
UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Corso di Laurea triennale in Economia e Commercio

# PREVISIONE DEL CHURN RATE DEI CLIENTI BCC: CONFRONTO TRA MODELLI ECONOMETRICI E MACHINE LEARNING

# PREDICTING BCC CUSTOMER CHURN RATE: A COMPARISON BETWEEN ECONOMETRIC AND MACHINE LEARNING MODELS

Relatore:
Prof. Riccardo Lucchetti

Rapporto Finale di:
Emma Guiducci

Anno Accademico 2025/2026

**INDICE**

# Capitolo 1 – INTRODUZIONE

Il rapporto tra una banca e i propri clienti si costruisce nel tempo attraverso la fiducia, la prossimità e la continuità del servizio. Per le Banche di Credito Cooperativo, istituti per definizione radicati nel territorio e orientati alla relazione di lungo periodo con famiglie e imprese locali, questa dimensione non è soltanto un valore identitario: è la principale fonte di valore economico e di sostenibilità. In questo contesto, la perdita di un cliente non si traduce solo in una riduzione dei ricavi correnti, ma implica la dissoluzione di un legame fiduciario costruito nel tempo e la rinuncia al valore futuro atteso da quel rapporto.

La letteratura sul tema documenta come il costo di acquisizione di un nuovo cliente superi significativamente quello della sua fidelizzazione, con stime che collocano il divario fino a sette volte superiore. Anticipare il rischio di abbandono, identificare cioè quali clienti stiano per uscire prima che lo facciano, consente di attivare interventi mirati di retention in modo proattivo, trasformando un dato descrittivo in uno strumento operativo. La previsione del churn rate si configura pertanto come un problema di classificazione binaria ad alto valore applicativo, al quale la ricerca accademica e il settore finanziario guardano con crescente interesse.

Il punto di partenza di questa tesi consiste nella definizione del contesto istituzionale e geografico in cui si radica l'analisi: le banche di Credito Cooperativo e nello specifico la realtà della BCC di Pergola e Corinaldo. In seguito il Capitolo 3 definisce il concetto di churn rate, ne illustra il ruolo strategico nel settore bancario e inquadra la letteratura sul tema, con particolare attenzione all'evoluzione degli strumenti di machine learning applicati alla churn analysis. Il quarto capitolo descrive il processo di acquisizione, pulizia e selezione del campione, nonché le variabili incluse nei modelli. Infine il Capitolo 5 rappresenta il fulcro dell'intera tesi, volto al confronto tra approcci predittivi di natura diversa: da un lato i modelli econometrici classici, il Modello di Probabilità Lineare stimato tramite OLS e il modello Logit binario, e dall'altro un algoritmo di machine learning, la Random Forest. Questa scelta riflette una duplice esigenza: valutare se i modelli tradizionali, ampiamente consolidati nella letteratura econometrica, siano sufficienti per un problema applicativo come questo; e verificare se e in quale misura l'adozione di tecniche di apprendimento automatico aggiunga un valore predittivo concreto rispetto alla loro maggiore complessità computazionale.

# Capitolo 2 – IL CREDITO COOPERATIVO E LA BCC DI PERGOLA E CORINALDO: EVOLUZIONE, STRUTTURA E TERRITORIO

## 2.1 LE BANCHE DI CREDITO COOPERATIVO NEL SISTEMA BANCARIO ITALIANO

Le Banche di Credito Cooperativo (BCC) rappresentano una componente storica del sistema bancario italiano. Nate dalle antiche Casse Rurali e Artigiane, hanno progressivamente assunto la forma moderna prevista dalla normativa bancaria, mantenendo però la natura mutualistica e il legame con il territorio. Le BCC operano come società cooperative senza finalità speculative, orientate a sostenere lo sviluppo dell'economia locale attraverso la raccolta del risparmio e la concessione del credito a famiglie, imprese, attività artigiane e agricole. Il loro inserimento nel sistema bancario nazionale è caratterizzato da tre elementi fondamentali:

- la mutualità prevalente come principio statutario e operativo
- la governance partecipativa fondata sul voto capitario
- il radicamento territoriale, che limita l'area operativa ma rafforza la conoscenza diretta del contesto economico e sociale

Le BCC mantengono un ruolo significativo in termini di presenza territoriale: risultano diffuse in un ampio numero di province e comuni, spesso costituendo l'unica realtà bancaria operativa in zone periferiche o a bassa densità di sportelli. Questa capillarità conferma la loro funzione di presidio finanziario e sociale nelle aree meno servite da banche commerciali di dimensioni nazionali, secondo dati istituzionali del movimento cooperativo del credito. Con la riforma del Credito Cooperativo, Legge 8 aprile 2016 n. 49, le BCC hanno subito un cambiamento radicale del proprio assetto: sono confluite nei Gruppi Bancari Cooperativi, mantenendo la propria autonomia mutualistica ma operando sotto la direzione della capogruppo. Sono due i gruppi bancari cooperativi che operano in Italia:

1. Gruppo Bancario Cooperativo Iccrea

2. Gruppo Bancario Cooperativo Cassa Centrale Banca

In seguito nel 2018, con il Decreto-legge n. 91, il processo riformativo è stato implementato, e sono stati posti come obiettivi principali "rafforzare la finalità mutualistica" ed il "radicamento territoriale delle BCC" (*La riforma si riforma 2018* 2018).

Nel loro funzionamento quotidiano, le Banche di Credito Cooperativo presentano una struttura aziendale analoga a quella degli altri istituti di credito, articolata in aree quali: gestione del credito, funzioni commerciali, controllo dei rischi, finanza, servizi operativi e amministrativi e sistemi informativi. La differenza principale risiede nel modello di governance: le BCC sono cooperative a mutualità prevalente, e pertanto perseguono obiettivi di utilità sociale,

sostenibilità e crescita locale, prima ancora della massimizzazione del profitto. Le filiali assumono un ruolo centrale: costituiscono il punto di contatto diretto con famiglie e imprese del territorio e favoriscono l'interpretazione delle esigenze locali, elemento distintivo rispetto ai grandi gruppi bancari commerciali.

## 2.2 LA BCC DI PERGOLA E CORINALDO

La Banca di Credito Cooperativo di Pergola e Corinaldo, contesto di riferimento del nostro studio, nasce come realtà radicata nel territorio, con una presenza storica significativa, e con una rete di filiali distribuite in diversi comuni dell'area marchigiana.

L'identità della banca trova realizzazione nel modello cooperativo che caratterizza l'intero movimento delle BCC: una forma d'impresa che coniuga attività bancaria, mutualità e attenzione allo sviluppo locale. La banca si presenta infatti, nello statuto ufficiale, come un soggetto orientato alla promozione del benessere economico e sociale delle comunità locali, mediante raccolta del risparmio, concessione del credito e sostegno alle iniziative di interesse collettivo (BCC Pergola e Corinaldo 2024).

Un passaggio determinante nell'evoluzione dell'istituto è rappresentato dalla fusione di due realtà bancarie complementari, la BCC di Pergola e quella di Corinaldo, che dal 1º gennaio 2018 sono un unico nucleo operativo(Angeletti 2017). L'integrazione ha portato a dei benefici non solo in senso quantitativo, ampliando l'area territoriale di competenza e la clientela, ma anche qualitativo, consentendo alla banca di rafforzare la propria presenza nella regione.

Questa espansione ha anche reso possibile una maggiore diversificazione della

clientela e una rete di filiali più articolata e coerente con le dinamiche socio-economiche dell'area. La banca opera all'interno di un territorio eterogeneo che comprende numerosi comuni appartenenti a diverse province: Pesaro e Urbino, Ancona e Macerata. L'area è caratterizzata da piccole e medie imprese, attività agricole e artigianali, realtà turistiche e comunità locali con forte coesione sociale. L'ampiezza del territorio in cui opera rispecchia la natura stessa della banca, che rappresenta un autentico presidio finanziario di prossimità per la comunità locale.

La BCC di Pergola e Corinaldo aderisce al Gruppo Bancario Cooperativo Iccrea, come previsto dalla riforma del Credito Cooperativo (d.l. 18/2016). L'appartenenza al gruppo consente all'istituto di garantire solidità patrimoniale, strumenti tecnologici d'avanguardia e una gestione centralizzata dei rischi, preservando al tempo stesso autonomia operativa, radicamento territoriale e finalità mutualistica. Attraverso questo primo sguardo d'insieme della banca, si vuole far comprendere al lettore il ruolo assunto dalle singole filiali nel quadro organizzativo ed economico complessivo.

# Capitolo 3 – IL CHURN RATE

## 3.1 DEFINIZIONE

Il Churn rate, detto anche tasso di abbandono o tasso di defezione, esprime la percentuale di clienti che smettono di utilizzare un prodotto o un servizio offerto da un'azienda in un dato periodo di tempo (GlossarioMarketing n. d.). Dal punto di vista operativo, il tasso di abbandono si calcola mettendo a rapporto il numero di clienti persi in un dato periodo, con il numero di clienti attivi all'inizio dello stesso periodo, moltiplicato per 100.

$$churnrate = \frac{clienti\,persi}{clienti\,iniziali} \times 100 \tag{3.1}$$

Si tratta di un indicatore critico che consente di misurare la salute di un business e aiuta a fare il benchmark della customer satisfaction, ovvero il grado di soddisfazione dei clienti (Castigli 2022). È uno dei principali indicatori di Marketing, i KPI (key performance indicator), metriche fondamentali che sono volte a misurare il raggiungimento degli obiettivi di marketing prefissati dall'azienda, l'efficacia delle strategie, nonché i rendimenti del budget allocato.
Ad oggi, l'evoluzione del marketing ha spostato il focus dalle tradizionali vanity metrics, come like e visualizzazioni, alle metriche di business concrete e

quantificabili, come il costo di acquisizione di un cliente (CAC), il customer lifetime value CLV e il tasso di abbandono. Si passa così da un approccio product centric a customer centric, che mette al centro di ogni decisione il cliente, con lo scopo di creare valore reciproco e instaurare relazioni di fiducia.

La fidelizzazione è diventata uno degli obiettivi principali perseguiti dalle aziende, proprio a causa dell'elevato costo che dovrebbero sostenere per acquisire nuovi clienti. In particolare, nel settore bancario resta cruciale poiché un cliente stabile tende a generare valore non solo attraverso i ricavi correnti, ma anche mediante l'utilizzo continuativo di servizi finanziari, la sottoscrizione di prodotti aggiuntivi e l'attivazione di un passaparola positivo sul territorio.

Oltre al fatto che stabilire una relazione solida con i clienti consente alle aziende di ottenere un vantaggio competitivo sul mercato.

Di fatto i consumatori, ad oggi, sono sempre più facilitati nelle loro decisioni finanziarie, grazie alla crescente conoscenza delle opportunità e dei rischi offerti, e grazie alla trasparenza delle condizioni a cui i servizi bancari sono offerti[1]. Questi due fattori, conoscenza e trasparenza, da un alto, ampliano la possibilità di confronto e di scelta del consumatore riguardo i prodotti offerti dagli intermediari, e dall'altro stimolano le banche a offrire prodotti migliori, agevolano la concorrenza e favoriscono l'innovazione(Tarantola 2007).

In tale contesto, il churn rate assume un ruolo cruciale, consentendo di quantificare il numero di clienti che, insoddisfatti del prodotto o servizio offerto, si

[1]La disciplina sulla trasparenza è costituita dal Testo Unico Bancario (TUB, Titolo VI), dalle delibere del Comitato Interministeriale per il Credito ed il Risparmio e dalle Istruzioni di vigilanza emanate dalla Banca d'Italia (Titolo X, Cap. 1). Per assicurare il rispetto della normativa la legge attribuisce alla Banca d'Italia e all'Ufficio Italiano dei Cambi (UIC) il potere di eseguire controlli e imporre sanzioni agli intermediari sottoposti alla sua vigilanza.

rivolgono spesso a intermediari più competitivi, consentendo di implementare misure che tengano i clienti legati all'azienda.

Il tasso di abbandono è strettamente legato, seppure in maniera opposta, al retention rate, che rappresenta la quota di clienti che continua a rimanere attivo in azienda: ridurre il churn rate aumenta direttamente la retention, e viceversa. L'obiettivo di un'impresa è quello di ridurre al minimo la defezione dei clienti e implementare le strategie di retention per mantenere alta la fedeltà dei clienti. Ed è proprio attraverso l'offerta di programmi fedeltà, sconti esclusivi, omaggi e contenuti personalizzati che le aziende si impegnano a soddisfare i clienti nel lungo termine.

Migliorare il tasso di abbandono, quindi, rappresenta una leva strategica fondamentale per ogni azienda. Secondo autorevoli studi (condotti da Bain&co e Harvard), infatti, un contenimento del churn rate di appena il 5% è in grado di innescare una crescita esponenziale dei profitti, che può oscillare tra il 25% e il 95% consolidando così la stabilità economica nel lungo periodo (Garabello 2023). Per monitorare questi fattori, e individuare le strategie di marketing efficienti ed efficaci, le aziende si avvalgono della churn analysis. In un mercato altamente competitivo, la churn analysis è uno strumento fondamentale di Customer Relationship Management (CRM) che permette alle banche di passare da un approccio reattivo (gestire il cliente che se ne va) a uno proattivo (prevenire l'abbandono). Ovvero individuare le campagne di marketing, le strategie per ridurre i costi operativi, e analizzare a fondo i clienti più propensi ad allontanarsi dalla banca.

## 3.2 L'IMPORTANZA DELLA CHURN ANALYSIS

L'attenzione verso il churn rate deriva dal fatto che tale indicatore consente di comprendere, misurare e anticipare la perdita di clientela in un contesto in cui la relazione con il cliente rappresenta uno dei principali fattori di creazione di valore (Majka 2024).

Per un'azienda, e in particolare per una banca, la perdita di clienti non comporta soltanto una riduzione quantitativa del numero di affiliati, ma implica una perdita di ricavi futuri, minori opportunità di vendita e un indebolimento della relazione fiduciaria costruita nel tempo.

In ambito bancario, inoltre, l'abbandono del cliente non coincide necessariamente con la sola chiusura del conto corrente, ma comprende più in generale tutte quelle situazioni in cui il cliente interrompe o riduce in modo significativo il proprio rapporto con l'istituto. Il fenomeno può manifestarsi, ad esempio, attraverso il trasferimento della liquidità verso altri intermediari, la cessazione nell'utilizzo di strumenti di pagamento, la rinuncia a prodotti accessori oppure una progressiva riduzione dell'operatività. In questa prospettiva, il churn bancario non rappresenta soltanto un evento finale, ma un processo graduale di disimpegno del cliente rispetto alla banca. Ecco, perciò, dei motivi che rendono la churn analysis centrale nel settore bancario.

Un primo aspetto riguarda la **redditività**: la letteratura e la prassi manageriale sottolineano come mantenere un cliente esistente sia generalmente meno costoso rispetto al procurarsene uno nuovo. A confermarlo è la ricerca condotta dalla Bain & Company, società globale di consulenza strategica, il cui

esito mostra che il costo di acquisizione di un nuovo cliente può arrivare ad essere superiore del 700%, rispetto al costo di retention di uno stesso. Le cause sono riconducibili ai costi commerciali e pubblicitari che vanno sostenuti per attrarre nuovi clienti all'azienda.

Altro aspetto riguarda il **valore del cliente nel lungo periodo**: nel settore bancario, infatti, la perdita di un cliente comporta non solo la rinuncia a ricavi correnti, ma anche al valore futuro atteso da quel cliente. Ogni abbandono cioè rappresenta una mancata opportunità di vendita incrementale in termini di upselling o cross-selling, limitando il margine di profitto(c'è una perdita di redditività marginale del cliente).

Infine, è importante perché ci permette di **salvaguardare la reputazione dell'impresa**: tassi di abbandono elevati possono significare problemi più profondi, come insoddisfazione, scarsa personalizzazione dell'offerta e custumer experience percepita insufficiente. Questi aspetti vanno previsti e contrastati per evitare che i clienti scontenti condividano la loro esperienza negativa e dissuadano i potenziali clienti (*Come creare un modello di abbandono dei clienti | Stripe* 2024).

## 3.3 LE CAUSE DI ABBANDONO

Comprendere le ragioni che spingono un cliente a interrompere il legame con il proprio istituto bancario è fondamentale per consentire ai marketers di intervenire per massimizzare la customer satisfaction. Un tasso di abbandono elevato può avere origine da diversi fattori, che riflettono sia l'esperienza del

cliente, che il panorama competitivo (Agarwal 2025). I principali motivi per cui i clienti abbandonano un servizio sono:

- Scarsa qualità del servizio clienti: nel momento in cui l'utente sperimenta un disservizio o un dubbio relativo al prodotto, il customer support rappresenta il principale, e talvolta unico, punto di contatto con l'istituto (Zeithaml, Berry e Parasuraman 1996). Se tale canale si dimostra lento, inefficace o sgradevole, l'iniziale frustrazione del cliente si traduce rapidamente nella ricerca di alternative sul mercato.

- Costi elevati: da un lato, l'applicazione di prezzi eccessivi o l'adozione di rincari periodici non giustificati da un reale potenziamento del servizio, tendono ad alienare la clientela storica. Dall'altro, le campagne di sconti aggressivi e sottocosto intraprese dai competitor riescono ad attrarre con estrema facilità i segmenti di consumatori caratterizzati da un'elevata sensibilità al fattore economico. La sfida che devono affrontare le imprese è gestire con attenzione le politiche di prezzo per mantenere la competitività, assicurando al contempo di continuare a fornire valore ai loro clienti.

- Pressioni competitive: le strategie commerciali messe in atto dai concorrenti esercitano una costante forza attrattiva sulla base clienti. Se l'offerta dei competitors sul mercato prevede prodotti qualitativamente superiori, a costi più bassi o percepiti come più accattivanti dai consumatori, quest'ultimi saranno più tentati ad abbandonare. Tale fenomeno si verifica soprattutto nei mercati caratterizzati da bassi costi di

transizione, in cui il passaggio da un'azienda all'altra non richiede sforzi burocratici o economici significativi da parte dell'utente. L'importante per le imprese è attuare strategie per mantenere il proprio vantaggio competitivo.

- Cause naturali: il tasso di abbandono può essere determinato da eventi biologici e fisiologici indipendenti dalla dinamica competitiva, primo fra tutti il decesso delle persone fisiche. A questi fattori si affiancano shock esogeni collegati al mutamento delle condizioni di mercato o delle preferenze dei consumatori. Durante i cicli economici recessivi, la necessità di contrarre la spesa costringe i soggetti a tagliare i consumi e i servizi ritenuti non essenziali. Parallelamente, i rapidi progressi dell'innovazione e il cambiamento nei gusti del pubblico possono determinare l'improvvisa obsolescenza di un'offerta commerciale, rendendola non più rispondente alle moderne necessità del mercato.

## 3.4 IL MACHINE LEARNING NELLA CHURN ANALYSIS

Nei paragrafi precedenti è stata messa in luce l'importanza del tasso di abbandono dei clienti, ma la vera sfida sta nel prevedere quali e quanti sono i clienti che lasceranno il servizio. Per contrastare questo fenomeno, la ricerca accademica e commerciale ha esplorato diverse metodologie predittive applicate ai dati transazionali e demografici delle banche. L'approccio standard si basa su modelli statistici tradizionali e di semplice implementazione come la regressione logistica classica e i singoli alberi decisionali. Questi garantisco-

no una chiara interpretabilità dei fattori di abbandono, pur mostrando limiti strutturali evidenti di fronte a grandi moli di dati e a relazioni complesse e non lineari. La recente evoluzione delle tecnologie di machine learning (tra il 2020 e il 2024) ha introdotto nel settore finanziario algoritmi d'insieme più robusti come il Random Forest, oltre a tecniche di boosting (XGBoost, CatBoost) e reti neurali deep. La letteratura scientifica ha evidenziato come non esista un modello predittivo universalmente ottimale, in quanto le prestazioni variano sensibilmente in funzione della natura strutturale dei dati bancari disponibili. Ad esempio, sebbene i modelli ensemble complessi offrano generalmente una precisione superiore su grandi volumi di dati non lineari, in presenza di forti componenti lineari un modello statistico tradizionale come la regressione logistica (Modello Logit) rappresenta ancora un'alternativa valida per la sua trasparenza e ridotta complessità computazionale. Il dibattito scientifico si concentra quindi sulla ricerca del compromesso ideale tra accuratezza predittiva ed esplicabilità del modello per supportare i processi decisionali di fidelizzazione.

Muovendo da queste premesse teoriche, il capitolo seguente traslerà l'analisi sul piano empirico attraverso lo studio del caso della BCC di Pergola e Corinaldo, dove i modelli verranno testati e confrontati sul dataset fornito dall'azienda, per decretare l'approccio più performante.

# Capitolo 4 – MATERIALI E METODI

## 4.1 ACQUISIZIONE DEI DATI

I dati empirici utilizzati per lo sviluppo e il confronto dei modelli predittivi di questa ricerca sono stati forniti dalla BCC di Pergola e Corinaldo, in forma rigorosamente anonima. Ricordiamo che l'obiettivo finale di questo studio è individuare i clienti che rischiano di abbandonare la banca. Inizialmente i dati forniti dall'istituto erano suddivisi in due database strutturalmente speculari, configurati come due rilevazioni fotonumeriche (cross-section) riferite rispettivamente alle date contabili del 1 Gennaio 2025 e del 1 Gennaio 2026. Grazie a questa articolazione temporale su base annuale si è potuto analizzare agevolmente il fenomeno del customer churn, osservando i comportamenti di persistenza o di abbandono dei clienti lungo un arco di dodici mesi. Il dataset nella sua formulazione originale, presenta dati per oltre 100.000 individui: la rilevazione relativa al 2025 registra le informazioni di ben 124.865 clienti, cifra salita a 126.318 osservazioni nel 2026. A ogni riga del dataset corrisponde un singolo codice NAG (Numero Anagrafica Generale), ovvero un codice numerico univoco e identificativo che la banca associa a ciascun cliente, socio o intestatario di conto, indipendentemente dal numero di rapporti (conti, mutui, depositi) aperti. Abbiamo inoltre 68 colonne, a ognuna delle quali corrisponde

una variabile differente, sia di natura quantitativa (età, saldi, data accensione) sia qualitativa (sesso, tipologia di NAG, comune di nascita). Il perimetro iniziale del campione si presenta tuttavia fortemente eterogeneo, includendo non solo le persone fisiche, ma anche società di capitali, ditte individuali, enti giuridici e altre forme di aggregazione istituzionale, oltre a presentare numerose variabili irrilevanti.

## 4.2 PROCESSO DI PULIZIA E SELEZIONE DEL CAMPIONE FINALE

Per rendere il campione di dati idoneo agli scopi della tesi e per evitare gravi distorsioni nei processi di stima statistica, è stato necessario attraversare una fase di data cleaning e di filtraggio delle osservazioni.

In primo luogo, si è proceduto a unire i due dataset in un unico file contenente di seguito i dati del 2024 e 2025, e alla riduzione della dimensionalità delle variabili, eliminando dal tracciato le colonne ridondanti, non informative o affette da eccessivi problemi di valori mancanti (missing values), mantenendo solo i predittori ritenuti teoricamente ed empiricamente rilevanti per la spiegazione del fenomeno del churn.

In secondo luogo, l'analisi è stata circoscritta al solo segmento retail delle persone fisiche, per cercare di comprendere le motivazioni psicologiche, comportamentali e di prezzo che guidano le scelte del singolo consumatore.

Successivamente, sul cluster delle persone fisiche è stato applicato un filtro anagrafico restrittivo, selezionando esclusivamente gli individui con un'età compresa tra 1 e 100 anni, così da depurare il campione da potenziali errori di

inserimento nei sistemi informativi o da anomalie anagrafiche.

Al termine di questo processo di scrematura e selezione, sono stati importati i dati su Gretl, software per l'analisi econometrica e statistica, dove l'orizzonte temporale della ricerca è stato segmentato considerando il tracciato del 2024 come l'anno di baseline. Questo significa che tutte le informazioni demografiche, anagrafiche e patrimoniali dei clienti utilizzate come variabili indipendenti (o predittori) fanno rigorosamente riferimento alla situazione consolidata al 31 dicembre 2024. Il campione finale su cui sono stati addestrati e testati i modelli si è assestato su un totale di 14.539 individui. Questo dataset finale, ripulito da rumore statistico ed eterogeneità strutturale, rappresenta la base empirica solida e robusta su cui si focalizzerà l'analisi descrittiva e la successiva applicazione della regressione Logit e dell'algoritmo di Random Forest.

Nel corso della sperimentazione è stata introdotta come caratteristica target la variabile "churn". Nel nostro caso, la variabile è stata formalizzata sfruttando l'informazione sul saldo del conto corrente registrato alla fine dell'anno successivo, rinominata *saldo25*. Come evidenziato dalla sintassi logica di Gretl *!ok(saldo25)*, la variabile churn è stata generata come una variabile dummy che assume il valore 1 se il saldo al 2025 risulta non pervenuto, nullo o mancante (segno di un rapporto estinto o strutturalmente interrotto), e il valore 0 qualora il cliente presenti ancora una regolare giacenza di conto corrente.

La variabile DESCR_COM_RES conteneva i nomi di oltre 360 comuni distinti, distribuiti su 14.539 clienti. Questa enorme dispersione causava anomalie statistiche, rendendo i modelli instabili (e distorcendo il test del Chi-Quadrato). I comuni sono stati quindi raggruppati e ricodificati in una nuo-

va variabile numerica discreta chiamata ZONA_GEO, contenente i tre poli geografici nettamente predominanti (che da soli coprono quasi un terzo del campione) e accorpando tutti gli altri in una categoria di controllo:

- Valore 1- Pergola: con 1.871 clienti, è l'area dell'entroterra pesarese-anconetano, storicamente baricentro operativo della banca.
- Valore 2- Senigallia: con 1.507 clienti, rappresenta il principale hub costiero e urbano.
- Valore 3-Corinaldo: con 1.451 clienti, è un ulteriore snodo strategico ad alta densità di clienti nel territorio d'origine della BCC.
- Valore 4- Altri Comuni: Categoria di controllo quantitativamente robusta che aggrega i restanti comuni del dataset, che contano 9.701 clienti.

Per consentire a OLS e Logit di elaborare correttamente l'informazione senza trattare i codici 1, 2, 3, 4 come numeri matematici, la variabile è stata infine convertita in variabili dicotomiche 0, se non è residente, e 1, se residente, ed escludendo la quarta classe dal computo delle regressioni per eleggerla a benchmark.

Statistiche descrittive delle principali variabili

| **Variabile** | **Media** | **Mediana** | **S.D.** | **Min** | **Max** |
|---|---|---|---|---|---|
| churn | 0,040 | 0,00 | 0,196 | 0,00 | 1,00 |
| SALDO | 13.375 | 2.644.20 | 35.524,00 | -864.480,00 | 958.550,00 |
| ETA | 51,91 | 52,00 | 18,89 | 1,00 | 100,00 |
| sq_ETA | 3.052,00 | 2.704,00 | 2.031,60 | 1,00 | 10.000,00 |
| Sesso_Femmina | 0,477 | 0,00 | 0,500 | 0,00 | 1,00 |
| ZONA_GEO | 3,307 | 4,00 | 1,0919 | 1,00 | 4,00 |

Come si può notare dalla tabella la media della variabile dipendente è pari a 0,040. Questo significa che soltanto il 4,0% dei clienti presenti nel dataset al 31/12/2024 ha registrato un abbandono (mancanza di saldo) nel corso del 2025.

La media dei saldi invece si attesta a 13.400 euro, mentre la mediana si ferma ad appena 2.640 euro. Tale divario, unito a una deviazione standard molto elevata (35.500 euro), riflette una classica distribuzione a "code pesanti", dove la maggior parte dei correntisti detiene piccole giacenze, mentre una ristretta minoranza possiede capitali ingenti (fino a un massimo di 959.000 euro). Di particolare interesse è il valore minimo, pari a -864.000 euro. Una siffatta giacenza negativa indica la presenza nel campione di posizioni con forti scoperti di conto o linee di credito concesse dall'istituto (affidamenti), un fattore che dal punto di vista relazionale vincola fortemente il cliente alla banca, riducendo la probabilità di churn.

Il campione analizzato mostra una clientela anagraficamente matura, con un'età media di 51,9 anni e una perfetta coincidenza con il valore mediano (52

anni). La deviazione standard di 18,9 anni indica una buona rappresentatività di tutte le fasce d'età, dai nuovi nati (minimo 1 anno, presumibilmente titolari di conti deposito minori o libretti di risparmio aperti dai genitori) fino ai centenari. La variabile sq_ETA (età al quadrato) tiene conto della potenziale non-linearità di questa componente. Per quanto riguarda la composizione di genere, la variabile Sesso_Femmina evidenzia una quota del 47,7%, delineando un campione sostanzialmente bilanciato.

Analogamente, per la variabile geografica (ZONA_GEO) si rimanda ai dettagli strutturali già discussi in precedenza, limitandosi qui a rilevare come il valore mediano pari a 4,00 confermi la robusta concentrazione della clientela nel cluster degli "Altri Comuni".

In seguito, dopo aver analizzato le statistiche descrittive, andiamo ad utilizzare uno dei principali strumenti di visualizzazione per individuare gli outlier nei dati: i box plot. I box plot sono uno strumento molto utile per individuare i valori minimi e massimi, la mediana e l'intervallo interquartile dei dati. I valori anomali nel grafico sono rappresentati come punti, il che li rende facilmente visibili.

La figura 4.1 ci mostra la distribuzione della variabile "SALDO", caratterizzata da una forte asimmetria e una polarizzazione estrema su entrambe le code della distribuzione:

- Coda superiore: Presenta una scia densa e quasi continua di outlier che si estende fino a un valore massimo di 958.550,00 €. Tale fenomeno riflette la presenza nel portafoglio della banca di una ristretta ma rilevante quota di clientela che detiene masse di liquidità ingenti, staccandosi nettamente

dalla massa retail, il cui valore mediano si attesta a soli 2.644,20 €.

- Coda inferiore: Mostra anomalie negative isolate ma profonde, capaci di raggiungere un picco minimo di -864.480,00 €. Queste osservazioni rappresentano i conti correnti strutturalmente in rosso o legati ad ampie linee di credito e affidamenti commerciali concessi dall'istituto.

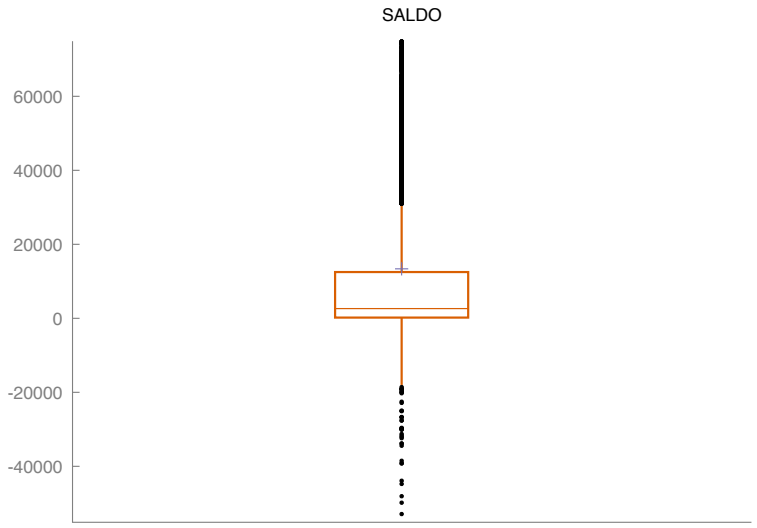


Figura 4.1: Boxplot della variabile SALDO per l'individuazione degli outlier.

La forte compressione della scatola centrale attorno allo zero evidenzia come la stragrande maggioranza dei correntisti presenti saldi contenuti, mentre la varianza complessiva del fenomeno sia dominata dai valori estremi. Sotto il profilo metodologico, questa specifica evidenza esclude la possibilità di normalizzare la variabile tramite semplici trasformazioni simmetriche e giustifica l'adozione di modelli non parametrici o basati su alberi di decisione (come la Random Forest), intrinsecamente robusti alla presenza di asimmetrie così marcate e di outlier non eliminabili dal dataset.

# Capitolo 5 – MODELLI PREDITTIVI DEL CHURN: OLS, LOGIT E RANDOM FOREST

## 5.1 INTRODUZIONE

Il presente capitolo costituisce il nucleo empirico del lavoro di ricerca e si pone l'obiettivo di modellare e stimare i determinanti del comportamento di churn della clientela all'interno dell'istituto di credito oggetto di studio. Al fine di garantire la massima robustezza statistica e di cogliere appieno le complesse relazioni sottostanti al fenomeno dell'abbandono, la scelta metodologica non si è fermata a un unico approccio, ma ha seguito un percorso progressivo e complementare che unisce l'econometria tradizionale alle potenzialità del machine learning.

Si è scelto di partire dal modello di probabilità lineare stimato tramite l'OLS classico, uno strumento intuitivo che funge da punto di partenza ideale per osservare la direzione dei legami e avere una prima misura immediata dell'impatto delle variabili. Subito dopo, l'indagine evolve verso il modello Logit binario, un passaggio naturale e indispensabile per correggere i limiti intrinseci della regressione lineare quando si lavora con variabili dipendenti dicotomiche, garantendo così che le probabilità teoriche stimate rimangano sempre comprese nell'intervallo reale tra lo zero e il cento per cento. A chiudere questo trittico

metodologico interviene infine la Random Forest, un algoritmo capace di muoversi dove i modelli strutturati faticano, intercettando relazioni non lineari, effetti soglia e interazioni nascoste tra le informazioni a nostra disposizione. Lungo tutta la trattazione, il funzionamento e la logica di questi tre strumenti verranno messi alla prova attraverso esempi pratici e concreti incentrati sulle variabili che si sono rivelate più significative nella fase descrittiva, come la consistenza dei risparmi dei correntisti, l'età e la nuova segmentazione geografica che abbiamo elaborato per valorizzare il legame con il territorio. Il capitolo si concluderà con un confronto diretto e serrato tra le diverse metodologie. Mettere specularmente a specchio la capacità interpretativa dell'OLS e del Logit con la forza predittiva della Random Forest permetterà non solo di individuare l'approccio più adatto alle reali esigenze strategiche della banca, ma anche di riflettere apertamente sui punti critici riscontrati durante le stime, sui limiti informativi del dataset e sui possibili margini di miglioramento per i futuri sviluppi della ricerca.

## 5.2 VARIABILI ESPLICATIVE

Poiché il profilo dettagliato e le statistiche descrittive dell'intero set di dati sono già stati ampiamente discussi nel capitolo precedente, in questa sede ci concentreremo esclusivamente sulle motivazioni teoriche ed econometriche che hanno guidato la selezione di questi specifici predittori all'interno dei nostri tre modelli. La scelta non è stata casuale, ma risponde all'esigenza di mappare tre macro-dimensioni fondamentali del comportamento umano e finanziario: la solidità della relazione economica, la componente demografica e il legame

con il territorio.

Il SALDO è stato isolato come la principale metrica di intensità del legame tra il cliente e l'istituto, un indicatore economico immediato di quanto un cliente sia "esposto" o "affezionato" finanziariamente alla banca. Sul fronte anagrafico, l'inserimento congiunto dell'ETA e della sua trasformazione quadratica sq_ETA risponde a una precisa necessità metodologica, quella di catturare le dinamiche non lineari legate alle diverse fasi del ciclo di vita biologico e professionale del cliente.

Per quanto riguarda il genere, non possiamo inserire contemporaneamente sia la variabile Sesso_Femmina sia quella speculare Sesso_Maschio. Se lo facessimo, incorreremo nel problema della multicollinearità e i modelli OLS e Logit fallirebbero, a differenza della Random Forest che invece è immune alla trappola delle dummy. Scegliendo di includere nel modello solo la dummy Sesso_Femmina, stiamo implicitamente eleggendo il genere maschile a nostro benchmark, ovvero a gruppo di controllo di riferimento. Il coefficiente che otterremo per la variabile femminile non descriverà un valore assoluto, ma indicherà lo scostamento, in termini di probabilità, delle donne rispetto agli uomini, a parità di tutte le altre condizioni.

Infine, la variabile ZONA_GEO è stata inserita per dare voce alla natura profondamente locale di una Banca di Credito Cooperativo. Anche in questo caso, poichè la variabile mappa il territorio dividendo il campione in quattro macroaree, siamo obbligati a sacrificarne una, escludendola dalla stima. Inserendo nel modello le tre dummy associate a Pergola, Senigallia e Corinaldo, stiamo trasformando la quarta categoria, quella degli "Altri Comuni", nello zero di

riferimento del nostro sistema. Di conseguenza, i tre coefficienti geografici che emergeranno dall'output ci diranno esclusivamente se risiedere in uno di questi tre hub storici aumenti o diminuisca il rischio di churn rispetto a chi vive in tutto il resto del territorio aggregato. La tabella seguente riassume la struttura logica e la tipologia di ciascuna variabile utilizzata.

Tabella 5.1: Quadro di sintesi delle variabili inserite nei modelli

| **Variabile** | **Tipo Variabile** | **Descrizione** |
|---|---|---|
| `churn` | Binaria | **Variabile Dipendente ($Y$):** identifica lo stato di abbandono del cliente (1) rispetto alla permanenza (0). |
| `SALDO` | Quantitativa continua | **Variabile Indipendente ($X$):** Saldo medio del conto corrente. |
| `ETA` | Quantitativa discreta | **Variabile Indipendente ($X$):** Età anagrafica del cliente. |
| `sq_ETA` | Quantitativa continua | **Variabile Indipendente ($X$):** Termine quadratico dell'età (cattura la non linearità). |
| `Sesso_Femmina` | Dicotomica / Dummy | **Variabile Indipendente ($X$):** Vale 1 se il cliente è donna (benchmark: genere maschile). |
| `ZONA_GEO` | Qualitativa categoriale | **Variabile Indipendente ($X$):** Configurazione territoriale originaria a 4 macro-aree. |
| `DZONA_GEO_1` | Dicotomica / Dummy | **Variabile Indipendente ($X$):** Indicatore specifico per l'hub di Pergola (1 = Sì, 0 = No). |
| `DZONA_GEO_2` | Dicotomica / Dummy | **Variabile Indipendente ($X$):** Indicatore specifico per l'hub di Senigallia (1 = Sì, 0 = No). |
| `DZONA_GEO_3` | Dicotomica / Dummy | **Variabile Indipendente ($X$):** Indicatore specifico per l'hub di Corinaldo (1 = Sì, 0 = No). |
| `DZONA_GEO_4` | Dicotomica / Dummy | **Benchmark:** Indicatore per l'area "Altri Comuni" (1 = Sì, 0 = No). Esclusa da OLS e Logit per evitare multicollinearità perfetta. |

## 5.3 IL MODELLO DI PROBABILITA' LINEARE: OLS

Il modello OLS (Ordinary Least Squares), o Minimi Quadrati Ordinari, costituisce lo strumento econometrico di base per la stima di relazioni lineari tra variabili. Dal punto di vista algebrico, l'obiettivo dell'OLS è trovare il vettore di coefficienti $\beta$ che minimizzi la funzione di perdita, definita dalla somma dei quadrati dei residui $SSR = e'e$, dove i residui rappresentano la differenza tra i valori osservati e quelli predetti (Lucchetti 2026). Quando la variabile dipendente è dicotomica ($Y \in 0, 1$), come nel nostro caso, il modello lineare assume la forma:

$$P(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{5.1}$$

In questo contesto, i coefficienti $\beta_j$ non rappresentano un semplice effetto geometrico, ma possono essere interpretati come effetti marginali, ovvero derivate parziali della funzione di regressione che indicano come varia mediamente la probabilità di churn a seguito di un aumento unitario della variabile esplicativa $X_j$, a parità di altre condizioni (ceteris paribus).

Lo stimatore OLS fornisce quindi una stima consistente, il che giustifica il ricorso al modello LPM come strumento esplorativo anche in presenza delle sue limitazioni strutturali.

Il principale limite del modello di probabilità lineare riguarda l'intervallo delle previsioni: trattandosi di una funzione lineare, il modello può generare valori di probabilità al di fuori dell'intervallo [0, 1], con previsioni negative o superiori all'unità prive di senso probabilistico.

Una seconda criticità è l'eteroschedasicità strutturale: la varianza degli errori

dipende sistematicamente dai valori dei regressori e non è costante, violando l'ipotesi di omoschedasicità richiesta dall'OLS classico. Ne consegue che gli stimatori OLS rimangono non distorti, ma perdono la proprietà di efficienza, rendendo necessaria la correzione degli errori standard per eteroschedasicità ai fini di un'inferenza statistica affidabile.

Nonostante tali limitazioni, il modello LPM rappresenta nella presente analisi il punto di partenza ideale: la sua semplicità interpretativa e la consistenza degli stimatori lo rendono prezioso per identificare la direzione e l'ordine di grandezza delle relazioni tra predittori e variabile dipendente, prima di procedere con le specificazioni più sofisticate offerte dal modello Logit e dall'algoritmo di Random Forest.

### 5.3.1 RISULTATI DELLA STIMA OLS

Le tabelle 5.2 e 5.3, riportano la stima del modello OLS, utilizzando tutte e 14539 le osservazioni e come variabile dipendente "churn".

Tabella 5.2: Risultati stima OLS/LPM

| **Variabile** | **Coefficiente** | **Std. Error** | **t-ratio** | **p-value** | **Sign.** |
|---|---|---|---|---|---|
| **const** | 0.0593928 | 0.0108613 | 5.468 | $4.62 \times 10^{-8}$ | *** |
| **SALDO** | $-4.617 \times 10^{-7}$ | $4.553 \times 10^{-8}$ | −10.14 | $4.39 \times 10^{-24}$ | *** |
| **ETA** | −0.0018815 | 0.0004253 | −4.424 | $9.74 \times 10^{-06}$ | *** |
| **sq_ETA** | $2.910 \times 10^{-5}$ | $3.959 \times 10^{-06}$ | 7.350 | $2.08 \times 10^{-13}$ | *** |
| **Sesso_Femmina** | −0.0079503 | 0.0032247 | −2.465 | 0.0137 | ** |
| **DZONA_GEO_1** | −0.0121345 | 0.0049091 | −2.472 | 0.0135 | ** |
| **DZONA_GEO_2** | 0.0027950 | 0.0053664 | 0.521 | 0.6025 | |
| **DZONA_GEO_3** | 0.0067201 | 0.0054775 | 1.227 | 0.2199 | |

Tabella 5.3: Statistiche di adattamento del modello OLS/LPM

| **Statistica** | **Valore** |
|---|---|
| **Osservazioni (n)** | 14.539 |
| **Media var. dipendente** | 0.039961 (tasso churn = $4,0\%$) |
| **$R^2$** | 0.022288 |
| **$R^2$ corretto** | 0.021817 |
| **F(7, 14531)** | 47.322 — p-value: $7.68 \times 10^{-67}$ |
| **Somma quadrati residui** | 545.350 |
| **S.E. della regressione** | 0.193727 |
| **AIC** | $-6.457,99$ |
| **BIC (Schwarz)** | $-6.397,31$ |

I risultati mostrano che il modello nel complesso è altamente significativo, supportato da un test F pari a 47,32 con un p-value associato virtualmente nullo ($7,68 \cdot 10^{-67}$), anche se l'$R^2$ pari a 0,022 indica che le variabili incluse spiegano il 2,2% della varianza di churn: un valore basso, ma atteso in modelli su variabili binarie con evento raro e fattori latenti non osservabili.

Dall'analisi dei singoli coefficienti emergono evidenze cruciali per il fenomeno in esame:

**SALDO**: Mostra un coefficiente pari a $-4,617 \cdot 10^{-7}$: la variabile più significativa del modello ($p < 0,01$). Tale valore negativo conferma che al crescere del saldo, la probabilità di abbandono diminuisce.

**ETA** e **sq_ETA**: L'inserimento del termine quadratico permette di isolare una relazione non lineare di tipo curvilineo. Il coefficiente di ETA è negativo ($-0.0019$), mentre quello di sq_ETA è positivo ($2,91 \cdot 10^{-5}$), entrambi significativi all'1%. Questa combinazione evidenzia una relazione a forma di U: la probabilità di fare churn tende a decrescere con l'avanzare dell'età, riflettendo la maggiore fedeltà della clientela matura, per raggiungere il punto di minimo

intorno ai 32 anni, dopodiché risale debolmente per le fasce anagrafiche più avanzate, verosimilmente in corrispondenza di eventi biografici quali pensionamento, ridistribuzione patrimoniale o trasferimento di residenza.

**Sesso_Femmina**: il coefficiente è pari a - 0,0080, significativo al 5% (p-value = 0,0137). A parità di saldo, età e zona geografica, le clienti di genere femminile presentano una probabilità di churn mediamente inferiore di circa 0,8 punti percentuali rispetto ai clienti maschi (categoria di riferimento).

**Variabili Territoriali**: Tra le dummy introdotte, solo l'hub geografico di Pergola (DZONA_GEO_1) mostra un effetto statisticamente rilevante (- 0.0121, $p<0.05$), indicando una maggiore fidelizzazione strutturale dei clienti di quell'area rispetto alla categoria di controllo "Altri Comuni". Le aree di Senigallia (DZONA_GEO_2) e Corinaldo (DZONA_GEO_3) non presentano invece variazioni significative rispetto al benchmark.

Nel complesso, il modello OLS fornisce un quadro esplorativo coerente e statisticamente robusto. Tuttavia, il rischio di previsioni al di fuori dell'intervallo [0, 1] e la limitata capacità di catturare relazioni non lineari complesse motivano il passaggio al modello Logit binario, illustrato nella sezione successiva.

## 5.4 IL MODELLO LOGIT

Il modello Logit binario rappresenta il principale strumento econometrico per la modellizzazione di variabili dipendenti dicotomiche. Come illustrato nel Capitolo 17 di Wooldridge 2013, l'idea fondamentale consiste nel sostituire la specificazione lineare dell'LPM con una funzione non lineare $G(\cdot)$ che garantisca

probabilità predette sempre comprese tra zero e uno:

$$P(Y = 1|X) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k) \tag{5.2}$$

Nel caso specifico del modello logistico, la funzione G è la distribuzione logistica cumulata: $G(z) = \exp(z)/[1+\exp(z)]$, una funzione strettamente crescente. La sua forma a "S" garantisce che qualunque combinazione lineare dei predittori produca sempre una previsione probabilistica compresa nell'intervallo [0, 1], superando la principale limitazione strutturale del modello LPM.

A differenza dell'OLS, i parametri del modello Logit vengono stimati attraverso il metodo della Massima Verosimiglianza (Maximum Likelihood Estimation, MLE). Questo approccio identifica il vettore di coefficienti $\beta$ che massimizza la funzione log-verosimiglianza del campione, ovvero la probabilità di osservare congiuntamente i valori di churn effettivamente registrati nel campione.

Lo stimatore MLE è consistente, asintoticamente normale ed efficiente sotto le ipotesi standard di corretta specificazione del modello.

Mentre per quanto riguarda l'interpretazione dei coefficienti, questi ultimi non indicano più l'effetto marginale diretto sulla probabilità, bensì l'effetto sui log-odds (il logaritmo del rapporto tra la probabilità dell'evento e la probabilità del non-evento), sebbene il segno del coefficiente mantenga lo stesso significato direzionale del modello MPL.

La significatività congiunta del modello viene verificata tramite il test del Rapporto di Verosimiglianza (Likelihood Ratio, LR), inoltre il tradizionale $R^2$ non è applicabile nel contesto Logit; in sua sostituzione, viene introdotto lo pseudo-

$R^2$ di McFadden, che esprime il guadagno in logverosimiglianza ottenuto aggiungendo le variabili esplicative rispetto al modello nullo. Valori compresi tra 0,2 e 0,4 sono generalmente considerati indicativi di un buon adattamento per modelli Logit applicati a fenomeni rari.

### 5.4.1 RISULTATI DELLA STIMA LOGIT

I risultati della stima logistica confermano la piena robustezza delle dinamiche isolate nel modello lineare, evidenziando una perfetta coerenza nei segni delle variabili esplicative. Il test del rapporto di verosimiglianza (Likelihood Ratio Test) pari a 404.99 (p<0,0000) certifica l'eccellente bontà di adattamento complessiva del modello.

Sotto il profilo dei singoli regressori, il SALDO si conferma l'esplicatore principale con un forte impatto negativo e un valore della statistica z pari a - 10,44. L'effetto curvilineo legato all'anagrafica trova ulteriore validazione: la variabile sq_ETA preserva la sua forte significatività (z=4.585), mentre la componente lineare ETA subisce un lieve riposizionamento statistico, rimanendo significativa solo al livello del 10% (z = -1,758, $p \approx 0.078$). Sia la componente di genere (Sesso_Femmina, z = -2,107) sia l'effetto territoriale positivo della zona di Pergola (DZONA_GEO_1, z = -2,032) mantengono inalterata la loro significatività ed influenza negativa sul tasso di abbandono.

L'analisi della performance predittiva rivela un'accuratezza complessiva apparentemente straordinaria, pari al 96.0% (con 13953 osservazioni correttamente classificate su 14539). Tuttavia, un esame critico condotto tramite la Confusion Matrix evidenzia il fallimento predittivo del modello sulla classe minoritaria: a

fronte di 581 casi reali di churn presenti nel dataset, il modello Logit classifica erroneamente come 0 (permanenza) la totalità di essi, predicendo il valore 1 soltanto per 5 soggetti che si rivelano poi falsi positivi. Questa evidenza, nota come problema dello sbilanciamento delle classi (class imbalance), deriva dal fatto che con un tasso di churn del 4% prevedere sistematicamente la classe maggioritaria garantisce di per sé un'accuratezza del 96% senza alcun reale potere predittivo sui clienti a rischio. Il Logit, massimizzando la verosimiglianza globale, tende a collocare la soglia di previsione in un punto che privilegia la classe dominante, risultando di fatto inutile ai fini dell'identificazione proattiva dei clienti che stanno per abbandonare la banca. La Figura 5.1 illustra visivamente questa criticità.

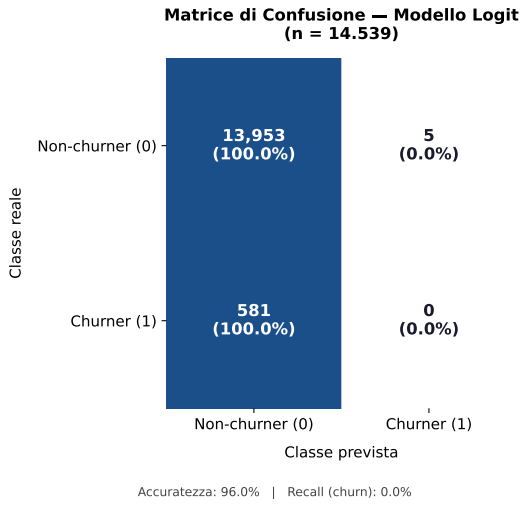


Figura 5.1: Matrice di Confusione: modello Logit

Questa limitazione motiva l'adozione di un approccio di machine learning come la Random Forest, intrinsecamente più robusto agli squilibri distributivi della variabile dipendente.

## 5.5 IL MODELLO RANDOM FOREST

Al fine di massimizzare la capacità predittiva e superare i limiti strutturali evidenziati dai modelli parametrici di fronte a dataset fortemente sbilanciati, l'analisi è stata estesa mediante l'applicazione della Random Forest.
Si tratta di un algoritmo di apprendimento automatico supervisionato per la classificazione e la regressione che costruisce un insieme (ensemble) di alberi decisionali e ne aggrega le previsioni tramite voto di maggioranza (James et al. 2013). Nel presente lavoro è applicata in modalità classificazione binaria: ogni albero emette un voto (churn o non churn) e la classe predetta è quella che raccoglie più voti tra i 500 alberi dell'ensemble. L'albero decisionale (unità base dell'algoritmo) suddivide ricorsivamente lo spazio dei predittori in regioni rettangolari, assegnando a ciascuna la classe modale delle osservazioni di addestramento. Ad ogni nodo viene scelta la variabile e la soglia che massimizzano la riduzione dell'indice di Gini:

$$G = \sum k \; \hat{p}mk(1 - \hat{p}mk) \tag{5.3}$$

dove $\hat{p}mk$ è la proporzione di osservazioni nel nodo m appartenenti alla classe k. Un valore di G prossimo a zero segnala un nodo omogeneo. La Figura 5.2 mostra un esempio di albero di classificazione: ogni nodo interno contiene la regola di suddivisione, i nodi terminali la classe predetta.

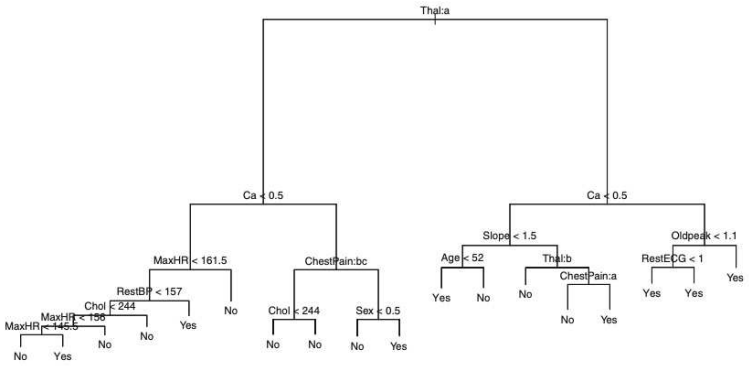


Figura 5.2: Esempio di albero decisionale di classificazione (Heart data, da James et al., 2013)

Un singolo albero soffre però di alta varianza. Il bagging (Bootstrap Aggregation) affronta questo problema addestrando B alberi su altrettanti campioni bootstrap e mediando i voti. Tuttavia, in presenza di un predittore dominante gli alberi risultano correlati tra loro, riducendo il beneficio della media.

La Random Forest introduce una decorrelazione forzata: ad ogni suddivisione estrae casualmente un sottoinsieme di m variabili (anziché tutte le p disponibili) e sceglie la migliore solo tra queste. Per la classificazione il valore standard è $m \approx \sqrt{p}$. Questo meccanismo diversifica gli alberi, riducendo più efficacemente la varianza dell'ensemble. La stima out-of-sample avviene tramite l'OOB error (Out-Of-Bag): ogni osservazione è predetta solo dagli alberi nei cui campioni bootstrap non è stata inclusa, fornendo una stima equivalente alla cross-validazione.

Due vantaggi strutturali distinguono la Random Forest dai modelli econometrici. Primo: la struttura ad albero cattura relazioni non lineari in modo automatico, rendendo superfluo il termine quadratico sq_ETA necessario in

OLS e Logit. Secondo: la Random Forest è immune alla multicollinearità perfetta tra dummy, consentendo di includere simultaneamente tutte e quattro le variabili geografiche (DZONA_GEO_1–4). L'importanza di ciascun predittore è misurata dal MeanDecreaseGini: somma delle riduzioni dell'indice di Gini prodotte dalle suddivisioni su quella variabile, mediata su tutti gli alberi.

### 5.5.1 RISULTATI DELLA RANDOM FOREST

La Random Forest è stata applicata in modalità classificazione binaria con 500 alberi e mtry = 2, addestrando il modello su 10.000 osservazioni (training set) e valutandolo su 4.539 (test set). Sotto il profilo delle performance complessive, l'algoritmo dimostra risultati eccellenti e un'ottima stabilità metodologica. La stima dell'errore OOB è pari al 2,4%, un valore strettamente speculare all'errore calcolato sul Test Set indipendente, pari al 2,82%. Il successo definitivo della Random Forest emerge dall'analisi comparativa della Confusion Matrix sul Test Set, dove l'accuratezza finale si stabilizza al 97,18%.

Il risultato più rilevante rispetto ai modelli econometrici precedentemente analizzati è la capacità di identificare correttamente 124 dei 182 churner presenti nel test set (recall = 68,1%). Il tasso di errore per i non-churner rimane contenuto all'1,6% (70 falsi positivi su 4.357). Il residuo 31,9% di churner non identificati (58 falsi negativi) rappresenta la sfida operativa più rilevante: sono i clienti per i quali un intervento proattivo di retention sarebbe stato più urgente.

Il grafico di importanza delle variabili (Figura 5.3) conferma la gerarchia identificata dai modelli econometrici: il SALDO è il predittore dominante (Mean-

DecreaseGini = 318,53), seguito dall'ETA (37,29). Le variabili geografiche e di genere presentano valori nettamente inferiori e tra loro simili (1,53–2,41), a indicare un contributo individuale marginale nella discriminazione tra churner e non-churner.

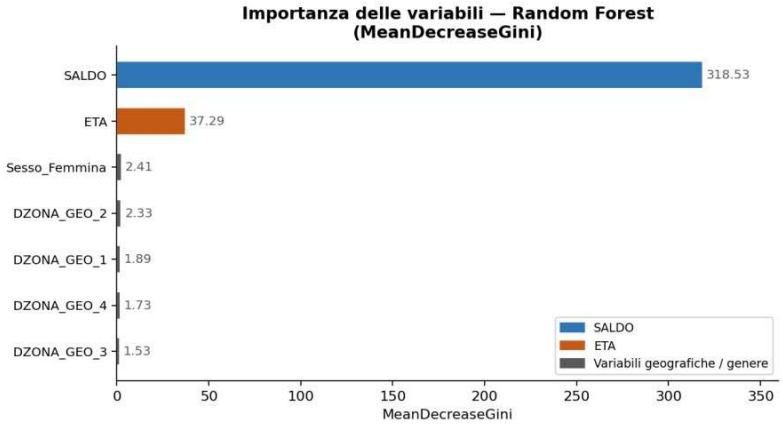


Figura 5.3: Importanza delle variabili (MeanDecreaseGini)

## 5.6 CONFRONTO TRA MODELLI

Il dataset oggetto di studio presenta una forte asimmetria distributiva, caratterizzata da un tasso di churn effettivo pari appena al 3,99% (581 clienti in abbandono su 14.539 osservazioni totali). Questa condizione di severo class imbalance rappresenta il principale banco di prova per le capacità predittive delle metodologie utilizzate. L'analisi ha impiegato tre approcci progressivamente più sofisticati per la previsione del churn nella BCC di Pergola e Corinaldo. Il confronto sistematico rivela non solo differenze di performance, ma una complementarità di ruoli: ciascun modello ha fornito un contributo specifico alla comprensione del fenomeno.

| Criterio | OLS (LPM) | Logit | Random Forest |
|---|---|---|---|
| Accuratezza complessiva | 97,8% (in-sample) | 96,0% | 97,18% |
| Churner identificati | N/D (no classif.) | 0 / 581 (0%) | 124 / 182 (68,1%) |
| Bontà adattamento | $R^2$ = 0,022 | McFadden $R^2$ = 0,083 | OOB err. = 2,4% |
| Non-linearità | No (richiede sq_ETA) | No (richiede sq_ETA) | Sì (automatica) |
| Multicollinearità | Problematica | Problematica | Immune |
| Interpretabilità | Alta | Media | Bassa (black box) |
| Ruolo nella tesi | Baseline esplorativo | Modello canonico | Predittore finale |

Figura 5.4: Confronto sintetico tra OLS, Logit e Random Forest. Note: l'accuratezza OLS è calcolata in-sample; Logit e Random Forest su campioni diversi (n = 14.539 e n = 4.539 rispettivamente).

Il modello OLS ha svolto il ruolo di strumento esplorativo: la significatività dei coefficienti e la loro diretta interpretabilità hanno permesso di identificare la struttura delle relazioni e la direzione degli effetti. L'$R^2$ contenuto (0,022) non inficia la validità delle stime dell'effetto parziale medio, atteso in modelli con variabile binaria ed evento raro.

Il modello Logit, pur migliorando la coerenza probabilistica delle previsioni e confermando i risultati dell'OLS, si è rivelato inutile ai fini dell'identificazione pratica dei clienti a rischio: zero churner identificati su 581. L'accuratezza del 96% maschera un fallimento classificativo completo sulla classe di interesse. Questo risultato mostra come le metriche di accuratezza globale siano fuorvianti in contesti con forte sbilanciamento tra le classi.

La Random Forest si afferma come il modello più performativo per lo scopo applicativo. È l'unica a identificare una quota significativa dei clienti a rischio

(68,1% di recall), senza modificare la composizione del dataset ma sfruttando la capacità intrinseca dell'algoritmo di modellare non-linearità e interazioni. La coerenza nella gerarchia delle variabili (SALDO dominante, ETA in seconda posizione) trasversale a tutti e tre i modelli rafforza la robustezza delle conclusioni sostanziali.

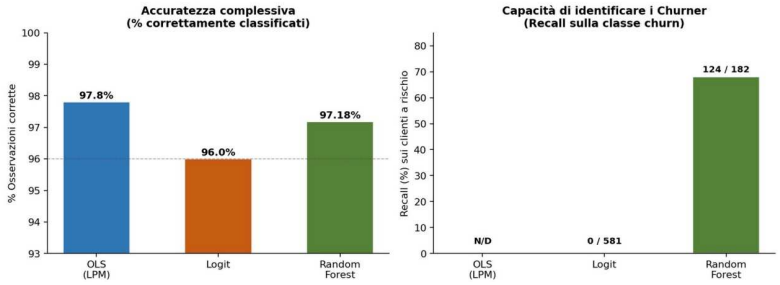


Figura 5.5: Confronto tra i modelli: accuratezza complessiva (sinistra) e recall sulla classe churn (destra).

# Capitolo 6 – CONCLUSIONI E SVILUPPI FUTURI

Dal confronto tra i diversi approcci, quello econometrico classico e quello computazionale è emerso che il modello che si presta meglio per questo tipo di task è la Random Forest.

Di fronte al problema del forte sbilanciamento delle classi, i modelli econometrici tradizionali (OLS e Logit) hanno mostrato un totale fallimento operativo. Al contrario, l'algoritmo di Random Forest, ha dimostrato una netta e indiscutibile superiorità predittiva, scardinando i vincoli imposti dal *class imbalance*, e confermandosi come l'unico strumento, tra quelli analizzati, dotato di reale efficacia operativa ai fini di una classificazione proattiva.

In un'ottica di *governance* bancaria e di *management* strategico, le implicazioni di questa tesi sono immediate. I risultati suggeriscono che l'adozione isolata di modelli statistici classici espone l'istituto al rischio di una totale cecità predittiva. Per implementare campagne di fidelizzazione e *retention* di successo, la banca deve necessariamente integrare algoritmi di Machine Learning nei propri sistemi di *Customer Relationship Management* (CRM). Disporre di un modello capace di intercettare oltre il 68% dei potenziali abbandoni consente di abbandonare logiche di marketing di massa, a favore di micro-interventi calibrati (es. offerte di tassi agevolati, consulenze personalizzate, azzeramento

temporaneo delle spese di tenuta conto) destinati esclusivamente ai segmenti a reale rischio di migrazione.

Naturalmente, il presente studio non è esente da limiti, come ad esempio la natura statica dei dati utilizzati, che fotografano il saldo medio come uno stock temporale fisso. Per cui sarebbe auspicabile che vengano condotti ulteriori approfondimenti in futuro, che tengano conto di informazioni aggiuntive e metodi più efficaci.

In primo luogo, sviluppi futuri potrebbero utilizzare tecniche specifiche per il trattamento dello sbilanciamento delle classi: metodi di sovracampionamento sintetico come SMOTE (Synthetic Minority Over-sampling Technique) o approcci di apprendimento cost-sensitive, che penalizzano asimmetricamente gli errori sulla classe minoritaria, potrebbero migliorare significativamente il recall sui churner senza aumentare la complessità del modello.

In secondo luogo, è consigliato il passaggio ad algoritmi di gradient boosting, in particolare XGBoost e LightGBM, che nella letteratura recente mostrano performance superiori alla Random Forest su dataset tabulari sbilanciati.

Infine, l'arricchimento del set informativo. Il dataset utilizzato in questa tesi include esclusivamente variabili patrimoniali, anagrafiche e geografiche. L'integrazione di informazioni comportamentali come la frequenza delle transazioni, l'utilizzo dei canali digitali, lo storico dei reclami e variazione del saldo nel tempo, potrebbe aumentare sostanzialmente il potere predittivo dei modelli, catturando segnali precoci di disaffezione che le sole variabili di stock non riescono a rilevare.

In conclusione, la tesi dimostra che il superamento della dicotomia tra eco-

nometria classica e Machine Learning risiede nella loro convergenza sinergica: la prima resta indispensabile per comprendere le determinanti strutturali e le logiche economiche del comportamento umano; il secondo si impone come l'architettura tecnologica imprescindibile per tradurre i dati in decisioni aziendali tempestive, efficienti e ad alto valore aggiunto.

# BIBLIOGRAFIA


Agarwal, Katyayani (2025). «Customer Churn Prediction Using Machine Learning Approaches». In: «International Journal of Scientific Research in Engineering and Management».

Garabello, Caroline (2023). «Artificial Intelligence in Sustainable Energy Industry: Prediction of Customer Churn Rate». Tesi di laurea mag. Roma: Luiss, p. 121.

James, Gareth et al. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer, pp. 303–331.

Lucchetti, Riccardo (2026). *Basic Econometrics*. Ancona: Università Politecnica delle Marche.

Majka, Marcin (2024). «Understanding Churn Rate». In: «ResearchGate».

Tarantola, A.M. (2007). «Dalla proprietà pubblica a quella privata: concorrenza ed efficienza del sistema bancario italiano». In: *Intervento alla Conferenza internazionale" The Perspective of the European banking and Financial Sector"(Mosca, 20 luglio 2007)*. URL: `https://www.bancaditalia.it/pubblicazioni/interventi-vari/int-var-2007/Tarantola_200707.pdf`.

Wooldridge, Jeffrey M. (2013). *Introductory Econometrics: A Modern Approach*. 5ª ed. Mason, OH: South-Western Cengage Learning, pp. 584–596.

Zeithaml, Valarie, Leonard Berry e A. Parsu Parasuraman (1996). «The Behavioral Consequences of Service Quality». In: «Journal of Marketing» 60.2, pp. 31–46.

## SITOGRAFIA


Angeletti, Veronique (2017). *Il Terzo Polo delle Bcc nelle Marche è la Bcc di Pergola e Corinaldo*. URL: https://www.civetta.tv/senza-categoria/2017/10/23/terzo-polo-delle-bcc-nelle-marche-la-bcc-pergola-corinaldo/.

BCC Pergola e Corinaldo (2024). *Statuto Sociale*. URL: https://www.pergolacorinaldo.bcc.it/doc2/scaricadoc.asp?iDocumentoID=2255633&iAllegatoID=0.

Castigli, Mirella (2022). *Cos'è il churn rate e come funziona?* URL: https://www.zerounoweb.it/big-data/cose-il-churn-rate-e-come-funziona/.

*Come creare un modello di abbandono dei clienti | Stripe* (2024). URL: https://stripe.com/it/resources/more/how-to-build-a-customer-churn-model-a-guide-for-businesses.

GlossarioMarketing (n. d.). *Churn rate: significato, definizione*. URL: https://www.glossariomarketing.it/significato/churn-rate/.

*La riforma si riforma 2018* (2018). URL: https://www.bccsiamo.it/riforma/la-riforma-si-riforma-2018/.