UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Scienze Economiche e Statistiche

Corso di Laurea Triennale in Statistica Per i Big Data

Prova finale in

Statistical Learning

MACHINE LEARNING FOR FRAUD DETECTION

Relatore:

Ch. mo Prof. Pietro Coretto

Candidato:

Davide Mascolo

matr. 0212800247

ANNO ACCADEMICO 2020/2021

Abstract

Con il termine frode o truffa, si definisce un comportamento avente come fine l'ottenimento di un vantaggio a scapito di un soggetto. Alla base di ogni truffa vi è un comportamento anomalo, ovvero un evento o osservazione rara, che non è conforme al metodo generale di distribuzione di quei dati e devia fortemente da ciò che è ritenuto normale. Un'anomalia, infatti, fa scattare un campanello d'allarme, sollevando il sospetto che ci sia qualcosa di sbagliato nel processo sottostante che sta generando quei dati. Nella seguente trattazione si pone particolare attenzione sul settore finanziario e le relative frodi causate dall'attacco a carte di credito con l'obiettivo di confrontare i risultati dei vari classificatori utilizzati.

Introduzione

Negli ultimi anni, l'utilizzo di carte di credito e di debito è notevolmente aumentato. Tuttavia, una parte significativa delle transazioni derivanti dall'utilizzo di tali metodi di pagamento sono fraudolente ed ogni anno vengono rubati miliardi di euro. I sistemi di fraud detection, ovvero *l'insieme dei processi ed analisi che consentono alle aziende di identificare e prevenire attività finanziare non autorizzate per impedire l'ottenimento di denaro o proprietà tramite falsi pretesti* (Alexander S. Gillis, 2019), sono diffusi da anni in molteplici settori industriali e proprio nel settore finanziario hanno conosciuto una grande popolarità. Il problema principale quando si affronta il tema della Fraud Detection, riguarda sicuramente lo *sbilanciamento delle classi* (Jason Brownlee, 2019) dato che il numero di transazioni quotidiane è molto elevato mentre la percentuale di frodi sul totale delle transazioni è molto bassa. È evidente, quindi, la difficoltà nell'individuare poche frodi localizzate in un insieme molto più vasto di transazioni. Se la classe più numerosa fosse anche quella di maggiore interesse, non sarebbe così grave assegnare ad essa tutti i soggetti dello studio; viceversa, la dominanza della classe di non interesse diventa un ostacolo quando la classe più importante è quella rara. Si necessita, quindi, di alcuni rimedi per gestire tale problematica in modo da indirizzare l'attenzione dell'algoritmo verso la classe d'interesse, visto che il processo di apprendimento, in presenza di una distribuzione della variabile di risposta estremamente sbilanciata è distorto, in quanto il metodo tende a focalizzarsi sulla classe prevalente ed ignorare gli eventi rari. Tuttavia, un servizio antifrode non deve limitarsi a voler individuare poche frodi localizzate, ma deve anche preoccuparsi del lasso di tempo in cui avviene tale azione. Infatti, è importante che l'individuazione della frode avvenga in tempo reale o almeno in un tempo che sia molto vicino al tempo reale, dato che la frode non sottende un arco temporale di lunga durata ma spesso consta di un singolo episodio, finalizzato all'ottenimento del massimo vantaggio economico nel minor tempo possibile. Da qui nasce l'esigenza di un sistema d'identificazione in grado di segnalare con la precisione più alta possibile eventi rari e soprattutto non ripetibili.

Presentazione del problema.

Il tema in esame è stato trattato con un approccio supervisionato e tenendo conto delle problematiche legate alla natura del fenomeno. Come anticipato nell'introduzione, il problema dello *sbilanciamento delle classi* per studi di classificazione binaria comporta gravi conseguenze sulla capacità di generalizzazione di un classificatore; quindi, dal momento che la classe d'interesse è quella con frequenza minore, bisogna adottare soluzioni differenti rispetto a studi di classificazione su classi bilanciate. Il problema della rappresentazione non uniforme dei dati tipica di numerose applicazioni nella vita reale, ha portato notevoli sviluppi nell'apprendimento da dati sbilanciati. Tuttavia, il problema della classificazione sbilanciata non è risolto, ma rimane un tema aperto in generale ed in pratica deve essere identificato ed affrontato in modo specifico per ciascun set di dati di addestramento (Bartosz Krawczyk, 2016). La tecnica utilizzata in questo studio, per affrontare tale problematica, è l'utilizzo di metriche di performance che tengono conto dello sbilanciamento tra le classi.

Metriche per la valutazione della performance.

La scelta della giusta metrica per la valutazione di un metodo rappresenta uno dei temi centrali di questo lavoro dato che l'utilizzo di una metrica che non tiene conto dello sbilanciamento tra le classi, con buona probabilità porterà a preferire un metodo scadente. Poiché la maggior parte delle metriche utilizzate comunemente per problemi di classificazione presuppongono una distribuzione delle classi bilanciata, la scelta di una metrica appropriata per problemi di classificazione sbilanciata è di fondamentale importanza e deve tener conto del *peso* degli errori che il classificatore compie. Esistono metriche standard come l'accuratezza della classificazione o l'errore di classificazione che funzionano bene sulla maggior parte dei problemi, motivo per cui sono ampiamente adottate, ma in problemi di *fraud detection* queste metriche diventano inaffidabili e fuorvianti. Partendo dai quattro possibili casi che il classificatore prevede, si definisce la prima *"metrica"*, ovvero la *matrice di confusione*, detta anche matrice di errore definita come una suddivisione delle previsioni in una tabella che mostra le previsioni corrette

sulla diagonale principale e gli errori commessi dal classificatore fuori dalla diagonale principale. Ogni riga della matrice rappresenta le istanze in una classe effettiva mentre ogni colonna rappresenta le istanze in una classe prevista o viceversa, entrambi le varianti si trovano in letteratura. Per superare il problema del *paradosso dell'accuratezza* (Tejumade Afonja, 2017) e partendo dai quattro possibili casi della matrice di confusione, si derivano metriche più fedeli per problemi con classi sbilanciate.

Sensitivity: Identifica la frazione di veri positivi correttamente classificati e quindi misura l'accuratezza predittiva del metodo per la classe positiva.

Precision: Rappresenta l'affidabilità e l'esattezza di un classificatore, ovvero la frazione dei classificati positivi che risultano essere veramente positivi.

Fβ – Score: Misura la precisione di un test ed è calcolata come media armonica ponderata di Precision e Sensitivity, e raggiunge il suo valore ottimale a 1 ed il suo valore peggiore a 0. β è un fattore reale positivo, ovvero un parametro determinato in modo tale che il recall sia considerato β volte più importante della precisione. Questa metrica *misura l'efficacia del recall rispetto ad un utente che attribuisce β volte più importanza al richiamo che alla precisione* (CJ Van Rijsbergen, 1979).

Coefficiente di Correlazione di Matthew: Questa metrica, nota anche come **coefficiente phi** è molto utilizzata nell'apprendimento automatico come misura della qualità di classificazioni binarie. Il coefficiente tiene conto di tutti e quattro i possibili casi della classificazione, ed è interpretato come il coefficiente di correlazione tra classi osservate e classi previste. Il suo utilizzo è molto consigliato per problemi di classificazione con classi sbilanciate e può essere utilizzato anche con classi aventi dimensioni molto differenti. Il vantaggio di questa metrica, a differenza dell'accuratezza o dell'F_1 sta nella sua definizione che considera tutte le componenti della matrice di confusione, offrendo quindi una valutazione più completa rispetto alle metriche sopra citate.

Kappa di Cohen: La metrica kappa di Cohen (***k***) è un coefficiente statistico che rappresenta il grado di accuratezza ed affidabilità di un classificatore. Spesso utilizzato per esaminare l'accordo tra due valutatori, è definito come un indice di concordanza che tiene conto della probabilità di concordanza casuale. Anche questa metrica è calcolata in base alla matrice di confusione ma contrariamente al calcolo dell'accuratezza complessiva, tiene conto della distribuzione sbilanciata tra le classi.

Receiver Operating Characteristic: La curva ***ROC*** è uno schema grafico per un classificatore binario, interpretabile anche come una curva di probabilità che traccia lungo i due assi la *sensibilità* ed *(1 – specificità)*, rispettivamente rappresentati da *True Positive Rate* (*TPR*, frazione di veri positivi) e *False Positive Rate* (*FPR*, frazione di falsi positivi) - anche conosciuto con il termine di *Fall-Out* – a vari valori di soglia e separa il "segnale" dal "rumore", permettendo di studiare i rapporti fra istanze effettivamente positive (*hit rate*) e falsi allarmi. Attraverso l'analisi delle curve si valuta la capacità discriminatoria di un classificatore tra un insieme del campione positivo ed uno negativo, calcolando l'area sottesa alla curva ROC, ovvero *l'Area Under Curve* (***AUC***). Il valore di AUC assume valori nel range [0; 1] ed *equivale alla probabilità che il risultato del classificatore applicato ad un'istanza estratta a caso dal gruppo dei positivi sia superiore a quello ottenuto applicandolo ad un'istanza estratta a caso dal gruppo dei negativi* (Donald Bamber, 1975; MH Zweig & G. Campbell 1993).

Discussione dei metodi.

Regressione Logistica

Il metodo logit, noto anche come metodo logistico o regressione logistica, è un metodo di regressione non lineare utilizzato quando la variabile dipendente è dicotomica, ovvero una variabile nominale con sole due modalità. La regressione logistica fa parte della famiglia dei modelli lineari generalizzati (Generalized Linear Model), ovvero metodi che sono una generalizzazione del classico metodo lineare nell'ambito della regressione lineare. La differenza

è che mentre il metodo lineare classico ipotizza che la variabile di riferimento sia distribuita normalmente, nei modelli generalizzati cade questa assunzione e la variabile dipendente può assumere qualsiasi distribuzione di variabile casuale appartenente alla famiglia esponenziale (Binomiale, Poisson, Gamma, ecc.) La regressione è utilizzata per classificare le osservazioni in base alle caratteristiche di riferimento nelle due categorie della variabile dipendente.

Analisi Discriminante Lineare

Il metodo LDA (*Linear Discriminant Analysis*) è utilizzato per trovare una combinazione lineare che caratterizza o separa due classi di oggetti o eventi. La differenza con i metodi logistici è che con la regressione logistica si assume che le *posterior class probability* (Adam Hayes, 2021) sono una qualche funzione delle features, ma non vi sono assunzioni sulla distribuzione di probabilità delle features. Anche l'LDA, come la regressione logistica, è un metodo di classificazione lineare poichè produce una superficie decisionale - detta anche *decision boundary* (Jason J. Corso, 2013) - lineare.

Analisi Discriminante Quadratica

Il metodo QDA (*Quadratic Discriminant Analysis*) a differenza della regressione logistica o del classificatore LDA, utilizza una superficie decisionale quadratica per separare le misurazioni di due o più classi di oggetti o eventi. Diversamente dall'assunzione di normalità per ciascuna classe k dell'analisi discriminante lineare, nell'analisi discriminante quadratica, *in ogni classe k la distribuzione congiunta delle features è una Normale multivariata centrata sul vettore medio* $\boldsymbol{\mu_k}$, *ed avente matrice di covarianza* $\boldsymbol{\Sigma_k}$ (Trevor Hastie & Robert Tibshirani, 2013). La separazione tra classi non avviene tramite rette, ma con archi di parabola, proprio perché il metodo utilizzato è quadratico e non lineare. La diversa forma della superficie decisionale prodotta da questi due metodi permette di comprendere che l'analisi discriminante quadratica è una generalizzazione dell'analisi discriminante lineare; ecco perché il metodo QDA contiene come caso innestato il metodo LDA, ovvero quando $\boldsymbol{\Sigma_k = \Sigma}$. I due metodi classificano in modo analogo, con la differenza che per l'analisi discriminante quadratica bisogna stimare più parametri e quindi è un

metodo più complesso, si adatta meglio ai dati di addestramento e dà più problemi quando si deve ottenere una stima accurata dell'error rate.

Albero di Decisione

Il metodo Decision Tree appartiene alla famiglia di metodi di apprendimento supervisionato. A differenza di altri metodi della stessa famiglia, un albero di decisione può essere utilizzato per risolvere sia problemi di regressione che di classificazione, come il caso in esame. Un albero di decisione è un metodo che prevede la suddivisione continua dei dati in base ad un determinato parametro ed è un modo utile per visualizzare un algoritmo che contiene solo istruzioni di controllo condizionale. Il motivo generale dell'utilizzo di un albero di decisione è quello di creare una regola di addestramento che può essere utilizzato per prevedere la classe o il valore della variabile di riferimento imparando da regole decisionali dedotte dai dati di addestramento. L'albero di decisione è composto da:

1. **Nodo**: verifica il valore di un determinato attributo.
2. **Bordi/Ramo**: corrisponde all'esito di un test e si connette al nodo o foglia successiva.
3. **Nodo Foglia**: nodo terminale che prevede il risultato e rappresenta etichette di classe o distribuzione di classi (Afroz Chakure, 2019).

Ogni nodo interno rappresenta una variabile, un arco verso un nodo figlio rappresenta un possibile valore per quella proprietà ed una foglia rappresenta il valore previsto per la variabile obiettivo a partire dai valori delle altre proprietà, che nell'albero è rappresentato dal cammino (*path*) dal nodo radice (*root*) al nodo foglia. Ogni nodo rappresenta un *"test"* da compiere su un attributo, ogni ramo rappresenta l'esito del test ed ogni nodo foglia rappresenta un'etichetta di classe, ovvero la decisione presa dopo aver calcolato tutti gli attributi; il processo consiste in una sequenza di test che comincia sempre dal nodo radice, ovvero il nodo genitore situato più in alto nella struttura, e procede verso il basso. A seconda dei valori rilevati in ciascun nodo, il flusso prende una direzione oppure un'altra e procede progressivamente verso il basso. La decisione finale si trova nei nodi foglia terminali, ovvero quelli più in basso. In questo modo, dopo aver analizzato le varie

condizioni, l'agente giunge alla decisione finale. Un albero di decisione può essere utilizzato come nel caso in esame, per problemi di classificazione; in questo caso la classificazione è descritta dall'albero, in cui i nodi foglia rappresentano le classificazioni e le ramificazioni rappresentano l'insieme delle proprietà che portano a quelle classificazioni. Di conseguenza ogni nodo interno risulta essere una macro-classe costituita dall'unione delle classi associate ai nodi figli. I percorsi dalla radice alla foglia rappresentano le regole di classificazione ed il predicato che si associa ad ogni nodo interno, sulla base del quale avviene la ripartizione dei dati, è chiamato *condizione di split*. Per il problema in esame si utilizza un albero di classificazione, dove il risultato è una variabile categoriale. Tale albero è costruito attraverso l'approccio del *partizionamento ricorsivo binario* (Afroz Chakure, 2019), ovvero un processo iterativo di suddivisione dei dati in partizioni e quindi ulteriori suddivisioni in ciascuno dei rami. Il partizionamento ricorsivo si avvale dell'approccio *divide et impera* poiché suddivide i dati in sottoinsiemi che vengono poi suddivisi ripetutamente in sottoinsiemi ancora più piccoli e così via fino a quando il processo si interrompe se l'algoritmo determina che i dati all'interno dei sottoinsiemi sono sufficientemente omogenei, o è stato soddisfatto un altro *criterio di arresto (haiting)*. In molte situazioni è utile definire un criterio di arresto o anche *criterio di potatura (pruning)* al fine di determinare la profondità massima dell'albero. Questo perché il crescere della profondità di un albero, ovvero della sua dimensione, non influisce direttamente sulla bontà del metodo. Infatti, una crescita eccessiva della dimensione dell'albero può portare solo ad un aumento sproporzionato della complessità computazionale. L'algoritmo divide et impera, combinato all'albero di decisione, prevede i seguenti step:

1. Seleziona un test per il nodo radice e crea un ramo per ogni possibile esito del test.
2. Divide le istanze in sottoinsiemi, uno per ogni ramo che si estende dal nodo.
3. Ripete l'operazione ricorsivamente per ogni ramo, utilizzando solo le istanze che raggiungono il ramo.
4. Arresta la ricorsione per un ramo se tutte le sue istanze hanno la stessa classe.

Un albero di classificazione è molto simile ad un albero di regressione, tranne per il fatto che nella classificazione l'output è qualitativo e non quantitativo. Per un albero di classificazione ogni osservazione appartiene alla classe di osservazioni di addestramento più comune nella regione a cui appartiene. Dal momento che si pianifica la classificazione per assegnare un'osservazione in una data regione, alla classe di osservazioni di addestramento più comune in quella regione, il tasso di errore di classificazione è semplicemente la frazione delle osservazioni di addestramento in quella regione che non appartengono alla classe più comune (Trevor Hastie & Robert Tibshirani, 2013).

$$classification\ error\ rate:\ E = 1 - max_k(\hat{p}mk)$$

Qui, $\hat{p}mk$ rappresenta la proporzione di osservazioni di addestramento nella regione $m-esima$ che provengono dalla classe $k-esima$. Tuttavia, risulta che l'errore di classificazione non è sufficientemente sensibile per la *"crescita"* degli alberi e nei casi pratici sono preferibili le misure che saranno definite in seguito. La sfida principale nell'implementazione dell'albero di decisione è identificare quali attributi si devono considerare come nodo radice ad ogni livello. Gestire questo problema significa conoscere la selezione degli attributi. Ci sono diverse misure di selezione degli attributi per identificare quello che può essere considerato come principale ad ogni livello, come:

1. Guadagno di informazioni (Gain)
2. Indice di Gini.

Se il set di dati è costituito da p attributi, decidere quale posizionare alla radice o ai diversi livelli dell'albero come nodi interni è un passaggio complicato. Selezionare casualmente qualsiasi nodo come root non è la soluzione al problema, infatti seguire un approccio casuale porterà a risultati poco precisi. I due criteri sopra citati calcolano i valori per ogni attributo; questi valori vengono poi ordinati e gli attributi vengono inseriti nell'albero di decisione seguendo l'ordine, ovvero

l'attributo con un valore più alto, in caso di guadagno di informazioni, viene posizionato alla radice. Utilizzando l'information Gain come criterio, si presuppone che gli attributi siano categoriali e, per l'indice di Gini si presume che gli attributi siano continui. Utilizzando il guadagno di informazioni come criterio, si cerca di stimare le informazioni contenute in ciascun attributo. Per misurare l'incertezza di una variabile casuale X si utilizza *l'Entropia*, il cui calcolo su ogni attributo permette di definire il rispettivo *guadagno di informazioni* che calcola la riduzione prevista dell'entropia dovuta all'ordinamento sull'attributo. Il guadagno di informazioni può essere definito come *la quantità di informazioni o segnale ottenuta su una variabile casuale dall'osservazione di un'altra variabile casuale*. Nella teoria degli alberi di decisione, questo termine viene anche utilizzato come sinonimo di *mutua informazione*, cioè come la misura della dipendenza reciproca tra due variabili. Più nel dettaglio, essa quantifica il contenuto informativo ottenuto su una variabile casuale osservando l'altra variabile casuale. Il concetto di mutua informazione è strettamente legato al concetto di *entropia* di una variabile casuale, definito come il *livello medio di informazione o incertezza inerente ai possibili risultati della variabile casuale* oppure come la quantità media di informazioni veicolate da un evento, considerando tutti i possibili esiti. Data una variabile casuale discreta X, con possibili esiti $x_1, \dots, x_n$, che si verificano con probabilità $P(x_1), \dots, P(x_n)$, l'entropia di X è formalmente definita come:

$$D = -\sum_{k=1}^{K} \hat{p}mk \log \hat{p}mk$$

L'indice di Gini è definito come:

$$G = \sum_{k=1}^{K} \hat{p}mk \left(1 - \hat{p}mk\right)$$

ed è una misura della varianza totale tra le classi K. Esso assume valori piccoli se tutti gli $\hat{p}mk$ sono prossimo a zero o uno. Per questo motivo, l'indice di Gini viene indicato come una misura

della *purezza del nodo*; un valore piccolo indica che un nodo contiene prevalentemente osservazioni di una singola classe. Poiché $0 \leq \hat{p}mk \leq 1$, ne deriva che $0 \leq -\hat{p}mk \log \hat{p}mk$. Si può dimostrare che l'entropia assume un valore vicino a zero se tutti i $\hat{p}mk$ sono vicini a 0 o vicini ad 1. Quindi, come l'indice di Gini, *l'entropia* assume un valore piccolo se il nodo è puro, infatti risulta che l'indice di Gini e l'entropia sono numericamente abbastanza simili. Quando si costruisce un albero di classificazione, l'indice di Gini o l'entropia sono utilizzati per valutare la qualità di una particolare suddivisione, poiché questi due approcci sono più sensibili alla purezza del nodo rispetto al tasso di errore di classificazione. È possibile utilizzare uno qualsiasi di questi tre approcci durante la fase di potatura dell'albero, ma il tasso di errore di classificazione è preferibile se l'obiettivo è l'accuratezza della previsione finale. Per problemi di classificazione viene utilizzata la funzione di costo di Gini poiché fornisce un'indicazione riguardo la purezza del nodo. Vi sono altri metodi per la selezione dei nodi come il *Chi-Quadrato* che permette di confrontare le proporzioni e quindi di creare alberi di decisioni verificando dove la proporzione è maggiore, o il *guadagno d'informazione* ovvero la ratio che permette di individuare l'attributo che restituisce il guadagno d'informazione maggiore nello stabilire l'appartenenza di un item ad una classe. Quando si hanno molte variabili, però, può capitare che la complessità dell'albero sia troppo alta per ottenere dei risultati accurati, in quanto l'albero è costituito da un numero di regole eccessivamente elevato e quindi ricade nel problema dell'*overfitting*, ossia l'eccessiva aderenza dei risultati al dataset di addestramento che quindi rende il metodo inutile ai fini predittivi. Per "sfoltire" l'albero si utilizzano delle tecniche di *pruning*, ossia di *"potatura"*, che consentono di eliminare i rami meno significativi ossia quelli che contengono pochi casi e sono poco rilevanti. La fase di *pruning* permette di mantenere un albero di decisione entro determinate dimensioni in modo da prevenire l'overfitting e ridurre al minimo gli errori di classificazione. Ci sono due tipologie di pruning:

1. **Pre-pruning**: consiste nel decidere prima di avviare la costruzione del metodo, di tagliare una parte dei dati.

2. **Post-pruning**: consiste nel generare prima l'albero e poi effettuare la potatura.

Un'alternativa utile per evitare l'overfitting sugli alberi di decisione ed alternativa al pruning, è fissare una *regola d'arresto*. Stabilire una regola d'arresto significa fissare il numero minimo di osservazioni per ogni nodo, il numero massimo di split oppure un valore minimo per il criterio di splitting. In conclusione, il vantaggio dell'utilizzo di un albero di decisione sta nell'economicità della costruzione e dell'utilizzo, infatti a differenza di altre soluzioni come le *reti neurali*, l'analista riesce a verificare come la macchina giunge alla decisione ed eventualmente dissentire. Ad esempio, per un albero di decisione applicato alla medicina che fornisce delle diagnosi, essendo una decisione importante per il paziente, è sempre opportuno che lo specialista verifichi il processo di classificazione che ha portato la macchina a prendere quella decisione. Per un professionista, quindi, è sicuramente più facile fare ciò leggendo un albero di decisione piuttosto che una rete neurale. Ciò non significa che la performance di un albero di decisione sia migliore rispetto alla performance di una rete neurale, infatti l'implementazione di una rete neurale potrebbe anche essere un criterio decisionale più efficiente, ma tuttavia più adatta alla logica della macchina e meno comprensibile per l'uomo, il che non è un aspetto da sottovalutare. Altro punto di forza è la velocità nel classificare le osservazioni sconosciute e la possibilità di gestire sia variabili numeriche che categoriali. Per contro, l'albero di decisione è un *weak learner* e può andare incontro ad errori di propagazione. Per un albero di decisione di grandi dimensioni, l'interpretazione è difficile e si complica il processo di decision making. Infine, un albero di decisione è instabile, il che significa che un piccolo cambiamento nei dati può portare ad una grande cambiamento nella struttura dell'albero di decisione ottimale ed è spesso relativamente impreciso, ma questo problema può essere risolto sostituendo un singolo albero di decisione con una *random forest* ovvero un metodo di *apprendimento d'insieme* utilizzato sia per la regressione che per la classificazione e che opera costruendo una moltitudine di alberi di decisione durante la fase di addestramento, dove l'output, per la classificazione, è la classe selezionata dalla maggior parte degli alberi. Per la regressione, invece, viene restituita la media o la previsione media dei

singoli alberi. Tuttavia, l'approccio basato sul *random forest* risolve il problema dell'overfitting a cui tende invece l'albero di decisione, ma non è facile da intepretare come un singolo albero di decisione, ciò fa capire che a priori non vi è una scelta migliore, in termini di classificatore nel problema del fraud detection, ma dipende strettamente dal caso in esame. Una soluzione che può essere adottata, e che è presente nel caso studio in esame, è quella di implementare diversi classificatori per confrontarne le performance.

Caso Studio

Il principale obiettivo del seguente caso studio è determinare le prestazioni di classificatori per il rilevamento di frodi con carta di credito su un set di dati aventi un numero diverso di osservazioni e di caratteristiche. A tal fine, i due differenti dataset utilizzati sono *"Small Card Data"* (SCD) ed *"European Card Data"* (ECD). Come in tutti i dataset relativi al problema del *fraud detection*, anche in questo caso, il numero delle osservazioni che registrano una transazione fraudolenta è molto inferiore rispetto al numero totale delle osservazioni, generando quindi un problema di *unbalanced class*. In tutti e due i dataset il valore della variabile di riferimento è *"0"* se la transazione effettuata è non fraudolenta oppure *"1"* se la transazione è fraudolenta.

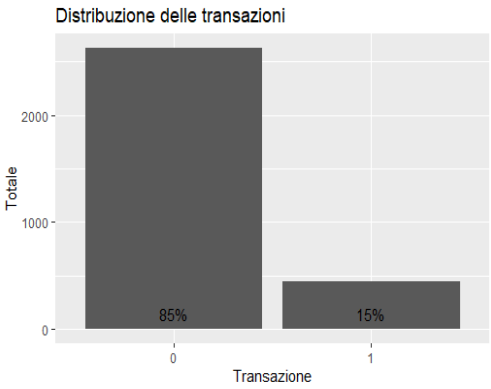
CASO 1

Il dataset *"Small Card Data"*, reperibile da *Kaggle*[1], è un set di dati contenente 3075 osservazioni per 12 variabili. La metà delle caratteristiche sono categoriali mentre l'altra metà numeriche. Di 3075 osservazioni il 14,6%, ovvero 448 transazioni, sono fraudolente come riportato nella tabella1 e nella figura1.

TABELLA 1: NUMERO DI OSSERVAZIONI TOTALE CON PERCENTUALE DI TRANSAZIONI FRAUDOLENTE E TRANSAZIONI NON FRAUDOLENTE.

Totale Transazioni	3075
Transazioni Non-Fraud	2627
Transazioni Fraud	448
% Transazioni Non Fraud	85.431
% Transazioni Fraud	14.569

FIGURA 1: DISTRIBUZIONE DELLE TRANSAZIONI.



Le variabili continue continue riportate in questo set di dati sono: il numero identificativo della carta, la data della transazione, l'importo medio della transazione per ogni giorno, l'importo della transazione, l'importo medio giornaliero del chargeback, ovvero il ritorno di denaro di una transazione disposto dalla banca che ha emesso la carta di pagamento del consumatore che annulla di fatto un trasferimento di denaro dal conto bancario; ancora l' importo medio del chargeback negli ultimi sei mesi e frequenza del chargeback negli ultimi sei mesi. Le variabili categoriali, invece, sono: una variabile che indica se la transazione è stata rifiutata o meno, numero di transazioni rifiutate ogni giorno, se la transazione effettuata è rivolta all'estero e se il Paese dove è diretta la transazione è ad alto rischio o meno. Infine, la variabile dipendente categoriale *isFradulent* indica se una transazione è fraudolenta o meno. La prima operazione di pre-processing dei dati riguarda la verifica dei valori mancanti all'interno del dataset.

TABELLA 2: VALORI MANCANTI PRESENTI ALL'INTERNO DEI DATI.

	FALSE	TRUE
NA	33825	3075

Si registrano 3075 valori mancanti. Da un'analisi più dettagliata, inoltre, emerge che i valori mancanti sono tutti relativi ad un'unica variabile, *Transaction.date* ovvero la data della transazione. Per tale motivo, al fine di rendere più agevole l'analisi si elimina tale variabile. Le variabili categoriali riportato i valori *"Y"* ed *"N"* e per tale motivo sono ricodificate rispettivamente con i valori *"1"* e *"0"* e trasformate in fattori a due livelli. Tale operazione è stata effettuata per le seguenti variabili: *Is.declined*, *isFradulent*, *isForeignTransaction* e *isHighRiskCoutry*. Successivamente si procede alla fase di analisi esplorativa che consente di catturare le principali caratteristiche dei dati. La prima analisi è stata svolta condizionando sulla variabile di riferimento e calcolando le principali statistiche descrittive, come riportate nella tabella3.

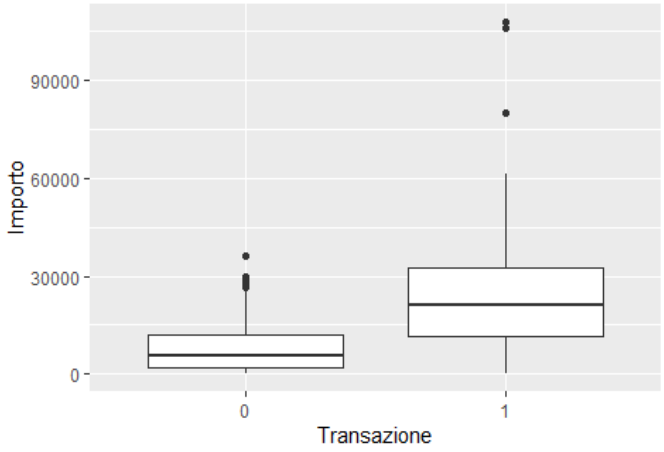
TABELLA 3: PRINCIPALI STATISTICHE DESCRITTIVE PER IL TIPO DI TRANSAZIONE.

Fraud	Total	Mean	Std	Min	Q1	Q2	Q3	Max
0	2.627	7662.996	6869.468	0	1997.918	5589.459	11752.53	36000
1	448	22855.441	15217.912	257.899	11.582,06	20944.744	32409.35	108000

Mediamente, una transazione fraudolenta ha un importo di *22855.441* dollari a fronte di un importo pari a *7662.996* dollari per la transazione non fraudolenta. Anche medianamente una transazione fraudolenta fa registrare un valore dell'importo decisamente maggiore rispetto ad una transazione non fraudolenta. Stesso discorso anche per l'importo massimo di una transazione fraudolenta che è pari a *108000* dollari contro i *36000* dollari di una transazione non fraudolenta.

Osservando i quartili al *25%* ed al *75%*, si nota che il *25%* delle transazioni fraudolente con importo maggiore superano i *32409.35* dollari mentre il *25%* delle transazioni non fraudolente con importo maggiore superano i *11752.53* dollari. Quanto emerso dalla tabella3 è riportato nella figura2 che conferma quanto detto in precedenza e riporta una visualizzazione, tramite *boxplot*, dell'importo delle transazioni condizionatamene alla classe di riferimento.

FIGURA 2: BOXPLOT DELL'IMPORTO DELLE TRANSAZIONI CONDIZIONATAMENTE AL TIPO DI TRANSAZIONE, OVVERO FRAUDOLENTA O NON FRAUDOLENTA.



Terminata l'analisi esplorativa dei dati, si procede alla modellizzazione utilizzando i metodi di classificazione citati in precedenza, ovvero la *regressione logistica*, l'*analisi discriminante lineare*, l'*analisi discriminante quadratica* e l'*albero di decisione*. Si divide il set di dati iniziale partizionandolo in un sottoinsieme contenente il 90% delle osservazioni, ovvero il set dei dati su cui il metodo viene addestrato, mentre il restante 10% delle osservazioni compone *l'external validation set*, ovvero un set di dati unseen tenuto fuori dal processo di stima. In questo caso, il set di addestramento contiene *2769* osservazioni mentre l'external validation set contiene le restanti *306* osservazioni. In problemi di unbalanced class, come il caso in esame, è importante che il partizionamento dei dati non sia casuale ma rispetti lo sbilanciamento della classe di riferimento che vi è nella struttura originale dei dati. Infatti, dopo aver partizionato i dati è necessario verificare la distribuzione delle classi.

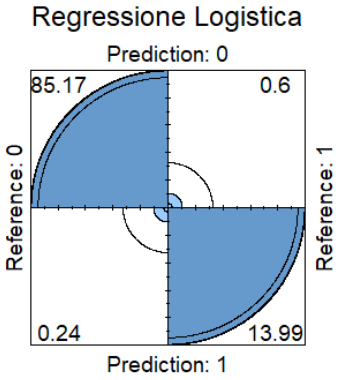
TABELLA 4: SBILANCIAMENTO DELLE OSSERVAZIONI NEL SET ORIGINALE DI DATI, NEL SET DI ADDESTRAMENTO E NELL'EXTERNAL VALIDATION SET.

	0	1
Dataset	85.431 %	14.569 %
Set Addestramento	85.410 %	14.590 %
External Validation Set	85.621 %	14.379 %

Ovviamente, le percentuali relative alla distribuzione della classe di riferimento non saranno esattamente le stesse, ma si vede come è rispettato il rapporto *85:15* presente nel dataset originale. Quindi, dopo aver partizionato i dati in set di addestramento ed external validation set ed aver confrontato la distribuzione della classe con quella originale, si specifica il metodo da addestrare per poi valutarne la performance. Per convalidare il metodo si utilizza il processo della *K-Folds Cross Validation con 10-Folds* ripetuta per *20 volte* sul set di addestramento, così da ottenere una media per ogni fold del valore delle metriche utilizzate; mentre l'external validation set viene utilizzato per calcolare il singolo valore delle metriche utilizzate su un set di dati che l'algoritmo non hai mai visto prima e scelto casualmente. Il primo classificatore utilizzato è la **regressione logistica** ed osservando la *Confusion matrix* e le metriche sopra citate, i risultati con la *KFCV* sono i seguenti:

TABELLA 5 – FIGURA 3: MATRICE DI CONFUSIONE IN KFCV DELLA REGRESSIONE LOGISTICA.

	Riferimento	
Previsione	**0**	**1**
0	85.17 %	0.6 %
1	0.24 %	13.99 %



Dalla tabella5 della matrice di confusione si evince che la *regressione logistica* prevede correttamente, in media, *l'85%* dei *veri negativi*, ovvero transazioni osservate come non fraudolente sono classificate come tali sbagliando solo lo *0.2%* delle volte ovvero la percentuale media di *falsi positivi* e quindi lo *0.2%* delle transazioni che sono osservate come non fraudolente sono erroneamente classificate come fraudolente. Tuttavia, per quanto detto in precedenza, i *falsi positivi* in problemi di fraud detection sono importanti per la valutazione generale del metodo ma non rappresentano la misura d'errore più importante che si vuole minimizzare. Infatti, è più importante minimizzare la percentuale di *falsi negativi* che in questo caso è pari allo *0.6%* vale a dire 6 falsi negativi ogni *1000* transazioni. Valutando altre misure di performance i risultati sono i seguenti:

TABELLA 6 – TABELLA 7: METRICHE DI PERFORMANCE DELLA REGRESSIONE LOGISTICA IN KFCV.

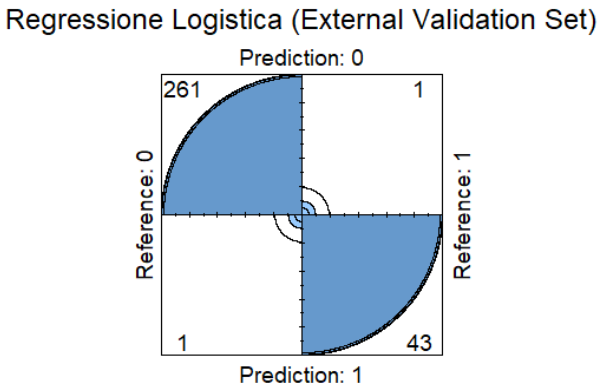
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.992	0.959	0.958	0.984	0.984

F1 Score	F2 Score	AUC	MCC	KCohen
0.970	0.963	0.978	0.966	0.966

Utilizzando l'external validation set per valutare l'errore commesso su un singolo set di dati scelto casualmente, la regressione logistica registra comunque delle ottime performance, infatti:

TABELLA 8 – FIGURA 4: MATRICE DI CONFUSIONE DELLA REGRESSIONE LOGISTICA SULL'EXTERNAL VALIDATION SET.

Riferimento		
Previsione	**0**	**1**
0	261	1
1	1	43



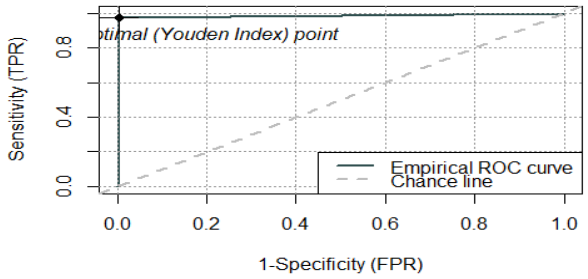
Anche con un set di dati esterno non utilizzato durante il processo di stima e scelto casualmente, la regressione logistica offre delle ottime performance generando un solo caso di *falso positivo* ed un solo caso di *falso negativo*. Valutando altre misure di performance i risultati sono i seguenti:

TABELLA 9 – TABELLA 10: METRICHE DI PERFORMANCE DELLA REGRESSIONE LOGISTICA SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.994	0.977	0.996	0.977	0.996

F1 Score	F2 Score	AUC	MCC	KCohen
0.977	0.977	0.987	0.973	0.974

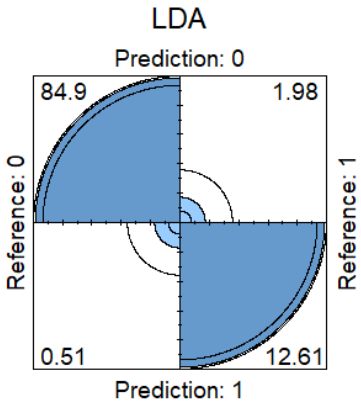
FIGURA 5: CURVA ROC DELLA REGRESSIONE LOGISTICA SULL'EXTERNAL VALIDATION SET. VALORE MOLTO VICINO AD 1, IL CHE INDICA UN'ELEVATA SENSITIVITY ED UN BASSO FALSE POSITIVE RATE.



Per l'**analisi discriminante lineare** i risultati con la *KFCV* sono i seguenti:

TABELLA 11 – FIGURA 6: MATRICE DI CONFUSIONE IN KFCV DELL'LDA

Riferimento		
Previsione	**0**	**1**
0	84.9 %	1.98 %
1	0.51 %	12.61 %



Rispetto alla regressione logistica si ha una percentuale media maggiore sia di *falsi negativi (2%)* che di *falsi positivi (0.5%)*. Anche valutando le altre metriche di performance, nonostante LDA offra ottime performance, esse sono comunque inferiore a quelle della regressione logistica, infatti:

TABELLA 12- TABELLA 13: METRICHE DI PERFORMANCE DELL'LDA IN KFCV.

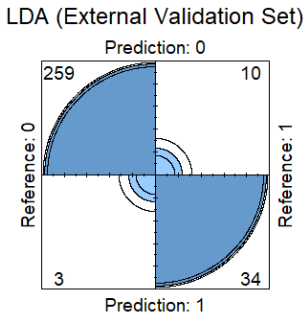
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.975	0.865	0.865	0.962	0.962

F1 Score	F2 Score	AUC	MCC	KCohen
0.910	0.882	0.933	0.898	0.895

Utilizzando l'external validation set, si nota come il classificatore LDA sul singolo campione generi *3 falsi positivi* e *10 falsi negativi*.

TABELLA 14 – FIGURA 7: MATRICE DI CONFUSIONE DELL'LDA SULL'EXTERNAL VALIDATION SET.

Riferimento		
Previsione	**0**	**1**
0	259	10
1	3	34



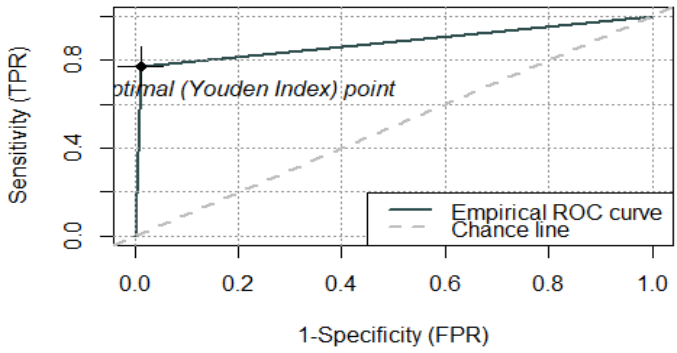
Valutando altre misure di performance i risultati sono i seguenti e si nota un calo legato ad alcune metriche come *Sensitivity* e F_2, rispetto al valore ottenuto con la KFCV, infatti:

TABELLA 15 – TABELLA 16: METRICHE DI PERFORMANCE DEL'LDA SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.958	0.773	0.988	0.929	0.963

F1 Score	F2 Score	AUC	MCC	KCohen
0.840	0.798	0.881	0.820	0.820

FIGURA 8: CURVA ROC DELL'LDA OTTENUTA SULL'EXTERNAL VALIDATION SET.

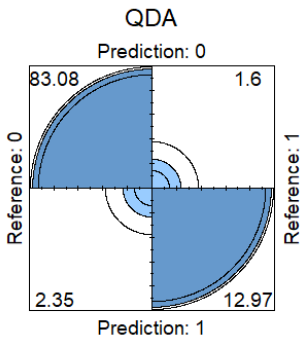


Dal confronto tra i primi due classificatori emerge la maggiore affidabilità della regressione logistica. Prendendo come riferimento le performance di questo classificatore, quindi, l'obiettivo è migliorarle utilizzando altri due classificatori, ovvero l'analisi discriminante quadratica e l'albero di decisione.

Per l'**analisi discriminante quadratica** i risultati con la *KFCV* sono i seguenti:

TABELLA 17 – FIGURA 9: MATRICE DI CONFUSIONE IN KFCV DELLA QDA.

Riferimento		
Previsione	**0**	**1**
0	83.08 %	1.6 %
1	2.35 %	12.97 %



Con questo classificatore la percentuale media di *falsi positivi (2.35%)* è maggiore sia rispetto a quella generata dal classificatore LDA sia rispetto alla regressione logistica, quindi il metodo QDA non si comporta meglio rispetto ai classificatori precedenti riguardo le transazioni non

fraudolente. Tuttavia, riguardo le transazioni fraudolente, fa registrare una percentuale media di *falsi negativi (1.6%)* minore rispetto a quella ottenuta con il classificatore LDA, ma maggiore rispetto alla regressione logistica. Anche valutando le altre metriche, si evince come le performance risultino essere inferiori rispetto alla regressione logistica, infatti:

TABELLA 18 – TABELLA 19: METRICHE DI PERFORMANCE DELLA QDA IN KFCV.

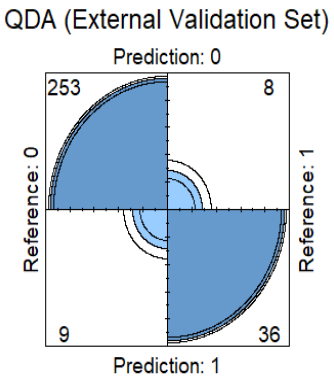
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.961	0.890	0.882	0.850	0.850

F1 Score	F2 Score	AUC	MCC	KCohen
0.868	0.881	0.931	0.846	0.845

Utilizzando l'external validation set la performance del classificatore QDA è la seguente:

TABELLA 20 – FIGURA 10: MATRICE DI CONFUSIONE DELLA QDA SULL'EXTERNAL VALIDATION SET

Riferimento		
Previsione	**0**	**1**
0	253	8
1	9	36



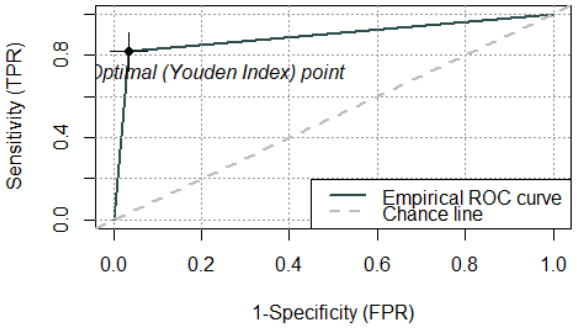
Rispetto ai risultati ottenuti con la regressione logistica sull'external validation set, si nota come sia il numero di *falsi positivi (9)* che il numero di *falsi negativi (8)* sono maggiori. Rispetto al classificatore LDA, invece, il numero di falsi positivi è maggiore, mentre per i falsi negativi il classificatore QDA si comporta meglio. Le altre misure di performance riportano i seguenti risultati:

TABELLA 21 – TABELLA 22: METRICHE DI PERFORMANCE DELLA QDA SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.944	0.818	0.966	0.800	0.970

F1 Score	F2 Score	AUC	MCC	KCohen
0.810	0.814	0.892	0.777	0.777

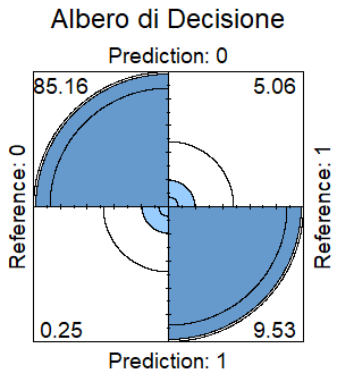
FIGURA 11: CURVA ROC DELLA QDA SULL'EXTERNAL VALIDATION SET.



Per l'**albero di decisione** i risultati con la *KFCV* sono i seguenti:

TABELLA 23 – FIGURA 12: MATRICE DI CONFUSIONE DELL'ALBERO DI DECISIONE IN KFCV.

Riferimento		
Previsione	**0**	**1**
0	85.16 %	5.06 %
1	0.25 %	9.53 %



L'albero di decisione genera una percentuale media di *falsi positivi(0.2%)* uguale a quella generata dalla regressione logistica e quindi inferiore a quella del classificatore LDA, quindi l'albero di decisione classifica abbastanza bene le transazioni non fraudolente. Tuttavia, rispetto alla regressione logistica genera una percentuale media di *falsi negativi (5%)* maggiore e quindi

non è il metodo da preferire. Valutando le altre metriche, l'albero di decisione offre performance inferiori rispetto alla regressione logistica.

TABELLA 24 – TABELLA 25: METRICHE DI PERFORMANCE DELL'ALBERO DI DECISIONE IN KFCV.

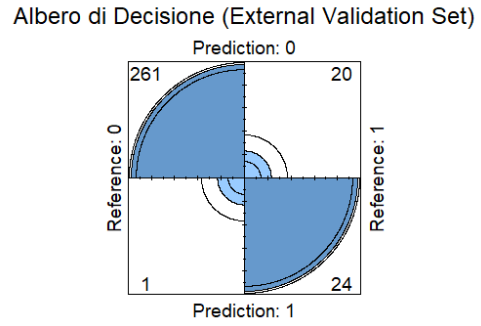
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.947	0.653	0.652	0.977	0.977

F1 Score	F2 Score	AUC	MCC	KCohen
0.778	0.597	0.825	0.772	0.750

Utilizzando l'external validation set la performance dell'albero di decisione è la seguente:

TABELLA 26 – FIGURA 13: MATRICE DI CONFUSIONE DELL'ALBERO DI DECISIONE SULL'EXTERNAL VALIDATION SET.

	Riferimento	
Previsione	**0**	**1**
0	261	20
1	1	24



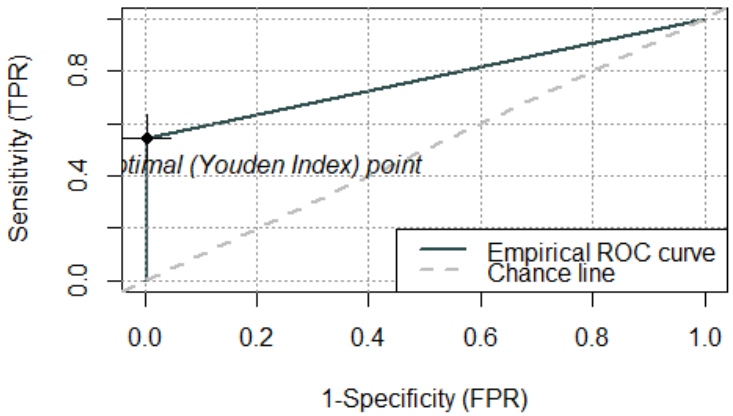
Rispetto ai risultati ottenuti con la regressione logistica sull'external validation set, si nota come il numero di *falsi positivi (1)* sia uguale , ciò significa che sul campione esterno di riferimento sia la regressione logistica che l'albero di decisione classificano in maniera quasi ottimale le transazioni non fraudolente. La differenza tra i due classificatori sta nei *falsi negativi*, che con l'albero di decisione diventano *20* a fronte di un solo falso positivo con la regressione logistica. Ciò influisce in maniera significativa sulle misure di performance, che riportano i seguenti risultati:

TABELLA 27 – TABELLA 28: METRICHE DI PERFORMANCE DELL'ALBERO DI DECISIONE IN SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.931	0.545	0.996	0.960	0.938

F1 Score	F2 Score	AUC	MCC	KCohen
0.706	0.607	0.771	0.694	0.660

FIGURA 14: CURVA ROC DELLL'ALBERO DI DECISIONE SULL'EXTERNAL VALIDATION SET.



Da una valutazione complessiva dei classificatori e delle metriche utilizzate per valutare la capacità previsiva, segue il quadro definitivo di ogni metodo con i rispettivi risultati ottenuti con la *KFCV*:

TABELLA 29: SINTESI DELLE METRICHE DI PERFORMANCE DI TUTTI I CLASSIFICATORI IN KFCV.

	F1	F2	Precision	Sensitivity	AUC	MCC	KCohen
LOGIT	***0.970***	***0.963***	***0.984***	***0.959***	***0.978***	***0.966***	***0.966***
LDA	0.910	0.882	0.962	0.865	0.933	0.898	0.895

QDA	0.868	0.881	0.850	0.890	0.931	0.846	0.845
DT	0.778	0.697	0.977	0.653	0.825	0.772	0.750

Ad esclusione dell'albero di decisione che restituisce i valori di performance più bassi, soprattutto per il punteggio F_2 e per la *Sensitivity*, sia il metodo LDA che QDA realizzano delle ottime performance; tuttavia, la regressione logistica massimizza sia la metrica F_2 che la *Sensitivity*, ovvero le metriche di riferimento per il problema in esame, ma anche tutte le altre metriche calcolate e ciò porta alla scelta della regressione logistica come miglior classificatore.

Per quanto riguarda le performance ottenute da ogni classificatore sull'external validation set, i risultati sono piuttosto concordanti con quelli ottenuti in precedenza, con la regressione logistica che massimizza tutte le metriche e si differenzia in maniera importante dagli altri classificatori.

TABELLA 30: SINTESI DELLE METRICHE DI PERFORMANCE DI TUTTI I CLASSIFICATORI SULL'EXTERNAL VALIDATION SET.

External Validation Set *306 samples, 10 predictor, 2 classes: "0", "1"*							
	F1	**F2**	**Precision**	**Sensitivity**	**AUC**	**MCC**	**KCohen**
LOGIT	**0.977**	***0.977***	***0.977***	***0.977***	***0.987***	***0.973***	***0.974***
LDA	0.840	0.798	0.929	0.773	0.991	0.820	0.820
QDA	0.810	0.814	0.800	0.818	0.892	0.777	0.777
DT	0.706	0.607	0.960	0.545	0.771	0.694	0.660

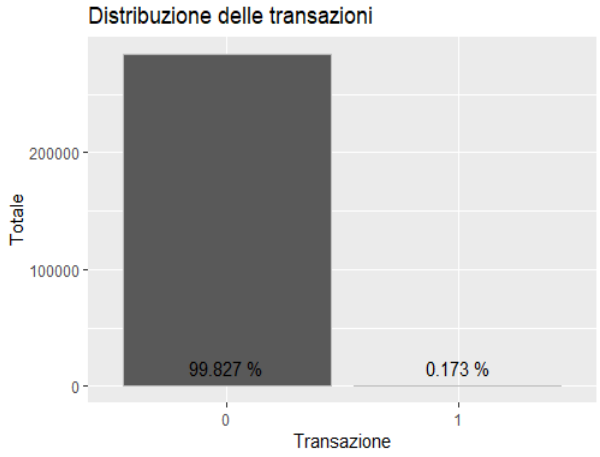
CASO 2

Il dataset "European Card Data", reperibile da *Kaggle*[2], è un set di dati contenente le transazioni effettuate con carte di credito nel settembre 2013 da titolari di carte europee e presenta le transazioni effettuate in due giorni, con lo 0.173% delle transazioni fraudolente, ovvero 492 osservazioni su un totale di 284807, come riportato nella tabella31 e nella figura15.

TABELLA 31: NUMERO DI OSSERVAZIONI TOTALE CON PERCENTUALE DI TRANSAZIONI FRAUDOLENTE E TRANSAZIONI NON FRAUDOLENTE.

FIGURA 15: DISTRIBUZIONE DELLE TRANSAZIONI.

N. Tot. di transazioni	284807
Transazioni Non-Fraud	284315
Transazioni Fraud	492
% Transazioni Non Fraud	0.173
% Transazioni Fraud	99.827



Il dataset contiene solo variabili numeriche che sono il risultato di una trasformazione utilizzando la PCA. Per motivi di riservatezza non è possibile fornire le caratteristiche originali e ulteriori informazioni di base sui dati. Le caratteristiche V1, V2, …, V28 sono le componenti principali ottenute con la PCA, le uniche caratteristiche non trasformate con la PCA sono "*Time*" ed "*Amount*". La variabile "*Time*" riporta i secondi trascorsi tra ogni transazione e la prima transazione nel set di dati. La variabile "*Amount*" è l'importo della transazione. "*Classe*" è la

variabile di risposta. Il dataset è stato raccolto ed analizzato durante una collaborazione di ricerca di $Worldline$ e del $Machine\ Learning\ Grup$ dell'$ULB(Universit$è $Libre\ de\ Bruxelles)$ sui big data mining e rilevamento delle frodi. Come per il caso precedente, la prima operazione di pre-processing dei dati è la verifica di valori mancanti che non sono presenti all'interno del dataset. Successivamente è stata formattata la variabile di risposta $Class$ come fattore. Nella fase successiva di analisi esplorativa, condizionatamente al tipo di transazione sono calcolate le principali statistiche descrittive per la variabile $Amount$, riportate nella tabella32 che segue:

TABELLA 32: PRINCIPALI STATISTICHE DESCRITTIVE PER IL TIPO DI TRANSAZIONE.

Class	Total	Mean	Std	Min	Q1	Q2	Q3	Max
0	284315	88.3	250	0	5.65	22	77	25691
1	492	122	257	0	1	9.25	106	2126

Mediamente una transazione fraudolenta ha un importo di $Amount$ pari a *122* dollari e maggiore rispetto all'importo medio di *88.3* dollari per una transazione non fraudolenta. Discorso opposto se si considera il valore mediano che è superiore per una transazione non fraudolenta rispetto ad una transazione fraudolenta ed anche per il valore massimo le transazioni non fraudolente fanno registrare un valore di $Amount$ maggiore. Come nel caso precedente, si divide il set di dati iniziale partizionandolo in un sottoinsieme contenente il 90% delle osservazioni utilizzate per la stima del metodo ed il restante 10% delle osservazioni utilizzate per *l'external validation set*. Il set di addestramento contiene *256326* mentre il validation set le restanti *28481* osservazioni. Anche in questo caso è importante che il partizionamento dei dati rispetti lo sbilanciamento della classe di riferimento che vi è nella struttura originale dei dati. Dopo aver partizionato i dati è necessario verificare la distribuzione delle classi.

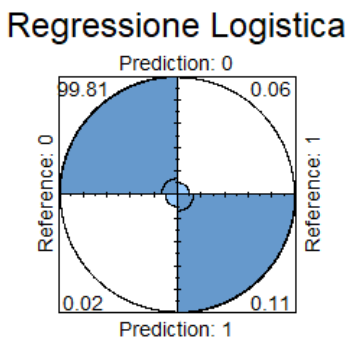
TABELLA 33: PRINCIPALI STATISTICHE DESCRITTIVE PER IL TIPO DI TRANSAZIONE.

	0	1
Dataset	99.827 %	0.173 %
Set Addestramento	99.827 %	0.173 %
External Validation Set	99.828 %	0.172 %

Si evince come è rispettato lo sbilanciamento presente nel dataset originale. Quindi, dopo aver partizionato i dati e confrontato la distribuzione della classe con quella originale, si specifica il metodo da addestrare per poi valutarne la performance. Per convalidare il metodo si utilizza prima il processo della *K-Folds Cross Validation con 10-Folds* ripetuta per *20 volte* sul set di addestramento, e successivamente l'external validation set. Il primo classificatore utilizzato è la **regressione logistica** ed i risultati con la *KFCV* sono i seguenti:

TABELLA 34 – FIGURA 16: MATRICE DI CONFUSIONE DELLA REGRESSIONE LOGISTICA IN KFCV.

Riferimento		
Previsione	**0**	**1**
0	99.81 %	0.06 %
1	0.02 %	0.11 %



Dalla tabella34 della matrice di confusione si evince come la regressione logistica preveda in media correttamente il *99.81%* dei *veri negativi* sbagliando mediamente nello *0.02%* dei casi, volte in cui si genera un *falso positivo*. La percentuale media di *falsi negativi* è pari a *0.06%*, vale a dire che su *10.000* transazioni osservate, in media, ci saranno *6* transazioni fraudolente classificate come non fraudolente. Valutando altre misure di performance i risultati sono i seguenti:

TABELLA 35 – TABELLA 36: METRICHE DI PERFORMANCE DELLA REGRESSIONE LOGISTICA IN KFCV.

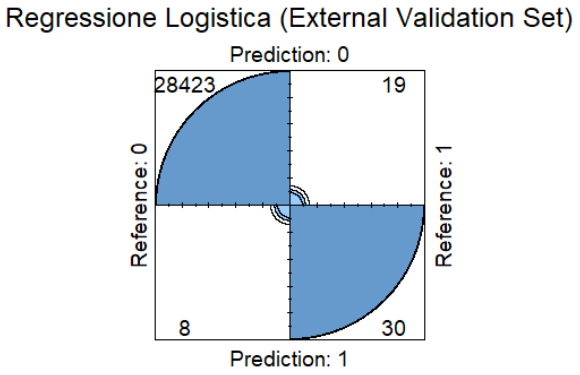
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.999	0.639	0.999	0.880	0.999

F1 Score	F2 Score	AUC	MCC	KCohen
0.737	0.674	0.820	0.747	0.737

Utilizzando l'external validation set, la regressione logistica riporta la seguente performance:

TABELLA 37 – FIGURA 17: MATRICE DI CONFUSIONE DELLA REGRESSIONE LOGISTICA SULL'EXTERNAL VALIDATION SET.

Riferimento		
Previsione	**0**	**1**
0	28423	19
1	8	30



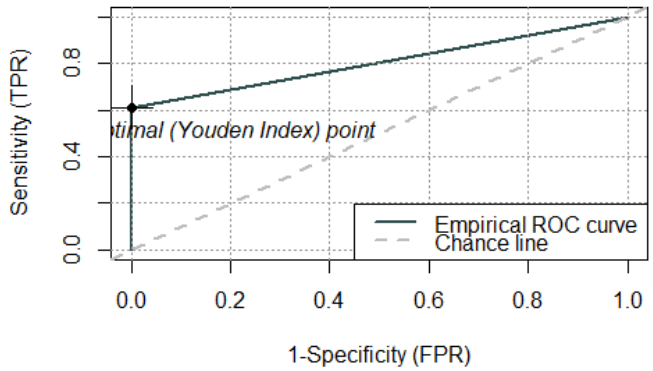
Con un set di dati esterno non utilizzato durante il processo di stima e scelto casualmente, la regressione logistica genera *8 falsi positivi* e *19 falsi negativi*. Ciò avrà un impatto significativo sulle misure riportate di seguito:

TABELLA 38 – TABELLA 39: METRICHE DI PERFORMANCE DELLA REGRESSIONE LOGISTICA SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.999	0.612	0.999	0.789	0.999

F1 Score	F2 Score	AUC	MCC	KCohen
0.690	0.641	0.806	0.695	0.699

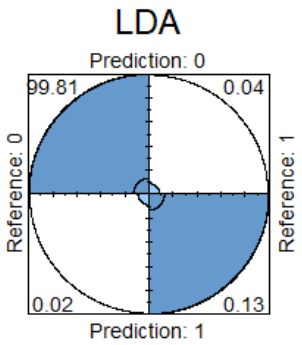
FIGURA 18: CURVA ROC DELLA REGRESSIONE LOGISTICA SULL'EXTERNAL VALIDATION SET.



Per l'**analisi discriminante lineare** i risultati con la *KFCV* sono i seguenti:

TABELLA 40– FIGURA 19: MATRICE DI CONFUSIONE DELL'LDA IN KFCV.

Riferimento		
Previsione	**0**	**1**
0	99.81 %	0.04 %
1	0.02 %	0.13 %



Rispetto alla regressione logistica si registra una percentuale media uguale di *falsi positivi(0.02%)* ed una riduzione della percentuale media di *falsi negativi(0.04%)*, il che fa

risultare il classificatore LDA più affidabile rispetto alla regressione logistica. Anche valutando le altre metriche di performance il classificatore LDA offre una performance migliore rispetto alla regressione logistica per ogni metrica considerata, infatti:

TABELLA 41– TABELLA 42: METRICHE DI PERFORMANCE DELL'LDA IN KFCV.

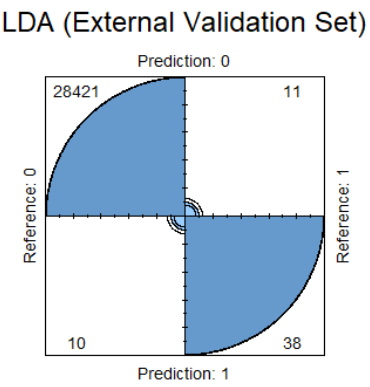
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.999	0.765	0.999	0.876	0.999

F1 Score	F2 Score	AUC	MCC	KCohen
0.815	0.784	0.882	0.820	0.814

Utilizzando l'external validation set, la performance del classificatore LDA è la seguente:

TABELLA 43– FIGURA 20: MATRICE DI CONFUSIONE DELL'LDA SULL'EXTERNAL VALIDATION SET.

Riferimento		
Previsione	**0**	**1**
0	28421	11
1	10	38



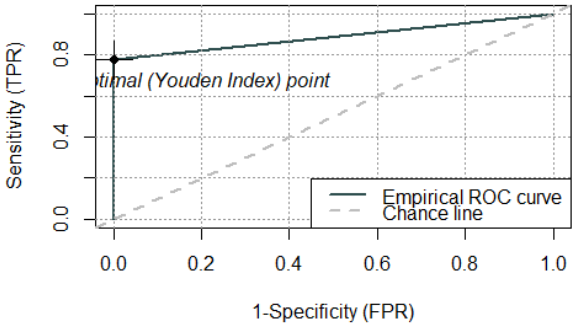
Anche sul singolo campione di validation, il classificatore LDA genera *11 falsi negativi* a fronte dei *19* generati dalla regressione logistica ed il numero dei *falsi positivi* subisce un leggero aumento, passando dagli *8* della regressione logistica ai *10* del classificatore LDA. Valutando altre misure di performance i risultati sono i seguenti:

TABELLA 44– TABELLA 45: METRICHE DI PERFORMANCE DELL'LDA SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.999	0.776	0.999	0.792	0.999

F1 Score	F2 Score	AUC	MCC	KCohen
0.784	0.779	0.888	0.820	0.783

FIGURA 21: CURVA ROC DELL'LDA SULL'EXTERNAL VALIDATION SET.

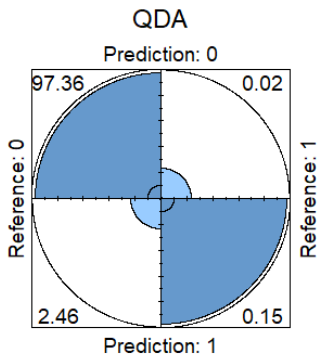


Dal confronto tra i primi due classificatori emerge la maggiore affidabilità dell'analisi discriminante lineare. Prendendo come riferimento le performance di questo classificatore, quindi, l'obiettivo è migliorarle utilizzando altri due classificatori, ovvero l'analisi discriminante quadratica e l'albero di decisione.

Per l'**analisi discriminante quadratica** i risultati con la *KFCV* sono i seguenti:

TABELLA 46– FIGURA 22: MATRICE DI CONFUSIONE DELLA QDA IN KFCV.

Riferimento		
Previsione	**0**	**1**
0	97.36 %	0.02 %
1	2.46 %	0.15 %



Con questo classificatore la percentuale media di *falsi positivi(2.46%)* è maggiore sia rispetto al classificatore LDA che alla regressione logistica, quindi il metodo QDA non fa un buon lavoro sulle transazioni non fraudolente. Tuttavia, per le transazioni fraudolente risulta addirittura migliore del classificatore LDA con una percentuale media di *falsi negativi* pari allo *0.02%*. D'altronde, una percentuale media di *falsi positivi* così alta non è un buon segnale per l'algoritmo, che tenderà a bloccare o quantomeno inviare un messaggio d'allerta a clienti che effettuano transazioni lecite, facendo risultare il sistema non in linea con la velocità della transazione che richiede il cliente. Anche valutando le altre metriche di performance con i modelli precedenti, il classificatore QDA non è da preferire.

TABELLA 47– TABELLA 48: METRICHE DI PERFORMANCE DELLA QDA IN KFCV.

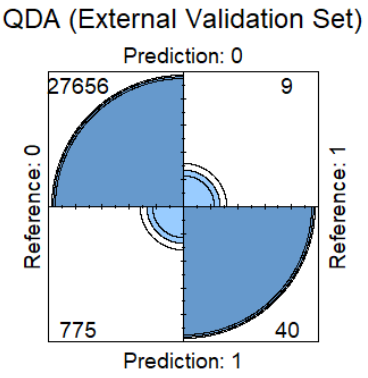
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.975	0.879	0.999	0.060	0.972

F1 Score	F2 Score	AUC	MCC	KCohen
0.110	0.230	0.930	0.222	0.106

Utilizzando l'external validation set la performance del classificatore QDA aumenta in termini di precision, ma mostra ancora un numero troppo elevato di *falsi positivi.*

TABELLA 49– FIGURA 23: MATRICE DI CONFUSIONE DELLA QDA SULL'EXTERNAL VALIDATION SET.

Riferimento		
Previsione	**0**	**1**
0	27656	9
1	775	40



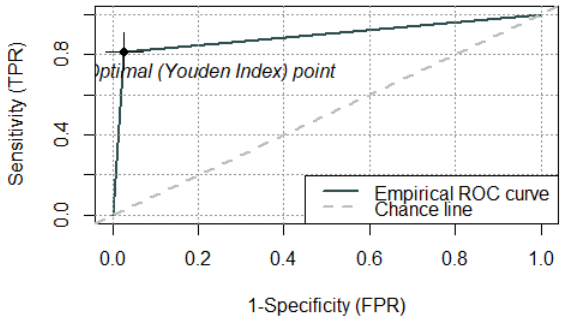
Valutando altre misure di performance i risultati sono i seguenti:

TABELLA 50– TABELLA 51: METRICHE DI PERFORMANCE DELLA QDA SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.973	0.816	0.973	0.050	0.999

F1 SCORE	F2 SCORE	AUC	MCC	KCOHEN
0.093	0.198	0.895	0.222	0.196

FIGURA 24: CURVA ROC DELLA QDA SULL'EXTERNAL VALIDATION SET.

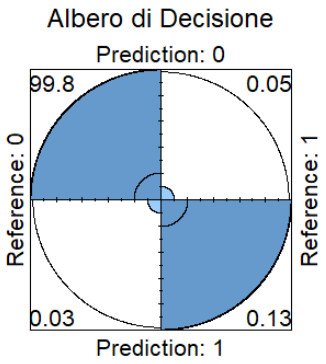


L'analisi discriminante lineare resta ancora il metodo che offre le migliori performance.

Per l'**albero di decisione** i risultati con la *KFCV* sono i seguenti:

TABELLA 52– FIGURA 53: MATRICE DI CONFUSIONE DELL'ALBERO DI DECISIONE IN KFCV.

	Riferimento	
Previsione	**0**	**1**
0	99.8 %	0.05 %
1	0.03 %	0.13 %



L'albero di decisione genera una percentuale media di *falsi positivi (0.03%)* uguale a quella ottenuta con la regressione logistica e leggermente superiore a quella ottenuta con il classificatore LDA. Anche per la percentuale media di *falsi negativi (0.05%)* la performance è leggermente

superiore rispetto al classificatore LDA e leggermente inferiore alla percentuale media di falsi positivi della regressione logistica. Valutando anche le altre metriche la performance è la seguente:

TABELLA 53– TABELLA 54: METRICHE DI PERFORMANCE DELL'ALBERO DI DECISIONE IN KFCV.

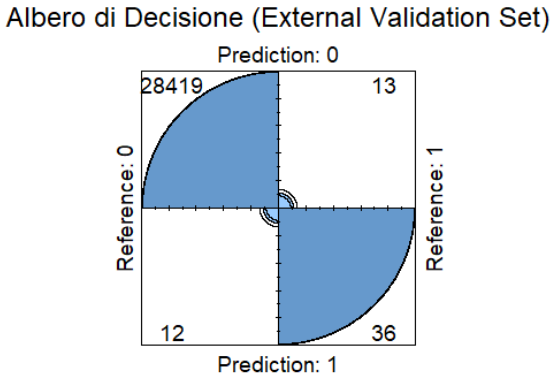
Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.999	0.730	0.726	0.830	0.820

F1 Score	F2 Score	AUC	MCC	KCohen
0.771	0.743	0.863	0.774	0.771

Utilizzando l'external validation set la performance dell'albero di decisione è migliore rispetto alla regressione logistica sempre sullo stesso set di dati, poiché genera un numero inferiore di *falsi negativi (13)* a fronte dei *19 falsi negativi* della regressione logistica. Questa riduzione ha un costo in termini di *falsi positivi* che aumentano da *8* a *12*. Nonostante la buona performance dell'albero di decisione, il classificatore LDA sull'external validation set offre ancora la migliore performance.

TABELLA 55– FIGURA 26: MATRICE DI CONFUSIONE DELL'ALBERO DI DECISIONE SULL'EXTERNAL VALIDATION SET.

Riferimento		
Previsione	**0**	**1**
0	28419	13
1	12	36



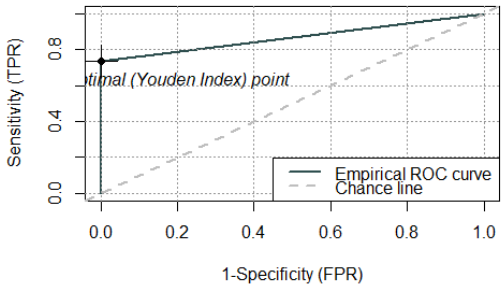
Valutando altre misure di performance i risultati sono i seguenti:

TABELLA 56– TABELLA 57: METRICHE DI PERFORMANCE DELL'ALBERO DI DECISIONE SULL'EXTERNAL VALIDATION SET.

Accuracy	Sensitivity	Specificity	Precision	Neg Pred Value
0.999	0.735	0.999	0.750	0.999

F1 Score	F2 Score	AUC	MCC	KCohen
0.742	0.738	0.867	0.742	0.742

FIGURA 27: CURVA ROC DELL'ALBERO DI DECISIONE SULL'EXTERNAL VALIDATION SET.



Considerando l'albero di decisione si cerca di migliorarne la performance andando a regolare gli *iperparametri*; in particolare si regola il *minisplit*, cioè il numero minimo di osservazioni nel nodo prima che l'algoritmo esegua una divisione, il *minbucket*, cioè il numero minimo di osservazioni presenti nella foglia e il *maxdepth*, cioè la profondità massima di qualsiasi nodo dell'albero finale con il nodo radice che viene trattato con profondità zero. La performance che ne segue è la seguente:

TABELLA 58: METRICHE DI CONFRONTO PER L'ALBERO DI DECISIONE CON I PARAMETRI OTTIMIZZATI.

F_1	F_2	AUC
0.725	0.714	0.842

Confrontando questi valori con quelli ottenuti dall'albero di decisione precedente, la scelta di questi iperparametri non giova alla performance e per questo motivo si sceglierà l'albero di decisione senza l'ottimizzazione degli iperparametri.

Da una valutazione complessiva dei classificatori e delle metriche utilizzate per valutare la capacità previsiva, questo è il quadro definitivo di ogni metodo con i rispettivi risultati ottenuti con la *KFCV* che fa ricadere la scelta sul classificatore LDA come miglior previsore.

TABELLA 59: SINTESI DELLE METRICHE DI PERFORMANCE DI TUTTI I CLASSIFICATORI IN KFCV.

	F1	F2	Precision	Sensitivity	AUC	MCC	KCohen
LOGIT	0.737	0.674	0.880	0.639	0.820	0.747	0.737
LDA	**0.815**	**0.784**	**0.876**	**0.765**	**0.882**	**0.820**	**0.814**
QDA	0.110	0.230	0.060	0.879	0.930	0.222	0.106
DT	0.771	0.743	0.830	0.730	0.863	0.774	0.771

Per quanto riguarda le performance ottenute da ogni classificatore sull'external validation set, i risultati sono piuttosto concordanti con quelli ottenuti in precedenza, con il classificatore LDA che massimizza tutte le metriche.

TABELLA 60: SINTESI DELLE METRICHE DI PERFORMANCE DI TUTTI I CLASSIFICATORI SULL'EXTERNAL VALIDATION SET.

External Validation Set *28480 samples, 31 predictor, 2 classes: "0", "1"*							
	F1	**F2**	**Precision**	**Sensitivity**	**AUC**	**MCC**	**KCohen**
LOGIT	0.690	0.641	0.789	0.612	0.806	0.695	0.699
LDA	**0.784**	**0.779**	**0.792**	**0.776**	**0.888**	**0.820**	**0.783**

QDA	0.093	0.198	0.050	0.816	0.895	0.222	0.196
DT	0.742	0.738	0.750	0.735	0.867	0.742	0.742

Conclusioni

Risulta chiaro come questo lavoro non sia una soluzione assoluta al problema del *fraud detection* che da un punto di vista pratico resta aperto e soprattutto dinamico nella risoluzione, ma proponga solo una delle strade per approcciare questo problema così diffuso. Per studi futuri, quindi, oltre alla scelta di utilizzare delle metriche che tengano conto dello sbilanciamento tra le classi, un'alternativa o un approccio che può essere fuso con l'utilizzo di queste metriche è il processo di **re-sampling**, ovvero tecniche che si basano sui dati e ne modificano la distribuzione. Tali tecniche cercano di alleviare o eliminare lo sbilanciamento tra le classi, ad esempio sottocampionando la classe maggioritaria (*undersampling*) o sovracampionando quella minoritaria (*oversampling*). La modifica della distribuzione della classe deve essere applicata solo al set di dati di addestramento visto che l'intento è quello di influenzare l'addestramento dei modelli. Un resampling più completo è noto con l'acronimo **SMOTE** (*Synthetic Minority Oversampling Technique*) che comporta un sovracampionamento sintetico di una minoranza, cioè anziché campionare una minoranza a partire da dati già esistenti e quindi non aggiungere informazioni utili al metodo, si sintetizzano degli esempi a partire da osservazioni già esistenti, dove sintetizzare nuove osservazioni significa selezionare esempi vicini nello spazio delle caratteristiche, disegnando una linea tra gli esempi nello spazio delle caratteristiche e disegnando un nuovo campione in un punto lungo la linea. Un altro approccio invece potrebbe essere l'utilizzo di tecniche che agiscono sui classificatori così da dare una prospettiva diversa al problema utilizzando delle penalità. La classificazione penalizzata impone un costo aggiuntivo al metodo per aver commesso errori di classificazione sulla classe di minoranza durante la fase di addestramento. Queste penalizzazioni possono spingere il metodo a prestare maggiore attenzione alla classe di minoranza. Lo svantaggio di questo metodo è la difficoltà nella selezione della

sanzione; di solito l'approccio che si utilizza è quello di provare una varietà di penalizzazioni e verificare quale penalizzazione offre le migliori performance per il caso specifico. Infine, l'approccio utilizzato in questo studio è stato di tipo supervisionato utilizzando i quattro classificatori descritti sopra, ma gli algoritmi supervisionati che possono portare soluzioni efficienti al problema in esame possono essere molteplici, tra cui il metodo SVM, il metodo random forest oppure una rete neurale che può fornire un risultato più accurato rispetto ad altro modelli considerando che si avvale del calcolo cognitivo. Allo stesso modo, oltre all'approccio supervisionato, il problema del fraud detection può essere affrontato anche con un approccio non supervisionato o spesso combinando i due approcci.

Riferimenti

- Alexander S. Gillis (2019), *"*Definition Fraud Detection*"*. *SearchSecurity – TechTarget*
- Jason Brownlee (2019), "A Gentle Introduction to Imbalanced Classification". *Machine Learning Mastery.*
- Bartosz Krawczyk (2016), "Learning from imbalanced data: open challenges and future directions. Springer Link.
- Tejumade Afonja (2017). "Accuracy Paradox". *Towards Data Science.*
- CJ Van Rijsbergen (1979). "Information Retrieval" (2nd ed). Butterworth - Heinemann.
- Donald Bamber (1975). "The area above the orfinal dominance graph and the area below the receiver operating characteristic graph". *Science Direct*. MH Zweig & G. Campbell (1993). "Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *PubMed – National Library of Medicine.*
- Adam Hayes (2021). "Posterior Probability Definition". *Investopedia – Math & Statistics*.
- J. Corso (2013). (https://cse.buffalo.edu/~jcorso/t/CSE555/files/quiz01_solutions.pdf).
- Afroz Chakure (2019). "Decision Tree Classification – An Introduction to Decision Tree Classifier". *Medium*.
- Trevor Hastie & Robert Tibshirani (2013). "An introduction to Statistical Learning".