ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI SCIENZE STATISTICHE “PAOLO FORTUNATI”

Corso di Laurea in

Finanza, Assicurazioni e Impresa

Previsione delle frodi finanziarie: Un confronto empirico tra alberi di classificazione e reti neurali

Relatore:

Prof.ssa Mariagiulia Matteucci

Presentata da:

Stefano Alberghini

Sessione 1/2024

Anno Accademico 2023/2024

«*Statistical thinking will one day be as necessary,*
for efficient citizenship as the ability to read and write.»

Samuel S. Wilks

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

In un mondo dominato dall'innovazione tecnologica e finanziaria, che ha indotto in modo esponenziale la crescita di pagamenti digitali, istituti bancari, operatori commerciali e individui si trovano di fronte ad una nuova sfida: come prevenire il rischio di frodi finanziarie? La problematica delle transazioni fraudolente, è diventato un serio ostacolo da affrontare, portando molteplici danni economici alle attività coinvolte. Secondo quanto riportato nel rapporto [1] della Banca Centrale Europea (BCE) di Maggio 2023, le perdite complessive dovute a transazioni fraudolente hanno raggiunto il livello più basso, ammontando a € 1.53 miliardi nel 2021, nell'area unica dei pagamenti in euro (SEPA). In Italia, secondo il rapporto di Dicembre 2023 [2] sui trend di mercato dell'osservatorio CRIF, l'importo medio è stato di € 4.845 nel primo semestre del 2023, in crescita del 3.1% rispetto all'anno precedente. Degli individui coinvolti, quelli appartenenti alla fascia d'età sotto i 30 anni, risultano essere i più colpiti.

Questo elaborato, cerca di approfondire la sfida della prevenzione delle frodi finanziarie, sfruttando le potenzialità degli strumenti statistici, per poter svelare schemi, anomalie e informazioni andando ad analizzare un set di dati contenente le informazioni di 1000 ipotetici consumatori che effettuano transazioni con 800 operatori commerciali. Nel corso di questo elaborato, esamineremo come sia possibile associare la conoscenza settoriale a decisioni basate sui dati al fine di migliorare il processo decisivo nell'ambito della rilevazione delle frodi. Considerando che ci confrontiamo con un problema di classificazione, esploreremo le prestazioni dell'albero di classificazione utilizzando la tecnica del pruning, un approccio noto per la sua semplicità, in contrasto con un metodo più avanzato come le reti neurali. L'obiettivo è valutare come questi due approcci si comportino nella risoluzione del problema delle transazioni fraudolente, tenendo conto delle differenze intrinseche tra un modello basato su alberi decisionali e un modello neurale, sia in termini di complessità che di capacità predittiva.

1. Dati e analisi esplorativa

1.1 Introduzione al set di dati

L'elemento principale di ogni analisi statistica è il set di dati, contenitore di molteplici informazioni nascoste, di grande importanza. Il set di dati in uso in questo elaborato è stato generato, utilizzando il generatore *Sparkov Data Generation* [3] di Brandon Harris, cercando di replicare comportamenti e caratteristiche reali. Il set di dati simulato metterà in evidenza la maggior parte delle sfide che gli analisti della rilevazione delle frodi affrontano utilizzando dati del mondo reale. In particolare, includeranno uno sbilanciamento delle classi, una combinazione di caratteristiche numeriche e categoriche (con caratteristiche categoriche che coinvolgono un numero molto elevato di valori), e relazioni tra le caratteristiche e scenari di frode dipendenti dal tempo. Il set di dati contenente le informazioni di 1000 consumatori che effettuano transazioni con 800 operatori commerciali, racchiude 25.000 transazioni simulate legittime e fraudolente, effettuate negli Stati Uniti con un intervallo temporale dal 1 Gennaio 2019 al 31 Dicembre 2020 tramite carta di credito. Inizialmente, come mostrato nella Tabella 1, sono presenti 9 variabili che presentano informazioni per ogni transazione effettuata. Infine, la variabile *is_fraud* sarà la nostra variabile obiettivo, e lo scopo principale di questo elaborato, sarà cercare di produrre previsioni in merito alla fraudolenza di transazioni basandoci su specifiche caratteristiche delle variabili in uso.

trans_date_trans_time	ora e giorno della transazione
category	categoria del negozio
amt	valore monetario della transazione
city	città del proprietario della carta di credito
state	stato del proprietario della carta di credito

	Tabella 1 continuo pagina precedente
city_pop	popolazione totale della città del proprietario della carta di credito
job	mestiere del proprietario della carta di credito
dob	data di nascita del proprietario della carta di credito
is_fraud	se la transazione è fraudolenta (1) o meno (0)

Tabella 1: Riepilogo variabili

1.2 Esplorando il set di dati: statistiche e grafici

Dopo aver fatto le opportune trasformazioni delle variabili in uso, possiamo andare a visualizzare la struttura del nostro set di dati, in particolare con la visuale illustrata nella Figura 1, come ci aspettavamo, le transazioni non fraudolente (17.494) risultano essere nettamente maggiori a quelle fraudolente (7.506).

Figura 1: Composizione dei dati

Iniziando l'esplorazione delle variabili, risulta evidente come il valore monetario delle transazioni contenute in questo set di dati, sia concentrato su importi relativamente bassi.

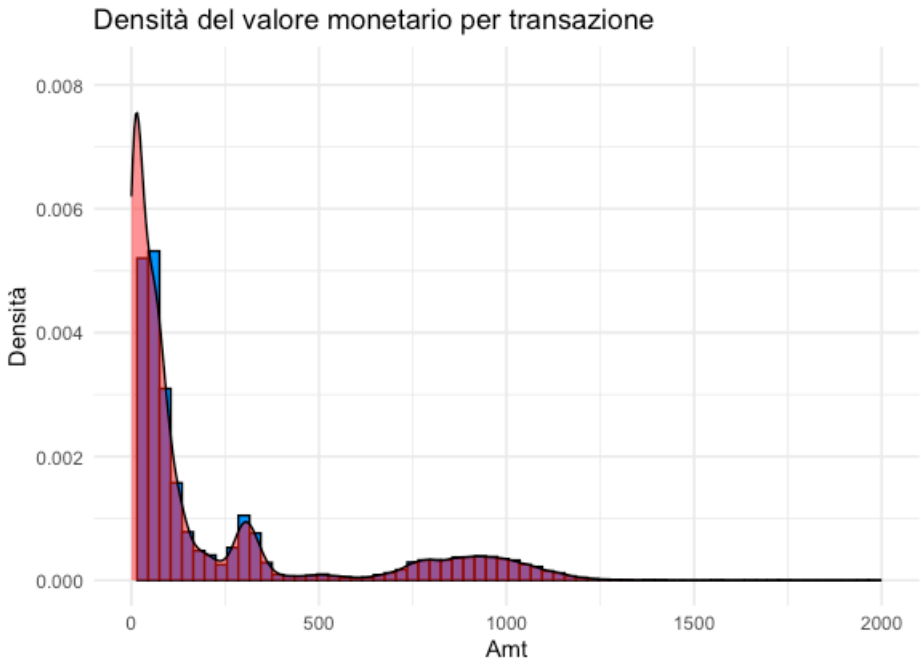

Figura 2: Densità del valore monetario

Nel processo di filtraggio delle transazioni fraudolente, emerge come l'importo medio è stato di $ 531,32 con un valore massimo di $ 1376.04, con una distanza tra l'importo minimo e l'importo massimo relativamente alta. Tuttavia, è interessante notare come la distanza tra la media e mediana non risulta essere particolarmente ampia, questo potrebbe essere un punto di partenza per la gestione e la rilevazione di valori anomali all'interno del nostro set di dati.

```
> summary(data_fraud$amt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.06  245.66  396.50  531.32  900.88 1376.04
```

Figura 3: Output di riepilogo transazioni fraudolente

Per l'utilizzo della variabile con l'albero di classificazione, andremo a creare dei gruppi in base alla grandezza monetaria della transazione effettuata, la nuova variabile sarà quindi chiamata *amt_category*, in Figura 4 è possibile visualizzare la suddivisione utilizzando i quartili.

```r
dati$amt_category <- cut(dati$amt,
                         breaks = quantile(dati$amt, probs = seq(0, 1, by = 1/4)),
                         labels = c("Molto Piccola", "Piccola", "Media", "Grande"),
                         include.lowest = TRUE)
```

Figura 4: Suddivisione per gruppi amt_category

Nel grafico successivo, sono rappresentate le frequenze relative a ciascun gruppo creato mediante l'implementazione della nuova variabile. Notiamo come i gruppi sono suddivisi ugualmente grazie alla suddivisione adottata.

Molto Piccola	Piccola	Media	Grande
0.25011	0.24991	0.24999	0.24999

Figura 5: Tabella di frequenza variabile amt_category

Esaminando la distribuzione delle transazioni fraudolente in relazione alla loro dimensione, emerge chiaramente che una considerevole percentuale di tali transazioni nel nostro set di dati è classificata come "Grande".

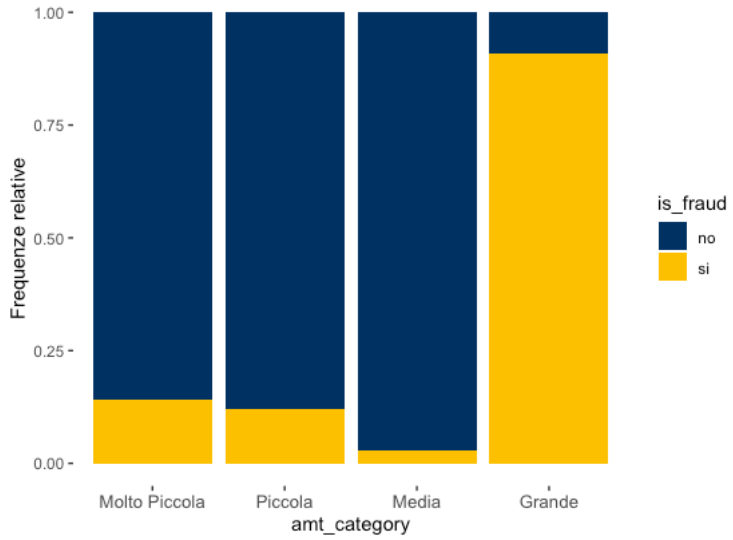

Figura 6: Relazione transazioni fraudolente con variabile variabile amt_category

Analizzando il test del chi-quadro in Figura 7, ci suggerisce data l'evidenza empirica del p-value, l'ipotesi di dipendenza con la variabile obiettivo.

In quanto nel nostro set di dati abbiamo la variabile *job* che contiene 494 lavori distinti, si è deciso per l'interpretazione successiva, la creazione di 9 raggruppamenti, per andare ad analizzare se la tipologia di lavoro può indicare una maggiore esposizione o meno. La nuova variabile sarà chiamata *jobs_group* che distingue la tipologia di lavoro come:

```
> chisq.test(dati$is_fraud, dati$amt_category)

        Pearson's Chi-squared test

data:  dati$is_fraud and dati$amt_category
X-squared = 10037, df = 2, p-value < 2.2e-16
```

Figura 7: Test di indipendenza per la variabile amt_category

Tecnico, Finanza, Assistenza sanitaria, Arte e ruoli creativi, Gestione e Servizi, Servizi e arte, Ingegneria e Amministrazione, Altro.

Altro	Arte e ruoli creativi	Assistenza Sanitaria
0.52132	0.00484	0.02400
Educazione	Finanza	Gestione e Servizi
0.03756	0.00784	0.04584
Ingegneria e Amministrazione	Servizi e arte	Tecnico
0.13852	0.09752	0.12256

Figura 8: Tabella di frequenza variabile jobs_group

Dall'analisi visiva riportata nella Figura 9, emerge chiaramente che le categorie di lavoratori più colpite sono quelle appartenenti ai settori dell'*arte e ruoli creativi*, dell'*assistenza sanitaria* e della *finanza*. Questa evidenza suggerisce un punto di partenza significativo per ulteriori indagini approfondite sulle cause che rendono queste specifiche categorie di lavoratori più vulnerabili alle frodi finanziarie.

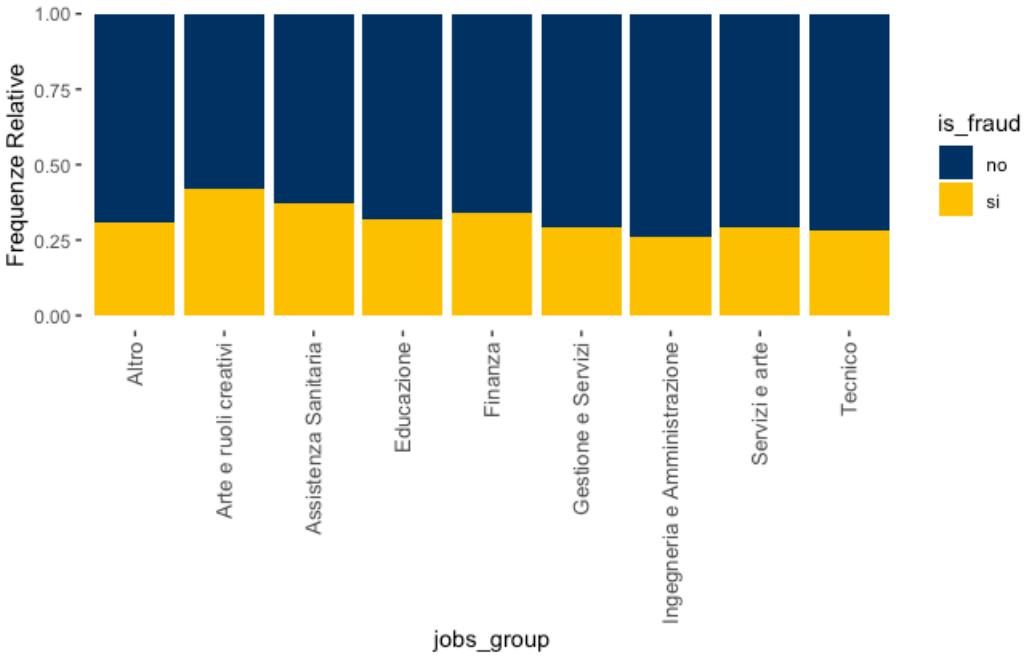

Figura 9: Relazione transazioni fraudolente con variabile jobs_group

Anche per questa variabile, in Figura 10 il test chi-quadro ci suggerisce la dipendenza con la variabile obiettivo.

```
> chisq.test(dati$is_fraud, dati$jobs_group)

        Pearson's Chi-squared test

data:  dati$is_fraud and dati$jobs_group
X-squared = 61.668, df = 8, p-value = 2.192e-10
```

Figura 10: Test di indipendenza variabile jobs_group

In seguito, analizzando la variabile *category* contenente la categoria del negozio dove è avvenuta la transazione, notiamo che diverse categorie sono denominate come "_pos, _net" questo ci suggerisce che sono avvenute con l'utilizzo di un terminale di pagamento (*POS*), quindi in un negozio fisico oppure in modalità online. Una nuova variabile sarà chiamata *categoria_trans* che distingue le transazioni avvenute via pos, online oppure diversamente (categoria di negozio dove non è stato specificato se online o fisico).

```
  altro internet      pos
0.49512  0.22076  0.28412
```

Figura 11: Tabella di frequenza variabile categoria_trans

Come previsto, l'analisi dei dati rivela che la categoria di transazioni più suscettibile a frodi è rappresentata da quelle effettuate tramite *internet*, rispetto alle categorie *pos* o *altro*. Questa osservazione conferma la tendenza comune in cui le transazioni online possono essere più esposte a rischi di frode rispetto a quelle condotte in modo fisico o attraverso altri mezzi.

```
> chisq.test(dati$is_fraud, dati$categoria_trans)

        Pearson's Chi-squared test

data:  dati$is_fraud and dati$categoria_trans
X-squared = 2642, df = 2, p-value < 2.2e-16
```

Figura 13: Test di indipendenza variabile categoria_trans

L'esame della dipendenza tra la variabile *categoria_trans* e la variabile dipendente in Figura 13 evidenzia una relazione significativa tra di esse.

Per quanto riguarda le variabili *city* e *city_pop*, possiamo decidere di creare una nuova variabile denominata *pop_group*, dove raggruppiamo le città in base alla dimensione

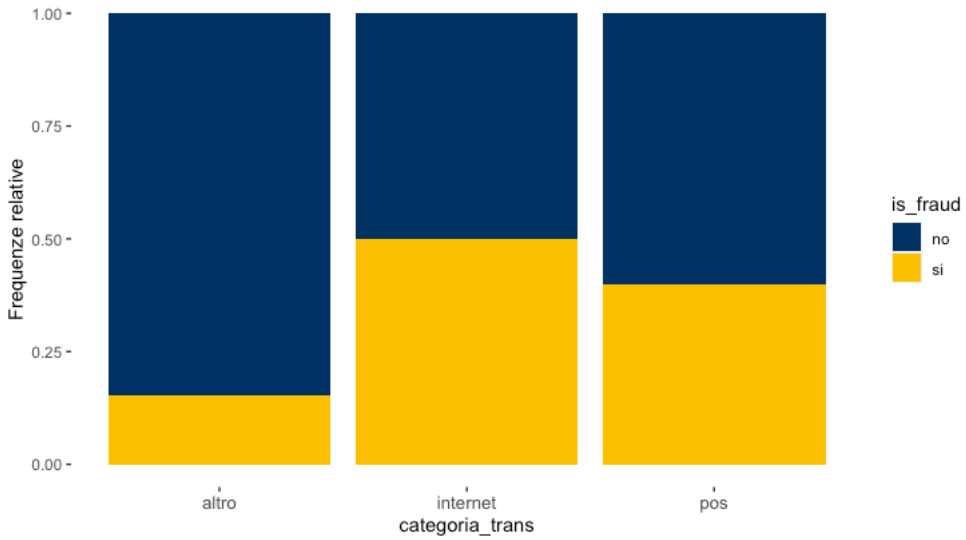

Figura 12: Relazione transazioni fraudolente con variabile categoria_trans

della popolazione. Il raggruppamento, verrà effettuato sulla base di una gerarchia di insediamento [4], definito dai gruppi seguenti: *Densità alta, Densità superiore, Densità medio-inferiore, Densità bassa, Densità molto bassa*.

```
soglie <- c(-Inf, 1000, 100000, 250000, 1000000, Inf)
etichette <- c("Densità molto bassa", "Densità bassa", "Densità medio-inferiore", "Densità superiore", "Densità Alta")
dati$pop_group <- cut(dati$city_pop, breaks = soglie, labels = labels)
```

Figura 14: Funzione di raggruppamento variabile pop_group

L'esame delle transazioni fraudolente in relazione alla dimensione della popolazione in cui sono state effettuate rivela in Figura 15 che, in una popolazione definita come *medio-inferiore*, la presenza di frodi risulta essere leggermente più elevata.

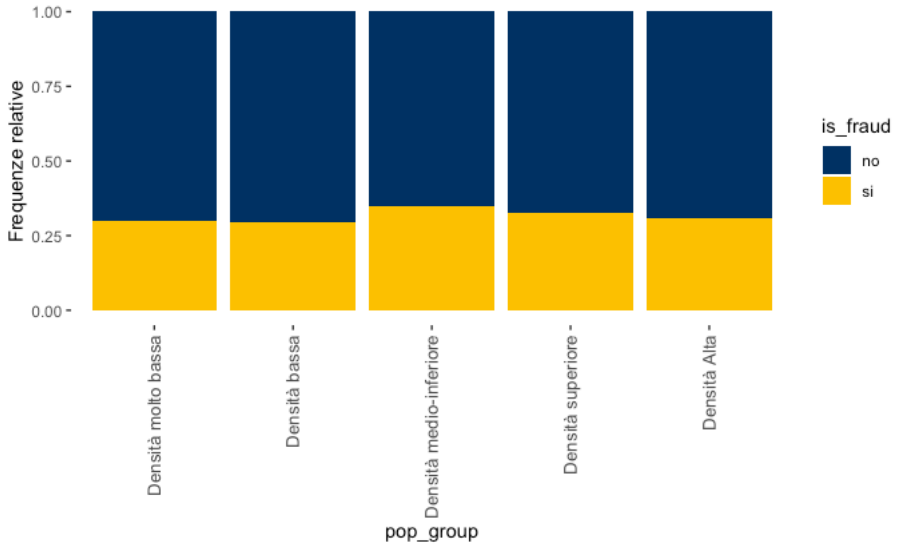

Figura 15: Relazione transazioni fraudolente con variabile pop_group

Analizzando il test chi-quadro in Figura 16, per valutare l'indipendenza, abbiamo per la variabile *pop_group* evidenza empirica di dipendenza con la nostra variabile obiettivo.

```
> chisq.test(dati$is_fraud, dati$pop_group)

        Pearson's Chi-squared test

data:  dati$is_fraud and dati$pop_group
X-squared = 24.446, df = 4, p-value = 6.5e-05
```

Figura 16: Test di indipendenza variabile pop_group

Successivamente, nel nostro set di dati, è presente la variabile *dob*, senza specificare accuratamente l'età dei soggetti. E' possibile calcolare l'età di ciascun individuo al 4 Febbraio 2024, e successivamente suddividerli in tre categorie distinte: giovani, adulti e anziani, creando una nuova variabile chiamata *age_category*. Un individuo verrà classificato come *giovane* se la sua età è minore di 35 anni, *adulto* se è compresa tra i 35 e 70 anni, ed infine *anziano* se è maggiore di 70 anni.

```
  adulto anziano giovane
 0.63324 0.14720 0.21956
```

Figura 17: Tabella di frequenza variabile age_category

L'analisi della distribuzione delle transazioni in base alle categorie di età, come mostrato nella Figura 17, evidenzia che la categoria definita come *adulta* è quella che registra la maggior parte delle transazioni nel nostro set di dati. Questa osservazione sottolinea l'importanza della fascia di età adulta come principale gruppo demografico coinvolto nelle attività di transazione.

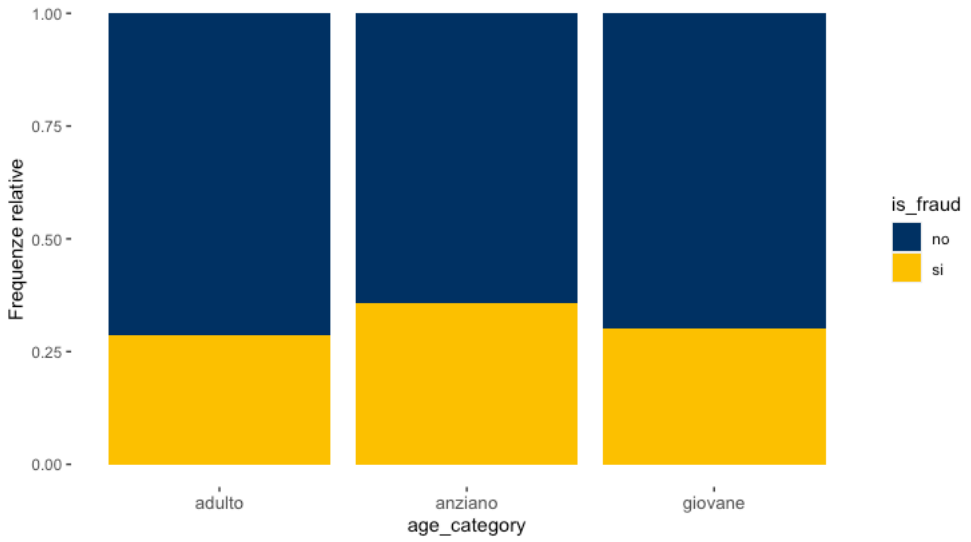

Figura 18: Relazione transazioni fraudolente con variabile age_category

L'analisi della relazione tra le transazioni fraudolente e le categorie di età indica che la categoria definita come *anziano* risulta essere la più suscettibile alle attività di frode. Questa osservazione suggerisce che gli individui anziani potrebbero essere più vulnerabili alle pratiche fraudolente, richiedendo una maggiore attenzione e misure preventive specifiche per proteggere questa fascia di utenti. Andando ad analizzare la dipendenza in Figura 19 con la variabile obiettivo, utilizzando il test del chi-quadro, abbiamo evidenza empirica dato il p-value per rifiutare l'ipotesi nulla di indipendenza al livello di significatività dell'1%. Quindi, la variabile *age_category* risulta essere associata alla variabile dipendente.

```
> chisq.test(dati$is_fraud, dati$age_category)

        Pearson's Chi-squared test

data:  dati$is_fraud and dati$age_category
X-squared = 79.846, df = 2, p-value < 2.2e-16
```

Figura 19: Test di indipendenza per la variabile age_category

Con la disposizione di informazioni sulla data e orario di ogni transazione effettuata, è possibile sfruttare tale informazione per analizzare in dettaglio le ore della giornata durante le quali si verificano più frequentemente transazioni non legittime, inoltre si è deciso la creazione di una nuova variabile *orario_bin*, allo scopo di categorizzare ogni

transazione come avvenuta nelle ore diurne (dopo le 06:00) o notturne (dopo le 18:00), per l'implementazione successiva nella fase di stima dei modelli successivi.

```
   giorno     notte
0.3769252 0.6230748
```

Figura 20: Tabella di frequenza variabile orario_bin

Analizzando la tabella di frequenza della variabile *orario_bin*, la maggioranza delle transazioni è stata effettuata in orari classificati come notturni. L'identificazione di questo modello temporale può essere significativa per la progettazione di strategie di prevenzione delle frodi. La maggior attività notturna potrebbe rappresentare un periodo critico in cui è necessario rafforzare le misure di sicurezza per mitigare i rischi di attività fraudolente. Le transazioni effettuate durante le ore notturne possono essere più difficili da monitorare e richiedere una particolare attenzione da parte dei sistemi di sicurezza finanziaria.

Confermando le aspettative, l'analisi dei dati in Figura 21 rivela che la maggioranza delle transazioni fraudolente nel nostro set di dati si verifica durante il periodo notturno. La concentrazione di transazioni fraudolente in questo periodo può essere attribuita a diversi fattori, per esempio, gli individui tendono ad essere meno disponibili, favorendo un maggior margine di tempo prima che l'individuo colpito si accorga dell'attività fraudolenta e adotti misure tempestive, come il blocco della carta di credito.

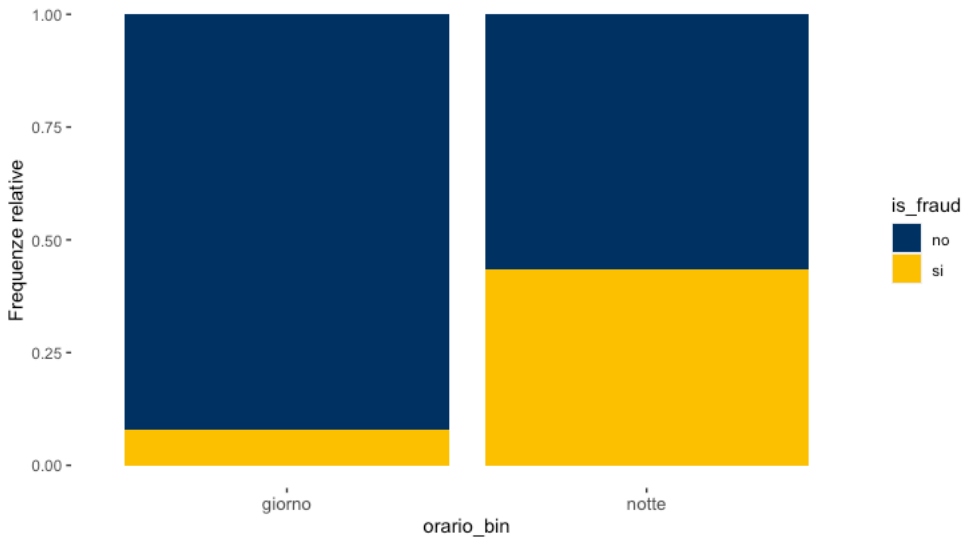

Figura 21: Relazione transazioni fraudolente con variabile orario_bin

Prendendo in riferimento il test del chi-quadro in Figura 22, anche in questo caso abbiamo evidenza empirica per rifiutare l'ipotesi nulla al livello di significatività dell'1%.

```
> chisq.test(dati$is_fraud,dati$orario_bin)

        Pearson's Chi-squared test with Yates' continuity correction

data:  dati$is_fraud and dati$orario_bin
X-squared = 3535.9, df = 1, p-value < 2.2e-16
```

Figura 22: Test di indipendenza per la variabile orario_bin

Un altro scenario interessante, è quello di analizzare la densità delle transazioni non legittime effettuate via internet, comunemente denominate *card-not-present* (CNP). Tale termine che si riferisce a tutte quelle situazioni dove la carta fisica non è presente al momento della transazione. Viceversa, il termine *card-present* (CP), si riferisce a tutte quelle transazioni avvenute utilizzando carta fisica durante l'atto effettivo [5].

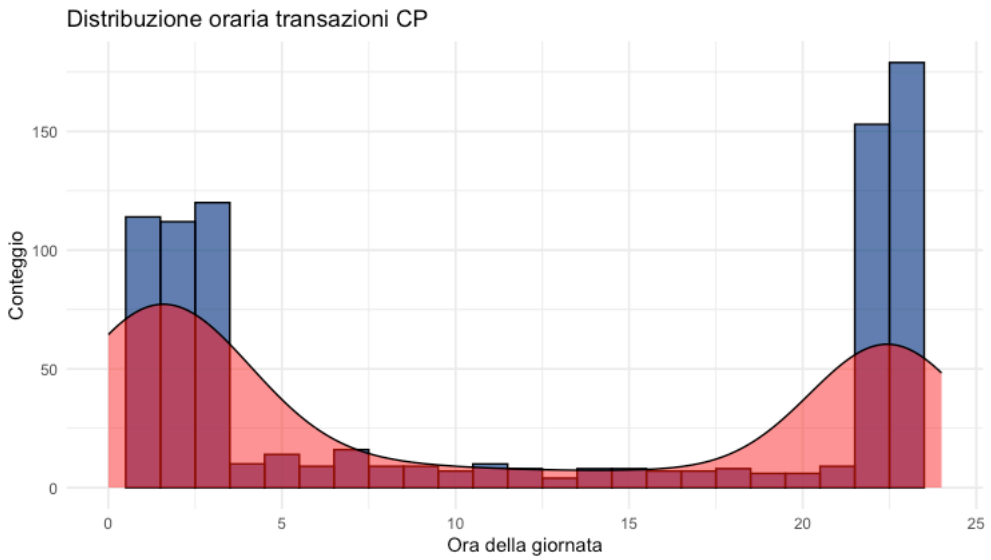

Figura 23: Densità transazioni fraudolente modalità CP

Sorprendentemente, Figura 23 rivela come le transazioni avvenute con l'utilizzo della carta fisica, sono avvenute con maggior frequenza nelle ore serali e notturne. Questo suggerisce la possibilità che tali transazioni siano avvenute in luoghi dove le attività commerciali operano 24 ore su 24 o tramite l'utilizzo di dispositivi POS in ambienti non fisicamente sorvegliati, come ad esempio le stazioni di rifornimento.

Per quanto riguarda le transazioni avvenute tramite la modalità CNP, l'andamento illustrato nel grafico in Figura 24 conferma le nostre aspettative, evidenziando maggiore frequenza nelle ore serali e notturne.

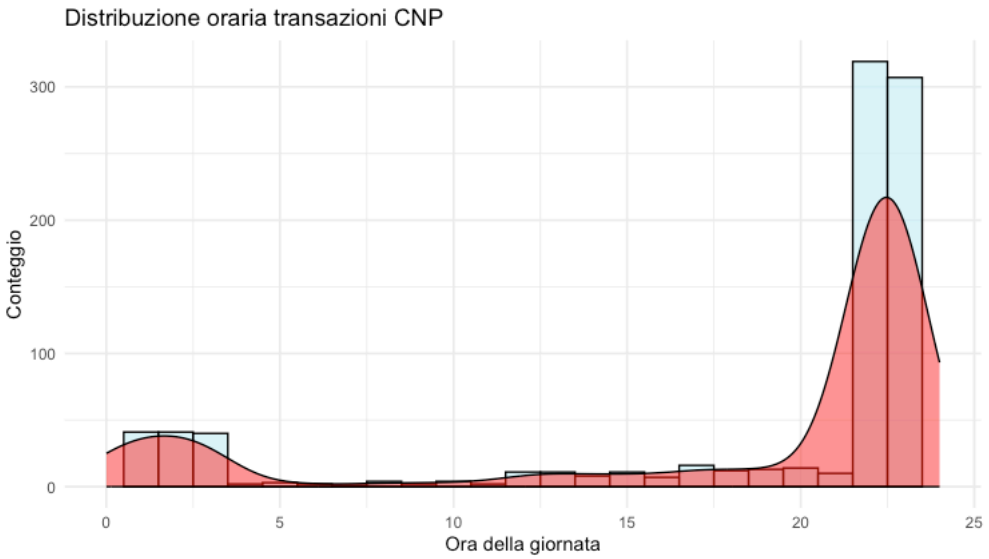

Figura 24: Densità transazioni fraudolente modalità CNP

Nelle figure successive, ci focalizziamo nelle transazioni fraudolenti avvenute nelle diverse categorie di negozi e nei diversi stati degli Stati Uniti. La categoria maggiormente colpita, con oltre il 5% di transazioni non legittime, risulta essere *grocery_pos*, cioè transazioni effettuate in un negozio alimentari tramite dispositivo POS, seguita da *shopping_net*, ovvero transazioni effettuate in rete. Questo suggerisce che i negozi appartenenti a queste due categorie, potrebbero essere caratterizzati da livelli di sicurezza relativamente bassi, specialmente nel caso della seconda categoria, dove la transazione non richiede la presenza fisica. Per quanto riguarda le categorie meno colpite, sembrano essere quelle caratterizzate da minor redditività o maggiore complessità, nel riciclare per denaro pulito, prodotti acquistati in modo non legittimo, come le categorie *food_dining, health_fitness, home, grocery_net, entertainment, kids_pets, personal_care, travel*, che rispettivamente indicano transazioni effettuate presso esercenti classificati come: ristorazione, vendita di prodotti o servizi salutari, articoli per la casa, spese online per generi alimentari, svago, prodotti per bambini o animali domestici, articoli per la cura personale e viaggi.

Nel complesso, analizzando le transazioni fraudolente per stato, emerge in Figura 26 che lo stato di Rhode Island è risultato essere il più frequentemente preso di mira, seguito dallo stato dell'Alaska. Questo risultato potrebbe apparire sorprendente, tuttavia è importante sottolineare che i criminali spesso adottano strategie per mascherare la loro posizione quando commettono attività illecite. Inoltre, è visibile un valore anomalo, in

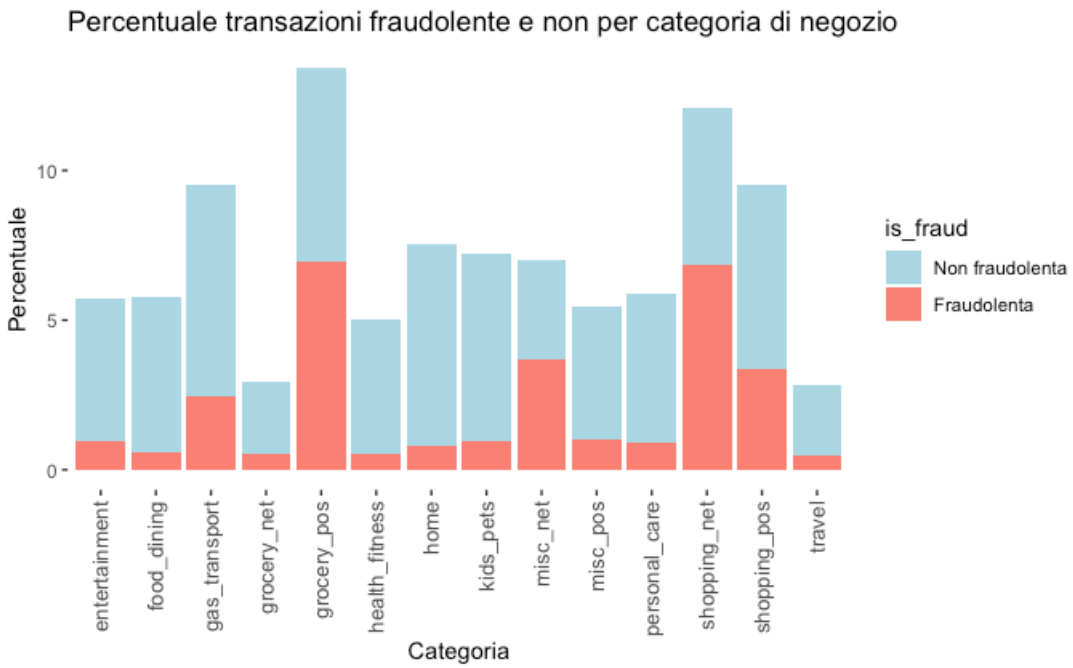

Figura 25: Percentuale transazioni fraudolente e non per categoria di negozio

quanto il set di dati, contiene solo 9 transazioni dello stato del Delaware, risultanti tutte fraudolente.

L'analisi dettagliata delle transazioni fraudolente a livello statale fornisce una panoramica più approfondita delle tendenze e delle vulnerabilità, consentendo ulteriori riflessioni su come affrontare e mitigare il rischio di frodi in contesti specifici.

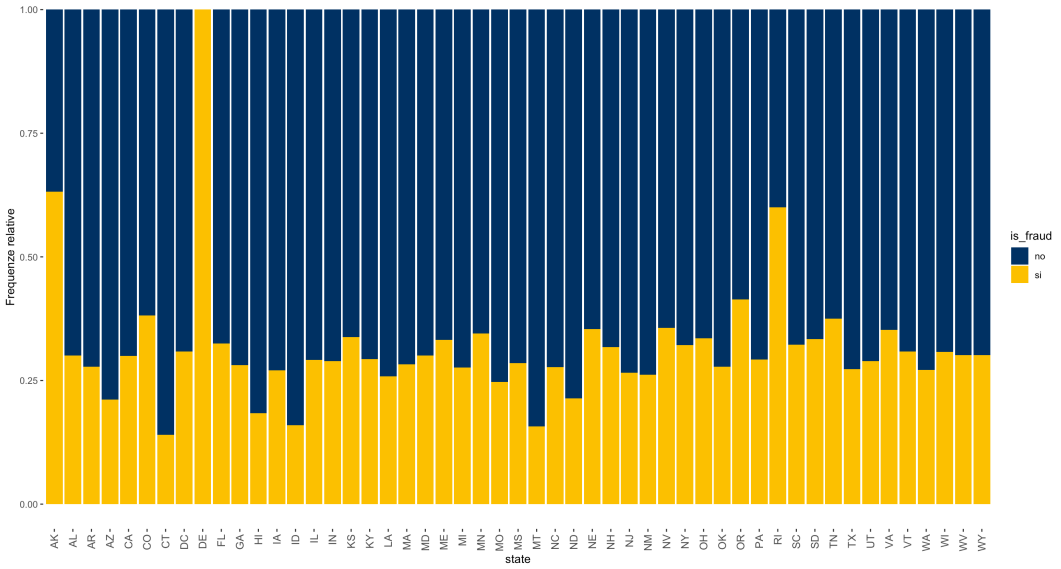

Figura 26: Transazioni fraudolente per stato

Per l'utilizzo della variabile nei nostri modelli successivi, si è deciso di raggruppare la

variabile *state* in base alla posizione geografica (Est, Nord, Sud, Ovest), così da poter facilitare l'interpretazione successiva.

Attraverso il nostro raggruppamento, è evidente che negli Stati del sud degli Stati Uniti si verificano un numero maggiore di transazioni. La concentrazione di transazioni negli Stati meridionali può essere influenzata da diversi fattori, tra cui le abitudini di spesa, la densità della popolazione, o anche specificità economiche e commerciali della zona.

```
      Est      Nord     Ovest       Sud
0.1958635 0.2819938 0.1464176 0.3757251
```

Figura 27: Tabella di frequenza variabile region

L'analisi della relazione delle transazioni fraudolente in base alla posizione geografica degli Stati Uniti non fornisce informazioni significative. Tale relazione sembra essere simile indipendentemente dalla posizione geografica. Questo risultato potrebbe suggerire che la distribuzione delle transazioni fraudolente non è strettamente legata alla collocazione geografica negli Stati Uniti. Al contrario, potrebbero essere altri fattori, come le caratteristiche demografiche, economiche o comportamentali degli utenti, a influenzare in modo più rilevante la presenza di attività fraudolente.

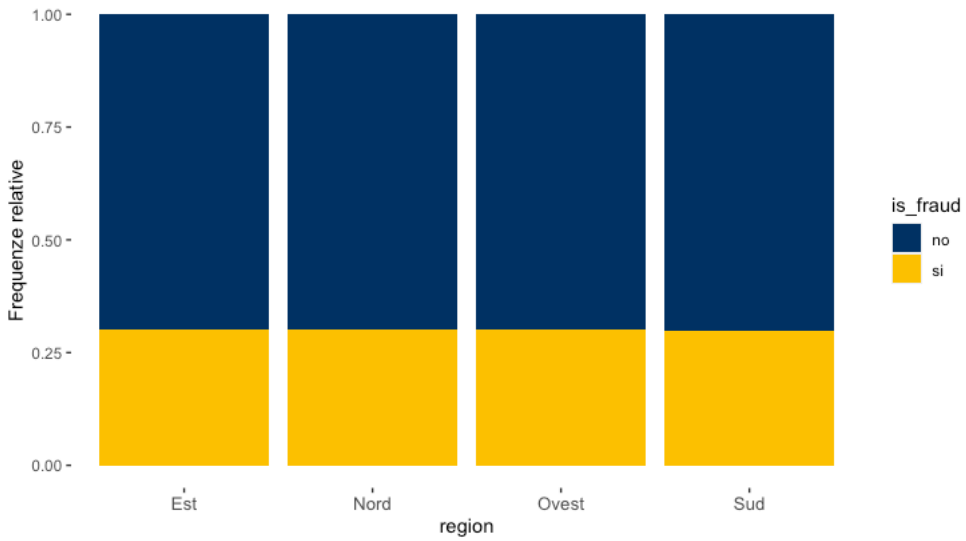

Figura 28: Relazione transazioni fraudolente con variabile region

Infatti, per la variabile *region*, il test del chi-quadro, risulta non rifiutare l'ipotesi nulla di indipendenza con la nostra variabile dipendente.

Infine, il set di dati viene suddiviso in *training set* e *test set*, con l'utilizzo della libreria R *caret*, che rispettivamente indicano il set di dati per l'addestramento dei modelli, e per

```
> chisq.test(dati$is_fraud, dati$region)

        Pearson's Chi-squared test

data:  dati$is_fraud and dati$region
X-squared = 0.40912, df = 3, p-value = 0.9384
```

Figura 29: Test di indipendenza per la variabile region

valutare la capacità predittiva del modello stimato. Per fare questo, utilizziamo il metodo *hold-out*, cioè il set di dati viene suddiviso in 75% delle unità nel set per l'addestramento, e il restante 25% delle unità per la valutazione del modello.

2. Alberi di classificazione

2.1 Introduzione e concetti fondamentali

Gli alberi di classificazione fanno parte delle tecniche di partizione ricorsiva CART (*Classification and Regression Trees*), che consentono di determinare l'appartenenza di ciascuna unità statistica alle classi definite dalla variabile obiettivo, nel nostro caso *is_fraud*. Questo avviene attraverso una suddivisione progressiva del collettivo in gruppi via via omogenei rispetto alla variabile obiettivo, e tale suddivisione si basa su ciascuna covariata. Gli alberi di classificazione forniscono decisioni strutturate a cascata, dove ogni nodo rappresenta una decisione basata su una covariata specifica. Questo approccio a struttura gerarchica consente di interpretare facilmente le decisioni del modello e di identificare i fattori che contribuiscono alla previsione della variabile obiettivo.

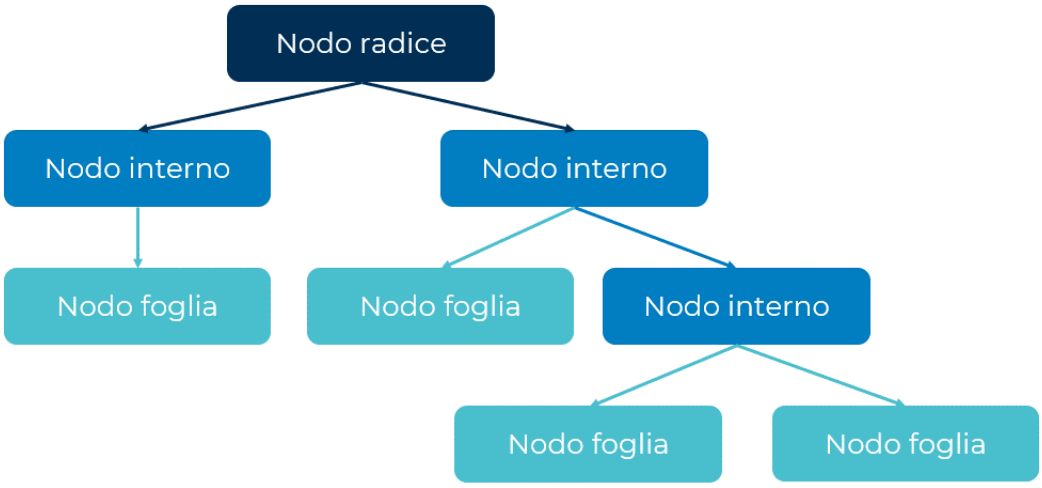

Figura 30: Architettura alberi di classificazione [6]

La segmentazione gerarchica mediante alberi di classificazione avviene in modo sequenziale. Durante il primo passo, l'insieme di *n* unità viene suddiviso in due o più sottoinsiemi disgiunti, definiti dalle modalità nello spazio delle covariate. I passi successivi si realizzano suddividendo ulteriormente i sottoinsiemi ottenuti al passo precedente.

Questo processo continua iterativamente fino a quando viene raggiunto un criterio di arresto predefinito, come ad esempio una profondità massima dell'albero o un numero minimo di osservazioni in ciascun nodo terminale. Questa sequenza di suddivisioni gerarchiche consente di creare una struttura ad albero che riflette le relazioni tra le covariate e la variabile obiettivo. Durante ciascun passo del processo di segmentazione, l'obiettivo è massimizzare l'omogeneità all'interno dei sottoinsiemi risultanti o massimizzare l'eterogeneità tra di essi. Questo viene valutato in base alla variabile obiettivo o ad altre misure di similarità. Un criterio comune utilizzato per questo scopo è la riduzione dell'impurità, dove si cerca di minimizzare la miscela delle classi all'interno di ciascun sottoinsieme. Per esempio, in un contesto binario come la previsione delle frodi, si potrebbe utilizzare l'indice di Gini o l'entropia come misura di impurità.

L'indice di eterogeneità di Gini per una variabile dipendente Y con J categorie, sia $p(j|t)$ la frequenza relativa della j-esima categoria nel t-esimo nodo, è composto come:

$$I(t) = 1 - \sum_{j=1}^{J} p(j|t)^2$$

Durante la costruzione dell'albero, la suddivisione viene scelta in modo che la riduzione dell'impurità sia massima. La misura del decremento di impurità del nodo t associata alla suddivisione s è definita come:

$$\Delta I(s,t) = I(t) - [I(t_l)]p_l - [I(t_r)]p_r$$

Dove p_l e p_r rappresentano la proporzione dei casi del nodo t che cadono, rispettivamente, nel nodo di sinistra (*left*) e nel nodo di destra (*right*).

La miglior scelta di suddivisione per un dato nodo t, sarà lo split s^* che produce la massima riduzione di impurità dell'albero.

$$\Delta I(s^*,t) = max\Delta I(s,t)$$

L'assenza di regole di stop nella segmentazione di un albero di classificazione può portare a una crescita eccessiva dell'albero, con nodi terminali che contengono solo casi appartenenti alla stessa classe della variabile dipendente. Per evitare questo fenomeno, vengono elencati successivamente alcuni classici criteri di arresto:

- **Dimensione minima del gruppo**: Si impone una soglia minima al numero di osservazioni necessarie in ciascun gruppo affinché le stime basate su tali gruppi siano affidabili e significative dal punto di vista statistico.
- **Minima disomogeneità del gruppo genitore**: Si basa sull'idea che un gruppo genitore, o nodo, con una bassa misura di disomogeneità sia considerato sufficientemente compatto o "puro", e quindi non più divisibile.
- **Minimo della capacità esplicativa della migliore suddivisione ad ogni passo**: Questo criterio richiede che, ad ogni passo della procedura di suddivisione, l'aggiunta di una nuova segmentazione debba portare a una significativa spiegazione della variazione o dell'eterogeneità all'interno dei gruppi, rispetto alla variabile dipendente. Se una potenziale segmentazione non contribuisce in modo sostanziale a spiegare questa disomogeneità, allora la procedura di costruzione dell'albero si interrompe.

2.2 La tecnica del pruning

La tecnica del pruning (potatura) è un'alternativa alle regole di arresto classiche nella costruzione degli alberi di classificazione. La procedura di potatura mira a semplificare un albero precedentemente costruito, rimuovendo alcuni dei suoi rami o nodi. L'obiettivo è evitare l'eccessiva complessità dell'albero e migliorare la sua capacità di generalizzazione su nuovi dati.
La procedura di potatura avviene in diverse fasi:

- **Costruzione dell'Albero**: Si costruisce un albero completo sulla base dei dati di addestramento.
- **Potatura**: Si rimuovono alcuni rami o nodi dall'albero completo, producendo una versione più semplificata.
- **Valutazione**: La versione potata dell'albero è quella che fornisce la stima del tasso

di errata classificazione migliore, tenendo conto della dimensione dell'albero di classificazione.

In pratica, viene utilizzata la tecnica del pruning selettivo che permette di identificare una sequenza ottimale di sottoalberi di dimensione decrescente. La sequenza ottimale di sottoalberi, viene identificata utilizzando una funzione di costo-complessità.

$$C_\alpha(T) = \hat{R}(T) + \alpha|\tilde{T}|$$

dove $\hat{R}(T)$ è la stima del tasso di errata classificazione tramite il metodo di risostituzione, α è il parametro di penalizzazione per gli alberi di grande dimensione, e $\tilde{T}$ è il numero di nodi terminali dell'albero. Fissato il valore α, vogliamo quello che minimizza la funzione di costo-complessità, ed è dimostrato che per ogni α, la sequenza di alberi ottimali viene identificata in modo univoco.

Poiché il numero di possibili sottoalberi può essere elevato, il pruning selettivo si concentra sull'individuare una sequenza di sottoalberi che massimizzano le prestazioni del modello, spostandosi gradualmente verso una struttura più semplice. I sottoalberi appartenenti alla sequenza ottimale vengono confrontati utilizzando una stima del tasso di errore di classificazione. Il sottoalbero ideale è quello per il quale la stima del tasso di errore di classificazione è più bassa. Questo criterio consente di identificare la struttura dell'albero che offre la migliore capacità predittiva, mantenendo contemporaneamente una complessità adeguata.

2.3 Applicazione alla rilevazione delle transazioni fraudolente

Per la ricerca sulla previsione delle transazioni fraudolente, ci avvarremo del metodo degli alberi di classificazione all'interno del nostro set di dati. Per implementare questo approccio, faremo uso delle librerie R *rpart* e *rpart.plot*. L'utilizzo di queste librerie non solo ci permetterà di condurre un'analisi efficace, ma anche di migliorare la presentazione visiva dell'albero di classificazione risultante, rendendo più accessibili e comprensibili i risultati ottenuti.

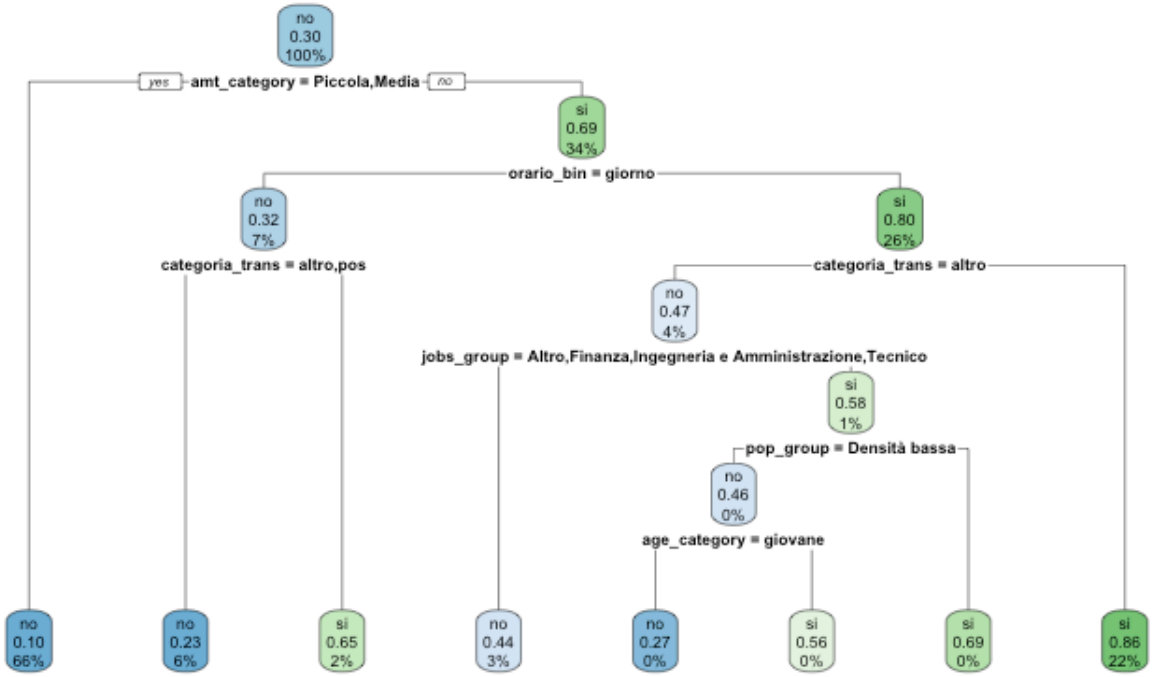

Figura 31: Classification Tree

Il primo albero di classificazione, rappresentato nella Figura 31, è stato generato utilizzando i criteri predefiniti della funzione *rpart*, visualizzabili attraverso *rpart.control()* con l'unica eccezione del parametro di complessità *cp*, impostato ad un valore uguale a 0.001 e utilizzando il set di dati di addestramento. Tale approccio ha prodotto un albero con otto nodi foglia, l'interpretazione dell'albero poco robusta data la presenza di nodi senza osservazioni presenti, come è possibile visualizzare nei tre nodi terminali con una percentuale approssimativamente uguale allo 0%. Inoltre, con l'utilizzo degli alberi di classificazione, il modello sceglierà automaticamente le variabili più importanti, rendendolo anche un modello in grado di aiutare con la scelta delle variabili più significative. La mancanza di restrizioni durante la creazione dell'albero può portare a una struttura dettagliata e intricata, che può risultare poco intuitiva per l'analisi visiva. Nelle prossime fasi della nostra analisi, esploreremo la strategia del pruning al fine di semplificare la struttura dell'albero e migliorare la sua interpretabilità, senza compromettere eccessivamente le prestazioni del modello nella predizione delle transazioni fraudolente.

La Figura 32 ci mostra la possibilità di confrontare utilizzando il metodo della validazione incrociata in relazione a ciascuna dimensione dell'albero, grazie all'utilizzo della tecnica del pruning selettivo. Questa metodologia consente di valutare l'impatto delle decisioni di

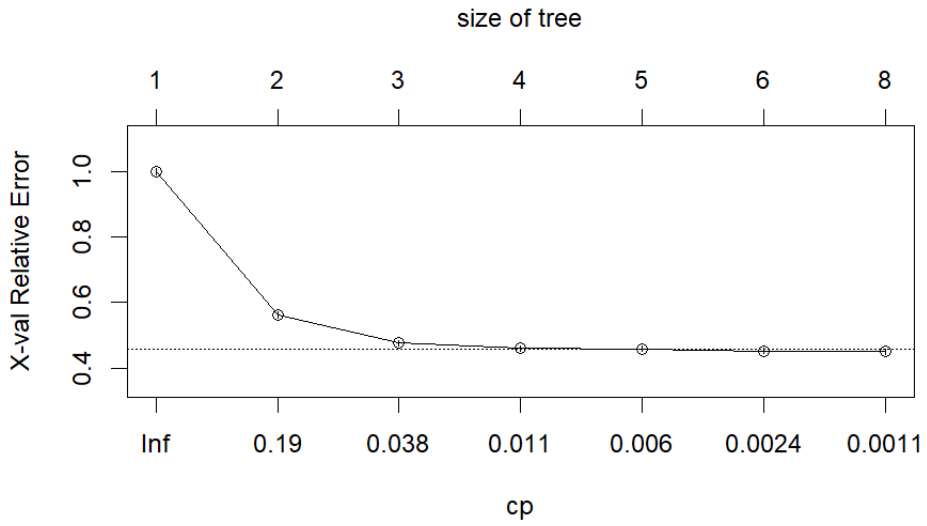

Figura 32: Tecnica del pruning

pruning su diverse parti dell'albero di classificazione. Esaminare l'errore relativo in questo contesto ci fornirà informazioni sulla trade-off tra la complessità dell'albero e la sua capacità predittiva. L'albero di classificazione dalla Figura 31 con 8 nodi terminali, ha un errore relativo pari a 0,4 circa. In aggiunta, l'albero con 5 nodi di classificazione mostra un errore relativo molto vicino a quello dell'albero senza pruning. Questa osservazione suggerisce che l'opzione di pruning potrebbe rappresentare un vantaggio significativo per migliorare l'interpretabilità dell'albero, riducendo la complessità senza sacrificare notevolmente le prestazioni predittive. Esplorare ulteriormente questa opzione potrebbe consentirci di ottenere un modello più chiaro e comprensibile senza compromettere eccessivamente l'efficienza della previsione. E' importante anche sottolineare, che un albero con pochi nodi terminali potrebbe non risultare utile per il nostro obiettivo di previsione. Per questo motivo, poichè l'errore di classificazione sembra essere molto vicino a l'albero con 5 nodi terminali, andremo ad optare per questa opzione. Complessivamente, questa scelta soggettiva mira a massimizzare la completezza e la comprensibilità del modello di classificazione, contribuendo così a una migliore interpretazione e utilizzo delle informazioni ottenute attraverso l'analisi dei dati.

Dalla Figura 33, è evidente come la riduzione della dimensione dell'albero attraverso la tecnica di pruning abbia notevolmente migliorato l'interpretabilità complessiva. In particolare, va notato che all'interno di ciascun nodo, è visualizzata la probabilità associata all'unità che ha raggiunto quel punto specifico dell'albero, indicando la quota di transazioni fraudolente. L'intensità del colore utilizzato nella rappresentazione grafica contribuisce

a fornire un'indicazione visiva immediata di questa probabilità.

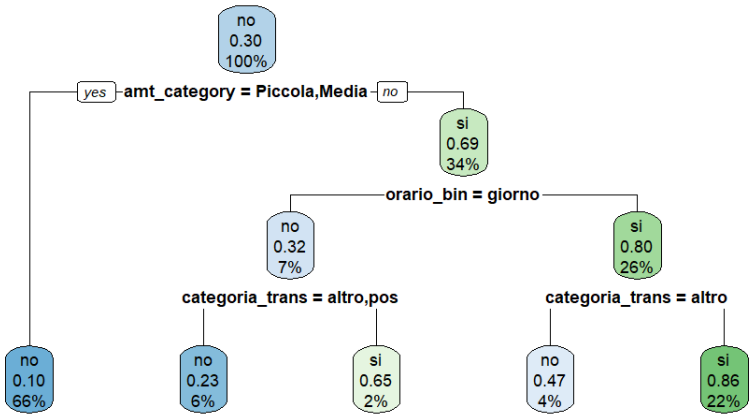

Figura 33: Albero di classificazione potato

Nella Tabella 2, per ogni nodo terminale, partendo da sinistra, viene descritto la sua sequenza degli split secondo lo spazio delle covariate, le unità classificate erroneamente, la previsione delle unità in quel nodo terminale ed infine la probabilità associata alle unità di essere una transazione fraudolente.

Sequenza degli split	Unità presenti	Unità classificate erroneamente	Probabilità transazione fraudolente	Previsione
amt_category = piccola, media	12.431 (66%)	1.239	0,1	no
amt_category = grande orario_bin = giorno categoria_trans = altro, pos	1.072 (6%)	247	0,23	no
amt_category = grande orario_bin = giorno categoria_trans = internet	298 (2%)	103	0,65	si
amt_category = grande orario_bin = notte categoria_trans = altro	768 (4%)	363	0,47	no
amt_category = grande orario_bin = notte categoria_trans = internet, pos	4.179 (22%)	594	0,86	si

Tabella 2: Caratteristiche nodi terminali albero potato

Per valutare la capacità predittiva dell'albero di classificazione, procederemo all'analisi mediante l'utilizzo della matrice di confusione. Questo strumento ci permetterà di valutare le performance del nostro modello nelle previsioni di transazioni fraudolente e non

fraudolente, consentendoci di calcolare metriche come la precisione, la specificità e la sensitività. L'esame della matrice di confusione fornirà un quadro completo e dettagliato delle prestazioni del nostro modello, agevolando la valutazione critica e la comprensione delle sue abilità predittive. In particolare, nel caso della prevenzione delle frodi, la metrica che ci interessa di più massimizzare risulta essere la sensitività [7]. Questo perchè, in un caso particolare come la rilevazione delle frodi, il mancato rilevamento di una transazione fraudolenta può comportare perdite finanziarie significative. L'albero di classificazione ottenuto tramite la tecnica del pruning, sembra essere un modello rilevante, evidenziando un tasso di corretta classificazione che si attesta intorno al 86%, con una sensitività pari circa al 65%. Questo risultato rappresenta un indicatore positivo della robustezza del nostro modello nella predizione delle transazioni fraudolente e non fraudolente.

```
> tab2
     prev
oss     no   si
  no  4150  223
  si   658 1218
> # precisione
> 1218 / (1218 + 223) * 100
[1] 84.52464
> # specificità
> 4150 / (4150 + 658) * 100
[1] 86.31448
> # sensitività
> 1218 / (1218 + 658) * 100
[1] 64.92537
```

Figura 34: Matrice di confusione e metriche albero pruning

Infine, analizzando la curva ROC (*Receiver Operating Characteristic*) in Figura 35 dell'albero di classificazione si avrà una misura grafica dell'accuratezza predittiva complessiva. Questa curva si basa sulla variazione dei livelli di cut-off nella matrice di confusione, consentendo di valutare le performance del modello attraverso la sensibilità e la specificità a diversi punti operativi. In altre parole, la curva ROC rappresenta il trade-off tra la capacità di un modello di classificare correttamente le transazioni fraudolente e la minimizzazione degli errori sui casi non fraudolenti, fornendo così una panoramica completa della sua efficienza predittiva in diverse condizioni operative. Per avere una valutazione quantitativa, si utilizza l'indicatore AUC (*Area Under the Curve*), che quantifica numericamente l'efficienza del modello. Un'area maggiore sotto la curva indica una

migliore capacità di discriminazione tra transazioni fraudolente e non fraudolente. Nel nostro modello, si ha un AUC intorno al 85%, cioè una prestazione complessiva ottima.

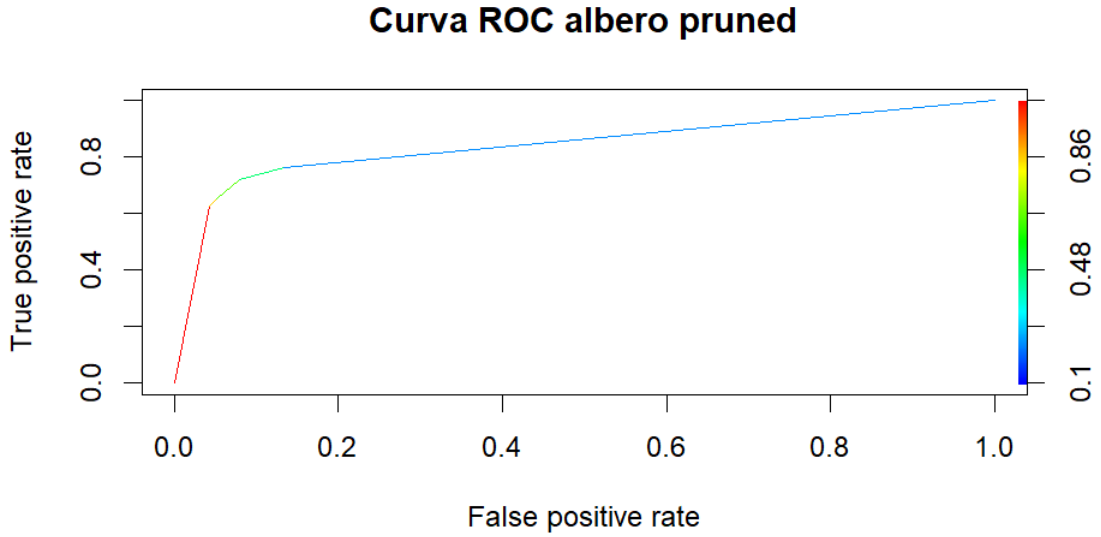

Figura 35: Curva ROC albero pruning

3. Reti Neurali

3.1 Introduzione alle reti neurali

Le reti neurali artificiali sono modelli statistici che insegnano al computer come elaborare dati, prendendo ispirazione dal funzionamento delle reti neurali biologiche e da come queste ultime elaborano le informazioni. Questi modelli sono parte del campo dell'apprendimento automatico (*machine learning*) e sono progettati per riconoscere modelli, fare previsioni e prendere decisioni basandosi su dati di input. Le reti neurali artificiali sono composte da strati di nodi (neuroni) interconnessi, ognuno con un peso associato, che vengono adattati durante il processo di apprendimento. Questi modelli sono ampiamente utilizzati in una varietà di applicazioni, inclusa la classificazione di immagini, il riconoscimento del linguaggio naturale e la previsione di serie temporali [8].
L'architettura di una rete neurale, comunemente è composta da tre livelli principali:

- ***Input Layer*** (Strato di ingresso): Questo è il primo livello della rete neurale, dove vengono inseriti i dati in ingresso al modello. Ogni nodo in questo layer rappresenta una caratteristica o una variabile di input. Ad esempio, in una rete neurale per il riconoscimento di immagini, ogni nodo potrebbe rappresentare un singolo pixel o una caratteristica specifica dell'immagine.
- ***Hidden Layer(s)*** (Strati nascosti): Questo è il cuore della rete neurale, dove avviene il processo di apprendimento. Ogni nodo in questi strati nascosti è collegato a tutti i nodi nel layer precedente e successivo. Gli strati nascosti sono responsabili di apprendere rappresentazioni più complesse e astratte dei dati di input attraverso l'elaborazione di pesi associati alle connessioni tra i neuroni. Una rete neurale può avere più di uno strato nascosto, a seconda della complessità del problema.
- ***Output Layer*** (Strato di uscita): Questo è l'ultimo livello della rete neurale, dove vengono prodotti i risultati o le predizioni. Il numero di nodi in questo layer dipende dalla natura del problema. Ad esempio, in un problema di classificazione binaria, come nel caso della prevenzione delle frodi, potrebbe esserci un solo nodo

che rappresenta l'output desiderato, mentre in una classificazione multiclasse ci sarebbe un nodo per ogni classe.

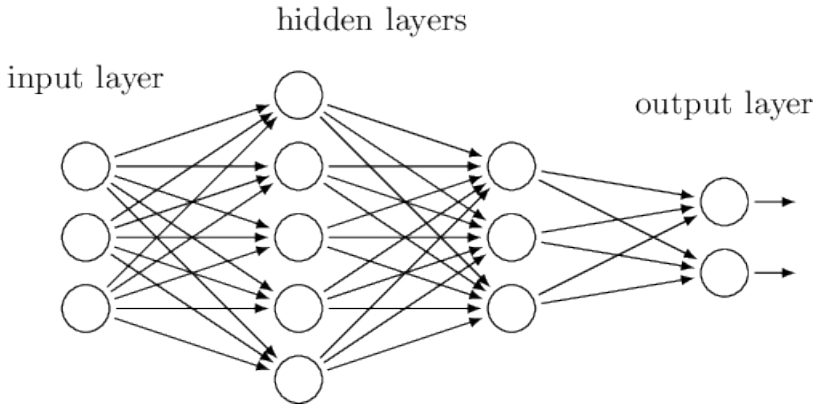

Figura 36: Architettura di una rete neurale

In termini di connessioni, ogni connessione tra neuroni ha un peso associato che viene adattato durante il processo di addestramento della rete neurale. Il flusso delle informazioni avviene dall'input layer, attraverso gli strati nascosti, fino all'output layer. Il processo di apprendimento avviene attraverso l'ottimizzazione dei pesi delle connessioni in modo che la rete neurale produca risultati desiderati per un determinato input. La stima dei parametri di una rete neurale coinvolge l'uso di algoritmi iterativi di minimizzazione numerica. L'obiettivo è trovare i valori ottimali dei pesi delle connessioni che minimizzano una funzione di perdita specifica del problema. Questa fase di ottimizzazione è fondamentale per l'addestramento della rete neurale e si basa su un processo iterativo che regola gradualmente il numero di neuroni *M* dello strato nascosto. Solitamente, è utilizzata la tecnica del *weigth decay*, cioè si ricerca un minimo vincolato che dipende da un parametro λ che regola la penalizzazione dell'entità delle stime dei parametri.

3.2 Applicazione alla rilevazione delle transazioni fraudolente

In questo elaborato, per l'implementazione pratica di una rete neurale con l'obiettivo di prevedere transazioni fraudolente, faremo uso delle librerie R *nnet* e *NeuralNetTools*.

A causa delle risorse computazionali limitate, adotteremo la tecnica del weight decay con un numero massimo di 5 neuroni (*M*) e la combineremo con i parametri λ di 0.0001, 0.001, e 0.01. Si procede con la valutazione delle diverse combinazioni per identificare quella che produce l'errore di previsione minore del modello utilizzando il set di dati per la fase di addestramento, e successivamente stimare l'errore di classificazione sul set di dati di valutazione. Questo processo consentirà di ottimizzare la configurazione della rete neurale per la previsione delle transazioni fraudolente.

```
size= 1 decay= 1e-04
[1] "er= 0.23"
size= 2 decay= 0.001
[1] "er= 0.23"
size= 3 decay= 0.01
[1] "er= 0.228"
size= 4 decay= 1e-04
[1] "er= 0.231"
size= 5 decay= 0.001
[1] "er= 0.227"
size= 1 decay= 0.01
[1] "er= 0.23"
size= 2 decay= 1e-04
[1] "er= 0.233"
size= 3 decay= 0.001
[1] "er= 0.23"
size= 4 decay= 0.01
[1] "er= 0.234"
size= 5 decay= 1e-04
[1] "er= 0.23"
size= 1 decay= 0.001
[1] "er= 0.23"
size= 2 decay= 0.01
[1] "er= 0.23"
size= 3 decay= 1e-04
[1] "er= 0.232"
size= 4 decay= 0.001
[1] "er= 0.224"
size= 5 decay= 0.01
[1] "er= 0.231"
```

Figura 37: Ricerca del numero di neuroni nascosti

Dalla Figura 37, osserviamo che il modello con 4 neuroni nascosto (*M*) e il parametro λ pari a 0.001 presenta l'errore di classificazione minore. Questo suggerisce che questa configurazione ottimizzata potrebbe essere la scelta migliore per il nostro modello neurale.

Attraverso l'analisi della Figura 38, è possibile osservare graficamente la rete neurale specificata. Come previsto, il modello è composto dalle variabili di input che costituiscono lo strato di ingresso, e presenta un diverso neurone nello strato nascosto. Questo neurone elabora le informazioni e fornisce la previsione nello strato di uscita. Nel

complesso, questa rete neurale è formata da 81 pesi, evidenziando la complessità del processo di apprendimento e adattamento del modello alle caratteristiche dei dati.

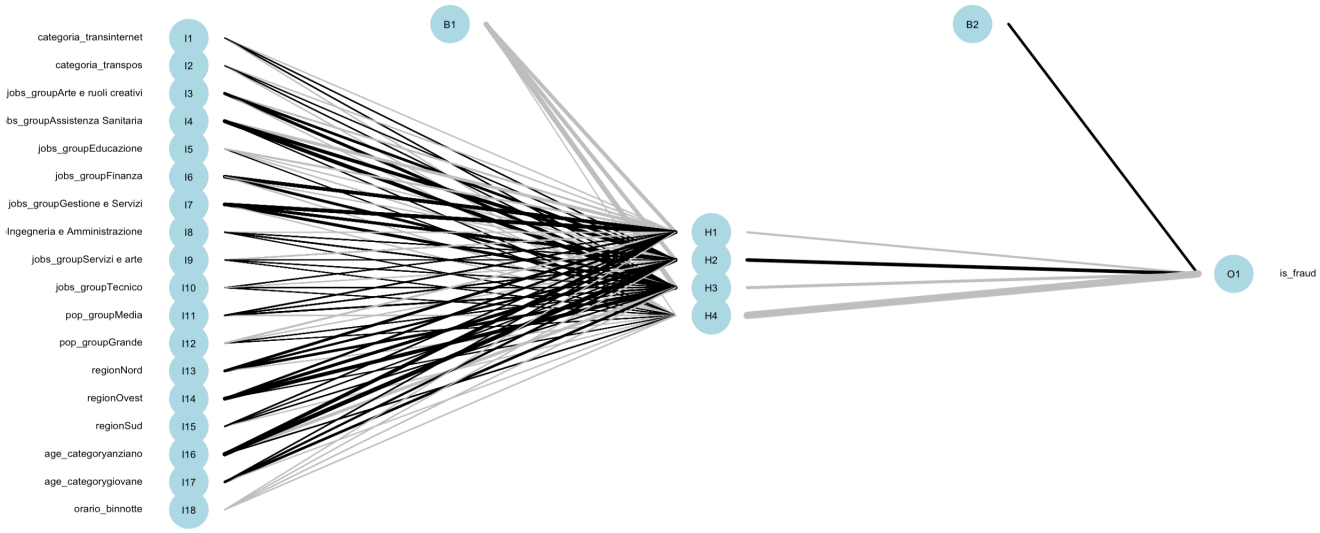

Figura 38: Rete neurale stimata

Analogamente all'approccio utilizzato per l'albero di classificazione, per valutare la capacità predittiva della rete neurale, facciamo ricorso alla matrice di confusione. Inoltre, daremo particolare rilevanza all'analisi della sensitività, mantenendo la considerazione per la motivazione precedentemente discussa. L'attenzione sulla sensitività ci permetterà di valutare la capacità della rete neurale nel rilevare accuratamente le transazioni fraudolente, aspetto critico nella gestione del rischio di frodi.

```
      no   si
  no 3621  752
  si  675 1201
> # precisione
> 1201 / (1201 + 752) * 100
[1] 61.49514
> # specificità
> 3621 / (3621 + 752) * 100
[1] 82.80357
> # sensitività
> 1201 / (1201 + 675) * 100
[1] 64.01919
```

Figura 39: Matrice di confusione e metriche rete neurale

La rete neurale stimata presenta una sensitività di circa il 64 %, risultato estremamente soddisfacente che evidenzia la sua capacità di rilevare efficacemente le transazioni fraudolente. Inoltre, la stima del tasso di corretta classificazione si attesta intorno al 77 %, suggerendo che la rete neurale, pur essendo un modello complesso, dimostra elevate capacità predittive.

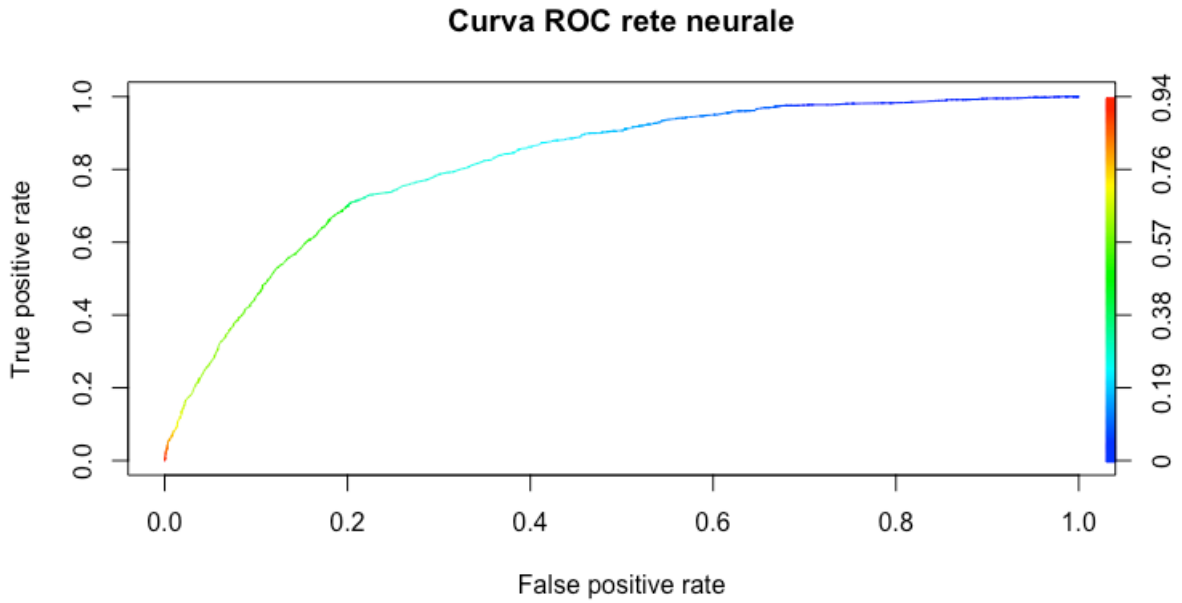

Figura 40: Curva ROC rete neurale

Anche la rete neurale presenta una curva ROC notevolmente soddisfacente per il modello nel suo complesso. L'Area Under the Curve (AUC) associata a questa curva si attesta intorno al 82 %, evidenziando un'eccellente capacità discriminativa del modello tra le transazioni fraudolente e non fraudolente

3.3 Vantaggi e problematiche

Le reti neurali presentano diversi vantaggi che le rendono uno dei metodi statistici più utilizzati in tempi recenti. Tra questi vantaggi si includono l'elevata velocità di calcolo, la robustezza alla presenza di valori anomali, la flessibilità e la capacità di adattamento, nonché l'assenza di ipotesi specifiche sulla distribuzione delle variabili. Queste caratteristiche le rendono adatte ad affrontare una vasta gamma di problemi complessi e ad adattarsi a gestire relazioni di qualsiasi natura.
Tuttavia, è importante notare che le reti neurali presentano anche alcuni svantaggi, tra cui la complessità nell'interpretazione dei risultati, difficoltà nella stima dei parametri e il rischio di overfitting in presenza di dati limitati.

Conclusioni

Nel corso di questo elaborato sulla prevenzione delle frodi, abbiamo esaminato due approcci distinti per la classificazione e la rilevazione delle frodi: l'albero di classificazione e il modello neurale. L'obiettivo principale era comprendere le prestazione di entrambi i modelli in termini di precisione, specificità, sensitività e capacità di adattamento del modello nel suo complesso, tramite l'utilizzo della curva ROC. Dopo un'analisi delle metriche precedentemente menzionate, emerge una parità di prestazioni tra i due modelli considerati. In situazioni simili, si potrebbe argomentare che l'albero di classificazione rappresenti il modello preferibile, poiché si propone come una soluzione semplice e facilmente interpretabile, a differenza della complessità associata alla rete neurale.

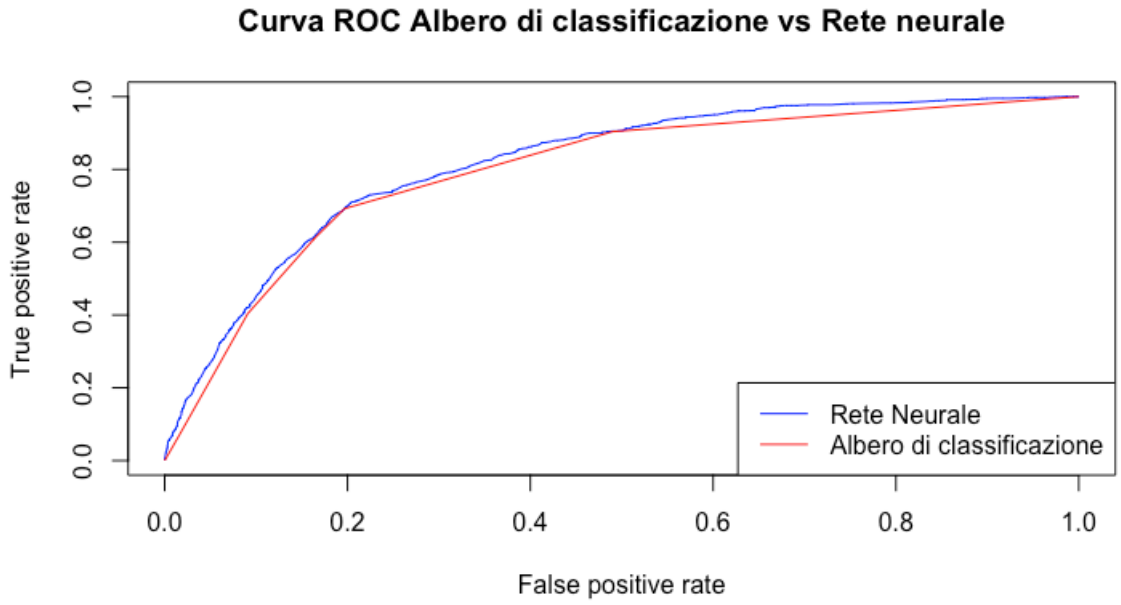

Figura 41: Confronto curva ROC

In generale, osserviamo come l'utilizzo della curva ROC suggerisca che un modello più complesso, come la rete neurale, possa essere considerato preferibile. Quando confrontiamo le AUC, notiamo che l'albero di classificazione presenta un 79%, mentre la rete neurale raggiunge un 82%. Questi risultati indicano una maggiore capacità di discriminazione della rete neurale, sottolineando il suo potenziale superiore nella classi-

ficazione delle transazioni fraudolente rispetto all'albero di classificazione. Nella Tabella 3, per agevolare il confronto, riportiamo i principali indicatori per i modelli utilizzati in questo elaborato.

Indicatore	Albero di Classificazione	Rete Neurale
Tasso di corretta classificazione	85,9%	76,9%
Precisione	84,5%	61,5%
Specificità	86,3%	82,8%
Sensitività	64,9%	64%

Tabella 3: Confronto indicatori principali

Dopo aver valutato le prestazioni dei modelli attuali, un passo successivo cruciale potrebbe riguardare l'espansione dell'analisi attraverso l'utilizzo di un campione di dati più ampio. Questo consentirebbe di ottenere una visione più completa delle capacità predittive dei modelli e di affrontare la sfida dello sbilanciamento dei dati, una problematica spesso presente nelle applicazioni di prevenzione delle frodi.

L'analisi di un campione di dati più esteso fornirebbe un quadro più accurato della generalizzazione dei modelli e della loro capacità di gestire scenari reali. In particolare, l'attenzione dovrebbe essere rivolta alla gestione efficace degli eventi di frode, tenendo conto della loro frequenza relativamente bassa rispetto alle transazioni normali.

Successivamente, potrebbe essere utile esplorare modelli di classificazione supervisionata aggiuntivi per confrontare le loro prestazioni con quelli già esaminati. Modelli come il modello logistico e il modello *random forest* offrono approcci diversi alla classificazione e potrebbero rivelarsi utili nel contesto specifico della prevenzione delle frodi. La valutazione di questi modelli consentirebbe di determinare quale approccio si adatta meglio ai requisiti specifici del problema in esame, considerando aspetti come la trasparenza del modello, la facilità di interpretazione e la capacità di gestire dati sbilanciati.

Inoltre, potrebbe essere interessante esplorare tecniche avanzate per affrontare lo sbilanciamento dei dati, come il sottocampionamento (*undersampling*) o il sovracampionamento (*oversampling*), al fine di migliorare ulteriormente le prestazioni del modello nella rilevazione delle frodi. Questa fase di analisi più approfondita contribuirebbe a fornire

una base solida per la selezione del modello ottimale nel contesto della prevenzione delle frodi finanziarie.

Bibliografia

[1] E. C. Bank, «Card fraud in Europe declined notably in 2021 amid the implementation of regulatory measures», mag. 2023. DOI: 10.2866/531174.

[2] CRIF. «Frodi creditizie in Italia: danni per oltre 83 milioni di euro nel I semestre 2023». (dic. 2023), indirizzo: https://www.crif.it/area-stampa/frodi-creditizie-in-italia-i-semestre-2023.

[3] J. Plotkin. «Sparkov_Data_Generation». (), indirizzo: https://github.com/namebrandon/Sparkov_Data_Generation.

[4] «Settlement hierarchy». (), indirizzo: https://en.wikipedia.org/wiki/Settlement_hierarchy.

[5] «Fraud Detection Handbook». (), indirizzo: https://fraud-detection-handbook.github.io/fraud-detection-handbook/Foreword.html.

[6] «L'AI prevede la domanda di mercato e aiuta la PMI». (), indirizzo: https://www.agendadigitale.eu/industry-4-0/intelligenza-artificiale-per-la-previsione-della-domanda-metodi-e-benefici-per-le-pmi/.

[7] «Precision and recall». (), indirizzo: https://en.wikipedia.org/wiki/Precision_and_recall.

[8] «Cos'è una Rete Neurale?» (), indirizzo: https://aws.amazon.com/it/what-is/neural-network/.

Ringraziamenti

A mio padre,
senza i tuoi sacrifici e le mattinate perse a portarmi a scuola, non sarei mai arrivato fino a questo punto. Hai sempre fatto di tutto per il mio bene. Spero di riuscire ad essere forte come te con i miei futuri figli.

A mia mamma,
che mi è sempre stata vicina, mi ha amato fin da piccolino, si è presa cura di me e non mi ha mai abbandonato.

A mia sorella,
la persona che mi è sempre stata accanto in ogni momento difficile, spingendomi a ragionare e a raggiungere i miei obiettivi.

A mio nonno,
la persona che mi ha sempre spinto e ha creduto in me fin dal primo giorno dei miei studi. Ricordo ancora quando la mattina presto mi portavi a scuola.

A Denise,
a.k.a Geronimo Stilton, cavalluccio marino splendente azzurro sopravvissuto in tavola, l'amore mio eterno. È grazie anche alla tua pazienza, dura come una vecchia roccia interminabile durante ogni esame, al tuo amore incondizionato, ai nostri infiniti viaggi e alle nostre stupide idee di business che sono arrivato qui. Spero di condividere con te il resto della mia vita e oltre.