Documento di analisi

Gruppo Clothster Analysis

Pisaniello Sara- Arguirov Kristofer

1° appello Statistical learning

Indice

1.Panoramica iniziale:

Si è riflettuto a lungo sulla realizzazione di un progetto di Data science nel quale potessimo realizzare un blog, grafici, analisi e rispondere a domande che potessero essere significative, e prepararci per un futuro lavorativo.

Il reperimento di dati sul web (già analizzati) o porsi domande generiche non finalizzate ad un contesto lavorativo non ci è sembrato entusiasmante.

Ci è sembrato invece interessante unire una realtà didattica e lavorativa, così da poter realizzare un progetto che fosse ben realizzato al fine dell'esame, ma anche ulteriore arricchimento per noi come statistici con poca esperienza al di fuori dell'università.

Questo elaborato verrà successivamente esposto in azienda e ci darà l'opportunità di un confronto con un pubblico di "non addetti ai lavori".

2.Contesto aziendale e scopo dell'analisi:

Il padre di uno dei componenti del gruppo gestisce una azienda di tovagliato PVC in lombardia.

Si è reso disponibile a fornirci un dataset relativo alle vendite 2018, escludendo informazioni sensibili come numero di clienti e prezzo applicato.

L'azienda è leader in qualità del prodotto e utilizzo di colorazioni e trame innovative.

Ogni anno vengono scartati il 30% di prodotti del catalogo precedente ed introdotte novità.

Nel processo di selezione annuale di tovaglie da escludere in catalogo, vengono quasi unicamente considerate: la quantità di quella merce venduta l'anno precedente, i clienti (alcuni richiedono stampe personalizzate in catalogo), prezzo applicato.

Inoltre le decisioni di marketing vengono prese in "linea generale" e non considerano le differenze per paesi che potrebbero aiutare a diversificare l'offerta e promuovere prodotti in maniera più efficace.

Ci è stato richiesto di effettuare un'analisi che potesse accomunare paesi e che permettesse all'azienda di conoscere anche questo aspetti di preferenze qualitative differenti, quali colori, novità etc. e rendere le azioni di marketing più customizzate.

3.Descrizione e creazione del dataset:

3.a. Creazione del dataset

Il dataset iniziale si componeva di 12mila osservazioni quali: data di ordine, codice prodotto, soggetto della tovaglia, quantità vendute per paese (nel caso italiano anche regione).

In base a queste informazioni (e servendoci del catalogo 2018), sono state inserite manualmente una serie di caratteristiche che non erano presenti: colore prevalente della tovaglia, colorazioni spente/accesa, tipologia di disegno (fotografico, geometrico,

monocromatico etc.), tipo di foglia (liscia o goffrata), se la tovaglia era una novità o meno in catalogo, stagione rappresentativa per la tovaglia (natalizia, estiva, neutrale etc.).

Sono stati, inoltre, stornati tutti i prodotti che venduti nel 2018, ma non presenti nel catalogo 2018.

Alcuni clienti infatti, avevano ordinato vecchie stampe (catalogo 2017).

Questa decisione è stata presa perché non si conoscevano le caratteristiche di questi prodotti (non avevamo il catalogo 2017 a disposizione) e avrebbe sviato dalla nostra analisi che era volta a descrivere i comportamenti di acquisto del catalogo 2018.

L'analisi non considera gli ultimi ordini di dicembre 2018 perché in quel periodo vengono ordinate tovaglie del catalogo 2019, e sono state escluse per la stessa motivazione.

Infine abbiamo eliminato la variabile codice prodotto perché informazione sensibile, ma anche non interessante ai fini della nostra analisi (ci è bastato numerare le tovaglie da 1 a 177).

Questo lavoro si è rilevato piuttosto lungo, ma ci ha fatto riflettere sul fatto che non sempre le analisi che vengono richieste sono abbinate ad dataset già pronto per l'utilizzo.

3.b. Preparazione del dataset ai fini dell'analisi

Dopo la "creazione" del dataset, è stata neccessaria una suddivisione ulteriore in 2 parti.

È stato immediatamente evidente che il 60% delle vendite avveniva in Italia e non all'Estero, il confronto non era equo, quindi si sono considerate nel primo dataset le vendite per regioni di Italia e nel secondo le vendite per paesi esteri in generale.

Ai fini della cluster analysis si è effettuata una modifica ulteriore. Infatti metodi di questo tipo difficilmente gestiscono dataset molto ampi in maniera efficiente, per cui non abbiamo preso in considerazione la data dell'ordine dei prodotti (non è stata rilevante in questa parte dell'analisi ai nostri scopi, l'abbiamo considerata in un'altra parte dell'analisi) e abbiamo espresso il dataset in 177 osservazioni per tovaglie con le relative caratteristiche e quantità di venduto per paese.

4.Metodi Statistici

La nostra indagine è stata sia di tipo esplorativo, che mediante l'utilizzo di una cluster analysis model based.

4.a Analisi esplorative

4.a.1 Propensione all'acquisto

Inizialmente si sono considerate per paesi/regioni di Italia:

1) la propensione all' acquisto del proprio prodotto preferito

2) la copertura di gamma (quanti prodotti vengono acquistati di ogni catalogo)

Si è verificato che vi sono paesi o regioni che sono più propense alla diversificazione e alcune meno, ma che il prodotto preferito comunque viene acquistato sostanzialmente di più rispetto agli altri.

Le preferenze di prodotto sono importanti nei comportamenti di acquisto, in prima analisi.

4.a.2 Preferenze sulle colorazioni accese/spente

È risultato evidente come in Italia si tenda ad acquistare colorazioni accese, mentre all'estero vi siano paesi con preferenze di colorazioni più spente

Ci siamo chiesti se potessero esserci dei fattori geografici e leggendo articoli su internet (es. Kazuko Sakamoto, *Difference in the Color Preference by a Geographical Factor*) ci siamo accorti che era proprio così.

Nel nostro dataset questa la condizione non sempre si verifica, infatti ad esempio non ci saremmo mai aspettati che un paese come la Norvegia acquistasse il 77% di tovaglie a colorazione accesa.

Tutto ciò è giustificato dal fatto che le preferenze a volte non sono solo causate da effetti geografici ma talvolta anche da fattori culturali, psicologici e dall'età dell'acquirente.

4.a.3 Serie storiche delle tovaglie per colore

Abbiamo verificato innanzitutto che i colori accesi sono molto più venduti durante tutto l'anno. Questo è probabilmente dovuto al fatto che L'Italia che ne è maggior acquirente, preferisce colorazioni di tipo acceso prevalentemente.

Risulta esserci inoltre picco ad agosto negativo, ed è dovuto alla chiusura aziendale, quindi pochissimi ordini vengono registrati in quel periodo.

Per quanto riguarda i colori prevalenti durante l'anno invece vi sono:

- colori che vengono acquistati in maniera costante durante tutto l'anno (es. beige, blu scuro)

- colori che vengono acquistati di più nel periodo estivo (verde, rosso, giallo)

- colori più nel periodo invernale (grigi, marroni).

4.a.4 Preferenze sull'acquisto di tovaglie pubblicizzate

Ci siamo posti la domanda sulle vendite delle tovaglie pubblicizzate (40% del catalogo).

La pubblicizzazione avviene tramite copertina del catalogo (quella del 2018 ha totalizzato un complessivo del 3,5% di vendite sul totale), prodotti in esposizione in fiera e merchandising in vista in fiera.

La predominanza dei paesi e regioni italiane risponde positivamente al marketing ed acquista un 10% in più mediamente di questi prodotti rispetto a quelli non sponsorizzati.

Tuttavia vi sono alcuni paesi che preferiscono i prodotti non sponsorizzati come Cile, Paesi Bassi, Lituania.

4.a.5 Preferenze sull'acquisto di novità in catalogo

Le novità introdotte nel catalogo (circa 30%) del 2018, tendono a vendere mediamente di più dei prodotti in generale pubblicizzati. Questo rispecchia un effetto positivo nella scelta dei nuovi prodotti 2018, e anche alcuni paesi che tendono a questa innovazione.

Mediamente si acquista il 20% in più di prodotti novità in catalogo rispetto agli altri.

4.a.6 Maggiori acquirenti in Italia e all'Estero

Abbiamo individuati i maggiori acquirenti in Italia:

1) Lombardia 27,25%
2) Marche 14,30%
3) Sicilia 10,70%
4) Sardegna9,06%
5) Veneto 8,47%

E all'estero

6) Ungheria 23,27%
7) Spagna 8,01%
8) Francia 6,04%
9) Russia 5,91%
10) Irlanda 5,54%

Questo può essere utile per poter individuare a chi rivolgersi maggiormente per la diversificazione dei prodotti e campagne di marketing.

4.b Analisi di cluster

Si è inizialmente provato ad effettuare un'analisi di cluster per Paese/regione e prodotti venduti, per individuare comportamenti affini.

Tuttavia questo non è stato possibile dal punto di vista di statistico. La scarsità delle osservazioni (32 paesi esteri e 11 regioni) ha portato a poca evidenza di clusterizzazione (verificata anche dalla statistica di Hopkins).

Quindi la nostra cluster è stata effettuata sui prodotti venduti (177) con le loro relative caratteristiche e quantità vendute per paese.

Questo è stato effettuato al fine di individuare prodotti con caratteristiche simili nelle vendite e nelle features qualitative che ci permettessero di dare un'indicazione di come disporle all'interno del catalogo.

4.b.1 Preprocessing

Il primo step è stato quello di utilizzare un'opportuna misura che permettesse di scalare e creare la matrice di distanza dei nostri dati.

Si è scelto di utilizzare la distanza di Gower, che rappresenta il pre-processing standard nel caso di dati misti (numerici e categoriali).

Anche in questo caso si è operato un controllo relativo alla tendenza di clusterizzazione.

Prima tramite rappresentazione visuale (VAT), ovvero heatmap che permettono di notare visivamente la presenza di cluster, e non dati "random".

Poi tramite la statistica Hopkins, che permette di misurare la probabilità che i dati siano generati da una variabile uniforme. In altre parole, testa la casualità spaziale dei dati.

Se il valore generato dalla statistica è prossimo a 0.5, i dati non sono clusterizzabili, ma hanno evidenza di casualità.

In questo caso i dataset risultavano avere un forte evidenza di clusterizzazione.

4.b.1 Scelta del modello di cluster

Si sono inizialmente utilizzati metodi classici quali agglomerativo e divisivo.

Per verificare il giusto metodo agglomerativo e divisivo ci si è avvalsi della distanza di cophenetic.

Ovvero il metodo indaga la correlazione tra la distanza di cophenetic operata dal metodo (es. average, single linkage, ward etc.) e matrice di distanza operata dai dati originali.

Una buona approssimazione è data da valori sopra il 75%.

Ci si è subito resi conto che questi modelli e diversi metodi individuavano un 20-25% di outliers in prodotti di tovaglie.

L'andamento dei dendrogrammi era "a cascata".

Le tovaglie che vendevano di più erano quelle che sbilanciavano la classificazione, veniva creato un unico gruppo e tanti gruppi da 1 (outliers).

Si è concluso che questi non fossero i metodi corretti di analisi e non cogliessero la natura dei nostri dati.

La cluster empirica può risultare molto sensibile a scale differenti e non essere in grado di creare gruppi in caso di disparità evidenti (20-25% tovaglie molto acquistate rispetto alle altre).

Infatti come è stato evidenziato nelle esplorative c'è un effetto forte di predilezione nei confronti di prodotti preferiti rispetto ad altri.

Si è provato un altro metodo empirico: K-means, che sembrava funzionare un po' meglio sulle numerosità dei cluster e ne individuava gruppi, ma era molto instabile.

Infatti con inizializzazioni diverse cambiava completamente i risultati dell'analisi e la numerosità e formazione dei cluster era molto volatile.

Si è provato con tecniche più avanzate come il DBSCAN che non convergeva perchè i dati non erano particolarmente densi in alcune zone, ma piuttosto sparsi. Anche questo non era il metodo adatto.

La nostra soluzione finale è stata quella di utilizzare una tecnica di cluster avanzata di tipo Model Based Clustering, che è, per altro, una generalizzazione di un k-means con assunzioni sul modello di generazione di dati.

Effettuando un confronto tra misure di stabilità esterna ed interna la Model Based è risultata non scostarsi troppo a livello di efficacia dai modelli empirici.

Quest'ultima evidenza ci ha convinti definitivamente.

4.b.2 Clustering Model Based

La clustering Model Based è un metodo di soft assigment, ovvero ciascuna osservazione è trattata con un principio probabilistico di appartenere o meno ad un determinato cluster.

Nel nostro caso la separazione era piuttosto netta (a parte 1 o 2 tovaglie su 177 indecise tra più cluster).

La stima avviene tramite massima verosomiglianza ed un EM step algorithm.

L'assunzione è che i dati provengano da una mistura di normale multivariata.

La forma dei cluster varia per: forma, orientamento e dimensioni (matrice di varianza e covarianza del modello)

Vengono computati una decina di modelli che consentono e capire quale matrice di varianze e covarianze si adatti meglio ai dati (es. cluster sferici, diagonali, uguale o diversa dimensione o uguale diverso orientamento), valutando con il BIC.

Un ulteriore vantaggio è quello che l'algoritmo implementato, oltre che valutare il miglior modello, seleziona automaticamente il numero ottimale di clusters.

L'output è stato particolarmente chiaro ed evidenziava gruppi di tovaglie con caratteristiche simili, che in effetti venivano vendute in egual misura all'interno di paesi e regioni.

5. Conclusioni e risultati

Il nostro progetto ci ha permesso di ottenere risultati che possono essere utili dal punto di vista delle strategie di marketing nell'azienda di tovagliato.

Infatti si è verificato che è importante diversificare la promozione dei prodotti:

1) Ciascun paese e regione italiana predilige un proprio prodotto in particolare, quindi è importante saper descrivere le preferenze qualitative che portano a queste decisioni
2) Non tutti i paesi preferiscono lo stesso tipo di colorazione (accesa/spenta)
3) Sarebbe più corretto differenziare il marketing sulle tovaglie durante l'anno (colori che vengono venduti di più in estate e colori che vengono venduti di più in inverno)
4) Le tovaglie pubblicizzate nel 2018 sono vendute di più mediamente il 10% in più), ma c'è molto margine di miglioramento.
5) Le tovaglie novità del catalogo sono state selezionate bene ed hanno portato ad una preferenza del 20% . Tuttavia alcune novità in catalogo potevano essere più pubblicizzate (visto il successo di vendita).

6) I 5 maggiori acquirenti in Italia e all'estero rappresentano un fortissima fetta di mercato e sono quelli verso i quali andrebbe più indirizzata la diversificazione
7) La cluster operata ha individuato gruppi di tovaglie con caratteristiche simili per venduto e qualità delle tovaglie, quindi potrebbero essere proposte insieme in una eventuale promozione, e disposte nelle stesse categorie in catalogo.

6. Strutturazione presentazione progetto

La struttura della presentazione si è basata sulla lezione "effective comunication", tenuta nel corso di Statistical Learning.

Si è cercato di coinvolgere il pubblico utilizzando un video introduttivo, per mostrare anche il processo di produzione sottostante delle tovaglie e tramite grafici e slides molto sintetiche che colpissero all'occhio subito.

Per la realizzazione del video si è utilizzato un video-editor, Filmora, in versione gratuita e abbiamo salvato il file utilizzato il tool bandicom.

Per la realizzazione delle slides ci si è basati su video youtube che consigliavano presentazioni coerenti, richiamassero colori simili, e utilizzassero grafici coinvolgenti.

Ci si è serviti di Flaticon e Unsplash, come strumento grafico ulteriore.

I grafici sono stati generati con Rstudio, Tableau, Excel o a mano (nel caso di cartine e bolle di cluster personalizzate con le tovaglie scaricate dal catalogo ed inserite una a una).

7. Bibliografia

[1] ChrisFraley, A.E.Raftery, T.B.Murphyand, L.Scrucca(2012). *mclustVersion 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. Technical Report No. 597, Department of Statistics, University of Washington. Pdf

[2] Chris Fraley and A. E. Raftery (2002). *Model-based clustering, discriminant analysis,anddensityestimation.* JournaloftheAmericanStatisticalAssociation 97:611:631.

[3] Alboukadel Kassambara (2017) *Multivariate Analysis, Practical Guide to Cluster Analysis in R, Unsupervised Machine Learning,* Published by STHDA (http://www.sthda.com)

[4] Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York

[5] Kazuko Sakamoto, *Difference in the Color Preference by a Geographical Factor*

[World Academy of Science, Engineering and Technology International Journal of Humanities and Social Sciences Vol:7, No:11, 2013]

[6] Ranjan Maitra. *Model-based clustering*. Department of Statistics Iowa State University. http://www.public.iastate.edu/~maitra/stat501/lectures/