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.

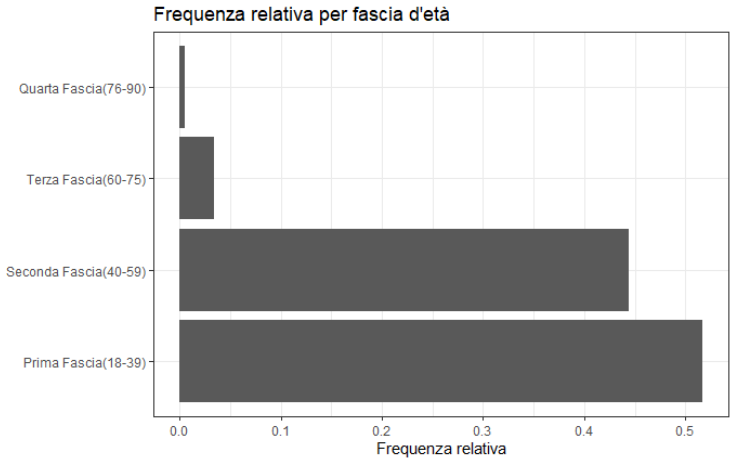

Figura 3.1: **Diagramma a barre delle frequenze relative per fascia d'età**

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.

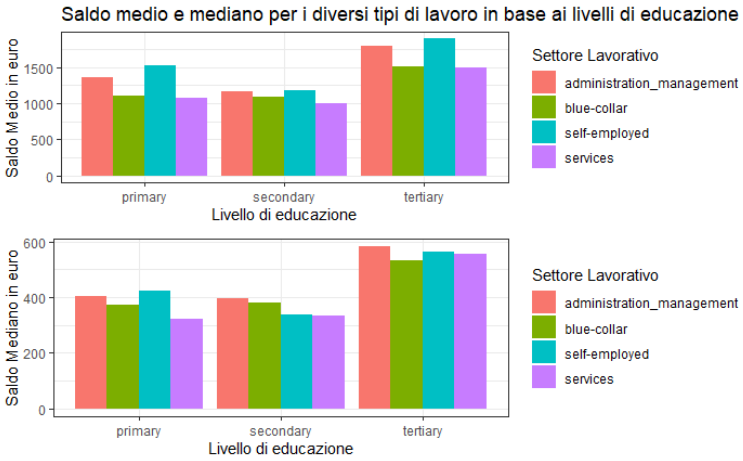

Figura 3.2: **Diagrammi a barre relativi al saldo medio e mediano per i tipi di lavoro in base ai livelli di educazione**

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.

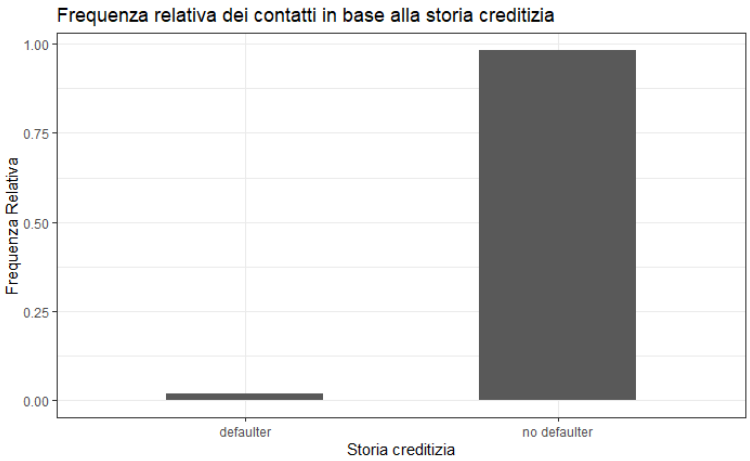

Figura 3.3: **Diagramma a barre della frequenza relativa in base alla storia creditizia**

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.

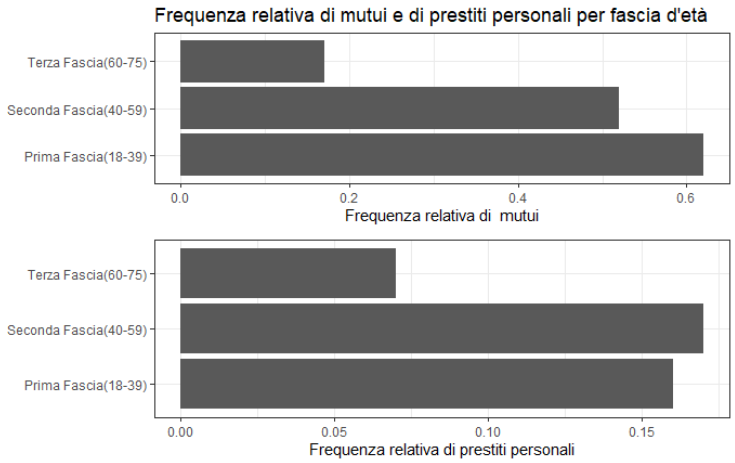

Figura 3.4: **Diagrammi a barre della frequenza relativa di mutui e di prestiti personali per fascia d'età**

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.

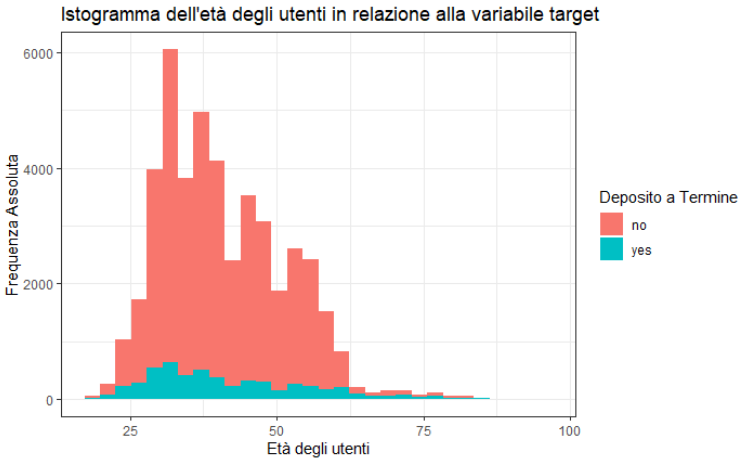

Figura 3.5: **Istogramma della distribuzione dell'età degli utenti in relazione alla variabile target**

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.

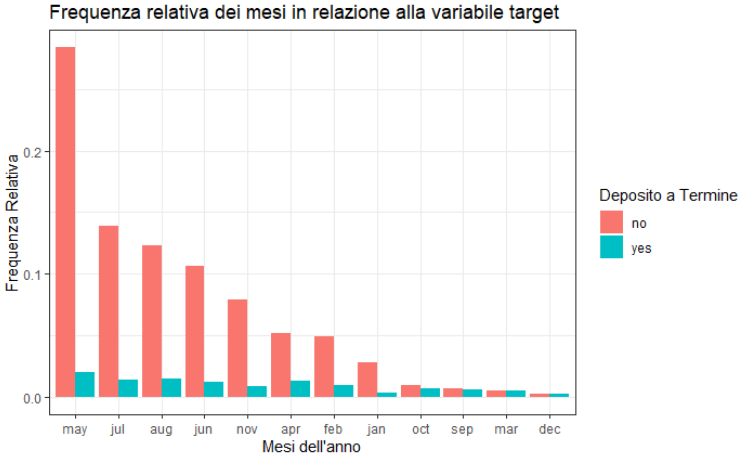

Figura 3.6: **Diagramma a barre della frequenza relativa dei mesi in relazione alla variabile target**

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.

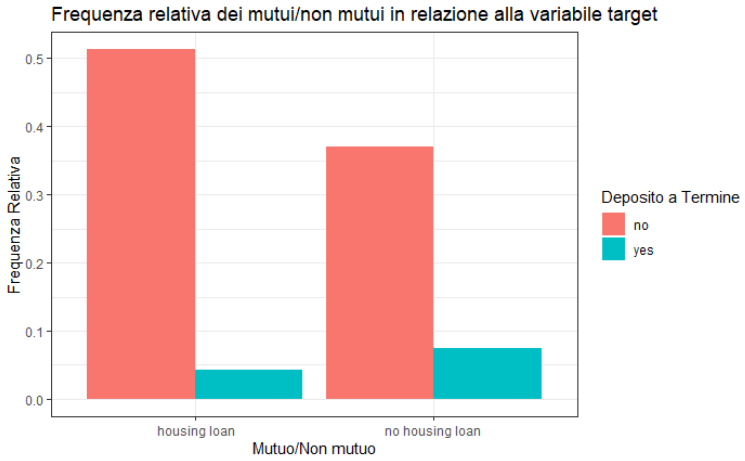

Figura 3.7: **Diagramma a barre della frequenza relativa dei mutui/non mutui in relazione alla variabile target**

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.

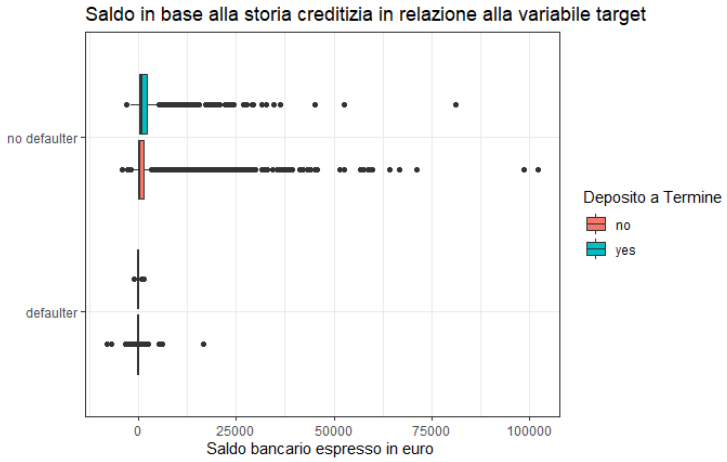

Figura 3.8: **Boxplot relativo al saldo in base alla storia creditizia in relazione alla variabile target**

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.

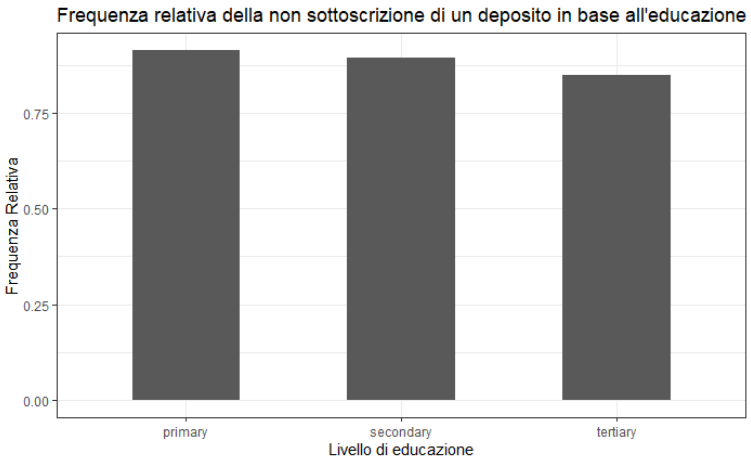

Figura 3.9: **Diagramma a barre della frequenza relativa della non sottoscrizione di un deposito in base al livello di educazione**

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).

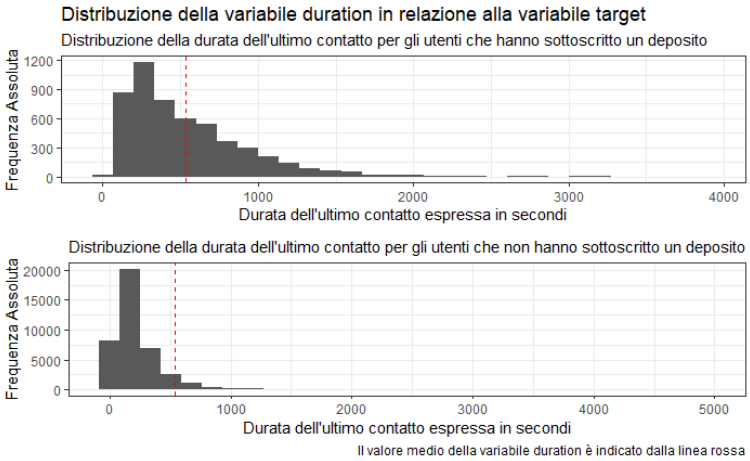

Figura 3.10: **Istogrammi della distribuzione della durata dell'ultimo contatto in relazione alla variabile target**

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".

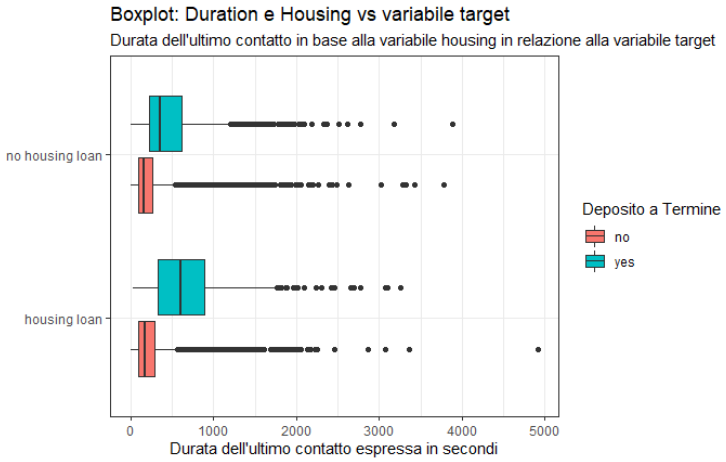

Figura 3.11: **Boxplot per la durata dell'ultimo contatto in base al fattore mutuo in relazione alla variabile target**

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.

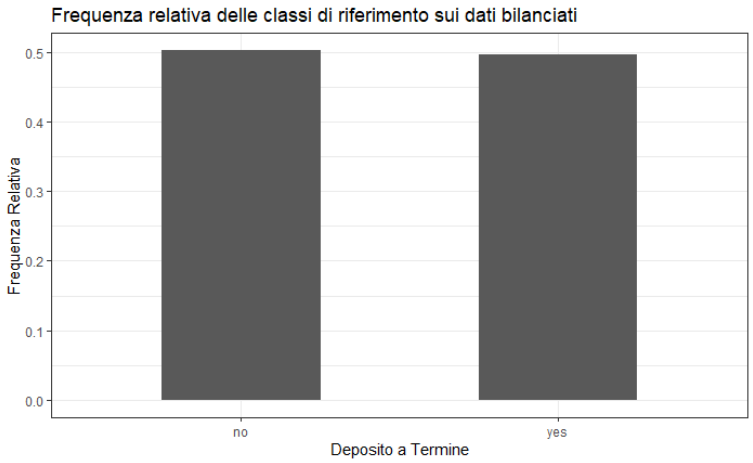

Figura 3.12: **Diagramma a barre della frequenza relativa delle classi di riferimento sui dati bilanciati appartenenti al training set**

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.

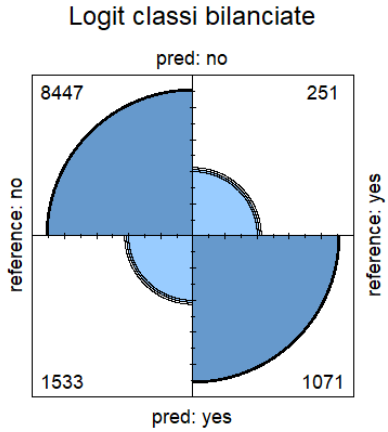

Figura 3.13: **Fourfold plots modello logit con classi bilanciate**

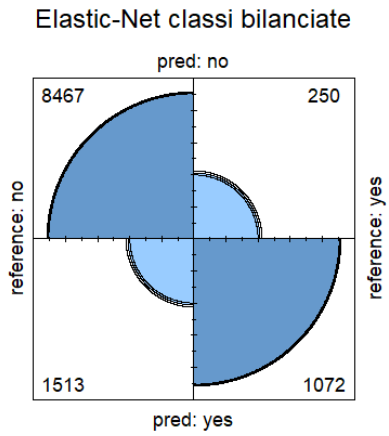

Figura 3.14: **Fourfold plots modello elastic net con classi bilanciate**

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.

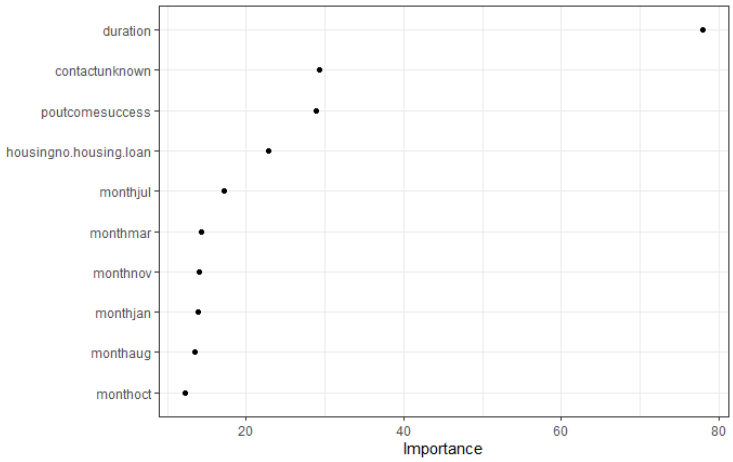

Figura 3.15: **Prime dieci variabili per importanza per il modello logit con classi bilanciate**

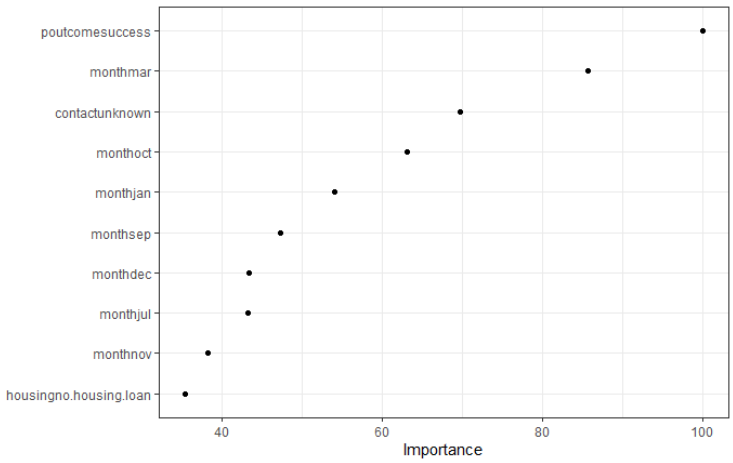

Figura 3.16: **Prime dieci variabili per importanza per il modello elastic net con classi bilanciate**

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 α 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.

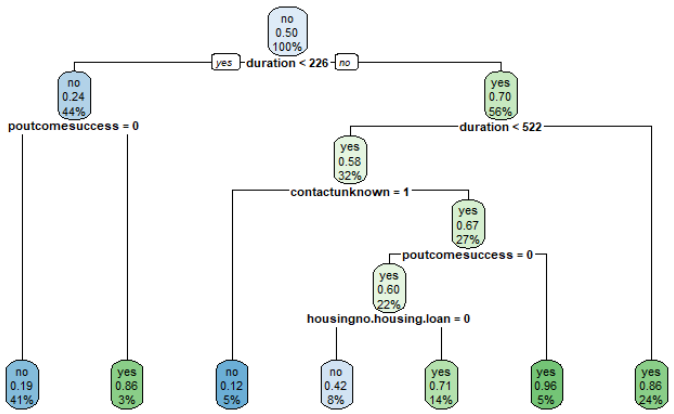

Figura 3.17: **Struttura dell'albero di decisione implementato sulle classi bilanciate**