
# UNIVERSITÀ DEGLI STUDI DI SALERNO

**DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE**

Corso di Laurea Triennale in Statistica Per i Big Data

Prova finale in
Modelli statistici

## L'UTILIZZO DEI BIG DATA IN AMBITO BANCARIO

Relatore:
Ch.mo Prof.
Marialuisa Restaino

Candidato:
Massimiliano Pastorino
Mat. 0212800235

ANNO ACCADEMICO 2020/2021

# Indice

# Elenco delle figure

# Elenco delle tabelle

# Abstract

Il mondo sta assistendo a uno dei periodi di sviluppo scientifico e tecnologico più esaltanti della storia. Grazie a tecnologie informatiche, social media, dispositivi mobili, cloud e big data, il modo di fare business è cambiato velocemente e profondamente. Al centro del business e del nuovo modo di competere c'è un nuovo tipo di consumatore, esigente e informato. È di fondamentale importanza per un'azienda superare la concezione ormai passata d'instaurare relazioni basate sull'offerta indifferenziata di beni e servizi.

Nella prima parte della tesi, l'obiettivo è comprendere le caratteristiche fondamentali dei big data e come l'analisi di grandi mole di dati può trasformarsi in un vantaggio competitivo per gli istituti finanziari. Capiremo quali sono i campi di applicazione con i relativi vantaggi e quali sono le principali sfide affinché un'organizzazione possa diventare data driven e quindi basare le decisioni di business sull'analisi dei dati. Grazie ai big data infatti aziende e istituti finanziari possono sapere molte più cose sul proprio modo di fare business e avere una visione completa sui gusti e le preferenze dei propri clienti. Questo si traduce direttamente in nuova conoscenza per rendere più efficace il processo decisionale e per ottenere performance più soddisfacenti.

Nella seconda parte introdurremo sia i principali modelli statistici utilizzati per la classificazione sia le metriche di performance utilizzate per valutare un classificatore. Inoltre, saranno illustrate alcune tecniche statistiche molto utili come la $k$-fold cross-validation, che è un modo popolare (ed efficace) di selezionare i parametri di ottimizzazione negli algoritmi di apprendimento statistico. Infine, saranno illustrate le principali tecniche utilizzate per bilanciare le classi di riferimento in un data set, in modo tale da superare la problematica delle *Imbalanced Classes* che rappresenta un problema piuttosto comune nei problemi di classificazione, e che può compromettere il buon addestramento di un classificatore.

Nella terza e ultima parte, sarà illustrato il caso studio. Il data set in esame, *Bank Marketing Data Set* è scaricabile dal seguente sito web (`https://archive.ics.uci.edu/ml/datasets/bank+marketing`). Tra gli obiettivi del caso studio, sicuramente rientra l'analisi descrittiva ed esplorativa, grazie alla quale potranno essere individuate le principali caratteristiche dei clienti a cui l'istituto bancario portoghese si rivolge, proponendo loro l'eventuale sottoscrizione di un deposito a termine.

Analizzare la tipologia di clientela è utile in quanto possiamo capire il target a cui l'istituito finanziario si rivolge, e quindi se la tipologia di clientela è piuttosto omogenea per caratteristiche come età, tipo di lavoro, tipo di educazione, e cosi via, o se essa è invece eterogenea e si contraddistingue per caratteristiche diverse. Ogni caratteristica può influenzare l'esito della sottoscrizione di un deposito a termine proposto dall'istituto finanziario. Per questo motivo sarà molto importante analizzare le principali caratteristiche di un cliente in relazione alla variabile target, che rappresenta la sottoscrizione o meno del deposito a termi-

ne. Così facendo, scopriremo quali caratteristiche incidono in modo positivo sul deposito, e quali invece in modo negativo.

Nella seconda parte del caso studio, saranno stimati i modelli statistici utilizzati per i problemi di classificazione, descritti dal punto di vista teorico nella seconda parte della tesi. In questo caso, l'obiettivo cardine sarà quello d'implementare il miglior modello di classificazione, ovvero quel modello che riesca a classificare con un certo grado di accuratezza coloro che hanno sottoscritto un deposito a termine, in modo tale da ridurre al minimo i possibili errori, e in modo particolare di ridurre al minimo il numero di falsi negativi, ovvero coloro che effettivamente non hanno sottoscritto il deposito a termine proposto dall'istituto finanziario.

# CAPITOLO 1

# L'era dei big data

## 1.1 L'origine dei big data: Brevi cenni storici

I big data sono dati di grosse dimensioni che non possono essere analizzati e archiviati con strumenti tradizionali. Infatti sono richieste competenze specifiche e tecnologie avanzate per poter estrarre informazioni utili dalle grandi mole di dati.

Secondo il report *How Much Information* del 2010, ognuno di noi genera in media 12 gigabyte di dati ogni giorno. Basti pensare al numero di post che vengono creati sui vari social media, e non solo.

Tanti autori, come ad esempio McKinsey, descrivono i big data come un insieme di dati il cui volume è talmente grande, da superare la capacità dei convenzionali strumenti di gestione di dati di raccoglierli, immagazzinarli, gestirli e analizzarli.

Una definizione simile viene fornita anche da O'Reilly Media, definendo i big data come un volume di dati troppo grande, che si sposta troppo velocemente o che non si adegua all'architettura dei database usati. Viene specificato, inoltre, che per riuscire a «estrarre valore da questi dati, bisogna scegliere un metodo alternativo per elaborarli».

Per molti anni la raccolta e l'immagazzinamento di grandi set di dati era affidato ai governi. Nel 1943 è stato sviluppato dal governo britannico il primo dispositivo di elaborazione di dati, Colossus, che era in grado di decifrare le comunicazioni del regime nazista ed era in grado d'identificare pattern all'interno dei messaggi a una velocità di 5mila caratteri per secondo.

Negli anni sessanta e settanta sono stati sviluppati i primi data center per raccogliere e archiviare dati e sono stati sviluppati i primi database relazionali.

Con la creazione del World Wide Web negli anni novanta, servivano strumenti più avanzati per la raccolta e lo scambio di dati. Per questo motivo nel 1992 nacque il Teradata DBC 1012 che è diventato il primo sistema in grado di memorizzare 1 Terabyte di dati che è l'equivalente di 1000 GB.

Nel 2005 O'Reilly Media utilizzò per la prima volta il termine "big data". A partire da quest'anno grazie all'avvento di social media come Facebook e You Tube, la produzione dei dati online ha subito una crescita esponenziale. Sempre nel 2005 è stato creato il framework open source Hadoop di Yahoo e poco più tardi Spark, framework molto utilizzati per la gestione e l'immagazzinamento veloce ed efficiente dei big data.

Con il progresso tecnologico, non solo gli uomini sono produttori di dati, infatti con lo sviluppo dell'internet of things anche oggetti e dispositivi sono connessi a Internet, raccogliendo dati sui modelli di utilizzo dei clienti e sulle prestazioni del prodotto generando dati

in tempo reale.

## 1.2 L'obiettivo e l'utilità dei big data

L'analisi dei big data, anche detta 'big data analytics", utilizza una serie di tecniche sofisticate di analisi su grandi volumi di dati(strutturati e non strutturati) con lo scopo d'individuare pattern, correlazioni e tendenze per trasformare i 'raw data", ovvero i dati grezzi, in informazioni utili con le quali prendere decisioni ottimali. L'obiettivo dell'analisi dei big data infatti è quello di aiutare le aziende a identificare nuove opportunità di business in modo tale da avere clienti più soddisfatti e profitti maggiori.

I Big Data sono utili nelle varie attività aziendali, ad esempio nello sviluppo del prodotto, aziende come Netflix e Procter & Gamble utilizzano i Big Data per anticipare la domanda dei clienti. Netflix costruisce modelli previsionali per le nuove serie da produrre, inoltre ogni azione di un cliente attraverso l'interfaccia e le applicazioni corrisponde a un dato. Netflix traccia un profilo comportamentale. Infatti non conosce soltanto cosa vediamo, ma conosce anche il modo in cui usiamo la piattaforma di streaming. Sa esattamente se vediamo compulsivamente un film o una serie tv, o se la vediamo a scaglioni, se rallentiamo, se ci fermiamo e questo rappresenta un vantaggio rispetto ai canali tradizionali in quanto netflix distribuisce film e serie tv attraverso internet e non via cavo o via satellite.

I Big Data sono utili anche nell'attività aziendale di manutenzione predittiva, infatti possiamo prevedere e prevenire guasti meccanici analizzando sia i dati strutturati come l'anno, la marca e il modello dell'attrezzatura, sia i dati non strutturati che contengono ad esempio dati trasmessi dai sensori o anche messaggi di errore. Analizzando questi dati le organizzazioni possono effettuare la manutenzione in modo più efficiente in termini di costi e allo stesso tempo massimizzare i tempi di attività delle apparecchiature.

## 1.3 L'integrazione, la gestione e l'analisi dei big data

### 1.3.1 La Data Integration

I Big Data offrono nuovi insight che aprono nuove opportunità e modelli di business. Per iniziare sono necessarie tre azioni chiave, esse sono l'integrazione, la gestione e l'analisi. L'integrazione di dati appartenenti a fonti diversificate è molto importante in quanto solo in questo modo possiamo avere una visione del cliente a 360 gradi.

Con il termine *Data Integration* facciamo riferimento a tutte quelle azioni che hanno come scopo quello di unificare diverse sorgenti di dati, in modo tale da creare una vista unica su un determinato processo.

Il processo d'integrazione dei dati è fondamentale per diversi motivi, tra i quali rientrano sicuramente la ricerca di pattern nascosti, relazioni tra diverse fonti o anche gestire diversi tipi di dati, quali dati semi-strutturati o non strutturati. La Data Integration è essenziale per sviluppare analisi avanzate dalle quali estrarre informazioni utili al processo di business.

Al giorno d'oggi è chiaro come la correlazione tra varietà di fonti, tipologie e formati di dati sia di fondamentale importanza, tuttavia il processo d'integrazione spesso per le aziende rappresenta una sfida e ciò è dovuto principalmente alla mancanza di tecnologie adeguate.

Prima di parlare dei quattro approcci d'integrazione di dati, è utile chiarire quali sono

le tipologie di dati che un'organizzazione possiede. Possiamo individuare cinque categorie, esse sono:

- *Dati machine to machine:* Sono tutti i dati generati dall'interazione di dispositivi elettronici, possiamo pensare a sensori, dispositivi RFID e tanto altro. Grazie all'avvento dell'internet delle cose la produzione di questi dati ha subito una crescita notevole negli ultimi anni.
- *Dati people to people:* Sono tutti i dati generati dall'interazione tra persone, possiamo pensare ai dati che quotidianamente vengono prodotti sui social network dalle persone, ad esempio immagini, video, gif e tanto altro.
- *Public admin data:* Sono tutti i dati presenti in database pubblici e includono i cosiddetti Open Data, ovvero dati disponibili gratuitamente in modo tale che chiunque possa scaricarli e analizzarli secondo un proprio interesse. Non sono presenti restrizioni, copyright, brevetti o altri meccanismi di controllo.
- *Enterprise data:* Sono quei dati presenti all'interno dei data warehouse aziendali.

I processi di Data Integration possono seguire approcci differenti, in particolare gli approcci maggiormente diffusi sono:

- *Silos:* Rappresenta il metodo tradizionale, all'interno di un'organizzazione abbiamo diversi dipartimenti, ognuno dei quali ha una propria linea di business con finalità differenti. In questo caso lo storage dei dati è organizzato per dipartimento aziendale dove sono presenti repository differenti, gli ambienti sono tra loro isolati e i dati non sono integrati. Nell'approccio tradizionale viene a mancare uno dei requisiti fondamentali, ovvero la comunicazione.
- *Data Warehouse (DWH):* È un archivio informatico, l'obiettivo principale è quello di raccogliere i dati dai sistemi operazionali aziendali integrandoli con i dati provenienti da fonti esterne. All'interno dei data warehouse vengono archiviati i dati strutturati. Prima di essere inseriti nell'archivio dati, essi vengono formattati in una struttura impostata seguendo lo schema che viene comunemente chiamato schema-on-write. Esempi di dati strutturati sono ad esempio il numero di carta di credito, l'indirizzo, l'età, lo stipendio di un dipendente, questi campi sono facilmente interrogabili con il linguaggio SQL, infatti i database relazionali contengono solo dati di tipo strutturato.
- *Data Lake (DL):* È un ambiente di archiviazione dei dati dove non è necessario predisporre una struttura, infatti essi sono archiviati nel loro formato nativo. Con l'utilizzo dei data lake possiamo integrare quantità di dati davvero notevoli di qualsiasi formato e provenienti da qualsiasi fonte. All'interno dei data lake vengono archiviati i dati non strutturati. I dati non strutturati utilizzano uno schema definito schema-on-read. Tra i dati non strutturati rientrano diversi formati di file come .pdf, .docx, e-mail, immagini satellitari e dati provenienti dai sensori IoT. La differenza più grande tra dati strutturati e non, è che i dati strutturati sono di utilizzo comune, infatti vengono spesso utilizzati dall'utente aziendale medio, mentre l'utilizzo dei dati non strutturati richiede una competenza maggiore, spesso essi sono analizzati da figure esperte come ad esempio i data scientist.

- *Modello Integrato:* È un modello utilizzato da diverse organizzazioni, esso permette di rispondere alle differenti esigenze di storage, gestione e analisi di qualsiasi tipologia di dato, combinando in maniera integrata sia i data warehouse sia i data lake.

Secondo uno studio fatto nel 2018 dall'Osservatorio Big Data Analytics sulle iniziative di Data Integration delle grandi aziende italiane, solo una minoranza di esse ha costruito un Data Lake (circa un'azienda su quattro), mentre le restanti organizzazioni utilizzano Data Warehouse tradizionali dove i dati sono suddivisi in Silos che non comunicano tra di loro.

Inoltre laddove sono stati costruiti sistemi di storage per archiviare dati non strutturati (con database di nuova generazione come i NoSQL), si è scoperto che non sempre essi sono in comunicazione con il Data Warehouse aziendale. Circa il 46% delle aziende protagoniste dello studio ha affermato che uno dei problemi più grandi e che rappresenta certamente un freno allo sviluppo di Analytics è proprio quello legato alla sfida dell'integrazione di dati provenienti da diversi fonti(web, social, sensori, open data, ecc.).

### 1.3.2 La gestione dei big data: Il framework open source Hadoop

Come detto in precedenza, nel corso degli anni sono stati sviluppati framework open source come ad esempio Hadoop per gestire la grande mole di dati che un'organizzazione possiede, esso è utilizzato in moltissime organizzazioni. Si tratta di un framework che funge da strato intermedio tra un sistema operativo e il software che lo utilizza, con esso possiamo lavorare con migliaia di nodi e petabyte di dati. Uno dei vantaggi di Hadoop è quello di riuscire a elaborare anche i dati non strutturati provenienti dal mondo digitale, sociale e da tutto il mondo sensorizzato appartenente all'internet delle cose. I componenti principali di Hadoop sono:

- *Hadoop Common Package:* È un set di librerie e utility che usano anche altri componenti. Contiene i file jar e gli scipt necessari per avviare Hadoop, inoltre contiene anche tutta la documentazione su come Hadoop funziona e anche una sezione contributi che include i progetti della comunità di Hadoop.

- *Hadoop Distributed File System (HDFS):* Ha il compito di memorizzare i dati in un formato facilmente accessibile. Supporta una directory e un file system gerarchici convenzionali, grazie a essi i file vengono distribuiti attraverso dei nodi di archiviazione che vengono chiamati Data Nodi, in un cluster Hadoop.

- *Hadoop MapReduce:* Elabora i dati mappando un esteso set di dati e filtrandoli per ottenere i risultati desiderati. Con questo modello di programmazione anziché trasferire i dati al sistema computazionale, HDFS elabora i data nodi e la trasformazione di questi dati viene fatta in loco, attraverso MapReduce. Ogni nodo elabora i dati in base alla richiesta e poi inoltra i risultati che vengono consolidati su un nodo master, il quale si occupa anche di memorizzare tutti i metadati associati alla gestione dei cluster.

- *YARN (acronimo di Yet Another Resource Negotiator):* Gestisce la schedulazione dei lavori, allocando le risorse del cluster alle applicazioni in esecuzione, decidendo la prioritizzazione nel caso ci sia una contesa per le risorse disponibili. La tecnologia traccia e monitora l'avanzamento dell'elaborazione dei lavori.

- *Hadoop Ozone e Hadoop Submarine:* Queste sono due tecnologie più recenti che offrono agli utenti rispettivamente un negozio di oggetti e un motore di apprendimento automatico.

Hadoop è molto diffuso in quanto è accessibile e facile da capire. Uno dei problemi della memorizzazione di grandi quantità di dati è dovuto ai costi di mantenimento delle risorse e dei sistemi hardware per gestire il carico di lavoro. Hadoop utilizza il cosiddetto "commodity hardware" che consente di utilizzare l'hardware in modo flessibile con sistemi a basso costo che si possono trovare comunemente in commercio, quindi non richiede sistemi brevettati o hardware personalizzati con costi decisamente proibitivi.

### 1.3.3 I tipi di data analytics

Dopo aver integrato e gestito la mole di dati che un'organizzazione possiede, per estrarre conoscenza dai dati si passa alla fase di analisi. È possibile distinguere quattro categorie principali di data analytics, a seconda dello scopo dell'analisi:

- *Analisi descrittiva:* Attraverso la costruzione di grafici e tabelle viene utilizzata per la realizzazione di report, l'obiettivo principale è quello di effettuare un'analisi riassuntiva e descrittiva degli eventi, in questo caso si cerca di rispondere alla domanda " Cosa è successo?".
- *Analisi predittiva:* É una delle tipologie di data analytics più utilizzate ed è un tipo di analisi che permette di delineare scenari di sviluppo futuri, in questo caso si cerca di rispondere alla domanda "Cosa potrebbe accadere in futuro ?". Nell'analisi predittiva si cerca d'identificare delle correlazioni o delle relazioni di causa-effetto fra i dati.
- *Analisi prescrittiva:* Grazie al machine learning e al supporto dell'intelligenza artificiale oggi non ci limitiamo solo a fare delle previsioni relative a determinati eventi, ma è possibile anche fare delle previsioni riguardo alle azioni ottimali da intraprendere. In questo caso si cerca di comprendere "Cosa succederebbe se scegliessimo l'opzione A ?", risparmiando così le risorse che verrebbero impiegate per provare tutte le soluzioni a disposizione e consentendo di effettuare la scelta (probabilmente) più efficace fin dall'inizio. In questo caso si cerca di rispondere alla domanda "Come potremmo rispondere a un evento futuro?".
- *Analisi diagnostica:* È un tipo di analisi che consente d'identificare nello specifico le cause che hanno portato alla situazione attuale e quindi il perché di qualcosa o di un dato evento. Le aziende usano tecniche come drill-down e data mining per determinare le cause di trend o avvenimenti, in questo modo le organizzazioni possono identificare, ripetere e ottimizzare le azioni che hanno dato dei risultati positivi. In questo caso si cerca di rispondere alla domanda "Perché qualcosa è successo?".

## 1.4 Le caratteristiche dei big data: Le 5 V

I big data presentano determinate caratteristiche, esse sono riassunte nelle cosiddette "5V". Esse sono:

- *Volume:* La prima caratteristica è sicuramente il volume, le organizzazioni raccolgono dati da diverse fonti, ad esempio dispositivi intelligenti (IoT), social media, apparecchiature industriali e altro ancora. Ogni secondo attraversano la rete più dati di quelli

che vi erano immagazzinati appena vent'anni fa, basti pensare che nel 2012 ogni giorno venivano creati circa 2.5 esabyte di dati, un numero che è raddoppiato più o meno ogni quaranta mesi. Oggi si parla ormai di zettabyte e addirittura di brontobyte di dati, quantità ben superiori al terabyte. Si tratta ovviamente di una quantità di dati che non può essere immagazzinata o elaborata dai sistemi convenzionali di gestione dei dati e che richiede delle tecnologie ad hoc.

- *Velocità:* La velocità di creazione dei dati viene considerata ancora più importante dei volumi in quanto permette a un'azienda di essere molto più agile dei concorrenti, pensiamo a Tag RFID, dispositivi intelligenti che generano flussi di dati in tempo reale che possono essere analizzati per prendere decisioni strategiche nel minor tempo possibile. Possiamo pensare al gruppo di ricercatori del MIT Media Lab, che hanno utilizzato i dati identificativi dei telefoni cellulari per stabilire il numero di persone che si trovavano nei parcheggi dei grandi magazzini Macy's nel giorno del "venerdì nero" che coincide con l'inizio delle vendite natalizie negli Stati Uniti. Ciò ha permesso alla direzione dell'azienda di stimare le vendite effettuate in quel giorno critico ancor prima che queste venissero contabilizzate. La rapidità quindi risulta fondamentale per prendere decisioni strategiche nel minor tempo possibile e per ottenere un vantaggio competitivo, pensiamo ad esempio agli analisti di Wall Street.

- *Varietà:* Per varietà invece si intende la diversa forma dei dati provenienti dalle diverse fonti, e come detto in precedenza, pensiamo sia ai dati strutturati (ovvero quei dati organizzati in tabelle, ad esempio dati finanziari o dati di vendita per tipi di prodotto, ecc.) sia ai dati non strutturati che rappresentano una buona parte dei dati generati al giorno d'oggi, pensiamo a fotografie, registrazioni vocali, video, e tanto altro ancora.

- *Veracità:* I big data sono utili in quanto ci aiutano a prendere decisioni. Partendo da questa base, si capisce come sia importante effettuare un "controllo qualità" adeguato dei dati. I dati devono essere affidabili, raccontare il vero. Come detto in precedenza, i dati provengono da fonti diversificate, l'operazione di collegamento, pulizia e trasformazione risulta non proprio semplice, tuttavia questo processo d'integrazione è di fondamentale importanza in quanto "Bad data is worse than no data" .

- *Valore:* Possiamo sfruttare i big data per generare nuova conoscenza, "trasformare" i dati in valore. Possiamo sfruttare i dati per fare delle previsioni che consentono di ottimizzare le decisioni di business, pensiamo ad esempio all'analisi dei dati relativi ai consumi per prevedere il comportamento di acquisto dei consumatori e sulla base di queste previsioni offrire nuovi prodotti e servizi. L'analisi dei dati permette di automatizzare alcuni processi aziendali, si possono fornire risposte più adeguate ai clienti in maniera veloce in base al loro comportamento online (e anche offline).

## 1.5 La figura professionale del Data Scientist

La figura professionale che ha il compito di analizzare questa mole di dati viene comunemente chiamata Data Scientist. È un professionista di alto livello, con una preparazione e una curiosità tale da poter fare scoperte nel mondo dei big data.

Negli ultimi anni la domanda di data scientist è cresciuta molto di più dell'offerta. Infatti in alcuni settori la scarsità di data scientist sta diventando un vincolo pesante.

Il compito principale di questa nuova figura professionale è quello di fare nuove scoperte

mentre navigano nel mare magnum dei dati; essi sono in grado di organizzare logicamente grandi quantità di dati destrutturati e di renderne possibile l'analisi. Man mano che fanno delle scoperte, comunicano quello che hanno appreso e presentano nuove implicazioni a livello di business.

La competenza di base comune a tutti i data scientist è la capacità di scrivere codice, ma la caratteristica dominante è sicuramente la fortissima curiosità, il desiderio di andare oltre la superficie di un problema. Questo richiede il cosiddetto pensiero associativo che caratterizza gli scienziati più creativi in qualunque campo.

La parola "scientist" è fondamentale per capire il ruolo di questa nuova figura professionale innovativa, infatti alcuni dei data scientist più brillanti hanno conseguito il PhD in materie esoteriche come l'ecologia e la biologia dei sistemi. Anche i fisici sperimentali devono progettare strumenti, raccogliere dati, effettuare esperimenti e comunicare i risultati. Di conseguenza le aziende che sono alla ricerca di persone in grado di lavorare sui dati fanno bene a selezionare esperti di fisica, di scienze sociali o anche esperti d'informatica, matematica o economia.

In generale è importante tenere in mente l'immagine dello scienziato perché la parola "data" può portare su una strada sbagliata. Infatti un analista quantitativo potrebbe essere molto bravo ad analizzare dati, ma non ad assemblare una massa d'informazioni destrutturate e a convertirle in una forma che può essere analizzata. A differenza delle professioni informatiche tradizionali, i data scientist devono possedere anche competenze sociali.

## 1.6 I Vantaggi dell'analisi dei big data in ambito bancario

Dopo la grande recessione del 2008 che ha colpito in maniera diffusa le banche globali, l'analisi dei big data ha goduto di una popolarità sempre più crescente nel settore finanziario. Molte delle organizzazioni BFSI(Banking, financial services and insurance) iniziarono a digitalizzare i loro processi operativi con l'acquisizione di nuove tecnologie come Hadoop e RDBMS(database relazionali) con l'obiettivo cardine di massimizzare i guadagni di business e ottenere un vantaggio competitivo sul resto delle istituzioni finanziarie.

La quantità di dati generati in questo settore è davvero notevole, ad esempio le banche possono utilizzare le informazioni transazionali di un cliente per monitorare nel corso del tempo il suo comportamento, questo permette di tracciare un profilo comportamentale.

L'offerta di servizi maggiormente personalizzati contribuisce alla fidelizzazione, la customer experience dunque diventa una questione di maggior rilievo, il cliente viene visto come l'attore principale del business. L'obiettivo del marketing diventa prevalentemente quello di fidelizzare la clientela acquisita piuttosto che quello di ampliare il proprio mercato a discapito della qualità del servizio offerto. La capacità dell'impresa nello stabilire, rafforzare e mantenere nel tempo relazioni con i consumatori diventa il core business ed è proprio questa la principale differenza con la visione classica del marketing in cui il cliente viene visto come una figura esterna.

Oltre al marketing personalizzato, un altro fronte importante è quello della fraud detection. Le frodi finanziarie sono in netto aumento e questo rappresenta un problema grande per le banche in quanto sono responsabili dei rimborsi. Analizzare i dati e rilevare in maniera istantanea variazioni comportamentali che si discostano dalla normalità assume un ruolo fondamentale nella prevenzione delle frodi. Questo rappresenta un buon biglietto da visita, la banca può far leva su ciò per attirare clienti sempre più preoccupati e che sono inclini a

favorire istituti bancari capaci di mantenere un alto livello di sicurezza. Qualsiasi comportamento fuori dal comune come un prelievo oneroso su un conto corrente che in media preleva una quantità di denaro inferiore innesca un processo di vigilanza con il quale opportune verifiche vengono avviate. Individuata la frode si procede con il conseguente blocco del conto corrente prima che venga prosciugato dal truffatore.

Anche la valutazione del rischio di credito assume un ruolo importante, modelli di machine learning possono confermare sulla base di determinate caratteristiche come il saldo, il tipo di lavoro, il tipo di educazione, la storia creditizia che certamente rappresenta uno dei parametri più importanti, la praticabilità di un prestito o la concessione di un mutuo. Le piattaforme di social lending di maggior successo come "Lending Club" quando devono concedere un prestito, assegnano un rating (punteggio) a ogni utente e di conseguenza assegnano un mercato per ogni prestito. Il rating viene calcolato tenendo in conto diversi fattori, spesso accade che un richiedente prestito non abbia una storia creditizia o che gli utilizzi delle carte di credito non siano registrate nelle varie banche di dati pubbliche. In questi casi, quando non si hanno dati finanziari attendibili, il merito di credito potrebbe basarsi esclusivamente sul rating di credito social frutto della raccolta dei dati dei profili social del richiedente incrociati con i pagamenti mobile, stile di vita, relazioni, desideri e acquisti con carte di credito. Un ulteriore esempio è quello di Zestfinance che è una fintech americana creata nel 2009 da ex dipendenti Google, essa da informazioni sulla capacità degli individui di ripagare i debiti mediante l'utilizzo di strumenti di big data analysis con funzione di autoapprendimento. Attraverso algoritmi e modelli predefiniti sono in grado mediamente di determinare il rischio d'insolvenza del soggetto con un indice di efficacia superiore al 40% rispetto ai metodi tradizionali di scoring con costi inferiori del 30%.

L'utilizzo e l'analisi dei big data è utile anche sul lato della gestione dei risparmi, infatti un consulente che conosce le abitudini di spesa, attraverso la raccolta dei dati sulle transazioni di pagamento e la propensione al rischio di un cliente, sarà in grado di proporre al cliente in questione prodotti capaci di rispondere al meglio alle sue esigenze. Su tale linea si sono mossi anche istituti bancari italiani come Intesa SanPaolo e UniCredit Banca, infatti entrambe attraverso l'uso dei big data hanno cercato di migliorare la qualità del servizio offerto per accrescere il grado di fidelizzazione spostando il focus principale dal successo nelle vendite alla qualità della relazione. La tecnologia ha di fatto reso possibile misurare la qualità della relazione mediante un indice tangibile, attraverso specifici modelli analitici che monitorano la frequenza, l'intensità e l'efficacia del rapporto con il cliente. Le banche si sono attrezzate per comprendere in maniera oggettiva come e quanto sta curando i propri clienti.

Un altro vantaggio è la gestione delle prestazioni, BNP Paribas è un gruppo bancario internazionale che sfrutta i big data per migliorare la propria produttività e risolvere man mano i problemi che si presentano. BNP può capire ad esempio quale filiale sta attraendo più clienti o quale filiale presenta il più alto livello di sicurezza(in termini di frodi) e di fidelizzazione, inoltre con l'utilizzo di giuste metriche di performance si può monitorare costantemente anche l'efficacia lavorativa dei singoli dipendenti.

## 1.7 Problematiche riguardo l'analisi dei big data in ambito bancario

### 1.7.1 Le infrastrutture legacy

L'analisi dei big data porta con se molti vantaggi, tuttavia ci sono sfide importanti da affrontare. Non tutti gli istituti bancari possono contare su un infrastruttura tecnologica adatta, spesso si ha necessità di aggiornare infrastrutture cosiddette legacy, ovvero sistemi informatici obsoleti che continuano a essere usati. I database strutturati su cui fino a poco tempo fa venivano immagazzinate quasi tutte le informazioni aziendali ora sono inadeguati a conservare e a processare i big data. Gli strumenti a disposizione sono sensibilmente migliorati negli ultimi anni e in linea generale queste tecnologie non sono proibitivamente costose e il software è quasi tutto open source, possiamo pensare a Hadoop che è uno dei framework più utilizzati.

Aggiornare le infrastrutture digitali è importante, pensiamo all'esempio della Commonwealth Bank of Australia (CBA) che nel 2005 aveva il peggior punteggio del settore in tema di soddisfazione dei clienti. Il nuovo amministratore delegato Ralph Norris che proveniva dal mondo dell'IT, scopri sistemi vecchi di decenni e questo influiva in modo negativo sulle prestazioni dei dipendenti. Studiando le lamentele dei clienti che riguardavano principalmente lunghe attese e prodotti scadenti, decise di sviluppare competenze digitali con l'obiettivo cardine di passare dall'ultimo al primo posto nella classifica della customer satisfaction. Con il progetto Finest Online, il servizio di banking venne modernizzato, grazie all'app mobile, CBA aiuta i clienti a capire se si possono permettere una determinata casa e avvia in sole ventiquattro ore le procedure per l'approvazione di un prestito richiesto online(contro i 14-21 giorni necessari in precedenza), inoltre ha costruito anche un sistema di videoconferenze in tutte le filiali, cosi facendo i clienti(sopratutto quelli che vivono nelle zone rurali) possono relazionarsi con i consulenti della banca in modo più semplice. La strategia di Ralph Norris ha contribuito a portare la banca al primo posto della classifica 2013 della customer satisfaction nel settore del retail banking, con un aumento delle azioni di oltre l'80%, dunque aggiornare e costruire strategie incentrate sui dati e sulle tecnologie moderne diventa una questione fondamentale.

Un'altra difficoltà sicuramente risiede nella raccolta dei dati, spesso capita di dover raccogliere dati provenienti da diversi reparti della banca e questo può causare complicanze, inoltre anche se la raccolta va a buon fine è necessario eliminare tutti i dati irrilevanti prima che diventino utilizzabili per l'elaborazione e l'analisi. Come già detto in precedenza la qualità dei dati è fondamentale, le operazioni di data wrangling che hanno come scopo quello di preparare i dati ricoprono un ruolo fondamentale nella vita di un data scientist.

### 1.7.2 La privacy dei dati raccolti

Un'altra delle sfide più importanti riguardo l'uso dei big data sicuramente risiede nella privacy e sulla proprietà di tali dati personali raccolti. Negli USA, nel 2009 la Bank of America ha registrato il brevetto di un metodo che consente ai consumatori di monetizzare l'utilizzo dei propri dati. Tutte le informazioni sono raccolte nella cosiddetta "banca delle informazioni", chi volesse utilizzarle pagherebbe una tariffa direttamente al possessore, che avrebbe anche traccia del loro utilizzo. Questo è un tipico esempio di come alcune banche americane affrontano il problema della privacy, gli utenti vengono pagati per l'utilizzo dei loro dati sensibili.

Gli istituti finanziari non si accontentano della mera raccolta dei dati sensibili, spesso le banche utilizzano i dati sensibili raccolti come fonte ulteriore di ricavi tramite la rivendita degli stessi a vari livelli e altre aziende di servizi. Molte banche e società emittenti di carte di credito vendono i propri dati grezzi relativi ai propri clienti attraverso intermediari specializzati come Cardylitcs.

Non tutte le organizzazioni sono aperte e trasparenti riguardo i metodi che adottano nella gestione dei dati, la maggior parte di esse preferisce tenere i consumatori all'oscuro. Le organizzazioni che adottano una politica di trasparenza sulle informazioni che raccolgono, che lasciano ai clienti il controllo sui loro dati personali saranno considerate affidabili e si guadagneranno un accesso sempre maggiore ai dati.

Nel 2014 è stato condotto uno studio per capire il tipo di atteggiamento che i consumatori avevano riguardo i dati, in particolare sono state intervistate 900 persone di cinque paesi diversi(Cina, Germania, Gran Bretagna, India, Stati Uniti) il cui mix è rappresentativo del popolo d'internet. Il 97% delle persone intervistate ha espresso preoccupazione per l'uso che le aziende o la pubblica amministrazione possono fare di queste informazioni, tra i principali timori figurano il furto d'identità, anche se non conoscono in modo preciso che tipo di dati stanno rivelando. Di solito, nei vari paesi le persone attribuiscono maggiore importanza alle informazioni relative ai documenti d'identità, alla situazione sanitaria e alle carte di credito, mentre poca importanza viene data alla collocazione geografica e ai dati anagrafici. In generale possiamo suddividere i dati in tre categorie, esse sono:

- *Dati dichiarati dal proprietario:* In generale queste informazioni vengono fornite spontaneamente, ad esempio l'indirizzo mail, il curriculum scolastico o lavorativo, l'età, il sesso e cosi via.

- *Scorie digitali:* Sono quei dati che ognuno di noi lascia in giro quando si usano dispositivi mobili e servizi web. Esempi di scorie digitali sono la cronologia di navigazione e la collocazione geografica.

- *Dati di profiling:* Sono quei dati che vengono utilizzati per creare previsioni su comportamenti e interessi individuali, essi si ricavano combinando insieme sia i dati dichiarati dal proprietario, sia le scorie digitali e altro ancora.

Dal caso studio è emerso che i dati dichiarati dal proprietario sono quelli a cui le persone attribuiscono il valore minore, a seguire abbiamo le scorie digitali e infine i dati di profiling. Oltre alle tre categorie di dati, sono state considerate anche tre tipologie di utilizzo dei dati, esse sono:

- *Migliorare un servizio o un prodotto:* Per migliorare un servizio o un prodotto possiamo pensare a una applicazione che offre percorsi ottimizzati in base all'ubicazione dell'utente.
- *Facilitare marketing o pubblicità mirate:* Per facilitare il marketing o pubblicità mirate possiamo pensare agli annunci pubblicitari basati sulla cronologia di navigazione.
- *Generare ricavi attraverso la rivendita:* Per generare ricavi attraverso la rivendita, possiamo pensare alle organizzazioni che vendono a terzi i dati degli acquisti fatti con la carta di credito.

Ciò che è emerso dal caso studio è che più le persone attribuiscono valore ai dati, più si aspettano di ricevere qualcosa dalle aziende in cambio della loro condivisione. I consumatori sono ben disposti a cedere i propri dati quando essi vengono utilizzati per migliorare un prodotto o un servizio, mentre il valore di ritorno atteso cresce se i dati vengono utilizzati per attività di marketing mirato. Infine il valore di ritorno atteso cresce se i dati sono rivenduti a terzi. In generale più un organizzazione ispira fiducia, più i consumatori saranno disposti a condividere i propri dati e la prova è data da un sondaggio effettuato tra i consumatori su 46 aziende di sette diversi settori in varie parti del mondo. I consumatori devono valutare le aziende seguendo la seguente scala:

- *Affidabili:* In cambio del servizio desiderato non avrebbero problemi a scambiare dati sensibili.
- *Inaffidabili:* In cambio di un servizio essenziale fornirebbero dati sensibili e solo se è necessario.
- *Completamente inaffidabili:* Con quell'azienda non condividerebbero mai dati sensibili.

Dai risultati del sondaggio è emerso che l'87% dei consumatori ha fiducia nei medici di famiglia, a seguire con una percentuale pari all'85% abbiamo società di pagamento e carte di credito come Paypal e Alipay che sono quelle che hanno ricevuto le valutazioni migliori. Le banche e le compagnie di assicurazione presentano una percentuale di fiducia pari all'76%. I colossi d'Internet come Google e Yahoo sono più indietro con una percentuale di fiducia pari all'68%. Facebook occupa l'ultimo posto con una percentuale di fiducia pari all'56%, forte all'inizio del 2018 dello scandalo Cambridge Analytica, società di consulenza britannica che aveva raccolto i dati personali di 87 milioni di account Facebook senza il loro consenso e li aveva usati per scopi di propaganda politica. Questo avvenimento ha provocato un forte calo del prezzo delle azioni di Facebook, inoltre si è chiesto anche una regolamentazione più rigorosa sull'uso dei dati personali da parte delle aziende tecnologiche.

L'utilizzo dei big data quindi con i suoi vantaggi e problematiche da gestire porta con se cambiamenti radicali nella cultura aziendale/bancaria. Prima dell'arrivo dei big data la domanda che un'organizzazione si poneva è "Cosa pensiamo?", oggi tutte le organizzazioni incentrate sui dati, le cosiddette organizzazioni data driven, si pongono la domanda "Cosa sappiamo?". Le organizzazioni data driven si lasciano guidare dai dati per prendere decisioni e sono più produttive delle organizzazioni che mettono in primo piano le HiPPO (Highest-Paid Person's Opinions), mediamente sono più produttive del 6% e più profittevoli del 5% rispetto ai concorrenti. L'approccio HiPPO è tipico del passato dove tutte le decisioni più

importanti venivano prese dalla persona più pagata, affidandosi esclusivamente sulla propria esperienza lavorativa e sul proprio intuito.

# CAPITOLO 2

# Modelli, metriche e tecniche utilizzate per i problemi di classificazione

## 2.1 I modelli di scelta binaria

Molte delle scelte compiute da individui e imprese sono del tipo " o questo o quello". L'interesse principale è spiegare le ragioni che sono alla base di una particolare scelta e individuare quali sono i fattori che entrano nel processo decisionale, inoltre si è interessati a capire anche in che misura ogni fattore influenza l'esito finale. Per questi motivi, nel corso degli anni sono stati sviluppati i cosiddetti modelli di scelta binaria, in cui vengono utilizzate variabili binarie (indicatrici) che rappresentano le particolari scelte effettuate dal soggetto in esame. La variabile binaria (che è la variabile dipendente) assume valore 1 se viene preferita una delle due alternative e assume valore 0 se viene preferita l'altra. Per rappresentare la scelta di un individuo usiamo la variabile indicatrice nel seguente modo:

$$y = \begin{cases} 1, & \text{Se la caratteristica è presente} \\ 0, & \text{Se la caratteristica è assente} \end{cases} \tag{2.1}$$

L'esito $y$ è ignoto fino al momento dell'estrazione del campione, dunque y è una variabile casuale. La funzione di probabilità di una variabile casuale binaria come questa è data da:

$$f(y) = p^y(1-p)^{1-y} \text{ dove } y \ = \{0, 1\}. \tag{2.2}$$

La probabilità $P(y = 1)$ è uguale a $p$ e rappresenta la probabilità che venga scelta la prima alternativa, di conseguenza la probabilità di $P(y = 0)$ è pari a $1 - p$, che rappresenta la probabilità che venga scelta la seconda alternativa. Questa è una variabile casuale discreta che ha una distribuzione di Bernoulli, e presenta le seguenti caratteristiche:

- $E[y] = p$.
- $Var[y] = p(1-p)$.

L'obiettivo è identificare quali fattori potrebbero influenzare la probabilità $p$ usando una funzione di regressione lineare, che in questo caso è chiamata modello di probabilità lineare:

$$E(Y) = p = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{2.3}$$

Possiamo scomporre il valore osservato di $y$ in due parti, la prima parte è la cosiddetta componente sistematica che è data da $E(y)$, che rappresenta il valore atteso di $y$, la seconda parte invece è composta dall errore casuale non prevedibile, ovvero e. Possiamo quindi considerare $y$ come:

$$y = E(y) + e = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e. \tag{2.4}$$

Una delle problematiche che si incontrano quando si utilizza questo modello per analizzare la scelta del soggetto in esame, e che le classiche ipotesi sul termine d'errore possono essere non valide e dato che la variabile dipendente $y$ può assumere solo due valori, lo stesso deve valere per il termine d'errore, di conseguenza la distribuzione del termine d'errore non può essere più descritta dalla solita curva a "campana" (distribuzione normale). La varianza del termine d'errore $e$ è data da:

$$Var(e) = p(1-p) = (\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)(1 - \beta_1 - \beta_2 x_2 - \cdots - \beta_K x_K). \tag{2.5}$$

L'errore non è omoschedastico, di conseguenza la formula classica dello stimatore dei minimi quadrati non è più valida. Un altro problema del modello a probabilità lineare è che i suoi valori previsti, $E(\hat{Y}) = \hat{p}$, possono cadere al di fuori dell'intervallo $[0, 1]$, di conseguenza i valori previsti non possono essere interpretati in termini di probabilità, in quanto la probabilità è compresa tra 0 e 1, estremi inclusi. Nonostante queste problematiche, vari studi hanno scoperto che quando $p$ non si avvicina troppo a 0 o 1, il modello di probabilità lineare riesce a fornire buone stime degli effetti marginali sulla probabilità di scelta $p$ quando le variabile esplicative $x_k$ subiscono delle variazioni. Per questi motivi, sono stati sviluppati dei modelli di scelta non lineari, come il probit o il logit, che assicurano che le probabilità cadano fra 0 e 1.

Per mantenere la probabilità di scelta $p$ all'interno dell'intervallo $[0, 1]$, si utilizza una relazione non lineare che ha una forma a S, fra $x$ e $p$. La funzione probit è data da:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-0{,}5\, u^2}\, du \tag{2.6}$$

dove:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-0{,}5\, z^2}. \tag{2.7}$$

che rappresenta la funzione di densità di $Z$ che è una variabile casuale che si distribuisce come una normale $N(0, 1)$, con media zero e varianza uno, infatti la funzione probit è collegata alla distribuzione di probabilità della normale standard. La funzione $\Phi(z)$ rappresenta la funzione di ripartizione, grazie alla quale calcoliamo le probabilità associate alla distribuzione normale, in termini geometrici infatti non facciamo altro che calcolare l'area al di sotto della funzione di densità della normale standardizzata a sinistra di $z$.

Con il modello probit, la probabilità $p$ che $y$ assuma valore 1 è data dalla formula:

$$p = P(Z \leq \beta_1 + \beta_2 x) = \Phi(\beta_1 + \beta_2 x). \tag{2.8}$$

Il modello probit è non lineare perché la (2.8) è una funzione non lineare di $\beta_1$ e $\beta_2$. Ovviamente, se $\beta_1$ e $\beta_2$ fossero già noti, potremmo già calcolare le varie probabilità grazie alla (2.8), tuttavia essi non sono noti e di conseguenza vanno stimati. A differenza dello stimatore dei minimi quadrati nel modello di regressione lineare, in questo caso non abbiamo

delle formule esplicite che ci consentono di calcolare i valori di $\tilde{\beta}_1$ e di $\tilde{\beta}_2$. Per ottenere queste stime ci affidiamo a tecniche di analisi numerica e a computer. Supponiamo per ipotesi di selezionare casualmente due individui tali che $y_1 = 1$ e $y_2 = 0$. In statistica la funzione che rappresenta la probabilità di osservare un particolare campione è detta funzione di verosomiglianza, in questo caso essa è pari a:

$$P(y_1 = 1, y_2 = 0) = \Phi[\beta_1 + \beta_2 x_1] \times [1 - \Phi[\beta_1 + \beta_2 x_2]] = L(\beta_1, \beta_2). \tag{2.9}$$

$L(\beta_1, \beta_2)$ è la funzione di verosomiglianza, essa dipende dai parametri ignoti $\beta_1$ e $\beta_2$. Qusta funzione è ottenuta combinando tra loro la (2.2) e la (2.8). Per stimare $\tilde{\beta}_1$ e $\tilde{\beta}_2$ come già detto in precedenza, ricorriamo a tecniche di analisi numerica e alla potenza computazionale dei computer. Di solito, invece di massimizzare la (2.9), si massimizza il logaritmo della funzione di verosomiglianza. In questo caso essa è data da:

$$\begin{aligned} \log\ L\ (\beta_1, \beta_2) &= \log\{\Phi[\beta_1 + \beta_2 x_1] \times [1 - \Phi[\beta_1 + \beta_2 x_2]]\} \\ &= \log \Phi[\beta_1 + \beta_2 x_1] + \log[1 - \Phi[\beta_1 + \beta x_2]]. \end{aligned} \tag{2.10}$$

È più semplice massimizzare la funzione di log verosomiglianza $\log L(\beta_1, \beta_2)$ perché riguarda una somma e non un prodotto di termini. Inoltre il logaritmo è una funzione monotona non decrescente; quindi i massimi delle funzioni (2.9) e (2.10) si verificano in corrispondenza degli stessi valori di $\beta_1$ e di $\beta_2$, dati da $\tilde{\beta}_1$ e da $\tilde{\beta}_2$. Questi valori massimizzano le funzioni (2.9) e (2.10) e sono chiamate stime di massima verosomiglianza.

Come detto in precedenza, uno degli aspetti più importanti da capire, è quello relativo agli effetti marginali per valutare come la probabilità cambi a seconda delle variazioni subite dalle variabili esplicative. Nel modello probit, l'effetto marginale di una variazione unitaria di $x$ sulla probabilità $y = 1$, è dato da:

$$\frac{dp}{dx} = \frac{d\Phi(t)}{dt} \times \frac{dt}{dx} = \phi(\beta_1 + \beta_2 x)\beta_2 \tag{2.11}$$

dove $t = \beta_1 + \beta_2 x$ e $\phi(\beta_1 + \beta_2 x)$ rappresenta la funzione di densità della normale standard valutata in $\beta_1 + \beta_2 x$. La (2.11) rappresenta l'effetto di un aumento di $x$ su $p$. Inoltre l'effetto dipende sia dal valore di $\beta_2$ sia dalla pendenza della funzione probit. Il segno di $\frac{dp}{dx}$ è determinato dal segno di $\beta_2$, in quanto il valore della funzione di densità $\phi(\beta_1 + \beta_2 x)$ è sempre positivo. Quindi se il segno di $\beta_2$ è positivo, $\frac{dp}{dx}$ sarà maggiore di 0. Il valore della funzione $\phi(\beta_1 + \beta_2 x)$ cambia in base al valore che la variabile $x$ assume. Ricordiamo che $\phi(z)$ è massima per $z = 0$ o analogamente per $\beta_1 + \beta_2 x = 0$, quindi in questo caso la probabilità che un individuo scelga la caratteristica 1 è pari a 0,5: infatti $p = \Phi(0) = 0, 5$.

Analogamente, la probabilità che un individuo scelga la caratteristica 0 è pari a 0,5, quindi in questo caso ci troviamo di fronte a una situazione di pari probabilità in cui l'effetto della variazione di $x$ è massimo. Se invece $\beta_1 + \beta_2 x$ è molto elevato, ad esempio pari a 3, allora la probabilità che un individuo scelga la caratteristica 1 è molto elevata e quasi prossima a 1. A differenza del caso precedente, l'effetto della variazione di $x$ avrà un effetto molto modesto. Situazione analoga invece se $\beta_1 + \beta_2 x$ è molto elevato ed è negativo, ad esempio pari a -3. Negli ultimi due casi l'effetto di una piccola variazione di $x$ è trascurabile in quanto l'individuo ha una forte preferenza per la caratteristica scelta, con una probabilità $p$, pari a 1 o 0.

Spesso, invece di utilizzare il modello probit che si basa sulla distribuzione normale e questo complica non di poco la stima del modello, viene utilizzato il modello logit che

differisce dal probit proprio per la particolare curva a S utilizzata per vincolare le probabilità all'interno dell intervallo $[0, 1]$. La funzione di densità della variabile casuale logistica $L$ è la seguente:

$$\lambda(l) = \frac{e^{-l}}{(1+e^{-l})^2} \; con - \infty < l < +\infty \tag{2.12}$$

mentre la funzione di ripartizione è data da:

$$\Lambda(l) = P(L \leq l) = \frac{1}{1+e^{-l}}. \tag{2.13}$$

La funzione di ripartizione associata a questa densità ha un'espressione in forma chiusa a differenza della distribuzione normale, e questo semplifica l'analisi. Con questo modello, la probabilità $p$ che $y$ assuma valore 1 è pari a:

$$p = P(L \leq \beta_1 + \beta_2 x) = \Lambda(\beta_1 + \beta_2 x) = \frac{1}{1+e^{-(\beta_1+\beta_2 x)}}. \tag{2.14}$$

Possiamo riscrivere questa espressione anche nel seguente modo:

$$p = \frac{1}{1+e^{-(\beta_1+\beta_2 x)}} = \frac{e^{(\beta_1+\beta_2 x)}}{1+e^{(\beta_1+\beta_2 x)}} \tag{2.15}$$

mentre la probabilità che y assuma valore 0 è data da:

$$1 - p = \frac{1}{1+e^{(\beta_1+\beta_2 x)}}. \tag{2.16}$$

La forma della funzione di densità logistica è diversa dalla funzione di densità di una normale, per questo motivo le stime di massima verosomiglianza di $\beta_1$ e di $\beta_2$ saranno differenti. Anche in questo caso, supponiamo per ipotesi di selezionare casualmente due individui tali che $y_1 = 1$ e $y_2 = 0$. In questo caso, la funzione di verosomiglianza è pari a:

$$P(y_1 = 1, y_2 = 0) = \Lambda[\beta_1 + \beta_2 x_1] \times [1 - \Lambda[\beta_1 + \beta_2 x_2]] = L(\beta_1, \beta_2). \tag{2.17}$$

Anche in questo caso, possiamo esaminare l'effetto marginale di una variazione unitaria di $x$ sulla probabilità $y = 1$, esso è dato da:

$$\frac{dp}{dx} = \frac{d\Lambda(t)}{dt} \times \frac{dt}{dx} = \lambda(\beta_1 + \beta_2 x)\beta_2. \tag{2.18}$$

A differenza del modello probit, non utilizziamo la funzione di densità normale, ma utilizziamo la funzione di densità della variabile casuale logistica come visto in (2.12).

## 2.2 La regressione penalizzata logistica

Il compito di determinare quali predittori sono associati a una data risposta non è affatto semplice. Quando selezioniamo ad esempio le variabili per un modello lineare, generalmente si considerano i singoli valori $p$-values. Questa procedura tuttavia può essere ingannevole. Ad esempio se le variabili sono altamente correlate, i valori $p$-values saranno elevati, spingendo il ricercatore a dedurre erroneamente che quelle variabili non sono predittori importanti. D'altra parte, variabili irrilevanti possono essere incluse nel modello, aggiungendo

così una complessità non necessaria, con la conseguente perdita d'interpretabilità del fenomeno oggetto dello studio.

Inoltre, se il numero di osservazioni $n$ non è di molto superiore al numero delle variabili, può esserci molta variabilità, con conseguente overfitting, e con previsioni più scadenti sulle osservazioni future non utilizzate nell'addestramento del modello.

Ci sono alcuni approcci per eseguire automaticamente la selezione delle variabili. Un approccio utilizzato è quello dello *shrinkage* o anche *regularization*, che prevede l'adattamento di un modello con tutti i predittori, ma dove i coefficienti stimati sono ridotti a zero rispetto alle stime classiche. Di conseguenza la varianza del modello costruito si riduce e la stima di alcuni dei coefficienti sarà pari a 0, in modo tale da individuare quali variabili sono irrilevanti per il fenomeno oggetto dell'analisi.

In generale, quando la relazione tra il logit in una risposta dicotomica e i predittori è quasi lineare, le stime di massima verosimiglianza avranno una bassa distorsione, ma potrebbero avere una varianza elevata come quando il numero di covariate è grande rispetto al numero di osservazioni o i predittori sono altamente correlati. In questo caso una piccola modifica dei dati di allenamento può causare un grande cambiamento nelle stime dei coefficienti.

Gli approcci *Ridge* e *Lasso*, due dei principali metodi di regressione penalizzata, gestiscono il *trade-off* tra *bias* e *variance*, scambiando un piccolo aumento in *bias* (distorsione) con una grande diminuzione della varianza delle previsioni, quindi questo approccio potrebbe migliorare l'accuratezza complessiva della previsione.

La regressione logistica *Ridge* è ottenuta minimizzando la funzione di verosimiglianza con un parametro penalizzato applicato a tutti i coefficienti tranne l'intercetta. La funzione di verosomiglianza del modello logit è data da:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} [y_i \ \log(\pi_i) + (1 - y_i) \ \log(1 - \pi_i)] \\ &= \sum_{i=1}^{n} \ \Big[ y_i \ \log\Big(\frac{\pi_i}{1 - \pi_i}\Big) \ + \log(1 - \pi_i) \Big] \\ &= \sum_{i=1}^{n} [y_i \ x_i\beta - \log(1 + e^{x_i\beta})]. \end{aligned} \tag{2.19}$$

Di conseguenza, nella regressione logistica *Ridge* si dovrà minimizzare la seguente funzione:

$$l_\lambda^R = \sum_{i=1}^{n} [y_i \ x_i\beta - \log(1 + e^{x_i\beta})] + \lambda \times \sum_{j=1}^{p} \beta_j^2. \tag{2.20}$$

Alla funzione di verosomiglianza del modello logit aggiungiamo un termine di penalizzazione $(L_2)$, in cui si ha la sommatoria dei beta al quadrato.

Lo stimatore *Ridge* riduce i coefficienti di regressione, in modo che le variabili, con un contributo minore al risultato, abbiano i loro coefficienti vicini allo zero. La *Ridge Regression* però ha il limite di includere nella selezione del modello tutte le variabili indipendenti e questo non permette di effettuare la selezioni delle variabili, di conseguenza se l'obiettivo è quello d'interpretare i coefficienti stimati, applicare una penalizzazione di questo tipo non è certamente consigliato. La stima dei parametri $\beta_p$ dipende dalla scelta del parametro $\lambda \geq 0$. All'aumentare del parametro $\lambda$, i coefficienti $\beta_p$ tendono a zero, anche se non saranno mai pari a 0.

Nella regressione logistica Lasso (Least Absolute Shrinkage and Selection Operator) invece, viene utilizzata una differente penalizzazione $(L_1)$ per la stima dei parametri $\beta_p$. In questo caso si dovrà minimizzare la seguente funzione:

$$l_\lambda^L = \sum_{i=1}^{n} [y_i \; x_i\beta - log(1 + e^{x_i\beta})] + \lambda \times \sum_{j=1}^{p} |\beta_j|. \tag{2.21}$$

Nella *Lasso Regression* la penalizzazione è uguale alla somma dei valori assoluti dei coefficienti, in questo caso, avviene sia una contrazione del valore dei parametri (per evitare l'*overfitting*) sia una selezione delle variabili, perché la penalizzazione azzera i coefficienti delle variabili collineari. Ciò comporta che il Lasso può anche essere visto come un'alternativa ai metodi di *features selection* per eseguire la selezione delle variabili al fine di ridurre la complessità del modello. Dal punto di vista dell'accuratezza previsiva, non esiste un metodo che domina sull'altro, in generale il *lasso* funziona meglio in un ambiente in cui abbiamo un numero piccolo di predittori con coefficienti sostanziali e i restanti predittori hanno coefficienti che sono molto piccoli o uguali a zero. La regressione *ridge* invece funziona meglio quando la variabile di risposta è funzione di molti predittori, tutti con coefficienti di dimensioni più o meno uguali. Tuttavia, il numero di predittori correlati con la variabile di risposta non è mai noto a priori per i set di dati reali. È possibile quindi utilizzare una tecnica come la convalida incrociata al fine di determinare quale approccio è migliore su un determinato set di dati.

L'utilizzo della *Lasso Regression* quindi risulta conveniente quando i modelli contengono molte variabili e tra queste poche hanno una significatività statistica per la predizione del valore della Y, mentre la *Ridge Regression* risulta profittevole quando tutte le variabili nel modello hanno una certa significatività. Quando un modello include molte variabili, può risultare difficile analizzare ogni singola variabile e trarre conclusioni generali su quale modello sia preferibile implementare.

Una soluzione intermedia alle due precedentemente analizzate è rappresentata dalla regressione *Elastic Net* la quale combina le proprietà delle regressioni *Ridge* e *Lasso* penalizzando il modello sia tramite la norma $L_1$ che la norma $L_2$, in questo caso si dovrà minimizzare la seguente funzione:

$$l_{(\lambda_L,\lambda_R)}^{EN} = \sum_{i=1}^{n} [y_i \; x_i\beta - \log(1 + e^{x_i\beta})] + \lambda_L \times \sum_{j=1}^{p} |\beta j| + \lambda_R \times \sum_{j=1}^{p} \beta_j^2 \tag{2.22}$$

o anche, con un altro tipo di parametrizzazione:

$$l_\lambda^{EN} = \sum_{i=1}^{n} [y_i \; x_i\beta - \log(1 + e^{x_i\beta})] + \lambda \times \Big(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2\Big). \tag{2.23}$$

Le penalizzazioni $L_1$ e $L_2$ assumono diversi valori di $\lambda$. Con la procedura di *cross-validation* è possibile stimare i valori ottimali di $\lambda_L$ e di $\lambda_R$. In particolare, con $\lambda_L = \lambda_R = 0$, non verrà introdotta alcuna penalizzazione, dunque si otterrà la classica regressione logistica. Con $\lambda_L = 0, \lambda_R > 0$ si otterrà una *Ridge Logistic Regression*, mentre con $\lambda_L > 0, \lambda_R = 0$ si otterrà una *Lasso Logistic Regression*. Infine, con $\lambda_L > 0, \lambda_R > 0$ si otterrà un modello ibrido che combina sia i vantaggi della penalizzazione *ridge*, sia i vantaggi della penalizzazione *lasso*, infatti con questo modello siamo in grado di lavorare bene in situazioni di correlazioni tra variabili. Ciò è dovuto al fatto che la penalizzazione *lasso*, mantiene un parametro tra le variabili correlate ed elimina le altre (la selezione di questo

parametro è del tutto casuale, per questo motivo è stata sviluppata una versione del lasso denominata *grouped lasso* che risolve questa problematica), mentre la penalizzazione *ridge* fa in modo che tutti i parametri correlati tendano a zero, in questo modo combinando $L_1$ e $L_2$, l'*Elastic Net Regression* mantiene solo alcune variabili fra quelle correlate e penalizza i loro coefficienti in modo che assumano valori prossimi a zero.

## 2.3 Decision Trees

Esistono molte metodologie per costruire alberi decisionali, ma il più noto è l'algoritmo dell'*albero di classificazione e regressione* (CART). Un albero decisionale di base suddivide i dati di addestramento in sottogruppi omogenei (ovvero gruppi con valori di risposta simili) e quindi adatta una semplice costante in ciascun sottogruppo (ad esempio per la regressione, la media dei valori di risposta all'interno del gruppo). I sottogruppi (chiamati anche nodi) sono formati ricorsivamente utilizzando delle partizioni binarie costruite ponendo semplici domande su ciascuna caratteristica (l'età è $< 18$?, il saldo è $> 1500$?). Questa operazione viene eseguita più volte fino a quando non viene soddisfatto un criterio di arresto adeguato (ad esempio viene raggiunta la profondità massima dell'albero). Dopo che tutto il partizionamento è stato eseguito, il modello prevede l'output in base ai valori di risposta medi per tutte le osservazioni che rientrano in quel sottogruppo (in caso di regressione) o la classe che ha la rappresentazione maggioritaria (in caso di classificazione).

Questi metodi *divide et impera* possono produrre regole semplici che sono facili da interpretare e da visualizzare con diagrammi ad albero. L'albero decisionale è composto da un primo sottogruppo situato nella parte superiore dell'albero ed è noto come nodo radice (*root*) e contiene tutti i dati di addestramento, mentre i sottogruppi nella parte inferiore dell'albero sono chiamati nodi terminali o foglie. Ogni sottogruppo intermedio viene definito nodo interno mentre le connessioni tra i nodi sono chiamate rami. Il CART utilizza il partizionamento ricorsivo binario (ricorsivo perché ogni divisione o regola dipende dalle divisioni sopra di essa) e l'obiettivo di ciascun nodo è trovare la caratteristica "migliore" ($x_i$) per partizionare i dati rimanenti in una delle due regioni (R1 e R2) in modo tale che l'errore complessivo tra la risposta effettiva ($y_i$) e la costante prevista ($c_i$) sia ridotto al minimo.

Per i problemi di regressione, la funzione obiettivo da minimizzare è l'SSE (*sum-of-squared- errors*):

$$SSE = \sum_{i \in R1} (y_i - c_1)^2 + \sum_{i \in R2} (y_i - c_2)^2. \tag{2.24}$$

Nei problemi di classificazione, dato che il modello prevede l'output in base alla classe che ha la rappresentazione maggioritaria in quella regione, viene utilizzato e si cerca di minimizzare il tasso di errore di classificazione, dato dalla formula:

$$\text{classification error rate} : E = 1 - max_k(\hat{p}_{mk}) \tag{2.25}$$

dove $\hat{p}_{mk}$ rappresenta la proporzione di osservazioni di addestramento nella regione $m$-esima che provengono dalla classe $k$-esima. Il *classification error rate* però, è una metrica non sufficientemente sensibile per gestire la "crescita" degli alberi, essa è preferibile se l'obiettivo è l'accuratezza della previsione finale. Per far "crescere" l'albero, si dividono i nodi ricercando la "variabilità entro nodo" minima, che generalmente viene misurata con due indici alterativi, essi sono l'indice di Gini e la *cross-entropia*.

L'indice di Gini è dato da:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.26}$$

mentre la *cross-entropia* è data da:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk}\ log\ \hat{p}_{mk}. \tag{2.27}$$

In entrambi i casi un valore piccolo rappresenta una misura di "purezza" (*purity*) del nodo, questo vuol dire che il nodo è composto in maniera predominante da osservazioni di una sola classe. La *decision boundary* di un albero di decisione è rappresentata da regioni rettangolari che racchiudono le osservazioni.

Sia negli alberi di regressione che in quelli di classificazione, l'obiettivo del partizionamento è ridurre al minimo la dissimilarità (diversità) nei nodi terminali. La costruzione di un albero generalmente segue un processo composto da due fasi: crescita (*growth*) e potatura (*pruning*).

Per la fase di crescita vale quanto detto in precedenza, si implementa un partizionamento ricorsivo binario con il quale i nodi vengono divisi e questo permette all'albero di crescere, si ricerca la "variabilità entro nodo" minima espressa attraverso l'indice di Gini o la *cross-entropia* nel caso di problemi di classificazione, mentre nei problemi di regressione si cerca di minimizzare la funzione obiettivo SSE.

Nella fase di potatura invece l'obiettivo è individuare un *subtree* ottimale, questo permette di evitare l'overfitting. Il sottoalbero ottimale nei problemi di regressione viene trovato minimizzando la funzione obiettivo SSE a cui viene aggiunto un termine di penalizzazione dato dal prodotto $\alpha \times |T|$, dove $\alpha$ rappresenta il parametro di complessità che controlla la dimensione dell'albero, mentre $T$ rappresenta il numero di split/nodi terminali nell'albero. Per un determinato valore di $\alpha$ troviamo il più piccolo albero potato che ha il più basso errore penalizzato.

Nei problemi di classificazione viene preso in considerazione il tasso di errata classificazione. Per migliorare l'accuratezza predittiva, si possono combinare più alberi per produrre una previsione aggregata. "Bagging", "random forests" e "boosting" sono alcuni approcci che implementano una tale strategia. Il prezzo da pagare per l'accresciuta accuratezza è una perdita in interpretabilità.

Come con i metodi di regolarizzazione, le penalità più piccole tendono a produrre modelli più complessi che si traducono in alberi più grandi, mentre penalità più grandi si traducono in alberi molto più piccoli. A maggiore profondità dell'albero corrisponde maggiore complessità del modello. In generale più un modello è complesso, più è facile mandarlo in *overfitting*, quindi è fondamentale durante la fase di addestramento implementare alberi di decisione non troppo profondi. Il parametro di complessità $\alpha$ tipicamente è stimato attraverso procedure di *cross-validation*.

## 2.4 Metriche per la valutazione della performance

Valutare le prestazioni di un classificatore attraverso metriche di performance appropriate è di fondamentale importanza. Quando abbiamo a che fare con algoritmi di classificazione è opportuno conoscere la nozione di matrice di confusione.

Una matrice di confusione ha una forma tabellare come si può notare dalla Figura 2.1.

| | | CLASSI PREVISTE | | |
|---|---|---|---|---|
| | | Negativo (-) | Positivo (+) | Somma |
| CLASSI VERE | Negativo (-) | ***TN*** True (-) | ***FP*** False (+) | ***N*** |
| | Positivo (+) | ***FN*** False (-) | ***TP*** True (+) | ***P*** |
| | Somma | ***N**** | ***P**** | |

Figura 2.1: **Struttura della matrice di confusione**

In questo caso, stiamo cercando di classificare, ovvero di mettere nella giusta "scatola" un'entità, che corrisponde a una determinata osservazione del data set in esame. La matrice di confusione conterrà sia i dati attuali, ovvero l'appartenenza di un'entità alla sua classe d'origine, sia i dati predetti, ovvero il modo in cui l'algoritmo ha classificato quell'entità. I valori predetti si trovano lungo le colonne, mentre quelli attuali si trovano sulle righe. Nella diagonale principale della matrice di confusione abbiamo le previsioni corrette, in questo caso il classificatore ha predetto correttamente la classe d'origine, viceversa all'esterno della diagonale principale abbiamo gli errori commessi dal classificatore. In una matrice di confusione abbiamo:

- *True positive (TP)*: È il caso di una classificazione corretta. Ad esempio quando si diagnostica una malattia a un paziente che è realmente affetto da quella malattia.
- *True negative (TN)*: È il caso di un classificazione corretta. Ad esempio quando nella diagnosi per un paziente, una certa malattia viene esclusa e realmente il paziente non soffre di quella malattia.
- *False positive (FP)*: È il caso di una classificazione non corretta. Ad esempio quando si diagnostica a un paziente una certa malattia, ma il paziente non è realmente affetto da quella malattia. Il falso positivo viene definito come errore di tipo 1.
- *False negative (FN)*: È il caso di una classificazione non corretta. Ad esempio quando non si diagnostica a un paziente una certa malattia, ma il paziente è realmente affetto da quella malattia. Il falso negativo viene definito come errore di tipo 2.

Dalla matrice di confusione si possono ottenere facilmente diverse metriche:

*Misclassification rate*: Il tasso di errore misura la percentuale di errore delle previsioni sul totale delle istanze. Varia da 0 (peggiore) a 1 (migliore). È dato dalla formula:

$$ERR = \frac{FP + FN}{TP + TN + FP + FN} \tag{2.28}$$

*Accuracy*: L'accuratezza misura la percentuale delle previsioni esatte sul totale delle istanze. È l'inverso del tasso di errore. Varia da 0 (peggiore) a 1 (migliore). È data dalla formula:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = 1 - Misclassification\ rate \tag{2.29}$$

*Sensitivity*(o anche *recall*): Rappresenta la frazione di (veri) positivi correttamente classificati ed è data dalla formula:

$$TPR = \frac{TP}{P} \tag{2.30}$$

Nel test delle ipotesi coincide con la nozione di "potenza del test" vista come 1 - Pr {Errore di II tipo}.

*Specificity*: È la frazione di (veri) negativi correttamente classificati ed è data dalla formula:

$$TNR = \frac{TN}{N} \tag{2.31}$$

*Precision*: Viene vista come una misura di esattezza/fedeltà del classificatore e rappresenta la frazione dei classificati positivi che risultano essere veramente positivi. È data dalla formula:

$$Precision = \frac{TP}{P^*} \tag{2.32}$$

dove $P^*$ rappresenta il totale delle osservazioni classificate come positive, mentre P rappresenta il totale delle osservazioni positive che si hanno nel data set di riferimento.

*False Positive Rate*: Il tasso dei falsi positivi è la percentuale delle previsioni positive errate (FP) sul totale delle istanze negative. Varia da 0 (migliore) a 1 (peggiore). È dato dalla formula:

$$FPR = \frac{FP}{TN + FP} \tag{2.33}$$

*F-score*: È una "media armonica" di Precision e Recall ed è data dalla formula:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.34}$$

Il valore più alto possibile di un punteggio F è 1.0, che indica una precision e una recall perfetta, mentre il valore più basso possibile è 0. L'F-score ignora completamente i veri negativi, quindi è fortemente discutibile in casi di forte sbilanciamento nelle classi.

*Coefficiente di Correlazione di Matthew(MCC)*: Rappresenta il coefficiente di correlazione tra classi osservate e previste ed è dato dalla formula:

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{2.35}$$

Ha valori sull'intervallo [-1,+1]. Un MCC $\rightarrow$ +1 indica una previsione perfetta mentre un MCC $\rightarrow$ -1 indica un totale disaccordo tra le classi previste e quelle osservate. Un MCC $\approx$ 0 indica una classificazione non migliore del "random guess". Il coefficiente di correlazione di matthew è utile quando si hanno classi fortemente sbilanciate, inoltre a differenza dell'accuratezza e dell'F-score, tiene conto di tutti e quattro i casi mostrati nella matrice di confusione.

*Receiver Operating Characteristic (ROC)*: La sensibilità e la specificità classificano gli utenti sulla base di un predefinito valore soglia, la curva ROC invece viene costruita considerando tutti i possibili valori soglia e per ognuno di questi si calcola la sensibilità e la proporzione di falsi positivi data dalla formula 1 – specificità. Congiungendo i vari punti che evidenziano la proporzione di veri positivi e di falsi positivi (coordinate) si ottiene una curva chiamata curva ROC come in Figura 2.2.

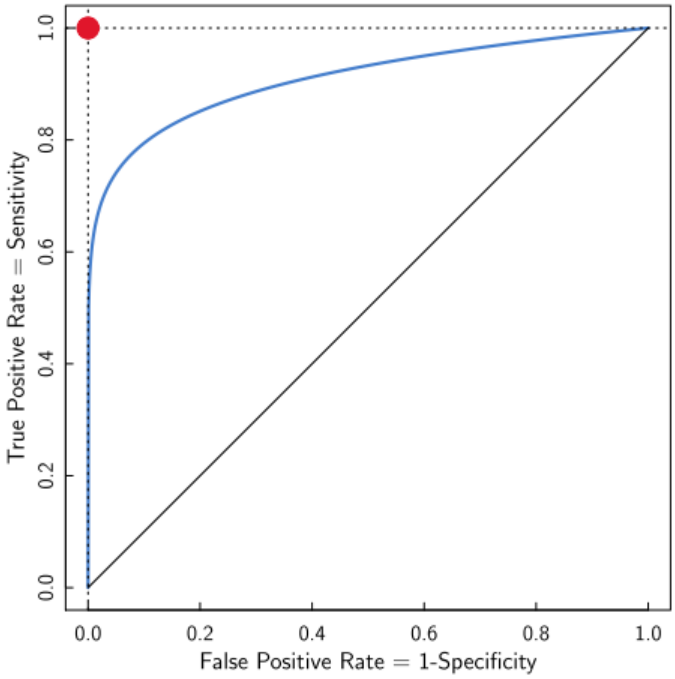


Figura 2.2: **Receiver Operating Characteristic curve (ROC)**

L'area sotto la curva ROC è detta AUC, ovvero *Area Under the Curve*. Più l'area sotto la curva è grande (quando la curva si avvicina al vertice del grafico in alto a sinistra) tanto maggiore è il potere discriminante. Lungo la retta a 45 gradi ci sono i classificatori equivalenti al *random guessing*. Sopra la retta a 45 gradi ci sono i classificatori migliori del

*random guessing*, sotto quelli peggiori. Il classificatore perfetto si collocherebbe nel punto rosso della Figura 2.2. Per l'interpretazione dei valori dell'area sottostante la curva ROC è possibile riferirsi alla classificazione proposta da Swets (1998):

- AUC = 0.5: Il classificatore non è informativo
- 0.5 < AUC $\leq$ 0.7: Il classificatore è poco accurato
- 0.7 < AUC $\leq$ 0.9: Il classificatore è moderatamente accurato
- 0.9 < AUC < 1: Il classificatore è altamente accurato
- AUC = 1: Il classificatore è perfetto

*Kappa di Cohen*: É un indice statistico che permette di valutare il grado di accordo tra due valutazioni qualitative effettuate sulle stesse unità statistiche ed è dato dalla formula:

$$k = \frac{Proporzione\ osservata - Proporzione\ dovuta\ al\ caso}{1 - Proporzione\ dovuta\ al\ caso} \tag{2.36}$$

Tiene conto della probabilità di concordanza casuale, dunque può essere visto come un indice che indica il grado di accuratezza e di affidabilità di un classificatore. Rispetto all'accuratezza tiene conto dello sbilanciamento esistente tra le classi. Il più alto valore che può assumere è pari a 1 e rappresenta il grado di concordanza ottimale, se invece $k$ assume valori inferiori a 0 la concordanza non esiste.

## 2.5 K-fold cross-validation

La $K$-fold cross-validation è un modo popolare (ed efficace) per selezionare i parametri di ottimizzazione negli algoritmi di apprendimento statistico. Di frequente utilizzo sono la five-fold cross-validation (con $k = 5$) e la ten-fold cross-validation (con $k = 10$). Nella $k$-fold cross-validation dividiamo il training set in $k$ fold di grandezza pressoché identica.

Nel primo step, il primo fold assumerà il ruolo di test set, mentre i restanti $k - 1$ fold svolgeranno il ruolo di training set. Il modello verrà addestrato sui $k - 1$ fold e successivamente verrà testato sui dati *unseen*, che in questo caso sono rappresentati dal primo fold.

Generalmente per i problemi di regressione si utilizza l'MSE (*Mean Squared Error*) come metrica di performance, mentre per i problemi di classificazione generalmente si fa riferimento al tasso di accurata classificazione. La logica per gli step successivi è la medesima. Nell'ultimo step, l'ultimo fold assumerà il ruolo di test set mentre i restanti $k - 1$ fold svolgeranno il ruolo di training set.

Alla fine del processo avremo $k$ metriche di performance ognuna relativa ai $k$ step effettuati. Le $k$ metriche di performance verranno aggregate per ottenere un'unica metrica finale. Nei vari step della $k$-fold cross-validation le osservazioni a turno faranno parte sia dei dati di addestramento sia dei dati unseen. La grandezza del train/test set è determinata da $k$.

In ogni fold ci sarà una frazione di $\frac{n}{k}$ dati, quindi $1 - \frac{n}{k}$ è la frazione di dati assegnata al training set mentre $\frac{n}{k}$ è la frazione di dati assegnata al test set. In generale i valori $k = 5$ e $k = 10$ rappresentano il giusto compromesso. Con $k = 5$ il training set contiene approssimativamente l'80% dei dati, mentre il test set contiene il rimanente 20%. Con $k = 10$ il training set contiene approssimativamente il 90% dei dati mentre il test set contiene il rimanente 10%.

In molti esperimenti si è dimostrato empiricamente come questi due valori rappresentino il giusto compromesso, tuttavia non esistono garanzie teoriche. In generale, quando $k$ è grande, i training sets avranno una dimensione maggiore e questo migliora la stima di $\hat{f}(\cdot)$, mentre le dimensioni dei test sets si riducono e questo riduce la qualità di stima di $\overline{L}_s$. L'overlap tra i training sets aumenta, e questo fa sì che la correlazione tra le $\overline{L}_s$ aumenti. Quando $k$ è piccolo accade esattamente l'opposto. $\{\overline{L}_1, \overline{L}_2, \overline{L}_s\}$ rappresentano la stima dell'errore di previsione sul test set per ogni split/fold. $\hat{f}(\cdot)$ è uno stimatore per $f(\cdot)$. $\hat{f}$ è costruito a partire da un training set $T$, quindi la casualità di $\hat{f}$ riflette l'informazione contenuta in $T$.

## 2.6 Oversampling e Undersampling

Quando si ha a che fare con dei problemi di classificazione può capitare d'imbattersi nel cosiddetto "sbilanciamento" delle classi e questo rappresenta un problema in quanto molto spesso capita che la classe di maggior interesse sia allo stesso tempo anche la classe minoritaria, ovvero la classe che ha la percentuale di osservazioni minore e questo può certamente influire sul processo di addestramento dei modelli implementati.

Tecniche per risolvere il problema dello sbilanciamento tra classi sono l' *undersampling* e l'*oversampling*.

Nell'*undersampling* (sotto-campionamento) si ridimensiona una classe prelevando da una popolazione un suo sottoinsieme. Tecniche di *undersampling* sono:

- *Undersampling randomico*: In questo caso si prendono casualmente degli elementi dalla popolazione target. In questo approccio non si hanno assunzioni teoriche di base e non rappresenta certo un metodo efficace e sicuro in quanto i dati estratti casualmente potrebbero non essere più rappresentativi della popolazione iniziale.

- *Undersampling stratificato*: In questo caso invece, a differenza del precedente, si decidono una o più variabili per cui stratificare il campionamento, così da creare un campione rappresentativo della popolazione iniziale. Questa tecnica permette di creare quindi dei campioni più fedeli della popolazione iniziale, ma allo stesso tempo il campionamento è molto condizionato dalla scelta della variabile per cui stratificare. Immaginiamo di avere nella popolazione iniziale $60\%$ donne e $40\%$ uomini, facendo un *undersampling* stratificato per la variabile sesso, creerò un sottoinsieme che avrà una distribuzione della variabile sesso analoga alla popolazione iniziale.

L'*undersampling* quindi agisce sulla classe più numerosa e cerca di ridimensionarla, questo è ben visibile nella Figura 2.3.

Nell'*oversampling* (sovra-campionamento) invece l'attenzione si sposta sulla classe meno rappresentata. A differenza dell' *undersampling*, l' *oversampling* crea dei nuovi dati (relativi alla classe minoritaria), i cosiddetti dati sintetici, tuttavia è bene ricordare di non eccedere con la "sintetizzazione" di nuovi dati perché si rischia d'inserire nel modello dei dati fittizi, cosi facendo rischiamo di allontanarci dalla realtà osservata. Lo scopo del sovracampionamento è quello di avere un modello di previsione migliore, questa tecnica dunque non è stata creata per scopi di analisi poiché ogni dato creato è sintetico, quindi è un promemoria.

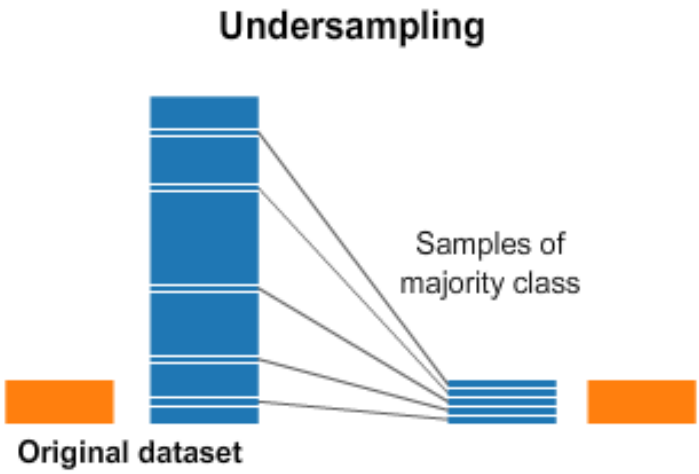


Figura 2.3: **Undersampling**

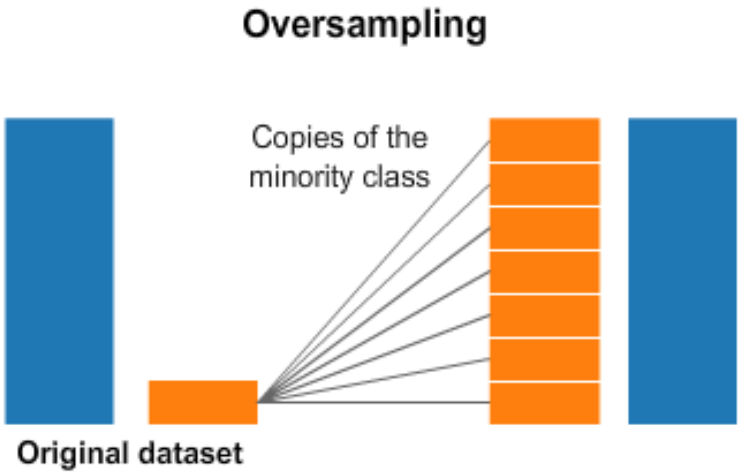


Figura 2.4: **Oversampling**

Nelle tecniche di sovracampionamento classiche, come si può vedere nella Figura 2.4, i dati appartenenti alla classe minoritaria vengono duplicati dalla popolazione dei dati di minoranza, tuttavia questo approccio non fornisce nuove informazioni.

La tecnica SMOTE (*Synthetic Minority Oversampling*) è una tecnica di *oversampling*, che funziona in modo diverso dal tipico *oversampling*, in quanto utilizza algoritmi di sovracampionamento più efficaci. In questo caso si vanno a prendere le osservazioni più vicine (con la distanza euclidea) tra quelle appartenenti alla classe minoritaria, si effettua la differenza tra i due vettori di features e si moltiplica questo valore per un numero casuale tra 0 e 1. In altre parole, si va ad applicare un perturbamento alla distanza tra due punti della classe minoritaria. Così facendo si creano osservazioni artificiali che accrescono il patrimonio di dati ma non ne modificano troppo il valore. La procedura viene ripetuta abbastanza volte fino a quando la classe di minoranza ha la stessa proporzione della classe di maggioranza. La Figura 2.5, illustra il funzionamento dell'algoritmo SMOTE.

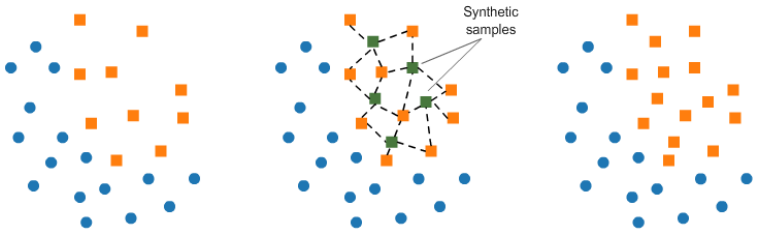


Figura 2.5: **SMOTE**

Lo SMOTE classico tuttavia funziona solo per le variabili continue. Trasformare i dati categoriali in numerici per applicare lo smote, porta alla creazione di dati che non hanno alcun senso, ad esempio per le variabili binarie che assumono valori 0 e 1 un potenziale sovracampionamento creerebbe dati come 0.20 o 0.75 e questo non ha alcun senso. Nel caso in cui nel data set di riferimento sono presenti variabili miste, si può pensare di applicare una variante dello smote, ovvero lo SMOTE-NC (Nominal and Continuous). In questo caso, indichiamo quali caratteristiche sono di tipo categoriale, così facendo smote ricampionerà i dati categoriali invece di creare dati sintetici.

Un'ulteriore tecnica di sovracampionamento è il campionamento sintetico adattivo (ADASYN), che crea dati sintetici in base alla densità dei dati. In questo approccio la generazione di dati sintetici sarebbe inversamente proporzionale alla densità della classe di minoranza. Significa che vengono creati più dati sintetici nelle regioni dello spazio delle caratteristiche in cui la densità degli esempi di minoranza è bassa, e meno o nessuno dove la densità è alta e questo è ben visibile nella Figura 2.6.

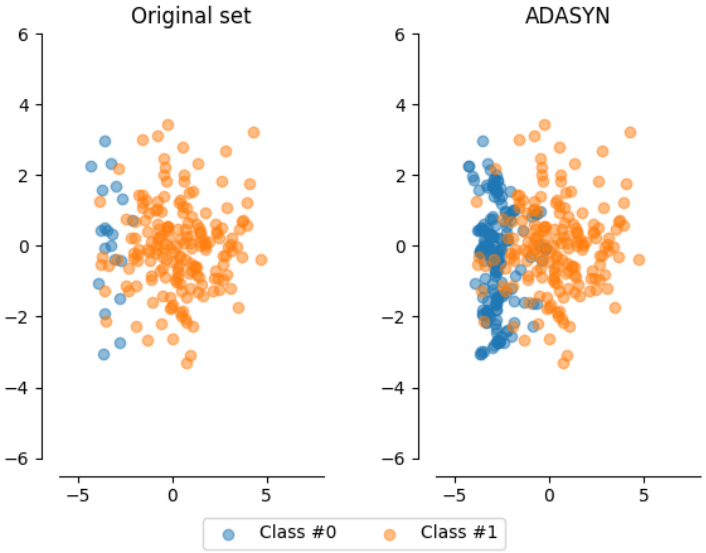


Figura 2.6: **ADASYN**

ADASYN si concentrerà sui dati di densità dove la densità è bassa. Spesso, i dati a bassa densità sono un valore anomalo. L'approccio ADASYN porrebbe quindi troppa attenzione su queste aree dello spazio delle caratteristiche, il che potrebbe comportare prestazioni peggiori del modello. Potrebbe essere meglio rimuovere il valore anomalo prima di utilizzare ADASYN.

# CAPITOLO 3

# Il caso studio: Bank Marketing data set

## 3.1 Descrizione Data set

Il data set in esame rappresenta il risultato di una campagna marketing effettuata da un istituto bancario portoghese e riporta il dettaglio degli utenti che hanno sottoscritto o meno un deposito a termine. Il dataset è composto da 45.211 osservazioni e da 17 variabili che possono essere suddivise in due grandi blocchi. Il primo blocco viene riportato nella Tabella 3.1 e racchiude tutte quelle variabili utili per estrarre informazioni sui clienti.

Tabella 3.1: **Informazioni sul cliente**

| Nome Variabile | Tipo | Descrizione |
|---|---|---|
| age | Numerica | Indica l'età dei clienti |
| job | Categoriale a 12 livelli | Indica il tipo di lavoro |
| marital | Categoriale a 4 livelli | Indica lo stato civile |
| education | Categoriale a 4 livelli | Indica il livello di educazione |
| default | Categoriale a 3 livelli | Indica se il credito è in default |
| balance | Numerica | Indica il saldo bancario espresso in euro |
| housing | Categoriale a 3 livelli | Indica l'eventuale presenza di un mutuo |
| loan | Categoriale a 3 livelli | Indica l'eventuale presenza di un prestito personale |
| poutcome | Categoriale a 4 livelli | Indica il risultato della precedente campagna di marketing |
| y | Categoriale a 2 livelli | Indica l'eventuale sottoscrizione di un deposito a termine(Variabile target) |

Il secondo blocco, riportato nella Tabella 3.2, racchiude tutte quelle variabili utili per estrarre informazioni sul comportamento della banca durante le varie campagne di marketing, sia quelle passate sia l'attuale, ad esempio il mese dell'ultimo contatto per capire se esistono dei mesi durante l'anno in cui la banca è più propensa ad avviare campagne, o anche il numero totale dei contatti avuti con i clienti, per analizzare quante volte la banca contatta un cliente, e cosa spinge la banca a contattare più volte un cliente rispetto ad'altri.

Tabella 3.2: **Informazioni sul comportamento della banca**

| Nome Variabile | Tipo | Descrizione |
|---|---|---|
| contact | Categoriale a 3 livelli | Indica la tipologia di contatto |
| day | Numerica | Indica il giorno dell'ultimo contatto |
| month | Categoriale a 12 livelli | Indica il mese dell'ultimo contatto |
| duration | Numerica | Indica la durata dell'ultimo contatto espressa in secondi |
| campaign | Numerica | Indica il numero di contatti avuti con il cliente durante la campagna marketing attuale |
| pdays | Numerica | Indica il numero di giorni trascorsi da quando il cliente era stato contattato per una precedente campagna marketing |
| previous | Numerica | Indica il numero di contatti avuti con il cliente prima dell'attuale campagna marketing |

Nei paragrafi successivi, con alcuni grafici, analizzeremo le principali caratteristiche degli utenti, e come queste caratteristiche impattano sull'eventuale sottoscrizione o meno del deposito a termine proposta dall'istituto finanziario.

In particolare, nel paragrafo 3.2, utilizzeremo le variabili presenti nella Tabella 3.1, in quanto sono utili per capire il target di riferimento a cui l'istituto finanziario si rivolge, mentre nel paragrafo 3.3, utilizzeremo sia le variabili presenti nella Tabella 3.1, sia le variabili presenti nella Tabella 3.2, che ricordiamo ci forniscono informazioni utili su come la banca si comporta nei confronti degli utenti. Ciò al fine di analizzare le variabili d'interesse in relazione alla variabile target, che rappresenta l'eventuale sottoscrizione o meno del deposito a termine proposto.

## 3.2 Analisi delle caratteristiche degli utenti

Utilizzando le variabili presenti nella Tabella 3.1, si vuole innanzitutto studiare e descrivere le principali caratteristiche degli utenti contattati dalla banca. L'età sicuramente è un fattore fondamentale in quanto potrebbe incidere in modo significativo sulla sottoscrizione di un deposito a termine. Raggruppare l'età secondo la classificazione temporale degli stadi del ciclo vitale umano come riportato nella Figura 3.1, può essere utile perché ci permette di capire se esiste un target preciso a cui la banca fa riferimento.

Dalla Figura 3.1 si nota facilmente come la banca si sia principalmente rivolta a un'utenza giovane/medio-giovane, di età compresa tra i diciotto e i cinquantanove anni. In particolar modo, la fascia maggiormente contattata è la prima, mentre pochi utenti appartenenti alla terza e alla quarta fascia sono stati contattati. Quest'ultimi quindi non rientrano nell'interesse della banca.

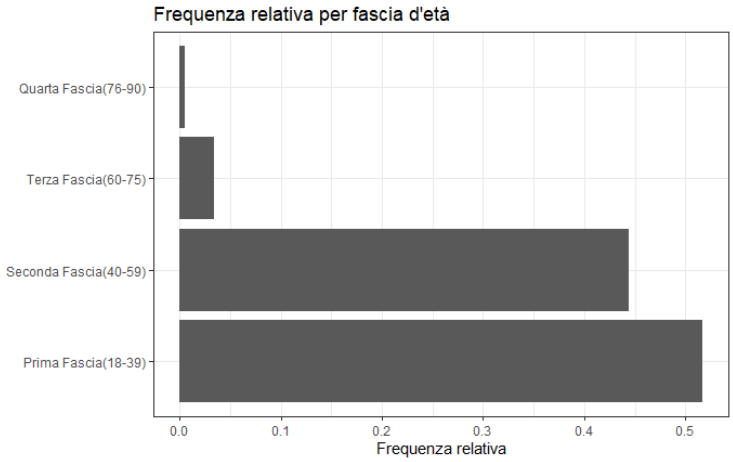


Figura 3.1: **Diagramma a barre delle frequenze relative per fascia d'età**

Altre caratteristiche importanti sono il livello di educazione e il tipo di lavoro. Per cercare di massimizzare le sottoscrizioni di un deposito a termine la banca potrebbe pensare di rivolgersi a un'utenza con un livello di educazione piuttosto alto, in quanto generalmente a un livello di educazione maggiore corrisponde un salario maggiore e quindi anche il tipo di lavoro svolto tende a essere più importante con maggiori responsabilità.

Dalla Figura 3.2 è chiaro come il livello di educazione influenzi il saldo. Infatti chi presenta il livello di educazione più alto ha un saldo medio maggiore e questo vale per tutti e quattro i settori lavorativi. Inoltre per tutti e tre i livelli di educazione notiamo che i lavoratori autonomi (Self Employed) e i lavoratori del settore Administration Management (amministrazione e gestione) presentano un saldo medio maggiore.

In basso viene riportato il livello mediano del saldo. La mediana è una misura robusta poco influenzata dalla presenza di dati anomali, in questo caso saldi molto elevati. Anche in questo caso gli utenti che hanno il livello di educazione più alto presentano un saldo mediano maggiore e questo vale per tutti e quattro i settori lavorativi. L'unica differenza si nota nel livello di educazione secondario. In questo caso il secondo lavoro con il saldo mediano maggiore è quello degli operai (Blue Collar) e non più quello relativo ai lavoratori autonomi. Ciò è dovuto alla natura robusta della mediana.

Alla luce dei risultati della Figura 3.2 si può affermare che non esistono grandi differenze in termini di saldo medio e mediano per chi presenta i primi due livelli di educazione, mentre come già accennato in precedenza chi possiede il livello di educazione più alto presenta un saldo maggiore e questo potrebbe influenzare positivamente la sottoscrizione di depositi a termine.

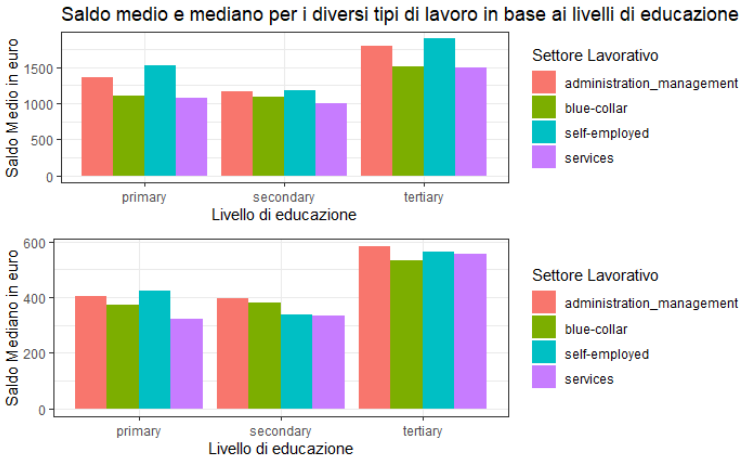


Figura 3.2: **Diagrammi a barre relativi al saldo medio e mediano per i tipi di lavoro in base ai livelli di educazione**

Un'altra caratteristica importante e influente è certamente la storia creditizia di un cliente. Essa può incidere in modo netto sulla praticabilità di un mutuo o di un prestito personale. In questo caso la banca potrebbe scartare a priori gli utenti che risultano inadempienti nei pagamenti.

La Figura 3.3 conferma quanto detto in precedenza. Il 98% degli utenti presenta una storia creditizia buona, ovvero non è un utente inadempiente nei pagamenti.

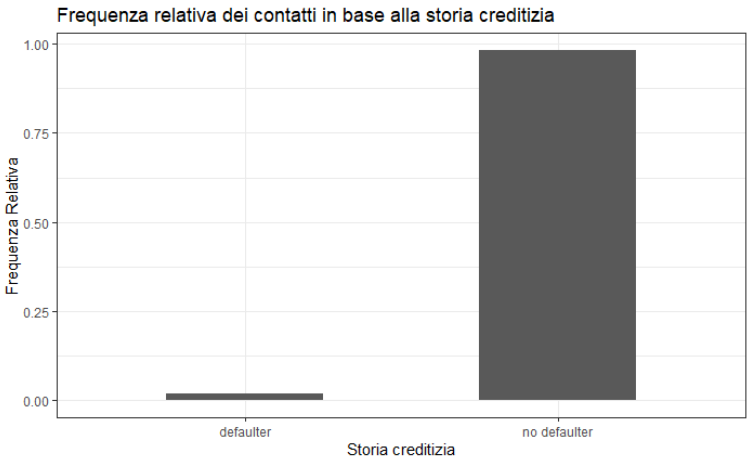


Figura 3.3: **Diagramma a barre della frequenza relativa in base alla storia creditizia**

Vista l'alta percentuale di utenti non inadempienti, possiamo pensare di analizzare caratteristiche come prestiti personali e mutui. Queste caratteristiche possono influenzare negativamente la sottoscrizione di un deposito a termine. Ad esempio pensiamo a quei clienti che hanno un mutuo. In questo caso risulta non proprio semplice investire una certa somma di denaro per il deposito considerata la "spesa" fissa a cui devono far fronte. Analogo discorso per gli utenti che usufruiscono di un prestito personale, per i quali risulta difficile pensare a un'effettiva sottoscrizione di un deposito.

Analizzare la presenza di queste caratteristiche per fascia d'età come nella Figura 3.4,

può essere molto utile in quanto sulla base dei risultati ottenuti la banca potrebbe diversificare la tipologia di sottoscrizione, ad esempio potrebbe proporre una somma "non proibitiva" da investire per coloro che usufruiscono di un prestito personale o di un mutuo, dato che nella maggior parte dei casi è richiesta una somma minima obbligatoria da versare.

Come si può facilmente notare dalla Figura 3.4, la terza fascia presenta la frequenza relativa minore di prestiti personali e di mutui e questo sembra essere ragionevole data l'età dei clienti. Come ci si aspettava, i clienti con un età giovane/medio-giovane presentano una frequenza relativa di mutui e di prestiti personali maggiore. Seguendo la logica precedente, gli utenti della terza fascia composta da pensionati o utenti prossimi alla pensione potrebbero avere una maggiore propensione a sottoscrivere depositi a termine in quanto la maggior parte di essi non ha "spese" a cui badare. Per spese si intende ad esempio l'eventuale mutuo.

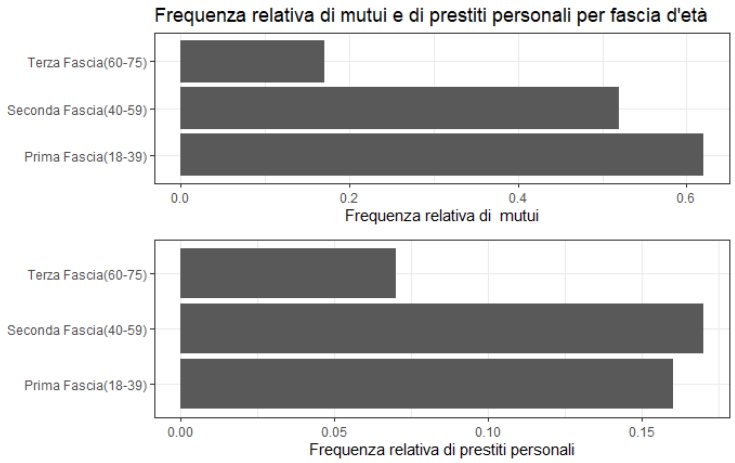


Figura 3.4: **Diagrammi a barre della frequenza relativa di mutui e di prestiti personali per fascia d'età**

## 3.3 Analisi delle variabili in relazione alla variabile target

Prendiamo in considerazione alcune delle variabili presenti nella Tabella 3.1 e nella Tabella 3.2 per esplorare le relazioni esistenti tra le variabili in esame e la variabile target.

Come già detto nel paragrafo precedente, l'età potrebbe incidere in modo significativo sulla sottoscrizione di un deposito a termine in quanto persone di età differenti potrebbero avere interessi differenti.

Come si può facilmente notare dalla Figura 3.5, persone di tutte le età possono sottoscrivere un deposito a termine. Tuttavia gli utenti appartenenti alla fascia d'età trenta-quaranta ne usufruiscono maggiormente.

La stessa fascia d'età però presenta il conteggio più alto anche tra coloro che non hanno sottoscritto un deposito, ma questo è ragionevole in quanto le persone presenti in questa fascia d'età sono anche le più contattate.

Dall'istogramma rappresentato nella Figura 3.5 è chiara la distribuzione asimmetrica dell'età degli utenti. In questo caso si nota un'asimmetria positiva con una lunga coda a destra, confermando ulteriormente quanto già detto nel paragrafo precedente: la banca si è rivolta principalmente a un'utenza giovane/medio-giovane.

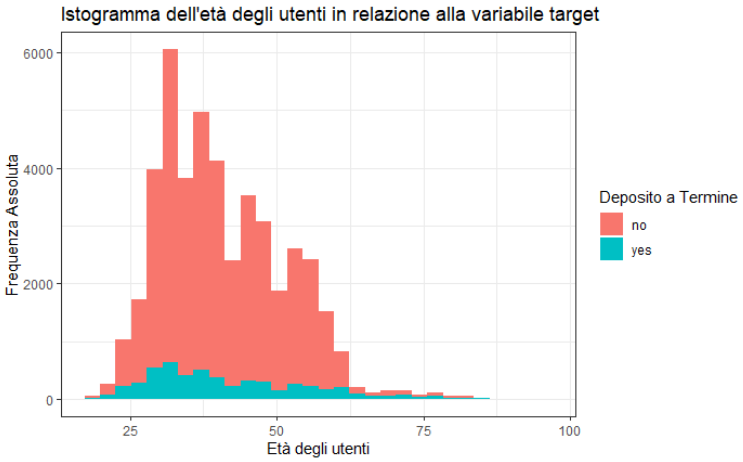


Figura 3.5: **Istogramma della distribuzione dell'età degli utenti in relazione alla variabile target**

Quando si avvia una campagna di marketing bisogna tenere in considerazione vari aspetti. Uno di questi sicuramente è il periodo dell'avvio che può risultare determinante in quanto in determinati periodi dell'anno un lavoratore può ricevere premi di produzione e questo può portare a investire in un deposito a termine la somma "bonus".

Dalla Figura 3.6 notiamo come la maggioranza degli utenti sia stata contattata durante i mesi estivi, in particolare nel mese di maggio.

Sebbene la maggior parte dei contatti avvenga nei mesi estivi, questo impatta solo in misura limitata sulla distribuzione della frequenza di sottoscrizione di un deposito a termine.

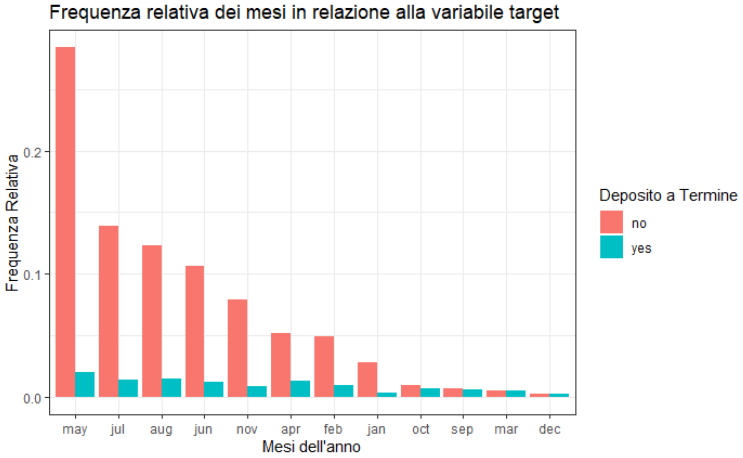


Figura 3.6: **Diagramma a barre della frequenza relativa dei mesi in relazione alla variabile target**

Nel paragrafo precedente abbiamo parlato di come la presenza di prestiti personali e di mutui potrebbe influenzare la sottoscrizione di un deposito a termine e di come la banca potrebbe diversificare le tipologie di sottoscrizioni in base al tipo di cliente.

Nella Figura 3.7 confrontiamo i clienti mutuatari con i clienti non mutuatari in relazione

alla variabile target per analizzare quanto il fattore mutuo incide sulle sottoscrizioni di depositi.

Dalla Figura 3.7 è chiaro come la maggioranza degli utenti contattati dalla banca abbia un mutuo. Tuttavia possiamo notare anche il conteggio maggiore di sottoscrizioni di depositi per coloro che non lo presentano.

Alla luce dei risultati ottenuti possiamo affermare come la presenza/assenza di un mutuo sia in grado d'influenzare la propensione di un cliente a investire. La banca potrebbe adottare una serie di agevolazioni per cercare di invogliare i clienti mutuatari a investire.

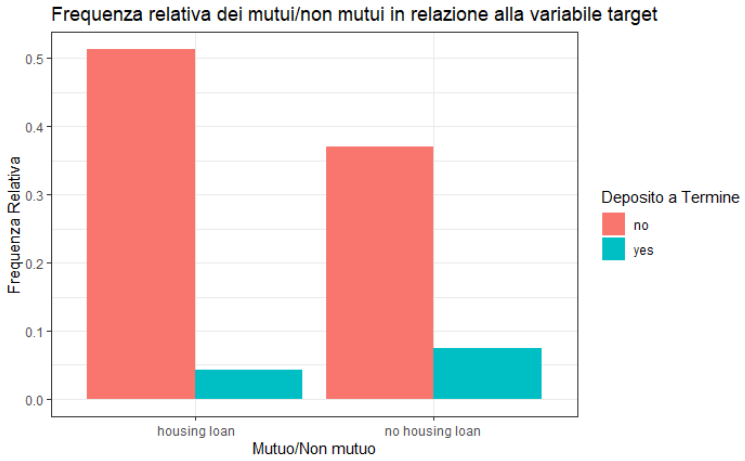


Figura 3.7: **Diagramma a barre della frequenza relativa dei mutui/non mutui in relazione alla variabile target**

La maggioranza degli utenti presenti nel data set non è inadempiente nei pagamenti. Questo fattore quindi può certamente influenzare l'eventuale sottoscrizione di un deposito e questo è chiaro se si analizza la Figura 3.8.

I boxplot sono non facilmente leggibili e ciò è dovuto alla presenza di code "pesanti" che riguardano la distribuzione dei saldi. Tuttavia chi paga le rate in tempo presenta un saldo maggiore e le persone con un saldo maggiore sono più propense a investire. Basta pensare che il saldo mediano degli utenti non inadempienti e che investono è pari a 755 euro, viceversa il saldo mediano degli utenti inadempienti che investono è pari a -2.5.

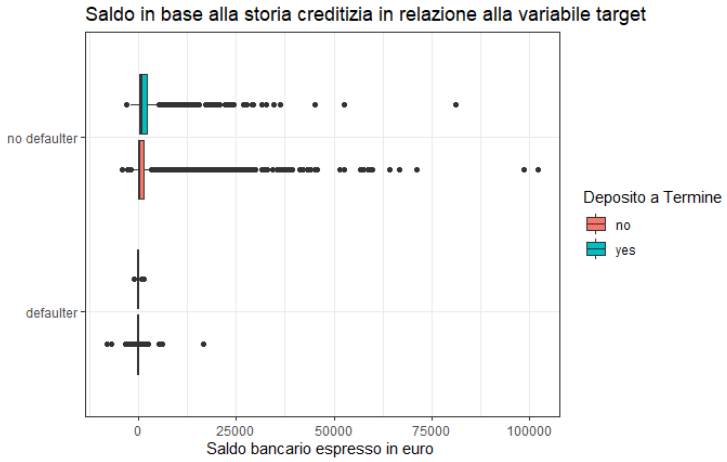


Figura 3.8: **Boxplot relativo al saldo in base alla storia creditizia in relazione alla variabile target**

Dalle analisi precedenti abbiamo scoperto che i saldi maggiori appartengono a coloro che presentano il livello di educazione più alto.

Dunque in linea di massima coloro che presentano un livello di educazione maggiore potrebbero essere più propensi a investire. La Figura 3.9 illustra quanto detto prendendo in considerazione la frequenza relativa di non sottoscrizioni in quanto risulta più semplice darne un'interpretazione.

Dalla Figura 3.9 notiamo come l'84% degli utenti con il livello di educazione massimo decida di non investire. Tuttavia questa percentuale è maggiore se si considerano gli utenti con un livello di educazione minore. In particolar modo possiamo notare l'altissima percentuale di non sottoscrizioni per gli utenti che presentano il livello di educazione primario, pari al 91%. In generale, la maggioranza degli utenti contattati dalla banca ha deciso di non sottoscrivere un deposito.

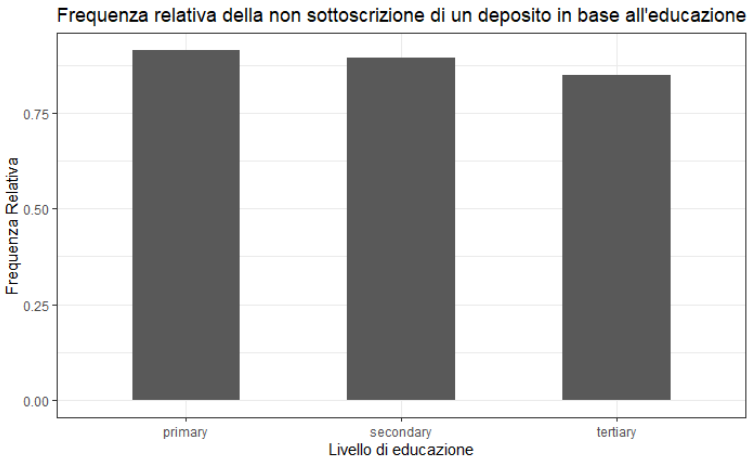


Figura 3.9: **Diagramma a barre della frequenza relativa della non sottoscrizione di un deposito in base al livello di educazione**

Concludiamo il nostro discorso con l'analisi della variabile duration che indica la durata (espressa in secondi) dell'ultimo contatto. Come si può immaginare, la durata del contatto rappresenta un fattore che può giocare un ruolo chiave nella sottoscrizione di un deposito, in quanto la banca in pochi minuti deve spronare il cliente ad avviare un possibile investimento. La Figura 3.10 analizza la distribuzione della durata dell'ultimo contatto in relazione alla variabile target.

Alla luce dei risultati ottenuti, possiamo affermare senza grandi sorprese che la distribuzione della variabile duration per gli utenti che hanno deciso di non sottoscrivere un deposito presenta una variabilità minore e questo sembra ragionevole in quanto chi decide di non investire tende a chiedere meno informazioni alla banca. Di conseguenza la durata complessiva del contatto tende a diminuire. Il discorso opposto invece può essere fatto per coloro che hanno deciso d'investire in un deposito. Dalla Figura 3.10 è chiara la variabilità maggiore della distribuzione della variabile duration. Le rette tratteggiate in rosso indicano il valore medio della durata dell'ultimo contatto. Gli utenti che decidono di non investire generalmente decidono entro i primi quattro minuti (221.1828 secondi), viceversa chi decide d'investire impiega generalmente 9 minuti (537.2946 secondi).

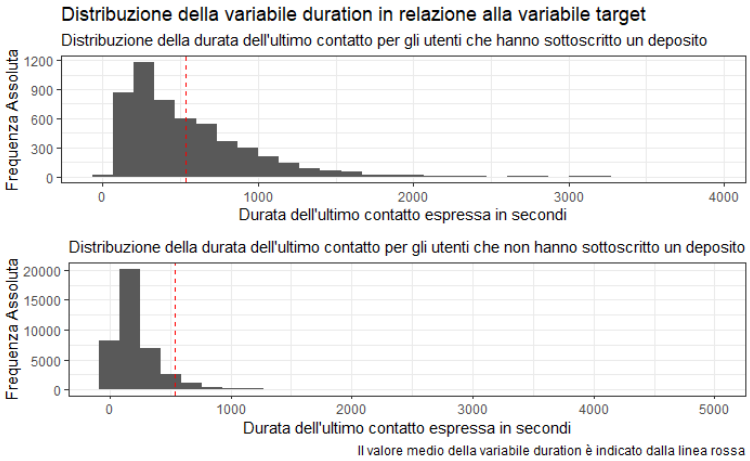


Figura 3.10: **Istogrammi della distribuzione della durata dell'ultimo contatto in relazione alla variabile target**

Un altro aspetto importante da analizzare è la durata dell'ultimo contatto in base alla presenza/assenza di un mutuo, in quanto come già detto in precedenza questo fattore riesce a influenzare la propensione di un utente a investire. La Figura 3.11 ci aiuta a capire quanto la durata in base al fattore mutuo incide sulla sottoscrizione di un deposito.

Dalla Figura 3.11 risulta chiaro quanto già detto in precedenza. I boxplot relativi agli utenti propensi a investire presentano una variabilità maggiore (espressa in secondi). Inoltre gli utenti che presentano un mutuo e che decidono d'investire sono coloro che impiegano più tempo in assoluto e questo sembra essere ragionevole data la loro situazione "delicata".

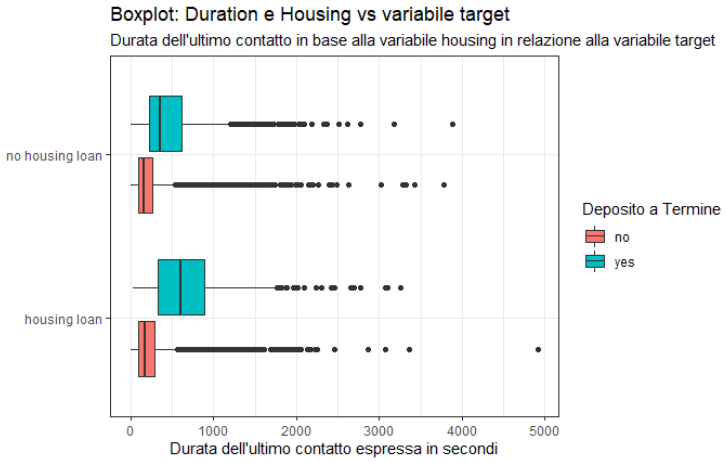


Figura 3.11: **Boxplot per la durata dell'ultimo contatto in base al fattore mutuo in relazione alla variabile target**

Nei paragrafi successivi, è presente la seconda parte del caso studio. Prima della sima dei modelli di classificazione, sarà necessario effettuare una serie di operazioni preliminari, come lo *split* dei dati in *training set* e in *test set*, e l'operazione di bilanciamento delle classi di riferimento del *training set*.

Successivamente, i modelli di classificazione implementati verranno confrontati in base alle proprie metriche di performance. Inoltre verranno effettuate una serie di considerazioni sull'importanza data dai modelli di classificazione alle varie variabili, con la conseguente interpretazione dei coefficienti stimati per le prime tre variabili per importanza per il modello in esame.

## 3.4 Stepwise Logistic Regression con classi sbilanciate e con classi bilanciate

Prima di stimare il modello logit, è necessario dividere il data set in due sottoinsiemi. Ricordiamo che la variabile target, è la sottoscrizione o meno del deposito a termine proposto dall'istituto finanziario.

Il primo sottoinsieme viene comunemente chiamato *Training Set*, mentre il secondo sottoinsieme viene comunemente chiamato *Test Set*. Il training set racchiude i dati di addestramento, mentre il test set racchiude i cosiddetti dati *unseen* sui quali il classificatore è chiamato a effettuare previsioni. Nel caso studio in esame, il training set è composto dall'75% delle osservazioni totali, mentre il test set è composto dall'25% delle osservazioni rimanenti.

Durante l'analisi esplorativa dei dati abbiamo scoperto il forte sbilanciamento esistente tra le classi di riferimento e questo è ben visibile dalla Tabella 3.3. Il forte sbilanciamento si ripercuote ovviamente anche nei due sottoinsiemi creati. Inoltre bisogna notare come il rapporto 88:11 presente nel data set iniziale sia stato rispettato.

In questi casi, il partizionamento effettuato non deve essere casuale ma deve rispettare lo sbilanciamento della classe di riferimento presente nel data set iniziale.

Tabella 3.3: **Percentuale di osservazioni per le classi di riferimento riguardanti il data set, il training set e il test set**

| | **no** | **yes** |
|---|---|---|
| **Data set** | 88.3% | 11.7% |
| **Training set** | 88.3% | 11.7% |
| **Test set** | 88.3% | 11.7% |

Come già detto nel Paragrafo 2.6, lo sbilanciamento può certamente influire sul processo di addestramento dei modelli implementati. Per questo motivo nel corso degli anni sono nate tecniche che cercano di risolvere questo problema di ricorrenza frequente.

Nel caso in esame, abbiamo fatto ricorso alla *"combination of over-sampling and under-sampling"*, che consiste in un mix delle due tecniche illustrate nel Capitolo 2 per bilanciare le classi di riferimento.

Così facendo avremo a disposizione sia i dati con le classi bilanciate, sia i dati con le classi sbilanciate, e questo permetterà di svolgere un'analisi parallela per capire quale classificatore implementato offre le maggiori garanzie. Ovviamente, il processo di bilanciamento riguarderà solo i dati di addestramento (training set) in quanto l'obiettivo è quello di analizzare il processo di apprendimento di un modello.

La Figura 3.12 indica la frequenza relativa delle classi di riferimento sui dati bilanciati appartenenti al training set. Come si può facilmente notare, il bilanciamento è andato a buon fine. È chiaro il rapporto 50:50 delle classi di riferimento, mentre in precedenza tale rapporto era pari a 88:11 come riportato nella Tabella 3.3.

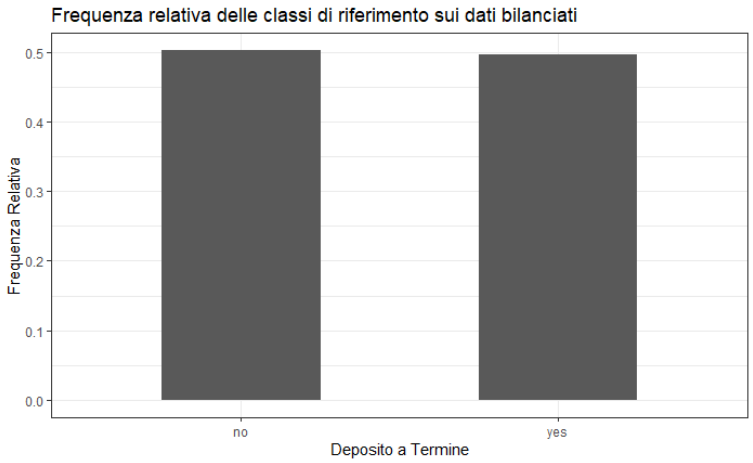


Figura 3.12: **Diagramma a barre della frequenza relativa delle classi di riferimento sui dati bilanciati appartenenti al training set**

Con la fine delle operazioni di "splitting" e bilanciamento dei dati, ci addentriamo ora nella cosiddetta fase di "Model Selection". In questa fase, utilizzeremo i dati del training set, e attraverso la procedura di cross-validation descritta nel Paragrafo 2.5 andremo a scegliere quel modello che ottimizza il bias-variance trade-off per evitare fenomeni come ad esempio l'overfitting. Per effettuare il processo di variable selection utilizziamo uno degli approcci Stepwise, in questo caso la Backward elimination.

La Backward elimination inizia con tutti i predittori presenti nel data set, stimando un modello completo che rappresenta il punto di partenza per l'algoritmo. Successivamente l'algoritmo rimuove in modo iterativo dal modello iniziale (completo) i predittori meno contributivi e si interrompe quando si dispone di un modello in cui tutti i predittori sono considerati rilevanti, nel senso che non si registra un peggioramento nella capacità contributiva delle variabili alla spiegazione del fenomeno.

Nel nostro caso rappresenta l'opzione migliore in quanto è efficace quando la dimensione del campione ($n$) è maggiore del numero di variabili ($p$). La Tabella 3.4 riporta le metriche di performance relative ai modelli ottimali scelti durante la fase di "Model Selection".

La procedura di selezione Backward elimination cerca di minimizzare una funzione-criterio di ottimalità. In questo caso la funzione criterio scelta è l'AIC (Akaike information criterion). È un metodo che permette di valutare e confrontare diversi modelli statistici, e fornisce una misura della qualità della stima di un modello prendendo in considerazione sia la bontà di adattamento, sia la complessità del modello. In generale, scegliamo quei modelli che presentano il valore più basso di AIC.

In questo caso, come si può notare dalla Tabella 3.4, durante la fase di "Model Selection" sono stati scelti i modelli che presentavano l'AIC più basso.

Tabella 3.4: **Metriche di performance per i modelli logit in base al tipo di classe**

| **Tipi di classe** | **AIC** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| sbilanciate | 16156 | 0.35 | 0.97 |
| bilanciate | 26374 | 0.81 | 0.85 |

Terminata la fase di "Model Selection", ci addentriamo ora in una delle fasi più importanti, ovvero la "Model Validation". In questa fase, utilizziamo i modelli scelti durante la fase di selezione per effettuare previsioni sui dati unseen (Test set). In questo caso ci interessa una stima accurata del suo livello di errore.

Attraverso l'analisi di varie metriche di performance, sceglieremo il modello che offre le maggiori garanzie, in questo caso quel modello che riesce a classificare con un certo livello di affidabilità coloro che hanno sottoscritto un deposito a termine. Risulta molto importante dunque minimizzare la percentuale di falsi negativi. Le Tabelle 3.5 e 3.6 riportano i dettagli delle matrici di confusione per i modelli logit stimati.

Dalla Tabella 3.5 notiamo che in media il logit sulle classi sbilanciate prevede bene l'86% dei veri negativi, ovvero coloro che non hanno sottoscritto un deposito a termine. Come già detto durante l'analisi esplorativa dei dati, la maggioranza degli utenti non ha sottoscritto un deposito a termine, e quindi non sorprende la buona capacità del classificatore di discriminare correttamente i veri negativi.

A causa del forte sbilanciamento che si ha tra le due classi di riferimento si fa fatica a classificare correttamente coloro che hanno sottoscritto un deposito a termine e questo rappresenta un problema in quanto si vuole implementare un classificatore capace di discriminare con un certo livello di fedeltà i veri positivi, in quanto rappresentano la classe di maggior interesse.

Dalla Tabella 3.6 invece, notiamo che in media il logit sulle classi bilanciate prevede bene il 9.5% dei veri positivi, in precedenza tale soglia era pari all'4.1%. La percentuale di falsi positivi è pari all'13.6% mentre in precedenza tale soglia era pari all'2.3%. Con il bilanciamento delle classi, la percentuale di falsi negativi si è ridotta all'2.2% mentre in precedenza tale soglia era pari all'7.6%.

Tabella 3.5: **Matrice di confusione del modello logit per classi sbilanciate**

| | | Reference | | |
|---|---|---|---|---|
| | | **no** | **yes** | Total |
| Prediction | **no** | 86% | 7.6% | 93.6% |
| | **yes** | 2.3% | 4.1% | 6.4% |
| | Total | 88.3% | 11.7% | 100% |

Tabella 3.6: **Matrice di confusione del modello logit per classi bilanciate**

| | | Reference | | |
|---|---|---|---|---|
| | | **no** | **yes** | Total |
| Prediction | **no** | 74.7% | 2.2% | 76.9% |
| | **yes** | 13.6% | 9.5% | 23.1% |
| | Total | 88.3% | 11.7% | 100% |

Partendo dalle Tabelle 3.5 e 3.6 ricaviamo varie metriche di performance, così facendo potremo giudicare quanto i modelli implementati fanno bene sui dati unseen. La Tabella 3.7 racchiude alcune delle metriche di performance utilizzate per valutare un modello statistico nei problemi di classificazione.

Il modello che fornisce le maggiori garanzie è il logit stimato sulle classi bilanciate, e questo è chiaro se si dà uno sguardo alla varie metriche di performance riportate nella Tabella 3.7.

Sensitività e precisione sono tra loro inversamente proporzionali, quindi è normale avere una precisione minore a fronte di una sensitività maggiore.

Il coefficiente di correlazione di Matthew è pari a 0.50. Ricordiamo che esso varia in un range che va da $[-1, +1]$. Un MCC $\rightarrow$ +1 indica una previsione perfetta.

Il kappa di Cohen è pari a 0.46 e prendendo in considerazione la classificazione dei valori di $k$ proposta da Landis JR e Koch GG (1977), possiamo affermare che esiste una moderata concordanza tra i due valutatori in quanto $0.41 \leq k \leq 0.60$.

L'AUC è pari a 0.83 e questo valore suggerisce che il classificatore implementato è moderatamente accurato, in quanto $0.7 \leq AUC \leq 0.9$.

Tabella 3.7: **Metriche di performance per il modello logit in base al tipo di classi**

| **Tipi di classe** | **Acc** | **Sens** | **Prec** | **MCC** | **F1-Score** | **Kappa** | **Auc** |
|---|---|---|---|---|---|---|---|
| sbilanciate | 0.90 | 0.33 | 0.63 | 0.42 | 0.43 | 0.39 | 0.65 |
| **bilanciate** | **0.84** | **0.81** | **0.41** | **0.50** | **0.54** | **0.46** | **0.83** |

## 3.5 Elastic net con classi sbilanciate e con classi bilanciate

In questo paragrafo stimiamo un modello di classificazione basato sulla penalizzazione, così come introdotto nel Capitolo 2. Saranno applicati i medesimi step visti nel paragrafo precedente.

Anche in questo caso, nella fase di *"Model Selection"*, utilizzeremo i dati del *training set*, e attraverso la procedura di *cross-validation* descritta nel Paragrafo 2.5 andremo a scegliere quel modello che ottimizza il *bias-variance trade-off* sulla base dei due iperparametri

appartenenti al modello elastic-net.

La Tabella 3.8 contiene gli iperparametri scelti con le relative metriche di performance.

Ricordiamo che il valore di alpha permette di avvicinarsi in modo elastico ai termini di penalizzazione Lasso e Ridge, mentre il valore di lambda gestisce la "grandezza" della penalizzazione. È fondamentale scegliere un modello non troppo complesso, in quanto bisogna seguire sempre il principio della parsimonia.

Dalle metriche di performance è chiaro che il modello elastic net addestrato sui dati bilanciati offre garanzie maggiori nonostante il livello di accuratezza leggermente minore. In caso di forte sbilanciamento tra le classi, l'accuratezza può essere fuorviante. In questo caso diamo maggiore importanza alla metrica kappa di cohen.

Tabella 3.8: **Iperparametri e metriche di performance per il modello elastic net sui dati sbilanciati e sui dati bilanciati**

| **Tipi di classe** | **Alpha** | **Lambda** | **Accuracy** | **Kappa** |
|---|---|---|---|---|
| sbilanciate | 0.55 | 0.002 | 0.90 | 0.39 |
| bilanciate | 0.1 | 0.004 | 0.83 | 0.67 |

Ora, utilizziamo i modelli scelti durante la fase di selezione per effettuare le previsioni sui dati unseen. Le Tabelle 3.9 e 3.10 riportano i dettagli delle matrici di confusione per i modelli elastic net implementati.

Come si può ben notare dalle Tabelle 3.9 e 3.10, le matrici di confusione dei modelli elastic net implementati sulle classi sbilanciate e sulle classi bilanciate sono quasi equivalenti alle matrici di confusione viste nel paragrafo precedente per i modelli logistici non penalizzati.

Alla luce di ciò, possiamo affermare che il modello elastic net e il modello logit offrono performance molto simili. In questo caso, valgono le medesime considerazioni fatte nel paragrafo precedente.

È chiaro che i modelli implementati sulle classi sbilanciate fanno fatica a classificare correttamente coloro che hanno sottoscritto un deposito a termine e questo non è un bene in quanto siamo interessati a un modello di classificazione capace d'individuare con un certo livello di fedeltà i veri positivi e che allo stesso tempo minimizzi la percentuale di falsi negativi.

Tabella 3.9: **Matrice di confusione del modello elastic net con classi sbilanciate**

| | | Reference | | |
|---|---|---|---|---|
| | | **no** | **yes** | Total |
| Prediction | **no** | 86.1% | 7.8% | 93.9% |
| | **yes** | 2.2% | 3.9% | 6.1% |
| | Total | 88.3% | 11.7% | 100% |

Tabella 3.10: **Matrice di confusione del modello elastic net con classi bilanciate**

| | | Reference | | |
|---|---|---|---|---|
| | | **no** | **yes** | Total |
| Prediction | **no** | 75% | 2.2% | 77.2% |
| | **yes** | 13.3% | 9.5% | 22.8% |
| | Total | 88.3% | 11.7% | 100% |

Nelle matrici di confusione sono riportate le percentuali di osservazioni appartenenti ai quattro casi possibili (descritti nel Capitolo 2, Paragrafo 2.4) che si possono avere nei problemi di classificazione, e come abbiamo visto esse sono quasi equivalenti.

Per cercare di carpire le differenze minime esistenti tra i modelli implementati analizziamo le Figure 3.13 e 3.14.

Come si può notare dalle Figure 3.13 e 3.14, le differenze tra i due modelli implementati sulle classi bilanciate sono minime e questo giustifica le percentuali simili viste nelle matrici di confusione precedenti. Analogo discorso può essere fatto per i modelli stimati sulle classi sbilanciate, ma essi non sono di nostro interesse in quanto offrono garanzie minori.

Dando uno sguardo più approfondito alle due Figure notiamo che ad esempio nella classificazione dei veri positivi e nella classificazione dei falsi negativi esiste uno scarto di una sola osservazione a favore del modello elastic net.

L'unica leggera differenza si nota nel conteggio dei veri negativi e dei falsi positivi. In questo caso in entrambi i casi esiste uno scarto di venti osservazioni. Quando le percentuali sono simili tra di loro, è utile valutare le frequenze assolute per evidenziare le differenze minime esistenti. Tuttavia in generale le percentuali sono utili in quanto le frequenze assolute vengono divise per il numero totale di osservazioni e questo permette di effettuare valutazioni più rapide.

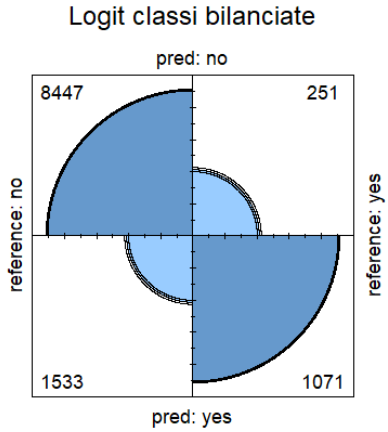


Figura 3.13: **Fourfold plots modello logit con classi bilanciate**

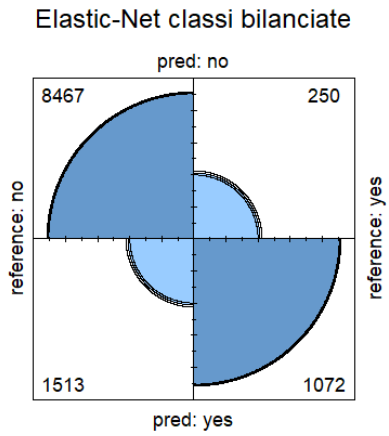


Figura 3.14: **Fourfold plots modello elastic net con classi bilanciate**

Prendendo in considerazione i due modelli stimati sulle classi bilanciate, sarebbe utile capire quali variabili presentano il maggior potere esplicativo. Inoltre è utile interpretare i coefficienti stimati relativi a tali variabili per effettuare una serie di valutazioni.

L'importanza delle variabili per i modelli regolarizzati fornisce un'interpretazione simile a quella della regressione lineare (o logistica). L'importanza è determinata dalla grandezza dei coefficienti standardizzati. Simile alla regressione lineare e logistica, la relazione tra le caratteristiche e la risposta è lineare monotona, tuttavia bisogna ricordare che se alla variabile di risposta è stata applicata una trasformazione logaritmica, $\log(Y)$, le relazioni stimate saranno ancora monotone ma non lineari sulla scala di risposta originale.

Nelle Figure 3.15 e 3.16 sono presenti le prime dieci variabili per importanza per ciascun modello.

Per il modello logit la variabile col maggior potere esplicativo è "duration", che indica la durata espressa in secondi dell'ultimo contatto avuto con la banca, mentre per il modello elastic net, la variabile col maggiore potere esplicativo è "poutcomesuccess" che rappresenta gli utenti che hanno accettato l'offerta proposta dalla banca nella scorsa campagna di marketing.

I due modelli considerati hanno differenti assunzioni teoriche alla base e hanno un diverso processo di stima, per questo motivo le variabili possono avere ordini differenti di importanza e questo è chiaro se si dà uno sguardo alle Figure 3.15 e 3.16.

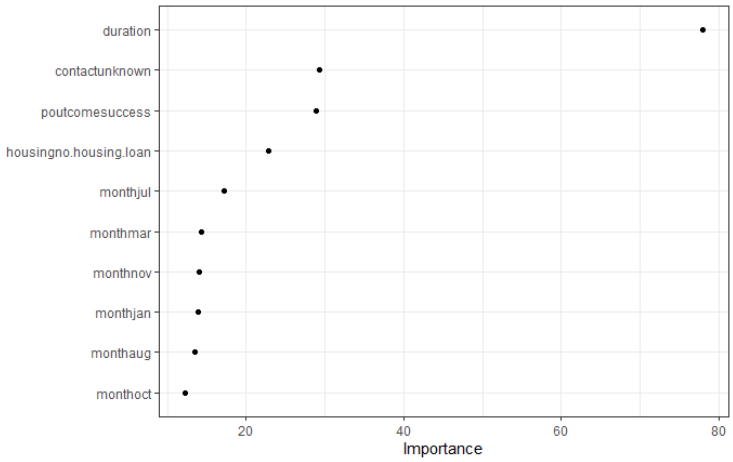


Figura 3.15: **Prime dieci variabili per importanza per il modello logit con classi bilanciate**

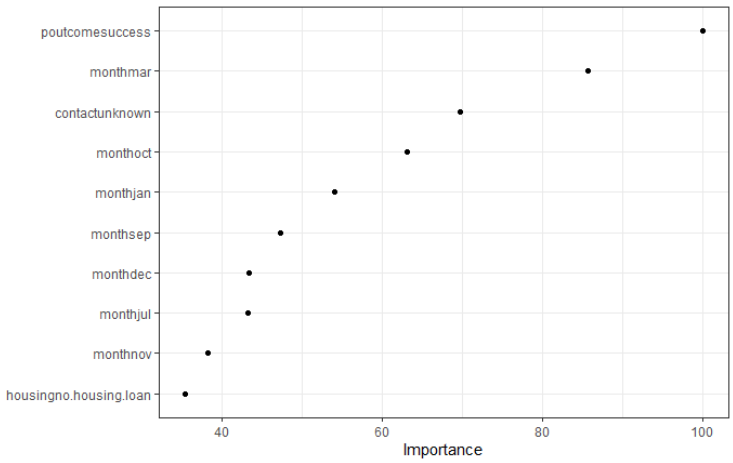


Figura 3.16: **Prime dieci variabili per importanza per il modello elastic net con classi bilanciate**

Come detto in precedenza, è utile non solo capire quali variabili presentano il maggior potere esplicativo, ma anche interpretare il segno dei relativi coefficienti stimati per capire la natura della relazione tra la variabile in esame e la variabile di risposta.

Le Tabelle 3.11 e 3.12 contengono le stime dei coefficienti relativi alle prime tre variabili per importanza di ciascun modello.

Partendo dalla Tabella 3.11, notiamo che i coefficienti stimati positivi indicano una relazione positiva tra la variabile d'interesse e la variabile di risposta, viceversa i coefficienti stimati negativi indicano una relazione negativa.

Per le variabili di tipo categoriale, come *contact unknown* e *poutcome success*, i coefficienti stimati devono essere interpretati in funzione del livello di riferimento scelto. Per evitare la trappola delle variabili *dummy* bisogna inserire nel modello un numero di *dummy* che è sempre pari a uno in meno rispetto ai vari livelli che la variabile categoriale in esame assume. Questo è importate per evitare il problema della multicollinearità.

La scelta del livello di riferimento avviene in modo casuale, di solito viene preso in considerazione il primo dei livelli che la variabile può assumere. Tuttavia si può anche decidere quale variabile *dummy* deve essere esclusa in base alle nostre esigenze. Nel seguente caso studio per ogni variabile categoriale è stato escluso il primo dei livelli che la variabile può assumere.

La stima del coefficiente associato alla variabile *duration* è pari a 0.006, ovvero mantenendo costanti i valori delle rimanenti variabili indipendenti, quando la durata dell'ultimo contatto subisce un incremento (una variazione unitaria), ci aspettiamo un incremento nel log dell'*odds* ratio pari a 0.006. L'*odds-ratio* è pari all'esponenziale del coefficiente, che è pari a 1.004, mentre la variazione percentuale dell'*odds* di $Y = 1$ (rappresenta la classe di riferimento, ovvero coloro che hanno sottoscritto un deposito a termine) rispetto alla variazione unitaria della variabile *duration* è pari a 0.60%, data dalla formula $(\exp(0.006) - 1 \times 100)$.

La stima del coefficiente associato alla variabile *contactunknown* è pari a -1.6. Questo significa che, mantenendo costanti i valori delle rimanenti variabili indipendenti, in media la probabilità di sottoscrivere un deposito a termine per coloro che sono stati contattati dalla banca via cellulare (rappresenta la modalità di contatto di riferimento) è maggiore rispetto agli utenti la cui modalità di contatto risulta sconosciuta. Il *log odds* è pari a -1.6, mentre il relativo *odds-ratio* è pari a 0.2. La variazione percentuale dell'*odds* di $Y = 1$ rispetto alla variabile *contactunknown* è pari a -79.8%.

Infine, la stima del coefficiente associato alla variabile *poutcomesuccess* è pari a 2.3. Ciò significa che, mantenendo costanti i valori delle rimanenti variabili indipendenti, in media la probabilità di sottoscrivere un deposito a termine per gli utenti che hanno accettato l'offerta proposta dalla banca nella scorsa campagna di marketing, è maggiore rispetto agli utenti che l'hanno rifiutata. Il *log odds* è pari a 2.3, mentre il relativo *odds-ratio* è pari a 9.9. La variazione percentuale dell'*odds* di Y = 1 rispetto alla variabile *poutcomesuccess* è pari a 897.41%.

Tabella 3.11: **Stime dei coefficienti associati alle prime tre variabili per importanza per il modello logit net con classi bilanciate**

| duration | contactunknown | poutcomesuccess |
|---|---|---|
| 0.006 | -1.6 | 2.3 |

Per quanto riguarda la Tabella 3.12, la stima del coefficiente associato alla variabile *poutcomesuccess* è pari a 2.1. Quindi possiamo dire che, mantenendo costanti i valori delle rimanenti variabili indipendenti, in media la probabilità di sottoscrivere un deposito a termine per gli utenti che hanno accettato l'offerta proposta dalla banca nella scorsa campagna di marketing è maggiore rispetto agli utenti che l'hanno rifiutata (categoria di riferimento).

Il coefficiente associato al mese di marzo è pari a 1.8. Quindi, mantenendo costanti i valori delle rimanenti variabili indipendenti, in media la probabilità di sottoscrivere un deposito per gli utenti contattati nel mese di marzo è maggiore rispetto agli utenti contattati nel mese di aprile (rappresenta il mese di riferimento).

La stima del coefficiente *contactunknown* è pari a -1.4. Mantenendo costanti i valori delle rimanenti variabili indipendenti, in media la probabilità di sottoscrivere un deposito a termine per coloro che sono stati contattati dalla banca via cellulare (rappresenta la modalità di contatto di riferimento) è maggiore rispetto agli utenti la cui modalità di contatto risulta sconosciuta.

Tabella 3.12: **Stime dei coefficienti associati alle prime tre variabili per importanza per il modello elastic net con classi bilanciate**

| poutcomesuccess | monthmar | contactunknown |
|---|---|---|
| 2.1 | 1.8 | -1.4 |

## 3.6 Decision Trees con classi sbilanciate e con classi bilanciate

Concludiamo le nostre analisi implementando il modello di classificazione *decison trees* (Albero di decisone). In questo caso, nella fase di *Model Selection* andremo a scegliere quel modello che sulla base dell'iperparametro alpha ottimizza il *bias-variance trade-off*.

La Tabella 3.13 contiene l'iperparametro scelto con le relative metriche di performance.

Come già detto nel Paragrafo 2.3, per un determinato valore di $\alpha$ troviamo il più piccolo albero potato che ha il più basso errore penalizzato. Nel nostro caso l'errore sarà il tasso di errata classificazione. Il valore di alpha è fondamentale in quanto regola il livello di complessità dell'albero. Alberi di decisione profondi generano modelli complessi e poco parsimoniosi.

Tabella 3.13: **Parametro di complessità alpha con le relative metriche di performance**

| Tipi di classe | Alpha | Sensitivity | Specificity |
|---|---|---|---|
| sbilanciate | 0.002 | 0.34 | 0.97 |
| bilanciate | 0.005 | 0.79 | 0.83 |

Utilizziamo i modelli scelti durante la fase di selezione per effettuare previsioni sui dati unseen. Le Tabelle 3.14 e 3.15 riportano i dettagli delle matrici di confusione per i modelli decision trees implementati.

Alla luce dei risultati ottenuti, possiamo affermare che lo sbilanciamento esistente tra le classi condiziona l'apprendimento dei vari classificatori. Come già detto in precedenza, si fa fatica a classificare chi sottoscrive un deposito a termine e questo lo si nota anche dall'alta percentuale di falsi negativi.

Dalla tabella 3.15, notiamo che in media l'albero di decisione implementato sulle classi bilanciate prevede correttamente l'8.7% dei veri positivi, mentre la percentuale di falsi negativi è pari all'3%, in precedenza, per i modelli elastic net e logit la percentuale di veri positivi era pari all'9.5%, mentre la percentuale di falsi negativi era pari all'2.2%.

Tabella 3.14: **Matrice di confusione del modello decision trees con classi sbilanciate**

| | | Reference | | |
|---|---|---|---|---|
| | | **no** | **yes** | Total |
| Prediction | **no** | 86.3% | 8% | 94.3% |
| | **yes** | 2% | 3.7% | 5.7% |
| | Total | 88.3% | 11.7% | 100% |

Tabella 3.15: **Matrice di confusione del modello deicison trees con classi bilanciate**

| | | Reference | | |
|---|---|---|---|---|
| | | **no** | **yes** | Total |
| Prediction | **no** | 73.6% | 3% | 76.6% |
| | **yes** | 14.7% | 8.7% | 23.4% |
| | Total | 88.3% | 11.7% | 100% |

L'albero di decisione dunque è il modello di classificazione che offre le performance peggiori come riportato nella tabella 3.16.

Tabella 3.16: **Metriche di performance per i modelli elastic net, logit e decision trees implementati sulle classi bilanciate**

| Modelli | Acc | Sens | Prec | MCC | F1-Score | Kappa | Auc |
|---|---|---|---|---|---|---|---|
| **Elastic-Net/Logit** | **0.84** | **0.81** | **0.41** | **0.50** | **0.54** | **0.46** | **0.83** |
| Decision trees | 0.82 | 0.75 | 0.37 | 0.44 | 0.50 | 0.40 | 0.80 |

Nonostante le performance peggiori, sarebbe utile visualizzare la struttura dell'albero di decisione per capire come esso ragiona in termini di classificazione, e questo è ben visibile nella figura 3.17.

Il primo nodo contiene il 100% delle osservazioni di training set ed è chiamato nodo radice (root). La classe prevista è quella dei "no", che rappresenta gli utenti che non hanno sottoscritto un deposito a termine.

Successivamente ha inizio il partizionamento ricorsivo binario descritto nel Paragrafo 2.3. In questo caso la variabile duration che indica la durata dell'ultimo contatto, crea la partizione che riduce al massimo la misura di variabilità scelta, di solito l'indice di Gini, sulla variabile dipendente.

Se la durata dell'ultimo contatto è minore di 226 secondi (3.7 minuti), allora viene creato un nodo contenente il 44% delle osservazioni iniziali e la classe prevista è "no", con una probabilità di sottoscrizione di un deposito a termine pari all'24%. Se la durata dell'ultimo contatto è maggiore di 226 secondi (3.7 minuti), allora viene creato un nodo composto dall'56% delle osservazioni rimanenti, la classe prevista è "yes", con una probabilità di sottoscrizione di un deposito a termine pari all'70%.

Successivamente la procedura descritta poc'anzi viene eseguita in modo ricorsivo. Infatti per ogni sottoinsieme creato si va alla ricerca della variabile da cui scaturisce la partizione che minimizza la misura di variabilità scelta sulla variabile dipendente. L'obiettivo del partizionamento è ridurre al minimo la dissimilarità (diversità) nei nodi terminali, che in questo caso sono sette.

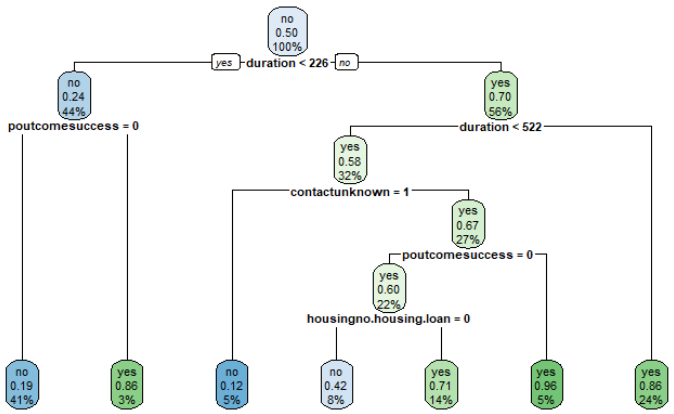


Figura 3.17: **Struttura dell'albero di decisione implementato sulle classi bilanciate**

## 3.7 Conclusioni

Alla luce dei risultati ottenuti, è chiaro come il bilanciamento delle classi riesca a influenzare positivamente il processo di addestramento dei vari classificatori implementati.

Per bilanciare le classi è stata utilizzata la libreria di R denominata ROSE (*Random Over-Sampling Examples*), in particolare la tecnica utilizzata è la *combination of over-sampling and under-sampling* come già detto nel paragrafo 3.4.

Per studi futuri, possiamo pensare di applicare differenti tecniche di bilanciamento per cercare quella che offre le maggiori garanzie, in particolare si può pensare di applicare l'algoritmo SMOTE-NC, facente parte della libreria SMOTE in R o anche si può pensare di applicare il campionamento sintetico adattivo(ADASYN), entrambi gli algoritmi sono descritti nel paragrafo 2.6.

Un altro approccio utile è utilizzare una serie di tecniche che agiscono sul classificatore attraverso l'utilizzo di penalizzazioni. Il classificatore da la stessa importanza ai clienti che hanno sottoscritto un deposito a termine e ai clienti che non hanno sottoscritto un deposito a termine. Nella realtà dei fatti questo approccio non è proprio corretto in quanto la perdita legata all'errata classificazione di un cliente che ha sottoscritto un deposito a termine è maggiore della perdita dovuta all'errata classificazione di un cliente che non ha sottoscritto un deposito a termine.

Per risolvere il seguente problema possiamo pensare di costruire una matrice di perdita, dando valori di perdita differenti in base ai due tipi di errore di errata classificazione, dando maggior "peso" ai clienti che hanno sottoscritto un deposito a termine mal classificati. L'approccio che possiamo seguire è quello di attribuire una serie di penalizzazioni differenti per poi verificare quale di essa offre le migliori performance.

Infine, si può pensare di applicare vari modelli di machine learning, sia appartenenti al mondo del learning supervisionato, sia quelli appartenenti al mondo del learning non supervisionato come il clustering, o anche si può pensare di combinare entrambi gli approcci. I modelli implementati nel caso studio appartengono al mondo del learning supervisionato ma non sono gli unici, infatti si può pensare di applicare modelli come il support vector machine(SVM) o anche il random forest, che rappresenta un ulteriore passo in avanti rispetto al semplice albero decisionale, di conseguenza modelli differenti appartenenti ai vari mondi del machine learning potrebbero offrire garanzie maggiori rispetto ai modelli implementati in questo caso studio.

# Bibliografia

[1] Andrew McAfee ed Erik Brynjolfsson. (2016). BIG DATA: LA RIVOLUZIONE MANAGERIALE. Harvard Business Review ITALIA. La grande sfida della trasformazione digitale(pag. 13-15, 22-23). Strategiqs Edizioni

[2] Thomas Davenport ed D.J. Patil. (2016). DATA SCIENTIST: IL LAVORO PIÙ INTERESSANTE DEL XXI SECOLO. Harvard Business Review ITALIA. La grande sfida della trasformazione digitale(pag. 43-46). Strategiqs Edizioni

[3] Darrel K.Rigby. (2016). FONDERE DIMENSIONE DIGITALE E MONDO FISICO. Harvard Business Review ITALIA. La grande sfida della trasformazione digitale(pag. 58-60). Strategiqs Edizioni

[4] Timothy Morey, Theodore "Theo" Forbath, Allison Schoop. I DATI DEI CLIENTI: PROGETTARE CREANDO TRASPARENZA E FIDUCIA. Harvard Business Review ITALIA. La grande sfida della trasformazione digitale(pag. 85-95). Strategiqs Edizioni

[5] R. Carter Hill, William E. Griffiths, Guay C. Lim. (2013). MODELLI PER VARIABILI DIPENDENTI QUALITATIVE E LIMITATE. PRINCIPI DI ECONOMETRIA (pag. 651-661). Zanichelli

# Sitografia

[6] Irene Di Deo. Data Integration: cosa significa, come farla e perché farla!. `https://blog.osservatori.net/it_it/data-integration-cosa-significa-come-farla`

[7] Laura Zanotti. Hadoop: perché è fondamentale per i big data. `https://www.zerounoweb.it/techtarget/searchdatacenter/hadoop-perche-e-fondamentale-per-i-big-data-quali-le-evoluzioni/`

[8] Inside Marketing. Big data cosa sono e perché sono importanti per le aziende. `https://www.insidemarketing.it/glossario/definizione/big-data/`

[9] BLOG Digital Banking. Big Data nel Futuro Digitale delle Banche.
https://socialbanca.it/la-rivoluzione-digital-dei-big-data-per-le-banche/

[10] Nazareno Lecis. Big Data nel settore bancario: tutto ciò che dovresti sapere.
https://financecue.it/big-data-nel-settore-bancario-tutto-cio-che-dovresti-sapere/13842/

[11] Affde. Big Data nel settore bancario: vantaggi, sfide e applicazioni.
https://www.affde.com/it/big-data-in-banking-advantages-challenges-and-applications.html

[12] Nodes. I Data Analytics nel settore bancario.
https://nodes.it/data-analytics-settore-bancario

[13] Alessandro Piva. Osservatori Digital Innovation della School of Management del Politecnico di Milano. Le 5V dei Big Data: dal Volume al Valore.
https://blog.osservatori.net/it_it/le-5v-dei-big-data

[14] ICHI.PRO. 5 tecniche SMOTE per sovracampionare i dati di squilibrio.
https://ichi.pro/it/5-tecniche-smote-per-sovracampionare-i-dati-di-squilibrio-202401874961077

[15] Davide Nardini. PULP LEARNING. Oversampling e Undersampling.
https://pulplearning.altervista.org/tecniche-undersampling-oversampling/