CAR PRICE ANALYSIS

Un'azienda automobilistica cinese, la Geely Auto, aspira ad entrare nel mercato statunitense installandovi la propria unità produttiva e producendo autovetture in loco per fare concorrenza alle controparti statunitensi ed europee.

Ha incaricato una società di consulenza automobilistica di comprendere i fattori da cui dipende il prezzo delle automobili. In particolare, vogliono capire quali sono i fattori che influenzano il prezzo delle auto nel mercato americano, poiché potrebbero essere molto diversi da quelli del mercato europeo.

Il nostro obiettivo è stato simile a quello di una società di consulenza, abbiamo effettuato una pulizia del dataset in modo tale da avere meno duplicati possibili e abbiamo selezionato le variabili quantitative che ci interessavano.

Tutte le variabili utilizzate hanno unità di misura americane

1) Regressione Lineare

Per prima cosa abbiamo effettuato una scalatura del dataset poiché gli ordini di grandezza di alcune variabili come il prezzo erano eccessivamente differenti rispetto ad altri ordini di grandezza.

Poi abbiamo effettuato la regressione lineare scegliendo come variabili indipendenti:

- Cavalli
- Lunghezza della macchina
- Rapporto di compressione
- Dimensione del motore

La scelta di queste variabili è stata del tutto empirica e dettata dall'intuito.

Notiamo che c'è un'ottima significatività per quanto riguarda i cavalli e la dimensione del motore rispetto alla variabile dipendente prezzo.

Vediamo che il vif assume tutto sommato valori inferiori a 5, questo significa che non c'è un'eccessiva collinearità tra le variabili e sia R-Quadro che Adj-R-Quadro hanno valori intorno allo 0.8, questo significa che abbiamo un buon modello che spiega molto bene la varianza teorica.

Il P-value del test-F è molto piccolo quindi rifiutiamo l'ipotesi nulla secondo cui tutti i beta siano uguali a zero.

Visualizzando la dipendenza lineare su un grafico bidimensionale vediamo che tutte le variabili si distribuiscono linearmente con il prezzo eccetto il RAPPORTO DI COMPRESSIONE

Cos'è il RAPPORTO DI COMPRESSIONE? DA COSA DIPENDE?

Il Rapporto di compressione è il rapporto tra (volume del cilindro quando il pistone si trova al punto morto nella fase di aspirazione) / (volume del cilindro quando il pistone si trova al punto morto nella fase di compressione). Notiamo che questo è un rapporto sempre >1 e che di fatto è indipendente dal prezzo della macchina bensì dal carburante che esso utilizza.

Le auto a benzina hanno un 0 < *compressionratio* < 9, Le auto a diesel hanno un *compressionratio* > 20, questo perché la mancanza della candela nei motori a diesel implica un rapporto di compressione maggiore che permette di aumentare la temperatura all'interno del cilindro in modo tale da accendere il gasolio per compressione.

2) CLUSTER ANALYSIS

La cluster analysis ha rafforzato ulteriormente la nostra tesi secondo cui la variabile prezzo potesse essere una buona variabile indipendente

Nel clustering gerarchico abbiamo utilizzato le seguenti variabili:

- *Wheelbase*
- *Lunghezza dell'auto*
- *Cavalli*
- *City mpg*
- *price*
- *Curb weight*

Nel cluster gerarchico abbiamo ottenuto quattro gruppi (utilizzando il metodo del gomito) tutto sommato di cardinalità omogenea. Uno di questi, che ha cardinalità minore, spicca sugli altri poiché costituito da auto più costose e prestigiose.

Nel caso del clustering non gerarchico e del metodo *K-means* abbiamo scelto di utilizzare tre gruppi.

Vediamo come il gruppo di destra (meno numeroso) è costituito da auto che hanno un prezzo maggiore, pesano di più e hanno un consumo maggiore e un numero di cavalli maggiore. Questo ordinamento lo ritroveremo anche nella *PCA*.

Abbiamo poi effettuato l'algoritmo *pam* per trovare un riscontro con la nostra analisi. Questo sceglie tra tutti i possibili algoritmi di clusterizzazione e seleziona quello ottimo, abbiamo poi valutato la *silhouette* per un numero di gruppi che vanno da 2 a 6 e notiamo che la migliore è proprio quella per tre gruppi.

Notiamo che la clusterizzazione effettuata con l'algoritmo pam è molto simile a quella che abbiamo effettuato con il *k-means*.

3) PCA

Effettuando l'algoritmo di *PCA* abbiamo visto come le variabili "*curbweight* price" e "*city mpg*" fossero correlate con la prima componente principale, questo ha senso infatti, spostandoci da sx verso dx il prezzo, il peso, la lunghezza aumentano e con essi aumentano anche i consumi. La variabile "city mpg" rappresenta le miglia per gallone che possono essere percorse dall'auto, vediamo che è inversamente correlata, questo suggerisce che all'aumento del prezzo il consumo aumenta.

Di conseguenza possiamo interpretare la prima componente principale come la "prestigiosità" della vettura o la "lussuosità".

La variabile *wheelbase* gioca un ruolo abbastanza importante nello spiegare la seconda componente principale. La *wheelbase* rappresenta la distanza tra l'asse di una ruota anteriore e l'asse della ruota posteriore posta sullo stesso lato, a differenza della pura *car-lenght* non è una semplice lunghezza, una *wheelbase* corta è associata ad una maggiore maneggevolezza della vettura e ad un migliore trasferimento di peso, al contrario, una *wheelbase* lunga conferisce maggiore stabilità in condizione di curve e soprattutto una maggiore abitabilità della vettura.

Di conseguenza la *wheelbase* può spiegare la seconda componente principale creando una nuova variabile che può essere interpretata come quanto la vettura si presta ad essere un'utilitaria.

CONCLUSIONI

Possiamo concludere dicendo che i fattori da cui dipende il prezzo delle auto americane sono sicuramente la dimensione dell'abitacolo, la dimensione dell'auto stessa , la potenza e il consumo. La Geely Auto dovrà puntare su auto di grandi dimensioni che consumano il giusto per far fronte alla concorrenza locale e venire incontro ai gusti degli statunitensi.

Alberto Sartini

Leonardo Galassi