Machine Learning and Data Mining progetto: Black Friday

Michele Belletti

Corso AA 2019-2020

1 Identificazione del problema

Data l'evoluzione esponenziale dei negozi online bisogna trovare soluzioni per rendere competitiva una piattaforma rispetto ad un altra. Le idee proposte sono una stima delle vendite e una stima della categoria che all'utente potrebbe interessare. Queste stime ci permettono di avere una conoscenza a priori del volume di vendita di tale piattaforma e anche ci permette di gestire al meglio le vendite potendo così ricavare un utile maggiore.

2 Indici di valutazione e prestazione

Il metodo che andremo ad utilizzare per la valutazione del modello sarà basata sulla percentuale di successo della predizione e del valore dell'errore RSME. Il rapporto fra i dati utilizzati per la creazione della modello e i dati utilizzati per la realizzazione della verifica saranno rispettivamente l'80 % dei dati a disposizione per il primo mentre il rimanente 20% per la verifica.

3 Soluzione Proposta

Supponendo il problema lineare, perciò andremo a risolvere un problema di regressione. Gli algoritmi che andremo ad utilizzare sono: regressione lineare, decision tree regression e random forest regression.

Data la presenza di valori non definiti (NaN) nella categoria dei prodotti 2 e 3, opteremo per 2 diversi approcci:

- elimineremo le categorie 2 e 3, ipotizzando che siano sottocategorie della prima, cioè categorie meno importanti per descrive il prodotto;
- assegneremo un valore scorrelato al sistema che descrivono, in modo che non facciano variare la stima, in più aggiungeremo dei dati che ci permetteranno di stimare più correttamente.

Poi andremo a determinare un algoritmo per la determinazione della categoria migliore di prodotti da consigliare ad un utente. Tale sistema sarà costruito in modo generale in modo da funzionare in qualsiasi condizione, cioè sia utenti già inscritti che utenti nuovi.

4 Valutazione sperimentale

Inizieremo ora a creare il sistema in base alle ipotesi sopra descritte.

4.1 Data

La prima cosa che si deve fare per la descrizione e la realizzazione di un modello è l'analisi dei dati a disposizione.

I dati a nostra disposizione sono gli ordini effettuati durante il Black Friday e sono contenuti in un file csv avente le seguenti informazioni:

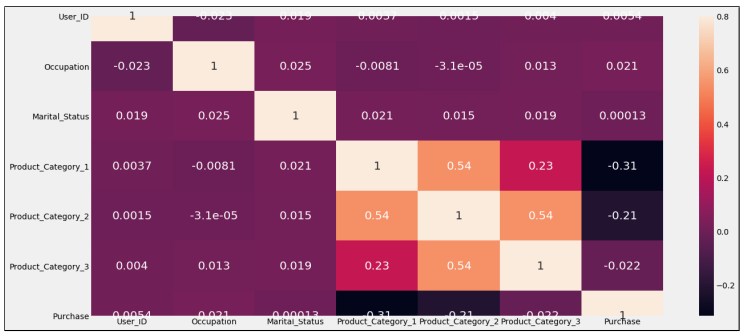
- L'utente attraverso il suo codice identificativo UserID;
- Il prodotto che l'utente ha comprato ProductID;
- L'eta dell'utente: Gender Age (valore che viene diviso ad intervalli di età 0-17, 18-25, 26-35, 36-45, 46-55, 55+;
- Il tipo di occupazione: Occupation (21 diverse occupazioni);
- La categoria della città: CityCategory (A,B,C che si suppone dipendano dalle dimensioni);
- Quanti anni sono passati da quando vive in quella città: StayInCurrentCityYears (0, 1, 2, 3, 4+);
- Stato civile: MaritalStatus;
- Le categorie del prodotto comprato: ProductCategory1, ProductCategory2, ProductCategory3;
- E la quantità dei soldi spesi: Purchase.

Ora analizzeremo i dati in funzione del numero di ordini (grafici 1-2-3-4-5-6): Si vede che gli ordini sono fatti principalmente da:

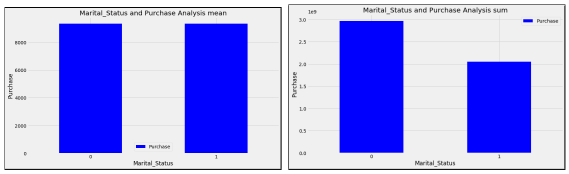
- persone di età tra i 26-35 anni, età nella quale si è più inclini ad utilizzare i mezzi della rete per acquistare i prodotti e si ha più disponibilità economica;
- le città di categoria B;
- i maschi e persone single;
- le persone con un occupazione di tipo 0, 4 e 7;

- e le persone che sono da circa un anno in città.

Analizzando la matrice di correlazione che esprime il grado di dipendenza tra due variabili si vede che non c'è una vera correlazione tra i dati a disposizione e il valore del purchase. Questo vale anche per il caso di voler suggerire una tipologia di prodotto, tranne che per le sottocategorie dove c'è una buona correlazione, ma non conoscendo neanche la prima categoria non possiamo neanche conoscere le sottocategorie il che ci porta a eliminarle nel calcolo della stima.



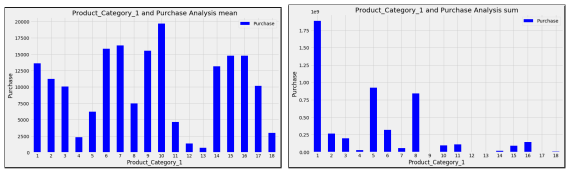
Andiamo ora ad analizzare le variabili indipendenti rispetto alla variabile Purchase che vogliamo prevedere:



Si ha che i grafici valutati con la media non differiscono tra i valori assegnati, l'unico grafico che non è uniforme in tal caso è quello che definisce la quantità del prodotto rispetto alla categoria. Dato che la media non appare come dato importante ma la loro quantità si, per migliorare la stima andremo anche a

tener conto della sua quantità. Questa aggiunta permette di avere una stima migliore, come vedremo in seguito.

Invece questo cambia nel caso della categoria del prodotto(grafici 9-10):



Analizzati i dati andremo a realizzare il modello:

4.2 Procedimento

I dati verranno tutti riportati in codice binario in modo che gli algoritmi possano gestirlo.

Come definito prima andremo ad utilizzare i seguenti algoritmi con le seguenti impostazioni:

- linear regression : normalizzazione attiva;
- DecisionTreeRegressor : max_ depth=15, min_ samples_ leaf=100;
- Random Forest: max_ depth=8, min_ samples_ leaf=150.

Adesso valuteremo questi 3 algoritmi con i diversi metodi proposti: il primo metodo utilizzato per predirre la quanti di soldi spesi sarà basato sull'eliminazione dei dati delle categorie dei prodotti 2 e 3. Elimineremo anche l'identificazione dell'utente perchè non influisce sulla predizione della spesa, così come il codice del prodotto. Queste due ultime eliminazioni permettono di stimare meglio l'aggiunta di un nuovo utente però peggiora la stima per un vecchio utente, dato che non c'è correlazione tra i suoi vecchi ordini. I risultati sono riportati nella prima tabella nella sezione risultati e discussione.

Il secondo metodo utilizza anche i dati della categoria 2 e 3, naturalmente si avrà una stima migliore dato che aggiungo altri dettagli all'ordine(13%). In più aggiungo anche il numero totale degli ordini con quel valore di eta, occupazione, tipo di categoria e identificazione del prodotto. Tali dati aggiuntivi permette di stimare meglio il risultato. I risultati sono riportati nella seconda tabella.

Fatto questo possiamo definire un sistema di raccomandazione per l'utente. Utilizzeremo i dati per l'ultimo metodo ed elimineremo ogni conoscenza rispetto alla categoria del prodotto. I risultati sono riportati nella terza tabella:

4.3 Risultati e discussione

I risultati sono i seguenti: usando [1]

Prima tabella:

Algoritmi	Linear Regression	Decision tree regression	random forest regression
Risultati	10.28%	63.90%	63.00%
RMSE	4708	2954	3008

Seconda tabella:

Algoritmi	Linear Regression	Decision Tree Regression	Random Forest Regression
Risultati	24.00%	70.16%	68.22%
RMSE	4344	2685	2802

Si ottiene che il migliore risultato è il secondo caso, cioè con i dati aggiuntivi definiti prima. In più l'algoritmo Decision tree regression ottiene una migliore predizione e un minore RMSE.

Terza tabella:

Algoritmi	Linear Regression	Decision tree Classifier	random forest Classifier
Risultati	10.71%	86.79%	82.72%
RMSE	3.552	2.56	2.807

Questi risultati dimostrano che il processo così costruito non fornisce una stima buona del tipo di prodotto consigliato. Per migliorare la stima si deve tener conto dell'utente, il che può essere fatto in due casi:

- l'aggiunta dell'utente all'algoritmo complessivo, il che porta ad avere un alto tempo di calcolo e il problema di quando si aggiunge un nuovo utente;
- oppure generare diversi algoritmi, uno per utente, che permetterà così di descrivere meglio le sue preferenze in più ridurrà il tempo di stima ma aumenta lo spazio in memoria perchè l'algorimo è definito in modo univoco.

Il codice è disponibile a `https://github.com/michbelle/Machine_learning_project.git`

Bibliografia

[1] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.