LA *DATA SCIENCE* NEL MONDO DI BANCHE E SOCIETA' FINANZIARIE

Il sempre maggior ricorso alle tecniche statistiche avanzate della *Data Science* da parte degli operatori del comparto bancario e finanziario è direttamente correlato all'affermazione delle metodologie BD&AA (*Big Data & Advance Analytics*) nell'ambito del panorama accademico internazionale. Tra queste, una menzione particolare la merita il *Machine Learning*, branca della *Data Science* particolarmente adatta per l'ottimizzazione di processi di diversa natura.

L'utilizzo di tali tecniche - che parte da un'approfondita analisi e pulizia del database originario - permette di invertire il paradigma statistico-informatico tradizionale: il modello sviluppato, che fino a ieri era un programma chiamato a stimare un output a partire da alcune variabili di input predeterminate, oggi studia e impara dai dati a disposizione in modo da replicare quanto appresso laddove i dati e le informazioni risultano parziali e/o assenti.

L'introduzione delle tecniche di *Machine Learning* ha permesso una profonda innovazione in merito all'utilizzo dei dati e delle serie storiche fino ad oggi osservate. La particolare capacità inferenziale dei nuovi modelli non parametrici permette infatti l'estrazione di un'informazione qualitativa e quantitativa (*Data Mining*), insieme a una modellizzazione delle dinamiche latenti sottostanti che determinano la variazione dei valori della variabile obiettivo.

Tali nuovi modelli hanno consentito la transizione da un'analisi descrittiva a un'analisi predittiva, capace di intercettare quelle che sono le dinamiche sottostanti al sistema. Il sempre maggior utilizzo delle tecniche di *Machine Learning* risulta oggi la chiave per la trasformazione tecnologica in tutti i campi scientifici, dalla ricerca medica ([1]) allo sviluppo dei servizi digitali ([2]), dalle decisioni strategiche aziendali ([3]) allo studio dei buchi neri ([4]).

A conferma della notevole rilevanza che tali metodologie di analisi iniziano a ricoprire all'interno del panorama bancario e finanziario, in merito all'utilizzo delle stesse si è espresso recentemente anche il Regolatore Europeo (*European Banking Authority* - EBA), il quale attraverso un documento dedicato, "*Final Report on Big Data and Advanced Analytics*", pubblicato il 13 gennaio 2020, ha rilevato un crescente interesse per l'utilizzo di soluzioni BD&AA in capo agli operatori di sistema: già oggi due banche su tre della Zona Euro utilizzano algoritmi di *Data Science* per ottimizzare i processi di gestione interna. Questi ultimi, conferma l'EBA, sono in grado di migliorare l'efficienza e la produttività purché si muovano all'interno di alcuni pilastri fondamentali - quali il *Data Management*, l'infrastruttura tecnologica, la metodologia analitica, l'organizzazione e la *governance* - e garantiscano una serie di "*element of trust*" individuati dal Regolatore.

Con riferimento al contesto bancario e finanziario - e in particolare nell'ambito della valutazione degli asset immobiliari in *leasing* o a garanzia dei finanziamenti erogati - il *Machine Learning* può essere utilizzato per soddisfare le esigenze di valutazione e monitoraggio delle garanzie immobiliari o, nel caso del *leasing*, dei beni iscritti a bilancio. Tale monitoraggio - che le *best practices* di sistema richiedono sia effettuato in maniera

([1]) Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "*Machine learning in medicine*." New England Journal of Medicine
([2]) https://ai.google/education/
([3]) https://argomenti.ilsole24ore.com/machine-learning.html
([4]) https://www.businessinsider.com/black-hole-photograph-event-horizon-telescope-algorithm-2016-6

continuativa, affidabile e statisticamente robusta - è inoltre propedeutico, come indicato dalla normativa prudenziale, alla possibilità di beneficiare di vantaggi in termini di minori assorbimenti patrimoniali.

Inoltre, il processo di analisi predittiva innescato a seguito dell'utilizzo di tali tecniche, fornisce agli operatori informazioni utili ai fini della fase di istruttoria del credito e dell'attività di *due diligence* del portafoglio, propedeutica per la cessione delle posizioni NPE (*Non Performing Exposures*) ad operatori specializzati.

In sintesi, il vantaggio tangibile derivante dall'utilizzo di questi modelli, nello specifico perimetro di analisi individuato, è quello di migliorare gli strumenti a disposizione degli operatori di mercato ai fini dell'attività di monitoraggio e valutazione degli asset immobiliari, sia con riferimento a quanto disciplinato dalla normativa di settore - si possono citare, a titolo esempio, i requisiti necessari per la valorizzazione degli asset immobiliari contenuti negli articoli 208 e 229 del CRR e le nuove *best practices* definite dall'EBA nel documento di consultazione in materia di linee guida per la concessione e il monitoraggio dei prestiti - sia al fine di efficientare i processi di gestione, ampliando e arricchendo il patrimonio informativo degli operatori nell'ambito dei processi, tra gli altri, di istruttoria del credito e *due diligence* di portafoglio.

OMI PRICE PREDICTION: BIG DATA E MACHINE LEARNING PER LA VALUTAZIONE IMMOBILIARE

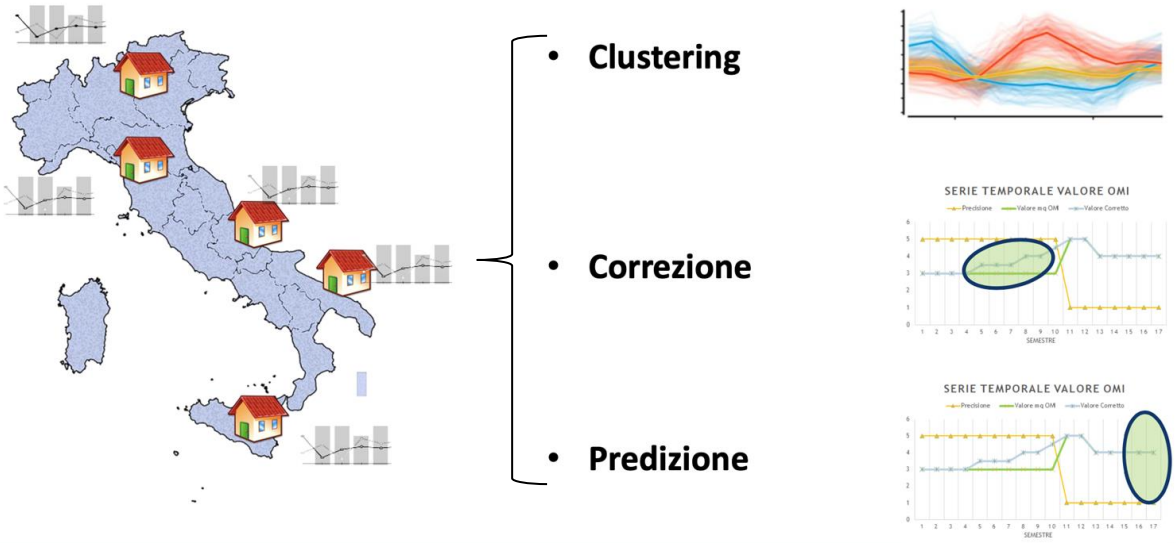
Il modello "OMI *Price Prediction*" (OMI PP) - realizzato da Moderari e Assilea con la collaborazione di studenti e dottorandi del dipartimento di *Data Science* dell'Università "La Sapienza" di Roma - è stato sviluppato a partire dalle serie storiche relative agli immobili non residenziali. La base dati è rappresentata dalle quotazioni immobiliari elaborate dall'Osservatorio del Mercato Immobiliare (OMI) dell'Agenzia delle Entrate ([5]). Il modello OMI PP intende migliorare le *performance* dell'elaborazione statistica che oggi Assilea mette a disposizione degli operatori, quest'ultima provvede ad associare ad ogni serie storica OMI una precisione decrescente man mano che si riduce la correttezza del valore stimato ([6]).

Nel caso di specie l'analisi predittiva ha riguardato l'andamento futuro delle serie storiche relative al valore degli immobili (analisi *forward looking*), in modo da permettere agli operatori un'analisi del c.d. "*what if scenario*" in base alla quale valutare l'adozione di differenti strategie gestionali.

In particolare, il modello in analisi, partendo dalle serie storiche delle quotazioni OMI distribuite sul territorio nazionale, genera tre output: 1) *clusterizzazione* delle serie storiche; 2) *correzione* del valore relativo al semestre corrente non avente precisione massima; 3) *predizione* del valore per i semestri futuri.

([5]) L'aggiornamento delle quotazioni OMI avviene con cadenza semestrale.
([6]) La precisione peggiora nel caso in cui le caratteristiche dell'immobile da stimare non siano presenti, o siano presenti solo in parte, all'interno del database OMI.

Figura 1: Output del modello.



Il modello si sviluppa su 3 processi:

- Raggruppamento (*clusterizzazione*) degli immobili per determinate caratteristiche di input: in fase di prima applicazione si è ritenuto corretto considerare la posizione geografica (latitudine/longitudine), la tipologia dell'immobile e il valore di riferimento OMI ([7]) (si veda figura 2)
- *Training* del modello predittivo statistico a partire dalla *clusterizzazione* effettuata ([8])
- Correzione e predizione del valore dell'immobile: il valore OMI viene corretto (sulla base delle quotazioni OMI similari con precisione massima) e ne viene previsto il valore futuro (stima previsionale del valore nel/i semestre/i successivo/i).

Le quotazioni OMI con precisione minore sono pertanto corrette grazie all'analisi dell'andamento delle serie storiche dei valori OMI relativi a immobili simili con precisione massima. Tale correzione è effettuata analizzando la somiglianza con altri immobili, ma non considerando solo la vicinanza geografica, caratteristica che ad esempio potrebbe risultare inadeguata per località isolate, bensì facendo leva sulla *clusterizzazione* delle serie storiche, che permette di individuare una similarità che implica una particolare somiglianza delle caratteristiche che hanno guidato l'identificazione dei *cluster*.

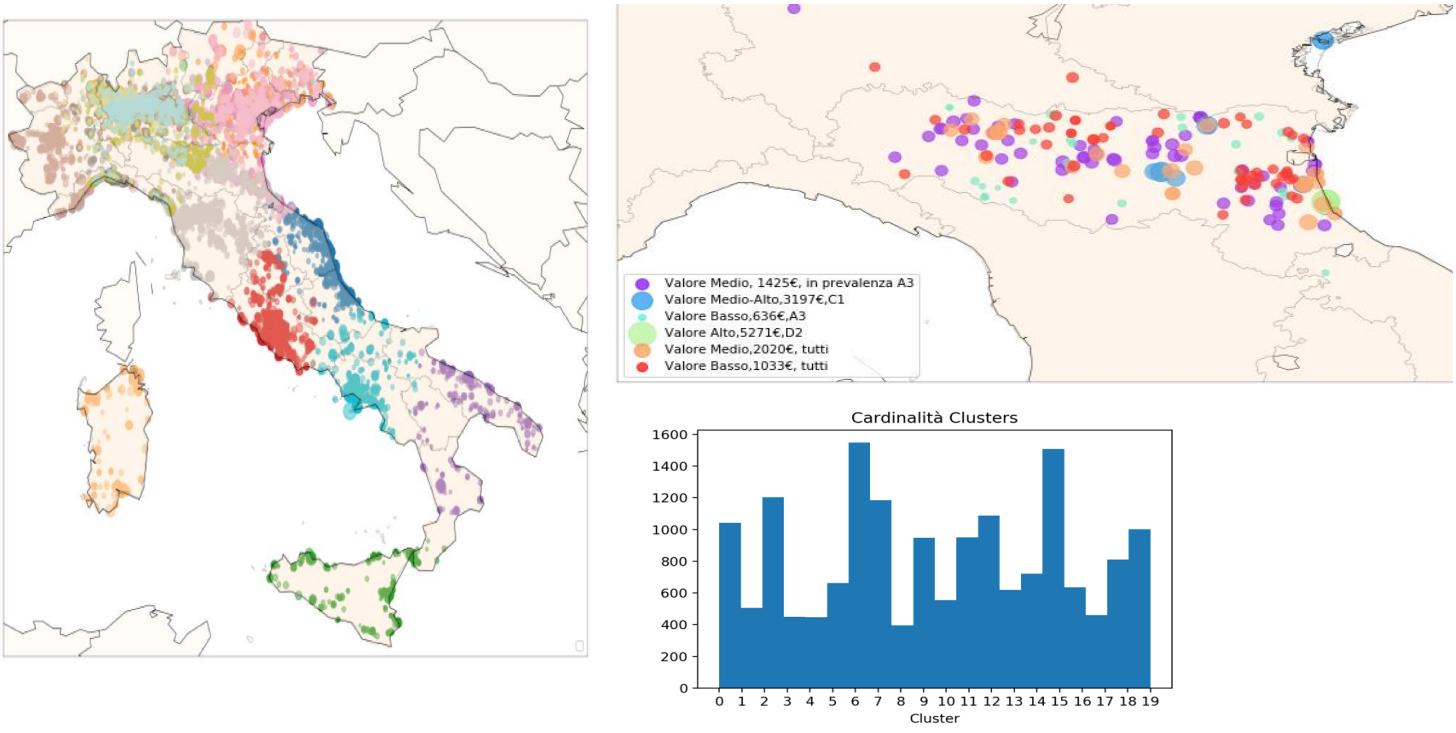
La *clusterizzazione* degli immobili può essere effettuata su scale diverse. Analizzando l'intero portafoglio degli immobili a disposizione in fase di test (circa 25mila immobili, figura di sinistra), si può notare un raggruppamento molto coerente con la geografia regionale al centro-sud (segno che la caratteristica geografica è risultata molto rilevante per l'algoritmo), mentre al nord i *cluster* tendono a sovrapporsi non rispettando precisamente i confini regionali (segno che l'algoritmo ha rilevato una maggiore significatività di altre caratteristiche per l'individuazione dei *cluster*). Da notare, inoltre, che la cardinalità dei gruppi rimane sempre

([7]) Il raggruppamento è stato effettuato attraverso l'uso di algoritmi di *clustering*, sulla base della dimensione, della tipologia dell'immobile e della zona geografica.

([8]) Per "allenare" il modello sono stati usati dei modelli ereditati dall'intelligenza artificiale, in particolare ci si è avvalsi di una rete neurale ricorrente, LSTM – *long short term memory Recurrent Neural Network*, per l'integrazione delle serie storiche dei prezzi con i meta-dati degli immobili, quali ad esempio la tipologia dell'immobile, la locuzione geografica etc.

superiore a 400 immobili (istogramma di destra), questo ci permette di considerare ogni cluster numericamente significativo.

Figura 2: *Clustering (applicazioni su scala nazionale e su scala regionale)*



Nel caso in cui, invece, si proceda all'analisi di un portafoglio di immobili più ristretto (circa 200 immobili, figura a destra) e concentrato quasi interamente nella stessa regione, si può notare come l'algoritmo consideri maggiormente rilevanti - ai fini dell'individuazione dei *cluster* - caratteristiche quali la categoria catastale e la quotazione OMI presente nelle serie storiche.

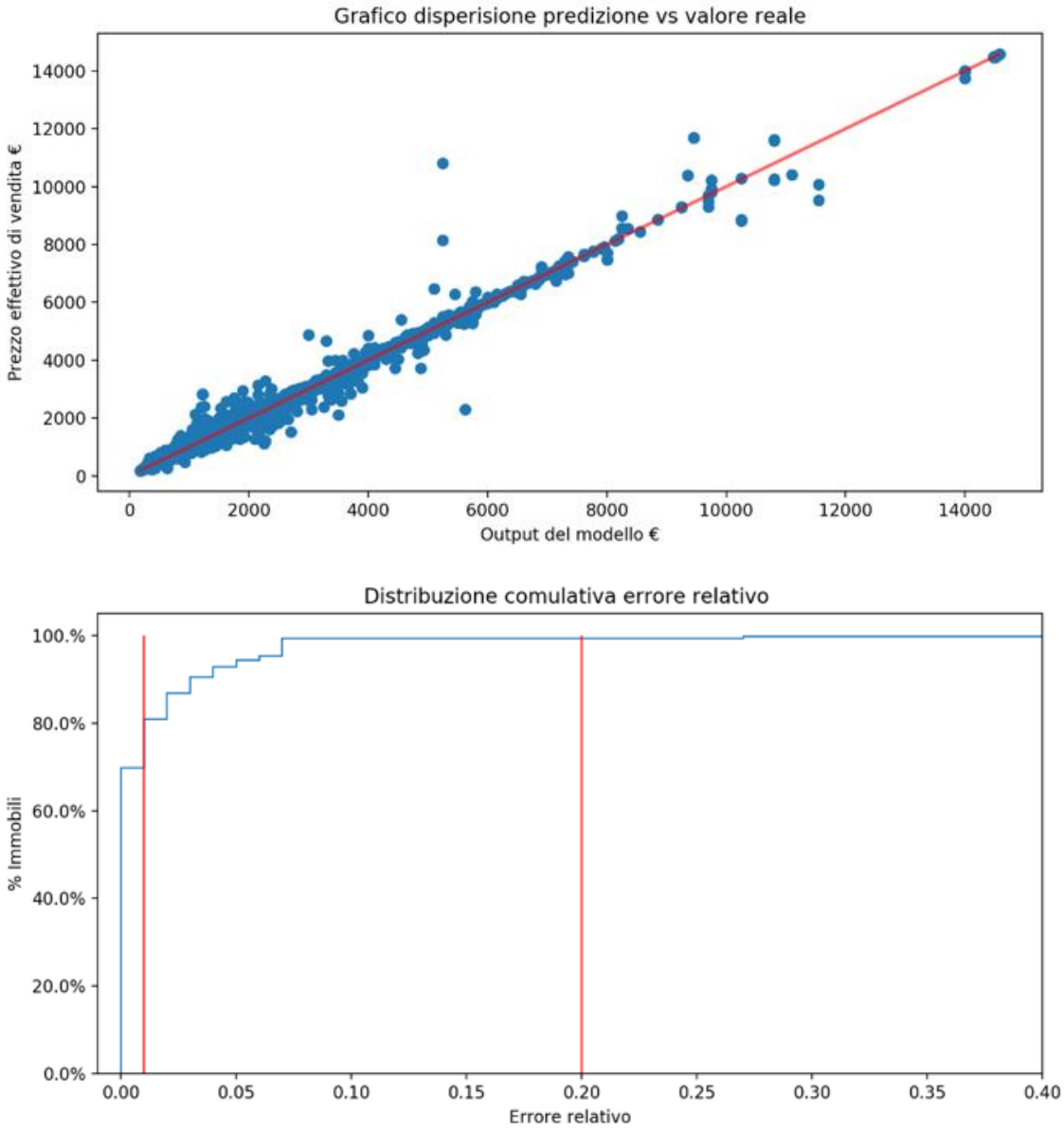
La validazione del modello è stata effettuata attraverso metodi statistici di *cross-validation.* In particolare, i dati utilizzati per la validazione del modello sono stati quelli di *benchmark* (cioè i dati con precisione massima restituiti dal database OMI).

I risultati, come si può osservare dalla figura 3, mostrano un elevato grado di affidabilità: sull'asse delle ascisse è riportato il valore predetto e sulle ordinate il valore OMI con precisione massima (valore reale, *benchmark*).

Lo *scatterplot* confronta il valore predetto in t+1 con il valore reale, quest'ultimo è calcolato su un gruppo di immobili non considerato in fase di allenamento e usato come *benchmark* in fase di validazione. La bisettrice rappresenta un predittore perfetto (valore predetto uguale al valore reale) e la distanza dei punti dalla bisettrice mostra l'errore delle stime effettuate dal modello. In particolare, come si può notare nel grafico della distribuzione dell'errore relativo, il 90% delle predizioni effettuate ha un errore relativo minore del 5%, e l'80% un errore relativo minore del 1%. La presenza di *outlier* in fase di predizione è imputabile a un andamento anomalo della serie storica relativa allo specifico immobile, spesso associato a "rumore" ([9]) presente nei dati.

([9]) Per "rumore" si intende, ad esempio, la presenza nel database di dati non corretti.

Figura 3: *Bontà predittiva del modello in t+1.*



Il modello descritto è stato ad oggi sviluppato in una prima versione test, gli obiettivi da raggiungere ai fini di un ulteriore miglioramento delle *performance* dell'algoritmo sono i seguenti: i) migliorare la precisione e ampliare la copertura geografica delle quotazioni ad oggi presenti all'interno del database OMI, ii) determinare valori prospettici di mercato che dimostrino affidabilità e coerenza per un periodo di almeno 4 semestri a partire dal periodo/data di elaborazione; iii) analizzare gli impatti sul modello derivanti dalle informazioni relative ad asset immobiliari venduti nell'ambito di cessioni di NPE (*Non Performing Exposures*); iv) inserire all'interno del modello dati macroeconomici esterni.

In particolare, con riferimento al punto iii), si provvederà a studiare la dinamica degli impatti sul valore degli asset immobiliari derivanti da operazioni di vendita massiva di NPE, in modo da ampliare la gamma di fenomeni che l'algoritmo rileva ai fini della stima previsionale della variabile output. L'obiettivo, infatti, è quello di prevedere l'effettivo deprezzamento dovuto a tale fenomeno ed elaborare una stima più accurata, in grado di isolare l'informazione sulle cessioni di NPE e dare la possibilità agli operatori di studiare il "*what if scenario*".

Per quanto concerne l'obiettivo descritto al punto iv), invece, si provvederà all'inserimento di fattori, da un lato, endogeni al sistema, dall'altro, correlati e determinanti per la stima del trend delle serie storiche. Tale implementazione avverrà attraverso una prima parte di *features selection*, finalizzata alla determinazione delle effettive le variabili esplicative; in seguito, si procederà all'inserimento delle variabili selezionate nel novero delle variabili di input del modello.

Le attività propedeutiche alla realizzazione degli obiettivi appena descritti si concluderanno orientativamente al termine del primo semestre 2020.