ANALISI DEL COMPORTAMENTO DEI CLIENTI DI UN E-COMMERCE PER PREVEDERE LE INTENZIONI D'ACQUISTO

Candidato: Simone Sinagra
Relatore: Luigi De Russis
Tutor aziendale: Michele Sonnessa

Introduzione

Il commercio elettronico è uno dei settori che sta maggiormente crescendo nel tempo grazie alla facilità di connessione, una maggiore fiducia negli acquisti online e alle consegne a domicilio sempre più efficienti.

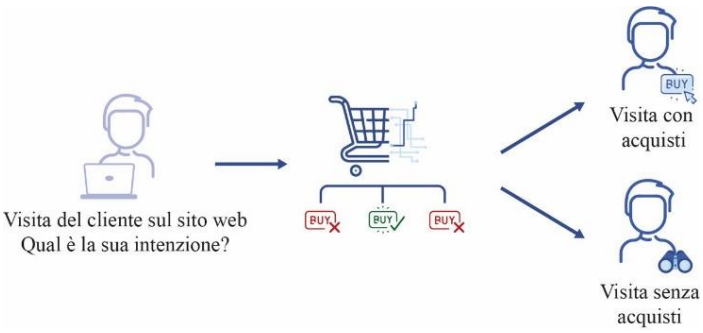
La Customer eXperience (CX) è il termine utilizzato per descrivere l'esperienza complessiva dei clienti che si relazionano con l'azienda, percorrendo il cosiddetto Customer Journey (CJ), ovvero il "viaggio" inteso come l'insieme di interazioni del cliente con l'azienda attraverso diversi punti di contatto (detti touchpoint). La cura e l'attenzione della CX assume un ruolo sempre più importante nelle scelte aziendali di marketing in quanto l'esperienza, positiva o negativa, di un cliente influisce notevolmente sul successo dell'azienda.

L'esigenza delle aziende che si interfacciano con il mondo digitale è quella di offrire ai clienti un'esperienza d'acquisto comparabile o migliore rispetto alla classica visita nel negozio fisico. Una delle strategie più ricercate nel mondo digitale, per poter offrire una CX più rilevante, è quella di mostrare contenuti che si adattano dinamicamente in base al cliente che sta utilizzando il sito web. Le modifiche alle pagine visualizzate dai clienti possono variare in base a diversi fattori del consumatore, tra cui le sue caratteristiche demografiche o del dispositivo che sta utilizzando, a seconda dei suoi comportamenti durante la sessione oppure alle interazioni che ha avuto nel passato con l'azienda.

La digitalizzazione ha permesso alle aziende di utilizzare strumenti di monitoraggio per raccogliere una grande quantità di dati generati durante la navigazione dei clienti sul sito web senza ostacolare l'esperienza di navigazione. In aggiunta, l'aumento della potenza di calcolo e dello spazio di archiviazione hanno permesso di analizzare efficientemente grandi volumi di dati per poter estrarre delle informazioni che aggiungono valore al business aziendale.

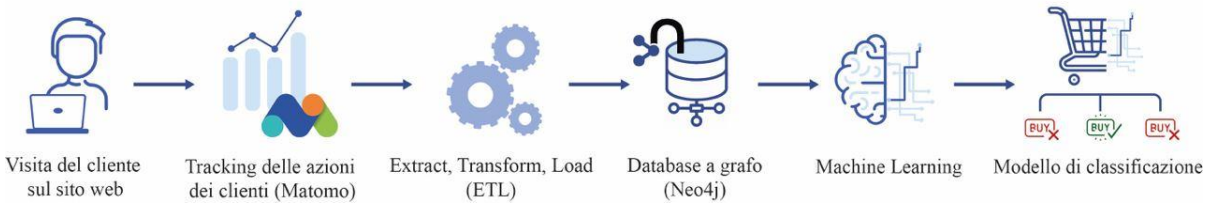
Obiettivo della tesi

Il progetto di tesi si focalizza principalmente sulla fase del CJ in cui il cliente si relaziona con l'azienda utilizzando il sito web dell'e-commerce per poter visualizzare ed acquistare i prodotti. L'obiettivo è quello di creare un buon modello predittivo che permetta di distinguere le visite dei clienti intenzionati ad acquistare dei prodotti dalle visite dei clienti che non effettuano alcun acquisto. Tra i risultati attesi, si auspica la generazione di un buon modello di classificazione attraverso l'utilizzo di un algoritmo di machine learning e le tracce che i clienti producono durante la navigazione sull'e-commerce.



Metodologia

Lo sviluppo del lavoro di tesi proposto è suddivisibile principalmente in due fasi, la prima riguarda l'acquisizione e la memorizzazione dei dati di navigazione dei clienti offerti da un software di tracking (Matomo), mentre la seconda comprende tutte le operazioni eseguite per la generazione del modello di classificazione delle visite.



Le piattaforme e-commerce, con l'ausilio di software di tracking, sono in grado di tracciare le attività dei loro clienti dal momento in cui accedono al sito fino al momento in cui lo abbandonano dopo aver acquistato o meno dei prodotti. Tuttavia, i dati di navigazione estratti dal software Matomo non risultano pronti per poter essere memorizzati direttamente nella base di dati in quanto necessitano di alcune modifiche per poter garantire la consistenza della base di dati. Tenendo conto di questa problematica, si è deciso di implementare un algoritmo basato sul processo "Extract, Transform, Load" (ETL) per poter estrarre i dati, sottoporli a delle modifiche per renderli compatibili con la struttura del database ed infine caricali nella base di dati.

Il progetto è stato sviluppato interamente in Python. Per la memorizzazione dei dati di navigazione dei clienti è stata scelta una base di dati a grafo (Neo4j), caratterizzata da un'elevata flessibilità ed efficienza nel gestire le relazioni tra i nodi delle azioni in ogni visita. Per poter interrogare e manipolare il database si è utilizzato il linguaggio Cypher.

Sviluppo del modello di classificazione

Dopo aver collezionato i dati provenienti dalle visite dei clienti sull'e-commerce in un periodo dalla durata di due mesi, si ottiene un dataset di partenza contenente 52.424 visite. Per poter generare il modello di classificazione è stato necessario eseguire una fase iniziale di visualizzazione dei dati seguita da una fase di pre-processing, in cui sono state effettuate diverse operazioni, tra cui la gestione dei valori mancanti, la trasformazione delle variabili categoriche, la riduzione della dimensionalità con la selezione delle feature più rilevanti (feature selection) e la progettazione di nuove funzionalità sfruttando la conoscenza del dominio (feature engineering).

In particolare, per poter arricchire il modello di classificazione con ulteriori informazioni delle visite del clienti sfruttando il processo di feature engineering, è stato necessario definire un numero di azioni che il cliente deve aver effettuato per poter classificare la sua visita. In questo modo nel momento in cui un cliente esegue N azioni, si hanno delle informazioni aggiuntive per poter addestrare il modello di classificazione (ad esempio la durata della visita, la quantità di articoli nel carrello…). Per poter scegliere il valore N, si è eseguita un'analisi della distribuzione delle lunghezze delle sessioni, intese come numero di azioni eseguite. Il risultato, mostrato in figura 2 attraverso un diagramma a scatola e baffi, mostra che il 25% delle visite non convertite hanno un numero massimo di azioni eseguite pari a quattro, mentre le visite convertite presentano un numero maggiore di azioni e hanno la mediana pari a circa undici azioni. In seguito a queste osservazioni si è deciso che il numero di azioni dopo cui l'algoritmo può classificare una visita è dieci, un numero sufficientemente grande per avere informazioni aggiuntive e allo stesso tempo non troppo grande per poter utilizzare il modello predittivo (ad esempio mostrando contenuti dinamici).

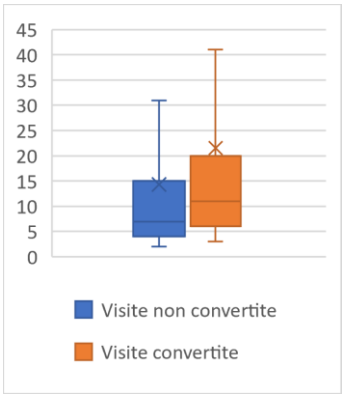


***Figura 1** - Diagramma della distribuzione dei valori di*

Dopo aver eseguito tutte le operazioni di pre-processing la dimensioni del dataset si è ridotte a 29.157 visite, di cui solo 2.010 relative a visite che presentano degli acquisiti. Rispetto al dataset di partenza è diminuita leggermente la condizione di sbilanciamento del dataset verso la classe delle visite non convertite; tuttavia, la differenza tra le due classi risulta ancora elevata e ciò avrà un impatto sul risultato finale.

Come classificatore si è deciso di utilizzare l'algoritmo di Random Forest (RF) perché si presta bene per risolvere i casi di classificazione come quello descritto in questo progetto di tesi ed inoltre è caratterizzato da una elevata robustezza al rumore e ad eventuali outliers, un'efficienza nella generazione del modello e un'accuratezza maggiore rispetto all'uso di un singolo albero decisionale. L'algoritmo RF può essere configurato attraverso diversi parametri ma non è possibile conoscere a priori quale sia la loro combinazione per poter ottenere il modello di classificazione più performante. Per poter ottenere dei risultati miglior, perciò, si è eseguita la tecnica di hyperparameter tuning, in cui sono stati testati un insieme di possibili valori ed è stata scelta la combinazione che ha generato il modello migliore.

Risultati e valutazioni

A partire dal dataset precedentemente descritto, sono stati utilizzati il 75% di dati per il train e il 25% di dati per il test. Il modello, generato utilizzando i dati di train, è stato testato utilizzando i dati di test. Per poter valutare la bontà del modello si sono considerate le misure dell'accuratezza, della precisione, del richiamo, e dell'F1-score. Quest'ultima misura è calcolata come la media armonica di precisione e recupero ed è particolarmente utilizzata nei casi in cui si utilizzano dataset con un'etichetta di classe predominante, come nel caso descritto in questo tesi (visite non convertite). I risultati ottenuti indicano un'accuratezza pari al 95.78% e una misura del punteggio F1 del 61.88%.

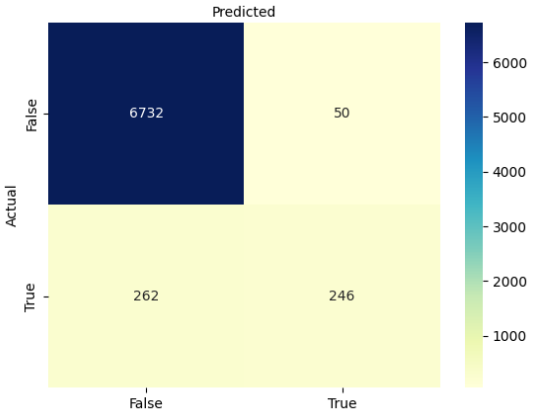


***Figura 2** - Matrice di confusione*

Osservando la matrice di confusione (figura 2) si osserva che la maggior parte delle predizioni sono fatte alla classe negativa e questo condiziona notevolmente il valore dell'accuratezza. I risultati ottenuti non mostrano una netta differenza tra i casi "true positive" e i casi "false negative", ovvero le sessioni dei compratori sono predette correttamente ed erroneamente in una quantità circa uguale.

Una delle caratteristiche negative di algoritmo di machine learning scelto è che il risultato della classificazione ottenuto è difficilmente interpretabile, in quanto una predizione può essere ottenuta da centinaia di alberi decisionali. Tuttavia, è possibile visualizzare e analizzare singolarmente gli alberi oppure ottenere una stima di quali siano le feature più importanti per la classificazione.

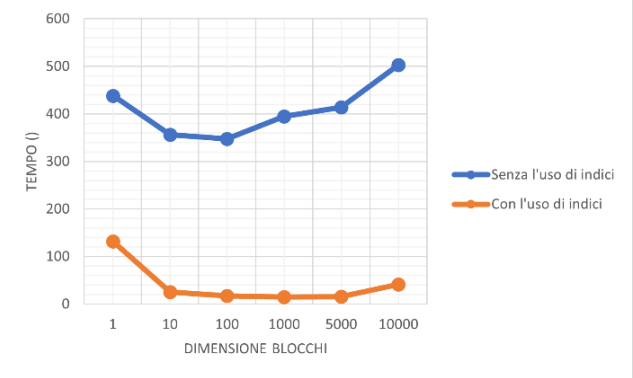


***Figura 3** - Tempi di memorizzazione nella base di dati a grafo*

Nello sviluppo del progetto di tesi si è rivolta particolare attenzione anche all'efficienza nell'implementazione del processo ETL, in particolare alla fase di caricamento dei dati nel database che richiede la creazione di molteplici nodi e relazioni. Sono stati provati diverse strategie, tra cui l'inserimento sequenziale e l'inserimento a blocchi. Quest'ultima ha mostrato tempi migliori, che sono stati ulteriormente migliorati introducendo la creazione degli indici sulle chiavi primarie dei nodi più ricercati. In figura 3 si può osservare la differenza dei tempi di inserimento, con o senza l'uso di indici, al variare della dimensione dei blocchi.

Conclusioni e lavori futuri

In questa tesi si è sviluppato un modello di classificazione delle visite dei clienti di un e-commerce utilizzando un algoritmo di machine learning. I risultati ottenuti hanno dimostrato che è possibile prevedere se una visita si concluderà con un ordine o meno analizzando le diverse caratteristiche implicite ed esplicite delle sessioni sul sito web. Nonostante il raggiungimento degli obiettivi prefissati, il modello presenta aspetti migliorabili negli sviluppi futuri, per esempio ricercando nuove feature provenienti da fonti dati diverse (es. il database aziendale contenente maggiori informazioni dei clienti).

I risultati hanno evidenziato il problema della natura del dataset, ovvero il fatto che il dataset era notevolmente sbilanciato verso la classe negativa, quella delle visite dei non compratori. Nonostante ciò, si è cercato di minimizzare il problema nella fase di pre-processing del dataset e il modello finale permette comunque di riconoscere circa il 50% delle visite dei compratori (classe di minoranza).

Il modello ottenuto potrà essere utilizzato in lavori futuri per prevedere le intenzioni d'acquisto dei clienti e modificare di conseguenza i contenuti in modo dinamico, per esempio mostrando delle call-to-action particolari per clienti non intenzionati a comprare, oppure guidando velocemente all'acquisto un cliente che è intenzionato a comprare.