RILEVAZIONE DI FRODI FINANZIARE TRAMITE RETE NEURALE

Progetto per l'insegnamento di Basi di Dati

Basato sugli appunti di
Stefano Zizzi

Università degli Studi di Urbino
Corso di Laurea in Informatica Applicata

Indice

Capitolo 1

Introduzione

Il presente progetto ha l'obiettivo di sviluppare un modello di machine learning per la rilevazione delle frodi finanziarie, con particolare riferimento alle transazioni con carta di credito. Nel contesto attuale, in cui il volume delle transazioni digitali continua a crescere, la capacità di identificare con precisione transazioni potenzialmente fraudolente è diventata una priorità per le istituzioni finanziarie e i provider di servizi di pagamento.

In questo progetto, partiamo da un dataset contenente informazioni su transazioni, alcune delle quali sono identificate come fraudolente. L'obiettivo principale è quello di addestrare un modello predittivo capace di distinguere le transazioni lecite da quelle fraudolente. Per raggiungere questo scopo, esploreremo e confronteremo diverse tecniche di machine learning, inclusi algoritmi di classificazione tradizionali e metodi avanzati basati su reti neurali.

Il processo di sviluppo prevede diversi passaggi, tra cui la raccolta dei dati, la pulizia e la preparazione del dataset, la selezione e ingegnerizzazione delle caratteristiche, l'allenamento del modello e la valutazione delle prestazioni tramite metriche appropriate come accuratezza, precisione e AUC-ROC. Un obiettivo secondario è anche quello di confrontare le prestazioni dei vari modelli per determinare quale approccio si dimostri più efficace nel rilevamento delle frodi.

Per l'implementazione pratica, viene utilizzato un notebook in Python che sfrutta librerie comuni di data science come Pandas, Scikit-learn, TensorFlow e Matplotlib. Il dataset, scaricato da Kaggle, viene elaborato e suddiviso in transazioni legittime e fraudolente, con l'intento di creare un sistema predittivo robusto e accurato.

Questo progetto fornirà uno strumento prezioso per affrontare una delle problematiche più pressanti nel settore finanziario, ovvero la lotta contro le frodi nelle transazioni elettroniche.

Capitolo 2

Analisi

L'obiettivo principale di questa sezione è esaminare il dataset delle transazioni e ottenere informazioni utili che possano facilitare il processo di rilevazione delle frodi. A tal fine, ho seguito i seguenti passi:

1. Comprendere il quadro generale del dataset e delle sue caratteristiche;
2. Caricare e pulire i dati;
3. Esplorare e visualizzare i dati per identificare tendenze, anomalie e pattern;
4. Preparare i dati per l'addestramento di algoritmi di machine learning;
5. Selezionare e addestrare un modello di machine learning;
6. Migliorare le prestazioni del modello attraverso tecniche di ottimizzazione e tuning.

2.1 Esplorazione del Dataset

Il dataset analizzato contiene oltre un milione di transazioni, ognuna caratterizzata da una serie di variabili come l'importo, il luogo, la categoria del commerciante, le informazioni geografiche e demografiche dell'utente, e l'etichetta `is_fraud`, che indica se la transazione è stata segnalata come fraudolenta. Un'analisi preliminare delle variabili ci fornisce un quadro iniziale delle informazioni contenute.

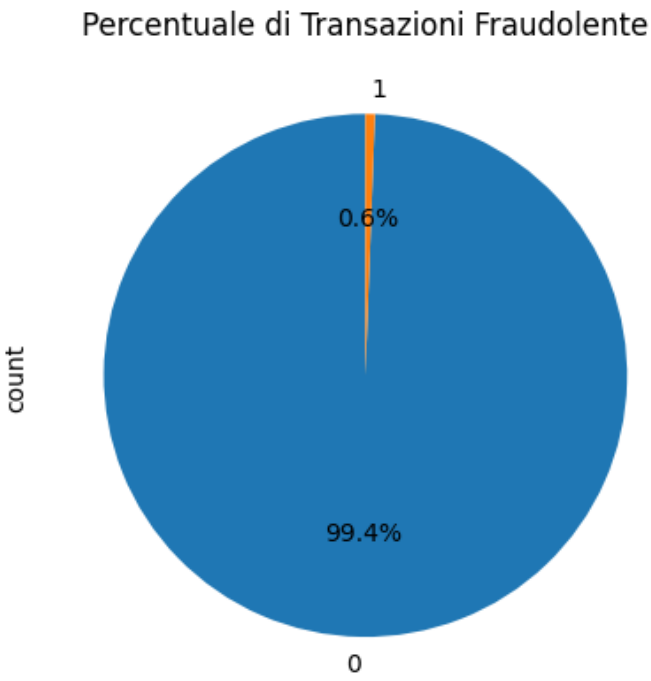
Il dataset contiene alcune colonne con valori nulli, come `merch_zipcode`, che richiederanno operazioni di pulizia e gestione dei dati mancanti. Una prima ispezione descrittiva ci permette di capire meglio la distribuzione dei valori, l'ampiezza del dataset e le dimensioni di alcune variabili chiave come l'importo della transazione (`amt`), la latitudine (`lat`), la longitudine (`long`) e la popolosità della città (`city_pop`).

	Unnamed: 0	trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender	street	city	state	zip	lat	long	city_pop	job	dob	trans_num	unix_time	merch_lat	merch_long	is_fraud	merch_zipcode
0	0	2019-01-01 00:00:18	2703186189652095	fraud_Rippin, Kub and Mann	misc_net	4.97	Jennifer	Banks	F	561 Perry Cove	Moravian Falls	NC	28654	36.0788	-81.1781	3495	Psychologist, counselling	1988-03-09	0b242abb623afc578575680df30655b9	1325376018	36.011293	-82.048315	0	28705.0
1	1	2019-01-01 00:00:44	630423337322	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23	Stephanie	Gill	F	43039 Riley Greens Suite 393	Orient	WA	99160	48.8878	-118.2105	149	Special educational needs teacher	1978-06-21	1f76529f8574734946361c461b024d99	1325376044	49.159047	-118.186462	0	NaN
2	2	2019-01-01 00:00:51	38859492057661	fraud_Lind-Buckridge	entertainment	220.11	Edward	Sanchez	M	594 White Dale Suite 530	Malad City	ID	83252	42.1808	-112.2620	4154	Nature conservation officer	1962-01-19	a1a22d70485983eac12b5b88dad1cf95	1325376051	43.150704	-112.154481	0	83236.0
3	3	2019-01-01 00:01:16	3534093764340240	fraud_Kutch, Hermiston and Farrell	gas_transport	45.00	Jeremy	White	M	9443 Cynthia Court Apt. 038	Boulder	MT	59632	46.2306	-112.1138	1939	Patent attorney	1967-01-12	6b849c168bdad6f867558c3793159a81	1325376076	47.034331	-112.561071	0	NaN
4	4	2019-01-01 00:03:06	375534208663984	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M	408 Bradley Rest	Doe Hill	VA	24433	38.4207	-79.4629	99	Dance movement psychotherapist	1986-03-28	a41d7549acf90789359a9aa5346dcb46	1325376186	38.674999	-78.632459	0	22844.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1296670	1296670	2020-06-21 12:12:08	30263540414123	fraud_Reichel Inc	entertainment	15.56	Erik	Patterson	M	162 Jessica Row Apt. 072	Hatch	UT	84735	37.7175	-112.4777	258	Geoscientist	1961-11-24	440b587732da4dc1a6395aba5fb41669	1371816728	36.841266	-111.690765	0	NaN
1296671	1296671	2020-06-21 12:12:19	6011149206456997	fraud_Abernathy and Sons	food_dining	51.70	Jeffrey	White	M	8617 Holmes Terrace Suite 651	Tuscarora	MD	21790	39.2667	-77.5101	100	Production assistant, television	1979-12-11	278000d2e0d2277d1de2f890067dcc0a	1371816739	38.906881	-78.246528	0	22630.0
1296672	1296672	2020-06-21 12:12:32	3514865930894695	fraud_Stiedemann Ltd	food_dining	105.93	Christopher	Castaneda	M	1632 Cohen Drive Suite 639	High Rolls Mountain Park	NM	88325	32.9396	-105.8189	899	Naval architect	1967-08-30	483f52fe67fabef353d552c1e662974c	1371816752	33.619513	-105.130529	0	88351.0
1296673	1296673	2020-06-21 12:13:36	2720012583106919	fraud_Reinger, Weissnat and Strosin	food_dining	74.90	Joseph	Murray	M	42933 Ryan Underpass	Manderson	SD	57756	43.3526	-102.5411	1126	Volunteer coordinator	1980-08-18	d667cdcbadaae83da3f4020e83591c83	1371816816	42.788940	-103.241160	0	69367.0
1296674	1296674	2020-06-21 12:13:37	4292902571056973207	fraud_Langosh, Wintheiser and Hyatt	food_dining	4.30	Jeffrey	Smith	M	135 Joseph Mountains	Sula	MT	59871	45.8433	-113.8748	218	Therapist, horticultural	1995-08-16	8f7c8e4ab7f25875d753b422917c98c9	1371816817	46.565983	-114.186110	0	59870.0

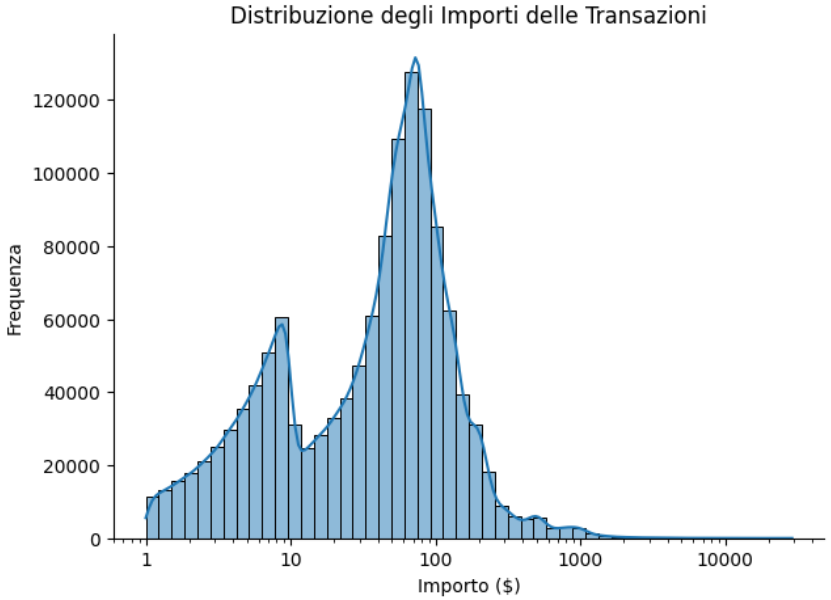
1296675 rows × 24 columns

2.1.1 Visualizzazione

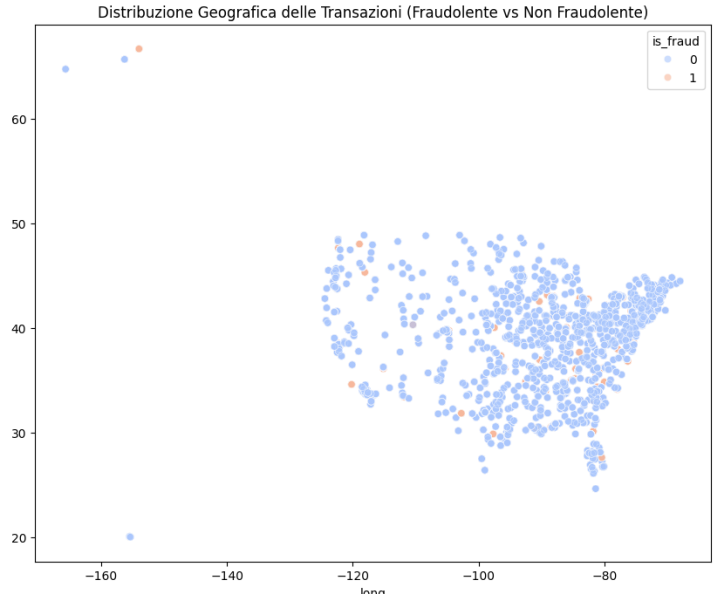
Un primo grafico a torta è stato utilizzato per visualizzare la percentuale di transazioni fraudolente nel dataset. Questo ci aiuta a comprendere l'entità del problema di classificazione sbilanciata.



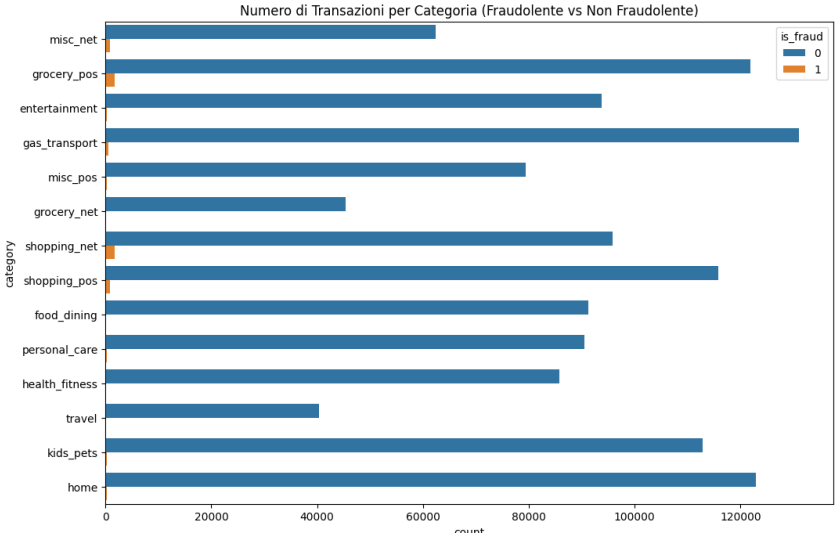
Successivamente, ho analizzato la distribuzione degli importi delle transazioni utilizzando un istogramma con scala logaritmica. Questo tipo di visualizzazione consente di vedere come la maggior parte delle transazioni siano di importo relativamente basso, con alcune transazioni che raggiungono valori molto elevati.



Un'altra analisi interessante è stata la visualizzazione delle transazioni su una mappa geografica, utilizzando le coordinate di latitudine e longitudine. Ho suddiviso le transazioni in fraudolente e non fraudolente, utilizzando colori diversi per rappresentare queste due categorie. Questo ci aiuta a individuare eventuali cluster geografici dove le frodi sono più frequenti.



Infine, ho esplorato la relazione tra la categoria dei commercianti e la frequenza delle transazioni fraudolente. Questo grafico a barre mostra come alcune categorie di commercianti siano maggiormente soggette a frodi rispetto ad altre. Ad esempio, categorie come il divertimento e il trasporto sembrano essere più vulnerabili alle frodi.



L'analisi esplorativa dei dati rivela alcune importanti osservazioni. Il dataset è altamente sbilanciato, con una percentuale molto ridotta di transazioni fraudolente rispetto a quelle legittime. Questo suggerisce che sarà necessario applicare tecniche specifiche per gestire il problema del class imbalance. Inoltre, la distribuzione degli importi delle transazioni e le informazioni geografiche potrebbero fornire indizi utili per la rilevazione delle frodi.

2.1.2 Elaborazione e Pulizia

Nella fase di preprocessing dei dati per l'analisi delle frodi sulle transazioni con carte di credito, si sono seguiti diversi step fondamentali per la preparazione e l'analisi del dataset. Ecco una descrizione dettagliata del processo.

Il dataset conteneva una colonna con la data di nascita degli utenti (denominata `dob`). Dal momento che la data di nascita è più difficile da usare per i modelli rispetto a un numero come l'età, è stata creata una funzione per calcolare l'età attuale di ciascun utente in base alla sua data di nascita. Questo è stato fatto convertendo la data di nascita in anni, utilizzando la data odierna come riferimento. La colonna

`dob` è stata quindi eliminata e sostituita con una nuova colonna `age`.

Per evitare di includere nel modello informazioni che potrebbero non essere rilevanti o che potrebbero creare bias, sono state eliminate numerose colonne contenenti dati sensibili e identificativi. Tra queste colonne figurano numeri di carta di credito, informazioni sul commerciante (come nome e indirizzo), e dettagli personali come il nome e l'indirizzo dell'utente. Questo passaggio è stato eseguito per focalizzarsi sui dati essenziali per il rilevamento delle frodi, come l'importo della transazione, la posizione geografica e la categoria di acquisto.

Dopo la pulizia iniziale, il dataset è stato riordinato per migliorare la leggibilità e l'accessibilità delle informazioni. Le colonne sono state disposte in un ordine logico che facilita l'analisi, con i dati temporali e spaziali (come il `timestamp` e le coordinate geografiche) posti all'inizio, seguiti dai dati finanziari e demografici (importo, età, genere, ecc.), fino ad arrivare alla variabile target (`is_fraud`), che indica se una transazione è fraudolenta o meno.

È stata quindi creata una funzione di reportistica per analizzare la struttura del dataset. Questo report includeva informazioni sui tipi di dati per ciascuna colonna (ad esempio numerico, categorico), il numero di valori mancanti e la percentuale di dati mancanti, oltre al numero di valori unici in ogni colonna. Questa analisi ha aiutato a identificare eventuali problematiche nei dati, come valori nulli, ridondanze o colonne con distribuzioni anomale di dati.

Questo passaggio di reportistica è cruciale per verificare l'integrità del dataset prima di procedere con fasi successive di modellazione. Il report ha mostrato che non c'erano valori mancanti nel dataset, il che ha facilitato le fasi successive di preparazione dei dati.

Il dataset conteneva alcune colonne categoriali, come `category` (categoria di acquisto), `gender` (genere) e `job` (occupazione). Poiché la maggior parte degli algoritmi di machine learning richiede che i dati siano in forma numerica, queste colonne categoriali sono state convertite in numeri attraverso un processo di encoding. È stata utilizzata una tecnica di *"Label Encoding"*, dove a ciascuna categoria è stato assegnato un numero intero. Questo ha consentito al modello di trattare queste informazioni in maniera efficace.

Dopo aver trasformato i dati, è stata eseguita un'analisi delle correlazioni tra le varie feature del dataset, incluso il target `is_fraud`. Utilizzando una matrice di correlazione, è stato possibile osservare come le diverse variabili fossero correlate tra loro. La visualizzazione attraverso una *heatmap* ha evidenziato eventuali forti correlazioni positive o negative, il che ha permesso di identificare eventuali ridondanze tra le colonne. Ad esempio, colonne come `merch_lat` e `merch_long` mostravano una forte correlazione, portando alla decisione di eliminarle per evitare ridondanza e migliorare l'efficienza del modello.

Dopo aver identificato forti correlazioni tra alcune colonne, come la latitudine e la longitudine del commerciante, si è proceduto con la loro rimozione. Queste colonne fortemente correlate avrebbero potuto influenzare negativamente il modello, aumentando il rischio di sovradattamento (*overfitting*). La loro eliminazione ha semplificato il dataset senza perdita significativa di informazione.

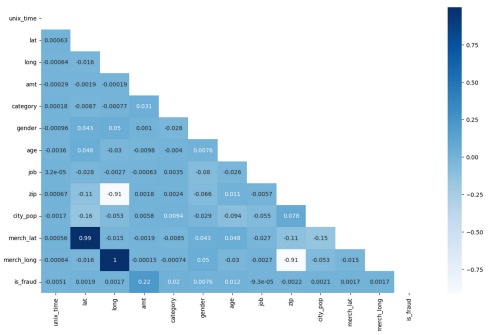


Figura 2.1: Prima delle rimozioni

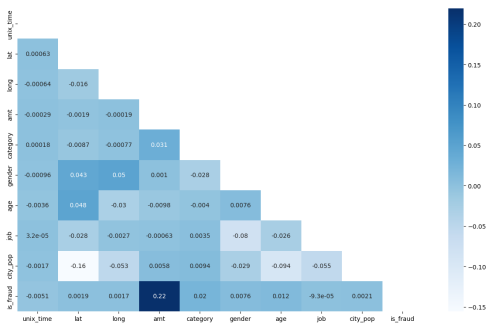


Figura 2.2: Dopo le rimozioni

Capitolo 3

Sviluppo

Il processo di sviluppo del modello per la rilevazione delle frodi nelle transazioni con carte di credito si basa su una rete neurale che sfrutta tecniche avanzate per migliorare la performance su un problema di classificazione binaria, ovvero la distinzione tra transazioni fraudolente e non fraudolente.

Il primo passo consiste nel suddividere il dataset in due parti: un set di addestramento (training) e uno di validazione (validation). Questo è stato fatto mantenendo l'80% dei dati per il training e il restante 20% per la validazione. La suddivisione dei dati serve a evitare che il modello impari in modo troppo specifico sui dati di addestramento (overfitting), permettendo di valutare la sua generalizzazione su nuovi dati non visti durante l'addestramento.

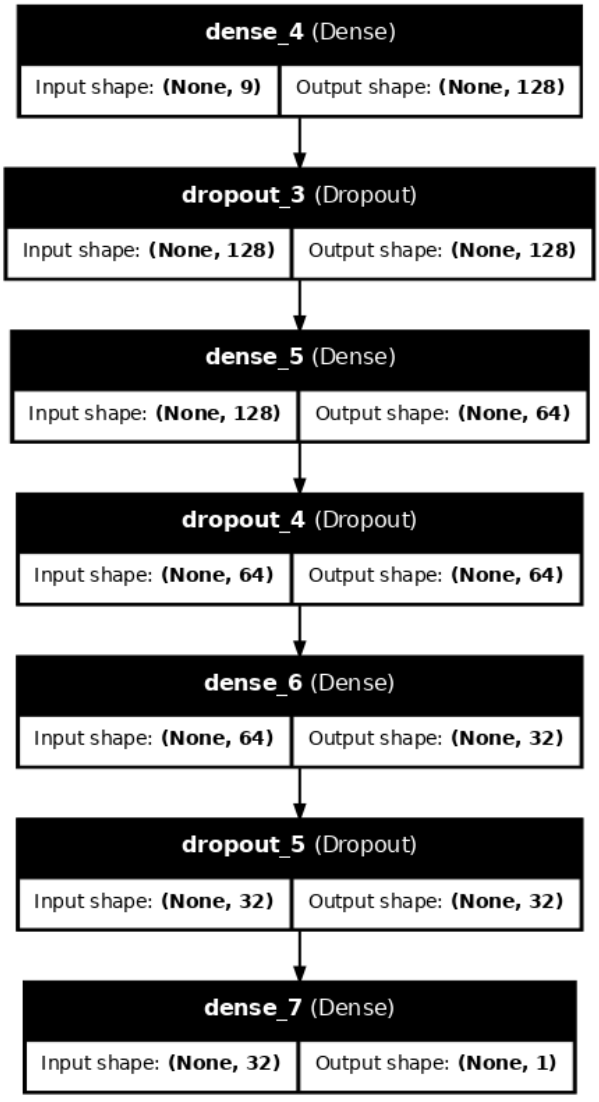
Le reti neurali sono sensibili alle scale delle variabili, quindi è fondamentale standardizzare i dati. In questo caso, è stato applicato un processo di standardizzazione per far sì che tutte le variabili abbiano una media pari a 0 e una deviazione standard pari a 1. Questo processo aiuta il modello a convergere più rapidamente durante l'addestramento e a migliorare la precisione.

La rete neurale sviluppata ha un'architettura sequenziale, composta da diversi strati:

- *Strato di Input*: Il primo strato contiene 128 neuroni e utilizza la funzione di attivazione $ReLU$, che è ideale per le reti profonde grazie alla sua capacità di introdurre non linearità.
- *Regolarizzazione e Dropout*: Per prevenire l'overfitting, è stato introdotto un livello di Dropout con una probabilità del 40%, il che significa che il 40% dei neuroni viene temporaneamente disattivato durante l'addestramento. Inoltre, è stata applicata una regolarizzazione L2 per imporre penalità sui pesi dei neuroni e migliorare ulteriormente la generalizzazione del modello.
- *Strati Intermedi*: Dopo il primo livello di neuroni, sono stati aggiunti ulteriori strati densi con un numero ridotto di neuroni (64,32), sempre con la funzione di attivazione $ReLU$ e con ulteriori livelli di Dropout e regolarizzazione L2.
- *Strato di Output*: Infine, il modello termina con uno strato di output a singolo neurone con attivazione *sigmoide*, che è ideale per problemi di classificazione binaria, poiché genera una probabilità di appartenenza a una classe (frode o non frode).

Poiché il dataset è caratterizzato da uno sbilanciamento tra le classi (con un numero molto inferiore di transazioni fraudolente rispetto a quelle lecite), per gestire il problema, durante l'addestramento del modello, sono stati assegnati pesi differenti alle classi per dare maggiore importanza alla classe fraudolenta (più rara). Questo aiuta il modello a prestare maggiore attenzione alle transazioni sospette.

L'addestramento della rete neurale è stato eseguito per un massimo di 100 epoche, ma con l'utilizzo di Early Stopping. Questa tecnica interrompe l'addestramento in modo automatico se il modello smette di migliorare sulle prestazioni di validazione, prevenendo l'overfitting. Durante l'addestramento, il modello ha monitorato la funzione di perdita sui dati di validazione per decidere quando fermarsi e conservare i pesi migliori ottenuti fino a quel punto.


dense_4 (Dense)
Input shape: (None, 9)
Output shape: (None, 128)
dropout_3 (Dropout)
Input shape: (None, 128)
Output shape: (None, 128)
dense_5 (Dense)
Input shape: (None, 128)
Output shape: (None, 64)
dropout_4 (Dropout)
Input shape: (None, 64)
Output shape: (None, 64)
dense_6 (Dense)
Input shape: (None, 64)
Output shape: (None, 32)
dropout_5 (Dropout)
Input shape: (None, 32)
Output shape: (None, 32)
dense_7 (Dense)
Input shape: (None, 32)
Output shape: (None, 1)

Capitolo 4

Valutazione e Conclusioni

Dopo l'addestramento, il modello è stato valutato sul set di validazione. Le metriche utilizzate per questa valutazione includono:

- **Accuracy**: Percentuale di transazioni correttamente classificate come frode o non frode.
- **Report di Classificazione**: Comprende precisione, recall e F1-score per entrambe le classi. Il recall è particolarmente importante in questo contesto poiché rappresenta la capacità del modello di identificare correttamente le frodi.
- **Confusion Matrix**: È stata generata una matrice di confusione per visualizzare il numero di veri positivi, falsi positivi, veri negativi e falsi negativi.

	precision	recall	f1-score	support
0	0.998	0.996	0.997	257815.0
1	0.479	0.661	0.556	1520.0
accuracy	0.994	0.994	0.994	0.994
macro avg	0.739	0.828	0.776	259335.0
weighted avg	0.995	0.994	0.994	259335.0

Sono stati generati dei grafici per visualizzare l'evoluzione dell'accuratezza e della funzione di perdita sia sui dati di training che su quelli di validazione durante il processo di addestramento. Questi grafici aiutano a comprendere meglio come si comporta il modello e se sta sovra-adattando (overfitting) o se sta ancora migliorando.

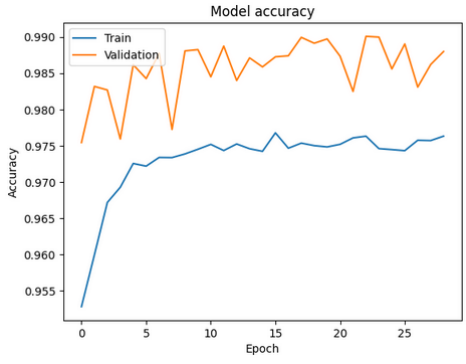


Figura 4.1: Precisione all'avanzare delle epoche

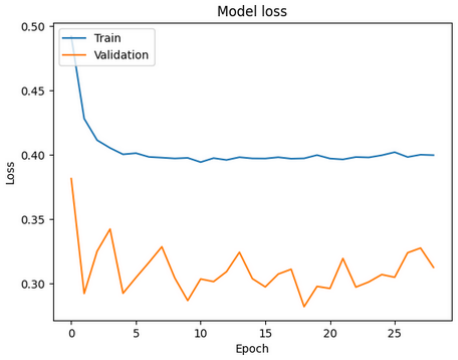
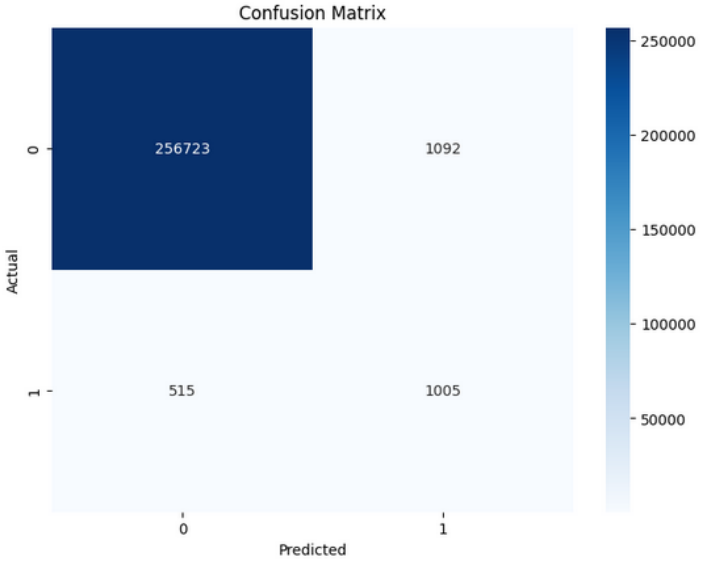


Figura 4.2: Perdita all'avanzare delle epoche

Infine, la *confusion matrix* è stata visualizzata graficamente per illustrare chiaramente come il modello classifica le transazioni, evidenziando eventuali errori di classificazione.


Confusion Matrix
256723
1092
515
1005
0
1
Actual
0
1
Predicted
250000
200000
150000
100000
50000

Queste dispense sono state stilate tramite LaTeX.
Il PDF è dotato di hyperlinking quindi è possibile cliccare sui nomi delle sezioni nell'indice per spostarsi tra le pagine, allo stesso modo, per tornare all'indice è sufficiente cliccare sul numero della pagina attuale.