

Predictive analytics: solo una questione di algoritmi?

Un'accattivante analisi sull'utilizzo dei modelli per la lettura e l'interpretazione dei dati da cui emerge la necessità di utilizzare sempre una grande capacità di critica, non solo nella raccolta del dato, ma anche nella scelta del modello e delle variabili.



Maurizio Poli

è SDA professor di Metodi Quantitativi. Le sue aree di interesse e di ricerca comprendono: modelling, tecniche di simulazione, tecniche e strumenti di data mining, previsione con tecniche e strumenti innovativi, business intelligence, controllo delle performance aziendali. maurizio.poli@sdabocconi.it



Renata Trinca Colonel

è SDA professor presso il Competence Centre di Metodi Quantitativi e coordinatore del percorso Business Data Analyst. I temi sui quali svolge attività di ricerca riguardano l'analisi statistica dei dati, data mining, ricerche di mercato, data visualization, business analytics, modelli statistici di scoring e rating, modelli predittivi. renata.trinca@sdabocconi.it

L'esplosione della quantità di dati e di informazioni giornalmente raccolte e la rapida evoluzione e diffusione dei *business analytics* nelle principali aziende hanno portato alla diffusione di sistemi e metodi matematico-statistici sempre più sofisticati e flessibili. In particolare, si è osservato un aumento consistente, nelle diverse realtà aziendali, delle tipologie e della complessità dei modelli predittivi che hanno favorito lo sviluppo di nuove tecniche e di procedure di selezione "automatica degli algoritmi migliori" nei principali software. Ma quale criterio seguire per definire un algoritmo "il migliore"?

Se da un lato è sempre più necessario che le analisi siano *data driven*, ovvero guidate primariamente dagli stessi dati analizzati, dall'altro è cruciale distinguere tra performance statistica e performance di business dei modelli, ovvero tra la capacità di un modello di "ottimizzare" indici di performance statistica e la possibilità dello stesso di risolvere efficacemente il problema aziendale che stiamo affrontando. La selezione del modello più corretto, in grado di supportare le decisioni manageriali in una specifica situazione e in un particolare contesto, non deve necessariamente essere fondata sulla mera valutazione degli in-

dicatori statistici di performance – minimizzazione dell'errore commesso dal modello nel "prevedere" i dati del passato – che si può, più o meno facilmente, ottenere da strumenti software avanzati. L'effettiva coerenza della soluzione trovata deve essere valutata in modo manageriale, come risultato del lavoro di un gruppo "misto" composto da analisti esperti e manager in grado contribuire in modo significativo alla valutazione del problema e della sua soluzione, definendo e delineando in ogni fase del processo i criteri più adatti per la scelta delle alternative da seguire. L'identificazione e lo sviluppo di un'alternativa tecnica piuttosto di un'altra dipende da diversi fattori; una buona analisi congiunta del problema e la conduzione congiunta del processo costituiscono due elementi critici che possono fare la differenza. Le potenzialità dei modelli predittivi in ambito manageriale sono indiscusse, ma non bisogna dimenticare che si tratta sempre di analisi *ad hoc* inserite in un determinato contesto, con precise finalità e non di esercizi condotti "in laboratorio".

Per provare a chiarire meglio questo pensiero proviamo a esaminare un problema concreto, per esempio i modelli di previsione basati su serie storiche, a cui è de-

1. T.C. Redman, "Algorithms make better predictions – except when they don't", HBR, settembre 2014.

dicata anche la case history che segue. I metodi matematico-statistici, alla base dell'analisi delle serie storiche, cercano le regolarità nel comportamento passato delle osservazioni per "proiettarle" nel futuro. I dati a disposizione per queste tipologie di problemi sono infatti semplici sequenze cronologiche che costituiscono la base empirica per lo studio e l'analisi di eventuali comportamenti sistematici e regolari nel tempo. Da tali serie storiche, gli algoritmi matematico-statistici, in modo più o meno sofisticato e complesso, vanno a identificare delle funzioni che consentono di estendere la serie in una previsione dello stesso fenomeno rivolta al futuro. Il punto cruciale non è la scelta dell'algoritmo in sé, ma la valutazione di altri fattori quali la rilevanza e la disponibilità dei dati storici, il livello di accuratezza desiderato, il periodo di tempo da prevedere o anche il fatto che per una previsione migliore potrebbero essere utilizzate altre informazioni o altri dati al momento non esistenti o costosi da rilevare. La definizione degli obiettivi e dei bisogni sottostanti spesso aiuta nella scelta dell'algoritmo che avrà potenzialità e limiti in linea con il contesto.

"Good algorithms make better predictions than people most of the time – except when they don't"¹ e possiamo aggiungere che spesso l'analista fa la differenza.

Modelli predittivi per valutare le potenzialità di mercato di un nuovo prodotto o servizio possono richiedere un livello di accuratezza diverso rispetto a modelli predittivi per misurare l'impatto sulle vendite di una determinata strategia commerciale. L'evoluzione e la diffusione delle tecniche predittive in ambiti diversi ha portato all'utilizzo di automatismi basati prevalentemente sull'identificazione del modello migliore in termini di performance "statistica", ovvero affidabilità e capacità esplicativa degli andamenti sto-

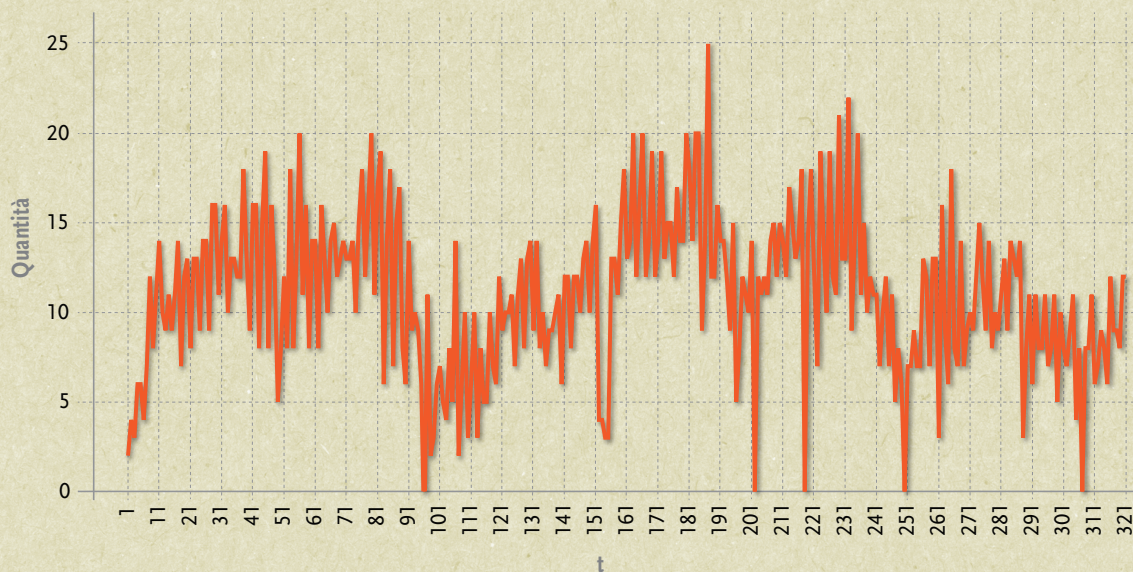
rici ("se sbaglio di meno nel passato, sbaglierò di meno nel futuro"). Il problema è che non sempre il modello che ha performance elevate sul passato manterrà le stesse performance nella previsione. Molto spesso modelli che riescono a ripercorrere l'andamento passato con un elevato grado di accuratezza e con errori contenuti (e quindi che presentano performance tecniche elevate) risultano essere modelli "più rigidi" che poco si adattano ai cambiamenti futuri. L'affidabilità delle previsioni può essere inoltre diversa nel breve e nel lungo periodo e quindi la definizione dell'orizzonte temporale non è un aspetto secondario nella selezione dell'algoritmo, ma anzi è un aspetto cruciale nella sua scelta.

La diffusione e il successo dei *predictive analytics* nelle aziende, oltre alla rapida misurabilità dei risultati, sono dovuti anche alla capacità di tali strumenti di considerare moltissime variabili contemporaneamente e di poter stimare l'impatto di variazioni anche minime di ognuna di esse. Il principale rischio e problema non sembra risiedere negli algoritmi, sempre più avanzati e flessibili, ma nelle stesse assunzioni del contesto, spesso implicite o mai definite apertamente dal management perché difficilmente traducibili in variabili.

■ Case history

Un'azienda italiana, specializzata nella preparazione e distribuzione di piatti pronti freschi per la grande distribuzione organizzata, ha utilizzato modelli di previsione basati su serie storiche per stimare in modo strutturato le quantità e l'assortimento ideali dei propri prodotti sull'intero territorio nazionale. La consegna dei prodotti, in termini di quantità e mix, effettuata giornalmente nei diversi punti vendita sul territorio è, infatti, un aspetto cruciale per l'azienda, che ha deciso di sviluppare

FIGURA 1 | SERIE STORICA DELLE QUANTITÀ VENDUTE DI UN PUNTO VENDITA



alcuni modelli predittivi per l'ottimizzazione delle consegne. Dopo una prima segmentazione dei punti vendita in quattro categorie in base alla tipologia e al potenziale di vendita, sono state studiate e analizzate le serie storiche del venduto e dei resi in termini di quantità e tipologia. Considerando l'alta deperibilità dei prodotti, obiettivo dell'analisi era stimare e prevedere la quantità e la tipologia di prodotto da consegnare in ogni punto vendita minimizzando i resi giornalieri.

Come si può notare con un primo sguardo al grafico 1 della serie storica di partenza, le oscillazioni giornaliere delle quantità effettivamente vedute (quantità consegnata al netto dei resi) in uno specifico negozio sono molto consistenti e l'elevata volatilità si riflette nella difficoltà di individuare un comportamento prevedibile, che è lo scopo della nostra analisi.

Gli algoritmi predittivi (appartenenti alla famiglia degli algoritmi *time series analysis*), con diverso livello di sofisticazione, sono in grado di separare le principali componenti (trend, ciclo e stagionalità) e di cogliere l'andamento delle oscillazioni legato a tali componenti. Ma solo spostando lo sguardo sul contesto e integrando l'analisi con le assunzioni e le conoscenze del management è stato possibile andare oltre, introducendo per esempio un impatto del tempo atmosferico ("quando piove o fa

freddo si vendono meno confezioni di alcuni piatti e più di altre"), definendo un orizzonte temporale molto breve per la previsione ("ora abbiamo bisogno di uno strumento che giorno per giorno ci dica quanti piatti preparare e dove consegnarli, non di una previsione di quanto venderemo complessivamente nel 2015") e pensando alla flessibilità di utilizzo ("quando la città si svuota per le vacanze scolastiche vendiamo meno in alcuni supermercati ma di più in altri, quelli in centro vicini agli uffici").

L'identificazione di nuove variabili di semplice rilevazione ha consentito di integrare le informazioni e di migliorare le performance dei modelli predittivi. Tali variabili sono inoltre facilmente "attivabili" per avere la stima corretta della quantità di prodotto da consegnare quotidianamente sulla rete di vendita.

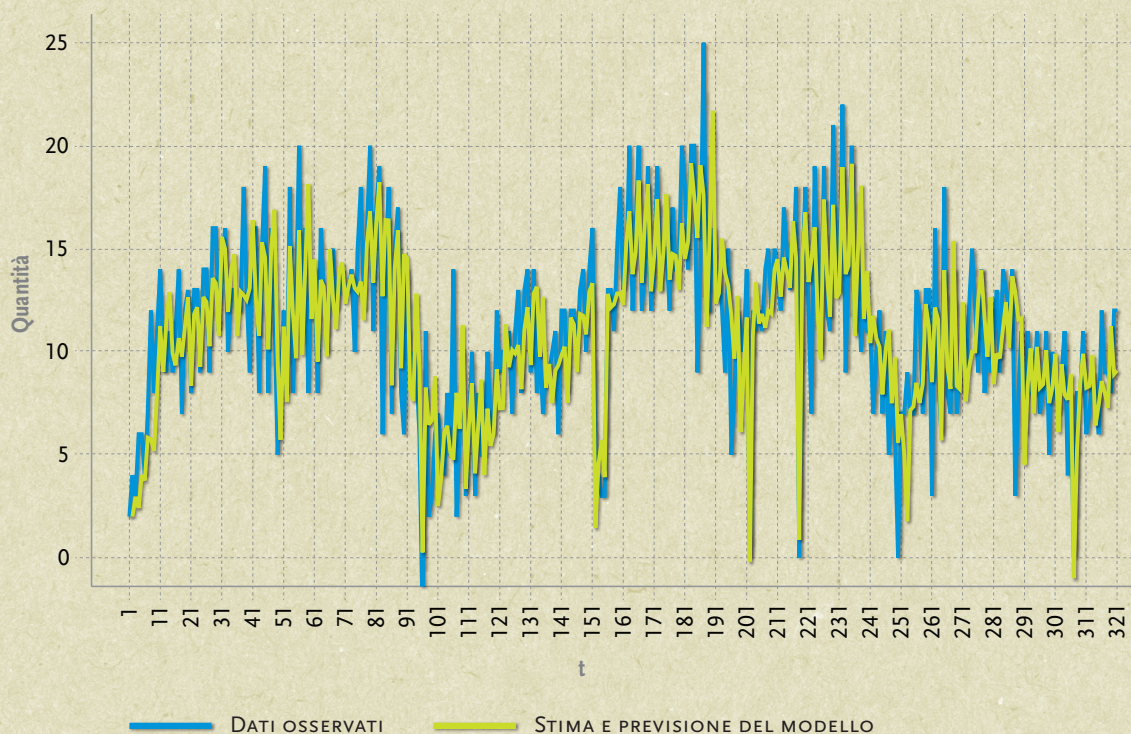
Focalizzando ora la nostra attenzione sugli algoritmi predittivi che sono stati sviluppati, qual è l'algoritmo "migliore"? È evidente che i principali algoritmi possono essere confrontati in modo che possa facilmente essere selezionato l'algoritmo migliore in base alle performance statistiche (anche in modo automatico utilizzando le procedure di selezione) e quindi alla capacità di corretta classificazione previsionale. Ma è possibile sostenere che il modello con l'errore minimo, in termini assoluti o percentuali, sia davvero sempre il modello migliore nella soluzione del

TABELLA 1 | CONFRONTO MODELLI PER LE QUANTITÀ VENDUTE DI UN PUNTO VENDITA

MODELLI	STATIONARY RSQUARE	RMSE	MAE	(OUTLIER)
Exponential Smoothing	0.471	3.828	3.003	0
Autoregressive Integrated Moving Average ARIMA (2,1,0)	0.728	2.781	2.076	7
Autoregressive Integrated Moving Average ARIMA (2,1,3)	0.721	2.838	2.093	7

NOTE: I MODELLI PRESENTATI SI RIFERISCONO AD ALCUNE DELLE PRINCIPALI METODOLOGIE (IN QUESTO CASO QUELLE CON PERFORMANCE SUPERIORI) A DISPOSIZIONE DEL BUSINESS ANALYST PER ANALIZZARE UNA SERIE STORICA (TIME SERIES) CON FINALITÀ PREDITTIVE. RMSE (ROOT MEAN SQUARE ERROR) E MAE (MEAN ABSOLUTE ERROR), RISPETTIVAMENTE L'ERRORE QUADRATICO MEDIO E L'ERRORE ASSOLUTO MEDIO, SONO DUE DEI PRINCIPALI INDICATORI DI "BONTÀ" DEL MODELLO; SI RIFERISCONO ALL'ERRORE (IN TERMINI QUADRATICI O ASSOLUTI) CHE IL MODELLO COMMITTE NEL "PREVEDERE" I VALORI DEL PASSATO (BACKCASTING). OVVIAMENTE MINORE È IL LORO VALORE, MIGLIORE È IL MODELLO. PER MAGGIORI DETTAGLI SI RIMANDA ALLA BIBLIOGRAFIA.

FIGURA 2 | MODELLO STIMATO DELLE QUANTITÀ VENDUTE DI UN PUNTO VENDITA



problema di business? È possibile concludere che prevedendo in modo ottimale la quantità di prodotto venduto giornalmente (quantità consegnata al netto dei resi) si vadano a ottimizzare le consegne?

Oltre alle tecniche e alla valutazione delle loro performance, indispensabili per prevedere la quantità venduta e l'impatto dei diversi indicatori, è stato necessario valutare le assunzioni di business e del contesto per evitare "un circolo vizioso" frutto del modello predittivo. In questo caso gli algoritmi sono riusciti a considerare in modo appropriato soltanto gli errori in un'unica direzione (sovrastima del modello, ovvero presenza di reso in caso di quantità consegnata superiore alla quantità effettivamente venduta), mentre non hanno potuto riconoscere gli altri errori (sottostima, ovvero quantità consegnata inferiore alla quantità di vendita potenziale) poiché non avendo reso non sono stati rilevati o segnalati da nessun indicatore (l'unica informazione è "zero resi", ovvero che si era in situazione di stock out, ma nessuna informazione sulla domanda potenziale inevasa). Minimizzare l'errore sulla quantità effettivamente venduta, oltre a una progressiva diminuzione dei resi, avrebbe potuto portare a una sottostima delle quantità da consegnare,

riuscendo a considerare solo marginalmente la potenzialità del punto di vendita. Congiuntamente alla valutazione delle performance statistiche, i modelli finali sono quindi stati integrati dal gruppo di lavoro (composto da analisti e manager) con alcune variabili supplementari per segnalare agli algoritmi la "rottura di stock" nel caso in cui non ci fossero resi oppure tali resi fossero inferiori al 10% rispetto alle consegne corrispondenti². Solo in questo modo è stato possibile trovare una soluzione ottimale, ovvero una soluzione statisticamente robusta, ma adeguata al bisogno per il quale è stata sviluppata.

Questo caso rappresenta un esempio concreto di come la soluzione di un problema aziendale possa essere trovata nel corretto utilizzo di una metodologia d'analisi. Tropo spesso si rinuncia a un approccio previsionale strutturato perché si ritiene che i dati raccolti siano troppo variabili e quindi "imprevedibili", e, in alcuni casi, si supporta questa opinione mostrando i risultati, inutili sotto il profilo business, di un algoritmo applicato utilizzando un software: il vero valore aggiunto è il modo in cui le tecniche vengono utilizzate, la consapevolezza di cosa si può ottenere e il processo con cui conseguire l'obiettivo. ■

2. Il motivo di tale soglia è che per il tipo di prodotto una quantità di resi molto bassa (porzioni rimaste sullo scaffale del frigorifero) è equivalente a uno stock out, in quanto l'appello commerciale di poche confezioni rimaste in un frigorifero altrimenti deserto è molto basso.

■ Riferimenti bibliografici

Akaike H., Kitagawa G. (2011), *The Practice of Time Series Analysis*, Springer, New York.

Anderson T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, New York.

Bee Dagum E. (1980), *The X-11-ARIMA seasonal adjustment method*, Statistics Canada, Catalogue n. 12-564.

Evans M.K. (2003), *Practical Business Forecasting*, Blackwell Publishing, Oxford.