# Rapporto Metodologico
## Analisi e Modellazione Dati CRM

## 1.0 Introduzione

---

Questo rapporto descrive in dettaglio la metodologia tecnica end-to-end che abbiamo utilizzato per analizzare i dati dei clienti, ingegnerizzare metriche di valore e sviluppare modelli di segmentazione e predittivi. L'approccio documentato parte dalla generazione di un dataset CRM realistico, procede con la creazione di feature significative e culmina nello sviluppo di modelli di machine learning per la segmentazione della clientela e la previsione della sostenibilità economica. Lo scopo di questo documento è fornire una documentazione chiara, trasparente e riproducibile per gli stakeholder tecnici, garantendo la piena comprensione dei passaggi analitici e dei modelli implementati.
Gli obiettivi principali di questo progetto di analisi sono stati i seguenti:

- **Generazione di un dataset CRM sintetico ma realistico:** Creare una base dati completa simulando anagrafiche, interazioni e transazioni dei clienti.
- **Ingegnerizzazione di feature e calcolo di metriche chiave:** Trasformare i dati grezzi in indicatori quantitativi come RFE (Recency, Frequency, Engagement), CLV (Customer Lifetime Value) e il rapporto LTV:CAC.
- **Sviluppo di un modello di segmentazione non supervisionato:** Utilizzare algoritmi di clustering per identificare gruppi di clienti omogenei sulla base dei loro comportamenti e del loro valore.
- **Creazione e validazione di un modello supervisionato:** Addestrare un classificatore per predire la sostenibilità di un cliente, definita in base alla sua redditività.
- **Analisi strategica dei segmenti e dei pattern di abbandono (churn):** Tradurre i risultati dei modelli in insight di business azionabili, profilando i segmenti e identificando i fattori associati al churn.

Il fondamento di ogni analisi affidabile risiede nella qualità dei dati di partenza. La sezione successiva illustra il processo di generazione e preparazione del dataset che ha costituito la base per tutte le elaborazioni successive.

## 2.0 Generazione e Preparazione del Dataset

La disponibilità di un dataset robusto e pulito è un prerequisito fondamentale per

---

qualsiasi iniziativa di data science. Per questo progetto, abbiamo generato un dataset

sintetico progettato per simulare un ambiente CRM complesso e realistico, includendo anagrafiche clienti, interazioni multicanale e cronologia degli acquisti.

---

### 2.1 Generazione Dati Sintetici

Il processo di generazione dei dati è stato orchestrato utilizzando la libreria Python Faker, che ha permesso di creare dati fittizi ma verosimili. Abbiamo generato tre DataFrame principali, le cui strutture e attributi chiave sono riassunti di seguito.

| DataFrame | Attributi Chiave Generati |
|---|---|
| `df_customers` | `CustomerID`, `FirstName`, `LastName`, `Email`, `Phone`, `Country`, `City`, `SignupDate`, `LeadStatus`, `Segment`, |
| `df_interactions` | `CustomerID`, `ContactDate`, `Channel`, `Notes`, `Clicks`, `TimeSpentSec`, `Replied` |
| `df_purch` | `CustomerID`, `PurchaseDate`, `Amount` |

Il volume di dati generati è stato dimensionato per consentire analisi statisticamente significative, risultando in **1000 clienti**, **7064 interazioni** e **4627 acquisti**.

---

### 2.2 Pulizia e Validazione Iniziale

A seguito della generazione, abbiamo eseguito operazioni preliminari di preparazione per garantire la coerenza dei dati. Il primo passo ha comportato la conversione delle colonne testuali contenenti date, come ContactDate e SignupDate, nel formato standard datetime, un passaggio essenziale per tutti i calcoli temporali successivi.

La gestione di anomalie più complesse, come importi di acquisto negativi e valori nulli, è stata posticipata e integrata nel flusso di lavoro dopo il calcolo delle metriche iniziali, come documentato nella sezione "Verifica consistenza e outlier". Questo approccio ha permesso di applicare le regole di pulizia su un dataset già arricchito di feature analitiche.

Una volta preparato il dataset di base, il passo successivo è stato quello di trasformare i dati grezzi in metriche quantitative, cuore dell'analisi comportamentale e di valore.

## 3.0 Ingegnerizzazione delle feature e calcolo delle metriche

---

Per comprendere a fondo il comportamento, il valore e il livello di coinvolgimento della clientela, abbiamo ingegnerizzato un set di metriche quantitative e interpretabili a partire dai dati grezzi transazionali e di interazione. Questa fase è cruciale per alimentare i successivi modelli di segmentazione e predittivi.

## 3.1 Calcolo delle Metriche RFE (Recency, Frequency, Engagement)

Il framework RFE è stato adattato per descrivere l'attività e il coinvolgimento del cliente basandosi sulle interazioni.

- **Recency:** Misura la freschezza della relazione. L'abbiamo calcolata come il numero di giorni trascorsi tra l'ultima interazione registrata (ContactDate) e una data di riferimento fissa (reference_date). Un valore basso indica un cliente recentemente attivo.
- **Frequency:** Quantifica l'intensità delle interazioni. L'abbiamo calcolata come il numero totale di interazioni registrate per ogni cliente.
- **Engagement:** Questa metrica composita è stata derivata per misurare la qualità dell'engagement, aggregando la somma del tempo totale speso nelle interazioni (TimeSpentSec) e il numero totale di volte in cui il cliente ha risposto (Replied).

## 3.2 Calcolo delle Metriche di Valore e Relazione (Monetary, Tenure)

Per arricchire il profilo del cliente con una dimensione economica e temporale, abbiamo calcolato le seguenti metriche.

- **Monetary:** Rappresenta il valore economico totale di un cliente, calcolato come la somma complessiva degli importi (Amount) di tutti gli acquisti effettuati.
- **Tenure:** Definisce la durata della relazione, calcolata come il numero di giorni intercorsi tra la data di iscrizione (SignupDate) e la reference_date dell'analisi.

## 3.3 Calcolo delle Metriche di Business Avanzate

Abbiamo sviluppato metriche più sofisticate per valutare la redditività e la sostenibilità di ogni cliente.

- **CLV_proxy:** Per ottenere una stima del valore del cliente, abbiamo calcolato una proxy basata sulla formula AvgPurchaseValue * Frequency. Tuttavia, poiché AvgPurchaseValue è definito come Monetary / Frequency, la formula si semplifica in (Monetary / Frequency) * Frequency, risultando matematicamente equivalente a Monetary. Sebbene questa proxy serva a stimare il valore basandosi sul comportamento storico, è funzionalmente equivalente alla spesa totale passata e non incorpora componenti predittive. Iterazioni future dovrebbero esplorare modelli CLV più avanzati, come quelli basati su distribuzioni probabilistiche (es. Beta-Geometric/Negative Binomial Distribution).
- **LTV:CAC Ratio:** Questo rapporto cruciale mette in relazione il valore del cliente con il costo sostenuto per acquisirlo (Acquisition_cost). È stato calcolato dividendo CLV_proxy per Acquisition_cost. Abbiamo utilizzato una soglia di

LTV:CAC >= 3 per definire un cliente come **"Sostenibile"**, identificando così le relazioni economicamente vantaggiose.

---

### 3.4 Sviluppo dell'Engagement Score Composito

Per sintetizzare le diverse dimensioni dell'engagement in un unico indicatore, abbiamo creato un punteggio composito normalizzato.

1. **Scalatura:** Le variabili Replies, TotalTimeSpent(min) e Frequency sono state scalate utilizzando un MinMaxScaler. Questa operazione ha normalizzato i loro valori in un range compreso tra 0 e 1, rendendoli confrontabili.
2. **Calcolo Ponderato:** L'EngagementScore è stato calcolato applicando una formula ponderata che attribuisce maggiore importanza alle risposte e al tempo speso: EngagementScore = 0.4 * Replies + 0.4 * TotalTimeSpent(min) + 0.2 * Frequency

Queste metriche arricchite hanno fornito una visione a 360 gradi del cliente, ponendo le basi per la successiva fase di segmentazione non supervisionata.

---

## 4.0 Segmentazione della Clientela tramite Clustering

L'obiettivo dell'analisi di clustering è stato quello di identificare gruppi di clienti naturalmente distinti e internamente omogenei sulla base delle metriche comportamentali e di valore calcolate. Questo approccio non supervisionato permette di scoprire strutture latenti nei dati senza la necessità di etichette predefinite.

---

### 4.1 Preparazione e Selezione del Modello di Clustering

Prima di applicare gli algoritmi, abbiamo eseguito un passaggio di pre-elaborazione fondamentale: le feature selezionate (Recency, Frequency, Monetary e Tenure) sono state standardizzate utilizzando StandardScaler per garantire che nessuna metrica dominasse le altre a causa della sua scala. Successivamente, abbiamo confrontato due algoritmi di clustering: KMeans e DBSCAN. La performance è stata valutata tramite il **Silhouette Score**.

| Algoritmo | Silhouette Score |
|---|---|
| DBSCAN | 508 |
| KMeans (k=5) | 0.243540 |

Nonostante il Silhouette Score superiore di DBSCAN, abbiamo selezionato in via preferenziale **KMeans**. Questa scelta è motivata da un requisito di business: per finalità di attivazione CRM, è necessario assegnare ogni singolo cliente a un segmento definito. Metodi partizionali come KMeans soddisfano questo requisito, mentre algoritmi basati

sulla densità come DBSCAN possono classificare i clienti meno clusterizzabili come rumore (outlier), escludendoli dall'analisi strategica.

Sebbene il Silhouette Score ottimale per KMeans sia stato ottenuto con k=5, abbiamo scelto di procedere con k=9. Questa decisione, pur rappresentando un compromesso rispetto alla metrica statistica, è stata presa per ottenere una maggiore granularità dei segmenti, consentendo una profilazione più dettagliata e azionabile a livello di business.

---

### 4.2 Definizione e Mappatura dei Segmenti

Il modello KMeans addestrato con k=9 ha assegnato a ogni cliente un'etichetta di cluster numerica. Per rendere questi cluster interpretabili, abbiamo creato una mappatura che associa a ogni ID un nome di segmento descrittivo.

| ID Cluster | Nome del Segmento |
|---|---|
| 0 | Champions |
| 1 | Loyal |
| 2 | Dormant |
| 3 | At Risk |
| 4 | New |
| 5 | Potential |
| 6 | Need Attention |
| 7 | Lost |
| 8 | Unclassified |

Questo processo ha trasformato l'output tecnico in una segmentazione strategica. Per un approccio complementare, abbiamo sviluppato anche un modello supervisionato per predire la sostenibilità dei clienti.

## 5.0 Modello Predittivo per la Sostenibilità del Cliente

Parallelamente alla segmentazione, abbiamo sviluppato un modello di apprendimento supervisionato con l'obiettivo specifico di classificare i clienti in base alla loro probabilità di essere economicamente sostenibili. Questo modello utilizza le metriche calcolate come predittori per identificare quali clienti generano un ritorno sull'investimento positivo.

---

### 5.1 Definizione di Feature e Variabile Target

La costruzione del modello ha richiesto una chiara definizione delle variabili di input (feature) e dell'output da predire (target).

- **Feature (X):** Le seguenti metriche sono state utilizzate come predittori per il modello:
    - Replies
    - Recency
    - Frequency
    - Monetary
    - AvgPurchaseValue
- **Variabile Target (y):** La variabile da predire è stata Sostenibile, derivata dalla soglia del rapporto LTV_CAC >= 3.

---

## 5.2 Confronto e selezione del modello

Per selezionare il classificatore più performante, abbiamo adottato un approccio rigoroso. Abbiamo utilizzato una Pipeline per concatenare la standardizzazione dei dati (StandardScaler) con l'algoritmo di classificazione. Sono stati confrontati tre modelli: **LogisticRegression**, **RandomForestClassifier** e **SVC**. Le performance sono state valutate tramite validazione incrociata (5-fold) utilizzando due metriche chiave: F1_macro_mean (per una valutazione bilanciata) e Recall_0_mean (la capacità del modello di identificare correttamente la classe negativa, ovvero tutti i clienti genuinamente 'Non sostenibile'), che è critica per minimizzare il rischio di investire in clienti non profittevoli.

| Modello | F1_macro_mean | Recall_0_mean |
|---|---|---|
| **RandomForest** | **994** | **1.00** |
| SVM | 725 | 0.94 |
| LogisticRegression | 696 | 0.94 |

Il RandomForestClassifier è stato selezionato come modello finale grazie alle sue performance nettamente superiori. Oltre ai punteggi eccellenti, la sua naturale resistenza all'overfitting e la capacità di gestire interazioni complesse tra le feature senza una definizione esplicita lo rendono una scelta robusta per questa tipologia di dati tabellari.

---

## 5.3 Salvataggio e Implementazione

Per la messa in produzione del modello, la fase di serializzazione è fondamentale. Tuttavia, il codice sorgente conteneva un errore critico: joblib.dump('RandomForest', model_path) salva la stringa di testo "RandomForest" invece dell'oggetto modello addestrato, rendendolo inutilizzabile. L'implementazione corretta per salvare la pipeline addestrata è: joblib.dump(models['RandomForest'].fit(X, y), model_path). Questo passaggio, se eseguito correttamente, salva il modello nel file modelli/random_forest_sostenibilita.pkl per un riutilizzo efficiente.

Abbiamo inoltre preparato una funzione di supporto, score_clients, che utilizza il modello salvato per assegnare non solo una classe predetta, ma anche una probabilità di sostenibilità e un'etichetta strategica ('Sostenibile', 'Non sostenibile', 'Borderline') a nuovi clienti.

# 6.0 Analisi Strategica dei Segmenti e del Churn

Questa sezione finale traduce i risultati tecnici della modellazione in insight di business azionabili. L'analisi si concentra sulla profilazione dettagliata dei segmenti di clientela e sull'esame dei pattern di abbandono (churn).

---

## 6.1 Profilazione e Analisi dei Segmenti

Per comprendere le caratteristiche distintive di ciascun segmento, abbiamo calcolato i KPI medi per ogni gruppo. La tabella seguente riassume le metriche chiave.

| Segmento | Recency_ | Frequency | Monetary | Tenure_ | EngagementSco | LTV_CAC | CustomerID |
|---|---|---|---|---|---|---|---|
| At Risk | 130.08 | 0.31 | 3932.20 | 1445.92 | 0.22 | 240.13 | 82 |
| Potential | 78.95 | 0.83 | 3757.98 | 1279.06 | 0.57 | 217.67 | 100 |
| Unclassifi | 876.66 | 0.13 | 2872.10 | 1071.25 | 0.11 | 169.01 | 102 |
| Lost | 68.74 | 0.75 | 1575.98 | 1310.50 | 0.50 | 80.14 | 101 |
| Loyal | 189.37 | 0.22 | 1373.56 | 1432.44 | 0.17 | 79.73 | 93 |
| Champio | 73.36 | 0.36 | 3903.06 | 441.54 | 0.24 | 69.45 | 100 |
| New | 437.51 | 0.21 | 2597.02 | 515.08 | 0.15 | 62.54 | 103 |
| Need | 94.05 | 0.30 | 1298.77 | 563.02 | 0.23 | 36.15 | 85 |
| Dormant | 60.91 | 0.85 | 1963.69 | 385.68 | 0.55 | 27.05 | 76 |

Dall'analisi emergono profili strategici. I **"Champions"** sono clienti ad altissimo valore (Monetary_mean di 3903.06) ma con una relazione ancora breve (Tenure_mean di 441.54) e bassa Recency (73.36). Rappresentano i nuovi clienti di maggior successo, sui quali focalizzare strategie di fidelizzazione. Al contrario, i **"Dormant"** sono molto attivi (Frequency_mean 0.85, EngagementScore_mean 0.55) ma non convertono questa interazione in valore economico (LTV_CAC_mean basso, 27.05). Sono clienti "impegnati ma non acquirenti", che richiedono iniziative di monetizzazione. I segmenti **"At Risk"** e **"Potential"** sono i più preziosi in termini di LTV:CAC, ma gli "At Risk" mostrano una Recency più alta, segnalando un rischio di abbandono imminente.

Sintetizzando, i segmenti possono essere raggruppati in categorie strategiche: **Alto Valore** (At Risk, Potential), **Clienti Core** (Champions, Loyal) e **Basso Valore/Rischio Churn** (Dormant, Lost, Unclassified), ciascuno richiedente un approccio di marketing differenziato.

### 6.2 Analisi del Churn

Per identificare i clienti a rischio, abbiamo definito una regola di churn. Un cliente è stato classificato come 'Churn' se una singola variabile booleana, costruita dal **OR logico** delle seguenti condizioni, risultava vera:

- L'ultima interazione risale a più di 365 giorni (Recency > 365).
- La durata della relazione è inferiore a una soglia minima (Tenure < tenure_thresh).
- Lo stato del lead è esplicitamente 'perso' (LeadStatus == 'perso').
- Il punteggio di engagement è inferiore a una soglia minima (EngagementScore < engagement_thresh).

L'analisi dei clienti in churn, raggruppati per canale di acquisizione, rivela i seguenti pattern.

| AcquisitionChannel | EngagementScore | Monetary | Tenure | Recency | Frequency |
|---|---|---|---|---|---|
| Ads | 0.3 | 2462.7 | 932.6 | 153.7 | 0.5 |
| Direct | 0.4 | 2504.8 | 944.4 | 153.6 | 0.5 |
| Referral | 0.4 | 2590.3 | 961.3 | 127.8 | 0.5 |
| SEO | 0.3 | 2428.1 | 906.7 | 165.9 | 0.5 |
| Social | 0.4 | 2384.3 | 893.9 | 128.1 | 0.5 |

I clienti in churn acquisiti tramite **SEO** e **Ads** mostrano un EngagementScore medio più basso (0.3). In particolare, il canale SEO è associato alla Recency media più alta (165.9), suggerendo che potrebbe attrarre clienti con una minore propensione all'engagement a lungo termine, richiedendo strategie di retention più aggressive post-acquisizione.

## 7.0 Conclusioni Metodologiche

Questo rapporto ha documentato un processo metodologico completo per l'analisi dei dati CRM. Partendo dalla generazione di un dataset sintetico, l'analisi è proseguita attraverso una rigorosa fase di ingegnerizzazione delle feature. Successivamente, abbiamo sviluppato e validato due approcci di modellazione complementari: un clustering non supervisionato per la segmentazione e un modello di classificazione supervisionato per predire la sostenibilità dei clienti. Infine, abbiamo tradotto i risultati in analisi strategiche, profilando i segmenti e identificando i pattern di churn.

Gli artefatti prodotti—il dataset arricchito RFM_CRM, il modello predittivo random_forest_sostenibilita.pkl e la segmentazione finale—costituiscono strumenti robusti e riutilizzabili per gli stakeholder tecnici. Essi forniscono una solida base analitica in grado di supportare future analisi, alimentare dashboard di monitoraggio e guidare decisioni di business data-driven, garantendo coerenza, trasparenza e riproducibilità.