Digital Business

Stefano Di Lena

2025

Indice

1 Introduzione

Con **business intelligence** si indica, invece, il processo aziendale e l'insieme di tecnologie usate per analizzare i dati, ricavandone conoscenza utile al miglioramento del proprio business. Le principali attività svolte sono: data preparation, data exploration e data mining (data analysis).

Il **business model** è uno schema che descrive come un'azienda genera, trasferisce ed ottiene valore. Un modello di business può essere descritto tramite alcuni elementi base (pilastri):

- customers (clienti);
- bundle prodotti/servizi;
- infrastrutture;
- fattibilità economica (il modello di business funziona se i ricavi sono maggiori dei costi).

1.1 The 9 Building Blocks

Figura 1.1: The Business Model Canvas

1.1.1 Value Proposition

[É il più importante elemento per i modelli di business]. Rappresentata dalla combinazione di prodotti e servizi (bundle) che genera valore per una specifica parte di clienti. É il motivo per il quale viene scelta un'azienda piuttosto che un'altra. La value proposition risolve un problema o soddisfa una richiesta. Può essere quantitativa (prezzo, tempo) o qualitativa (design, esperienza del cliente); quindi può corrispondere a qualcosa di misurabile o meno.

Alcuni esempi:

- *Novelty* - crea nuovi bisogni (il primo telefono, il primo mp3, il primo smartwatch);
- *Performance* - migliorare i prodotti già esistenti o aggiungere nuove caratteristiche (chip nuovi per laptop che li rendono più veloci rispetto ai competitors);
- *Customization* - creazione in collaborazione con il cliente (il sito Nike permette al cliente di scegliere modello, colore ed altri dettagli riguardo alle scarpe da acquistare);
- *"getting the job done"* - fare ciò che viene richiesto nei tempi e costi previsti (agenzie di consulenza);
- *Design* - migliore aspetto in confronto ai competitors (BMW, Apple);
- *Brand/status* - alcuni prodotti comunicano l'apparenza ad un certo status (Rolex per il lusso, sportwear per gli sportivi);
- *Prezzo* (inteso come rapporto qualità/prezzo) - molte imprese puntano su prezzi bassi ma bisogna sempre valutare la qualità del servizio offerto;
- *Costs reduction* - sconti (Amazon non fa pagare i costi di spedizione per acquisti superiori a 20$);
- *Risk reduction* - vendere prodotti compresi di garanzia o altri tipi di assistenza;
- *Convenience/usability* - realizzare prodotti o servizi facili da usare.

1.1.2 Customer Segments

É il gruppo di clienti (o aziende) che si decide di raggiungere e servire. Tutti i clienti all'interno di un segmento si comportano in maniera simile, sono omogenei. [É molto più economico comunicare con un gruppo di clienti piuttosto che col singolo cliente (per questo motivo conviene individuare i segmenti)]. Due segmenti diversi invece sono eterogenei. Esistono diversi approcci:

- *Mass Market* (mercato di massa) - non ci sono differenze tra i segmenti di clienti, tutti hanno gli stessi bisogni (mercato delle caramelle);

- *Niche Market* (mercato di nicchia) - basato sulle realazioni fornitore/cliente (produzione di componenti specifici);
- *Segmented Market* - divide il mercato in segmenti secondo caratteristiche differenti specifiche (possiamo trovare gusti di Coca-Cola diversi a seconda della nazione in cui andiamo a comprare la bevanda). Possiamo avere due diversi tipi di mercato segmentato.
  - *Diversified*: offro prodotti diversi in base ai segmenti (Amazon offre i servizi di vendita ma possiede per i clienti interessati anche un servizio di cloud storage);
  - *Multi-sides platforms*: ha bisogno di più segmenti per funzionare (le carte di credito hanno bisogno sia di clienti che la posseggono ma anche delle attività commerciali che le accettano come metodo di pagamento).

1.1.3 Channels

Definiscono come un'impresa (firm) raggiunge il segmento di clienti e trasmette loro la value proposition. Ci sono diversi tipi di canali: comunicazione, distribuzione e vendita. Gli obiettivi di questi sono:

- introdurre prodotti e servizi ai clienti;
- aiutare i clienti a dare un senso alla value proposition;
- rendere possibile la vendita del prodotto o servizio;
- consegnare la value proposition ai clienti;
- dare assistenza post-acquisto.

Tutti i canali hanno 5 fasi: consapevolezza, valutazione, acquisto, spedizione ed after sale. I canali possono essere ancora distinti in:

- *diretti* (dipendente della Apple che vende un iphone);
- *indiretti* (dipendente MediaWorld, che vende più band di telefoni, che ti propone un iphone);
- *property* (negozio Samsung);
- *partner* (stand Samsung in un negozio Euronics).

Possono capitare anche delle combinazioni di questi tipi (dipendente Samsung in uno stand presente in negozio Expert che ti vende un prodotto).

1.1.4 Customer Relationships

Tipo di rapporto stabilito con ogni segmento di clienti. Le relazioni tra clienti possono essere *personal* (ci si confronta direttamente con un'altra persona) o *automatic* (si parla con un bot o si cercano informazioni su un forum).

Queste relazioni servono a raggiungere determinati obiettivi:

- acquisire nuovi clienti;
- trattenere i clienti;
- upselling (aumentare le vendite).

I possibili tipi di customer relationiships sono:

- *Assistenza personale* (call center);
- *Personale dedicato* (consulente bancario);
- *Self-service* (Frequently Asked Questions, FAQ, su un determinato prodotto o sevizio);
- *Servizi automatici* (suggerimenti di film e serie tv di Netflix);
- *Communities* (forum e blogs);
- *Co-creation* (recenzioni su un determinato prodotto o servizio).

1.1.5 Revenue Streams

I flussi di ricavi sono i soldi ottenuti da ogni segmento di clienti (considerando solo il profitto). Può essere:

- *One-time payment* - pago una sola volta (acquisto di un'auto);
- *Recurring payment* - pago a rate (leasing di un'auto);

e può essere generato in diversi modi.

- *Asset sale*: ricavi provenienti dalla vendita di un prodotto o servizio;
- *Usage fee*: ricavi che dipendono dall'utilizzo di un prodotto o servizio (in un hotel paghiamo in base al numero di notti);
- *Subscription fees*: pagamento di un'abonamento per l'uso (in palestra paghi una tassa fissa per andare, indipendentemente dal numero di giorni in cui vai o dal numero di attrezzi che utilizzi al suo interno);
- *Lending/renting/leasing*: pago una determinata quota in maniera ricorrente per possedere un prodotto o servizio;
- *Licensing*: pago per usare la proprietà intellettuale di un'altra azienda o persona [brevetti (patent), SIAE].

- *Brokerage fees*: esiste un intermediario che ottiene una percentuale sulla transazione (tiketone, ebay, agenzie immobiliari);
- *Advertising*: attività commerciali che pagano per pubblicizzare i propri prodotti o servizi ed i flussi sono generati se si acquista tramite banner o link.

Ogni flusso ha un meccanismo di prezzo differente (fisso o dinamico).

Fixed Menu Pricing Predefined prices are based on static variables		**Dynamic Pricing** Prices change based on market conditions	
List price	Fixed prices for individual products, services, or other Value Propositions	*Negotiation (bargaining)*	Price negotiated between two or more partners depending on negotiation power and/or negotiation skills
Product feature dependent	Price depends on the number or quality of Value Proposition features	*Yield management*	Price depends on inventory and time of purchase (normally used for perishable resources such as hotel rooms or airline seats)
Customer segment dependent	Price depends on the type and characteristic of a Customer Segment	*Real-time-market*	Price is established dynamically based on supply and demand
Volume dependent	Price as a function of the quantity purchased	*Auctions*	Price determined by outcome of competitive bidding

Figura 1.2: Pricing Mechanisms

1.1.6 Key Resources

Il primo elemento del lato dei costi sono le risorse chiave, ovvero tutte le risorse, le attività ed i mezzi necessari affinché il modello di business funzioni. Queste possono essere proprietarie dell'azienda, ma anche comprate o affittate dai partners chiave. Le key resources possono essere:

- materiali, qualcosa di quantitativo e che possiamo toccare (per produrre un auto abbiamo bisogno di stabilimenti produttivi e punti vendita);
- finanziarie, non hanno bisogno di stabilimenti produttivi o punti vendita ma la risorsa finanziaria è la risorsa chiave (banche);
- intellettuali, brevetti (la value proposition di Coca-Cola è il gusto, ma per poterla produrre oltre agli stabilimenti produttivi abbiamo bisogno del brevetto);
- umane (Leonello Cucinelli, imprenditore italiano, che assunse donne esperte nel lavoro a maglia per produrre prodotti in cachemire).

1.1.7 Key Activities

Sono tutte le attività necessarie per far funzionare il modello di business. Possono essere:

- *Production* (Fiat);

- *Problem solving* (agenzie di consulenza);
- *Platform/Network* (tutti gli e-business, Facebook).

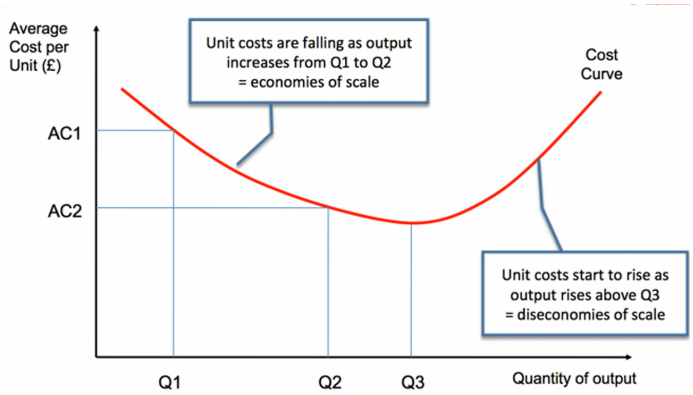
1.1.8 Key Partners

La rete di fornitori e partner necessari a far funzionare un modello di business. Possono essere:

- *not competitor alliances* (Mc Donald ha bisogno di bevande da servire insieme al cibo e si serve di Coca-Cola, per quest'ultima è conveniente questa alleanza perché verrà bevuta da milioni di persone);
- *competitor alliances* (Apple e Samsung sono competitor, perché vendono prodotti simili per gli stessi consumatori, ma Apple compra schermi Samsung per i propri Iphone.);
- *joint venture for new business development* (Starbucks che abituato a vendere le sue bevande nei propri store decide di entrare nella grande distribuzione, nei supermarket. Ma non avendo una linea di produzione, distribuzione e vendita si è affiancata a Pepsi).

Queste partnership iniziano per:

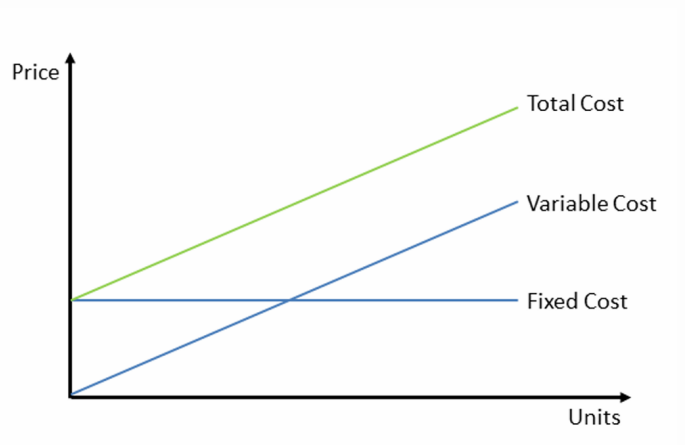
1. *ottimizzare il business con le economie di scala*, se aumenta il numero di ciò che produci e vendi, si riduce il costo marginale di produzione (a Pepsi conviene raddoppiare la produzione per permettere a Starbucks di distribuire il frappuccino, perché così facendo riducono il costo marginale di ogni singola bottiglia).



2. *ridurre il rischio e l'incertezza* (esistevano prima due standard di supporto video, i Blu-ray e gli HD-Disk, le compagnie dovevano scegliere formato supportare per la distribuzione dei loro contenuti; si unirono tutte in una partnership con la Blu-ray per ridurre la scelta sul mercato).
3. *acquisire risorse o attività* (sempre l'esempio di Starbucks e Pepsi).

1.1.9 Cost Structure

La struttura dei costi è la lista di tutti i costi associati al modello di business. [I costi principali del business model sono quelli associati alle key resources, activities e partnership].



I *costi fissi* sono quelli che non cambiano nel tempo e non dipendono dalla quantità della produzione, i *costi variabili* sono quelli associati alla quantità prodotta. La somma di questi due costi fornisce il *costo totale*.

2 Patterns

Per business model patterns intendiamo degli schemi o modelli che si ripetono e vengono adottati da molte compagnie.

2.1 Un-Bundling Business Model

Per Bundle intendiamo una combinazione di prodotti e servizi, quindi con Un-Bundle intendiamo una divisione di questi; ovvero una separazione del business model.

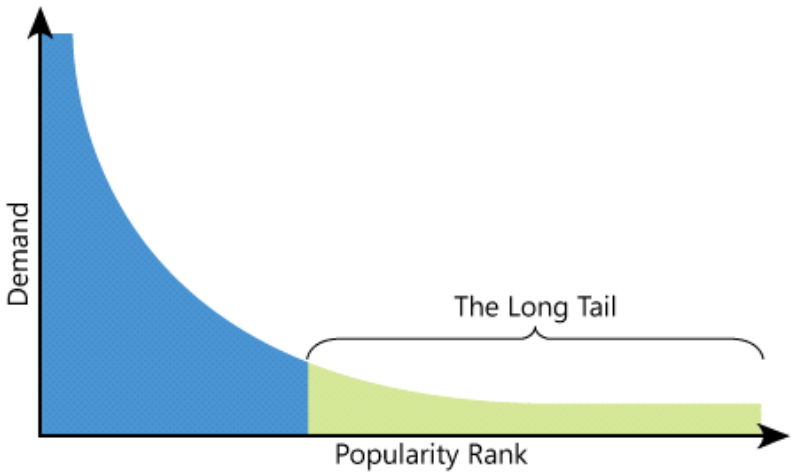
In generale ogni impresa segue tre diversi obiettivi di tipo economico, competitivo e culturale:

- gestire le relazioni con i clienti;
- innovare i prodotti o servizi;
- gestire le infrastrutture.

Nella realtà è difficile seguirli tutti, quindi le compagnie solitamente seguono 1, o al massimo 2 di questi obiettivi (solo le compagnie grandi riescono a gestire e portare avanti tutti e 3 gli obiettivi). Per completare i tre obiettivi alcune imprese sono costrette a fare *outsourcing*.

2.2 Long Tail Business Model

É basato sulla vendita di molti prodotti di nicchia (ma non esclusivamente questi).



I ricavi derivanti dalla vendita dei prodotti di nicchia potrebbe essere superiore rispetto a quelli dovuti ai bestseller.

Questo tipo di business model ha bisogno di bassi costi di inventario ed una piattaforma solida (che abbia accesso veloce a tutti i prodotti).

Sui prodotti di nicchia, quindi difficili da trovare, il margine di ricavo è più alto (le persone sono disposte a spendere di più per avere quel bene o servizio).

Ragioni per cui è stata possibile la creazione di Long Tail Business Model:

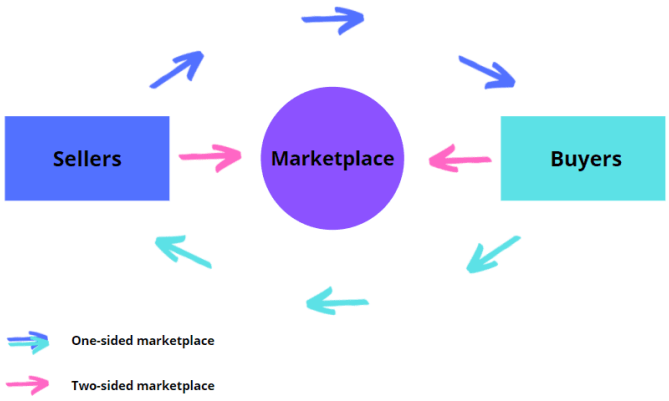
1. **Democratizzazione degli strumenti di produzione** (al girno d'oggi è molto più semplice produrre un prodotto o servizio, per creare una canzone ad esempio basta avere uno strumento ed una connessione ad internet)

2. **Democratizzazione della distribuzione** (ricollegandoci all'esempio di prima non abbiamo bisogno di uno studio di registrazione e produzione o di un'etichetta discografica ma possiamo distribuire la canzone su youtube).

3. **Riduzione dei costi per il collegamento di richieste e offerte** (grazie ad un algoritmo la tua canzone raggiungerà le persone interessate al tuo lavoro, anche se si trovano dall'altra parte del mondo).

2.3 Multi-Sided Platforms

Questo modello di business connette differenti customer segments che sono interdiendenti tra loro. Genera valore, semplificando le interazioni tra i segmenti. All'aumentare del numero di utenti aumeterà anche il valore del modello (*network effect*).



Il problema principale di questo modello è come attrarre i segmenti (offrendo degli sconti magari), quale segmento attrarre per primo e a che prezzo.

Alcuni esempi di multi-sided plattforms di successo sono: Uber, airbnb, Google play.

Un caso di insuccesso invece è la Playstation 3. I costumers della PS3 erano gli sviluppatori dei giochi e i gamers; questa console rendeva possibile la connessione tra questi segmenti. In questo caso la Sony decise di attrarre prima i gamer, abbassando i prezzi della console, però non avevano calcolato che i videogiocatori avrebbero potuto piratare i giochi e ciò si traduce in mancati

ricavi (non si è riuscito ad attrarre un numero sufficiente di utenti che acquistava giochi originali). Dalla PS4 hanno risolto questo problema aggiungendo un abbonamento per poter giocare online.

2.4 Free as a Business Model

É un modello in cui c'è almeno un segmento che usa il servizio gratuitamente.

2.4.1 Advertising-based

Questo è possibile perché c'è un altro segmento che paga per ottenere un servizio differente (spazio pubblicitario).

2.4.2 Freemium

Oppure perché esistono due versioni di un servizio o prodotto (free + premium). Ovviamente la versione base deve essere accattivante per attrarre nuovi utenti, ma non deve avere troppe caratteristiche altrimenti nessuno deciderà di passare alla versione premium (la quale dovrà aggiungere delle feature importanti). Solitamente la maggior parte degli utenti (90%) continua ad utilizzare una versione base. I pochi utenti che pagano il premium sostengono il business model. Questo è possibile grazie ai bassi costi marginali per utente.

2.4.3 Bait & Hook

Infine, esiste la tecnica esca ed uncino, in cui un'impresa regala qualcosa che per funzionare ha bisogno di altro. Un esempio potrebbe essere quella di operatori telefonici che regalano telefoni per chi effettua un passaggio di abbonamento (copriranno le spese del telefono con le spese mensili fisse).

2.5 Open Business Model

Fondata sulla collaborazione.

Può essere:

- *outside in* - la conoscenza viene da fuori nella nostra compagnia;
- *inside out* - La nostra conoscenza è fornita all'esterno.

L'idea alla base è che si possono guadagnare più soldi condividendo la conoscenza rispetto ad appropriarsi di essa.

3 Strategy

La *strategia* descrive il piano a lungo termine usato per definire e coordinare azioni utili al raggiungimento di un obiettivo.

La *strategia aziendale* consente di definire l'identità desiderata di un'impresa in uno specifico ambiente ed il piano che permetterà di raggiungerla.

3.1 Plan

1. Descrivere il business model (canvas) esistente;
2. Definire l'identità desiderata, cioè definire gli obiettivi;
3. Analizzare l'ambiente (interno ed esterno);
4. Analisi SWOT (Strenghts, Weaknesses, Opportunities, Threats) per ogni elemento del business model;
5. Identificare cosa inserire, cosa rimuovere, cosa aumentare e cosa diminuire;
6. Modificare il business model o creare un nuovo business model.
7. Scegliere tra merging o splitting, ovvero unire il nuovo business model con il precedente o meno.

3.1.1 Enviroment

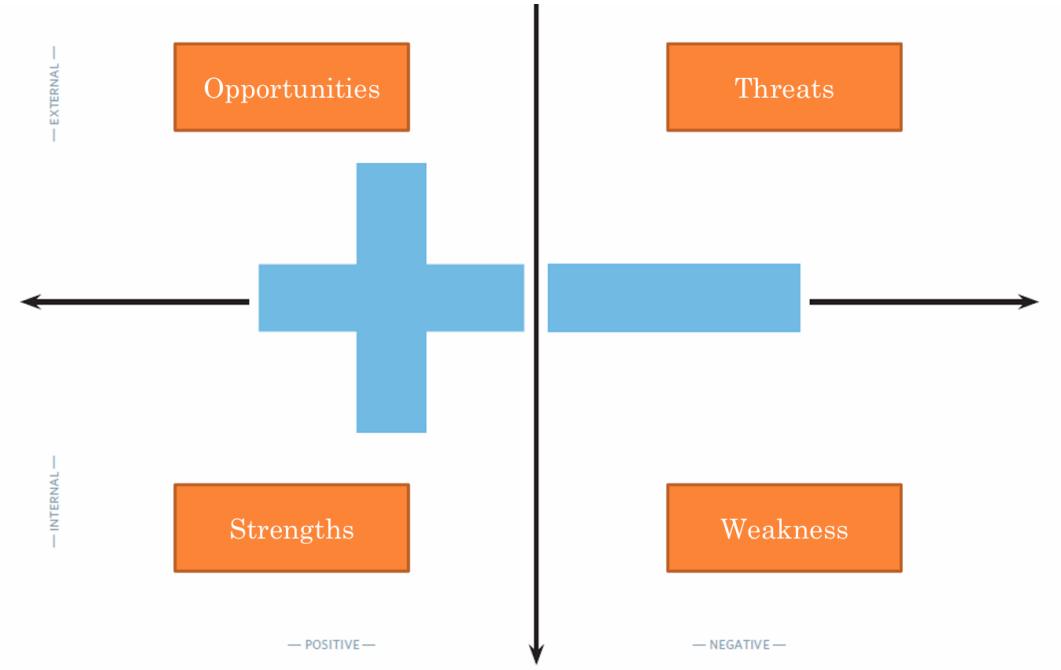
L'ambiente non deve limitare il business model, ma lo influenza. Dopo aver effettuato un'analisi dell'ambiente bisogna prendere consapevolezza delle informazioni ottenute e costruire un business model coerente con esse.

Ogni modello è progettato per essere eseguito in uno specifico *environment*. L'ambiente può cambiare frequentemente, quindi bisogna monitorarlo costantemente per capire cosa modificare per adattare il modello.

L'ambiente può essere definito da una serie di fattori:

- **market forces** - problemi del mercato, segmenti di mercato, bisogni e richieste, costi di spostamento da un'azienda alla sua competitor, attrattività in termini di profitto:
- **industry forces** - competitors e i loro punti forza, nuovi concorrenti, prodotti sostitutivi. fornitori, stakeholders (attori che possono influenzare il modello di business e l'impresa);
- **key trends** - tendenza tecnologica, tendenze normative, tendenze sociali e culturali, tendenza socioeconomica;
- **macro-economic forces** - condizioni del mercato globale, mercati capitali, materie prime ed altre risorse, infrastruttura economica.

3.1.2 SWOT Analysis

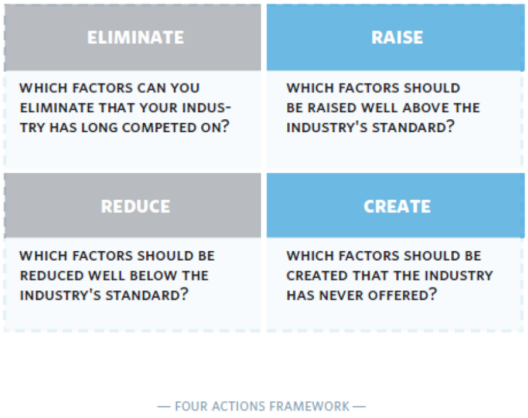


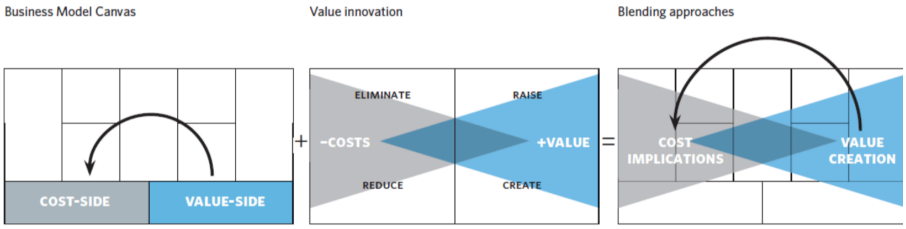
- **Internal**: identificare i punti di forza e debolezza.
- **External**: identificare le opportunità e le minacce.

3.1.3 Prospective

Dopo aver effettuato l'analisi SWOT abbiamo quindi una lista di caratteristiche che dovremmo incrementare, ridurre, eliminare o creare dal nostro business model. Ci sono tre approcci di modifica:

1. iniziare dai customer segments, quindi modificare il lato destro ed andare poi a controllare gli altri riquadri del canvas di conseguenza;
2. iniziare dalla value proposition e modificare il resto di conseguenza;
3. partire dai costi e successivamente aggiornare anche il resto del canvas.





3.1.4 Managing Multiple Business Models

Dopo aver creato il nuovo modello di business, la domanda da porci è: devo eliminare il vecchio modello? Conviene unire i due modelli in uno solo o è meglio separarli?

Quando un modello di business è adattabile a quello vecchio per quasi tutti gli elementi, la migliore soluzione è unirli.

Quando divido, ottengo un nuovo business quindi nuove risorse attività e partnership, è necessario farlo quando i due modelli sono diversi e non riesco ad usare delle risorse in comune.

4 Business Intelligence

Un *knowledge worker* è un lavoratore tenuto a prendere decisioni importanti e strategiche, con effetti di breve-lungo termine, all'interno di organizzazioni complesse. La maggior parte delle decisioni venivano fatte in maniera intuitiva, usando l'esperienza e le informazioni disponibili. Questo oggi non è possibile e vengono prese decisioni critiche, usando metodi analitici e modelli matematici.

L'obiettivo della business intelligence è quello di fornire ai knowledge workers gli strumenti e le metodologie che permettono loro di prendere decisioni effettive in modo rapido.

- **Data:** codifica strutturata di entità primarie singole, nonché di Transazioni che coinvolgono due o più entità primarie.
- **Information:** esito delle attività di estrazione e trattamento dei dati.
- **Knowledge:** trasformazione dell'informazione in conoscenza, usata per migliorare il processo decisionale.

4.1 Architecture

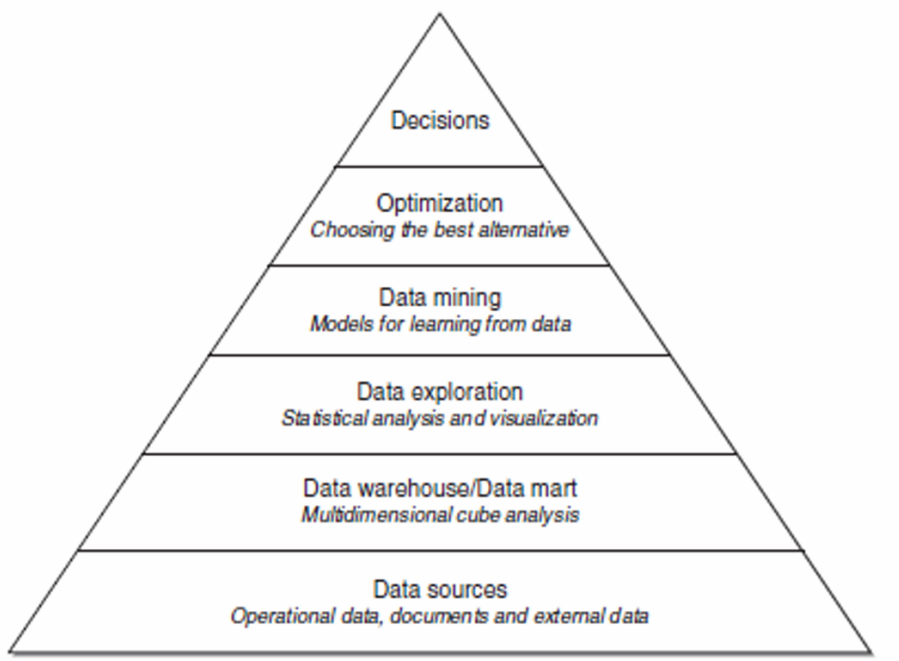


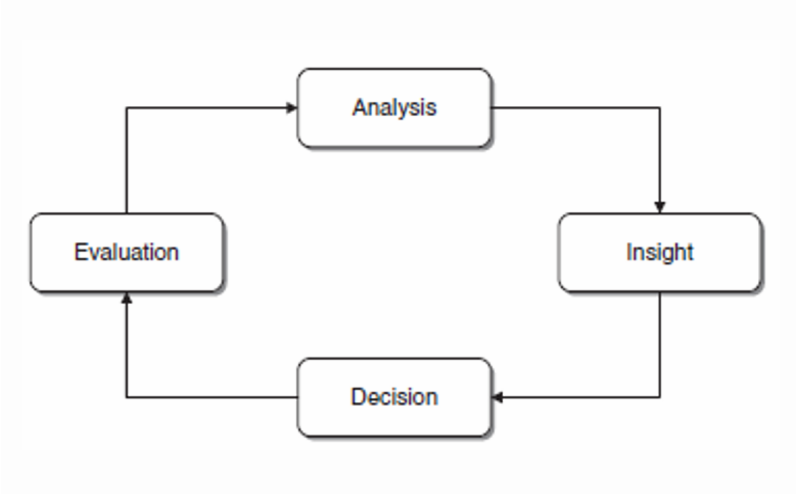
Figura 4.1: Business Intelligence Architecture.

- **Data Sources:** raccogliere ed unificare i dati provenienti da fonti eterogenee primarie e secondarie.
- **Data Warehouse and Mart:** i dati provenienti da fonti differenti sono conservati in database che supportano la BI attraverso strumenti di estrazione e trasformazione.

- **Data Exploration:** analisi passiva (query, reporting system, statistical methods). I decision makers generano delle ipotesi a priori ed usano strumenti per trovare risposte per confermare la tesi originaria.
- **Data Mining:** metodologie attive (estrazione delle informazioni e conoscenza dai dati). Non richiede la formulazione di ipotesi a priori, il suo scopo è espandere la conoscenza.
- **Optimization:** modelli che determinano la migliore soluzione, scegliendo tra un set di azioni alternative.
- **Decision:** effettuata dagli operatori, basando la decisione finale sui risultati delle analisi precedenti

Le metodologie di business intelligence possono essere trovate principalmente in tre dipartimenti: marketing and sales, logistics and production, accounting and control.

4.2 Cycle of a Business Intelligence Analysis



- *Analysis:* individuare il problema.
- *Insight:* maggior apprendimento del problema.
- *Decision:* prendiamo decisioni basandoci sulla conoscenza ottenuta dalle fasi precedenti. Queste diventano azioni.
- *Evaluation:* valutare le prestazioni.

5 Data Mining

Processo iterativo che si focalizza sull'analisi di grandi databases, con l'obiettivo di trarre informazioni e conoscenza che possano essere utili nel problem solving dei knowledge workers.

É basao sui *learning methods*, quindi il principale scopo è quello di trarre conclusioni partendo da un sample di vecchie osservazioni e generalizzare queste rispetto all'intera popolazione, facendo in modo che siano il più accurato possibile.

Le attività di data mining principali si dividono in due tipologie:

- *interpretation* - identificare dei patterns regolari nei dati ed esprimere questi attraverso regole e criteri che possano essere compresi facilmente dagli esperti del dominio applicativo. Le regole devono essere originali e non-trivial (non banali) per aumentare effettivamente il livello di conoscenza del sistema di interesse.

- *prediction* - anticipare il valore che una variabile assumerà nel futuro oppure stimare la likelihood degli eventi futuri.

Il Data Mining trova applicazione in svariati settori: Marketing Relazionale (identificare segmenti di clienti, predire risposte a campagne, analizzare comportamento d'acquisto, market basket analysis). Rilevamento Frodi (telefonia, assicurazioni, banche). Valutazione del Rischio (es. concedere prestiti). Text Mining (classificare documenti). Image Recognition (riconoscere caratteri, facce, comportamenti sospetti). Web Mining (analizzare clickstream). Diagnosi Medica (rilevare malattie dai risultati di test clinici).

Come dati di Input abbiamo nelle righe osservazioni passate e come colonne le informazioni disponibili ad ogni osservazione.

Gli attributi possono essere:

- *Categorici* se hanno un numero finito di valori distinti e rappresentano proprietà qualitative di un entità. Possono assumere, quindi, valori booleani/binari (1 o 0, vero o falso).

- *Numerici*, quando hanno un numero finito o infinito di valori che possono assumere. Queste possono essere discrete (valori finiti, controllabili) o continue (valori non controllabili).

Transaction ID	**Amount**	**Quantity**	**Payment by Credit Card**	**Store**	**Moment of the Day**
1	10	3	1	On-line	Morning
2	23	3	1	Store1	Morning
3	1	2	0	Store1	Evening
4	45	1	1	Store2	Afternoon
5	65.5	7	0	Store1	Afternoon
6	12.6	8	0	Store2	Morning

Nell'esempio in Tabella, ogni riga è una transazione ed ogni colonna una caratteristica correlata a quella determinata transazione. Degli attributi categorici possono assumere anche due o più valori e si dicono in questo caso *nominale* (i vari store); oppure *ordinale* se questi valori hanno un ordine fisso (momento della giornata).

Le attività di Data Mining si dividono in due tipi principali in base all'esistenza o meno di una variabile target.

- *Supervised Learning*: è presente un attributo target che rappresenta la classe a cui appartiene ogni record o una quantità misurabile. L'obiettivo è la predizione e l'interpretazione rispetto a questo attributo target. Esempi includono la lealtà dei clienti o il valore totale delle chiamate future. Le tecniche principali sono Classificazione (per target categorico) e Regressione (per target numerico continuo).

- *Unsupervised Learning*: non è guidato da un attributo target predefinito. L'obiettivo è scoprire pattern ricorrenti, affinità o raggruppamenti nel dataset. Esempi includono l'identificazione di cluster di clienti con comportamenti omogenei. Le tecniche principali sono Association Rules e Clustering.

Possiamo identificare sette tasks basici del data mining.

1. **Caratterizzazione e Discriminazione:** è di tipo supervised. Dividiamo i consumatori differenti tra loro, viene eseguita prima di costruire un modello di classificazione.

2. **Classificazione:** è di tipo supervised. Identifichiamo un utente per dire se è rilevante o meno nell'analisi.

3. **Regressione:** è di tipo supervised. Rende possibile la predizioni dei valori di un attributo. Come la classificazione ma fatto con variabili continue (nella classificazione erano categoriche).

4. **Time Series Analysis:** è di tipo supervised. Viene utilizzata quando l'attributo target si evolve nel tempo ed è associato a periodi adiacenti sull'asse temporale. L'obiettivo è prevedere il valore della variabile target per uno o più periodi futuri

5. **Association Rules:** è di tipo unsupervised. Mirano a identificare associazioni interessanti e ricorrenti tra gruppi di record in un dataset.

6. **Clustering:** è di tipo unsupervised. Segmenta una popolazione eterogenea in un dato numero di sottogruppi composti da osservazioni che condividono caratteristiche simili. Viene utilizzata per scoprire pattern ricorrenti e affinità nel dataset.

7. **Descrizione e Visualizzazione:** è di tipo unsupervised. Rappresentazioni grafiche, come istogrammi.

6 Data Preparation

La qualità dei dati in ingresso, spesso, è molto bassa. Quindi dobbiamo effettuare la preparazione dei dati prima di compiere qualsiasi analisi.

6.1 Data Cleaning

La qualità scarsa dei dati deriva da:

- *Incompletezza:* i dati hanno molti valori mancanti. Per gestire questo tipo di dai si effettua:
  - Eliminazione: scartare tutte le registrazioni per le quali mancano i valori di uno o più attributi. Questa opzione, però, può comportare una sostanziale perdita di informazione se vengono scartati molti record o se la distribuzione dei valori mancanti varia in modo irregolare tra gli attributi.
  - Sostituzione tramite ispezione: analisi manuale di ciascun valore mancante da parte di esperti del dominio applicativo.
  - Identificazione: utilizziamo un valore convenzionale per codificare ed identificare i valori mancanti. Ad esempio possiamo inserire -1 per indicare i valori mancanti nel caso di attributi che assumono solo valori positivi; così facendo riconosco subito i missing value.
  - Sostituzione automatica: tramite il software WEKA, nel momento in cui non posso eliminare i dati mancanti né sostituirli con valori standard, possiamo ottenere dei valori sostitutivi specifici. Per gli attributi numerici si usa la media delle osservazioni rimanenti. Per gl attribui categorici si usa il valore più frequente. Nelle analisi supervisionate si usa o la media o il valore più frequente, calcolato solo per i record con la stessa classe target. Un metodo più complesso e accurato è la stima del Maximum Likelihood Value, tramite modelli di regressione.
- *Noise:* valori errati o anomali (outlier). Dei prezzi ad esempio risultano troppo alti o bassi rispetto alla media. Può essere dovuto a malfunzionamenti dei dispositivi di misurazione e trasmissione. Possiamo identificarli tramite la dispersione dei dati (conoscendo la media e la deviazione standard individuiamo quali sono le osservazioni al di fuori del nostro intervallo);oppure tramite cluster analysis.
- *Inconsistenza:* discrepanze nei dati. Gli errori possono includere valori inaccettabili (una lettera in uno spazio numerico) o valori differenti con significato simile (spesso dovuti ad un errore, scriviamo BAII invece di BARI). Gli errori con valori inaccettabili portano alla cancellazione della righa, quelli con valori simili richiedono una correzzione.

6.2 Data Transfomation

- Discretizzazione: trasformare gli attributi continui in attributi categorici o ridurre il numero di valori possibili per attributi categorici. Questo riduce l'accuratezza ma aumenta la semplicità e l'intuitività del modello. Può essere bassata su soglie soggettive (l'esperienza), automatiche (algoritmi in base alla distribuzione o gerarchie (città-regione).
- Feature Extraction: generare nuovi attributi a partire da quelli esistenti.
- Standardizzazione: uniformare le scale di variazione dei diversi attributi per renderli confrontabili. Abbiamo diversi metodi:
  - *decimal scaling*, ovvero spostare la virgola. Matematicamente:

$$x'_{ij} = \frac{x_{ij}}{10^h}$$

  - *min-max*, in cui misurando i valori max e minimi iniziali ed inserendo dei valori max e minimi desiderati, possiamo cambiare ogni valore di un attributo in un range specifico. Matematicamente:

$$x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}(x'_{\max,j} - x'_{\min,j}) + x'_{\min,j}$$

6.3 Data Reduction

L'analisi di un numero eccessivo di istanze può richiedere molto tempo inutilmente, per questo motivo è utile ridurre il numero di attributi o istanze nel dataset.

- *Sampling:* selezionare un sottogruppo di osservazioni.
  - Semplice: estrazione casuale di una percentuale dei dati totali (adatto ad analisi non supervisionate).
  - Stratificato: estrazione dei dati in modo da preservare le percentuali o distribuzioni del dataset originale rispetto ad un attribuo categorico considerato critico (utile nelle analisi supervisionate, perché viene mantenuta la distribuzione della variabile target).
- *Feature selection:*
  - Qualitativa: eliminazione degli attributi non significativi per l'analisi, con un solo valore o con distribuzioni anomale (effettuata da esperti).
  - Quantitativa:
    - Filter Methods: Selezionano gli attributi più significativi, spesso i più correlati con la variabile target. Sono adatti per grandi dataset.

* Wrapper Methods: Trovano il sottoinsieme di variabili che porta alla massima precisione. Sono specifici per classificazioni o regressioni. Metodi comuni includono l'inclusione forward (bottom-up) e l'esclusione backward (top-down).
* PCA (Principal Component Analysis): Riduce il numero di attributi mantenendo la maggior parte delle informazioni. Identifica le variabili correlate (ridondanti) e si concentra su quelle con maggiore varianza. Il risultato sono componenti ortogonali ordinate per varianza spiegata.

7 Data Exploration

É descritta come un'attività di Business Intelligence passiva, in cui i decision maker generano ipotesi preventive e utilizzano strumenti come sistemi di query e reporting o metodi statistici per trovare risposte e confermare le loro intuizioni originali.

7.1 Univariate

L'analisi univariata studia un singolo attributo. L'obiettivo è descrivere l'attributo, disegnare conclusioni sul suo contenuto informativo ed evidenziare anomalie o outlier.

7.1.1 Analisi Grafica

Utilizzo di grafici per visualizzare la distribuzione di un attributo. Per gli attributi categorici, si usano grafici a barre verticali o istogrammi, dove l'asse verticale rappresenta le frequenze (numero di osservazioni per ciascun valore) e l'asse orizzontale i valori assunti dall'attributo. Per attributi continui, si possono definire delle classi.

7.1.2 Analisi di Tendenza

Misurazione di valori centrali che riassumono la distribuzione dell'attributo. la Media (sensibile agli outlier):

$$\bar{\mu} = \frac{x_1 + x_2 + \cdots + x_m}{m} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

la Media Ponderata (utilizzata per ridurre l'effetto degli outlier):

$$\bar{\mu} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_m x_m}{w_1 + w_2 + \cdots + w_m} = \frac{\sum_{i=1}^{m} w_i x_i}{\sum_{i=1}^{m} w_i}$$

la Mediana (il valore centrale delle osservazioni, meno influenzato dai valori estremi). É calcolata con due formule a seconda se m è:

- *dispari:*

$$x^{med} = x_{(m+1)/2}$$

- *pari:*

$$x^{med} = \frac{x_{m/2} + x_{(m+1)/2}}{2}$$

La Moda: il valore che si presenta con maggiore frequenza. Indica il picco più alto della distribuzione.

In una distribuzione simmetrica: media, mediana e moda tendono a coincidere; in una asimmetrica, invece, assumono valori diversi.

7.1.3 Analisi di Dispersione

Misurazione di quanto i valori dell'attributo sono distribuiti o concentrati. Include il Range (differenza tra valore massimo e minimo):

$$x^{range} = x^{\max} - x^{\min}$$

La Mean Absolute Deviation:

$$MAD = \frac{1}{m}\sum_{i=1}^{m}|s_i| = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{\mu}|$$

un MAD basso indica che i valori vicini alla media hanno dispersione bassa.
La Varianza (dà maggiore importanza agli errori e alle dispersioni più evidenti):

$$\bar{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m}s_i^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{\mu})^2$$

La Deviazione Standard:

$$\bar{\sigma} = \sqrt{\bar{\sigma}^2}$$

Per gli attributi categorici, si possono usare indici come l'Indice di Gini:

$$G = 1 - \sum_{h=1}^{H} f_h^2$$

L'Entropia:

$$E = -\sum_{h=1}^{H} f_h \log_2 f_h$$

che misurano l'eterogeneità o l'omogeneità di una distribuzione.

Vengono spesso normalizzati per avere un range da 0 (completamente omogenea, es. tutti gli utenti nella stessa città) a 1 (massima eterogeneità, es. utenti distribuiti uniformemente in diverse città).

Le formule normalizzate (nel range 0, 1) sono:

$$G_{rel} = \frac{G}{(H-1)/H} \qquad E_{rel} = \frac{E}{\log_2 H}$$

La varianza con distribuzione normale:

- Contiene circa il 68% dei valori osservati ($\bar{\mu} \pm \bar{\sigma}$)
- Contiene circa il 95% dei valori osservati ($\bar{\mu} \pm 2\bar{\sigma}$)
- Contiene circa il 100% dei valori osservati ($\bar{\mu} \pm 3\bar{\sigma}$)

La varianza con distribuzione arbitraria:

- *Teorema di Chebyshev:* dato un numero $\gamma \geq 1$ ed un gruppo di m valori, una proporzione di almeno $(1 - 1/\gamma^2)$ dei valori cadrà nell'intervallo $(\bar{\mu} \pm \gamma\bar{\sigma})$

Coefficiente di variazione (utilizzato per confrontare due o più gruppi di dati con diverse distribuzioni):

$$CV = 100\frac{\bar{\sigma}}{\bar{\mu}}$$

7.1.4 Analisi della Distribuzione di frequenza

Utilizzo di indici per capire la forma della distribuzione, come l'Indice di Asimmetria (Skewness):

$$I_{skew} = \frac{\bar{\mu}_3}{\bar{\sigma}_3}$$

dove $\bar{\mu}_3 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^3$. Indica se la distribuzione è normale ($I_{skew} = 0$), distorta verso destra ($I_{skew} > 0$) o sinistra ($I_{skew} < 0$).

L'Indice di Kurtosis, misura l'intensità della dispersione o l'appiattimento della distribuzione.

$$I_{kurt} = \frac{\bar{\mu}_4}{\bar{\sigma}_4} - 3$$

dove $\bar{\mu}_4 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^4$. Se ($I_{kurt} = 0$) la distribuzione è normale, quando ($I_{kurt} < 0$) è hypo-normale [maggiore dispersione] o ancora, se ($I_{kurt} > 0$) è hyper-normale [minore dispersione].

7.2 Bivariate

Studio dell'intensità della relazione tra una coppia di attributi. Si possono presentare tre casi: entrambi gli attributi numerici, uno numerico e uno categorico, o entrambi categorici.

7.2.1 Analisi Grafica

Utilizzo di rappresentazioni come lo Scatterplot (diagramma di dispersione) per visualizzare il legame tra due attributi numerici. Altri grafici possono essere usati a seconda del tipo di attributi (es. grafici a barre). È importante notare che l'analisi bivariata può mostrare la correlazione tra attributi, ma non stabilisce un legame di causalità.

7.2.2 Correlazione

Misurazione statistica dell'intensità e della direzione della relazione lineare tra due attributi numerici.

Covarianza:

$$v_{jk} = cov(a_j, a_k) = \frac{1}{m-2} \sum_{i=1}^{m} (x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$$

valori positivi indicano concordanza tra i due attributi (se uno aumenta, l'altro anche), mentre valori negativi indicano discordanza (se uno aumenta, l'altro tende a diminuire).

Coefficiente di Pearson:

$$r_{jk} = corr(a_j, a_k) = \frac{v_{jk}}{\bar{\sigma}_j \bar{\sigma}_k}$$

varia nel range [-1, +1].

- Se $r > 0$ c'è concordanza tra gli attributi. Le osservazioni mostrano una tendenza lineare orientata verso l'alto. Più il valore si avvicina ad 1 e più la relazione tra i due attributi approssima una retta crescente. Se $r = 1$ i punti si dispongono esattamente lungo una retta.
- Se $r < 0$ c'è una discordanza tra gli attributi. Le osservazioni mostrano una tendenza a disporsi lungo una retta orientata verso il basso. Più il valore si avvicina a -1 e più la relazione approssima una retta decrescente.
- Se $r = 0$ non si manifesta alcun legame di natura lineare tra gli attributi.

Tabelle di Contingenza: utilizzate per analizzare la relazione tra due attributi categorici.

7.3 Multivariate

Studio delle relazioni tra gruppi di attributi contemporaneamente. Spesso si ricorre a rappresentazioni grafiche, come una matrice di scatterplot che visualizza le relazioni a coppie per tutti gli attributi numerici.

8 Regression

La regressione è un metodo di data mining ampiamente utilizzato per i problemi di stima. Il suo obiettivo principale è identificare la relazione tra una variabile target e un sottoinsieme di altre variabili. Si basa sull'analisi di un dataset contenente osservazioni passate per le quali sono noti i valori degli attributi esplicativi (indipendenti) e il valore della variabile target.

É fondamentale notare che i modelli di regressione sono utilizzati quando la variabile target è continua e numerica. Se la variabile target è categorica, devono essere utilizzati modelli di classificazione.

8.1 Regressioe Lineare Semplice

Questo modello considera una sola variabile indipendente ($n = 1$).

La relazione tra la variabile dipendente (target Y) e la variabile indipendente (X) è espressa come:

$$Y = wX + b + \epsilon$$

dove w è il coefficiente angolare e b l'intercetta. Nella realtà è difficile avere una relazione lineare pura tra due variabili, per questo motivo è introdotto il termine di errore ϵ.

'errore è rappresentato dai residui, che sono i segmenti verticali che mostrano la discordanza tra i valori sulla retta e le osservazioni effettive. Un modello è accurato se l'ammontare dei residui è piccolo. L'errore deve essere casuale, risultato dell'azione di variabili indipendenti escluse con effetto trascurabile. Se l'errore non è casuale (ad esempio, aumenta all'aumentare di Y e X), potrebbe esserci una relazione con variabili non considerate.

I coefficienti w e b sono determinati minimizzando la Sum of Squared Errors:

$$SSE = \sum_{i=1}^{m} e_i^2 \sum_{i=1}^{m} [y_i - f(x_i)]^2 = \sum_{i=1}^{m} [y_i - wx_i - b]^2$$

8.2 Regressioe Lineare Multipla

Modello utilizzato quando il numero di variabili indipendenti è $n > 1$.

$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n + b + \epsilon$$

coefficienti angolari (w_j) nella regressione multipla rappresentano l'effetto marginale di una singola variabile esplicativa X_j sul target Y, assumendo che tutte le altre variabili rimangano invariate. Ci sono alcune limitazioni:

1. Il valore di ciascun coefficiente dipende dall'intero set di variabili esplicative; l'introduzione o rimozione di variabili cambia tutti i coefficienti.

2. La scala dei valori degli attributi predittivi influenza il valore del coefficiente corrispondente. È quindi utile procedere a una standardizzazione preliminare di tutte le variabili indipendenti.

Per gestire variabili categoriali nella regressione multipla, è necessario trasformarle in variabili dummy:

- 1, se l'istanza assume un valore specifico per quella categoria;
- 0, altrimenti.

Per evitare problemi di multicollinearità, si trasformano in dummy solo N-1 valori della variabile categoriale (dove N è il numero di valori possibili).

8.3 Valutazione dei Modelli di Regressione

Esistono vari criteri per valutare la qualità e l'accuratezza predittiva del modello.

8.3.1 Significance of the Coefficients

Un coefficiente non è significativo se il suo intervallo di confidenza contiene il valore 0. Per verificare la significatività si può controllare se il $t - value > |2|$ o se il $p - value < 0,05$.

L'intercetta del modello non deve necessariamente essere significativa.

8.3.2 Analysis of Variance

Si valuta se la varianza dei residui è inferiore alla varianza della variabile dipendente. Si valuta se $F - value > 1$ o $Pr < 0,0001$. Questo assicura che l'errore sia meno rilevante del fenomeno e che gli errori siano bassi e casuali.

8.3.3 Coefficient of Determination

Chiamato anche R^2, rappresenta la percentuale di varianza totale della variabile target spiegata dalle variabili predittive presenti nel modelli. Varia tra 0 e 1. Un R^2 elevato indica un buon modello (un valore più vicino a 1 indica che il modello spiega una porzione maggiore della varianza totale del fenomeno). Tuttavia, il valore di R^2 può aumentare semplicemente all'aumentare del numero di osservazioni o di variabili predittive, portando a una potenziale sovrastima. Per questo motivo, a volte si preferisce utilizzare l'*adjusted* R^2, che fornisce una descrizione più veritiera del fenomeno.

8.3.4 Multi-collinearity of the Independent Variables

Le variabili predittive non devono essere linearmente correlate tra loro. Se esiste una correlazione lineare significativa tra due o più variabili indipendenti, il modello presenta multicollinearità. Questa rende la stima dei coefficienti di regressione inaccurata e compromette la significatività complessiva del modello. Un segnale di multicollinearità può essere un elevato valore del coefficiente di determinazione (vicino a 1), ma con coefficienti di regressione dei predittori che non sono significativamente diversi da 0. La soluzione a questo problema è la selezione di un sottoinsieme di variabili non collineari oppure eliminare alcune variabili che sono collineari con altre.

8.4 Selezione delle Variabili Predittive

1. **Forward Inclusion:** partendo da un insieme vuoto di predittori, si aggiunge una variabile alla volta. In ogni iterazione, si sceglie la variabile esclusa che fornisce il maggiore aumento del potere predittivo (criteri: aumento dell'R^2 aggiustato o del valore della F-statistic), a condizione che l'incremento superi una soglia minima. È un processo iterativo e potenzialmente lungo.

2. **Backward Exclusion:** partendo includendo tutte le variabili predittive nel modello, si rimuove una variabile alla volta. In ogni iterazione, si seleziona per esclusione la variabile la cui rimozione causa il maggiore aumento del potere esplicativo del modello. I criteri sono analoghi a quelli dell'inclusione forward.

3. **Exhaustive Search:** analizza tutti i possibili sottoinsiemi di predittori per identificare il modello ottimale. È applicabile solo quando il numero di variabili esplicative è basso, a causa della crescita esponenziale del numero di combinazioni.

9 Classification

Consideriamo la classificazione per l'apprendimento supervisionato. A differenza dei modelli di regressione, che prevedono il valore di un attributo numerico, i modelli di classificazione sono utilizzati per prevedere il valore di un attributo target di tipo categorico.

L'obiettivo principale dei modelli di classificazione è, partendo da un insieme di osservazioni passate la cui classe target è nota, generare un insieme di regole che consentano di prevedere la classe target di esempi futuri. Questo implica l'identificazione di relazioni ricorrenti tra le variabili esplicative (attributi predittivi) che descrivono gli esempi appartenenti alla stessa classe. Queste relazioni vengono poi tradotte in regole di classificazione.

I dati di input per un modello di classificazione consistono in un dataset (D) contenente m osservazioni (anche dette esempi, casi, istanze o record) descritte da n attributi esplicativi (o predittivi, che abbiamo a disposizione nel nostro dataset), che possono essere sia categorici che numerici, ed un attributo target (o classe) che deve essere per forza categorico (avere un numero finito di valori).

La classificazione può essere:

- *Binaria:* l'attributo che voglio prevedere ha solo due valori.

- *Multiclass:* ci sono più di due classi per il target.

Il processo di classificazione tipicamente include le seguenti fasi:

1. *Training*: Un algoritmo di classificazione viene applicato a una porzione del dataset, chiamata training set, per derivare le regole che assegnano la classe target a ciascuna osservazione. Il training set è usato per

apprendere la relazione funzionale tra la variabile target e le variabili esplicative.

2. *Test*: Le regole generate durante il training vengono utilizzate per classificare osservazioni non incluse nel training set, che fanno parte del test set. La classe target per queste osservazioni è nota, consentendo di valutarne l'accuratezza confrontando la classe prevista dal classificatore con la classe reale di ciascuna istanza. L'uso di training e test set disgiunti è cruciale per evitare di sovrastimare l'accuratezza del modello.

3. *Prediction*: Le regole generate nel training vengono applicate alle variabili esplicative di nuove osservazioni (per cui la classe target non è nota) per assegnare loro una classe target prevista.

Alcuni algoritmi utilizzano anche un *tuning set* per identificare il valore ottimale di alcuni parametri. Molti classificatori generano anche una funzione score, un valore reale associato a ciascuna osservazione. Questo può essere interpretato, dopo la standardizzazione, come una stima della probabilità che la classe prevista dal classificatore sia corretta.

9.1 Valutazione dei modelli di classificazione

- **Accuratezza (Accuracy):** la percentuale di istanze classificate correttamente. È una misura chiave per confrontare modelli.
- **Velocità (Speed):** il tempo di calcolo, importante per gestire problemi di dimensioni maggiori.
- **Robustezza (Robustness):** la stabilità delle regole generate e dell'accuratezza al variare del training e test set.
- **Scalabilità (Scalability):** l'abilità di apprendere da grandi dataset, spesso usando tecniche di campionamento.
- **Interpretabilità (Interpretability):** la semplicità e comprensibilità delle regole generate per esperti del dominio.

9.2 Holdout Method

Divide il dataset in due sottoinsiemi disgiunti, training set (T) e test set (V), valutando l'accuratezza sul test set. La suddivisione è spesso casuale, con T tipicamente tra 1/2 e 2/3 delle osservazioni totali. L'accuratezza può dipendere dal test set selezionato.

9.2.1 Repeted Random Sampling

Possiamo replicare l'Holdout Method r volte, per ottenere una stima più robusta ed affidabile.

Per ogni ripetizione, viene estratto un campione casuale indipendente Tk per il training set, e l'accuratezza corrispondente viene valutata sul test set rimanente. Alla fine della procedura, l'accuratezza viene stimata utilizzando la media campionaria delle accuratezze ottenute.

Però, non si ha controllo sul numero di volte in cui ciascuna osservazione può apparire nel training set o nel test set. Osservazioni contenenti valori anomali (outliers) potrebbero causare effetti indesiderati sulle regole di classificazione generate e sulla stima dell'accuratezza.

9.3 Cross-Validation

É un altro metodo per addestrare e valutare un modello.

Si basa su una partizione del dataset in r sottoinsiemi disgiunti. La procedura richiede r iterazioni. Ad ogni iterazione, uno dei sottoinsiemi viene selezionato come test set, e l'unione di tutti gli altri sottoinsiemi viene utilizzata come training set. L'algoritmo di classificazione viene applicato r volte, utilizzando ciascuno dei r training set a turno e valutando l'accuratezza ogni volta sul corrispondente test set.

Ogni osservazione del dataset appare lo stesso numero di volte nei training set e esattamente una volta nei test set. Alla fine della procedura, l'accuratezza complessiva viene calcolata come la media aritmetica delle r accuratezze individuali.

Valori più alti di r sono preferiti per ottenere una stima più robusta dell'accuratezza. Una scelta popolare nella pratica è la *tenfold cross-validation*, in cui il dataset viene partizionato in 10 sottoinsiemi.

Una variante è il *Leave-One-Out*, in cui ciascuno dei test set include una sola osservazione. Questo richiede un maggiore sforzo computazionale ma l'operazione di training viene effettuata con un numero maggiore di osservazioni.

9.4 Confusion Matrices

La matrice di confusione è uno strumento per valutare l'accuratezza di un classificatore in problemi di classificazione binaria (dove la classe target può assumere due valori, ad esempio -1, 1 oppure 0, 1).

Possiamo considerarla (nel caso di classificazione binaria) come una matrice 2x2 dove le righe rappresentano i valori osservati (reali) e le colonne i valori previsti dal modello di classificazione. Gli elementi di questa matrice saranno:

- *True Negative* (p), numero di previsioni corrette per esempi negativi (valore osservato -1, valore previsto -1);
- *False Positive* (q), numero di previsioni errate per esempi negativi (valore osservato -1, valore previsto +1);
- *False Negative* (u), numero di previsioni errate per esempi positivi (valore osservato +1, valore previsto -1);

- *True Positive* (v), numero di previsioni corrette per esempi positivi (valore osservato $+1$, valore previsto $+1$).

Consente di analizzare gli errori commessi e il loro tipo. I valori sulla diagonale principale rappresentano le previsioni corrette, mentre quelli fuori dalla diagonale secondaria rappresentano gli errori.

Dalla confusion matrix si possono calcolare varie metriche per valutare la performance del classificatore.

Accuratezza:

$$Accuracy = \frac{p+v}{m}$$

True Negative rate:

$$TN = \frac{p}{p+q}$$

False Negative rate:

$$FN = \frac{u}{u+v}$$

False Positive Rate

$$FP = \frac{q}{p+q}$$

True Positive Rate (efficacia):

$$TP = \frac{v}{u+v} = Recall$$

Efficienza:

$$Precision = \frac{v}{q+v}$$

F-measure (media ponderata tra Precision e Recall):

$$F = \frac{(\beta^2 - 1) \cdot TP \cdot Precision}{\beta^2 \cdot Precision + TP}$$

È possibile assegnare una matrice di costi per gli esempi classificati erroneamente. Regolando adeguatamente il peso relativo dei costi di misclassificazione (False Negatives e False Positives), è possibile indirizzare un classificatore verso gli obiettivi effettivi dell'applicazione. Ad esempio, nella diagnosi medica, il costo dei False Negatives è molto maggiore del costo dei False Positives; nella previsione dell'abbandono (Churn), il costo dei False Positives è molto maggiore del costo dei False Negatives.

		predictions		
		-1 (negative)	$+1$ (positive)	total
examples	-1 (negative)	p	q	$p+q$
	$+1$ (positive)	u	v	$u+v$
	total	$p+u$	$q+v$	m

9.5 Curva ROC (Receiver Operating Characteristic)

Permettono di valutare visivamente l'accuratezza del classificatore e confrontare diversi modelli. Graficano il tasso di TP contro il tasso di FP al variare della soglia di classificazione. L'Area Under the Curve (AUC) è una misura concisa per confrontare i classificatori; un'area maggiore indica una performance migliore.

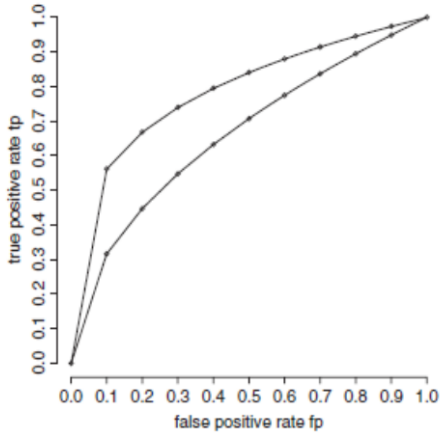


Figura 9.1: ROC curve.

Si cerca di selezionare il miglior trade-off, che aumenti i True Positives senza modificare eccessivamente i False Positives. Il punto migliore è quello che si avvicina maggiormente al punto (0,1) o all'asse Y. Punti Chiave:

- (0,1) rappresenta il classificatore ideale, senza errori di previsione → FP=0, TP=1.
- (0,0) corrisponde a un classificatore che prevede sempre la classe -1.
- (1,1) corrisponde a un classificatore che prevede sempre la classe +1.

9.6 Classificatio Trees

Gli alberi di classificazione sono popolari per la loro semplicità, facilità d'uso, velocità, robustezza e interpretabilità. Un albero di classificazione è composto da un nodo radice (Root Node), nodi intermedi e nodi foglia (Leaf Node). Ogni nodo, ad eccezione delle foglie, rappresenta una condizione o un controllo su un attributo. Lo sviluppo di un albero corrisponde alla fase di training.

Il processo di costruzione di un albero di classificazione segue una procedura ricorsiva (top-down induction of decision trees):

> Le osservazioni inizialmente nel nodo radice vengono suddivise in sottoinsiemi disgiunti che confluiscono nei nodi discendenti (branching). Ad ogni nodo, si verificano le condizioni per interrompere lo sviluppo (Stopping Criteria). Se le condizioni sono soddisfatte, il nodo diventa una foglia; altrimenti, si procede a un'ulteriore

suddivisione. Al termine, ogni nodo foglia viene etichettato con il valore della classe a cui appartiene la maggioranza delle osservazioni in quel nodo (majority voting). La suddivisione in ciascun nodo avviene tramite una regola di separazione (Splitting Rule).

Per assegnare la classe target a una nuova osservazione durante la fase di previsione, si segue un percorso dal nodo radice a un nodo foglia, applicando la sequenza di regole basate sui valori degli attributi della nuova osservazione. La classe target prevista coincide con la classe con cui è stata etichettata la foglia raggiunta durante la fase di sviluppo.

La maggior parte dei classificatori, inclusi gli alberi di classificazione, genera una *score function*. Negli alberi di classificazione, a ciascuna osservazione in una foglia è associata la proporzione più alta della classe target per le osservazioni contenute in quella foglia. Questo può essere interpretato come una stima della probabilità che la classe prevista sia corretta.

9.6.1 Splitting Rules

Criteri per identificare la regola ottimale per suddividere le osservazioni e creare i nodi discendenti. Gli alberi possono essere binari (con al massimo due rami per nodo) o multi-split trees (con un numero arbitrario di rami). La scelta dipende dal tipo di attributo (binario, categorico con più di due classi, numerico).

9.6.2 Stopping Criteria

Insieme di regole utilizzate in ogni nodo per determinare se continuare ricorsivamente lo sviluppo o considerare il nodo una foglia. Ci sono due ragioni principali per limitare la crescita: evitare l'overfitting (l'albero riflette eccessivamente le peculiarità del training set, riducendo la capacità di generalizzazione) e limitare la proliferazione di foglie e regole di classificazione profonde.

Esempi includono: la dimensione minima del nodo, la purezza (proporzione di osservazioni della stessa classe sopra una soglia) e il guadagno minimo da una possibile suddivisione.

9.6.3 Pruning Rules

Criteri applicati per evitare una crescita eccessiva dell'albero durante lo sviluppo (pre-pruning) o per ridurre il numero di nodi dopo che l'albero è stato generato (post-pruning).

Il pruning aiuta a ridurre le regole e l'eccessiva aderenza al training set. Il post-pruning valuta il vantaggio della rimozione di un ramo confrontando l'accuratezza predittiva.