Master Universitario di I livello in: "ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE" A.A. 2022/2023

Titolo della tesi: Tecniche di data mining applicate al controllo di gestione industriale

Autore: Alessandro Caboni

Abstract

Attraverso l'utilizzo di tecniche di data mining si è tentato di individuare e di dare spiegazione alla relazione tra elementi costitutivi di un pezzo prodotto (n. di componenti, peso, ore uomo, ore macchina, etc…) e la marginalità dello stesso. Si è costruito un dataset di variabili ricavate tramite la lettura della distinta base e del ciclo dei pezzi prodotti. Si è effettuata un'analisi statistica descrittiva delle variabili per comprendere la natura del dataset e si è cercato di individuare una relazione tra margine e le variabili del dataset ricorrendo ai modelli di regressione lineare e di classificazione. La regressione è stata sviluppata per gradi: si è scelto di trovare la relazione tra variabili e costo, poi tra variabili e prezzo di vendita, infine tra variabili e margine. L'attività di regressione ha imposto un'accurata scelta delle variabili per evitare problemi di multicollinearità e per massimizzare l'efficacia del modello: durante il lavoro è emersa la possibilità che si stessero omettendo variabili significative o che la relazione tra margine e predittive non fosse lineare. Si è quindi ricorso alla combinazione di più variabili , alla loro trasformazione e alla stratificazione del dataset: questi passaggi hanno permesso di osservare che è possibile individuare una relazione in particolare tra il margine, il valore della tecnologia utilizzata e la complessità del pezzo prodotto. È un risultato parziale, poiché limitato a una porzione di dataset troppo ridotta, ma incoraggia a svolgere ulteriori approfondimenti e suggerisce di ampliare il dataset in futuro a variabili non solo legate alla "ricetta" di produzione (distinta e cicli) ma anche a variabili che riflettano le matematiche del disegno e il rapporto con il cliente, così da ridurre l'impatto di eventuali variabili omesse. Dopo questo primo risultato, frutto di parametrizzazioni manuali, ho deciso di ricorrere all'analisi delle componenti principali (PCA) per ridurre il numero delle variabili e provare una via alternativa. Tramite la PCA sono state individuate due componenti principali, interpretate come "complessità organizzativa" e "complessità tecnica" , due ambiti legati alla produzione di un pezzo. La regressione effettuata sulle componenti principali ha permesso di eliminare il problema della multicollinearità nella spiegazione di costo e prezzo di vendita e di comprendere in maniera più strutturata e rigorosa la natura delle variabili incluse nel dataset. Questo passaggio non ha permesso però di costruire un modello più robusto ed esplicativo nella spiegazione del margine poiché, anche combinando le varabili originarie con le componenti principali, non si è riusciti ad aggirare il problema delle presunte variabili omesse e della non linearità della relazione tra margine e predittive a disposizione. Per questo motivo si è successivamente deciso di cambiare la tecnica di analisi e di ricorrere alla classificazione, attraverso il metodo degli alberi decisionali. La scelta

muoveva dal fatto che , essendo molte le variabili discrete nel dataset, un modello basato su una sequenza di "if-then" avrebbe potuto essere più efficace. La scelta si è rilevata corretta ma ha mostrato che dietro ad ogni tipo di parametrizzazione utilizzata (max depth, Gini, entropia), era alto il rischio di over-fitting. Durante il lavoro svolto sono state condotte anche due piccole analisi minori: la market basket analysis sul venduto e una simulazione di comportamento tra R quadro, VIF e correlazione in sede di regressione lineare. La market basket analysis mi ha aiutato a capire se ci sia una relazione di trascinamento del venduto a cliente tra famiglie prodotto: è emerso che la vendita del "tubo" e quella del "filtro aria" hanno un legame spiegato da regole di confidenza e lift relativamente elevati, tanto da meritare un approfondimento in ambito di strategia commerciale. Nel secondo caso ho utilizzato la simulazione monte carlo per comprendere meglio un fenomeno che altrimenti avrei affrontato solo a livello di concetto teorico. Concludendo: l'attività di regressione e la PCA hanno messo in luce due aspetti importanti. Il primo è quello di approcciare secondo un framework più strutturato il lavoro d'analisi: distinguere tra complessità tecnica e complessità organizzativa, è una sottigliezza, non scontata a priori, che ha aiutato a strutturare l'indagine critica. Il secondo è quello che, arricchendo il dataset di dati non legati al manufatto ma al rapporto con il cliente, si possa arrivare a costruire un modello solido. La classificazione ha mostrato invece che la predizione tra margine alto/basso e le caratteristiche del pezzo è già effettuabile con i dati a disposizione, seppur nei limiti dell'over fitting: è un risultato promettente che induce a raccogliere ulteriori dati per arrivare ad un modello solido di previsione.