***Progetto Machine Learning A.A. 2023-2024***

**Credit Card Fraud Detection**

Il progetto presentato in questa relazione è stato svolto con l’obiettivo di sviluppare un modello di Machine Learning in grado di individuare, dall’analisi di un dataset contenete informazioni riguardo transazioni bancarie eseguite con carta di credito, le *transazioni fraudolente*.

**EDA**

Prima di costruire il modello di apprendimento è stato necessaria una fase di *EDA* (Exploratory Data Analysis), per poter raccogliere il maggior numero di informazioni possibili sul dataset esaminato.

Le prime informazioni riguardo il dataset riguardano le *features* presenti al suo interno: oltre alle feature *Time* (secondi trascorsi tra ciascuna transazione e la prima transazione del dataset), *Amount* (importo della singola transazione) e *Class* (“1” se la transazione è una frode, “0” altrimenti) sono presenti le features V1, ..., V28 (***principal components*** ottenute da una *PCA*). Per motivi di privacy non è stato possibile utilizzare le features originali né ricavare altro tipo di informazioni.

Da una prima analisi emerge immediatamente che il dataset è estremamente ***sbilanciato***: le frodi costituiscono solo lo *0.17%* del totale delle transazioni presenti nel dataset (492 frodi su 284807 transazioni totali). In questa situazione il dataset è inutilizzabile se si vogliono costruire modelli predittivi e algoritmi senza rischio di ***overfit***: essi potrebbero assumere che la maggior parte delle transazioni non siano frodi e, dunque, non saper riconoscere quelle effettivamente fraudolente.

Le prime manipolazioni del dataset riguardano la rimozione dei duplicati (lavoriamo su 283726 transazioni invece che su 284807) e il conteggio di valori mancanti (assenti nel caso del nostro dataset).

Di seguito si è cercato di *visualizzare la distribuzione delle osservazioni* del dataset e di individuare ***correlazioni tra*** ***le features*** usando *pair plots* e una ***matrice di correlazione*** (sfruttando la *correlazione di Pearson*). Essendo le features V1-V28 ottenute eseguendo una PCA sul dataset originale esse sono totalmente scorrelate. Moderati livelli di correlazione si rilevano invece tra le seguenti coppie di features: "V3-Time" (-0.42),"V3-Amount" (-0.21),"V2-Amount" (-0.53),"V7-Amount" (0.4),"V20-Amount" (0.34). Tramite analisi più accurate si osserva che queste correlazioni non sono rilevanti al fine del corretto funzionamento di un successivo classificatore.

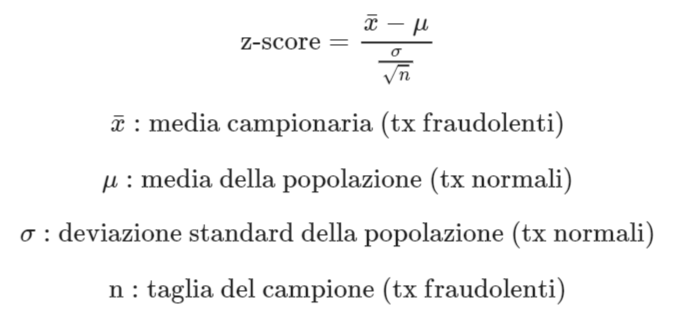
È utile poi visualizzare quando avvengono le frodi, per valutare se sono presenti eventuali pattern. Per fare ciò è necessario convertire il tempo da *secondi* a *giorni* / *ore* / *settimane.* Da questa analisi risulta che la maggior parte delle transazioni non fraudolente avviene nella fascia oraria 10-23 (molto più che nella fascia 0-9), mentre le transazioni fraudolente sembrano meglio distribuite nel tempo rispetto a quelle legittime. È, però, impossibile individuare delle tendenze per le transazioni fraudolente come accadeva invece per le transazioni legittime poiché il dataset è ancora sbilanciato.

Alcune valutazioni sono state effettuate sull’importo delle transazioni fraudolente per valutare se potesse essere un indicatore utile per la costruzione del successivo classificatore. Per fare ciò è stato creato un ***test d’ipotesi*** per verificare se le transazioni fraudolente avessero importo maggiore rispetto alle transazioni legittime:

H0 AmountFraud AmountNormal

H1 AmountFraud > AmountNormal

Utilizzando un *livello di significatività* del 99% calcoliamo la statistica del test. Come stimatore per la media della popolazione usiamo la media campionaria:



Dato che la statistica del test è maggiore di 2.33 (il quantile Z(1%)) rifiutiamo l'ipotesi nulla H0. Segue che c'è il 99% di probabilità che l'importo speso in transazioni fraudolente sia maggiore di quello speso in transazioni legittime. Eseguiamo il test d’ipotesi per ognuna della 28 features della PCA per vedere se il valore delle transazioni fraudolente è significativamente diverso da quello della transazione normale. Utilizziamo sempre un livello di significatività del 99% e rifiutiamo se la statistica del test (ora in valore assoluto) è maggiore del quantile Z(0.5%) ossia 2.58. Dal risultato di questa analisi possiamo dire se una feature è *significativa* (valori delle features diversi per i diversi tipi di transazione) oppure no (features *NON significative*, possono essere scartate). Ricaviamo che le features V13, V15, V22, V23, V25, V26 non sono molto utili per differenziare tra transazioni fraudolente e legittime. Probabilmente rimuovere tali features potrebbe essere utile per costruire classificatori migliori.

Per poter costruire un modello predittivo efficace e in grado di riconoscere correttamente dei pattern è necessario prelevare un sotto campione del dataframe per avere un numero bilanciato di transazioni fraudolente e transazioni legittime.

**Data Splitting e Random Under-Sampling**

Prima di procedere con il *Random Under-Sampling* si deve separare il database originale in *training* e *test set*. Per addestrare i modelli quando si implementano tecniche di *Random Under* / *Over Sampling* è necessario usare i set costruiti sui dati originali e non sui dati creati da tali tecniche.

La tecnica del *Random Under-Sampling* consiste nel rimuovere osservazioni, in modo randomico, per ottenere un dataset più bilanciato ed evitare modelli in overfit. Noto il numero di transazioni fraudolente (492) si riduce il numero di transazioni legittime (scegliendole in modo randomico, eseguendo uno shuffle) per avere un 50% di transazioni legittime (ossia 492) ed un 50% di transazioni fraudolente (ossia 492). Per quanto questa tecnica sia utile ed efficace, si corre il rischio di avere modelli di classificazione non accurati a causa di una *perdita di informazioni* causata dalla riduzione estrema del numero delle transazioni legittime (da 284315 a 492).