DIPARTIMENTO DI STUDI PER L'ECONOMIA E L'IMPRESA

CORSO DI LAUREA MAGISTRALE IN MANAGEMENT E FINANZA

TESI DI LAUREA

PREVISIONE DELLE INTENZIONI DI ACQUISTO SU UN SITO WEB - UN APPROCCIO ROBUSTO

Relatore:

Prof. Aldo Goia

Correlatore:

Prof.ssa Clementina Bruno

Candidato:

Davide Visin

ANNO ACCADEMICO 2023-2024

A mamma e papà

INDICE

INTRODUZIONE

Negli ultimi anni, l'*e-commerce* ha rivoluzionato le abitudini di acquisto dei consumatori, che sempre più spesso si rivolgono a piattaforme *online* specializzate per soddisfare le proprie esigenze.

In questo contesto, per le aziende è diventato cruciale instaurare una connessione profonda e personalizzata con il consumatore, al fine di migliorare l'esperienza di acquisto e incentivare la finalizzazione di una transazione economica.

Per riuscire ad offrire all'utente un'esperienza fortemente personalizzata, il *Machine Learning* gioca un ruolo fondamentale.

Il *Machine Learning*, o apprendimento automatico, è una branca dell'intelligenza artificiale che consente ai sistemi di apprendere automaticamente dai dati e migliorare le proprie performance nel tempo. Questa tecnologia permette alle aziende di analizzare grandi quantità di dati e prevedere comportamenti futuri, come la probabilità che un consumatore completi un acquisto.

L'applicazione di tecniche di *Machine Learning* permette di analizzare e comprendere i comportamenti, le preferenze e le necessità dei consumatori in modo sempre più preciso, consentendo alle aziende di offrire un'esperienza d'acquisto personalizzata e ottimizzata, con l'obiettivo di massimizzare il tasso di conversione e la fidelizzazione.

Gli algoritmi di *Machine Learning* offrono una vasta gamma di tecniche per comprendere e prevedere i comportamenti dei consumatori. Tra questi, i modelli di regressione logistica rappresentano degli strumenti utilizzabili per classificare e prevedere se un consumatore finalizzerà o meno un acquisto, basandosi su una serie di variabili osservabili come il comportamento di navigazione e le interazioni con il sito.

La regressione logistica è particolarmente efficace quando le relazioni tra le variabili indipendenti e l'evento da predire sono lineari e le osservazioni non presentano anomalie significative.

Tuttavia, nel mondo reale i dati del comportamento dei consumatori possono essere altamente eterogenei e influenzati da fattori esterni, come fluttuazioni stagionali, eventi imprevedibili o dati anomali. In questi casi, l'utilizzo di una tecnica robusta applicata al modello di regressione logistica tradizionale diventa essenziale, poiché questo approccio è più resistente alla presenza di *outlier* (valori anomali) garantendo previsioni più accurate e affidabili.

Grazie a questi modelli, le aziende possono non solo prevedere con maggiore precisione quali consumatori sono più propensi all'acquisto, ma anche identificare le strategie più efficaci per

migliorare la loro esperienza d'acquisto, personalizzando le offerte e ottimizzando i processi decisionali.

Data l'importanza della tematica introdotta, l'obiettivo di questo studio risulta essere quello di fornire una serie di procedure e strumenti utili per determinare con precisione i principali *driver* che spingono il consumatore ad effettuare un acquisto su un sito web.

Il presente studio si articola in tre aree principali: la prima parte riguarda una revisione della letteratura esistente, la seconda è dedicata all'analisi dei dati del caso in studio e l'applicazione della regressione logistica standard, l'ultima parte invece, è incentrata sull'applicazione di una tecnica robusta alla regressione logistica standard per ovviare alle problematiche che possono insorgere per la presenza di dati anomali.

La prima parte, costituita dalla revisione della letteratura, è suddivisa in tre sezioni:

Nella prima sezione, si esamina l'importanza del concetto di *Customer Experience* nel processo decisionale del consumatore, approfondendo anche il tema dei *Big Data* (ossia le enormi "moli" di dati che vengono generate da sistemi informatizzati) per sottolineare la rilevanza di queste tematiche nell'*e-commerce*.

La seconda sezione presenta un'analisi di come altri studi hanno affrontato problematiche simili a quelle introdotte e gli strumenti utilizzati.

Infine, la terza sezione, esplora l'uso di strumenti di terze parti, come le metriche fornite da Google, che possono contribuire alla previsione delle intenzioni di acquisto online.

Nella seconda parte dello studio, si introduce il caso studio utilizzato, il quale prevede l'analisi di un dataset che raccoglie diverse variabili di contesto legate ai comportamenti dei visitatori di un sito web di *e-commerce*, verranno forniti dettagli sulla sua origine e saranno descritte le variabili che lo compongono. Inizialmente, verrà svolta un'analisi di tipo descrittiva, accompagnata da rappresentazioni grafiche delle variabili, per poi approfondire lo studio delle correlazioni tra le variabili e l'intenzione di acquisto.

In questa parte verranno introdottie sia le metodologie con cui le problematiche del dataset del caso in studio saranno trattate, sia l'applicazione del modello di regressione logistica standard.

La terza parte si focalizza sull'introduzione e applicazione dei modelli di *Machine Learning* robusti, inizialmente verranno confrontate le performance dei modelli standard e robusto su dati simulati per poter verificare la bontà dell'approccio proposto in questo studio, successivamente verrà applicato al

dataset il modello robusto per poi confrontare i risultati ottenuti con quelli ottenuti con un approccio standard.

Le performance dei modelli applicati verranno confrontate utilizzando strumenti di diagnostica appositi come la matrice di confusione, sia sul dataset oggetto di studio sia su una simulazione creata *ad-hoc* per dimostrare l'impatto dei valori anomali sulla capacità predittiva dei modelli.

Il presente lavoro si concluderà con un capitolo conclusivo in cui verranno illustrate le implicazioni manageriali derivanti dall'applicazione dei modelli, in ottica strategica per un'azienda che vuole gestire al meglio il proprio sito di *e-commerce* attraverso l'implementazione di modelli di classificazione; le implicazioni manageriali saranno frutto di un'attenta lettura dei risultati generati dai modelli.

Per svolgere le analisi viene utilizzato il linguaggio di programmazione R, per ulteriori informazioni riguardo al codice utilizzato fare riferimento al capitolo Appendice di questo studio.

EXCURSUS BIBLIOGRAFICO

L'evoluzione tecnologica e la crescente centralità del consumatore hanno trasformato profondamente le dinamiche competitive delle aziende, specialmente nel settore dell'*e-commerce*.

In un mercato sempre più globalizzato, la capacità di distinguersi non è più legata unicamente a fattori tradizionali come prezzo e qualità, ma si fonda sempre più sull'abilità di creare esperienze coinvolgenti e memorabili per i clienti. È in questo contesto che emerge l'importanza della *Customer Experience*, un approccio che integra aspetti sensoriali, emotivi e relazionali per costruire un legame duraturo tra consumatore e azienda.

Parallelamente, i *Big Data* hanno rivoluzionato il modo in cui le aziende gestiscono e utilizzano le informazioni.

Grazie alla crescente capacità di raccogliere e analizzare enormi quantità di dati, le imprese possono comprendere meglio il comportamento dei consumatori e adattare le proprie strategie in tempo reale. Nell'*e-commerce*, i *Big Data* consentono, ad esempio, di migliorare l'efficacia del processo decisionale e la comprensione dei principali *driver* che trasformano un semplice visitatore in cliente.

Infine, l'adozione di tecniche di *Machine Learning* ha ulteriormente potenziato la capacità delle imprese di anticipare e rispondere alle esigenze dei consumatori. Attraverso modelli predittivi basati sui dati comportamentali raccolti online, le aziende possono identificare in anticipo le intenzioni di acquisto dei clienti e personalizzare l'offerta in modo più efficace. Questi strumenti consentono non solo di aumentare i tassi di conversione, ma anche di offrire un'esperienza al cliente fluida e intuitiva, riducendo le frizioni durante la navigazione. L'utilizzo di strumenti di terze parti come Google Analytics, inoltre, arricchisce ulteriormente la capacità di analisi, integrando dati provenienti da fonti esterne fornendo *insight* preziosi per ottimizzare le strategie digitali.

Customer experience, e-commerce e Big Data

In un contesto sempre più globalizzato, risulta molto più complesso per le aziende creare un vantaggio competitivo duraturo e redditizio.

Lo studio proposto da [1], analizza il ruolo della *Customer Experience* nel marketing moderno e fornisce un modello interpretativo utile per le aziende che intendono creare un'esperienza di valore per i consumatori.

La crescente competitività dei mercati globali ha portato molte aziende a spostare la loro attenzione dai tradizionali fattori di differenziazione, come prezzo e qualità, verso un approccio più centrato

sull'esperienza del cliente, riconoscendo in essa una leva strategica fondamentale per ottenere un vantaggio competitivo sostenibile.

Una strada possibile per rimanere competitivi è focalizzarsi sull'esperienza del consumatore, infatti, circa l'85% dei *senior manager* ritiene che focalizzarsi solo sui tradizionali elementi di marketing come prezzo, prodotto e qualità non sia più sufficiente.

Questo approccio appena descritto, si rende necessario in quanto i modi con cui un potenziale cliente riesce ad entrare in contatto con l'azienda sono incrementati a dismisura, e dunque, risulta fondamentale per un'azienda di successo monitorare e fornire miglioramenti nella relazione cliente-azienda.

In questa prospettiva, si sviluppa il concetto di CRM (*Customer Relationship Management*), il quale implica la considerazione non solo degli aspetti razionali tradizionalmente studiati nel marketing che vede il consumatore come un *problem-solver* perfettamente razionale (e quindi focalizzato, ad esempio, su prezzo e qualità del prodotto) ma bensì di tutti quegli aspetti irrazionali che stanno emergendo solo negli ultimi anni (ad esempio le emozioni, elementi intangibili legati all'azienda e al prodotto) che possono creare valore per il cliente e costituire fonte di un vantaggio competitivo importante.

Il concetto di *Customer Experience* nasce a metà degli anni '80 del Novecento, in contrasto con la letteratura esistente fino a quel momento, che vedeva il consumatore come perfettamente razionale nelle scelte di acquisto, come già accennato; alla dimensione razionale nella scelta di un prodotto si aggiunge la dimensione esperienziale che pone in risalto le emozioni che il prodotto o l'azienda possono suscitare in un individuo.

L'obiettivo principale di un'azienda in un ottica esperienziale, è quello di fornire al consumatore non tanto la vendita di esperienze "memorabili" ma rendere unica l'esperienza che il consumatore vive dall'inizio alla fine con l'azienda in modo tale da soddisfare tutte le sue aspettative nel miglior modo possibile così da "legare" il cliente all'azienda non solo attraverso le leve tipiche del marketing ma anche coinvolgere tutto il suo "bagaglio" emozionale.

Nasce così il concetto di *Customer Experience* che viene definito come "*un insieme di interazioni tra un consumatore e un prodotto, un'azienda o una parte di un'organizzazione che suscitano una reazione*", che pone in risalto tutte quelle variabili che erano state sempre tralasciate negli approcci tradizionali.

Il concetto di *Customer Experience* non è univoco e, come evidenziato dallo studio sopracitato, può essere suddiviso in diverse componenti esperienziali.

Componente sensoriale

È fondamentale coinvolgere i sensi del consumatore, offrendo stimoli che possano essere percepiti a livello sensoriale. Ad esempio, un sito web ben progettato può trasmettere una sensazione di bellezza all'utente.

Componente emozionale

Questa componente mira a suscitare stati d'animo, sentimenti e emozioni nel consumatore, contribuendo a creare un legame affettivo con l'azienda.

Componente cognitiva

Consiste nel facilitare la formazione di connessioni mentali tra il consumatore e l'azienda, creando così degli "standard" di riferimento nella mente del cliente.

Componente pragmatica

Si riferisce all'offerta di un'interazione con l'azienda che sia semplice e fluida, priva di frizioni e quanto più intuitiva possibile.

Componente dello stile di vita

Questa componente implica la costruzione di valori autentici legati all'azienda, nei quali il consumatore si identifica, contribuendo così alla creazione di un vantaggio competitivo per l'azienda.

Componente relazionale

Si concentra sul coinvolgimento del consumatore e della sua rete sociale, favorendo la creazione di relazioni solide tra l'azienda e i clienti. Ad esempio, si possono sviluppare comunità online in cui gli utenti discutono dei prodotti offerti da un determinato sito web.

Dall'analisi delle principali componenti fondamentali della *Customer Experience,* discendono importanti implicazioni manageriali per un'azienda, le quali sono scomponibili nei seguenti punti:

Sviluppare innovazioni guidate da esperienze: le innovazioni che puntano maggiormente sulle esperienze offerte all'utente hanno maggiori possibilità di successo rispetto a quelle che si basano solo su aspetti tradizionali.

Considerare le caratteristiche funzionali della propria offerta commerciale: in quanto per creare un vantaggio competitivo duraturo diventa fondamentale studiare oltre agli aspetti tradizionali anche gli aspetti emozionali del consumatore come la sua percezione dei prodotti e le emozioni suscitate.

Una visione integrata della *Customer Experience*: permette di coinvolgere il cliente in tutte le fasi della relazione con l'azienda, un esempio può essere la proposta di un sito *web* di offrire prodotti personalizzati su misura per il cliente che siano unici per ognuno di essi.

Le componenti della Customer Experience possono risultare molto differenti a seconda delle caratteristiche di un determinato prodotto.

Nell'ultimo decennio con il rapidissimo progredire delle tecnologie digitali si è assistito ad una crescita rapidissima dell'*e-commerce:* secondo le analisi condotte da [2], nel 2023 oltre il 19% delle vendite a livello *retail* mondiali sono state effettuate attraverso attraverso Internet e siti di *e-commerce.*

Di pari passo con la crescita di Internet e dell'attenzione verso il consumatore, si è verificata la crescita esponenziale di dati che possono essere raccolti ed elaborati dalle aziende durante la loro attività; nasce così il concetto di *Big Data,* enormi moli di dati che possono essere manipolati grazie alla crescente potenza di calcolo disponibile sul mercato e al continuo progresso delle tecniche di elaborazione dei dati stessi.

Lo studio presentato da [3] evidenzia come l'analisi dei *Big Data* possa migliorare le strategie e le operazioni nell'*e-commerce*, ponendo in risalto, inoltre, il valore commerciale derivante dalla loro implementazione.

Le imprese operanti nel settore *e-commerce* che utilizzano tecniche di analisi dei *Big Data* integrate nelle loro catene del valore registrano una produttività in media più alta del 5%-6% rispetto ai competitor che non integrano questo tipo di strategie.

Negli Stati Uniti, principale economia mondiale, il 91% delle compagnie presenti nell'indice americano *Fortune 1000* investe in programmi aziendali di gestione dei *Big Data.*

Nel mondo *e-commerce,* strategie basate sui *Big Data* permettono alle aziende di conoscere i comportamenti e le preferenze di ogni singolo individuo che transita sul loro sito, permettendo così una gestione molto più approfondita di ciò che il consumatore gradisce e/o non gradisce sul sito *web* e permettono anche di comprendere quali siano i principali *driver* che portano il semplice visitatore del sito a diventare un cliente. Inoltre i *Big Data* rendono estramamente più rapidi e precisi i processi decisionali aziendali.

Dunque risulta abbastanza scontato il motivo per cui le aziende che operano nell'*e-commerce* siano quelle che hanno i tassi di crescita più alti per quanto riguarda le adozioni di strategie nell'ambito *Big Data.*

Nello studio viene evidenziato come nel settore *e-commerce* siano cruciali i sistemi basati su *Big Data* che presentano dei benefici informativi immensi soprattutto legati, ad esempio, all'assistenza in tempo reale dei clienti al *pricing* dinamico dei prodotti ed alle offerte personalizzate al controllo delle preferenze degli utenti).

Gli autori dello studio forniscono un elenco di caratteristiche legate ai Big Data a cui un'azienda operante nel mondo *e-commerce* deve tener conto per sfruttare al massimo i vantaggi forniti dai dati raccolti senza commettere errori, che vengono riportate di seguito.

Volume

Le analisi dei *Big Data* mettono "in gioco" una quantità immensa di informazioni, bisogna assicurarsi che le informazioni siano quanto più pulite possibile e pronte all'uso, in quanto devono essere della qualità più alta possibile, pena i cattivi risultati che possono essere ottenuti a fini predittivi/modellistici.

Varietà

Con questa caratteristica si denota il fatto che i *Big Data* non siano costituiti da dati proveninenti da una sola fonte, bensì da innumerevoli fonti che originano a loro volta tipologie differenti di dati con caratteristiche diverse fra loro. Un esempio può essere fornito dall'unione di dati anagrafici che un utente fornisce al sito e *pattern* di consumo regionali nella regione in cui vive quel determinato utente.

Velocità

Questa caratteristica fa riferimento alla frequenza di generazione dei dati e dunque alla necessità di sincronia tra la generazione di nuovi dati e i processi decisionali aziendali, infatti tanto più i dati sono raccolti velocemente e tanto più rapidamente si riescono a prendere decisioni *data-driven,* tanto maggiori saranno le opportunità potenzialmente vantaggiose per l'azienda.

Veridicità

Prima di impostare una strategia basata sui *Big Data* è necessario assicurarsi che i dati generati/raccolti siano il più attendibili possibile; l'azienda, quindi, deve essere in grado di saper distinguere e riconoscere i dati di alta qualità da quelli di qualità bassa in quanto l'utilizzo di dati di bassa qualità non apporta nessun vantaggio.

I *Big Data* vengono trattati all'interno del sistema azienda secondo un approccio RBV (*Resource Based View*) proposto da [4] che vede le fonti interne dell'azienda come uno dei principali fattori di successo che devono essere valorizzati ed impiegati adeguatamente

Lo scopo principale dell'analisi dei Big Data è dunque quello di generare *insight* portatori di benefici economici per l'azienda e può essere visto come l'insieme dei benefici transazionali, informativi e strategici che un'azienda può ricavare dall'elaborazione dei *Big Data*.

Machine learning e applicazioni nell'e-commerce

Nel mondo *e-commerce* le aziende devono "fronteggiare" sia dati strutturati come ad esempio nomi, età, indirizzi, sia dati non strutturati come ad esempio click su un sito, intervalli di tempo trascorsi; è necessario dunque riconcilliare tutti questi dati attorno al singolo utente.

Come già detto in precedenza, vi è una marcata importanza specialmente in ambito *digital* dei dati riguardanti il comportamento del consumatore, infatti secondo quanto riportato nello studio condotto da [4] emerge che la comprensione fin dall'inizio delle intenzioni di acquisto di un consumatore in un sito è la chiave per una strategia di *digital marketing* di successo.

Uno dei principali modi per ottenere un vantaggio competitivo è quello di essere in grado di poter prevedere le intenzioni di acquisto del consumatore utilizzando dati relativi al suo comportamento su un sito.

Per poter prevedere le intenzioni di acquisto vengono utilizzati dei modelli basati su ML (*Machine Learning*) che utilizzano dati di contesto (ad esempio la tipologia di browser, il tempo trascorso sul sito, il giorno in cui viene effettuato l'accesso…), raccolti al termine di ogni sessione, che permettono la comprensione dell'intenzione o meno di acquisto.

I principali modelli utilizzati in letteratura sono algoritmi di classificazione e si registrano differenti approcci a questa tipologia di problemi, come ad esempio l'utilizzo combinato di più algoritmi di classificazione o l'utilizzo di singoli algoritmi.

Precedentemente si accennava al concetto di sessione, che può essere definita come l'intervallo di tempo che un utente trascorre su un sito *web* sulla base della descrizione fornita da [5].

Una sessione è costituita da log di utilizzo del sito *web* che permettono di riscostruire l'attività che l'utente esegue sul sito in studio. La buona costruzione di un modello di ML dipende dalla corretta estrazione dai log delle variabili importanti che siano correlate con l'intenzione di acquisto.

I dati provenienti dai siti web vengono definiti come un vero e proprio "termometro" che permette di controllare in tempo reale lo stato di salute di un sito *web* ([6], Capitolo 1).

Le informazioni che possono essere raccolte sono innumervoli: per evitare problemi legati a quantità eccessive di dati da elaborare, viene consigliato dalla letteratura, di non vedere il proprio sito *web* come un semplice "silo" di informazioni, bensì integrare il proprio sito all'interno della propria strategia aziendale a 360 gradi.

Riprendendo lo studio proposto da [4], viene posto in evidenza un *framework* che possa permettere, solo sulla base di fattori contestuali, la previsione dell'intenzione di acquisto da parte di utenti senza un *background* nel sito, che, dunque, non sono provvisti di tutta una serie di informazioni storiche sul loro conto nel sito oggetto di studio; si tratta di utenti che eseguono quindi un "*cold start*".

Il sopracitato approccio prevede la raccolta di dati provenienti dalla piattaforma *e-commerce* sulla quale il sito internet si basa, una fase di *pre-processing* dei dati per creare un dataset "pulito" ed utilizzabile. Dopodichè, dopo aver generato un dataset che contenga tutti i dati delle varie sessioni registrate sul sito, si scelgono le variabili più importanti per il modello di ML utilizzato che possano essere in grado di spiegare l'intenzione di acquisto.

I modelli più diffusi per questo tipo di analisi secondo la rassegna bibliografica fornita da [5] sono: Alberi Decisionali, Reti Neurali, Reti Neurali Ricorrenti e Regressione Logistica.

Una volta che il modello è stato addestrato sul set di variabili contestuali ritenute più idonee, si procede con la valutazione delle *performance* del modello (o dei modelli) attraverso l'utilizzo di strumenti come la matrice di confusione, che permette di valutare la corretta identificazione da parte del modello degli utenti che con maggior probabilità effettueranno un acquisto.

Nel dataset oggetto di studio da parte di [4] emerge che fattori come il tipo di sistema operativo, la posizione geografica dell'utente, la valuta utilizzata siano fattori che possono modellare correttamente l'intenzione di acquisto del consumatore.

Da ciò discendono importanti implicazioni manageriali che riguardano l'applicazione del citato *framework,* che permette lo studio di strategie *ad-hoc* per segmentare e proporre contenuti personalizzati agli utenti sulla base di dati contestuali, riuscendo ad ottenere così un maggior numero di conversioni sul sito.

Nel caso proposto da [5] il modello di ML viene addestrato su variabili come: il numero di prodotti visitati dall'utente durante una sessione, il numero di prodotti cliccati, il tempo trascorso visitando la pagina di un prodotto, se la visita è avvenuta o meno durante un weekend.

Nelle analisi trattate in [4], [5] il dataset presentava per ovvie ragioni uno sbilanciamento delle classi della variabile risposta (che è costituita da una variabile dicotomica che assume valore 1 se l'utente acquista, 0 viceversa) in quanto le persone che acquistano sono un numero estremamente più basso rispetto a quelle che non acquistano durante una visita di un sito web; per questa ragione i dati legati all*'e-commerce* utilizzati per le analisi di classificazione spesso sono dati "sbilanciati", infatti secondo Shopify [7], noto operatore nel settore *e-commerce*, il tasso medio di conversione per un sito è di circa del 3%.

Questa criticità è stata trattata attraverso tecniche di *oversampling* che permettono la creazione di nuove osservazioni attraverso tecniche statistiche senza alterare le caratteristiche originarie del dataset; sono presenti anche approcci basati su *undersampling* che permettono di riequilibrare le classi senza creare nuove osservazioni.

Non esistono in letteratura approcci basati solamente su algoritmi di *Machine Learning* di classificazione, bensì, attraverso tecniche di *clustering,* è possibile segmentare i visitatori in gruppi omogenei sulla base di dati contestuali utilizzando algoritmi di apprendimento non supervisionati che possono far emergere dei *pattern* latenti nel comportamento degli utenti sul sito (ad esempio il browser utilizzato) secondo l'analisi del caso studio proposto da [8].

Strumenti di terze parti: Google Analytics

Una strategia basata solo su dati contestuali fondata solamente su dati interni può non essere ottimale in tutti i casi, infatti esistono diversi strumenti di terze parti (ad esempio Google Analytics di Google) che permettono di integrare nella propria strategia *digital* ulteriori indicatori che possono prevedere l'intenzione di acquisto del consumatore ([6], Capitolo 1).

Si identificano tre principali fonti di informazioni nel caso di un sito *web*:

Metriche *Offsite*

Sono metriche disponibili *online* provenienti da una miriade di fonti e facilmente accessibili ad esempio siti che si occupano di fornire dettagli e dati su altri siti *web.* I dati *offsite* sono difficilmente utilizzabili e conciliabili con dati raccolti *onsite.*

Metriche *Onsite*

Metriche e dati disponibili internamente, ad esempio dati di log forniti da un *server* su cui è ospitato il sito.

Metriche di terze parti legate al sito

Sono tutte quelle metriche che vengono fornite internamente attraverso *software* di terze parti, i quali arricchiscono i dati potenzialmente raccoglibili *onsite.* Il precedentemente citato Google Analytics fa parte di questa categoria.

Google Analytics, dunque, diventa un prezioso alleato per poter ricavare ulteriori informazioni. Esempi di metriche interessanti possono essere: il tasso di rimbalzo, comprendere quali sono le pagine che contribuiscono maggiormente a creare valore per il cliente, capire se il sito aiuta il cliente a costruire una relazione di fiducia con l'azienda.

È utile sottolineare nuovamente, anche in ottica Google Analytics, il concetto già visto in precedenza collegato alle criticità dei *Big Data* emerse nello studio condotto da [3], sostenendo che raccogliere troppe informazioni solo perché sono facilmente disponibili porta solo a maggiore confusione e incapacità decisionale.

Inoltre Google Analytics integra al suo interno dei modelli di ML che possono contribuire fin dall'inizio a prevedere gli acquisti dei visitatori.

UNA PRIMA ESPLORAZIONE DEL DATASET UTILIZZATO PER IL CASO STUDIO

Per lo svolgimento del caso studio, si è deciso di utilizzare un dataset adatto allo scopo, disponibile sul sito di UCI Machine Learning Repository [10].

Il dataset raccoglie 12330 sessioni registrate su un sito *web* che si occupa di *e-commerce B2C (Business to Consumer)*, gli autori hanno estratto il log dal sito in maniera tale da poter collegare tutte le variabili di contesto raccolte con l'intenzione o meno di acquisto attraverso una variabile dicotomica.

L'estrazione è avvenuta in maniera tale che, ciascuna delle 12330 sessioni disponibili, appartenga ad un solo utente per evitare le visite ripetute di uno stesso utente che possano in qualche maniera "falsare" i dati estratti creando ad esempio delle tendenze inesistenti.

Non si registrano valori mancanti o variabili con osservazioni parziali.

Per ciascuna osservazione, sono state raccolte 18 variabili in grado di spiegare il comportamento che l'utente manifesta nel sito oggetto di studio. Sono stati raccolti sia dati di contesto legati al comportamento sul sito, sia dati provenienti da fonti terze (Google Analytics).

Sono state raccolte le seguenti variabili di contesto, di seguito riportate con il loro nome originale.

Administrative

La variabile raccoglie il numero di pagine di natura amministrativa totali visitate dall'utente durante la sessione. Si intendono, ad esempio, quelle pagine legate alla gestione dell'account utente.

Administrative Duration

Contiene il tempo (in secondi) che l'utente trascorre nelle pagine legate alla gestione dell'account sul sito.

Informational

La variabile rappresenta il numero di pagine di natura informativa visitate dall'utente durante la sessione. Rientrano sotto questa categoria tutte quelle pagine in cui è possibile trovare informazioni legate al sito, ad esempio: l'indirizzo dell'azienda e le condizioni di vendita.

Informational Duration

È costituita dal tempo (in secondi) che l'utente trascorre nelle pagine legate alla parte informativa del sito.

Product Related

Rappresenta il numero di pagine legate ai prodotti che l'utente visita. Rientrano in questa classificazione le pagine che descrivono le caratteristiche del prodotto, il prezzo.

Product Related Duration

La variabile rappresenta il tempo (in secondi) che l'utente trascorre nelle pagine legate ai prodotti in vendita sul sito.

Special Day

Questa particolare variabile, mostra la vicinanza ad uno specifico giorno dell'anno (come ad esempio la Festa della Mamma, il giorno di San Valentino). Il valore di questa variabile è determinato considerando le dinamiche tipiche di un sito di *e-commerce* come il tempo di attesa tra la realizzazione dell'ordine e l'effettiva consegna di quanto ordinato. Un esempio, può essere fornito dal giorno di San Valentino (14 Febbraio): la variabile assume un valore diverso da zero nei giorni compresi tra il 2 Febbraio e il 12 Febbraio, oltre questo intervallo, assume valore 0 salvo che non vi sia l'immediata prossimità di un altro "giorno speciale".

Operating System

Tipologia di sistema operatitvo utilizzato dall'utente per collegarsi al sito. In questo studio si sceglie di non considerare questa variabile in quanto non sono forniti dettagli utili per determinare con esattezza il tipo di sistema operativo utilizzato.

Browser

Tipologia di browser utilizzato dall'utente per collegarsi al sito. Anche in questo caso non sono forniti dettagli approfonditi, perciò si sceglie di omettere questa variabile.

Region

Indica da quale regione proviene la visita al sito, non essendo forniti dettagli specifici sulle regioni si sceglie di omettere la variabile.

Traffic Type

Rappresenta la sorgente del traffico con cui l'utente giunge sul sito. Non è disponibile un dettaglio adeguato delle fonti, perciò si sceglie di omettere anche questa variabile nel presente studio.

Visitor Type

Rappresenta la tipologia di visitatore del sito, identifica 3 tipologie di utenti: “Nuovo Visitatore”, “Visitatore non Nuovo”, “Altro”.

Weekend

Indica se la sessione è avvenuta in un weekend o durante un giorno settimanale.

Month

Indica in quale mese dell’anno è avvenuta la sessione di un utente sul sito. Sono stati registrati tutti i mesi ad eccezione di Gennaio e Aprile.

Per quanto riguarda le variabili raccolte attraverso lo strumento Google Analytics si rilevano:

Bounce Rates

La variabile illustra la frequenza media di rimbalzo delle pagine vistate dall’utente, può essere vista come il rapporto tra le sessioni composte da una sola pagina e il numero totale di sessioni [11]. Di fatto non esiste un valore ideale a cui un gestore di un sito web dovrebbe aspirare, dipende dall’obbiettivo che si vuole raggiungere. Ad esempio, un basso valore della frequenza di rimbalzo sarebbe auspicabile in un caso come quello oggetto di questo studio in cui si desidera che l’utente, visiti ed esegua azioni su numerose pagine al fine di acquistare un prodotto.

Exit Rates

Rappresenta la frequenza media di uscita delle pagine visitate dall’utente nel sito oggetto di studio. Il valore fornito da Google Analytics[12] permette di comprendere, in percentuale, quante volte una determinata pagina è stata oggetto di un evento di uscita dal sito da parte dell’utente.

Page Values

La variabile è costituita dal valore medio (in termini monetari) delle pagine che un utente visita prima di completare una transazione come consultabile in [13]. Ad esempio, se una pagina non è stata coinvolta in nessuna transazione sul sito, il suo valore sarà zero. Questa metrica, permette di capire quali sono le pagine di un sito *web* che contribuiscono maggiormente alle conversioni e che quindi, forniscono maggior valore per l’utente. Questa metrica viene calcolata ripartendo il valore della transazione eseguita per le singole pagine che l’utente ha visitato prima di finalizzare la transazione. Per poter visualizzare questa metrica è necessario che la funzionalità *e-commerce* di Google Analytics sia abilitata come nel presente caso in studio.

Infine il dataset presenta la variabile risposta:

Revenue

Rappresenta attraverso una variabile dicotomica l'intenzione o meno di acquisto da parte di un utente del sito. Sarà la variabile risposta dei successivi modelli di classificazione che verranno descritti nei capitoli successivi per modellare l'intenzione di acquisto.

ESPLORAZIONE DETTAGLIATA DEL DATASET

In questo capitolo verrà svolta un'approfondita analisi del dataset in studio, nella prima parte del capitolo ci si concentrerà sull'analisi esplorativa del dataset, che consiste nello studio delle singole variabili che compongono il dataset, affiancando sia indicatori numerici come i quantili, medie, valori minimi e massimi, sia rappresentazione grafiche idonee a rappresentare la variabile in studio.

Nella seconda parte invece, ci si concentrerà nello studio delle correlazioni che le variabili hanno in relazione alla variabile risposta dei futuri modelli di classificazione che verranno costruiti in seguito.

Analisi esplorativa del dataset

Attraverso questa analisi sarà possibile identificare possibili fonti di distorsione nei dati che permetteranno di cogliere e di implementare le opportune azioni correttive oltre a comprendere come siano strutturate le variabili a livello di distribuzione delle osservazioni.

Per ogni variabile si sceglie la rappresentazione grafica più idonea alla tipologia di variabile (discreta/continua/categoriale).

Variabili discrete: si sceglie la rappresentazione attraverso istogrammi o grafici a barre che riportano sull'asse Y le frequenze assolute osservate e sull'asse X classi di valori osservati.

Variabili continue: vengono rappresentate attraverso i *density plot*, particolari grafici che permettono di visualizzare la distribuzione di una variabile numerica. Sull'asse Y viene riportata la densità di probabilità mentre sull'asse X i valori osservati.

Variabili categoriali: la rappresentazione avviene mediante grafici a barre che riportano sull'asse Y le frequenze assolute osservate, sull'asse X le etichette delle modalità osservate. Nel caso di variabile dicotomica, ovvero che presenta solo due modalità, si sceglie la rappresentazione mediante grafico a torta.

Per quanto riguarda le variabili discrete/continue, si fornisce una tabella riassuntiva dei principali indici numerici di sintesi che riassumono le informazioni numeriche principali della variabile in studio.

Invece, per le variabili categoriali, si fornisce una tabella che indica il numero assoluto di osservazioni rilevate per ogni modalità.

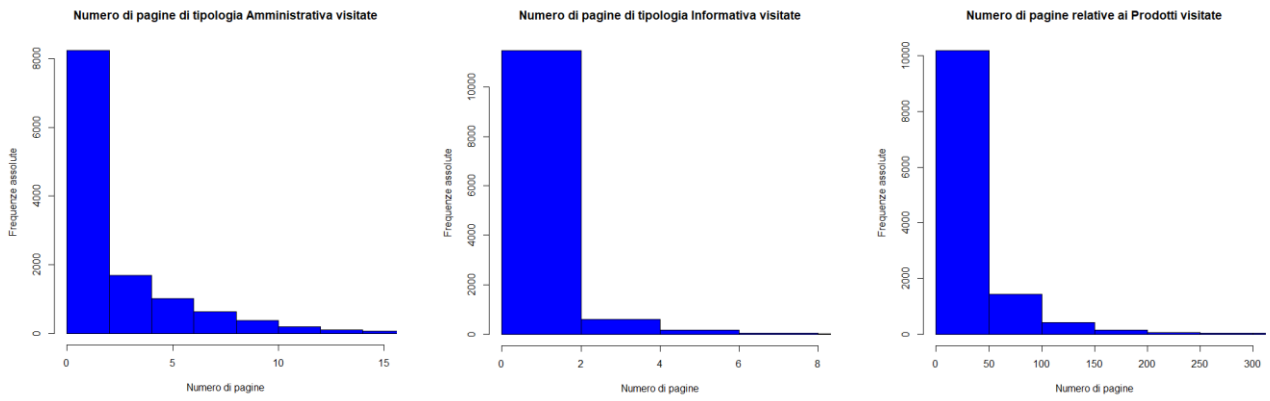

Figura 1 Istogrammi tagliati sull'asse X delle variabili Administrative, Informational e Product Related

	Administrative	Informational	Product Related
Valore Minimo	0	0	0
1° Quartile	0	0	7
Mediana	1	0	18
Media	2.315	0.5036	31.73
3° Quartile	4	0	38
Valore Massimo	27	24	705
Deviazione Standard	3.3218	1.2702	44.4755

Tabella 1 Indici numerici riassuntivi variabili Administrative/Informative/Product Related

Le variabili Administrative, Informative e Product Related sono discrete in quanto rappresentano il numero assoluto di pagine visitate delle rispettive categorie da un utente durante una sessione. Dalla Figura 1 emerge come tutte e tre le variaibli presentino una forte asimmetria che le comprime verso sinistra. Osservando i dati della Tabella 1 si può entrare nel dettaglio di ogni variabile in studio, in particolare:

Variabile Administrative: presenta moltissime ossservazioni con valore nullo, ciò viene confermato dall'osservazione del primo quartile, pari a 0, che permette di concludere che almeno il 25% delle osservazioni presenti un valore nullo. La media, assume un valore superiore della mediana (che rappresenta il 50% della popolazione) fornendo un'indicazione di asimmetria nella distribuzione della variabile. Si registra inoltre una certa dispersione attorno alla media, per via del valore della deviazione standard, pari a 3.3218, insieme ad un valore massimo pari a 27, suggerisce la presenza di outlier, ossia di valori anomali, il cui numero è pari a 404.

Variabile Informational: la distribuzione risulta fortemente concetrata su valori nulli; infatti, dall'osservazione dei quartili, emerge che almeno il 75% delle osservazioni manifestano un valore

nullo. La media superiore alla mediana (che ricordiamo, rappresenta il 50% delle osservazioni), fornisce un indicatore di asimmetria. Il valore massimo osservato pari a 24, insieme ad una forte variabilità dei dati rispetto alla media, rappresentata dalla deviazione standard pari a 1.2702, porta a ipotizzare la presenza di valori anomali, che sono infatti, 2631.

Variabile Product Related: anche per questa variabile la distribuzione è asimmetrica, in quanto il valore della media è superiore alla mediana. Dai quartili emerge che il 75% delle osservazioni sia compreso tra 0 e 38 e almeno il 25% delle osservazioni ha visto 7 pagine legate ai prodotti. Si registra un valore massimo pari a 705, insieme alla deviazione standard molto elevata pari a 44.4755, fornisce indizi su una distribuzione che contiene valori anomali, i quali risultano essere 987.

Da questa analisi e dai grafici della Figura 1, emergono dunque, criticità legate a distribuzioni fortemente asimmetriche e presenza di outlier che possono inficiare delle prestazioni di modelli di classificazione che verranno in seguito utilizzati, per modellare l'intenzione di acquisto di un utente.

I dati osservati risultano comunque ragionevoli, in quanto è lecito pensare che un utente durante la sua navigazione sul sito non navighi sempre in tutte le sezioni del sito (dunque il fatto che vi siano delle distribuzioni fortemente concentrate su zero è dovuto a questo).

Per quanto riguarda gli *outlier* possono essere dovuti ad utenti che visitano un numero elevato di pagine prima di acquistare (o valutare un acquisto).

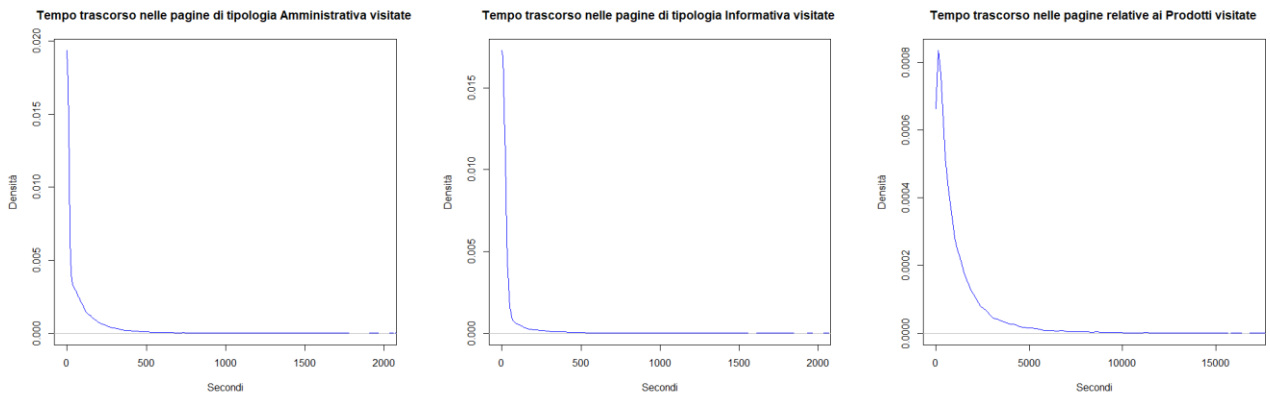

Figura 2 Density Plot tagliati sull'asse X delle variabili Administrative Duration, Informational Duration e Product Related Duration

	Administrative Duration	Informational Duration	Product Related Duration
Valore Minimo	0	0	0
1° Quartile	0	0	184.1
Mediana	7.5	0	598.9
Media	80.82	34.47	1194.8
3° Quartile	93.26	0	1464.2
Valore Massimo	3398.75	2549.38	63973.5
Deviazione Standard	176.7791	140.7493	1913.669

Tabella 2 Indici numerici riassuntivi variabili Administrative Duration/Informative Duration/Product Related Duration

Le variabili Administrative Duration, Informational Duration e Product Related Duration sono continue in quanto rappresentano il tempo (in secondi) che un utente trascorre nelle varie sezioni del sito durante una sessione. Dai *density plot* osservabili nella Figura 2, è possibile comprendere come anche in questo caso, vi sia una forta asimmetria nella distribuzione delle variabili. Entrando nel dettaglio degli indici numerici presenti in Tabella 2, si possono trarre le seguenti conclusioni per ogni variabile:

Variabile Administrative Duration: la distribuzione presenta numerose osservazioni nulle, in quanto osservando il primo quartile si deduce che almeno il 25% delle osservazioni è nullo. Si registra una forte differenza nei valori della mediana pari a 7.50 e della media pari a 80.82, che ci permette di concludere che vi siano dei valori molto alti che "trascinano" la media verso l'alto, come ad esempio, il valore massimo osservato pari a 3398.75. Osservando anche la deviazione standard pari a 176.7791 si ipotizza la presenza di numerosi valori anomali, che sono in questo caso pari a 1171.

Variabile Informational Duration: anche in questa variabile la distribuzione è fortemente asimmetrica e concentrata su valori nulli, infatti il primo quartile, terzo quartile e la mediana, sono valori nulli, permettendoci di concludere che, almeno il 75% delle osservazioni hanno un valore nullo. La media pari a 34.47 risulta molto più alta della mediana che ci permette di concludere che ci siano dei valori che spostano verso l'alto la media, come, ad esempio, il valore massimo pari a 2549.38. La deviazione standard pari a 140.7493 permette di concludere che vi sia una marcata dispersione attorno alla media. Sono presenti 2405 valori anomali.

Variabile Product Related Duration: la distribuzione, in questo caso, risulta essere sempre asimmetrica ma leggermente meno rispetto alle due variabili precedenti, infatti il primo quartile risulta essere diverso da zero pari a 184.10 permettendoci di concludere che in questa variabili le osservazioni nulle siano sicuramente un numero inferiore rispetto alle altre due variabili. La media pari 1194.80, assume un valore nettamente più alto rispetto alla mediana 598.90, fornendo anche in questo caso indizi sulla presenza di valori che spostano la media verso l'alto, infatti si osserva un valore massimo di ben 63973.50 che insieme ad una forte dispersione attorno alla media data dalla deviazione standard pari a 1913.669 segnala la presenza di outlier, che risultano essere 961.

In questo caso, si può notare, anche paragonando i grafici della Figura 1 e della Figura 2, che le distribuzioni sia del numero di pagine visitato che della rispettiva durata trascorsa dall'utente nelle varie pagine seguano dinamiche analoghe con distribuzioni fortemente asimmetriche.

Risulta comunque lecito ritenere attendibili i valori osservati nella Tabella 2, in quanto, è ragionevole pensare, che un utente che visita un sito di *e-commerce* probabilmente passi più tempo cercando i prodotti di suo gradimento rispetto a visitare le parti informative del sito oppure quelle legate alla gestione dell'account, infatti da quanto emerge osservando la Figura 2 si nota come la distribuzione legata alla durata trascorsa sulle pagine legate ai prodotti sia sensibilmente più simmetrica e meno concentrata su zero rispetto a quella amministrativa e informativa.

I valori estremi, invece, possono essere dovuti ad utenti che fanno un intenso utilizzo del sito, ad esempio, perché fortemente indecisi e la loro scelta richiede molto tempo, oppure banalmente sono utenti che lasciano aperta la sessione sul loro PC.

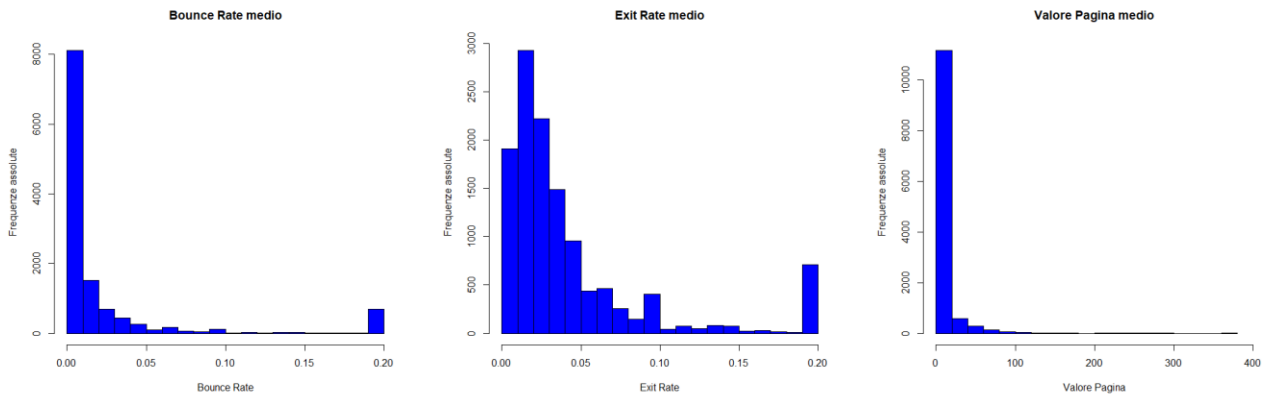

Figura 3 Istogrammi delle variabili Bounce Rates, Exit Rates e Page Values

	Bounce Rates	Exit Rates	Page Values
Valore Minimo	0	0	0
1° Quartile	0	0.0143	0
Mediana	0.0031	0.0252	0
Media	0.0222	0.0431	5.889
3° Quartile	0.0168	0.05	0
Valore Massimo	0.2	0.2	361.764
Deviazione Standard	0.0485	0.0486	18.5684

Tabella 3 Indici numerici riassuntivi variabili Bounce Rates/Exit Rates/Page Values

Le variabili Bounce Rates, Exit Rates e Page Values sono variabili discrete e rappresentano rispettivamente: il tasso di rimbalzo medio delle pagine visitate dall'utente, il tasso di uscita medio delle pagine visitate dall'utente, e il valore pagina medio delle pagine visitate dall'utente.

Queste variabili, come già detto in precedenza, sono frutto dell'estrazione dei dati forniti da Google Analytics, i quali permettono di aggiungere variabili legate alle *perfomance* del sito che possono potenzialmente fornire informazioni preziose sull'intenzione di acquisto.

Dagli istogrammi della Figura 3 emerge una certa asimmetria nelle variabili, entrando più nel dettaglio analizzando anche i principali indici numerici della Tabella 3 si osserva:

Variabile Bounce Rates: la maggior parte dei valori risulta concentrarsi verso valori prossimi allo zero come viene evidenziato dai quartili e dalla mediana, che ci permettono di comprendere che il 50% dei valori osservati siano compresi tra 0 e 0.0031, l'asimmetria viene evidenziata dal fatto che la media pari a 0.0222, assume un valore più elevato della mediana pari a 0.0031. Si registra un'ampia

variabilità rispetto alla media, confermata dal valore della deviazione standard pari a 0.0485. Vengono registrati ben 1551 outlier.

Variabile Exit Rates: osservando la Figura 3, si può cogliere una minor asimmetria rispetto alle altre due variabili in studio in questa parte (Bounce Rates e Page Values). I valori osservati risultano molto piccoli, osservando i quartili si nota che il 75% della popolazione è compresa tra 0 e 0.05. La minor asimmetria, è dovuta al minor scostamento tra mediana pari a 0.0252 e media pari a 0.0431, rispetto alle altre variabili. Si registra comunque, una deviazione standard pari a 0.0486, che risulta essere alta rispetto alla media. Inoltre l'osservazione di un valore massimo elevato pari a 0.20, permette di ipotizzare l'esistenza di valori anomali che risultano essere 1099.

Variabile Page Values: la variabile presenta una distribuzione estremamente asimmetrica, concentrata su valori nulli sulla base dell'osservazione dei quartili, infatti, dal terzo quartile si deduce che almeno il 75% delle osservazioni è nulla. Inoltre la media pari a 5.889 è estremamente più alta della mediana pari a 0, insieme ad un valore massimo di 361.764 e una deviazione standard di 18.5684, suggerisce la presenza di outlier che sono ben 2730.

Dall'analisi svolta sopra, emergono delle variabili fortemente asimmetriche con un elevato numero di outlier, sono risultati che non stupiscono dato che riguardano metriche di un sito web.

Per esempio, dei valori fortemente concentrati verso zero sono assolutamente plasuibili ed auspicabili per un sito di *e-commerce* come quello in studio (in particolar modo per le variabili Exit Rates e Bounce Rates) in quanto sono indice di un sito che è in grado di catturare l'attenzione dell'utente ed evitare che l'utente interrompa la sua visita. Discorso opposto invece, vale per la variabile Page Values, in questo caso, è auspicabile che l'utente visiti pagine che hanno un alto valore pagina, in quanto, le pagine che presentano valori pagina più elevati, sono tutte quelle pagine che hanno effettivamente contribuito maggiormente nelle conversioni precedenti al periodo in studio, risultando più attrattive. Le pagine che hanno valore zero, sono tutte pagine che non hanno mai contribuito nel processo di conversione dell'utente, dunque, non dovrebbero essere considerate in una strategia *digital* efficace. Si osserva che almeno il 75% delle osservazioni registrate sul sito, abbia un Valore Pagina medio pari a zero, suggerendo quindi dei margini di miglioramento in tema di *Customer Experience* del sito.

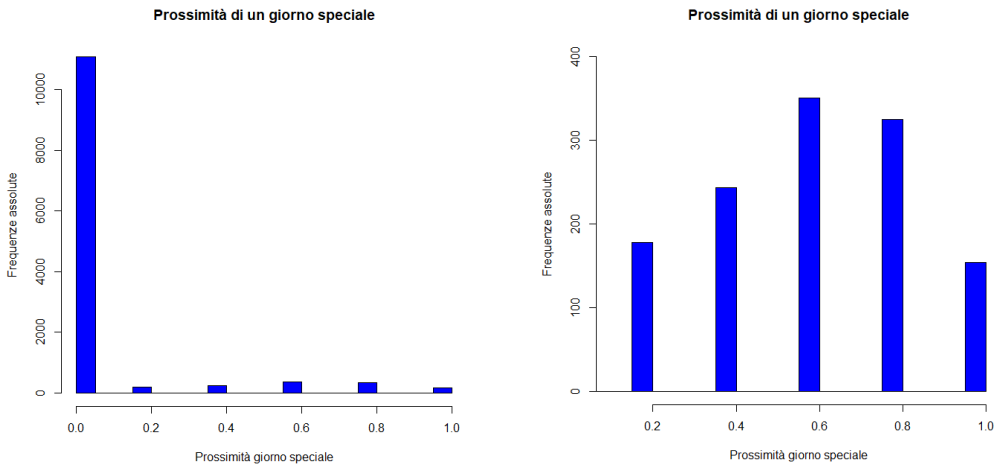

Figura 4 Grafico a barre della variabile Special Day e grafico a barre tagliato per 0 della variabile

	Special Day
Valore Minimo	0
1° Quartile	0
Mediana	0
Media	0.0614
3° Quartile	0
Valore Massimo	1
Deviazione Standard	0.1989

Tabella 4 Indici numerici riassuntivi variabile Special Day

La variabile Special Day è una variabile discreta che rappresenta la vicinanza ad un giorno "speciale" che può essere ritenuto importante dal consumatore per effettuare un acquisto. Come già detto in precedenza, la variabile assume un valore tanto più alto tanto più il "giorno speciale" si avvicina.

Dal grafico in Figura 4, si osserva come la maggior parte delle osservazioni del dataset si concentri su valori nulli, ci si può aspettare ciò in quanto la maggior parte dei giorni dell'anno non sono giorni che presentano festività o ricorrenze particolari, viene inoltre confermato dall'osservazione del terzo quartile in Tabella 4 che permette di stabilire che almeno il 75% delle visite totali sul sito in studio non avviene in giorni particolari. La mediana pari 0 risulta nettamente inferiore alla media pari a 0.0614, insieme all'osservazione grafica della Figura 4, si può osservare la presenza di asimmetria nella variabile. Il valore elevato assunto dalla deviazione standard pari a 0.1989 porta ad evidenziare la presenza di *outlier* che risultano essere 1251.

.

Per completezza, qui di seguito vengono riportati nello stesso ordine di trattazione in precedenza delle variabili, i grafici di tipo *box-plot* disegnati con il *metodo di Tuckey* delle variabili numeriche presenti nel dataset.

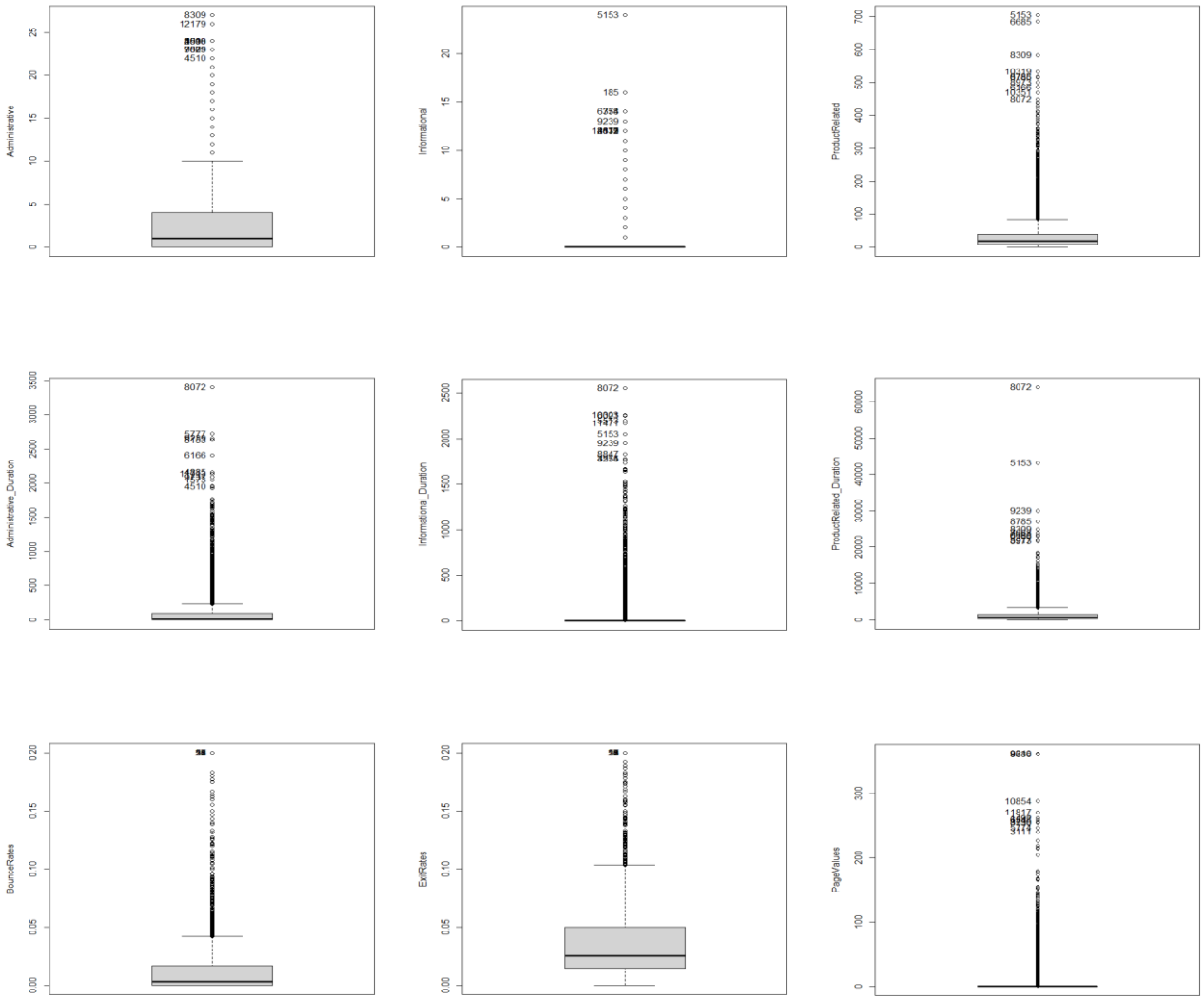

Figura 5 Grafici box-plot di tutte le variabili numeriche studiate

Come viene evidenziato dalla Figura 5, tutti i box-plot delle variabili di contesto numeriche studiate, contengono un elevatissimo numero di outlier (posti in risalto dai "pallini" che si trovano al di sopra del baffo superiore del box-plot), confermando quanto emerso nella precedente fase di studio delle singole variabili. La presenza di valori anomali in maniera così "massiccia", sarà opportunamente trattata nei prossimi capitoli.

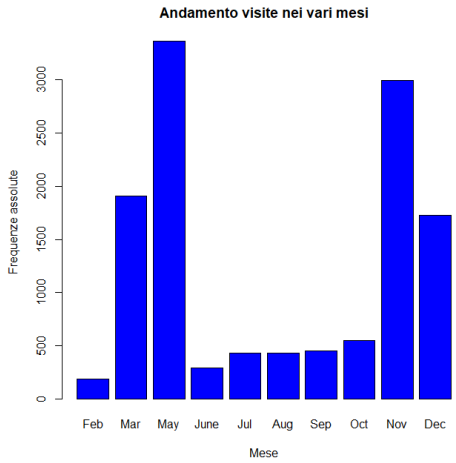

Figura 6 Grafico a barre della variabile Month

	Month
Febbraio	184
Marzo	1907
Maggio	3364
Giugno	288
Luglio	432
Agosto	433
Settembre	448
Ottobre	549
Novembre	2998
Dicembre	1727

Tabella 5 Numero di osservazioni per modalità osservata della variabile Month

La variabile Month è una variabile categoriale, attraverso la Figura 6 è possibile osservare la distribuzione delle sessioni del sito nell'arco dei dieci mesi rilevati nel dataset. Si osservano dei picchi di visualizzazioni nei mesi di Maggio, Novembre seguiti da Marzo e Dicembre, esistono due periodi nel corso dell'anno (Marzo-Maggio e Novembre-Dicembre) in cui si registra un maggior afflusso di visitatori.

Nella Tabella 5 vengono riportate nel dettaglio il conteggio delle sessioni per ogni mese.

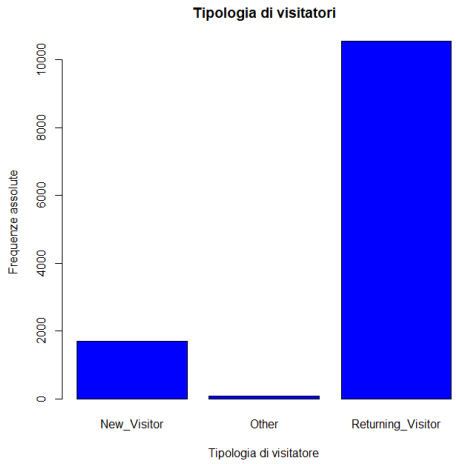

Figura 7 Grafico a barre della variabile Visitor Type

	Visitor Type
Nuovo Visitatore	1694
Altro	85
Visitatore non Nuovo	10551

Tabella 6 Numero di osservazioni per modalità osservata della variabile Visitor Type

La variabile Visitor Type è una variabile categoriale che rappresenta la tipologia di visitatore del sito classificandolo in tre categorie: Nuovo visitatore, Altro e Visitatore non Nuovo.

Da quanto emerge dalla Figura 7, si può notare che la maggior parte dei visitatori del sito rientrano sotto la categoria “Visitatori non Nuovi” mentre i “Visitatori Nuovi” e “Altro” risultano essere una minoranza. Il sito in studio dunque, è prevalentemente frequentato da visitatori “fedeli”.

Nella Tabella 6 è possibile osservare nel dettaglio il numero di visitatori per tipologia.

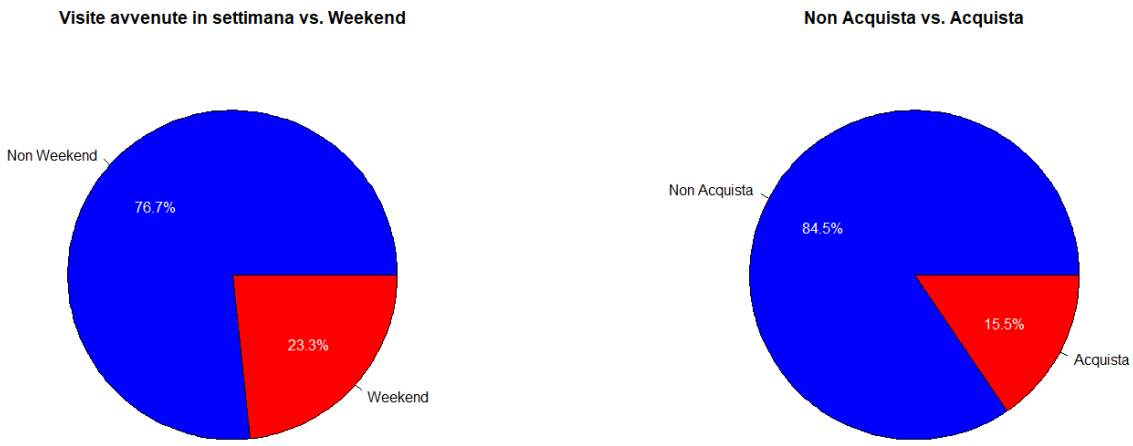

Figura 8 Grafici a torta delle variabili Weekend e Revenue

	Weekend	Revenue
No	9462	10422
Sì	2868	1908

Tabella 7 Numero di osservazioni per modalità osservata delle variabili Weekend e Revenue

Le variabili Weekend e Revenue sono delle variabili categoriali dicotomiche, ovvero possono assumere solo valori pari a 0 o 1.

Variabile Weekend: la variabile assume valore 0 se la sessione sul sito è avvenuta in un giorno compreso tra Lunedì-Venerdì, mentre 1, indica che la sessione è avvenuta nel weekend. Osservando la Figura 8 e la Tabella 7 emerge chiaramente che la maggior parte delle sessioni, pari al 76,7% sul totale, non avviene nel weekend.

Variabile Revenue: rappresenta la variabile risposta del modello di classificazione, indica se l'utente finalizzerà o meno un acquisto sul sito. Assume valore 0 se l'utente non acquista nulla sul sito, 1 viceversa. Per le ragioni già viste in letteratura, è normale aspettarsi uno sbilanciamente tra le classi della variabile risposta in quanto solo una piccola parte dei visitatori di un sito finalizza un acquisto. Osservando la Figura 8 e la Tabella 7, rispetto a quanto visto in [7], il sito presenta un ottimo tasso di conversione, pari al 15,5%, indicatore di un generale apprezzamento da parte dei visitatori dei prodotti in vendita sul sito e del sito stesso.

PREVISIONE DELLE INTENZIONI DI ACQUISTO

In questo capitolo verrà appplicato il modello di regressione logistica standard per prevedere le intenzioni di acquisto di un utente sul sito in studio.

Nella prima parte di questo capitolo verrà spiegato come sarà gestito il dataset ed il relativo sbilanciamento nelle classi della variabile risposta Revenue e la successiva suddivisione in dati di addestramento e dati di test.

Nella seconda parte verranno studiate nel dettaglio le correlazioni tra le varie variabili di contesto raccolte nel dataset e la variabile risposta Revenue.

Per quanto riguarda la terza parte, verrà introdotto il modello di classificazione logistica Logit che verrà costruito utilizzando le variabili ritenute di interesse nella seconda parte, dopodichè saranno commentati i risultati ottenuti.

Gestione dello sbilanciamento delle classi e suddivisione del dataset.

Come già evidenziato nei capitoli dedicati all'analisi esplorativa del dataset, la variabile risposta Revenue presenta uno sbilanciamento molto forte, solo il 15,5% (1908 osservazioni) degli utenti presenta un valore pari a 1 (ovvero che appartiene al gruppo di chi acquista) mentre il restante 84,5% (10422 osservazioni) presenta un valore pari a 0, ovvero non acquista.

Questo sbilanciamento, se non opportunamente gestito, causa seri problemi predittivi al modello di classificazione, falsando in maniera significativa i risultati di predizione, in quanto il modello sarà in grado di riconoscere perfettamente chi non acquista, mentre le pochissime osservazioni di chi acquista, verranno scarsamente riconosciute dal modello che è stato addestrato su dati con la classe 0 prevalente.

Per ovviare al problema dello sbilanciamento, come visto in letteratura in [4], [5], è possibile utilizzare delle apposite tecniche di *oversampling* che generano delle nuove osservazioni della classe minoritaria mantenendo inalterate le caratteristiche del dataset (medie, varianze…), oppure tecniche di *undersampling* che bilanciano le classi senza aggiungere nuove osservazioni rimuovendo le osservazioni della classe maggioritaria in maniera casuale.

In questo studio, si decide di utilizzare la tecnica dell'*undersampling,* in quanto non aggiunge osservazioni "artificiali" a quelle realmente osservate.

Viene quindi generato un nuovo dataset costituito da da 3816 osservazioni casuali, costituito da 1908 osservazioni della classe di chi acquista e 1908 osservazioni di chi non acquista.

Siccome il dataset presenta anche numerosissimi *outlier*, risulterebbe troppo limitante creare un solo dataset attraverso l'*undersampling* con osservazioni raccolte casualmente dal dataset, in quanto si potrebbe generare un dataset che presenta pochissimi o moltissimi outlier, andando a falsare in maniera significativa le capacità descrittive e predittive dei modelli.

Questo problema appena descritto, è risolvibile iterando un determinato numero di volte la generazione del dataset, per ogni iterazione viene addestrato il modello sul dataset appena generato.

Viene scelto un numero di iterazioni pari a 50, dunque verranno generati 50 diversi dataset attraverso l'*undersampling,* di conseguenza verranno addestrati 50 modelli differenti.

I modelli di ML con finalità predittive, come quelle oggetto di studio, non vengono solamente addestrati sui dati disponibili, bensì vengono anche "testati" su osservazioni su cui non sono mai stati addestrati, a tal fine si adopera un approccio basato sulla "*cross-validation*" ossia una suddivisione (*split*) del dataset in due parti: nel caso in studio, viene utilizzato l'80% del dataset generato attraverso l'*undersampling* per addestrare il modello (*training set*), il 20% del dataset invece, come test del modello (*test set*).

Anche lo *split* in *traning set* e *test set* viene iterato 50 volte di pari passo con la generazione del dataset ottenuto mediante l'*undersampling* così da ridurre il rischio di incorrere in dataset poco rappresentativi del fenomeno in studio.

Un'immagine può aiutare a comprendere l'algoritmo implementato:

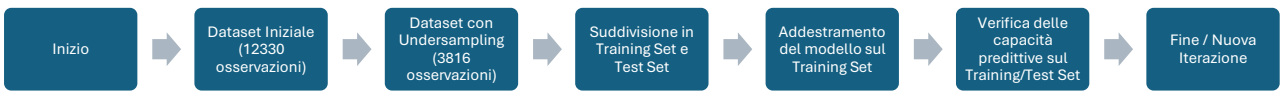

Figura 9 Funzionamento dell'algoritmo implementato per l'undersampling e suddivisione in training/test set

Individuazione dei *driver* che spingono all'acquisto

L'obiettivo di questo studio, è quello di determinare i principali *driver* che spingono gli utenti del sito in studio ad effettuare un acquisto attraverso l'utilizzo di modelli di classificazione, risulta importantissimo, per un corretto funzionamento del modello, un'accurata scelta delle variabili da inserire nel modello.

Per la creazione di modelli efficaci ed efficienti, bisogna rispettare il criterio della "parsimonia", ossia utilizzare il minor numero possibile di variabili esplicative per catturare una relazione con la variabile risposta. Di seguito verranno riportati due casi "limite" applicati al criterio della parsimonia.

Nel primo caso, utilizzando un numero troppo basso di variabili esplicative, il modello diventa incapace di cogliere le relazioni che esistono tra i dati in studio, creando così, una situazione di *underfitting*.

Nel secondo caso invece, l'utilizzo di un numero troppo alto di variabili esplicative, fa si che si verifichi una situazione di *overfitting* sui dati di addestramento, ovvero il modello si adatta perfettamente ai dati osservati in fase di addestramento, ma non è in grado di riconoscere e classificare correttamente osservazioni su cui non è stato addestrato

La variabile risposta dei modelli di classificazione che verranno usati è rappresentata dalla variabile Revenue, che come già detto in precedenza, rappresenta l'acquisto o meno dell'utente al termine della sessione.

Il dataset in studio presenta due macro-tipologie di variabili: le variabili numeriche la cui correlazione con la variabile risposta viene studiata del punto di vista numerico attraverso l'indice R^2 (l'indice assume valori compresi tra 0 ed 1), il quale, permette di calcolare la quantità di informazione che le singole variabili apportano alla variabile risposta, ottenuto mediante la creazione di modelli lineari che hanno come variabile risposta Revenue e come unico regressore la variabile di cui si vuole valutare l'impatto, dal punto di vista grafico invece, tramite dei grafici di tipo *box-plot* tracciati per gruppi che possano mettere in risalto la correlazione con la variabile risposta.

Le correlazioni tra le variabili categoriali e la risposta invece, vengono studiate attraverso il test di indipendenza del *chi-quadrato* e l'indicatore *V di Cramer*.

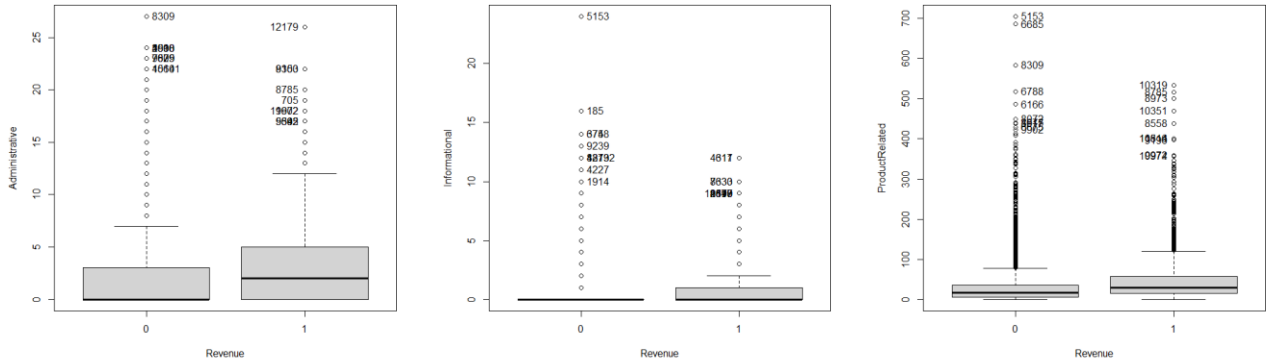

Figura 10 Boxplot tracciati per gruppo Revenue delle variabili Administrative, Informational e Product Related

	Indice di correlazione (R^2)
Administrative	0.0193
Informational	0.0091
Product Related	0.0251

Tabella 8 indici di correlazione lineare di Pearson delle variabili Administrative, Informational e Product Related

Osservando i boxplot della Figura 10, emerge quanto già osservato nell'analisi esplorativa, infatti tutte le variabili hanno una forte presenza di outlier rappresentati dai "puntini" al di sopra del terzo baffo del boxplot.

Entrando più nel dettaglio attraverso l'analisi della Figura 10 e della Tabella 8 si osservano nelle singole variabili le seguenti correlazioni con la variabile Revenue:

Variabile Administrative: presenta un valore dell'indice R^2 pari a 0.0193, indice di una correlazione debole tra il numero di pagine di natura amministrativa visitate e la probabilità che l'utente esegua un acquisto sul sito. I boxplot prsentano molta zona grigia.

Variabile Informational: l'indice R^2 pari a 0.0091 mostra una correlazione molto debole con la probabilità che l'utente esegua un acquisto.

Variabile Product Related: il valore dell'indice R^2 pari a 0.0251 mostra una moderata correlazione con la probabilità che l'utente effettui un acquisto sul sito. Il boxplot presenta comunque una notevole zona grigia.

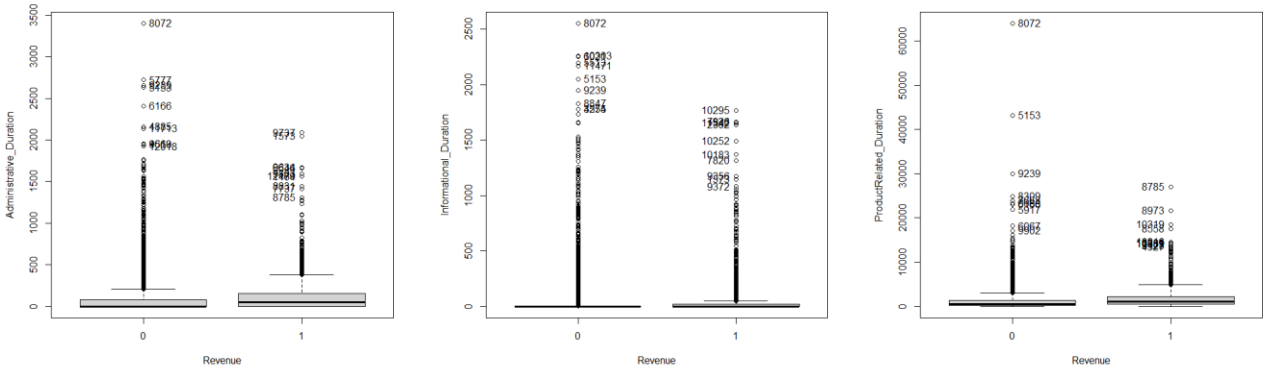

Figura 11 Boxplot tracciati per gruppo Revenue delle variabili Administrative Duration, Informational Duration e Product Related Duration

	Indice di correlazione (R^2)
Administrative Duration	0.0088
Informational Duration	0.0049
Product Related Duration	0.0232

Tabella 9 indici di correlazione lineare di Pearson delle variabili Administrative Duration, Informational Duration e Product Related Duration

Osservando i boxplot della Figura 11, emerge chiaramente la presenza di numerosi outlier in tutte le variabili. Attraverso lo studio della Figura 11 e della Tabella 9 emergono le seguenti considerazioni:

Variabile Administrative Duration: presenta un valore dell'indice R^2 pari a 0.0088 che evidenzia una debolissima correlazione con la probabilità di acquisto. I boxplot presentano molta zona grigia.

Variabile Informational Duration: presenta un piccolissimo valore di R^2 pari a 0.0049 che evidenzia uno scarso legame con la probabilità di acquisto. I boxplot presentano una fortissima zona grigia.

Variabile Product Related Duration: presenta una moderata correlazione con la probabilità di acquisto dimostrata dall'indice R^2 pari a 0.0232. I boxplot presentano una zona grigia marcata.

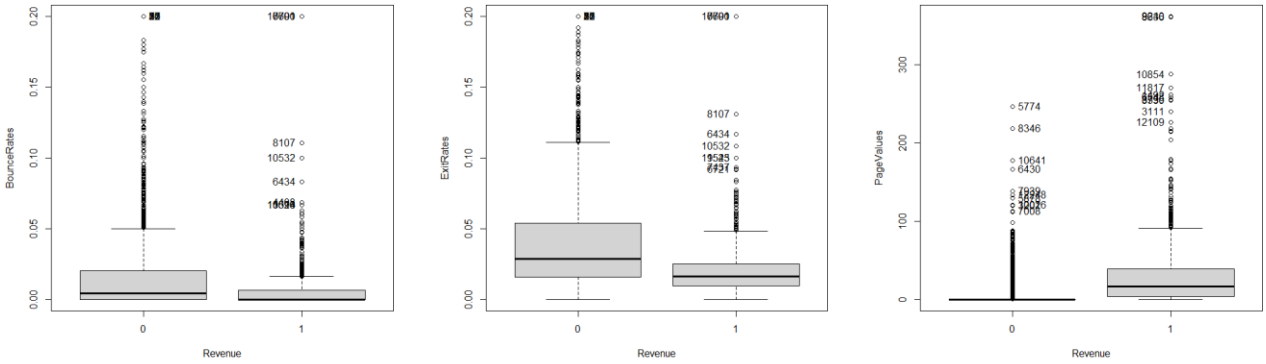

Figura 12 Boxplot tracciati per gruppo Revenue delle variabili Bounce Rates, Exit Rates e Page Values

	Indice di correlazione (R^2)
Bounce Rates	0.0227
Exit Rates	0.0429
Page Values	0.2426

Tabella 10 indici di correlazione lineare di Pearson delle variabili Bounce Rates, Exit Rates e Page Values

Osservando i boxplot della Figura 12, risulta evidente la presenza di outlier, ricordiamo ancora una volta che queste variabili sono frutto dell'estrazione di dati legati al sito attraverso Google Analytics, nel dettaglio, anche grazie agli indici calcolati in Tabella 10 si osserva:

Variabile Bounce Rates: presenta un valore dell'indice R^2 pari a 0.0227 che evidenzia una moderata correlazione con la probabilità di acquisto. I boxplot presentano comunque una zona grigia evidente.

Variabile Exit Rates: il valore dell'indice R^2 è pari a 0.0429 ci permette di concludere che esiste una buona correlazione con la probabilità di acquisto. I boxplot presentano una discreta zona grigia.

Variabile Page Values: il valore dell'indice R^2 pari a 0.2426 mostra una forte correlazione con la probabilità di acquisto. Il boxplot presenta una zona grigia meno estesa rispetto alle altre variabili.

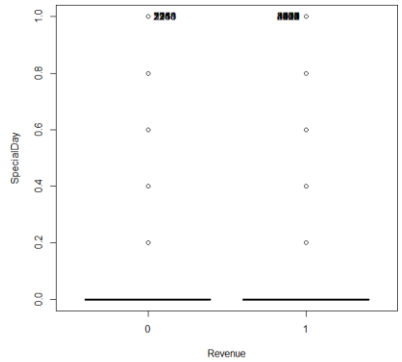

Figura 13 Boxplot tracciato per gruppo Revenue della variabile Special Day

	Indice di correlazione (R^2)
Special Day	0.0068

Tabella 11 indice di correlazione lineare di Pearson della variabile Special Day

Attraverso l'osservazione del boxplot di Figura 13 emerge chiaramente la presenza di numerosi outlier e soprattutto una zona grigia molto forte che rende i boxplot delle due classi praticamente identici e indistinguibili fra loro. L'indice di correlazione R^2 visibile in Tabella 11 mostra una correlazione debolissima con la probabilità di acquisto.

Si realizza una tabella riepilogativa che pone in ordine di importanza le variabili sulla base del valore dell'indice R^2 in relazione alla variabile risposta Revenue:

	Indice di correlazione (R^2)
Page Values	0.2426
Exit Rates	0.0429
Product Related	0.0251
Product Related Duration	0.0232
Bounce Rates	0.0227
Administrative	0.0193
Informational	0.0091
Administrative Duration	0.0088
Special Day	0.0068
Informational Duration	0.0049

Tabella 12 Classifica delle correlazioni delle variabili numeriche con la variabile risposta Revenue

Dall'osservazione della Tabella 12 emerge come le principali variabili da prendere in considerazione per la costruzione di un modello

Le variabili numeriche "candidate" per il modello di classificazione sulla base dell'indice R^2 sono le seguenti: Page Values, Exit Rates, Product Related, Product Related Duration e Bounce Rates. Le variabili con un valore dell'indice inferiore a 0.02 vengno trascurate per via dello scarso apporto di informazione che forniscono alla variabile risposta Revenue.

Si nota, che le variabili Exit Rates e Bounce Rates sono fortemente correlate tra loro, infatti entrambe sono metriche che ci permettono di ricavare il tasso medio di uscita delle pagine visitate dall'utente, dunque, viene studiata la correlazione tra le due variabili ed emerge un indice R^2 pari a 0.8336 che permette di concludere che Bounce Rates "spiega" l'83% circa della variabilità di Exit Rates, dunque è opportuno rimuovere la variabile Bounce Rates per via della ridondanza di informazione apportata.

Discorso analogo vale per le variabili Product Related e Product Related Duration, che attraverso l'osservazione dell'indice R^2 pari a 0.7412 si può comprendere che la variabile Product Related Duration apporti moltissima informazione alla variabile Product Related. Appare una conclusione ragionevole, in quanto risulta abbastanza scontato, che, all'aumentare delle pagine relative ai prodotti visitate, aumenti di pari passo il tempo che l'utente trascorre sulle pagine legate ai prodotti. Si sceglie anche in questo caso di rimuovere la variabile Product Related Duration.

Per quanto riguarda le variabili categoriali/dicotomiche, come già accennato nella parte introduttiva, si studia la correlazione attraverso il test del chi quadrato (χ^2), che, permette di verificare la presenza di dipendenza tra la variabile in studio e la variabile risposta (Revenue) ed eventualmente accettare o rifiutare l'ipotesi di assenza di dipendenza statistica tra le variabili. Per avere invece un'idea della "forza" della correlazione viene utilizzata la V di Cramer costituita da un indice numerico che può assumere valori tra 0 ed 1.

Di seguito si riporta una tabella riepilogativa dei risultati ottenuti:

	Chi-quadrato (χ^2)	**P-Value**	**V Cramer**
Month	384.93	< 0.05	0.177
Visitor Type	135.25	< 0.05	0.105
Weekend	10.391	< 0.05	0.029

Tabella 13 Chi-quadrato e V Cramer delle variabili categoriali in relazione a Revenue

Dalla Tabella 13, emerge che per tutte le variabili in studio, si può rifiutare l'ipotesi di assenza di dipendenza statistica con la variabile risposta Revenue grazie ai valori di significatività del *P-Value* ottenuti inferiori alla soglia del 5%, rifiutando quindi, l'ipotesi nulla. La V di Cramer invece permette di capire la forza della correlazione che le variabili hanno con Revenue; sempre dalla Tabella 13, emerge che la variabile Month presenta un valore pari a 0.177 indice di una modesta correlazione con la risposta, dunque è opportuno inserire questa variabile nel modello. Si trascurano dal modello le variabili Visitor Type e Weekend in quanto presentano un valore della V di Cramer troppo basso.

In conclusione, sulla base dei risultati ottenuti in questo capitolo, si decide di utilizzare le seguenti variabili per i modelli di classificazione: Page Values, Exit Rates, Product Related e Month.

Descrizione e applicazione del modello Logit

Il modello Logit è un modello di classificazione che permette di stimare la probabilità che la variabile risposta (Revenue nel caso in studio) assuma valore pari ad 1, ossia che l'utente acquisti sul sito.

Il modello Logit può essere così descritto:

$$P(Y = 1|X_1, X_2, \dots, X_q) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_qX_q}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_qX_q}}$$

Dove:

q rappresenta il numero di predittori.

$\boldsymbol{P(Y = 1|X_1, X_2, \dots, X_q)}$ è la probabilità che la variabile risposta Y assuma il valore 1 (ossia che l'utente acquisti sul sito) date le variabili esplicative $X_1, X_2, \dots, X_q$.

$\boldsymbol{\beta_0}$ rappresenta la probabilità di base quando tutte le variabili esplicative $X_1, X_2, \dots, X_q$ sono uguali a zero, indica l'influenza in termini di probabilità che non è spiegata dalle variabili indipendenti sulla variabile risposta Y.

$\boldsymbol{\beta_1, \beta_2, \dots, \beta_q}$ sono i coefficienti associati alla variabili esplicative $X_1, X_2, \dots, X_q$, rappresentano l'effetto di ciascuna X_q sulla probabilità che l'evento accada.

Il modello Logit, può essere riscritto per pemettere la linearizzazione della relazione tra i coefficienti delle variabili esplicative e la risposta, ciò può essere descritto come:

$$\log\left(\frac{P(Y = 1|X_1, X_2, \dots, X_q)}{1 - P(Y = 1|X_1, X_2, \dots, X_q)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_qX_q$$

Dove:

$\boldsymbol{\log\left(\frac{P(Y = 1|X_1, X_2, \dots, X_q)}{1-P(Y = 1|X_1, X_2, \dots, X_q)}\right)}$ rappresenta il logaritmo del cosiddetto "*odds ratio"* ossia il logaritmo della probabilità che accada un determinato evento sulla probabilità che non accada.

A questo punto, risulta di fondamentale importanza introdurre il concetto di funzione di perdita, il quale sarà richiamato anche nel capitolo successivo per poter mostrare le "debolezze" di un approccio tradizionale al caso in studio.

La funzione di perdita è un concetto fondamentale in ambito *Machine Learning*, utilizzato per valutare quanto un modello dal punto di vista predittivo operi correttamente rispetto ai dati osservati.

In sostanza, essa misura la discrepanza tra le previsioni generate dal modello e i valori reali osservati, consentendo di quantificare l'errore delle capacità predittive del modello.

Il modello Logit, appena descritto, risulta essere molto "influenzato" rispetto agli *outlier*, in quanto si basa sulla cosiddetta funzione di perdita di log-verosomiglianza, che, nel caso di una variabile risposta dicotomica (come nel caso in studio), può essere esplicitata con la seguente equazione[14]:

$$\log L = \sum_{i=1}^{n} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$$

Dove:

$\mathbf{\log L}$ rappresenta la funzione di perdita della log-verosomiglianza

$\boldsymbol{n}$ è il numero di osservazioni del dataset

$\boldsymbol{y_i}$ è il valore osservato della variabile risposta per l'osservazione i-esima

$\boldsymbol{\hat{p}_i}$ rappresenta la probabilità che la variabile risposta per l'osservazione i-esima registri un valore pari a 1, viene stimata mediante il modello Logit.

La funzione di perdita può essere così interpretata:

Se $\boldsymbol{y_i = 1}$: la somma si riduce a $log\,(\hat{p}_i)$, dunque se il modello stima una probabilità molto alta (quindi vicina ad 1), la funzione di perdita registrerà una bassissimo valore prossimo allo zero; qualora la probabilità predetta fosse bassa (quindi prossima a zero), il termine $log\,(\hat{p}_i)$ assume un valore negativo molto forte, aumentando significativamente il valore della funzione di perdita.

Se $\boldsymbol{y_i = 0}$: la somma in questo caso, si riduce al termine $log\,(1 - \hat{p}_i)$, se il modello stima correttamente una probabilità bassa (quindi vicina a zero), il termine risulta essere vicino a zero contribuendo poco alla perdita complessiva; se la probabilità predetta fosse alta, il termine $log\,(1 - \hat{p}_i)$ assume un importante valore negativo, aumentando di conseguenza la perdita totale.

La funzione di perdita complessiva, dunque, risulta essere composta dalla somma di tutti i termini per tutte le osservazioni del dataset, l'obiettivo come già detto in precedenza, è quello di minimizzare questa somma.

La funzione di log-verosimiglianza è particolarmente sensibile agli *outlier*. Infatti, la presenza di dati anomali può compromettere la capacità del modello di stimare correttamente le probabilità. Se, a causa degli *outlier*, la probabilità predetta $\widehat{\boldsymbol{p}_i}$ differisse notevolmente dal valore reale $\boldsymbol{y_i}$, la funzione di perdita assegna un valore molto negativo a quella previsione, influenzando significativamente la perdita totale del modello. Anche un piccolo numero di *outlier* può avere un impatto importante, poiché le penalizzazioni per previsioni estremamente errate sono molto severe.

Si procede ora all'applicazione del modello Logit standard al caso in studio.

Applicando il modello Logit sul dataset, utilizzando le quattro variabili esplicative scelte nella parte relativa allo studio della correlazione, si ottengono le seguenti stime dei valori medi dei coefficienti, in quanto, la stima del modello viene iterata 50 volte seguendo l'algoritmo descritto nel paragrafo precedente; non vengono riportate statistiche inferenziali essendo una media.

Nome del coefficiente	**Stima del coefficiente medio**
$\hat{\beta}0$ (Intercetta)	-0.4675
$\hat{\beta}1$ (Exit Rates)	-14.7911
$\hat{\beta}2$ (Page Values)	0.1023
$\hat{\beta}3$ $(Dicembre)$	-0.6197
$\hat{\beta}4$ $(Febbraio)$	-2.4976
$\hat{\beta}5$ $(Luglio)$	0.1157
$\hat{\beta}6$ $(Giugno)$	-0.5351
$\hat{\beta}7$ $(Marzo)$	-0.6023
$\hat{\beta}8$ $(Maggio)$	-0.8288
$\hat{\beta}9$ $(Novembre)$	0.5489
$\hat{\beta}10$ $(Ottobre)$	-0.1339
$\hat{\beta}11$ $(Settembre)$	0.1564
$\hat{\beta}12$ $(Product\ Related)$	0.0052

Tabella 14 Stima dei coefficienti medi del modello Logit

Osservando la Tabella 14, è possibile osservare le stime medie dei coefficienti delle singole variabili esplicative utilizzate; si noterà che nonostante le variabili scelte per il modello siano quattro, i coefficienti medi stimati (senza contare l'intercetta) risultano essere dodici, in quanto, la variabile Month è una variabile categoriale che rappresenta i 10 mesi dell'anno rilevati nel dataset; vengono quindi create nove variabili *dummy* per ogni mese in studio, ogni mese avrà dunque un suo coefficiente ad eccezione del mese di Agosto che risulta essere il mese base del modello, il cui coefficiente risulta essere implicito nel modello.

Per semplificare la lettura della Tabella 14 viene fornito un breve esempio di come un coefficiente impatti sulla probabilità di acquisto. Ad esempio, se il valore di uscita medio (Exit Rates) delle pagine visitate da un utente aumenta di 0.10 (ossia del 10%), a parità di tutto il resto, il logaritmo dell'*odds ratio* della probabilità di acquisto diminuisce in media di 1.47911 (0.1*-14.7911); risultando in una diminuizione in media del logaritmo dell'*odds ratio* della probabilità di acquisto se il valore medio di uscita delle pagine visitate da un utente aumenta del 10%.

Sotto il punto di vista delle capacità predittive, il modello viene valutato attraverso la matrice di confusione, la quale permette, di valutare la capacità del modello di riconoscere i casi negativi (chi non acquista) e i casi positivi (chi acquista).

La matrice di confusione può essere schematizzata come segue:

	Previsione Positiva (Acquista)	**Previsione Negativa (Non Acquista)**
Reale Positivo (Acquista)	Veri Positivi	Falsi Negativi
Reale Negativo (Non Acquista)	Falsi Positivi	Veri Negativi

Tabella 15 Esempio di matrice di confusione

La Tabella 15 permette di valutare la capacità del modello di identificare e classificare correttamente l'intenzione di acquisto dell'utente sul sito.

Con finalità diagnostiche, vengono utilizzati due indicatori che possono essere facilmente ricavati dalla matrice di confusione e sono:

Sensitività: permette di misurare la proporzione di veri positivi identificati correttamente dal modello rispetto al totale dei positivi reali. Tanto più il valore si avvicina ad 1, tanto migliori saranno le capacità del modello di identificare le osservazioni positive. Basandosi sulla Tabella 15 risulta:

$$Sensitività = \frac{Veri\ Positivi}{Veri\ Positivi + Falsi\ Negativi}$$

Accuratezza: permette di misurare la proporzione del totale delle previsioni corrette del modello (quindi sia positive che negative) rispetto al totale delle osservazioni. Sulla base della Tabella 15 risulta:

$$Accuratezza = \frac{Veri\ Positivi + Veri\ Negativi}{Veri\ Positivi + Veri\ Negativi + Falsi\ Positivi + Falsi\ Negativi}$$

Il modello Logit ottiene le seguenti performance:

	Training Set	**Test Set**
Sensitività Media	0.8517	0.8452
Accuratezza Media	0.8105	0.8060

Tabella 16 Valori medi di sensitività e accuratezza del modello Logit

Osservando la Tabella 16, è possibile valutare le performance del modello Logit; viene osservato un valore pari a 0.8517 di sensitività medio sul *training set,* che permette di concludere che il modello riesce a riconoscere l'85% circa delle persone che effetivamente acquistano sul sito; per quanto riguarda il riconoscimento medio di chi acquista al di fuori del set di addestramento, si registra un valore pari a 0.8452, leggermente più basso rispetto al valore ottenuto in fase di addestramento. Sotto il punto di vista dell'accuratezza media, si registra un valore di 0.8105 sul *training set* che permette di concludere che il modello, in fase di addestramento, è in grado di classificare correttamente l'81% circa delle osservazioni sia positive che negative, sul totale delle osservazioni; sul *test set* questo valore invece, scende a 0.8060.

Il modello Logit in conclusione, ottiene dei buoni risultati di classificazione delle intenzioni di acquisto sia sul *training* che sul *test set.* Nel successivo capitolo verrà spiegato e applicato un metodo robusto al modello appena utilizzato e saranno valutate le sue *perfomance*.

INTRODUZIONE DELLA TECNICA ROBUSTA

Nel corso di questo studio, l'analisi esplorativa dei dati ha rivelato in modo costante e evidente la presenza di valori anomali, noti anche come *outlier*, sia tramite l'osservazione visiva dei grafici *box-plot*, sia attraverso lo studio delle correlazioni tra le variabili. Questi valori anomali, se non trattati adeguatamente, possono influenzare in modo significativo le stime dei modelli statistici, generando distorsioni che compromettono l'affidabilità dei risultati.

In particolare, i modelli di regressione tradizionali, come quelli basati sul metodo della log-verosomiglianza trattato in precedenza, sono estremamente sensibili agli *outlier*.

Infatti, un singolo *outlier* può influire in modo sproporzionato su questo processo, trascinando la stima del modello verso valori che non riflettono accuratamente il comportamento generale dei dati. Il risultato può essere una stima distorta dei coefficienti e una rappresentazione imprecisa del fenomeno in studio.

Questo tipo di errore porta, a sua volta, a previsioni fuorvianti, che possono avere conseguenze significative quando i modelli vengono utilizzati per prendere decisioni operative o strategiche.

L'introduzione di tecniche robuste offre una soluzione efficace a questo problema.

Rispetto ai metodi tradizionali, le tecniche robuste sono progettate per attenuare l'influenza degli *outlier*, proteggendo il modello dalla distorsione delle stime.

In pratica, le tecniche robuste utilizzano funzioni di perdita o algoritmi che assegnano un peso minore agli *outlier*, riducendone così l'effetto negativo. Questo approccio consente di ottenere stime più affidabili e rappresentative dell'intera popolazione, fornendo modelli che catturano meglio le relazioni sottostanti tra le variabili.

Oltre alla gestione degli *outlier*, un ulteriore vantaggio delle tecniche robuste è la loro capacità di ridurre l'impatto degli errori di specificazione. Tali errori si verificano quando il modello omette variabili importanti o include variabili irrilevanti. In un contesto tradizionale, la mancanza di una variabile cruciale può portare a una distorsione delle stime, poiché il modello non è in grado di catturare pienamente il fenomeno in studio. Le tecniche robuste, d'altro canto, tendono a essere meno sensibili a questo tipo di errore, fornendo stime più accurate anche in presenza di specificazioni imperfette.

Un altro aspetto fondamentale dell'uso di tecniche robuste riguarda la stabilità delle stime nel tempo, in molti contesti applicativi, inclusi quelli trattati in questo studio, i dati raccolti provengono da ambienti soggetti a cambiamenti dinamici e imprevedibili.

I modelli tradizionali, fortemente influenzati dagli *outlier* e da eventuali errori di specificazione, possono risultare particolarmente instabili quando la struttura dei dati cambia nel tempo. Di contro, i modelli robusti mostrano una maggiore resilienza: le loro stime dei coefficienti sono meno sensibili alle variazioni nei dati, consentendo di mantenere una maggiore coerenza nelle previsioni e nell'interpretazione dei risultati.

In conclusione, l'adozione di un approccio robusto si rivela essenziale in studi caratterizzati dalla presenza di *outlier* o da un elevato grado di complessità nei dati. Non solo consente di migliorare l'accuratezza delle stime, ma garantisce anche una maggiore stabilità e affidabilità nel tempo, qualità imprescindibili per una corretta comprensione e modellizzazione del fenomeno in esame.

Irrobustimento del modello logistico

Prima di entrare nei dettagli tecnici del modello robusto, è importante richiamare il concetto di funzione di perdita, già introdotto in precedenza. Questa funzione misura l'efficacia del modello dal punto di vista predittivo, con l'obiettivo di minimizzarla per ottenere i parametri che meglio si adattano ai dati in esame.

Per una gestione ottimale dei valori anomali, in questo studio viene utilizzato l'approccio robusto per gli *outlier* proposto in [15].

Nello studio appena citato, viene evidenziato come sia possibile rendere robusta la stima dei parametri in presenza di dati fortemente asimmetrici e contententi *outlier* che possono rendere inaffidabile l'addestramento del modello.

Nel caso binomiale, viene proposta una funzione di perdita basata sulla *quasi-verosomiglianza di Mallows*, che permette una gestione migliore degli *outlier* estremi.

Prima di spiegare l'equazione di stima però, può essere opportuno introdurre i concetti proposti sempre in [15], legati ai *residui di Pearson* e alla *funzione di Huber*.

Per quanto riguarda i *residui di Pearson*, possono essere definiti come una misura delle differenza tra i valori osservati e attesi in un modello, vengono utilizzati per valutare quanto correttamente un modello riesca ad adattarsi ai dati. Possono essere definiti come segue:

$$r_i = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}$$

Dove:

$\boldsymbol{r_i}$ rappresenta il residuo di Pearson per l'osservazione i-esima.

$\boldsymbol{y_i}$ rappresenta il valore osservato dell'osservazione i-esima

μ_i è il valore atteso per l'osservazione i-esima

$\boldsymbol{V}(\mu_i)$ è la varianza del valore atteso μ_i

Il numeratore dell'equazione rappresenta lo scostamento tra il valore osservato ed atteso, mentre il denominatore ha funzione di normalizzare lo scostamento, rendendo il tutto una misura standardizzata.

La funzione di Huber $\psi_c(r)$, rappresenta la funzione di perdita che viene introdotta nel modello Logit per ridurne la sensibilità agli *outlier*.

Essa, può essere definita nel seguente modo:

$$\psi_c(r) = \begin{cases} r & se|r| \leq c \\ c\ sign(r) & se\ |r| > c \end{cases}$$

Dove:

$\boldsymbol{r}$ rappresenta il residuo di Pearson.

$\boldsymbol{c}$ è una costante che assume segno positivo e determina il "punto di passaggio" della funzione da quadratica a lineare.

$\boldsymbol{sign(r)}$ è la funzione segno che restituisce 1 se $r > 0$, -1 se $r < 0$, 0 se $r = 0$.

Dunque, è possibile identificare due situazioni distinte osservando la funzione di Huber:

Per casi in cui $|r| \leq c$: la funzione di Huber restituisce un valore pari al residuo r, il che permette di concludere che le osservazioni che registrano un valore dei residui molto basso, vengono trattate come in una normalissima funzione di perdita quadratica.

In casi in cui $|r| > c$: la funzione di Huber diventa lineare rispetto al residuo r, rendendo meno influenti gli outlier che registrano valori molto alti.

È possibile ora introdurre la funzione di stima robusta, sarà utilizzata nel modello Logit robusto; sulla base di quanto emerge in [15]:

$$\sum_{i=1}^{n}\left[\psi_c(r_i)\,\omega(x_i)\,\frac{1}{\sqrt{V(\mu_i)}}\,\mu_i' - a(\beta)\right] = 0$$

Dove:

$\boldsymbol{\psi_c(r_i)}$ rappresenta la funzione di Huber applicata al residuo di Pearson i-esimo

$\boldsymbol{\omega(x_i)}$ rappresenta il peso specifico per l'osservazione i-esima, attraverso la tecnica robusta è possibile assegnare ad ogni osservazione un peso, che permette di ridurre l'impatto sulla stima di eventuali osservazioni che registrano un valore anomalo.

$\frac{\mathbf{1}}{\sqrt{\boldsymbol{V(\mu_i)}}}$ è l'inverso della radice quadrata della varianza dei valori attesi μ_i

μ_i' è la derivata della media rispetto ai parametri del modello.

$\boldsymbol{a(\beta)}$ è un termine di "aggiustamento" del modello, costituito dalla media pesata delle derivate

Utilizzo del modello robusto su dati simulati

In questo paragrafo, viene proposta una simulazione creata per dimostrare la bontà dell'approccio robusto utilizzando il modello Logit robusto descritto nel paragrafo precedente, rispetto all'utilizzo di un modello Logit standard.

Per fare ciò, è stato generato un dataset apposito costituito da 200 osservazioni totali, 100 appartenenti al gruppo 1, 100 appartenenti al gruppo 0.

Le variabili, sono state generate in maniera casuale seguendo una distribuzione normale, avendo cura di modificare la media delle singole variabili in base al gruppo di appartenenza dell'osservazione, così da costruire delle correlazioni tra il valore assunto dalle variabili e il gruppo di appartenenza.

Lo scopo di questa simulazione, è quello di confrontare il modello Logit Standard con il modello Logit robusto, sia nel caso in cui il dataset non presenti *outlier* rilevanti, sia nel caso in cui il dataset presenti numerosi *outlier* che potrebbero impattare le stime del modello.

Le prestazioni dei modelli, verranno sempre valutate attraverso i parametri di accuratezza e sensitività ricavati dalla matrice di confusione.

Esempio su dati simulati

In questo paragrafo, si procede ad una rapida descrizione attraverso dei grafici *boxplot* tracciati per gruppo delle singole variabili, per poter fornire un'idea di come sia strutturato il dataset.

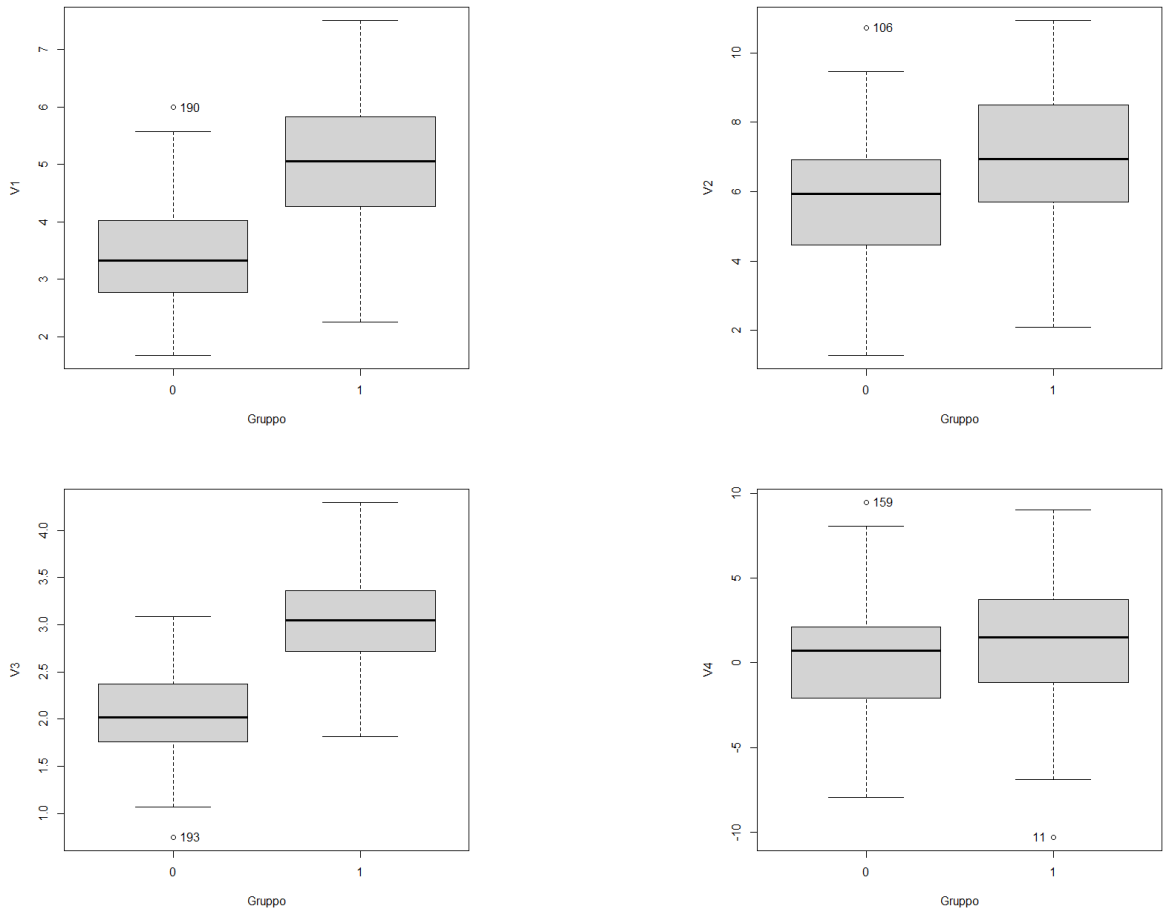

Figura 14 Grafici boxplot tracciati per gruppi delle variabili V1, V2, V3, V4

Attraverso la Figura 14, è possibile osservare le correlazioni che esistono tra i valori assunti dalle variabili e l'appartenenza ai due gruppi.

Nel caso di V1 si nota una chiara correlazione positiva tra il valore assunto dalla variabile e l'appartenza al gruppo 1, con una zona grigia quasi inesistente. Per quanto riguarda V2, si registra una correlazione positiva anche se è presenta una certa zona grigia. Nel caso di V3, si nota una chiarissima correlazione positiva con una zona grigia inesistente. Infine V4 presenta una debole correlazione positiva ed un'importante zona grigia.

In questo caso, le variabili verranno tutte inserite nel modello, in quanto lo scopo è quello di dimostrare il corretto funzionamento dell'approccio robusto, a differenza del dataset in studio in cui l'obbiettivo è quello di ottenere un modello predittivo il più "parsimonioso" possibile.

Si ripete lo stesso tipo di analisi, attraverso i grafici a boxplot tracciati per gruppi con l'inserimento, nello stesso dataset, di diversi *outlier*, generati casualmente con distribuzione normale.

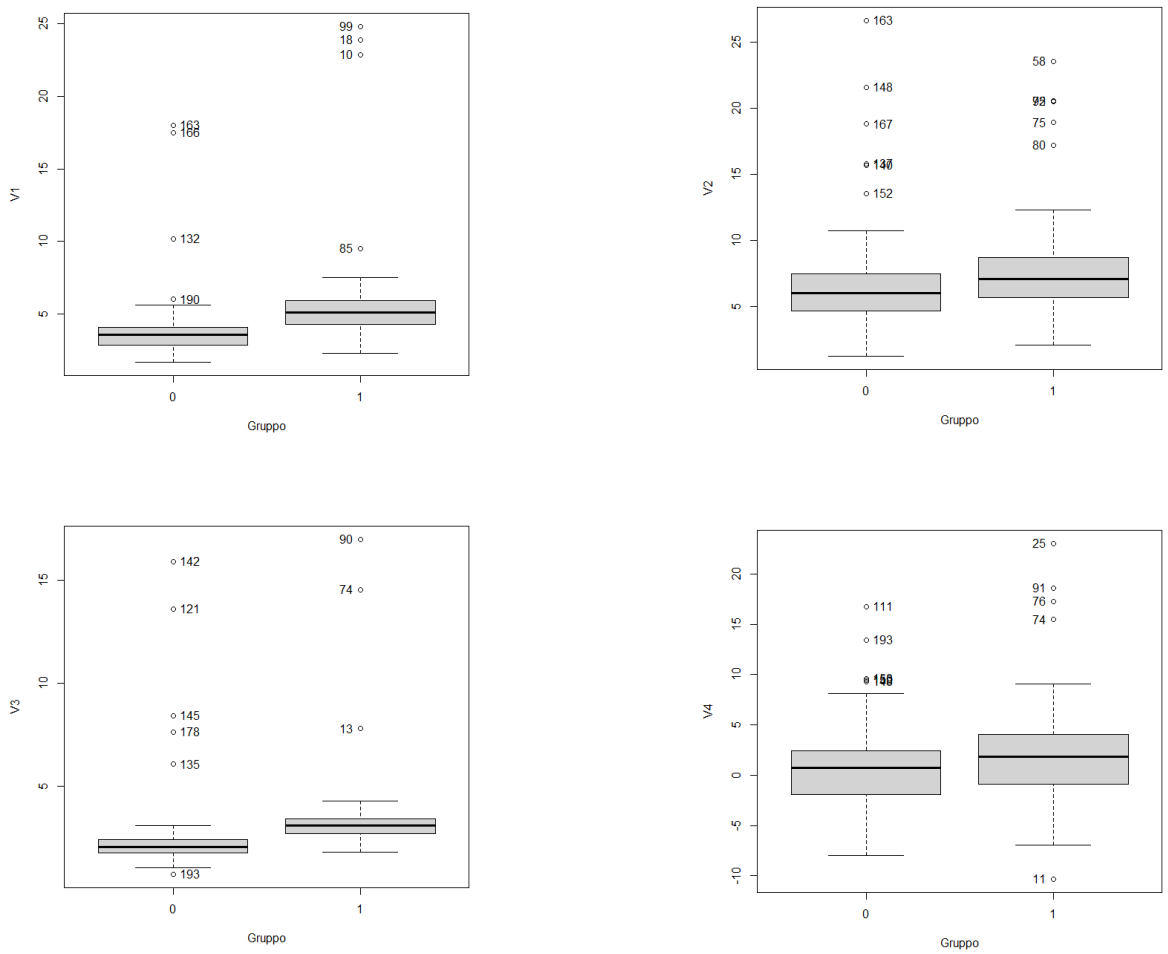

Figura 15 Grafici boxplot tracciati per gruppi delle variabili V1, V2, V3, V4 con inserimento di outlier

Osservando la Figura 15, è possibile vedere come i boxplot in ogni variabile risultino più "schiacciati" rispetto alla Figura 14, questo è dovuto alla presenza di *outlier*.

In un caso come quello riportato in Figura 15, un approccio basato sul modello Logit standard senza nessuna tecnica robusta, potrebbe portare a significativi cali di *performance* del modello

La valutazione delle performance dei due modelli nei due casi separati sarà oggetto di valutazione del prossimo paragrafo.

Performance dei modelli nella simulazione

Dopo aver descritto in maniera sintetica nei paragrafi precedenti le logiche dietro alla costruzione del dataset utilizzato in questa simulazione, si procede ora alla creazione dei modelli (e successiva valutazione) di classificazione: il modello Logit standard (già applicato sul dataset in studio e commentanto in maniera approfondita ad inizio capitolo) ed il modello Logit robusto.

Nel caso della simulazione, non viene eseguito lo *split* in *train* e *test* set, in quanto ai fini della simulazione, non risulta utile per comprendere come i modelli si comportino al di fuori delle osservazioni su cui vengono addestrati; dunque, si utilizza l'intero dataset per l'addestramento del modello.

I due modelli registrano le seguenti *perfomance* sul dataset utilizzato nella simulazione senza l'inserimento degli *outlier*:

	Modello Logit Standard	**Modello Logit Robusto**
Sensitività	0.9223	0.9314
Accuratezza	0.935	0.94

Tabella 17 Valori di sensitività e accuratezza del modello standard e robusto sul dataset senza outlier

Dalla Tabella 17, è possibile osservare che entrambi i modelli ottengono dei risultati molto simili sul dataset registrano delle prestazioni di sensitività e accuratezza molto elevate.

Anche se non sono presenti outlier impattanti, si nota come il modello robusto ottenga comunque delle prestazioni sensibilmente migliori rispetto al modello Logit standard.

Aggiungendo invece, 40 outlier distribuiti normalmente al dataset della simulazione, si ottengono i seguenti risultati:

	Modello Logit Standard	**Modello Logit Robusto**
Sensitività	0.8229	0.8704
Accuratezza	0.81	0.9

Tabella 18 Valori di sensitività e accuratezza del modello standard e robusto sul dataset con outlier

Osservando i risultati presenti nella Tabella 18, risulta evidente la differenza prestazionale tra il modello standard e quello robusto in presenza di outlier.

Il modello Logit standard, perde circa il 10% sotto il punto di vista della sensitività e il 12% circa in termini di accuratezza (rispetto al caso senza *outlier*), registrando così, un significativo deterioramento delle *performance* causato dalla presenza di valori anomali.

Risultano evidenti invece, i benefici apportati dal modello robusto in termini di prestazioni, perdendo solo circa il 5% in termini di sensitività e il 4% circa in termini di accuratezza (rispetto al caso senza *outlier* riportato in Tabella 17).

Sempre dall'osservazione della Tabella 18, è possibile riscontrare come il modello robusto riconosca in media il 5% circa in più di casi appartenenti al gruppo 1 (sensitività) rispetto al modello standard ed ha un'accuratezza in media superiore del 9% rispetto al modello standard, riconoscendone correttamente dunque, il 9% in più.

Dai risultati, risulta chiaro che il modello Logit robusto potrebbe apportare dei miglioramenti alle prestazioni ottenute dal modello Logit standard sul dataset in studio.

Nel paragrafo successivo, verrà dunque implementato il modello robusto.

Applicazione del modello robusto al caso studio

In questo paragrafo verrà applicato al dataset in studio il modello Logit robusto, per poter beneficiare di un incremento nelle prestazioni a fini predittivi.

Come nel caso del modello Logit standard, il modello sarà sempre addestrato sulle quattro variabili scelte in precedenza nella fase di studio della correlazione con la risposta (Page Values, Exit Rates, Product Related e Month).

Anche in questo caso, lo *split* in *training set* e *test set* del dataset ribilanciato (mediante *undersampling*) verrà iterato 50 volte; di conseguenza anche l'addestramento del modello e successiva verifica sul *test set* sarà iterata 50 volte.

Vengono qui di seguito riportati i coefficienti medi del modello robusto:

Nome del coefficiente	**Stima del coefficiente**
$\hat{\beta}0$ (Intercetta)	-0.8672
$\hat{\beta}1$ (Exit Rates)	-11.2286
$\hat{\beta}2$ (Page Values)	0.3838
$\hat{\beta}3$ $(Dicembre)$	-0.5731
$\hat{\beta}4$ $(Febbraio)$	-3.0226
$\hat{\beta}5$ $(Luglio)$	0.2593
$\hat{\beta}6$ $(Giugno)$	-0.4585
$\hat{\beta}7$ $(Marzo)$	-1.0228
$\hat{\beta}8$ $(Maggio)$	-1.7097
$\hat{\beta}9$ $(Novembre)$	0.6919
$\hat{\beta}10$ $(Ottobre)$	-0.3567
$\hat{\beta}11$ $(Settembre)$	0.2847
$\hat{\beta}12$ $(Product\ Related)$	0.0050

Tabella 19 Stima dei coefficienti medi del modello Logit robusto

Dalla Tabella 19 si può notare come i coefficienti medi stimati differiscano sensibilmente dai coefficienti medi stimati dal modello Logit standard in Tabella 14, questo accade, per via dei pesi utilizzati e della differente funzione di perdita adoperata dal modello robusto per ridurre l'impatto degli *outlier*.

Il modello Logit robusto, ottiene le seguenti prestazioni:

	Training Set	Test Set
Sensitività Media	0.8746	0.8710
Accuratezza Media	0.8433	0.8396

Tabella 20 Valori medi di sensitività e accuratezza del modello Logit robusto

Ossevando la Tabella 20, è possibile comprendere come il modello Logit robusto "lavori" sia sui dati di addestramento che su quelli di verifica.

Il modello, registra ottimi valori di sensitività sia in fase di addestramento che in fase di verifica, sostanzialmente presentando una differenza minima tra i due valori, il che permette di concludere che il modello robusto riconosce correttamente circa l'87% delle osservazioni positive (ovvero chi acquista sul sito) sia in fase di addestramento che in fase di verifica.

Sotto il piano dell'accuratezza media invece, il modello registra valori prossimi all'84% circa sia in fase di addestramento che in fase di verifica, indice del fatto che il modello abbia ottime capacità di riconoscimento delle osservazioni positive e negative sia nel campione che fuori campione.

Vengono riepilogate la performance dei due modelli qui di seguito:

	Training Set (Modello Logit standard)	Training Set (Modello Logit Robusto)	Test Set (Modello Logit standard)	Test Set (Modello Logit robusto)
Sensitività Media	0.8517	0.8746	0.8452	0.8710
Accuratezza Media	0.8105	0.8433	0.8060	0.8396

Tabella 21 Confronto tra le performance ottenute sul training/test set dal modello Logit standard e Logit robusto

Dalla Tabella 21, è possibile visionare i risultati ottenuti dal modello Logit standard e Logit robusto a confronto.

In fase di addestramento, si nota come il modello robusto ottentga un 2% circa in più sotto il punto di vista della sensitività media; fuori campione (sul *test set*) il modello robusto ottiene invece un 2,6% in più, indicando che il modello robusto è in grado di riconoscere correttamente in media il 2,6% dei casi positivi (ovvero chi acquista) in più rispetto al modello standard.

Sotto l'aspetto dell'accuratezza media invece, il modello robusto in fase di addestramento ottiene il 3% circa in più rispetto all'approccio standard, in fase di test invece, il modello robusto ottiene un 4% in più rispetto al modello standard.

Risultano scontati, dunque, i benefici apportati dalla tecnica robusta sul dataset in studio, a parità di modello sottostante utilizzato (il modello Logit), attraverso una tecnica di stima robusta, è possibile incrementare sensibilmente le prestazioni di previsione sul dataset.

Avendo verificato la bontà della tecnica robusta, il modello robusto verrà utilizzato per lo studio delle implicazioni manageriali che ne derivano sul dataset in studio.

CONCLUSIONI

Grazie alla tecnica robusta, si è potuto osservare un sensibile miglioramento rispetto ad un approccio standard, permettendo di ottenere dei risultati più precisi sia in fase di addestramento che in fase di test; in un caso come quello in studio, risulta di fondamentale importanza la capacità del modello di poter prevedere in anticipo in modo preciso l'intenzione di acquisto del visitatore del sito.

Osservando i coefficienti medi del modello robusto calcolati in Tabella 19, è possibile ottenere le informazioni che seguono.

L'intercetta rappresenta la probabilità di acquisto (ovvero quando Y = 1) in termini di *odds ratio* nel caso in cui tutte le variabili esplicative siano pari a zero, ossia quando: le pagine vistate dall'utente hanno una media di tasso di uscita (Exit Rates) pari a 0, un valore pagina medio pari a 0, il numero di pagine relative al prodotto visitato dal visitatore sia zero e infine la visita sia avvenuta nel mese di Agosto (che rappresenta il mese base del modello).

Il coefficiente medio stimato della variabile Exit Rates, ha segno negativo, indice della correlazione negativa tra la probabilità di acquisto e il tasso medio di uscita delle pagine visitate dall'utente. Ad esempio, se il tasso medio di uscita delle pagine visitate aumentasse del 20% (quindi 0.2), il logaritmo dell'*odds ratio* diminuirebbe in media di 2.2457.

Nel caso del coefficiente medio stimato della variabile Page Values, per ogni incremento unitario del valore delle pagine medie visitate da un utente, a parità di tutto il resto, il logaritmo dell'*odds ratio* aumenta in media di 0.3838. Il segno del coefficiente inoltre è positivo sottolineando una correlazione positiva tra il valore medio delle pagine visitate e l'intenzione di acquisto.

Passando al coefficiente medio della variabile Product Related (ossia il numero di pagine relativo al prodotto visitate), si nota come anche in questo caso, il coefficiente abbia segno positivo per via della correlazione positiva con l'intenzione di acquisto. Ad esempio, se il numero di pagine relative ai prodotti visitate aumentasse di 1 (a parità di tutto il resto), il logaritmo dell'*odds ratio* della probabilità di acquisto aumenterebbe in media di 0.0050.

Per quanto riguarda lo studio del coefficiente medio stimato del mese in cui avviene la visita, si sceglie di descrivere solo alcuni coefficienti medi che riportano un valore assoluto alto (i quali impattano maggiormente in termini di probabilità di acquisto) per evitare ripetizioni.

Si sceglie di studiare, a titolo di esempio, i seguenti mesi: Febbraio, Maggio, Settembre, Novembre.

Nel caso di Febbraio, attraverso l'osservazione del coefficiente medio, si può concludere che se la visita avviene nel mese di Febbraio, la probabilità di acquisto espressa dal logaritmo dell'*odds ratio* diminuirebbe mediamente di 3.0026, indicando una scarsa propensione all'acquisto di chi visita il sito nel mese di Febbraio.

Per quanto riguarda il mese di Maggio, il coefficiente stimato medio, fornisce informazioni sulla presenza di una correlazione negativa tra la probabilità di acquisto e la visite avvenute nel mese di Maggio. Il logaritmo dell'odds ratio in questo caso diminuirebbe in media di 1.7097 a parità di tutto il resto.

Il coefficiente medio stimato del mese di Settembre invece, pari a 0.2847, permette di comprendere che esiste una certa correlazione positiva tra la probabilità di acquisto e le visite avvenute nel mese di Settembre, infatti se la visita avviene in questo mese, il logaritmo dell'*odds ratio* della probabilità di acquisto, in media, aumenterebbe di 0.2847 a parità di tutto il resto.

Il mese di Novembre, presenta un coefficiente medio stimato pari a 0.6919, indice di una correlazione positiva tra la probabilità di acquisto e le visite avvenute nel mese di Novembre, infatti se la visita avvenisse in questo mese, il logaritmo dell'*odds ratio* aumenterebbe in media di 0.6919.

Riepilogando, dunque, dallo studio dei coefficienti medi del modello robusto, emergono importanti considerazioni sia dal punto di vista manageriale, sia per quanto riguarda la gestione del sito.

Sotto il punto di vista del tasso di uscita, sulla base del coefficiente medio osservato, è possibile concludere che le pagine che presentano un tasso medio di uscita elevato durante il "percorso" dell'utente sul sito in studio, può, di fatto, far perdere al sito un potenziale acquisto per via dell'abbandono delle pagine da parte dell'utente, risulta chiaro dunque, sul perché un gestore di un sito deve concentrarsi sulla creazione di pagine e contenuti ingaggianti, che possano migliorare la *customer experience* dell'utente durante la navigazione; infatti, come evidenziato dello studio dei coefficienti medi, basta anche solo rialzo lieve del tasso medio di uscita per far "crollare" le probabilità di acquisto.

Il sito inoltre, dovrebbe incentivare l'utente a seguire i percorsi più "caldi" sul sito, ossia spingere il visitatore a visitare tutte quelle pagine che presentano un Valore Pagina elevato, ovvero tutte quelle pagine che in passato hanno contribuito maggiormente alle conversioni, da ciò è possibile ricavare informazioni preziose per quanto riguarda la costruzione di un percorso sul sito che possa garantire una *customer experience* di altissimo livello.

Il coefficiente medio con segno positivo della variabile Product Related, fornisce preziose indicazioni sul fatto che l'utente che visita numerose pagine relative ai prodotti e, dunque, trascorre più tempo su queste pagine, ha una probabilità maggiore di acquistare, dunque risulta opportuno creare un sito che presenti i prodtti in maniera "gradevole" e che spinga l'utente a trascorrere molto tempo sul sito.

Sotto il punto di vista mensile invece, è possibile osservare i mesi che riportano un coefficiente medio positivo (dunque che manifestano una correlazione positiva con la probabilità di acquisto) sono: Luglio, Settembre e Novembre.

Gli altri mesi rilevati (ad eccezione di Agosto che rappresenta l'anno base), presentano dei coefficienti medi negativi, indice di una correlazione negativa con la probabilità di acquisto.

Dunque, il gestore del sito dovrebbe concentrarsi sull'eseguire delle promozioni mirate, nei mesi che presentano coefficiente medio negativo, per provare ad incrementare le vendite in quei mesi che presentano delle conversioni basse, praticando delle strategie aggressive per attirare clientela come sconti, concorsi.

Lo studio proposto, permette di trarre importanti conclusioni sui *driver* che "spingono" il consumatore a finalizzare un acquisto su un sito *web*.

Dai risultati emersi dallo studio si trova riscontro positivo con quanto già esposto in letteratura da [4],[5] sull'importanza dei fattori contestuali in ottica predittiva dell'intenzione di acquisto del consumatore; emerge inoltre una marcata rilevanza delle metriche di terze parti legate al sito (fornite da Google Analytics) come quanto visto in [6].

In particolare nel caso in studio si rileva l'importanza dei seguenti fattori contestuali: mese e numero di pagine relative ai prodotti visitate.

Risultano molto impattanti nel modellare l'intenzione di acquisto anche i dati non contestuali forniti da terze parti che sono: tasso di uscita medio delle pagine visitate dall'utente, valore medio delle pagine visitate dall'utente.

Emerge chiaramente l'importanza (e la necessità) di possedere un sito *web* che possa attrarre e incuriosire chi lo visita e sappia fornire una *Customer Experience* di alto livello come visto in [1].

Il sito web ottimale per svolgere attività legate all'*e-commerce* è dunque, quel sito, che riesce a intrattenere l'utente, infatti nel caso in studio, l'utente che si intrattiente di più sul sito (e che visita un numero di pagine legate ai prodotti maggiore) ha più probabilità di acquistare; l'importanza di un sito web ben strutturato è evidenziato anche dall'importanza delle metriche di terze parti, infatti se si riesce a "far percorrere" all'utente un percorso che tocchi le pagine che presentano un valore della

metrica Valore Pagina elevato risulta molto più probabile che effettuino un acquisto; le pagine come già detto nel corso di questo studio, devono sapere catturare l'attenzione dell'utente che le visita spingendolo a restare sul sito senza abbandonare la navigazione come testimoniato dalla forte correlazione negativa tra l'intenzione di acquisto e il tasso medio di uscita delle pagine.

Nello studio, inoltre, risalta la bontà dell'approccio robusto proposto per gestire al meglio gli *outliers* che sono emersi nella fase dell'analisi esplorativa, dimostrando come una corretta gestione e comprensione delle problematiche legate ai *Big Data* sia di fondamentale importanza.

APPENDICE

In allegato a questo studio è possibile trovare il seguente materiale:

online_shoppers_intention.csv: dataset utilizzato nello studio in formato CSV.

File Ambiente Tesi.R, .Rhistory, .RData: file generati da RStudio, sowftware utilizzato per la scrittura ed esecuzione del codice utilizzato per svolgere l'intera analisi.

RINGRAZIAMENTI

A conclusione di questo lavoro di tesi di Laurea Magistrale desidero ringraziare chi mi è stato vicino per tutta la durata di questo splendido percorso.

Innanziatutto vorrei ringraziare i miei genitori per avermi sempre supportato e sostenuto nel mio percorso accademico e aiutato nei momenti difficili.

Ringrazio inoltre i miei compagni di corso con cui ho condiviso parecchi momenti insieme, in particolare Carlotta e Vanessa per l'incredibile supporto durante l'intero percorso.

Colgo l'occasione per ringraziare il mio relatore Prof. Aldo Goia e la correlatrice Prof.ssa Clementina Bruno per avermi seguito con cura e attenzione per l'intero svolgimento del lavoro di tesi dedicandomi parecchio tempo per risolvere ogni mio dubbio e/o problema.

BIBLIOGRAFIA

[1] C. Gentile, N. Spiller, e G. Noci, «How to Sustain the Customer Experience»:, *Eur. Manag. J.*, vol. 25, fasc. 5, pp. 395–410, ott. 2007, doi: 10.1016/j.emj.2007.08.005.

[2] «Global e-commerce share of retail sales 2027», Statista. Consultato: 13 giugno 2024. [Online]. Disponibile su: https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/

[3] S. Akter e S. F. Wamba, «Big data analytics in E-commerce: a systematic review and agenda for future research», *Electron. Mark.*, vol. 26, fasc. 2, pp. 173–194, mag. 2016, doi: 10.1007/s12525-016-0219-0.

[4] R. Esmeli, M. Bader-El-Den, e H. Abdullahi, «An analyses of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain», *J. Bus. Res.*, vol. 147, pp. 420–434, ago. 2022, doi: 10.1016/j.jbusres.2022.04.012.

[5] R. Esmeli, M. Bader-El-Den, e H. Abdullahi, «Towards early purchase intention prediction in online session based retailing systems», *Electron. Mark.*, vol. 31, fasc. 3, pp. 697–715, set. 2021, doi: 10.1007/s12525-020-00448-x.

[6] B. Clifton, *Advanced Web metrics with Google Analytics*, 3. ed. Chichester: John Wiley, 2012.

[7] «What's a Good Average Ecommerce Conversion Rate in 2024? - Shopify». Consultato: 21 giugno 2024. [Online]. Disponibile su: https://www.shopify.com/blog/ecommerce-conversion-rate

[8] C. J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. Del Jesus, e S. García, «Web usage mining to improve the design of an e-commerce website: OrOliveSur.com», *Expert Syst. Appl.*, vol. 39, fasc. 12, pp. 11243–11249, set. 2012, doi: 10.1016/j.eswa.2012.03.046.

[9] «Tecnologia e integrazioni di Analytics - Analytics», Google Marketing Platform. Consultato: 14 giugno 2024. [Online]. Disponibile su: https://marketingplatform.google.com/intl/it/about/analytics/features/

[10] Y. K. C. Sakar, «Online Shoppers Purchasing Intention Dataset». [object Object], 2018. doi: 10.24432/C5F88Q.

[11] «Frequenza di rimbalzo - Guida di Analytics». Consultato: 11 maggio 2024. [Online]. Disponibile su: https://support.google.com/analytics/answer/1009409?hl=it

[12] «[GA4] Entrances and exits - Analytics Help». Consultato: 11 maggio 2024. [Online]. Disponibile su: https://support.google.com/analytics/answer/11080047?hl=en

[13] P. Koks, «How Page Value in Google Analytics Can Improve Your Insights», Online Metrics. Consultato: 12 giugno 2024. [Online]. Disponibile su: https://online-metrics.com/page-value/

[14] W. N. Venables e B. D. Ripley, *Modern Applied Statistics with S*. in Statistics and Computing. New York, NY: Springer New York, 2002. doi: 10.1007/978-0-387-21706-2.

[15] E. Cantoni e E. Ronchetti, «Robust Inference for Generalized Linear Models», *J. Am. Stat. Assoc.*, vol. 96, fasc. 455, pp. 1022–1030, 2001.