Team Data sci-analyst

Metodologia di lavoro

I dataset (DS) fornitici riportano dati sul settore merceologi “scarpe”, sia da uomo che da donna:

n.3 dataset riportanti dati di scarpe da donna

n.1 dataset riportanti dati di scarpe da donna

Un primo scopo dell’analisi dei dati è quello di trovare delle correlazioni tra gli stessi che ci potessero dare delle indicazioni interessanti sullo spazio di mercato in cui fare la nostra proposta di lancio di un nuovo brand di scarpe, in ottica di estensione di brand.

Usando i notebook Jupyter con libreria pandas, abbiamo fatto una prima analisi della tipologi di dati presenti nei dataset. Da questa prima generica analisi, abbiamo estrapolato le seguenti informazioni: i data set riportano i dati di modelli di scarpe distribuiti attraverso il canale online, nel mercato statunitense; quest’ultima informazione è stata supportata dall’analisi delle valute presenti nel DS.  
Abbiamo quindi deciso di eliminare le righe che non riportavano la valuta USD; le righe eliminate sono risultate in quantità risibile, rispetto alla totalità delle righe dei DS.

Dall’analisi del contenute delle colonne, si è deciso di procedere lavorando solo con i dati di nostro interesse, estrapolando quindi solo le colonne che riportano i dati che abbiamo ritenuto utili per le nostre analisi, eliminando tutte quelle colonne che riportavano informazioni verosimilmente usate interamente dai diversi distributori, per la catalogazione dei diversi articoli; abbiamo altresì eliminato tutte le colonne che riportavano dati in quantità poco rilevante rispetto alla totalità degli items del DS.

Tra le colonne che riteniamo più interessanti abbiamo:

* Index, che ci dà un codice univoco per ogni articolo (differenziandolo anche per colore, taglia, prezzo, recensione)
* Prezzo max, che è il prezzo massimo a cui l’articolo è venduto
* Prezzo min, che è il prezzo minimo a cui quell’articolo è ventuto
* Colore
* Categoria
* Recensione

La colonna “recensioni” , anche se non presente in tutti i DS, ci è sembrata, in questa prima fase, quella più interessante per estrarre una prima indicazione sulla categoria di scarpe che ha destato più interesse tra quelle presenti, a prescindere che il giudizio sull’articolo sia negativo o positivo. Non avendo, infatti, dati di vendita, abbiamo ritenuto che la presenza delle recensioni ci desse intanto un’indicazioni di quali articoli sono stati venduti e quindi di maggiore interesse.  
Sia che il giudizio sia negativo che positivo, abbiamo quindi individuato uno spazio di azione; nel primo caso, possiamo agire sul miglioramento dell’articolo, nel secondo caso, possiamo inserirci in diretta concorrenza.

Per quanto riguarda l’analisi più approfondita delle colonne “categorie”, “colori” e “recensioni”, essendo queste colonne con struttura annidata di diverso tipo (dizionari, liste di stringhe), stiamo ancora valutando la strategia migliore per l’estrapolazione di tali dati, per averne una chiara lettura., ma dopo aver avuto indicazioni dal team MKTG su quali categorie di scarpe dobbiamo indirizzare la nostra ricerca all’interno dei dati.

Seguirà Notebook Jupyter con markdown e codice per avere un quadro più esplicito del lavoro che è stato e verrà fatto.