26 Febbraio 2021

Modelli AI per il credit scoring

Un business case con i dati PSD2

Tommaso Giordani – Chief Risk Officer Gruppo Sella

Benefici AI nel credit scoring

Contesto

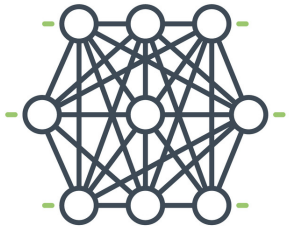
La manipolazione dei dati è diventata sempre più indispensabile a fronte di una messa a disposizione di dati massiva e granulare. In questo contesto le adozioni di metodologie di ML, anche se non recentissime, ad esempio il Random Forest('95), hanno prodotto risultati significativi.

L'intelligenza artificiale è ora un tema centrale in quanto, negli ultimi anni, si è creato un ambiente favorevole alla sua crescita: l'incremento esponenziale della **potenza dei processori**, il potenziamento delle **strutture hardware** e lo sviluppo sempre più **imperante** e **massivo di avanzati algoritmi matematici**.

- **Maggiore accuratezza:** in qualsiasi contesto, i modelli AI hanno dimostrato una migliore capacità nel predirre e/o classificare.
- **Sfruttare un know-how internazionale:** in un mondo sempre più globalizzato, dove le distanze sono nulle, l'attenzione è orientata su questi metodi, sfruttati da tutte le FAGM (facebook, amazon, google e microsoft). Risulta facilitata la curva di apprendimento.
- **Uso di open-source: «Python» ed «R»** sono software gratuiti
- **AI,** mettere le basi per un ambiente «futuro» dove i modelli si miglioreranno da soli «nell'on-going»

Metodologie di Machine learning e AI

Reti neurali



*Un modello "**Artificial Neural Networks**" simula il design di processo del cervello basato su scienze cognitive, che si traduce, in termini algorimici, su elaborazione a parallelismo distribuito.*

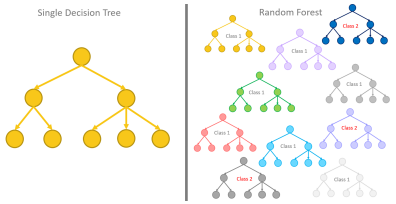
Strenghness

- ❑ ***Capacità predittiva**, grazie alla capacità di learning dagli eventi*

Weackness

- ❑ ***Interpretabilità,***
- ❑ ***Complessità***
- ❑ ***Condizionato alla numerosità campionaria***

Random Forest



*Un modello "**RandomForest**" è uno modello che permette di sfruttare output di una molteplicità di alberi decisionali su campioni di traning e validation cercando di massimizzare **information GAIN**.*

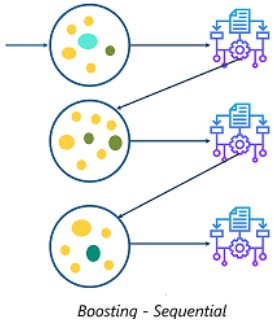
Strenghness

- ❑ ***Calcola l'importanza delle variabili***
- ❑ ***Basso Overfitting***
- ❑ ***Coglie relazione non lineari***

Weackness

- ❑ ***Complessità***

XGBoost



*Un modello «**XGB**» si basa sempre su una logica di albero decisionale con la capacità di «imparare» dal suo albero precedente*

Strenghness

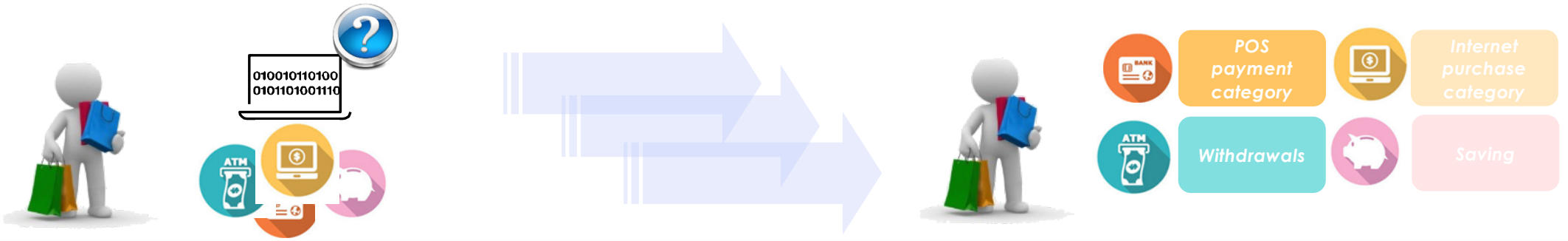
- ❑ ***Calcola l'importanza delle variabili***
- ❑ ***Basso Overfitting***
- ❑ ***Performance significative***

Weackness

- ❑ ***Complessità***

Rivoluzione dei dati di movimentazione: Behavioral paths e consumption Lifestyle

Cambio di paradigma dei dati transazionali



Ad oggi, nei modelli di rating «tradizionali» informativa dei c/c viene rileva in modo **«silent sound»** , ovvero, le informazioni tracciate e imputate a modello sono i valori **omnicomprensivi e blank** di spesa e/o di incasso, ad esempio importi dare.

In altre parole, non si legge la tipologia di movimento

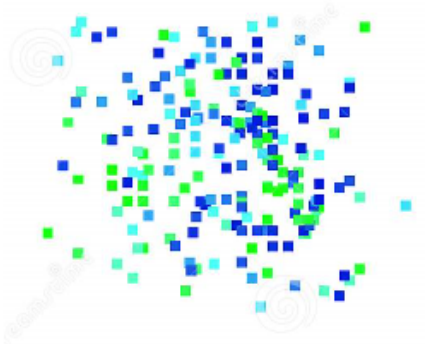
Grazie alla **PSD2** e alle avanzate **metodologie di ML** si procede alle decomposizione/decriptazione, **attraverso il categorizzatore,** massiva dell'informativa aggregata derivante dai nostri c/c e carte di credito.

Se questo si pensa, accompagnato anche dall'informativa, laddove concessa, di terzi flussi finanziari, si possono tracciare veri e propri **comportamenti di spesa dell'individuo dai quali si possono disegnare i lifestyles cogliendo relazioni di rischio ad oggi precluse.**

Modelli AI per il credit scoring

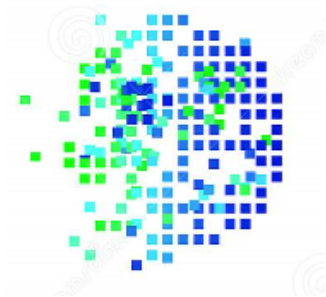
Data management

Big Data



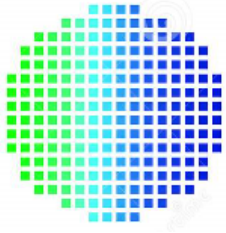
Partendo dagli oltre **260 milioni** di movimentazioni categorizzati

Analytics



Attività di selezione e ordinamento dei dati che ha portato ad una significativa riduzione dimensionale arrivando a **76 mln di movimenti**

Sample



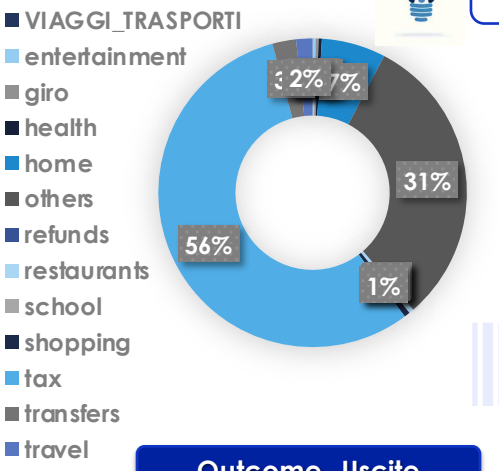
Schiacciando questi movimenti a livello di singolo cliente abbiamo ottenuto circa **700k controparti su cui addestrare i modelli.**

Variabili PSD2 – Macro categorie

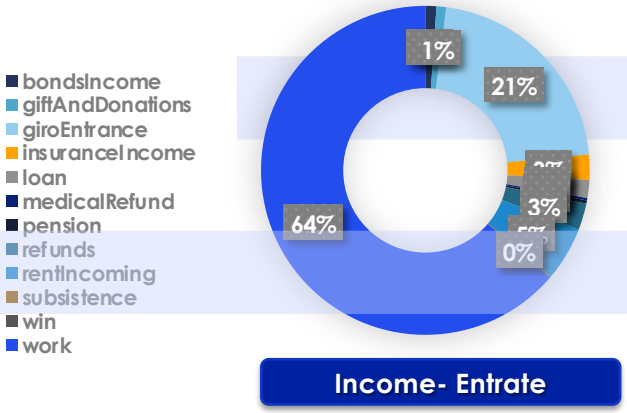
Sintetizzazione Dati

I dati provenienti dal motore semantico, di uscite e di entrate , vanno ulteriormente categorizzati e affinati al fine di ottenere in maniera organica dei set informativi utili allo sviluppo .

Categorizzatore (*)



Outcome- Uscite



Income- Entrate

Ambiti

- Reliability
- Management
- Lifestyle
- Capacity
- Timing

(*)Evidenze del campo categoria sul sottoinsieme del campione di training non privati

Modelli AI per il credit scoring

Outcomes modelli – privati

Performance

Modello Machine Learning - XGB

Campione	Accuracy Ratio
Training	82%
Out-of-sample	82%

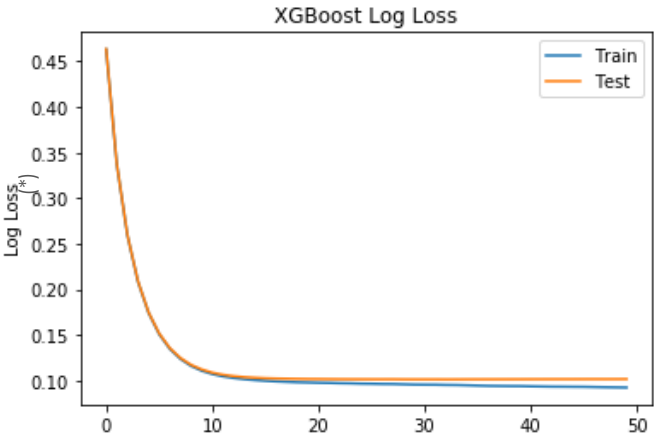
Modello Tradizionale

Campione	Accuracy Ratio
Training	79%

Importance

Features	Importance
Saldi medi	57.1%
Tasso di variazione giacenze	9.1%
Elasticità entrate su uscite	5.6%
Esposizione media in tasse	5.5%
Media entrate permanenti	4.8%
Presenza pagamenti more	4.1%
Indice di sostenibilità finanziaria	3.3%
Esposizione media in mortgages	2.0%
Esposizione media in "spesa virtuosa"	1.6%
Trend bonifici in entrata	1.4%
Esposizione media in ristoranti	1.1%
Volatilità spese carte	0.9%
Saldo gestione correnti	0.7%
Spese media in viabilità	0.7%
Clustering supermarket	0.4%
Trend bonifici in uscita	0.4%
Presenza abbonamenti media	0.3%
Indice di greenness	0.3%
Entrate da investimento	0.2%
Presenza spese in charity	0.2%
Presenza spesa in cultura	0.1%

Learning Curve



LA curva mostra un «break-even», con l'incrementare degli stimatori, superati i 15 il modello non apprende più.

(*) $$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-p(y_i))$$

Legenda

- Esito Positivo AR >= 50
- Esito Accettabile, 50% < AR < =35%
- Esito Negativo AR < 35%