$g(\pi_i) = \beta_1 x_{i1} + ... + \beta_p x_{ip} = x_i\beta \qquad i = 1, ..., N \tag{2.2}$$

Con funzione legame logistica $g(\cdot)$ tale che $g : [0, 1] \to \mathbb{R}$, il modello assume la forma:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = x_i\beta \tag{2.3}$$

Una stima del vettore dei parametri $\beta = (\beta_1, ..., \beta_p)$ può essere ottenuta con un approccio frequentista attraverso la stima di massima verosimiglianza:

$$\hat{\beta} = \arg\max_{\beta} l(\beta; y) \tag{2.4}$$

Per una trattazione completa si veda Salvan et al. (2020) e McCullagh & Nelder (1989).
Visto l'elevato numero di covariate presenti nel dataset a disposizione, per definire un modello più parsimonioso, viene effettua la procedura di *stepwise forward* per una selezione automatica delle variabili. In breve, tale metodo prevede, partendo dal modello con la sola intercetta, l'aggiunta delle variabili che migliorano l'adattamento secondo un criterio definito, in questo caso l'*AIC (Akaike information criterion).*

2.1.3 Classificazione mediante Regressione Lineare

Come descritto nel dettaglio da Azzalini & Scarpa (2012) nel caso in cui la variabile risposta Y_i ha due classi 0 e 1, si può decidere di utilizzare un modello di regressione lineare per ottenere la regola:

$$\hat{y}_i = x_i^T \hat{\beta} \tag{2.5}$$

con la quale classificare un cliente nel gruppo 1 se $\hat{y}_i > s$ o 0 se $\hat{y}_i < s$. In questo caso le regioni di classificazione sono separate dall'iperpiano: $x^T\hat{\beta} = s$, dove s rappresenta la soglia di discriminazione scelta per la previsione delle due categorie. I parametri $\hat{\beta}$ sono stimati con i minimi quadrati ordinari. In queste analisi è stato scelto il modello lineare nella forma più semplice:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \tag{2.6}$$

Si noti che le realizzazioni della variabile casuale ϵ_i nella scrittura dell'equazione (2.6) sommate alla parte deterministica devono dare 0 oppure 1.
Si può estendere tale procedimento ponendo nel predittore lineare funzioni non lineari delle variabili esplicative. Questa procedura risulta poco ottimale difatti tale modello viene usato come *benchmark*, ossia una sorta di termine di paragone, una soglia al di sotto del quale è difficile scendere.
Anche in questo caso per definire un modello più parsimonioso, si effettua la procedura *stepwise forward* per la selezione automatica delle variabili utilizzando come criterio di selezione l'AIC.
Di seguito vengono riportati nella Tabella 2.2 i risultati dei modelli stepwise logistico e lineare per il segmento Grandi Imprese nel mese di riferimento dicembre 2019. Sono riportate le stime dei coefficienti con il relativo *standard error* e il p-value associato. Per semplicità di lettura vengono riportate le prime sette variabili selezionate da tale procedura e che risultano maggiormente significative.
L'intercetta nel modello logistico è pari a -5.194 ed esprime un effetto negativo sul logaritmo della quota, ossia su $\frac{\pi_i}{1-\pi_i}$, fermo restando il valore delle ulteriori esplicative del modello. L'intercetta nel modello lineare è pari a 0.019 e indica il livello medio della variabile risposta ed esprime un effetto positivo sulla risposta a parità delle altre covariate. Nel modello logistico le covariate hanno un effetto additivo sul logaritmo della quota a differenza del modello lineare dove le variabili esplicative hanno un

effetto additivo sulla risposta Y_i.

	Modello Logistico			*Modello Lineare*		
	$\hat{\beta}_{GLM}$	*Std. Error*	$Pr(>\|z\|)$	$\hat{\beta}_{LM}$	*Std. Error*	$Pr(>\|t\|)$
Intercetta	-5.194	0.143	2e-16	0.019	0.001	2e-16
*Dare_TrasfDenaro_**sett_2019***	0.559	0.110	3e-14	0.009	0.001	2e-16
*Dare_RateMutuo_Media_T_**sett_2019***	0.878	0.108	3e-14	0.021	0.001	3e-15
*Dare_RateMutuo_Media_T_**dic_2019***	-0.678	0.086	3e-14	-0.021	0.002	2e-16
*Dare_MateriePrime_**sett_2019***	–0.441	0.138	3e-14	-0.003	0.001	3e-12
*Avere_TrasfDenaro_Media_T_**dic_2019***	-0.118	0.228	3e-14	-0.005	0.002	2e-16
*Dare_Prestiti_Media_T_**sett_2019***	-0.675	0.127	1e-02	-0.011	0.001	3e-03
*Dare_Prestiti_Media_T_**dic_2019***	0.220	0.080	0.005	0.006	0.002	2e-08

Tabella 2.2: Risultati dei modelli stepwise logistico e lineare per il segmento Grandi Imprese, mese di riferimento dicembre 2019.

É d'intresse stabilire di quanti trimestri 'andare indietro nel tempo' per trovare relazioni significative tra le covariate e la variabile risposta. Dalla Tabella 2.2 si nota come la procedura stepwise in entrambi i modelli ha selezionato come significative nel prevedere lo stato di *Early Warning* di un cliente le variabili relative ai due periodi che precedono il trimestre in cui il cliente è passato allo Stadio 2.
Per verificare se, prendendo mesi differenti da quello finora considerato (dicembre 2019), le variabili esplicative che risultano significative nello spiegare la risposta sono sempre quelle relative ai due periodi precedenti al mese di rilevazione, vengono presi i valori della variabile $Target$ nel mese di settembre 2019 (avendo a disposizione rilevazioni ogni tre mesi delle variabili).
É d'interesse quindi prevedere la probabilità che un *CustomerID* sia in stato di *Early Warning* nel mese di settembre 2019 anziché dicembre 2019 utilizzando gli stessi modelli. Per svolgere quest'ultima analisi si conducono le stesse operazioni preliminari descritte nella Sezione 2.1.1, con la sola differenza che ora si fa riferimento a settembre 2019. Vengono in seguito eliminate le covariate di dicembre 2019 per non utilizzare, nel prevedere la risposta, informazioni disponibili successivamente.
I risultati ottenuti da quest'ultima analisi e riportati nella Tabella 2.3, dimostrano che anche in questo caso la maggioranza delle variabili selezionate sono relative ai due

periodi precedenti rispetto al trimestre in cui il cliente passa allo Stadio 2. Si può inoltre notare che variabili relative alla stessa tipologia di transazione ma riferite ai due differenti trimestri che precedono il trimestre d'insolvenza hanno segno opposto, fermo restando il valore delle altre esplicative. Questo può indicare un cambiamento del comportamento dei movimenti bancari del cliente in prossimità del trimestre di insolvenza a parità delle altre covariate.
Trovare tuttavia una corretta interpretazione dei coefficienti in questo caso risulta difficile, anche se siamo di fronte a modelli semplici. Quest'ultima problematica può essere risolta mediante un'analisi con modelli funzionali, i quali forniscono un'interpretazione più semplice e intuitiva dell'andamento dei coefficienti nel tempo.

	Modello Logistico			*Modello Lineare*		
	$\hat{\beta}_{GLM}$	*Std. Error*	$Pr(>\|z\|)$	$\hat{\beta}_{LM}$	*Std. Error*	$Pr(>\|t\|)$
Intercetta	-6.075	0.262	2e-16	0.021	0.001	2e-14
*Dare_MateriePrime_**giu_2019***	-1.143	0.181	3e-11	-0.008	0.001	2e-17
*Dare_RateMutuo_Media_T_**giu_2019***	0.860	0.123	3e-11	0.018	0.001	2e-17
*Dare_RateMutuo_Media_T_**sett_2019***	-0.351	0.077	3e-11	-0.015	0.001	2e-17
*Dare_Prestiti_Media_T_**giu_2019***	-0.706	0.129	3e-11	-0.008	0.004	3e-06
*Dare_TrasfDenaro_**sett_2019***	0.253	0.073	3e-10	0.019	0.001	2e-17
*Dare_Prestiti_Media_T_**sett_2019***	0.152	0.079	0.001	0.007	0.001	1e-04
*Avere_TrasfDenaro_Media_T_**giu_2019***	0.125	2e-07	0.001	0.010	0.001	2e-16

Tabella 2.3: Risultati ottenuti dai modelli stepwise logistico e lineare per il segmento Grandi Imprese, mese di riferimento settembre 2019.

2.1.4 Regressione Logistica Lasso

La regressione Lasso (*Least Absolute Shrinkage and Selection Operator*) è un metodo di *shrinkage*. Tale procedura sebbene fornisca delle stime dei coefficienti distorte permette di contrarre i coefficienti verso lo zero riducendo la variabilità di questi ultimi. Il modello preso come riferimento assume la forma della 2.3. Sia quindi $l(\beta)$

la funzione di log-verosimiglianza cambiata di segno per dati binomiali:

$$l(\beta) = -\sum_{i=1}^{N}[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)] = \sum_{i=1}^{N}[y_i \, x_i \, \beta - \log(1 + e^{x_i \beta})] \quad (2.7)$$

Sia inoltre $L_\lambda(\beta, \lambda)$ la funzione di perdita da minimizzare tale che:

$$L_\lambda(\beta, \lambda) = l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2.8)$$

La stima dei parametri $\beta = (\beta_0, ..., \beta_p)$ nel lasso dipende dal valore del parametro di regolarizzazione $\lambda \geq 0$ scelto via *convalida incrociata*. La linea tratteggiata del grafico di Figura 2.1 della convalida incrociata mostra il valore ottimale di $\lambda = \lambda_{1se}$.

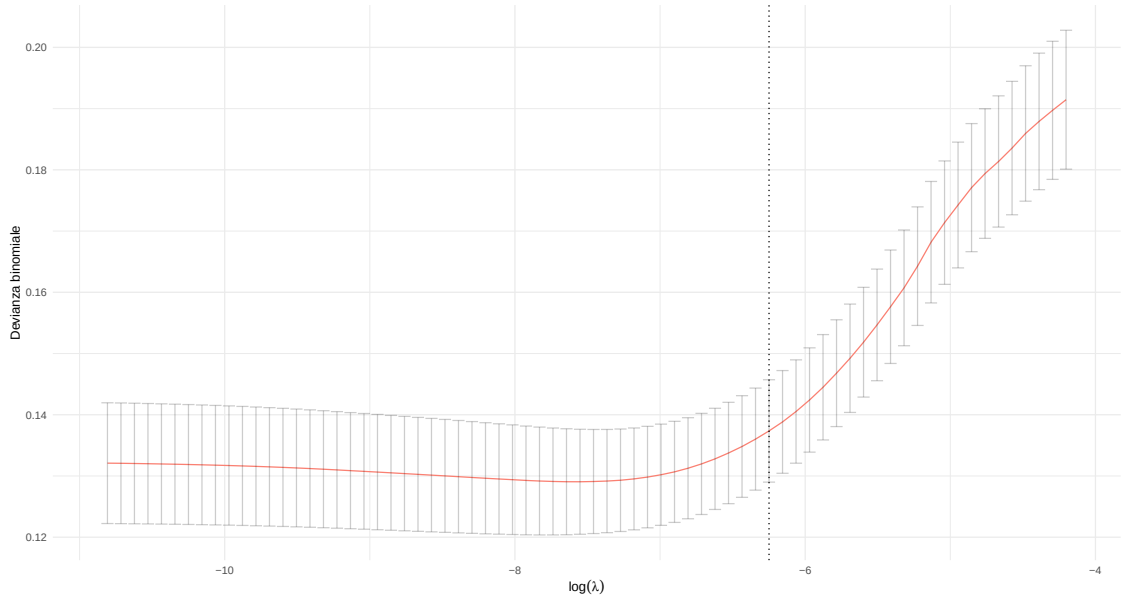

Figura 2.1: Convalida incrociata del lasso logistico, segmento Grandi Imprese, mese di riferimento dicembre 2019

Una stima di β del modello è ottenuto risolvendo il seguente problema di minimo:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} L_\lambda(\beta, \lambda) \quad (2.9)$$

Un'importante vantaggio del Lasso consiste nel fatto che le stime di alcuni coefficienti sono esattamente pari a zero se il parametro λ è sufficientemente grande.

Il Lasso produce quindi una selezione delle variabili oltre a ridurre la variabilità dei coefficienti stimati. I risultati delle analisi sono stati ottenuti tramite il pacchetto `R` **glmnet** (Friedman et al. 2017) il quale utilizza il metodo di ottimizzazione *Proximal Newton* per minimizzare la 2.8. Per una trattazione completa di tale metodo si veda Friedman et al. (2010).
Nella Figura 2.2 vengono riportate le variabili selezionate dal Lasso logistico, ponendo particolare attenzione ai segni dei coefficienti, i quali sono stati evidenziati in arancione quelli con segno negativo e in verde quelli con segno positivo, trovando conferma con quanto precedentemente detto.

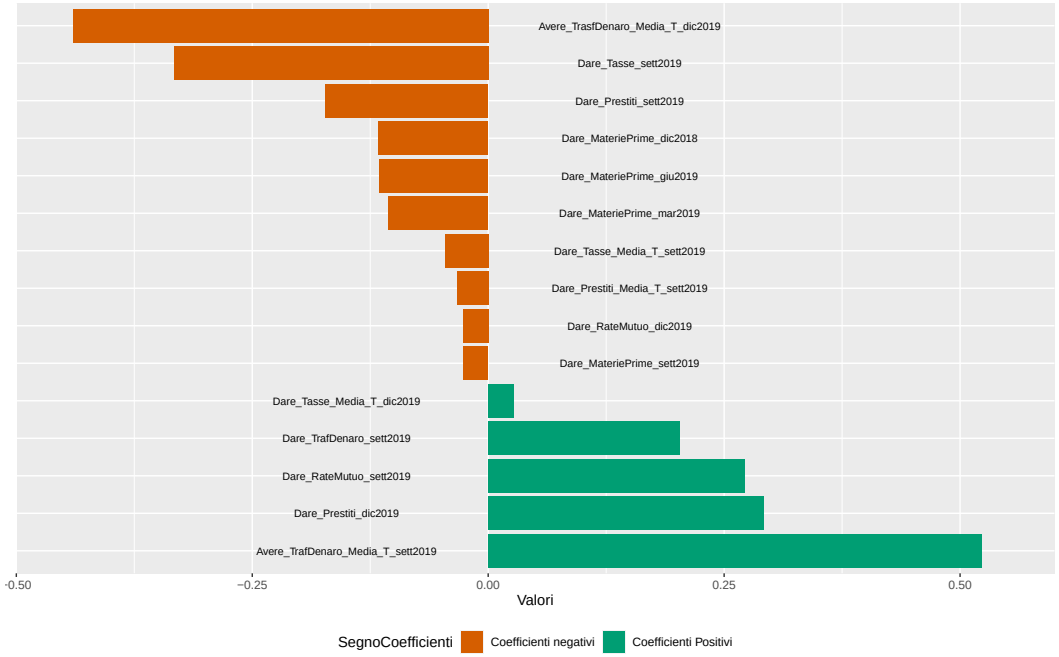

Figura 2.2: Grafico delle variabili selezionate dal lasso logistico con coefficiente positivi (in verde) e negativo (in arancione) per il segmento Grandi Imprese mese di riferimento dicembre 2019.

Anche in questo caso il modello ha selezionato principalmente le variabili relative ai due periodi che precedono il trimestre nel quale il cliente passa allo Stadio 2.
Anche in questo caso i coefficienti delle covariate relative alla stessa tipologia di transazione e riferite a settembre 2019 e dicembre 2019 mostrano segni opposti. Questo comportamento negli ultimi due periodi può essere spiegato con un effetto di compensazione tra i due trimestri. Ciò può indicare un cambiamento nei comportamenti del

cliente che si riflette in una variazione dei suoi movimenti bancari nei due trimestri che precedono il periodo di insolvenza.

2.1.5 Splines di Regressione Multidimesionali Adattive

Le *Multivariate Adaptive Regression Splines* ($MARS$) introdotte da Friedman (1991) sono una particolare specificazione iterativa delle *spline* di regressione che risulta efficiente quando il numero delle covariate è elevato. Nel dettaglio, questo modello è una combinazione lineare di splines con nodo nel punto ζ. Tali funzioni base sono definite nell'insieme:

$$\begin{aligned} C = \{x_j,\ (x_j - \zeta)_+, (\zeta - x_j)_+ :\ \zeta \in \{x_{i1}, x_{i2}, ..., x_{ip}\} \\ i = 1, 2, .., N,\ j = 1, 2, .., p\} \end{aligned} \tag{2.10}$$

La notazione $(x)_+$ indica la parte positiva di x. Il $MARS$ nei problemi di classificazione con funzione legame logistica è un modello che assume la seguente forma:

$$\text{logit}(\pi_i) = f(x) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(x) \tag{2.11}$$

Le funzioni di base $h_k(x)$ sono funzioni in C o prodotti di due o più funzioni in C. Il numero delle coppie di funzioni di base da includere nel modello è rappresentato da K.
La stima del vettore dei parametri β e la definizione delle funzioni h_k può essere ottenuta attraverso la seguente procedura ricorsiva:

- *Crescita:* si indica con M l'insieme delle funzioni incluse nel modello e C l'insieme di quelle candidate
  - si include inizialmente nel modello la sola intercetta inizializzando $K = 0$ e $h_0 = 1$
  - ad ogni passo successivo $K+1$ si includono nel modello la coppia di funzioni che massimizza un criterio di adattamento ai dati:

$$\hat{\beta}_{k+1} h_l(x) \cdot (x_j - \zeta)_+ + \hat{\beta}_{k+2} h_l(x) \cdot (\zeta - x_j)_+ \qquad h_l \in M \tag{2.12}$$

  i coefficienti $\hat{\beta}_{k+1}$ e $\hat{\beta}_{k+2}$ sono stimati attraverso i minimi quadrati;

- *Potatura:* il modello fino a qui ottenuto è volutamente sovradattato. Mediante perciò un altro criterio come quello della *convalida incrociata generalizzata* vengono selezionate le funzioni di base che saranno incluse nel modello finale.

Nella Tabella 2.5 sono riportati i risultati ottenuti dopo aver adattato il modello $MARS$ ai dati nel mese di riferimento dicembre 2019.

2.1.6 Gradient Boosting

Il *gradient boosting* è un algoritmo di *machine learning* utilizzato sia per problemi di regressione che di classificazione. Sia $f(x)$ la funzione che si vuole approssimare, rappresentata come combinazione lineare di funzioni di base:

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m) \tag{2.13}$$

dove β_m sono i coefficienti ignoti della combinazione lineare mentre $b(x, \gamma) \in \mathbb{R}$ sono funzioni semplici con agomento multivariato x e caratterizzate dai parametri γ anch'essi ignoti. Una stima dei parametri può essere ottenuta minimizzando la seguente funzione di perdita che cambia a seconda della natura della variabile risposta:

$$\underset{\{\beta_m, \gamma_m\}_i^M}{\arg\min} = \sum_{i=1}^{N} L\left(y_i,\ \sum_{m=1}^{M} \beta_m\ b(x; \gamma_m)\right). \tag{2.14}$$

Essendo troppo onerosa da un punto di vista computazionale la minimizzazione della (2.14), si procede minimizzando una singola funzione di base alla volta:

$$\min_{(\beta,\gamma)} = \sum_{i=1}^{N} L\left(y_i,\ \beta\ b(x; \gamma)\right) \tag{2.15}$$

Per minimizzare la 2.14 si utilizza l'algoritmo di *Forward Stepwise Additive Modelling* dove ad ogni m-esima iterazione per $m = 1, ..., M$ dopo aver stimato $b(x, \gamma_m)$, si ricava il corrispondente β_m da sommare alla funzione del passo precedente ovvero $f_{m-1}(x)$. Si ottiene infine la stima della funzione all'm-esimo passo, $f_m(x)$, attraverso la seguente minimizzazione:

$$L\big(y_i,\ f_{m-1}(x_i) + \beta\ b(x_i, \gamma)\big) = \big(y_i - f_{m-1}(x_i) - \beta\ b(x_i, \gamma)^2\big) \tag{2.16}$$

Dove $y_i - f_{m-1}(x_i)$ sono i residui r_{im} all'm-esimo passo.
L'espansione 2.13 può dare luogo a vari modelli, come le reti neurali (Rumelhart et al. 1986) o MARS (Friedman 1991), a seconda delle forme particolari che essa può assumere. Si considera ora il caso in cui le funzioni di base scelte $b(x, \gamma_m)$ sono gli alberi di regressione che possono essere scritti formalmente come:

$$T(x, \Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j) \tag{2.17}$$

dove $\Theta = \{R_j, \gamma_j\}_{j=1}^{J}$ rappresentano i parametri ignoti rispettivamente indicanti le regioni terminali e i parametri. La combinazione di diversi alberi è data da

$$f_M(x) = \sum_{m=1}^{M} T(x, \Theta_m) \tag{2.18}$$

dove

$$\Theta_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L\left(y_i,\ f_{m-1}(x_i) + T(x_i, \Theta_m)\right). \tag{2.19}$$

La 2.19 viene risolta numericamente attraverso l'algoritmo *Gradient Tree Boosting* descritto in Alg.1: Per una trattazione esaustiva di tale procedura si veda ad esempio Hastie et al. (2001).
Nello specifico il pacchetto R **gbm** (Ridgeway et al. 2013), utilizzato nelle analisi, è stato implementato secondo tale algoritmo con una modifica proposta da Friedman (2002). Il parametro relativo al *tasso di apprendimento* della procedura di boosting è stato fissato pari a $v = 0.1$ mentre il numero di iterazioni è stato scelto pari a $M = 1000$ dopo varie prove. La funzione di perdita in questo caso è la devianza di una distribuzione binomiale.
Attraverso i risultati ottenuti dall'adattamento del modello ai dati, si giunge alle stesse conclusioni già discusse nelle sezioni precedenti. Dal grafico di importanza relativa del *gradient boosting* in Figura 2.3 si è infatti notato che le variabili maggiormente importanti nel prevedere la risposta sono anche in questo caso quelle che fanno riferimento ai due periodi che precedono il trimestre in cui il cliente è insolvente.

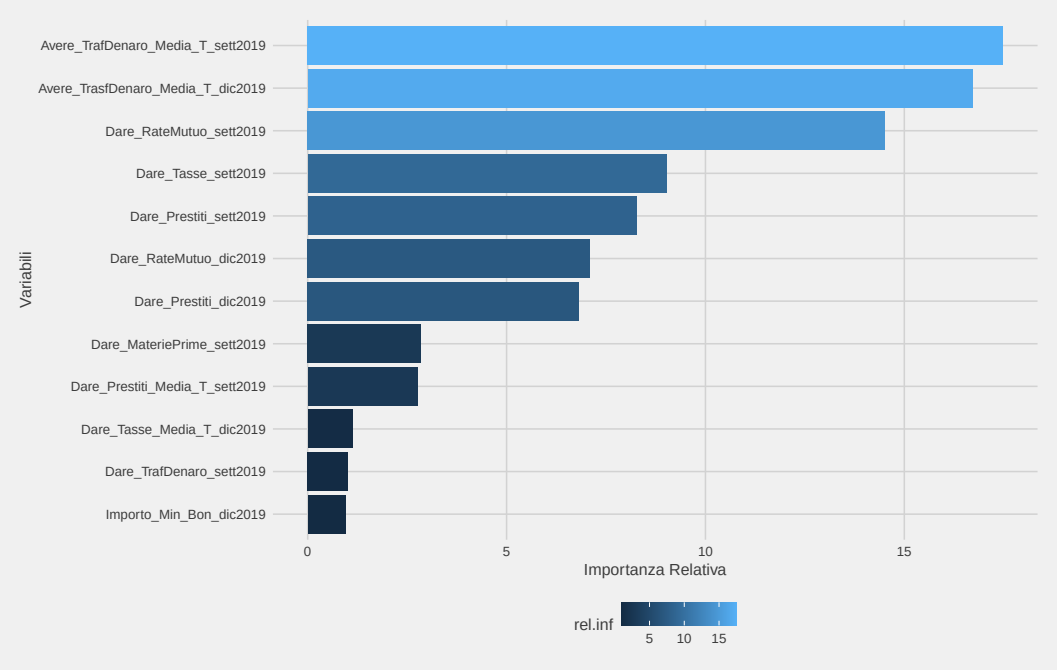

Figura 2.3: Grafico importanza relativa variabili gradient boosting

Algorithm 1 Gradient Tree Boosting

1. Inizializzazione $f_0(x) = \arg\min_\gamma \frac{1}{n} \sum_{i=1}^{N} L(y_i, \gamma)$

2. Per m=1 a M:

   (a) Per i=1 a N calcolare: $\frac{\partial \sum_{i=1}^{N} L(y_i; f(x_i))}{\partial f(x_i)} = f(x_i) - y_i = r_{im}$
   Dove $f(x_i) - y_i$ sono i residui interpretabili come gradiente negativo.

   (b) Adattare un albero di regressione a r_{im} e ottenere le regioni terminali R_{ij}, $j = 1, ..., J_m$

   (c) Per $j = 1, 2, ..., J_m$ calcolare:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_j \in R_{jm}} L(y_i,\ f_{m-1}(x_i) + \gamma)$$

   (d) Aggiornare $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

3. Ottenere $\widehat{f(x)} = f_M(x)$

2.2 Confronto tra modelli

Per verificare la capacità previsiva dei modelli adattati e per confrontare le loro prestazioni viene utilizzato l'*insieme di verifica*. Dal momento che le classi della variabile risposta Y sono sbilanciate, per non perdere ulteriori informazioni, invece di ribilanciare l'insieme di stima, si è ritenuto opportuno scegliere una *soglia* appropriata che separi la frontiera di classificazione. Come valore per tale soglia è stato scelto il valore proporzione della classe in minoranza (0.02). In questo modo si assegna un costo maggiore alla classificazione dell'evento raro e non si rischia di dare maggior peso alla modalità più frequente. Nelle analisi si sono scelti tre valori per la soglia (0.01, 0.02, 0.05) al fine di avere un confronto completo della capacità previsiva dei modelli.
Data la *matrice di confusione* nella Tabella 2.4, per confrontare le procedure di classificazione, tenendo in considerazione anche lo sbilanciamento tra classi, sono utilizzate le seguenti misure di *adeguatezza* delle previsioni:

- *Precisione*: frazione di veri positivi sul totale dei predetti positivi;
- *Richiamo*: i veri positivi sul totale di quelli effettivamente positivi;
- *False discovery rate (F1)*: rappresenta la media armonica di precisione e richiamo;
- *Tasso Errata Classificazione*: frazione della somma dei falsi positivi e falsi negativi sul totale di osservazioni nell'insieme di verifica.

Previsti	*Osservati*	
	-	+
-	*Veri Negativi*	*Falsi Negativi*
+	*Falsi Positivi*	*Veri Positivi*

Tabella 2.4: Matrice di errata classificazione.

Un altro strumento utilizzato per valutare la bontà di previsione dei modelli è la curva di *ROC*. Essa fornisce un grafico della qualità previsiva del modello nel suo complesso. In particolare l'area sotto la curva (AUC) rappresenta una misura della qualità globale del modello. Tale indicatore non dipende dalla soglia selezionata a differenza degli indici prima descritti.

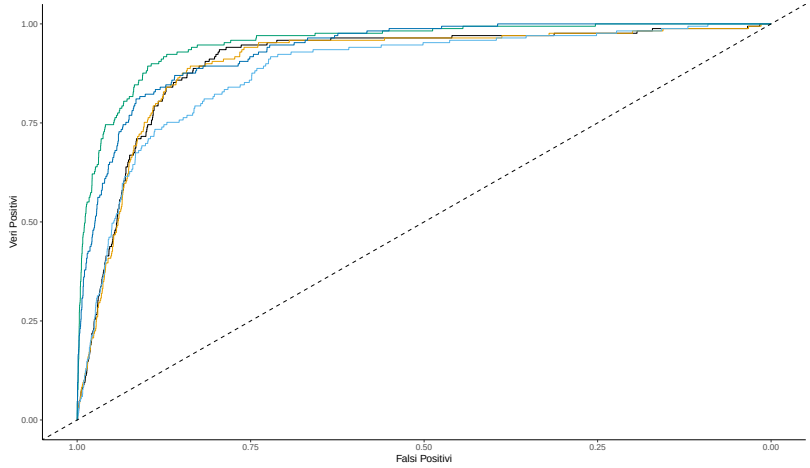

Figura 2.4: Curve di ROC: (verde) gradient boosting, (blu) MARS, (giallo) lasso logistico, (nero) logistico, (azzurro) modello lineare

Dai risultati riportati nella Tabella 2.5 e dal confronto delle curve di ROC in Figura 2.4 si osserva che tutti i modelli considerati presentano un ottimo adattamento ai dati. Le informazioni derivanti dalle transazioni possiedono quindi una buona capacità esplicativa nel prevedere la risposta. Le prestazioni per tutti i modelli, tuttavia, non sono particolarmente soddisfacenti considerando gli indici F1, Precisione o Richiamo. Questo è dovuto al fatto che le classi della risposta sono molto sbilanciate e di conseguenza i modelli hanno difficoltà nel classificare correttamente la classe dell'evento raro.

Tra tutti, il *gradient boosting* è il modello preferibile in termini previsivi, sia in termini di *AUC* che per l'indice F1 per la soglia 0.05. Tale metodo risulta di difficile interpretazione, tuttavia in questo caso, come affermato precedentemente, l'interpretabilità risulta complessa anche per un modello più semplice come quello logistico o lineare. É pertanto preferibile una tecnica che porta a previsioni migliori come il *gradient boosting* o il MARS.

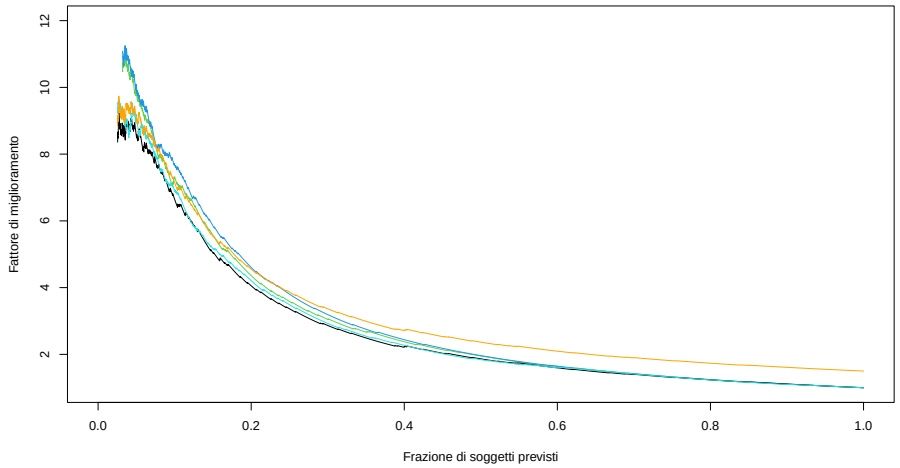

Figura 2.5: Curve di Lift: (verde) gradient boosting, (blu) MARS, (giallo) lasso logistico, (nero) logistico, (azzurro) modello lineare

In Figura 2.5 viene riportato il confronto tra curve Lift per ciascun modello. La curva lift fornisce un'indicazione di quanto un certo metodo migliora le previsioni rispetto al metodo che opera casualmente. Il *lift*, o fattore di miglioramento, è definito per una certa soglia come il rapporto tra la previsione ottenuta dal modello e quella che si otterrebbe utilizzando le proporzioni iniziali. Se il fattore di miglioramento è ad esempio 8 vuol dire che se viene preso il 10% della popolazione il modello prevede 8 volte meglio rispetto al metodo casuale.

	Soglia	Precisione	Richiamo	F1	Errore	AUC
Lineare	*0.01*	0.109	0.123	0.110	0.020	0.869
	0.02	0.135	0.127	0.128		
	0.05	0.160	0.233	0.187	0.022	
Lineare ridotto	*0.01*	0.110	0.122	0.111	0.021	0.871
	0.02	0.137	0.128	0.130	0.020	
	0.05	0.182	0.213	0.189	0.019	
Logistico	*0.01*	0.137	0.130	0.138	0.021	0.889
	0.02	0.175	0.221	0.182	0.015	
	0.05	0.246	0.253	0.252	0.021	
Logistico Ridotto	*0.01*	0.140	0.137	0.142	0.021	0.891
	0.02	0.186	0.231	0.187	0.013	
	0.05	0.247	0.253	0.254	0.021	
Lasso Logistico	*0.01*	0.101	0.101	0.102	0.021	0.909
	0.02	0.192	0.201	0.224	0.020	
	0.05	0.231	0.201	0.232	0.021	
MARS	*0.01*	0.132	0.214	0.163	0.023	0.912
	0.02	0.154	0.243	0.185	0.016	
	0.05	0.276	0.253	0.263	0.019	
Gradient Boosting	*0.01*	0.231	0.211	0.200	0.023	0.923
	0.02	0.359	0.310	0.370	0.016	
	0.05	0.381	0.321	0.381	0.010	

Tabella 2.5: Tabella di confronto degli indici di Precisione, Richiamo, F1, Tasso Errata Classificazione e AUC, segmento Grandi Imprese, mese di riferimento dicembre 2019

Capitolo 3

Modelli per Dati Funzionali

3.1 Caratteristiche dei dati funzionali

In diverse applicazioni di analisi dei dati, sono disponibili misurazioni ripetute nel tempo per lo stesso processo aleatorio. I dati considerati in questa tesi, ossia le transazioni bancarie effettuate da ciascun cliente periodicamente, ne sono un esempio.
Un approccio che fornisce strumenti utili per l'analisi di questi oggetti complessi è la FDA (*Functional Data Analysis*). Tale approccio possiede un'importante differenza da quello tradizionale: ogni unità statistica non è più uno scalare o un vettore, bensì una funzione. Le osservazioni disponibili sono quindi considerate realizzazioni di variabili casuali che giacciono in uno spazio funzionale. Nella pratica, tali dati sono osservati in corrispondenza di un insieme finito di punti del dominio $t_1, ..., t_T \in \mathcal{T} \subseteq \mathbb{R}$. Nel contesto considerato in questa tesi, le unità statistiche, ovvero le varie tipologie di transazioni bancarie di ogni cliente, sono delle serie storiche trimestrali. Nello specifico, ciascuna curva viene osservata in istanti temporali equispaziati: le misurazioni avvengono ogni tre mesi nell'arco di un anno ($T = 5$).
Ciascuna osservazione è generata da un processo così definito:

$$x_i(t) = f_i(t) + \epsilon_i(t), \qquad i = 1, ..., N, \qquad t = t_1, ..., t_T \tag{3.1}$$

in cui $\epsilon_i(t)$ rappresenta l'errore di misura, $x_i(t)$ è la generica osservazione, N indica il numero di curve disponibili, in questo caso $N = 23624$ per le varie tipologie di transazione. La generica funzione è rappresentata da $f_i : \mathcal{T} \to \mathbb{R}$. In Figura 3.1 è riportato un esempio di 20 funzioni grezze estratte casualmente e relative alle

transazioni effettuate dai clienti per l'avvio di prestiti.

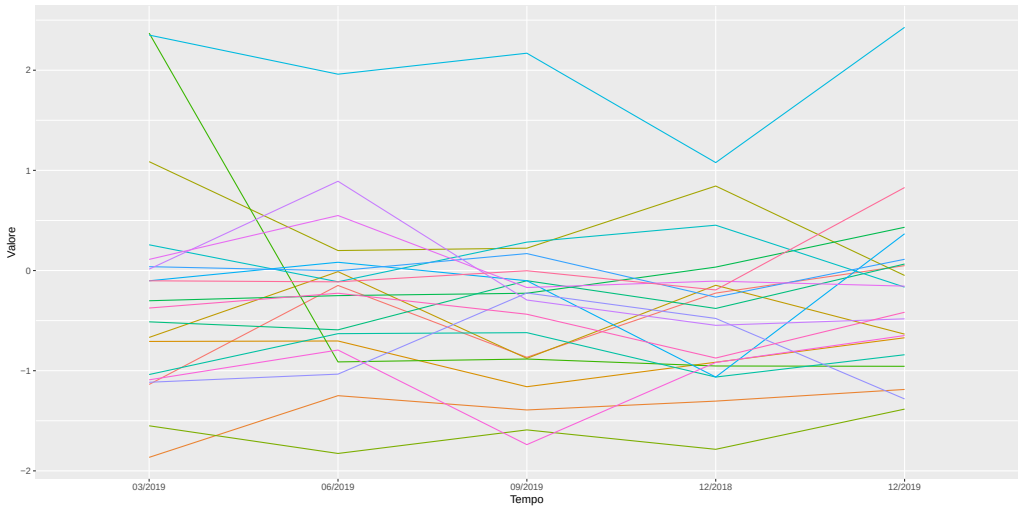

Figura 3.1: Funzioni grezze delle transazioni *Dare_Prestiti* di 20 *CostumerID* con $T = 5$. Segmento Grandi Imprese.

3.1.1 Lisciamento dei dati funzionali

Per ottenere una rappresentazione accurata di $f_i(t)$, l'approccio maggiormente utilizzato in letteratura è quello dell'espansione in funzioni di base. Queste ultime permettono di rappresentare tali oggetti complessi, in quantità finito dimensionali tramite combinazioni lineari di funzioni di basi note e linearmente indipendenti $\phi_k(t)$ pesate con i coefficienti ignoti c_k per $k = 1, ..., K$. Come affermato da Ramsay & Silverman (2005), per K sufficientemente elevato, un sistema di funzioni di base riesce ad approssimare in maniera adeguata qualsiasi funzione. Applicando l'approccio appena descritto, la generica funzione $f(t)$ può essere rappresentata nel seguente modo:

$$f(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \Phi^T \mathbf{c} \tag{3.2}$$

La definizione di un sistema di funzioni di base opportuno per la funzione oggetto di studio è cruciale per ottenere un'approssimazione sufficientemente accurata di quest'ultima. A seconda della natura dei dati a disposizione, si possono scegliere diverse tipologie di funzioni di base. Nello specifico, per dati periodici usualmente la scelta ricade sulle basi di Fourier, mentre per dati senza periodicità sulle *splines*, le quali

approssimano localmente la funzione mediante un polinomio di grado m. In particolare, si considera una specificazione di queste ultime ovvero le *B-splines*, utilizzate per approssimare le curve anche in queste analisi. Per una trattazione completa si veda Ramsay & Silverman (2005). Tali funzioni di base sono molto flessibili e permettono di approssimare la maggior parte delle forme delle funzioni. In Figura 3.2 viene rappresentata una base di funzioni *B-splines* di ordine 4 con 1 nodo interno.

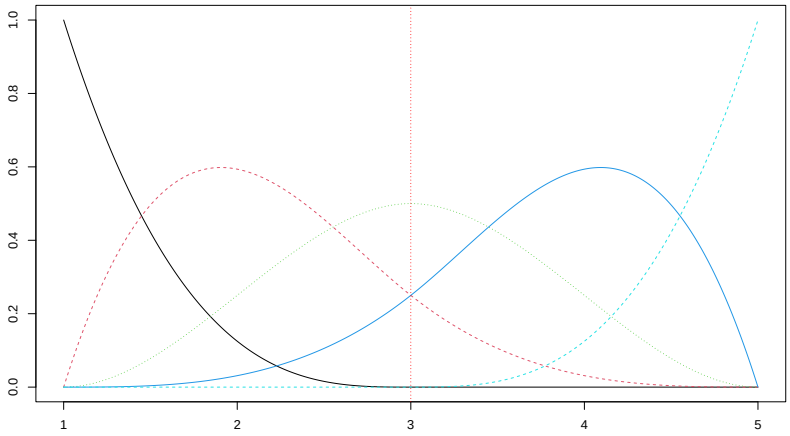

Figura 3.2: Base di funzioni *B-splines* di ordine $J = 4$ con $G = 1$ nodo interno

Una volta stabilito il sistema di funzioni di base, ossia le *B-splines* in questo caso, l'obiettivo è quello di approssimare le N funzioni disponibili lisciandole ed eliminando l'errore. Usando una notazione matriciale e indicando con $\boldsymbol{x} = [x_1, ..., x_N]$ la matrice $T \times N$ contenente i valori osservati, la 3.2 può essere scritta nel seguente modo:

$$\boldsymbol{x} = \Phi\mathbf{C} + \epsilon \tag{3.3}$$

in cui $\Phi = [\phi_1, ..., \phi_K]$ è la matrice $T \times K$ delle funzioni di base per ogni istante di osservazione, tale che $\phi_K = (\phi_K(t_1), ..., \phi_K(t_T))^T$. I vettori colonna per l'i-esimo soggetto $\mathbf{c}_i = (c_{1i}, ..., c_{Ki})^T$ della matrice $\mathbf{C} = [c_1, ..., c_N]$ dei coefficienti ignoti di dimensione $K \times N$, vengono stimati mediante i minimi quadrati ordinari:

$$\hat{\boldsymbol{c}}_i = (\Phi^T\Phi)^{-1}\Phi^T x_i \tag{3.4}$$

É possibile utilizzare anche metodi di regolarizzazione con penalizzazione, ad esempio

L_2 al fine di penalizzare curve poco lisce, ossia "l'irregolarità" della funzione:

$$\text{PEN}_2 = \int D^2[f(t)]^2 dt \tag{3.5}$$

Tale penalizzazione agisce sull'integrale della derivata seconda della funzione f.
La stima dei coefficienti per l'i-esimo cliente, utilizzando il criterio dei minimi quadrati penalizzati e usando la 3.5 come penalità, si ricava in questo caso nel seguente modo:

$$\hat{\boldsymbol{c}}_i = (\Phi^T\Phi + \lambda R)^{-1}\Phi^T x_i \tag{3.6}$$

Dove R rappresenta la matrice di penalizzazione tale che $R = \int D^2\Phi(t)D^2\Phi^T(t)dt$. Il parametro di lisciamento λ e il numero di funzioni di base vengono scelti solitamente via convalida incrociata. Nel dettaglio, le curve delle transazioni sono state lisciate tramite *B-splines* cubiche con il numero K di funzioni di base scelto pari a 5 con 3 nodi equispaziati. Si è provato a scegliere il parametro λ via convalida incrociata generalizzata, tuttavia il parametro selezionato da tale procedura portava ad avere delle curve lisciate che non approssimavano in maniera adeguata l'andamento delle funzioni grezze, per questo motivo si è scelto λ con il cosiddetto 'metodo ad occhio' come suggerito da Azzalini & Scarpa (2012) pari a 0.01.
Il vettore delle osservazioni lisciate per l'i-esimo soggetto è pari a:

$$\hat{x}_i = \Phi\hat{\boldsymbol{c}}_i = \Phi(\Phi^T\Phi + \lambda R)^{-1}\Phi^T x_i \tag{3.7}$$

In Figura 3.3, si mostrano le stesse curve della Figura 3.1 di 10 *CostumerID*, ma in questo caso vengono riportate le corrispondenti funzioni lisciate tramite le funzioni di base *B-splines* di ordine 4, $\lambda = 0.01$.

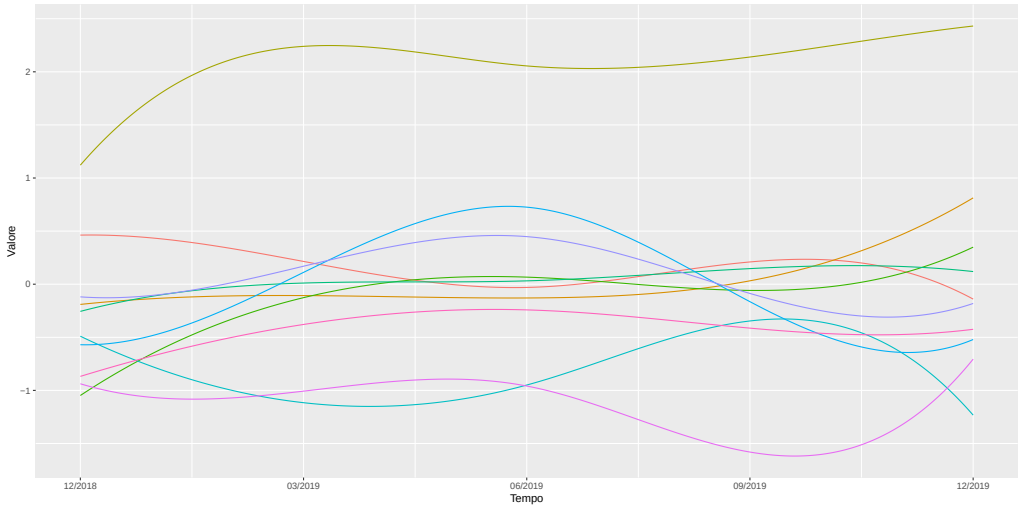

Figura 3.3: Funzioni lisciate delle transazioni *Dare_Prestiti* di 10 *CostumerID*.

3.1.2 Statistiche descrittive per dati funzionali

Gli indici di posizione e variabilità, come media e varianza, possono essere generalizzati anche al caso funzionale. La media funzionale è definita come la media degli N valori osservati, valutata in ogni tempo di osservazione t:

$$\bar{x}(t) = \frac{1}{N} \sum_i x_i(t) \tag{3.8}$$

In maniera analoga la varianza si ottiene

$$\text{var}(t) = \frac{1}{(N-1)} \sum_i [x_i(t) - \bar{x}(t)]^2 \tag{3.9}$$

In Figura 3.4 e 3.5 , vengono confrontate le medie funzionali di due covariate relative alle transazioni in *Dare* e *Avere*, dei *CostumerID* che sono in stato di 'allarme' a dicembre 2019 (arancione), con coloro che non lo sono (azzurro). Dal grafico viene confermato quanto osservato in fase esplorativa: le curve medie per i due gruppi si differenziano in modo evidente negli ultimi due periodi di osservazione. Nel dettaglio, per i *CostumerID* che sono in *Early Warning* a dicembre 2019, si osserva una decrescita nelle curve medie delle transazioni in *Dare* (crescita in *Avere*) nei sei mesi che precedono il trimestre di insolvenza. Questo può indicare un inizio di 'sofferenza' di liquidità. Un cambiamento nell'andamento delle stesse curve medie, con una crescita

in *Dare* (descrescita in *Avere*), emerge tre mesi prima del periodo di passaggio allo Stadio 2. Questo comportamento negli ultimi due periodi per ciascuna covariata, può essere spiegato con un effetto di compensazione tra i due trimestri. Esplorando gli stessi grafici per le rimanenti covariate, viene osservato che l'andamento delle funzioni medie dipende dalla tipologia di transazione e non dal flusso finanziario.

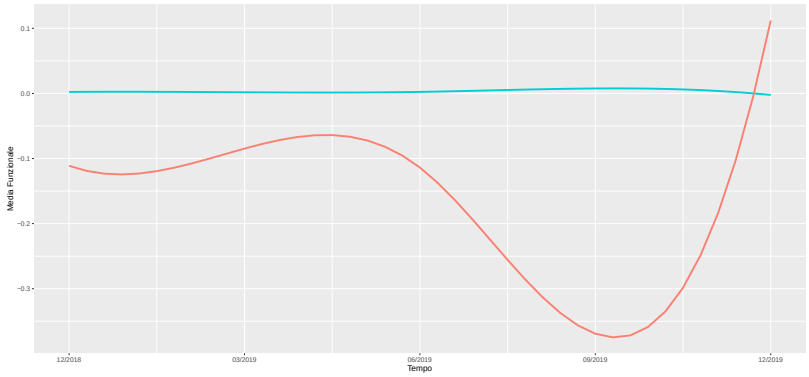

Figura 3.4: Funzioni medie delle transazioni *Importo_Min_Bon* per i due gruppi. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

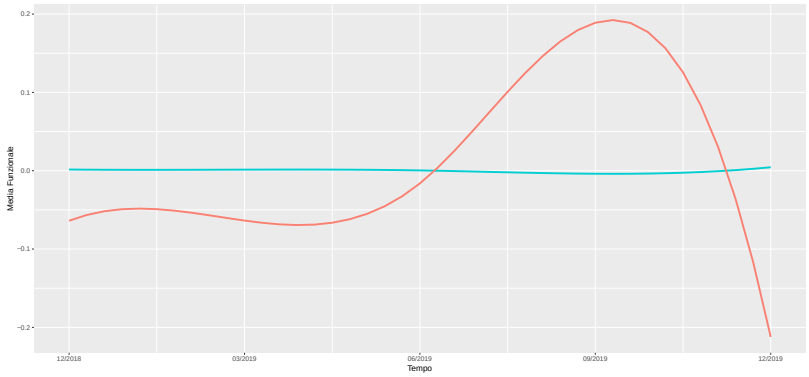

Figura 3.5: Funzioni medie delle transazioni *Avere_Vendite_Media_T* per i due gruppi. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

3.2 Previsione della risposta con predittori funzionali

Nelle analisi che seguono, l'informazione proveniente dalla dipendenza temporale tra le transazioni effettuate dallo stesso *CostumerID* in tempi differenti viene utilizzata per identificare i clienti a maggior rischio di insolvenza. La natura longitudinale dei dati non viene quindi ignorata, ma considerata parte integrante nella costruzione dei modelli che vengono adottati. Considerare tale caratterista nelle metodologie, può portare ad avere una più completa comprensione del fenomeno d'interesse.
L'obiettivo dell'analisi è la classificazione, attraverso le transazioni periodiche, dell'i-esimo cliente nella classe 1 se esso è in *Early Warning* nel mese di dicembre 2019 oppure nella classe 0. Si vuole perciò trovare una relazione tra la variabile risposta scalare, ovvero la variabile $Target$, indicata con Y_i, e le covariate funzionali disponibili.
Analogamente all'analisi precedente, sono state considerate le informazioni delle covariate fino a dicembre 2019. Le operazioni preliminari sono uguali a quelle riportate nella Sezione 2.1.1.
Per non *sovradattare* il modello ai dati, il dataset è stato diviso per il 60% in *insieme di stima* utilizzato per la stima dei modelli e il restante 40% in *insieme di verifica* impiegato per valutare le prestazione dei modelli.
Per le stesse motivazioni mezionate nel Capitolo 1, si riporta l'analisi del solo segmento Grandi Imprese.

3.2.1 Modello di Regressione Lineare funzionale con risposta scalare

La regressione funzionale costituisce un'ambito di ricerca in continua evoluzione e d'interesse per molti ricercatori. Un primo approccio nello studio di questi tipi di modelli è stato introdotto da Ramsay & Dalzell (1991), i quali hanno sviluppato il modello lineare funzionale (FLM). Questo modello può essere visto come un'estensione del tradizionale modello lineare multivariato. Tuttavia il modello FLM possiede una rilevante differenza da quest'ultimo: le covariate non sono dei vettori ma delle funzioni. Il modello in questione con risposta scalare Y_i e covariate funzionali viene

così definito:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \int x_{ij}(t)\beta_j(t)dt + \epsilon_i \qquad i = 1, ..., N \tag{3.10}$$

Nella 3.10 si considera l'integrale su tutto il dominio dell'osservazione funzionale come indicatore riassuntivo. Nel dettaglio, $x_{ij}(t)$ rappresenta l'i-esima osservazione del j-esimo predittore funzionale per l'istante di osservazione t. Nel caso in esame sono disponibili 11 covariate funzionali, $p = 11$, delle varie tipologie di transazioni.
L'approccio utilizzato per giungere ad una stima dei coefficienti ignoti consiste nel rappresentare le quantità funzionali $x_{ij}(t)$, con un sistema di funzioni di base come quello descritto nella 3.2. In particolare si assume che i coefficienti c_{ij} dell'i-esimo soggetto della j-esima variabile esplicativa, siano stimati in una fase precedente come riportato nella 3.6. Le funzioni $\beta_j(t)$ sono rappresentate:

$$\beta_j(t) = \Theta_j(t)\mathbf{b}_j \qquad j = 1, ..., p \tag{3.11}$$

Dove $\Theta_j(t)$ rappresenta la matrice $T \times K_\beta$ delle funzioni di base e $\mathbf{b}_j$ è il vettore K_β-dimesionale dei coefficienti ignoti. Tipicamente viene scelto lo stesso sistema di funzioni di base sia per le covariate che per i coefficienti funzionali, in questo caso le *B-spline*. Una stima delle quantità $\mathbf{b}_j$ può essere ricavata attraverso i minimi quadrati penalizzati:

$$\text{SSE}_\lambda = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \int \beta_j(t)x_{ij}(t)dt \right)^2 + \sum_{j=1}^{p} \lambda_j \int [L_j\beta_j(t)]^2 dt \tag{3.12}$$

in cui usualmente $L_j = D^2$. Indicando la generica osservazione per la covariata j-esima con $x_{ij} = \int x_{ij}(t)\Theta(t)dt$ e con Z_i il vettore riga della matrice Z tale che $Z_i = [1 \quad x_{i1} \quad ... \quad x_{ip}]$, i coefficienti ignoti $\boldsymbol{\zeta} = [\beta_0 \quad \boldsymbol{b}_1 \quad ... \quad \boldsymbol{b}_j]^T$ vengono stimati attraverso l'equazione:

$$\hat{\boldsymbol{\zeta}} = (Z^T Z + R_\theta)^{-1} Z^T y \tag{3.13}$$

dove R_θ è la matrice di penalizzazione i cui elementi nella diagonale sono costituiti dai termini $\lambda_j R_{j\theta}$ con $R_{j\theta} = \int L_j\Theta_j(t)L_j\Theta_j(t)^T dt$.
Come descritto nel dettaglio da Azzalini & Scarpa (2012) e come precedentemente visto nel Capitolo 2, nel caso in cui la variabile risposta Y_i ha due classi 0 e 1, si può

decidere di utilizzare un modello di regressione lineare per ottenere la regola:

$$\hat{y} = Z\hat{\boldsymbol{\zeta}} \tag{3.14}$$

con la quale classificare una volta ottenuta la stima del vettore degli N scalari $\hat{y}$, il cliente i-esimo nel gruppo 1 se lo scalare $\hat{y}_i > s$ o 0 se $\hat{y}_i < s$, dove s è la soglia di discriminazione scelta. Tale procedura, come affermato nel Capitolo 2, risulta poco ottimale difatti tale modello viene usato come *benchmark*, ovvero una sorta di termine di paragone, una soglia al di sotto del quale è difficile scendere.
I parametri di lisciamento λ_j per i coefficienti $\beta_j(t)$ sono stati selezionati via convalida incrociata generalizzata.
Nelle Figure 3.6 e 3.7 vengono rappresentati i coefficienti funzionali di due covariate con i relativi intervalli di confidenza al 95% (linea tratteggiata). Più nel dettaglio si tratta di intervalli di confidenza punto a punto che danno una valutazione della variabilità locale della curva. Come si può notare, entrambi i coefficienti sono rilevanti nel prevedere la risposta negli ultimi due trimestri di osservazione, dal momento che gli intervalli non comprendono lo zero.
Nel modello FLM l'andamento dei singoli coefficienti nel tempo esprime un effetto negativo (o positivo) a seconda dell'istante di tempo a cui ci si condiziona sulla risposta Y_i, fermo restando il valore delle ulteriori esplicative del modello.

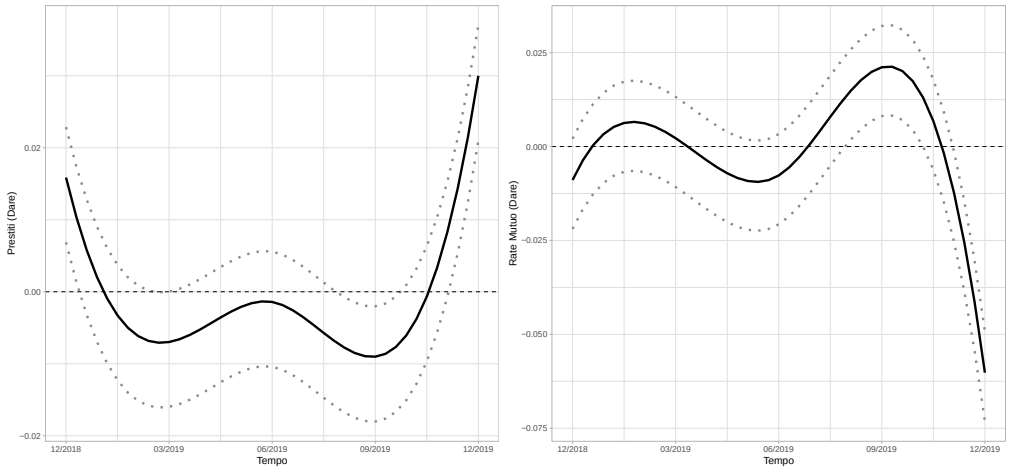

Figura 3.6: Coefficiente funzionale modello lineare funzionale (FLM) delle transazioni *Dare_Prestiti* e *Dare_RateMutuo_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

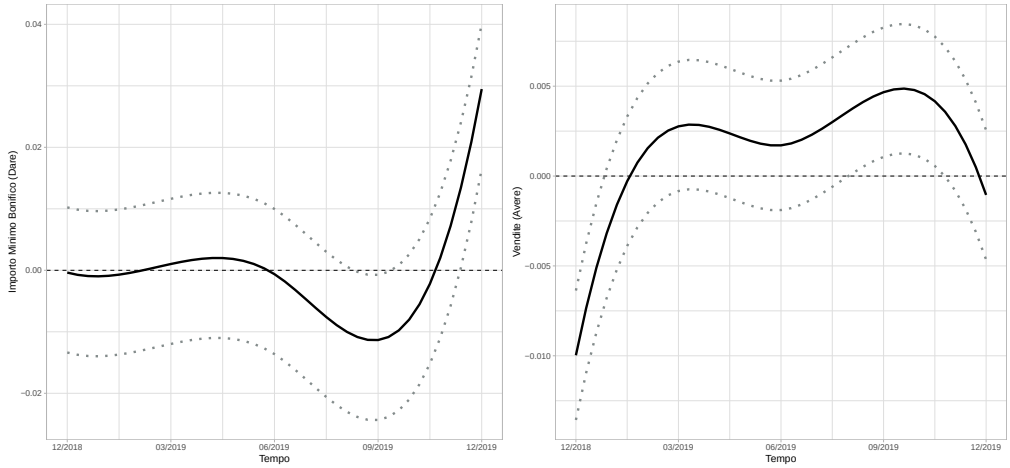

Figura 3.7: Coefficiente funzionale modello lineare funzionale (FLM) delle transazioni *Importo_Min_Bon* e *Avere_Vendite_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

3.2.2 Modello di Regressione Logistica funzionale

Un'estensione del modello lineare funzionale è il modello lineare funzionale generalizzato FGLM introdotto da Müller (2005). Considerando la funzione di legame logistica la 3.10 può essere modificata nella:

$$\log \frac{\mathbb{P}(Y_i = 1|\tilde{X}_i)}{\mathbb{P}(Y_i = 0|\tilde{X}_i)} = \beta_0 + \sum_{j=1}^{p} \int x_{ij}(t)\beta_j(t)dt + \epsilon_i \qquad i = 1, ..., N \tag{3.15}$$

in cui $\tilde{X}_i$ è il vettore delle covariate funzionali per l'i-esimo cliente per $j = 1, .., p$. La distribuzione condizionata di Y_i, date le covariate funzionali, è una Bernoulli. Come dimostrato da Escabias et al. (2007) per garantire l'identificabilità del modello 3.15 è necessario assumere che la variabili funzionali siano derivabili fino al secondo ordine e gli autovalori delle loro covarianze siano diversi da zero.

Analogamente a quanto visto per il modello FLM, si rappresentano le quantità funzionali, ossia le covariate e i coefficienti funzionali con un sistema di funzioni di basi, come riportato nelle 3.2 e 3.11. In queste analisi sono state scelte le *B-splines* per entrambe le quantità funzionali.

Le stime dei coefficienti ignoti delle espansioni di base dei $\beta_j(t)$ possono essere ottenute tramite i minimi quadrati pesati iterati (IWLS), sfruttando la funzione di verosimiglianza del modello associato come descritto in Müller & Stadtmüller (2005).

Solo il parametro relativo al numero di funzioni di basi dei $\beta_j(t)$ è stato selezionato via convalida incrociata generalizzata.
Nelle Figure 3.8 e 3.9, sono riportati i coefficienti funzionali di quattro covariate con gli intervalli di confidenza punto a punto al 95% (linea tratteggiata) calcolati via *bootstrap*. Anche in questo caso si può osservare, fatta eccezione della covariata riguardante le vendite i cui intervalli di confidenza non comprendono lo zero anche nel primo periodo, che le transazioni relative agli ultimi due trimestri sono rilevanti nel prevedere la risposta. Questi risultati confermano quanto precedentemente affermato: la risposta è influenzata in particolar modo dai due periodi che precedono il trimestre di insolvenza con andamenti differenti in base alla tipologia di transazione. L'andamento dei coefficienti nel tempo per lo stesso tipo di transazione è uguale per entrambi i modelli FLM e FGLM, cambiano solo la grandezza delle stime dei coefficienti.
Nel modello FGLM l'andamento dei singoli coefficienti nel tempo esprime un effetto negativo (o positivo) a seconda dell'istante di tempo a cui ci si condiziona sul logaritmo della quota, ossia su $\log \frac{\mathbb{P}(Y_i=1|\tilde{X}_i)}{\mathbb{P}(Y_i=0|\tilde{X}_i)}$, a parità delle altre esplicative.

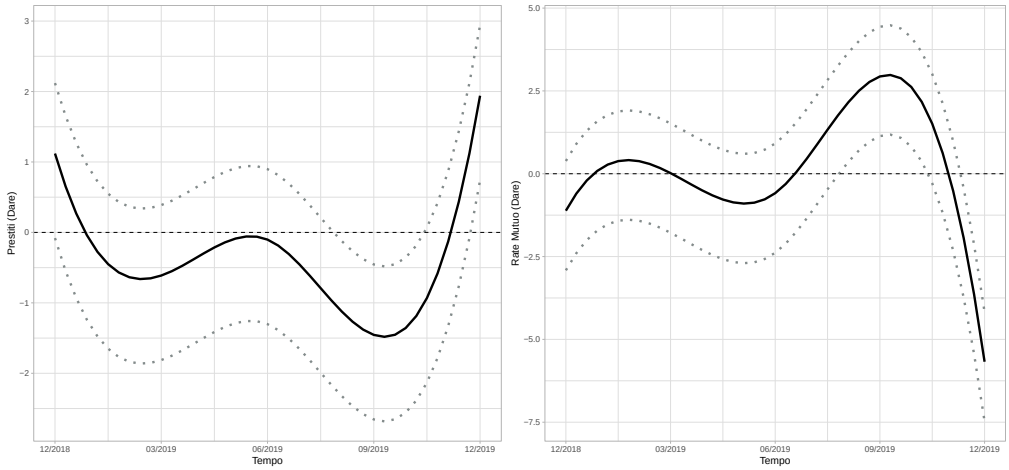

Figura 3.8: Coefficiente funzionale modello logistico funzionale (GFLM) delle transazioni *Dare_Prestiti* e *Dare_RateMutuo_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

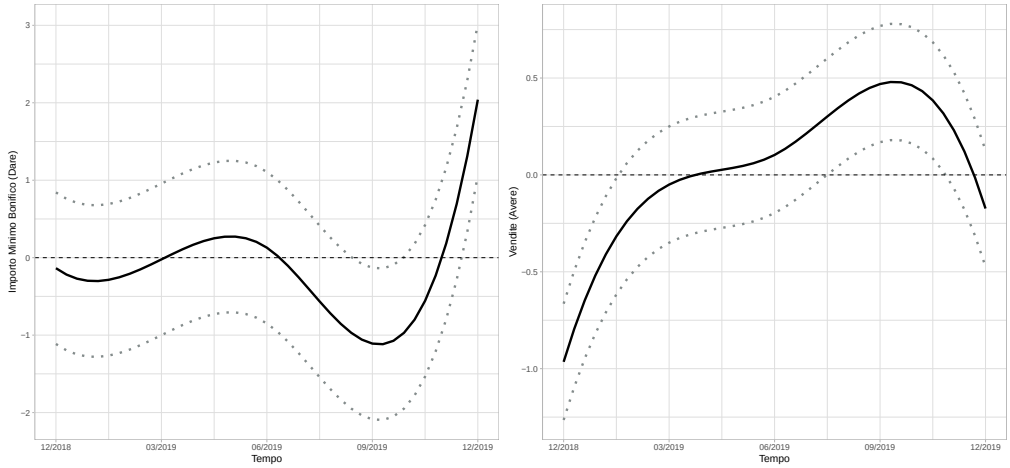

Figura 3.9: Coefficiente funzionale modello logistico funzionale (FGLM) delle transazioni *Importo_Min_Bon* e *Avere_Vendite_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

3.2.3 Functional Linear Regression That's Interpretable (FLiRTI): caso multivariato

James et al. (2009) propongono la metodologia FLiRTI (Functional Linear Regression That's Interpretable) per la stima dei coefficienti funzionali di un modello di regressione FLM o più in generale FGLM. L'idea alla base di tale metodologia è uguale a quella presente nel lasso (Tibshirani 1996) ovvero la selezione automatica delle variabili. Il principale vantaggio del FLiRTI sta nel fatto che pone esattamente pari a zero il coefficiente $\beta(t)$ nelle regioni dove non vi è alcuna relazione tra quest'ultimo e la variabile risposta scalare Y_i. Questo metodo, tra l'altro, produce una stima interpretabile e accurata dei coefficienti funzionali.
Nella metodologia esposta nell'articolo si fa riferimento al solo caso univariato e in questa tesi tale procedura verrà estesa anche al caso multivariato.
Gli autori considerano il modello FLM della 3.10 nel caso univariato con risposta scalare e una covariata funzionale, dove il coefficiente funzionale $\beta(t)$ viene rappresentato tramite un sistema di funzioni di base. La curva $\beta(t)$ può quindi essere rappresentata nel seguente modo:

$$\beta(t) = \boldsymbol{B}(t)^T \boldsymbol{\eta} + \epsilon(t) \tag{3.16}$$

in cui $\mathbf{B}(t) = [b_1(t), ..., b_K(t)]$ è il vettore K-dimensionale delle funzioni di base, $\boldsymbol{\eta}$ è il vettore dei parametri ignoti da stimare e $\epsilon(t)$ rappresenta il termine d'errore. I coefficienti ignoti di $\beta(t)$ sono stimati mediante la procedura FLiRTI che si basa sul lasso o sul selettore di Dantzig (Candes & Tao 2007) Sostituendo la 3.16 al modello 3.10 si ottiene:

$$Y_i = \beta_0 + \boldsymbol{X}_i^T \boldsymbol{\eta} + \epsilon_i^* \tag{3.17}$$

dove $\boldsymbol{X}_i = \int X_i(t)\boldsymbol{B}(t)dt$ e $\epsilon_i^* = \epsilon + \int X_i(t)\epsilon(t)dt$. Pur non assumendo sparsità nel vettore $\boldsymbol{\eta}$, si ipotizza che uno o più ordini delle sue derivate siano pari a zero. In altre parole, $\beta^{(d)}(t) = 0$ per uno o più valori di $d = 0, 1, 2, ...$ e per un'ampia regione di t. Se $\beta^{(0)}(t) = 0$ allora $X(t)$ non ha effetti su Y. Condizionandoci ora al caso in cui $d = 2$ viene definito il vettore A:

$$A = [D^2\boldsymbol{B}(t_1), D^2\boldsymbol{B}(t_2), ..., D^2\boldsymbol{B}(t_T)]^T \tag{3.18}$$

dove D^d è l'operatore della d-esima differenza, in questo caso differenza seconda, tale che $D^2\boldsymbol{B}(t_j) = T^2[\boldsymbol{B}(t_j) - 2\boldsymbol{B}(t_{j-1}) + \boldsymbol{B}(t_{j-2})]$. Si ottiene quindi:

$$\boldsymbol{\gamma} = A\boldsymbol{\eta} \tag{3.19}$$

in cui $\gamma_j = T^2[\boldsymbol{B}(t_j)^T\eta - 2\boldsymbol{B}(t_{j-1})^T\eta + \boldsymbol{B}(t_{j-2})^T\eta]$ rappresenta un'approssimazione di $\beta^{(2)}(t_j)$. Assumendo la sparsità di γ_j, si avrà che $\beta^{(2)}(t_j) = 0$ in molti dei punti. Si otterranno di conseguenza delle stime di $\beta(t)$ lineari tranne nei punti in cui si annulla il coefficiente, corrispondenti ai valori nulli in γ_j. Posto $\boldsymbol{\eta} = A^{-1}\boldsymbol{\gamma}$ e si ottiene il modello FLiRTI:

$$\boldsymbol{Y} = V\boldsymbol{\gamma} + \epsilon^* \tag{3.20}$$

Dove $V = [1|XA^{-1}]$ è la matrice che incorpora 1 per l'intercetta β_0. Si ottiene $\hat{\boldsymbol{\gamma}}$ adattando il FLiRTI al modello 3.20, utilizzando il lasso (o il selettore di Dantzig) dopo aver standardizzato V. Dopo aver ricavato le quantità $\hat{\boldsymbol{\gamma}}$, si ottiene la stima FLiRTI per $\beta(t)$:

$$\hat{\beta}(t) = \boldsymbol{B}(t)^T\hat{\boldsymbol{\eta}} = \boldsymbol{B}(t)^T A^{-1}\hat{\boldsymbol{\gamma}}_{\boldsymbol{(-1)}} \tag{3.21}$$

in cui $\boldsymbol{\gamma}_{(-1)}$ rappresenta la stima di $\hat{\boldsymbol{\gamma}}$ dopo aver sottratto la stima dell'intercetta β_0. Il modello FLiRTI può essere facilmente esteso al caso in cui non si assuma la gaussianità nella distribuzione della risposta attraverso il GLM FLiRTI, il quale utilizza

il selettore di Dantzig Escabias et al. (2007) generalizzato nel caso di risposta non gaussiana nel seguente modo:

$$\min_\beta ||\boldsymbol{\beta}||_1 \quad s.t. \; |\boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{\mu})| \leq \lambda \quad j = 1, .., T \tag{3.22}$$

in cui $\mu = g^{-1}(\boldsymbol{X}\boldsymbol{\beta})$ con g funzione di legame. La 3.22 è risolta attraverso un'approccio iterativo come descritto in James et al. (2009).
In questa tesi, dal momento che si dispongono di p covariate, viene estesa la metodologia FLiRTI al caso multivariato mediante l'utilizzo dell'algoritmo *backfitting* come descritto nell'Algoritmo 2. In breve, viene ottenuta la stima di ogni $\hat{\boldsymbol{\beta}}_j(t)$ adattando il modello FLiRTI sui residui calcolati non considerando quest'ultimo coefficiente. Tale procedura viene ripetuta fintantoché le stime dei parametri non si stabilizzano, ovvero fino a quando la differenza tra le stime dei parametri all'iterazione $m-1$ e m è più piccola di una quantità stabilita. Viene utilizzato l'algoritmo implementato per la stima del modello FLiRTI per il caso FLM, adottando lo stesso approccio descritto nella Sezione 3.2.1.

Algorithm 2 FLiRTI multivariato con Backfitting

1.Inizializzazione:

$\hat{\beta}_0 = \sum_{i=1}^N y_i/N$

$\hat{\boldsymbol{\beta}}_j(t) = 0$ per ogni $j = 1, ..., p$ e $t = t_1, ..., t_T$

2. **while** $tol > \epsilon$ **do** $j = 1, ..., p$

3. $\hat{\boldsymbol{\beta}}_j(t) = FLiRTI\left(y_i - \hat{\beta}_0 - \int \sum_{k \neq j} \hat{\boldsymbol{\beta}}_k(t)\, x_{ik}(t) dt\right)$ ▷ Passo di Backfitting

4. **end while**

Nelle Figure 3.10 e 3.11 sono riportate le stime dei coefficienti funzionali delle stesse covariate analizzate per i modelli FLM e FGLM, stimate in questo caso con il FLiRTI multivariato. Dai grafici è evidente come il lasso funzionale pone esattamente pari a zero, a differenza dei modelli precedentemente adattati, la stima del coefficiente $\beta_j(t)$ nelle regioni, corrispondenti ai trimestri di osservazione iniziali, dove non dovrebbe esserci alcuna relazione tra quest'ultimo e la variabile risposta scalare Y_i. L'andamento delle curve stimate è simile a quello ottenuto con i modelli FLM e FGLM. Le

linee tratteggiate indicano gli intervalli di confidenza punto a punto al 95% ottenuti via *bootstrap*.

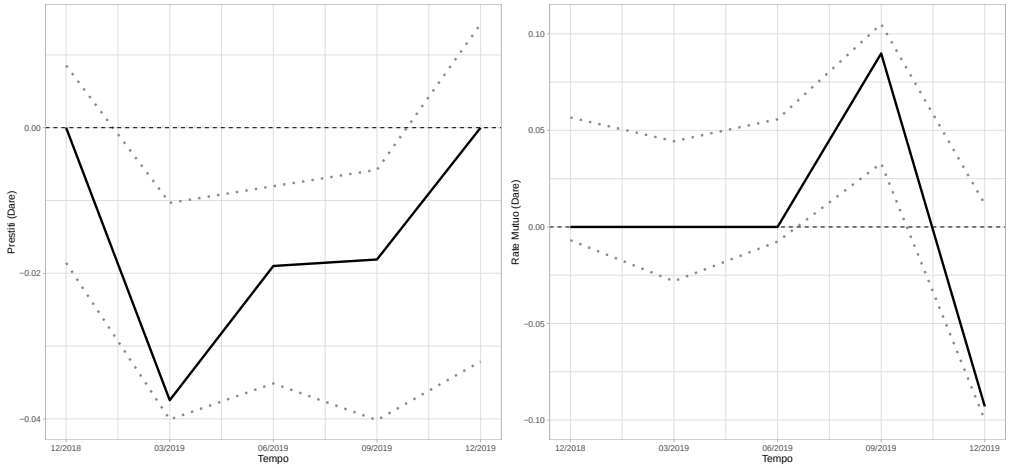

Figura 3.10: Coefficiente funzionale modello FLiRTI multivariato delle transazioni *Dare_Prestiti* e *Dare_RateMutuo_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

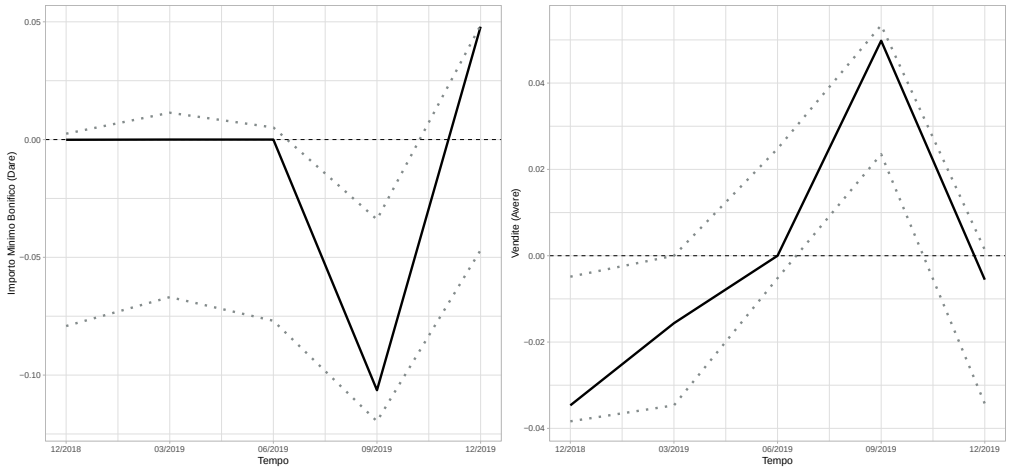

Figura 3.11: Coefficiente funzionale modello FLiRTI multivariato delle transazioni *Importo_Min_Bon* e *Avere_Vendite_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

3.2.4 Modello di regressione funzionale basato sul boosting

Brockhaus et al. (2018) propongono il modello di regressione funzionale basato sul boosting *FDboost*. Tale modello risulta molto flessibile in quanto la stima dei coefficienti è ottenuta mediante l'algoritmo del *gradient boosting* modificato. Il principale vantaggio di *FDboost* sta nel fatto che opera una selezione automatica delle variabili, ottenendo dei buoni risultati in termini di adattamento, in particolare quando il numero delle covariate è elevato.
Il modello alla base di questa metodologia è rappresentato nel seguente modo:

$$g(Y_i) = \sum_{j=1}^{J} h_j(x_i) \tag{3.23}$$

Dove $g(\cdot)$ nel caso specifico è la funzione legame logistica e $h_j(x_i)$, denominate *base-learner*, sono le funzioni che rappresentano gli effetti di una singola covariata o di più esplicative. L'intercetta scalare viene rappresentata ponendo $h_1(x) = \beta_0$, mentre per una generica covariata funzionale $h_j(x_i, t) = \int x_{ij}(t)\beta_j(t)dt$. Con la seguente specificazione dei termini additivi ci si riconduce a un modello dalla forma analoga a quella dell'equazione 3.15. Le quantità funzionali $x_{ij}(t)$ e $\beta_j(t)$ vengono rappresentate, anche in questo caso, mediante espansioni in funzioni di base:

$$\int x_{ij}(t)\beta_j(t)dt = \int x_{ij}(t) \sum_{k=1}^{K_j} \Phi_j(t)\boldsymbol{\theta}_j dt = \boldsymbol{b}_j(\boldsymbol{x}_i)^T\boldsymbol{\theta}_j \tag{3.24}$$

Nello specifico, per i $\beta_j(t)$ vengono scelte le *P-splines* di grado 3 con 1 nodo interno e con parametro di regolazione $\lambda = 0.01$ scelto via convalida incrociata.
Per gli effetti $h_j(x_i, t)$ si ottiene quindi la seguente rappresentazione:

$$h_j(\boldsymbol{x}_i) = \boldsymbol{b}_j(\boldsymbol{x}_i)^T\boldsymbol{\theta}_j \tag{3.25}$$

in cui $\boldsymbol{b}_j(\boldsymbol{x}_i)^T$ è il vettore riga della matrice di dimensione $N \times K_j$ per il j-esimo effetto di ogni osservazione, mentre $\boldsymbol{\theta}_j = (\theta_{j1}, ..., \theta_{jK_j})$ sono i coefficienti ignoti da stimare. Gli effetti sono regolarizzati da un termine di penalità quadratica: $\boldsymbol{\theta}_j^T\boldsymbol{P}_j(\boldsymbol{\lambda})\boldsymbol{\theta}_j$, in cui $\boldsymbol{P}_j(\boldsymbol{\lambda}) = \lambda_j\boldsymbol{P}_i$ rappresenta la matrice di penalità per la covariata j-esima con il coefficiente di regolarizzazione λ_j.

La procedura descritta da Brockhaus et al. (2018), ha come obiettivo la minimizzazione della seguente funzione di perdita attesa:

$$\hat{\boldsymbol{h}} = \arg\min_{h} \frac{1}{N} \sum_{i=1}^{N} L(y_i; \boldsymbol{h}(\boldsymbol{x}_i)) \tag{3.26}$$

Nello specifico, la 3.26 viene minimizzata attraverso il *gradient-descendent* con un approccio di tipo *stepwise*. Ad ogni passo del *boosting*, ciascuna *base-learner* viene stimata separatamente mediante il *gradient-descendent*. Solo le *based-learns* che migliorano l'adattamento del modello per un determinato criterio entrano a far parte di quest'ultimo, le rimanenti vengono escluse. Per questo motivo *FDboost* opera, tra l'altro, una selezione delle variabili. Scegliendo come funzione di perdita l'errore quadratico medio, il gradiente negativo corrisponde ai residui.
Per ottenere, inoltre, una stima adeguata dei coefficienti del modello, i parametri vengono selezionati via convalida incrociata e risultano pari a $\lambda = 0.01$, $M = 2000$ relativo al numero di iterazioni e $v = 0.1$ riguardante il *tasso di apprendimento*.
Nelle Figure 3.12 e 3.13 vengono riportate le stime di quattro covariate funzionali con gli intervalli di confidenza punto a punto al 95% (linea tratteggiata) calcolati via *bootstrap*. L'andamento delle curve dei coefficienti è simile a quello ottenuto con gli altri modelli adattati. La metodologia *FDboost* produce, inoltre, stime meno variabili rispetto agli altri metodi poiché l'ampiezza degli intervalli in alcune covariate è più ristretta.

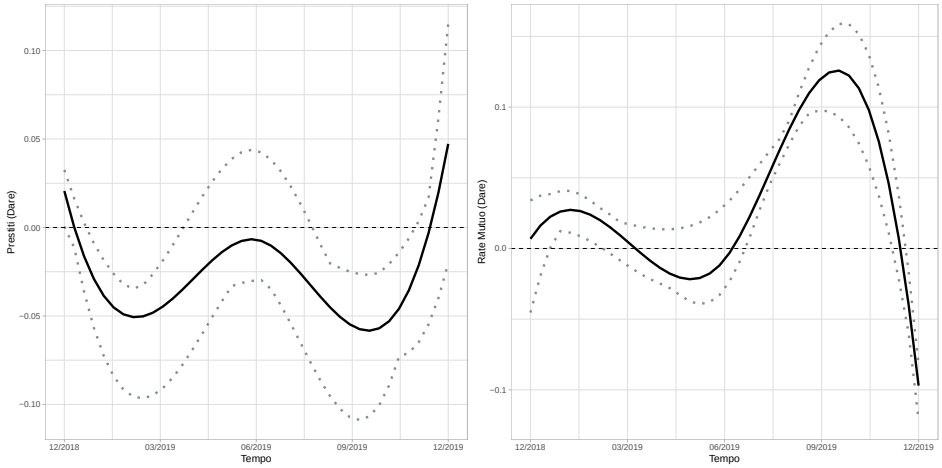

Figura 3.12: Coefficienti funzionali modello *FDboost* delle transazioni *Dare_Prestiti* e *Dare_RateMutuo_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

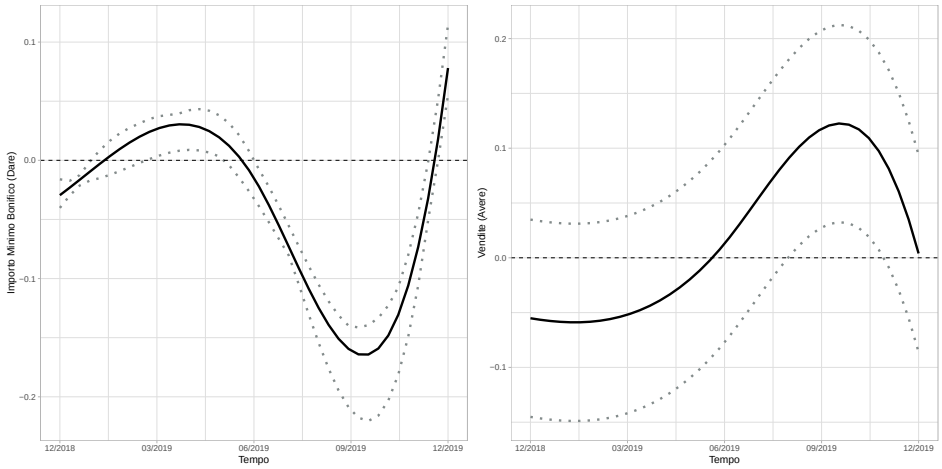

Figura 3.13: Coefficienti funzionali modello *FDboost* delle transazioni *Importo_Min_Bonifico* e *Avere_Vendite_Media_T*. Segmento Grandi Imprese, mese di riferimento dicembre 2019.

3.3 Confronto tra modelli

Per verificare la capacità previsiva dei modelli adattati e per confrontare le loro prestazioni viene utilizzato *l'insieme di verifica*. Dal momento che le classi della variabile risposta Y sono sbilanciate, come descritto nella Sezione 2.2, viene scelta una soglia

appropriata che separi la frontiera di classificazione pari al valore proporzione della classe in minoranza (0.02). Al fine di avere un confronto completo della capacità previsiva dei modelli, nelle analisi sono stati scelti nello specifico tre valori per la soglia (0.01, 0.02, 0.05). Nella tabella 3.1 sono riportati i valori degli indici considerati, mentre nella 3.14 sono riportate le curve di ROC. Considerando l'indice AUC, emerge che tutte le metodologie hanno un buon adattamento ai dati. Le prestazioni per tutti i modelli, tuttavia, non sono particolarmente soddisfacenti considerando gli indici F1, Precisione o Richiamo. Questo è dovuto al fatto che le classi della risposta sono molto sbilanciate e di conseguenza i modelli hanno difficoltà nel classificare correttamente la classe dell'evento raro. Il FLiRTI multivariato risulta il migliore tra tutti considerando l'indice F1 e soglia pari a 0.05. In Figura 3.15 viene riportato il confronto tra curve di Lift per ciascun modello.

	Soglia	Precisione	Richiamo	F1	Errore	AUC
Lineare Funzionale (FLM)	*0.01*	0.101	0.120	0.109	0.021	0.870
	0.02	0.131	0.126	0.126	0.020	
	0.05	0.162	0.234	0.188	0.022	
Logistico Funzionale (FGLM)	*0.01*	0.136	0.127	0.136	0.021	0.891
	0.02	0.173	0.220	0.181	0.020	
	0.05	0.257	0.256	0.262	0.021	
FLiRTI multivariato	*0.01*	0.105	0.113	0.111	0.019	0.901
	0.02	0.189	0.141	0.162	0.016	
	0.05	0.200	0.457	0.278	0.020	
FDboost	*0.01*	0.101	0.113	0.100	0.023	0.910
	0.02	0.134	0.121	0.145	0.016	
	0.05	0.189	0.201	0.194	0.010	

Tabella 3.1: Tabella di confronto degli indici di Precisione, Richiamo, F1, Tasso Errata Classificazione e AUC, segmento Grandi Imprese, mese di riferimento dicembre 2019

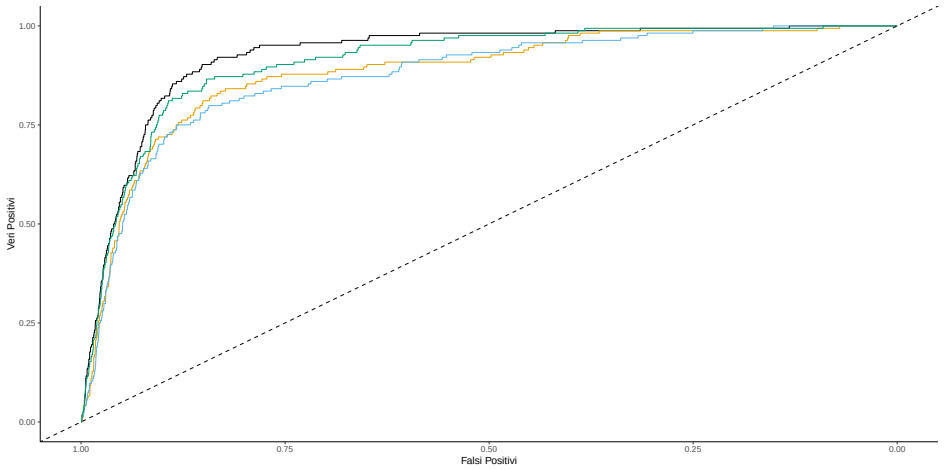

Figura 3.14: Curve di ROC: FLM (azzurro), GFLM (giallo), FLiRTI multivariato (verde) e *FDboost* (nero).

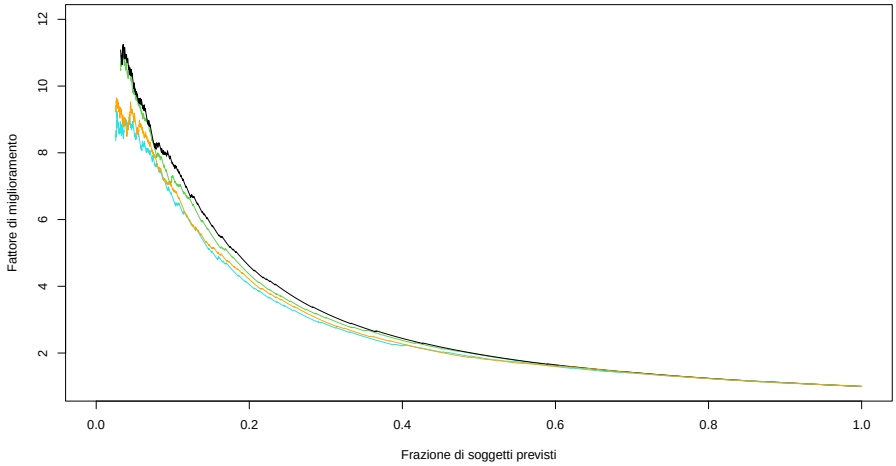

Figura 3.15: Curve di lift: FLM (azzurro), GFLM (giallo), FLiRTI multivariato (verde) e *FDboost* (nero).

3.4 Confronto complessivo

In Tabella 3.2 vengono riportati i valori degli indici per tutti i modelli adattati.Viene riportato solo il valore 0.05 per la soglia, poichè è il valore, tra tutti quelli considerati, che fornisce un indice F1, Richiamo e Precisione più elevati.

	Soglia	Precisione	Richiamo	F1	Errore	AUC
Lineare Funzionale (FLM)	*0.05*	0.162	0.234	0.188	0.022	0.870
Modello Lineare	*0.05*	0.160	0.233	0.187	0.022	0.869
Logistico Funzionale (FGLM)	*0.05*	0.257	0.256	0.262	0.021	0.891
Modello Logistico	*0.05*	0.246	0.253	0.252	0.021	0.889
FLiRTI multivariato	*0.05*	0.200	0.457	0.278	0.020	0.901
Lasso Logistico	*0.05*	0.231	0.201	0.232	0.020	0.909
FDboost	*0.05*	0.189	0.201	0.194	0.010	0.910
gradient boosting	*0.05*	0.381	0.321	0.381	0.010	0.923
MARS	*0.05*	0.276	0.253	0.263	0.010	0.912

Tabella 3.2: Tabella di confronto complessivo degli indici di Precisione, Richiamo, F1, Tasso Errata Classificazione e AUC, segmento Grandi Imprese, mese di riferimento dicembre 2019

Dai risultati ottenuti dalle analisi si può affermare che le informazioni derivanti dalle transazioni possiedono una buona capacità esplicativa nel prevedere la risposta per le varie tipologie di modelli considerati. A causa dell'elevato sbilanciamento delle classi della risposta, tuttavia, i modelli riscontrano alcune difficoltà nel classificare correttamente la classe dell'evento raro come si può notare dal valore non molto elevato dell'indice F1 per alcuni valori della soglia, come riportato nelle Tabelle 2.5 e 3.1. La scelta di un opportuno valore per tale soglia di discriminazione è dunque fondamentale per ottenere una corretta classificazione dello stato di insolvenza del cliente.
Da un punto di vista previsivo il *gradient boosting* risulta essere tra tutti il modello preferibile sia in termini di AUC che per l'indicatore $F1$. Tuttavia è un metodo, come precedentemente detto, di difficile interpretazione.
L'utilizzo di modelli per dati funzionali d'altra parte permette di comprendere l'effetto delle variabili esplicative sulla variabile risposta nel tempo. Come spesso accade,

non esiste una metodologia che prevale sull'altra, ma una combinazione di approcci contribuisce ad avere una comprensione più accurata del fenomeno in esame. Si può dunque concludere che entrambe le classi di modelli utilizzati permettono di prevedere l'evento d'interesse in modo soddisfacente tramite le transazioni, tuttavia da un punto di vista interpretativo, è preferibile utilizzare i modelli per dati funzionali per una migliore interpretazione del fenomeno.

Capitolo 4

Analisi di Sopravvivenza e Rischio di Credito

É possibile studiare i dati disponibili anche mediante l'analisi della sopravvivenza. Questo metodo è un ulteriore strumento a disposizione per una visione più completa del fenomeno. Dal momento che il rischio di credito è un fenomeno che si presenta nel tempo, è d'interesse studiare entro quanto tempo un cliente è soggetto all'evento di *Early Warning*.

4.1 Analisi di sopravvivenza nel contesto del rischio di credito

L'analisi della sopravvivenza ha come obiettivo lo studio dell'incidenza di un evento in un determinato arco temporale, dove con incidenza si indica il numero dei soggetti che sperimentano tale evento. La funzione di sopravvivenza, indicata con $S(t)$, è la probabilità che l'evento di interesse non si verifichi (o la probabilità che un individuo sopravviva) fino a un tempo t. Indicando con T il tempo in cui si verifica l'episodio per un soggetto, la curva di sopravvivenza può essere rappresentata nel seguente modo:

$$S(t) = Pr(T > t) \tag{4.1}$$

in cui $0 \leq S(t) \leq 1$ e $T \geq 0$.
Lo studio di queste funzioni è principalmente utilizzato in ambito medico per valuta-

re la probabilità di sopravvivenza di un paziente o un gruppo di soggetti. Tuttavia, poiché è un metodo molto flessibile, è possibile adattarlo a contesti differenti. In questa tesi, infatti, si vuole utilizzare tale metodologia per valutare la probabilità che i clienti della banca facenti parte dei tre gruppi (Grandi Imprese, Piccole e Medie Imprese e Privati) sopravvivano all'evento d'interesse che in questo caso è identificato come *la prima volta* che in un soggetto si verifica un *Early Warning*, nel periodo che va da dicembre 2018 a dicembre 2019, mentre chi non li riceve figura come censura. Per semplicità non si sono considerati i casi in cui un cliente può avere più episodi, nella fattispecie più episodi di *Early Warning*. In questa tesi ci si è focalizzati solo sulla prima volta che il *CostumerID* ha sperimentato l'evento d'interesse non considerando il caso di eventi ricorrenti.
Per l'analisi si utilizzano metodi non parametrici come lo stimatore di Kaplan-Maier, il quale valuta quando il cliente nel corso del tempo di osservazione ha sperimentato l'*Early Warning* o meno e metodi semi-parametrici quali il modello di Cox, per il quale si prevede l'inserimento delle transazioni come covariate.

4.2 Indicatore anticipato

Prima di procedere nella stima di $S(t)$, è d'interesse valutare se il Sistema di Allarme Preventivo adottato dalla banca fornisce dei dati affidabili per trarre conclusioni ragionevoli dalle seguenti analisi.
Nello specifico, l'istituto di credito, come descritto nel Capitolo 1, attraverso il Sistema di Allarme Preventivo, costruisce un *indicatore anticipato*. Quest'ultimo è una sorta di regola la quale, classificando un cliente in stato di *Early Warning*, ci suggerisce implicitamente che quest'ultimo sarà insolvente nel trimestre successivo e che, se non viene attuata nessuna misura cautelativa, esso andrà in *default* nei successivi mesi. In altre parole, la regola dichiara che se il cliente presenta un 'allarme' in un determinato mese allora nei successivi mesi prevederà un rischio di credito elevato con conseguente default, salvo attuazione di misure cautelative. Per verificare l'affidabilità di tale *indicatore anticipato*, tenendo presente quanto descritto nella Sezione 1.3 per la variabile *Target*, vengono calcolati i Veri Positivi e i Falsi Positivi mediante i seguenti criteri:

- *Vero Positivo*: se la variabile $Target$ per un i-esimo *CostumerID* assume il valore 1 al mese t e NA nel mese di osservazione $t+1$ è un vero positivo. Dal momento che al tempo $t+1$ non è avvenuta alcuna rilevazione tale cliente in quel mese non presenta una situazione creditizia regolare e si assume sia andato in default;
- *Falso Positivo*: se la variabile $Target$ per un i-esimo *CostumerID* assume il valore 1 al mese t e 0 nel mese di osservazione $t+1$ è un falso positivo, poichè al tempo $t+1$ il cliente presenta una situazione creditizia regolare.

In Tabella 4.1 è riportato un esempio su come vengono costruiti, a partire dalla procedura appena descritta, gli indici relativi ai Veri Positivi e Falsi Positivi.

Suddivisione	*Target_dic_2018*	*Target_mar_2019*	*Target_giu_2019*	*Target_sett_2019*	*Target_dic_2019*
Vero positivo	0	0	1	NA	NA
Falso positivo	0	0	1	0	0
Falso positivo	0	1	0	0	0
Vero positivo	1	NA	NA	NA	NA

Tabella 4.1: Esempio su come vengono suddivisi i Falsi Positivi e Veri Positivi per alcuni clienti.

La regola è affidabile se c'è una proporzione elevata di Veri Positivi e bassa di Falsi Positivi. Tali quantità vengono calcolate per tutti e tre i dataset, ottenendo i seguenti risultati:

	Veri Positivi	Falsi Positivi
Grandi Imprese	85%	15%
Piccole e Medie Imprese	80%	20%
Privati	87%	13%

Tabella 4.2: Risultati veri positivi e falsi positivi

Dai risultati ottenuti in Tabella 4.2, si può considerare quindi l'*indicatore anticipato*

dell'istituto di credito una regola ragionevolmente affidabile. In definitiva, calcolare la probabilità che i clienti sopravvivano all'evento d'interesse, definito come la prima volta che un soggetto abbia un *Early Warning*, può essere uno strumento utile per fornire informazioni sulla probabilità che i clienti non sperimentino l'evento default.

4.3 La curva di sopravvivenza Kaplan-Meier

Lo stimatore Kaplan-Meier è un metodo non parametrico utilizzato per stimare la curva di sopravvivenza $S(t)$. Una stima di $S(t)$ mediante questo metodo è data dalla seguente equazione:

$$\hat{\boldsymbol{S}}(t) = \prod_{t_k<t} \left(1 - \frac{d_k}{n_k}\right) \tag{4.2}$$

in cui n_k è il numero di clienti a rischio di sperimentare l'evento al tempo t_k e d_k è il numero di soggetti nei quali si verifica l'episodio d'interesse.
La curva Kaplan-Meier tiene in considerazione anche delle osservazioni censurate, in particolare della censura a destra, che si verifica se un cliente non viene più rilevato, ossia se esce dal campione prima che si osservi l'esito finale. In questo caso si considerano censurati tutti quei clienti nei quali ad un tempo t la variabile $Target$ assume il valore 0 e al tempo successivo $t + 1$ non vi è più alcuna rilevazione. Si assume che i soggetti in questo caso abbiano abbandonato il campione per motivi diversi dal default, quindi i 'non default' figurano come censure.
Per stimare quindi $S(t)$ tramite questa metodologia è stato necessario reperire, a partire dai tre dataset disponibili, le quantità d'interesse:

- *Durata*: tempo di sopravvivenza o di censura per ciascun cliente;
- *Stato*: indica se la durata in questione è censurata o meno, ovvero 0 se un individuo è stato censurato, 1 altrimenti.

In Tabella 4.3 viene riporato un esempio su come sono state ricavate le quantità necessarie per la stima di $S(t)$ a partire dai tre dataset disponibili.

Stato	Durata	*Target_dic_2018*	*Target_mar_2019*	*Target_giu_2019*	*Target_sett_2019*	*Target_dic_2019*
1	2	0	1	NA	NA	NA
1	3	0	0	1	NA	NA
0	5	0	0	0	0	0
1	2	0	1	0	1	0
0	4	0	0	0	0	NA
0	2	0	0	NA	NA	NA
1	3	0	0	1	0	1

Tabella 4.3: Esempio su come vengono ricavate le quantità: Stato e Durata per alcuni clienti.

In Figura 4.1 vengono riportate le curve di sopravvivenza stimate con Kaplan Meier per ciascuno dei tre gruppi con i relativi intervalli di confidenza al 95%. Per la stima delle curve viene mantenuta una numerosità uguale per i tre segmenti. Si è perciò selezionato casualmente un campione di 24000 clienti in ciascun dataset.

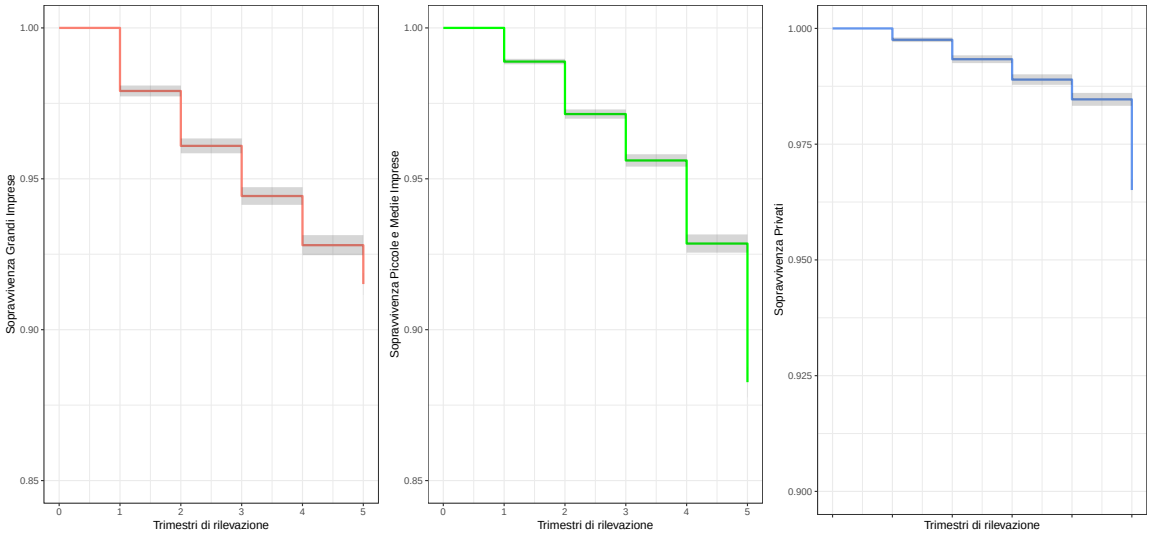

Figura 4.1: Curve di sopravvivenza per i tre segmenti: Grandi Imprese (rosso), Piccole e Medie Imprese (blu) e Privati (verde) calcolati tramite lo stimatore Kaplan-Meier.

In Figura 4.2 vengono riportate nello stesso grafico le curve di sopravvivenza della

Figura 4.1 con i relativi intervalli di confidenza al 95%. Dal grafico emerge come i tre gruppi differiscano per durata dell'evento, in particolare la curva di sopravvivenza relativa ai Privati si discosta in maniera evidente dagli altri gruppi. Viene notato inoltre come il segmento delle Grandi Imprese abbia una probabilità di sopravvivenza più bassa rispetto agli altri due gruppi.
Per analizzare se le differenze tra i tre segmenti sono significative viene calcolato il Logrank test. Si tratta di un test che, nel caso di ipotesi nulla H_0, ipotizza le tre curve di sopravvivenza uguali e quindi anche le loro distribuzioni. La statistica test ha distribuzione χ^2 con numero di gradi di libertà pari al numero di gruppi meno uno. Il risultato del test conferma una differenza tra le curve.

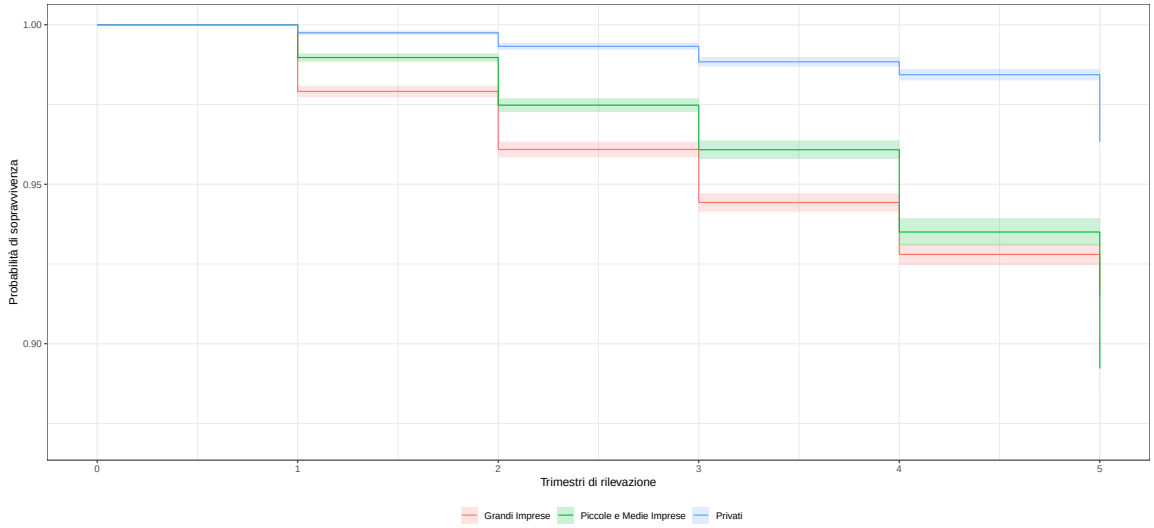

Figura 4.2: Curve di sopravvivenza stratificata per i tre segmenti: Grandi Imprese (rosso), Piccole e Medie Imprese (blu) e Privati (verde) calcolati tramite lo stimatore Kaplan-Meier.

Dalla Figura 4.2 si può verificare graficamente se l'ipotesi di proporzionalità dei rischi è soddisfatta o meno, andando a vedere se lo sono le curve di sopravvivenza. Dal momento che quest'ultime non si incrociano, l'assunto di proporzionalità viene rispettato. Questo permette di adattare il modello di Cox.

4.4 Il Modello di Cox

Il modello di Cox è un modello semi-parametrico che permette di analizzare il rapporto tra un fattore di rischio e l'incidenza di un determinato evento. Questo modello

permette l'inserimento anche di p covariate. In questo caso viene inserita la covariate tempo-dipendente relativa ai *Prestiti*, poichè è l'unica variabile esplicativa comune in tutti e tre i dataset.
Il rischio che l'individuo i-esimo, con vettore di covariate che dipende dal tempo $X_i(t)$, sperimenti l'evento d'interesse al tempo t può essere definito nel seguente modo:

$$\hat{\boldsymbol{h}}_i\big(t|X_i(t)\big) = h_0(t)\, \exp\left(\beta_1^T x_{i1}(t) + ... + \beta_p^T x_{ip}(t)\right) \tag{4.3}$$

dove $\hat{\boldsymbol{h}}_i\big(t|X_i(t)\big)$ è la funzione di rischio, $h_0(t) \geq 0$ è la parte non parametrica che rappresenta la funzione rischio di base mentre $\exp\left(\beta_1^T x_{i1}(t) + ... + \beta_p^T x_{ip}(t)\right) > 0$ indica l'effetto delle covariate sul rischio di base ed è la parte parametrica. Si osservi che solo le covariate dipendono dal tempo, infatti i coefficienti β_j sono costanti al variare di t.
Al fine di ricavare le quantità necessarie per la stima di questo modello, vengono effettuate delle operazioni preliminari in ciascuno dei tre dataset che possono essere riassunte nei seguenti punti:

(1) per ciascun cliente viene diviso l'episodio originale in sotto-episodi contigui in base al tempo di cambiamento di stato delle covariate. In questo caso si sono costruiti sotto-episodi dalla durata di 1 periodo (corrispondente a un trimestre);

(2) viene ricostruita la variabile di censura per ciascun sotto-episodio;

(3) assegnazione del valore delle covariate tempo-dipendenti a ciascun sotto-episodio.

In Tabella 4.4 vengono riportati i risultati delle stime ottenute dal modello di Cox, ponendo come gruppo di riferimento le Grandi Imprese, dei coefficienti relativi al gruppo Piccole e Medie Imprese, Privati e alla covariata relativa ai *Prestiti*.

Covariate	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$	$\text{se}(\hat{\beta}_j)$	*valore-p*
Gruppo: Piccole e Medie Imprese	0.09259	1.09701	0.03329	0.002
Gruppo: Privati	-1.07311	0.34194	0.04462	0.001
Prestiti	-0.23859	0.78774	0.01512	0.001

Tabella 4.4: Risultati modello di Cox

Dai risultati ottenuti si nota come la covariata sulle transazioni inerenti ai prestiti e i due gruppi hanno un effetto significativo nello spiegare il fenomeno d'interesse. Il modello conferma quanto emerso dai grafici, ossia che far parte del gruppo dei Privati diminuisce il rischio di sperimentare l'evento d'interesse. Anche la covariata relativa ai *Prestiti* ne diminuisce il rischio.
In Figura 4.3 è riportato il grafico delle curve del rischio cumulato, stimate a partire dal modello di Cox con covariata Prestiti per i tre gruppi. Le curve confermano quanto appena emerso dall'analisi, ossia che il segmento delle Grandi Imprese in tutti i periodi considerati, è quello che ha un rischio maggiore di sperimentare l'evento d'interesse rispetto agli altri due gruppi.

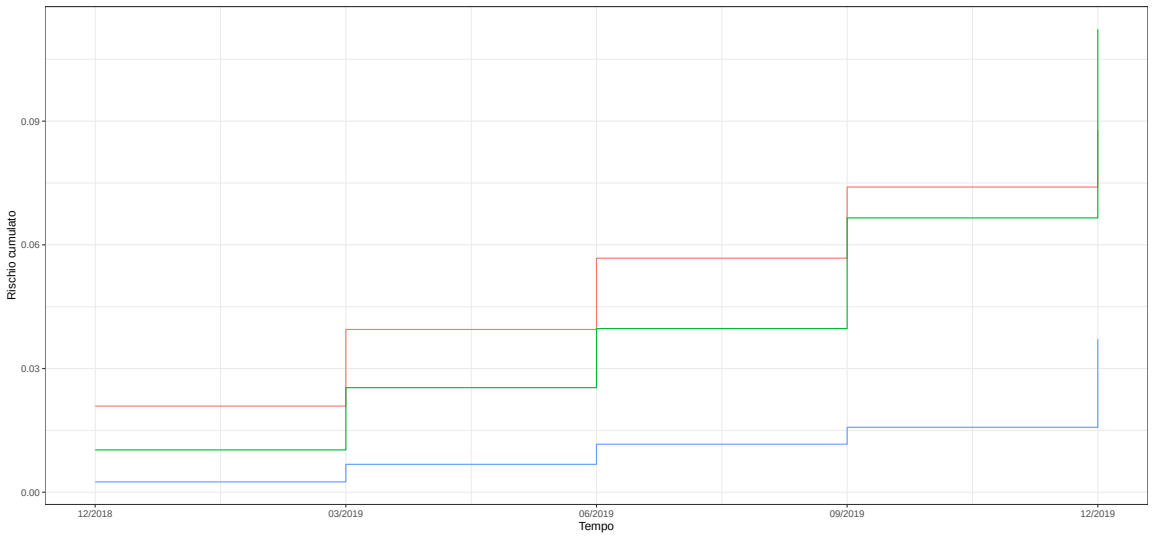

Figura 4.3: Output modello di Cox con covariata *Prestiti*. Curve di rischio cumulato stratificata per i tre segmenti: Grandi Imprese (rosso), Piccole e Medie Imprese (blu) e Privati (verde).

Conclusioni

Lo scopo di questa tesi era quello di prevedere con largo anticipo i default creditizi del portafoglio clienti di un'importante banca del panorama italiano su tre segmenti: Privati, Piccole e Medie Imprese e Grandi Imprese. I modelli sinora utilizzati dall'istituto di credito, per prevedere lo stato d'insolvenza di un cliente, prendono in considerazione informazioni statiche nel tempo, quali ad esempio dati socio-demografici, dati di bilancio o saldi statici mensili (come conto corrente o carte). A differenza di tale approccio, in questa tesi sono state considerate informazioni dinamiche nel tempo, ovvero le serie storiche trimestrali delle transazioni come covariate nei modelli adattati. A tal fine sono state utilizzate due principali tipologie di modelli statistici: i modelli di data mining e una classe di modelli alternativa e più sofisticata, ossia i modelli per dati funzionali. Dai risultati ottenuti dalle analisi, si può affermare che le informazioni derivanti dalle transazioni possiedono una buona capacità esplicativa nel prevedere la risposta per entrambe le tipologie di modelli considerati. Inoltre, sulla base dei risultati ottenuti, la combinazione di queste due classi di modelli permette di affrontare in maniera più completa il problema in esame. In particolare, l'utilizzo di modelli per dati funzionali, i quali tengono in considerazione la natura longitudinale delle osservazioni, permette di avere una migliore interpretazione dell'effetto delle variabili esplicative sulla variabile risposta nel tempo, fornendo interpretazioni più accurate del fenomeno. I modelli di data mining invece non consentono di intuire immediatamente i differenti comportamenti dei clienti in base ai valori dei coefficienti delle transazioni. É stata infine condotta un'analisi della sopravvivenza sui clienti dei tre segmenti per analizzare l'incidenza dell'evento d'interesse nell'arco temporale considerato, utilizzando strumenti usualmente impiegati in ambito medico, quali la curva di sopravvivenza stimata attraverso lo stimatore Kaplan-Meier e il modello di Cox con covariate tempo-dipendenti (in questo caso le transazioni). Da tale analisi si è potuto constatare che i clienti a maggior rischio di insolvenza sono quelli che ap-

partengono al segmento Grandi Imprese; al contrario, i Privati sono coloro che hanno una minor probabilità di sperimentare l'evento d'interesse nel periodo di osservazione considerato. Possibili miglioramenti futuri potrebbero riguardare il metodo di raccolta dei dati da parte della banca. I dati disponibili sulle transazioni erano infatti delle serie storiche trimestrali dove venivano sommati i valori delle transazioni del mese di osservazione oppure veniva calcolata la media trimestrale di quest'ultime. Per usare tutte le informazioni disponibili nelle analisi si è deciso di utilizzare tutte le covariate, pur sapendo che le variabili contenenti informazioni trimestrali sono una media dei tre mesi soggetta a variabilità. Se nei tre mesi i clienti hanno sostenuto spese di importo molto differente, la media è un indicatore approssimativo. Non sarebbe quindi un approccio 'raffinato' quello che è stato seguito. Per avere un'analisi più accurata si potrebbero raccogliere i dati con frequenza mensile, in modo tale da eliminare la possibile variabilità nei tre mesi, per ottenere così delle serie storiche delle transazioni con più istanti di osservazione e con più informazioni relative ai singoli mesi oltre che una minore variabilità.

Appendice A

Risultati Integrativi

Nelle Tabelle A.1 e A.2 vengono riportati i risultati dei modelli lineare, logistico, lasso logistico, MARS e *gradient boosting* per i segmenti Piccole e Medie Imprese e Privati mese di riferimento dicembre 2019.
Nelle Tabelle A.3 e A.4 vengono riportati i risultati dei modelli FLM, GFLM, FLiRTI multivariato e *FDboost* per i segmenti Piccole e Medie Imprese e Privati mese di riferimento dicembre 2019.

	Soglia	Precisione	Richiamo	F1	Errore	AUC
Lineare	*0.01*	0.112	0.104	0.120	0.023	0.871
	0.02	0.145	0.136	0.134	0.022	
	0.05	0.191	0.203	0.180	0.021	
Lineare ridotto	*0.01*	0.115	0.127	0.123	0.023	0.872
	0.02	0.177	0.158	0.180	0.020	
	0.05	0.212	0.253	0.201	0.018	
Logistico	*0.01*	0.123	0.140	0.157	0.020	0.890
	0.02	0.180	0.231	0.198	0.013	
	0.05	0.266	0.274	0.281	0.022	
Logistico Ridotto	*0.01*	0.154	0.142	0.165	0.020	0.891
	0.02	0.198	0.241	0.193	0.013	
	0.05	0.247	0.258	0.259	0.021	
Lasso Logistico	*0.01*	0.106	0.112	0.112	0.020	0.902
	0.02	0.182	0.211	0.225	0.020	
	0.05	0.254	0.211	0.246	0.021	
MARS	*0.01*	0.142	0.217	0.182	0.021	0.911
	0.02	0.189	0.223	0.196	0.016	
	0.05	0.286	0.255	0.274	0.010	
Gradient Boosting	*0.01*	0.221	0.201	0.201	0.027	0.920
	0.02	0.352	0.340	0.380	0.016	
	0.05	0.395	0.323	0.390	0.010	

Tabella A.1: Tabella di confronto degli indici di Precisione, Richiamo, F1, Tasso Errata Classificazione e AUC, segmento Piccole e Medie Imprese, mese di riferimento dicembre 2019

	Soglia	Precisione	Richiamo	F1	Errore	AUC
	0.01	0.109	0.126	0.117	0.022	0.880
Lineare	*0.02*	0.135	0.129	0.132		
	0.05	0.165	0.232	0.192	0.022	
	0.01	0.116	0.127	0.115	0.025	0.882
Lineare ridotto	*0.02*	0.149	0.138	0.142	0.020	
	0.05	0.182	0.224	0.201	0.012	
	0.01	0.135	0.139	0.140	0.021	0.890
Logistico	*0.02*	0.180	0.231	0.201	0.013	
	0.05	0.259	0.253	0.259	0.021	
	0.01	0.143	0.140	0.145	0.029	0.892
Logistico Ridotto	*0.02*	0.192	0.241	0.121	0.013	
	0.05	0.242	0.269	0.259	0.012	
	0.01	0.106	0.112	0.112	0.020	0.908
Lasso Logistico	*0.02*	0.182	0.211	0.228	0.020	
	0.05	0.221	0.211	0.245	0.021	
	0.01	0.112	0.224	0.173	0.021	0.915
MARS	*0.02*	0.156	0.247	0.197	0.016	
	0.05	0.273	0.250	0.266	0.019	
	0.01	0.235	0.221	0.203	0.022	0.929
Gradient Boosting	*0.02*	0.369	0.320	0.390	0.016	
	0.05	0.375	0.331	0.396	0.011	

Tabella A.2: Tabella di confronto degli indici di Precisione, Richiamo, F1, Tasso Errata Classificazione e AUC, segmento Privati, mese di riferimento dicembre 2019

	Soglia	Precisione	Richiamo	F1	Errore	AUC
	0.01	0.109	0.123	0.111	0.020	0.880
Lineare Funzionale (FLM)	*0.02*	0.132	0.129	0.130	0.020	
	0.05	0.169	0.238	0.190	0.022	
	0.01	0.141	0.131	0.146	0.021	0.901
Logistico Funzionale (FGLM)	*0.02*	0.175	0.227	0.191	0.020	
	0.05	0.258	0.259	0.270	0.021	
	0.01	0.107	0.116	0.119	0.029	0.906
FLiRTI multivariato	*0.02*	0.201	0.151	0.172	0.016	
	0.05	0.201	0.501	0.308	0.021	
	0.01	0.109	0.116	0.112	0.020	0.911
FDboost	*0.02*	0.154	0.131	0.155	0.011	
	0.05	0.191	0.203	0.201	0.010	

Tabella A.3: Tabella di confronto degli indici di Precisione, Richiamo, F1, Tasso Errata Classificazione e AUC, segmento Piccole e Medie Imprese, mese di riferimento dicembre 2019

	Soglia	Precisione	Richiamo	F1	Errore	AUC
	0.01	0.105	0.123	0.110	0.021	0.892
Lineare Funzionale (FLM)	*0.02*	0.132	0.127	0.129	0.020	
	0.05	0.132	0.244	0.198	0.022	
	0.01	0.137	0.128	0.139	0.021	0.900
Logistico Funzionale (FGLM)	*0.02*	0.174	0.223	0.182	0.022	
	0.05	0.259	0.249	0.270	0.021	
	0.01	0.109	0.112	0.120	0.018	0.903
FLiRTI multivariato	*0.02*	0.190	0.142	0.172	0.017	
	0.05	0.201	0.447	0.288	0.021	
	0.01	0.110	0.116	0.121	0.021	0.918
FDboost	*0.02*	0.136	0.129	0.151	0.016	
	0.05	0.199	0.221	0.204	0.011	

Tabella A.4: Tabella di confronto degli indici di Precisione, Richiamo, F1, Tasso Errata Classificazione e AUC, segmento Privati, mese di riferimento dicembre 2019

Elenco delle figure

Elenco delle tabelle

Bibliografia

Azzalini, A. & Scarpa, B. (2012), *Data analysis and data mining: An introduction*, OUP USA.

Brockhaus, S., Fuest, A., Mayr, A. & Greven, S. (2018), 'Signal regression models for location, scale and shape with an application to stock returns', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(3), 665–686.

Candes, E. & Tao, T. (2007), 'The dantzig selector: Statistical estimation when p is much larger than n', *The annals of Statistics* **35**(6), 2313–2351.

Escabias, M., Aguilera, A. M. & Valderrama, M. J. (2007), 'Functional pls logit regression model', *Computational Statistics & Data Analysis* **51**(10), 4891–4902.

Febrero-Bande, M. & de la Fuente, M. O. (2012), 'Statistical computing in functional data analysis: The r package fda. usc', *Journal of statistical Software* **51**, 1–28.

Friedman, J. H. (1991), 'Multivariate adaptive regression splines', *The annals of statistics* **19**(1), 1–67.

Friedman, J. H. (2002), 'Stochastic gradient boosting', *Computational statistics & data analysis* **38**(4), 367–378.

Friedman, J., Hastie, T., Simon, N., Tibshirani, R., Hastie, M. T. & Matrix, D. (2017), 'Package 'glmnet."', *Journal of statistical software* **33**(1), 1–22.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software* **33**(1), 1.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), 'The elements of statistical learning. springer series in statistics', *New York, NY, USA* .

James, G. M., Wang, J. & Zhu, J. (2009), 'Functional linear regression that's interpretable', *The Annals of Statistics* **37**(5A), 2083–2108.

McCullagh, P. & Nelder, J. A. (1989), 'Generalized linear models. chapman and hall', *London, UK* .

Müller, H.-g. (2005), 'Functional modelling and classification of longitudinal data', *Scandinavian Journal of Statistics* **32**(2), 223–240.

Müller, H.-G. & Stadtmüller, U. (2005), 'Generalized functional linear models', *the Annals of Statistics* **33**(2), 774–805.

Ramsay, J. O. & Dalzell, C. (1991), 'Some tools for functional data analysis', *Journal of the Royal Statistical Society: Series B (Methodological)* **53**(3), 539–561.

Ramsay, J. O. & Silverman, B. W. (2005), 'Functional data analysis'.

Ridgeway, G., Southworth, M. H. & RUnit, S. (2013), 'Package 'gbm"', *Viitattu* **10**(2013), 40.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), 'Learning representations by back-propagating errors', *nature* **323**(6088), 533–536.

Salvan, A., Sartori, N. & Pace, L. (2020), Modelli lineari generalizzati, *in* 'Modelli Lineari Generalizzati', Springer, pp. 67–119.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.