AI 4 Fraud Detection

CODICE	**DT0274**
DURATA	**2 gg**
PREZZO	**1.350,00 €**
EXAM	

DESCRIZIONE

Un problema tipico delle aziende di e-commerce, di emissione di carte di credito, di telecomunicazioni e di assicurazioni (e non solo) è il rilevamento delle transazioni fraudolente, ed in generale delle frodi. Si ritiene che molte aziende perdano il 5% del loro fatturato anno a causa di frodi di vario tipo. In passato questo rilevamento era basato su regole deterministiche (del tipo *if-then-else*), abbastanza facilmente applicabili anche in Excel, e sovraccaricava gli uffici aziendali di investigazione (*auditing*) con una mole di segnalazioni; inoltre, il rilevamento deterministico non era elastico ai nuovi tipi di frodi.

Oggi, grazie a moderne tecniche di Intelligenza Artificiale e Machine Learning predittivo basate sui dati e sulla statistica avanzata, - in particolare le tecniche semi-supervisionate, il Text Mining e la Social Network Analysis - se un'azienda possiede una serie storica di transazioni/frodi e di documenti collegati può rilevare in modo probabilistico (in modo automatico in tempo reale) le "nuove" transazioni o richieste che probabilmente sono fraudolente, ordinarle per probabilità, ed indirizzare così in modo mirato le investigazioni più accurate (manuali) da parte degli uffici competenti, ottimizzando il loro lavoro in base al budget / tempo disponibili. Moderne tecniche analitiche permettono oggi di scoprire i "pattern anomali" nel comportamento del cliente e distinguere efficacemente tra transazioni anomale (*outlier*) e patologiche (frodi vere e proprie).

Inoltre, il corso affronta due tipici problemi della Fraud Detection: a) la scarsità di transazioni investigate (frode sì/no) a causa del costo del rilevamento, b) la scarsità di transazioni fraudolente sul totale (meno dell1%, tipicamente), e suggerisce degli approcci ad hoc per superare questi due problemi.

Infine, il corso tratta gli aspetti economici della Fraud Detection, come l'*Expected Fraud Loss* (l'importo totale frodato atteso) e la misurazione del suo *Return On Investment (ROI)*, e fornisce dei benchmark prestazionali usati internazionalmente.

Le tecniche illustrate nel corso sono a supporto decisionale degli uffici ispettivi dell'azienda, complementandone le capacità, e non li sostituiscono. Il corso illustrerà con dati reali (anonimizzati) una framework generale di indagine e rilevamento frodi applicabile ai suddetti casi, implementata con il software.

Il corso è completamente *hands-on* sul codice Python (od R), senza slide. Tutti i casi d'uso illustrati nel corso sono implementati in codice ben commentato in italiano.

OBIETTIVI RAGGIUNTI

Al termine del corso i partecipanti saranno in grado di:

- conoscere i diversi casi d'uso della AI nel campo del rilevamento frodi
- applicare le tecniche di AI a problemi reali di rilevamento frodi nelle aziende
- utilizzare il codice software fornito nel corso come punto di partenza per adattarlo alla propria realtà aziendale.

TARGET

Il corso si rivolge a Data Scientist, Analisti di rischio, Project Manager, Sviluppatori e chiunque voglia acquisire una competenza effettiva e pratica sul tema del rilevamento delle frodi tramite le nuove tecniche di AI.

PREREQUISTI

Conoscenze di base effettiva del linguaggio Python od R

CONTENUTI

Pre-elaborazione dei dati non adatti all'analisi frodi, in particolare

- gestione dei valori mancanti;
- gestione delle componenti sotto-rappresentate;
- standardizzazione dei dati numerici;
- categorizzazione;
- anonimizzazione;
- in generale, la manipolazione dei dati.

Analisi descrittiva dei dati aziendali passati

- misure di statistica descrittiva;
- la legge di Benford;
- la regola "box plot";
- i grafici avanzati;
- le serie temporali?
- la collaborazione con l'esperto di dominio.

Tecniche non supervisionate di *outlier ranking*

- le deviazioni dalla “normalità”: lo strumento storicamente di maggior successo nella Fraud Detection (*red flag*);
- le transazioni “sospette” (*outlyingness*);
- lo stato dell’arte: il *local outlier factor* (LOF);
- l’outlier ranking basato sul *clustering*;
- confronti tra tecniche.

Previsioni di frode supervisionate (classificazione probabilistica)

- la dimensione del campione necessaria;
- partizionare i dati in due o tre subset (training, validazione, test);
- il *fraud probability ranking*;
- la scelta dei predittori con tecniche tradizionali: correlazione, variable selection automatica, score di Fisher, Information Value, Cramer’s V;
- la scelta dei predittori con tecniche RIDIT e PRIDIT applicate alla Fraud Detection;
- le previsioni di frode con la tecnica *Bayesiana* del Naive Bayes;
- le previsioni di frode con la tecnica *ensemble* del Boosting;
- le previsioni di frode con la tecnica (tradizionale) della Regressione Logistica;
- Reti Neurali e SVM per la Fraud Detection?
- la scelta dei predittori con la Principal Component Analysis (PCA) applicata alla Fraud Detection;
- confronti tra tecniche;
- la previsione dell’importo frodato (anziché della probabilità di frode).

Benchmark internazionali di riferimento nella Fraud Detection probabilistica: che capacità predittiva è lecito attendersi?

La valutazione della qualità del ranking (per l’investigazione)

- accuratezza ed errore;
- i falsi positivi ed i falsi negativi;
- le matrici di confusione;
- costi differenti degli errori: cosa fare;
- misure di valutazione: Precision & Recall, Lift Chart, curve ROC/AUC;
- la valutazione delle transazioni non-etichettate

Confrontare le previsioni ex-ante con gli effettivi ex-post: il back-testing nella Fraud Detection

La gestione delle classi “sbilanciate” (poche frodi nel dataset)

- metriche di valutazione modificate;
- sotto-campionamento e sovra-campionamento dei dati di training;
- il classico metodo *SMOTE*;
- il bilanciamento empirico delle classi;
- il recente approccio *Likelihood*;
- il learning *cost-sensitive*;
- la strategia di “campionamento stratificato” dei dati di test.

Tecniche predittive innovative (semi-supervisionate) nel caso di poche transazioni investigate

- self-training;
- semi-supervised clustering;
- T-SVM.

Il Text Mining e la Social Network Analysis per la previsione delle frodi

Quali output fornire agli uffici investigativi aziendali?

La prospettiva economica della Fraud Detection: l’*Expected Fraud Loss* ed il Ritorno dell’Investimento

Dataset utilizzati: report commerciali fraudolenti (400.000 record), altri.