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

Metodi Statistici Multivariati per Il Credit Scoring

Prof.ssa Mariagiulia Matteucci

Stefano Alberghini

Università di Bologna
Finanza, Assicurazioni e Impresa

18 Luglio, 2024

Panoramica

1. Dati e analisi esplorativa
2. Alberi di classificazione
3. Reti Neurali
4. Conclusioni

Introduzione

- **Impatti Economici**: Secondo il rapporto della BCE, le perdite per transazioni fraudolente nell'area SEPA nel 2021 sono state € 1.53 miliardi, con un aumento dell'importo medio in Italia nel primo semestre 2023
- **Fattori Demografici**: I giovani sotto i 30 anni sono particolarmente vulnerabili, come indicato dal rapporto CRIF di dicembre 2023
- **Metodologia**: L'approccio include l'analisi di un set di dati contenente 25.000 transazioni simulate di 1000 ipotetici consumatori e 800 operatori commerciali, confrontando le prestazioni tra alberi decisionali e reti neurali per la classificazione delle transazioni fraudolente e non

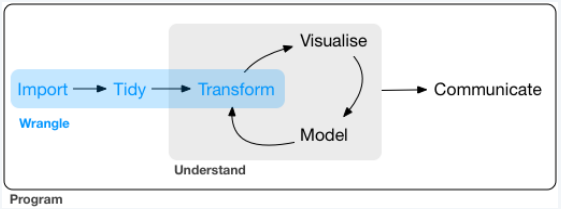
Dati e analisi esplorativa

Struttura del set di dati

Variabile	Descrizione
trans_date_trans_time	ora e giorno della transazione
category	categoria del negozio
amt	valore monetario della transazione
city	città del proprietario della c.c.
state	stato del proprietario della c.c.
city_pop	popolazione totale città del proprietario della c.c.
job	mestiere del proprietario della c.c.
dob	data di nascita del proprietario della c.c.
is_fraud	se la transazione è fraudolenta (1) o meno (0)

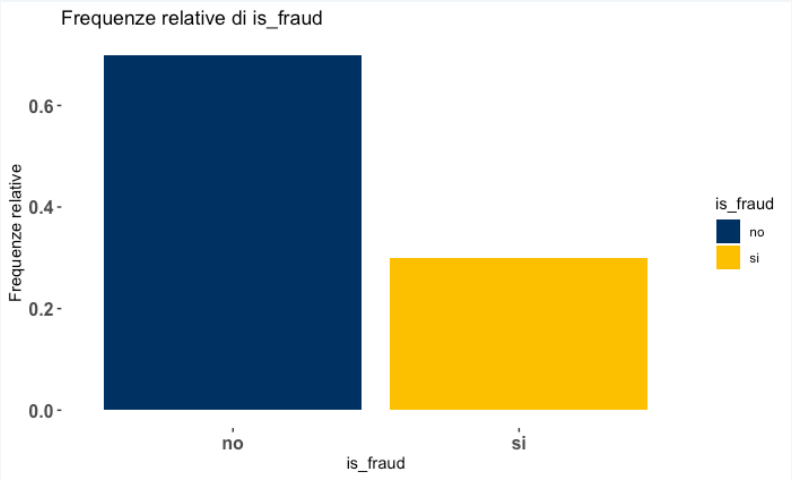
Table: Riepilogo variabili

Data Wrangling

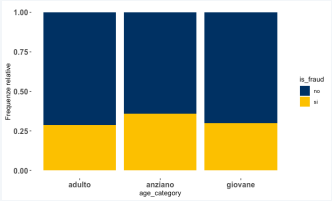


- **Trasformazione**: Filtraggio e creazione di nuove variabili per l'utilizzo nei modelli successivi tramite l'ausilio della libreria *dplyr*
- **Visualizzazione**: Visualizzazione grafica della distribuzione delle variabili con l'uso della libreria *ggplot2*
- **Modellazione**: Processo di stima dei modelli tramite il linguaggio R e impiego delle librerie *rpart*, *rpart.plot*, *nnet*, *NeuralNetTools*

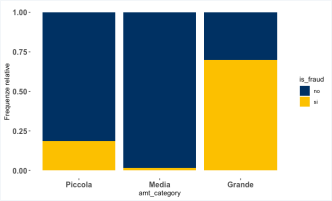
Visualizzazione dei dati



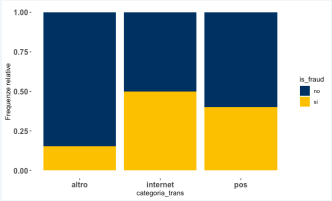
Visualizzazione dei dati (2)



(a) **age_category**

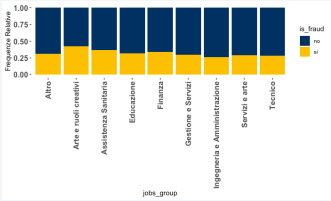


(b) **amt_category**

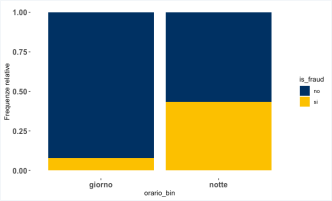


(c) **categoria_trans**

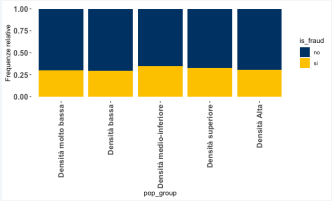
Visualizzazione dei dati (3)



(a) **jobs_group**

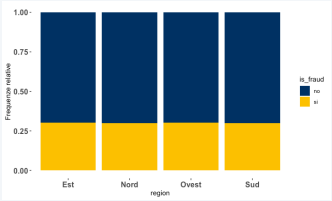


(b) **orario_bin**

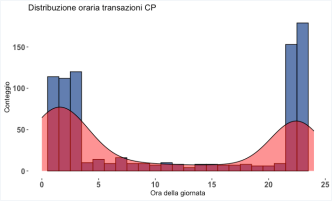


(c) **pop_group**

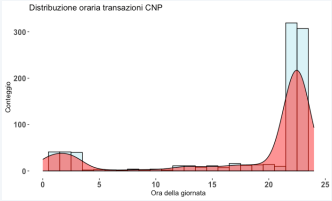
Visualizzazione dei dati (4)



(a) **region**



(b) **Transazioni fraudolente CP**



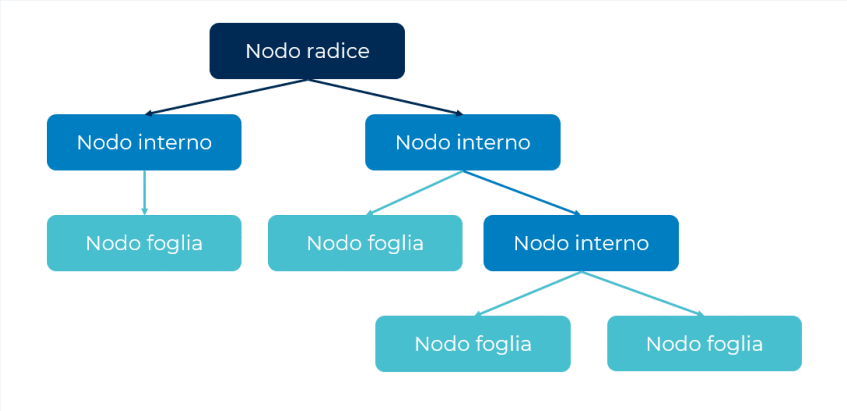
(c) **Transazioni fraudolente CNP**

Test Chi-Quadrato

Variabile	P-value
age_category	2.2e+16
amt_category	2.2e+16
categoria_trans	2.2e+16
jobs_group	2.192e-10
orario_bin	2.2e-16
pop_group	6.5e-05
region	0.9384

Alberi di classificazione

Architettura alberi di classificazione



Tecnica del Pruning

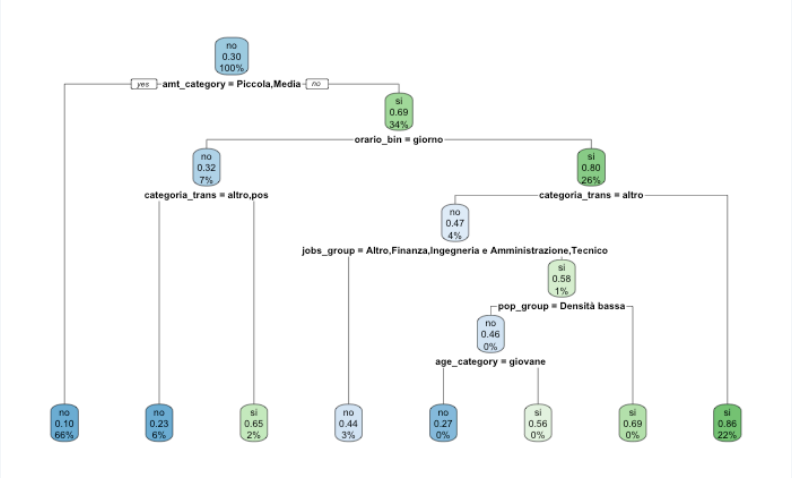
La procedura di potatura avviene in diverse fasi:

1. **Costruzione dell'Albero**: Si costruisce un albero completo sulla base dei dati di addestramento.
2. **Potatura**: Si rimuovono alcuni rami o nodi dall'albero completo, producendo una versione più semplificata.
3. **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.

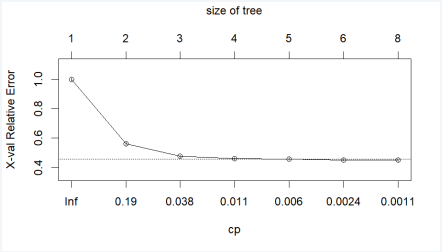
- **Pruning Selettivo**: Sequenza ottimale di sottoalberi, viene identificata utilizzando una funzione di costo-complessità:

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

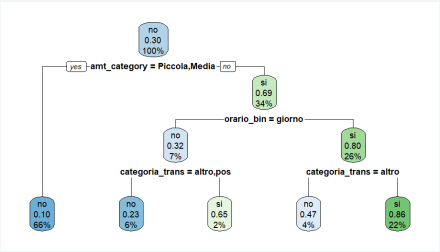
Risultati: Albero di Classificazione



Analisi della tecnica del pruning



(a) **Tecnica del pruning**



(b) **Albero di classificazione potato**

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

Reti Neurali

Architettura Rete Neurale

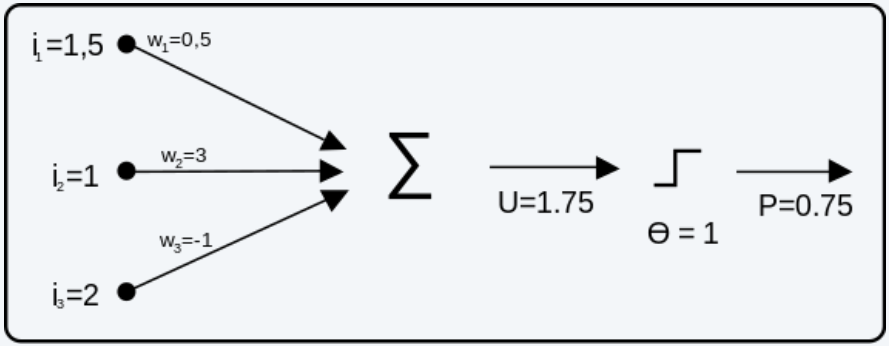
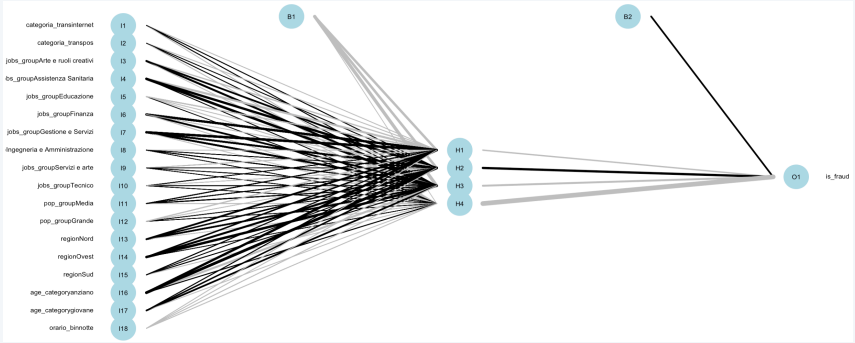


Figure: Rete Neurale di McCulloch e Pitts

Risultati: Rete Neurale

Size	1	2	3	4	5	1	2	3	4	5	1	2	3	**4**	5
λ	0.0001	0.001	0.01	0.0001	0.001	0.01	0.0001	0.001	0.01	0.0001	0.001	0.01	0.0001	**0.001**	0.01
Errore di classificazione	0.23	0.23	0.228	0.231	0.227	0.23	0.233	0.23	0.234	0.23	0.23	0.23	0.232	**0.224**	0.231

Table: Ricerca del numero di neuroni nascosti (Weight Decay)



Conclusioni

Confronto della capacità di classificazione

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%

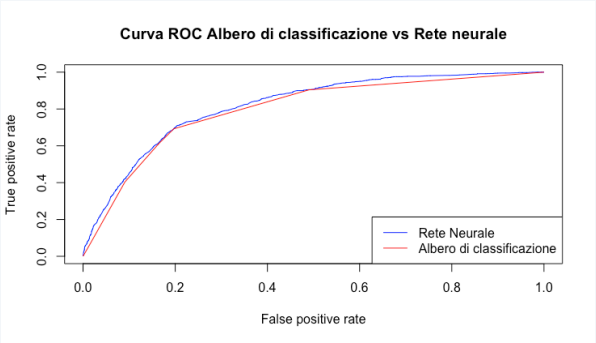


Figure: AUC Rete Neurale: 82% - AUC Albero di classificazione: 79%

Grazie per la vostra attenzione

Prof.ssa Mariagiulia Matteucci

Università di Bologna
Finanza, Assicurazioni e Impresa

Stefano Alberghini

18 Luglio, 2024