

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea



STRATEGIA DI ACQUISIZIONE DELLA CLIENTELA :

UN MODELLO PREDITTIVO NEL MONDO DEL FASHION LUXURY ONLINE

Relatore: Lucia Dalla Pellegrina

Tesi di Laurea di:

Lorenzo Mauri

Matr. N. 807306

Anno Accademico 2018/2019

Desidero ricordare e ringraziare tutti coloro che mi hanno aiutato durante la stesura di questo lavoro di tesi, in particolare la prof.ssa Lucia Dalla Pellegrina senza la quale questo elaborato non avrebbe potuto prendere forma.

Ringrazio di cuore il team della società Nustox.com, nello specifico il Founder&CEO Matteo Milione, Co-founder&CDO Gianluca Bogialli, il responsabile di Data Analytics Alessandro Patuzzo ed il CTO Luca Campana per avermi dato preziosi consigli durante la definizione degli obiettivi del progetto.

Last but not least.... Un ringraziamento speciale ai miei genitori, senza il loro consiglio e sostegno non avrei mai potuto intraprendere questo percorso triennale, a Martina che mi incoraggia sempre nei momenti difficili e a tutti gli amici più cari.

Grazie mille.

Lorenzo

INDICE

i. INTRODUZIONE	6
1. DIGITAL MARKETING	8
1.1 Terminologia	8
1.2 Strategia SEO	9
1.3 Marketing Funnel	14
1.4 Metriche di Web Analytics	15
1.4.1 Cos è una sessione e quanto dura	16
1.4.2 Cos è una sessione e quanto dura Sessioni naturali, a pagamento e comprehensive	18
S1) SUMMARY	19
2. UNA PANORAMICA SUL MERCATO DEL LUSO & CASE STUDY NUSTOX	20
2.1 Il settore e le caratteristiche del mercato del lusso	20
2.2 Nustox	28
2.2.1 Modello di Business	29
S2) SUMMARY	30
3. WEB SCRAPING PER LA RACCOLTA DATI	32
3.1 Struttura del dataset	32
3.2 Web crawler	33
3.2.1 Elementi per l'implementazione in R	35
S3) SUMMARY	40

4. REGRESSIONE LINEARE MULTIPLA PER LA STIMA DEL NUMERO DI SESSIONI PAID	41
4.1 Specificazione del modello e stima dei parametri	44
4.2 Verifica e diagnostica del modello	48
4.3 Analisi dei residui	53
4.4 Utilizzo del modello : previsione	57
S4) SUMMARY	59
 5. IDENTIFICAZIONE DI UN MODELLO SARIMA PER LA PREVISIONE DEL NUMERO DI SESSIONI NATURAL.....	 60
5.1_Proprietà dei processi stocastici.....	60
5.2 Verifica di anomalie nei dati & verifica della stazionarietà.....	64
5.3 Identificazione & diagnostica del modello	68
5.4 Analisi dei residui	72
5.5 Utilizzo del modello : previsione	75
S5) SUMMARY	78
 6. BUSINESS PLAN & PREVISIONE DEL NUMERO DI ORDINI DI PRODOTTO.....	 78
6.1 Approcci per la stima del funnel di marketing.....	79
6.2 Alcune formule utili	82
 CONCLUSIONE	 84
 FONTI	 86

INTRODUZIONE

Nel corso degli ultimi anni e ancora oggi la tecnologia sta avendo un incredibile sviluppo e sta occupando gradualmente sempre più spazio nelle nostre vite.

Milioni di persone possiedono accesso ad Internet e si collegano regolarmente in rete, sia per svago, sia per motivi lavorativi o di studio.

Gli studenti infatti sono sempre più “digitalizzati” e nelle scuole od università è ormai divenuto indispensabile avere un computer personale così come possedere uno smartphone ed essere registrati sulle principali piattaforme social.

Di fatto il modo di comunicare è cambiato radicalmente, sono cambiate le nostre abitudini e persino la nostra quotidianità è stata stravolta dal mondo virtuale.

A sostegno di tale ipotesi, si stima che solo nel 2018 siano stati venduti circa 1.56 miliardi di smartphone, contro le 680 milioni di unità del 2012.

Il fenomeno cui siamo di fronte viene indicato dagli esperti con il termine “Digital Revolution” per sottolineare il passaggio dalla tecnologia meccanica ed elettronica analogica a quella elettronica digitale.

“Rivoluzione digitale” riporta oggi giorno alla mente una trasformazione di tipo multisetoriale, diventando sinonimo di cambiamento anche in campo economico, sociale e politico.

In particolare i consumatori sono passati dagli store fisici di qualche decennio fa ai siti di e-commerce o ai grandi marketplace, inducendo le aziende ad adattarsi alla nuova domanda e alle nuove opportunità offerte da Internet.

Nello specifico le imprese hanno dovuto affrontare nuove sfide dettate dal cambiamento e hanno scoperto le potenzialità del Digital Marketing, cambiando totalmente modello di business e approccio al mercato.

Prestando quindi particolare attenzione alle metodologie di ottimizzazione dei motori di ricerca proposte dal marketing digitale, questo elaborato ha lo scopo di fornire in ultima

analisi un modello statistico in grado di prevedere quantità utili alla crescita del business rivolto al cliente finale (B2C) . Utilizzato con scopi previsivi, quest'ultimo aiuterà l'azienda nei processi decisionali.

Durante il nostro "viaggio" verrà proposta inoltre una panoramica sul mercato del Fashion Luxury Online e un'implementazione di un semplice algoritmo per la raccolta dati in ambiente R.

Con il termine Digital Marketing identifichiamo tutte le azioni - che si avvalgono di un qualunque supporto digitale come il web, smartphone o social media - volte alla commercializzazione di un bene o servizio.

Si tratta quindi di una materia che si occupa di tutte le attività legate alla comunicazione delle aziende con i mercati e i clienti, attraverso l'integrazione dei nuovi canali digitali con quelli tradizionali.

Nel seguito faremo uso di terminologie specifiche perciò, per facilitare la comprensione, forniamo ora alcune definizioni.

1.1) Terminologia

- **Marketplace** : tradotto dall'inglese 'luogo di mercato' , indica generalmente siti internet di intermediazione per la compravendita di un bene o un servizio.

Il marketplace si differenzia da un e-commerce perché al suo interno sono presenti normalmente più venditori ed è, per certi versi, la stessa differenza riscontrabile tra un negozio di prossimità e un centro commerciale.

L'ente gestore del marketplace si fa solitamente carico del catalogo prodotti dei venditori e si preoccupa della sua sponsorizzazione digitale.

Il marketplace è preferibile poiché offre all'utente finale maggiori garanzie durante la finalizzazione dell'acquisto e la successiva ricezione del prodotto acquistato.

- **Search Engine Result Page** (acronimo SERP): pagina dei risultati di un motore di ricerca contenente un elenco di tutte le pagine web, ordinate secondo criteri specifici.
- **PageRank** : algoritmo di analisi che assegna una ponderazione numerica a ciascun elemento di un insieme di documenti con collegamenti ipertestuali.

- **Traffico paid** : traffico ottenuto mediante inserzioni sponsorizzate e campagne pubblicitarie online. E' doveroso sottolineare che se da una parte il traffico paid non porta vendite nell'immediato, dall'altra può condurre all'acquisizione di nuovi clienti nel lungo periodo. Solitamente il 30% del traffico totale è paid.
- **Traffico organico** : contrariamente, il traffico naturale rappresenta l'utenza più importante poichè, avendo un'intenzione d'acquisto chiara e precisa, presenta maggiori probabilità di conversione. Solitamente il 70% del traffico totale è organico.

1.2) Strategia SEO

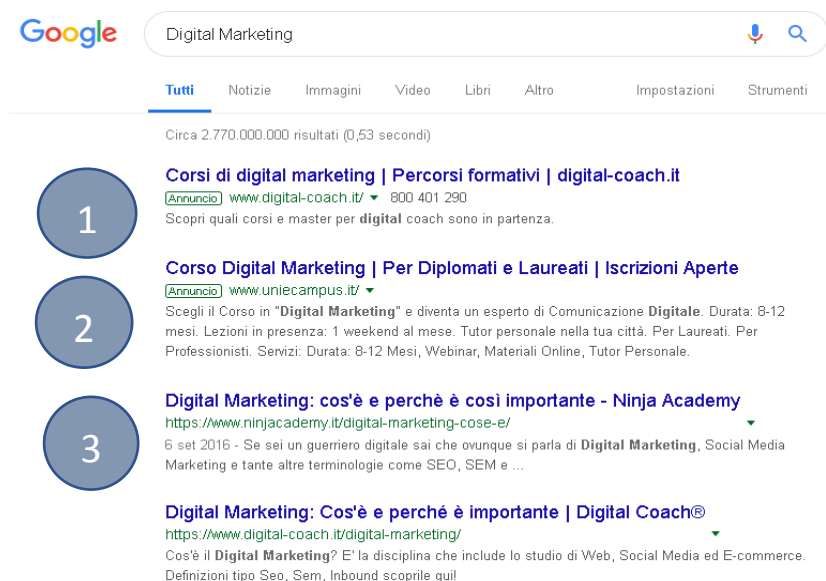
Partiamo dalla definizione di SEO proposta da Ian Dodson in uno dei suoi libri¹, la quale ci darà lo spunto per parlare di ottimizzazione dei motori di ricerca.

Con il termine "Search Engine Optimization" intendiamo il processo di affinamento di un sito web che utilizza metodi sia on-page che off-page in modo che il sito sia indicizzato e classificato con successo dai motori di ricerca.

Il SEO quindi riguarda una classificazione di documenti e indicizzazione degli stessi nel mondo online. In altri termini, quando utilizziamo il termine "SEO" ci riferiamo ad una strategia atta a migliorare continuamente la posizione del nostro sito web nei risultati forniti dal motore di ricerca (acronimo inglese di SERP ,Search Engine Results Page in *Figura 1.1*). I siti vengono ordinati secondo criteri precisi che tengono conto, ad esempio, della pertinenza e autorevolezza del sito.

¹ " L'arte del marketing digitale. Guida per creare strategie e campagne di successo" , Ian Dodson , Milano 2016

Figura 1. 1



La logica competitiva, in questo caso, è analoga a quella che sussiste tra i negozi fisici : se l'avere un'attività su una piazza significava avere un quantitativo più ampio di clienti rispetto ad un negozio situato nelle vie interne del paese, con l'avvento di internet la competizione si è spostata in un'altra direzione e l'obiettivo è diventato quello di posizionare il proprio sito più in alto nella SERP.

Chiaramente, ad una posizione più alta corrisponde un volume maggiore di traffico e quindi di consumatori.

La strategia si articola principalmente in 4 fasi² :

1. **Definizione degli obiettivi** : in una prima fase l'azienda si pone degli obiettivi di medio-lungo periodo. Di conseguenza, prima di intraprendere questa strada è necessario valutare gli aspetti positivi e negativi in funzione del traguardo finale.
2. **Strategie dirette** : riguarda tutte le azioni di ottimizzazione effettuate sul proprio sito. Questa fase comprende la keyword research, cioè la ponderazione dell'investimento economico sulle parole che generano più traffico (e che quindi sono quelle più cercate su Google).

² " L'arte del marketing digitale. Guida per creare strategie e campagne di successo" , Ian Dodson , Milano 2016

3. **Strategie indirette** : riguarda tutte le azioni di ottimizzazione effettuate al di fuori del sito come le strategie di back-link (che spiegheremo nel seguito, pagina 12)
4. **Analisi delle performance** : l'ultima fase è dedicata all'analisi dei dati e miglioramenti in termini di posizionamento ottenuti. E' possibile monitorare le campagne pubblicitarie tramite Google Adwords e le statistiche del sito attraverso Google Analytics.
In questo momento è doveroso, qualora necessario, effettuare una revisione.
Il processo è comunque ciclico, pertanto lo step successivo sarà ancora il punto 1.

Ottimizzare il posizionamento di un sito significa quindi agire in modo diretto e indiretto tramite due strategie :

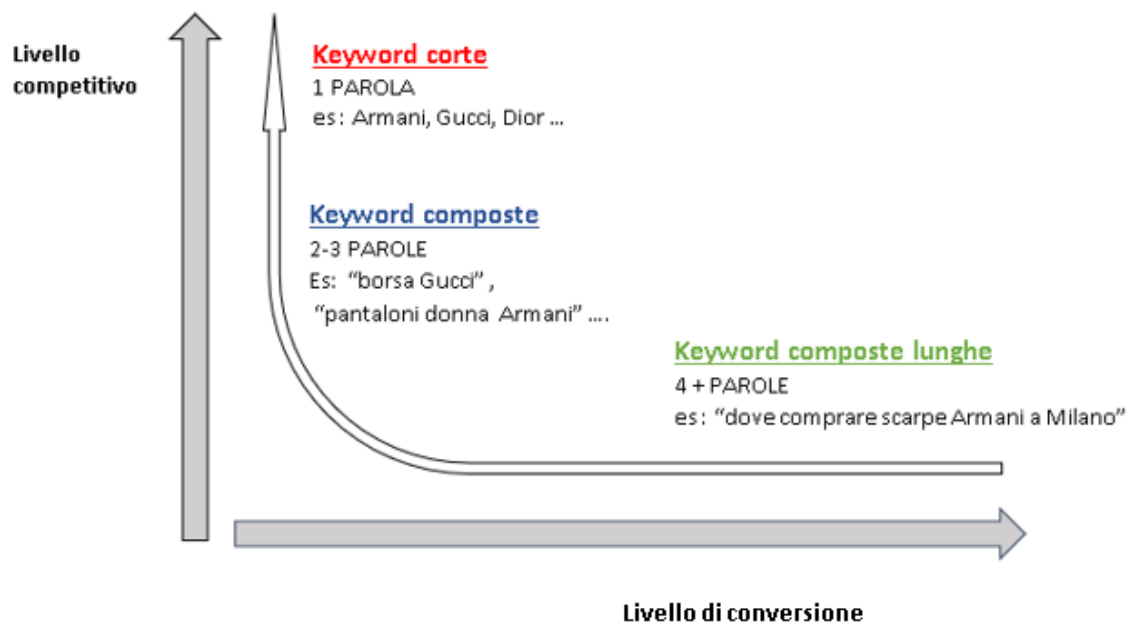
- La **Keywords strategy** è una strategia diretta alla pagina web che riguarda l'individuazione dei termini di ricerca più digitati. Questi ultimi dovranno essere inseriti sapientemente all'interno dei contenuti del sito in modo tale da essere letti dall'algoritmo di Google e posizionati correttamente nella SERP.

Le parole chiave si distinguono prevalentemente per la lunghezza :

- Le keyword corte sono termini di una sola parola e sono molto competitive in quanto i volumi di ricerca che generano sono elevatissimi. Basti pensare alle parole "Armani", "Gucci" o "Dior" che vengono cercate milioni di volte al mese.
Pertanto, *in linea teorica*, l'azienda che sarà meglio indicizzata su questi termini godrà certamente di un numero elevato di clienti. Nella pratica è bene notare che il nome del brand è "posseduto" dal brand stesso quindi per un'altra azienda sarà quasi impossibile collocarsi più in alto nella SERP.
- Le keyword composte sono stringhe di due o tre parole, come ad esempio "borsa Gucci" oppure "pantaloni donna Armani". Generando meno traffico rispetto a quelle corte risultano meno competitive ma offrono una più facile indicizzazione.
- Le keyword composte lunghe, seppur poco ricercate sono le più interessanti in quanto scarsamente competitive e con un alto livello di conversione. Infatti un consumatore che ricerca "dove comprare scarpe Armani a Milano" avrà un'intenzione d'acquisto più forte rispetto al cliente che digita soltanto "Armani".

- La strategia adottata da molte aziende è quindi quella di scegliere un paniere di keyword lunghe proprio per il fatto che un insieme di questi termini è capace di generare molto più traffico rispetto alle keyword singole.

Figura 1.2 : categorie di Keyword



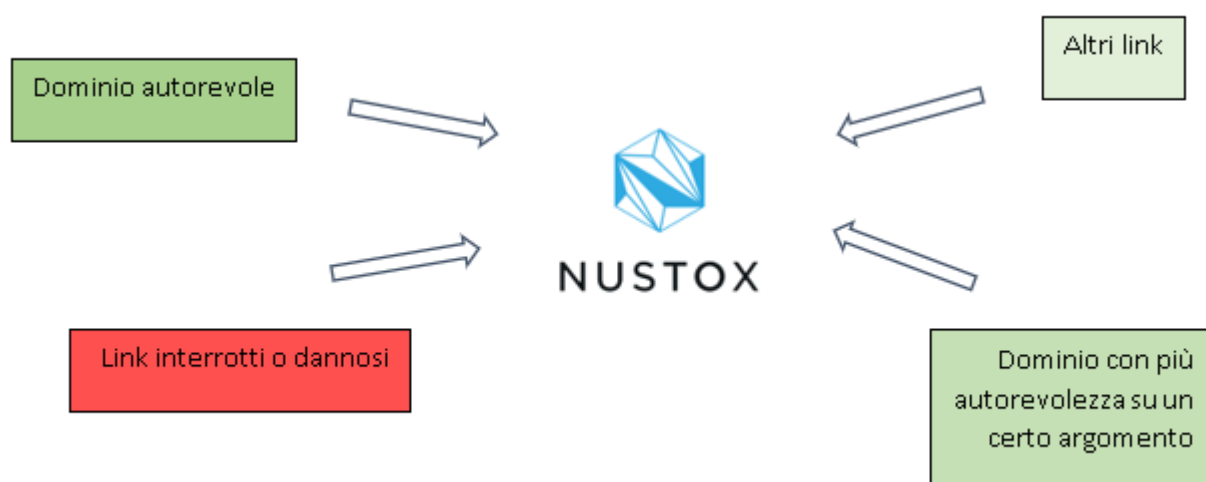
- La **Backlink strategy** è una strategia indiretta esterna che si basa sui collegamenti link presenti nelle pagine web. I motori di ricerca, durante il calcolo della SERP, attribuiscono uno score positivo per ogni link che punta alla nostra pagina quindi per ogni persona che “parla” positivamente di noi. Ad esempio, se un dominio con un punteggio di autorevolezza elevato ha una buona opinione della nostra azienda, allora Google tenderà a proporre i nostri contenuti prima di altri.

I link vengono distinti principalmente in esterni oppure interni : un link viene definito *in ingresso*³ quando il consumatore visualizza la nostra pagina grazie ad un sito web esterno, mentre al contrario, un link interno porta il cliente a visualizzare altre pagine.

³ “ L’arte del marketing digitale. Guida per creare strategie e campagne di successo” , Ian Dodson , Milano 2016

Tuttavia quelli *interrotti* a causa di mancati aggiornamenti o quelli *dannosi*, come le fonti di spam, vengono penalizzati e influiscono negativamente sul posizionamento.

Figura 1. 3



- Perché proprio Google ?

In questo elaborato si è deciso di considerare unicamente il motore di ricerca Google perché è quello maggiormente utilizzato : infatti le statistiche confermano che da desktop Google gode di una quota di mercato dell'80.5 % su scala globale⁴.

Inoltre mette a disposizione alcuni tools che rendono l'analisi SEO molto più immediata e intuitiva (come Adwords ed Analytics).

Google sfrutta, come altri motori di ricerca, un algoritmo complesso per posizionare i siti web all'interno della SERP. Mediante tale procedimento sistemico di calcolo, il motore di ricerca riesce a fornire all'utente dei risultati che siano pertinenti con l'interrogazione effettuata.

⁴ "Google domina il mondo della ricerca", Il Sole 24 Ore (2017)

<http://www.infodata.ilsole24ore.com/2017/05/10/google-domina-mondo-della-ricerca-soprattutto-console/>

I parametri con cui le pagine web vengono classificate sono innumerevoli, come ad esempio :

- a. la frequenza e la posizione della keyword all'interno della pagina
- b. la longevità della pagina web in esame
- c. il numero di pagine web che sono collegate alla pagina in esame

1.3) IL MARKETING FUNNEL

La conversione del traffico online avviene secondo lo schema del funnel di marketing online, ossia una struttura a imbuto che cerca di descrivere il processo di acquisizione della clientela.

Figura 1. 3



Partendo dall'uppermost stage, ovvero lo strato più alto, troviamo i potenziali clienti raggiunti attraverso blog, webinar, canali social e mail, oltre alle suddette campagne di marketing e annunci sponsorizzati. Grazie a questi strumenti di marketing l'utente forma la propria consapevolezza circa il prodotto proposto dall'azienda.

Gli utenti nello strato immediatamente sottostante si informano e iniziano a nutrire interesse per il prodotto, sebbene limitato.

La fase successiva permette di inviare contenuti più specifici al consumatore in quanto viene reso noto un maggiore coinvolgimento dello stesso. Si possono utilizzare mail automatizzate, prove gratuite o specifiche sui prodotti.

L'utente può considerarsi veramente interessato quando dimostra di voler comprare un prodotto di uno specifico brand, ad esempio quando aggiunge un articolo al carrello online. Il penultimo step è il più critico di tutto il processo di acquisizione poichè l'acquirente si trova di fronte ad una scelta e valuta se effettuare l'acquisto o meno. L'obiettivo finale del reparto marketing e vendite è ovviamente la conversione quindi, una volta giunti allo step finale, il cliente verrà indotto a comprare.

1.4) METRICHE DI WEB ANALYTICS

Gli strumenti di digital analytics sono molteplici e possono essere raggruppati in 3 macrocategorie¹: web analytics, ad server e tool di attribution.

I servizi di web analytics più conosciuti dagli addetti ai lavori sono oggi Google Analytics e Adobe Analytics e servono principalmente per il tracciamento degli utenti online (ovviamente in forma anonima).

Per fare ciò ogni dispositivo web è identificato da un codice univoco (stringa di testo chiamata "cookie") inviato da un web server ad un browser e poi rimandato indietro ogni volta che l'utente compie una specifica azione.

La piattaforma open-source Analytics mette a disposizione diverse KPI (Key Performance Indicators) e statistiche del sito web, utili per monitorare il business.

Tradotto dall'inglese, un "indicatore chiave di performance" è una metrica quantificabile, facilmente aggiornabile nel breve periodo e legata strettamente agli obiettivi aziendali.

Il numero di tali indicatori è generalmente ristretto e può variare a seconda del settore e dell'azienda, in quanto a realtà diverse corrispondono necessità e obiettivi differenti.

Premesso che non esistono criteri assoluti per la scelta degli indicatori, possiamo però attenerci a delle best practices basate sulla tipologia di market in cui l'azienda opera e sulla propria longevità.

Nel caso del mercato del luxury, le cui caratteristiche saranno esaminate nel capitolo successivo, l'attenzione delle KPI è rivolta alla qualità delle visite degli utenti e alla fedeltà degli stessi.

Il tasso di conversione (rapporto tra numero di transazioni e utenti), ad esempio, servirà per monitorare la velocità di crescita del business.

I servizi ad server (come DCM, AdForm e Sizmek) sono invece dedicati al tracciamento dell'attività di web marketing mentre i tool di attribution sono volti a migliorare l'interazione tra diversi canali di acquisizione.

Per l'analisi che condurremo nei prossimi capitoli abbiamo scelto le seguenti metriche :

- Numero di sessioni complessive
- Numero di sessioni naturali
- Numero di sessioni a pagamento
- Investimenti in pubblicità Google
- Investimenti in pubblicità in Facebook

1.4.1) COS E' UNA SESSIONE E QUANTO DURA

Vediamo ora cosa si intende con "Sessione" e la sua durata.

Come mostrato in *Figura 1* una sessione è un insieme di interazioni (anche diverse tra loro) effettuate con un determinato sito web in un arco temporale.

Con il termine "interazioni" intendiamo eventi (ad esempio quando un utente aggiunge un prodotto al carrello), transazioni, interazioni social e visualizzazioni di pagina.

Figura 1



Il conteggio del servizio di Google avviene secondo criteri ben precisi basati principalmente sui fattori di tempo e di cambio di campagna.

Distinguiamo infatti una sessione dall'altra in base al tempo di inattività, dal cambio data e dal tipo di inserzione pubblicitaria che ha permesso all'utente di interagire con il sito.

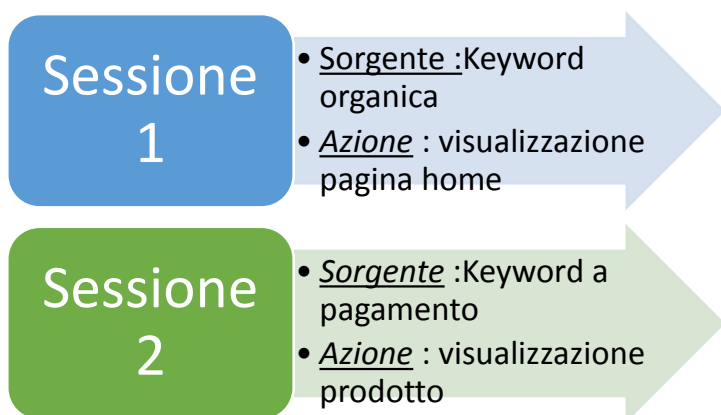
Teoricamente una sessione può durare anche 24 ore ma solitamente termina prima.

Per fare un esempio concreto, quando un utente si allontana dal pc e non effettua azioni per 30 minuti la sessione viene chiusa. Va ricordato però che se nel mentre si dovesse interagire con il web, Analytics sposterebbe il tempo di scadenza.

Anche la sorgente del traffico è importante poichè quando il valore della campagna viene aggiornato, Analytics apre una nuova campagna.

Avremo pertanto due diverse sessioni a seconda che l'utente sia arrivato al nostro sito tramite un annuncio piuttosto che da un altro, come mostrato in *Figura 2*.

Figura 2



- **Quando una campagna viene aggiornata ?**

In genere l'aggiornamento avviene quando la visualizzazione del sito è dovuta ad un motore di ricerca, url codificato o un sito web referente.

In altri termini, viene creata una nuova campagna ogni volta che l'utente fa click su un link che rimanda alla nostra pagina web⁵.

1.4.2) SESSIONI NATURALI, A PAGAMENTO E COMPLESSIVE

Google Analytics mette a disposizione tantissime metriche che certe volte possono sembrare ridondanti. Talvolta piuttosto che essere d'aiuto, confondono chi si ritrova ad analizzarle.

Per i fini di questo elaborato, come detto in precedenza, ci avvaliamo dei dati inerenti alle sessioni naturali, a pagamento e complessive.

Le sessioni organiche sono le più importanti in quanto delineano una forte propensione degli utenti all'acquisto.

Tuttavia esse dipendono fortemente dalla stagionalità, dalle condizioni del mercato e da fattori esogeni come, ad esempio, le principali festività.

Un compratore è infatti più stimolato all'acquisto durante le feste di Natale o durante i saldi piuttosto che in altri periodi dell'anno e di conseguenza, in queste circostanze, il traffico spontaneo subirà inevitabilmente un aumento (quindi aumenteranno anche le sessioni).

Di contro, quelle a pagamento sono le più semplici da ottenere poichè legate linearmente a componenti endogene come l'investimento in attività pubblicitarie. Un aumento del budget spendibile in marketing, infatti, porta quasi sempre ad un aumento delle sessioni.

Infine, le sessioni complessive sono generate dalla somma delle precedenti e

⁵ https://support.google.com/analytics/answer/2731565?hl=it&ref_topic=1012046

presentano una misura generale dell'andamento del business.

S1) SUMMARY

In questo primo capitolo “Digital Marketing” abbiamo introdotto la strategia di ottimizzazione dei motori di ricerca utile a migliorare il posizionamento della webpage aziendale. Se correttamente applicata, tale tecnica permette ad un'azienda di incrementare la propria visibilità online, quindi di generare più traffico e di conseguenza aumentare il numero delle vendite.

Per monitorare il processo di conversione delle utenze, descritto dal marketing funnel, gli analisti aziendali sfruttano i tools offerti da Google, primo tra tutti Analytics .

Sebbene vi siano infinite metriche e KPI suggerite da questo strumento, il focus della nostra discussione verterà sul **numero di sessioni complessive, naturali e a pagamento, investimenti in pubblicità Google e Facebook.**

Nel prossimo capitolo analizzeremo le caratteristiche del mercato del lusso, in particolare ci soffermeremo sul compartimento dell'abbigliamento e capiremo le logiche che muovono i suoi operatori.

Infine presenteremo un caso di studio aziendale.

2.1) IL SETTORE E LE CARATTERISTICHE DEL MERCATO DEL LUSO

Prima di definire le peculiarità del mercato luxury, è doveroso comprendere quando un brand si possa definire 'di lusso' e quando invece non appartiene a tale categoria.

Purtroppo non vi è un'accezione assoluta tant'è che alcune volte capita di pensare al brand di lusso come ad un marchio unico che impone il suo stile all'interno di una nicchia, mentre altre di riferirsi ad un brand che comunica un modo di vivere caratterizzato da eleganza, raffinatezza e benessere.

Ebbene, possiamo soltanto definire delle linee generali ricordandoci sempre che non esiste un settore specifico del lusso, bensì un settore che abbraccia più categorie di prodotto.

La qualità e l'innovazione sono caratteristiche basilari in questo mercato : se si pensa alla nicchia dell'abbigliamento, gli stilisti pongono grandissima attenzione ai materiali e al design, perseguendo un certo gusto artistico.

Al cliente piace pensare che il prodotto sia artigianale cioè realizzato a mano con estrema cura per i dettagli, anche se di fatto la maggior parte della produzione viene realizzata in modo automatizzato. Inoltre un prodotto costoso è idealmente anche un prodotto pregiato e per questa ragione il prezzo deve necessariamente mantenersi sopra certe soglie standard.

Un'altra caratteristica importante è rappresentata dal canale distributivo : esso deve obbligatoriamente essere la botique fisica oppure un e-commerce / marketplace dedicato in modo tale da conferire ai prodotti un alone di esclusività, molto apprezzato dal cliente che vuole sentirsi unico.

Per fare qualche esempio, Armani e Dior possono definirsi brand di lusso ma non possono fare lo stesso Zara e H&M a causa del prezzo medio dei loro prodotti e dei canali di distribuzioni poco selettivi.⁶

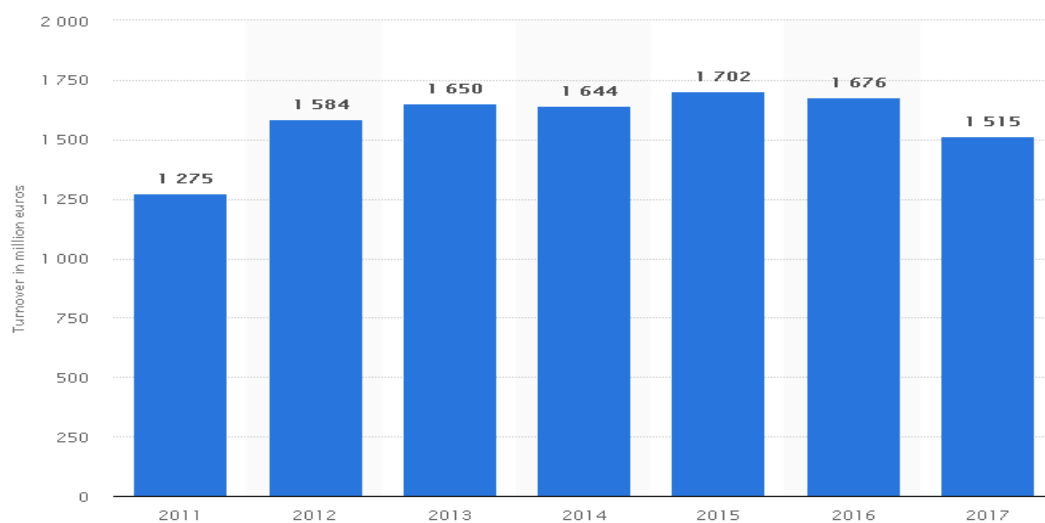
⁶ "L'economia delle aziende di abbigliamento", Elisa Giacosa, G.Giappichelli EDITORE, 2011

Analogamente, il mercato di nicchia del fashion luxury online si distingue principalmente per la tipologia di clientela e per gli attori economici al suo interno.

Alessandro Perego, direttore Scientifico degli Osservatori Digital Innovation del Politecnico di Milano, conferma che *“ L’Abbigliamento è uno dei comparti merceologici più dinamici dell’eCommerce B2C italiano per almeno tre ragioni: ritmo di crescita superiore a quello medio del commercio elettronico, offerta eterogenea e in continuo fermento, e, infine, spiccata propensione all’innovazione”*

Un case study interessante, a livello nazionale, è quello del gruppo Armani s.p.a che ha registrato nel solo 2017 un fatturato di ben 1.515 milioni di euro⁷ rimandendouna delle società più redditizie sul territorio italiano.

Grafico 2. 1



La crescita (+18.82%) dell’Emporio Armani tra 2011 e 2017 è sinonimo che i consumatori del settore rispondono positivamente alle proposte del mercato. A sostegno di tale affermazione, anche l’Osservatorio indica una crescita annua con un tasso del 30% circa cioè riporta un dato coerente con quanto appena detto.

⁷ <https://www.statista.com/statistics/694400/turnover-of-italian-fashion-company-giorgio-armani/>

La spesa dei clienti del comparto del fashion online sarebbe infatti passata da 1.47 mld nel 2015 contro gli 1.83 mld del 2016.⁸

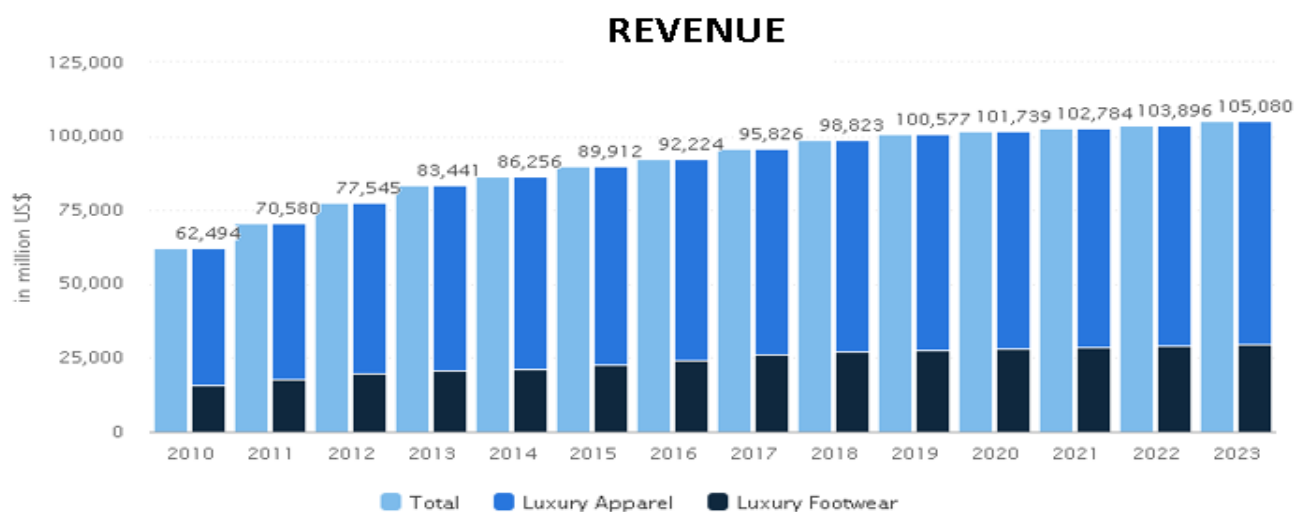
Dal lato dell'offerta, invece, i web shopper effettuano - nel 46% dei casi - acquisti sui marketplace generalisti.

Seguono, con un peso del 36%, gli acquisti 'luxury' realizzati sui siti delle grandi Dot Com, delle vendite private, dei produttori high fashion del Made in Italy e delle boutique multi-brand⁹.

Per quanto concerne i ricavi, Statista.com è ottimista circa il futuro del mercato mondiale tanto da stimare una crescita del 4.47 % tra il 2019 e il 2023.¹⁰

Il *Grafico 2.2* mostra anche i ricavi derivanti dal comparto abbigliamento e calzature.

Grafico 2.2



Source: Statista, March 2019

La crescita economica, in questo mercato, dipende sostanzialmente dall'andare di diversi fattori come :

- PIL : potendo essere considerato un indicatore del benessere di una nazione, le variazioni del pil influenzano la domanda di beni di lusso. Teoricamente infatti la relazione è direttamente proporzionale.

⁸ <http://ecommerce.moda/statistiche-ecommerce/dati-statistiche-ecommerce-moda-fashion-2016-italia/>

⁹ https://www.osservatori.net/it_it/osservatori/comunicati-stampa/la-online-nel-fashion-un-canale-che-fa-tendenza

¹⁰ <https://www.statista.com/outlook/21030000/100/luxury-fashion/worldwide#market-revenue>

- Tassi di cambio : il cambio di valuta influenza i ricavi delle aziende del settore per quanto riguarda i prodotti venduti all'estero.
- Mercati emergenti : le opportunità offerte dai Paesi emergenti rappresentano un potenziale notevole di crescita per i vari brand.
- Internet : l'utilizzo delle tecnologie emergenti permette di vendere ed acquistare online, incrementando il fatturato. Se qualche decennio fa il digital era poco conosciuto, oggi sta diventando il canale distributivo principale delle grandi aziende.

I principali operatori del fashion luxury sono riportati in *Figura 2.1* classificati secondo il traffico generato dai propri siti web nell'arco di un mese e con le rispettive quote.

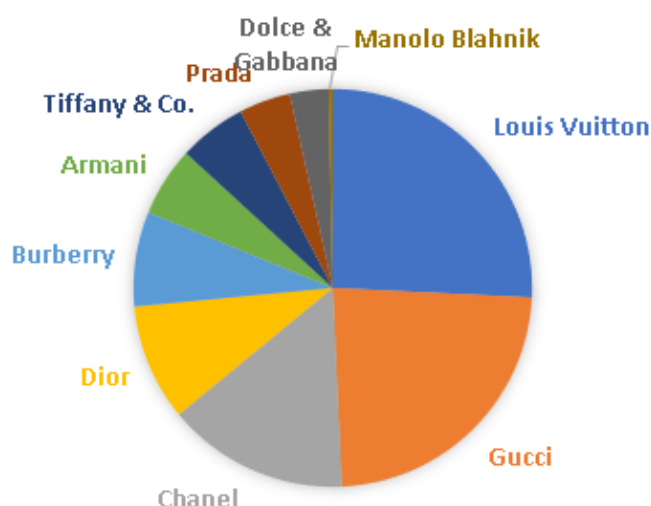
Il grafico a torta ci mostra invece una ripartizione del mercato.

Figura 2. 1

BRAND	TRAFFICO TOTALE (mese)	Market share
Louis Vuitton	8.700.000	25,8%
Gucci	7.940.000	23,5%
Chanel	4.980.000	14,7%
Dior	3.200.000	9,5%
Burberry	2.600.000	7,7%
Armani	1.900.000	5,6%
Tiffany & Co.	1.860.000	5,5%
Prada	1.410.000	4,2%
Dolce & Gabbana	1.070.000	3,2%
Manolo Blahnik	104.850	0,3%
Totale	33.764.850	1

Grafico 2. 3

TOP 10 LUXURY MARKET



Tutte le aziende facenti parte di questo mercato sono Customer-Oriented, ovvero mirano alla soddisfazione della clientela.

L'attenzione è quindi rivolta, in generale, a :

1. Customer Experience
2. Customer Journey
3. Livello di notorietà della propria marca (Brand Awareness)
4. Percezione della marca (Brand Identity) e comunicazione visiva (Brand Image)

1) Customer Experience

La Customer Experience è definita come il modo in cui i clienti percepiscono l'insieme delle loro interazioni con l'azienda ¹¹.

Per Luciano D'Arcangelo, Analytics e Customer Intelligence Solution Manager SAS, *“si parla di economia dell'esperienza a indicare una forte richiesta da parte della domanda non più solo di prodotti ma anche di esperienze uniche, personalizzate, addirittura ‘memorabili’ attraverso servizi e contenuti “.*

¹¹ Outside In, H.Manning, K. Bodyne, 2012, Forrester Research

In altre parole il cliente del mercato di lusso, oltre ad essere definito “alto spendente”, è anche un consumatore non razionale poichè le sue scelte d’acquisto sono spesso condizionate da una componente emozionale.

Egli è infatti alla ricerca di un’esperienza che assume valore in quanto soggettiva ed irripetibile, quindi non replicabile né da altri prodotti né da altri individui.

L’emozione in questo caso supera quindi il valore intrinseco del capo, la quale contribuisce alla percezione che il cliente ha del prodotto.

Tuttavia sussistono ancora dubbi e perplessità nelle aziende sul come disegnare una CX di successo.

Lo rivelano alcuni risultati dello studio “Lessons From The Leading Edge Of Customer Experience Management” condotto e pubblicato da Harvard Business Review Analytics Services in collaborazione con SAS nel 2014.

Il 45 % del campione di 403 executives vede la CX come una priorità strategica ma sempre la stessa percentuale riscontra difficoltà sui seguenti punti :

- a. Integrazione con i sistemi aziendali (41%)
- b. Complessità dell’omnicanalità (37%)
- c. Incanalazione degli insight (33%)

Queste statistiche rivelano che vendere un’esperienza assieme al prodotto è un compito tutt’altro che semplice, il quale può talvolta portare ad errori di valutazione da parte dei manager.

La CX rischia di essere considerata dai dirigenti d’azienda come semplice cambiamento estetico del sito web o di un evento ma è in realtà molto di più : la sfida allora diventa quella di capire quali siano i fattori per una comunicazione digitale innovativa.

Una possibile soluzione arriva da uno studio condotto da un ente di ricerca in collaborazione con SAS, il quale rivela che l’integrazione tra più canali e la capacità di avere una vista unica e univoca dell’utente costituiscono i fattori chiave per una comunicazione di successo ¹².

¹² Dati dallo studio “Data Elevates The Customer Experience” in collaborazione con SAS.

2) Customer Journey

Il Customer Journey è un termine usato in ambito marketing per indicare tutti i punti di contatto (diretti e indiretti), le interazioni del consumatore con il sito aziendale.

Il “viaggio del consumatore” permette di capire le preferenze dei clienti, gli interessi e le intenzioni di acquisto in modo da migliorare l’esperienza di consumo.

3) Brand Awareness

La brand awareness costituisce un elemento fondamentale nelle teorie di marketing poiché è strettamente legata alla notorietà presso il pubblico target. Infatti la BA rappresenta il grado di conoscenza (o consapevolezza) che i consumatori hanno del marchio, non solo in termini quantitativi ma anche qualitativi.

Una grande boutique di moda è quindi non solo capace di farsi riconoscere ma anche di trasmettere valori e concetti che hanno lo scopo di portare il cliente alla fidelizzazione.

A tal proposito David Aaker, celebre economista statunitense noto per numerosi elaborati riguardanti la gestione del marchio, propone una struttura piramidale della conoscenza del brand nella quale il cliente passa dall’ignorare totalmente l’esistenza della marca (base della piramide) al ricordarne valori, principi, filosofie, prodotti e servizi (apice della piramide in

Figura 2.2).

Lo scopo aziendale è costituito, come si può intuire, dal raggiungimento della “vetta” ovvero quella situazione nella quale la realtà aziendale rappresenta, per il pubblico, la prima scelta.

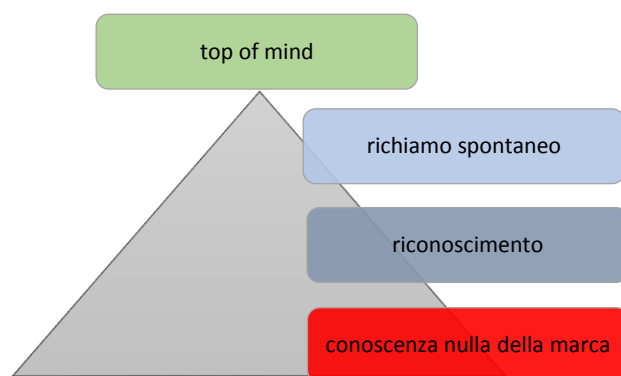


Figura 2. 2

4) Brand Identity & Brand Image

Il concetto di identità del marchio è collegato alla percezione che i clienti hanno del brand. Il primo elemento su cui si basa la BI è il nome dell'azienda, a cui sono legati altri elementi di minore importanza come lo slogan e aspetti grafico-comunicativi.

Mentre il nome può essere astratto o descrittivo ma non necessariamente corto, lo slogan deve riassumere un messaggio in poche parole ed essere facile da ricordare.

Celebri slogan sono ad esempio quelli di Nike "Just do it" oppure "Think different" di Apple. Per l'azienda informatica di Cupertino fondata da Steve Jobs e Wozniak l'obiettivo era quello di creare un'identità riconoscibile¹³ cioè un logo innovativo e dotato di semplicità comunicativa che potesse venire ricordato facilmente dalla collettività, indipendentemente da età e ceto sociale.

Scelsero per questo motivo una mela morsicata, riconducibile nella letteratura a Newton e all'episodio biblico di Adamo ed Eva.

"Le mele , d'altra parte, sono anche portatrici di sensazioni positive e vengono associate alla buona salute. Infatti, un vecchio proverbio recita : una mela al giorno toglie il medico di turno ". (Dipartimento legale Apple Computer Inc.) ¹⁴

Allo stesso modo le boutique dell'alta moda che si affacciano al mondo digital desiderano essere identificate positivamente come uniche ed insostituibili e mantenere inalterata la propria immagine di lussuosità (proprio come è percepibile negli store fisici). Mirano inoltre, attraverso diverse strategie di marketing, alla fidelizzazione del cliente piuttosto che al compratore one-shot.

¹³ "Apple : storia della mela più sexy del mondo. La lussuria del marchio secondo Steve Jobs" , Marco Giamberini, 2012, p.11

¹⁴ "Emozione Apple. Fabbricare sogni nel XXI secolo" , A. Dini , Milano , Il Sole 24 ORE S.p.a.,2008 [2007], p.188

2.2) NUSTOX

La stesura di questo elaborato è stata possibile grazie alla collaborazione con la startup Nustox.com, una realtà innovativa ed emergente operante nel settore del fashion luxury online.

Durante l'esperienza triennale ho avuto infatti la fortuna di incontrare i professionisti di questa società e di entrare a stretto contatto con i processi aziendali che la caratterizzano. Il percorso intrapreso, della durata complessiva di sedici settimane, è stato particolarmente utile sia per la mia crescita personale che lavorativa poiché ha favorito l'applicazione dei concetti teorici appresi in università.

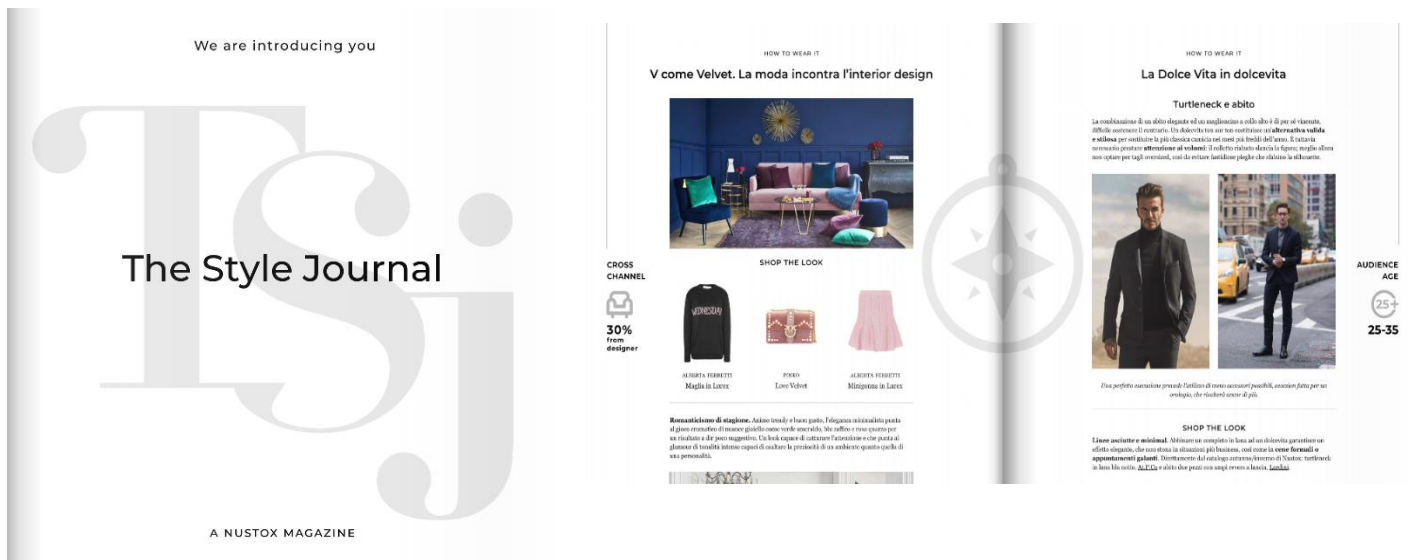
La società - fondata nel 2016 dalla passione, voglia e intraprendenza di Matteo Milione, Francesco Tetro e Gianluca Bogialli – nasce come startup tecnologica che utilizza il suo knowhow per supportare le operazioni quotidiane e le sfide digitali dei rivenditori al dettaglio.

La piattaforma è un marketplace esclusivo di moda del lusso, attraverso la quale portare online le migliori Fashion Boutique italiane ed internazionali.

La strategia di Nustox è infatti quella di creare partnership reciprocamente gratificanti con rivenditori e marchi leader del settore e, aspetto fondamentale della strategia di marketing, è quella di ingaggiare l'utente attraverso la creazione di contenuti originali, che raccontino la moda, i designers e il lifestyle. Sono infatti contenuti attraverso i quali gli utenti possono identificarsi e rispecchiarsi il più possibile nei loro bisogni.

Tutto ciò è possibile grazie al The style Journal¹⁵, un magazine online che tratta proprio le suddette tematiche.

¹⁵ "The style Journal" : <https://www.flipsnack.com/nustox/the-style-journal.html>



2.2.1) Modello di business

Nustox.com adotta un modello di business B2B2C, schematizzato in *Figura 2.3*.

Figura 2. 3



Il marketplace ha la peculiarità di non possedere un magazzino con i prodotti di aziende esterne, bensì quella di svolgere il ruolo di intermediazione tra le esigenze dei due operatori economici. In aggiunta il servizio di dropshipping rende più veloci tutti i processi di spedizione e ricezione coinvolti.

- Catalogo dei prodotti

Il catalogo dei prodotti delle boutique che aderiscono al marketplace viene sponsorizzato da Nustox.com.

Il processo di caricamento, che avviene prevalentemente in due momenti dell'anno, è piuttosto lungo ed è affidato al team IT.

Ad ogni prodotto viene associato un codice identificativo (SKU number), le specifiche, il nome del marchio e una breve descrizione, tradotto anche in lingua inglese.

- Logistica

Al momento dell'ordine, il servizio di dropshipping permette di affidare la logistica ai corrieri della società DHL, i quali si preoccupano di trasportare il prodotto direttamente dalla boutique al domicilio del cliente. In tal senso il consumatore effettua un acquisto in modo tradizionale, esattamente come farebbe in un negozio fisico.

- ✓ Vantaggi :

Rispetto ai loro Competitor quali Farfetch.com, Giglio.com, Luisaviaroma,

Nustox.com offre un canale di vendita più flessibile e centralizzato sull'analisi dei dati.

Retailer e Designer che collaborano con l'azienda possono infatti usufruire della reportistica generata dagli Analytics del sito per conoscere le esigenze dei propri consumatori e le relative preferenze.

Risulta pertanto conveniente, da parte delle case di moda, affidarsi ad una realtà dedicata piuttosto che rivolgersi ai canali più conosciuti.

S2) SUMMARY

Nel corso del capitolo "Una panoramica sul mercato del lusso & Nustox.com" abbiamo descritto le caratteristiche principali degli operatori economici del settore.

Il cliente del mercato di lusso è nettamente diverso da un consumatore tradizionale, soprattutto per le preferenze e i marchi come Louis Vuitton, Gucci o Chanel, oltre ad avere una presenza fisica con negozi sul territorio, dominano anche il mondo virtuale in termini di traffico.

Ci siamo inoltre soffermati sulla descrizione di una realtà aziendale : diversamente dagli e-commerce, Nustox.com è un marketplace ed in quanto tale aggrega più retailer preoccupandosi esclusivamente di pubblicizzare i prodotti degli stessi.

Infine la logistica permette di non avere costi di stock e di garantire al cliente tempistiche di consegna più brevi.

Nella prossima parte “Web scraping per la raccolta dati” vedremo come poter implementare un algoritmo di navigazione simulata in ambiente R tramite l’utilizzo di apposite librerie (Rvest per dati statici e RSelenium per lo scraping dinamico).

Al centro della trasformazione digitale cui stiamo assistendo in questi ultimi anni vi è sostanzialmente internet ed il dato online, il quale risulta sempre più interessante e significativo da studiare.

In ambito marketing, ad esempio, i social network e le piattaforme di analytics offrono quantità enormi di dati utili alla profilazione della clientela e per questo motivo la domanda che ci poniamo in questo capitolo è la seguente :

“Come ottenere il dato ‘intrappolato’ nel web ?”

A questo proposito, una tecnica informatica molto diffusa è il web scraping, utile per recuperare dati da liste statiche così come da liste dinamiche.

Nel secondo caso abbiamo infatti la necessità di indagare strutture complesse e di affidarci ad una tecnica di simulazione di navigazione che sia in grado di emulare il comportamento umano.

L’algoritmo che andremo a descrivere è stato realizzato grazie alla supervisione di Alessandro Patuzzo, responsabile di Data Analytics e del CTO Luca Campana.

3.1) STRUTTURA DEL DATASET

Il dataset in esame è stato ottenuto grazie ad una piattaforma di analytics, la quale permette, attraverso le proprie funzionalità, di monitorare l’andamento delle pagine web, il posizionamento in base alle varie keyword, le interazioni ricevute attraverso i canali social e di ricavare dati statistici riguardanti i propri competitor.

Categoria 1° Livello	Categoria 2° Livello	Dominio	Appartenenza Categoria (%)	Domain Authority (%)	Traffico Mese Totale
Shopping	Abbigliamento	www.hm.com	74	68	70.520.000
Shopping	Sport	www.nike.com	43	64	61.900.000
Shopping	Abbigliamento	www.asos.com	65	67	57.860.000
Shopping	Abbigliamento	www.zara.com	72	71	48.140.000
Shopping	Sport	www.garmin.com	11	57	28.900.000
Shopping	Sport	store.nike.com	36	66	28.100.000
Shopping	Sport	www.rei.com	28	56	19.800.000
Shopping	Sport	www.sportsdirect.com	35	54	18.990.000
Shopping	Sport	www.next.co.uk	11	55	18.600.000
Shopping	Sport	www.adidas.com	22	56	16.500.000

Il campione conta 176 siti dei più importanti brand del settore e per ciascuno mostra le categorie di appartenenza (primo e secondo livello) e le metriche inerenti al traffico organico e paid.

In particolare, troviamo le seguenti variabili (*Tabella 3.1*):

- **Appartenenza alla categoria (%)** : esprime il grado di appartenenza del sito alla categoria specificata.
- **Domain Authority** : è una metrica sviluppata dal tool di Moz¹⁶ e studiata per analizzare l'indicizzazione di un sito nella SERP. Tale punteggio è calcolato sulla base di un algoritmo di machine learning ed essendo su scala logaritmica passare da un valore di 70 a 80 (ad esempio) è molto più difficile rispetto al passare da 10 a 20.¹⁷
In altri termini, la DA può essere considerata come indicatore dello “stato di salute” di un sito web e indicatore del suo rendimento nelle ricerche su Google.
- **Traffico Mese Totale** : traffico paid mensile complessivo in Italia

3.2) WEB CRAWLER

Concentriamoci ora sulla parte applicativa in cui spiegheremo nel dettaglio lo script R utilizzato.

Dopo aver organizzato i dati nel modo corretto e averli preparati per l'analisi, carichiamo come prima cosa nell'environment di R tutte le librerie di cui avremo bisogno.

¹⁶ Moz è una società con sede a Seattle che si occupa di strumenti software dedicati al marketing

¹⁷ “What is domain authority ?” <https://moz.com/learn/seo/domain-authority>

```
library(RSelenium)
library(seleniumPipes)
library(stats)
library(base)
```

Prima di proseguire vediamole brevemente, facendo sempre riferimento alla documentazione ufficiale.

Per automatizzare il browser localmente o collegandosi a un server remoto, Rselenium sfrutta l'API (Application Programming Interface) Selenium Webdriver .

Il grande vantaggio di webdriver è l'automazione del browser : quando navighiamo utilizzando tale applicazione il browser carica tutte le risorse di rete come i file scritti in linguaggio Java, i selettori css in html e le immagini e li esegue tramite il protocollo JSON (Java Script Object Notation) permettendo di testare step-by-step il proprio algoritmo. Uno svantaggio di WebDriver¹⁸ è invece l'allocazione della memoria in quanto, quando si utilizza l'applicazione, l'intero web browser viene caricato sulla memoria del sistema e ciò comporta conseguenze in termini di velocità, risorse di sistema e sicurezza.

Certe pagine web non permettono lo scraping e hanno al loro interno dei sistemi di sicurezza che intercettano gli algoritmi di estrazione dati bloccandoli per un periodo di tempo limitato o in modo definitivo. Anche per questo motivo si raccomanda di effettuare lo scraping responsabilmente nel rispetto delle normative vigenti.

Il pacchetto Rvest permette di manipolare il codice HTML, quindi di selezionare i vari elementi (selettori CSS, classi, attributi, ecc...) di una pagina web e di estrarre informazione. Una volta stabilita la connessione con il server di Selenium, la libreria SeleniumPipes è in grado di fornire istruzioni all'algoritmo, cioè di "scrivere" il comportamento del browser automatico.

¹⁸ "SELENIUM DOCUMENTATION – Web Driver " <https://seleniumhq.github.io/docs/wd.html>

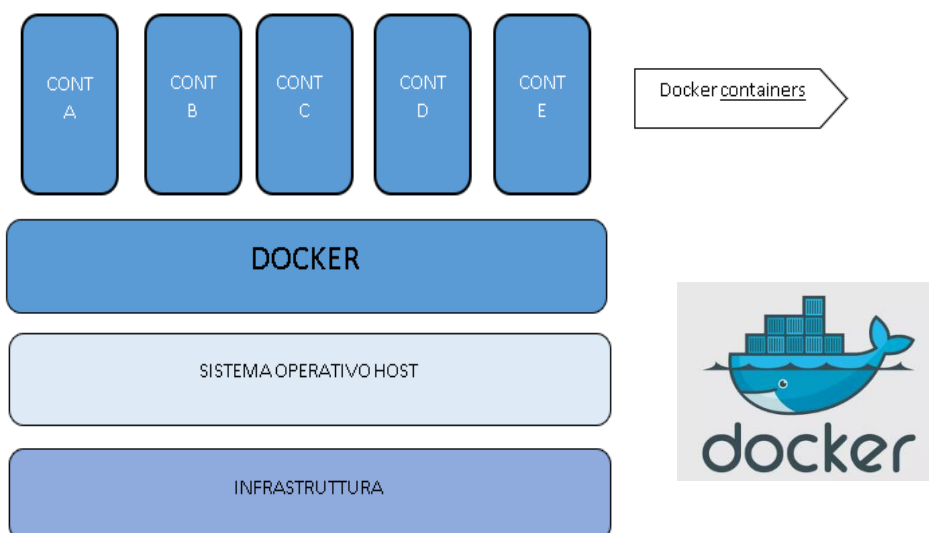
Con le proprie funzioni a disposizione è possibile simulare l'esperienza di navigazione umana, svolgendo in modo del tutto automatizzato le azioni più comuni.

lptools infine contiene una serie di funzioni che agiscono sull'etichetta numerica (IP) che identifica univocamente il dispositivo (in termini più tecnici "host") collegato alla rete informatica. Tale dispositivo utilizza l'Internet Protocol¹⁹ come protocollo di rete.

3.2.1) ELEMENTI PER L'IMPLEMENTAZIONE IN R

I valori della variabile "Traffico Mese Italia" in *Tabella 3.1* sono stati ottenuti, come detto poco sopra, tramite l'utilizzo di un algoritmo di scraping (altresì detto web crawler), il quale è stato progettato grazie anche all'utilizzo dell'applicazione multipiattaforma TightVNC Viewer e a Docker, un progetto open-source inizialmente pensato per il sistema operativo di Linux. In *Figura 3.1* vediamo l'architettura semplificata di Docker, la quale prevede più contenitori aggreganti file (denominati "immagini") necessari all'esecuzione di codici e applicazioni.

Figura 3. 1



¹⁹ Github Documentation – JsonWireProtocol <https://github.com/SeleniumHQ/selenium/wiki/JsonWireProtocol>

Dopo aver inizializzato Docker (*Figura 3.2*)²⁰, al fine di virtualizzare l'esecuzione del browser Chrome utilizziamo l'immagine "selenium/standalone-chrome-debug"²¹ (*Figura 3.3*)

Figura 3. 2

```
MINGW64:/c:/Program Files/DockerToolbox
```

```
docker is configured to use the default machine with IP 192.168.99.100  
For help getting started, check out the docs at https://docs.docker.com
```

Start interactive shell

```
loren@LAPTOP-1TKFL5TH MINGW64 /c:/Program Files/DockerToolbox  
$ █
```

Figura 3. 3

```

MINGW64;C:/Program Files/DockerToolbox
Start interactive shell

loren@LAPTOP-1TKFL5TH MINGW64 /c/Program Files/DockerToolbox
$ docker pull selenium/standalone-chrome-debug
Using default tag: latest
latest: Pulling from selenium/standalone-chrome-debug
84ed72f608f: Pull complete
be2bf1c4a48d: Pull complete
a5bdc6303093: Pull complete
e9055237d68d: Pull complete
695f285c122c: Pull complete
37b3644f34d9: Pull complete
d2736f6d19f5: Pull complete
08c14aa9b728: Pull complete
c5a8c2734444: Pull complete
89e9e85ba652: Pull complete
3e0b6e795d77: Pull complete
a264082708ae: Pull complete
2261986fe222: Pull complete
6d407c1d443b: Pull complete
2de86487068f: Pull complete
d9f7f5c6a822: Pull complete
cb3fa0da4e86: Pull complete
e9d8fcb5608d: Pull complete
642b8a11d69e: Pull complete
bfa78cccccae5: Pull complete
f938ce7b161f: Pull complete
9e0c37861f9c: Pull complete
65eba50d0f11: Pull complete
92be42c4286c: Pull complete
8d3677441583: Pull complete
6f856a6f29f5: Pull complete
88ead391768d: Pull complete
bb6a20ddc8c9: Pull complete
61eb6a5f438e: Pull complete
efc41ea9d193: Pull complete
57803214f3e8: Pull complete
0f78910124cd: Pull complete
7faae97d858: Pull complete
Digest: sha256:29a26eda894a2283f5fff559374d53bab247f5d37587eed35974c56b11228bda
Status: Downloaded newer image for selenium/standalone-chrome-debug:latest

```

²⁰ <https://seleniumhq.github.io/docs/wd.html>

²¹ <https://hub.docker.com/search?q=selenium%20standalone&type=image>

Ora che abbiamo caricato in memoria gli strumenti necessari per la virtualizzazione, rimane da collegarsi al server remoto di Selenium. Basterà eseguire il comando in *Figura 3.4*.

Il termine “port” (tradotto dall’inglese “porta”), in telematica, è lo strumento che permette ad una macchina di collegarsi contemporaneamente con altri calcolatori e di scambiare dati. I pacchetti sono quindi identificati da 4 caratteristiche :

- Indirizzo IP sorgente
- Indirizzo IP di destinazione
- Porta sorgente
- Porta di destinazione

Va sottolineato che la porta di destinazione è univoca, mentre la porta sorgente viene assegnata in modo totalmente casuale dalla macchina. In questo caso il collegamento tra mittente e destinatario è univocamente identificato.

Figura 3. 4

```
loren@LAPTOP-1TKFL5TH MINGW64 /c/Program Files/DockerToolbox
$ docker run -d -p 4445:4444 -p 5901:5900 selenium/standalone-chrome-debug:latest
73f35e7087ab94d71009f0d750067d6a603e05b3203c1d484a33e9e6b783335b
```

New TightVNC infine si occupa di stabilire la connessione, previa autenticazione (*Figura 3.5*)

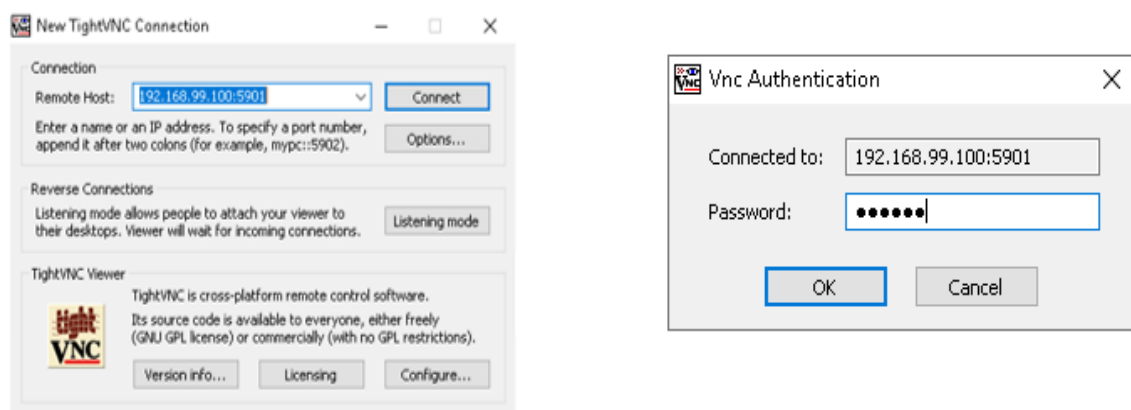
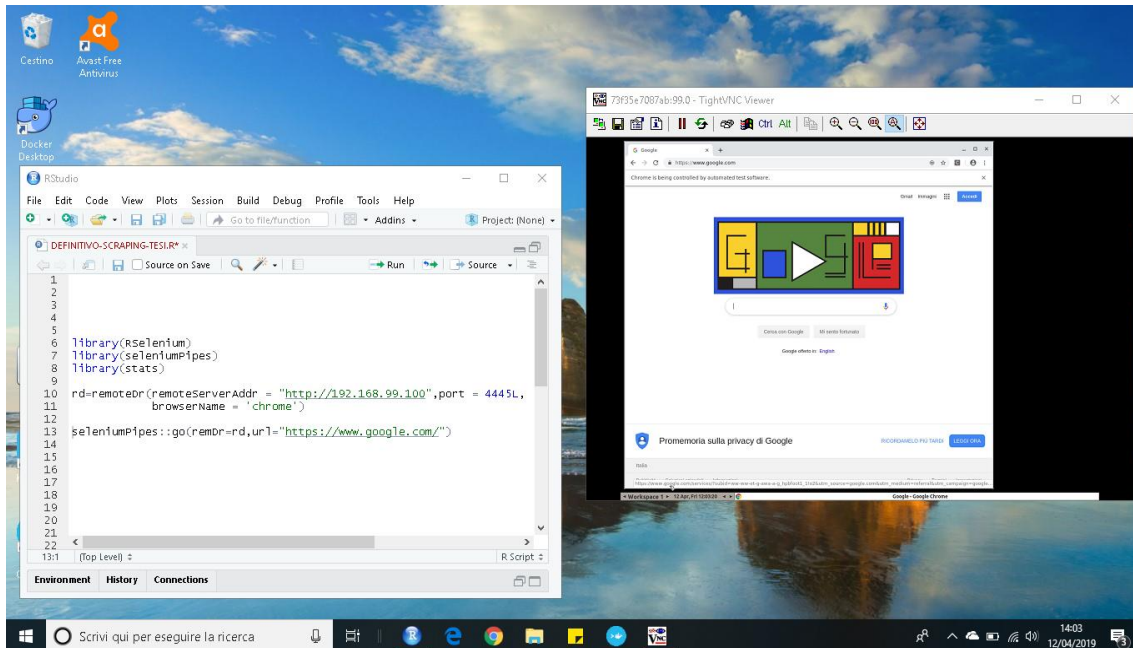


Figura 3. 5

Finalmente possiamo gestire e controllare lo script, testandolo passo dopo passo grazie al server Selenium (Figura 3.6).

Figura 3. 6



Passiamo quindi all'implementazione dell'algoritmo importando le librerie necessarie in R e creando il driver remoto con la funzione `remoteDr()`.

Per una più facile comprensione, commentiamo il codice integrale.

```
rd=remoteDr(remoteServerAddr = 'http://192.168.99.100' , port = 4445L,  
            browserName = 'chrome')
```

Il modo più facile e intuitivo per simulare la navigazione umana “avanti e indietro” è quello di usare un ciclo `for`. Inizializziamo la variabile ‘traffico’ che conterrà le informazioni che stiamo cercando di raccogliere.

```
traffico=vector()
```

```
dominio=c('www.carpisa.it', 'www.vans.it', 'www.bridesire.it', 'www.yamamay.com', 'ww  
w.guess.eu')
```

```
for (i in 1:length(dominio)){
```

Il comando go() permette di collocarci sul motore di ricerca o su una pagina web più generica. In questo caso, per comodità, è Google.it ma si potrebbe partire da qualsiasi altro sito.

Successivamente selezioniamo l'elemento sfruttando il selettore css = "engagementInfo-valueNumber", corrispondente al valore che vogliamo estrarre.

```
seleniumPipes::go(remDr = rd,url = "www.google.it")  
el=seleniumPipes::findElement(remDr = rd,using = "id",value = "js-swSearch-  
input")
```

Alcune pagine web potrebbero non caricarsi subito a causa di diversi fattori come scarsa connessione o versione del browser non aggiornata quindi è buona norma sospendere per un breve periodo di tempo l'esecuzione dell'algoritmo con il comando Sys.sleep. La funzione runif() garantisce invece che l'intervallo temporale di sospensione sia sempre diverso.

```
Sys.sleep(runif(1,0,1.5))
```

```
seleniumPipes::elementSendKeys(webElem = el,dominio[i],key="enter")  
Sys.sleep(runif(1,0,1.5))
```

```
elemento=seleniumPipes::findElement(rd,using = "class name",value =  
"engagementInfo-valueNumber")  
Sys.sleep(runif(1,0,1.5))
```

Memorizziamo nella variabile "traffico" il valore di nostro interesse

```
traffico[i]=seleniumPipes::getElementText(elemento)  
}
```

Infine chiudiamo la sessione di navigazione con il comando deleteSession()

`deleteSession(rd)`

Una volta che le informazioni sono state inserite nella variabile dedicata, è possibile esportarle su un file .csv direttamente da R

S3) SUMMARY

In questo capitolo abbiamo appena visto come collegare Rstudio a Google Chrome e navigare in modo automatico il web. Ricordiamo che la tecnica utilizzata può e deve essere sfruttata nel rispetto dei vincoli posti dai siti esterni e cercando di non danneggiare terze parti.

Nel corso del capitolo successivo esamineremo i concetti teorici ed applicativi per analizzare i dati aziendali di Nustox, cioè cercheremo di prevedere il numero di sessioni paid con la creazione di una regressione lineare multipla.

Vedremo più avanti che tali previsioni contribuiranno a guidare le scelte strategiche aziendali.

REGRESSIONE LINEARE MULTIPLA PER LA STIMA DEL NUMERO DI SESSIONI PAID

Il dataset in esame riporta i valori delle metriche di Analytics registrati giorno per giorno tra il 03 maggio 2018 e 31 marzo 2019. Le variabili presenti, mostrate in *Tabella 4.1* , sono tutte quantitative (gli investimenti in pubblicità sono misurati in euro).

Tabella 4. 1

Giorno	Numero di sessioni complessive	Numero di sessioni naturali	Numero di sessioni a pagamento	Investimenti in pubblicità Google	Investimenti in pubblicità FB
03/05/2018	50	6	44	3,52	6,62
04/05/2018	27	5	22	0,65	8,75
05/05/2018	76	13	63	4,41	8,59
06/05/2018	29	8	21	0,1	6,09
07/05/2018	85	22	63	4	4,14
08/05/2018	55	13	42	0,86	1,23
09/05/2018	371	79	292	23,43	15,83
10/05/2018	311	35	276	16,79	28,2
11/05/2018	340	44	296	8,67	25,81
12/05/2018	333	34	299	10,17	32,93

Data la natura delle variabili e il numero di osservazioni disponibili, si è deciso di effettuare una previsione del numero delle sessioni naturali su base giornaliera, aggregandole poi su base mensile. Otterremo quindi alla fine di questo capitolo una trend line del periodo apr 19 – mar 20.

Come detto poc'anzi il numero delle sessioni a pagamento è fortemente legato al budget che l'azienda è disposta ad investire nel marketing e quindi è difficilmente modellizzabile attraverso un processo stocastico. Si è ritenuto adatto, invece, un modello lineare che mettesse in relazione la variabile di interesse con le quote investite su Google e Facebook.

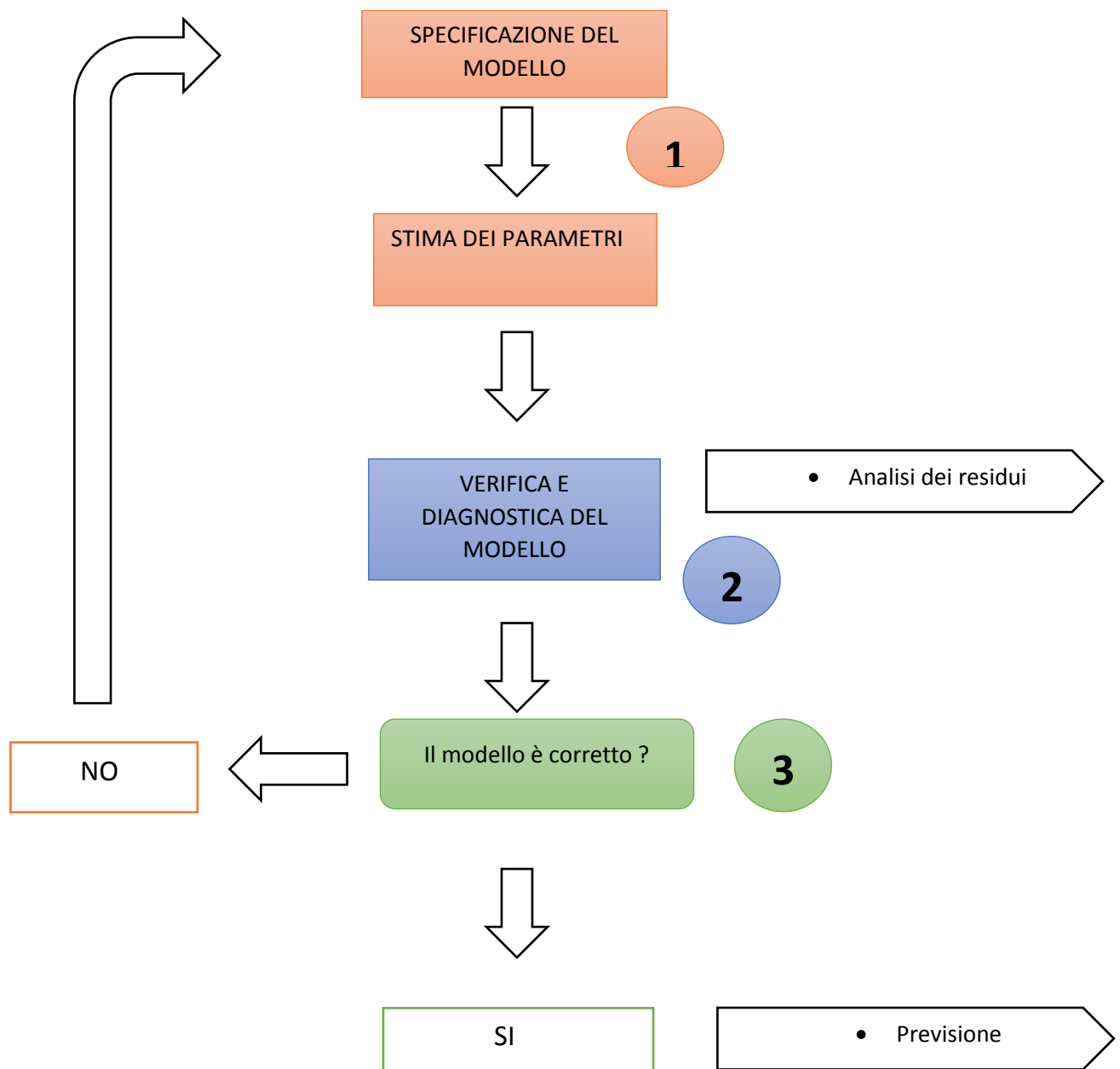
Il modello di regressione lineare multipla rappresenta una generalizzazione del modello lineare semplice e si presta alla modellizzazione di molti fenomeni, in quanto è frequente avere $r > 1$ regressori.

Se da un lato si ottiene, in questo modo, una rappresentazione migliore del fenomeno in esame, dall'altro nascono nuovi problemi come la scelta delle variabili, multicollinearità e test multipli ²²

In questo frangente l'algebra delle matrici risulta molto comoda per specificare la forma analitica e le ipotesi classiche in quanto il modello viene espresso sotto la forma del seguente prodotto matriciale : $Y = X\beta$

Per poter specificare correttamente un modello è necessario seguire alcune fasi, che riportiamo di seguito :

²² "Statistica" , di Domenico Piccolo, Il Mulino, 2010



Le ipotesi classiche²³ sugli errori che andremo ad analizzare in questa parte sono :

1. *Linearità del modello* : $\underline{Y} = \underline{X} \underline{\beta} + \underline{\epsilon}$
2. *Media degli errori nulla* : $E(\underline{\epsilon}) = \underline{0}$
3. *Omoschedasticità* : $\text{Var}(\underline{\epsilon}) = E(\underline{\epsilon}' \underline{\epsilon}) = \sigma^2 \underline{I}_n$
4. *Assenza di multicollinearità* : $r(\underline{X}) = k + 1$

Dove, dato il numero di covariate k e n il numero di osservazioni, \underline{Y} è il vettore risposta di dimensione $n \times 1$, $\underline{\beta}$ vettore coefficienti $(k+1) \times 1$, \underline{X} matrice del disegno di dimensione $n \times (k+1)$ e $\underline{\epsilon}$ è il vettore degli errori $n \times 1$.

La matrice di varianza/covarianza è una matrice diagonale, come indicato nell'ipotesi 3) mentre l'assenza di multicollinearità del punto 4) indica che la matrice del disegno non deve contenere una o più colonne che siano combinazione lineare delle altre (deve essere quindi di rango pieno).

Per la stima dei parametri, invece, viene utilizzato il metodo dei minimi quadrati, ovvero si andrà a cercare quel vettore $\underline{\beta}$ tale per cui la somma degli scarti al quadrato (differenza tra valori osservati e valori stimati) sia minima.

La soluzione unica è data dalla seguente equazione : $\underline{\beta}^* = (\underline{X}'\underline{X})^{-1} \underline{X}' \underline{y}$

4.1) SPECIFICAZIONE DEL MODELLO e STIMA DEI PARAMETRI

Prima di specificare il modello è doveroso osservare i dati : un buon modello infatti non può prescindere da un'organizzazione chiara e pulita del dataset.

Come tutti i modelli statistici, lo scopo della regressione è quello di trovare relazioni tra variabili ; nel nostro caso la forte correlazione tra numero di sessioni e investimenti trova dimostrazione empirica in quanto risulta pari a

²³ "Statistica" , di Domenico Piccolo, Il Mulino, 2010

```
## [1] 0.7150057
```

Il dataset a nostra disposizione, comprendente un campione di numerosità 334, mostra per ogni giorno le quote investite in pubblicità su Facebook e Google e le relative sessioni.

##	day	total.invest	invest.fb	invest.google	session
## 40	2018-05-10	44.99	28.20	16.79	276
## 41	2018-05-11	34.48	25.81	8.67	296
## 42	2018-05-12	43.10	32.93	10.17	299
## 43	2018-05-13	34.17	19.77	14.40	262
## 44	2018-05-14	20.37	6.47	13.90	141
## 45	2018-05-15	20.27	11.36	8.91	141
## 46	2018-05-16	43.47	18.26	25.21	306
## 47	2018-05-17	25.65	15.19	10.46	293
## 48	2018-05-18	36.35	16.15	20.20	324
## 49	2018-05-19	36.40	21.79	14.61	265
## 50	2018-05-20	31.44	17.77	13.67	184

I *Grafico 4.1* e *Grafico 4.2* rivelano che la relazione sessione-investimenti è caratterizzata da una forte variabilità, infatti maggiore è l'investimento e maggiore è la variazione del numero di sessioni che viene registrato.

Una sponsorizzazione su Google, come suggerito dai grafici, genera più sessioni che una pubblicità su Facebook in quanto, a parità di investimento, i livelli delle sessioni sono più alti. Tale affermazione risulta sensata se pensiamo che gli utenti presenti sui social sono meno propensi all'acquisto di un prodotto rispetto a coloro che effettuano query su un motore di ricerca.

In altre parole Facebook agisce sugli strati più alti del funnel di marketing (consumatori con meno engagement), mentre Google attira la clientela più vicina alla conversione.

Grafico 4. 1

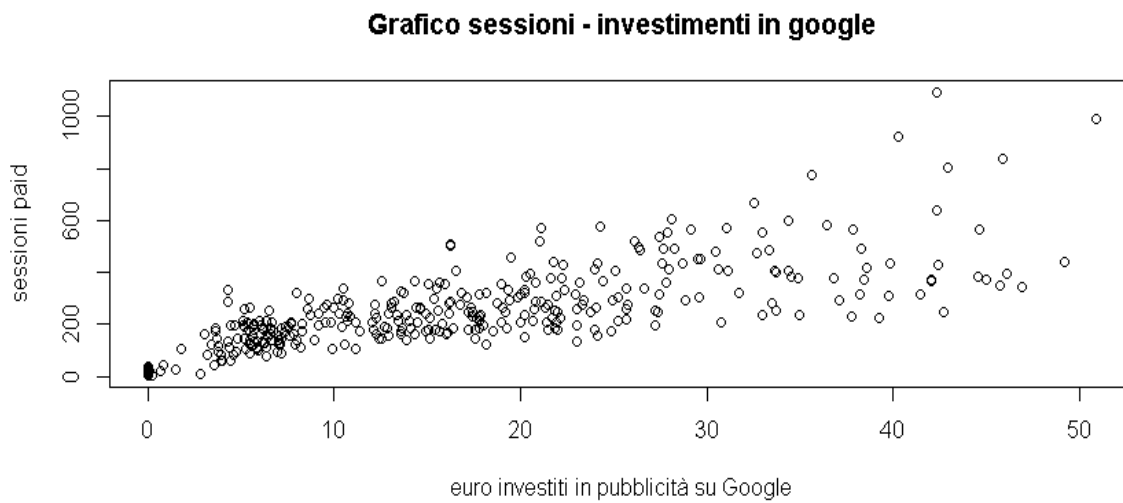
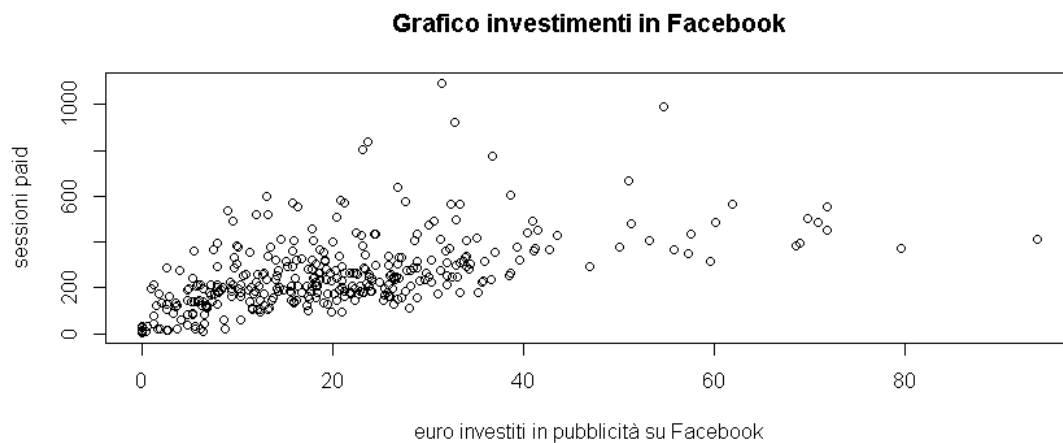


Grafico 4. 2



Tornando all'analisi statistica, l'equazione del modello da stimare è la seguente :

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \varepsilon_i$$

, dove β_0 , β_1 e β_2 sono i coefficienti di regressione, x_1 e x_2 le esplicative (variabili indipendenti) e y la variabile dipendente. In particolare la risposta corrisponde al numero

delle sessioni mentre le covariate corrispondono agli investimenti in adv Google e Facebook.
 ε_i è la componente erratica

In questo caso, per correggere la variabilità, l'applicazione di una trasformata logaritmica appare una scelta ragionevole.

La trasformazione dati attraverso la funzione logaritmo comporta infatti molteplici vantaggi tra cui la normalizzazione della distribuzione quando siamo in presenza di asimmetria positiva e, nel caso di variabili continue, è appunto utile per rendere omogenea la varianza quando essa aumenta al crescere della media.

Tuttavia è noto che tale funzione non ammette valori negativi o nulli e quindi, come conseguenza, dobbiamo prima sincerarci che non vi siano dati di questo tipo. A tal proposito incrementiamo di una costante $C=1$ tutti i valori del dataset.

La trasformazione sarà quindi del tipo : $Y = \log(X + C)$

Il modello è quindi :

$$\log(y) = \beta_0 + \beta_1 * \log(x_1) + \beta_2 * \log(x_2) + \varepsilon_i$$

e la relazione funzionale implementata in R diventa `log(session) ~ log(invest.google) + log(invest.fb)`. e il summary ottenuto viene riportato qui sotto.

```
##
## Call:
## lm(formula = log(session) ~ log(invest.google) + log(invest.fb),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45581 -0.26908  0.01346  0.29872  1.26394
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.99151    0.06489  46.101  < 2e-16 ***
## log(invest.google) 0.70548    0.03887  18.152  < 2e-16 ***
```

```
## log(invest.fb)      0.17479      0.03955      4.419 1.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4183 on 362 degrees of freedom
## Multiple R-squared:  0.8058, Adjusted R-squared:  0.8047
## F-statistic: 750.9 on 2 and 362 DF,  p-value: < 2.2e-16
```

La prima cosa che osserviamo è l'indice di determinazione multipla e la significatività dei parametri.

L'osservazione di queste informazioni è necessaria per accogliere il modello ma non è sufficiente. Se infatti si attribuisse un valore assoluto alla significatività dei coefficienti sarebbe possibile accettare un modello mal specificato oppure rifiutarne uno corretto, commettendo quindi un errore.

L'indice R^2 è uno degli indici più diffusi e utilizzati in ambito statistico per la sua semplicità interpretativa e rappresenta una misura globale di accostamento normalizzata tra 0 e 1 costituita dal rapporto tra la variabilità spiegata dai regressori e la variabilità complessiva.

In altri termini è un indice della bontà del modello.

Nel nostro caso la porzione di variabilità spiegata dal logaritmo dell'advertising in Facebook e Google è di circa 80%.

4.2) VERIFICA E DIAGNOSTICA DEL MODELLO

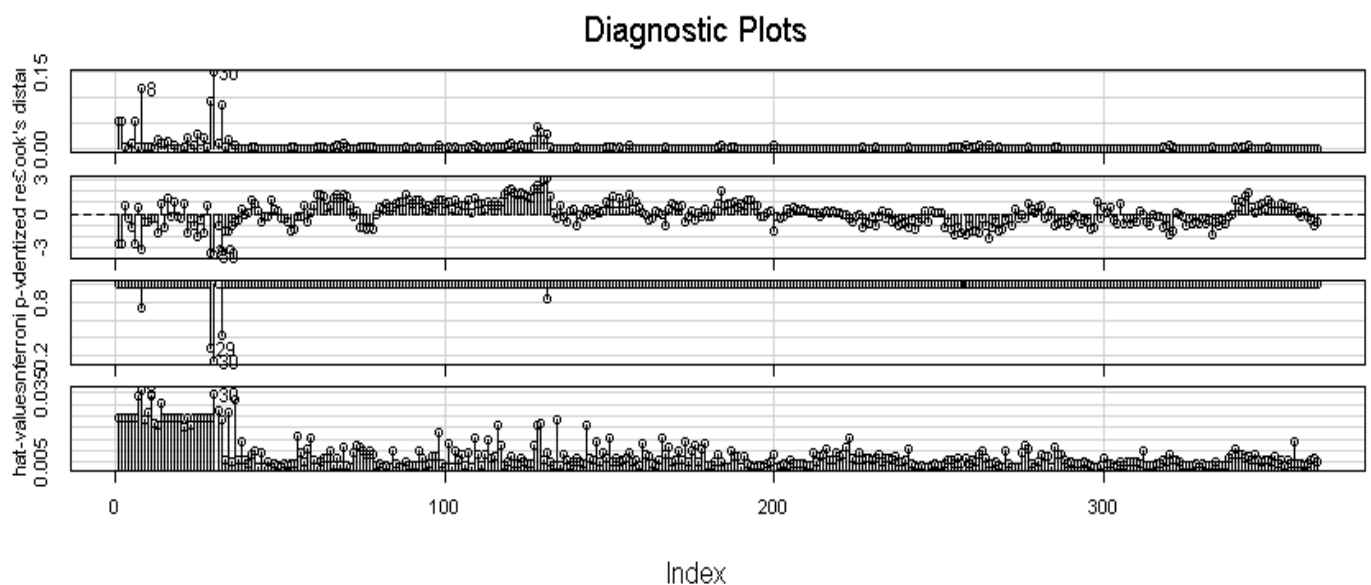
La fase di verifica e diagnostica del modello riguarda le ipotesi classiche sul modello ed errori di varia natura. Possiamo classificare gli errori principalmente in due grandi categorie

- a. *Errori grossolani* : sono errori relativi la raccolta dati o derivanti dal troncamento errato di variabili continue.
- b. *Punti anomali* : sono osservazioni che si distinguono dalla massa di dati o che influenzano il fitting del modello. I punti anomali si classificano a loro volta in :

- Outliers
- Punti di leva
- Punti influenti

Il *Grafico 4.3*, creato in R con la funzione `InfluenceIndexPlot()`, fornisce una panoramica grafica contenente tutte le informazioni riguardanti le osservazioni anomale. E' utile per avere una prima idea dei punti a rischio presenti nel dataset.

Grafico 4. 3



Un'analisi più approfondita con test statistici e strumenti grafici ad hoc rivela che l'osservazione 30 è potenzialmente un punto anomalo e punto influente, mentre l'osservazione 8 è un punto di leva.

b.1) Outliers

Gli outlier sono classificati come osservazioni anomale caratterizzate da un alto residuo.

Per testare la loro presenza è utile la funzione `outlierTest()` di R contenuta nel package "car" il cui output, oltre a mostrare irregolarità nei dati, tiene anche conto della correzione di Bonferroni.

Nel caso di test multipli il metodo di Bonferroni è fondamentale poichè riesce a trattare la propagazione dell'errore di primo tipo (cioè il rifiuto dell'ipotesi nulla quando questa è vera).

In particolare Bonferroni dimostra che utilizzando per i singoli test un livello alfa diviso per la loro numerosità si ottiene la garanzia di avere un errore che ha come limite superiore quello nominale (usualmente 0.05).

Entrando nello specifico, l'ipotesi nulla afferma che la i -esima osservazione è un outlier e si basa su una semplice regola decisionale comprendente i residui studentizzati (jack-knifed) e la distribuzione T.

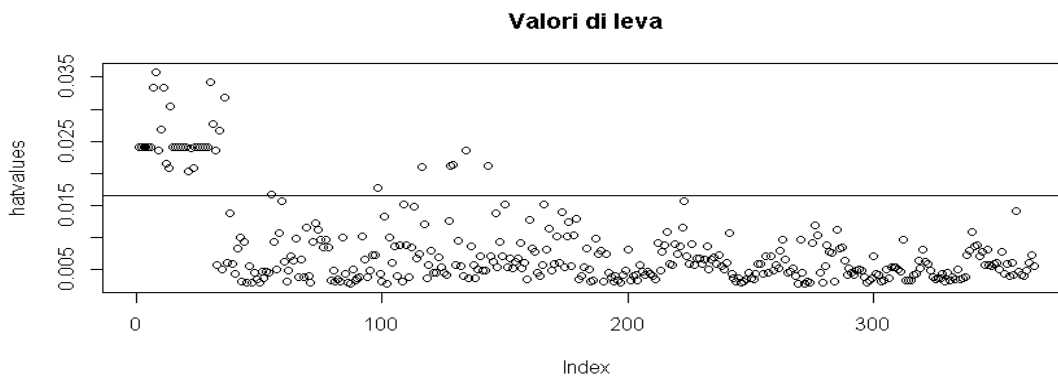
Il risultato del test indica la 30esima osservazione come un potenziale outlier.

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferonni p
## 30 -3.598933      0.00036413      0.13291
```

b.2) Punti di leva

I valori di leva, in ambito statistico, sono quei punti non allineati con i valori fittati dal modello.

Il *Grafico 4.4*, ottenuto in R grazie al package “stats” e alla funzione `hatvalues()`, evidenzia le osservazioni che si trovano al di sopra della soglia (pari a due volte la media dei leverage) e che quindi andrebbero analizzate attentamente. Risulta dall'analisi che il punto 8 è di leva e pertanto andrà tenuto in considerazione nel seguito.



```
## [2] "Punto di leva maggiore : "  
## 8
```

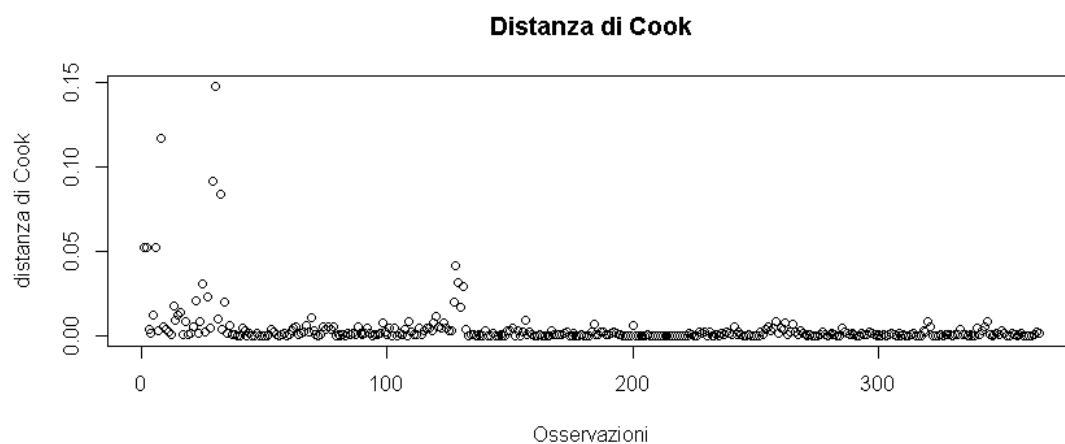
b.3) Punti influenti

I punti influenti sono definiti tali se la loro rimozione dal dataset stravolge il fitting del modello. Diremo quindi che un punto è sicuramente influente se è sia un outlier che un punto di leva.

In generale, però, è buona norma rimuovere l'osservazione caratterizzata da distanza di cook più elevata e ristimare il modello per valutarne l'impatto.

In ambiente R la funzione `cook.distance()` del package "stats" si rivela estremamente importante (*Grafico 1.5*).

Grafico 4.5



```
## [1] "Distanza di Cook maggiore : "  
## 30
```

Decidiamo di rimuovere la 30esima osservazione ristimando il modello e riportando nuovamente il summary.

```
##  
## Call:  
## lm(formula = log(session) ~ log(invest.google) + log(invest.fb),  
##     data = data)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1.41417 -0.27152  0.01458  0.29045  1.23843   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    3.02361    0.06447  46.903  < 2e-16 ***  
## log(invest.google) 0.71989    0.03845  18.723  < 2e-16 ***  
## log(invest.fb)    0.15106    0.03947   3.827 0.000153 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.4116 on 361 degrees of freedom  
## Multiple R-squared:  0.8081, Adjusted R-squared:  0.807   
## F-statistic: 759.9 on 2 and 361 DF,  p-value: < 2.2e-16
```

E' interessante notare che il fitting è migliorato : l'indice R^2 , seppur di poco, è aumentato.

- **Interpretazione dei parametri**

Per interpretare i parametri dobbiamo ricordare l'applicazione della trasformata alle covariate : 0,71989 è infatti l'aumento medio nel log delle sessioni utente a fronte dell'incremento proporzionale – e non assoluto - del log dell'investimento in Google (analogamente per Facebook).

Allora un incremento dell'1% degli investimenti in Google comporterà, fermo restando le altre esplicative, un aumento medio delle sessioni di $(1.01)^{0.71989}$, cioè del 0.7118 %

Analogamente, un incremento di un punto percentuale degli investimenti su Facebook comporterà, fermo restando le restanti esplicative, un aumento medio delle sessioni del 0,1504 %.

Quanto detto dagli strumenti grafici trova quindi conferma empirica : il numero di sessioni derivanti da advertising su Google cresce (in media) più velocemente di quello generato dalla pubblicità su Facebook.

4.3) ANALISI DEI RESIDUI

Verifichiamo ora, come anticipato, le ipotesi classiche sugli errori, come al solito sia per via grafica che tramite opportuni test statistici.

Affronteremo nello specifico :

- i. Normalità dei residui
- ii. Omoschedasticità
- iii. Forma strutturale del modello

Normalità dei residui

Per valutare la normalità dei residui poniamo a confronto la loro distribuzione con quella di una distribuzione Normale standard.

Il Normal QQ plot è la rappresentazione grafica più adatta per fare ciò poichè è una raffigurazione dei quantili teorici della distribuzione Normale standard e di quelli empirici dei residui derivati dalla regressione. Possiamo quindi avere un'intuizione riguardo la normalità se i quantili empirici approssimano la retta rossa del *Grafico 4.6*.

Un altro modo per sincerarci della normalità è il confronto dell'istogramma dei residui , sempre con una distribuzione Normale standard.

Nel nostro caso anche il *Grafico 4.7* sembra confermare tale ipotesi ; procediamo con il test di Shapiro- Wilk (ipotesi nulla H_0 : i dati provengono da una Normale). Per poter interpretare il valore della statistica test W dobbiamo ricordare che è data dal rapporto tra la stima parametrica di σ^2 e la sua stima usuale. Un valore di *0.99419* suggerisce che le due stime sono molto vicine tra loro e fa propendere quindi per l'accettazione dell'ipotesi H_0 .

Grafico 4. 6

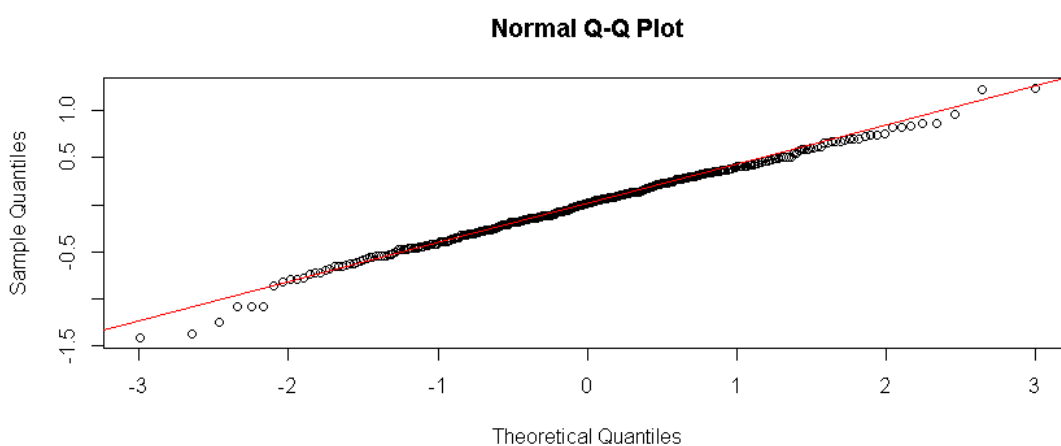
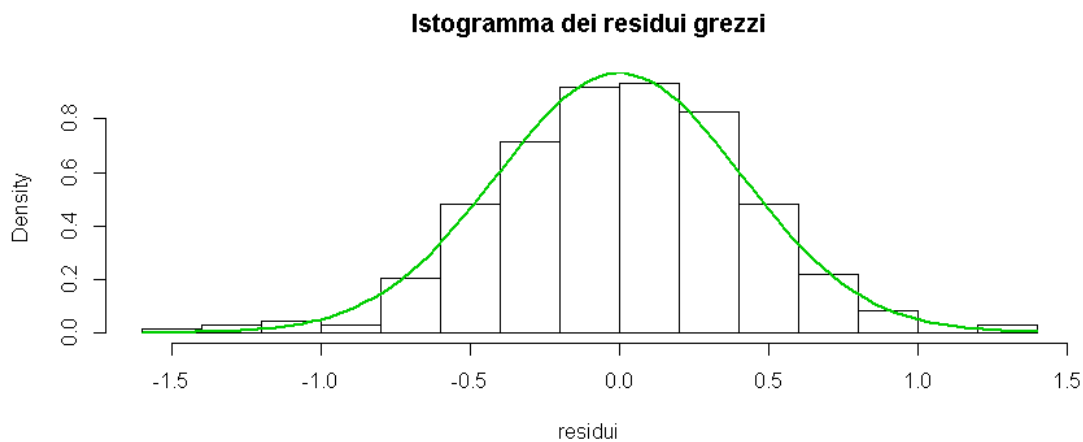


Grafico 4.7



```
##  
## Shapiro-Wilk normality test  
##  
## data:  residui  
## W = 0.99419, p-value = 0.1803
```

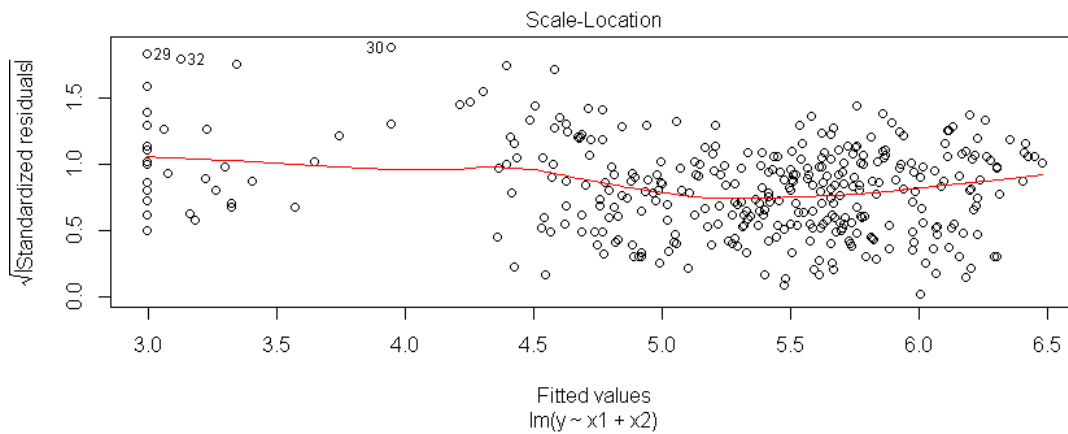
Il p-value sopra la soglia usuale toglie ogni dubbio circa la normalità.

Omoschedasticità

Un'altra ipotesi classica riguarda la variabilità dei residui. La diagnostica grafica di default di R è in questo caso molto utile : lo scale location plot non è altro che uno strumento grafico che mette in relazione i valori fittati dal modello e la radice quadrata dei residui

standardizzati. La linea rossa del [Grafico 4.8](#) indica variabilità costante attraverso i dati e di conseguenza omoschedasticità.

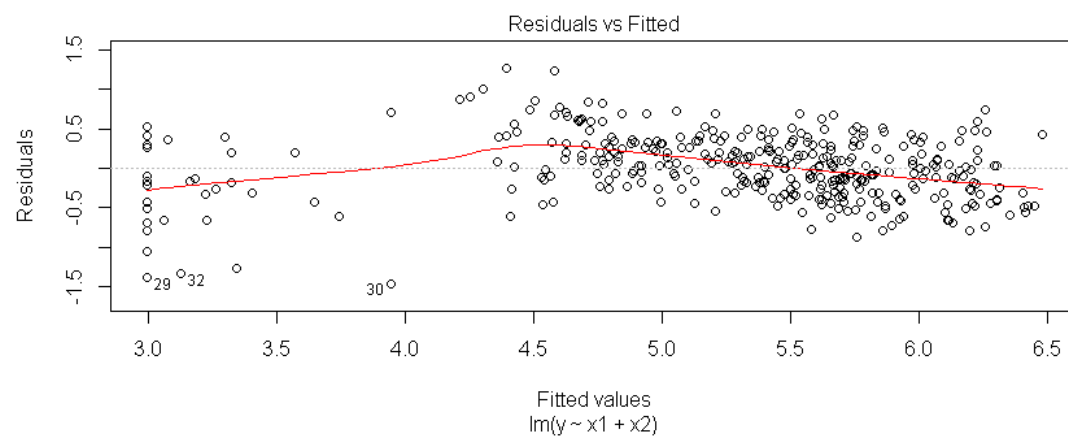
Grafico 4.8



Forma strutturale del modello

Infine vediamo quale strumento grafico utilizzare per la forma strutturale del modello, ovvero la linearità. Ora, la linea rossa del [Grafico 4.9](#) ci dà un'idea dei possibili pattern nei dati.

Grafico 4.92



Il grafico non presenta andamenti sistematici di nessun tipo.

Il modello è corretto ?

Alla luce di quanto detto, possiamo ritenere che il modello specificato sia valido per la spiegazione del fenomeno di nostro interesse.

4.4) UTILIZZO DEL MODELLO : PREVISIONE

Il modello di regressione lineare multipla viene usato generalmente per due fini :

1. Interpretativi
2. Previsivi (puntuale o intervallare)
 - I. Per la risposta media
 - II. Per la singola unità statistica

Ci concentriamo, nello specifico, sulla parte previsiva per la singola unità statistica creando una funzione che, date le quote investite in advertising Google e Facebook, restituisce la previsione per il numero di sessioni a pagamento e il relativo intervallo di confidenza a livello $1-\alpha = 0.95$.

Di seguito il codice R utilizzato.

```
forecast_paid = function(Google_invest=0,Facebook_invest=0){  
  data=data.frame('day' = day,'total invest'=tot+1,'invest fb'=fb+1,'invest goo  
gle'=google+1,'session'=sessions+1)  
  y=log(session)  
  x1=log(invest.google)  
  x2=log(invest.fb)  
  modello = lm(y~x1+x2,data)  
  x.new=data.frame(x1=log(Google_invest),x2=log(Facebook_invest))  
  #nella seguente riga è possibile decidere il livello dell'intervallo di confi  
denza e il #tipo di intervallo (previsivo su una singola unità statistica o c
```

onfidenza)

```
print(exp(predict.lm(modello,x.new,level = .95,interval = "prediction")))
```

#è importante specificare il tipo di intervallo per effettuare una previsione sulla singola unità# statistica piuttosto che sulla risposta media

```
}
```

Use case NUSTOX : l'azienda decide di investire 10.000 euro in pubblicità su Google e 5.000 euro su Facebook. Secondo il nostro modello, la previsione è di 58.567 sessioni.

Per poter comprendere meglio il dato, alla misura puntuale viene affiancato un intervallo di confidenza.

```
forecast_paid(Google_invest = 10000,Facebook_invest = 5000)
```

```
##          fit          lwr          upr
## 1 58566.74 24439.07 140351.6
```

Osservazione: all'inizio dell'elaborato ci eravamo prefissati di ottenere una previsione su base mensile.

Nel proseguo ci occuperemo quindi di distribuire il numero di sessioni previsto su 12 mesi futuri, ipotizzando che esso avrà la stessa distribuzione di frequenza relativa dell'anno passato.

Se da una parte questa ipotesi può risultare vincolante, dall'altra dobbiamo tenere in considerazione che il business di Nustox (ma in generale quello dell'intero settore fashion) è fortemente stagionale.

Il *Grafico 4.10* mostra infatti alcuni picchi di traffico a Novembre (Black Friday), a Marzo (in corrispondenza del caricamento online del nuovo catalogo) e a Luglio (saldi estivi).

Segue che, a meno di eventi straordinari nel mercato, questa ipotesi è del tutto accettabile.

In *Tabella 2* viene riportata la frequenza relativa registrata nella finestra temporale aprile 2018 – marzo 2019 e la relativa distribuzione di frequenza della nostra previsione.

Tabella 4. 2

Mese	Frequenza relativa 2018	Previsione Sessioni Paid 2019/2020
Aprile	0,6%	329
Maggio	6,7%	3909
Giugno	9,9%	5771
Luglio	14,1%	8264
Agosto	5,2%	3031
Settembre	4,7%	2768
Ottobre	9,0%	5271
Novembre	12,0%	7007
Dicembre	6,7%	3934
Gennaio	8,9%	5216
Febbraio	7,2%	4193
Marzo	15,2%	8873
Totale	100,0%	58567

3

Grafico 4. 10



S4) SUMMARY

Ci siamo proposti di effettuare un forecast delle sessioni derivanti da attività di marketing per il periodo aprile 19- marzo 20 e abbiamo appena tracciato una trend line.

Inoltre abbiamo scoperto che questo tipo di sessioni sono facilmente condizionabili dalla quota investita in advertising.

Nel capitolo 5, vedremo invece come stimare la componente naturale. Diversamente da quanto appena fatto, utilizzeremo le tecniche statistiche per lo studio di una serie storica poiché il numero di sessioni spontanee si presta meglio a questo tipo di analisi.

IDENTIFICAZIONE DI UN MODELLO SARIMA PER LA PREVISIONE DEL NUMERO DI SESSIONI NATURAL

Analogamente a quanto visto nel capitolo precedente, in questa parte ci occuperemo della previsione del numero di sessioni naturali.

Diversamente dalla regressione lineare però (dove l'obiettivo era quello di trovare la migliore interpolazione possibile nei dati) l'approccio alle serie storiche è differente. Siccome la singola serie storica è una delle tante possibili realizzazioni di un processo stocastico ed è una serie di variabili casuali indicizzate nel tempo, lo scopo dell'analisi sarà quello di trovare un modello che meglio descrive tale processo.

Ancora una volta le teorie inferenziali saranno di fondamentale importanza.

5.1) PROPRIETA' DEI PROCESSI STOCASTICI

Le caratteristiche dei processi che ci permetteranno di analizzare la maggior parte delle serie sono sostanzialmente :

1. Stazionarietà
 - i. In senso stretto
 - ii. In senso debole
2. Ergodicità

3. Invertibilità

Il concetto di stazionarietà rientra nella stima dei parametri del modello e riguarda l'invarianza rispetto ad una traslazione arbitraria lungo l'asse temporale e di determinate caratteristiche distributive del processo.

La stazionarietà in senso stretto di verifica se : $F(X_{t1}, X_{t2} \dots X_{tn}) = F(X_{t1+k}, X_{t2+k}, \dots, X_{tn+k})$,

Per ogni n-upla $(t1, t2, \dots, tn)$, k e n interi

$F(.)$ è la funzione di distribuzione n-dimensionale del processo e $\{X_{t1}, X_{t2} \dots X_{tn}\}$ un insieme finito di variabili casuali , k e n interi

Diremo invece che il processo è stazionario in senso debole se tutti i momenti congiunti fino all'ordine n esistono e sono invarianti rispetto all'origine temporale (media e varianza costanti e covarianza che dipende solo dall'intervallo di tempo considerate).

Quindi se :

- $E(X_t) = \mu \quad \forall t$
- $E(X_t - \mu)^2 = \gamma(0)$
- $E(X_{ti} - \mu)(X_{tj} - \mu) = \gamma(k) \quad \text{dove } k = ti - tj$

La proprietà di ergodicità è fondamentale per effettuare previsioni e riguarda la possibilità di estendere le caratteristiche di una serie storica osservata a tutte le altre realizzazioni del processo.

Quindi un processo $\{X_t\}$ con $t \in T$ (spazio parametrico) si dice ergodico rispetto alla funzione momento $E[g(X_t)]$ se il momento campionario converge in media quadratica alla funzione momento, cioè se :

$$\frac{1}{N} \sum_{t=1}^N g(X_t) \xrightarrow{mq} E[g(X_t)]$$

Infine l'invarianza è la proprietà che permette di esprimere una variabile casuale X_t come funzione delle precedenti più un errore a_t (rumore bianco stocastico), cioè :

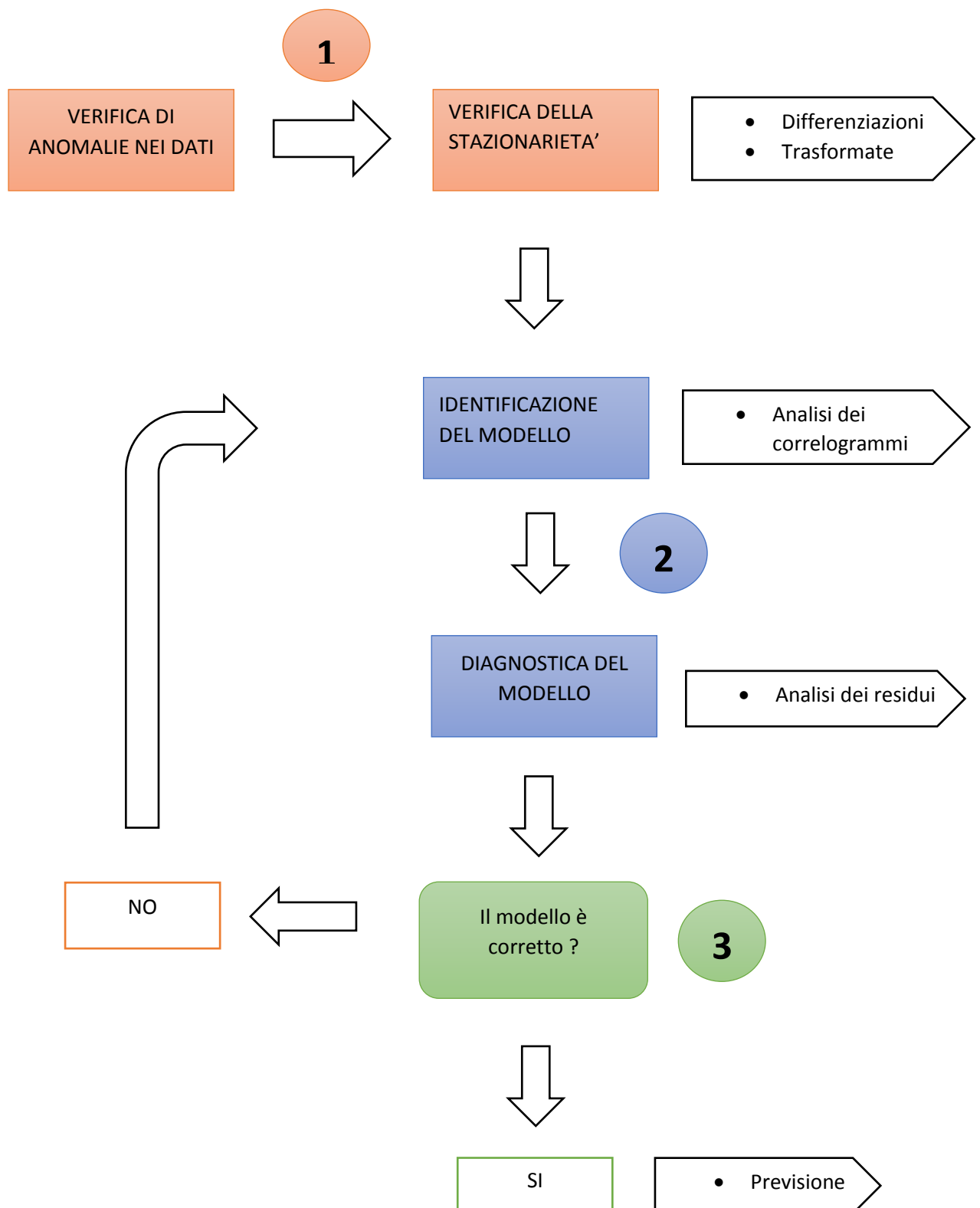
$$X_t = f(X_{t-1}, X_{t-2}, \dots ; a_t)$$

Il grafico sottostante riporta la serie del numero di sessioni naturali .

Grafico 5. 1



Procediamo ora con lo studio che sarà strutturato secondo il metodo di Box e Jenkins.



5.2) VERIFICA DI ANOMALIE NEI DATI & VERIFICA DELLA STAZIONARIETA'

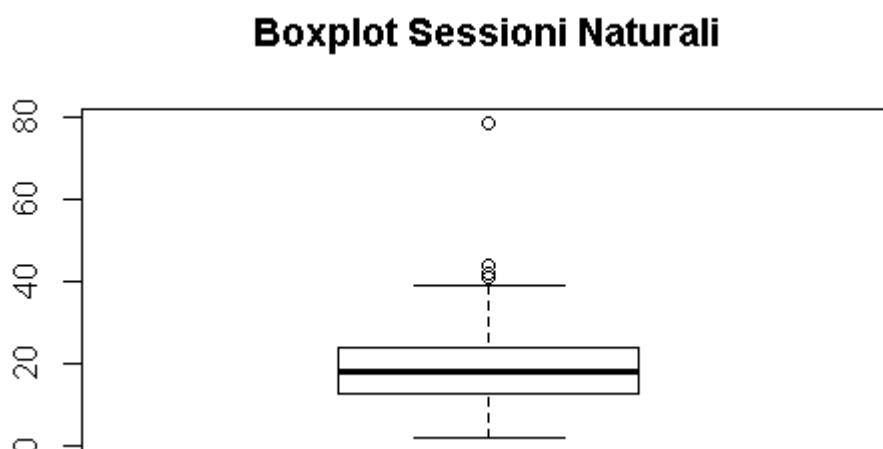
Il metodo di Box e Jenkins prevede nello STEP 1 l'osservazione del grafico della serie e l'adozione di trasformate atte a stabilizzare l'andamento e la variabilità.

In questo ambito le trasformate più utilizzate sono le differenziazioni in quanto permettono di rendere stabile la serie e controllare l'eventuale stagionalità.

Nel caso in esame, osserviamo che vi sono alcuni picchi inusuali. Infatti il boxplot (Grafico 5.2) rivela che essi sono valori in corrispondenza delle date del 9 e 11 maggio 2018 e del 10 gennaio 2019 (tabella qui sotto)

##	giorno	Sessioni Naturali
## 41	2018-05-11	44
## 285	2019-01-10	44
## 39	2018-05-09	79

Grafico 5. 2



Tali anomalie possono essere giustificate da particolari eventi del mercato come la festa della mamma oppure i saldi immediatamente successivi alle feste natalizie, nei quali i clienti sono soliti visitare il sito con maggiore frequenza rispetto ad altri periodi dell'anno. Per questi motivi possiamo ritenere ragionevole l'eliminazione di tali outlier dal dataset, in quanto rappresentano shock esogeni del mercato che potrebbero alterare sensibilmente i risultati ottenuti durante l'analisi.

Il grafico senza outliers è il seguente :

Grafico 5.3



Una volta rimossi gli elementi di disturbo, il grafico mostra un andamento più regolare. Nonostante questo, la serie presenta sempre non stazionarietà : decidiamo quindi di applicare una differenziazione e successivamente di valutare la variabilità.

Ricordiamo che la non stazionarietà di un processo stocastico si manifesta quando una o più radici del polinomio delle componenti autoregressive si trovano sul cerchio dell'unità (quindi, in valore assoluto, pari a 1).

Differenziando per $d=1$, otteniamo un andamento stazionario in media associato al correlogramma sottostante (Grafico 5.4)

Grafico 5. 4

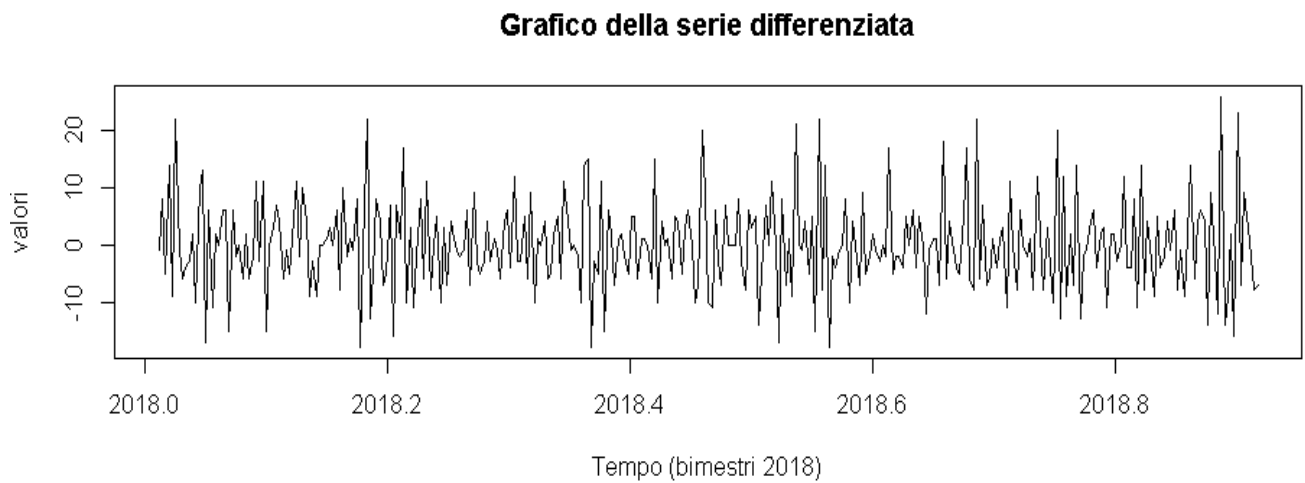
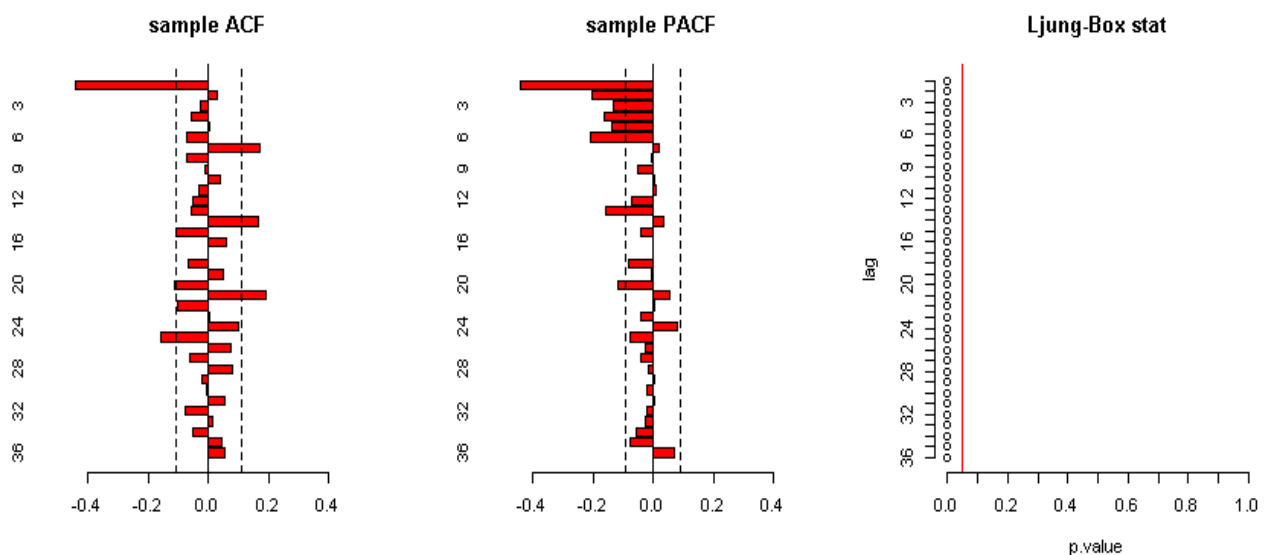


Grafico 5. 5



L'evidenza empirica suggerisce stagionalità, infatti le bande del grafico della funzione di autocorrelazione sono significative al al primo, settimo, quattordicesimo e ventunesimo lag.

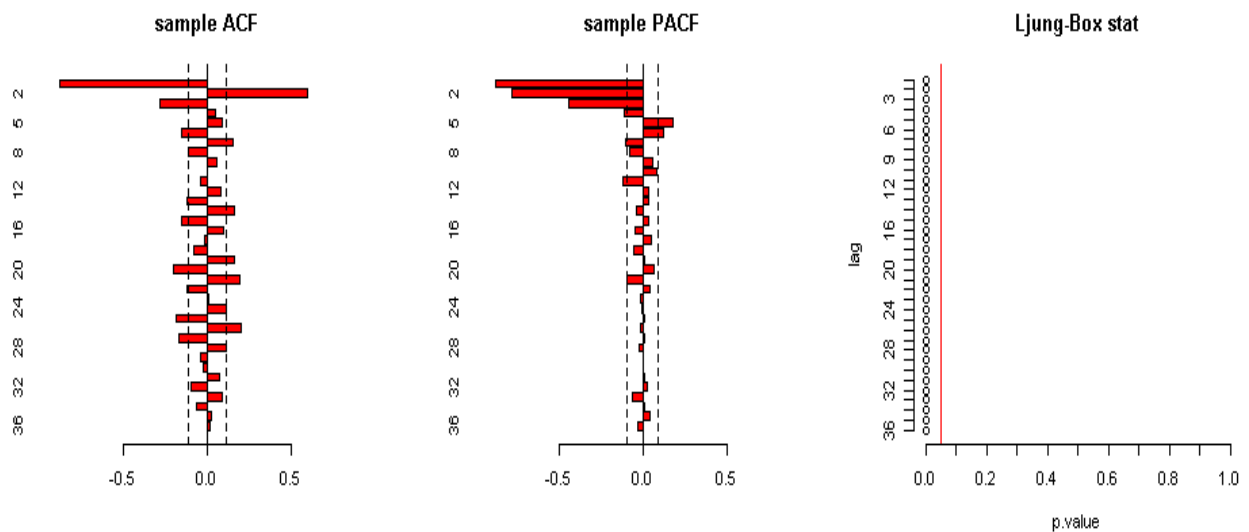
Ciò significa che il numero di sessioni naturali è molto influenzato dal periodo dell'anno, in particolare è "legato" ai mesi di gennaio e luglio, in corrispondenza dei saldi invernali ed estivi.

Continuando l'analisi, a questo punto, la serie sembra essere stazionaria in media ma non in varianza e pertanto risulta opportuna un'altra differenziazione, questa volta stagionale di periodo 7. Guardiamo ancora il grafico e il rispettivo correlogramma :

Grafico 5. 6



Grafico 5. 7



Prima di proseguire con la nostra analisi ripetiamo il test di Dickey-Fuller per trovare un'ulteriore conferma riguardo la stazionarietà.

Testiamo H_0 : presenza di radici unitarie. La statistica test di DF è un numero minore di zero : più è negativo e più "forte" sarà il rifiuto dell'ipotesi nulla.

In R il package “tseries” fornisce la funzione `adf.test` ;

Il risultato del test è riportato di seguito :

```
adf.test(x = serie.storica)

##
## Augmented Dickey-Fuller Test
##
## data:  serie.storica
## Dickey-Fuller = -3.7658, Lag order = 6, p-value = 0.02105
## alternative hypothesis: stationary
```

Il $p\text{-value} = 0.02105 < 0.05$ (usuale soglia di significatività) perciò accettiamo l'ipotesi alternativa di stazionarietà.

5.3) IDENTIFICAZIONE & DIAGNOSTICA DEL MODELLO

La fase di identificazione consiste nell'associare l'autocorrelazione e l'autocorrelazione parziale stimate a quelle teoriche dei modelli ARMA studiati.

Lo STEP 2 prevede l'analisi del correlogramma della serie differenziata e la stima di alcuni modelli che saranno poi confrontati tra loro sulla base della significatività delle rispettive statistiche test e dei valori degli aic.

La *Tabella 5.1* mostra come associare l'andamento dei grafici di ACF e PACF con i vari modelli per la parte non stagionale.²⁴

²⁴ Dispensa di Jack Lucchetti (<http://www2.econ.univpm.it/servizi/hpp/lucchetti/didattica/matvario/procstoc.pdf>)
Wei (2006) Time Series Analysis, Univariate and multivariate
Methods, 2nd edition, Addison-Westley. Hamilton (1994)
Time Series Analysis, Princeton University Press.

Tabella 5. 1

PROCESSO	ACF	PACF
AR(p)	decesce esponenialmente o combinazioni di onde sinusoidali nulla dopo il lag q decesce	nulla dopo il lag p
MA(q)		decesce esponenialmente o combinazioni di onde sinusoidali decesce
ARMA(p,q)		

Analogamente la Tabella 5.2 riporta l'associazione per la componente stagionale (Tabella 2)

Tabella 5. 2

PROCESSO	ACF*	PACF*
AR(P)s	decesce ai lags ks	si annulla dopo il lag Qs
MA(Q)s	si annulla dopo il lag Qs	decesce ai lags ks
ARMA(P,Q _s)s	decesce ai lags ks	decesce ai lags ks
	k=1,2,...	k=1,2,...

Come detto in precedenza, sceglieremo il modello che meglio rappresenta il processo generatore e che contiene meno parametri possibili. Infatti scegliere l'AIC più basso significa penalizzare il modello con più parametri, quindi quello più complesso.

Per la parte non stagionale, in questo caso, la funzione di autocorrelazione si annulla dopo il terzo lag mentre la funzione di autocorrelazione parziale decresce lentamente. Sulla base di tale osservazione pensiamo quindi a modelli misti ARMA(1,1),ARMA(1,2) E ARMA(2,1). Analogamente per la componente stagionale, la funzione di autocorrelazione decresce e quella di autocorrelazione parziale si annulla. Proviamo a stimare perciò un modello autoregressivo di ordine 2,3 o 4.

I valori delle rispettive statistiche test sono i seguenti:

```

arimaest(serie.storica,nsorder = c(1,1,1),sorder = c(2,1,0),vf=T,periodo = 7)$`t statistics`
##      ar1      ma1      sar1      sar2
##  1.917485 -10.185678 -13.602652 -7.703433
arimaest(serie.storica,nsorder = c(1,1,2),sorder = c(3,1,0),vf=T,periodo = 7)$`t statistics`
##      ar1      ma1      ma2      sar1      sar2      sar3
## 13.955161 -17.240615  5.611479 -13.188392 -7.619831 -3.018791
arimaest(serie.storica,nsorder = c(2,1,1),sorder = c(4,1,0),vf=T,periodo = 7)$`t statistics`
##      ar1      ar2      ma1      sar1      sar2      sar3
##  6.538105  4.173972 -106.249791 -13.254766 -7.828130 -3.796880
##      sar4
## -2.550743
arimaest(serie.storica,nsorder = c(2,1,1),sorder = c(2,1,0),vf=T,periodo = 7)$`t statistics`
##      ar1      ar2      ma1      sar1      sar2
##  6.752217  3.858078 -120.702697 -12.618092 -6.897912
arimaest(serie.storica,nsorder = c(1,1,2),sorder = c(3,1,0),vf=T,periodo = 7)$`t statistics`
##      ar1      ma1      ma2      sar1      sar2      sar3
## 13.955161 -17.240615  5.611479 -13.188392 -7.619831 -3.018791
arimaest(serie.storica,nsorder = c(1,1,2),sorder = c(4,1,0),vf=T,periodo = 7)$`t statistics`
##      ar1      ma1      ma2      sar1      sar2      sar3
## 14.721290 -17.852811  6.038742 -13.513892 -8.038971 -4.030277
##      sar4
## -2.685502

```

Escludiamo subito il primo modello, il quale presenta coefficienti non significativi (inferiori, in valore assoluto, al valore del quantile della normale 1.96).

Procediamo ora alla valutazione dei vari modelli tramite criterio dell'AIC :

```

arimaest(serie.storica,nsorder = c(1,1,2),sorder = c(3,1,0),vf=T,periodo = 7)$stime
##
## Call:
## arima(x = dati, order = nsorder, seasonal = list(order = sorder, period = periodo),
##      include.mean = vf)
##
## Coefficients:
##      ar1      ma1      ma2      sar1      sar2      sar3
##      0.8158 -1.4799  0.4799 -0.7129 -0.4678 -0.1643
## s.e.  0.0585  0.0858  0.0855  0.0541  0.0614  0.0544
##
## sigma^2 estimated as 44.9:  log likelihood = -1180.63,  aic = 2375.26

```

```

arimaest(serie.storica,nsorder = c(2,1,1),sorder = c(4,1,0),vf=T,periodo = 7)$stime
##
## Call:
## arima(x = dati, order = nsorder, seasonal = list(order = sorder, period = periodo),
##       include.mean = vf)
##
## Coefficients:
##          ar1      ar2      ma1      sar1      sar2      sar3      sar4
##      0.3445  0.2187 -1.0000 -0.7150 -0.5116 -0.2474 -0.1391
## s.e.  0.0527  0.0524  0.0094  0.0539  0.0654  0.0652  0.0545
##
## sigma^2 estimated as 44.42:  log likelihood = -1179.29,  aic = 2374.58

arimaest(serie.storica,nsorder = c(2,1,1),sorder = c(2,1,0),vf=T,periodo = 7)$stime
##
## Call:
## arima(x = dati, order = nsorder, seasonal = list(order = sorder, period = periodo),
##       include.mean = vf)
##
## Coefficients:
##          ar1      ar2      ma1      sar1      sar2
##      0.3556  0.2017 -1.0000 -0.6388 -0.3540
## s.e.  0.0527  0.0523  0.0083  0.0506  0.0513
##
## sigma^2 estimated as 46.45:  log likelihood = -1186.43,  aic = 2384.86

arimaest(serie.storica,nsorder = c(1,1,2),sorder = c(3,1,0),vf=T,periodo = 7)$stime
##
## Call:
## arima(x = dati, order = nsorder, seasonal = list(order = sorder, period = periodo),
##       include.mean = vf)
##
## Coefficients:
##          ar1      ma1      ma2      sar1      sar2      sar3
##      0.8158 -1.4799  0.4799 -0.7129 -0.4678 -0.1643
## s.e.  0.0585  0.0858  0.0855  0.0541  0.0614  0.0544
##
## sigma^2 estimated as 44.9:  log likelihood = -1180.63,  aic = 2375.26

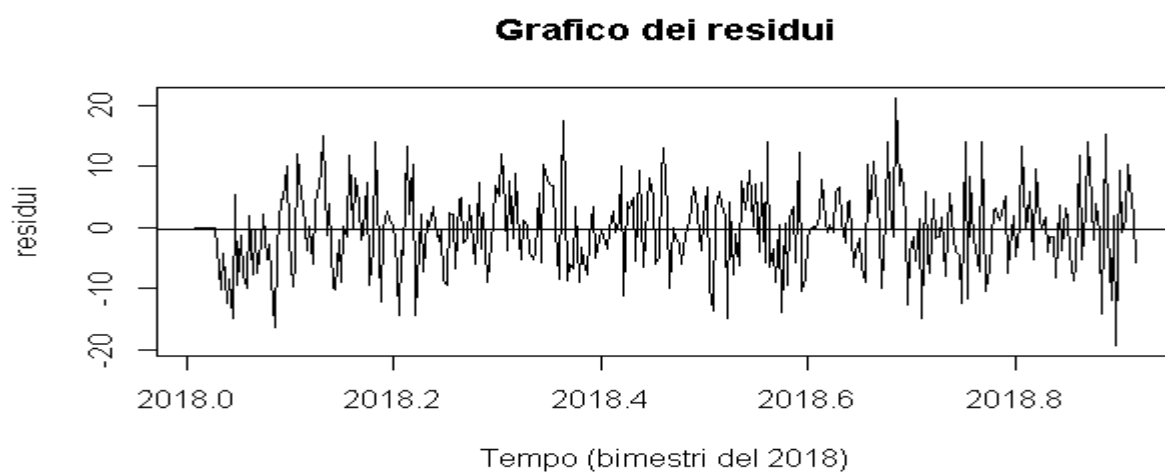
arimaest(serie.storica,nsorder = c(1,1,2),sorder = c(4,1,0),vf=T,periodo = 7)$stime
##
## Call:
## arima(x = dati, order = nsorder, seasonal = list(order = sorder, period = periodo),

```

```
## include.mean = vf)
##
## Coefficients:
##          ar1          ma1          ma2          sar1          sar2          sar3          sar4
##          0.8315      -1.5080      0.5080      -0.7416      -0.5427      -0.2694      -0.1470
## s.e.    0.0565      0.0845      0.0841      0.0549      0.0675      0.0668      0.0547
##
## sigma^2 estimated as 43.91:  log likelihood = -1177.08,  aic = 2370.15
```

Scegliamo quindi l'ultimo modello con $aic = 2370.15$ e passiamo alla verifica delle ipotesi sui residui.

Grafico 5.8



5.4) ANALISI DEI RESIDUI

L'analisi dei residui è volta a verificare le ipotesi di

- Omoschedasticità
- Media nulla
- Incorrelazione
- Normalità

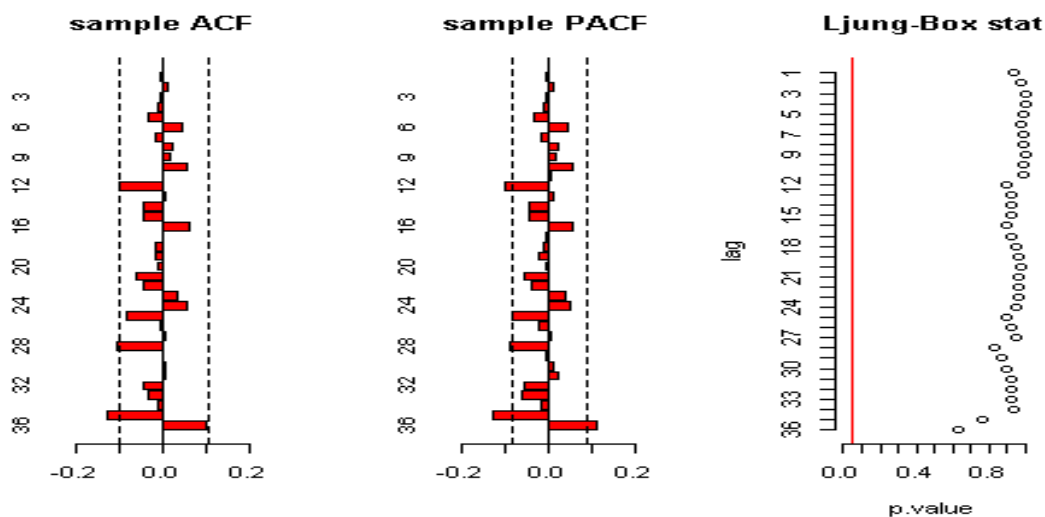
Cioè verificare se lo shock presente nel modello possa essere considerato un processo white-noise o meno.

Possiamo ritenerci abbastanza soddisfatti riguardo l'ipotesi di omoschedasticità (cioè varianza costante per ogni istante temporale) e media quasi nulla.

Valutiamo ora l'incorrelazione osservando il correlogramma dei residui : ci attendiamo che le bande rientrino tutte (o almeno la maggior parte) all'interno delle bande di significatività.

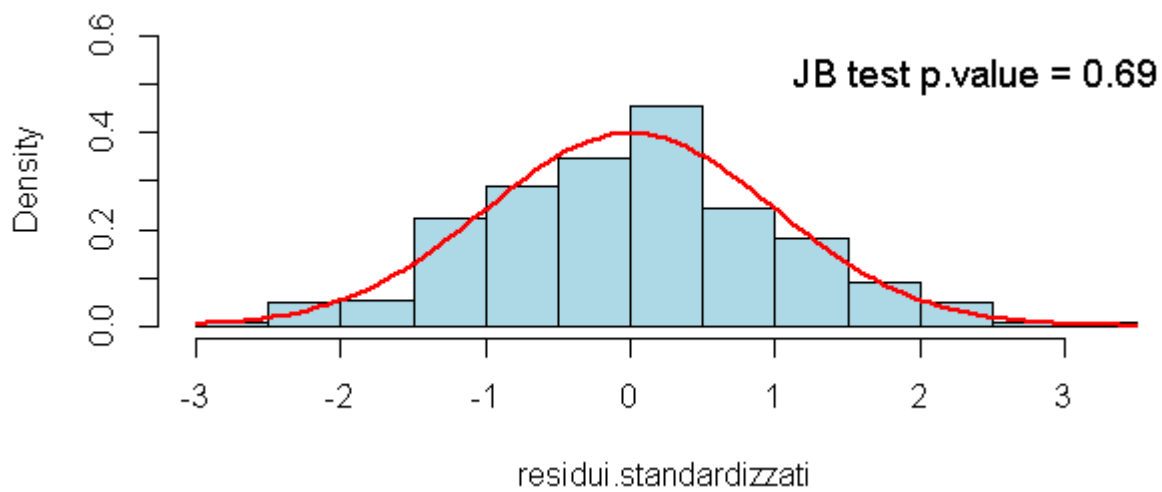
Anche in questo caso possiamo considerarci soddisfatti poichè, come atteso, le barre del grafico rientrano nella regione di significatività. In aggiunta, la statistica test di Ljung-Box conferma l'ipotesi di incorrelazione in quanto tutti i p-value associati sono superiori alla soglia usuale pari a 0.05 (Grafico 5.9)

Grafico 5. 9



Infine verifichiamo la normalità con il test di Jarque-Bera e plottando il grafico dei residui teorici di una normale standard con quelli empirici.

Grafico 5. 10

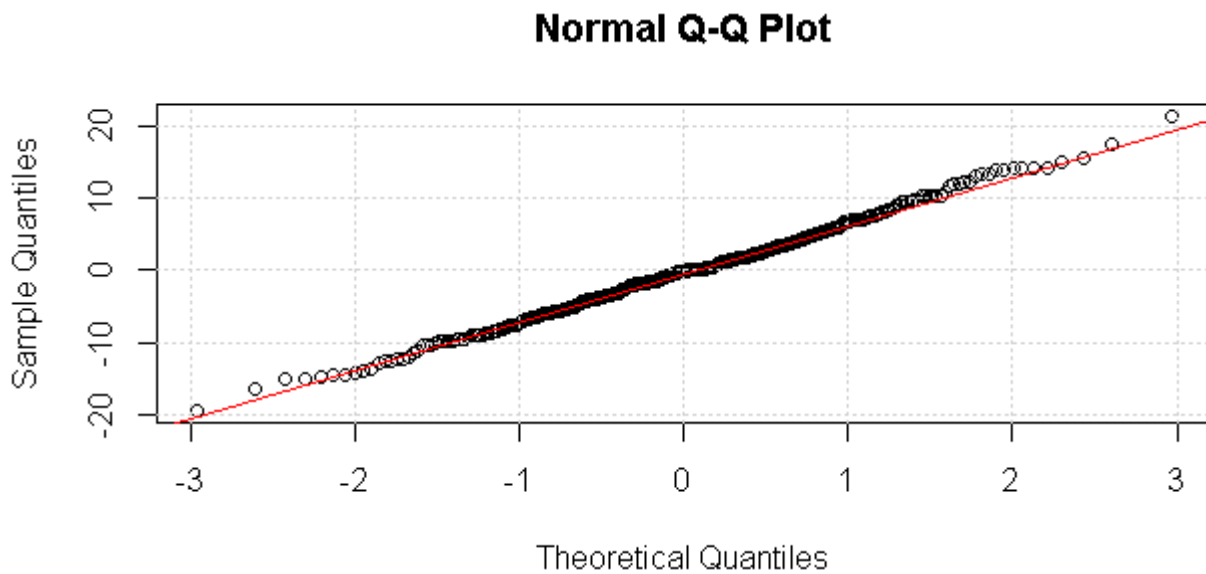


```
## Jarque Bera Test
## data:  res.stand
## X-squared = 7.3845, df = 2, p-value = 0.02492
```

Nonostante il test faccia propendere per il rifiuto dell'ipotesi di normalità ($p.value = 0.025 < 0.05$), sia l'istogramma che il qqnorm(*Grafico 5.11*) evidenziano un buon adattamento alla normale standard.

Queste piccole distorsioni sono dovute, come detto in precedenza, a valori inusuali all'interno del dataset pertanto possiamo supporre che anche la normalità dei residui venga rispettata.

Grafico 5. 11



In conclusione, scegliamo di modellizzare i dati con il seguente modello SARIMA

$$(1 - B)(1 - B^7)(1 - 0.83B)(1 + 0.74B^7 + 0.54B^{14} + 0.27B^{21} + 0.15B^{28})Z_t \\ = (1 - 1.51B + 0.51B^2)a_t$$

5.5) UTILIZZO DEL MODELLO : PREVISIONE

Procediamo con la parte previsiva.

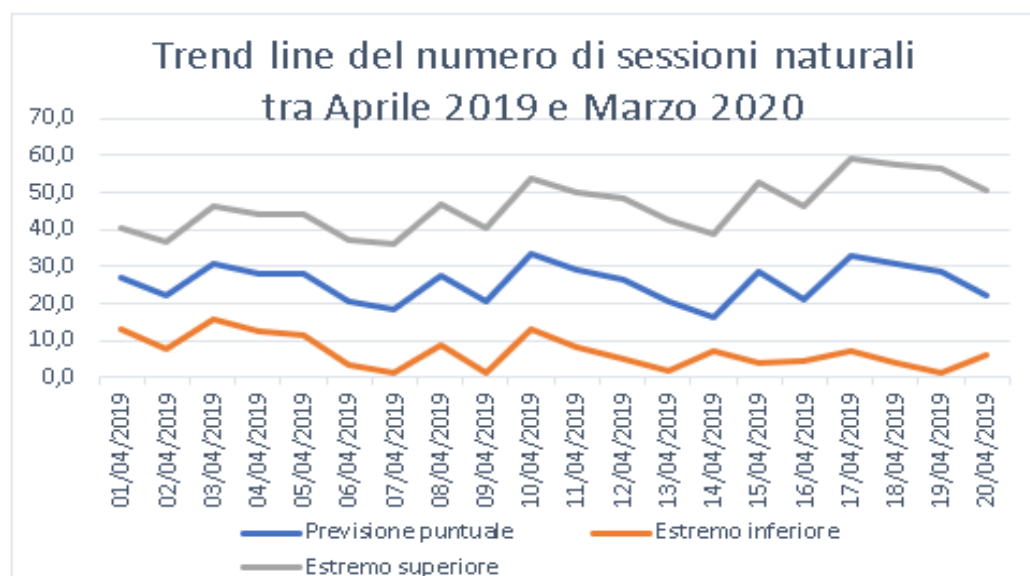
Il nostro scopo, ora, è quello di ottenere previsioni puntuali delle sessioni naturali su base giornaliera tra aprile 19 e marzo 2020, affiancate da un intervallo di confidenza al livello $1-\alpha = 0.95$.

Riportiamo solamente le prime 20 stime future, per non appesantire l'elaborato e il grafico relativo

Tabella 5. 3

Giorno	Previsione puntuale	Estremo inferiore IC (0.95)	Estremo superiore IC (0.95)
01/04/2019	26,8	13,2	40,4
02/04/2019	22,2	7,7	36,7
03/04/2019	30,9	15,8	46,0
04/04/2019	28,2	12,5	43,9
05/04/2019	27,8	11,5	44,1
06/04/2019	20,3	3,4	37,2
07/04/2019	18,5	1,1	35,9
08/04/2019	27,6	8,7	46,5
09/04/2019	20,7	1,0	40,5
10/04/2019	33,3	12,8	53,7
11/04/2019	29,1	8,0	50,1
12/04/2019	26,6	4,9	48,4
13/04/2019	20,4	2,0	42,8
14/04/2019	16,0	7,0	39,0
15/04/2019	28,4	3,9	52,8
16/04/2019	21,1	4,2	46,4
17/04/2019	33,0	6,9	59,1
18/04/2019	30,9	4,1	57,8
19/04/2019	28,8	1,2	56,3
20/04/2019	22,1	6,2	50,4

Grafico 5. 12

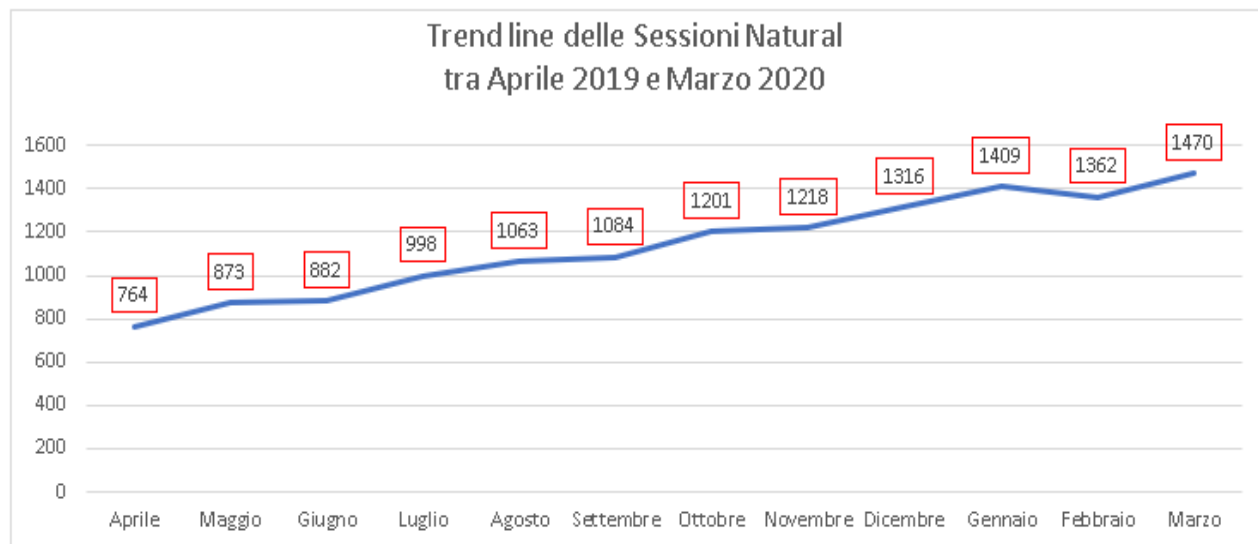


A questo punto, analogamente a quanto fatto nel capitolo precedente, distribuiamo il totale previsto sui 12 mesi dell'anno ottenendo la *Tabella 4.3*

Tabella 4. 3

Mese	Frequenza relativa	Previsione Sessioni Natural 2019/2020
Aprile	5,6%	764
Maggio	6,4%	873
Giugno	6,5%	882
Luglio	7,3%	998
Agosto	7,8%	1063
Settembre	7,9%	1084
Ottobre	8,8%	1201
Novembre	8,9%	1218
Dicembre	9,7%	1316
Gennaio	10,3%	1409
Febbraio	10,0%	1362
Marzo	10,8%	1470
Totale	100,0%	13642

Grafico 5. 13



S5) SUMMARY

Il quinto capitolo è stato dedicato a qualche cenno teorico riguardante i processi stocastici e al forecast del numero di sessioni naturali tramite l'analisi dei dati storici aziendali.

La trend line ottenuta rivela che il prossimo anno Nustox accrescerà il proprio traffico online nei prossimi 12 mesi.

Il capitolo finale, invece, riporterà il business plan dell'azienda ricavato a partire dalle previsioni appena ottenute. Commenteremo due possibili scenari e qualche indice economico.

BUSINESS PLAN & PREVISIONE DEL NUMERO DI ORDINI DI PRODOTTO

In questa parte conclusiva vedremo come è strutturato il business plan aziendale e cercheremo di capire la logica che permette di acquisire nuova clientela, date le previsioni delle sessioni naturali e a pagamento del Capitolo 4 e Capitolo 5.

Questa fase, di norma, segue lo schema del marketing funnel presentato nel Capitolo 1

Ricordiamo in breve di cosa si tratta :

Il funnel è una struttura a imbuto che descrive come la quantità di traffico generata dal website si ripartisce nei vari step che portano all'acquisizione di nuova clientela.

Nella sua parte più alta troviamo il traffico complessivo generato dal marketplace mentre nello strato inferiore i consumatori che acquistano il prodotto (cioè, in termini tecnici, le conversioni). Negli strati intermedi, invece, i potenziali acquirenti eseguono una serie di azioni come la visualizzazione dei dettagli di un certo prodotto, l'aggiunta dello stesso al carrello oppure l'esecuzione di un ordine.

Quest'idea può essere applicata al traffico online ma anche estesa al conteggio delle sessioni.

6.1) APPROCCI PER LA STIMA DEL FUNNEL DI MARKETING

Nustox in questa fase considera due scenari possibili :

A. Top-down :

L'approccio top-down prevede la stima delle sessioni complessive tramite strumenti statistici analoghi a quelli utilizzati nei capitoli precedenti.

L'idea è quella di stabilire delle quote a priori (in %) per le sessioni naturali e a pagamento e ripartire così la stima delle sessioni overall.

Di conseguenza, tramite percentuali note (fissate in base all'esperienza dei professionisti del settore) , vengono stimate le revenue , le conversioni e altri indici che vedremo in dettaglio nel seguito.

B. Bottom- up :

L'approccio bottom-up calcola la stima delle sessioni complessive sommando le stime delle sessioni naturali e a pagamento.

Conseguentemente vengono ricavate le revenue, conversioni e altri indici

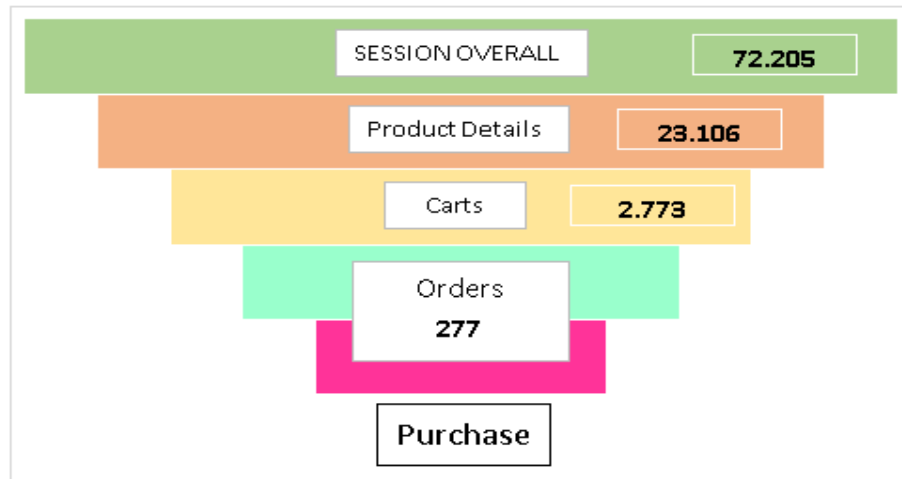
Scenario B : bottom-up

Con riferimento allo *Schema 6.1* le sessioni complessive vengono suddivise negli strati intermedi secondo le percentuali in *Tabella 6.1*

Tabella 6. 1

Layer del Funnel	Quota Sessioni Complessive
Product Details	32%
Carts	12%
Orders	10%

Schema 6. 1



Per avere un quadro generale della situazione, guardiamo la [Tabella 6.2](#) : in blu sono evidenziate le previsioni delle sessioni naturali e a pagamento.

Notiamo che le sessioni naturali costituiscono il 19% del totale e quelle paid il 81% , coerentemente con quanto detto nei capitoli precedenti. Da qui è possibile calcolare diverse KPI come il prezzo medio del prodotto, le revenue, il tasso di crescita del prezzo, il ricavo marginale e altri rapporti riportati qui sotto.

- Conversion Rate / Sessione
- Conversion Rate / Dettaglio del prodotto
- Conversion Rate/ Carrello
- Conversion Rate degli ordini / sessione

Tabella 6. 2

Assumpt & Fall	KPI/Economics	apr-19	mag-19	giu-19	lug-19	ago-19	set-19	ott-19	nov-19	dic-19	gen-20	feb-20	mar-20	Tot 12m
Trough	Session Overall*	1.093	4.782	6.653	9.262	4.094	3.852	6.472	8.225	5.250	6.625	5.555	10.343	72.205
19%	Ow/Natural	764	873	882	998	1.063	1.084	1.201	1.218	1.316	1.409	1.362	1.470	13.640
81%	Ow/Paid	329	3.909	5.771	8.264	3.031	2.768	5.271	7.007	3.934	5.216	4.193	8.873	58.565
32%	Product Details	350	1.530	2.129	2.964	1.310	1.233	2.071	2.632	1.680	2.120	1.778	3.310	23.106
20%	Ow/Natural	70	306	426	593	262	247	414	526	336	424	356	662	4.621
80%	Ow/Paid	280	1.224	1.703	2.371	1.048	986	1.657	2.106	1.344	1.696	1.422	2.648	18.485
12%	Carts	42	184	255	356	157	148	249	316	202	254	213	397	2.773
20%	Ow/Natural	8	37	51	71	31	30	50	63	40	51	43	79	555
80%	Ow/Paid	34	147	204	285	126	118	199	253	161	204	171	318	2.218
10%	Orders	4	18	26	36	16	15	25	32	20	25	21	40	277

Gli indicatori più importanti, in questo caso, sono il prezzo medio del prodotto, del prodotto inserito nel carrello, i loro tassi di crescita e il ROAS.

Il “Return On Advertising Spent” è una KPI fondamentale nel web marketing poiché misura il profitto per ogni euro speso in pubblicità e viene utilizzato per la valutazione di ogni singola campagna, ad esempio di Adwords.

Esso viene calcolato tramite la seguente formula :

$$ROAS (\%) = \frac{Gross\ Margin}{Mkg\ Investment} \times 100$$

E’ doveroso notare che tale indicatore deve essere contestualizzato e risulta informativo se osservato nel lungo periodo piuttosto che nel breve termine.

Un ROAS alto non significa automaticamente che l’azienda è in crescita o che sta avendo profitti alti infatti, se il prodotto sponsorizzato avesse costi di produzione elevati, i profitti sarebbero ugualmente bassi.

Nel caso di Nustox, i costi di produzione sono sostenuti direttamente dalle case di moda e non vi sono costi di stock quindi possiamo affermare che un valore del 180% su 12 mesi è un buon risultato.

6.2 ALCUNE FORMULE UTILI

Dato l'ammontare dell'investimento in web marketing (nel nostro caso 15.000 euro), la quota da investire per l'i-esimo mese viene così calcolato :

$$Mkg\ Investment = \frac{Sessioni\ Paid\ i - esimo\ mese}{Totale\ Sessioni\ Paid\ 12\ mesi} \times (Investimento)$$

Il margine netto del mese i-esimo, è ottenuto tramite la seguente equazione :

$$Gross\ Margin = (B2C\ Revenues\ del\ mese\ i - esimo) \times 22\%$$

CPV ("Cost Per View") indica quante sessioni abbiamo ottenuto con 1 euro speso in pubblicità.

$$cpv = \frac{Mkg\ Investment}{Sessioni\ Paid}$$

Tabella 6. 3

AVG Product Price	320,0	326,4	331,9	331,9	331,9	339,3	347,1	356,4	365,0	375,9	383,8	387,7	349,8
Growth Price		2,00%	1,70%	0,00%	0,00%	2,20%	2,30%	2,70%	2,40%	3,00%	2,10%	1,00%	1,76%
Item per Cart	1,1	1,1	1,1	1,1	1,1	1,2	1,2	1,3	1,3	1,3	1,4	1,4	1,2
Growth Price		2,40%	2,04%	0,00%	0,00%	2,64%	2,76%	3,24%	2,88%	3,60%	2,52%	1,20%	2,12%
B2C Revenues	1.476,7	6.751,6	9.746,8	13.569,7	5.998,7	5.919,7	10.455,7	14.089,2	9.474,0	12.757,9	11.196,8	21.308,9	122.745,7

Tabella 6. 4

22%	Gross Margin	324,9	1.485,3	2.144,3	2.985,3	1.319,7	1.302,3	2.300,3	3.099,6	2.084,3	2.806,7	2.463,3	4.688,0	27.004,1
CpV		0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
15000	Mkg Investment	84,1	1.001,2	1.478,0	2.116,6	776,4	708,9	1.350,0	1.794,7	1.007,6	1.336,0	1.073,9	2.272,6	15.000,0
	ROAS	386%	148%	145%	141%	170%	184%	170%	173%	207%	210%	229%	206%	180%
	CR Product Detail /Session	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%
	CR Cart/Product Detail	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%
	CR Orders/Cart	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%
	CR Orders/Session	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%

Scenario A : top-down

In questo caso la logica del funnel rimane la stessa ma cambia la stima di partenza: ora viene identificato un modello SARIMA per le sessioni overall e fissate delle quote di comodo per le sessioni naturali e paid, rispettivamente 15% e 85%.

Si ricavano come prima i valori degli strati del funnel e gli indicatori di prestazione tra cui il ROAS.

Le percentuali dei vari strati non cambiano (*Tabella 6.1*)

Tabella 6. 5

Assumpt & Fall Trought	KPI/Economics	apr-19	mag-19	giu-19	lug-19	ago-19	set-19	ott-19	nov-19	dic-19	gen-20	feb-20	mar-20	Tot 12m
	Session Overall*	1134	5065	6598	9904	4389	3878	6512	8984	5326	6732	4365	12577	75.464
15%	Ow/Natural	170	760	990	1.486	658	582	977	1.348	799	1.010	655	1.887	11.320
85%	Ow/Paid	964	4.305	5.608	8.418	3.731	3.296	5.535	7.636	4.527	5.722	3.710	10.690	64.144
32%	Product Details	363	1.621	2.111	3.169	1.404	1.241	2.084	2.875	1.704	2.154	1.397	4.025	24.148
20%	Ow/Natural	73	324	422	634	281	248	417	575	341	431	279	805	4.830
80%	Ow/Paid	290	1.297	1.689	2.535	1.124	993	1.667	2.300	1.363	1.723	1.117	3.220	19.319
12%	Carts	44	194	253	380	169	149	250	345	205	259	168	483	2.898
20%	Ow/Natural	9	39	51	76	34	30	50	69	41	52	34	97	580
80%	Ow/Paid	35	156	203	304	135	119	200	276	164	207	134	386	2.318
10%	Orders	4	19	25	38	17	15	25	34	20	26	17	48	290

Tabella 6. 6

	B2C Revenues	1.532,8	7.150,8	9.666,7	14.510,3	6.430,3	5.959,9	10.520,8	15.389,4	9.611,3	12.963,5	8.798,3	25.911,5	128.445,5
22%	Gross Margin	337,2	1.573,2	2.126,7	3.192,3	1.414,7	1.311,2	2.314,6	3.385,7	2.114,5	2.852,0	1.935,6	5.700,5	28.258,0
	Mkg Investment	246,9	1.102,7	1.436,4	2.156,2	955,5	844,3	1.417,7	1.955,9	1.159,5	1.465,6	950,3	2.738,1	16.428,9
	ROAS	137%	143%	148%	148%	148%	155%	163%	173%	182%	195%	204%	208%	172%
	CR Product Detail /Session	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%	32%
	CR Cart/Product Detail	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%
	CR Orders/Cart	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%
	CR Orders/Session	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%	0,38%

Rispetto alla strategia bottom-up notiamo uno scostamento di 8 punti percentuali nel ROAS.

Due risultati comunque non molto distanti tra loro.

CONCLUSIONE

Nel corso di questo elaborato abbiamo introdotto le tematiche del marketing digitale, il caso di studio Nustox, un'applicazione di web scraping, esposto i concetti teorici ed applicativi per la costruzione di modelli statistici e stilato un piano di business.

Effettuare proiezioni con modelli ad hoc rappresenta, per una startup, un vantaggio competitivo non indifferente poiché una buona allocazione delle risorse è capace di minimizzare le perdite e quindi massimizzare gli utili.

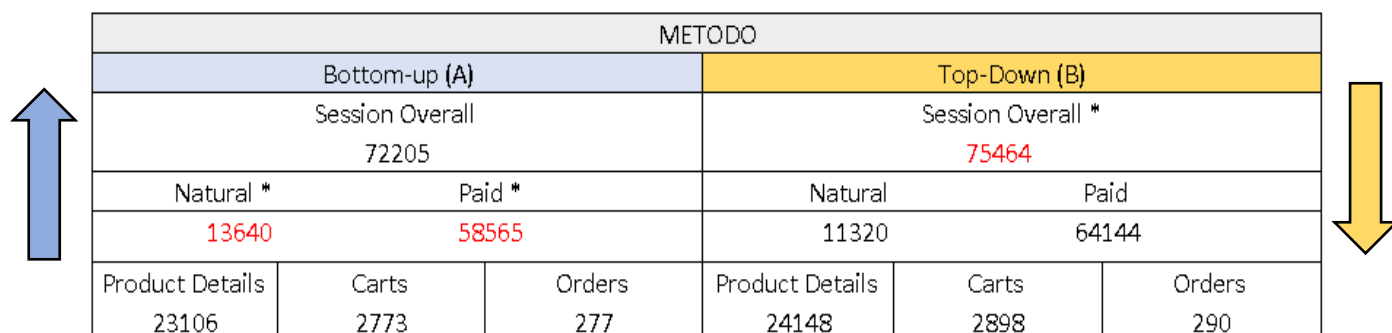
Il modello di regressione lineare multivariata porta alla luce un risultato interessante quale la correlazione positiva tra il numero delle sessioni e l'investimento in pubblicità, in particolare investire su Google.com garantisce un miglior risultato – in termini di sessioni – rispetto all'investimento su Facebook. Sebbene questo dato fosse già intuibile in precedenza (perché la propensione all'acquisto degli utenti è più forte sul motore di ricerca e più debole sul social) , ora la nostra tesi è supportata da un'evidenza empirica.

Al fine di prevedere il quantitativo di ordini di prodotto, sono stati proposti due approcci, rispettivamente denominati top-down e bottom-up.

Sebbene siano in contrasto tra loro, entrambi i metodi portano allo stesso risultato ed il vantaggio del loro utilizzo congiunto risiede nell'avere due termini di confronto in modo tale da avere un'accuratezza maggiore circa i calcoli effettuati in precedenza.

Riportiamo in sintesi i risultati (*Tabella C.1*) evidenziando in rosso le stime empiriche.

Tabella C.1



METODO					
Bottom-up (A)			Top-Down (B)		
Session Overall 72205			Session Overall * 75464		
Natural *		Paid *	Natural		Paid
13640		58565	11320		64144
Product Details 23106	Carts 2773	Orders 277	Product Details 24148	Carts 2898	Orders 290

Ricordiamo che l'approccio bottom-up è così chiamato per il fatto che si parte da due misure distinte (sessioni natural e paid) per poi arrivare ad un unico risultato aggregato (sessioni overall) , mentre la strategia top-down prevede la ripartizione della misura complessiva in due parti secondo percentuali stabilite a priori.

Da qui, i risultati sono stati inseriti nel piano B2C di business aziendale permettendo il calcolo di diverse KPI.

Una raccomandazione per eventuali ricerche future potrebbe essere quella di utilizzare strumenti di calcolo statistico più avanzato o quella di spostare il focus su metriche più interessanti. Ricordiamo che la scelta delle KPI deve essere funzionale al business e pertanto rimane a cura dell'analista.

La diffusione di algoritmi "intelligenti" quali le reti neurali permetterebbero di migliorare notevolmente la precisione della previsione evitando di aggiornare il modello manualmente. In studi futuri sarebbe pertanto interessante approfondire queste tematiche, confrontare i vari risultati e scegliere la tecnica migliore a beneficio delle scelte aziendali.

FONTI :

Capitolo 1 :

- “ L’arte del marketing digitale. Guida per creare strategie e campagne di successo” , Ian Dodson , Milano 2016
- “Google domina il mondo della ricerca”, Il Sole 24 Ore (2017)
<http://www.infodata.ilsole24ore.com/2017/05/10/google-domina-mondo-della-ricerca-soprattutto-console/>
- Google Analytics :
https://support.google.com/analytics/answer/2731565?hl=it&ref_topic=1012046

Capitolo 2 :

- <https://www.statista.com/statistics/694400/turnover-of-italian-fashion-company-giorgio-armani/>
- “L’economia delle aziende di abbigliamento” , Elisa Giacosa , G.Giappichelli EDITORE, 2011
- <http://ecommerce.moda/statistiche-ecommerce/dati-statistiche-ecommerce-moda-fashion-2016-italia/>
- https://www.osservatori.net/it_it/osservatori/comunicati-stampa/la-online-nel-fashion-un-canale-che-fa-tendenza
- Outside In, H.Manning, K. Bodyne, 2012, Forrester Research
- Dati dallo studio “Data Elevates The Customer Experience” in collaborazione con SAS.

- “Apple : storia della mela più sexy del mondo. La lussuria del marchio secondo Steve Jobs” , Marco Giamberini, 2012, p.11
- “Emozione Apple. Fabbricare sogni nel XXI secolo” , A. Dini , Milano , Il Sole 24 ORE S.p.a.,2008 [2007], p.188
- “The style Journal” : <https://www.flipsnack.com/nustox/the-style-journal.html>

Capitolo 3 :

- “What is domain authority ?” <https://moz.com/learn/seo/domain-authority>
- “SELENIUM DOCUMENTATION – Web Driver “ <https://seleniumhq.github.io/docs/wd.html>
- Github Documentation – JsonWireProtocol
- <https://github.com/SeleniumHQ/selenium/wiki/JsonWireProtocol>
- <https://seleniumhq.github.io/docs/wd.html>
- <https://hub.docker.com/search?q=selenium%20standalone&type=image>

Capitolo 4 :

- “Statistica” , di Domenico Piccolo, Il Mulino, 2010

Capitolo 5 :

- Dispensa di Jack Lucchetti
(<http://www2.econ.univpm.it/servizi/hpp/lucchetti/didattica/matvario/procstoc.pdf>)
Wei (2006) Time Series Analysis, Univariate and multivariate
Methods, 2nd edition, Addison-Westley. Hamilton (1994)
Time Series Analysis, Princeton University Press.