

Cattedra Customer intelligence e logiche di analisi dei big data

**Anticipare e ridurre la propensione alla
cancellazione di una prenotazione alberghiera:
i modelli predittivi di Machine Learning e le
strategie comunicative adottabili**

RELATORE

CORRELATORE

CANDIDATO

Anno Accademico

Abstract

Con il presente elaborato si vuole comprendere se sia possibile prevedere, attraverso la costruzione di un modello predittivo di *machine learning*, le cancellazioni delle prenotazioni alberghiere e se un'ideale comunicazione con il cliente, attivata dalla struttura alberghiera, può diminuire la propensione alla cancellazione.

In particolare, si cerca di realizzare ciò attraverso la costruzione di un modello di *random forest*, creato utilizzando il linguaggio di programmazione R, ed applicato ad un *dataset* contenente circa 119.000 osservazioni riferite a due hotel portoghesi.

La previsione della cancellazione delle prenotazioni è fondamentale nel settore alberghiero, poiché le cancellazioni impattano fortemente sulla previsione della domanda e di conseguenza sui ricavi. Per cercare di evitare una perdita eccessiva, le strutture alberghiere impiegano, infatti, politiche di cancellazioni restrittive e/o politiche di *overbooking* che se da un lato possono limitare le perdite, dall'altro determinano effetti negativi sulla reputazione dell'hotel attraverso una *e-WOM* negativa.

Nonostante ciò, la letteratura non si è mai soffermata molto su questo aspetto, trattando la previsione delle cancellazioni quasi esclusivamente in ambito delle prenotazioni aeree.

È quindi di notevole importanza sviluppare un modello in grado di anticipare una cancellazione in modo da mettere in atto strategie comunicative volte a ridurre la propensione dei clienti alla stessa e fissare, al contempo, un tasso di *overbooking* ottimale.

Nel corso dell'elaborato si cerca, quindi, di rispondere a due domande: se è possibile avere un modello predittivo affidabile in grado di anticipare la volontà di cancellazione del cliente e se l'attivazione di un'efficace comunicazione può attenuare la percentuale della cancellazione.

Nello specifico, dopo aver costruito un modello predittivo accurato ed aver dimostrato che la comunicazione con il cliente può ridurre la propensione dello stesso alla cancellazione, si elabora un ragionamento sui *benefit* e/o sul tipo di comunicazione applicabile in base alla generazione di appartenenza del cliente (*baby boomers*, *Gen X*, *Gen Y* e *Gen Z*) e in base alla diversa categorizzazione, cliente *business* o *leisure*.

Quest'ultimo aspetto rappresenta un fattore di novità in quest'ambito, poiché, non è presente in letteratura l'applicazione di una comunicazione differenziata per cercare di ridurre la propensione alla cancellazione.

Relativamente agli sviluppi futuri è necessario però, limitatamente alle strategie comunicative differenziate, eseguire ulteriori ricerche ed approfondimenti poiché dai dati in possesso non è stato possibile risalire ai dati demografici dei clienti e di conseguenza non è stato possibile dimostrare, attraverso una ricerca quantitativa, l'efficacia delle diverse strategie personalizzate che sono state delineate.

Sommario

INTRODUZIONE	3
CAPITOLO 1	5
CUSTOMER JOURNEY E TOUCHPOINT DEL CLIENTE	6
TRAVEL MARKETING STRATEGY	9
I MICRO-MOMENTI NEL TURISMO	10
IL MARKETING NEL SETTORE TURISTICO ALBERGHIERO.....	16
STRATEGIE ATTUABILI.....	19
L'IMPORTANZA DI UNA CORRETTA GESTIONE DELL'OVERBOOKING	21
AZIONI PER EVITARE LA CANCELLAZIONE	24
CAPITOLO 2	25
LA DEMAND FORECASTING, PERCHÉ È UTILE NEL BUSINESS.....	25
DEMAND AND SALES FORECASTING DIFFERENZE	28
LE METODOLOGIE CLASSICHE DI DEMAND FORECAST	28
I NUOVI METODI DI DEMAND FORECASTING	34
REVENUE MANAGEMENT (RM).....	40
CAPITOLO 3	43
INTRODUZIONE ALL'INTELLIGENZA ARTIFICIALE.....	43
TEST DI TURING.....	44
LA NASCITA DELL'AI E I PRIMI SISTEMI DI INTELLIGENZA ARTIFICIALE	46
IL MACHINE LEARNING	50
LE DIFFERENTI TIPOLOGIE DI MACHINE LEARNING.....	52
<i>Metodi supervisionati</i>	53
<i>Metodi non supervisionati</i>	54
<i>Metodi semi-supervisionati</i>	56
<i>Reinforcement Learning</i>	57
DIFFERENTI METODOLOGIE DI APPRENDIMENTO DELL'AI	57
CAPITOLO 4	60
DESCRIZIONE DEL DATASET.....	60
DATA VISUALIZATION	63
ANALISI DELLA CORRELAZIONE.....	76
PREVISIONE DELLA CANCELLAZIONE DELLA DOMANDA	77
COMUNICARE CON IL CLIENTE	82
ATTUARE STRATEGIE COMUNICATIVE PERSONALIZZATE IN BASE ALLA GENERAZIONE DI APPARTENENZA	84
CONCLUSIONI	88
APPENDICE	90
BIBLIOGRAFIA	98
RIASSUNTO	101

Introduzione

La presente trattazione si occupa di esaminare il settore dei viaggi, in particolare il mondo alberghiero, cercando di capire se esiste una metodologia per prevedere le cancellazioni di una prenotazione e se gli esiti risultano essere affidabili.

In aggiunta, si vuole capire anche se una comunicazione con il cliente attivata dalla struttura alberghiera può diminuire la propensione alla cancellazione. Si elaborerà poi, a fine trattazione, un approfondimento sui vari *benefit* che possono essere offerti al cliente per cercare di ridurre la volontà di quest'ultimo di effettuare la cancellazione. Sarà possibile effettuare ciò poiché, durante l'elaborato, saranno approfondite le caratteristiche proprie del mercato oltre che le caratteristiche e le peculiarità dei micro-momenti tipici del cliente viaggiatore. Prendendo a riferimento l'approccio del marketing generazionale, cioè la segmentazione della clientela in base alla fascia d'età, ho analizzato quali possono essere gli strumenti e le strategie migliori da applicare nel mondo dell'*hospitality* prendendo a riferimento le varie generazioni: *baby boomers*, generazione X, generazione Y ed infine generazione Z. È proprio sulla base di questa segmentazione che ho poi ideato delle possibili strategie personalizzate, per fasce d'età appunto, che possono essere applicate per cercare di ridurre la propensione alla cancellazione dei clienti.

Se, infatti, è stato condotto qualche studio sulla generazione di modelli in grado di prevedere la cancellazione e lo studio dell'effetto della comunicazione, non sono state effettuate particolari ricerche sui diversi *benefit* che possono essere offerti al cliente per cercare di diminuire la sua propensione alla cancellazione.

Ho scelto di trattare questo argomento poiché il mondo dei viaggi mi ha sempre affascinato e ancor di più mi hanno sempre incuriosito le strategie applicabili a questo. È, infatti, un mercato per lo più stagionale che quindi deve esercitare sforzi strategici durante tutto il corso dell'anno, ma i cui risultati ed effetti saranno visibili quasi esclusivamente durante l'arco di una stagione, invernale o estiva che sia.

Inoltre, con riguardo al cliente e ad eccezione di alcuni casi, è anche una delle voci di spesa più importanti durante il corso dell'anno. Basti pensare che secondo un'indagine Federalberghi condotta con riguardo al 2021, la spesa media complessiva che gli italiani hanno dedicato ai viaggi è stata di 876€ a persona¹.

Penso, inoltre, che in un periodo di incertezza e crisi nel mercato ad oggetto, cercare di capire in che modo si possa migliorare la gestione dello stesso, e più nello specifico del *cycle revenue management*, sia di vitale importanza per la salvaguardia di questo settore.

Dopo aver analizzato quali sono le metodologie tradizionali e basate sul *machine learning* per studiare la domanda prima e anticipare la cancellazione poi, ho deciso di provare a costruire dei modelli predittivi basati sul linguaggio di programmazione R.

Ho dapprima provato a costruire un modello di previsione sfruttando l'algoritmo del *decision tree*. Quest'ultimo non ha, però, prodotto risultati soddisfacenti a livello di accuratezza. Così, per cercare di

¹ <https://www.federalberghi.it/comunicati/lestato-degli-italiani.aspx>; ultima consultazione: 01/06/2022

aumentare l'accuratezza, ho scelto di utilizzare l'algoritmo *random forest* che, grazie alla costruzione di una numerosità di alberi decisionali, permette una maggiore precisione nell'analisi.

Dopo varie prove ed operazioni, tra cui una *features selection* ed un'analisi della correlazione, ho costruito il modello definitivo di previsione della cancellazione basato, appunto, sull'algoritmo di *random forest*.

Per la costruzione e l'esecuzione del modello ho adoperato, dividendo il *dataset* in due parti come prevede la teoria, dei dati riferiti a delle prenotazioni di una catena alberghiera relativamente a due differenti hotel dislocati in Portogallo: un *city hotel* ed un *resort hotel*.

Il *dataset* è stato poi utilizzato anche per la scrittura e l'esecuzione di codici aventi ad oggetto grafici *ggplot* che hanno consentito di compiere analisi descrittive, facilmente leggibili e interpretabili, sulle prenotazioni.

L'elaborato è articolato in quattro capitoli, ognuno dei quali ha dei sotto capitoli, più una breve conclusione e un'appendice in cui è riportato il codice di programmazione R. Segue alla conclusione anche un riassunto dell'intero elaborato.

Il primo capitolo presenta, oltre ad un'analisi generale del mercato alberghiero e della *customer journey*, degli approfondimenti riguardo ai micro-momenti specifici del cliente tipo.

Inoltre, fornendo come prova dati provenienti da varie indagini, si delineano quelle che sono le linee guida da seguire per lo sviluppo di una comunicazione efficace da parte della struttura.

In questo capitolo, si presenta anche la categorizzazione della popolazione secondo generazione e si illustrano le sue peculiarità.

Nel secondo capitolo, attraverso lo studio della letteratura precedente e della teoria, si illustrano quali sono le metodologie di previsione della domanda e della sua cancellazione. Nello specifico, si analizzano prima le metodologie maggiormente tradizionali e classiche, per poi passare ad analizzare le metodologie basate su algoritmi di *machine learning*. È, inoltre, presentato il sistema del *revenue management* proprio del settore alberghiero.

Una spiegazione sul *machine learning*, in particolare sulle modalità applicabili e sulla sua differenza rispetto all'intelligenza artificiale e al *deep learning*, è fornita nel capitolo 3. In questo capitolo si analizzano anche le differenti categorizzazioni di *machine learning* effettuate in base alla modalità di apprendimento e di addestramento del modello.

Nel quarto capitolo sono, invece, presentati e commentati i risultati delle analisi condotte, anche prendendo in considerazione gli *output* prodotti da R. Dopo aver esaminato i risultati ho sviluppato delle possibili strategie che possono essere applicate prendendo a riferimento la divisione generazionale illustrata nel capitolo 1.

Come si vedrà nella parte finale del lavoro di tesi è possibile, da parte del *management* della struttura e attraverso l'uso di modelli predittivi, anticipare la cancellazione di una prenotazione alberghiera. Si potrà, inoltre, verificare che la presenza di una comunicazione, attivata dallo *staff* della struttura e rivolta al cliente, può ridurre la propensione di quest'ultimo alla cancellazione con risvolti positivi sul *revenue management*.

CAPITOLO 1

Il mercato dell'*hospitality* è un mercato in continua evoluzione e che negli anni ha subito notevoli cambiamenti anche per via dei vari avanzamenti tecnologici.

Mentre, infatti fino a qualche anno fa, le prenotazioni si effettuavano quasi esclusivamente attraverso le agenzie di viaggio ora si può optare tra una più ricca varietà di opzioni.

Certamente occorre effettuare un discorso generazionale, è chiaro che le fasce d'età più giovani opteranno quasi esclusivamente per i canali di prenotazione online, OTA o di proprietà della struttura, mentre le altre fasce d'età potrebbero ancora preferire i canali di prenotazione tradizionali.

Come appena accennato la prenotazione attraverso l'utilizzo di internet è possibile effettuarla sia direttamente sul sito dell'hotel prescelto che sugli OTA cioè gli *Online Travel Agency*.

Questi ultimi sono siti online che permettono la comparazione di varie strutture per una certa località; i più famosi sono ad esempio: Booking, Expedia e Airbnb.

Gli OTA fino a qualche anno fa avevano il quasi monopolio sulle prenotazioni online, ma negli ultimi anni le cose stanno cambiando. Come, infatti, svelato da SiteMinder, una delle principali piattaforme di commercio alberghiero, la scelta dell'utilizzo del sito di proprietà della struttura prescelta è balzato dal quarto posto nel 2018 al secondo nel 2021 (SiteMinder, 2022). Questa classifica è riferita all'Italia ma è in linea con le tendenze globali.

Questo successo è da attribuire sicuramente al maggior numero di mezzi utilizzabili dagli albergatori e dal costo ormai contenuto degli stessi, ma anche una maggior fiducia degli stessi verso i nuovi mezzi tecnologici. La tecnologia non interviene unicamente nella fase di prenotazione ma un cliente consulta internet per lasciarsi ispirare, per cercare la struttura perfetta, per confrontare le strutture e leggere le recensioni sulle stesse.

È quindi di notevole importanza considerare e prestare attenzione alla *Customer Journey*.

Fenomeno delle cancellazioni che vista la sempre più forte propensione a prenotare online e la presenza della formula "cancellazione gratuita" è sempre più frequente.

Il momento storico vissuto in questi ultimi due anni non ha sicuramente migliorato la situazione poiché vista l'incertezza del periodo si tende ad effettuare prenotazioni solo quando queste presentano la garanzia della cancellazione gratuita.

Se, infatti, da un lato la cancellazione gratuita è una caratteristica molto vantaggiosa per il cliente, la stessa cosa non la si può dire per l'albergatore o, in generale, per colui che affitta uno spazio dove soggiornare.

Questo porta, quindi, a una discriminazione di tutte quelle strutture che per policy aziendale decidono di non offrire questa funzionalità.

Per questo motivo, ma anche per garantire, soprattutto in questo periodo storico, alle strutture alberghiere di raggiungere un risultato economico il più possibile positivo, che diventa prioritario mettere in atto politiche di studio e analisi della domanda e della successiva cancellazione.

Posto ciò, la stima della domanda è uno dei passaggi fondamentali nella realizzazione del processo di *revenue management*.

Per capire come stimare la domanda e la successiva cancellazione, non si può prescindere da un attento esame circa i processi di generazione della medesima e dei relativi fattori scatenanti.

Occorre, di conseguenza, studiare il processo di *decision making*.

Capire in modo ottimale quali sono i fattori chiave che portano un soggetto alla prenotazione permetterà successivamente anche di capire perché un soggetto potrebbe voler disdire.

Questo perché, la cancellazione di una sistemazione non implica, necessariamente, che il soggetto non si recherà più in quel luogo ma anche semplicemente che potrebbe aver trovato una sistemazione che meglio si addice alla sua idea di “posto perfetto” o a condizioni migliori.

La politica, sempre più frequente, della cancellazione gratuita e gli strumenti facilmente consultabili online di monitoraggio dei prezzi e delle recensioni, può infatti portare il soggetto, che ha già prenotato presso una struttura, a monitorare la stessa o i suoi diretti competitor in modo da capire se ci dovessero essere cambiamenti.

Una situazione tipica potrebbe essere rappresentata dal caso in cui un soggetto per ottenere il prezzo più conveniente prenota nella struttura con molto anticipo, ma nel periodo successivo rendendosi conto che le recensioni sono cambiate o trovando un’offerta last minute in un'altra struttura potrebbe decidere di disdire la prima prenotazione.

Seppur come delineato, cercare di predire una cancellazione è di vitale importanza per il management di una struttura alberghiera e per il conseguente *revenue management cycle*, non molta letteratura si è soffermata su questo aspetto.

Se, infatti, sono stati sviluppati molteplici modelli così come sono state fatte molte ricerche, per cercare di predire la cancellazione nel mondo dei viaggi aerei la stessa cosa non si può dire per il mondo dell’*hospitality*; anche se questa tendenza si sta modificando negli ultimi anni.

Se, quindi, le cancellazioni impattano fortemente sulla gestione della domanda e sulle successive strategie, come le rigide politiche di cancellazione o l’*overbooking*, è importante sviluppare un modello che prevenga o quanto meno anticipi le stesse.

Il modello di ML, prima di essere applicato necessita di un addestramento che può essere effettuato attraverso i dati del *property management system*. Quest’ultimo è il database posseduto dalla struttura alberghiera in cui è contenuto lo storico delle prenotazioni e delle eventuali cancellazioni.

Se il modello sarà costruito e addestrato correttamente, questo dovrebbe riuscire a prevedere una cancellazione studiando quali sono le caratteristiche alla cui presenza la cancellazione avviene più spesso.

Customer Journey e touchpoint del cliente

Come già detto il mercato dell'*hospitality* si è evoluto e continua ad evolversi negli ultimi anni, vuoi per i cambiamenti tecnologici, di tendenza ma anche semplicemente per l'entrata nel mercato di nuovi e differenti *competitor*.

È quindi estremamente importante cercare di capire nel migliore dei modi quali siano le esigenze del cliente per cercare di instaurare un rapporto con lo stesso che sia forte e durevole nel tempo.

In generale, in tutti i mercati e settori, per capire quale sia il comportamento del cliente e come lo stesso prenda la decisione di acquisto si può utilizzare la *customer journey* del relativo mercato.

La *customer journey* analizza come i clienti interagiscono con l'azienda. Le varie interazioni, *online* ed *offline*, sono chiamate *touchpoint* e sono quindi i mezzi attraverso i quali il cliente entra in contatto con l'azienda.

Le esperienze che si genereranno da queste interazioni saranno poi fondamentali per l'obiettivo finale dell'azienda, cioè l'acquisto del prodotto/servizio da parte del cliente.

In presenza di un'esperienza non positiva è chiaro che il cliente non andrà a scegliere quella specifica azienda preferendone un'altra, viceversa se l'interazione è andata per il meglio. Analizzando il comportamento dei consumatori si potranno poi costruire prodotti ed esperienze differenziate a seconda dei bisogni e dei desideri del cliente.

La *customer journey* si è evoluta nel tempo così come sono evoluti e aumentati a dismisura i vari *touchpoint*.

La *customer journey* originale prevede 5 differenti fasi: *awareness*, *familiarity*, *consideration*, *purchase* e per finire *loyalty* ed è solitamente rappresentata come un imbuto. La forma è dovuta al fatto che mentre molti clienti sono consapevoli dell'esistenza di quel brand/azienda, un numero sicuramente di molto inferiore sarà leale.

Andando nello specifico ad analizzare le differenti fasi:

- *Awareness* indica la fase in cui il cliente riconosce un certo bisogno ed è consapevole di poterlo soddisfare attraverso i prodotti/servizi forniti dalle varie aziende presenti nel mercato;
- *Familiarity*: in questa fase il cliente sempre cercando di soddisfare il suo bisogno, familiarizza con le possibili soluzioni presenti nel mercato. Andrà nello specifico, se quel bisogno è già sorto o per via di un *advertising* impressivo, a cercare la soluzione tra tutti i brand e i prodotti che sono a lui più familiari e quindi più riconoscibili;
- *Consideration*: è la fase in cui il soggetto inizia a reperire informazioni sui prodotti per poi effettuare il confronto. Sono, quindi, di estrema importanza in questa fase la percezione delle caratteristiche del prodotto e il relativo prezzo;
- *Purchase*: rappresenta il primo obiettivo dell'azienda cioè l'acquisto da parte del cliente del proprio prodotto. In questa fase, infatti, il cliente dopo aver effettuato i confronti e le varie valutazioni sui prodotti presenti nel mercato decide quale soddisfi meglio il proprio bisogno e conclude l'acquisto;
- *Loyalty*: questo è il secondo obiettivo dell'azienda cioè fidelizzare la clientela. Fidelizzare la clientela è di enorme importanza poiché un cliente fedele al sorgere nuovamente del bisogno scatenante non ripercorre l'intera *customer journey* ma ricomprerà il brand con il quale ha instaurato la relazione di

fedeltà. È quindi di notevole importanza non solo sviluppare un prodotto con caratteristiche notevoli ma anche offrire un'assistenza post-vendita che permetta al cliente di essere soddisfatto dell'acquisto e dell'esperienza di acquisto vissuta.

Tuttavia, come già accennato questa tipologia di *customer journey* lineare, anche detta a *funnel*, non è più considerata molto valida poiché il consumatore non è più passivo nel rapporto con il brand ma risulta avere un ruolo maggiormente attivo, anche critico, aiutato anche dai numerosi *touchpoint* a sua disposizione.

Nel 2009 McKinsey ha introdotto una nuova tipologia di *customer journey* che, invece di essere rappresentata in forma lineare, è di forma circolare.

Anche in questo caso, l'*input* è fornito dalla presenza di un bisogno che il consumatore dovrà soddisfare ma il *panel* sarà differente.

Abbiamo, anche in questo caso, differenti fasi:

- *Consider*: è la fase di partenza del processo decisionale. Poiché deve soddisfare un bisogno il consumatore cercherà immediatamente nella propria memoria una soluzione a questo e andrà a richiamare tutta una serie di brand o prodotti/servizi che potrebbero aiutarlo. Affinché il brand sia considerato occorre, però, che questo sia presente nel *consideration set* cioè nell'idea che lui ha nella sua mente. In questo caso, quindi, i brand che avranno precedentemente svolto un'attività di marketing di successo avranno il vantaggio di essere immediatamente richiamati nella mente del consumatore.
- *Evaluation*: si ha la fase di valutazione delle alternative e ricerca delle informazioni. In questa fase si hanno cambiamenti anche nel *consideration set*, poiché il consumatore cercando informazioni attraverso i vari *touchpoint* potrebbe venire a conoscenza di un nuovo brand così come rendersi conto che un certo prodotto presente nel *consideration set* non è più di suo gusto. Anche in questa fase si riesce a percepire la bravura del *management*, i vari *touchpoint* devono catturare l'attenzione del cliente e fargli capire l'unicità del proprio prodotto.
- *Buy*: si ha poi la fase dell'acquisto che, come accadeva nella *customer journey* precedentemente analizzata, è soltanto il primo obiettivo del brand.
- *Experience*: da questa fase in poi inizia la parte di *customer journey* successiva all'acquisto e quindi che si basa, oltre che sulla funzionalità del prodotto, anche sulla qualità dei servizi post-vendita offerti o sulla garanzia offerta.

Da questa fase in poi si può originare il *loyalty loop*. Questo *loop* si genererà esclusivamente se l'esperienza che si è avuta è stata positiva. Il *loyalty loop* permette, al sorgere nuovamente del bisogno, di non ripercorrere l'intero processo decisionale ma, vista l'esperienza positiva che si è avuta e la relazione forte che si è instaurata con il brand, di riacquistare il prodotto/servizio. Una cattiva esperienza, invece, implica un'uscita da questo ciclo di fidelizzazione così come una potenziale uscita anche dal *consideration set*. Tuttavia, la possibilità di rientrata nel ciclo di fidelizzazione a seguito di questa cattiva esperienza non è facile poiché per convincere

nuovamente il soggetto occorrerebbe cambiare l'opinione che il soggetto ha nella propria mente e questo, ad esempio, attraverso un riposizionamento del marchio.

La particolarità di questo *loop*, è che spesso questo, a seconda del mercato, potrebbe non esserci. Ci sono, infatti, alcuni mercati in cui la disponibilità e la presenza di molteplici mezzi con cui visualizzare e confrontare i possibili competitor porta i soggetti a non fidelizzarsi ad uno specifico brand poiché sono portati a confrontare e provare le varie alternative offerte dal mercato.

Travel marketing strategy

Come già detto precedentemente, la tecnologia ha cambiato enormemente il mondo dell'*hospitality* ed in particolar modo il modo con cui si effettua una prenotazione.

Se, infatti, fino a qualche anno fa la strategia più comune era quella di recarsi presso un'agenzia di viaggi, ora è possibile visualizzare le varie offerte e prenotare semplicemente attraverso uno *smartphone*.

Tuttavia, quando si fa ricorso ad internet per visualizzare le varie opzioni possibili, si viene "sommersi" da una miriade di alternative diverse che rendono quindi la scelta del consumatore più difficoltosa.

Al contempo questo porta al *management* l'esigenza di dover convincere il consumatore che la propria struttura sia la migliore, sia quella che meglio si addice a lui.

Per far questo, occorre sicuramente ben coordinare gli *advertising*; se, ad esempio, attraverso l'analisi dei *big data* si è a conoscenza che il soggetto viaggerà per lavoro occorrerà mostrare in un Google Ads, ad esempio, la foto di una sala conferenza presente nella struttura e non la foto dello spazio dedicato ai bambini di cui dispone la struttura.

Ma più importante occorre pubblicizzare la propria struttura e contattare i clienti in specifici momenti nei quali essi sono più vulnerabili ad azioni di questo tipo. Si parla in particolare di micro-momenti.

Mentre fino ad una decina di anni fa si utilizzava internet in modo più sporadico attualmente, visti comunque anche i nuovi mezzi tecnologici a disposizione, siamo perennemente connessi.

Ora, infatti, si consulta il web moltissime volte al giorno spesso per motivi professionali o comunque più seri ma il più delle volte anche come passatempo con ricerche futili.

I micro-momenti sono proprio questo, rivolgersi al web, attraverso una varietà differente di dispositivi, affinché questo ci presenti una soluzione ad un bisogno o un desiderio che abbiamo in quel momento.

Ci sono diverse tipologie di micro-momenti, in particolare:

- *I want to know moments*: è la fase in cui l'utente effettua una ricerca online poiché vuole reperire delle informazioni. In questa fase quindi il cliente è ancora in fase esplorativa, e non si ha, per ora, alcuna intenzione di acquisto. Sono proprie di questo momento le ricerche effettuate per soddisfare una curiosità che si può essere scaturita da un qualsiasi stimolo esterno come una pubblicità ma anche la semplice visione di un prodotto.
- *I want to go moments*: questo micro-momento è seguente alla nascita della voglia da parte del soggetto di acquistare un certo prodotto/servizio. Proprio perché è nata in lui quest'intenzione di acquisto,

questo si rivolge al web per capire dove può reperire quello specifico prodotto/servizio nei dintorni di dove si trova.

- *I want to buy moments*: è un micro-momento simile al precedente, poiché anche in questa fase l'utente è convinto di effettuare un certo acquisto e, quindi, ricerca online come e dove concludere l'acquisto.
- *I want to do moments*: l'utente in questo caso vuole imparare a fare qualcosa e per far questo effettua delle ricerche online in grado di aiutarlo in ciò.

Questi sono i quattro principali micro-momenti che sono solitamente individuati; tuttavia, sarebbe possibile indicarne di ulteriori considerato che il ricorso alla ricerca *online* è molto forte e avviene per moltissime altre differenti motivazioni.

Attualmente, è importantissimo cercare di individuare e sfruttare al massimo questi micro-momenti poiché in questi il brand può riuscire a convincere il consumatore ad effettuare l'acquisto e a prediligere il proprio brand piuttosto che quello di un competitor.

I micro-momenti nel turismo

Da uno studio effettuato da Google e Ipsos è emerso che il 56% di coloro che hanno effettuato delle ricerche online da smartphone ha preso in considerazione di acquistare un brand che normalmente non avrebbe considerato, grazie alle informazioni lette durante la ricerca (Google I. , 2015).

I micro-momenti li si hanno anche quando si parla di programmazione di un viaggio, e nello specifico essi sono:

- 1) *I want to get away moments*: è la fase esplorativa, in cui il soggetto non ha ancora delle idee in mente ben precise ma cerca ispirazione navigando online;
- 2) *Time to take a plan moments*: è la fase della programmazione, in questo caso il soggetto ha stabilito quale sarà la sua meta ma deve ancora organizzare e pianificare il tutto. In questo micro-momento deciderà ad esempio le date in cui partire, il posto in cui soggiornare ed anche come raggiungere la località designata;
- 3) *Let's book it moments*: dopo aver programmato tutto nel precedente micro-momento si passa alla fase di prenotazione vera e propria
- 4) *Can't wait to explore moments*: questo micro-momento, diversamente dagli altri, lo si ha quando l'esperienza del viaggio è già iniziata. È la fase in cui il soggetto vive il viaggio.

Questi momenti corrispondono a differenti motivazioni che spingono il soggetto ad effettuare la ricerca online. Il consumatore deve essere accompagnato nel viaggio tra i differenti punti, esso nel primo momento vorrà essere ispirato e informato, nel secondo vorrà programmare, nel terzo prenotare e nell'ultimo condividere. In tutti questi micro-momenti è importante che il brand sia presente e soprattutto si "renda utile". In questi momenti, infatti, il soggetto vorrà anche essere aiutato e di conseguenza la presenza di informazioni ai suoi dubbi è un passo fondamentale per ottenere la fiducia, e la probabile conversione, del soggetto.

Un esempio lampante di questo è il brand alberghiero Red Roof Inn; questo sfruttando il fatto che moltissimi voli vengono cancellati e posticipati è andato incontro al bisogno del consumatore in particolare nel micro-momento *let's book it*. Incrociando infatti i dati delle cancellazioni degli aerei, la strategia del brand prevedeva di mostrare ai soggetti una *ads* contenente il nome del proprio hotel e la vicinanza dall'aeroporto. Questa strategia ha sicuramente giovato al brand, che ha guadagnato in *share of voice* e CTR (Association, 2014), e lo ha fatto cercando di risolvere un problema che affliggeva il consumatore in uno specifico micro-momento. In questo modo il brand è, inoltre, riuscito a creare anche una relazione, fin dall'inizio forte con il proprio consumatore, poiché lo stesso si è sentito aiutato dalla catena alberghiera.

Ma ciò che sta cambiando è anche il mezzo attraverso il quale si percorre questa nuova *customer journey* caratterizzata dai sopra citati micro-momenti. Se, infatti, fino a pochi anni fa la prenotazione, quando avveniva per mezzo online, avveniva attraverso mezzi *desktop* attualmente il mezzo maggiormente utilizzato è lo *smartphone*. Quando, infatti, sentiamo l'impulso di vivere uno dei quattro micro-momenti lo strumento più facile e più a portata di mano è quasi sempre il telefono e di conseguenza utilizziamo questo mezzo per effettuare le ricerche online così come per prenotare.

In una ricerca effettuata prendendo in considerazione più di un migliaio di viaggiatori provenienti da tutto il mondo è emerso che il 77% dei viaggiatori *business* e il 53% dei viaggiatori *leisure* hanno effettuato la propria prenotazione attraverso lo *smartphone* (Digital, 2019).

Questo a dimostrazione proprio che è importante per i brand alberghieri essere presenti online ed essere presenti soprattutto nei micro-momenti esposti precedentemente.

Proprio perché lo strumento dello *smartphone* permette con maggior facilità di cercare informazioni online e anche soprattutto visto che il soggetto effettua queste ricerche anche in momenti di pausa che lo stesso cerca rapidità nella risposta e nella consultazione. Da un'analisi effettuata da Google e avente ad oggetto il periodo maggio 2014 - maggio 2015 si è dimostrato che il tempo trascorso per sessione sui siti di viaggio mobile è sceso del 7% rispetto al periodo precedentemente analizzato.

È, quindi, di notevole importanza anche riprogettare i siti in modo che siano di facile consultazione usando lo *smartphone*, molto spesso infatti i siti, essendo stati progettati inizialmente per una versione *desktop*, non sono molto fruibili da *mobile* ma questo come appena evidenziato dai dati è un fattore molto pregiudicante.

I *touchpoint* che un soggetto ha durante la pianificazione di un viaggio sono molteplici, e possono dunque anche provenire da differenti mezzi: *mobile* o *desktop*. In questo caso, come in un qualsiasi settore di vendita, occorre assicurare un'esperienza omnicanale; un'esperienza che quindi è ottimizzata per tutti i differenti e vari punti di contatto che il soggetto può avere con il brand.

Inoltre, vista la rapidità con cui si vogliono avere dei risultati dal motore di ricerca, per i siti di proprietà dell'hotel, è anche molto importante soddisfare questa rapidità di consultazione, ad esempio, attraverso il generatore di preventivi che permette inserendo esclusivamente la data e la tipologia di stanza di ottenere un primo preventivo.

Se quindi, come già anticipato ad inizio capitolo, i siti di proprietà della struttura alberghiera sono saliti al secondo posto delle modalità maggiormente utilizzate dai clienti per la prenotazione, un primo ostacolo all'utilizzo di questi potrebbe essere riconducibile ad una preferenza dei consumatori.

Dallo stesso studio analizzato poco fa effettuato da Travelport Digital è, infatti, emerso che i consumatori preferiscono effettuare la propria prenotazione attraverso *app* di prenotazione scaricate sul proprio *smartphone*. Nello specifico, per i clienti *leisure* il 49% degli intervistati ha utilizzato il web per effettuare la propria prenotazione mentre il 51% un'*app* di prenotazione; ma la preferenza è molto più marcata per i clienti *business* per i quali le percentuali sono rispettivamente del 70% e del 30% (Digital, 2019).

Le ragioni dietro l'utilizzo delle *app*, sempre per lo stesso studio, sono numerose: possibilità di effettuare *check-in* da *app*, la presenza delle notifiche o ancora la possibilità di utilizzare la stessa senza connessione internet.

Andando nello specifico ad analizzare i singoli micro-momenti, il primo su cui porre l'attenzione è: *I want to get away moments*.

In questa fase, come già anticipato, il soggetto è ancora in una fase esplorativa e spesso non avendo neanche scelto la propria destinazione utilizza i canali *online* come mezzo attraverso il quale trarre un'ispirazione.

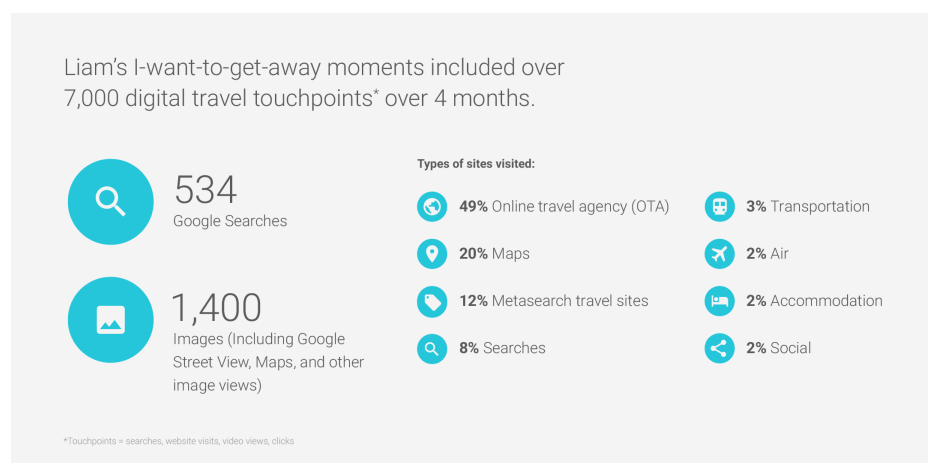


Figura 1: Touchpoint del consumatore e quantità di ricerche; Fonte: Google partnered with Luth to analyze the digital activity of its opt-in panel participants during a one-month, live-tracking study (2015)

Prendendo in considerazione la fig. 1 è possibile verificare il gran numero di interazioni che un soggetto compie quando deve pianificare un viaggio. Nello specifico, i dati sono riferiti ad uno specifico soggetto, Liam, che sta pianificando tre differenti viaggi. Ipotizzando che per ogni differente viaggio si impieghino la stessa quantità di ricerche, si può ipotizzare che in media si effettuano 178 ricerche e si visualizzano più di 460 immagini.

È proprio questa ricerca delle informazioni a modellare la scelta del consumatore, da una ricerca condotta da Google con Ipsos Connect è emerso, infatti, che il 67% dei viaggiatori è più propenso a prenotare con un brand di viaggi che gli ha fornito (Google I. , 2015), nella prima fase di *discovery* ma anche successivamente, informazioni che si sono rilevate rilevanti e che li hanno aiutati nel loro processo di scelta anche risolvendogli dubbi o paure.

Inoltre, in questa fase puramente di ispirazione, un ruolo chiave viene svolto dai contenuti visivi. Lasciarsi ispirare da un'immagine raffigurante un mare limpido è, infatti, più facile rispetto alla semplice descrizione di quella stessa spiaggia. Occorre, quindi, che gli stessi brand sviluppino contenuti visivi che permettano al soggetto di poter essere ispirato e scegliere di conseguenza il brand.

I contenuti visivi possono essere di differente natura, immagini o brevi video, e questi devono essere condivisi *online* e sui vari *social network* su cui è presente il brand.

Il contenuto dal più forte impatto, in questa fase, potrebbe essere rappresentato da una foto/ un video che oltre ad ispirare il soggetto lo informa o gli permette di rompere delle barriere che esso ha su quella destinazione.

Il secondo micro-momento è il *Time to take a plan moments*, in questo caso il soggetto ha ben chiara qual è la sua destinazione e per questo inizia ad effettuare delle ricerche per capire come raggiungere quel luogo o anche dove soggiornare. Ma in questa fase, molto di programmazione, il soggetto inizia a decidere e programmare anche quali attività svolgere così come quali luoghi visitare.

Spesso i soggetti, se non hanno già visitato quei luoghi e non hanno costruito una relazione forte e di *loyalty* con uno specifico brand, effettueranno delle ricerche per capire dove soggiornare.

In questo caso, quindi, il soggetto ricercherà una sistemazione utilizzando delle *query* sul motore di ricerca. Le *query* maggiormente utilizzate in questi casi sono ad esempio: "hotel a Roma" o "miglior hotel a Roma centro".

In questo caso, occorrerà quindi ben calibrare le *ads online* in modo che a seguito della ricerca di queste *query* al soggetto venga sottoposto il proprio brand.

Il penultimo micro-momento è il *Let's book it moments*, in questa fase avviene la prenotazione vera e propria. Infatti, dopo aver deciso dove andare il soggetto è pronto ad acquistare e fermare la sistemazione.

Kendra's let's-book-it moments

One of Kendra's search paths

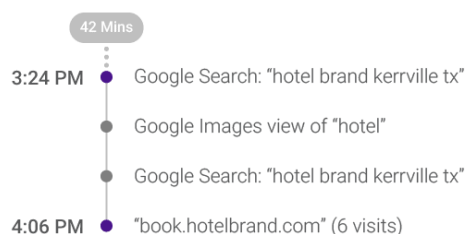


Figura 2: Esempio di ricerche effettuate online; Fonte: Google partnered with Luth to analyze the digital activity of its opt-in panel participants during a one-month, live-tracking study (2015)

Solitamente in questo micro-momento oltre ad aver già programmato la destinazione e i vari luoghi da visitare, si è già deciso anche il brand alberghiero che si vorrà prenotare. Come è possibile vedere in fig. 2, infatti, un tipico percorso di ricerca prevede di scrivere già sul motore di ricerca il nome del brand.

Tanto che, come si può vedere sempre nella fig. 2, il tempo necessario per prenotare è relativamente breve, circa una quarantina di minuti. Questo tempo dipende sicuramente anche dal soggetto che compie la prenotazione; i viaggiatori *business*, vista la frequenza maggiore con cui effettuano viaggi e prenotazioni, impiegheranno meno tempo di un soggetto che invece solitamente non viaggia molto.

Questo tempo breve è però anche dovuto alla caratteristica di questo micro-momento e del precedente; infatti, la maggior parte delle ricerche sono state già effettuate nel precedente step e di conseguenza il tempo necessario per prendere la decisione finale ed effettuare la prenotazione è minore.

È necessario effettuare una precisazione che riguarda questo micro-momento. Se il soggetto dovesse incontrare problemi durante la prenotazione è chiaro che lo stesso passerà oltre, ed andrà a prenotare un brand diverso.

È quindi importante sviluppare il proprio sito in modo che sia ottimizzato sia per dispositivi mobili che per computer. Questo anche perché il consumatore ricerca sempre più l'esperienza *omnichannel* ed anche per effettuare una prenotazione potrebbe utilizzare prima l'uno poi l'altro dispositivo.

L'ultimo micro-momento è invece il *Can't wait to explore moments*, che si ha quando la prenotazione è già stata effettuata.

Questa fase è collegata alle attività che il soggetto potrebbe voler fare una volta raggiunta la destinazione. Anche in questo caso, sono le informazioni a giocare un ruolo fondamentale.

Da una ricerca "*Think with Google*", è infatti emerso che circa l'85% dei viaggiatori di piacere sceglie solo una volta arrivati a destinazione le attività da svolgere (Google T. w., 2016).

È quindi essenziale che il brand fornisca al soggetto delle informazioni, anche sulle attrazioni locali e meno turistiche, che possano aiutare lo stesso cliente a decidere le attività da compiere. Quest'azione svolta dal brand sarà percepita dal cliente come un aiuto nei loro confronti, e questo porterà sicuramente a una migliore e più forte relazione.

Ma le informazioni e gli aiuti che si possono fornire al cliente non sono solo ed esclusivamente a livello turistico o nel momento di arrivo del soggetto in struttura.

In questo ne è un esempio la catena alberghiera Hilton.

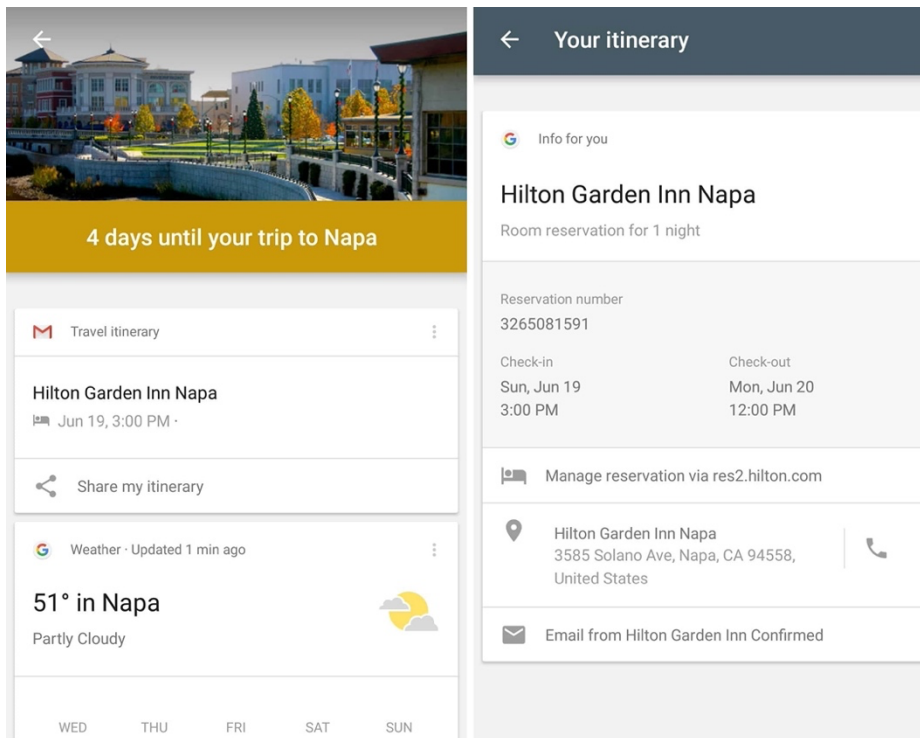


Figura 3: Hilton Can't wait to explore moments; Fonte: "How Micro-Moments Are Reshaping the Travel Customer Journey" Think with Google (2015), link: <https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/micro-moments-travel-customer-journey/>

Come si può vedere in fig.3, la nota catena alberghiera fornisce informazioni ed aggiornamenti circa la prenotazione effettuata dal proprio cliente attraverso Google o attraverso l'app di sua proprietà.

In questo modo, il cliente si sente maggiormente curato e di conseguenza si instaura quella relazione maggiormente forte e leale che dovrebbe essere l'obiettivo di ogni brand.

Al cliente, ad esempio, vengono forniti aggiornamenti circa il meteo in modo che lo stesso sia preparato sul tempo e sulla temperatura che troverà giunto a destinazione.

Pur essendo queste informazioni facilmente reperibili in rete, il poter accedere alle stesse senza dover effettuare una specifica ricerca, magari anche più volte al giorno, consente al cliente di risparmiare tempo.

La catena alberghiera si è spinta oltre grazie alla propria app. Infatti, agli HHonors members permette, attraverso l'app, di poter scegliere la stanza in cui pernottare, effettuare il *check-in* e una volta in loco aprire la stanza prescelta senza necessità di una chiave ma solo attraverso l'app.

Questo rappresenta per il cliente, soprattutto business, un enorme vantaggio in quanto permette di ottimizzare i tempi del soggiorno. Peraltro, la possibilità di effettuare il *check-in* autonomamente e aprire la porta attraverso l'app permettono al soggetto di avere una maggiore flessibilità riguardo all'orario di arrivo in struttura.

Tutti questi accorgimenti fanno sì che il soggetto, in presenza di un altrettanto piacevole esperienza durante il soggiorno, valuti positivamente l'esperienza e sia disposto a tornare nuovamente nella struttura poiché ha percepito un'attenzione verso di sé che lo spinge ad essere fedele al brand.

Il marketing nel settore turistico alberghiero

Un brand può commercializzare beni o servizi. Ovviamente non tutti i beni sono uguali poiché questi differiscono per caratteristiche. È chiaro che l'oggetto venduto da un brand alberghiero, ad esempio una stanza, non ha le stesse caratteristiche di un prodotto venduto, invece, da un'azienda di alimenti.

Per questo, quando si parla di prodotti, si può far riferimento alla *Sec Taxonomy*.

La *sec taxonomy* prevede di classificare i prodotti in tre differenti classi: *search goods*, *experience goods*, *credence goods*.

I primi sono tutti quei prodotti che possono essere valutati antecedentemente all'acquisto attraverso, ad esempio, un'ispezione diretta del prodotto. Appartengono a questa categoria i prodotti d'abbigliamento come anche l'arredamento.

La seconda classe, l'*experience goods*, si riferisce, invece, ai prodotti che possono essere valutati esclusivamente dopo che li si è acquistati, successivamente, quindi, ad avere vissuto appunto l'esperienza scaturita dal prodotto/servizio acquistato. Ed è proprio a questa classe che appartiene il prodotto oggetto dell'elaborato cioè il soggiorno presso una struttura alberghiera.

L'ultima categorizzazione si riferisce, invece, a tutti quei prodotti e/o servizi che sono difficili da valutare precedentemente all'acquisto ma anche successivamente allo stesso. Esempio di prodotti appartenenti a questa classe sono, ad esempio, gli integratori alimentari o le prestazioni mediche. In questi casi, è infatti alquanto difficile riuscire a capire l'utilità finale che si è ottenuta dal consumo di quel prodotto o dalla fruizione del servizio.

Come delineato poco fa il pernottamento presso una struttura è un *experience goods* poiché è necessaria l'esperienza per capire se il prodotto acquistato, in questo caso un servizio, sia di gradimento del cliente.

Il consumatore solo dopo aver soggiornato potrà effettivamente capire se l'esperienza che ha vissuto è stata positiva e di conseguenza anche se il brand acquistato è stato di suo gradimento.

È, quindi, fondamentale costruire e migliorare al meglio la relazione con il cliente proprio perché questa è usata dallo stesso per valutare la soddisfazione del servizio.

Ma la relazione, che è migliorabile attraverso differenti metodi e modalità, non è legata solo all'esperienza vissuta durante il soggiorno in struttura ma anche a tutto ciò che accade precedentemente come, ad esempio, al percorso di prenotazione.

La relazione e la *brand equity* di un marchio alberghiero si compone infatti di due differenti tipi di esperienze: l'esperienza di prenotazione, online o offline che sia, e l'esperienza vissuta durante il soggiorno.

È importante, quindi, che l'esperienza e gli strumenti adoperati dal *management* migliorino entrambe le situazioni.

Quindi, l'attenzione al cliente si deve garantire sia prima che successivamente alla prenotazione, poiché l'esperienza e il giudizio finale del soggetto sulla struttura sarà l'insieme dei giudizi dello stesso sia sulla fase pre-acquisto che post-acquisto.

Per costruire, dapprima, e migliorare, poi, la relazione con la propria clientela possono essere utilizzati una moltitudine di strumenti comunicativi. La necessità di utilizzare questi strumenti è sempre più forte poiché l'ambiente dell'*hospitality* è sempre più ampio e competitivo visti i mezzi tecnologici in continua evoluzione. Ma gli strumenti adoperabili per costruire azioni di marketing strategico e processi di marketing operativo sono vari poiché vari sono i momenti in cui questi possono essere applicati così come sono vari i soggetti verso i quali questi vengono applicati.

Per quanto riguarda il primo punto, a momenti diversi del ciclo di acquisto del cliente corrisponderanno strumenti ottimali differenti. Se infatti, nei primi momenti sono necessari strumenti che permettano di mostrare e valorizzare il proprio brand tra le molteplici scelte presenti nei motori di ricerca, nelle fasi successive si devono individuare strumenti volti alla fidelizzazione dello stesso.

Nella pratica, nella prima fase del ciclo di acquisto, si possono adoperare strumenti come Google ads o Facebook ads volti a mostrare al soggetto il proprio brand.

Questo sarà possibile poiché questa tipologia di strumenti traccia le ricerche e la presenza online di un soggetto ed andrà quindi a consigliare strutture che sono coerenti a questa.

Se, infatti, si effettua una ricerca online relativa ad un luogo o ancora si cercano *query* del tipo "hotel a Roma" in automatico gli strumenti di tracciamento capteranno la nostra esigenza di un alloggio e attraverso pubblicità sui motori di ricerca, sui blog o ancora sui *social network* e ci mostreranno *ads* relative a quel luogo.

In questo caso, sarà quindi importante impostare in modo ottimo le *keywords* in modo che la struttura sia tra le prime ad essere consigliate al soggetto.

Ma è necessario anche consolidare la propria presenza online poiché a seguito di queste *ads* i soggetti vorranno verificare e confrontare le varie strutture che gli sono state sottoposte attraverso le inserzioni o che hanno trovato in rete.

Per questo sarà importante non solo sviluppare un sito accessibile e ottimizzato anche per i dispositivi *mobile*, come già detto precedentemente, ma anche curare e aggiornare le proprie pagine dei *social media* in modo da offrire al cliente il maggior numero di strumenti con cui esaminare il brand e la struttura ed, eventualmente contattarla.

Per quanto riguarda la fase successiva a quella di soggiorno, quindi quella più prettamente indirizzata al mantenimento e alla fidelizzazione del soggetto, in questo caso uno degli strumenti più utilizzati resta la *newsletter* oppure la classica cartolina di augurio inviata dal brand utilizzata anche come veicolo per mostrare le novità dei propri servizi oppure il listino prezzi aggiornato.

La scelta dell'uno o dell'altro strumento, così come in generale la scelta di prediligere il marketing tradizionale o quello digitale, dipende principalmente dal tipo di clientela che il brand possiede.

Per questo motivo, è importante conoscere a fondo la propria clientela e pensare di conseguenza a strumenti che possano essere idonei alla stessa.

In generale, si può dividere la popolazione, secondo l'anno di nascita, in 4 fasce: *baby boomers*, generazione X, generazione Y e generazione Z. A queste differenti generazioni corrispondono chiaramente, per via delle abitudini di acquisto o dei diversi bisogni da soddisfare, strumenti e tecniche di marketing differenti.

Per quanto riguarda i *baby boomers*, questi sono i nati tra il 1946 e il 1964. Si tratta di persone che, solitamente, godono di una ricchezza materiale maggiore e che, nonostante l'età, vogliono comunque essere considerati giovani e rilevanti (Rowson, 2016). Proprio per questa peculiarità non sono da escludersi completamente i mezzi tecnologici e i social media poiché questi soggetti sono presenti e spesso molto attivi su questi.

Sono però una generazione che predilige molto le informazioni e di conseguenza il sito dovrebbe essere sviluppato con molto testo e qualche immagine e prediligendo un *layout* chiaro e magari una dimensione del carattere leggermente più grande del solito (Emilie Slootweg, 2018). Si deve, quindi, disegnare una comunicazione che sia più accessibile e leggibile ma senza che passi il messaggio che è stata sviluppata in questo modo vista l'età dei soggetti.

Per quanto riguarda la generazione X, in questo caso si fa riferimento a quella fascia della popolazione nata tra il 1965 ed il 1980. Secondo molti studi, questa è la generazione, da un punto di vista del marketing, più difficile da stigmatizzare poiché racchiude persone che, per via del differente anno di nascita, sono spesso molto differenti nelle caratteristiche.

Lo strumento perfetto per questa generazione, che si rivela invece non utile per le altre, è l'e-mail. Se infatti per le persone più giovani l'e-mail è vista come uno strumento datato e i *baby boomers* potrebbero non usarle molto spesso, per questa fascia d'età l'e-mail è uno strumento molto efficace. Si parla, infatti, di soggetti che sono attivi a livello lavorativo e che quindi consultano l'e-mail varie volte durante la giornata.

È quindi uno strumento ideale per questa generazione la *newsletter*. Quest'ultima è uno strumento che può essere utilizzato al fine di mantenere viva la relazione con il cliente, anche informandolo di eventuali promozioni o offerte, e che risulta essere un valido alleato al fine di fidelizzare lo stesso.

Il problema di questo strumento riguarda la frequenza; una frequenza eccessiva potrebbe, infatti, produrre il risultato opposto cioè quello di annoiare il cliente e di portarlo ad allontanarsi.

La generazione X racchiude soggetti nati e/o cresciuti nell'epoca dell'introduzione del divorzio in Italia e delle successive liti familiari e per questo ricercano la stabilità e la fiducia verso un brand. Sono anche la generazione che potenzialmente ha il maggior reddito e che di conseguenza potrebbe richiedere maggiori servizi. Per questo motivo una struttura alberghiera, per i propri messaggi pubblicitari, potrebbe far leva sulla comodità delle stanze e sui servizi offerti cercando sempre di comunicare un'immagine di se come brand affidabile e di cui potersi fidare (Emilie Slootweg, 2018).

Così come per i *baby boomers*, anche per questa generazione non sono da escludersi i *social media*, poiché anche in questo caso questi sono strumenti utilizzati dalla popolazione di questa fascia d'età.

La generazione Y comprende quei soggetti nati tra il 1981 e il 1995, e che quindi hanno assistito in pieno all'evoluzione della tecnologia. Sono, quindi, molto *confident* con gli strumenti digitali così come con i *social media* ed è su questi che si può basare la comunicazione. Sono, inoltre, molto attenti alle opinioni altrui sia

quando si tratta di opinioni di conoscenti che quando si tratta di recensioni lette online (Emilie Slootweg, 2018). Sempre secondo gli stessi autori, è quindi importante cercare di fidelizzare il cliente, soprattutto nei suoi primi anni di interessamento al mercato, per cercare di generare in lui un senso di fiducia che permetta allo stesso di essere veicolo di un passaparola positivo. Inoltre, si aggiunge, che la generazione Y è molto attenta al nome del brand, di conseguenza potrebbe scegliere una sistemazione invece che un'altra per il solo motivo del grande nome di brand dietro e apprezza molto i contenuti personalizzati.

Di conseguenza, per attirare questo tipo di generazione occorre pensare ad una pubblicità basata principalmente sui canali online e su contenuti personalizzati.

Un'ultima generazione analizzabile è quella della generazione Z. Si tratta, in questo caso, di soggetti nati tra il 1996 e il 2012. In questo caso si parla di soggetti che sono nati nell'era tecnologica e dei social media, per cui questi mezzi rappresentano quasi l'unico metodo utilizzabile per effettuare *advertising* su questa generazione. Contrariamente, infatti, alle generazioni precedenti, in cui i mezzi online possono essere combinati a mezzi quali televisione o radio, questa generazione non è molto avvicinabile con questi mezzi preferendo, nella maggior parte dei casi, servizi streaming.

Altro punto di distacco rispetto alle generazioni precedenti è rappresentato anche dalla ricerca di valori nel brand. Infatti, questa generazione ricerca i brand che comunicano valori, che esprimono quindi una loro identità. Questa ricerca è più evidente e forte rispetto alle generazioni precedenti, tanto da spingere i soggetti a non comprare un certo brand se non si esprime su un certo valore oppure se comunica un valore differente dai propri. Predilige inoltre un linguaggio diretto, spontaneo e creativo e soprattutto i contenuti in cui gli è data la possibilità di interagire ed esprimere la propria opinione.

Seppur come appena delineato ogni generazione considerata ha proprie caratteristiche peculiari che rendono necessaria l'ideazione di una campagna pubblicitaria *ad hoc* non è detto che una struttura ricettiva abbia unicamente un target di riferimento. Potrebbe infatti accadere che una struttura in un certo periodo della stagione si riferisce maggiormente ad una fascia d'età e in un altro ad un'altra fascia d'età. In questo caso, visto che comunque la pubblicità effettuata per mezzo online o social è comunque indicata per qualsiasi generazione presa in considerazione, è necessario ben configurare l'*ads*.

Per i baby boomers si strutturerà una *ads* maggiormente informativa, mentre per la generazione Z, ad esempio, una *ads* maggiormente diretta e creativa. Dopo di che, sarà Google stesso, se si sta parlando di Google *ads*, a utilizzare l'uno o l'altro *format* in base alle informazioni del soggetto.

Strategie attuabili

Riuscire a prevedere con una buona accuratezza la presenza di cancellazioni permette allo stesso tempo una buona politica di overbooking così come una politica di comunicazione e fidelizzazione del cliente.

Il *management* della struttura, sapendo la percentuale di cancellazioni, può infatti fissare un tasso di *overbooking* tale da permettere di affittare le camere cancellate da altri clienti non perdendo quindi dei ricavi.

Tuttavia, una seconda politica applicabile a seguito della conoscenza della possibile cancellazione è la comunicazione anticipata con il cliente. Il management sapendo infatti che quella prenotazione rischia di essere cancellata, e prima che ciò avvenga, può comunicare con il cliente offrendo incentivi in grado di evitare ciò. Esempi di incentivi che possono essere offerti sono: upgrade della camera, sconto, omaggi...

È chiaro che, ogni albergatore conoscendo il prototipo di cliente della propria struttura dovrà optare per un vantaggio che sia adeguato a quest'ultimo. Far leva su un incentivo errato non riuscirebbe infatti a risolvere la situazione ma anzi potrebbe peggiorarla.

In uno studio condotto sul mercato alberghiero a Taiwan, (Chiang-Ming Chen, 2015) è risultato che, naturalmente, il primo fattore che influenza la domanda e di conseguenza la prenotazione è il prezzo. Se, infatti, altri fattori che incidono sulla domanda sono maggiormente forti per un tipo di cliente, più giovane/meno giovane, il prezzo è un fattore che produce effetto su un cliente di qualsiasi età e tipo, business e no. Avremo quindi che all'aumentare del prezzo, diminuirà la domanda.

Mentre il fattore età incide, e non poco, sul tipo di hotel prescelto. All'aumentare dell'età aumenterà anche, infatti, la propensione a scegliere una sistemazione che presenti una qualità dei servizi migliori.

Ulteriore fattore che incide sulla scelta del tipo di sistemazione è la durata del soggiorno. A seconda, infatti, della durata del soggiorno i clienti tendono a scegliere tipi differenti di sistemazioni. In particolare, probabilmente per motivi di budget, all'aumentare dei giorni si tende maggiormente a non scegliere catene internazionali ma sistemazioni a gestione più familiare.

Dall'analisi è risultato sorprendentemente che il reddito non ha effetto sul tipo di sistemazione scelto; in particolare, a reddito maggiore non corrisponde un maggior livello di qualità dei servizi richiesti dai clienti. Tuttavia, questo risultato potrebbe essere falsato dal tipo di vacanza effettuata a Taiwan, oggetto dello studio considerato. Infatti, solitamente si associa Taiwan a un tipo di vacanza più rurale, e questo potrebbe essere il motivo del perché, non necessariamente, un turista più facoltoso vuole prenotare in sistemazioni più di lusso. In generale, esistono principalmente due differenti tipi di viaggio: viaggi di piacere e i viaggi di lavoro comunemente denominati di business.

È chiaro che a questi due tipi di viaggi corrispondano bisogni differenti e di conseguenza caratteristiche richieste alla sistemazione differenti.

Secondo una ricerca di Yavas et al. condotta prendendo a campione un gruppo di abitanti del Southeast (USA), precisamente 341, che per almeno due volte negli ultimi 12 mesi avevano soggiornato in un hotel per motivi d'affari o per piacere (Ugur Yavasa, 2005).

Ne è risultato che, per i clienti business è essenziale la presenza in struttura di svariati servizi quali ad esempio la presenza di connessione wi-fi, sale riunioni così come centri fitness. Questo è il fattore più importante ricercato dai viaggiatori business.

Procedendo in ordine di preferenza segue: la facilità nella prenotazione e il check-in / check-out espressi.

Ulteriore elemento apprezzato da questo tipo di clientela risulta essere la posizione della struttura mentre non è essenziale la grandezza della stanza.

Per quanto riguarda, invece, i viaggiatori di piacere i risultati sono in parte differenti.

Se infatti anche per i viaggiatori di piacere il primo fattore più importante è, come per i viaggiatori business, la presenza dei servizi per quanto riguarda i fattori successivi le cose cambiano.

Nel campione dei viaggiatori per il tempo libero, la seconda dimensione più importante è, invece, legata alla sicurezza della struttura e del luogo in cui essa si trova. Questo è un punto che differenzia i due target; infatti, per quanto riguarda i viaggiatori business questo è un fattore che non è stato minimamente accennato da questi ultimi e che di conseguenza non rappresenta per gli stessi un fattore importante o considerato nella scelta.

Segue poi un ulteriore fattore, marginalmente meno importante per il segmento che è quello della comodità.

In questo torna quindi utile la grande distinzione effettuata poco sopra. Infatti, al cliente *business* e a quello *leisure* corrisponderanno incentivi da poter sfruttare differenti.

Se per il primo un incentivo utile potrebbe essere la presenza di un servizio espresso di trasporto da/per la stazione, lo stesso non si può dire per il cliente che viaggia per piacere. Il cliente business è, infatti, un cliente che notoriamente non è sensibile al prezzo e di conseguenza tutto ciò che è attinente allo stesso potrebbe non avere successo come deterrente alla cancellazione.

Discorso completamente opposto vale per i clienti di piacere che invece sono più sensibili al prezzo e ad eventuali benefit e upgrade della struttura; per questi gli incentivi possono quindi variare dallo sconto fino all'entrata omaggio nel servizio spa della struttura.

Questa appena delineata è però una divisione del cliente molto grossolana, è chiaro che, ad esempio, il cliente dell'hotel a 2 stelle avrà caratteristiche e di conseguenza leve su cui agire differenti rispetto al cliente dell'hotel a 5 stelle.

Così come saranno sensibili ad incentivi differenti anche i clienti dello stesso hotel che hanno però caratteristiche differenti; ad esempio, viaggiano in coppia o in famiglia.

In questo sta all'expertise del management riuscire a capire quale incentivo funziona meglio su quel cliente rispetto ad un altro.

Soluzione ed incentivo più universale che allo stesso tempo potrebbe funzionare in modo migliore anche all'albergatore per assicurarsi comunque il ricavo derivante dalla prenotazione è quella di offrire, sempre attraverso la comunicazione diretta con il cliente, la possibilità di rinviare la propria prenotazione in una data diversa da quella inizialmente prenotata.

In tutto ciò è quindi essenziale sviluppare un modello che sia accuratamente in grado di anticipare e predire le cancellazioni.

[L'importanza di una corretta gestione dell'overbooking](#)

Come già accennato precedentemente, è di estrema importanza cercare di anticipare le cancellazioni effettuate dal cliente in modo da porre in essere conseguenti strategie per non perdere troppo fatturato.

Lo studio della domanda e delle caratteristiche in presenza delle quali la cancellazione avviene in maniera più frequente è sicuramente fondamentale.

Costruita la *buyer* persona, e annotate le proprie caratteristiche, si può infatti capire quando mettere in atto la strategia di *overbooking*.

Se da un lato la pratica dell'*overbooking* risulta essere una prassi consolidata per far sì che l'albergatore non perda possibili fonti di ricavo, dall'altra può essere una strategia molto pericolosa a livello di fidelizzazione del cliente.

La gestione di un *overbooking*, con le relative azioni applicate, possono infatti influenzare enormemente il giudizio del soggetto sulla struttura alberghiera presso cui doveva soggiornare.

ha scritto una recensione a giu 2012
San Francisco, California • 1001 contributi • 136 voti utili

●●●○○

Vittima dell'overbooking: senza camera ma con compensazione soddisfacente

«Avrei dovuto alloggiare in questo albergo per una notte: avevo una prenotazione confermata con l'indicazione che sarei arrivato in tarda serata. Quando sono arrivato, poco prima di mezzanotte, l'albergo non aveva più camere disponibili a causa di overbooking: inconvenienti del genere non fanno mai piacere e non ho gradito il fatto di non essere stato contattato preventivamente per informarmi del problema e della soluzione.
Soluzione peraltro accettabile: mi è stata offerta una notte a titolo gratuito in un altro albergo in zona, a poco meno di dieci minuti d'auto. L'albergo nel quale ho finito col pernottare era forse meno recente e più rumoroso, dunque non del tutto comparabile, ma una notte gratis è più che accettabile come compensazione. Il giorno successivo il direttore dell'albergo mi ha contattato per scusarsi di persona, sincerarsi che fossi soddisfatto almeno della soluzione di ripiego, e offrire 5000 punti Hilton Honors come ulteriore compensazione.
Voto: tre punti su cinque - uno di penalizzazione per l'overbooking in sé (che appartiene alla logica gestionale degli hotel ma non è mai gradevole quando se ne cade vittime), e uno per non avermi avvertito preventivamente, indicandomi telefonicamente la sistemazione alternativa.»

Mostra meno ▲

Data del soggiorno: maggio 2012
Tipo di viaggio: Ha viaggiato per affari
●●●○○ Posizione ●●●○○ Servizio

Figura 4: Recensione di un albergo relativamente alla situazione *overbooking*; Fonte: screen sito review

Come si può vedere dalla fig. 4 che rappresenta una recensione presa sul web, l'*overbooking* potrebbe essere classificato, come gravità, al pari di un'errata o assente comunicazione preventiva da parte della struttura.

Per quanto, applicando metodi di ML e di studio della domanda e, calcolando il corretto tasso di *overbooking*, queste situazioni dovrebbero essere abbastanza limitate, occorre capire preventivamente come gestire situazioni simili per non incorrere in situazioni negative.

La consuetudine vorrebbe che al cliente fosse offerto, a spese dell'hotel, una sistemazione di pari livello o superiore e che sia il più possibile vicina alla propria struttura, sia a livello di posizione geografica che di servizi offerti.

Tuttavia, non bisogna tralasciare la necessità di avvisare preventivamente il soggetto della criticità in modo che lo stesso possa decidere di cancellare la prenotazione o meno.

Una corretta e preventiva comunicazione potrebbe portare anche il cliente a spostare semplicemente il proprio soggiorno presso la struttura ed in questo modo, è chiaro, le conseguenze per la struttura, sia a livello di immagine che di ricavi, sarebbero minime.

Viceversa, una gestione errata della questione porterebbe, oltre che alla perdita di quel cliente, a una WOM negativa con conseguenti riflessi negativi anche sui potenziali clienti che, leggendo le recensioni online, potrebbero decidere di non proseguire con il processo di prenotazione.

Se infatti, anche in passato, un cattivo passaparola poteva portare degli effetti negativi al brand, negli ultimi anni attraverso internet ogni recensione può essere letta da milioni di utenti ed è chiaro, quindi, che gli effetti di una cattiva recensione saranno ampliati rispetto al passato.

Proprio, l'aver a disposizione internet per poter consultare in qualsiasi momento le recensioni di clienti passati ha fatto sì che si è passati dal semplice fenomeno dello WOM all'eWOM cioè *eletronic word of mouth*. Dove l'eWOM può essere definito come la condivisione e lo scambio di informazioni da parte dei consumatori su un prodotto o un'azienda tramite Internet, i social media ed in generale, attraverso la comunicazione mobile (Chu, 2021).

Questo perché come detto ampiamente dalle teorie dell'influenza sociale, le persone si lasciano influenzare molto dalle decisioni degli altri o dall'opinione che essi hanno.

L'influenza dell'eWOM, sia positiva che negativa, ha sul settore dell'ospitalità un impatto molto forte poiché si parla di servizi immateriali; cioè servizi che sono difficili da valutare prima dell'acquisto. Per questo motivo un cliente non potendo valutare il servizio personalmente prima dell'acquisto, si rivolge alle recensioni online. È quindi di estrema importanza cercare di soddisfare l'intera clientela in modo da possedere recensioni che sono il più possibile positive. Per far questo è quindi di estrema importanza saper gestire bene la domanda così come il tasso di overbooking.

In uno studio condotto da Montgomery et al. nel 2018 ed avente ad oggetto la scelta di un resort da parte di alcuni studenti universitari è emerso che, sorprendentemente, il prezzo non è stato un criterio di scelta predominante (Rhonda Montgomery, 2018). L'influenza sociale, invece, rappresentata dalle differenti recensioni dei viaggiatori ha avuto invece un forte effetto sulle scelte dei consumatori. Inoltre, sempre secondo detto studio, questo risultato è in linea con la letteratura precedente che ha sempre affermato l'importanza delle informazioni. Tuttavia, lo studio compie un ulteriore passo avanti rispetto alla letteratura precedente, cercando di indagare la ragione di questa differente importanza dei fattori prezzo e recensioni; ragione che viene individuata nell'effetto salienza. Numerosi studi circa l'effetto salienza affermano infatti che le informazioni che presentano una spiegazione letterale, come le recensioni, risultano avere maggiore influenza rispetto ai prezzi. Questo è dovuto al fatto che i prezzi sono esclusivamente rappresentati sotto forma di elenco mentre, appunto, le recensioni come se fossero quasi un racconto. Questo è anche dimostrato da uno studio compiuto nel 2014, in cui al consumatore venivano sottoposte alternativamente delle recensioni e delle valutazioni aggregate; ne è risultato che le prime hanno avuto un potere di influenza nettamente maggiore (Noone, 2014). Ciò che è emerso dallo studio condotto da Montgomery et Al. è che anche una sola recensione negativa può ridurre di molto la forza di influenza positiva esercitata da altre recensioni anche super positive.

È per questo importante gestire ogni cliente nel modo migliore possibile, anche quando si tratta di *overbooking*. Poiché l'impatto che la sua recensione negativa può avere sul *business* della struttura è più elevato di ogni altro eventuale costo che si può sostenere per cercare di compensare il cliente di un eventuale overbooking.

Azioni per evitare la cancellazione

Una delle azioni che si possono mettere in atto, preventivamente alla prenotazione, è la scelta riguardante le politiche di cancellazione e rimborso.

Qualora si scelga una prenotazione non rimborsabile, allora il soggetto sarà più propenso ad effettuare il soggiorno poiché viceversa perderebbe tutto il prezzo della vacanza.

Da una ricerca Booking si è, infatti, verificato che l'aggiunta di una tariffa non rimborsabile diminuisce le cancellazioni di almeno il 9% ed aumenta anche le prenotazioni del 5%².

Sebbene possa sembrare strano che la presenza di una tariffa non rimborsabile possa portare ad un aumento delle prenotazioni, questo di norma accade poiché le tariffe non rimborsabili, proprio per questa loro caratteristica, presentano un prezzo inferiore rispetto alle tariffe *standard*.

Tuttavia, occorre effettuare una precisazione: le prenotazioni in presenza di una tariffa non rimborsabile aumentano solo se si pone un'alternativa rimborsabile o con date flessibili.

Questo perché, in questo modo, si assicurano anche i soggetti indecisi o che comunque vogliono essere rassicurati sull'esistenza di un rimborso.

Inoltre, la presenza di più tariffe riferite alle diverse tipologie di prenotazione, rimborsabili non rimborsabili e con data flessibile, permettono anche al soggetto di capire qual è il vantaggio, a livello di prezzo, derivante dal prenotare una tariffa non rimborsabile rispetto ad una tariffa rimborsabile.

² <https://partner.booking.com/it/aiuto/guide/come-capire-meglio-prevenire-e-ridurre-le-cancellazioni>

CAPITOLO 2

La demand forecasting, perché è utile nel business

L'obiettivo fondamentale di un'impresa è quello di massimizzare il profitto, e cioè, la differenza tra i ricavi ed i costi in un determinato periodo. Per rendere il profitto massimo, un'impresa è quindi alla continua ricerca di metodologie per ridurre il più possibile gli sprechi e dunque i costi.

In questo può aiutare l'utilizzo di metodologie di *demand forecasting* che permettono di effettuare una previsione sulla domanda e la conseguente programmazione dell'approvvigionamento delle materie prime così come del fattore umano necessario per la produzione dell'output. La conoscenza delle risorse necessarie per soddisfare la domanda consente peraltro di evitare gli sprechi che comportano comunque un costo maggiore per l'impresa.

Inoltre, effettuando una corretta analisi e previsione della domanda e attraverso un buon coordinamento tra funzioni, si può migliorare anche notevolmente il processo aziendale andando a sviluppare ad esempio piani di marketing più targettizzati e temporalmente accurati così come l'entrata nel mercato di un nuovo prodotto. La demand forecasting sta, negli ultimi periodi, acquisendo sempre più importanza data la sempre maggiore incertezza dei mercati e anche alla luce del sempre più breve ciclo di vita dei prodotti. Come, infatti, afferma Bass et al. "la previsione più importante è la previsione prima del lancio" (F.M. Bass, 2001). Se infatti, tutte le previsioni delle domande sono astruse, la previsione prima del lancio di un nuovo prodotto è sicuramente la più difficile e importante non avendo alcun dato storico a supporto.

Quando si parla però di demand forecasting, occorre effettuare una prima differenza. Non tutte le aziende, infatti, operano secondo lo stesso modello di business; ci sono aziende che operano esclusivamente su commessa ed altre che invece, tipiche dei prodotti a largo consumo, operano su previsione. Già dalla nomenclatura è possibile capire che nell'ultimo caso è necessaria una previsione maggiore. Quando si opera per previsione è infatti necessario effettuare una *forecast* che riguarda sia le materie prime sia i prodotti finiti. Caso differente quando si opera per commissione in cui, invece, la previsione o non la si effettua o la si effettua esclusivamente sulle materie prime.

Tuttavia, come già detto, la previsione la si può effettuare per differenti motivi e su differenti archi temporali, e da questi due fattori dipendono anche l'accuratezza e la maggior facilità di previsione.

La prevedibilità di un evento o di una quantità dipende da diversi fattori, (Hyndman, 2018), tra cui:

1. quanto bene comprendiamo i fattori che contribuiscono ad esso;
2. quanti dati sono disponibili;
3. se le previsioni possono influenzare la cosa che stiamo cercando di prevedere.

È chiaro che questi fattori, influenzando la prevedibilità dell'evento, andranno ad influenzare anche l'accuratezza della stessa previsione. Inoltre, più si effettueranno previsioni a breve termine, più i fattori sopra elencati saranno disponibili e più precisi. Di contro nell'ipotesi di una previsione di lungo periodo, sembra

logico pensare che, disponendo di minori e meno precise informazioni, queste saranno meno accurate rispetto ad una previsione effettuata su un intervallo di tempo più ristretto.

A seconda del differente orizzonte temporale preso in considerazione, abbiamo tipi di previsioni differenti a cui corrisponderanno obiettivi perseguibili differenti.

Per quanto riguarda il breve termine, come affermato da Hyndman, le previsioni sono necessarie per la programmazione del personale, della produzione e del trasporto.

Mentre le previsioni a medio termine sono necessarie per determinare quelli che sono i requisiti che serviranno in futuro all'organizzazione per poter operare in modo produttivo. Di conseguenza occorrerà verificare, anche in base agli obiettivi che vuole perseguire l'impresa, quelli che sono i requisiti a livello di risorse, in modo da sviluppare un piano di acquisto che le coinvolga. Occorre, quindi, effettuare questo tipo di previsione per cercare di capire qual è il fabbisogno di materie prime, di personale o attrezzature e predisporre di conseguenza delle azioni per acquisirle.

Per ultime, ed in questo caso si parla di previsioni che superano i 2/3 anni, ci sono le previsioni a lungo termine. Questo tipo di previsioni sono usate principalmente per la pianificazione strategica e quindi al fine di fissare gli obiettivi a lungo termine che l'azienda vuole perseguire e fissare di conseguenza il piano d'azione necessario per raggiungerli. Ciò che è importante sottolineare è la forte influenza e considerazione che, durante la fase di sviluppo della pianificazione strategica, è riservata all'ambiente esterno; si deve, infatti, prestare molta attenzione e verificare quali sono i fattori ambientali e sociali dell'ambiente esterno così come occorre tener in considerazione i vari *trend* e le opportunità del mercato.

Per quanto riguarda gli obiettivi perseguibili con la previsione della domanda, è chiaro che questi sono differenti a seconda della funzione aziendale presa in considerazione. Per quanto riguarda la funzione marketing, questa utilizzerà le previsioni per analizzare eventuali *trend* così come per analizzare gli effetti di eventi promozionali/eventi comunicativi sulle vendite.

Per effettuare una corretta previsione della domanda, si devono seguire principalmente 5 *step* (Hyndman, 2018), che sono: definizione del problema, raccolta delle informazioni, analisi preliminare di natura esplorativa, scelta e selezione dei modelli ed infine utilizzo e valutazione finale del modello di previsione. Il primo passo da attuare è, appunto, la definizione del problema; questo è uno step fondamentale poiché da questo si deve capire con quali fini sarà utilizzata la previsione e da quali funzioni. È, quindi, necessario che in questa fase coloro che sono responsabili della realizzazione della previsione si coordinino e interloquiscano con le funzioni aziendali che andranno a collaborare nella realizzazione della stessa e che utilizzeranno poi, in futuro, quella stessa previsione.

Di importanza vitale per la corretta realizzazione della previsione, è anche il secondo step che ha come obiettivo quello di raccogliere informazioni in modo da poter sviluppare in maniera idonea la previsione. Si potranno raccogliere sia i dati storici sia le esperienze di coloro che lavorano giornalmente sul fenomeno su cui si sta lavorando. In questo modo, si basa la previsione sia su dati che sull'esperienza di esperti.

Per quanto riguarda i dati, si può scegliere di evitare di considerare i dati di più vecchia data poiché, sicuramente, negli anni si sono avuti cambiamenti strutturali nel mercato che hanno reso gli stessi meno utili o addirittura pregiudizievoli. Tuttavia, qualora si abbia un buon modello predittivo, non sarà necessario fare ciò poiché questo gestisce automaticamente e tiene in considerazione questi cambiamenti.

Essenziali nel terzo step, cioè l'analisi esplorativa, sono gli elementi di *data visualization*. Questi ultimi, infatti, permettono di verificare gli andamenti, le tendenze della domanda così come la presenza di eventuali trend stagionali. Qualora si visualizzino *outliers* corrispondenti a valori anomali che si sono registrati nel corso del tempo, allora sarà necessario effettuare una ricerca per capire la natura degli stessi.

Il quarto step consiste nella considerazione e poi successiva scelta del modello. Si dovrà valutare la numerosità dei dati storici così come la tipologia degli stessi. Inoltre, occorrerà anche verificare se quel modello permetterà di raggiungere l'obiettivo prefissato di previsione; poiché, come già detto in precedenza, la previsione che si effettua può essere utilizzata per obiettivi differenti.

L'ultimo step riguarda la previsione vera e propria, in questa fase, infatti, mediante l'utilizzo del modello più rispondente ai dati in possesso si passa alla previsione per il periodo futuro. Per quanto riguarda l'applicazione si deve prima costruire il modello sui dati storici in possesso per poi andare ad effettuare la previsione per il periodo futuro. Appartiene a questa fase, anche la valutazione dei risultati del modello e di conseguenza dell'accuratezza dello stesso.

All'interno di un'organizzazione nel momento in cui si vuole effettuare una previsione della domanda si devono, quindi, coinvolgere differenti funzioni poiché, naturalmente, per ottenere un buon risultato si ha bisogno il più possibile di informazioni.

È chiaro che quando una previsione è effettuata prendendo in considerazione soltanto le informazioni appartenenti ad una specifica funzione, questa sarà sicuramente meno accurata della stessa previsione ma prodotta prendendo in considerazione maggiori informazioni provenienti da diverse funzioni aziendali.

All'interno del processo previsionale possono interloquire differenti funzioni poiché, come già detto, più si hanno informazioni maggiori sono le possibilità che la previsione sia corretta. Qualora questo non sia fattibile occorre almeno coinvolgere le funzioni: finanza, *marketing*, *supply chain* e *sales*.

Questo perché la vendita non dipende esclusivamente dall'azione di una singola funzione, ma è l'insieme di varie azioni messe in atto dalle differenti funzioni.

In particolare, l'interazione con le vendite e il marketing è necessaria per determinare il livello della domanda futura prevista anche in base alle promozioni, all'impatto dei prodotti competitivi o l'entrata di nuovi prodotti (John T. Mentzer, 2008).

Per capire l'impatto di questi ultimi fattori è necessario, infatti, che ci sia un'interazione tra le varie funzioni aziendali in modo da stimare bene l'impatto di un'eventuale promozione, ad esempio, prendendo dati storici ed informazioni appartenenti alle diverse funzioni.

Ogni funzione dovrà poi comunicare eventuali azioni che possano avere effetti e conseguenze sulla disponibilità della domanda o sull'aumento/diminuzione della stessa.

Ad esempio, il reparto marketing dovrà informare le altre funzioni delle modalità esecutive e creative degli advertising così come dell'eventuale presenza di offerte promozionali.

Demand and sales forecasting differenze

Seppur molto simili e spesso confuse tra di loro la previsione della domanda e delle vendite risultano essere differenti tra di loro.

La *sales forecasting* risulta essere la previsione più facilmente effettuabile da parte del business. Questo perché, spesso, è lo stesso sistema utilizzato per registrare le vendite a fornire la loro evoluzione futura. Tuttavia, il problema sta nel fatto che si analizzano e prendono in considerazione esclusivamente le vendite effettuate; di conseguenza, non si tiene conto di un eventuale inventario in eccesso o di una mancata vendita dovuta all'assenza di scorte.

Questo implica che la previsione futura di vendite sarà effettuata prendendo in considerazione solo le vendite ma non l'eventuale mancanza di scorte che di conseguenza comporterebbe una errata previsione di vendite anche per l'anno successivo.

Inoltre, prendendo in considerazione le vendite queste saranno influenzate anche dal semplice posizionamento a scaffale; è chiaro, infatti, che un posizionamento più in basso nello scaffale comporterà minori vendite. Altri fattori da cui dipenderà la vendita saranno anche il prezzo del prodotto così come la presenza di promozioni non solo nel negozio considerato ma in generale nella totalità dei punti vendita.

Come detto, quindi, la previsione delle vendite non considera l'inventario e le scorte ne tantomeno la domanda futura; per questo motivo, per eseguire un'analisi che sia il più possibile accurata, si può ricorrere all'analisi della domanda e alla successiva previsione della stessa.

Come dicono le parole stesse, la previsione delle vendite esamina i dati di vendita mentre la previsione della domanda i dati relativi alla domanda di quello specifico prodotto/servizio.

Nella *demand forecast* oltre alla domanda che realmente si è convertita in vendita, si considera anche la domanda che per svariati motivi non si è convertita in vendita. Questo perché dietro a questa conversione potrebbero esserci problemi legati al business come appunto un inventario assente.

Inoltre, la previsione della domanda deve considerare accuratamente anche gli eventi promozionali per cercare di capire quale sia la domanda base, cioè la domanda che ci sarebbe comunque stata senza la quota aggiuntiva dovuta alla promozione del prodotto/servizio.

Le metodologie classiche di Demand Forecast

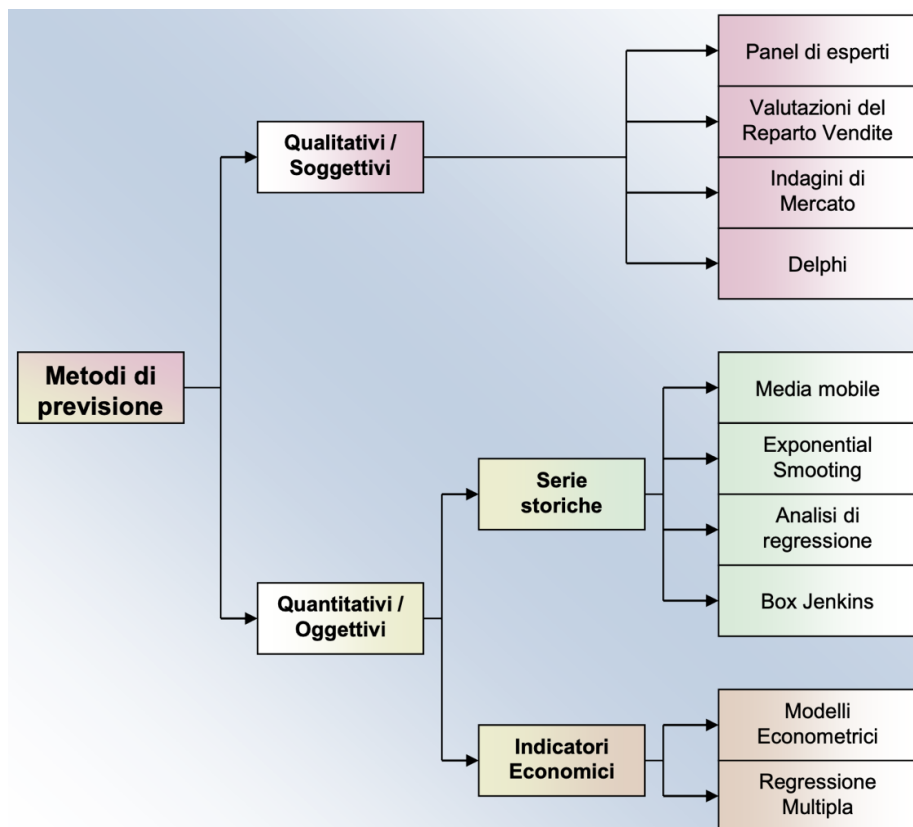


Figura 5: Differenti metodologie; Fonte: dispense prof. Melloni Università di Modena e Reggio Emilia "Previsione della domanda- contenuti di base"

Tuttavia, come si può vedere dalla fig.5, non esiste un'unica metodologia per procedere alla demand forecasting. La prima distinzione riguarda se il metodo è basato su metodi qualitativi, e quindi a base soggettiva, o segue un approccio *data driven*. È chiaro pensare che, nel secondo caso la demand forecast sarà più accurata essendo l'analisi più profonda e maggiormente incentrata sui dati.

Approfondendo nello specifico le metodologie definite "*judgmental*", queste sono quelle metodologie che non utilizzano una procedura *evidence-based* e che di conseguenza lavorano secondo un giudizio soggettivo. Come visibile nella fig.5, esistono differenti tipi di analisi ma tutte si basano sul giudizio di un soggetto ritenuto esperto; soggetto la cui funzione, a seconda del metodo considerato, varia all'interno dell'organizzazione aziendale.

Le previsioni della domanda possono essere effettuate, come è possibile verificare nella figura, dal reparto vendita. In questo caso, l'esperto in questione, che dovrà attraverso il proprio giudizio effettuare la previsione, è l'agente di vendita. Questo tipo di previsione si basa sul presupposto che, essendo gli agenti di vendita i soggetti più a contatto con i clienti sono anche quelli che conoscono meglio gli stessi. È chiaro che, visto il presupposto di conoscenza alla base di questo tipo di metodo predittivo, ogni agente di vendita effettuerà la previsione per la zona di sua competenza; dopo di che, ottenute le varie previsioni effettuate dai vari agenti, queste potranno essere aggregate per dar vita ad una previsione più generale, che possa ad esempio riguardare l'intero mercato di vendita del prodotto/servizio o una specifica area geografica.

Altri tipi di metodi come il panel di esperti o il metodo Delphi sono invece sviluppati entrambe dai manager. Nel primo metodo, una serie di esperti appartenenti a differenti funzioni all'interno della realtà aziendale, come finanza, marketing, supply chain, *sales*, si riuniscono per elaborare la previsione sulla domanda sulla

base dei dati di ogni funzione. Il problema di questo metodo riguarda la gerarchia. Poiché, intervengono negli incontri differenti soggetti e manager appartenenti a livelli gerarchici differenti e di conseguenza la previsione potrebbe essere influenzata da ciò. Come è possibile immaginare, infatti, il manager di più alto livello può più facilmente far rispettare la propria idea e al contempo un soggetto potrebbe dar maggior credito ad una certa previsione poiché proveniente dal manager più anziano o non intervenire per evitare situazioni di conflitto.

Il metodo Delphi ha cercato di superare il limite del dominio di *leadership* attraverso l'anonimato. In particolare, questo è un metodo d'indagine iterativo che sottopone cioè agli esperti una serie di questionari più o meno strutturati. In generale, si hanno tre fasi: fase esplorativa, fase analitica e per ultima la fase valutativa. Più si va avanti nel processo e maggiormente il questionario sarà strutturato. Per quanto riguarda il primo questionario, quello della fase esplorativa, in questo caso lo stesso sarà poco strutturato e con domande per lo più aperte. In questo modo, si permette agli esperti, sempre appartenenti a differenti funzioni della realtà aziendale, di esprimere la propria opinione e la propria idea in modo libero. Una volta superata questa fase, si comunicano i punti di vista e le opinioni emerse agli esperti, sempre garantendo l'anonimato, attraverso un sunto nel secondo questionario. In questa seconda somministrazione, le domande saranno maggiormente dettagliate e basate su quanto emerso nella prima fase. In questo modo, pur non essendoci, si può ricreare una sorta di confronto *face-to-face* che permette in modo costruttivo di generare degli *insights*. Successivamente al secondo questionario si iniziano ad ottenere dei primi dati quantitativi che saranno usati poi per la costruzione del terzo questionario. È prassi nel terzo questionario elencare, sulla base di quanto emerso precedentemente, delle tendenze che gli esperti dovranno valutare per cercare di effettuare la previsione.

Questo come detto prima, è solitamente lo schema adottato, ma un metodo Delphi può variare anche a seconda della complessità del fenomeno analizzato o dell'omogeneità delle opinioni degli esperti coinvolti. Questo perché è chiaro che, in quest'ultimo caso, saranno necessarie più somministrazioni in caso di opinioni molto divergenti.

Si può, infatti, dire concluso il processo solo quando dal questionario emerge un'unica idea condivisa tutti gli esperti.

Come si può capire, il metodo Delphi è un metodo qualitativo, poiché basato sul giudizio di un soggetto esperto, abbastanza strutturato che proprio per questa sua caratteristica consente di ottenere buoni risultati predittivi. Allo stesso tempo è anche veritiero che, viste le numerose fasi di somministrazioni e il grande numero di soggetti coinvolti, è un metodo non frequentemente utilizzato per via del suo alto costo di realizzazione.

Come detto precedentemente, il metodo Delphi e il metodo del panel degli esperti sotto molti punti di vista sono metodi molto simili. In entrambi i metodi, infatti, si effettua la previsione basandosi sul giudizio di soggetti esperti appartenenti a differenti funzioni, ma ciò che differenzia queste due metodologie riguarda l'anonimato. Nel metodo Delphi, infatti, per evitare che la gerarchia tra i manager influisca sull'analisi, le opinioni degli esperti vengono anonimizzate.

In tutto ciò, quindi, la demand forecast sarà sviluppata da un esperto che, basandosi sulle informazioni in suo possesso e sulla sua esperienza, formulerà la previsione. È logico pensare che più la situazione che si sta analizzando è simile ad una precedente, più la previsione sarà corretta. Molto dipenderà però, anche, dalla disponibilità delle informazioni e dai *bias* del soggetto esperto.

Nonostante i metodi basati sul giudizio degli esperti non siano i metodi più accurati per effettuare previsioni, questa è una delle metodologie più utilizzate nelle organizzazioni per effettuare la *demand forecast* anche visto, ad eccezione fatta del metodo Delphi, del loro costo e della loro facilità di realizzazione.

Alternativamente, un altro metodo utilizzabile è rappresentato dalle ricerche di mercato; questo tipo di indagine può essere sviluppata direttamente dall'azienda stessa oppure da aziende terze che offrono appositamente il servizio. Questo tipo di analisi seppur può essere utilizzata per effettuare previsioni, è il più delle volte adottata per cercare di individuare e studiare il mercato di riferimento per un nuovo prodotto e poi la relativa possibile domanda. L'indagine viene condotta su un gruppo di soggetti che sono rappresentativi del target del prodotto, se si sta parlando di un prodotto inedito, ovvero sui clienti se l'analisi ha ad oggetto un prodotto già presente sul mercato. È, quindi, una metodologia prettamente qualitativa che, per la sua natura esplorativa, può essere combinata con ulteriori metodi di previsione.

Le metodologie quantitative basandosi, invece, su criteri oggettivi ed effettuando le analisi sui dati sono maggiormente accurate.

In questo caso occorre però effettuare una distinzione; questo perché, si può operare con due differenti metodi: seguendo metodologie basate sulle serie storiche oppure metodologie basate sugli indicatori economici.

Quando si effettua la previsione basandosi sulle serie storiche, si può procedere attraverso differenti metodi: media mobile, *exponential smoothing*, analisi di regressione e per ultimo attraverso il *Box Jenkins*.

Il metodo della media mobile è maggiormente indicato quando la media è stazionaria poiché si basa l'analisi sulle ultime osservazioni registrate. Inoltre, il metodo, a seconda delle modalità di calcolo, può essere: media mobile semplice, media mobile ponderata o la media mobile esponenziale.

Il metodo della media mobile semplice utilizza il principio classico della media, dove, cioè, considerando un certo numero di osservazioni le si sommano per poi ridividere la stessa per il numero di osservazioni che si sono considerate. È chiaramente il metodo più semplice e veloce da utilizzare, il problema è, però, che ogni osservazione ha il medesimo peso e di conseguenza il primo valore considerato vale quanto l'ultimo.

Il limite detto precedentemente si supera utilizzando il metodo della media ponderata; in questo caso, infatti, ogni valore viene pesato. Nello specifico, il primo valore considerato sarà pesato 1, il secondo 2 e così via fino ad arrivare all'ennesimo valore considerato che sarà pesato n. Con questa differenza rispetto al metodo precedente pur migliorando la sensibilità e sensitività della misurazione, non si riesce comunque a cogliere in modo istantaneo i reali trend del mercato.

Ulteriore media considerabile è la media mobile esponenziale, questo metodo è particolarmente utile per la definizione e l'individuazione dei *trend*. Come accadeva per la media ponderata precedentemente analizzata, anche in questo caso si pesano i diversi dati dando maggior rilievo ai dati più recenti. Ciò che differenzia i due

metodi riguarda proprio il tipo di peso; mentre, infatti, nel precedente metodo si attribuiva un peso in modo lineare, in quest'ultimo metodo lo si attribuisce assegnando un valore progressivo esponenziale. Di conseguenza, con i pesi che variano in modo esponenziale, l'importanza data alle osservazioni più recenti risulta essere maggiore rispetto al metodo della media mobile ponderata.

Il metodo di *exponential smoothing* prevede come nel metodo precedentemente esposto, una ponderazione esponenziale delle osservazioni; il peso sarà crescente al crescere della recentezza dell'osservazione. Si opera, quindi, uno *smoothing*, cioè lisciatura, esponenziale.

Anche in questo caso, esistono differenti tipologie di *smoothing* esponenziale.

La prima metodologia è rappresentata dal *single exponential smoothing*, che è indicata in assenza di trend o stagionalità. Si ha un unico fattore di *smoothing*, *alpha*, che indica il tasso di decadimento esponenziale da attribuire alle osservazioni. Essendo un coefficiente il suo valore è di prassi compreso tra 0 ed 1.

Più il coefficiente avrà un valore vicino all'uno, maggiormente considererà e darà più peso alle osservazioni più recenti. Si può affermare che maggiormente il coefficiente è vicino ad 1, più l'apprendimento sarà veloce, cioè prenderà maggiormente a riferimento le più recenti osservazioni; al contrario, più il coefficiente tenderà allo 0 e più i dati storici influenzeranno la previsione.

La seconda tipologia di *exponential smoothing*, cioè il *double exponential smoothing*, cerca di superare il limite del modello precedente includendo e studiando anche eventuali trend presenti.

In questo modello, in aggiunta ad *alpha*, viene considerato un ulteriore coefficiente, *beta*, che attraverso i pesi controlla le tendenze presenti nei dati.

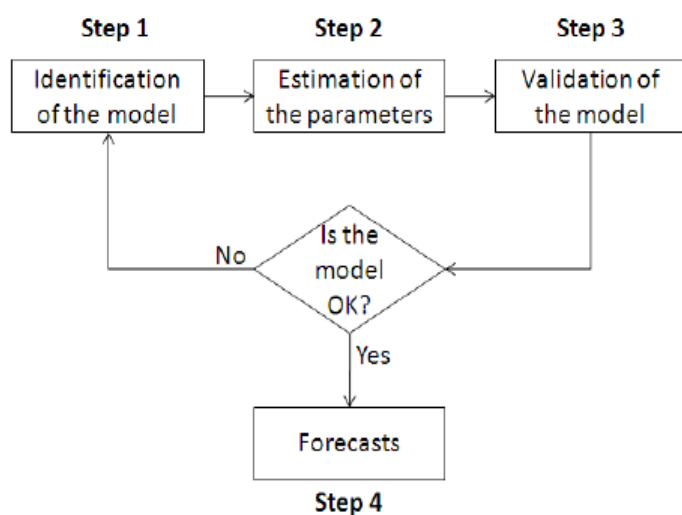


Figura 6: Il metodo Box Jenkins. Fonte: *A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting*; Alvarez et al. (2015)

Il metodo *Box Jenkins*, ha sempre l'obiettivo di prevedere i valori futuri di una serie, in questo caso della domanda, basandosi sull'analisi e lo studio delle serie storiche. Nello specifico, il metodo si basa sull'estrapolazione delle serie storiche mediante una media mobile autoregressiva a due parametri. La media mobile autoregressiva valuta la domanda attraverso la somma pesata delle domande passate e di componenti casuali che non sono prevedibili.

L'ultima tipologia di metodo, di natura quantitativa, applicabile è l'analisi di regressione.

La regressione si utilizza quando si vuole predire come varia il valore di una variabile indipendente al variare di una o più variabili indipendenti. La regressione può essere di due tipi: lineare o logistica. Alle due diverse tipologie di analisi corrispondono due obiettivi di analisi differenti; con la prima si cerca di predire il valore di una variabile dipendente continua mentre con la regressione logistica si vuole prevedere il valore di una variabile dipendente dicotomica. È quindi chiaro che la regressione logistica sarà utilizzata principalmente per risolvere problemi di classificazione.

Viceversa, la regressione lineare, vista la natura dell'analisi, può essere utilizzata per effettuare predizioni sulla domanda.

$$y^i = \beta_0 + \beta_1 x^i + \epsilon^i$$

Con la regressione lineare, come è visibile in figura, si vuole predire il valore di una variabile dipendente, y , prendendo in considerazione la presenza e la variazione di una variabile indipendente che funge da coefficiente della x . Quando si ha più di una variabile indipendente si parla di regressione multipla. Nell'equazione è poi presente β_0 che, invece, rappresenta l'intercetta della retta.

In particolare, attraverso la costruzione del modello, ed analizzando i p -value dei vari coefficienti si vuole capire se le variabili indipendenti considerate sono dei buoni predittori per la variabile dipendente oggetto d'esame. In particolare, si considera il cosiddetto test d'ipotesi dove:

H_0 : *coefficiente=0*

H_1 : *coefficiente diverso da 0*

Se il p -value dovesse risultare minore della soglia prescelta, solitamente 0,05, allora si potrà rigettare H_0 andando ad accettare H_1 . In quest'ultimo caso si ha che, essendo il coefficiente diverso da 0, la variabile indipendente collegata sarà un buon predittore della variabile dipendente considerata.

Tuttavia, per effettuare una buona predizione occorre identificare il modello migliore, cioè il modello capace di minimizzare il numero di errori tra misure e previsioni. Per verificare l'errore si possono considerare due differenti indicatori: la somma degli errori al quadrato e *R-squared*.

La prima metrica, SSE, ottenibile effettuando la sommatoria dei valori al quadrato del vettore dei residui, esaminando gli errori può valutare la bontà del modello. Naturalmente, più il valore sarà vicino allo zero meno errori saranno registrati dal modello. Di conseguenza, il modello costruito maggiormente riuscirà a spiegare il fenomeno e ad effettuare delle predizioni.

Mentre l'*R-squared* è una misura statistica che rappresenta la proporzione della varianza di una variabile dipendente che è spiegata da una o più variabili indipendenti. Contrariamente a SSE, in questo caso ci auspichiamo che l'*R-squared* tenda il più possibile ad uno. Questo perché la metrica varia da 0 ad 1, dove 0

sta ad indicare un modello che non riesce per nulla a spiegare il fenomeno studiato, mentre 1 indica che il modello spiega perfettamente il fenomeno.

Lo scopo, come già detto, è costruire un modello che rispecchi la realtà in modo da effettuare buone previsioni; tuttavia, l'inserimento nel modello di troppe variabili indipendenti potrebbe comunque non essere positivo. Questo perché un modello dovrebbe semplificare la realtà, obiettivo non perseguito se si complica eccessivamente il modello.

Tutti i metodi quantitativi finora analizzati sono basati sul principio delle serie storiche; tuttavia, questo implica che tutti i metodi presuppongono che l'andamento della domanda futura segua il pattern della domanda passata. Cioè, si stima il futuro basandosi sull'andamento passato.

Ma, esistono, come è possibile verificare nella figura 1, ulteriori metodi non basati sull'analisi dei dati passati ma sugli indicatori economici.

Questi ultimi prevedono di mettere in relazione la domanda di un prodotto/servizio con degli indicatori economici come, ad esempio, il reddito nazionale lordo o il reddito pro capite. Facendo ciò si analizza e si studia la domanda in relazione a come variano questi indici economici di riferimento.

La regressione multipla, come anticipato precedentemente, cerca di predire, invece, il valore di una variabile dipendente considerando come variano più variabili indipendenti.

I nuovi metodi di Demand Forecasting

Tuttavia, negli ultimi anni, vista la mole crescente di dati a nostra disposizione e di sistemi sempre più potenti, le tecniche di *demand forecasting* sopra evidenziate sono state, in parte, surclassate da tecniche più avanzate basate sul *Machine Learning*.

La prima precisazione da effettuare riguarda gli algoritmi utilizzabili. Come scritto di seguito, infatti, si possono adoperare differenti e molteplici algoritmi per effettuare la *forecast*; ognuno dei quali può essere *fit* con una certa situazione mentre non essere adatto in altre. Tutto ciò per dire che, quindi, non esiste la superiorità di un algoritmo su un altro, ma occorrerà valutare caso per caso quale degli algoritmi meglio si adatta a quella specifica situazione.

I primi due algoritmi utilizzabili appartengono alla famiglia degli algoritmi basati sulle serie storiche; questi sono: ARIMA e SARIMA.

L'ARIMA, media mobile integrata autoregressiva, è spesso utilizzato complementariamente ai modelli di *smoothing* esponenziale. Tuttavia, mentre i modelli di *smoothing* esponenziale si basano su una descrizione della tendenza e della stagionalità nei dati, i modelli ARIMA mirano a descrivere le autocorrelazioni nei dati (Hyndman, 2018). Questo tipo di algoritmo è maggiormente indicato per le previsioni a più breve termine.

Come si può vedere dallo *spelling* della parola stessa, ciò che si aggiunge è indicato dalla lettera S. Se, infatti, AR indica l'autoregressione, I l'integrato e A la media mobile, a queste lettere si aggiunge la S che sta ad indicare la stagionalità. Infatti, con SARIMA si intende: media mobile autoregressiva integrata stagionale. Ciò

che differenzia, quindi, l'ARIMA da SARIMA è la considerazione della stagionalità e dunque il riuscire a supportare dati di serie temporali uni-variate che coinvolgono backshift del periodo stagionale (Jain, 2022). Ulteriore algoritmo utilizzabile è la regressione lineare. La regressione lineare è una metodologia statistica che prevede l'utilizzo dei dati passati per effettuare previsioni sul futuro studiando il rapporto lineare esistente tra la variabile dipendente e quella indipendente.

Si effettua ciò, nello specifico, andando a stimare i valori attesi della variabile dipendente attraverso la considerazione di come variano le variabili indipendenti. Quando si effettua la regressione, il primo passo da compiere a livello di dati, sarà quello di dividere gli stessi in due parti attraverso lo *split ratio*, cioè seguendo una proporzione che solitamente è 70/30. Infatti, in questo modo si potrà andare a costruire il modello di regressione su una parte dei dati, definita *train set* e che è composta dalla maggioranza degli stessi, per poi andare ad effettuare la previsione sul *test set*. Pur dividendo il dataset attraverso il rapporto predefinito dato dallo *split ratio*, la proporzione dei dati, per la variabile di nostro interesse, esistente nel dataset originario rimarrà invariato. Operando in questo modo, il modello viene eseguito su dati differenti rispetto a quelli con i quali è costruito ed in questo modo si assicura una previsione maggiormente affidabile poiché sarà lo stesso modello ad essere più accurato.

Sarà cura del soggetto che costruisce il modello, andare a valutare quali variabili inserire nello stesso, verificando la loro significatività; le variabili non significative, non dovranno infatti essere considerate per il modello in questione.

Un modello sarà, quindi, differente a seconda di quali variabili si prendono in considerazione nello stesso. Per valutare quale sia il modello migliore, si devono confrontare gli scarti esistenti tra la previsione e il valore reale. L'obiettivo di questa parte di analisi è infatti quello di identificare quale sia il modello migliore, cioè il modello capace di minimizzare il numero di errori tra misure e previsioni. Per far questo si potrà, ad esempio, andare a considerare, come già detto precedentemente, la somma dei quadrati degli errori o l'*R-squared*.

Contrariamente ad SSE, in questo caso ci auspichiamo che l'*R-squared* tenda il più possibile ad uno. Questo perché, questa metrica può variare da 0 ad 1, dove 0 sta ad indicare un modello che non riesce per nulla a spiegare il fenomeno studiato, mentre 1 indica che il modello spiega perfettamente il fenomeno. È molto difficile, se non impossibile, che questo assuma il valore 1, poiché vorrebbe dire che il fenomeno è spiegato in modo perfetto dal modello costruito. L'obiettivo è però, come già accennato prima, che questo si avvicini il più possibile al valore 1 e quindi alla perfezione.

L'operazione di verifica della bontà del modello per una corretta riuscita deve essere eseguita sia per quanto riguarda il *training set* che il *test set*.

Alternativo al metodo della regressione lineare, è quello della *random forest*, un modello non parametrico appartenente alla più vasta categoria dei modelli *bagging*.

In questo modello, si effettua la predizione dei valori, attraverso l'aggregazione di diversi modelli di alberi decisionali. Quest'ultimo si costruisce attraverso varie operazioni di *splitting* dei valori delle variabili indipendenti, il tutto senza fare assunzioni di linearità. Il risultato sarà, come dice la parola stessa, un albero

che mostrerà in maniera più facilmente leggibile rispetto ad una regressione l'output della predizione. Per predire il valore in uscita, cioè il valore della variabile dipendente, possiamo seguire gli *splits* nei vari nodi dell'albero fino a raggiungere l'elemento foglia, cioè l'elemento finale dal quale non partono successive diramazioni. In ogni nodo sarà presente una condizione che farà capire quale diramazione, cioè direzione nell'albero, prendere per capire la predizione finale per quella specifica osservazione. A seconda degli *splitting*, cioè delle divisioni che si effettuano, si avrà una differente profondità dell'albero decisionale. Di conseguenza, per aumentare o diminuire la profondità occorrerà effettuare un numero crescente o decrescente di split. C'è da precisare che maggiore sarà la profondità dell'albero maggiore sarà la complessità del modello. Occorre ben calibrare il numero di *split*, poiché una situazione in cui si effettuano pochi split non è positiva poiché la previsione coprirà un range molto elevato. Tuttavia, non è positivo anche il caso contrario, cioè, effettuare *split* in modo reiterato poiché si arriverebbe ad avere un problema di *overfitting*. L'*overfitting* produce l'effetto che il modello funziona perfettamente sull'insieme dei dati sui quali è stato costruito, ma non funziona in modo egregio su altri dati, come potrebbero essere quelli del *test set*.

Elemento essenziale nella costruzione di un corretto albero decisionale è, pertanto, capire quanti *split* effettuare e soprattutto cosa porre come condizione.

Ulteriore metodologia applicabile è quella delle reti neurali artificiali (ANN). Le ANN sono un sistema computazionale basato sulla replica artificiale del cervello umano ed in particolare imitano il modo con cui i neuroni biologici si inviano i segnali.

L'idea di fondo delle reti neurali artificiali, modelli basati sull'apprendimento profondo, è quindi quella di imitare il comportamento del cervello umano.

L'apprendimento profondo consente al modello di apprendere partendo dalla rappresentazione dei dati, il tutto, come già detto, prendendo ispirazione dalla neurobiologia.

Biologicamente parlando, i neuroni sono collegati tra loro in modo ponderato e quando vengono stimolati trasmettono elettricamente il loro segnale attraverso l'assone. Nel neurone ricevente i vari input sono riassunti in un unico impulso.

A seconda di come il neurone è stimolato dall'input cumulato, il neurone stesso emette o meno un impulso.

Non appena il segnale accumulato supera un certo valore (chiamato valore di soglia), il neurone attiva un impulso elettrico che viene poi trasmesso ai neuroni collegati a quello corrente.

Si ha una vasta rete di miliardi di neuroni poiché ogni neurone è tipicamente collegato a migliaia di altri neuroni.

Anche le computazioni molto complesse, non eseguibili facilmente attraverso altri algoritmi, possono essere eseguite da una rete di neuroni anche abbastanza semplici.

Quando due neuroni interagiscono frequentemente, formano un legame che permette loro di trasmettere facilmente e in modo più accurato le informazioni.

Al contrario, quando due neuroni interagiscono raramente, la trasmissione è spesso incompleta e porta a una memoria difettosa o a nessuna memoria.

Il *Perceptron* è una delle più semplici architetture di Rete Neurale Artificiale, inventata nel 1957 da Frank Rosenblatt. Esso si basa su un neurone artificiale chiamato Threshold Logic Unit (TLU).

In questa architettura, gli ingressi e le uscite sono numeri ed ogni connessione di ingresso è associata a un peso.

Il TLU calcola una somma ponderata dei suoi ingressi

$$(z = w_1x_1 + w_2x_2 + \dots + w_nx_n = x^T w)$$

Poi, applica una funzione di passo, *step function*, a questa somma e produce il risultato:

$$hw(x) = step(z), \quad dove z = x^T w$$

Addestrare una TLU in questo caso significa trovare i giusti valori per i pesi in modo che questa dia vita ad un risultato veritiero.

Esistono differenti tipi di reti neurali a seconda del numero di strati che si considerano.

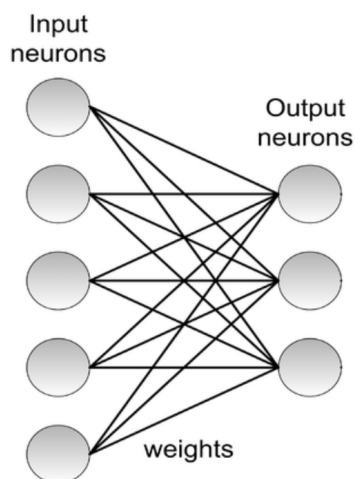


Figura 7: Single Layer; Fonte: *Neuromorphic Spiking Neural Networks and their Memristor- CMOS Hardware Implementations*. Mesa et al. (2019)

Il tipo più semplice di rete neurale è la *single layer*, che consiste in un singolo strato di nodi di uscita in cui gli ingressi sono alimentati direttamente alle uscite attraverso una serie di pesi.

Quando tutti i neuroni di uno strato sono connessi ad ogni neurone dello strato precedente (cioè i suoi neuroni di ingresso), lo strato è chiamato uno strato completamente connesso, o uno strato denso.

Come detto precedentemente, quando un neurone biologico attiva spesso un altro neurone, la connessione tra questi due neuroni si rafforza e così funziona anche la rete artificiale. Infatti, il peso della connessione tra due neuroni tende ad aumentare quando questi si accendono insieme ed inoltre, la regola di apprendimento del *Perceptron* rinforza le connessioni che aiutano a ridurre l'errore.

Il *Perceptron* è alimentato da un'istanza di allenamento alla volta, e per ogni istanza fa le sue previsioni. Per ogni neurone di uscita che ha prodotto una previsione sbagliata, rinforza i pesi di connessione dagli ingressi che avrebbero contribuito alla previsione corretta.

Quando si ha a che fare con più strati, invece, si parla di Multilayer Perceptron (MLP).

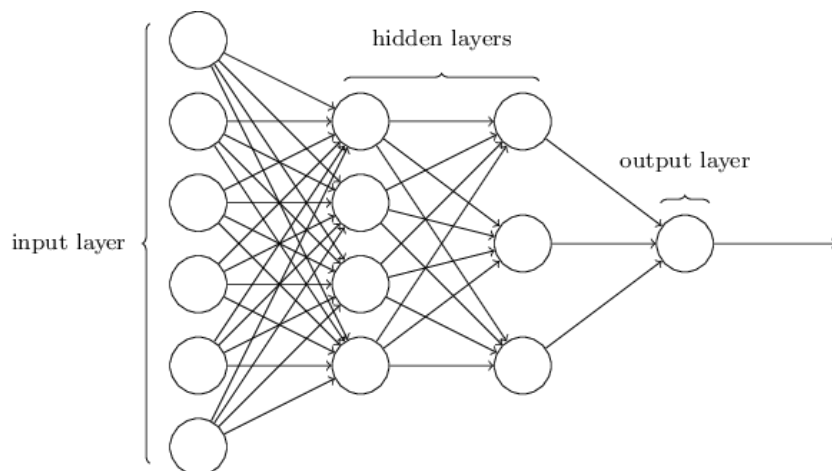


Figura 8: Rappresentazione di una rete neurale artificiale di 4 strati; Nielsen, *Neural Networks and Deep Learning* (2015)

Come è possibile verificare nella fig. 8, una rete neurale con più strati è composta:

- dagli *input layer*, ovvero i dati di ingresso
- uno o più *hidden layers*, che rappresentano gli strati dove avviene l'elaborazione e la lavorazione delle informazioni
- per finire con un *output layer*, che contiene il risultato finale.

Se una rete neurale contiene una profonda pila di strati nascosti, allora si chiama una rete neurale profonda (DNN).

Come detto ogni connessione tra neuroni è pesata da un valore che, inizialmente, è attribuito casualmente per poi essere modificato con le varie elaborazioni degli output e anche ogni qualvolta siano introdotte nuove informazioni.

Le reti neurali artificiali possono essere basate sia su tecniche di apprendimento supervisionato che non supervisionato ma anche rinforzato.

In caso di apprendimento supervisionato, attraverso il *training set*, si assegna alla rete neurale un insieme di dati *input* con i relativi output; in questo modo, la rete neurale può apprendere come siano collegate le due parti e poter poi, somministrando i nuovi dati appartenenti al *test set*, effettuare la previsione. Trattandosi di apprendimento supervisionato, affinché la rete neurale costituisca un buon modello, occorre modificare i vari pesi delle connessioni in modo da premiare la corretta generazione dell'output penalizzando invece la rete nel caso in cui l'output generato sia errato. In questo modo, attraverso questo meccanismo definito *error back propagation*, si può addestrare in maniera ottimale la rete neurale.

Caso differente quando si tratta di apprendimento non supervisionato, in questo caso, infatti, sarà la stessa rete neurale a dover modificare i pesi poiché non c'è alcun controllo da parte di un soggetto terzo. Ciò che differenzia l'apprendimento supervisionato da quello non supervisionato, oltre all'assegnazione dei pesi,

riguarda il fatto che mentre nel primo oltre agli *input* si assegnano anche gli *output*, nel secondo si assegnano esclusivamente gli *input*.

Mentre quando si ha a che fare con l'apprendimento per rinforzo, la rete neurale impara interagendo con l'ambiente e imponendosi dei disincentivi per ogni azione errata e un incentivo per le azioni che, invece, portano ad un risultato positivo. In questo modo, l'algoritmo comprende autonomamente quale è la strategia migliore, cioè la *policy*, da adottare per cercare di ottenere la miglior ricompensa possibile.

Riassumendo, quando si deve effettuare una previsione e si vogliono utilizzare i metodi di demand forecasting basati sul *Machine Learning* occorre prima di tutto verificare come e per raggiungere quale scopo la previsione viene effettuata per poi andare a verificare, attraverso l'utilizzo di metodi di *data visualization*, se sono presenti particolari andamenti. Dopo questi passaggi preparatori occorre verificare quale algoritmo utilizzare, questo è un passo molto importante poiché la corretta scelta della metodologia da utilizzare porta a risultati migliori. Tuttavia, fare questo risulta spesso non di facile realizzazione poiché, come già detto in precedenza, non esiste una metodologia che sia in assoluto superiore ad un'altra ma occorre valutare caso per caso.

Per capire quale sia il metodo di previsione migliore per il caso in oggetto sarà, infatti, necessario valutare oltre che naturalmente l'accuratezza della previsione anche altri fattori come: la distorsione, la varianza e non per ultimo l'interpretabilità dei risultati (Nikolopoulos, 2016)

Dopo aver scelto la metodologia da utilizzare occorre costruire il modello e addestrarlo attraverso l'utilizzo dei dati a disposizione. Però, per far questo, e per non inficiare la buona costruzione del modello, sarà necessario dividere i dati presi in considerazione in due parti in modo che il modello venga addestrato su una di queste parti mentre la previsione avvenga attraverso l'utilizzo dell'altra parte.

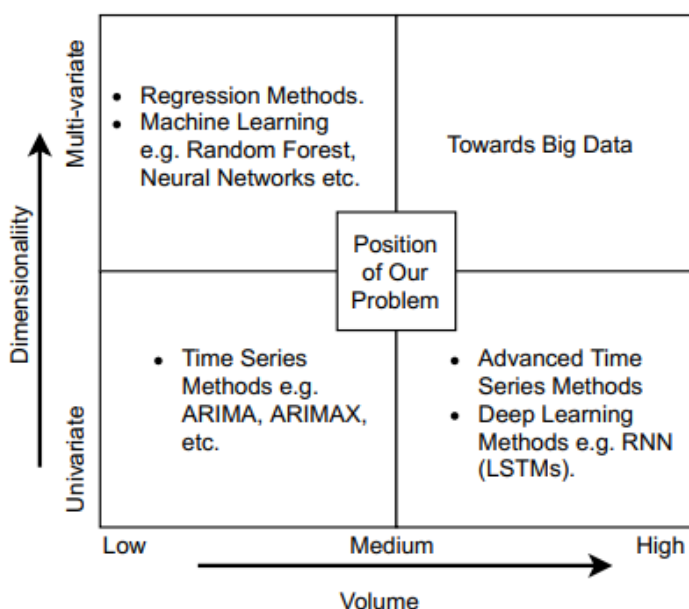


Figura 9: Classificazioni di metodi di previsione in base alle caratteristiche dei dati; Fonte: Nikolopoulos et al. *Forecasting Branded and Generic Pharmaceuticals* (2016)

Come illustrato nella fig.9, i metodi di previsione possono essere classificati attraverso due dimensioni: il volume e la dimensionalità dei dati.

Questa classificazione e lo studio dei dati in possesso, può aiutare il soggetto che intende effettuare la previsione a compiere una scelta per quanto riguarda il modello.

Da come si può verificare nella figura, una prima differenza riguarda le dimensioni coinvolte nel processo di analisi. A seconda, infatti, del numero di variabili prese in considerazione, solo una nel caso della dimensionalità univariata, si avrà un differente posizionamento nella tabella. In particolare, quando si sta trattando un'unica dimensione e al tempo stesso un volume non elevato di dati, allora si può risolvere il problema di previsione attraverso metodi basati sulle serie storiche, quindi ARIMA e ARIMAX. Quando, invece, la mole di dati è più elevata allora occorre o utilizzare metodi basati sul *deep learning*, come ad esempio le reti neurali ricorrenti, o metodi pur sempre basati sulle serie storiche ma di natura più avanzata.

Quando, invece, si sta cercando di risolvere un problema di previsione della domanda basandosi su più variabili, allora il discorso fatto sopra non è più valido. Quando, infatti, si ha a che fare con più variabili casuali che variano simultaneamente, gli algoritmi descritti sopra non sono più consigliati.

In questo caso, quando si posseggono numerosi dati i classici algoritmi di *machine learning* potrebbero non essere più in grado di effettuare la previsione per via della numerosa potenza di calcolo richiesta dal processo. In questi casi, occorre quindi ragionare e costruire modelli sulla base dei *Big Data*.

Quando, invece, pur lavorando con le multivariate si hanno moli di dati ancora gestibili ottimamente da parte degli algoritmi di *machine learning* allora si può ricorrere alternativamente: a metodi di regressione, alla *Random Forest* o ancora alle reti neurali.

Come già detto in precedenza, ogni metodo performa in maniera migliore sotto alcune condizioni; tuttavia, in uno studio condotto sulla previsione della domanda nella vendita al dettaglio, si è evidenziato come performano i vari metodi in questo campo.

Considerando il canale online, il metodo migliore, primo il metodo sviluppato dagli autori, risulta essere il LSTM mentre il peggiore è l'ARIMAX che non riesce a modellare ottimamente in presenza di informazioni temporali non lineari, cosa invece differente per quanto riguarda il metodo LSTM. Quando, invece, si considera il mondo offline allora in questo caso il metodo migliore, sempre non considerando il modello proposto dagli autori, è la random forest. (Punia, Nikolopoulos, Prakash Singh, Madaan, & Litsiou, Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail, 2020)

Revenue Management (RM)

Il *Revenue Management* è tradizionalmente definito come l'arte e la scienza di prevedere la domanda e contemporaneamente di regolare il prezzo e la disponibilità dei prodotti per soddisfare quella particolare domanda (Mehmet Erdem, 2016).

Il *Revenue Management (RM)* è applicato, principalmente, in tre differenti settori che condividono tra loro caratteristiche simili: compagnie aeree, alberghi e auto noleggi.

Infatti, tutti i loro prodotti: sono deperibili, la domanda varia significativamente nel tempo ed hanno grandi costi fissi mentre i costi variabili sono piccoli nel breve periodo (Wen-Chyuan Chiang, 2006).

La *demand forecast* rientra tra i 5 differenti problemi del *revenue management* assieme: alla determinazione dei prezzi, alle aste, al controllo della capacità e alla tecnica dell'*overbooking* (Wen-Chyuan Chiang, 2006).

L'obiettivo del pricing è quello di rispondere alla domanda su come determinare il prezzo per vari gruppi di clienti e come variare i prezzi nel tempo per massimizzare i ricavi.

Per quanto riguarda la determinazione dei prezzi, l'obiettivo è, naturalmente, massimizzare i profitti e di conseguenza si procede, per far questo, modificando i prezzi nel tempo e a seconda del tipo di clientela. L'asta è, invece, un modo "nuovo" di aggiustamento dei prezzi che grazie all'avvento delle prenotazioni online è entrato a far parte del *revenue management*.

L'obiettivo del controllo della capacità è, invece, similmente alla determinazione del prezzo, quello di massimizzare i ricavi cercando di allocare le risorse a disposizione in modo da raggiungere tale obiettivo.

L'*overbooking* consiste nel vendere più capacità di quella di cui si dispone, accettando nel mondo alberghiero/ricettivo maggiori prenotazioni, in modo da bilanciare le cancellazioni ed eventuali *no-show*.

Dove con *no-show* si intende il fenomeno secondo il quale i soggetti prenotati presso una struttura non disdicono la propria prenotazione ma non si presentano al *check-in*. L'obiettivo di tale processo del RM sarà, quindi, quello di cercare di trovare un livello di *overbooking* tale da permettere di controbilanciare perfettamente le disdette/*no-show* dei prenotati.

Secondo Mehmet et al., proprio come la maggior parte delle altre pratiche di gestione, la RM alberghiera si è evoluta nel corso degli anni (Mehmet Erdem, 2016). Si è, infatti, passati dall'essere semplicemente una pratica di gestione tattica ad avere un ruolo strategico nelle organizzazioni alberghiere, che incorpora anche le strategie di vendita e di marketing e la gestione del canale.

Quando si parla di *Revenue Management*, occorre effettuare una distinzione, poiché, così come per la strategia aziendale, abbiamo due livelli dello stesso: lo *strategic revenue management* e il *tactical revenue management*.

Per quanto riguarda lo *strategic revenue management*, questo rappresenta un processo aziendale altamente strutturato e multidisciplinare che coinvolge differenti funzioni che appartengono all'organizzazione tra cui il marketing e le vendite (Buckhiester, 2011). L'obiettivo di questo processo è quello di garantire un risultato il più possibile ottimo andando ad analizzare attraverso il dialogo di varie funzioni aziendali: i competitor, le strategie di *pricing* così come la *demand forecast*.

Invece, il *tactical revenue management* si propone di determinare quali pacchetti di servizi rendere disponibili per la vendita in vari momenti (Baker, 2002).

Tuttavia, vista la grande importanza relativa alla *demand forecast*, e la necessità di effettuare una previsione il più possibile vicina ai reali valori della domanda, occorre prendere in considerazione anche le cancellazioni. Basti pensare che secondo uno studio condotto su una grande catena alberghiera del Regno Unito, che ha

registrato circa 240.000 prenotazioni tra il 2004 e il 2006, circa il 20% sono state cancellate (Dolores Romero Morales, 2010). Questa percentuale, inoltre, è in continua crescita visti i nuovi mezzi di prenotazione e le nuove *policy* adottate dalla maggioranza dei siti di prenotazione online, di cui un esempio lampante è la formula “cancellazione gratuita”. Di conseguenza, quando si analizza la domanda occorre calcolare e di conseguenza valutare la cosiddetta domanda netta, cioè la domanda totale depurata dalle cancellazioni.

Oltre alle cancellazioni è, inoltre, frequente anche il caso denominato *no-show*, dove cioè l'ospite, pur non cancellando la propria prenotazione, non si presenta nella struttura ricettiva prenotata.

È, quindi, di fondamentale importanza cercare di trovare un modo per far sì che l'albergatore, per via delle cancellazioni e dei *no-show*, non sia troppo penalizzato a livello economico ma possa prevedere gli stessi ed attuare di conseguenza una strategia.

Negli anni molte strutture hanno applicato politiche di *overbooking* e cancellazione restrittiva; tuttavia, entrambe le politiche portano conseguenze non sempre positive (Nuno Antonio, 2019). Sempre gli stessi autori affermano, infatti, che queste due politiche oltre a produrre effetti negativi nel brevissimo periodo, per via dei costi di risarcimento o riallocazione, producono effetti anche nel lungo periodo visto che le probabili lamentele online produrrebbero una *word of mouth* (WOM) negativa e di conseguenza una perdita nei ricavi futuri.

È, quindi, di notevole importanza sviluppare modelli di *demand forecasting* e modelli che studino la cancellazione della domanda in modo da anticipare la stessa. Effettuare la *demand forecast* ed essere a conoscenza anticipatamente della possibile domanda è, infatti, essenziale al fine di definire anche gli altri punti del *revenue management cycle* come ad esempio l'ottimizzazione del prezzo. Poiché, se si è a conoscenza precedentemente che in un dato periodo si registra solitamente maggiore domanda o si registra un tasso di cancellazione della stessa molto elevato, anche i prezzi dovranno essere impostati di conseguenza.

Tuttavia, come è ipotizzabile pensare, prevedere con esattezza l'occupazione, cioè la domanda, in un dato periodo può non essere semplicissimo. Infatti, politiche efficaci di *revenue management* richiedono la previsione della domanda per segmento di mercato, per tipo di camera o *range* tariffario, oltre ad altri fattori chiave, alcuni già nominati, come cancellazioni e *no-show* come anche *over stay*, *under stay* e *walk-in* (Zvi Schwartz, 2016).

Come già detto precedentemente, con *no-show* si intende il fenomeno secondo il quale coloro che hanno una prenotazione presso la struttura alberghiera non si presentano il giorno del *check-in* non disdicendo però la loro prenotazione. Mentre con *over stay* ed *under stay* si intende, rispettivamente, la permanenza nella struttura per più tempo/meno tempo del tempo inizialmente prenotato. Con il termine *walk-in* si intende, invece, il fenomeno secondo il quale i soggetti si presentano presso la struttura senza prenotazione in cerca di un alloggio in cui soggiornare.

Come affermato dagli stessi autori, oltre ad effettuare *demand forecast*, i sistemi di *revenue management* necessitano della valutazione delle probabilità associate. Infatti, gli stessi per effettuare raccomandazioni, ad esempio, sulle tariffe o sul corretto livello di *overbooking* da praticare devono valutare e considerare la probabilità che le previsioni effettuate sulla domanda specifica del segmento si avverino (Zvi Schwartz, 2016).

CAPITOLO 3

Introduzione all'intelligenza artificiale

Attualmente si sente sempre più spesso parlare dell'intelligenza artificiale tanto che possiamo ormai vedere attorno a noi tantissimi esempi applicativi basati sulla stessa.

Ma nello specifico cos'è l'intelligenza artificiale?

Per definire l'intelligenza artificiale occorre prima di tutto prendere in considerazione i quattro grandi filoni di pensiero seguiti dagli studiosi. Questi differenti filoni si differenziano per come diversamente valutano il successo di un modello. Di conseguenza, si può pensare all'intelligenza artificiale e categorizzarla come (Stuart Russell, 2005):

- 1) Sistemi che pensano come esseri umani
- 2) Sistemi che pensano razionalmente
- 3) Sistemi che agiscono come esseri umani
- 4) Sistemi che agiscono razionalmente

I primi due approcci rivolgono maggiormente l'attenzione ai processi di pensiero e ragionamento, mentre gli ultimi due più sul comportamento.

Come anticipato, a questi differenti approcci corrisponderanno criteri per valutare il successo del modello differenti.

Per i punti 1) e 3) si verificherà il successo confrontando la macchina di AI con il comportamento umano, si andrà cioè a verificare la somiglianza tra i due. Nei restanti punti si andrà, invece, a confrontare i risultati dell'AI con i possibili risultati di un'intelligenza perfetta ed ideale.

Il primo approccio è quello proprio della scienza cognitiva. Questa si pone l'obiettivo non solo di replicare il comportamento umano ma di replicare anche i passaggi e la sequenza dei passaggi che sono seguiti dall'umano per arrivare a quel risultato. Appartiene a questa categoria il *General Problem Solver*.

Il secondo approccio basa tutto il suo funzionamento sull'inferenza e sul ragionamento simbolico. Si cerca cioè di produrre un risultato ottimo attraverso un sistema di inferenze logiche. Questo approccio è stato quello maggiormente seguito dagli scienziati nel primo periodo di sviluppo dell'intelligenza artificiale.

Il terzo approccio, invece, ha l'obiettivo di produrre un risultato che sia simile a quello prodotto da un soggetto umano. Diversamente però dalla scienza cognitiva, in questo caso non importa come ci si arrivi al risultato. Non si vogliono, cioè, replicare i passaggi ma importa raggiungere solo ed esclusivamente il risultato. A questo tipo di sistema può essere applicato il test di Turing. Questo tipo di test serve è stato sviluppato con l'intento di verificare se l'interlocutore con cui si sta interagendo sia una macchina o un umano.

L'ultimo tipo di approccio è, invece, riferibile a tutti quei sistemi che, portando a termine il compito assegnato, cercano di produrre un risultato che sia soddisfacente. Questo tipo di sistemi, quindi, non hanno l'obiettivo di raggiungere il risultato migliore ma hanno esclusivamente il compito di raggiungere il risultato.

Test di Turing

Come detto precedentemente, il test di Turing permette di verificare se l'interlocutore con cui si sta interagendo sia una macchina computazionale o un umano. Altro non è, quindi, che un test per verificare se la macchina può imitare o meno il comportamento umano tanto da essere "scambiata" per uno di loro.

Questo tipo di test è stato pensato e ideato da Alan Turing nel 1950. È stato quindi formulato ancora prima della nascita dell'intelligenza artificiale; nascita che è fissata al 1956.

Il matematico prese ispirazione da un gioco chiamato "gioco dell'imitazione". In questo gioco sono presenti tre partecipanti: un uomo, una donna e un terzo soggetto che è tenuto lontano dai due.

L'obiettivo del gioco è che il terzo soggetto deve cercare di capire quale dei due soggetti, che lui non vede, sia l'uomo e quale la donna. Ma in tutto ciò anche i soggetti hanno dei rispettivi scopi. Infatti, un soggetto dovrà mentire in modo da far effettuare una previsione errata al terzo soggetto, mentre l'altro partecipante dovrà aiutare il terzo soggetto fornendogli informazioni che gli permettano di effettuare una previsione corretta.

Naturalmente il terzo soggetto, l'intervistatore, non sa quale soggetto sia sincero e quale no e dovrà, quindi, formulare una previsione su chi sia l'uomo e chi la donna. Potrà fare ciò, sottoponendo i soggetti a delle domande a cui gli stessi risponderanno attraverso una macchina da scrivere/computer.

Il funzionamento del test di Turing è identico. Esistono però due differenti modi di eseguire questo test: a una o a due fasi.

Il classico test di Turing prevede l'articolazione dello stesso in due fasi. Ciò che differenzia le due fasi è la presenza della macchina; se infatti, nella prima fase tutti e tre i partecipanti sono umani nella seconda fase la macchina sostituisce uno dei due soggetti intervistati.

Si eseguirà, quindi, dapprima il test con sia l'intervistatore che gli intervistati esseri umani per poi cambiare uno di questi ultimi con una macchina. L'intervistatore dovrà capire chi dei due soggetti è un uomo e chi una donna. Si ripeterà n volte il test per poi calcolare il tasso di errore dell'intervistatore.

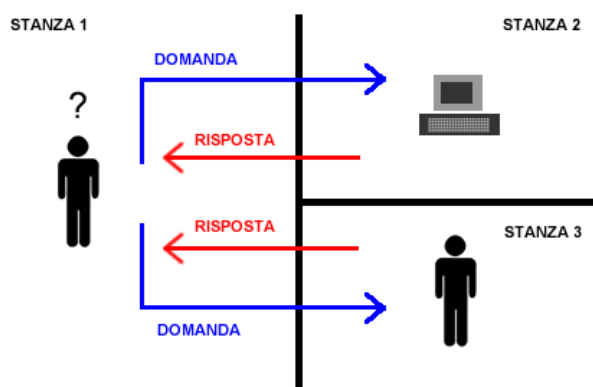


Figura 10: Seconda fase del Test di Turing; Fonte: <https://www.andreaminini.com/data/andreaminini/test-di-turing-5-andrea-minini.gif>

Si passa poi alla seconda fase che prevede come è possibile verificare in fig.10, la presenza della macchina come uno dei due intervistati. L'obiettivo in questo caso sarà quello di individuare quale sia l'uomo e quale la macchina. Si segue il medesimo procedimento della prima fase e si andrà a ricalcolare il tasso di errore per questa seconda fase.

Se la percentuale di errore della seconda fase del gioco è simile o inferiore a quella della prima fase, allora il Test di Turing viene superato e la macchina può essere categorizzata come intelligente.

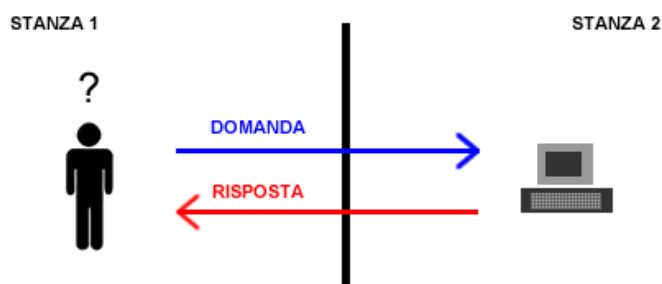


Figura 11: Test di Turing semplificato; Fonte: <https://www.andreaminini.com/data/andreaminini/test-di-turing-6-semplificato-andrea-minini.gif>

L'altra tipologia di test di Turing, visibile in fig.11, prevede, invece, un'unica fase in cui l'intervistatore deve capire se sta interloquendo con un soggetto umano o una macchina. Questa modalità pur essendo la più nota è anche la modalità in cui, in modo più probabile, una macchina anche "non intelligente" potrebbe riuscire ad ingannare l'umano.

Seppur il film "The imitation game" ha portato alla ribalta la conoscenza di questo test, è probabilmente più conosciuto, anche se inconsapevolmente l'anti test di Turing. L'anti test di Turing altro non è che il test Captcha.

Captcha è, infatti, l'acronimo di "Completely Automated Public Turing test to tell Computers and Humans Apart" ed è una misura di sicurezza volta a proteggere gli utenti dagli spam e dalle decriptazioni delle password. Questo test è definito "anti Turing" poiché il soggetto che sta compiendo una determinata azione per cui è richiesto il Captcha deve dimostrare di essere un umano e non una macchina.

Il test è alquanto semplice: al soggetto viene mostrata una stringa di testo, colorata e/o all'interno di un'immagine distorta, e il soggetto deve indicare quali lettere e/o numeri riconosce. Se indovina la stringa allora il test è stato passato dal soggetto che è dunque riuscito a dimostrare che lui è un umano.

Ma oltre a questa modalità più "classica" sono previste ulteriori varianti: individuare in un'immagine suddivisa in quadrati la presenza di un certo oggetto oppure utilizzare un file audio.

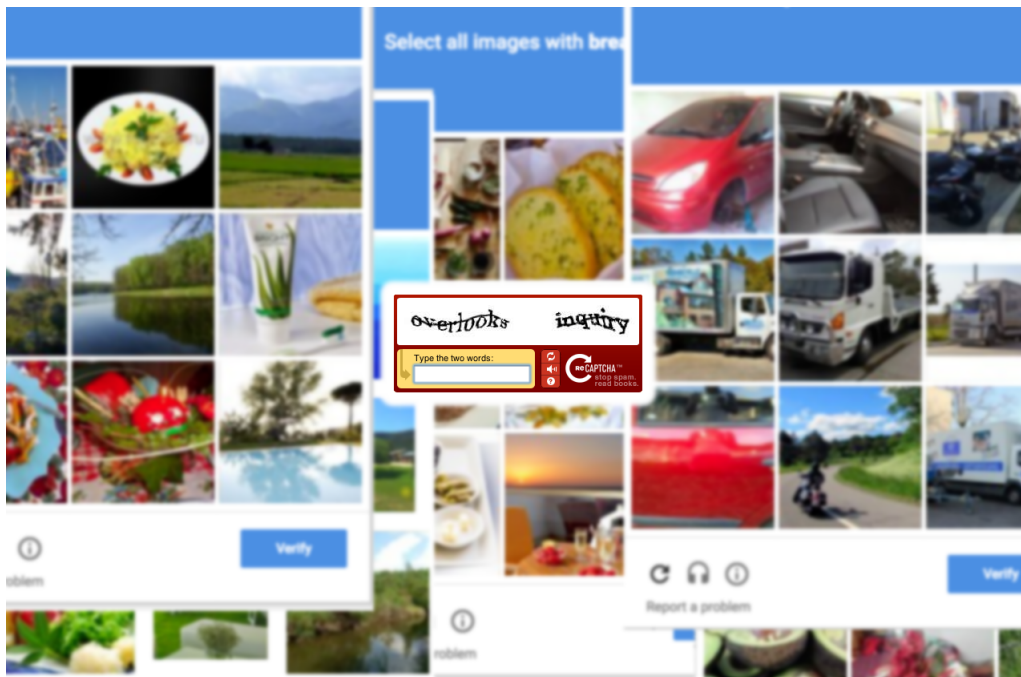


Figura 12: Esempi di Captcha; Fonte: <https://media.kasperskydaily.com/wpcontent/uploads/sites/89/2016/04/05234935/recaptcha-featured.png>

La nascita dell'AI e i primi sistemi di intelligenza artificiale

Come detto nel precedente capitolo elemento essenziale al fine di effettuare la previsione per la cancellazione è la costruzione di un modello che, imparando dai dati storici, effettui la predizione.

Tuttavia, non si può parlare di modelli di predizione senza effettuare un *excursus* sul percorso che ha portato ai moderni strumenti di *machine learning* e *deep learning*.

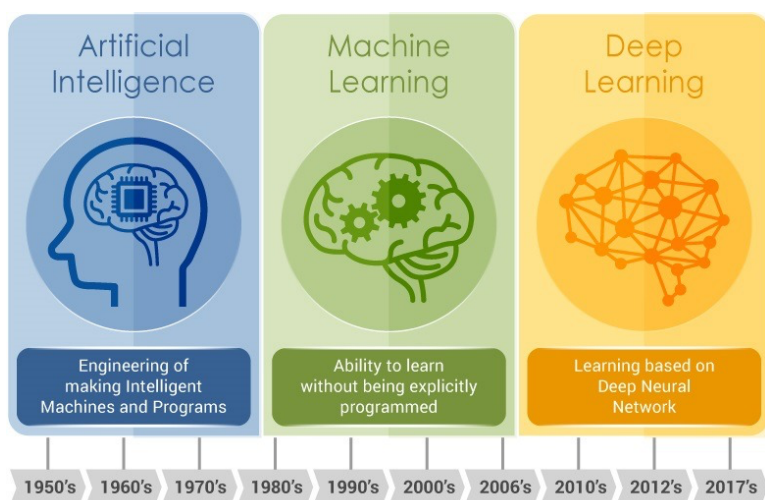


Figura 13: Linea del tempo delle tecniche di apprendimento automatico; Fonte: https://j3fau2wsnsolqkiarz7keon5-wpengine.netdna-ssl.com/wpcontent/uploads/2018/05/History_of_artificial_intelligence.jpg

Come si può vedere nella fig.13, l'intelligenza artificiale è nata negli anni '50, quando un gruppo di informatici iniziò a porsi la domanda se le macchine potessero pensare.

Il primo uso del termine "intelligenza artificiale" può essere attribuito a John McCarthy, che nel manifesto di una conferenza estiva di Dartmouth del 1956 disse:

“Lo studio deve procedere sulla base della congettura che ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza può in linea di principio essere descritta in termini tanto rigorosi da rendere possibile programmare una macchina in grado di simularli. Verrà fatto un tentativo di trovare come fare in modo che le macchine usino il linguaggio, formino astrazioni e concetti, risolvere tipi di problemi ora riservati agli umani, e migliorare sé stesse”

È alla data ufficiale di questa conferenza che possiamo attribuire la nascita della moderna intelligenza artificiale e dei suoi filoni di pensiero.

Per molti anni, in particolare fino agli anni '80, la corrente di pensiero predominante è stata quella di credere che le macchine potessero raggiungere il livello dell'intelligenza umana, e svolgere compiti umani, solo attraverso la creazione di molteplici regole scritte, una per una, dagli sviluppatori.

Questo periodo è stato, poi, denominato: GOFAI cioè *Good Old-Fashioned Artificial Intelligence*.

Questo periodo, e i suoi esponenti, seguivano come corrente di pensiero la logica simbolica.

Per spiegare la logica simbolica si può far riferimento ad un ragionamento molto semplice, sviluppato nel IV secolo a.C. da Aristotele, e allo stesso tempo celebre, che è:

“Socrate è un uomo”,

“Tutti gli uomini sono mortali”

→ *“Socrate è mortale”*

È nello stesso anno della nascita dell'IA che Herbert Simon e Allen Newell presentano il primo vero programma di IA, cioè: il *Logic Theorist*.

Il programma costruito sotto forma di albero aveva lo scopo di dimostrare alcuni principi matematici. In particolare, si partiva dalla radice, in cui era presente la tesi, per poi attraverso i nodi, che presentavano le ipotesi matematiche, andare a dimostrare il teorema attraverso una serie di proposizioni logiche.

Già da questo programma si iniziò a capire il grande potenziale dell'intelligenza artificiale che permetteva ad una macchina non solo di effettuare semplici calcoli ma di effettuare delle dimostrazioni di complicati principi matematici.

Ulteriore svolta nel campo è stata data da Arthur Samuel. Questo è noto per aver sviluppato il primo gioco di dama in cui la macchina è stata in grado di sconfiggere un avversario umano.

Tuttavia, è alla fine degli anni '60 che l'AI inizia ad avere alcuni fallimenti. Infatti, l'AI simbolica sebbene si dimostrò adatta a risolvere problemi logici ben definiti, come il gioco degli scacchi, non riusciva a risolvere problemi più complessi, come ad esempio la classificazione delle immagini o la traduzione di un testo o una frase.

Per quanto il sistema del linguaggio dei simboli sia un campo ormai per lo più abbandonato dalla ricerca, essendosi sviluppati modelli più avanzati, c'è ancora una branca dell'AI che utilizza questo tipo di approccio: il *planning*.

Questo sottocampo dell'intelligenza artificiale si occupa di affrontare problemi che richiedono la formulazione di una serie di passi per realizzare un certo risultato.

Esempi di applicazione sono ad esempio il fornire linee guida su come cucinare delle ricette o come raggiungere una destinazione o ancora come assemblare dei pezzi per costruire un certo oggetto.

Si ha, quindi, uno stato iniziale ed uno o più stati finali desiderati. Il sistema opererà e formulerà poi un'indicazione dei passaggi in modo da minimizzare il numero di passaggi o il tempo richiesto.

In tutto ciò, c'è però un problema. Infatti, il numero dei possibili passaggi che possono portare alla soluzione finale può essere molto elevato e di conseguenza, potrebbe non essere possibile esaminare tutte le opzioni disponibili. Si parla in questo caso di esplosione combinatoria.

L'esplosione combinatoria può essere risolta facendo ricorso alle euristiche.

Un esempio di queste è la *Greedy Heuristics* cioè "euristica avida". Quest'euristica prevede di selezionare ad ogni passaggio l'azione che nell'immediato dà il guadagno maggiore.

A seguito dei primi fallimenti dei sistemi basati sul linguaggio dei simboli, negli anni '80, sorse una nuova classe di sistemi basati sulla conoscenza e per questo definiti "sistemi di conoscenza" o anche "sistemi esperti". L'idea di fondo di questo tipo di sistemi era quella di "catturare" e riassumere la conoscenza e competenza umana in una forma computabile.

In questo modo si voleva rendere la conoscenza umana su un certo fenomeno disponibile a tutti ed anche in modo più economico.

L'idea di base era che tutti i soggetti, anche i meno abbienti, potevano ricevere una sorta di consulenza da un esperto in modo più economico.

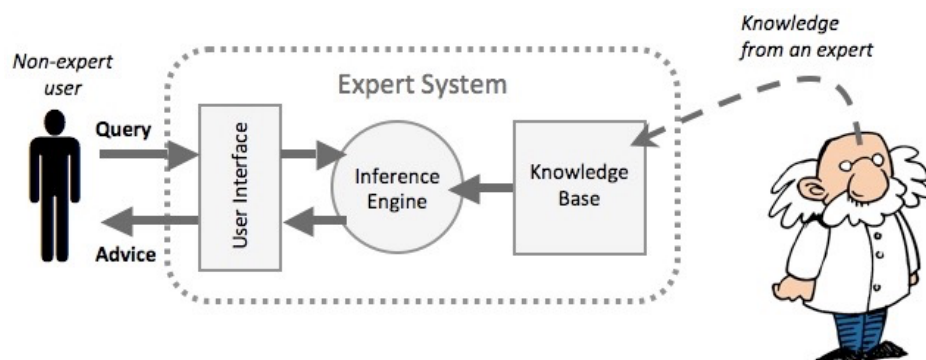


Figura 14: I sistemi esperti; https://miro.medium.com/max/635/1*FwCvqrUP2djtz2Xk4NFIpA.png

Come si può vedere dalla fig. 14, si raccoglieva la conoscenza del soggetto esperto attraverso questionari o interviste e le si racchiudevano poi all'interno di una prima componente del sistema, denominata "*knowledge base*". La *knowledge base* altro non è che una collezione di fatti, regole e relazioni su uno specifico dominio di interesse rappresentato in forma simbolica (Mokbel, 2021).

È poi presente l'*Inference Engine* cioè il motore di inferenza. Il motore di inferenza serve a descrivere come manipolare e combinare questi simboli (Mokbel, 2021).

Sia i fatti che le regole erano rappresentati in modo esplicito, e questo permetteva un più facile e rapido aggiornamento oltre che una maggior facilità di modifica dello stesso anche quando veniva incorporata nuova conoscenza.

L'intero funzionamento del sistema esperto segue la regola "*if-then*", cioè: se succede questo/se si ha questo allora si ha questa conseguenza/si è in questa situazione.

Uno dei primi sistemi esperti che è stato sviluppato è stato il sistema DENDRAL. Questo sistema è stato sviluppato da Edward Feigenbaum, ed aveva l'obiettivo di mappare la struttura molecolare del composto partendo dall'analisi spettrale della molecola che veniva fornita.

Facendo di nuovo riferimento alla storia, negli anni '60 fu sviluppato un primo modello di rete neurale artificiale, denominato *Perceptron*.

Questo era composto unicamente da due strati: quello di entrata e di uscita. La rete era allenata sfruttando l'algoritmo di *back-propagation*. L'algoritmo, che rappresenta quindi un algoritmo di apprendimento, confronta il valore in uscita del sistema con il valore obiettivo, cioè quello in ingresso, e sulla base della differenza di questi valori, calcola la funzione di perdita ed assegna i pesi in modo che i valori in uscita convergano verso i valori in entrata.

Si calcola, quindi, l'errore presente nella rete ed in base a questi, l'algoritmo andrà a modificare i pesi sinaptici della rete neurale.

Tuttavia, visti gli studi ancora nelle prime fasi, in un articolo del 1967 Minsky e Papert definirono le reti neurali inadeguate ad ogni applicazione pratica (Kelemen, 2007). Questo, unito ai vari studi ed esperimenti fallimentari portarono ad abbandonare le reti neurali che furono riprese solo verso la fine degli anni '80.

Nel 1980, c'è stato, invece, il passaggio dall'intelligenza artificiale pura al *machine learning*. Sebbene, questi due fenomeni sono spesso confusi e i relativi nomi usati in modo alternativo, tra i due esiste una differenza ben marcata.

L'obiettivo del *machine learning* è infatti "quello di progettare algoritmi che possano migliorare automaticamente attraverso l'esperienza" (Jonas Wälter, 2020).

La differenza con l'intelligenza artificiale è quindi abissale.

Nei primi anni 2000 si è poi fatto un ulteriore passo avanti. Si è passati ai cosiddetti modelli di *deep learning*.

Il *deep learning* basa il proprio funzionamento su una serie di strati successivi di rappresentazioni sempre più significativi. L'apprendimento profondo coinvolge decine di strati successivi di rappresentazioni che sono tutti appresi automaticamente dall'esposizione ai dati di addestramento attraverso le reti neurali.

Il *deep learning* è riuscito a risolvere anche i problemi più complessivi proprio attraverso questa serie di strati successivi di rappresentazioni.

La profondità del modello sarà data dal numero di strati di cui è composto il modello. È chiaro che all'aumentare del numero di strati aumenterà anche la profondità del modello.

Il Machine Learning

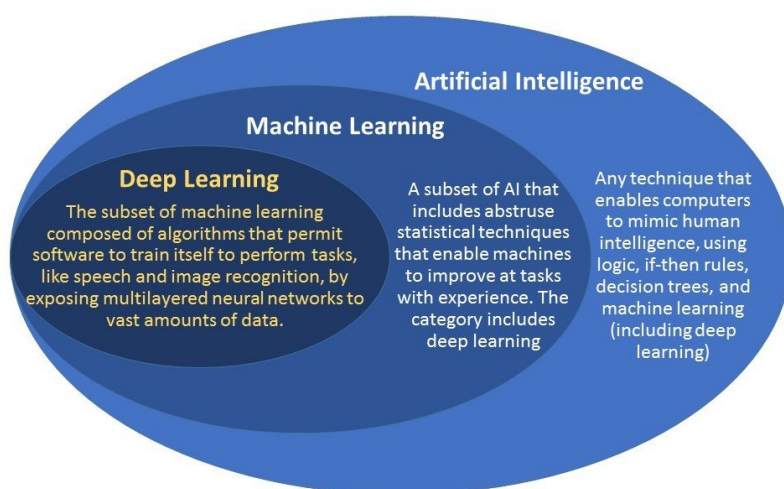


Figura 15: Differenza tra Artificial Intelligence, Machine Learning, Deep Learning; Fonte: <https://hcs-pharma.com/wp-content/uploads/2019/07/MLIADL.jpg>

Come si può vedere in fig. 15, ciò che differenzia il ML dall'intelligenza artificiale è proprio l'obiettivo perseguito. Se, infatti, l'obiettivo del *machine learning* (ML) è come detto poco fa quello di progettare algoritmi che apprendono autonomamente, l'obiettivo dell'*artificial intelligence* è quello di costruire algoritmi che siano in grado di pensare ed agire come un essere umano ma non ci si pone l'obiettivo o il problema dell'apprendimento autonomo dello stesso.

Quindi, come si può vedere sempre nella fig. 15, il *machine learning* è un sottoinsieme dell'intelligenza artificiale. Se, infatti, gli algoritmi di *machine learning* fanno parte della più grande categoria degli algoritmi di *artificial intelligence* la stessa cosa non si può dire per gli algoritmi di *artificial intelligence*.

Nell'intelligenza artificiale è il programmatore ad immettere regole in funzione delle quali l'algoritmo analizza e lavora sui dati immessi per poi produrre delle risposte. Questo processo lo si può vedere dalla fig.7. Come è possibile verificare, il primo passo da compiere è quello della scrittura delle regole dopo di che si valuta il modello e a seconda del risultato che si ottiene, soddisfacente o meno, si lancia l'algoritmo oppure si analizzano gli errori dello stesso per cercare di modificare e/o adattare il modello.

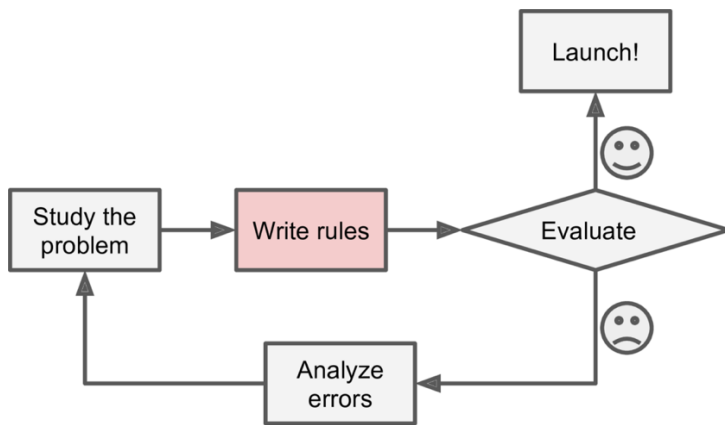


Figura 16: Programmazione tradizionale; Fonte: https://pankajchaudhary5.files.wordpress.com/2019/09/mlst_0101.png?w=1024

Nell'apprendimento automatico, invece, il programmatore immette i dati, i risultati attesi da quei dati e l'algoritmo in base a questi si crea autonomamente delle regole.

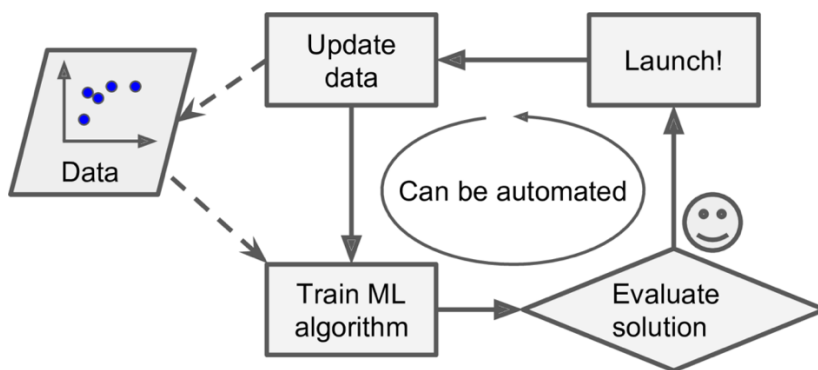


Figura 17: Machine learning; Fonte: https://pankajchaudhary5.files.wordpress.com/2019/09/mlst_0103.png?w=1024

Come si può vedere dalla fig.17, l'input di partenza del sistema di *machine learning* sono i dati. In base a questi ed in particolare in base al *training set* si addestra l'algoritmo che poi verrà valutato. Come già detto, in questo caso sarà l'algoritmo stesso a costruire le proprie regole senza che queste vengano scritte dal programmatore. Dopo di che, naturalmente, occorrerà valutare l'algoritmo e i relativi risultati. Se il controllo sarà positivo l'algoritmo verrà lanciato.

Ciò che differenzia però questo caso dal precedente sono i dati aggiornati. Se infatti nel modello tradizionale qualora si aggiungessero nuovi dati occorrerebbe riscrivere le regole e riaddestrare il modello, nel *machine learning* questo processo avviene autonomamente. Infatti, l'algoritmo apprenderà e scriverà nuove regole per gestire i nuovi dati.

Uno degli esempi maggiormente conosciuti, ed anche uno dei primi esempi maggiormente applicati di *machine learning*, è il filtro antispam.

Il filtro antispam è anche l'esempio perfetto per spiegare il funzionamento del ML.

Per la costruzione dell'algoritmo si parte da un *training set* in cui sono contenute varie mail classificate come spam o meno; in base a queste, l'algoritmo costruisce le proprie regole.

Di conseguenza, andrà a classificare come e-mail spam, le mail contenenti alcune parole che potrebbero essere contenute in queste; ad esempio, la parola “free”.

Tuttavia, è logico pensare che coloro che vogliono inviare e-mail spam notando che l’inserimento di alcune parole produce la classificazione in spam potrebbero decidere di evitare di inserire quella specifica parola.

Se l’algoritmo non apprendesse autonomamente le nuove tipologie di mail spam, queste, in seguito, non sarebbero classificate come tali.

Invece, visto che l’algoritmo di *machine learning* apprende autonomamente, anche grazie alla classificazione e segnalazione della mail operata dai soggetti, le regole si aggiornano e di conseguenza si aggiorneranno anche i criteri e le parole che allertano l’algoritmo.

Nel caso del filtro spam, per esempio, dopo che una serie di persone ha etichettato delle mail contenenti una certa parola allora l’algoritmo inizierà ad indicare mail contenenti quella parola come probabili spam.

In questo modo, quindi, l’algoritmo ha aggiornato il proprio funzionamento sulla base dei nuovi dati.

In sostanza, l’apprendimento automatico è ottimo per risolvere problemi che richiederebbero un lungo elenco di regole, come l’esempio della spam. Ma anche problemi complessi per i quali l’approccio dell’intelligenza artificiale non produce un risultato soddisfacente.

Le differenti tipologie di Machine Learning

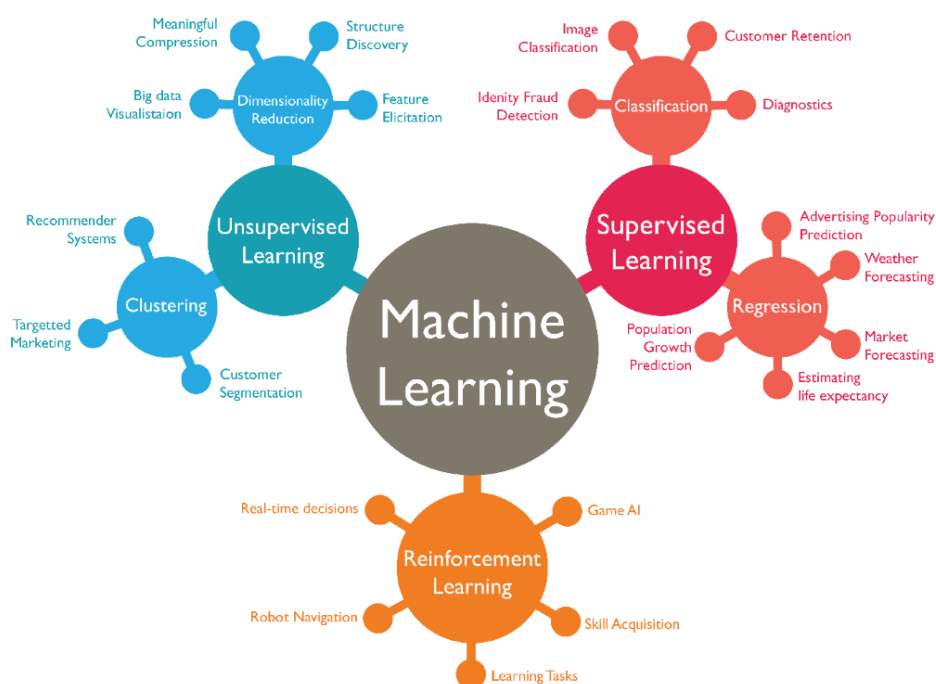


Figura 18: I differenti tipi di ML; Fonte: <https://static.cwi.it/wp-content/uploads/2020/01/algoritmi-machine-learning-1024x732.png>

Come è possibile verificare in fig. 18, non esiste un’unica tipologia di machine learning ma questo si può dividere in tre grandi filoni.

- 1) Supervised Learning
- 2) Unsupervised Learning
- 3) Reinforcement Learning

Ciò che differenzia queste differenti tipologie di apprendimento automatico è la diversa tipologia di apprendimento e addestramento a cui si sottopone il modello.

Metodi supervisionati

Nel *machine learning* supervisionato si ha l'intervento del programmatore che etichetta i dati. In particolare, il programmatore inserisce i dati nella macchina per poi fornire anche le etichette. Con l'assegnazione delle etichette si intende che il programmatore per ogni variabile *input* fornisce il risultato, che è rappresentato dall'etichetta. In questo modo, con la presenza per ogni dato del valore *input* e *output* la macchina impara e inizia ad elaborare un proprio modello.

Il compito della macchina sarà quindi quello di dover stimare una funzione ipotesi che approssimi nel migliore dei modi, e anche nel modo meno astruso, i punti generati dal processo di etichettatura. Sarà, infatti, presa in considerazione la funzione più semplice che meglio si adatta ai dati. Questo è il cosiddetto *Ockham's razor*. Dopo di che, l'algoritmo potrà essere utilizzato al fine di effettuare previsioni sui dati futuri.

Come è immaginabile però il lavoro di addestramento attraverso questa modalità, cioè attraverso l'inserimento da parte del programmatore di tutte le etichette, è molto costoso a livello di tempo.

Le tecniche di *machine learning* supervisionate possono essere utilizzate per risolvere due differenti problemi: i problemi di classificazione e i problemi di regressione.

L'obiettivo perseguito quando si deve risolvere un problema di classificazione è quello di suddividere i dati in due o più classi. Si cerca, cioè, di categorizzare le osservazioni per capire a quale classe appartengano.

La classificazione può essere lineare o non lineare; ad entrambe le tipologie di classificazione corrispondono pro e contro.

La classificazione lineare, basandosi su un processo di apprendimento alquanto semplice, è molto veloce nell'operare la classificazione, tuttavia, si potrebbe incorrere nell'*underfitting*. L'*underfitting* si ha quando si prendono in considerazione pochi parametri per operare la classificazione, proprio per assicurare un processo semplice e veloce, ma in questo modo si generano molti *bias*. Si hanno cioè osservazioni classificate in maniera errata.

Sono, invece, più precisi i classificatori non lineari. Questi proprio prendendo in considerazione molteplici parametri e funzioni, anche non lineari appunto, riescono a compiere una classificazione più precisa. Il problema in questo caso è la lentezza del processo di classificazione e l'*overfitting*. Con *overfitting* si intende, invece, indicare la situazione in cui il modello, essendo molto sensibile ai dati con i quali è stato addestrato, potrebbe cambiare classificazione ad un medesimo punto ogni qualvolta che si modificano i dati di training, cioè i dati con i quali si è addestrato il modello.

Ulteriore problema che può essere risolto attraverso le tecniche di apprendimento supervisionato è la regressione.

Attraverso la regressione si può, studiando la relazione esistente tra variabili dipendenti ed indipendenti ed effettuando un'interpolazione dei dati, stimare il valore della variabile dipendente.

Anche la regressione può essere eseguita seguendo una metodologia lineare o non lineare così come accadeva per la classificazione. Anche in questo caso, i problemi legati alla scelta dell'uno o dell'altro metodo sono identici; si avranno, infatti, i problemi legati all'*overfitting* e all'*underfitting*.

Ma la regressione, in particolare la regressione logistica, può essere utilizzata anche al fine di risolvere un problema di classificazione.

Infatti, la regressione logistica può essere adoperata per prevedere il valore di una variabile dipendente dicotomica. Andando a prevedere il valore di una variabile dicotomica non si fa, infatti, altro che classificare quella variabile.

Il valore della predizione sarà infatti dato dalla probabilità che la variabile dipendente appartenga ad una determinata categoria e quindi classe.

Metodi non supervisionati

In questo caso, diversamente dal caso esposto precedentemente, nei modelli non supervisionati non c'è alcun intervento umano volto ad etichettare i dati di addestramento. Quindi, non si fornisce un *target*, cioè l'*output* corrispondente, ad ogni singolo *input* immesso nel sistema.

Si devono, quindi, esplorare interamente i dati per capire le strutture intrinseche presenti negli stessi.

Questi tipi di algoritmi di *machine learning* sono utilizzati o per effettuare la *cluster analysis* oppure per effettuare una riduzione delle dimensioni dei dati.

L'obiettivo dell'analisi di clusterizzazione è quello di raggruppare le osservazioni in gruppi in modo che le osservazioni appartenenti allo stesso cluster siano tra loro simili. Si vuole, quindi, raggruppare le osservazioni in base alla loro similarità in modo quindi che tutte le osservazioni che sono tra loro simili si trovino in uno stesso cluster.

Si ha una corretta clusterizzazione quando si minimizza la distanza *intra-cluster* e si massimizza la distanza *inter-cluster*. Quando, in pratica, la distanza tra le osservazioni appartenenti allo stesso cluster è minimizzata mentre è massimizzata la distanza esistente tra i vari differenti cluster.

Si parla in termini di distanza poiché il metodo della *cluster analysis* basa il proprio funzionamento e la successiva divisione in gruppi proprio sulla base della distanza esistente tra le varie osservazioni.

È chiaro che, più due osservazioni saranno tra loro distanti più queste saranno differenti tra di loro non presentando molte similarità.

Per quanto riguarda la distanza si possono scegliere molteplici differenti formule di calcolo della stessa, distanza di Lagrange o distanza media quadratica, anche se la più utilizzata è quella Euclidea.

Esistono differenti metodi da poter seguire quando si esegue un processo di clusterizzazione.

Infatti, si può scegliere un metodo partizionale o un metodo gerarchico.

Il metodo gerarchico produce un insieme di cluster annidati che possono essere visualizzati con l'utilizzo del dendrogramma. Il dendrogramma altro non è che un diagramma ad albero che evidenzia tra quali osservazioni è avvenuta la clusterizzazione e a che altezza. Inoltre, quando si esegue un cluster gerarchico si possono optare per due differenti tipologie di cluster: agglomerativo e divisivo. Con la prima tipologia di algoritmo di clusterizzazione si parte da una situazione in cui tutte le osservazioni sono separate fino ad arrivare al caso limite opposto cioè quello in cui tutte le osservazioni fanno parte di un unico cluster. Nella clusterizzazione divisiva invece è l'esatto opposto, si parte da una situazione in cui tutte le osservazioni appartengono al medesimo cluster per arrivare alla situazione in cui tutte le osservazioni sono singole.

Ma tornando alla divisione precedente, cioè metodo gerarchico o partizionale, ciò che differenzia questi è la diversa tempistica con cui occorre decidere il numero dei cluster.

Se, infatti, nel metodo gerarchico il numero dei cluster lo si può decidere dopo che si è già effettuato il processo di clusterizzazione e quindi prendendo come aiuto nella decisione anche il dendrogramma, nel cluster partizionale il numero dei cluster deve essere deciso prima dell'analisi.

In quest'ultimo caso occorre, infatti, per iniziare l'analisi fissare un numero "k" di cluster desiderati, le clusterizzazioni saranno poi eseguite fin quando non si raggiungerà il numero k prescelto.

Nonostante il problema del dover fissare a priori il numero di cluster, questo tipo di clusterizzazione è maggiormente indicata in caso di *dataset* molto grandi per i quali la costruzione di una struttura gerarchica risulterebbe lunga e computazionalmente complessa.

La *cluster analysis* ha applicazioni in ogni campo.

Un campo applicativo è, ad esempio, quello della comunicazione, poiché la *cluster analysis*, grazie alla divisione in gruppi in base alla similarità, permette di azionare campagne pubblicitarie maggiormente mirate e personalizzate.

Esempio aggiuntivo e forse di più facile visione è Netflix. Quando, infatti, la piattaforma di *streaming online* suggerisce un contenuto da vedere altro non fa che consigliare un contenuto che è piaciuto a persone che appartengono al nostro stesso *cluster*.

Ulteriore problema risolvibile e rientrante nei metodi di *machine learning* non supervisionati è la riduzione della dimensionalità.

Visti i nuovi e sempre maggiormente aggiornati mezzi di informazione, ogni fenomeno che si analizza è descritto e caratterizzato da un'enorme mole di dati. È chiaro, quindi, che per analizzarlo sarà necessario non solo molto tempo ma anche mezzi molto avanzati.

Inoltre, i *patterns* nei dati potrebbero essere in parte nascosti o comunque non immediatamente visibili. Per questo motivo si può ricorrere alla riduzione della dimensionalità.

Quest'ultima permette di ridurre la mole informativa inerente un certo fenomeno senza però comprometterne eccessivamente la qualità. La riduzione dovrebbe, infatti, riguardare esclusivamente quelle informazioni che non sono rilevanti per l'analisi che si sta compiendo.

Ma la riduzione non implica esclusivamente l'eliminazione di dimensioni non rilevanti ma anche la combinazione di tutte quelle informazioni che sono correlate e ridondanti.

Queste operazioni di riduzione possono essere compiute attraverso due differenti metodologie: estrazione delle caratteristiche e selezione delle caratteristiche. La prima considera tutte le informazioni originali ma le trasforma, combinandole in modo lineare e non, in modo che la dimensione finale sia minore.

La selezione delle caratteristiche, invece, seleziona unicamente un sotto-insieme delle informazioni originali. Se, quindi, il primo metodo permette di considerare, anche se in modo ridotto, tutte le caratteristiche, quest'ultimo consente di sceglierne soltanto un piccolo sotto-insieme.

Tuttavia, anche la trasformazione delle caratteristiche ha dei punti negativi.

Con la trasformazione delle caratteristiche, infatti, si perdono tutte le unità di misura delle caratteristiche stesse. Se quindi, si vogliono estrarre regole significative che prendono in considerazione anche le unità di misura, si può procedere unicamente con la selezione delle caratteristiche.

Inoltre, altra occasione in cui si può effettuare unicamente la selezione delle caratteristiche è quando si ha a che fare con caratteristiche non numeriche; cosa molto probabile nel campo dell'apprendimento automatico.

Metodi semi-supervisionati

Come già detto, le tecniche di *machine learning* supervisionate basandosi sull'etichettatura effettuata dal programmatore sono molto costose in termini di tempo e di conseguenza anche di costo.

Per questo, volendo sempre in parte effettuare un processo di etichettatura, si può ricorrere ai metodi semi-supervisionati.

Il metodo semi-supervisionato permette, infatti, di etichettare solo una parte dei dati di addestramento, lasciando l'altra parte da etichettare.

In questo modo il modello inizialmente si addestra e lavora sulla parte dei dati etichettati, cioè sui dati di cui sa a priori l'output per ogni input, per poi procedere l'addestramento sulla restante parte dei dati. Per quanto riguarda quest'ultima fase l'algoritmo procederà a lavorare per inferenza.

Tuttavia, prima di procedere con l'esecuzione dei dati, ed applicare l'inferenza, occorre verificare se siano soddisfatti i seguenti tre presupposti (Nuzzo, 2021):

- 1) Presupposto di continuità, i punti che sono vicini hanno una più alta probabilità di rappresentare lo stesso output
- 2) Presupposto del cluster, i dati appartenenti allo stesso cluster hanno una più alta probabilità di rappresentare lo stesso output
- 3) Presupposto molteplice, i dati sono rappresentativi anche in uno spazio di dimensioni inferiore a quello di input, cioè si può effettuare la previsione anche con meno colonne di quelle presenti nel dataset

Reinforcement Learning

Nei metodi di *machine learning* con e senza supervisione, la macchina impara dai dati che possono essere a seconda del differente tipo di metodo etichettati o meno.

Nel *reinforcement learning*, invece, la macchina impara dall'ambiente e dall'interazione con lo stesso.

In particolare, l'addestramento avviene con il modello di AI che esegue delle azioni e riceve in cambio dei premi o delle sanzioni. Riceverà dei premi quando l'azione effettuata risulta essere giusta e porta ad un output corretto, mentre riceverà una sanzione quando l'azione svolta risulta essere errata.

Il rinforzo quindi altro non è che il segnale di ricompensa che la macchina riceve a seguito di un'azione.

Di conseguenza, non sarà necessario eseguire la raccolta e la successiva pre-elaborazione dei dati, come invece è essenziale nelle altre tipologie di *machine learning*, ma sarà necessario che la macchina operi ed effettui moltissime interazioni con l'ambiente poiché esclusivamente in questo modo questa può addestrarsi ed imparare.

Le azioni corrette, che l'agente deve scegliere in una determinata situazione, vengono chiamate *policy*. Le *policy* però non sono imposte dall'esterno ma è l'agente di AI che deve capirle attraverso l'interazione con l'ambiente in particolare con il processo definito *trial and error*. Quest'ultimo, altro non rappresenta che l'insieme delle azioni messe in campo dalla macchina di AI per raggiungere un certo obiettivo, il tutto cercando di massimizzare il segnale di ricompensa.

Se vengono fissati correttamente sia le sanzioni che i premi la macchina sarà in grado di imparare autonomamente senza necessitare di alcun intervento umano.

Si utilizza questo approccio quando si ha a che fare con ambienti molto incerti e variabili non prevedibili.

Vista la capacità di imparare in modo autonomo e la possibilità di essere adoperato per ambienti incerti il *reinforcement learning* si dimostra ottimo per i sistemi di guida autonoma.

Differenti metodologie di apprendimento dell'AI

Finora si è effettuata la distinzione riguardante l'addestramento dei sistemi di ML.

Si è, in particolare, effettuata la distinzione tra i sistemi supervisionati, non supervisionati, semi supervisionati e di *reinforcement learning*.

Tuttavia, si può effettuare un'ulteriore distinzione che riguarda la differente capacità del sistema di *machine learning* di imparare in modo incrementale.

Possiamo individuare, infatti, sistemi basati sull'*online learning* e sistemi basati sul *batch learning*.

Attraverso l'*online learning* l'AI impara continuamente ed autonomamente dai dati che riceve. Sarà quindi un'ottima metodologia di apprendimento in tutti quei casi in cui il sistema riceve i dati in modo continuo. L'applicazione dell'apprendimento online è indicata in tutti quei sistemi che hanno la necessità di adattarsi rapidamente e che sono quindi riferiti ad aree merceologiche in continuo cambiamento. Tanto che uno dei campi in cui maggiormente è utilizzato questo tipo di apprendimento è il campo finanziario.

Il problema è che imparando in modo continuativo dai dati che riceve, l'AI potrebbe apprendere in modo errato se nei dati che riceve sono presenti degli errori.

L'algoritmi di apprendimento online, con la continua ricezione di nuovi pacchetti di dati, può anche essere utilizzato per addestrare dei sistemi con una mole di dati molto elevata ma senza che questi siano presenti nella memoria principale della macchina.

Se però un pacchetto di dati, ricevuto dalla macchina, presenta errori al suo interno è chiaro che l'AI, basando l'apprendimento su questi dati, imparerà in modo errato.

Per questo motivo occorre monitorare sempre il sistema poiché in caso di riduzione delle prestazioni dovute ad errori, si può bloccare l'apprendimento e ritornare allo stato di apprendimento precedente all'immissione dei nuovi dati.

Inoltre, l'aggiornamento dell'apprendimento lo si può regolare anche attraverso il *learning rate*. Il *learning rate* è un parametro che rappresenta il peso attribuito ai nuovi dati, esso indica, in pratica, la velocità con cui il sistema di apprendimento si adatta ai nuovi dati che riceve.

Colui che ha impostato il sistema di AI può scegliere se dotarlo di un *learning rate* alto o basso, anche in base al tipo di dati che il sistema riceverà.

Ad un alto *learning rate* corrisponderà un adattamento rapido del sistema ai nuovi dati, cioè il sistema darà molto peso ai nuovi dati e di conseguenza cambierà anche di molto ciò che ha appreso fino a quel momento poiché tenderà a "dimenticare" molto più velocemente i vecchi dati.

Viceversa, in presenza di un *learning rate* basso; il sistema sarà dotato di maggior inerzia e dunque imparerà più lentamente al sopraggiungere dei nuovi dati.

Ulteriore metodologia di apprendimento è rappresentata dal *batch learning*. In questo caso il sistema non è progettato per apprendere in modo continuativo con la ricezione di nuovi dati. Esso, infatti, apprenderà solo in fase di addestramento.

Se si vuole aggiornare il sistema con ulteriori dati rispetto a quelli con cui è stato costruito, occorrerà ricostruire e riaddestrare il sistema partendo da zero. Non si può, quindi, aggiornare il sistema esclusivamente con i nuovi dati ma occorrerà, in pratica, costruire un altro sistema e addestrarlo con i vecchi e i nuovi dati.

Questo tipo di apprendimento non sarà quindi indicato nelle aree merceologiche che sono incerte e in continua evoluzione; poiché, appunto, l'aggiornamento di un sistema è un processo lungo e costoso.

Nell'apprendimento basato sull'istanza, l'*instance based learning*, il sistema memorizza e impara dagli esempi, per poi generalizzare ai nuovi casi utilizzando una misura di somiglianza.

Il sistema confronta quindi i dati appartenenti al *training set* con cui è stato addestrato con i nuovi dati a disposizione e li classifica in base alla loro reciproca somiglianza.

In questo tipo di apprendimento il sistema assegna le etichette ai nuovi dati in base alla vicinanza degli stessi con gli elementi già etichettati.

Ultima metodologia di apprendimento e aggiornamento di un sistema è il *model based learning*.

Questa metodologia consiste nel costruire e addestrare un modello con degli specifici dati, rappresentati dal *training set*, per poi effettuare, attraverso il modello costruito, previsioni sui nuovi dati.

In pratica, si costruisce il modello e poi si sottopongono i nuovi dati allo stesso al fine di effettuare delle previsioni sui nuovi dati.

Un modello è una versione semplificata della realtà. Proprio perché è una versione semplificata occorre, quando si costruisce un modello, non inserire troppe variabili nello stesso poiché questo si complicherebbe inutilmente aggiungendo dettagli ed informazioni non utili al caso.

Inserire, infatti, anche informazioni molto specifiche e superflue potrebbe portare a una maggiore difficoltà di generalizzazione. Quindi, il modello pur funzionando bene con i dati del *training set*, cioè con i dati con i quali è stato costruito, non generalizzerebbe e di conseguenza non funzionerebbe in modo corretto in presenza di nuovi dati.

CAPITOLO 4

Al fine di mostrare realmente come applicare modelli di *machine learning* per prevedere la cancellazione della domanda e, conseguentemente, capire se l'applicazione di strategie di marketing evita o perlomeno diminuisce questa probabilità ho applicato questi procedimenti su dati realmente provenienti da una realtà alberghiera.

Il *dataset* è stato scaricato dal sito Kaggle e si riferisce ad uno studio effettuato nel 2019 da Nuno Antonio et al. e pubblicato sul *Data Science Journal* con il titolo “*An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations*”.

Il dataset preso in esame si riferisce ad una catena alberghiera portoghese che ha condiviso i dati relativi a due differenti hotel 4 stelle: un resort hotel ed un hotel situato in città.

I dati si riferiscono all'arco temporale luglio 2015 – agosto 2017 e provengono direttamente dal *Property Management System* (PMS) della catena alberghiera.

Come già detto anche nei precedenti capitoli il PMS è il programma utilizzato dalle strutture alberghiere per gestire le prenotazioni, il *planning* delle camere così come anche la fatturazione.

In particolare, nel PMS si possono anche annotare le caratteristiche legate al cliente come, ad esempio, la nazionalità o attraverso quale strumento esso ha prenotato. Si utilizzerà lo stesso anche per annotare se la prenotazione è stata cancellata, se c'è stato il *check-in* o ancora se c'è stato un *no-show*.

Il *no-show*, come già anticipato, lo si ha quando il soggetto pur non disdicendo la prenotazione non si presenta comunque in struttura il giorno di arrivo previsto da prenotazione.

Descrizione del dataset

Il *dataset* è composto da 119.390 osservazioni e 32 variabili. Sono presenti alcuni valori mancanti, classificati sul *file csv* e su R come *na*. Vista la presenza dei *missing value* nel *dataset*, che potrebbero compromettere la corretta analisi e visualizzazione dei dati, ho effettuato un'analisi di pulizia relativamente a questi valori.

Ho effettuato ciò attraverso la funzione *na.omit* che crea un nuovo oggetto contenente tutti i dati del dataset originario ad eccezione delle osservazioni con valori mancanti. In particolare, il nuovo dataset, avrà 119.386 osservazioni; di conseguenza, le osservazioni eliminate sono state 4.

Di seguito l'elenco, in ordine di presenza sul *file csv*, e la descrizione delle variabili presenti nel *dataset*.

- *Hotel*: variabile *character*, presenta unicamente due valori: *city* o *resort*. Questa variabile indica, quindi, se quell'osservazione si riferisce all'hotel di città o al *resort*.
- *Is canceled*: variabile categorica, presenta unicamente due valori: 1 (prenotazione cancellata) o 0 (prenotazione non cancellata).
- *Lead time*: variabile *integer*, indica quanti giorni sono trascorsi dal momento dell'inserimento nel sistema della prenotazione al momento di arrivo in struttura. Indica, di conseguenza, con quanti giorni di anticipo è stata effettuata la prenotazione.
- *Arrival date year*: variabile *integer*, indica l'anno di arrivo del cliente in struttura.

- *Arrival date month*: variabile categorica, indica il mese di arrivo in struttura. Il mese, come è intuibile dalla natura della variabile, è indicato non in formato numerale ma letterale.
- *Arrival date week number*: variabile *integer*, indica il numero della settimana in cui si ha l'arrivo del cliente in struttura
- *Arrival date day of month*: variabile *integer*, indica il numero del mese in cui è previsto l'arrivo del cliente in struttura
- *Stays in weekend nights*: variabile *integer*, indica il numero di notti del weekend, quindi sabato e domenica, che sono comprese nella prenotazione del soggetto
- *Stays in week nights*: variabile *integer*, indica il numero di notti totali di soggiorno in struttura del cliente
- *Adults*: variabile *integer*, indica il numero degli adulti compresi nella prenotazione effettuata
- *Children*: variabile *integer*, indica il numero dei bambini compresi nella prenotazione effettuata
- *Babies*: variabile *integer*, indica il numero dei neonati compresi nella prenotazione effettuata
- *Meal*: variabile categorica, indica il tipo di pensione prenotata. Può assumere vari valori: *undefined* (nessun pasto compreso), BB (bed & breakfast), HB (mezza pensione), FB (pensione completa)
- *Country*: variabile categorica, indica lo stato di provenienza del cliente
- *Market segment*: variabile categorica, indica il tipo di segmento di mercato. Può assumere differenti valori: *aviation, complementary, corporate, direct, groups, offline TA/TO, online TA/TO*. Dove TA indica agenzie di viaggio e TO *tour operator*.
- *Distribution Channel*: variabile categorica, indica la tipologia di canale distributivo. Può assumere differenti valori: *corporate, direct, GDS, TA/TO*.
- *Is repeated guest*: variabile categorica, indica se la prenotazione è stata effettuata da un soggetto che è già stato ospite in struttura. Può assumere due valori: 1 (è un cliente ricorrente), 0 (non è un cliente ricorrente).
- *Previous cancellations*: variabile *integer*, indica il numero di prenotazioni che eventualmente sono state già cancellate dal soggetto
- *Previous bookings not canceled*: variabile *integer*, indica il numero di prenotazioni non cancellate che il soggetto ha avuto presso la struttura.
- *Reserved room type*: variabile categorica, indica la tipologia di stanza riservata per la prenotazione. Può assumere differenti valori: A, B, C, D, E, F, G, H, I, K, L, P
- *Assigned room type*: variabile categorica, indica la tipologia di stanza assegnata. Può assumere differenti valori: A, B, C, D, E, F, G, H, I, K, L, P. La tipologia di stanza assegnata può variare dal tipo di stanza riservata per ragioni interne all'hotel o per richieste speciali da parte del cliente.
- *Booking changes*: variabile *integer*, indica quanti cambiamenti sono stati effettuati alla prenotazione dal momento di inserimento nel sistema al momento di arrivo in struttura

- *Deposit type*: variabile categorica, indica la tipologia di deposito collegata alla prenotazione. Può assumere tre valori: *no deposit* (nessun deposito), *non refund* (non rimborsabile, deposito che copre l'intera somma della prenotazione), *refundable* (rimborsabile, è stato versato un deposito il cui valore è inferiore al costo dell'intero soggiorno).
- *Agent*: variabile categorica, indica il codice dell'agenzia che ha effettuato la prenotazione. I dati sono rappresentati sottoforma di ID per garantire l'anonimato.
- *Company*: variabile categorica, indica la società che ha effettuato la prenotazione o che è responsabile del pagamento. I dati sono rappresentati sottoforma di ID per garantire l'anonimato.
- *Days in waiting list*: variabile *integer*, indica quanti giorni la prenotazione è stata in *waiting list* prima che questa fosse confermata al cliente.
- *Customer type*: variabile categorica, indica il tipo di prenotazione. Può assumere 4 valori: *contract* (contratto di *allotment*), *group* (prenotazione è associata a un gruppo), *transient*, *transient party* (associata a un'altra prenotazione *transient*).
- *Adr*: variabile numerica, indica l'*average daily rate*.
- *Required car parking spaces*: variabile *integer*, indica il numero di parcheggi richiesti dal cliente
- *Total of special requests*: variabile *integer*, indica il numero di richieste speciali effettuate dal cliente
- *Reservation status*: variabile categorica, indica lo *status* della prenotazione. Può assumere 3 valori: *canceled* (prenotazione cancellata), *check-out* (il cliente ha pernottato), *no-show* (il cliente non si è presentato al *check-in*).
- *Reservation status date*: formato data, indica la data ultima in cui il *reservation status* è stato modificato.

Data visualization

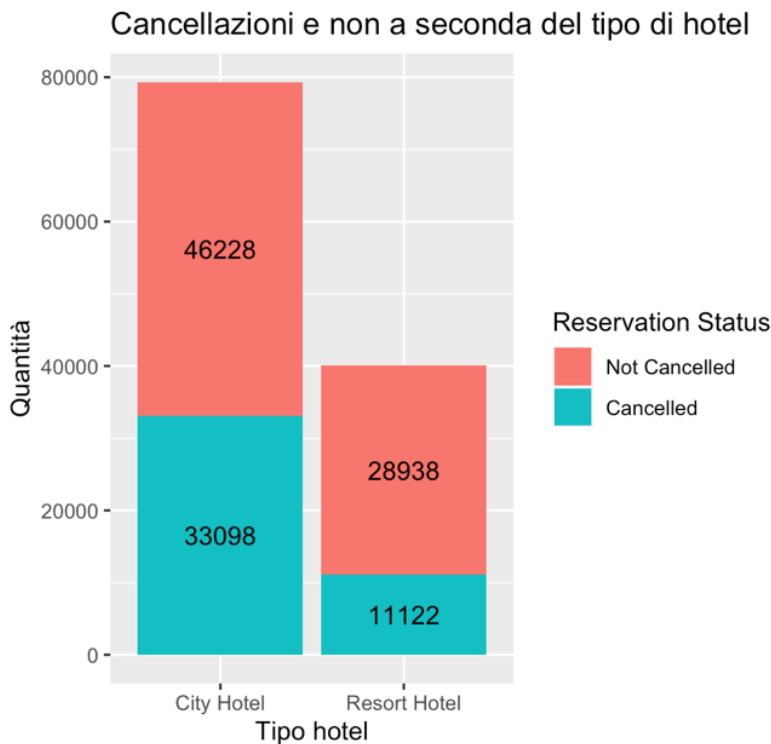


Figura 129: Come variano le cancellazioni al variare del tipo di hotel; Fonte: realizzazione propria

Come è possibile verificare in fig.19, il Resort Hotel presenta un numero minore di prenotazioni rispetto al city hotel per cui si hanno rispettivamente 40.059 prenotazioni e 79.326 prenotazioni.

Calcolando le percentuali si può verificare che la percentuale di cancellazioni è più elevata nel city hotel.

Abbiamo, infatti, un 41% di cancellazioni per il city hotel mentre solo il 27% per il resort hotel.

Per cercare di capire una ragione dietro questa differenza, ho analizzato la variabile *lead time*. Questa variabile, come già anticipato, indica quanti giorni sono passati dal momento della prenotazione al momento di arrivo, o presunto tale, in struttura; esso, in pratica, indica con quanti giorni di anticipo si è effettuata la prenotazione. Per far questo, ho analizzato separatamente le osservazioni appartenenti ai due differenti tipi di hotel studiando, attraverso la funzione *summary*, la media e la mediana della variabile *lead time*.

Per quanto riguarda il *resort hotel*, è emerso che la media, per la variabile studio in oggetto, è 93 giorni con un minimo di 0, che fa pensare ad un arrivo senza prenotazione, ad un massimo di 737 giorni. Quest'ultimo dato risulta però essere un caso limite poiché, considerando il terzo quartile questo è pari a 155 giorni.

Analizzando il *city hotel*, che presenta un tasso di cancellazione maggiore, abbiamo che la media è pari a 110 giorni, con un minimo anche in questo caso di 0 giorni ed un massimo di 629 giorni. Anche in questo caso, quest'ultimo rappresenta un caso limite poiché il terzo quartile è pari a 163 giorni.

Di conseguenza, si può desumere che le prenotazioni per quanto riguarda il *city hotel* vengono in media effettuate con più giorni di anticipo.

Di conseguenza, i giorni di differenza delle medie del *lead time* riferite ai due tipi di hotel è di 18 giorni mentre la differenza percentuale della cancellazione è pari a 14 punti percentuali.

Come varia la cancellazione in base al tipo di hotel

Ed in base al lead time

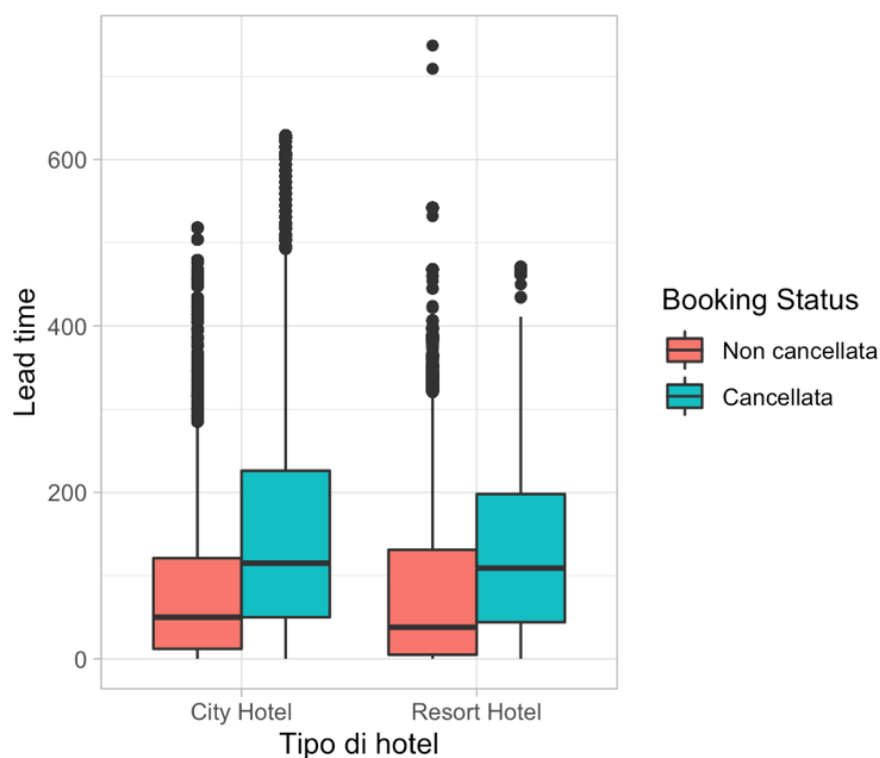


Figura 130: Come varia la cancellazione in base al tipo di hotel e al lead time; Fonte: realizzazione propria

Tuttavia, volendo effettuare un'analisi maggiormente approfondita si può far ricorso al grafico denominato *box plot*. Questo è raffigurato in fig.20. Come si può notare, in questo grafico si può verificare come varia il *lead time* rispetto al tipo di hotel ma anche, e soprattutto, in relazione al *booking status*.

Emerge, quindi, che le prenotazioni che sono state cancellate sono anche quelle che sono state effettuate con maggiore anticipo e questo vale sia per il city hotel che per il resort hotel. Tuttavia, per il city hotel questa differenza è maggiormente amplificata rispetto al resort hotel, come è maggiormente alta la percentuale generale di cancellazione per questo tipo di hotel.

Ulteriore elemento che potrebbe influenzare la percentuale di cancellazione è rappresentato dal tipo di deposito, e quindi di garanzie a fronte della prenotazione, richiesta.

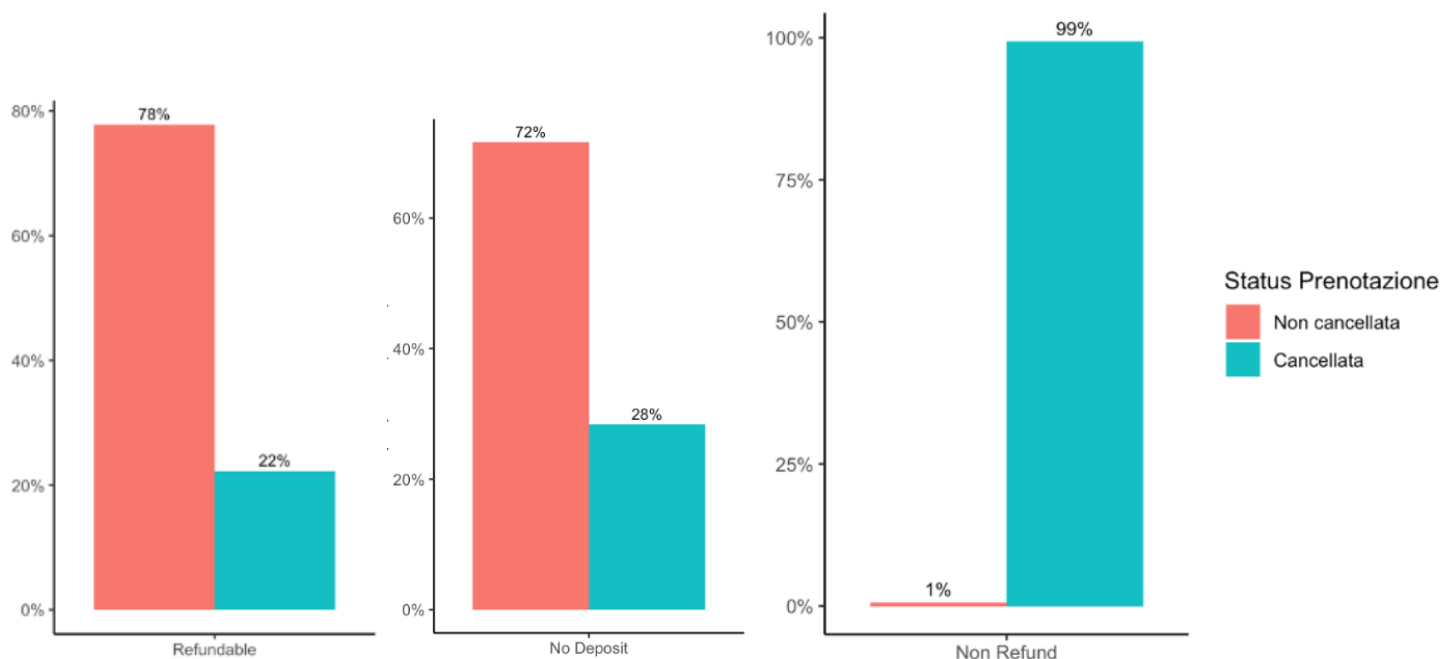


Figura 21: Il deposito influenza la cancellazione?; Fonte: realizzazione propria

Come si può notare dalla fig. 21, un ragionamento *data driven* sul nesso cancellazione – polizza rimborsabile o meno non può essere effettuato; si può infatti notare che quando la cancellazione risulta essere non rimborsabile le cancellazioni superano di molto le prenotazioni confermate. Ciò è, naturalmente, non fattibile poiché è naturale che quando la tariffa è non rimborsabile la percentuale di cancellazione sia inferiore a quando la tariffa è rimborsabile.

Questo strano accadimento è spiegato dagli stessi autori del *dataset* e può essere verificato sugli stessi dati. Tutte queste prenotazioni non *refundable* ma cancellate provengono quasi interamente da canali OTA *online* e sono utilizzate per fini prettamente legali. Per richiedere il visto in Portogallo è necessario, infatti, esibire una prenotazione presso una struttura alberghiera o simile e di conseguenza, in molti utilizzano gli OTA, fornendo poi carte di credito fasulle per aggirare questo problema del visto. Le prenotazioni sono poi chiaramente cancellate dalla stessa struttura, che si vede rifiutato il corrispettivo dovutogli, ma nel frattempo queste prenotazioni permangono nel sistema.

Nonostante, quindi, non si abbiano prove oggettive sul presente *dataset* è logico pensare che la percentuale di cancellazione per le prenotazioni non rimborsabili “reali” sia nettamente inferiore a quella delle prenotazioni rimborsabili.

Di questo se ne può in parte vedere esempio confrontando le tipologie di prenotazioni rimborsabili e quelle senza deposito. Nelle prime, infatti, la percentuale di cancellazioni risulta essere superiore alla seconda di 6 punti percentuali.

Questo è dovuto semplicemente al fatto che un soggetto è più propenso ad effettuare una cancellazione quando non ha versato alcun deposito rispetto a quando, pur essendo rimborsabile, lo ha versato.

Naturalmente, per effettuare delle decisioni consapevoli e per capire in modo ottimale quali possono essere le strategie attuabili, il *management* deve prima di tutto analizzare l'andamento della domanda.

Studiare l'andamento della domanda permette di verificare se sono presenti momenti di picco della stessa o momenti più stagnanti.

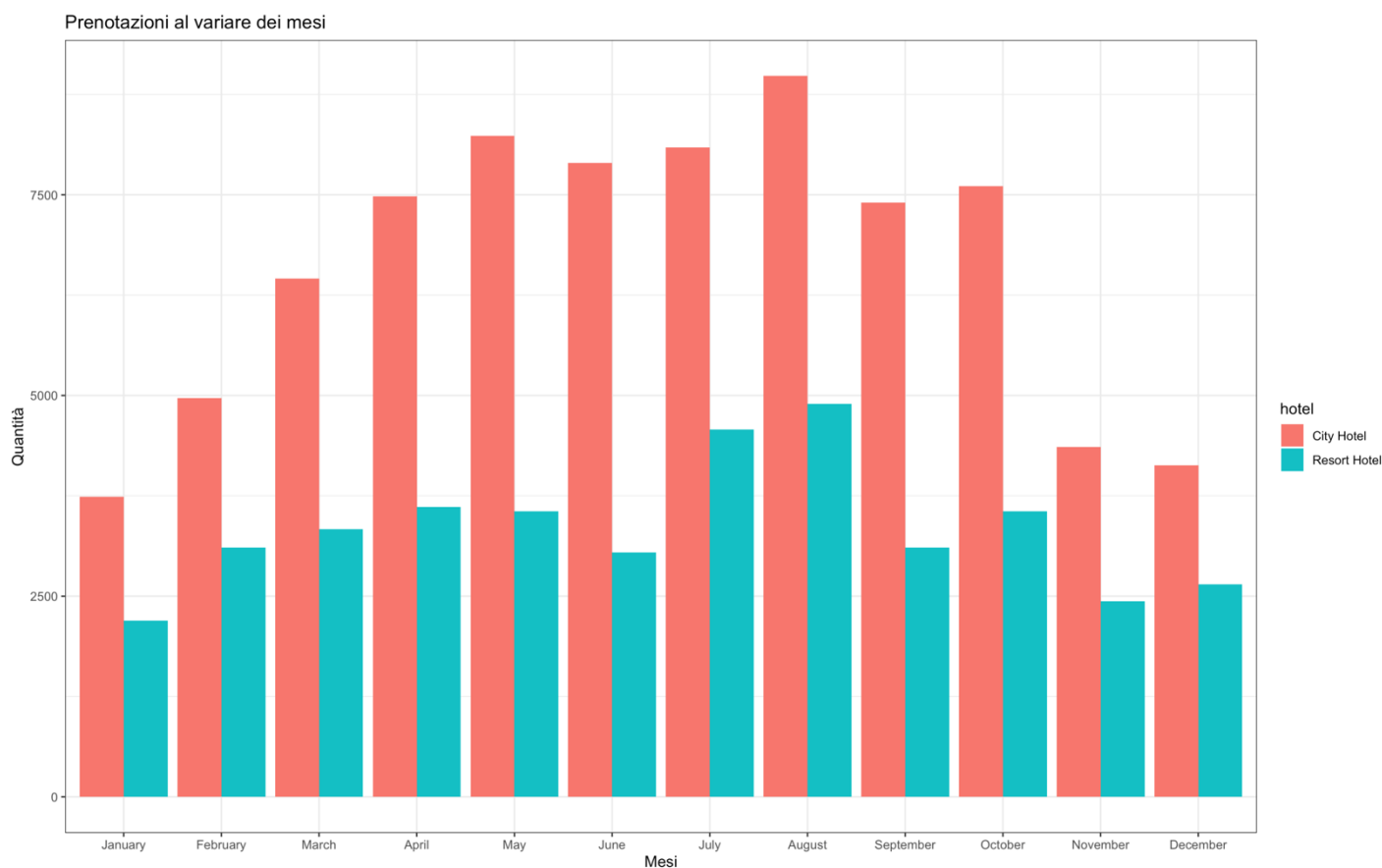


Figura 22: Andamento della domanda; Fonte: realizzazione propria

Come si può vedere dalla fig. 22, nei mesi estivi si ha un picco delle prenotazioni. Questo picco è maggiormente accentuato per il city hotel, mentre per il resort hotel pur essendoci un maggior numero di prenotazioni la domanda risulta essere abbastanza stabile.

Ulteriore analisi effettuabile, sempre considerando l'andamento della domanda, è analizzare congiuntamente a questa se le cancellazioni sono maggiormente presenti in un certo periodo dell'anno piuttosto che in un altro.

Reservation Status al variare del mese

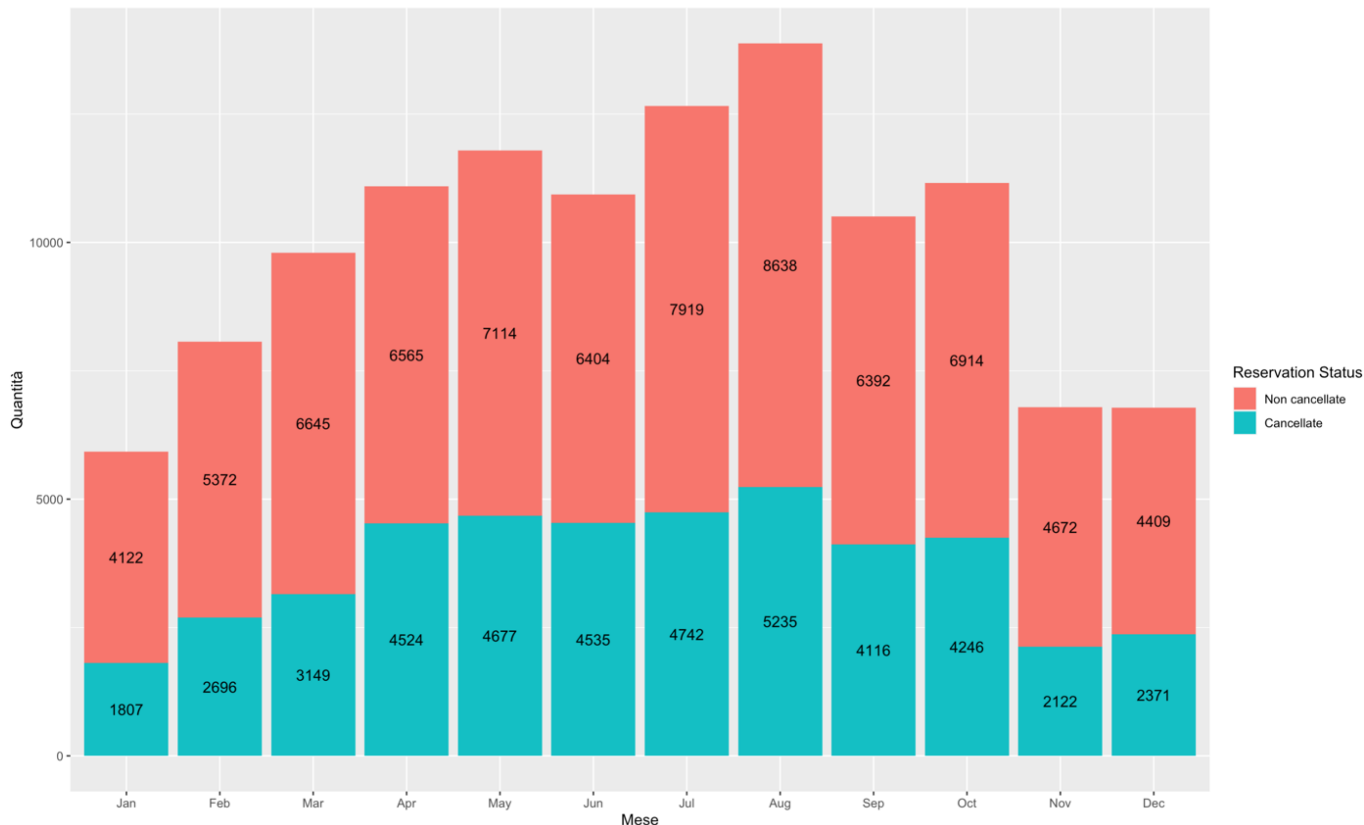


Figura 23: Come si compongono le prenotazioni; Fonte: realizzazione propria

Come si può verificare dalla fig. 23 non è presenta un picco di cancellazioni in un dato periodo ma l'andamento delle cancellazioni segue l'andamento della domanda.

Ma per studiare in modo ancora migliore le cancellazioni e per capire quali potrebbero essere le motivazioni dietro a queste, ulteriore analisi effettuabile è quella che riguarda lo studio dei canali distributivi.

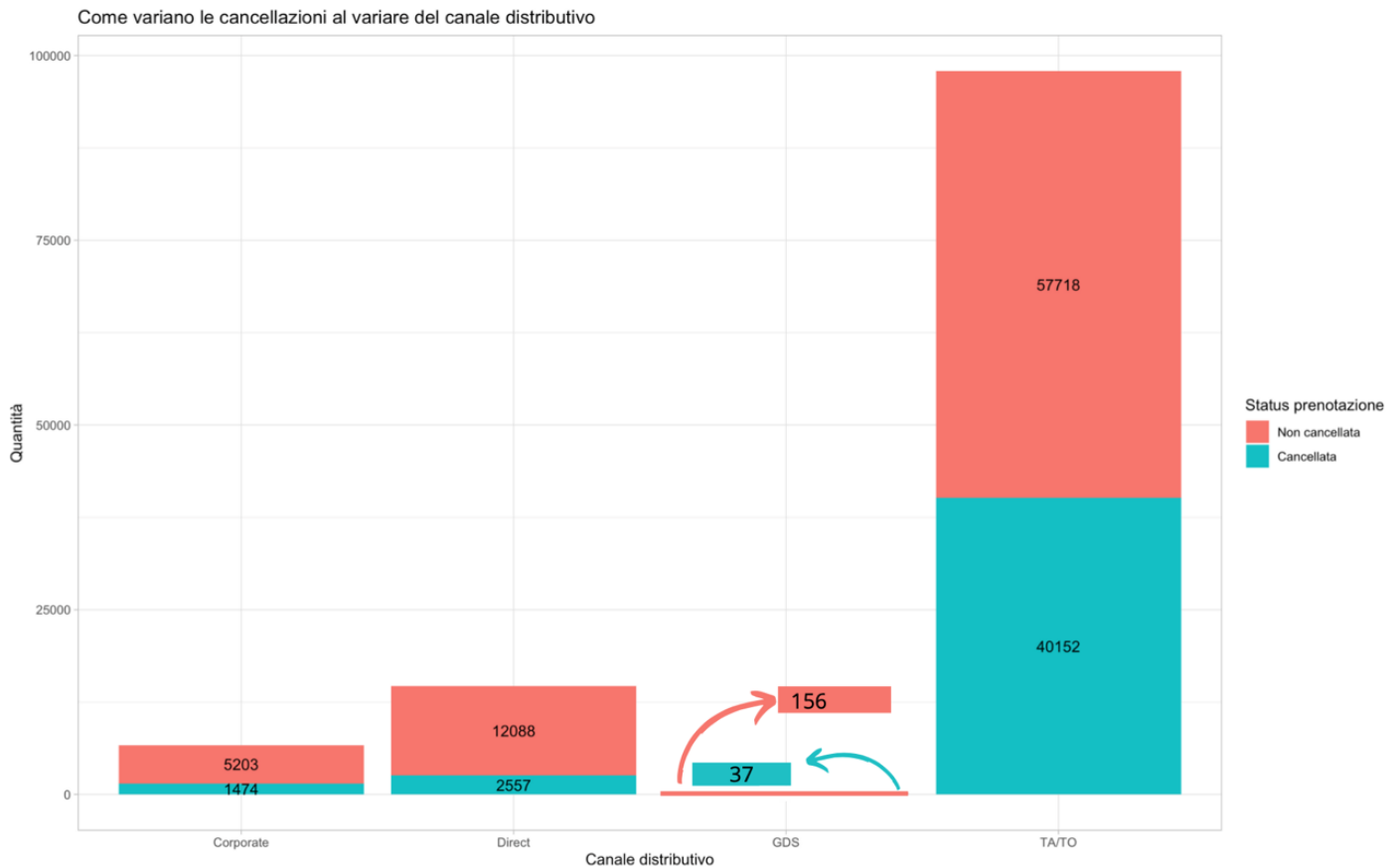


Figura 24: Prenotazioni per canale di vendita; Fonte: realizzazione propria

Come è possibile verificare dalla fig.24, esiste una differente probabilità di cancellazione considerando i vari canali di vendita.

Se, infatti, quando si considera il canale *corporate* si ha una percentuale di cancellazione pari al 22% quando invece si considera il canale TA/TO, cioè quello delle agenzie di viaggio e dei tour operator, allora la percentuale sale raggiungendo un valore percentuale pari al 41%.

La percentuale minore la si ha con il canale della prenotazione diretta per cui si ha una percentuale di cancellazione pari al 17%. Si registra una percentuale di cancellazioni accettabile, pari al 19%, anche per quanto riguarda il canale dei GDS cioè dei *Global Distribution System*, che cura le prenotazioni B2B.

Si potrebbe pensare che questa differenza sia dovuta principalmente alle modalità di rimborso poiché, solitamente, quando si prenota attraverso un TA online, ad esempio, si possono scegliere tra differenti tipi di tariffa mentre solitamente quando si effettua una prenotazione in modo diretto si ha l'obbligo di versare una caparra al fine di garantire la prenotazione.

Comparazione dei canali distributivi

Secondo il tipo di deposito

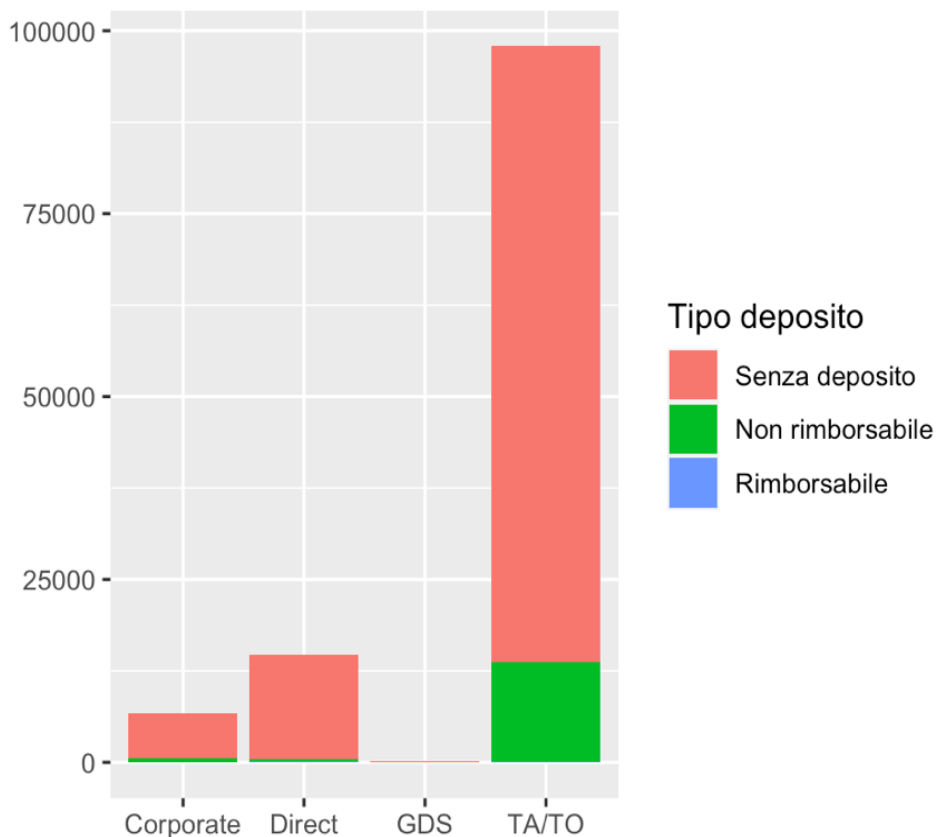


Figura 25: La differente presenza di deposito; Fonte: realizzazione propria

Tuttavia, come si può vedere in fig.25 ciò non si è verificato per il caso in oggetto. Infatti, per tutti i canali almeno il 75% delle prenotazioni non risulta garantita da alcun deposito iniziale.

Di conseguenza, la maggior probabilità di cancellazione legata al canale dei TA e dei TO non dipende dal differente tipo di rimborso applicabile bensì è da far risalire alla stessa tipologia di mezzo di prenotazione.

Questo potrebbe in parte essere spiegato dal fatto che una prenotazione effettuata *online* attraverso siti OTA è più in linea con una prenotazione immediata e veloce.

Si visualizza, ad esempio, un'offerta per un certo hotel ed il soggetto per non perdere l'occasione, non essendo richiesto alcun deposito, effettua la prenotazione d'impulso. Successivamente, si potrebbe recare sul sito dell'hotel o su siti di recensioni e potrebbe non valutare più come adatto quell'hotel oppure ancora potrebbe trovare un'offerta maggiormente conveniente oppure più in linea con le proprie attitudini.

Mentre, una prenotazione effettuata sul sito dell'hotel richiede una ricerca maggiore, si è ad esempio passati da siti OTA al sito stesso dell'hotel magari anche consultando più di un sito di recensioni, è quindi un processo che richiede maggior tempo e di conseguenza potrebbe portare ad una decisione più ponderata.

Ulteriore aspetto indagabile è la presenza di una *waiting list*.

In questo caso si può indagare come varia la cancellazione al cambiare dei giorni di lista d'attesa. In questo caso ci si potrebbero aspettare due differenti fenomeni. Si potrebbe assistere all'aumento della percentuale di cancellazione all'aumentare dei giorni di attesa, poiché il soggetto non vedendosi confermata la prenotazione potrebbe cambiare idea o essere costretto a cancellare la prenotazione non trovando posto.

Ma allo stesso tempo la cancellazione potrebbe essere meno probabile poiché il soggetto avendo aspettato svariati giorni e non avendo preso la decisione di cancellare la prenotazione, nel frattempo, potrebbe essere indice di decisione ferma sulla struttura.

Come variano le cancellazioni Al variare dei giorni in waiting list

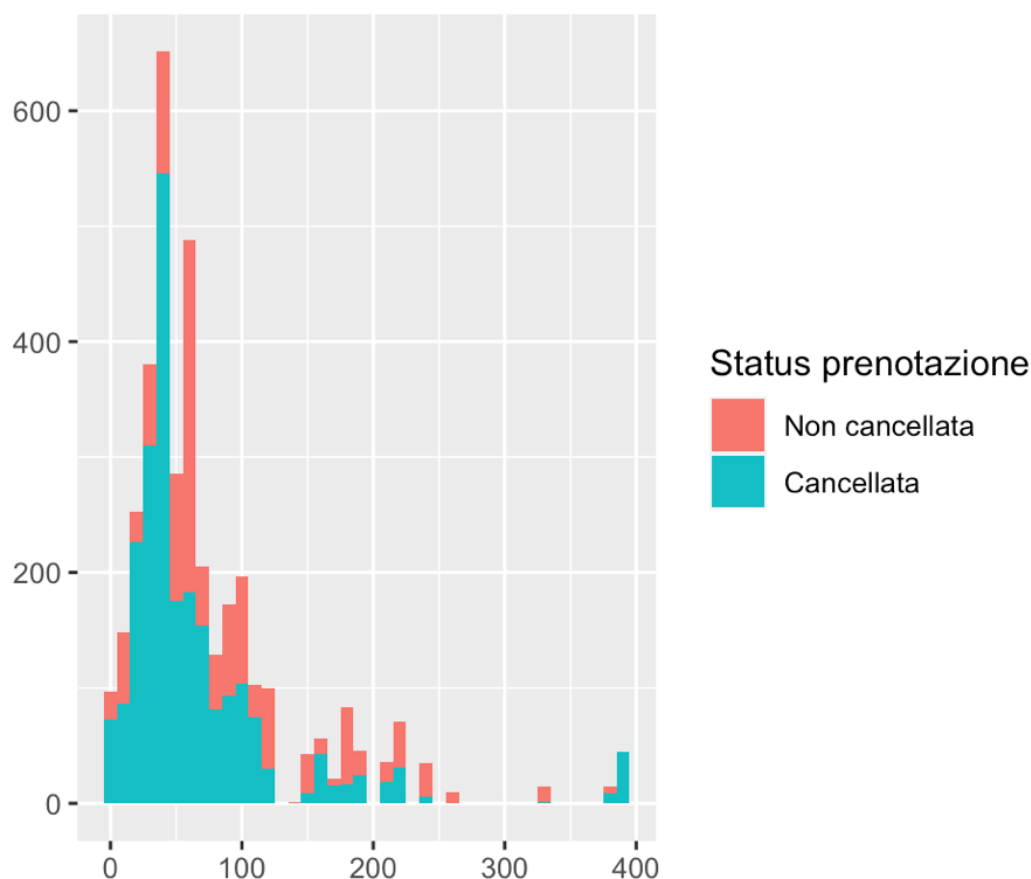


Figura 26: Come variano le cancellazioni; Fonte: Realizzazione propria

Come si può vedere in fig.26 ad eccezione di quando la *waiting list* è vicina ai 400 giorni in cui si registrano più prenotazioni cancellate che non, per tutti gli altri casi si registrano più prenotazioni confermate che cancellate.

La particolarità la si registra quando la *waiting list* è di circa 60 giorni. In questa situazione, infatti, si registra un picco nelle prenotazioni non cancellate mentre le prenotazioni cancellate sono minori. Questo è il punto del grafico, insieme alla *waiting list* di 150 giorni, in cui c'è la percentuale di cancellazione minore. Tuttavia, mentre nel primo caso le prenotazioni sono molteplici e quindi il dato risulta essere maggiormente affidabile nel caso della *waiting list* di 150 giorni le prenotazioni sono inferiori, circa 50, e di conseguenza la differenza tra prenotazioni cancellate e non cancellate potrebbe maggiormente essere dettata dal caso.

Per studiare la domanda e la sua eventuale cancellazione si può, inoltre, andare a studiare lo stato di provenienza degli ospiti. Questo perché da questa analisi potrebbe emergere che un soggetto proveniente da un certo stato sia maggiormente incline alla cancellazione rispetto ad un altro.

Inoltre, lo studio della provenienza potrebbe fornire anche utili spunti per effettuare azioni di *advertising* oltre che di fidelizzazione.

Tuttavia, le nazioni da cui provengono le prenotazioni sono molteplici e di conseguenza un grafico che li rappresenti tutti non è molto esplicativo.

Per questo motivo, per esigenze grafiche, ho scelto di impostare come filtro almeno 1500 prenotazioni. In questo modo, e creando un nuovo oggetto, si prenderanno in considerazione solo gli stati con maggiori prenotazioni.

Però, gli stati sono rappresentati nel *file csv* da sigle per questo motivo ho convertito queste ultime nel nome per intero dello stato.

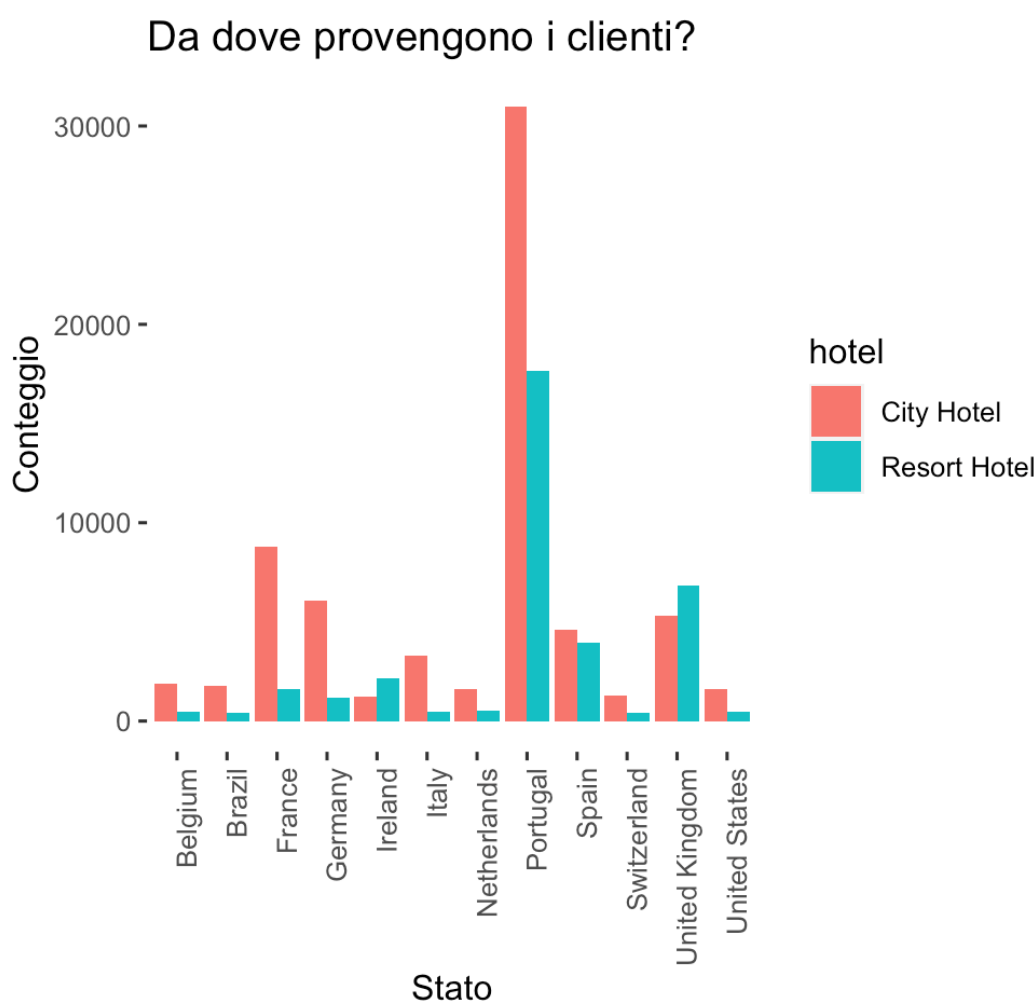


Figura 27: Da dove provengono i clienti?; Fonte: realizzazione propria

Come si può vedere dalla fig. 27, naturalmente vista la localizzazione degli hotel, la maggior parte delle prenotazioni provengono dal Portogallo. Si ha poi il Regno Unito, la Spagna, la Francia e la Germania.

Vista la notevole differenza di numeri di questi stati, è logico pensare che l'hotel sia perlopiù frequentato da una clientela nazionale e di conseguenza, eventuali azioni di marketing perseguibili, sono da incentrare in questa zona.

Ma l'analisi può essere ancora migliorata quando si mettono in relazione le cancellazioni.

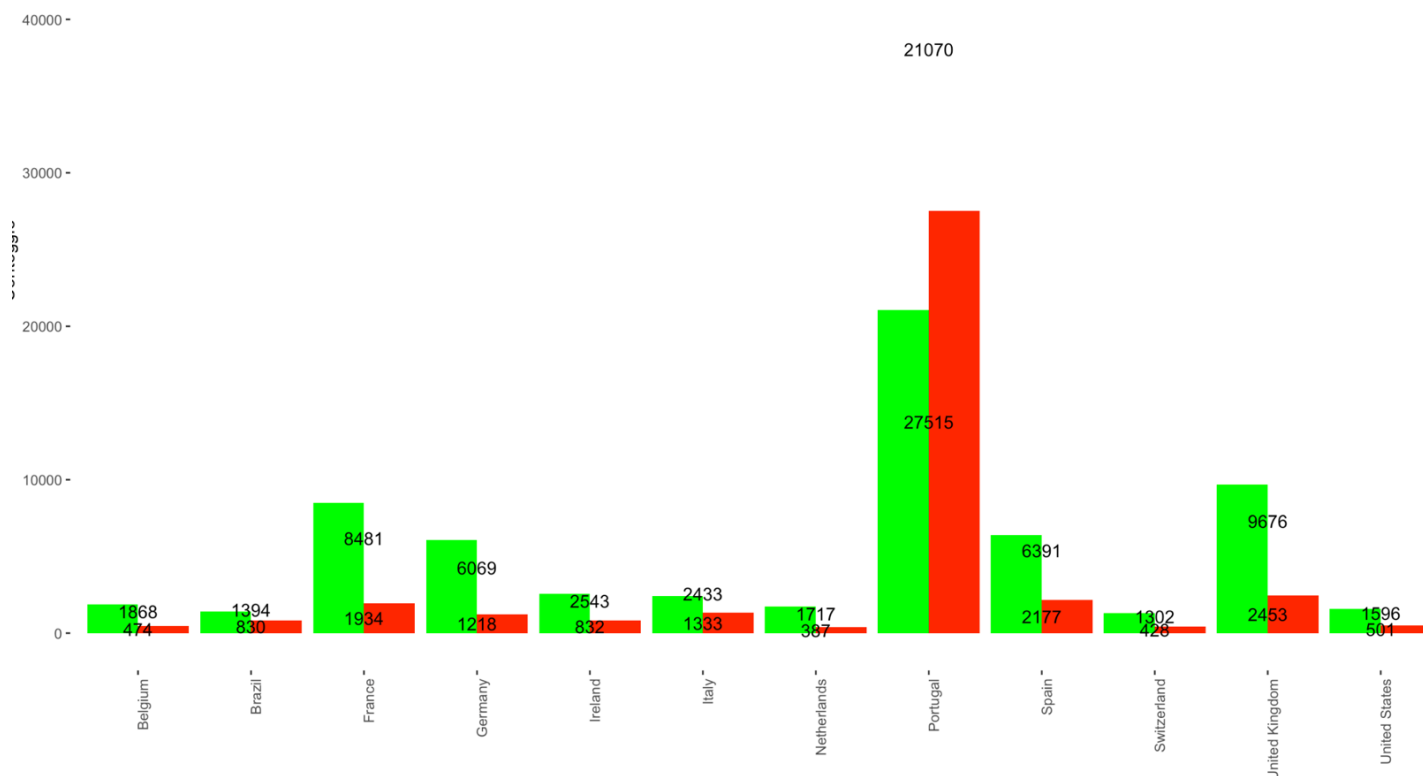


Figura 28: Come variano le cancellazioni al variare dello stato; Fonte: realizzazione propria

Per il dataset in oggetto si ha però un problema di analisi, poiché gli stati sono molteplici e di conseguenza un grafico che permetta di rappresentarli tutti non è di facile lettura. Motivo per cui ho scelto di impostare dei filtri che permettano di analizzare solo alcuni stati che verifichino la condizione imposta.

Per questo ho imposto due differenti filtri: più di 1500 prenotazioni, meno di 1500 prenotazioni ma più di 1000.

Come si può vedere dalla fig.28 a seconda dello stato considerato si ha una differente probabilità di cancellazione; il grafico riguarda le nazioni con più di 1500 prenotazioni.

Il Portogallo, sede dei due hotel considerati, è lo stato che ha maggiori prenotazioni ma anche quello in cui si hanno più prenotazioni cancellate che non cancellate; nello specifico il 57% delle prenotazioni. Questo analizzando i vari stati considerati risulta essere l'unico caso in cui accade.

Ci sono nazioni in cui le prenotazioni cancellate e non si quasi eguagliano l'Italia, ad esempio, come anche il Brasile ma non si ha la supremazia delle prenotazioni cancellate come accade per il Portogallo.

Si hanno poche cancellazioni in confronto alle prenotazioni per i clienti provenienti dalla Francia (18%) e dalla Germania (17%) così come anche dal Regno Unito (20%).

Già questo grafico permette di effettuare un primo ragionamento circa il tasso di *overbooking* e la gestione delle prenotazioni.

Se, infatti, la maggioranza delle prenotazioni in un certo periodo provengono dal Portogallo, che è lo stato per cui si registrano più prenotazioni cancellate che non, è chiaro che il tasso di *overbooking* sarà più elevato rispetto a quando la maggioranza delle prenotazioni proviene, ad esempio, dal Regno Unito. Questo dato lo si può unire anche all'analisi compiuta inizialmente sullo studio delle cancellazioni al variare dei canali di

prenotazione. Tutto ciò permette, infatti, di andare ad indagare e costruire il profilo della persona che con più probabilità potrebbe effettuare la cancellazione della propria prenotazione.

Ulteriore elemento indagabile riguarda la tipologia di prenotazione.

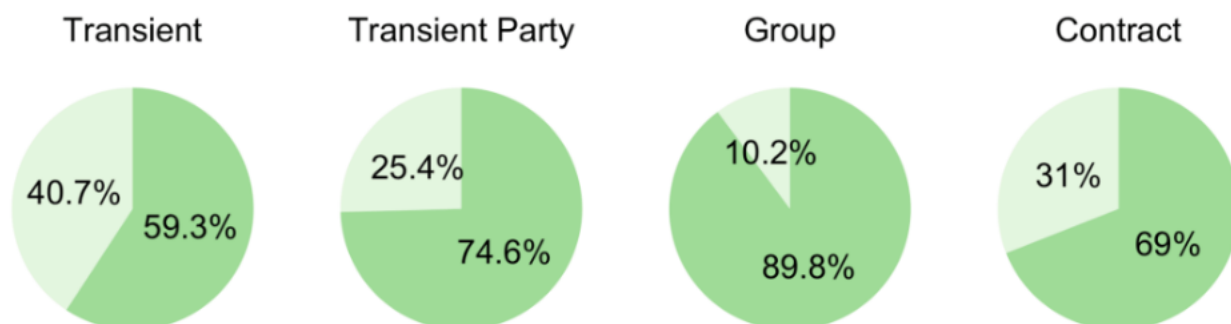


Figura 29: Tipologia di prenotazione con associata la percentuale delle prenotazioni cancellate (in chiaro); Fonte: realizzazione propria

Come si può notare, dalla fig.29, le percentuali di cancellazione variano, anche di molto, a seconda della tipologia di prenotazione che è stata effettuata.

La percentuale più bassa la si registra per le prenotazioni di gruppo, cioè quelle prenotazioni che sono state effettuate da un gruppo di persone che viaggia assieme. In questo caso, infatti, si è registrata una percentuale di cancellazione del solo 10%; questo potrebbe essere in parte associato al fatto che la prenotazione, e la volontà di soggiornare presso la struttura, proviene da una decisione di gruppo e non è quindi soltanto a descrizione di un soggetto l'eventuale cancellazione.

La seconda percentuale più bassa, pari al 25%, è associata alla tipologia di prenotazione *transient party*.

Le prenotazioni classificate come *transient* appartengono a quella classe di clienti che non prenotano in gruppo e né sono parte di un contratto di prenotazione. In particolare, però, si indicano come *transient party* quelle prenotazioni *transient* che sono associate ad almeno un'altra prenotazione sempre *transient*.

Si tratta, quindi, di soggetti che hanno in essere più di una prenotazione. Di conseguenza, questo farebbe pensare maggiormente a soggetti che viaggiano molto e quindi a prenotazioni associate a viaggi di lavoro.

Inoltre, essendo presenti nel sistema più di una prenotazione, questo potrebbe essere indice di soggetti che hanno già soggiornato presso la struttura tanto da essere fedeli ed effettuare presso la stessa più di una prenotazione.

Per quanto riguarda, invece, le prenotazioni classificate come *contract* queste si riferiscono alle prenotazioni effettuate in *allotment*. L'*allotment* è un contratto stipulato tra la struttura e un *tour operator* che prevede il blocco, quindi la prenotazione, da parte di quest'ultimo di un gruppo di camere che poi sarà lui stesso a commercializzare.

In questo caso la cancellazione è pari al 31% e potrebbe essere dovuta al fatto che non sempre i *tour operator* riescono a vendere in *toto* tutto il lotto di stanze da loro acquistate e di conseguenza effettuano la cancellazione. Ma la percentuale maggiore di prenotazioni cancellate la si è registrata con la tipologia di prenotazioni *transient*. Queste come già detto, sono tutte quelle prenotazioni non effettuate in gruppo o attraverso un TO,

ma effettuate singolarmente da ogni soggetto che quindi potrebbe decidere in più libertà, contrariamente alle altre due tipologie di prenotazione, di effettuare la cancellazione.

Per effettuare un'analisi maggiormente approfondita sulla tipologia di prenotazione *transient* che rappresenta, appunto, quella per cui si registra un tasso di cancellazione più alto, si può prendere in considerazione il segmento di mercato.

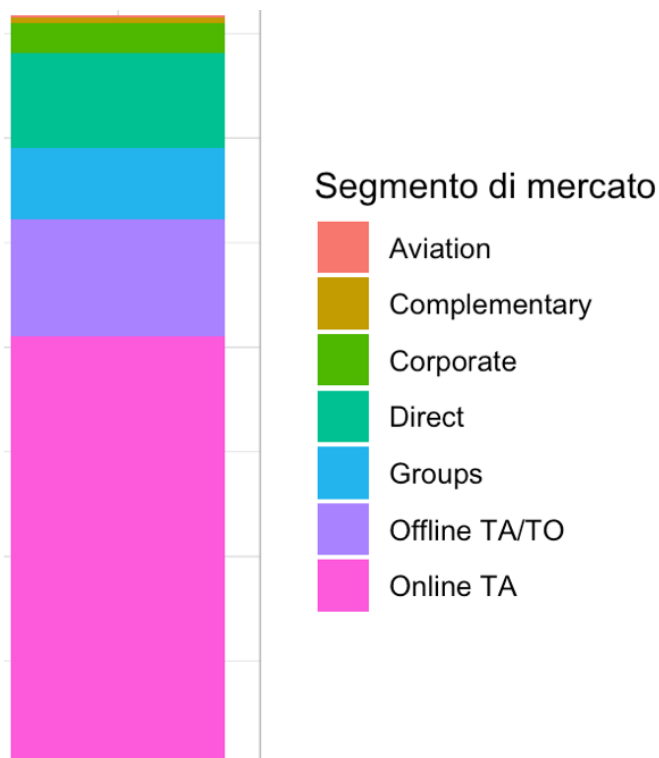


Figura 30: Composizione delle prenotazioni transient; Fonte: realizzazione propria

Come si può verificare dalla fig.30 la maggior parte delle prenotazioni proviene da *travel agency (TA) online*. Mentre soltanto una piccola porzione proviene dal segmento dell'*aviation* e della clientela *corporate*.

Si classificano come *aviation* tutte quelle prenotazioni che provengono da compagnie aeree che per via di ritardi nei voli riserva delle camere per i propri clienti.

Mentre chiaramente con clientela *corporate* si indicano i viaggiatori *business*.

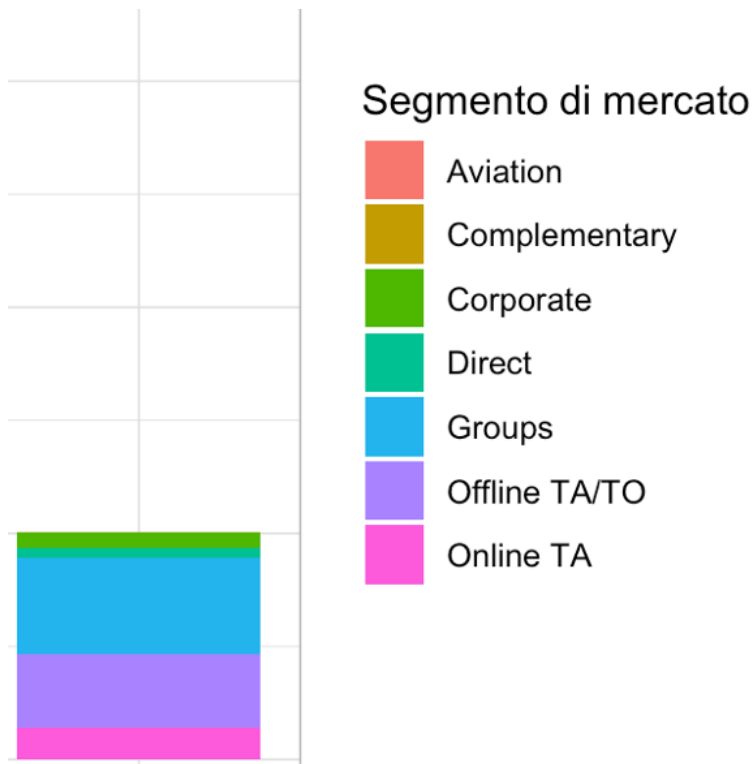


Figura 141: Composizione *Transient-party*; Fonte: realizzazione propria

Per quanto riguarda invece le composizioni *transient-party*, come si può vedere dalla fig.31, queste sono per la maggior parte composte da prenotazioni effettuate da un gruppo e da TA *offline*. Non si hanno, invece, molte prenotazioni provenienti da TA *online* e da clientela *corporate*.

C'è da specificare che con prenotazioni di gruppo si identificano tutte quelle prenotazioni che, cumulativamente, non superino le 10 camere. Questo perché, qualora si superi questo limite, si parlerà di prenotazioni classificate come *Group* e non più come *transient*.

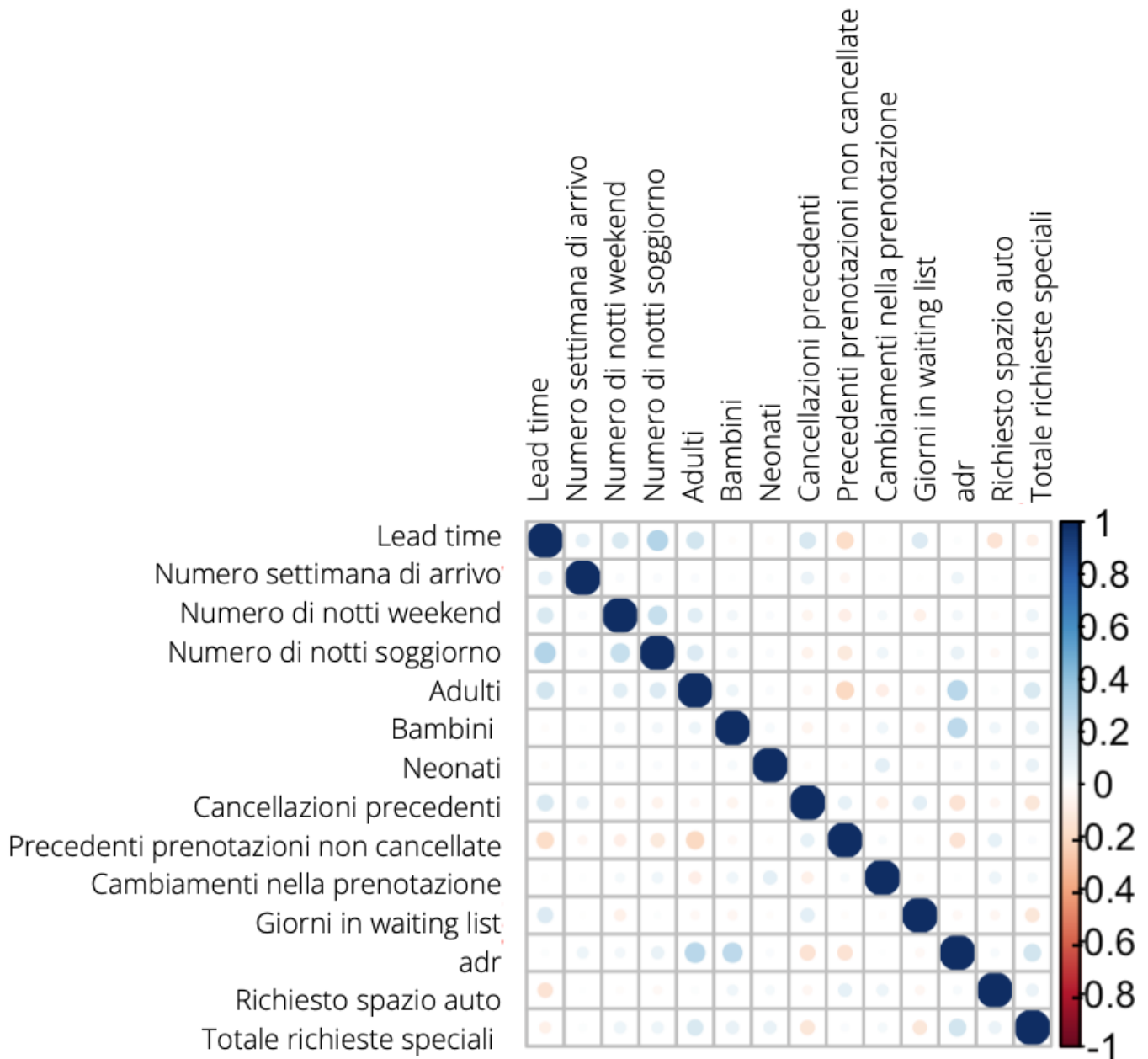


Figura 32: Analisi delle correlazioni; Fonte: realizzazione propria

Attraverso la matrice visibile in fig.32, si può verificare il rapporto e la forza di correlazione tra tutte le variabili numeriche.

Come si può evincere dalla matrice, non sono presenti variabili che presentano una forte correlazione né positiva né negativa.

Una delle correlazioni maggiormente forti, 0,30, è quella tra la variabile “lead time” e la variabile “numero notti soggiorno”.

La variabile “lead time” è anche correlata negativamente, -0,18, con la variabile “precedenti prenotazioni non cancellate”.

Riportando l'attenzione sulla variabile "numero di notti soggiorno", questo oltre naturalmente ad essere correlata con la variabile "numero di notti weekend" è anche correlata positivamente, seppur in modo debole, con la variabile "adulti".

Quest'ultima variabile è, invece, anche correlata, con un indice pari a 0.27, alla variabile "adr". L'*adr* indica l'*average daily rate* cioè la tariffa giornaliera media pagata per una stanza.

Si ha una correlazione pari a 0.27 anche quando si analizza la relazione "adr" e "bambini"

Queste due correlazioni risultano essere giustificabili dal fatto che maggiori saranno gli occupanti di una stanza, maggiore sarà il ricavo derivante dalla stessa.

Previsione della cancellazione della domanda

La previsione della cancellazione della domanda, così come la previsione della domanda stessa, è possibile effettuarla attraverso differenti metodologie.

Per quanto riguarda le tecniche di *machine learning*, le principali metodologie sono rappresentate: dalla *random forest*, dal *decision tree* o dalla regressione logistica. Si deve infatti risolvere, per effettuare la previsione, un problema di classificazione. Infatti, le osservazioni, che corrispondono alle prenotazioni, dovranno essere classificate secondo l'appartenenza a due diverse classi rappresentabili, ad esempio, come: probabile cancellazione e non probabile cancellazione.

Per quanto riguarda il presente elaborato, ho dapprima costruito e studiato un modello di *decision tree* per poi scegliere di procedere con la costruzione di un modello di *random forest* poiché, come si vedrà di seguito, presenta un'accuratezza oltre che una specificità e sensitività migliori.

Il codice per la generazione di questi modelli può essere visto in appendice.

Inizialmente, ho provato a sviluppare un modello di predizione basato sul metodo del *decision tree*.

Per costruire correttamente un modello, che sia un *decision tree* o un modello di *random forest*, occorre preventivamente suddividere il *dataset* in due parti. In questo modo, infatti, si potrà costruire il modello soltanto su una parte dei dati per poi applicarlo sulla restante parte. In questo modo si verifica il funzionamento del modello su dati che sono differenti rispetto a quelli con i quali esso è stato costruito. Si può effettuare questa operazione attraverso la funzione *split* che dividerà, appunto, il *dataset* originario in due sottoinsiemi, *training set* e *test set*, secondo un rapporto predefinito dato dallo *split ratio*. Nella divisione si rispetterà anche la proporzione esistente nel *dataset* originario per quanto riguarda la variabile oggetto di studio.

Per quanto riguarda lo *split ratio*, ho fissato un tasso di 0.7 che rappresenta il tasso solitamente utilizzato per questo tipo di operazione. Nelle varie prove effettuate per migliorare l'accuratezza, che si vedrà non è troppo alta, ho provato a cambiare lo stesso non avendo però risultati migliori; motivo per cui ho deciso di fissare nuovamente lo stesso al livello 0.7.

Tuttavia, prima di effettuare la costruzione e l'applicazione del modello, ho verificato l'accuratezza del caso base.

0 1
 75165 44220

Figura 33: Caso base; Fonte: realizzazione propria

Come si può vedere in fig. 33, sono presenti 44.220 osservazioni che risultano essere cancellate (1) e 75.165 che risultano essere non cancellate (0). Dal momento che le prenotazioni non cancellate sono più frequenti delle prenotazioni cancellate, il nostro caso base sarà dato dal predire sempre le prenotazioni non cancellate. Di conseguenza, si avrà un'accuratezza del caso base del 63%, ottenuta dividendo i casi in cui si hanno le prenotazioni non cancellate con la totalità delle osservazioni.

Decision Tree

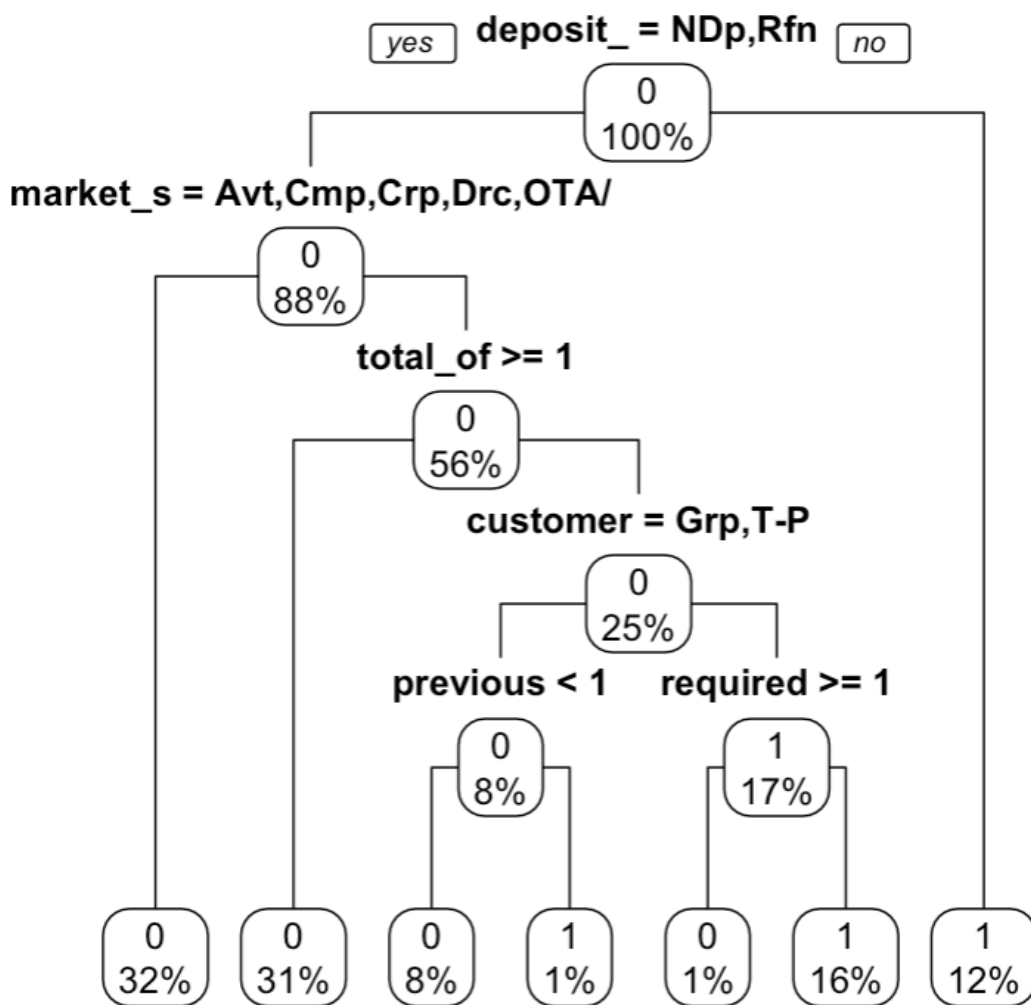


Figura 34: Decision Tree; Fonte: realizzazione propria

La fig. 34 rappresenta l'albero decisionale che è stato creato. Per la sua creazione è stato imposto un *min bucket* pari a 15. Tuttavia, come si può vedere dalla figura successiva questo tipo di modello non produce risultati ottimi.

Reference		
Prediction	0	1
0	20195	5264
1	2354	8002

Figura 35: Confusion matrix decision tree; Fonte: realizzazione propria

In fig. 35 è rappresentata la *confusion matrix* ottenuta dall'esecuzione del *decision tree* sul *test set* con una soglia pari a 0,5.

Come si può vedere dalla matrice, i casi in cui correttamente viene predetto che la prenotazione non verrà cancellata sono 20.195; mentre i casi in cui, correttamente, vengono predette le cancellazioni sono 8.002.

Questi appena delineati sono i casi in cui le predizioni risultano corrette, quelle cioè per cui si predice una classe di appartenenza che si rivela essere realmente quella.

Il problema sono gli altri due casi, quelli cioè in cui si predice erroneamente l'appartenenza ad una classe.

Nello specifico 2.354 rappresentano i falsi negativi; cioè i casi in cui è stata predetta la classe negativa, cioè la cancellazione, ma in realtà si trattava di prenotazioni che non sono state cancellate.

Viceversa, 5.264 rappresenta i casi falsi positivi cioè quelli in cui si è predetta la classe positiva, quindi la non cancellazione, ma in realtà si trattava di osservazioni appartenenti alla classe negativa.

Tuttavia, questo metodo non si è dimostrato all'altezza poiché, pur inserendo un numero elevato di variabili, e ottenendo un'accuratezza del 79% la *sensitivity* e la *specificity* risultavano essere molto sbilanciate tra di loro con un valore, rispettivamente, dell'89% e del 60%.

Visti i risultati non soddisfacenti dell'albero decisionale per cercare di migliorare le prestazioni ho proseguito elaborando un modello di *random forest*. La *random forest* basando la classificazione non su un unico albero ma su molteplici, qualora correttamente sviluppato, permette infatti una migliore prestazione.

Sempre effettuando la creazione del modello sul *training set* ed effettuando la verifica delle prestazioni sul *test set*, ho inizialmente sviluppato il modello prendendo in considerazione tutte le variabili ad eccezione di alcune quali: *agent* e *company* poiché non utili ai fini delle analisi.

Ho dovuto escludere anche la variabile *country* poiché presenta più di 53 categorie che rappresenta il numero massimo di categorie gestibili da R per la creazione di un modello.

Ho scelto di includere tutte le variabili nel modello in modo poi da operare la *feature selection*; quest'analisi consente, infatti, di calcolare la *mean decrease accuracy*. Questa metrica esprime di quanto l'accuratezza diminuisce escludendo dal modello le varie variabili oggetto di studio.

Di seguito i risultati.

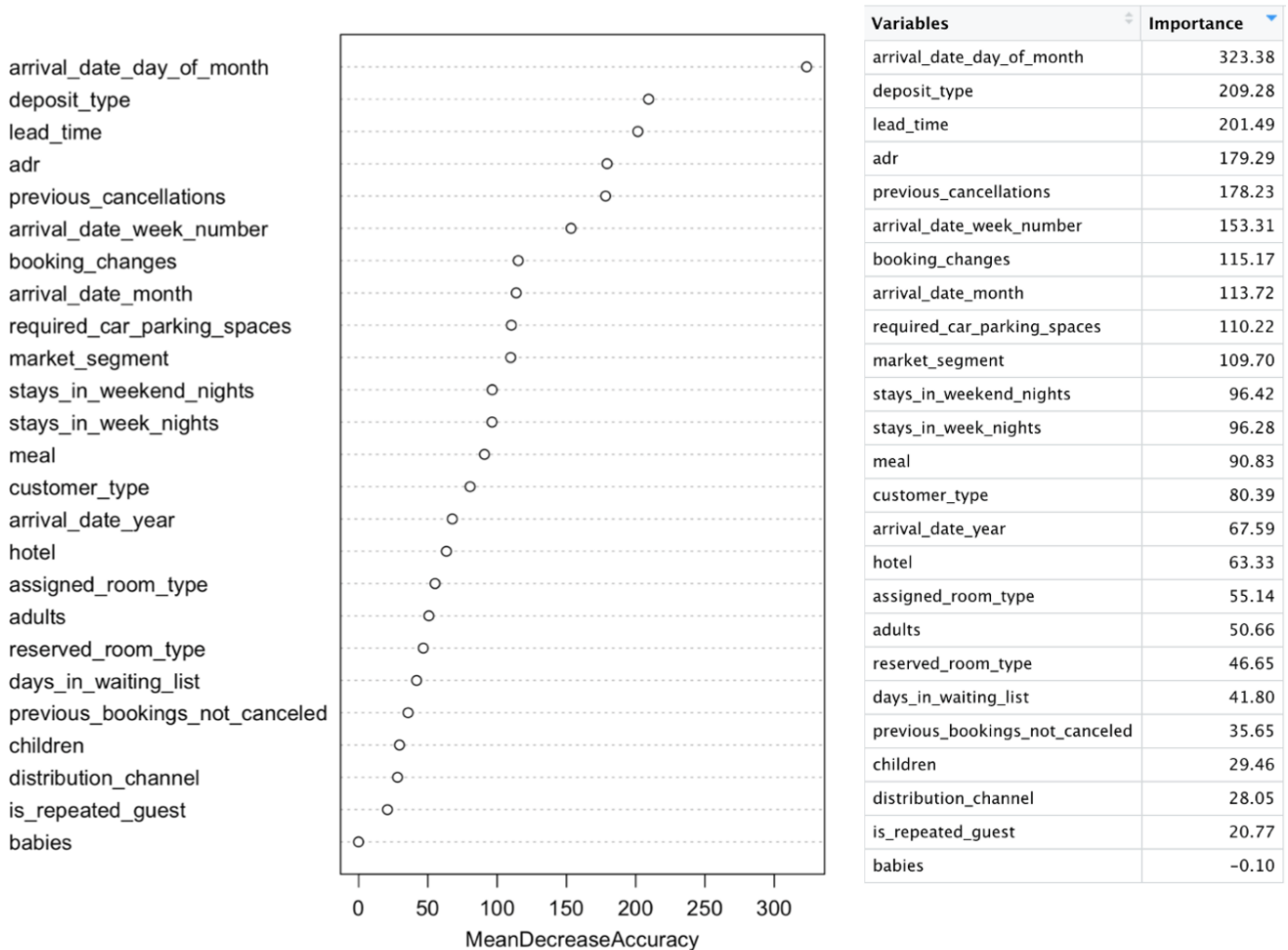


Figura 36: Feature selection; Fonte: realizzazione propria

Come si può vedere dalla fig. 36 la variabile *babies* comporta grazie alla sua esclusione dal modello comporta addirittura un aumento nell'accuratezza.

Mentre altre variabili come, ad esempio, *is repeated guest* e *distribution channel* pur avendo un'importanza maggiore rispetto alla variabile esaminata precedentemente non comportano comunque una perdita moto elevata di accuratezza.

Poiché un modello dovrebbe essere rappresentativo della realtà ma senza complicarla eccessivamente, accuratezza permettendo, è sempre positivo riuscire ad eliminare dal modello l'utilizzo di una o più variabili. Ho dunque eliminato le variabili: *babies*, *is repeated guest*, *distribution channel*, *previous bookings not canceled*, *days in waiting list*, *reserved room type*, *adults*, *assigned room type*, *hotel*, *arrival date year*.

Inoltre, quando si costruisce un modello si può anche analizzare la matrice di correlazione.

Due variabili molto correlate tra loro, sia positivamente che negativamente portano, infatti, a risultati nel modello non troppo soddisfacenti o comunque a risultati simili a quelli che si otterrebbero considerando solo una delle due variabili correlate.

Tuttavia, le correlazioni esistenti tra le variabili non sono molto forti, essendo la più alta, tra *lead time* e notti soggiorno, di solo 0.30. Ho quindi provato ad effettuare nuovamente la costruzione del modello, eliminando

la variabile *stays in week nights*. Ho scelto di eliminare questa di variabile e non *lead time* poiché quest'ultima ha, secondo il grafico analizzato precedentemente, maggiore importanza.

Di seguito i risultati ottenuti dalle varie fasi di costruzione del modello.

Prima feature selection

		Reference	
Prediction		0	1
0	20164	3224	
1	2385	10042	

Accuracy : 0.8434
Sensitivity : 0.8942
Specificity : 0.7570

26 variabili

Dopo feature selection

		Reference	
Prediction		0	1
0	20083	3514	
1	2466	9752	

Accuracy : 0.833
Sensitivity : 0.8906
Specificity : 0.7351

15 variabili

Dopo analisi correlazione

		Reference	
Prediction		0	1
0	20020	3551	
1	2529	9715	

Accuracy : 0.8302
Sensitivity : 0.8878
Specificity : 0.7323

14 variabili

Figura 37: Confronto accuratezza tra modelli; Fonte: realizzazione propria

Come è possibile verificare nella fig. 37, la scelta del modello da utilizzare è ricaduta sull'ultimo modello creato, cioè quello che segue sia alla *feature selection* che all'analisi della correlazione. Come si può verificare, le variabili comprese in quest'ultimo modello sono 14, dodici in meno rispetto alla prima versione di modello. Tuttavia, nonostante una diminuzione notevole delle variabili considerate, l'accuratezza è diminuita di solo 0.1, la *specificity* di 0.2 e la *sensitivity* è rimasta pressoché invariata.

Tutto ciò a dimostrazione che le variabili eliminate dal modello non avevano un'importanza molto elevata nella previsione della cancellazione e che quindi il loro non inserimento è stata una scelta corretta.

		Reference	
Prediction		0	1
0	20020	3551	
1	2529	9715	

Figura 38: Confusion matrix random forest; Fonte: realizzazione propria

Passando all'analisi del modello prescelto, come si può vedere dalla matrice rappresentata in fig. 38 i casi in cui correttamente viene predetto che la prenotazione non verrà cancellata sono 20.020; mentre i casi in cui, correttamente, vengono predette le cancellazioni sono 9.715.

Per valutare però l'accuratezza del modello occorre analizzare oltre che i *true positive* che i *false negative* anche i casi in cui la predizione non si è rivelata corretta.

Nello specifico 2.529 rappresentano i falsi negativi; cioè i casi in cui è stata predetta la classe negativa, cioè la cancellazione, ma in realtà si trattava di prenotazioni che non sono state cancellate.

Viceversa, 3.551 rappresenta i casi falsi positivi cioè quelli in cui si è predetta la classe positiva, quindi la non cancellazione, ma in realtà si trattava di osservazioni appartenenti alla classe negativa.

L'accuratezza è dell'83% mentre la specificity del 73% e la sensitivity dell'89%.

Per semplificare ulteriormente il modello, ho provato a considerare meno delle 14 variabili del modello sopra enunciato; tuttavia, l'accuratezza ne ha risentito. In particolare, ho effettuato preliminarmente una nuova operazione di *feature selection* a cui è seguita la generazione di nuovi modelli predittivi creati eliminando le variabili che risultavano meno importanti. Già dalla considerazione di 13 variabili l'accuratezza è diminuita seppur di poco. Ho comunque provato a costruire un modello composto da sole 10 variabili, le più importanti secondo la *feature selection*. Il problema di questo modello è che l'accuratezza è diminuita ancora di 2 punti percentuali così come è diminuita di 3 punti percentuali la *sensitivity*. Per compensare la riduzione dell'accuratezza, così come della *sensitivity*, è necessario che il modello si semplifichi di molto altrimenti la perdita risulta essere vana. Alla luce dei risultati dei sopra evidenziati ho scelto, quindi, di non considerare questo modello ma quello precedentemente analizzato che garantisce migliori risultati.

Comunicare con il cliente

Come visto nei precedenti paragrafi, applicare modelli di ML può essere molto utile all'interno di una realtà alberghiera.

In particolare, attraverso l'utilizzo di metodi di *random forest* e/o *decision tree*, si può stabilire se una prenotazione può o meno essere suscettibile di cancellazione.

Dopo aver sviluppato un modello in grado di predire ciò, occorre fissare il tasso di *overbooking* in modo tale che le prenotazioni cancellate si bilancino con le prenotazioni extra che sono state prese.

Tuttavia, un'ulteriore strada perseguibile per cercare di limitare al meglio le cancellazioni è quella di comunicare attivamente con il cliente.

Infatti, il modello costruito, una volta messo in funzione, andrà ad attribuire delle classificazioni alle varie prenotazioni. Una prenotazione potrà, infatti, essere classificata come non suscettibile di cancellazione oppure come possibile cancellazione. Tuttavia, questa classificazione può variare nel tempo poiché il sistema è in grado di apprendere continuamente attraverso l'immissione di nuovi dati, corrispondenti alle nuove prenotazioni. Ciò produce una classificazione che varia nel tempo, anche giornalmente.

Questo è il punto fondamentale per cui risulta essere conveniente sviluppare un modello di *machine learning*. Poiché il suo funzionamento e la sua classificazione varierà e si perfezionerà all'aumentare dei dati e al variare degli stessi.

Per questo motivo, nello studio di Nuno Antonio et al. presentato nella 16° IEEE *International Conference on Machine Learning and Applications (ICMLA)* nel 2017 con il titolo di "*Predicting hotel bookings cancellation with a machine learning classification model*", gli autori hanno utilizzato una metrica per capire la frequenza con cui è stata predetta la cancellazione (Nuno Antonio A. D., 2017).

La metrica utilizzata è stata quella della frequenza minima (MF). Questa metrica è stata calcolata dividendo i giorni in cui la prenotazione è stata classificata come probabile cancellazione con i giorni totali.

Ciò significa che, maggiore è la metrica più volte la prenotazione è stata classificata come probabile cancellazione.

Lo stesso studio ha poi proseguito analizzando l'effetto che una comunicazione ha sulla possibilità di cancellazione della prenotazione.

Il modello utilizzato dallo studio per effettuare la previsione presenta un'accuratezza abbastanza elevata, pari all'84%. Il modello utilizzato da Nuno Antonio et al. oltre ad aver ottenuto una buona accuratezza, non sembra neanche essere incline all'*overfitting* o alla classificazione eccessiva degli arrivi futuri come probabili cancellazioni.

Nello studio, si è poi analizzato come la percentuale delle cancellazioni vari in presenza di una comunicazione con il cliente attivata da parte della struttura.

In prima istanza il *management*, anche sotto consiglio degli autori, aveva deciso di contattare telefonicamente i clienti classificati come "probabile cancellazione" offrendo sconti e servizi aggiuntivi.

Tuttavia, a detta della struttura, questa operazione era troppo onerosa sia in termini di tempo che di spesa.

Infatti, il dover contattare telefonicamente uno per uno i clienti era dispendioso ed inoltre, spesso si verificava che il cliente, a cui era stato offerto un particolare *upgrade*, ne chiedesse in aggiunta anche un altro.

Per tutto ciò, si è cambiata strategia. Si è passati, infatti, a un contatto effettuato via mail e senza l'offerta di sconti o pacchetti di *upgrade* ma esclusivamente chiedendo informazioni riguardanti il soggiorno.

In particolare, si richiedevano informazioni come l'ora del *check-in* o il tipo di letto che il cliente prediligeva. Alla comunicazione sono conseguite due differenti reazioni: o i soggetti sentendosi curati da parte della struttura non disdicevano più la prenotazione oppure disdicevano comunque la prenotazione. Tuttavia, anche quest'ultimo caso può in un qualche modo essere classificato come positivo poiché, avendo la certezza della cancellazione, la sistemazione può essere riassegnata ad un altro cliente.

Inoltre, sfruttando il contatto, si potrebbe anche provare ad offrire al cliente servizi aggiuntivi ad un prezzo scontato in modo da mettere in atto tecniche di *upselling*.

Hotel	Action	MF ≥ 0% (all bookings)			MF ≥ 50%		
		Canceled	Not canceled	% Canceled	Canceled	Not canceled	% Canceled
H1	No	471	1,429	24.8%	125	153	45.0%
	Yes	12	97	11.0%	6	70	7.9%
H2	No	1,010	2,905	25.8%	269	325	45.3%
	Yes	15	181	7.7%	9	111	7.5%

Figura 39: Tabella riassuntiva dei risultati; Fonte: Nuno Antonio et al. "An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations" (2019)

Nella fig. 39, sono riassunti gli effetti ottenuti dal contatto con il cliente sulle cancellazioni.

Nei casi in cui si è avuto il contatto del cliente, identificate da “yes”, il tasso di cancellazione è diminuito in tutti i casi.

Come delineato dagli stessi autori, per quanto riguarda le prenotazioni in generale, quelle con un $MF \geq 0$, e in riferimento al resort hotel (H1), la percentuale di cancellazione a seguito dell'azione intrapresa dalla struttura è diminuita di circa 13 punti percentuali con una diminuzione relativa delle cancellazioni pari al 56%. Per quanto riguarda invece il *city* hotel, questa è diminuita di 18 punti percentuali, che si traduce in una diminuzione relativa delle cancellazioni del 70%. È stato inoltre condotto dagli autori un test *Chi-square* che ha dato esito positivo, e quindi la differenza è statisticamente significativa per entrambi gli hotel.

Per le prenotazioni per le quali è stata prevista la probabilità di cancellazione in almeno la metà dei giorni in cui il modello è stato applicato alle stesse, le differenze sono ancora più marcate. Si ha, infatti, una differenza di 37 punti percentuali per entrambi gli hotel; anche in questo caso la differenza risulta essere significativa.

Ulteriore elemento da indagare riguarda l'*Odds Ratio*. Quest'indice è il rapporto tra la frequenza con la quale un evento si verifica in un gruppo e la frequenza con la quale lo stesso evento si verifica in un gruppo di controllo. È quindi, una misura di efficacia dell'azione intrapresa.

L'*Odds Ratio* risulta essere pari ad 9.3 per H1, cioè per il resort hotel, e pari a 10 per H2 e quindi per il *city* hotel quando si prendono in considerazione le prenotazioni con un $MF \geq 50\%$.

Questo indica che, non contattare un cliente comporta un fattore di aumento della cancellazione pari a 9.3 per H1 e 10 per H2 (Nuno Antonio A. d., *Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior*, 2019).

Sempre secondo quanto studiato dagli autori del *paper* appena esaminato, il ricavo delle camere che non è stato perso a causa delle cancellazioni è di 16k euro per H1 e di 22k euro per H2.

Di conseguenza, la comunicazione porta dei vantaggi e non solo sull'evitare la cancellazione.

Quando, infatti, si comunica con il cliente questo potrebbe disdire la sua prenotazione, il che a primo impatto potrebbe sembrare negativo tuttavia, in questo modo, si è a conoscenza della cancellazione e di conseguenza la struttura potrà procedere alla riassegnazione della stanza.

[Attuare strategie comunicative personalizzate in base alla generazione di appartenenza](#)

Si è dimostrato, quindi, che una comunicazione con il cliente può portare lo stesso a non disdire la propria prenotazione con conseguenze abbastanza evidenti anche in ambito finanziario.

Tuttavia, nel primo capitolo si è esaminato come i soggetti, appartenenti a differenti fasce d'età, oltre a prediligere strumenti comunicativi differenti, ricercano anche nella struttura *benefit* differenti.

Di conseguenza, per attuare una migliore comunicazione che porterebbe a migliori risultati per quanto riguarda la percentuale di cancellazioni e di conseguenza sui ricavi, occorrerebbe prendere in considerazione questo aspetto.

Ma prima di concentrarsi sulla generazione e sui *benefit* maggiormente efficaci, occorre effettuare un'ulteriore distinzione.

Come già detto nei precedenti capitoli, generalmente si effettua un viaggio per due differenti motivazioni: o per piacere o per lavoro.

È chiaro che, essendo soggiorni con caratteristiche differenti, anche i *benefit* su cui far leva per cercare di convincere il cliente a non cancellare la propria prenotazione dovranno essere differenti.

I viaggiatori *business*, come già detto, non sono molto sensibili al prezzo e di conseguenza offrire uno sconto potrebbe non funzionare. Viceversa, questo tipo di viaggiatori, ricercherà la velocità nei servizi, il *check-in* da app offerto dalla catena Hilton, ad esempio, e per questo occorrerà offrire agli stessi servizi che lo facilitano nel soggiorno.

Un *benefit* che potrebbe essere adatto a questo tipo di clientela, e potrebbe diminuire la percentuale di cancellazione rispetto a questo segmento, è ad esempio offrire il servizio di *transfert* veloce da e per la struttura.

In questo modo, il soggetto potrà utilizzare il tempo che avrebbe impiegato per raggiungere la struttura lavorando o riposando.

Il discorso è differente, invece, per quanto riguarda i clienti *leisure*.

Questa tipologia di clientela è, infatti, maggiormente sensibile al prezzo e di conseguenza come incentivo per evitare la cancellazione si potrebbero proporre sconti o omaggi.

Tuttavia, mentre la clientela *business* ha al suo interno esigenze e caratteristiche abbastanza omogenee, la stessa cosa non può dirsi per la clientela *leisure*. La categorizzazione dei viaggiatori per piacere, invece, può contenere al suo interno soggetti che presentano differenti bisogni poiché, ad esempio, alcuni potrebbero viaggiare in famiglia altri con amici.

È soprattutto per questa clientela che può essere quindi utile seguire la categorizzazione per fasce d'età.

Sia per quanto riguarda la comunicazione e la pubblicizzazione nella fase iniziale della *customer journey*, che nella fase finale della stessa a seconda dell'età si possono, infatti, avere *touchpoint* e/o *benefit* che possono meglio funzionare di altri.

Approfondendo e facendo un discorso generazionale, i *baby boomers*, cioè i soggetti nati tra il 1946 e il 1964, sono i soggetti che maggiormente ricercano informazioni sul viaggio e sulla struttura. Da un punto di vista comunicativo questi prediligono i siti che contengono molte informazioni e spiegazioni (Emilie Slootweg, 2018); ed è proprio su questa ricerca di informazioni che si potrebbe far leva.

Visto che, come detto precedentemente, anche i risultati basati sulla ricerca di Nuno et al. del 2019 poggiano quasi esclusivamente su comunicazioni non offerenti *benefit* monetari, per questa generazione potrebbe essere una buona strategia contattare il cliente anche solo semplicemente per fornire qualche informazione aggiuntiva sul soggiorno e/o sulla struttura e rendersi disponibili a rispondere a qualsiasi dubbio.

Per quanto riguarda il mezzo comunicativo, seppur per rendersi visibile al soggetto nella parte iniziale della *customer journey* non si devono escludere mezzi di comunicazione tecnologici, in questo caso il mezzo

migliore potrebbe essere rappresentato dal telefono. Questo perché, per fornire informazioni e rendersi disponibili al soggetto il telefono sembra essere lo strumento migliore.

Per quanto riguarda i nati tra il 1965 e il 1980, cioè la generazione X, questi ricercano molto nella struttura stanze comode e una molteplicità di servizi a loro disposizione (Emilie Sloomweg, 2018).

Inoltre, sono anche i soggetti che potenzialmente presentano il maggior reddito.

Per tutti questi motivi, in questo caso, una buona strategia potrebbe essere quella di offrire, qualora disponibile, un *upgrade* della camera, gratuitamente o con un piccolo *extra*, oltre che uno sconto per un servizio offerto in struttura.

Questo perché, oltre a costruire una migliore relazione con lo stesso, che potrebbe portare ad una non cancellazione della prenotazione, rappresenta anche il modo per effettuare strategie di *up-selling* e di *cross selling*.

Se, infatti, al soggetto viene offerta come *upgrade* una stanza di qualità superiore, questo potrebbe decidere la prossima volta che soggiognerà presso la struttura di prenotare nuovamente quella stanza; poiché, come detto, sono la generazione che presenta potenzialmente il maggior reddito e che pretende una stanza con tutte le comodità. Di conseguenza, qualora il soggetto avesse pagato un piccolo *extra* per la stanza di qualità superiore, non si manifesterebbe solo un *up-selling* immediato ma anche una crescita del valore potenziale della relazione.

Per quanto riguarda il mezzo comunicativo da prediligere questo è sicuramente l'e-mail, poiché è uno dei mezzi maggiormente consigliati per raggiungere questa generazione.

Viceversa, offrendo sconto per un servizio si potrebbe generare una strategia di *cross selling*, in grado di aumentare il potenziale guadagno dato da quella relazione e quella prenotazione.

Per quanto riguarda la generazione Y, nati tra il 1981 e il 1995, questi sono i soggetti che maggiormente ricercano e danno importanza al nome del brand e alla personalizzazione della comunicazione e dei servizi (Emilie Sloomweg, 2018).

Occorre, quindi, in questo caso puntare tutto sulla personalizzazione dell'esperienza del cliente. Chiaramente questa dipenderà dai servizi e dalla tipologia di struttura. Tuttavia, esistono personalizzazioni applicabili ovunque e su qualsiasi categoria di hotel che presentano anche un basso, se non nullo, costo di realizzazione. Ad esempio, si potrebbe chiedere al soggetto quale tipologia di cuscino preferisce o a che piano e in che posizione preferisce la stanza.

È logico che, qualora si abbiano a disposizione molti servizi offerti in struttura, la personalizzazione può coinvolgere questi ed essere maggiormente presente e forte.

La generazione Z racchiude, invece, i nati dopo il 1996. Questa è la generazione tecnologica per eccellenza e che predilige come strumenti comunicativi tutto ciò che è basato sui *social* e sull'*online*.

È la generazione che, a differenza di tutte le altre, è attenta ai valori comunicati dai brand che acquista tanto da non effettuare una compera se il valore comunicato non è coerente con la sua persona.

Per questo, un'ottima strategia potrebbe consistere nell'inviare a questa categoria di clienti un'informativa, attraverso i *social* o attraverso messaggistica istantanea, di quali sono gli accorgimenti adottati dalla struttura per proteggere un certo valore. In altre parole, inviare un'informativa in cui si esorta il soggetto a non portare contenitori di plastica poiché tutto il necessario per il loro soggiorno, come ad esempio i saponi, verrà fornito loro in modo sfuso attraverso dei *dispenser*. In questo modo, si fornisce un doppio vantaggio al cliente; infatti, si dimostrerà la propria attenzione verso l'ambiente e al contempo si migliorerà il rapporto con lo stesso. Queste appena delineate sono solo alcune delle tante strategie applicabili per cercare di migliorare la relazione con il cliente e di conseguenza evitare una cancellazione della domanda.

Di conseguenza, occorrerebbe effettuare un'analisi verificando quale strategia è più efficace.

Tuttavia, per l'applicazione di queste strategie personalizzate e basate sulla fascia d'età è richiesta la conoscenza dei dati demografici dei clienti prima dell'arrivo in struttura e del *check-in* e questa potrebbe rivelarsi una difficoltà.

Infatti, mentre questi dati, quando si effettuano prenotazioni senza l'utilizzo di intermediari, possono essere richiesti al momento di conferma della richiesta di prenotazione da parte del cliente, quando la prenotazione proviene da OTA *online* questo processo potrebbe essere più complicato.

In tutto ciò, il *management* deve sempre e comunque aggiornare i propri profili *social* poiché, come è stato possibile verificare, questi vengono utilizzati ormai da tutti senza distinzione d'età.

I *social* non sono, infatti, solo un mezzo utilizzato per condividere foto o la propria esperienza ma rappresenta anche un modo per informarsi e tenersi aggiornati.

Conclusioni

Dal lavoro di tesi e dalle analisi effettuate è emerso che è possibile effettuare la previsione di una cancellazione alberghiera anche con una precisione abbastanza elevata.

Inoltre, è stato possibile verificare attraverso l'analisi e lo studio di *paper* di ricerca che la comunicazione con il cliente, classificato come "probabile cancellazione", porta risultati positivi sulla percentuale di cancellazione. Infatti, quando si ha la comunicazione con il cliente, che può avere ad oggetto l'offerta di *benefit* o semplicemente una richiesta di informazioni, la percentuale delle cancellazioni si riduce e questo è sintomo che la strategia funziona correttamente.

Ma ripercorrendo il lavoro e l'intero elaborato, si è visto che si può procedere con differenti modalità nel prevedere questa cancellazione.

Si potrà procedere, con riguardo al *machine learning* e come si è esaminato durante la trattazione con due differenti modalità: attraverso la generazione e l'utilizzo del *decision tree* o alternativamente del *random forest*.

Il *decision tree* sarebbe l'algoritmo da prediligere in questi casi poiché presenta una più veloce esecuzione e un *output* maggiormente leggibile.

Tuttavia, non sempre questo algoritmo produce risultati soddisfacenti a livello di accuratezza e quindi, in questi casi occorre considerare l'algoritmo del *random forest*.

Questo è cioè che è successo nell'analisi condotta; infatti, il *decision tree* costruito non ha avuto risultati soddisfacenti né per quanto ha riguardato l'accuratezza né per quanto ha riguardato la *specificity* e la *sensitivity*.

Per cercare di aumentare l'accuratezza ho, quindi, proceduto ad applicare l'algoritmo di *random forest*. Questo promette, visto il numero variabile di alberi che si possono includere nello stesso, una maggiore accuratezza rispetto al semplice albero decisionale.

Tuttavia, per cercare di diminuire le variabili incluse nel modello, ho effettuato una *feature selection*.

Questa ha permesso, infatti, di studiare l'importanza delle variabili in modo da capire quali variabili siano più importanti e come la loro eliminazione impatta sull'accuratezza.

Si è visto, ad esempio, che l'eliminazione della variabile *babies* non comporta una riduzione dell'accuratezza ma un aumento della stessa. Sono state, inoltre, eliminate ulteriori variabili.

In questo modo, il modello prodotto conserva un'accuratezza pressoché invariata, rispetto al modello di *random forest* sviluppato inizialmente, ma con un numero nettamente inferiore di variabili.

In particolare, il modello predittivo, che prende in considerazione 14 variabili, ha un'accuratezza pari all'83% mentre la *specificity* è del 73% e la *sensitivity* dell'89%.

Il compito del modello è, appunto, quello di classificare ogni singola prenotazione che verrà immessa nel sistema di gestione per cercare di capire se questa è suscettibile o meno di essere cancellata.

La presente ricerca apre nuove strade per approfondire ulteriormente il tema della previsione della cancellazione. In particolare, disponendo di maggiori dati a disposizione occorrerebbe sviluppare un modello

che, oltre alle variabili prese in oggetto, consideri anche fattori quali l'età o il genere. Lo studio della cancellazione prendendo ad oggetto anche questi fattori appena delineati potrebbe, infatti, permettere di costruire un *identikit* del soggetto che maggiormente potrebbe essere incline alla cancellazione.

Naturalmente, sapendo questo, le azioni di *marketing* e di comunicazione dovranno essere incentrate maggiormente su questi soggetti.

Ulteriore aspetto da indagare riguarda gli effetti sulla cancellazione dei *benefit* proposti a fine del quarto capitolo. Occorrerebbe infatti studiare, attraverso una ricerca qualitativa e/o quantitativa, se i *benefit* proposti per ogni differente fascia generazionale funzionano o ne sono presenti ulteriori che producono un miglior effetto. Non è stato possibile, infatti, indagare questo aspetto nella presente ricerca poiché non si aveva alcun dato demografico a disposizione.

Appendice

```
#####Importo librerie #####
library(readr)
library(ggplot2)
library(viridis)
library(dbplyr)
library(scales)
library(tidyverse)
library(lubridate)
library(countrycode)
library(gridExtra)
library(plotly)
library(caret)
library(e1071)
library(zeallot)
library(countrycode)
library(ISLR)
library(rpart)
library(rpart.plot)
library(rattle)
library(GoodmanKruskal)
library(arm)
library(randomForest)
library(rpart.plot)
library(foreach)
library(doParallel)
library("PerformanceAnalytics")
library(corrplot)
library(correlation)
library(caTools)
library(rpart)
library(rpart.plot)
```

```
##### Imposto la directory e carico il file csv #####
setwd("~/Desktop/Tesi/R")
hotel_bookings <- read_csv("hotel_bookings.csv")
H1 <- read_csv("H1.csv")
H2 <- read_csv("H2.csv")
```

```
Booking <- hotel_bookings %>%
  mutate(arrival_date_month = factor(
    arrival_date_month,
    levels = c(
      "January",
      "February",
      "March",
      "April",
      "May",
      "June",
      "July",
      "August",
      "September",
      "October",
      "November",
      "December"
    ),
    ordered = TRUE
  ))
```

```
Booking1 <- na.omit(Booking)
Booking2 = Booking1 [Booking1$distribution_channel!= "Undefined",]
```

```
summary(Booking2)
summary(H1)
summary(H2)
#####OPERAZIONI ULTERIORI#####
```

```
hotelbycountry <- Booking2 %>% group_by(country) %>% summarise(n = n())
hotelbycountry$country_name <- countrycode(hotelbycountry$country, origin = "iso3c", destination = "country.name")
```

```
Booking2<-Booking2%>%
```

```
mutate(
  hotel=as.factor(hotel),
  is_canceled=as.factor(is_canceled),
  meal=as.factor(meal),
  country=as.factor(country),
  market_segment=as.factor(market_segment),
  distribution_channel=as.factor(distribution_channel),
  is_repeated_guest=as.factor(is_repeated_guest),
  reserved_room_type=as.factor(reserved_room_type),
  assigned_room_type=as.factor(assigned_room_type),
  deposit_type=as.factor(deposit_type),
  customer_type=as.factor(customer_type),
  reservation_status=as.factor(reservation_status),
  agent=as.factor(agent),
  company=as.factor(company),
  arrival_date_day_of_month=as.factor(arrival_date_day_of_month),
  arrival_date_month=as.factor(arrival_date_month),
  arrival_date_year=as.factor(arrival_date_year)
)
```

```
#####GGPLOT#####
```

```
#Prenotazioni e percentuali cancellati/non cancellati
```

```
ggplot(data = Booking2,
  aes(
    x = hotel,
    y = prop.table(stat(count)),
    fill = factor(is_canceled),
    label = scales::percent(prop.table(stat(count)))
  )) +
  geom_bar(position = position_dodge()) +
  geom_text(
    stat = "count",
    position = position_dodge(.9),
    vjust = -0.5,
    size = 3
  ) +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Reservation Status", x = "Tipo hotel", y = "Percentuale") +
  theme_classic() +
  scale_fill_manual( values = c("green", "red", "green", "red"),
    name = "Reservation Status",
    breaks = c("0", "1"),
    labels = c("Not Cancelled", "Cancelled")
  )
)
```

```
# Quantità prenotazioni cancellate e non
```

```
ggplot(Booking2, aes(hotel, fill = factor(is_canceled))) +
  geom_bar() +
  geom_text(stat = "count",
    aes(label = ..count..),
    position = position_stack(vjust = 0.5)) +
  scale_fill_discrete(
    name = "Reservation Status",
    breaks = c("0", "1"),
    label = c("Not Cancelled", "Cancelled")
  ) +
  labs(title = "Cancellazioni e non a seconda del tipo di hotel", x = "Tipo hotel", y = "Quantità") +
  scale_x_discrete(labels = c("City Hotel", "Resort Hotel"))
```

```
# Prenotazioni cancellate e non al variare del mese
```

```
ggplot(Booking2, aes(arrival_date_month, fill = factor(is_canceled))) +
  geom_bar() +
  geom_text(stat = "count",
    aes(label = ..count..),
    position = position_stack(vjust = 0.5)) +
  scale_fill_discrete(
    name = "Reservation Status",
    breaks = c("0", "1"),
    label = c("Non cancellate", "Cancellate")
  )
```

```

) +
labs(title = "Reservation Status al variare del mese", x = "Mese", y = "Quantità") +
scale_x_discrete(labels = month.abb)

#Come variano le prenotazioni al variare dei mesi
ggplot(Booking2, aes(arrival_date_month, fill = hotel)) +
  geom_bar(position = position_dodge()) +
  labs(title = "Prenotazioni al variare dei mesi",
        x = "Mesi",
        y = "Quantità") + theme_bw()

#Boxplot del booking status, del lead time ed in base al tipo di hotel
ggplot(data = Booking2, aes(
  x = hotel,
  y = lead_time,
  fill = factor(is_canceled)
)) +
  geom_boxplot(position = position_dodge()) +
  labs(
    title = "Come varia la cancellazione in base al tipo di hotel",
    subtitle = "Ed in base al lead time",
    x = "Tipo di hotel",
    y = "Lead time"
  ) +
  scale_fill_discrete(
    name = "Booking Status",
    breaks = c("0", "1"),
    labels = c("Non cancellata", "Cancellata")
  ) + theme_light()

#Come varia la cancellazione in base al tipo di distribuzione
ggplot(Booking2, aes(x=distribution_channel, fill = is_canceled)) +
  geom_bar() +
  geom_text(stat = "count",
            aes(label = ..count..),
            position = position_stack(vjust = 0.5)) +
  scale_fill_discrete(
    name = "Status prenotazione",
    breaks = c("0", "1"),
    label = c("Non cancellata", "Cancellata")
  ) +
  labs(title = "Come variano le cancellazioni al variare del canale distributivo", x = "Canale distributivo", y = "Quantità") + theme_light()

#Come varia la scelta di deposito/non in base al tipo di canale di distribuzione
ggplot(data = Booking2) +
  geom_bar(mapping = aes(x = distribution_channel, fill = deposit_type)) +
  scale_fill_discrete(
    name = "Tipo deposito",
    label = c("Senza deposito", "Non rimborsabile", "Rimborsabile")
  ) +
  labs(title = "Comparazione dei canali distributivi", subtitle = "Secondo il tipo di deposito")

#Come varia la cancellazione in base alla presenza di waiting list
Booking2%>%
  filter(days_in_waiting_list>1)%>%
  ggplot(aes(x=days_in_waiting_list,fill=is_canceled))+
  geom_histogram(binwidth = 10) +
  scale_fill_discrete(
    name = "Status prenotazione",
    label = c("Non cancellata", "Cancellata")
  ) +
  labs(title = "Come variano le cancellazioni", subtitle = "Al variare dei giorni in waiting list")

#Come varia la cancellazione in base alla presenza o meno del deposito
Refundable <- subset(Booking2, Booking2$deposit_type == "Refundable")
No_deposit <- subset(Booking2, Booking2$deposit_type == "No Deposit")
Non_refund <- subset(Booking2, Booking2$deposit_type == "Non Refund")

ggplot(data = Refundable,
  aes(
    x = deposit_type,
    y = prop.table(stat(count)),
    fill = factor(is_canceled),

```

```

    label = scales::percent(prop.table(stat(count)))
  )) +
  geom_bar(position = position_dodge()) +
  geom_text(
    stat = "count",
    position = position_dodge(.9),
    vjust = -0.5,
    size = 3
  ) +
  scale_y_continuous(labels = scales::percent) +
  theme_classic() +
  scale_fill_discrete(
    name = "Status Prenotazione",
    breaks = c("0", "1"),
    labels = c("Non cancellata", "Cancellata"))

ggplot(data = No_deposit,
  aes(
    x = deposit_type,
    y = prop.table(stat(count)),
    fill = factor(is_canceled),
    label = scales::percent(prop.table(stat(count)))
  )) +
  geom_bar(position = position_dodge()) +
  geom_text(
    stat = "count",
    position = position_dodge(.9),
    vjust = -0.5,
    size = 3
  ) +
  scale_y_continuous(labels = scales::percent) +
  theme_classic() +
  scale_fill_discrete(
    name = "Status Prenotazione",
    breaks = c("0", "1"),
    labels = c("Non cancellata", "Cancellata"))

#Come variano i numeri delle prenotazioni dei due hotel rispetto alla nazione
Booking3 <- Booking2 %>% group_by(country) %>% filter(n() > 1500)
Booking3$country_name <- countrycode(Booking3$country,
  origin = "iso3c",
  destination = "country.name")

ggplot(Booking3, aes(country_name, fill = hotel)) + geom_bar(stat = "count", position = position_dodge()) +
  geom_text(stat = "count",
    aes(label = ..count..),
    position = position_stack(vjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), panel.background = element_blank()) +
  labs(title = "Da dove provengono i clienti?", x = "Stato", y = "Conteggio")

#Come variano le cancellazioni in base allo stato
ggplot(Booking3, aes(country_name, fill = is_canceled)) + geom_bar(stat = "count", position = position_dodge()) +
  geom_text(stat = "count",
    aes(label = ..count..),
    position = position_stack(vjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), panel.background = element_blank()) +
  labs(title = "Cancellazioni in base allo stato", x = "Stato", y = "Conteggio") + scale_fill_manual(values = c("green", "red"))

Booking4 <- Booking2 %>% group_by(country) %>% filter(n() > 500) %>% filter(n() < 1500)
Booking4$country_name <- countrycode(Booking4$country,
  origin = "iso3c",
  destination = "country.name")

ggplot(Booking4, aes(country_name, fill = is_canceled)) + geom_bar(stat = "count",
  position = position_dodge()) +
  geom_text(stat = "count",
    aes(label = ..count..),
    position = position_stack(vjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), panel.background = element_blank()) +
  labs(title = "Cancellazioni in base allo stato", x = "Stato", y = "Conteggio") +
  scale_fill_manual(values = c("green", "red"))

#Come varia il segmento di clientela in base al tipo di distribuzione

```

```
ggplot(data = Booking2) +
  geom_bar(mapping = aes(x = distribution_channel, fill = market_segment)) +
  labs(title = "Comparazione dei canali distributivi", subtitle = "In base al segmento di mercato")
```

```
#Composizione della clientela nei due tipi di hotel
ggplot(data = Booking2) +
  geom_bar(mapping = aes(x = hotel, fill = market_segment)) +
  labs(title = "Come è composta la clientela in base al segmento")
```

```
ggplot(data = Booking2) +
  geom_bar(mapping = aes(x = hotel, fill = customer_type)) +
  labs(title = "Come è composta la clientela in base al segmento")
```

```
#Prenotazioni Transient come sono composte
ggplot(data = Booking2) +
  geom_bar(mapping = aes(x = customer_type == "Transient", fill = market_segment)) +
  labs(title = "Prenotazioni Transient", subtitle = "Composizione") +
  scale_fill_discrete(
    name = "Segmento di mercato") + theme_light()
```

```
#Prenotazioni Non Transient come sono composte
ggplot(data = Booking2) +
  geom_bar(mapping = aes(x = customer_type == "Transient-Party", fill = market_segment)) +
  labs(title = "Prenotazioni Transient-Party", subtitle = "Composizione") +
  scale_fill_discrete(
    name = "Segmento di mercato") + theme_light()
```

```
#plot customer type cancellation
dfc.cancelation <- Booking2 %>%
  mutate( booking_cancelations = ifelse(is_canceled == 0, "Non cancellata", "Cancellata")
  )
```

```
ThemePie <- theme(axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  axis.text.x = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank(),
  legend.position = "none",
  plot.title = element_text(size=20, hjust = 0.5),
  plot.subtitle = element_text(size=15, hjust = 0.5)
)
```

```
Pie1 <- dfc.cancelation %>%
  group_by(booking_cancelations) %>%
  summarize(bookings_count = n()) %>%
  mutate(Percent = round(bookings_count / sum(bookings_count), 3)) %>%
  ggplot(aes(x = "", y = bookings_count, fill = booking_cancelations)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(Percent*100, "%")), position = position_stack(vjust=.5), size = 7) +
  labs(title = "Prenotazioni cancellate e non",
    subtitle = "Colore più chiaro indica le cancellazioni") +
  theme_minimal() +
  ThemePie +
  scale_fill_brewer(palette = "Blues")
```

```
#2. Plot 2 by Transient customers
Pie2 <- dfc.cancelation %>%
  filter(customer_type == "Transient") %>%
  group_by(booking_cancelations) %>%
  summarize(bookings_count = n()) %>%
  mutate(Percent = round(bookings_count / sum(bookings_count), 3)) %>%
  ggplot(aes(x = "", y = bookings_count, fill = booking_cancelations)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(Percent*100, "%")), position = position_stack(vjust=.5), size = 7) +
  labs(title = "Transient") +
  theme_minimal() +
  ThemePie +
  scale_fill_brewer(palette = "Greens")
```

```

#3. Plot 3 by Transient Party customers
Pie3 <- dfc.cancelation %>%
  filter(customer_type == "Transient-Party") %>%
  group_by(booking_cancelations) %>%
  summarize(bookings_count = n()) %>%
  mutate(Percent = round(bookings_count / sum(bookings_count), 3)) %>%
  ggplot(aes(x = "", y = bookings_count, fill = booking_cancelations)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(Percent*100, "%")), position = position_stack(vjust=.5), size = 7) +
  labs(title = "Transient Party") +
  theme_minimal() +
  ThemePie +
  scale_fill_brewer(palette = "Greens")

#4. Plot 4 by Group customers
Pie4 <- dfc.cancelation %>%
  filter(customer_type == "Group") %>%
  group_by(booking_cancelations) %>%
  summarize(bookings_count = n()) %>%
  mutate(Percent = round(bookings_count / sum(bookings_count), 3)) %>%
  ggplot(aes(x = "", y = bookings_count, fill = booking_cancelations)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(Percent*100, "%")), position = position_stack(vjust=.5), size = 7) +
  labs(title = "Group") +
  theme_minimal() +
  ThemePie +
  scale_fill_brewer(palette = "Greens")

#5. Plot 5 by Contract customers
Pie5 <- dfc.cancelation %>%
  filter(customer_type == "Contract") %>%
  group_by(booking_cancelations) %>%
  summarize(bookings_count = n()) %>%
  mutate(Percent = round(bookings_count / sum(bookings_count), 3)) %>%
  ggplot(aes(x = "", y = bookings_count, fill = booking_cancelations)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(Percent*100, "%")), position = position_stack(vjust=.5), size = 7) +
  labs(title = "Contract") +
  theme_minimal() +
  ThemePie +
  scale_fill_brewer(palette = "Greens")

#6. Arrange plots by grid
grid.arrange(Pie1, Pie2, Pie3, Pie4, Pie5, ncol = 5)

#### CORRELAZIONE ####
#Seleziono le sole colonne numeriche del dataframe
Booking_gerico = unlist(lapply(Booking2, is.numeric))

#Creiamo un dataframe più piccolo che contiene le sole colonne numeriche
Booking_gerico = Booking2[,Booking_gerico]

#Usiamo la funzione cor per calcolare le correlazioni
corData = cor(Booking_gerico, method = c("spearman"))

#Plot delle correlazioni
corrplot(corData)

#### ALBERO DECISIONALE####
as.factor(Booking2$sis_canceled)
Booking2$reservation_status <- NULL
table(Booking2$sis_canceled)
(75165/(75165+44220))*100

set.seed(3000)
spl = sample.split(Booking2$sis_canceled, SplitRatio = 0.7)
Train = subset(Booking2, spl==TRUE)
Test = subset(Booking2, spl==FALSE)

```



```

DecisionTree = rpart(is_canceled ~ hotel + deposit_type + previous_cancellations + is_repeated_guest + arrival_date_month
+ stays_in_weekend_nights + stays_in_week_nights + adults + children + babies + meal + market_segment
+ previous_bookings_not_canceled + reserved_room_type + assigned_room_type +
  booking_changes + days_in_waiting_list + customer_type + adr + required_car_parking_spaces +
  total_of_special_requests + distribution_channel,
data = Train, method="class", minbucket=15)

prp(DecisionTree)
prp(DecisionTree, type=1, extra=100, fallen.leaves = TRUE, main="Decision Tree")

# Effettuare le predizioni

PredictVal = predict(DecisionTree, newdata = Test)
PredictVal
PredictVal [,2]
table(Test$sis_canceled, PredictVal [,2] > 0.5)

PredictCART = predict(DecisionTree, newdata = Test, type = "class")
PredictCART

testCanc = Test$sis_canceled

table(testCanc)

table(testCanc, PredictCART)
confusionMatrix(as.factor(PredictCART), Test$sis_canceled)

#Calcolo accuratezza (predizioni verificate/ predizioni totali)
(20195+8002)/(20195+8002+2354+5264) #0,78

#Specificity (TN/(TN+FP))
(8002/(8002+5264)) #0,60

#Sensitivity (TP/(TP+FN))
(20195/(20195+2354))#0,89

##### RANDOM FOREST #####
as.factor(Booking2$sis_canceled)
Booking2$reservation_status <- NULL

set.seed(3000)
spl = sample.split(Booking2$sis_canceled, SplitRatio = 0.7)
Train = subset(Booking2, spl==TRUE)
Test = subset(Booking2, spl==FALSE)

training_set1_features <- Train[c('hotel','is_canceled','lead_time','adults','children','babies','meal',
  'market_segment','distribution_channel','is_repeated_guest',
  'previous_cancellations','previous_bookings_not_canceled','reserved_room_type',
  'deposit_type','days_in_waiting_list','customer_type','adr',
  'required_car_parking_spaces', 'arrival_date_year', 'arrival_date_month',
  'arrival_date_week_number', 'arrival_date_day_of_month', 'stays_in_weekend_nights',
  'stays_in_week_nights', 'booking_changes', 'assigned_room_type')]

rf_model3<-randomForest(is_canceled~.,
  data=training_set1_features,
  ntree=500,
  cutoff=c(0.5,0.5),
  mtry=8,
  importance=TRUE)

rf_model3

rf_pred_prob<-predict(rf_model3,Test,type="prob")[,2]
rf_pred_class<-predict(rf_model3,Test,type="class")
confusionMatrix(as.factor(rf_pred_class),Test$sis_canceled)

##### FEATURE SELECTION
varImpPlot(rf_model3)
importance = importance(rf_model3)
varImportance = data.frame(Variables = row.names(importance),
  Importance =round(importance[, 'MeanDecreaseAccuracy'],2))

#dopo feature selection

```

```

training_set2_features <- Train[c('is_canceled','lead_time','meal',
                                'market_segment','previous_cancellations',
                                'deposit_type','customer_type','adr',
                                'required_car_parking_spaces','arrival_date_month',
                                'arrival_date_week_number','arrival_date_day_of_month','stays_in_weekend_nights',
                                'stays_in_week_nights','booking_changes')]

rf_model2<-randomForest(is_canceled~.,
                        data=training_set2_features,
                        ntree=500,
                        cutoff=c(0.5,0.5),
                        mtry=8,
                        importance=TRUE)

rf_model2

rf_pred_prob1<-predict(rf_model2,Test,type="prob")[,2]
rf_pred_class1<-predict(rf_model2,Test,type="class")
confusionMatrix(as.factor(rf_pred_class1),Test$is_canceled)

# dopo analisi correlazione
training_set3_features <- Train[c('is_canceled','lead_time','meal',
                                'market_segment','previous_cancellations',
                                'deposit_type','customer_type','adr',
                                'required_car_parking_spaces','arrival_date_month',
                                'arrival_date_week_number','arrival_date_day_of_month','stays_in_weekend_nights',
                                'booking_changes')]

rf_model1<-randomForest(is_canceled~.,
                        data=training_set2_features,
                        ntree=500,
                        cutoff=c(0.5,0.5),
                        mtry=8,
                        importance=TRUE)

rf_model1

rf_pred_prob2<-predict(rf_model1,Test,type="prob")[,2]
rf_pred_class2<-predict(rf_model1,Test,type="class")
confusionMatrix(as.factor(rf_pred_class2),Test$is_canceled)

#prova con 10 variabili
varImpPlot(rf_model1)
importance = importance(rf_model1)
varImportance = data.frame(Variables = row.names(importance),
                            Importance =round(importance[, 'MeanDecreaseAccuracy'],2))
training_set4_features <- Train[c('is_canceled','lead_time',
                                'market_segment','previous_cancellations',
                                'deposit_type','adr',
                                'required_car_parking_spaces',
                                'arrival_date_week_number', 'arrival_date_day_of_month',
                                'booking_changes')]
rf_model4<-randomForest(is_canceled~.,
                        data=training_set4_features,
                        ntree=500,
                        cutoff=c(0.5,0.5),
                        mtry=8,
                        importance=TRUE)

rf_pred_prob4<-predict(rf_model4,Test,type="prob")[,2]
rf_pred_class4<-predict(rf_model4,Test,type="class")
confusionMatrix(as.factor(rf_pred_class4),Test$is_canceled)

```

Bibliografia

- Association, M. M. (2014). Red Roof Inn Turns Flight Cancellations into Customers.
- Baker, T. M. (2002). Service package switching in hotel revenue management systems. *Decision Sciences*, 109-132.
- Buckhiester, B. (2011). Revenue Management as a Multi-disciplinary Business Process. *Journal of Hospitality Financial Management*,, 91-103.
- Chiang-Ming Chen, Y.-C. T.-H. (2015). The decision-making process of and the decisive factors in accommodation choice. *Current Issues in Tourism*.
- Chu, S.-C. (2021, Settembre 22). *Oxford Bibliographies*. Retrieved from Oxford Bibliographies: <https://www.oxfordbibliographies.com/view/document/obo-9780199756841/obo-9780199756841-0267.xml>
- Dallari, F. (s.d.). *Liuc*. Tratto da my.liuc:
https://my.liuc.it/MatSup/2008/Y71015/GPI_previsioni.pdf
- Digital, T. (2019). How travelers are using mobile in 2019; End traveler and industry research.
- Dolores Romero Morales, J. W. (2010). Forecasting Cancellation Rates for Services Booking Revenue Management Using Data Mining. *European Journal of Operational Research*, 554-562.
- Emilie Slootweg, B. R. (2018). My generation: A review of marketing strategies on different age groups. *Hospitality Management*, 85-92.
- F.M. Bass, K. G. (2001, Giugno 1). Forecasting Diffusion of a New Technology Prior to Product Launch. *INFORMS Journal on Applied Analytics* 31, pp. 82-93.
- Google, I. (2015). "Il consumatore connesso". Retrieved from <https://www.thinkwithgoogle.com/intl/it-it/tendenze-e-insight/tendenze-di-consumo/come-catturare-i-momenti-che-contano/>
- Google, T. w. (2016). *Think with Google*. Retrieved from <https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/micro-moments-travel-customer-journey/>
- Hyndman, A. (2018). *Forecasting: Principles and Practice*. Australia.
- Jain, A. P. (2022, Marzo 1). *neptuneblog*. Retrieved from neptuneblog:
<https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide>
- John T. Mentzer, T. P. (2008). SUPPLY CHAIN MANAGEMENT AND ITS RELATIONSHIP TO LOGISTICS, MARKETING, PRODUCTION, AND OPERATIONS MANAGEMENT . *JOURNAL OF BUSINESS LOGISTICS*.
- Jonas Wälter, F. d. (2020). Aiding vehicle scheduling and rescheduling using machine learning. *International Journal of Transport Development and Integration*, 308-320.
- Kelemen, J. (2007). From Artificial Neural Networks to Emotion Machines with Marvin Minsky. *Acta Polytechnica Hungarica*.
- Mehmet Erdem, L. J. (2016). An overview of hotel revenue management research and emerging key patterns in the third millennium. *Journal of Hospitality and Tourism Technology*, 300-312.
- Melloni, P. R. (n.d.). *Previsione della domanda*. Modena.

- Mokbel, R. (2021, Giugno 10). *LinkedIn*. Retrieved from <https://www.linkedin.com/pulse/machine-reasoning-taking-artificial-intelligence-from-rita-mokbel>
- Nielsen, M. (2015). *Neural Networks and Deep Learning*.
- Nikolopoulos, K. S. (2016). Forecasting Branded and Generic Pharmaceuticals. *International Journal of Forecasting*, 344–357.
- Noone, B. M. (2014). Effects of price and user-generated content on consumers' prepurchase evaluations on variably priced services. *Journal of Hospitality & Tourism Research*, 562-581.
- Nuno Antonio, A. d. (2017). Predicting Hotel Bookings Cancellation With a Machine Learning Classification Model. *16th IEEE International Conference on Machine Learning and Applications*, 1049-1054.
- Nuno Antonio, A. D. (2017). Predicting Hotel Bookings Cancellation With a Machine Learning Classification Model. In *16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1049-1054). Cancun: IEEE.
- Nuno Antonio, A. d. (2019). Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hospitality Quarterly*, 1-22.
- Nuno Antonio, A. d. (2019). Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hospitality Quarterly*, 298-319.
- Nuzzo, M. D. (2021). *Data Science e Machine Learning: Dai Dati alla Conoscenza*. Indipendente.
- Punia, S., Nikolopoulos, K., Prakash Singh, S., Madaan, J. K., & Litsiou, K. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research*, 4964-4979.
- Punia, S., Nikolopoulos, K., Prakash Singh, S., Madaan, J. K., & Litsiou, K. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research*, 4964-4979.
- Review, P. (n.d.). *About Us*. Retrieved from Product Review: <https://www.productreview.com.au/i/about-us>
- Rhonda Montgomery, L. A. (2018). Online Traveler reviews as sOcial influence: Price is nO lOnger King. *Journal of Hospitality & Tourism Research*, 445–475.
- Rowson. (2016). *Modern hotel operations management*.
- SiteMinder. (2022, gennaio 25). Retrieved from SiteMinder: <https://www.siteminder.com/it/notizie/top-hotel-booking-revenue-makers-2021/>
- Stuart Russell, P. N. (2005). *Intelligenza artificiale. Un approccio moderno*. Pearson.
- Ugur Yavasa, E. B. (2005). Dimensions of hotel choice criteria: congruence between business and leisure travelers. *Hospitality Management*, 359-367.
- Wen-Chyuan Chiang, J. C. (2006). An overview of research on revenue management: current issues and future research . *International Journal of Revenue Management*, 97-128.

Zvi Schwartz, M. U. (2016). Hotel daily occupancy forecasting with competitive sets: a recursive algorithm. *International Journal of Contemporary Hospitality Management*, 267-285.

RIASSUNTO

Obiettivo perseguito dalla tesi e descrizione del lavoro svolto

Con il presente elaborato si cerca di comprendere come, attraverso la costruzione di un modello predittivo di *machine learning*, sia possibile prevedere le cancellazioni delle prenotazioni alberghiere. Ci si prefigge anche il compito di capire se un'ideale comunicazione con il cliente, attivata dalla struttura alberghiera, può diminuire la propensione del cliente alla cancellazione.

Nel corso della trattazione si cercherà di individuare anche i possibili *benefit*, diversi in base alle generazioni e alle loro caratteristiche, che possono essere utilizzati per convincere il soggetto a non disdire la propria prenotazione.

Nonostante l'importanza del settore, basti pensare che secondo un'indagine Federalberghi condotta con riguardo ai dati dell'anno 2021, la spesa media complessiva che gli italiani hanno dedicato ai viaggi è stata di 876 € a persona³, non molti studi sono stati compiuti circa l'anticipazione della cancellazione della domanda. Dopo aver delineato la teoria inerente alla domanda e all'anticipazione della stessa ed aver fornito accenni sul mercato in oggetto, andrò a costruire un modello, attraverso il linguaggio di programmazione R, che cercherà di anticipare la cancellazione.

Effettuerò ciò utilizzando un *dataset* contenente i dati relativi alle prenotazioni di due hotel portoghesi. Si tratta di un *dataset* di 32 variabili contenente circa 119.000 osservazioni.

Dapprima, si costruirà un modello basato sull'algoritmo del *decision tree*, e successivamente per ottenere una maggiore accuratezza si procederà con la costruzione, e l'applicazione, di un modello basato sull'algoritmo di *random forest*. Quest'ultimo grazie alla costruzione di una numerosità di alberi decisionali, nel mio caso 500, promette infatti risultati migliori.

L'applicazione di modelli predittivi è molto importante per il *management* della struttura e per assicurare un corretto e positivo *revenue management cycle*.

Il *Revenue Management (RM)* è tradizionalmente definito come l'arte e la scienza di prevedere la domanda e contemporaneamente di regolare il prezzo e la disponibilità dei prodotti per soddisfare quella particolare domanda (Mehmet Erdem, 2016). La *demand forecast* rientra tra i 5 differenti problemi del *revenue management cycle* assieme: alla determinazione dei prezzi, alle aste, al controllo della capacità e alla tecnica dell'*overbooking* (Wen-Chyuan Chiang, 2006).

Effettuare la *demand forecast* ed essere a conoscenza anticipatamente della possibile domanda è, quindi, essenziale al fine di definire anche gli altri punti del *revenue management cycle* come ad esempio l'ottimizzazione del prezzo.

Tuttavia, è al contempo essenziale sviluppare modelli predittivi della cancellazione che permettano di identificare quali e quante prenotazioni probabilmente saranno cancellate.

³ <https://www.federalberghi.it/comunicati/lestate-degli-italiani.aspx>; ultima consultazione: 01/06/2022

Successivamente a questa classificazione, lo *staff* della struttura potrebbe decidere di contattare i soggetti che, sulla base del modello, sono maggiormente identificati come “probabili cancellazioni” in modo da poter diminuire la propensione alla disdetta di questi offrendo sconti e/o omaggi.

Ma, in caso di contatto e in assenza di un alto *budget*, lo *staff* potrebbe anche solo chiedere conferma della prenotazione o di alcuni dati inerenti alla stessa.

Come è stato, infatti, dimostrato in un *paper* di ricerca anche il semplice contatto produce effetti positivi sulla riduzione della cancellazione della domanda. In particolare, non contattare il cliente produce un aumento delle cancellazioni di quasi 10 volte rispetto al caso in cui lo stesso, invece, viene contattato (Nuno Antonio A. d., 2019).

Descrizione e caratteristiche del mercato oggetto di studio

Il mercato alberghiero, ma in generale dell'*hospitality*, ha subito negli ultimi anni cambiamenti radicali che richiedono l'introduzione e l'applicazione di nuovi metodi comunicativi e di gestione da parte del *management*.

Come in un qualsiasi altro mercato, anche il mercato oggetto di studio, ha infatti subito al suo interno numerosi cambiamenti dettati dallo sviluppo di nuovi strumenti digitali e di nuove tecnologie.

L'era del digitale ha, infatti, comportato l'entrata nel mercato di nuovi competitor, come ad esempio le agenzie di viaggio online (OTA) che hanno rivoluzionato le modalità di prenotazione.

Se però, per quanto riguarda le prenotazioni *online*, fino a qualche anno fa, gli OTA erano lo strumento principale con cui effettuare le prenotazioni, attualmente si sta, invece, assistendo ad una crescita esponenziale delle prenotazioni sul sito di proprietà della struttura. Questa tendenza è sicuramente imputabile ai costi, ormai contenuti, della costruzione e gestione di un sito internet.

Ulteriore motivazione a fronte di questo aumento delle prenotazioni sui siti di proprietà è dovuta anche all'abolizione della pratica del *Parity Rate* a seguito dell'approvazione nel 2017 del DDL Concorrenza.

Questa pratica prevedeva l'obbligo per le strutture che avevano sottoscritto un contratto con gli OTA di non offrire sul proprio sito, o in qualsiasi altra modalità, tariffe inferiori rispetto a quelle praticate sui primi.

Vista questa tendenza, per cercare di perfezionare la conversione del cliente è però necessario sviluppare il sito di proprietà seguendo alcuni accorgimenti.

Il sito dovrà, infatti, essere accessibile tramite *smartphone* poiché è uno degli strumenti maggiormente utilizzati dai consumatori sia in fase esplorativa che di prenotazione.

In una ricerca effettuata prendendo in considerazione più di un migliaio di viaggiatori provenienti da tutto il mondo è, infatti, emerso che il 77% dei viaggiatori *business* e il 53% dei viaggiatori *leisure* ha effettuato la propria prenotazione attraverso lo *smartphone* (Digital, 2019).

Possedere un sito non accessibile da *mobile* o che richieda un lungo periodo di caricamento pregiudicherà quindi l'esperienza del consumatore che potrebbe anche decidere di non voler più prenotare presso la struttura. In tutto ciò, è di notevole importanza anche sviluppare un *booking engine*.

Quest'ultimo è lo strumento attraverso il quale il cliente autonomamente, sul sito della struttura, può verificare se c'è disponibilità nel periodo di suo interesse e il costo del soggiorno. Infine, attraverso l'utilizzo dello stesso, può anche perfezionare l'acquisto.

Il *booking engine* altro non è, quindi, che una forma di *call to action* che permette al soggetto di acquistare il servizio. Essendo quindi una forma di *call to action*, la sua presenza è necessaria poiché senza di essa non si avrebbe la conversione del cliente.

Ma il predetto strumento permette anche una migliore gestione delle risorse appartenenti all'organizzazione. Verificando autonomamente la disponibilità e la relativa tariffa si evitano, infatti, scambi di *mail* o chiamate tra il cliente e la struttura; il che comporta un risparmio di tempo per lo *staff*.

Tuttavia, per far emergere la propria struttura si possono sviluppare, anche, campagne di Google Ads che attraverso la selezione di alcune *keywords* ottimamente scelte, permettano di risultare tra i primi risultati nel motore di ricerca.

Questa digitalizzazione del mercato alberghiero è andata di pari passo ad una digitalizzazione, sempre maggiore, anche dei consumatori. Si può, infatti, assistere attualmente ad una *customer journey* notevolmente differente rispetto a quella che poteva esserci qualche anno fa.

Grazie alla crescente presenza *online* delle strutture così come dei clienti, si sono sviluppati, ed evoluti, anche i micro-momenti. Questi sono momenti in cui il soggetto si rivolge al web, anche attraverso una varietà differente di dispositivi, affinché questo gli presenti una soluzione ad un bisogno o un desiderio che il soggetto ha in quel momento.

La rete non viene usata, turisticamente parlando, unicamente per effettuare la prenotazione ma anche per lasciarsi ispirare, per informarsi e, *last but not least*, per condividere con gli altri la propria esperienza.

Visto il crescente utilizzo degli strumenti *online* e dei momenti in cui tutti noi ci rivolgiamo ad esso, esistono una pluralità di micro-momenti ma, volendo effettuare una categorizzazione, questi possono essere riassunti principalmente in 4 micro-momenti: *I want to get away moments*, *Time to take a plan moments*, *Let's book it moments* e *Can't wait to explore moments* (Google, 2016).

Si parte, quindi, dalla fase in cui il soggetto utilizza il web per prendere ispirazione circa il prossimo viaggio, fino ad arrivare al momento in cui condivide la propria avventura. Le fasi intermedie sono quelle in cui il soggetto pianifica e prenota.

È importante che in questo percorso la struttura sia presente, attraverso svariati strumenti, in modo da comunicare con il cliente e di assicurare poi una conversione dello stesso.

In particolare, occorre che la struttura sia utile al soggetto, poiché, come afferma la definizione stessa di micro-momento, il soggetto in questi casi fa ricorso al web per trovare una soluzione ad un problema ed è quindi di fondamentale importanza riuscire ad aiutare lo stesso in questa circostanza.

Se, da un lato però, questa digitalizzazione ha portato sicuramente dei vantaggi ai clienti così come anche agli stessi operatori del mercato, dall'altro lato, come in un qualsiasi altro aspetto, ha portato anche degli svantaggi.

Uno dei principali svantaggi riguarda la maggior facilità, anche grazie agli strumenti disponibili in rete, di effettuare una cancellazione.

Una delle possibili cause di questa crescita delle cancellazioni riguarda anche la maggior facilità con cui si possono verificare le recensioni *online*.

Se, infatti, non si avevano fino a qualche anno fa strumenti in grado di informare il cliente sulla qualità della struttura prenotata, fatta eccezione per il *word of mouth (WOM)* di parenti o amici, attualmente sono disponibili una numerosità di siti che consentono di effettuare questo controllo.

Il soggetto potrebbe, infatti, aver eseguito la prenotazione “vittima” del principio di scarsità. Il principio di scarsità è molto utilizzato su tutte le piattaforme OTA e prevede l’inserimento, lungo il processo di scelta e di prenotazione, di frasi del tipo “Rimangono solo 2 stanze” che portano il soggetto ad affrettare la prenotazione per paura di perdere la struttura prescelta.

Questa tecnica di marketing viene utilizzata per far aumentare il valore percepito dal cliente rispetto a quella struttura oltre che ad evitare il posticipo della fase di prenotazione. Il soggetto, viste le tante alternative presentategli, potrebbe infatti voler rimandare la prenotazione per prendere una decisione maggiormente ponderata, ma alla luce del principio di scarsità, questo avviene meno frequentemente in quanto il soggetto ha paura di perdere l’occasione e conclude comunque il processo di prenotazione.

Celerità alla prenotazione facilitata anche dalla presenza delle tariffe rimborsabili o della formula “paga in struttura” che consentono, rispettivamente, il rimborso completo della caparra o la possibilità di non pagare alcun acconto.

In tutti questi casi, il soggetto è maggiormente spinto ad effettuare la prenotazione.

Tuttavia, in un secondo momento, il soggetto potrebbe recarsi su un sito di recensioni, o sul sito stesso della struttura prenotata, e rendersi conto che l’albergo non è in linea con quelle che sono le sue aspettative o i suoi bisogni e di conseguenza, effettua la cancellazione.

Se, infatti, la presenza di tariffe rimborsabili rappresentano una garanzia per il cliente che prenota, la medesima cosa non la si può dire per quanto riguarda la struttura.

La struttura, infatti, potrebbe ricevere da un giorno all’altro la richiesta di cancellazione da parte del cliente senza poter, naturalmente, fare nulla. Come è logico pensare, questa pratica però comporta una perdita di ricavi per la struttura poiché la stessa, nel caso in cui non dovesse trovare un altro affittuario per la stanza in oggetto, registrerebbe una perdita.

Per questo motivo è stata introdotta all’interno delle pratiche manageriali alberghiere, ma non solo, la pratica dell’*overbooking*.

Detta tecnica prevede di accettare più prenotazioni di quante siano le stanze disponibili in un dato periodo, consci del fatto che qualche prenotazione verrà disdetta.

Tuttavia, la pratica dell’*overbooking* è un’arma a doppio taglio in quanto, se da un lato, potrebbe garantire maggiori ricavi andando a limitare i danni di una cancellazione e assicurando una prenotazione ad ogni stanza, dall’altro, se non gestita correttamente, potrebbe essere causa di un malcontento nei propri clienti.

Se si dovesse, infatti, praticare un tasso di *overbooking* troppo elevato allora la struttura si troverebbe prenotate più stanze di quante non ne ha a disposizione, dovendo, quindi, attuare operazioni di disdetta.

Sono diverse le azioni esercitabili a seguito di un *overbooking* ma tutte, chi più chi meno, portano ad un malcontento nel cliente. Si può, infatti, scegliere di pagare al cliente una struttura differente oppure chiedere di spostare il proprio soggiorno.

Ma qualsiasi azione applicabile porta, comunque, alla nascita di un sentimento non positivo nel cliente. Il sentimento negativo insinuatosi nel cliente porta infatti a un passaparola negativo da parte dello stesso che, soprattutto nella nuova era tecnologica, può portare ad una WOM, denominata e-WOM, negativa.

È quindi di notevole importanza riuscire ad effettuare una previsione della domanda che cerchi di anticipare la stessa e al contempo cerchi di anticipare anche l'eventuale cancellazione.

Riuscire, infatti, a prevedere la cancellazione delle prenotazioni permette al *management* della struttura alberghiera di gestire e modificare, all'evenienza, il tasso di *overbooking*.

Sapendo, infatti, che in un certo periodo e sotto alcune precise condizioni si ha un tasso di cancellazione maggiormente alto allora il *management*, per assicurare una massimizzazione dei posti occupati e dei ricavi, dovrà fissare per quel periodo di riferimento un tasso di *overbooking* più alto rispetto al solito.

Costruzione del modello predittivo

Il *dataset* che è stato oggetto di analisi si riferisce ad uno studio effettuato nel 2019 da Nuno Antonio et al. e pubblicato sul *Data Science Journal* con il titolo “*An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations*” e disponibile per il *download* sul sito Kaggle⁴.

Il *dataset* contiene le informazioni riguardanti 119.390 prenotazioni nell'arco temporale luglio 2015 – agosto 2017 riferiti a due hotel 4 stelle, un *city hotel* ed un *resort hotel*, appartenenti ad una catena alberghiera portoghese. I dati sono stati originariamente estratti dal *Property Management System* (PMS) della catena alberghiera, cioè dal programma utilizzato dalle strutture alberghiere per gestire le prenotazioni.

Per condurre le analisi ho utilizzato il programma R Studio ed il relativo codice scritto è in appendice.

Per effettuare la previsione, e quindi la classificazione delle osservazioni rispetto a due classi “probabile cancellazione” e “non probabile cancellazione” ho dapprima costruito un modello basato sull'algoritmo del *decision tree*.

Questo è stato costruito prendendo in considerazione prima solo alcune variabili, ma i risultati in questo caso non si sono rilevanti soddisfacenti, e poi tutte le variabili presenti nel dataset (ad eccezione: del *reservation status date*, della *company* e dell'*agent*). Tuttavia, anche in quest'ultimo caso l'accuratezza era pari al 79% e la *sensitivity* e la *specificity* risultavano essere, rispettivamente, dell'89% e del 60%. Per questo motivo, sapendo che la *random forest*, vista la numerosità degli alberi decisionali presi in considerazione, promette un'accuratezza migliore ho proceduto a costruire un modello predittivo basato su questo algoritmo e avente

⁴ <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>

ad oggetto tutte le variabili prese in considerazione nel *decision tree*. Ho, però, dovuto escludere anche la variabile *country* poiché presenta più di 53 categorie, cioè il numero massimo di categorie gestibili da R per la creazione di un modello.

Ho scelto di includere tutte le variabili nel modello in modo poi da operare la *feature selection*; quest'analisi consente di calcolare e verificare la *mean decrease accuracy*. Questa metrica esprime, cioè, di quanto l'accuratezza diminuisce escludendo dal modello le varie variabili oggetto di studio.

Ad esempio, con l'eliminazione della variabile *babies* dal modello si ha un aumento nell'accuratezza dello stesso.

Mentre altre variabili come, ad esempio, *is repeated guest* e *distribution channel* pur avendo un'importanza maggiore rispetto alla variabile esaminata precedentemente, non comportano comunque una perdita elevata di accuratezza una volta eliminate dal modello.

Poiché un modello dovrebbe essere rappresentativo della realtà ma senza renderla eccessivamente complicata è sempre positivo riuscire ad eliminare dal modello l'utilizzo di una o più variabili.

Ho dunque eliminato le variabili: *babies*, *is repeated guest*, *distribution channel*, *previous bookings not canceled*, *days in waiting list*, *reserved room type*, *adults*, *assigned room type*, *hotel*, *arrival date year*.

Inoltre, quando si costruisce un modello si può anche analizzare la matrice di correlazione.

Due variabili molto correlate tra loro, sia positivamente che negativamente portano, infatti, a risultati nel modello non troppo soddisfacenti o comunque a risultati simili a quelli che si otterrebbero considerando solo una delle due variabili correlate.

Tuttavia, le correlazioni esistenti tra le variabili prese in oggetto non sono molto forti, essendo la più alta, tra *lead time* e *stays in week nights*, di solo 0.30. Ho però provato comunque ad effettuare nuovamente la costruzione del modello, eliminando la variabile *stays in week nights*. Ho scelto di eliminare questa variabile e non *lead time* poiché quest'ultima ha, secondo l'analisi di *feature selection*, maggiore importanza.

Di seguito i risultati ottenuti dalle varie fasi di costruzione del modello.

Prima feature selection

		Reference	
Prediction		0	1
0	20164	3224	
1	2385	10042	

Accuracy : 0.8434
Sensitivity : 0.8942
Specificity : 0.7570

26 variabili

Dopo feature selection

		Reference	
Prediction		0	1
0	20083	3514	
1	2466	9752	

Accuracy : 0.833
Sensitivity : 0.8906
Specificity : 0.7351

15 variabili

Dopo analisi correlazione

		Reference	
Prediction		0	1
0	20020	3551	
1	2529	9715	

Accuracy : 0.8302
Sensitivity : 0.8878
Specificity : 0.7323

14 variabili

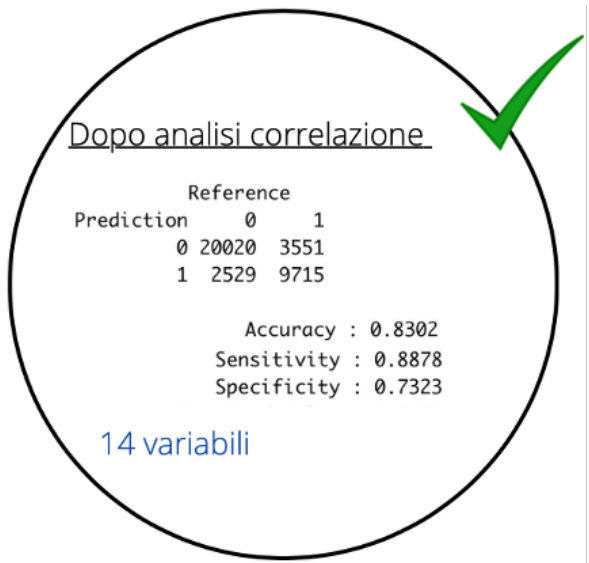


Figura 40: Confronto accuratezza tra modelli; fonte: realizzazione propria

Come è possibile verificare nella fig. 40, la scelta del modello da utilizzare è ricaduta sull'ultimo modello creato, cioè quello che segue sia alla *feature selection* che all'analisi della correlazione. Di conseguenza, anche l'eliminazione dal modello della variabile *stays in week nights* si è rivelata corretta.

Analisi e risultati

Ho condotto dapprima delle analisi prettamente descrittive anche attraverso l'utilizzo dei grafici *ggplot*.

Da queste analisi, e attraverso lo studio dei grafici, ho verificato che il 41% delle prenotazioni riguardanti il *city hotel* sono state cancellate mentre per quanto riguarda il *resort hotel* la percentuale scende fino a raggiungere il 27%.

Comparando la variabile *lead time*, che rappresenta i giorni di anticipo con cui è avvenuta la prenotazione, con le cancellazioni che si sono verificate è emerso che le prenotazioni effettuate con maggior anticipo sono state anche quelle maggiormente cancellate.

Infatti, il *city hotel* presenta una percentuale di cancellazioni di 14 punti percentuali maggiore rispetto al *resort hotel* e la differenza per quanto riguarda la media dei giorni di anticipo è di 18 giorni. Questa relazione è visibile anche attraverso il grafico *box plot* (fig.41) presentato di seguito.

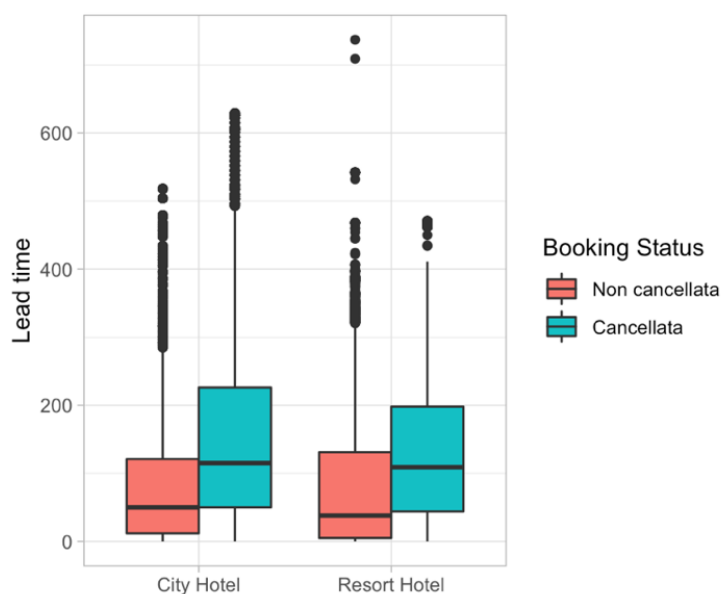


Figura 41: Come variano le cancellazioni al variare del tipo di hotel e del lead time; Fonte: realizzazione propria

Sulla base dei grafici *ggplot* ho analizzato anche l'andamento della domanda e delle cancellazioni.

Si è verificato che si ha un picco delle prenotazioni nel periodo estivo che è, infatti, il periodo di alta stagione. Per quanto riguarda le cancellazioni, queste seguono l'andamento della domanda, e di conseguenza le maggiori cancellazioni nel periodo estivo sono esclusivamente da ricondurre al picco registrato nella domanda. Non è stato possibile individuare periodi in cui le cancellazioni hanno subito un aumento non dipendente dall'andamento della domanda.

Inoltre, la cancellazione dipenderà anche dal tipo di canale distributivo considerato, come è possibile verificare in fig.42.

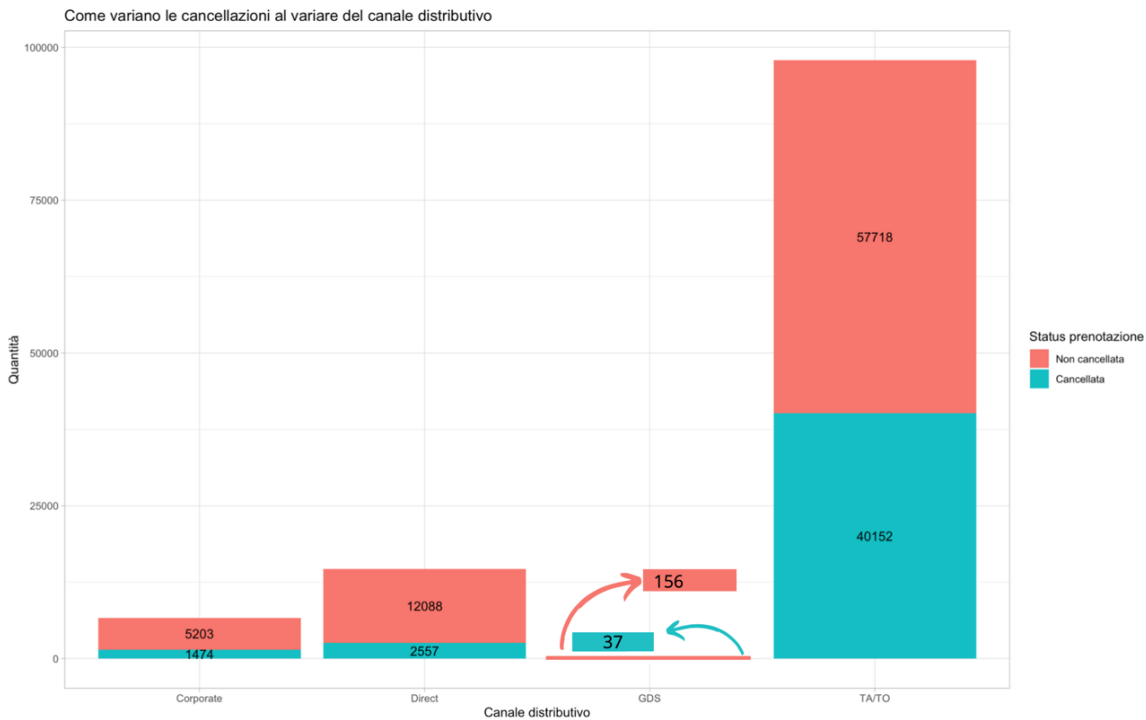


Figura 415: Come variano le prenotazioni al variare del canale distributivo considerato; fonte: realizzazione propria

Infatti, se si considera il canale *corporate* si ha una percentuale di cancellazione pari al 22% che sale al 41% se si considera il canale TA/TO, cioè quello delle agenzie di viaggio e dei tour operator (anche *online*).

La percentuale minore la si ha con il canale delle prenotazioni dirette per cui si ha una percentuale di cancellazione pari al 17%. Si registra, invece, una percentuale di cancellazioni pari al 19% quando si considera il canale dei GDS cioè dei *Global Distribution System*, che curano le prenotazioni B2B.

Si potrebbe pensare che questa differenza sia dovuta principalmente alle modalità di rimborso poiché, solitamente, quando si prenota attraverso un TA *online* si possono scegliere tra differenti tipi di tariffa mentre generalmente quando si effettua una prenotazione in modo diretto si ha l'obbligo di versare una caparra al fine di garantire la prenotazione. Tuttavia, da un'ulteriore analisi è emerso che per tutti i canali almeno il 75% delle prenotazioni non risulta garantita da alcun deposito iniziale.

Di conseguenza, la maggior probabilità di cancellazione legata al canale dei TA e dei TO non dipende dal differente tipo di rimborso applicabile bensì è da far risalire alla stessa tipologia di prenotazione.

Questo potrebbe in parte essere spiegato dal fatto che una prenotazione effettuata *online* attraverso siti OTA è più in linea con una prenotazione immediata e veloce.

Ulteriore fattore da cui può dipendere la propensione alla cancellazione, oltre al canale distributivo appena analizzato, è rappresentato dalla nazionalità. Per il dataset in oggetto si ha però un problema di analisi per quanto riguarda questa variabile; questo perché, gli stati sono molteplici e di conseguenza un grafico che permetta di rappresentarli tutti non è di facile lettura. Motivo per cui ho scelto di impostare dei filtri che permettano di analizzare esclusivamente gli stati che verificano una certa condizione.

In particolare, ho analizzato, attraverso due differenti grafici, le nazioni che presentavano più di 1500 prenotazioni e le nazioni che avevano un numero di prenotazioni compreso tra 1000 e 1500.

Dall'analisi effettuata è risultato che il Portogallo, tra gli Stati presi in considerazione attraverso i filtri, risulta essere lo Stato con più cancellazioni pari al 57% del totale delle prenotazioni. La minor percentuale di cancellazione la si registra per: la Germania (17%), la Francia e l'Austria (con il 18%).

È chiaro, quindi, che quando la prenotazione perverrà da cittadini portoghesi, questa più facilmente potrebbe essere oggetto di cancellazione. Perciò quando si ha a che fare con un numero elevato di prenotazioni provenienti dal Portogallo, per un dato periodo, occorrerà in quel frangente praticare un tasso di *overbooking* maggiore.

Inoltre, la percentuale di cancellazioni che si registra varia anche a seconda della tipologia di prenotazione che è stata effettuata e questo è possibile verificarlo attraverso la fig.43.

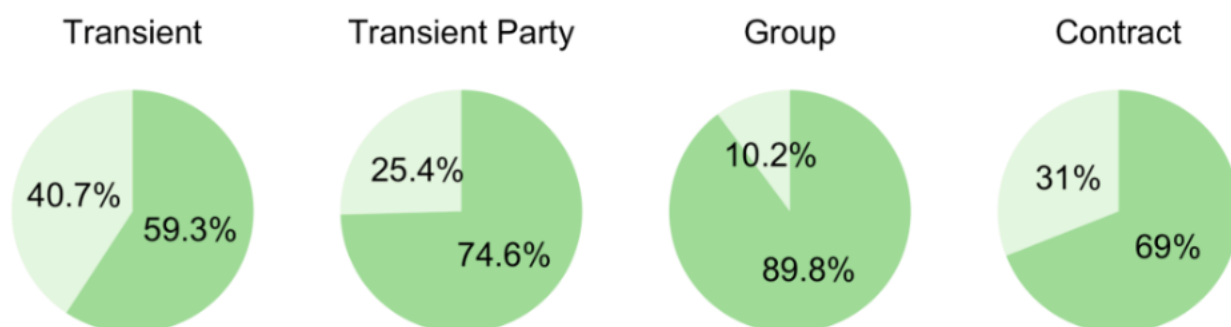


Figura 416: Tipologia di prenotazione con associata la percentuale delle prenotazioni cancellate (in chiaro); fonte: realizzazione propria

Infatti, per le prenotazioni di gruppo la percentuale di cancellazione è soltanto del 10% mentre per le prenotazioni *transient*, cioè tutte quelle prenotazioni non in gruppo e che non sono parte di un contratto di prenotazione, si registra una percentuale pari al 41%. La percentuale per le prenotazioni *contract*, cioè le prenotazioni effettuate da agenzie attraverso contratti di *allotment*, è invece del 31%. L'*allotment* è il contratto stipulato tra la struttura e un *tour operator* che prevede il blocco, quindi la prenotazione, da parte di quest'ultimo di un gruppo di camere che poi sarà lui stesso a commercializzare.

In questo caso la cancellazione è pari al 31% e ciò potrebbe essere dovuto al fatto che non sempre i *tour operator* riescono a vendere in *toto* tutto il lotto di stanze da loro acquistate e di conseguenza effettuano la cancellazione.

La seconda percentuale più bassa, pari al 25%, è associata alla tipologia di prenotazione *transient party*.

Le prenotazioni classificate come *transient*, come già detto, appartengono a quella classe di clienti che non prenotano né in gruppo né sono parte di un contratto di prenotazione. In particolare, però, si indicano come *transient party* quelle prenotazioni *transient* che sono associate ad almeno un'altra prenotazione sempre *transient*.

Analisi ulteriori e applicazione del modello

Per verificare la relazione esistente tra le variabili ho condotto un'analisi di correlazione. Da questa sono emerse correlazioni non molto forti, sia positivamente che negativamente parlando, tra le variabili oggetto di studio. Una delle correlazioni maggiormente alte, 0.30, è quella tra la variabile “*lead time*” e la variabile “numero notti soggiorno”. Seconda correlazione, in ordine di forza, è quella tra la variabile *adr* e la variabile adulti; dove *adr* indica l'*average daily rate* cioè la tariffa giornaliera media pagata per una stanza.

L'analisi di correlazione non è utile esclusivamente per capire la relazione esistente tra le variabili ma anche, come già chiarito, per la costruzione del modello.

Per quanto riguarda l'analisi di previsione, e come già precedentemente detto, ho dapprima costruito ed utilizzato un modello basato sull'algoritmo del *decision tree* per poi scegliere di procedere con un modello basato sull'algoritmo del *random forest* per via della maggior accuratezza riscontrabile in questo.

		Reference	
Prediction		0	1
0	20020	3551	
1	2529	9715	

Figura 417: Confusion matrix del modello predittivo scelto

Come si può vedere dalla *confusion matrix* rappresentata in fig. 44 i casi in cui correttamente viene predetto che la prenotazione non verrà cancellata sono 20.020; mentre i casi in cui, correttamente, vengono predette le cancellazioni sono 9.715.

Per valutare però l'accuratezza del modello occorre analizzare sia i *true positive* che i *false negative*, cioè i casi in cui la predizione non si è rivelata corretta.

Nello specifico 2.529 rappresentano i falsi negativi; cioè i casi in cui è stata predetta la classe negativa, cioè la cancellazione, ma in realtà si trattava di prenotazioni che non sono state cancellate.

Viceversa, 3.551 rappresenta i casi falsi positivi cioè quelli in cui si è predetta la classe positiva, quindi la non cancellazione, ma in realtà si trattava di osservazioni appartenenti alla classe negativa.

L'accuratezza è dell'83% mentre la *specificity* del 73% e la *sensitivity* dell'89% e il modello si basa su 14 variabili.

Effetto della comunicazione sulla percentuale di cancellazione

Nello studio di Nuno Antonio et al., da dove è stato preso il *dataset* oggetto di studio, si è analizzato anche l'effetto che una comunicazione ha sulla percentuale di cancellazione di una prenotazione.

Il modello utilizzato dallo studio per effettuare la previsione presenta un'accuratezza abbastanza elevata, pari all'84%.

In prima istanza il *management*, anche sotto consiglio degli autori, aveva deciso di contattare telefonicamente i clienti classificati come “probabile cancellazione” offrendo sconti e servizi aggiuntivi.

Tuttavia, ad avviso della struttura, questa operazione era troppo onerosa sia in termini di tempo che di spesa.

Si è passati, quindi, a un contatto effettuato via mail e senza l'offerta di sconti o pacchetti di *upgrade* ma esclusivamente chiedendo informazioni riguardanti il soggiorno.

In particolare, si richiedevano informazioni come l'ora del *check-in* o il tipo di letto che il cliente prediligeva. Alla comunicazione sono conseguite due differenti reazioni: o i soggetti sentendosi curati da parte della struttura non disdicevano più la prenotazione oppure disdicevano comunque la prenotazione. Tuttavia, anche quest'ultimo caso può essere classificato come positivo poiché, avendo la certezza della cancellazione, la sistemazione può essere riassegnata ad un altro cliente.

Hotel	Action	MF \geq 0% (all bookings)			MF \geq 50%		
		Canceled	Not canceled	% Canceled	Canceled	Not canceled	% Canceled
H1	No	471	1,429	24.8%	125	153	45.0%
	Yes	12	97	11.0%	6	70	7.9%
H2	No	1,010	2,905	25.8%	269	325	45.3%
	Yes	15	181	7.7%	9	111	7.5%

Figura 418: Effetto dell'azione comunicativa sulle cancellazioni; fonte: Nuno Antonio et al. "An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations" (2019)

Nella fig. 45, sono riassunti gli effetti ottenuti dal contatto con il cliente sulle cancellazioni.

Come è possibile verificare, nei casi in cui si è avuto il contatto del cliente, identificati da "yes", il tasso di cancellazione è diminuito in tutti i casi.

La metrica MF identifica la frequenza minima. Questa metrica è stata calcolata dividendo i giorni in cui la prenotazione è stata classificata come probabile cancellazione con i giorni totali di esecuzione del modello.

Ciò significa che, maggiore è il valore della metrica più volte la prenotazione è stata classificata come probabile cancellazione.

Come delineato dagli stessi autori, per quanto riguarda le prenotazioni con un $MF \geq 0$, la percentuale di cancellazione a seguito dell'azione intrapresa dalla struttura è diminuita, rispettivamente, di circa 13 punti percentuali e di 18 punti percentuali per quanto riguarda il resort hotel ed il *city* hotel. Inoltre, questa differenza è statisticamente significativa.

Per le prenotazioni per le quali è stata prevista la probabilità di cancellazione in almeno la metà dei giorni in cui il modello è stato applicato alle stesse, le differenze sono ancora più marcate. Si ha, infatti, una differenza, significativa, di 37 punti percentuali per entrambi gli hotel.

Ulteriore elemento da indagare riguarda l'*Odds Ratio*. Quest'indice è il rapporto tra la frequenza con la quale un evento si verifica in un gruppo e la frequenza con la quale lo stesso evento si verifica in un gruppo di controllo. È quindi, una misura di efficacia dell'azione intrapresa.

L'*Odds Ratio* risulta essere pari a 9.3 per H1, cioè per il resort hotel, e pari a 10 per H2 e quindi per il *city* hotel; questo prendendo in considerazione le prenotazioni con un $MF \geq 50\%$.

Questo indica che, non contattare un cliente comporta un fattore di aumento della cancellazione pari a 9.3 per H1 e 10 per H2 (Nuno Antonio A. d., 2019) .

Sempre secondo quanto studiato dagli autori del *paper* appena esaminato, il ricavo delle camere che non è stato perso a causa delle cancellazioni è di 16k euro per H1 e di 22k euro per H2.

Si è dimostrato, quindi, che una comunicazione con il cliente può portare lo stesso a non disdire la propria prenotazione con conseguenze abbastanza evidenti anche in ambito economico.

Nella comunicazione, come detto, si possono o semplicemente chiedere informazioni/conferme oppure offrire degli sconti e/o omaggi.

Tuttavia, occorre distinguere la clientela *business* da quella *leisure* poiché, essendo soggetti che viaggiano per motivi differenti, anche i *benefit* su cui far leva dovranno essere differenti.

I viaggiatori *business* non sono molto sensibili al prezzo e di conseguenza offrire uno sconto potrebbe non funzionare. Viceversa, questo tipo di viaggiatori, ricercherà la velocità nei servizi e per questo occorrerà offrire agli stessi servizi che li aiutino nel soggiorno. Un *benefit* che potrebbe essere adatto a questo tipo di clientela, e potrebbe diminuire la percentuale di cancellazione rispetto a questo segmento, è ad esempio offrire il servizio di *transfert* veloce da e per la struttura.

Il discorso è differente, invece, per quanto riguarda i clienti *leisure*. Questa tipologia di clientela è, infatti, maggiormente sensibile al prezzo e di conseguenza come incentivo per evitare la cancellazione si potrebbero proporre sconti o omaggi.

Tuttavia, mentre la clientela *business* ha al suo interno esigenze e caratteristiche abbastanza omogenee, la stessa cosa non può dirsi per la clientela *leisure*. La categorizzazione dei viaggiatori per piacere, infatti, può contenere al suo interno soggetti che presentano differenti bisogni poiché, ad esempio, alcuni potrebbero viaggiare in famiglia altri con amici.

È soprattutto per questa clientela che può essere, quindi, utile seguire la categorizzazione per fasce d'età.

Questa categorizzazione non è, infatti, solo utile per quanto riguarda la comunicazione e la pubblicizzazione nella fase iniziale della *customer journey*, ma può essere d'aiuto anche nella fase finale della stessa. Difatti, a seconda dell'età si possono individuare differenti *touchpoint* e/o *benefit* che possono meglio funzionare di altri al fine di diminuire la propensione alla cancellazione.

Approfondendo e facendo un discorso generazionale, i *baby boomers*, cioè i soggetti nati tra il 1946 e il 1964, sono i soggetti che maggiormente ricercano informazioni sul viaggio e sulla struttura. Da un punto di vista comunicativo questi prediligono i siti che contengono molte informazioni e spiegazioni (Emilie Slootweg, 2018); ed è proprio su questa ricerca di informazioni che si potrebbe far leva.

Visto che, come detto precedentemente, anche i risultati basati sulla ricerca di Nuno et al. del 2019 poggiano quasi esclusivamente su comunicazioni non offerenti *benefit* monetari, per questa generazione potrebbe rivelarsi una buona strategia contattare il cliente anche solo semplicemente per fornire qualche informazione aggiuntiva sul soggiorno e/o sulla struttura e rendersi disponibili a rispondere a qualsiasi dubbio.

Per quanto riguarda i nati tra il 1965 e il 1980, cioè la generazione X, questi ricercano nella struttura stanze comode e una molteplicità di servizi a loro disposizione (Emilie Slootweg, 2018). Inoltre, sono anche i soggetti che potenzialmente presentano il maggior reddito.

Per tutti questi motivi, in questo caso, una buona strategia potrebbe essere quella di offrire, qualora disponibile, un *upgrade* della camera, gratuitamente o con un piccolo *extra*, oltre che uno sconto per un servizio offerto in struttura. Questo perché, oltre a costruire una migliore relazione con lo stesso, che potrebbe portare ad una non cancellazione della prenotazione, in questo modo si effettuano anche strategie di *up-selling* e di *cross selling*. Se, infatti, al soggetto viene offerta come *upgrade* una stanza di qualità superiore, questo potrebbe decidere la prossima volta che soggiognerà presso la struttura di prenotare nuovamente quella stanza; poiché, come detto, sono la generazione che presenta potenzialmente il maggior reddito e che pretende una stanza con tutte le comodità. Di conseguenza, qualora il soggetto avesse pagato un piccolo *extra* per la stanza di qualità superiore, non si manifesterebbe solo un *up-selling* immediato ma anche una crescita del valore potenziale della relazione. Per quanto riguarda il mezzo comunicativo da prediligere questo è sicuramente l'e-mail, poiché è uno dei mezzi maggiormente consigliati per raggiungere questa generazione. Viceversa, offrendo uno sconto per un servizio si potrebbe generare una strategia di *cross selling*, in grado di aumentare il potenziale guadagno dato da quella relazione e da quella prenotazione.

Per quanto riguarda la generazione Y, cioè i nati tra il 1981 e il 1995, questi sono i soggetti che maggiormente ricercano e danno importanza al nome del brand e alla personalizzazione della comunicazione e dei servizi (Emilie Slootweg, 2018). Occorre, quindi, in questo caso puntare tutto sulla personalizzazione dell'esperienza del cliente. Chiaramente questa dipenderà dai servizi e dalla tipologia di struttura. Tuttavia, esistono personalizzazioni applicabili ovunque e su qualsiasi categoria di hotel che presentano anche un basso, se non nullo, costo di realizzazione. Ad esempio, si potrebbe chiedere al soggetto quale tipologia di cuscino preferisce o a che piano e in che posizione preferisce la stanza. È logico che, qualora si abbiano a disposizione molti servizi offerti in struttura, la personalizzazione può essere maggiormente presente e forte.

La generazione Z racchiude, invece, i nati dopo il 1996. Questa è la generazione tecnologica per eccellenza e che predilige come strumenti comunicativi tutto ciò che è basato sui *social* e sull'*online*.

È la generazione che, a differenza di tutte le altre, è attenta ai valori comunicati dai brand tanto da non effettuare una compra se il valore comunicato non è coerente con la sua persona.

Per questo, un'ottima strategia potrebbe consistere nell'inviare a questa categoria di clienti un'informativa, attraverso i *social* o attraverso messaggistica istantanea, per comunicare quali sono i valori propri della struttura e quali sono gli accorgimenti adottati dalla stessa per applicare e diffondere quel valore.

Queste appena delineate sono solo alcune delle tante strategie applicabili per cercare di migliorare la relazione con il cliente e di conseguenza evitare una cancellazione della domanda.

In tutto ciò, il *management* deve sempre e comunque aggiornare i propri profili *social* poiché, come è stato possibile verificare nel corso della trattazione, questi vengono utilizzati ormai da tutti senza distinzione d'età.

Inoltre, questi non sono più solo un mezzo utilizzato per condividere foto o la propria esperienza ma rappresentano anche un modo per informarsi e tenersi aggiornati.