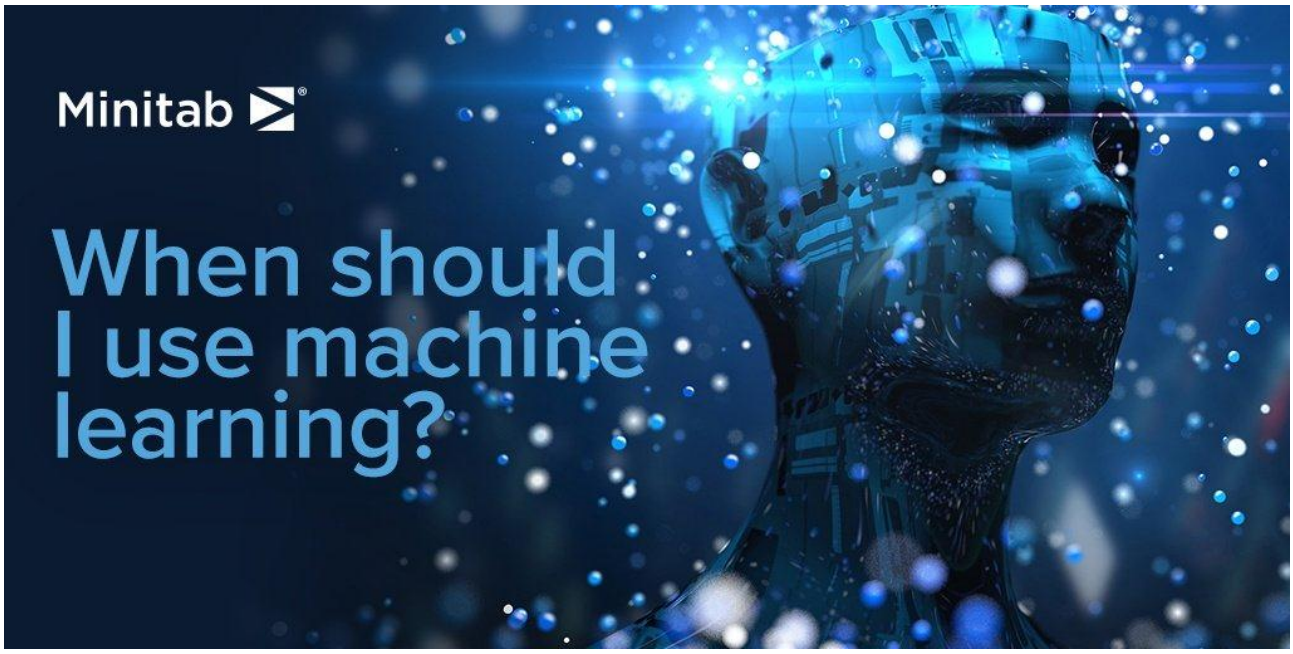
GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

Quando dovrei usare il Machine Learning?

https://blog.minitab.com/blog/bruno-scibilia/when-should-i-use-machine-learning
Versione italiana a cura di Luca Biasibetti.

Il vantaggio potenziale posseduto dai dati memorizzati sui server è enorme. Banche, compagnie assicurative, compagnie di telecomunicazioni, produttori e in realtà le organizzazioni di tutti i settori dovrebbero fare buon uso dei dati che possiedono per migliorare le loro operazioni, comprendere meglio i loro clienti e trovare un vantaggio competitivo.
Con l'avvento dell'Industry 4.0 e dell'Industry Internet of Things (IIoT), le informazioni arrivano in streaming da molte fonti: dalle apparecchiature sulle linee di produzione, dai sensori nei prodotti, dai dati di vendita e molto altro. Essere in grado di raccogliere e analizzare questi dati sta diventando sempre più complicato dal momento in cui le aziende utilizzano gli stessi per raccogliere informazioni al fine di migliorare la loro efficienza ed efficacia sul mercato.
Questa situazione presenta molte nuove opportunità ma comporta anche alcune sfide significative.



NUOVE SFIDE

Le grandi quantità di dati prodotti da questi sistemi moderni presentano sfide uniche che non sono percepibili con piccoli dataset.
I grandi dataset possono contenere un numero elevato di predittori e/o un numero elevato di righe. È anche prassi comune che i dati ottenuti da osservazioni siano più complessi da analizzare rispetto ai dati ottenuti da esperimenti attentamente progettati. In questo breve articolo vedremo come le sopracitate criticità possano influenzare l'analisi dei dati.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

PRESENZA DI UN GRAN NUMERO DI PREDITTORI

I tradizionali strumenti di modellizzazione statistica come la regressione e la regressione logistica si basano sui p-value per rilevare gli effetti significativi di un determinato predittore. Spesso, in particolare, sosteniamo che un predittore con un p-value inferiore a 0,05 è statisticamente significativo. Tuttavia, con questo benchmark di 0,05 stiamo accettando un tasso di errore del 5% ovvero che un predittore su venti sarà significativo solo per caso. Con molti predittori, infatti, fare affidamento unicamente ai p-value può portare ad una modellizzazione del rumore casuale.
Per mostrare ciò, abbiamo simulato casualmente 100 colonne di dati normalmente distribuiti con 15 osservazioni ciascuna. Una regressione stepwise mostra che non meno di 13 colonne su 99 variabili hanno un effetto statisticamente significativo sull'ultima colonna (p-value molto vicini a 0) e il valore di R^2 è estremamente alto (100%). Ovviamente, questi, sono tutti effetti spuri, ossia causati esclusivamente da fluttuazioni casuali (vedi immagine di seguito).

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	13	9.18626	0.706635	4.90579E+10	0.000
C20	1	0.88607	0.886067	6.15149E+10	0.000
C26	1	0.02061	0.020608	1.43072E+09	0.000
C41	1	0.02580	0.025798	1.79102E+09	0.000
C48	1	0.79167	0.791671	5.49615E+10	0.000
C50	1	0.00365	0.003648	2.53252E+08	0.000
C52	1	0.09610	0.096097	6.67151E+09	0.000
C55	1	0.00000	0.000000	7071.30	0.008
C69	1	0.09485	0.094849	6.58483E+09	0.000
C71	1	0.10882	0.108822	7.55491E+09	0.000
C74	1	0.00026	0.000263	18273084.54	0.000
C87	1	0.00853	0.008526	5.91932E+08	0.000
C96	1	0.00001	0.000006	450602.95	0.001
C99	1	0.73619	0.736187	5.11095E+10	0.000
Error	1	0.00000	0.000000		
Total	14	9.18626			

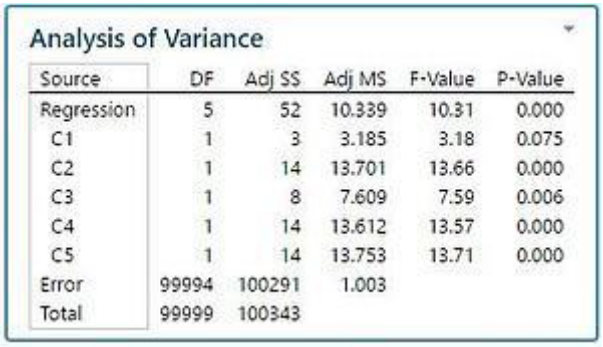
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0000038	100.00%	100.00%	100.00%

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

PRESENZA DI UN GRAN NUMERO DI OSSERVAZIONI: POTENZA vs SIGNIFICATIVITA'

Dimensioni di campione molto grandi, migliorano la potenza e la capacità di rilevare termini statisticamente significativi anche quando sono molto piccoli, tuttavia tali effetti statisticamente significativi non implicano necessariamente un significato pratico. Con dataset di grandi dimensioni, il p-value può diventare eccessivamente sensibile ai piccoli effetti, portando quindi a un modello finale molto complesso contenente quasi tutti i predittori iniziali.
Anche se questi termini possono essere statisticamente significativi, la maggior parte di essi avrebbe effettivamente poco significato pratico.
Per illustrare questo fatto, abbiamo simulato 6 colonne con 100.000 osservazioni ciascuna, abbiamo inserito un modello di modo che le 5 colonne in input abbiano solo un piccolo effetto sull'ultima colonna (impatto reale ma molto piccolo). Il valore di R^2 è, come ci si aspetta, molto basso (vicino allo 0%) ma i p-value indicano che gli effetti sono molto significativi dal punto di vista statistico.



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	52	10.339	10.31	0.000
C1	1	3	3.185	3.18	0.075
C2	1	14	13.701	13.66	0.000
C3	1	8	7.609	7.59	0.006
C4	1	14	13.612	13.57	0.000
C5	1	14	13.753	13.71	0.000
Error	99994	100291	1.003		
Total	99999	100343			

COMPLESSITA' DOVUTA AD EFFETTI NON LINEARI, OUTLIER E VALORI MANCANTI

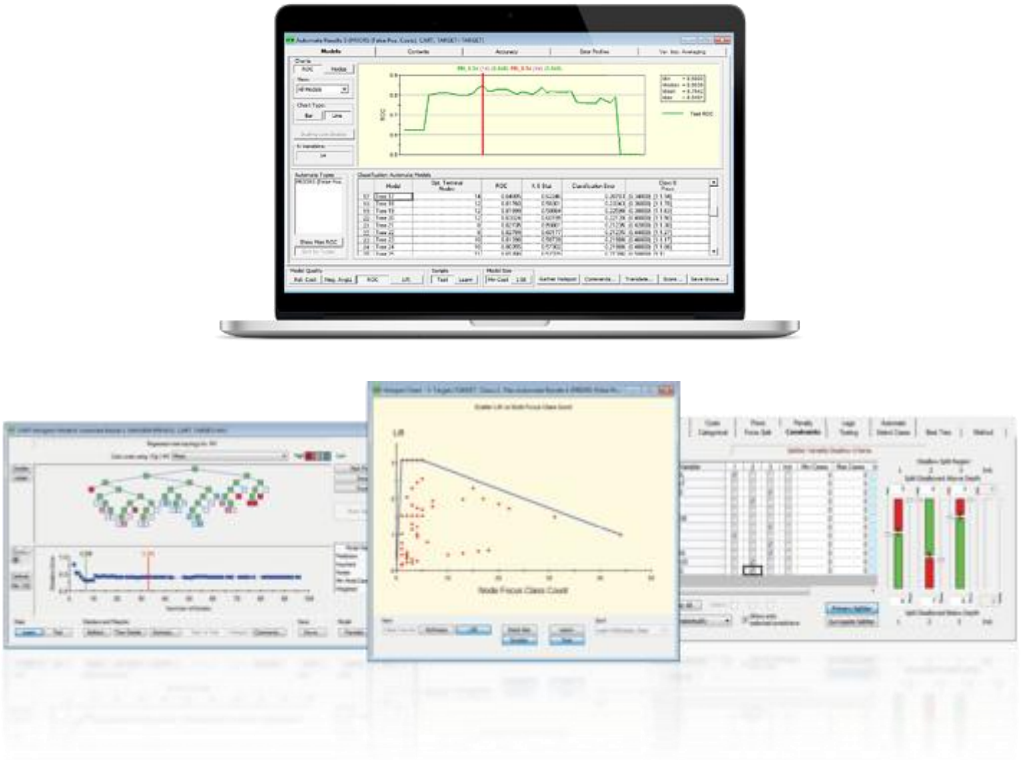
Su un intervallo di valori più ampio e un periodo di tempo prolungato, è probabile che le variabili seguano andamenti non lineari. Valori mancanti e outlier hanno maggiore probabilità di essere presenti influenzando l'efficienza degli strumenti di modellizzazione statistica (a causa di un singolo valore mancante, ad esempio, l'intera riga di osservazioni potrebbe non essere presa in considerazione).
Ad esempio, una banca che prova ad identificare transazioni fraudolente dovrà analizzare un numero molto elevato di predittori e osservazioni, con molti effetti non lineari complessi, valori mancanti e outlier. Lo stesso accade per siti di produzione che mirano ad identificare i fattori che abbattono i rendimenti o per aziende che monitorano i registri delle attrezzature di manutenzione per prevenire i guasti.

DATA MINING E ANALISI PREDITTIVA

Alcuni potenti algoritmi di Machine Learning come CART, Random Forests, TreeNet Gradient Boosting, e Multivariate Adaptive Regression Splines sono delle utili aggiunte agli strumenti posseduti da qualsiasi professionista, in particolare di fronte a dataset di grandi dimensioni. Queste tecniche basate su regole ben precise sono meno influenzate dalle limitazioni sopra

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

descritte, poiché non si basano su soglie di significatività statistica dei p-value e si basano invece su alberi decisionali con regole condizionate da IF, AND, OR che isoleranno gli outlier e "imputeranno" i valori mancanti.



I giorni dei dati di piccole e medie dimensioni, comunque, sono tutt'altro che finiti. Strumenti di modellizzazione statistica come il Design of Experiments rimarranno popolari ancora a lungo nell'ambito dell'ingegneria di processo, R&D, Qualità e Validazione, per ottimizzare attività e processi. Black Belt e Master Black Belt continueranno ad implementare strumenti di analisi dati Six Sigma per la root cause analysis, il miglioramento della qualità e dell'efficienza a tutti i livelli e, ai fini aziendali, utilizzeranno p-value per identificare predittori statisticamente significativi.

CONCLUSIONE

Ci stiamo addentrando in un futuro in cui viene richiesto alle varie organizzazioni aziendali di estrarre informazioni da una quantità sempre crescente di dati. Diventa dunque ancora più importante garantire che vengano scelti gli strumenti giusti per essere in grado di analizzare dati di dimensioni e complessità variabili.
I moderni strumenti di Machine Learning come CART, TreeNet, Random Forests e MARS offrono un'ottima scelta per dati di grandi dimensioni e / o relazioni più complesse, mentre le tecniche di modellizzazione più tradizionale come la regressione continueranno ad essere gli strumenti di scelta quando vengono soddisfatte le ipotesi di modello e l'obiettivo è trovare un'equazione semplice ed interpretabile.
Gli approcci della statistica e del Machine Learning svolgono un ruolo fondamentale nella ricerca di informazioni fruibili. Saranno poi la collaborazione e la comunicazione tra queste due discipline basate sui dati che consentiranno alle aziende di prendere decisioni migliori e ottenere un vantaggio competitivo.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

Per i clienti che si interfacciano con la gestione di grandi dataset complessi, l'offerta dei prodotti Minitab si è evoluta, integrando alcune piattaforme veloci e accurate per il data mining e l'analisi predittiva tra cui CART, TreeNet, Random Forests, MARS e altre metodologie.

[Hai difficoltà a gestire, comprendere e sfruttare a pieno i tuoi dati? Parla con GMSL]

Inizia ad esplorare il concetto di Machine Learning guardando i seguenti Webinar.